When will we ever use this?

Making Decisions Using Advanced Mathematics

$$7x_1 + 7x_3 + 7x_4 + 7x_6 \leq 1470$$

$$5x_1 + 15x_2 + 4x_3 + 10x_4 + 7x_5 + 4x_6 \leq 2250$$

Volume II: Probability and Decision Making

$$x_i \geq 0, \text{ for each } i = 1, 2, \ldots, 6$$

WAYNE STATE UNIVERSITY

NC STATE UNIVERSITY

UNCCHARLOTTE

A partnership between North Carolina State University, Wayne State University and University of North Carolina Charlotte.

Project Principal Investigators

Dr. Kenneth Chelst
Professor, Department of Industrial & Manufacturing Engineering
Wayne State University

Dr. Thomas Edwards
Associate Professor, Teacher Education Division
Wayne State University

Dr. Karen Keene
Assistant Professor, Department of STEM Education
North Carolina State University

Dr. Karen Norwood
Associate Professor, Department of Math STEM Education
North Carolina State University

Dr. David Pugalee
Professor, Center for Science, Technology, Engineering, and Mathematics (STEM) Education
University of North Carolina - Charlotte

Dr. Robert Young
Professor, Edward P. Fitts Department of Industrial & Systems Engineering
North Carolina State University

Project Managers
Dr. Molly Purser
North Carolina State University

Mr. Jay Johnson
Wayne State University

Researchers
<u>North Carolina State University</u>
Ms. Amy Craig
Ms. Richelle Dietz
Ms. Krista Holstein
Mr. Bill Scott

<u>Wayne State University</u>
Mr. Saman Alaniazar
Mrs. Hatice Ucar
<u>University of North Carolina – Charlotte</u>
Ms. Sarah Johnson

MINDSET Project

Mathematics **IN**struction using **D**ecision **S**cience and **E**ngineering **T**ools

Finally, an answer to the question all mathematics teachers fear…

"How will we ever use this?"

Mathematics INstruction using Decision Science and Engineering Tools (MINDSET) is a NSF funded collaboration between mathematics educators, industry professionals, engineers, and mathematicians at NC State University, Wayne State University and UNC Charlotte. Partnering with high school teachers, the MINDSET project team is working to create, implement, and evaluate a new curriculum. The intent is to reinforce standard mathematics concepts using math-based decision-making tools from Operations Research (OR) and Industrial Engineering (IE) for a non-calculus fourth-year mathematics course. The curriculum is presented to high school students as a series of real-world problems with the purpose of making the underlying mathematics more relevant to them. One goal of MINDSET is to improve students' mathematical abilities and attitudes by building on skills learned in Algebra II in relatable problem contexts. A second goal is to improve their problem solving skills and, especially, to interpret the results of that problem solving activity.

The course is comprised of two sections, one section of deterministic content and the other probabilistic. Problem contexts covered in the deterministic curriculum include linear programming, the critical path method, facility location problems, transportation problems, and multi-criterion decision making. In the probabilistic curriculum, topics include probability distributions, decision trees, quality control, Program Evaluation Review Technique (PERT), Markov chains, and queuing theory. But…MINDSET is different. These techniques and tools are woven into real-world problem contexts, where the student starts with the contexts and proceeds to math skill development.

Acknowledgements

The following teachers have piloted the MINDSET materials and provided valuable feedback to improve the material. We are grateful for their support of the project.

WHAT IS OPERATIONS RESEARCH?

Operations research (OR) is a scientific way to analyze problems, make decisions, and improve processes. OR professionals try to provide a sound basis for decision-making. These decisions may focus on day-to-day operations that arise in a manufacturing plant. Or, they may involve long-range issues such as designing new environmental regulations or establishing minimum prison sentence guidelines.

Operations researchers attempt to understand and structure complex situations. They develop mathematical and computer models of a system of people, machines, and procedures. If you have ever played the game Sim City, you have manipulated a computer model.

Operations researchers often use numerical, algebraic, and statistical techniques. Then they manipulate their models to study the behavior of the system. They use this understanding to predict how the system will behave under different rules and policies. Then they work to improve system performance.

Unlike most disciplines, we can point to specific events that mark the birth of operations research. OR was born in the years just prior to World War II. The British anticipated an air war with Germany. In 1937, they began to test radar. By 1938, they were studying how to use the information radar provides to direct the operations of their fighter planes.

Until this time, the word experiment usually meant a scientist carrying out a controlled experiment in a laboratory. In contrast, this radar-fighter plane project used a multi-disciplinary team of scientists. They studied actual operating conditions in the field instead of in the laboratory. Then they designed experiments in the field of operations, and the new term "operations research" was born. Their goal was to understand the operation of the complete system of equipment, people, and environmental conditions (e.g. weather, nighttime). Then they tried to improve it. Their work was an important factor in winning the Battle of Britain. OR eventually spread to all of the military services. Several of the leaders of this effort eventually won Nobel Prizes in their original fields of study.

All branches of the US Armed Forces later formed similar groups of interdisciplinary scientists. These groups worked to protect naval convoys, search for enemy convoys, enhance anti-submarine warfare and improve the effectiveness of bombers. To do so, they collected data by directly observing operations. Then they built a mathematical model of the system. Next, they used the model to recommend improvements. Finally, they obtained feedback on the impact of the changes.

Today, every branch of the military has its own operations research group. These OR groups include both military and civilian personnel. They play a key role in long-term strategy and weapons development. They also direct the operation of actions such as Operation Desert Storm. In addition, the National Security Agency has its own Center for Operations Research.

In the 1950s, national professional organizations were formed. These organizations published research journals. Universities added OR departments. All of this raised operations research to the level of a profession. The use of OR expanded beyond the military to include other government organizations and private companies. The petroleum and chemical industries were early users of OR. They improved the performance of plants, developed natural resources and planned strategy. Today, OR plays important roles in industry and government such as in:

- airline industry – scheduling planes and crews, pricing tickets, taking reservations, and planning the size of the fleet;
- pharmaceutical industry – managing research and development and designing sales territories;
- delivery services – planning routes and developing pricing strategies;
- financial services – credit scoring, marketing, and internal operations;
- lumber industry – managing forests and cutting timber;
- local government – deploying emergency services; and
- policy studies and regulation – environmental pollution, air traffic safety, AIDS, and criminal justice policy.

As the field of OR grew, there was less stress on interdisciplinary teams. Today, the focus is on using mathematical models. These models can be deterministic or probabilistic. Mathematical programming, routing and network flow are examples of deterministic models. Examples of probabilistic models include queuing, simulation and decision trees. Mathematical modeling is the core of the OR curriculum. OR programs are found in either engineering or business schools. Most mathematics departments also offer introductory OR courses at the junior or senior undergraduate level.

Volume II: Probabilistic Decision Making

May 2011

Table of Contents

Section 11.0: Introduction

We all make decisions in our jobs, our communities and in our personal lives that involve significant uncertainty. Some examples are:

- How much technology should be invested in a plant when product demand is uncertain?
- How much car collision insurance should be purchased when you do not know whether or not you will have an accident?
- Should a company build a new power plant in a foreign country and, if yes, in which country?
- On a personal financial level, how much should you invest in a particular stock or mutual fund?
- Should millions be invested in a new drug that has proven effective in animal tests?
- How much time should you invest in studying when the questions on the exam are uncertain and payback from more study-time is hard to predict?
- What career should you pursue when the economy and job market are uncertain?

Specific stocks and the stock market as a whole demonstrate much uncertainty from day to day and year to year. Individuals experience different types of random accidents. There is uncertainty in the demand for power and the stability of developing countries. The link between animal drug trials and drug effectiveness in humans is far from perfect. Decision analysis is an operations research modeling tool used to select the best decision in the presence of uncertainty.

- What specific uncertainties do you face in the next day, week, or month?
- What about other members of your family or friends?
- What uncertainties will affect your planning over the next year?

The oil industry was one of the earliest users of decision analysis and continues to lead in its application. Pharmaceutical companies apply decision trees, which are an extension of probability trees, to make research and development decisions. Industrial giants such as DuPont, Xerox, and Kodak have used decision trees to plan new products and production capacity. The US Forest Service uses decision trees to plan controlled forest fires. The Decision Analysis Affinity Group (www.daag.org) is an organization that runs conferences at which corporate users of decision analysis share experiences. There is also the Decision Analysis Society which is an organization affiliated with INFORMS. They maintain a homepage at http://faculty.fuqua.duke.edu/daweb/.

The decision tree methodology involves accounting for every possible decision and outcome. The best alternative generally maximizes the expected value of profit or minimizes the expected value of cost. Sometimes the variable optimized is not financial. The goal might be minimize time or maximize the number of people attending an event. Modern software such as Precision Tree, an Excel add-on, helps with the analysis and offers graphical representations of the results. These enable a decision maker to explore the strengths and weaknesses of the alternatives.

As decision analysis developed, the leaders in the field recognized two critical psychological and practical issues that needed to be addressed in order to make the tool of greater practical value. First, the models required estimates of probabilities that were often not easy to obtain with a detailed analysis of data. Therefore, subject matter experts were interviewed in order to estimate the probabilities. Decision analysts, along with mathematical psychologists, became leaders in the effort to understand biases and misconceptions that individuals display when asked to make a forecast. They developed interview protocols in order to get expert opinion in a way that reduces the likely bias.

Second, the expected value does not capture the fact that people are often fearful of taking risks, especially large ones. This risk aversion is the foundation for all of the insurance industry and the huge market in extended warranties. Decision analysts became leaders in researching attitudes towards risk and designing a methodology called utility theory that captures this behavior.

Section 11.1: Planning a Wedding

Chris and Mary Ann are planning their own wedding. They have hired a wedding planner to help them make the arrangements. Chris and Mary Ann have determined that they can afford to have no more than 240 people attend their wedding reception. They have decided that their main goal is to maximize the number of people attending their reception without going over their limit of 240.

As is usually done, Chris and Mary Ann will mail out wedding invitations requesting an RSVP indicating the invitee's acceptance of the invitation or not. Their wedding planner, Amy Smith, tells them that based on her experience, the probability that one-half of the invitees will accept the invitation is 0.2, the probability that two-thirds will accept is 0.5, and the probability that three-fourths will accept is 0.3. Ms. Smith says that although these probabilities are estimates, they reflect what she has seen in her 25 years as a wedding planner.

Furthermore, Ms. Smith knows from experience that there will always be "no shows", people who accept the invitation, but for some reason do not attend. Based on her experience, she estimates that the probability of a 5% rate of no shows is 0.6 and the probability of a 10% no show rate is 0.4. She again explains to Mary Ann and Chris that these probabilities are estimates based on her many years of experience.

11.1.1 How Many People Should We Invite?

Chris and Mary Ann decide to explore what might happen if they invite 240 people. They realize that Ms. Smith's estimates lead to multiple possibilities, and that each possible outcome is composed of two parts. They decide to diagram the situation to help them better understand it. Their diagram appears in Figure 11.1.1.
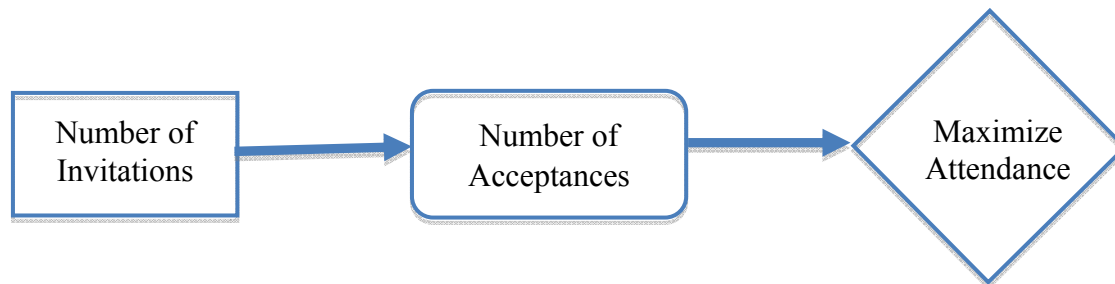
Figure 11.1.1: Chris and Mary Ann diagram their wedding scenario.

It seems to Chris and Mary Ann that something is missing from their diagram. Then Ms. Smith points out that her estimates of the percentages of acceptances and no shows do not appear in the diagram.

Q1.   Which box or boxes in the diagram are affected by the percentage of people invited who accept the invitation?

Q2.   Which box or boxes in the diagram are affected by the percentage of no shows?

After discussing the situation with Ms. Smith, Chris and Mary Ann revise their diagram by adding two uncertain events. Their new diagram is shown in Figure 11.1.2.
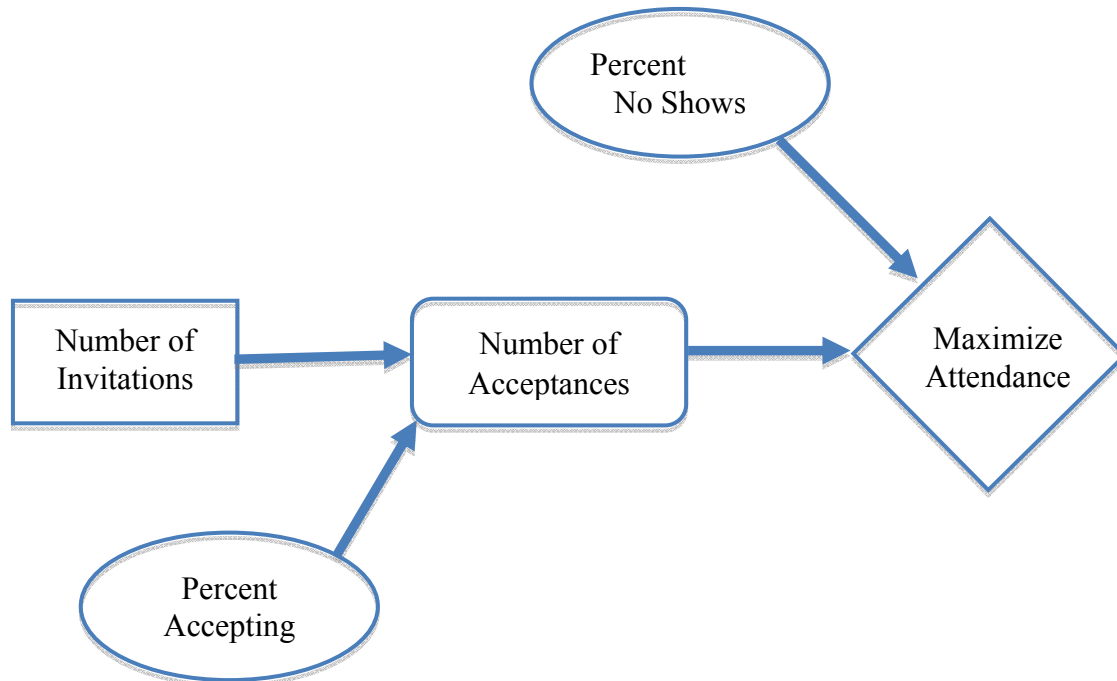
Figure 11.1.2: The wedding scenario showing rates of acceptance and no shows

After analyzing their revised diagram, Chris and Mary Ann can see two factors that will affect the number of people they invite who attend their wedding reception: the per cent who accept their invitation and the per cent of no shows. Next, they try to determine how many possibilities there would be if they used Ms. Smith's estimates, but are unable to agree.

Mary Ann says that she thinks there are six, but Chris is skeptical. Mary Ann explains her thinking this way:

> First, Amy Smith gave us three possible rates of acceptance. Then she gave us two possible percentages of no shows. The two no show percentages would apply to each of the three acceptance rates. So that would be six possibilities. For example, the acceptance rate might be one-half, and the no show percentage might be 5%.

Then Ms. Smith offers to draw a diagram of Mary Ann's idea. (See Figure 11.1.3.)
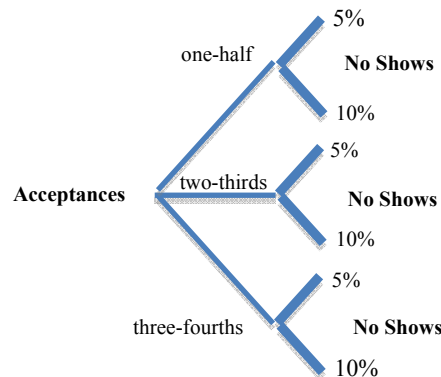
Figure 11.1.3: Ms. Smith's diagram showing six possibilities

Q3.   Identify the portion of the diagram in Figure 11.3 that represents an acceptance rate of one-half and 5% no shows.

Q4.   List the other five possibilities, based on Amy Smith's estimates.

Mary Ann applied the **Fundamental Principle of Counting** to determine how many possibilities there are in this scenario. This principle states that if there are 3 possible outcomes for one event (e.g., the percentage of acceptances) and 2 possible outcomes for a second event (e.g., the rate of no shows), then there are $3 \cdot 2 = 6$ ways in which both events might occur. More generally,

> ***The Fundamental Principle of Counting:*** If there are m possible outcomes for one event and n possible outcomes for a second event, then there are mn possible ways in which both events can occur.

The diagram that Amy Smith drew to clarify Mary Ann's thinking is called a **tree diagram**. A tree diagram is useful for identifying all of the possibilities when several things occur. It is also a useful tool when determining the probability of a **compound event**; that is, an event that is the result of more than one outcome. For example, the event that the acceptance rate will be one-half *and* the percentage of no shows will be 5% is a compound event, because it consists of two separate outcomes.

Once Chris and Mary Ann agree on the number of outcomes, they begin to wonder how many people are likely to attend. This time Chris takes the lead. He reasons that if one-half of the people who are invited accept the invitation, that's 120 people. Then if 5% do not show up, 5% of 120 is 6 no shows. Finally, if there are 6 no shows out of the 120 people who accepted the invitation, then 114 people will actually attend.

Q5.   Do you think that Chris's reasoning is sound? Why/why not?

Q6.   Complete a table similar to the one below to estimate the number of people who attend Chris and Mary Ann's wedding reception for each of the 6 possible outcomes that they identified based on Amy Smith's estimates.

Acceptance rate	Percentage of no shows	Number of people likely to attend
One-half	5%	(240)(0.5)(0.95)=114
r	p	

Table 11.1.1: Estimating Attendance for Six Possible Outcomes.

Now Chris and Mary Ann decide that they need to know the average number of people who will attend. They have calculated a number for each of 6 possible outcomes, but now they wonder if they can just average those six numbers. Chris says that he thinks they can, *if* the 6 possible outcomes are equally likely to occur.

Q7.   Do you think that Chris is right about averaging the six numbers if they are equally likely? Why/why not?

Q8.   Do you think that the 6 possible outcomes are equally likely?

Mary Ann then says that she thinks she knows how they can determine the probability of each of the 6 possible outcomes. She argues that if the probability that only one-half of the people invited is 0.2, that's the same as $\frac{2}{10}$. Then, if the probability that there will be 5% no shows is 0.6, that's like $\frac{6}{10}$. Finally, she says that they need to find $\frac{6}{10}$ of $\frac{2}{10}$, and that would be $\frac{12}{100}$.

Q9.   How did Mary Ann come up with $\frac{12}{100}$?

Q10.   Do you think Mary Ann's argument is correct?

Now Chris and Mary Ann wonder what a probability of $\frac{12}{100}$ means for their wedding plans. Amy Smith, their wedding planner, says it means that out of every 100 weddings that she has planned, on average there were 12 times when the acceptance rate was one-half and the percentage of no shows was 5%. So, in terms of Mary Ann and Chris's wedding, there are 12 chances in every hundred that 114 of the 240 invitees will actually attend. Then Amy adds her probability estimates to the tree diagram she drew earlier. She tells Mary Ann and Chris that the revised diagram will help them to compute the probabilities of the other five possibilities.
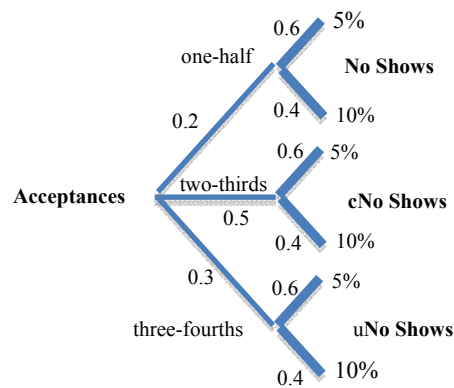
Figure 11.1.4: Amy Smith's tree diagram with her probability estimates

Q11.   Complete a table similar to the one below by finding the probability for each of the 6 possible outcomes.

Acceptance rate	% of no shows	Likely # attending	Probability
One-half	5%	114	0.12
One-half	10%		
Two-thirds			
Two-thirds			
Three-fourths			
Three-fourths			

Table 11.1.2: Finding the Probability of Each of the Six Possible Outcomes

Q12.   Which of the 6 possible outcomes is the most likely to occur? How many people would attend if that possibility occurred?

Q13.   Which of the 6 possible outcomes is the least likely to occur? How many people would attend if that possibility occurred?

Q14.   What is the sum of the six probabilities in your table? Why does that make sense (or not)?

Q15.   What is the sum of the two probabilities for which the acceptance rate was one-half? … two-thirds? … three-fourths? Why do these sums make sense (or not)?

Q16.   Do results similar to those in question 15 occur when you consider the three probabilities for which the percentages of no shows were 5% and 10%?

Q17.   Using the results in your table, what is the probability that the likely number of people attending will be less than 110? … greater than or equal to 110?

Q18.   Using the results in your table, what is the probability that the likely number of people attending will be less than 120? … greater than or equal to 120?

Q19.   Using the results in your table, what is the probability that the likely number of people attending will be less than 150? … greater than or equal to 150?

When you answer these questions, you may find it helpful to sort the possible outcomes from the lowest to highest number of attendees. This has been done in Table 11.1.3

Acceptance rate	% of no shows	Likely # attending	Probability	Cumulative Probability
One-half	10%			
One-half	5%	114	0.12	.20
Two-thirds				
Two-thirds				
Three-fourths				
Three-fourths				

Table 11.1.3: Ordering the outcomes by the number attending

The method that Mary Ann used to find the probability that one-half of the invitees will accept the invitation and that there will be 5% no shows is called the **Multiplication Rule for Compound Events.** In order to apply the *multiplication rule* to a *compound* event, all of the separate outcomes that make up the compound event must be **independent**. Independent events are events for which the probabilities of the outcomes do not affect on one another. For example, consider the following.

Every year medicine makes progress in understanding the role of genetics in medical disorders. It has long been known that some genetic defects are independent of the gender of the individual and others are linked to the gender of the individual. For example, cystic fibrosis occurs equally in males and females. In contrast, color blindness is much more common amongst males than females. In the United States, an estimated 7% of the males have color blindness and only 0.4% of the females. If a name is selected at random, the event that the individual is a male and the event that the individual has cystic fibrosis are independent. In contrast, if a name is selected at random, the event that the individual is a male and the event that the individual is color-blind are not independent. In other words, once you identify the individual as male, your estimate of the probability of color blindness increases significantly.

As a second example of events that are independent or not, suppose that, on the morning of a big test in your mathematics class, you were involved in an automobile accident on your way to school. No one was injured, but there was some damage to your car.

Q20.   Would the event of being involved in an accident on the way to school affect *your* performance on the test? In other words, would *you* consider these two events independent, or not?

Q21.   Do you think that *everyone* would answer the previous question the same way?

Q22.   Can you think of someone whose performance on the test would be affected by having been in an accident on the way to school?

Q23.   Can you think of someone whose performance would *not* be affected?

Q24.   Give an example of a pair of events that might for some people be independent, but for others not.

11.1.2 Should We Invite 240 People?

Mary Ann and Chris return to the question of how many people they should expect to attend if they invite 240. They have identified 6 possible outcomes, they know how many people would attend for each outcome, but they have learned that the 6 outcomes are not equally likely. They are sure that they cannot just average the 6 possible numbers of people attending, because they are not equally likely to occur. But they are not sure what to do. Once again, they consult Amy Smith. She suggests that, because the 6 possibilities have different probabilities, they could use a weighted average. That is, each of the 6 possible numbers of attendees could be weighted by its probability. Amy applied her weighted average idea like this:

$$\text{Average number attending} = 114(0.12)+108(0.08)+152(0.30)+144(0.20)+171(0.18)+162(0.12)$$
$$= 13.68+8.64+45.6+28.8+30.78+19.44$$
$$= 146.94$$

The weighted average that Amy Smith calculated is called the **expected value** of a **random variable**. The *random variable* in this case is the number of people who attend. This number is a variable, because it can have more than one value. It is a *random* variable, because its value depends on random events. In this case, the random events are the acceptance rate and the percentage of no shows. Finally, the *expected value* of the number of people who attend is 146.94. Notice that this expected value is not an integer.

Q25.   Why is it acceptable to have a non-integer for the expected value of the number of people who attend?

Q26.   If we rounded 146.94 to 147, based on the data in the preceding tables, could we actually observe 147 people attending the wedding based on Amy Smith's estimates

Q27.   If 146.94 is the expected value of the number attending out of 240 people who are invited, what is the expected value of the *percentage* of invited guests who will attend?

It is important to remember that the expected value of a random variable is really a weighted average. Each of the possible values of the random variable is weighted by the probability that it occurs. Because probability can be thought of as a percentage, there is no need to divide the sum of the weighted values, because division is built into the probabilities.

11.1.3 Should We Invite More Than 240 People?

Chris and Mary Ann decide to explore what might happen if they invite more guests than their maximum. Amy Smith, their wedding planner, suggests that they consider the expected number attending for several different numbers of invitees. They decide to consider inviting 270, 300, and 330 guests. Table 11.3 shows the likely number attending and its probability for each of the six possible outcomes if 270 guests are invited.

Q28.   What is the probability that the likely number of people attending is less that 175 if they invite 270 guests?

Q29.   Complete a table similar to Table 11.1.4 if 300 or 330 guests are invited. Then use the likely number of attendees and their respective probabilities to find the expected value of the number of attendees for each of the three cases.

Acceptance rate	% of no shows	Likely # attending out of 270[*]	Probability
One-half	10%	122	0.08
One-half	5%	129	0.12
Two-thirds	10%	162	0.20
Two-thirds	5%	171	0.30
Three-fourths	10%	183	0.12
Three-fourths	5%	193	0.18

* Fractional numbers have been rounded up to the next integer in each case.
Table 11.1.4: Likely number attending and probability if 270 are invited

Q30.   If they decide to invite 300 guests, what is the probability that the likely number attending is less than 200? How does this probability change if they decide to invite 330?

Recall the goal that Mary Ann and Chris have: To maximize attendance at their wedding without exceeding 240 attendees. They have now considered inviting 240, 270, 300, and 330 guests.

Q31.   Does the expected value of the number attending exceed 240 for any of these cases?

Q32.   If the expected number attending is less than 240, does that mean that the actual number attending will not exceed 240? Explain.

Q33.   Is there any case for which Chris and Mary Ann can be sure that the number attending will not exceed 240? Explain.

Q34.   How many guests do you think that Mary Ann and Chris should invite? Explain.

Section 11.2: Investment in Automation

Boss Controls (BC) is an automotive supplier that manufactures integrated cup holders with temperature controls for car-makers throughout the world. Their new model is to be made available on one million new luxury cars worldwide. Initial estimates that car buyers will select this option are as high as 50% of the time or as low as 30% of the time. Because of a general decline in the global economy, BC marketing estimates there is a slightly higher chance that demand will be 30% rather than 50%. Management assigns a 0.55 probability that only 30% of the luxury car buyers will request their temperature controlled cup holders. Since the only other possibility they are considering is that 50% of luxury car buyers will request this option, a 0.45 probability is assigned to that possibility.

Q1.   Why did they have to assign a 0.45 probability to the second possibility?

11.2.1 The Problem

Like most companies, BC wants to maximize its profit. BC's revenue comes from selling their cup holders to car companies at a fixed price of $60 apiece. The company can control manufacturing costs by how heavily it invests in automation. An investment of $13 million in high-speed equipment would lower the cost of producing each cup holder to $12. With a more modest investment of $8 million, the cost would be $27 per cup holder. Therefore, the net profit per unit after a high or low investment in automation will be $48 or $33, respectively. This per unit profit does not yet include the investment cost.

Q2.   What decision must be made in this situation?

Q3.   What is the **chance event** that BC faces? (i.e., an event whose likelihood must be predicted using probability and is outside the control of BC's management)?

11.2.2 Building a Decision Tree to Solve the BC Problem

In situations such as this, managers need to take a systematic approach to determine the best strategy for achieving their goal. We will build a **decision tree** similar to the probability tree in the last section to model the situation and show all the possible outcomes, their probabilities and their consequences.

Decision trees are made of *nodes* and *arcs*. The nodes on a decision tree represent decisions or random events. The arcs connect these nodes, showing paths that represent possible scenarios, depending on the outcomes of decisions and random events. To distinguish between the two types of nodes, decision and random chance, they are represented differently in a decision tree. Decision nodes are represented as rectangles and random chance nodes as circles, as in Figure 11.2.1.
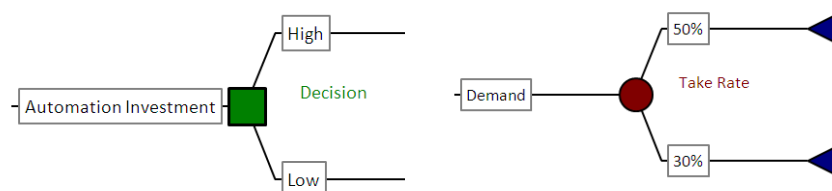

Figure 11.2.1: Example of decision node and chance node

In the BC scenario, whether there will be a high or low level of investment is a *decision* that management must make. However, whether the demand for their product will be 50% or 30% is strictly a random

event, and BC's management has no control over it. The two elements, decision and random chance, can be combined into a decision tree that models the entire Boss Control scenario as shown in Figure 11.2.2.
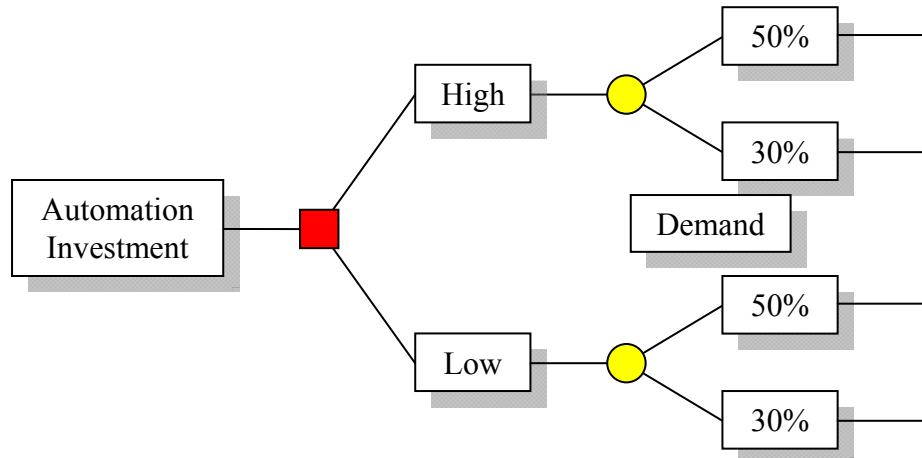
Figure 11.2.2: Decision tree for BC automation investment

This decision tree shows all the possible outcomes for every combination of decisions and random events. Now we must add the appropriate data to determine the expected profit of each path through the decision tree.

Recall that the low investment in automation was going to cost BC $8 million and the high investment was going to cost $13 million. This information can be added to the decision tree by placing the numbers along the appropriate branches, below the arcs. (The space above the arcs is reserved for another quantity, as you will see shortly). In this problem, the units on all such numbers will be millions of dollars. The numbers below are negative because they represent expenses BC must pay, not revenue earned.

Recall also that the unit profit after a low investment in automation is $33 whereas the unit profit after a high investment is $48. The "30%" and "50%" after the chance nodes represent the proportion of car buyers who select the luxury cup holder option. This is referred to as the *take rate*. In BC's case, the estimated low take rate is 30% and the estimated high take rate is 50%.

Q4.   What is the estimated probability that the take rate will be 30%?

Q5.   What is the estimated probability that the take rate will be 50%?

Q6.   What is the base of these take rates? In other words, what is the total number of potential sales?

The net revenue for each case can be calculated by multiplying the projected number of buyers times the net profit per unit. The projected net revenue with a 50% take rate and *high* investment is
        Net Revenue = 1,000,000*.5*$48 = $24 million

Q7.   What is the projected net revenue with a 30% take rate and *high* investment?

Q8.   What is the projected net revenue with a 50% take rate and *low* investment?

Q9.   What is the projected net revenue with a 30% take rate and *low* investment?

In Figure 11.2.3, the decision tree has been updated to include the information above. As before, the numbers beneath the arcs are in millions of dollars. They represent cost or net revenue. The numbers assigned to arcs coming from a random event show the probability of that path.
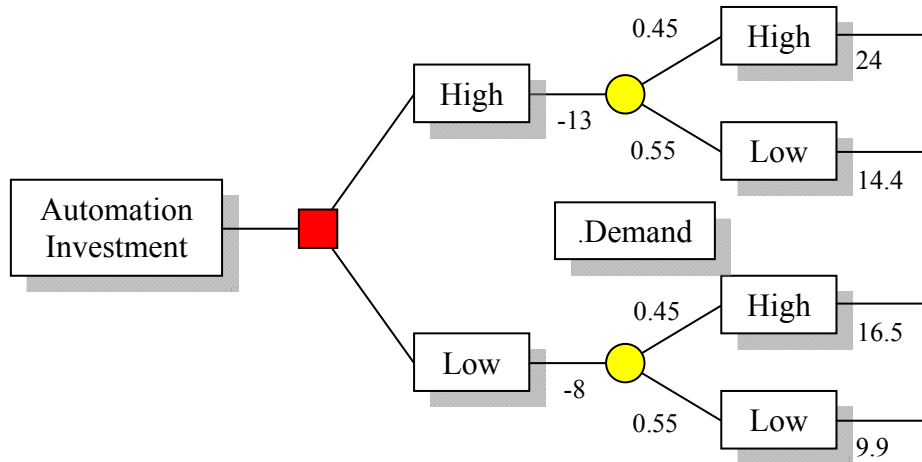
Figure 11.2.3: Decision tree showing dollar values and probabilities

An end node, represented by a triangle, is placed at the end of each path through the decision tree. Next to the triangle, we record the net profit for that path. The net profit is determined by subtracting the investment cost from the net revenue. For example, with a High Investment of $13 million dollars and a 50% take rate, the net profit is:

$$\$24 \text{ million} – \$13 \text{ million} = \$11 \text{ million}.$$

Figure 11.2.4 has been updated to show each of the four possible net profit values to the right of the end nodes.
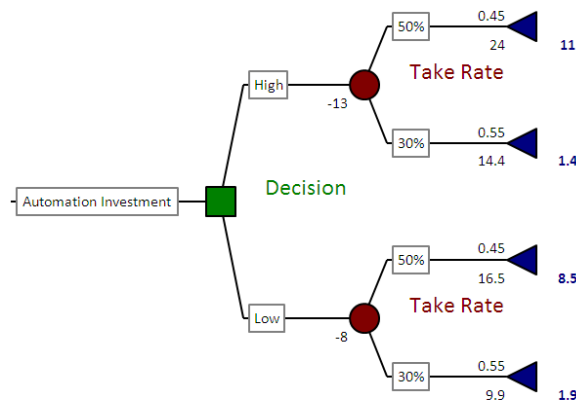
Figure 11.2.4: Decision tree showing probabilities and expected values at end nodes

The next step is to determine the expected value of the net profit for each decision. We simply use the basic formula for expected value. The expected value of the net profit for a high investment is calculated as follows:

If we make the high investment decision, there is a 0.45 probability that the net profit will be $11 million. There is a 0.55 probability that the net profit will be only $1.4 million. Thus, the expected value of the net profit for the High Investment decision is:
      Expected Net Profit = (0.45)($11 million) + (0.55)($1.4 million) = $5.72 million

Q10.   What is the expected value of the net profit for a low investment?

Q11.   Is it clear from your answers whether BC should make a low or high investment in automation? Why or why not?

Q12.   Which of the two options, high or low investment in automation, has the larger expected value for net profit? How much larger than the other option is it?

Q13.   Suppose the vendor for the High Speed automation has just announced a $500,000 price increase from $13 million to $13.5 million. Would this affect the preferred decision? What if the increase was a million?

11.2.3 Boss Controls Sensitivity Analysis

The management at Boss Controls now wonders how sensitive the solution to their automation problem is to their probability estimates of the take-rate. Recall that they have estimated the probability of a 50% take-rate to be 0.45 and the probability of a 30% rate to be 1–0.45=0.55. But, they wonder, "What if our estimate of the probability of a 50% take-rate is too high? Could that change which alternative has the larger expected value of profit? If so, at what point does the larger expected value change from high investment to low investment?" In other words, they want to learn how low the probability of a 50% take-rate could be and still have the high investment in automation as the alternative producing the larger expected value of the profit.

Q14.   Notice that the BC managers are not wondering what would happen if their estimate of the probability of a 50% take-rate is too low. Why not?

The decision to make either a high or low investment in automation rests on the expected value of the profit for each of the alternatives. Recall that those expected values are given by:
      Expected value of profit for high investment = $11 million(0.45)+$1.4 million(1–0.45)
      Expected value of profit for low investment = $8.5 million(0.45)+$1.9 million(1–0.45)

Now, for each alternative, we can define a function to calculate the expected value of the profit based on varying probability estimates.

      Let: x = the probability of a 50% take-rate,
              y_1 = the expected value of profit for high investment, and
              y_2 = the expected value of profit for low investment

      Then, $1-x$ = the probability of a 30% take-rate,
              $y_1 = 11(x)+1.4(1-x)$, and
              $y_2 = 8.5(x)+1.9(1-x)$.

Notice that in both cases we have dropped the "$" and the "million." We can do this, because we are using the same unit of measure, millions of dollars, in each case.

Figure 11.2.5 shows the set-up and calculator graphs of these two functions on the same coordinate axes. Set-up and graph this system on a graphing calculator.
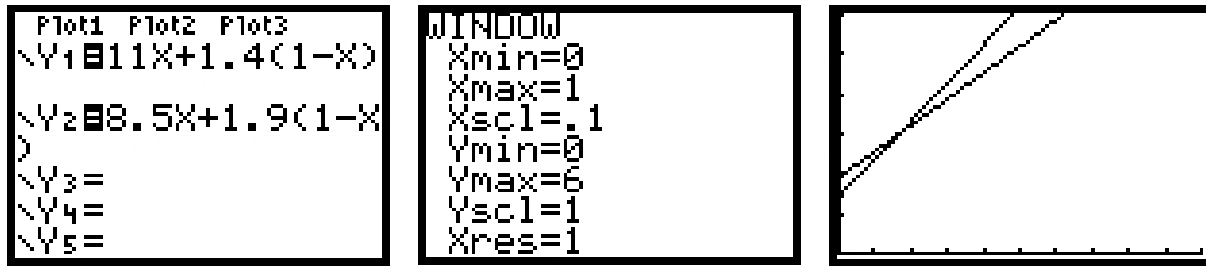
Figure 11.2.5: Expected value functions, window, and graphs for sensitivity analysis.

Q15.   The viewing window for the graphs has been set up so that x varies between 0 and 1, only. Why does that make sense for this problem?

Q16.   What do the y-values of the functions represent? What are their units of measure?

Use the TRACE feature of your calculator to trace on the graph of y_1 until the value of x is 0.45 when rounded to two decimal places.

Q17.   When $x \approx 0.45$, is y_1 above or below y_2? What does that mean in the context of the problem?

Q18.   Use the calculator to find the expected value of profit for high investment if the probability of a 50% take-rate is 0.4? What is the expected value of profit for low investment for this probability of a 50% take rate?

Q19.   What are the slopes of y_1 and y_2, the expected value of profit lines? What do these slopes mean in the context of the problem?

Q20.   Notice that the graphs of the two functions intersect. What is the significance of the point of intersection?

Q21.   Use the calculator to find the x-value of the point of intersection, rounded to two decimal places. What does this x-value tell you?

Q22.   For what probabilities of a 50% take-rate does the high investment in automation produce the higher expected value of profit? For what probabilities does the low investment produce the higher expected value?

Section 11.3: Green Tree Energy—Location a New Plant

Green Tree Power, Inc (GTP) is planning to expand their energy company by building a new power plant in a developing country. After much consideration, they have narrowed their possibilities to the countries of Cassedonia and Kisanthia. Each country can provide the required land and utilities for GTP to build and run their new power plant. In turn, the selected country will gain the benefits of the new energy technology that GTP can provide. Choosing the ideal location relies on several key pieces of information.

In order to build the new power plant in Cassedonia, GTP estimates that the investment cost will be $50 million. However, there is significant uncertainty with regard to increased demand for power. As a result, the predicted total net revenue over the next five years is uncertain. Experts project that revenues could be as low as $80 million or as high as $110 million. The specific probabilistic forecast is that five-year total net revenues will be $80, $90, or $110 million with a 30%, 40%, and 30% chance, respectively. The political structure in Cassedonia is in transition. There are multiple political parties fighting for control of the country. These political parties have very different social and economic plans. Thus, the leadership of GTP believes there is a 20% chance that a Cassedonian government will take over the new power plant. If so, the government will simply repay the original investment cost of $50 million with no interest. In that case, GTP would have no net revenue gain from its investment.

If GTP builds the new power plant in Kisanthia, the investment costs are still $50 million. The population of Kisanthia is slightly lower than in Cassedonia but demand is still uncertain. GTP estimates the total net-revenue for a five-year period in Kisanthia, after the operation costs, will be $66, $80, or $90 million with a 30%, 40%, and 30% chance, respectively. The country of Kisanthia is a long established stable democracy that is committed to encouraging foreign investments. While the total forecasted revenue in Kisanthia is significantly lower than in Cassedonia, it has a major advantage. There is little chance that the Kisanthian government will take over the new power plant.

Where should GTP build their new power plant?

11.3.1 Developing the Decision Tree

Joe Riden, a senior risk analyst from GTP, has been charged with making this decision. In order to weigh the options carefully and objectively, Joe will create a decision tree to analyze the situation and determine the best choice for GTP.

In our situation, we have one decision to make with two possible options—GTP's new power plant can be built in Cassedonia or Kisanthia. Joe decides to develop the decision tree in stages. First he places the two alternative decisions at the first decision node. The uncertainty with regard to Kisanthia involves just the net revenue. He therefore, adds the random event, net revenue, to the Kisanthian branch of the tree. He includes all of the critical information at the appropriate places. He places a -50 on the decision branch to represent the investment cost. The net revenue random event has three branches, high, medium, and low. For each branch, Joe inputs the probability and the net revenue values. For the High revenue branch the probability is 0.3 and its projected revenue for Kisanthia is $90 million. He also inputs the corresponding numbers for Medium (0.4, $80 million) and Low (0.3, $66 million). This first stage of Joe's tree construction is presented in Figure 11.3.1.
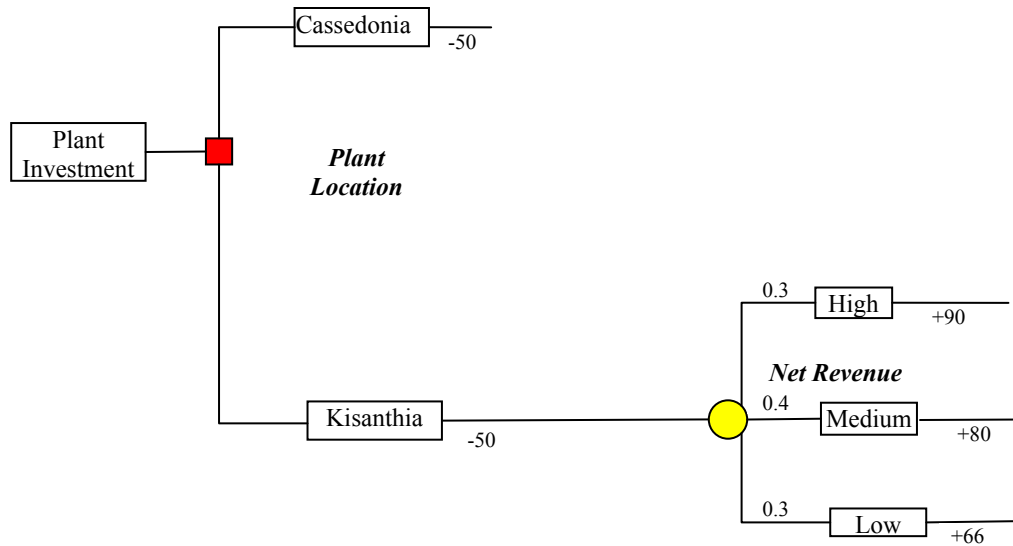
Figure 11.3.1: GTP—Kisanthian Alternative

Next Joe adds the uncertainties related to building a plant in Cassedonia. Recall there is a 20% chance that the Cassedonian government will take over the power plant after it is completed. We therefore need a chance node off of Cassedonia. The random event, a government takeover, has two branches: yes and no. If the government seizes control of the power plant, they will repay the $50 million dollar investment and GTP will no longer be involved with the plant. This return of $50 million is included along the Yes branch for a government takeover.

However, there is an 80% chance that the government will not take over the new power plant and GTP can move on to producing power, and thus revenue. Down this branch, there is another random event, uncertain net revenue. Joe attaches another random node with three branches for the net revenue. Again, each branch has a probability and dollar amount for net revenue: High (0.4, $110 million) Medium (0.4, $90 million) and Low (0.3, $80 million). Joe added the probability of a takeover and the net revenue uncertainty as presented in Figure 11.3.2.
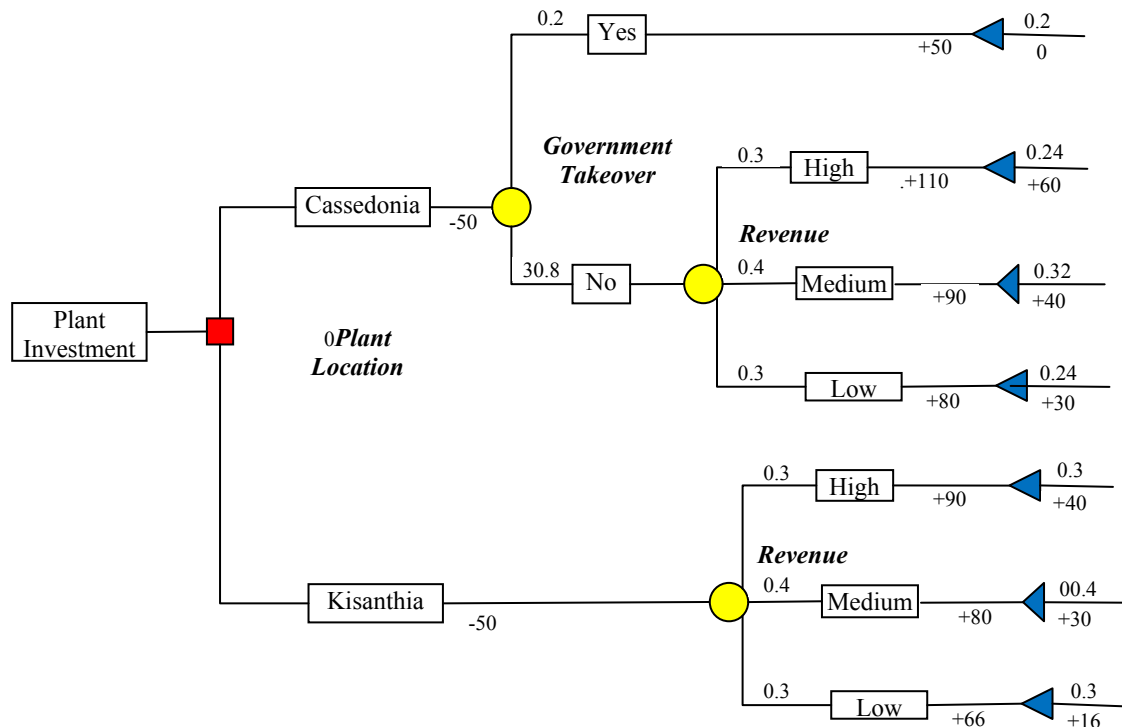
Figure 11.3.2: GTP – Cassedonian Alternative Added

In order to make the decision, we need to perform some calculations. At each end node, we want to note two important values: the probability of following that path from start to finish and the total profit for GTP if they follow that path. For example, if GTP were to build in Kisanthia and net revenues turn out to be high, the net profit would be $40 million: $90 million in net revenue minus $50 million investment. The probability of that happening is just 0.3. He then adds the probabilities and net profits for the other two branches.

For Cassedonia, if the government takes over the plant, the net profit is zero. This has a probability of 0.2 of occurring. If the government does not takeover the plant and the net revenue is high, the net profit is $60 million ($110-$50). The likelihood of this happening is the product of two probabilities. Mr. Riden multiplies the probability of no government takeover, 0.8, by the probability of high net revenue, 0.3. The product is 0.24. He repeats this process for the other two possible branches for net revenue for Cassedonia. Figure 11.3.3 contains all of the final dollar values and probabilities placed alongside the end nodes. The probability is placed above the net profit value.
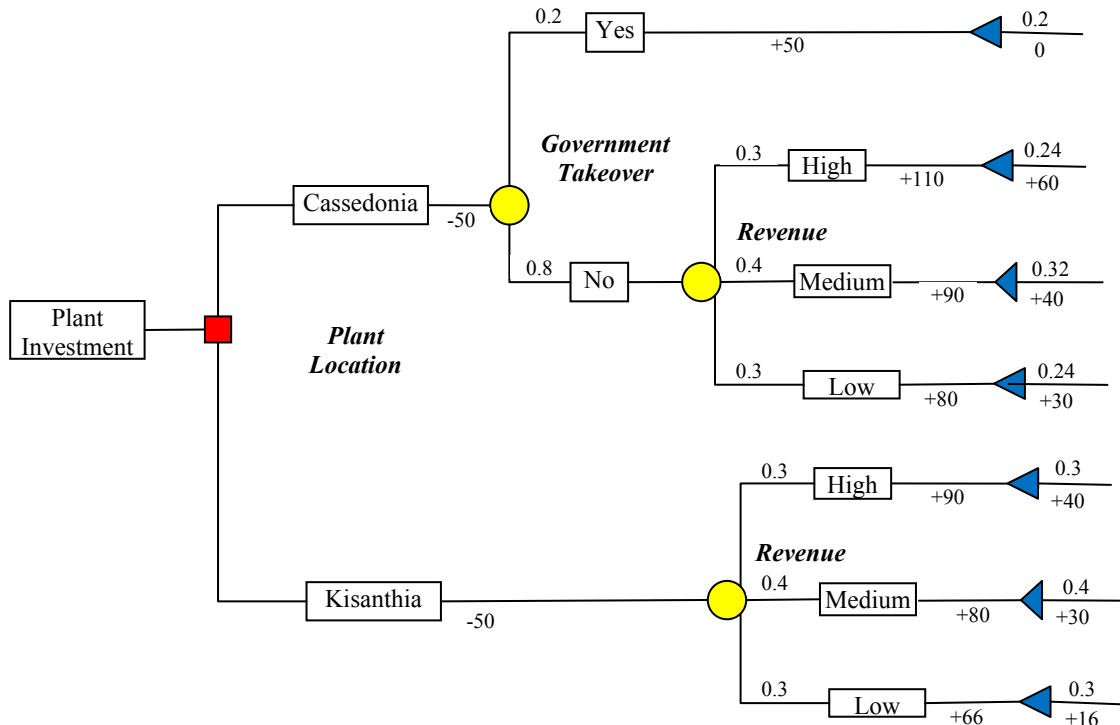
Figure 11.3.3: GTP Decision Tree

11.3.2 Computing the Expected Value

The final step in making the decision is to determine the expected value of the net profit for each alternative decision. We need the expected value of the net profit for building the power plant in Cassedonia and the expected value of the net profit for building in Kisanthia. The country with the larger expected value will be the best option for GTP.

The expected value for net profit of building the plant in Kisanthia is determined by taking the weighted sum of the different possible net profit values. There is a 30% chance of making $40 million, a 40% chance of making $30 million, and a 30% chance of making $16 million. So, the expected value of net profit is:

$$(0.3)(\$40 \text{ million}) + (0.4)(\$30 \text{ million}) + (0.3)(\$16 \text{ million}) = \$28.8 \text{ million}$$

Consider the option of building in Cassedonia. There is a 20% chance of making $0, a 24% chance of making $60 million, a 32% chance of making $40 million, and a 24% chance of making $30 million. Thus, the expected value of net profit for building the plant in Cassedonia is:

$$(0.2)(0) + (0.24)(\$60 \text{ million}) + (0.32)(\$40 \text{ million}) + (0.24)(\$30 \text{ million}) = \$34.4 \text{ million}$$

Q1.   Based on the expected values, which country is preferred for the investment?

Q2.   What is the minimum net profit for Cassedonia? How likely is that to occur?

Q3.   What is the minimum net profit for Kisanthia? How likely is that to occur?

Q4.   How might the issue of risk affect your preferred decision?

11.3.3 GTP, Inc. Considers Insurance

Freud's of London understands the psychology of risk. It offers specialty risk insurance for large projects. Freud's is prepared to offer GTP insurance against a possible government takeover in Cassedonia. They are prepared to charge GTP a $3.5 million upfront premium to insure against a government takeover. If the government takes over GTP plant, Freud's will pay GTP $10 million dollars. We will analyze this insurance policy both from GTP's perspective and that of Freud's. The GTP part of the tree is revised and presented in Figure 11.3.4. a negative $3.5 million dollars for the insurance premium is added to the cost of the Cassedonia branch. That total cost is now $53.5 million. As a result each end value for NO government takeover decreases by $3.5 million. However, the end node for the YES government takeover now has a net profit of $6.5 million dollars. They receive $50 million back from government plus a $10 million payment from the insurance company for total revenue of $60 million against costs of $53.5 million. This yields a net revenue of $6.5 million.
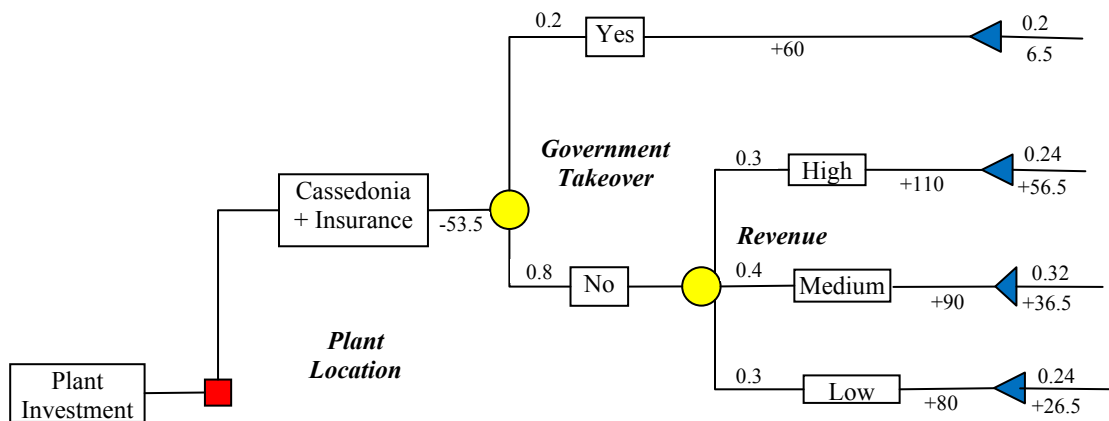
Figure 11.3.4: GTP Cassedonia with Insurance

The new expected value of net revenue for the Caledonian option can be calculated:
  (0.2)($6.5 million)+(0.24)($56.5 million)+(0.32)($36.5 million)+(0.24)($26.5 million) = $32.9 million

The expected net profit has declined by $1.5 million. However, GTP is assured now of making at least $6.5 million from its $50 million investment.

Q5.   Would you recommend buying the insurance?

The above analysis focused on GPT's perspective. Let's look at the decision from the insurer's perspective of Freuds of London. If they offer insurance, they face only one uncertainty, a government takeover. They are uninterested in the uncertain event, revenue. Their decision tree is presented in Figure 11.3.5. If they sell no insurance, they gamble no money and they make no profit. If there is a government takeover, Freud's must pay GPT $10 million. Their net loss would be $6.5. With no takeover, their profit is the premium they charged.

Q6.   How much risk does Freud's face?

Q7.   What is the expected value of the profit Freud's will earn?
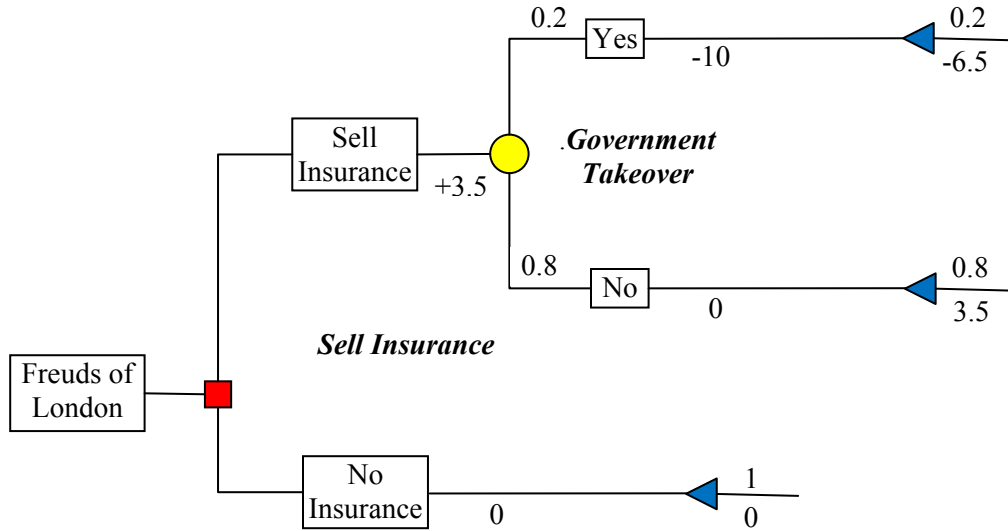
Figure 11.3.5: Freud's of London Insurance Decision

The early decision analysts recognized that expected value alone did not represent how many individuals and companies deal with risk. Many of us are "risk averse." This means we would be willing to accept a reduced expected value in exchange for more certainty. This is the reason people buy insurance. They are willing to pay money to avoid risk. The cost of the insurance is always more than the expected value of the loss. We pay hundreds of dollars to insurance against a relatively unlikely catastrophic risk to our homes. We buy family medical insurance for thousands of dollars to cover the cost of a major surgery and a long hospital stay that could cost more than a hundred thousand dollars. As individuals living one life and facing one situation, we cannot rely on long range expected value. However, insurance companies can tolerate these types of risks and live by the expected value. They can pool the risk across thousands of customers each year. Their financial performance will approximate the expected value.

Decision analysts developed a concept called utility theory to quantify this concept of risk aversion. The mathematics of utility theory is beyond the scope of the course. Instead, we will present tradeoffs between reduced expected value and more certainty and let you judge your preference. In the next example of automobile collision insurance, we will challenge you to revisit your own attitude towards risk when making smaller insurance decisions

Section 11.4: Purchasing Collision Insurance

Jee Min is a high school junior at Cassidy High School in Thomasville, Michigan. He has been an excellent driver for one year. With the help of his parents, he has just purchased a 2005 Chevrolet Cavalier. Jee drives his car to school and to his part-time job on the weekends and after school. He is considering purchasing collision insurance for this car. It has a *Kelly Blue Book* value of $6600.

After gathering quotes from insurance companies through the internet, Jee has learned that the lowest six-month premium for collision insurance with a $500 deductible is $1700. He is not sure that he can afford that much, so he decides to investigate the cost of collision insurance with a $1000 deductible. The six-month premium for collision insurance with a $1000 deductible is $1500. The deductible represents the maximum amount of loss the owner incurs in the case of an accident. With a $500 deductible, the owner must absorb the cost of the first $500 of damages. For example, if the damages were $250, he bears the whole cost. If the damages were $6600, he absorbs the $500 and the insurance company pays him $6100. Jee Min is also considering the possibility of not carrying any collision insurance. In order to make the best decision, he must consider all of the consequences of each possibility.

To examine his options, Jee created a decision tree showing the three collision insurance possibilities. Each branch of the tree is labeled. Let's examine his decision tree. The tree begins with a decision node, because Jee must decide whether to purchase collision insurance with a $500 deductible, a $1000 deductible, or no collision insurance at all. The next node on each of the three branches is a probability event: a collision in the next 6 months. He chose a 6 month time period for this event, so that it matches the six-month period covered by the premium quotes that he obtained. Next, Jee Min needed an estimate of the probability that he would be involved in a collision in the next six months.

After some Internet research, Jee Min learned that according to the National Highway Safety Administration, the probability that a male teenage driver in the U.S. will have an accident in any six-month period is 30%. He decided to use this probability estimate. Then he attached branches to the probability nodes to account for the possibilities that he will have an accident or not during a given six-month period.

Q1.　　How do you think the National Highway Safety Administration determined the probability a male teenage driver will have an accident in a six-month period?

Q2.　　Why was 70% assigned to the branches representing the event that Jee Min will not have an accident in the next 6 months?

Let's explore the structure of the tree (See Figure 11.4.1).

Q3.　　There are three branches that leave the main rectangular decision node. What do these three branches represent?

Q4.　　Each of these three branches leads into a random circle node. What does that random node represent? Why are there two branches coming out of this node?

If Jee Min does have an accident, various damage amounts are possible. He realized that even if he rounded damage amounts to the nearest dollar, there would still be 6,600 possible damage amounts. Therefore, he decided to list these amounts in ranges. He also investigated the probabilities that the damage amount of an accident falls within that range. Jee Min also collected this probability data online from the National Highway Safety Administration. He organized this information in table 11.4.1.

Damage Amount Ranges	Probability
Less than or equal to $500	45%
Greater than $500 but less than or equal to $1000	15%
Greater than $1000 but less than or equal to one-half the value of the car	25%
Greater than one-half the value of the car but less than or equal to $6,600, the total value of the car	15%

Table 11.4.1: Damage Estimate Probabilities

Based on this data, Jee Min added this information to the decision tree. He added another probability node to each branch of the tree that represents having or not having an accident in the next 6 months. Then he added a branch for each of the possibilities in Table 11.4.1 to each new node. Finally, he labeled each new branch with the probability of incurring that amount of damage. Note the number of sequences of branches in Jee Min's tree. The probability of that sequence of random events is listed at the end of each sequence of branches.

Q5.   Each branch labeled YES contains another random node? What does this represent? Why are there four branches leaving this second node?

Let's explore the probabilities on the tree. There is a probability assigned to the end node of each sequence of branches. For example, the top end node has a 0.135 assigned to it.

Q6.   How was this end probability determined?

Q7.   Which end nodes have the highest probabilities and why?

Q8.   The probabilities for the top five branches in the picture sum to 1. Why is this the case?

Let's discuss the dollar amounts of the tree and the end node values.

Jee Min realized that for each complete decision branch of his tree, he must assign a cost to each branch. However, in order to do these calculations, he realizes that he needs *individual* damage amounts and not *ranges* of damage amounts. Jee Min decides to use a single number within each range to represent the damage amount for that branch. He listed those in the following table.

Damage Amount Ranges	Representative Amount
Less than or equal to $500	$250
Greater than $500 but less than or equal to $1000	$750
Greater than $1000 but less than or equal to one-half the value of the car	$2150
Greater than one-half the value of the car but less than or equal to the total value of the car	$6600

Table 11.4.2: Damage Estimate -Representative Amounts

For the three lowest damage ranges, Jee Min decided to use the midpoint of the range: $250 for accidents having damage less than or equal to $500, $750 for accidents having damage greater than $500 and less than or equal to $1000, and $2150 for accidents having damage greater than $1000 and less than or equal to one-half the $6600 value of his car. For the last range, damage greater than one-half value and less than or equal to the full value of his car, he learned that insurance companies almost always "total" the car when the damage falls within this range. Therefore, he decided to use $6,600 for accidents in that range.

All four of these dollar amounts are entered into the appropriate branches of the tree for the no collision insurance decision. Now let's look at the section of the tree with the decision, $1000 deductible, and the random event the damage. The dollar amounts for these four branches do not match the values in table.

Q9.   Which two branches match the table and which two do not? Why?

Q10.   For the section of the tree with a $500 deductible, there are three branches with a $500 value. Why is this the case?
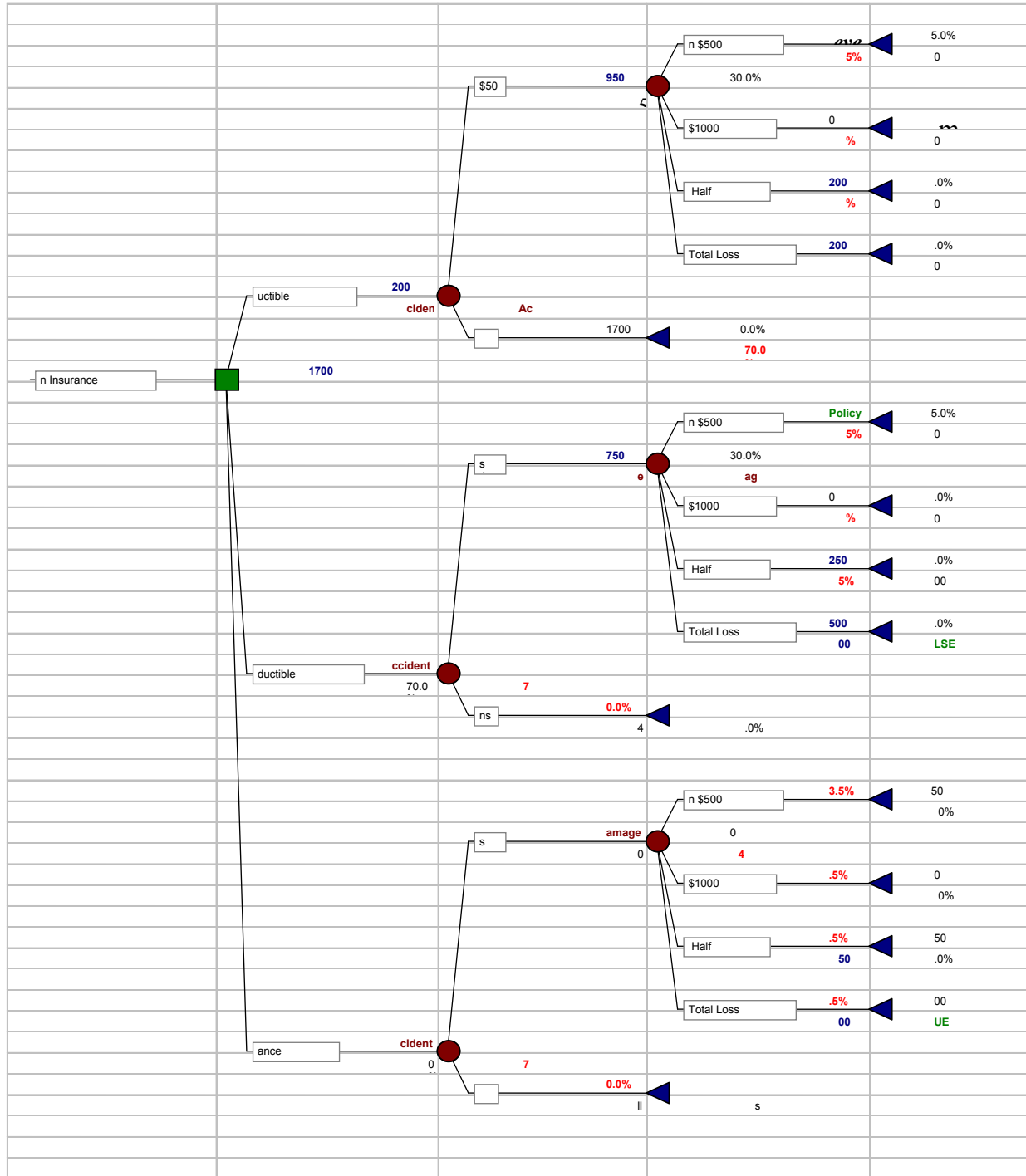
Figure 11.4.1: Collision Insurance Decision Tree

Consider only the portion of the entire decision tree that represents the decision to purchase $500 deductible collision insurance. Recall that the premium for this deductible amount is $1700. For the top-

most branch, the cost is $250. Thus, the end node value is the sum of these two costs, $1950. The second branch, has a $500 cost. The end node value is therefore $2,200.

Q11.   Why are the end node values for the third and fourth branch the same $2,200 as for the second branch?

In a similar way, Jee Min adds the cost of the premium with his liability for damages to determine the end node value for each sequence of branches.

Q12.   What is the expected cost to Jee Min of the decision to purchase $500 deductible collision insurance?

Q13.   Examine each of the remaining decision options and the total cost. In a similar way, determine the expected cost to Jee Min of making that decision. Enter each of the expected costs in the table below.

Decision	Expected Cost
$500 Deductible	
$1000 Deductible	
No Insurance	

Table 11.4.3: Expected Costs

Q14.   Based on this analysis, which option has the smallest expected cost?

Q15.   Should Jee Min base his decision only on this analysis? Explain why or why not.

The basis for the insurance industry recognizes the concern people have with incurring huge costs associated with relatively infrequent events. Every reduction in premium fees can be directly subtracted from the total expected value to determine the impact.

Q16.   Given your own attitude towards risk, would you be willing to pay more than the expected value calculated for the no insurance options to reduce your risk? If so, how much would you be willing to pay to have a policy with a $1,000 deductible?

11.4.1 A Revised Estimate of the Probability of Having an Accident

Over the past six years, Michigan and other states in the U.S. have instituted graduated driver training programs for new teenage drivers. In fact, Jee Min participated in such a program. The establishment of these programs has resulted in safer teenage drivers.

What if the National Highway Safety Administration has now determined the probability that a teenage driver in the U.S. will have an accident in any six-month period is 22%.

Q17.   Will this new estimate of the probability of having an accident affect the expected cost for each decision?

Q18.   Using this new estimate, what is the probability of *not* having an accident in the six-month period?

The insurance companies are slowly responding to this reduced rate of accidents. They are considering significant reductions in the premiums charged.

Q19.   If the premiums were reduced by 30%, what would be the expected values for each of the insurance policies?

Q20.   Recalculate the expected cost for no insurance option. Should Jee Min change his decision? Explain.

Section 11.5: Chapter 11 (Decision Trees) Homework Questions

1. The probability of selling a dress in a store is 25 percent each week.
    a) Construct a probability tree to determine all of the possible of outcomes over a three-week period.
    b) What is the probability that the dress will not be sold at the end of the third week?

2. A contestant on a TV show must pass four stages to win a big prize. The probabilities of winning in stage 1, 2, 3, and 4 are 0.8, 0.6, 0.4, and 0.3 respectively. The contestant wants to know the probability of winning the big prize.
    a) Construct a probability tree to determine the possible outcomes of the game.
    b) What is the probability that he wins the prize?
    c) What is the probability that he makes to stage 4 but does not win the prize?

3. A TV cable company has a technical support department to solve customers' problems by phone. In this department the staff is categorized in four levels based on their ability to solve customer problems. The company first assigns a problem to level one; if they cannot solve it, someone at level two is assigned. This process is repeated until it finally reaches the most experienced staff for one last attempt at solving the problem. The probability that a staff person is able to solve the problem at each level is 0.50, 0.75, 0.85, and 0.95 consecutively.
    a) List all of the possible outcomes.
    b) Construct a probability tree to determine the probability of each outcome.
    c) What is the probability that a customer's problem is unsolved?
    d) What is the probability that the problem is solved by someone at level 3?
    e) Why is this probability less than the probability that a level 1 individual solves the problem?

4. A manager at Wayne State football games must decide ten days in advance which product to order for the stadium vendors to sell. Each product will have the university logo. The three options are sun visors, umbrellas, and ponchos. He will stock only one of the items. Sales and the resultant profit will depend upon the weather on the day of the game. The long-range weather forecast is 35% chance of rain, 25% chance of overcast skies, and 40% chance of sunshine. The manager estimates that the following profits will result from each decision and each weather condition.

	Weather Condition		
Decision	*Rain(.35)*	*Overcast(.25)*	*Sunshine(.40)*
Sunglass	-600	-300	1600
Umbrella	2100	0	-800
Poncho	1800	500	-600

    a) What is the best decision for each weather condition?
    b) Draw the associated decision tree needed to make the best decision.
    c) What decision should be made if he desires to maximize the expected value?

5. The owner of a restaurant is considering two ways to expand operations: open a drive-thru window or serve breakfast. There are increased annual costs which each option and a one-time cost associated with the drive-thru. Labor and marketing costs are annual costs that the restaurant has to pay each year that include hiring new staff and placing more ads in local media. Redesigning the restaurant is a one-time cost that is paid at the beginning and does not repeat each year. The detail is provided in the following table.

Decision	Costs		
	Annual		One Time
	Labor	Marketing	Restaurant Redesign
Drive-thru window	28,000	10,000	50,000
Breakfast	38,500	5,000	-

The forecasted increase in income resulting from these proposed expansions depends on whether a competitor opens a restaurant down the street or not. Based on the restaurants evaluation, the manager is sure that the competitor won't open a new restaurant with 60 percent of certainty. Based on the competitor action, the restaurant's profit will be different for each decision. Following table provides estimation in increase of income based on the competitor's action.

Decision	Competitor	
	Open(0.4)	Not Open (0.6)
Drive-thru window	110,000	130,000
Breakfast	80,000	120,000

The owner of the restaurant is focused just on next year. He therefore decided to consider the one-time cost for the redesign the same as all of the labor and marketing costs that are ongoing.

a) Calculate the profit of each decision when considering the competitor's action.
b) What is the best alternative if no competitor opens nearby? What is the best alternative if a competitor opens nearby?
c) Draw the associated decision tree.
d) What decision should the company follow?
e) Let p represent this probability that the competitor will open a restaurant down the street. Write an equation to calculate the expected value for each decision as a function of p.
f) For what value of p are these two expected values equal?
g) Graph the equations of the expected values to determine their intersection point. What does this intersection point represent?
h) Recall that the owner treated the design change and marketing cost the same as operating costs. Would the decision change if he considered only 50% of these costs this year (design and marketing)?

6. A company is about to launch its new fast food for sale in supermarkets throughout Arkansas. The research department is convinced that a special type of chicken wings will be a great success. The marketing department wants to launch an intensive advertising campaign. The advertising campaign will cost $1,000,000 and if successful will produce $4,800,000 profit. If the campaign is unsuccessful (25% chance), the profit is estimated at only $1,800,000. If no advertising is used, the revenue is estimated at $3,500,000 with probability 0.6 if customers are receptive and $1,500,000 with probability 0.4 if they are not.

a) Draw the associated decision tree.
b) What course of action should the company follow in launching the new product if they want to maximize the expected value?
c) Write an equation to calculate the expected value for each decision as a function of the probability that the major advertising campaign will be effective?
d) Graph the equations of the expected values to determine their intersection point. What does this intersection point represent?

7. A discount clothing store uses an interesting strategy to attract customers to return each week to shop. They tell the customers that every 7 days they reduce by 25% the original price of any unsold dress. On each dress there is a label of its original price and the date it was hung on the rack. Thus customers know that a $40 dress placed out on Nov. 7th will be priced only $30 on Nov. 14th if it is not sold before then. It will be reduced by another $10 on Nov. 21st if it is still unsold. After three weeks, any unsold dress is sent to a local charity. Each week, there is a .60 probability that the dress will be sold.

   a) Nancy Drew saw a dress she really liked and knows she can get the almost identical dress for $50 online. The current store price is $40. Construct a decision tree to determine whether or not she should buy the dress now or gamble and wait a week and buy it next week if it remains unsold. (If when she comes back next week, Nancy finds the dress has been sold, she will buy it online.)
   b) Just before finalizing her decision, she found another place online that sells the same dress for $45. Why might a lower price online affect her purchase decision in this store? Should she buy the dress now or gamble and buy it in the second week if available?
   c) She just saw a more expensive dress for sale at $80. These more expensive dresses have only a 40% chance of being sold each week and again they tell the customers that every 7 days they reduce by 25% the original price of any unsold dress. If she would be able to buy a similar dress with $90 online, construct a full tree for 3 weeks.

8. A contestant on a TV show has to decide whether to stop or try to answer another question. The contestant is first asked a question about US Geography. If the contestant answers correctly, she earns $700. Historically three out of four contestants answer the first question correctly. If answered incorrectly, the game is over. If answered correctly, the contestant can leave with $700 or go on and answer a question about US presidents. If answered correctly, the contestant wins an additional $1000. If the answer is incorrect, the contestant loses all previous earnings and is sent home. Historically, two out of three contestants answer this question correctly. The third question is about rock 'n' roll music. This question is worth $1500, and the same rule applies. The chance of answering this question correctly is 50-50.
   a) Draw a decision tree that can be used to determine how to maximize a contestant's expected earnings. What is the best decision and what are the expected earnings in this case?
   b) Some contestants may feel more or less knowledgeable about the third question category. Let p represent the probability that a contestant will answer the third question correctly. Write an equation to calculate the expected value for attempting to answer the third question in terms of p.
   c) Based on the previous question, what is the cutoff value of p such that a contestant should attempt the third question?
   d) The TV show is considering changing the reward for answering the 3rd question correctly. Let m represent the amount of money a contestant will earn for answering the third question correctly. Write an equation to calculate the expected value for the last decision as a function of m.
   e) Graph the equations of the expected values to determine the intersection point for the last stage. What does this intersection point represent?

9. SSS Company, a software company, is considering submission of a bid for a state government contract to install their software on 30,000 computers. The government would use their software to oversee the management of tens of thousands of large and small contracts the government signs every year. There is only one other potential bidder for this contract, Complexo Computers, Inc. Complexo has a long record and reputation with this kind of contract. As a result of its lesser experience, to win the bid SSS's bid must be at least $5 less per computer installation than Complexo's. Complexo Computers is certain to bid and is generally more expensive than SSS. SSS management believes that it is equally likely that Complexo will bid $100, $90, or $80 per computer installation.

a) What are the possible bids that SSS should consider?

SSS's bidding decision is complicated by the fact that it is currently working on a new process to install software remotely through the internet. If this process works as hoped, then it may substantially lower the cost of installations. However, there is some chance that the new process will actually be more expensive than the current installation process. Unfortunately, SSS will not be able to determine the cost of the new process without actually using it to install the software. The higher SSS bids the more money it makes if it wins the contract. However, the higher the bid, the less likely it is to win the contract. If SSS decides to bid, it will cost $20,000 to prepare all of the relevant documents required to submit the bid. SSS will incur this expense regardless of whether it wins or loses the bidding competition. With the proposed new installation process, there is a 0.25 probability that the cost will be $50 per computer and a 0.50 probability that the cost will be $75 per computer. Unfortunately, there is also a 0.25 probability that the cost will be $85 per computer.

b) Construct a decision tree to model this situation.
c) Based on your decision tree, do you recommend SSS Company to submit a bid, and if so, what should they bid per installation?
d) Under the optimal policy, what is the probability they will win the contract?
e) What is the overall expected value if they bid on the contract?
f) If they win the contract, what is their expected value of profit?

10. A group of high school students has decided to start a summer business. The work that they are thinking about is designing and coloring T-shirts and selling them to clothing stores in their community. For mass production of colored T-shirts, they need special equipment which they can buy or rent. After negotiating with a company about equipment, they figure out that they have three options to start their business:

- They can buy all the equipment and do the design and printing themselves. In this case they have to pay for equipment but they can recover part of the money at the end of summer by reselling the equipment. The cost of buying equipment is $8,100, and they can resell it at 50% of the original price. The cost of printing will be $1 per T-shirt.
- The second option is renting the equipment and returning it at the end of summer. The renting cost is $1,500 for the whole summer and a variable cost of $1.50 per print.
- The third option is outsourcing the printing. In this case they do the designs themselves but send them to a company for printing. The company charges them $2 per T-shirt. Note that in option one each T shirt costs them 1 dollar.

The fact that the market demand for colored T-shirts is not certain makes the decision making difficult. After doing some market evaluation, they summarized their expectation in following table.

Demand (Number of T-shirts)	Probability
2000	15%
5000	50%
8000	35%

a) If they can sell each T-shirt for $5, construct a decision tree to help them make their decision.
b) What is the best option if the demand is 2,000 T-shirts?
c) What is the best option if the demand is 5,000 T-shirts?
d) What is the best option if the demand is 8,000 T-shirts?
e) Which option is the best for them? What is the expected profit if that decision is made?

11. A software company released a beta version of a software package. It expects a large number of requests from the users for fixing potential bugs in the software. These include crashing, lock up, and incompatibility errors. The company has established a help desk to handle telephone requests. The company trained two groups of software specialists to support the software. Group 1 has just been hired and trained; meanwhile specialists in Group 2 are senior technicians very capable of solving the problems. The senior specialists solve the problems with 100 percent certainty, but their salaries are much higher than other specialists. The payment system of the company for the specialists is problem based and the company pays them based on the number of the problems that they attempt to solve. Group 1 salaries are $20 per problem and Group 2 salaries are $35 per problem. The software company always has a dilemma as to which specialist to assign in order to minimize the cost of the support. If they assign a problem to a Group 1 specialist and he is not able to solve the problem, they reassign it to a senior specialist. In this case pays both specialists are paid. This costs the company $55 per problem. To address this issue, they developed an automatic system to predict the chance of solving a problem by group 1 based on previous cases.

   a) A crashing problem was just received, and the prediction software forecasts a 70 percent chance of success for a Group 1 specialist. Draw a decision tree for this problem.
   b) Based on the decision tree, what kind of specialist should be assigned to the problem first?
   c) Another problem, compatibility error, was received and the prediction software forecasts a 50 percent chance of success for a Group 1 specialist. Draw a decision tree for this decision.
   d) Based on the decision tree, what kind of specialist should be assigned to the problem?
   e) Let p represent the probability that the Group 1 specialist will be able to solve the problem. Write an equation to calculate the expected value of the cost for each decision as a function of p.
   f) Graph the equations of the expected values to determine their intersection point.
   g) They want to know what should be the cutoff value of the probability to assign directly to an expert instead of assigning the task to a group-one specialist. Use both a graphical representation of part e) and an algebraic representation of part f) to find that probability.
   h) In the previous question, what was the role of the two salaries in determining the beak even value of p? Assume that the salary of group one specialists is x and the salary of group two workers is y. Write an equation to calculate the expected value of the cost for each decision as a function of p, x and y.
   i) Find the value of p as a function of x and y that leads to the same expected value of the cost regardless of whether the problem is first assigned to a Group 1 or Group 2 specialist.

12. (Continuing the previous problem) After finishing the first phase, management figured out that they need a Group 2 category of specialists that are more knowledgeable than Group 1 but not necessarily experts. They are to be paid at $28 because of their higher success rates than Group 1.

   a) A problem was just received and the prediction software forecasts 70 percent chance of success for a group-one specialist and 85 percent chance of success for a group-two specialist. Experts can solve the problem for sure. Draw a decision tree for this problem. (Assume that if a group-two specialist fails to fix the problem, the company won't assign it to group one)
   b) Based on the decision tree, what kind of specialist should be assigned to the problem first?
   c) What if the probability of success for Group 1 would be 60% and still 85% for Group 2?

13. An automotive part has to go through two different processes by metal lathes to be shaped properly. Each process has a cost associated with the type of lathe. Each step in processing has a risk of ruining the part and turning it into scrap. For example, when Lathe 1 is processes a part, the cost is $100 and the risk of being scrapped is 10%. Each part that successfully processed by both lathes is sold for $450. The net profit is equal to the number of parts sold minus the cost of processing all parts. The

cost of processing includes both finished and scrapped parts. The following table shows the cost and risk of each lathe.

	Cost	Risk of scrap
Lathe 1	$100	10%
Lathe 2	$150	20%

There is no recycling value if the part is ruined; scrapped parts are worthless.

a) What is the probability that a part will end up being scrapped? Does it make a difference as to the order of the processes?
b) The processes can be done in either order. Draw a decision tree to determine the optimal sequence of processes which maximizes the expected value of the net profit per part.
c) Which process should be done first?
d) If the cost of processing by Lathe 2 changed, for what cost would the optimal strategy change? What is the percentage change?

14. Continuing the previous problem, suppose there are three processes that must be done by three different lathes. Parts that are made with these 3 processes are sold for $800 each. Each lathe has an associated cost and risk of ruin as follows:

	Cost	Risk
Lathe 1	$100	10%
Lathe 2	$150	20%
Lathe 3	$200	25%

a) How many different sequences need to be considered?
b) What is the probability that a part is ruined?
c) Draw a decision tree to determine the optimal sequence of processes which maximizes the expected value of the net profit per part.
d) In which order should the processes be done?
e) Explain how you can use a pair-wise comparison among processes to find the optimal sequence?
f) If there were four processes, how many pair-wise comparisons would need to be made to find the optimal sequence?

Chapter 11 Summary

What have we learned?

We have learned that decision-making often involves uncertainty. It may be easy to choose between two events such as taking a job that pays $8 per hour or a job that pays $10 per hour. However it becomes more difficult to choose if the benefits are less certain. What if there was a chance the higher paying job involved a lower base pay but you could earn tips so your pay was $6 per hour with tips ranging from $0 to $4 per hour. Your decision would need to be based on how likely those amounts were.   In order to choose between multiple options, we need to be able to identify not only the benefit of different possibilities but also the likelihood of the different possibilities occurring. A decision tree can help us to combine those together to find the expected values of different options. The option we should choose is the one that has the highest expected value. It is interesting to note that often, the expected value is not actually a possible value.

Using probability trees is a multiple step process:

1.   Create a decision tree including decision nodes, chance nodes and end nodes
2.   Identify the potential cost or benefit for each branch off a node
3.   Identify the probability for each branch off a chance node
4.   Find expected value for each branch
5.   Combine the expected values for each branch of the decision node
6.   Identify the decision node branch with the best expected value

Terms

Probability Tree   A diagram used to show all the possible outcomes for a combination of events

Fundamental Principle of Counting   If there are m possible outcomes for one event and n possible outcomes for a second event then there are $m \times n$ possible ways in which both events can occur

Compound Event   An event that is the result of more than one outcome

Multiplication Rule   If A and B are independent events then the probability of A and B both happening equals the probability of A times the probability of B

Independent   Events A and B are independent if the outcomes for events A and B do not effect each other

Random Variable   A variable whose value depends on random events

Expected Value   The weighted average found by multiplying the possible results of an random variable by the probability of the random variable having that value

Chance Event   An event whose likelihood must be predicted using probability and is outside anyone's control

Decision Tree   A diagram similar to a probability tree to model a situation representing possible scenarios depending on decisions in addition to random events

Node   The boxes (representing decisions), circles (representing random chance), and triangles (representing the end of a path) on the decision tree

Arcs   The connections between nodes on a decision tree showing possible scenarios

Take Rate   The proportion of customers that select a particular option

Risk Averse   A desire to avoid risks even when the potential benefit of an action is large

Chapter 11 (Decision Trees) Objectives

You should be able to:

- Create a decision tree including decision nodes, chance nodes and end nodes

- Identify the potential cost or benefit for each branch off a node

- Identify the probability for each branch off a chance node

- Use the fact that probabilities must add up to 1

- Find value for each branch of the decision tree

- Find the expected value for each option

Chapter 11 Study Guide

1. What is a decision tree and for what is it used?

2. What are nodes on a decision tree? What shape is used for each type of node?

3. How do you find the end node value for a branch of a decision tree? Where do you put that value on the decision tree?

4. Once you find the value for each branch of a decision tree, how do you find the expected value for each option of the decision nodes?

5. What must be true about the sum of the probabilities for each branch after a decision node?

6. What does risk averse mean and how does it affect the results of decision tree analysis?

7. Give an example of risk aversion.

Section 12.0: Introduction

The world around us is filled with uncertainty, risk, and variability that complicate day-to-day decisions and the development of long-term plans. This applies to personal decisions as well as decisions made by companies and government agencies. When you work on college applications, you cannot be sure which schools will accept you. In a rush to get a meal between classes, you will face variability in the time it takes to be served at the school cafeteria. The school newspaper editor is concerned about how many members of the writing team will meet their deadlines. While reviewing alternative car insurance plans, the student driver struggles to decide on the size of the collision insurance deductible. Obviously, no one plans to have an accident, but the risk of an accident is always present. Companies that provide insurance look at the same problem. They come up with pricing strategies for insurance that pool the risk of an individual with large groups of similar people.

Companies launching a new product or service must deal with uncertain demand. Police patrol supervisors must consider random fluctuations in the demand for service when planning how many patrol officers are needed on each shift. Plant managers and school officials must cope with workers who randomly do not show up for work due to illness.

This text is designed to provide decision-making guidance in the presence of uncertainty. One of the challenges in learning basic concepts of probability is that many of us do not have good intuition about randomness. We will address this problem while developing probabilistic decision-making skills. The text is therefore designed to first develop your ability to recognize and understand patterns of random events. We will do so by having you simulate random experiments first with a coin flip, then with the random function in your calculator, and lastly using the random number generator in Excel to develop and analyze a large sample. In introducing basic concepts, we will routinely use the concept of relative frequency as an estimate of probability. Thus, our introductory examples will use data rather than the counting methods that you may have seen in other courses.

In the succeeding chapters we introduce mathematical formulas that can be used to describe different patterns of randomness. These formulas will be applied to two different types of variables: discrete and continuous. A discrete variable is countable. For example, the number of crimes in a day and the number of people absent from work are discrete variables. On the other hand, a continuous variable cannot be counted and is often measured instead. For example, the time to complete an exam and the height of an individual are continuous variables. These formulas are necessary building blocks for later chapters that cover many probabilistic modeling tools. You will also learn fundamental principles of managing processes so as to reduce their variability.

In order to fully understand probabilistic modeling, you will need a good sense of the nature of random behavior. All of us have basic intuition about probability and randomness. This intuition develops over time from our experiences, what we watch on television, observe at sporting events, and read on the internet. The problem is that research has shown that many of our intuitions about randomness and probability are incorrect. For example, a manager being told that each and every one of his ten suppliers is 95% reliable often believes the system he has in place is in good shape. In the chapter that follows, you will learn why the manager needs to be concerned.

Because of the many misconceptions, the development of probability analysis skills requires demonstrating the flaws in our intuition. Only after this is accomplished, can you move on to learn formal approaches to probability.

The first example we explore involves understanding what is often our preconceived notion of a random pattern. Specifically, we will ask you to imagine and write down a typical sequence of heads and tails when flipping a coin. You and your classmates will then be asked to actually flip a coin. By comparing the lists, we hope to dislodge a misconception you may have about random events.

One of the primary goals of this section and all future sections is to begin understanding how much variability there may be in a simple random process. This will be accomplished by comparing and contrasting the results of your coin flipping experiment with those of your neighbors and those of the class as a whole. For example, we will explore how much difference there is in the percentage of heads and tails in your list and those of each and every one of your class mates. We will then look at the average for the entire class.

Section 12.1: The Super Bowl—Conference Dominance?

Every year since 1967, the American and National Conference champions of the National Football League compete in a championship game know as the Super Bowl. Table 12.1.1 identifies by conference the winner of each of the 45 Super Bowls from I through XLVII.

The question is: What is it about the sequence of conference wins that has led most sports writers to believe that one conference or the other was superior at different extended periods in time?

Super Bowl					
Year	**Winning Conference**	**Year**	**Winning Conference**	**Year**	**Winning Conference**
1967	National	1982	National	1997	National
1968	National	1983	National	1998	American
1969	American	1984	American	1999	American
1970	American	1985	National	2000	National
1971	American	1986	National	2001	American
1972	National	1987	National	2002	American
1973	American	1988	National	2003	National
1974	American	1989	National	2004	American
1975	American	1990	National	2005	American
1976	American	1991	National	2006	American
1977	American	1992	National	2007	American
1978	National	1993	National	2008	National
1979	American	1994	National	2009	American
1980	American	1995	National	2010	National
1981	American	1996	National	2011	National

Table 12.1.1: Super Bowl Winners by Conference

Q1.   What do you notice about these results?

Q2.   Does one conference appear stronger than the other? Why or why not?

Q3.   If neither conference is superior to the other, what number would it make sense to use as the probability that the winner will come from a given conference?

Q4.   What was the overall percentage of wins by each conference? Each column represents a 15-year period. What was the winning percentage in each 15-year period?

In order to explore your conception of randomness, we will model the Super Bowl results since 1967. Assume that the two conferences are equally strong. If so, each conference winner has the same 50% chance of winning the Super Bowl in any year. We will use a coin flip to represent the 50% chance. First you will imagine and write down a typical sequence of heads and tails. Only afterwards, will you actually flip a coin.

Q5.   On a sheet of paper, list the numbers from 1 to 45 to represent each of the Super Bowls from 1967 to 2011. Table 12.1.2 shows how to set up the list before entering the Hs and Ts. You will use an H to represent an AFC win and a T for an NFC win. Now imagine a random sequence of wins assuming the two conferences are equally strong. Record next to each number, an H or a T.

1		16		31	
2		17		32	
3		18		33	
4		19		34	
5		20		35	
6		21		36	
7		22		37	
8		23		38	
9		24		39	
10		25		40	
11		26		41	
12		27		42	
13		28		43	
14		29		44	
15		30		45	
Number of Heads		Number of Heads		Number of Heads	
Percent of Heads		Percent of Heads		Percent of Heads	

Table 12.1.2: A format for recording simulated coin flips

After recording all forty three Hs and Ts, fill in the bottom rows of the table: the number of heads and percent in each column. Also calculate the percent of Hs for all forty-three imagined events.

Q6.   Compare the percent of heads in your three columns and the overall percent. How much difference is there among the column percents?

Q7.   Compare the percent of heads in your list with those of 2 or 3 other students sitting close by. Are your results similar, or is there a lot of difference among them?

Next, make another list just like the one in Table 12.1.2 to record the results of your actual coin flips. After recording all forty three Hs and Ts, fill in the bottom rows of the table: the number of heads and percent in each column. Also calculate the percent of Hs for all forty-three flips.

Q8.   Compare the percent of heads in your three columns and the overall percent. How much difference is there among the column percents?

Q9.   Compare the percent of heads in your list with those of 2 or 3 other students sitting close by. Are your results similar, or is there a lot of difference among them?

In the following questions, you will compare your imagined sequence with the sequence of the actual coin flips. Be sure to label your lists, so that you'll know which is which.

Q10.   Was there more variability in the column percents among the imagined lists or the coin flip lists?

Q11.   What were the maximum and minimum percents for you and your neighbors for each column for the imagined lists and coin flip lists? Are you surprised at how far these numbers are from 50% in the coin flip lists?

Q12.   What were the maximum and minimum percents of Hs and Ts for you and your neighbors for all forty-three imagined events and coin flip? What differences do you notice about the range?

Q13.   For the entire class, what was the minimum percent of Hs observed when flipping the coins forty-three times? What was the maximum? Are you surprised at how far these numbers are from 50%?

Q14.   Did anyone in the class, have a percentage that low or that high in the imagined list?

Q15.   For the entire class, what was the overall percent of Hs and Ts for the coin flip random experiment?

We will now return to the actual data for the Super Bowl. Find the total the number of times in each of the three columns that the National Conference won and calculate the percent.

Q16.   Look at the percentage of National Conference wins in the first column of the actual Super Bowl data. In your coin flip table, was there any percentage of Hs that low? Did anyone in the entire class have a percentage of Hs that low in a column of coin flips?

Q17.   Look at the percentage of National Conference wins in the second column of the actual Super Bowl data. In your coin flip table, was there any percentage of Hs that high? Did anyone in the entire class have a percentage of Hs that high in a column of coin flips?

Q18.   Based on your comparisons between your actual coin flips and the Super bowl wins do the percentages support the news writers' claims of dominance in each of the fifteen year periods?

Up to this point we have explored different aspects of the percent of Hs and Ts. We have compared the imagined sequence with that generated by actual coin flips. We have also compared minimum and maximum percents for columns and for the total. As you looked across a larger and larger set of data, you should have seen a wider range of percents for the column percents and less variability for the percent of the total. The total percent for the entire class would likely be close to 50%.

Now we are going to look at the sequences of consecutives Hs or Ts. These sequences are called strings. Based on numerous class room experiments, it has been found that lists of imagined sequences of Hs and Ts tend to have shorter strings and rarely have any long strings of four more. We are going to see if your experience and those of your classmates replicate the research.

Record the lengths of every string of consecutive heads or tails in your imagined list. Then do the same for your coin flip list. For example, Table 12.1.3 records a sequence of H's and T's as well as the length of each string.

T	2	T	1	F	1
T		F	1	T	1
F	1	T	1	F	2
T	1	F	1	F	
F	2	T	2	T	1
F		T		F	3
T	1	F	2	F	
F	3	F		F	
F		T	1	T	2
F		F	2	T	
T	3	F		F	1
T		T	2	T	4
T		T		T	
F	2	F	1	T	
F		T	1	T	

Table 12.1.3: A list of H's and T's with the length of each string

Q19.   Compare the lengths of the strings in your two lists. Which list has the longest string? How long is that string?

Q20.   Compare your answers with those of two or three neighbors. In your group, what is the length of the longest string? Which list did it come from?

Q21.   Compare your answers with the entire class. What is the length of the longest string? Which list did it come from?

Q22.   In the Super Bowl results, what was the longest string of wins for one conference? Did any of the flip coin lists have a string this long?

Q23.   Based on your comparisons do the data support the news writers' claims of dominance?

Q24.   Have you discovered any misconceptions in your own thinking about randomness?

12.1.1 Percentage, Long Strings, and Probabilistic Thinking

Every alternative sequence of 45 heads and tails has the exact same probability of occurrence, one in 8.8 trillion. Thus the likelihood of seeing an exact sequence of alternating heads and tails has the same as the probability of seeing a sequence of all heads or all tails. However, the likelihood of producing exactly 22 heads out of the total of 45 is orders of magnitude higher than the chance of 45 heads in a row. That is due to the fact that there are many sequences of heads and tails that result in exactly 21 heads but there is only one way to obtain 45 heads in a row. Thus, 45 heads in a row might lead you to believe that the coin is two-headed.

This same principle points to an interesting phenomenon. When random flips of a coin are carried out, the sequence often includes multiple examples of long runs of four or more heads or four or more tails. However, individuals recording their own made up sequence have a tendency to believe that randomness means frequently alternating back and forth between heads and tails and therefore tend to create lists that have few runs of heads or tails. As a result an experienced probabilistic thinker can often identify which of two lists was generated by flipping a coin and which was generated by a human mind with a tendency to look at the list as it develops and work to balance the sequence of heads and tails.

In addition although runs of heads or tails are likely, the probability in a sequence of 45 flips there will be a random run of 13 or more heads in a row is small. Thus the sequence of 13 national conference wins in a row although possibly the result of pure random flips of 50-50, is more likely to support the claims of sports writers. During this period, the National conference was the better conference and the chance of it winning the super bowl during much of the time period was much higher than 50-50.

One of the primary goals of probabilistic analysis is to attempt to assess whether a pattern is simply random or if some factor is contributing to the pattern. It is never possible to resolve this issue with certainty but an understanding of probability theory can help decide which driving force is most likely. This concept of pattern differentiation is a critical element of statistical process controls in manufacturing. The workers track and chart the sequence of parts that are within and outside acceptable specifications in order to determine when and if a manufacturing process has become misaligned, out of control, and requires corrective action. An analogous situation arises when a credit card company tracks an individual's purchases and attempts quickly to identify any purchase that looks extremely suspicious, out of the range of the wide variability of purchases of the individual. When such a purchase falls outside some pattern of location, magnitude, type of purchase and/or frequency, the credit card company considers marking it as possible fraud. It might then attempt to contact the card holder for clarification or in extreme cases cancel the card on the spot.

12.1.2 Next Steps

On the pages that follow, you will be introduced to a wide range of real world problems that operations researchers solve every day. They approach these real world problems using an array of techniques that apply principles of probability and statistics. You will probably already be familiar with many of these concepts, but some of them might be new. We believe that a major difference between our use of probability and statistics in this textbook and what you may have seen in the past is that we will use probability and statistics in the context of making decisions.

Along the way, you will also learn about the role of computers in solving real world problems by actually using computers. Using a computer allows us to simulate a situation rather than attempt to collect large amounts of data. In some cases, it will also allow us to realistically model the behavior of a system in order to learn more about the system. When we use mathematics in this way, more as a modeling tool and less as a computational tool, your role, much like that of an OR professional, becomes to formulate the model and interpret the results that the model returns.

Studying mathematics in this way may require you to develop a new MINDSET about what mathematics is, how it can be used, and the best ways to learn it. Whatever mathematics is, it is certainly not a spectator sport. That is why a large portion of this book looks very similar to the Super Bowl simulation problem that you just read. Many questions were asked, but few were answered in the text. Where will the answers come from? You will provide those answers with the help of your classmates and teacher. In doing so, you will also discover that applying mathematics to solve real-world problems can be a relevant, creative, and exciting team sport. We hope that all of this will help you begin to form that new MINDSET.

Section 12.2: Managing and Deciding in an Uncertain World

It is not possible to make 100% accurate predictions as to who will win the Super Bowl, how many people will show up for work tomorrow, how many student newspaper writers will complete their assignments on time, how many cars a dealership will sell, or how long it is going to take get through the cafeteria line. However, probability theory and complex probability models enable us to identify individual occurrences that are more or less likely, as well as estimate long-term averages. This knowledge and understanding is essential for planning and making critical decisions in advance and adjustments on the spot. A plant or office manager must decide how many spare workers to have on-duty to address a problem of absenteeism. The newspaper editor will also have to plan what to do if only eight of ten writers have completed their assignments. Similarly, an automobile dealership needs to know how much inventory to carry. Often we must balance the added costs of having reserves against the impact of running short.

12.2.1 Are Fluctuations Around the Mean Random or the Result of Problem?

One challenge managers and we personally face in a variety of situations involves understanding and interpreting fluctuations around a long-term average. Is a rise in absenteeism or lateness simply a reflection of a random fluctuation or an indication of a developing problem that needs to be addressed? Biostatisticians regularly are on the lookout for occurrences of disease that are out of the "normal" range. A report comes in that a small city has twice the national average rate of a specific type of cancer. Given the tens of thousands of communities, is it likely or unlikely purely by random fluctuation to see one or more cities with a rate twice the national average. If it is likely, no action may be taken, no detailed study initiated. If however, probability theory suggests this high rate for a small city is extremely rare, public health officials may commission an expensive study to gather more data to try to identify the contributing factors.

You will explore the pattern of which conference won the Super Bowl in an attempt to draw a conclusion about whether the data suggest that one conference is superior to the other. You will also look at data for a customer service telephone line to determine whether there is sufficient evidence that the service is not meeting the timeliness standard established by management. You will assist the faculty advisor to the school newspaper to better understand why they are having problems publishing the paper on time. You will also explore the cost of being late and how to reduce that cost. Finally, you will assist a plant manager to decide how many spare workers to have on duty so as to maintain productivity to cope with random fluctuations in the number of workers who are absent from work each day.

12.2.2 Advance Planning

Imagine you and your extended family of 50 people have planned a major outdoor dinner barbecue? Do you have a plan for rain? What if a week in advance you are told there is a 5% chance of rain? What if it is 50% or 80%? What options do you have if you start planning a week in advance that you may not have the morning of the event? What communication challenges do you face? How can you plan for them to ensure as much as possible that everyone knows what to do? The situation a manager faces is similar, when he is trying to schedule work to be completed and he is not sure how many workers will be absent on any given day or how many of them will complete their assigned tasks on time. The more he understands about random phenomena, the better able he is to plan for the future.

12.2.3 A Simple Uncertainty Model – Two Possible Random Outcomes

The simplest kind of uncertainty to model is one that involves only two possible equally likely outcomes, such as a coin flip. There are many other uncertain situations with only two possible random outcomes: a worker or student is present or absent, an assignment is handed in on time or not, or a team wins or loses a game. However, in these cases the two possible outcomes are not necessarily equally likely. In these cases, the critical number, often called a parameter, is the probability of the outcome occurring. It would be 0.1 if there is a ten percent chance a worker will be absent on any given day. It might be 0.95, if a student is extremely reliable in turning in his or her assignments on time. In both cases, the two outcomes are both **mutually exclusive** and **complementary**. They are mutually exclusive, because both outcomes cannot happen at the same time. A flip of a coin is either heads or tails but not both. They are complementary, because if one does not happen, the other must. If the coin is not a head, then it must be a tail. If a worker shows up for work, then he is not absent and vice versa. Complementary events are two events that are also **collectively exhaustive**. The two complementary events exhaust all of the possible outcomes.

This simple situation forms the basis for all of the examples in this chapter and the next. We expand on this two-alternative context to study what happens when the same situation is repeated three, five, ten, twenty, or more times. Here we are interested in the total number of times one outcome or the other occurs. In the worker situation, we are interested in how many workers are absent each day due to the potential impact on productivity.

The starting point in modeling these uncertain situations is to determine the probability, p, of a specific occurrence which we arbitrarily call a success. For a coin that is fair, we assume there it is a 0.5 probability of a Head and a 0.5 probability of a Tail. For the worker case, the manager would look at historical data and determine the proportion of days missed by the workers in his organization. The challenge is to use this value to understand what happens when the same situation is repeated multiple times. For example, how many heads will occur if we flip a coin 12 times? It is extremely unlikely if we repeat the 12 coin flips two, three, or four times that we will see the same exact sequence of heads and tails. However, in this chapter and the next we will explore what types of predictions we can make, and with what degree of likelihood.

Anytime we gather data about an uncertain situation, we have information about how things were. What we do not know is how likely this situation is to be repeated. What happens if we change something, what will the random patterns look like then? For example, our small office of ten workers may have extensive data on the random number of workers who are absent from work each day during flu season. How would we predict from this information what the situation is likely to be if we expanded to fourteen workers or cut back to seven? A mathematical probability model can assist us in this analysis.

12.2.4 Simulation

The simplest and most flexible method of modeling probability is to **simulate** the situation. For example, if the probability of reaching a customer service representative right after opening time were 50%, we could let "heads" represent reaching customer service and "tails" represent not reaching customer service. We can simulate a work-week by flipping the coin five times and recording the results. Then we can count the number of times in the "week" that customer service was open on time. In this way, by flipping a coin, we can get a sense of what the results of actually calling might have been.

A 50-50 probability can be simulated by flipping a coin, but what about a random event in which the chance of success is 0.7 and the chance of failure is 0.3? In this chapter we will use the random number

generator found on your calculator and also in EXCEL to simulate these types of situations. To simulate a 50-50 probability, we can generate any two random integers to represent the two equally likely outcomes. Whenever the two events are not equally likely, we need to be more creative. To simulate a 0.7 chance of success, we will generate random integers between 1 and 10. If the number is 7 or less, we record a success and if the number is 8 or more, we record a failure. To simulate a 0.95 chance, we will generate random integers between 1 and 100. If the probability is recorded to three decimal places, we will use the integers from 1 to 1000. With this simple modeling tool we will be able explore, study and come to understand two-outcome random phenomenon with different values of p, the probability of success.

Section 12.3: Customer Service at Koala Foods

Like most companies, Koala Foods, a wholesale distributor of foods to supermarkets, is concerned about its customer service. Koala Foods assures, the supermarkets it serves that their needs can be addressed on weekdays beginning at 8:00 a.m. by calling a toll-free customer service hotline. The president of Koala Foods wants to increase customer satisfaction, so he hires a company that will use probability techniques to monitor the customer service department. Specifically, the monitoring company will initially check that the customer service department is actually operational by 8:00 a.m. daily. Later the monitoring company will track individual calls to assess the quality of service.

12.3.1 Forming a Strategy

AGB Company, a customer service consulting firm, has been contracted by Koala Foods to monitor its customer service department. There are various ways of measuring quality in this situation. Mr. Smith, the President of Koala Foods, proposes checking to see if the customer service department is operational by 8:02 a.m., at 8:05 a.m., and lastly at 8:10 a.m. He would be satisfied if the customer service department was ready and answering calls by 8:02 at least 50% of the time. He sets a higher standard of 70% for 8:05 a.m. He feels that by 8:10 a.m. the service should almost always be open and sets a standard of 95%. This is equivalent to the service being ready to answer calls at 8:10 a.m. 19 days out of 20.

The first three days AGB Company calls the Koala Foods customer service department at 8:02 a.m. Each day they encounter an "hours of service" recording rather than a customer service representative.

Q1.   Does Koala Foods have a problem?

Q2.   What are some possible reasons a call would not be answered at 8:02 a.m.?

Q3.   Is this enough information to decide that there is a problem in meeting the 50% standard?

12.3.2 Simulating the Situation

We are interested in better understanding the significance of failing to reach customer service three days in a row. We need to decide whether this is sufficient evidence of a problem. If so, Koala Foods management will have to introduce greater supervision to ensure the standard is met. The alternative is to continue calling each morning to collect more data to see if there truly is a problem.

In order to gain the necessary understanding of this sequence of three consecutive failures, AGB Company decides to simulate three consecutive calls a large number of times. They want to know how likely is it to see three failures in a row if the 50% standard was being met overall. You and your classmates will assume the role of AGB and simulate this 50-50 random experiment by flipping a coin. Let the "heads" side of the coin represent a call that receives a recorded message and the "tails" side of the coin represent a call that gets through to a customer service representative.

Q4.   Do you think that your first three consecutive flips will all be heads? If you repeat this process a total of ten times, do you think you will ever see three heads in a row?

Q5.   Do you think that anybody in the class will ever get all three heads?

Q6.   Flip the coin three times and record the results. Repeat this process nine more times. How many times did you observe a series of three heads?

Q7.   Did any of your classmates flip a series of three heads? Count the total number of times that you
        and your classmates observed three heads in a row. What is the proportion of sets of three flips
        that were three heads in a row?

Q8.   Think back to Koala Foods Customer Service Department and meeting the 50% standard. Does it
        seem unusual that AGB failed to reach customer service at 8:02 a.m. three days in a row if the
        50% standard was being met?

Another method of simulation is the random number generator in your graphing calculator. It is capable
of simulating three repetitions of this random event with one single command. It is also very easy to
repeat this process as many times as you want. Within the MATH menu, the PRB submenu contains the
random integer command, randInt(. If you enter randInt(0,1,3), your calculator will return a set of three
random integers that are either "0" or "1". Each time you push ENTER, your calculator will generate and
display a new set of three values. Let "0" represent a call that receives a recorded message and "1"
represent a call that gets through to a customer service representative.

Q9.   Enter randInt(0,1,3) into your calculator ten times and count the number of times the string {0 0
        0} occurs.

Q10.   Compare your answer to #9 with those of your classmates.

Q11.   Would you reconsider your answer to #4 and #5?

Q12.   Should you reconsider your response to #1?

After seeing the results of the simulation, Mr. Smith decides the data presented by AGB Company was
inconclusive. He now knows that he would need data for more than three days to conclude with
confidence that the standard of 50% is not being met.

12.3.3 Dealing with a Higher Standard

Recall that the standard for reaching customer service at 8:05 a.m. was 70%.

Q13.   Does the 70% standard for calls placed at 8:05 a.m. seem like a reasonable expectation if you are:
        a.   a manager at Koala Foods?
        b.   a customer service representative at Koala Foods?
        c.   a customer of Koala Foods who needs service?

AGB also called at 8:05 a.m. and three days in a row no one answered. Mr. Smith wonders if there is any
reason why these three days of data might be conclusive about meeting this higher standard. He decides to
ask AGB to conduct another simulation. Once again, you and your classmates will play the role of AGB.

Q14.   Will a coin be adequate to simulate this scenario? Why or why not?

To use the calculator, Enter randInt(1,10,3). This will generate three random integers between 1 and 10
inclusive. Assign the numbers "1" through "7" as calls that are answered by customer service. The
numbers "8" through "10" represent calls that receive an answering machine.

Q15.   Why does this method accurately model the 70% standard?

Q16.   You and your classmates will perform this simulation ten times each. Do you think there will be more, the same, or fewer simulated instances of receiving the recorded three times in a row as compared to the first simulation?

Execute this simulation on your calculator ten times. For each set of three, count and record the number of times a number between "8" and "10" occurs.

Q17.   Did you have any lists of three numbers in which all three numbers were 8, 9, or 10?

Q18.   Did any of your classmates observe a set of three numbers that were all 8, 9, or 10? Count the total number of times in the class that occurred. What is the proportion of total simulations did that occur?

Q19.   Think back to Koala Foods Customer Service Department and meeting the 70% standard. Does it seem unusual that AGB failed to reach customer service at 8:05 a.m. three days in a row if the 70% standard was being met?

12.3.4 Dealing with an Even Higher Standard

Recall that the standard for reaching customer service at 8:10 a.m. was much higher. It was set at 95%. AGB also called at 8:10 a.m. This time the results were different. One out of the three times, they reached customer service. The other two times, they received a recorded message. Once again Mr. Smith wonders if these results indicate conclusively that the standard of 95% is not being met. He decides to ask AGB to conduct another simulation. Once again, you and your classmates will play the role of AGB.

Q20.   Determine the beginning and the end of an appropriate random integer interval that could be used to simulate this situation. Assume again a three-day monitoring period.
          randInt(_____,_____,_____)

Q21.   What range of integers represents a call that is answered by a customer service representative?

Q22.   What range of integers represents a call that receives a recording?

As you run your simulation and observe three numbers, you will need to record how many times a number representing an answered call appears. Run your simulation ten times.

Q23.   For your ten simulations, how many times were 0, 1, 2, and 3 calls answered by customer service?

Q24.   In your class, how many times were 0, 1, 2, and 3 calls answered by customer service?

Mr. Smith wonders whether only one out of three answered calls is sufficient evidence to be concerned. If so none out of three answered calls is even stronger evidence.

Q25.   Did any member of the class observe either zero or one calls being answered? How many times did this happen in the whole class? What proportion is this?

Q26.   With regard to the 95% standard being met, what do you think Mr. Smith should conclude from AGB's three phone calls in which only one was answered at 8:10 a.m.?

Section 12.4: Getting *The Lancer* to Press

Each month Al Mitchell, the faculty advisor for a school newspaper, oversees the production of the newspaper, *The Lancer*. He has 10 student writers, five of whom are editors and five are staff writers. In order to get to press on time, it is necessary that Al's students finish their articles by the required deadline. That is, he needs each and every student to meet his or her deadline.

12.4.1 The Problem

Currently, Al finds that each of his student writers meets his or her deadline only 95% of the time. On the surface, this does not seem bad, but *The Lancer* frequently goes to press late. He wants to change that. Al recalls something from his high school probability class called the **multiplication principle**: The probability that a set of **independent** events will occur can be found by multiplying the probabilities of each event occurring. Formally stated, the multiplication principle says that:

> If *X* and *Y* are *independent* events, then the probability of both *X* and *Y* occurring is equal to the product of the probabilities of *X* occurring and *Y* occurring:
> $$P(X \text{ and } Y) = P(X) \cdot P(Y)$$

Al wonders what independent events would mean in this context. Since one particular writer meeting a deadline does not affect the likelihood of any other writer meeting the deadline, the events that writers X and Y meet the deadline are *independent*. Therefore, the multiplication principle can be used. If each of his 10 writers meets the deadline 95% of the time, the probability that all of them meet the deadline is:
$$(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95) = 0.95^{10} \approx 0.599$$

Now, for the paper to go to press on time, every one of the ten writers must meet the deadline. If even one of the ten does not meet the deadline, the printing of the paper is delayed. This calculation shows that *The Lancer* would go to press on time approximately 60% of the time! No wonder it has been going to press late so often. Al wants to change this situation to ensure the paper goes to press on time 90% of the time.

Each writer meeting deadlines 95% of the time seems to be a reasonable expectation. However, having each of the10 students meet the deadlines 95% of the time results in a big problem for *The Lancer*. To improve this situation, Al believes he can get his editors to meet the deadlines 99% of the time. They have the most experience and are likely to work better under pressure compared to the staff writers who have just joined *The Lancer*'s staff.

Al's idea is to have his five editors meet the deadlines 99% of the time, and the less experienced staff writers can continue to meet the deadlines 95% of the time. Using the multiplication principle, the probability of going to press on time would then be:
$$(0.99)(0.99)(0.99)(0.99)(0.99)(0.95)(0.95)(0.95)(0.95)(0.95) = (0.99)^5(0.95)^5 \approx 0.736$$

While going to press on time about 74% of the time is far better than the 60% it was previously, he would still not achieve his 90% goal.

Al goes back to the drawing board and believes that he can push his editors to be on time 100% of the time. He feels the editors can handle that level of responsibility. If the five editors meet the deadlines 100% of the time and the staff writers meet the deadlines 95% of the time, the probability of going to press on time would be:
$$(1)(1)(1)(1)(1)(0.95)(0.95)(0.95)(0.95)(0.95) = (1)^5(0.95)^5 \approx 0.775$$

Getting the editors to meet their deadlines 100% of the time barely made a difference!

Al sees that the only way to get *The Lancer* to press on time is to push also the staff writers to meet their deadlines at a higher rate. If the editors meet the deadlines 100% of the time, what must the probability be for the staff writers in order for *The Lancer* to go to press on time 90% of the time?

Q1.   If each staff writer meets her/his deadline 96% of the time, would The Lancer go to press on time 90% of the time? If not, what if the staff writers meet their deadlines 97% of the time?

Al begins to wonder whether it is reasonable to have a goal of 90% for *The Lancer* to go to press on time. If the editors meet their deadlines 100% of the time, Al wonders what standard should be required of each staff writer. We can take an algebraic approach to determine this standard.

Let *x* represent the standard for each staff writer meeting the deadline. The probability of the paper going to press on time at least 90% of the time can be represented by the following inequality. Then we need to solve the inequality for *x*:

$$1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot x \cdot x \cdot x \cdot x \cdot x \geq 0.90$$

$$1^5 \cdot x^5 \geq 0.90$$

$$x^5 \geq 0.90$$

$$x \geq \sqrt[5]{0.90}$$

Now, taking the fifth-root of a number is equivalent to raising the number to the $\dfrac{1}{5}$ power, and this can be done with a calculator:

$$x \geq \left(0.90\right)^{\frac{1}{5}}$$

$$x \geq 0.979$$

The solution shows us that for *The Lancer* to go to press on time 90% of the time, each of the five staff writers would have to make deadline almost 98% of the time! Al wonders if this is a realistic goal. In the next chapter, Al will evaluate a change in policy. He will consider going to press on time if one or two writers are late. He will simply just leave their columns out of the paper.

12.4.2 Complementary Events

Let's consider the timely publication of *The Lancer* from a slightly different perspective. Instead of focusing on how frequently *The Lancer* goes to press on time, let's consider how often it goes to press late.

Going back to the original case, the editors and staff writers originally each met the individual deadlines 95% of the time. This led to *The Lancer* going to press on time roughly 60% of the time.

Q2.   That being the case, how often did *The Lancer* go to press late?

If they were on time 60% of the time, it makes sense to say that they were late 40% of the time. Mathematically, this happens because going to press on time and going to press late are **complementary events**. Complementary events are both **mutually exclusive** and **collectively exhaustive**. If two events are mutually exclusive, that means they cannot occur at the same time (e.g., when rolling a die, the event

of getting a three and the event of getting a five are mutually exclusive, but getting a 3 and getting an odd number are not). Two events are collectively exhaustive if together they are the only possible outcomes. This implies that one of them must occur (e.g., when rolling a die, the event of rolling an odd number and the event of rolling an even number are collectively exhaustive, but rolling an odd number and rolling a 2 are not).

Q3.   Keeping with our die-rolling context, list three pairs of complementary events.

When we consider publishing a student newspaper, one pair of complementary events is publishing on time or not publishing on time (i.e., publishing late). Because an event and its complement include all the possible outcomes of an event, the probability that one or the other of them will occur is equal to one (i.e., one of them is certain to occur). Now, because they cannot both occur (why not?), you can find the probability of an event's complement by subtracting the probability of the original event from one. This is significant because calculating the probability of the complement of an event is sometimes easier than finding the probability of the event directly. Thus, in the case of *The Lancer*,

> P(publishing on time) + P(not publishing on time) = 1
> $\therefore$ P(not publishing on time) = 1 – P(publishing on time).

You might think that all of the writers on time and all of the writers not on time are complementary events. All of the writers on time corresponds directly to the event that all ten writers meet their individual deadlines. The complementary event that the paper is late is not equivalent to *all* writers being late. It only takes *one* writer being late for publication of *The Lancer* to be delayed. Thus the event the paper is late includes all of the following possibilities: 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 writers being late. Calculating directly the probability of the paper being late requires calculating the probability of each of these one through ten possibilities. The calculation of each of these is in itself complex and requires the advanced concepts discussed in the next chapter.

12.4.3 If Late, How Late?

Al recognizes and understands the challenge he faces publishing *The Lancer* in a timely fashion. He is also concerned that if the paper is going to be late, how frequently it will be delayed one, two, or more than two days. In looking back over the past year, Al notices that when one or more of the editors do not meet the deadline, printing is delayed just one day. However, when a staff writer is late, tracking him down to fulfill his commitment delays the paper two days. In the worst case scenario, when both an editor and a staff writer are late, the whole system gets messed up and printing is delayed more than two days while sorting out the problem. For the analysis that follows, let's assume that Al has worked with the editors and they have improved their performance. Each editor is meeting the deadline 99% of the time. The staff writers are each still meeting the deadline only 95% of the time.

We will use a capital letter such as E to represent the occurrence of an event. The complement of an event will include a superscript c as in E^c.

Define:
>   E= the event all of the *editors* meet the deadline;
> then E^c = the complementary event that not all of the *editors* meet the deadline.

Remember that "not all of the *editors* meeting the deadline" involves many possibilities.

Calculate P(E) = $.99^5$ = .951

Using the principle of the complement that $P(E) + P(E^c) = 1$

   $P(E^c) = 1 - .951 = .049$

So the probability that not all of the *editors* meet the deadline is about 5%.

Similarly define
   S= the event that all of the *staff writers* meet the deadline;
   then S^c = the complementary event that not all of the *staff writers* meet the deadline.

Calculate $P(S) = .95^5 = .774$

Using the principle of the complement that $P(S) + P(S^c) = 1$

   $P(S^c) = 1 - .774 = .226$

So the probability that not all of the *staff writers* meet the deadline is about 23%.

You now have enough information to calculate the probability that the newspaper will be on time, one day late, two days late and more than two days late.

Using the multiplication rule, what is the probability of E and S occurring? This is the probability that *The Lancer* goes to press on time.
   $P (E \text{ and } S) = (.99)^5 (.95)^5 = (.951)(.774) = .736$

So the probability *The Lancer* will go to press on time is 74%.

Now let's calculate the probability The Lancer will be one day late. This occurs if *not all* of the editors meet the deadline (E^c) and *all* of the staff writers do meet the deadline (S). We combine the principle of the complement and the multiplication rule to determine the probability of this happening.

   $P(E^c \text{ and } S) = P(E^c) P(S) = (1-.951)(.774) = .038$

Thus, *The Lancer* will go to press *one* day late almost 4% of the time.

Similarly, we can calculate the probability that the paper will be two days late. This occurs if *all* of the editors meet the deadline (E) and *not all* of the staff writers do meet the deadline (S^c).

   $P(E \text{ and } S^c) = P(E) P(S^c) = (.951) (1-.774) = .215$

Thus, *The Lancer* will go to press **two** days late almost 21.5% of the time.

Lastly, we can calculate the probability that the paper will be more than two days late. This occurs if *not all* of the editors meet the deadline (E^c) and *not all* of the staff writers do meet the deadline (S^c).

   $P(E^c \text{ and } S^c) = P(E^c) P(S^c) = (1-.951) (1-.774) = .011$

Thus, *The Lancer* will go to press *more than two* days late 1% of the time.

Let's summarize all of these probabilities:

        P(on time) = .736
        P(one day late) = .038
        P(two days late) = .215
        P(more than two days late) = .011

These probabilities sum to one since these four events represent all of the possible outcomes.

12.4.4 Cost of Being Late

The printing company Al uses imposes an added charge if the paper goes to press late. There is a $50 charge for one day late, a $120 total charge for two days late and $300 for more than two days late. However, if Al gives the company an *accurate* statement of the number of days late, the printer can reschedule his other work. In that case he cuts the fee in half. Al needs to include in his budgets the cost of paying late fees. To do so sensibly, he needs to calculate the average late fee. He incurs no late fee when paper is published on time. There is a $50 late fee for one day late, a $120 late fee for two days late, and a $300 late fee for more than two days late. Al realizes he cannot simply average these four numbers. This would not account for the different likelihoods. To get an accurate prediction he must weight the late fee by the probability of its occurrence. This is called the expected value of the late fee. The expected value of a random variable is determined by taking the weighted sum of every possible outcome multiplied by its probability.

Let L = late fee
        $E(L) = 0\ P(L=0) + 50\ P(L=1) + 120\ P(L=2) + 300\ P(L=3)$

        $E(L) = 0\ (.736) + 50(.038) + 120\ (.215) + 300\ (.011) = \$31$

Thus Al must budget an average of $31 per issue of the school paper to account for late fees.

The student editor-in-chief has left a message on Al's voice mail that the paper is going to be late but without any specifics. Recall that the publisher has offered to reduce the late fees, if Al lets him know exactly how late the paper will be. Al needs to call the printer immediately. He wonders how many days late he should tell the printer. Looking at the respective probabilities for being late, it is clear to Al that the paper will most likely be two days late.

Since Al saves money only if he is correct, he wonders how probable that is? What information does Al have? He knows the paper is going to be late but he is not sure how late. At first, he thinks the probability is 21.5%. However, Al realizes this percentage relates all of the times the paper is published, whether late or on time. He knows that cannot be right, because he knows the paper is going to be late. Al realizes that what he needs to calculate is the proportion of times the paper is exactly two days late out of the total proportion of times the paper is late. He knows the probability of the paper going to press on time is 0.736, the chance of it being late is just (1-0.736) or 0.264.

Therefore, , the proportion of times it is exactly two days late out of the total proportion of the time that the paper is late is:

$$\frac{0.215}{0.264} = 0.814$$

Thus, if the paper is going to be late, there is a better than 81% likelihood that he will accurately predict the number of days to be two. This will save $60 out of the $120 when he is correct.

To check this result, Al decides to calculate the proportions for being one day and more than two days late.

Of that proportion, the proportion of time it is one day late is:

$$\frac{0.038}{0.264} = 0.144$$

The proportion of time it is more than two days late is:

$$\frac{0.011}{0.264} = 0.042$$

Notice again that .144 +. 814 +.042 = 1.

This should add up to one since when the paper is late there are only these three possibilities and one of them must occur.

Section 12.5: Worker Absenteeism at BT Auto Industries

Unscheduled absenteeism is the bane of any manager of a manufacturing or service facility. When a worker is absent either some work does not get done, or other workers have to increase their productivity, or substitute workers must be called in.

12.5.1 The Problem

According to The Centers for Disease Control and Prevention (CDC), the flu season usually ranges from November through March. Historically, it is a ten week period from mid-January through mid-March that sees the most flu outbreaks. The CDC recommends that anyone who contracts the flu stay home to prevent spreading it to others.

BT Auto Industries is a small manufacturer based in Detroit, Michigan with twelve workers. During flu season, they experience a larger than usual number of workers who miss work due to illness. In particular, an average of one out of every ten workers is absent every day during the flu season! In order to maintain their level of productivity, BT Auto Industries hires spare workers to cover the shifts of the employees who have called in sick. Due to tight budget constraints this year, the management wants to hire just enough spare workers to ensure the probability that 95% of the time the absent workers are covered. Anything less and productivity is impacted.

12.5.2 Simulating the 10-Week 50 Workday Flu Season

In order investigate the number of spare workers to hire on standby, the management runs a computer simulation of the number of workers who show up each day. The simulation will use random numbers to replicate each worker's attendance during the 10-week peak flu season. They will need to simulate the attendance of each of the twelve workers on each of 50 workdays. An analysis of attendance records for the past five years indicates that an average of one in ten workers is absent every day during these ten weeks. They will, therefore, use a 10% absenteeism rate of each of the 12 workers. They assigned a value of 1 to mean that the worker was absent on a given day. An integer between 2 and 10 represented a worker who was present that day. Table 12.5.1 illustrates the simulation of three days. Each of the random numbers takes on a value between 1 and 10. Table 12.5.2 illustrates how these numbers are then translated into whether or not a worker is absent or present. On day 1, the number 1 appears under worker A. There was each a total of 1 absence that day. On day 2, the number 1 appears under worker C, thus there was one worker absent that day. On day 3, the simulation generated the number 1 under workers E, I, and L. There were 3 workers on day 3.

	Workers												Total
Day	A	B	C	D	E	F	G	H	I	J	K	L	per day
1	1	6	6	7	2	8	8	4	9	6	9	9	1
2	3	8	1	4	7	9	9	8	8	10	3	10	1
3	6	10	7	6	1	4	2	5	1	3	7	1	3

Table 12.5.1: Simulation of first three days.

	Individual Worker												Total
Day	A	B	C	D	E	F	G	H	I	J	K	L	per day
1	A	P	P	P	P	P	P	P	P	P	P	P	1
2	P	P	A	P	P	P	P	P	P	P	P	P	1
3	P	P	P	P	A	P	P	P	A	P	P	A	3

Table 12.5.2: Translate number into Absent (A) or Present (P)

Table 12.5.3 summarizes the frequency distribution of the number of absent workers for the fifty day simulation. Table 12.5.4 provides the detailed simulated data for each worker each day for 50 days. It also includes daily totals and individual worker totals for the simulated 10-week period of fifty days. Use the information in the Table 12.5.3 and 12.5.4 to answer the following questions.

Absent Workers	Frequency	Relative Frequency
0	13	26.0%
1	22	44.0%
2	11	22.0%
3	3	6.0%
4	1	2.0%
5	0	0.0%

Table 12.5.3: Frequency Distribution of the Simulated Number of Absent Workers

Q1.   For what percentage of the 50 simulated days would one spare worker be enough?

Q2.   For what percentage of the 50 simulated days would two spare workers be enough?

Q3.   If the goal is to have enough spare workers 95% of the time, based on this simulation how many spare workers are needed?

Q4.   If the goal is to have enough spare workers 99% of the time, based on this simulation how many spare workers are needed?

Q5.   Can you explain why there are gaps in the actual percentages in questions 1-4?

Q6.   Based on the Table 12.5.4, what is each worker's actual rate of absenteeism? What is the minimum percentage rate and what is the maximum?

Q7.   Why is each worker's actual rate not equal to 10%?

Q8.   What is the overall average absenteeism rate for the 12 workers? Why is this rate closer to the one in ten average as compared to many of the individual workers?

Q9.   In setting up the simulation, what assumption have we made about each worker's rate of absenteeism? Do you think the assumption is valid?

Day	A	B	C	D	E	F	G	H	I	J	K	L	Total per day
							Workers						
1	1	6	6	7	2	8	8	4	9	6	9	9	**1**
2	3	8	1	4	7	9	9	8	8	10	3	10	**1**
3	6	10	7	6	1	4	2	5	1	3	7	1	**3**
4	10	5	3	7	9	8	9	6	9	9	7	1	**1**
5	9	10	8	2	5	4	6	5	9	10	3	5	**0**
6	10	7	5	1	2	2	1	4	6	6	10	3	**2**
7	8	4	9	4	8	5	3	7	10	10	5	9	**0**
8	9	5	6	8	2	8	4	10	9	10	7	7	**0**
9	1	8	2	9	2	3	2	4	6	10	8	7	**1**
10	6	4	7	4	10	9	2	4	7	10	6	5	**0**
11	6	4	2	6	7	2	3	5	9	10	4	3	**0**
12	6	1	8	4	9	6	10	7	9	7	9	8	**1**
13	9	7	3	9	8	5	9	5	6	6	3	10	**0**
14	3	8	4	6	7	8	10	9	1	5	1	8	**2**
15	4	9	5	2	7	5	5	5	8	5	1	5	**1**
16	5	8	2	10	9	3	7	3	6	2	6	1	**1**
17	7	10	10	4	4	3	6	3	7	8	7	9	**0**
18	6	7	9	2	2	6	8	8	5	9	7	2	**0**
19	9	9	8	3	10	7	7	10	9	1	3	4	**1**
20	3	9	10	6	7	4	10	8	9	3	9	8	**0**
21	8	5	8	7	6	1	4	4	8	2	9	7	**1**
22	9	4	7	6	8	9	2	10	10	7	8	7	**0**
23	10	2	7	1	4	9	8	1	5	4	9	4	**2**
24	2	1	5	7	3	1	5	3	4	1	8	4	**3**
25	5	8	7	9	1	10	9	6	7	5	7	2	**1**
26	3	2	5	5	2	6	8	5	3	9	9	3	**0**
27	10	2	8	3	1	10	5	5	9	3	1	2	**2**
28	4	2	6	9	7	3	7	2	6	7	2	1	**1**
29	1	2	3	1	10	10	10	4	2	9	2	7	**2**
30	9	3	5	8	8	1	4	1	10	3	4	5	**2**
31	3	2	8	1	3	1	5	7	8	1	1	7	**4**
32	10	3	5	10	6	3	6	2	7	5	1	6	**1**
33	5	1	9	10	6	10	6	1	10	6	2	10	**2**
34	3	2	10	7	9	2	8	9	4	4	1	5	**1**
35	4	9	5	7	1	8	2	4	5	6	4	8	**1**
36	10	3	9	2	9	3	6	3	1	3	6	6	**1**
37	5	6	4	4	6	4	7	3	4	2	3	6	**0**
38	1	6	9	5	7	4	8	10	8	4	7	6	**1**
39	7	4	1	8	4	6	3	9	4	9	2	6	**1**
40	8	9	5	10	1	5	1	1	6	8	10	4	**3**
41	6	9	8	7	7	5	10	1	3	7	5	1	**2**
42	5	2	3	9	9	6	7	5	3	9	6	9	**0**
43	9	6	3	10	5	7	2	8	9	1	6	8	**1**
44	6	5	1	7	4	4	1	7	10	2	9	3	**2**
45	8	7	4	1	6	7	8	5	1	2	7	9	**2**
46	8	7	7	3	1	7	5	5	8	3	6	6	**1**
47	10	1	5	3	3	5	5	7	7	2	4	1	**2**
48	1	6	6	5	8	7	3	4	8	10	4	6	**1**
49	2	4	5	2	1	3	10	4	5	4	8	8	**1**
50	5	1	6	10	10	6	4	6	3	6	9	9	**1**
Absent	**5**	**5**	**3**	**5**	**7**	**4**	**3**	**5**	**4**	**4**	**6**	**6**	

1 = absent; 2 thru 10 = present
Table 12.5.4: Daily Results of Computer Simulation of Worker Absenteeism

12.5.3 Creating Your Own Simulation

Because this is a random process, the results from repeated simulations can be different. To be confident about their spare worker policy, BT Auto's management decides to replicate this fifty day period many times over. They also would like to understand how much variability there might be from year to year in achieving their 95% goal. Each member of the class will create a spreadsheet to simulate the scenario of a 12-person workforce with a 10% absenteeism rate. The model will show a 50-day period.

Open Excel and create a table with a column labeled for each worker and a row for each day. Within the table, in the cell for your first worker on the first day, you need to enter a formula whose output will indicate whether that worker was absent on that day. The formula is similar to the "randInt(" command on the graphing calculator. In Excel the formula is "RANDBETWEEN." As you might imagine, this formula will return an integer value between any two integers you input.

For the situation where each worker is absent 10% of the time, we need to think of the absentee rate as a probability. A 10% absenteeism rate means that on any given day a worker may be absent with a probability of one out of ten. Thinking about it this way, we can choose a random number between 1 and 10 and assign an output of 1 to represent a worker being absent and the outputs 2 through 10 to represent a worker being present.

To do this, enter the following formula into the appropriate cell.

    =RANDBETWEEN(1,10)

Once this formula is entered, an integer between 1 and 10 should appear in the cell. To fill the rest of your spreadsheet, drag the formula across first and then down to copy it into the remaining cells.

Your spreadsheet should now be populated with integers from 1 to 10 for each of your 12 workers (columns) for 50 days (rows). You will need to save and prevent these randomly generated numbers from constantly changing. To do this, you will need to copy all of the numbers using a *copy*, *paste special*, *value* command. You can and paste these values right into the same cell locations they originally appear in.

Now we need to determine how many workers were absent each day and how many times each worker was absent over the twenty days. Excel has a formula that enables us to find this information.

The "COUNTIF" command will search a range of cells and count how many times a particular value occurs. We are concerned with absences, which are indicated by a zero in a cell. We can give Excel a command to search each row and each column for zeros, and it will report back how many times the number zero occurred. The information you need to provide the formula is the range of cells to search and the value it should count.

Enter the following formula in the cell at the end of the row showing the attendance for day one.

=COUNTIF([*drag range of cells here*],1)

As before, drag this formula down the column to day 50 to copy the formula. You will need to enter the formula again at the bottom of the first column, using the appropriate cells, and copy it across.

		Workers												Total
Worker Absenteeism														
		A	B	C	D	E	F	G	H	I	J	K	L	per day
	1	1	6	6	7	2	8	8	4	9	6	9	9	1
	2	3	8	1	4	7	9	9	8	8	10	3	10	1
	3	6	10	7	6	1	4	2	5	1	3	7	1	3
	4	10	5	3	7	9	8	9	6	9	9	7	1	1
	5	9	10	8	2	5	4	6	5	9	10	3	5	0
	6	10	7	5	1	2	2	1	4	6	6	10	3	2
	7	8	4	9	4	8	5	3	7	10	10	5	9	0
	8	9	5	6	8	2	8	4	10	9	10	7	7	0
Day	9	1	8	2	9	2	3	2	4	6	10	8	7	1
	10	6	4	7	4	10	9	2	4	7	10	6	5	0
	11	6	4	2	6	7	2	3	5	9	10	4	3	0
	12	6	1	8	4	9	6	10	7	9	7	9	8	1
	13	9	7	3	9	8	5	9	5	6	6	3	10	0
	14	3	8	4	6	7	8	10	9	1	5	1	8	2
	15	4	9	5	2	7	5	5	5	8	5	1	5	1
	16	5	8	2	10	9	3	7	3	6	2	6	1	1
	17	7	10	10	4	4	3	6	3	7	8	7	9	0
	18	6	7	9	2	2	6	8	8	5	9	7	2	0
	19	9	9	8	3	10	7	7	10	9	1	3	4	1
	20	3	9	10	6	7	4	10	8	9	3	9	8	0
Total per Worker		2	1	1	1	1	0	1	0	2	1	2	3	

1 = absent
other = present

Figure 12.5.1: 20-Day Spreadsheet

Q10.   Compare your row and column totals with those of your neighbors. What do you notice?

Q11.   What could explain the similarities and differences you noted?

Q12.   Can you find anyone else with a simulation spreadsheet that is exactly like yours?

Managers at BT Auto with different responsibilities will look at the data presented in Table 12.5.3 from diverse perspectives. The payroll manager is concerned with the details as to who showed up for work each day. He has to submit that information so as to ensure each individual is paid appropriately and the worker's absences are recorded against his sick leave allowance. The Human Resource (HR) manager may be interested in the pattern of absences for individual workers. Is a specific worker absent an unusual amount of time? Is another worker always there and perhaps eligible for a bonus? The HR manager will be interested in the column totals for each worker. In the next section we look at the variability in the simulated data from the perspective of each worker's record.

The plant manager is primarily concerned with getting the work done. This official is most interested in how many workers showed up for work. This assumes that all of the workers can do all of the jobs and there is no specialization. The manager will, therefore, focus on the row totals. In a later section, we take the plant manager's perspective and explore the randomness in the number of workers who were absent. This analysis is critical to determining the need and cost of having spare workers to cover for absentees.

Before proceeding let's explore your gut instinct. Since there is 0.1 probability a worker will be absent on any given day, workers will be absent on average five days out of fifty.

Q13.   However, what do you think the chances are that a worker will not miss a single day of the fifty days? Is it one in ten? One in a hundred? One in thousand? Or less than one in a thousand?

Q14.   Also, what do you think the chances are that a specific worker will be absent ten or more days during the flu season?

Q15.   Similarly, what do you think the chances are that all twelve workers will show up for work on a given day?

Individual Worker Absences—Your Individual and Class Pooled Data
The Assistant Vice President of Human Resources is interested in identifying the most reliable workers. These are workers who show up every day no matter what. He is also interested in monitoring those workers who are frequently out sick.

Q16.   From the column totals, identify the minimum and maximum number of days a worker was absent. How many workers were not absent even once during a 50 day period in the spreadsheet in Table 12.5.4? What about the workers in your spreadsheet?

Q17.   Which worker in your simulation had the best attendance record? How much better than the average was that worker?

Q18.   Which worker had the worst attendance record? How much worse was that worker than the average?

Q19.   Compare your answers with your neighbors? Were the workers with the best and worst attendance records the same for you and your neighbors?

Q20.   What are the minimum and maximum values for the whole class?

Q21.   Justify whether or not the worker with best attendance in the simulated dataset should be rewarded for good attendance.

You are going to use the class minimum and maximum values to develop a frequency distribution table of the number of days a worker was absent during the 50-day time period. You will first do this just for your own data. Then the class will pool all the data to obtain larger sample and thus a better estimate of the relative frequency of different values.

Create a sequence of integer values beginning at the class minimum and ending at the class maximum. Next to each integer value, record the number of times in *your* spreadsheet that each value appeared in the column totals.

Q22.   How many workers in your spreadsheet were absent the class minimum and class maximum number of days?

Q23.   Now you will pool the frequency distribution results for the whole class and then determine the *relative frequency* for the whole class. Find the observed relative frequency for each value and record it in a class pooled table. To start, count the total number of times the minimum value appeared in the pooled data. Divide this by 12 times the number of students in the class. The 12 represents the number of workers in each spreadsheet. Repeat this for every value in the table.

Q24.   For class pooled data, what is the relative frequency of a worker being present each and every day during a 50 day period? How does this compare to your gut feel estimate noted in question 13?

Q25.   In the class pooled data, how many times was a worker absent exactly once during a 50 day period? What is its relative frequency?

Q26.   In the class pooled data, how many times was a worker absent ten or more times during a 50 day period? What is its relative frequency? How does this compare to your gut feel estimate noted in question 14?

Total Number of Workers Absent Each Day
In this section, you will take the plant manager's perspective. The focus will be on the frequency distribution of the number of workers absent each day.

Q27.   In your spreadsheet look at the row totals. What were the minimum and maximum number workers that were absent?

Q28.   How many days experienced no absences?

Q29.   What are the minimum and maximum values for the whole class?

You are going to use the class minimum and maximum values to develop a frequency distribution table of the number of workers absent during each day of the 50-day time period. You will first do this just for your own data. Then the class will pool all the data to obtain larger sample and thus a better estimate of the relative frequency of different values.

Create a sequence of integer values beginning at the class minimum and ending at the class maximum. Next to each integer value, record the number of times in your spreadsheet that each value appeared in the row totals. Calculate the relative frequency by dividing by 50.

Q30.   In your row totals, how many days was exactly one worker absent during the 50-day period?

Q31.   Now tabulate the results for whole class. How many days was exactly one worker absent during the 50-day period? Calculate this proportion for the combined total of workers simulated by the whole class.

Q32.   Repeat this calculation of relative frequency for each value ranging from the class minimum to the class maximum.

Q33.   How do these relative frequencies compare to the data in Table 12.5.3

Cost of Spare Workers
The company is considering hiring three steady spare workers to show up each day for work. They are on standby to replace absent workers. To encourage them to show up for work, the company established a compensation policy. These standby workers are paid $55 just for coming to work even if they do no work. However, if the individual actually works, the worker is paid an additional $70 for a total of $125.

Assume 3 spare workers showed up for work.

Q34.   What is the cost to the company on a day that only one worker was absent. Just one spare worker was needed and the other two were sent home?

Q35.   What is the cost to the company on a day that two workers were needed and only one was sent home?

Q36.   On any day when the number of absent workers is more than 3, what is the cost for standby workers that day?

Q37.   In general let N be the number of standby workers who actually work. With 3 workers on regular standby, write an equation for the total cost as a function of N.

In Table 12.5.5 we added a column that includes the cost for standby workers for each day in the simulation. It is the next to last column on the right.

Q38.   What is the simulated total 50-day cost for the current policy of 3 spares on standby? What is the average daily cost?

New Policy under Consideration
Management is considering paying one worker $100 per day everyday no matter what because most days there is at least one worker absent. The other two workers are paid as before $55 or $125. If the first worker is not needed for the production line, he is given a special task such as cleaning up some set of files or counting inventory. It is not a high value job but it keeps him busy. We have added another column to our original table and recorded each day's cost for this new policy.

Q39.   On days with zero absences, what does this new policy cost? Does it cost more or less than before?

Q40.   On days with one absence, how much money does this new policy cost? Is it more or less than before?

Q41.   On days with two absences, how much money does this new policy cost? Three absences?

Q42.   Under what days does the policy cost more money and when does it save money?

Q43.   What is the most common dollar amount under the new policy? Explain why occurs so often.

Q44.   Based on Table 12.5.5, how many days out of the fifty did the new policy save money?

Q45.   Would you recommend the new policy or the current policy? Explain your answer

					Workers								Absent	Cost	
Day	A	B	C	D	E	F	G	H	I	J	K	L	per day	Current	New
1	1	6	6	7	2	8	8	4	9	6	9	9	1	$235	$210
2	3	8	1	4	7	9	9	8	8	10	3	10	1	$235	$210
3	6	10	7	6	1	4	2	5	1	3	7	1	3	$375	$350
4	10	5	3	7	9	8	9	6	9	9	7	1	1	$235	$210
5	9	10	8	2	5	4	6	5	9	10	3	5	0	$165	$210
6	10	7	5	1	2	2	1	4	6	6	10	3	2	$305	$280
7	8	4	9	4	8	5	3	7	10	10	5	9	0	$165	$210
8	9	5	6	8	2	8	4	10	9	10	7	7	0	$165	$210
9	1	8	2	9	2	3	2	4	6	10	8	7	1	$235	$210
10	6	4	7	4	10	9	2	4	7	10	6	5	0	$165	$210
11	6	4	2	6	7	2	3	5	9	10	4	3	0	$165	$210
12	6	1	8	4	9	6	10	7	9	7	9	8	1	$235	$210
13	9	7	3	9	8	5	9	5	6	6	3	10	0	$165	$210
14	3	8	4	6	7	8	10	9	1	5	1	8	2	$305	$280
15	4	9	5	2	7	5	5	5	8	5	1	5	1	$235	$210
16	5	8	2	10	9	3	7	3	6	2	6	1	1	$235	$210
17	7	10	10	4	4	3	6	3	7	8	7	9	0	$165	$210
18	6	7	9	2	2	6	8	8	5	9	7	2	0	$165	$210
19	9	9	8	3	10	7	7	10	9	1	3	4	1	$235	$210
20	3	9	10	6	7	4	10	8	9	3	9	8	0	$165	$210
21	8	5	8	7	6	1	4	4	8	2	9	7	1	$235	$210
22	9	4	7	6	8	9	2	10	10	7	8	7	0	$165	$210
23	10	2	7	1	4	9	8	1	5	4	9	4	2	$305	$280
24	2	1	5	7	3	1	5	3	4	1	8	4	3	$375	$350
25	5	8	7	9	1	10	9	6	7	5	7	2	1	$235	$210
26	3	2	5	5	2	6	8	5	3	9	9	3	0	$165	$210
27	10	2	8	3	1	10	5	5	9	3	1	2	2	$305	$280
28	4	2	6	9	7	3	7	2	6	7	2	1	1	$235	$210
29	1	2	3	1	10	10	10	4	2	9	2	7	2	$305	$280
30	9	3	5	8	8	1	4	1	10	3	4	5	2	$305	$280
31	3	2	8	1	3	1	5	7	8	1	1	7	4	$445	$420
32	10	3	5	10	6	3	6	2	7	5	1	6	1	$235	$210
33	5	1	9	10	6	10	6	1	10	6	2	10	2	$305	$280
34	3	2	10	7	9	2	8	9	4	4	1	5	1	$235	$210
35	4	9	5	7	1	8	2	4	5	6	4	8	1	$235	$210
36	10	3	9	2	9	3	6	3	1	3	6	6	1	$235	$210
37	5	6	4	4	6	4	7	3	4	2	3	6	0	$165	$210
38	1	6	9	5	7	4	8	10	8	4	7	6	1	$235	$210
39	7	4	1	8	4	6	3	9	4	9	2	6	1	$235	$210
40	8	9	5	10	1	5	1	1	6	8	10	4	3	$375	$350
41	6	9	8	7	7	5	10	1	3	7	5	1	2	$305	$280
42	5	2	3	9	9	6	7	5	3	9	6	9	0	$165	$210
43	9	6	3	10	5	7	2	8	9	1	6	8	1	$235	$210
44	6	5	1	7	4	4	1	7	10	2	9	3	2	$305	$280
45	8	7	4	1	6	7	8	5	1	2	7	9	2	$305	$280
46	8	7	7	3	1	7	5	5	8	3	6	6	1	$235	$210
47	10	1	5	3	3	5	5	7	7	2	4	1	2	$305	$280
48	1	6	6	5	8	7	3	4	8	10	4	6	1	$235	$210
49	2	4	5	2	1	3	10	4	5	4	8	8	1	$235	$210
50	5	1	6	10	10	6	4	6	3	6	9	9	1	$235	$210
Absent	5	5	3	5	7	4	3	5	4	4	6	6		$12,240	$11,900

Table 12.5.5: Workers absences and standby worker costs

Section 12.6: Chapter 12 (Basic Probability) Homework Questions

12.6.1 Complement
1. Two Complementary events
   a. Describe complementary events with regard to the outcome of a baseball game
   b. Be careful to describe complementary events with regard to the outcome of a football or soccer game. How is this answer different than the answer to a?
   c. Describe complementary events with regard to performance on a test.
   d. Describe complementary events with regard to the launch of the Space Shuttle

2. Provide examples of events that are mutually exclusive but are not complementary in the context of:
   a. The outcome of a soccer game or college football game.
   b. Performance on a test
   c. Timeliness of completing a project

12.6.2 Simulation
3. Simulate by hand flip coin – Two coins
   a. Flip two coins. List all of the possible outcomes of the experiment.
   b. In the coin flip experiment if you only record the number of heads, list all of the possible outcomes of the experiment?
   c. Provide an example of complementary events with regard to the outcome of flipping two coins
   d. Flip two coins 10 times. Record your results into the table below. (Number of Heads equals 0 means experiment resulted in all Tails.)

Number of Heads	Experiments										Frequency	Relative Frequency
	1	2	3	4	5	6	7	8	9	10		
0												
1												
2												

Table 12.6.1

   e. How many of the times did you get one Head and one Tail in 10 experiments? Write your answer into the corresponding row under the frequency column?
   f. Write the frequency of other possibilities into the table.
   g. Calculate the relative frequency of each possibility and compare relative frequency of each outcome with two of your friends.

4. Simulate by calculator flip coin – Two coins
   a. Go to the APPS menu of your calculator, press 0 for probability simulation, and then press 1 to toss coins. Press on the SET button: change the number of trial set to 50, change the number of coins to 2, and press the OK button. Now, press on the TOSS button. You will see a graph with bars showing the frequency of each possible outcome. What do you observe in the graph?
   b. Now, go to the DATA menu and count the frequency of each outcome.
   c. What are the relative frequencies of each outcome? Compare your results with your friends.

5. Calculate Probabilities – Use Multiplication Rule (Two coins)
   a. Calculate the probability of two heads when two coins are flipped
   b. Calculate the probability of zero heads when two coins are flipped

    c. Use complement to calculate the probability of exactly one head when two coins are flipped
    d. Compare the theoretical probabilities with the observed relative frequencies you found in questions 3 and 4.

6. Simulate by hand flip coin – Three coins
    a. Flip three coins. List all of the possible outcomes of the experiment.
    b. If you only record the number of heads, list all of the possible outcomes of the experiment?
    c. In the three coin flip experiment, you can have either all (three) Heads or less than three (zero, one, or two) Heads. The probability of having three Heads or less than three Heads add up to 1. What are the other complementary events with regard to the outcome of flipping three coins?
    d. In the three coin flip experiment, one Head and two Tails and three Heads and zero Tails are mutually exclusive events and they are not complementary. Give an example of two mutually exclusive events that are not complementary.
    e. Flip three coins 10 times. Record the number of times the outcome is all Heads and its complement. Calculate the relative frequency. Do relative frequencies add up to 1?
    f. Now record the number of Heads you observed in 10 trials into the table below. Calculate relative frequencies.

Number of Heads	Experiments										Frequency	Relative Frequency
	1	2	3	4	5	6	7	8	9	10		
0												
1												
2												
3												

Table 12.6.2

    g. How many of the times did you get only one Head in 10 experiments? Write your answer into the corresponding row under the frequency column.
    h. Write the frequency of other possibilities into the table.
    i. From the table above look at the relative frequencies of each outcome. Do they add up to 1? Do mutually exclusive events have to add up to one as complementary events?
    j. Chances of getting only one Head or only one Tail in flipping three fair coins must be equal. Is your result a representation of this statement?

7. Simulate by calculator flip coin – Three coins
    a. Go to the APPS menu of your calculator again. Press 0 for probability simulation, and then press 1 to toss coins. Press on the SET button: change the number of trial set to 50, change the number of coins to 3, and press the OK button. Now, press on the TOSS button. You will see a graph with bars showing the frequency of each possible outcome. What do you observe in the graph?
    b. Now, go to the DATA menu and count the frequency of each outcome.
    c. What are the relative frequencies of each outcome? Compare your results with your friends.

8. Calculate Probabilities – Use Multiplication Rule (Three coins)
    a. Calculate the probability of three heads when three coins are flipped
    b. Calculate the probability of zero heads when three coins are flipped
    c. Explain why the probability of getting one head and two tails should be the same as the probability of getting two heads and one tail.
    d. Use Complement and the answer to part c to calculate the probability of exactly one head. Exactly two heads.

    e.   Compare the theoretical probabilities with the observed relative frequencies you found in questions 6 and 7.

9. Physical simulation – Pick one
    a.   Take 2 red and 4 blue M&Ms, and put them into a bag. Then, blindly withdraw an M&M. Record your outcome. Put the M&M that you picked back into the bag and repeat this experiment 6 times. How many times did you pick a blue M&M in 6 trials?
    b.   What is the frequency of picking R? What is the frequency of picking B?
    c.   Do the relative frequencies you found in part (b) make sense? Why/Why not?
    d.   Compare the relative frequency of picking Rs to 1/3.
    e.   Repeat part (a) 9 times. Namely you will blindly withdraw an M&M, put it back into the bag, and put a sign into the table below if you observe an R; leave it blank otherwise. Repeat 6-tuple withdrawals 9 times.

Trial #	1	2	3	4	5	6	# of Rs	Relative Freq. of Rs
1								
2								
3								
4								
5								
6								
7								
8								
9								

Table 12.6.3

    f.   How often are your relative frequencies exactly 1/3 in 9 trials?
    g.   Pool your results from all trials. Is the relative frequency of picking Rs close to 1/3? Explain why it should be closer to 1/3.

10. Physical simulation – Pick two with replacement
    a.   You have 2 red and 4 blue M&Ms in the bag. Blindly withdraw an M&M, look at its color. Put it back into the bag. Withdraw a second M&M, look at its color and put it back into the bag. What did you get in each trial?
    b.   Repeat this experiment 20 times. Record your results into the table.

Outcomes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Frequency	Relative Frequency
RR																						
RB																						
BR																						
BB																						

Table 12.6.4

c.   Which was the most frequent outcome? Which was the least frequent outcome?
d.   Now, record your results as "number" of Rs into the table below.

Number of Rs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Frequency	Relative Frequency
0																						
1																						
2																						

Table 12.6.5

e. Calculate theoretical probabilities of having 0, 1, or 2 Rs in the two M&M pick experiment.
f. Compare the relative frequencies you observed in part (d) with the theoretical probabilities.

11. Using MS Excel Random Generator
a.   Repeat this experiment 50 times using Randbetween function of Excel. Set the Excel sheet as in the snapshot. Write the formula in cell B2 as the Formula Bar and copy it to other cells.

B2				f_x	=IF(RANDBETWEEN(1,6)<=2,"R","B")	
	A	B	C	D	E	F	G
1	Trial	1st coin	2nd coin	# of Rs			
2	1	R					
3	2						
4	3						
5	4						

Figure 12.6.1

b.   Record how many times you observed 0, 1, or 2 Rs into the table below.

Number of Rs	Frequency	Relative Frequency
0		
1		
2		

Table 12.6.6

c.   Compare the relative frequencies you observed in part (b) with the theoretical probabilities. Are the relative frequencies closer to the theoretical probabilities compared to the part (d) of question 10?

12. Physical simulation – Pick two without replacement
a.   Again you have 2 red and 4 blue M&Ms in the bag. You blindly withdrew an M&M, which was red. This time you didn't put it back into the bag and withdrew a second M&M. How likely are you pick a red M&M again in the second trial?

13. Multiple Choice Exam

A student has to take an exam, but he has not studied at all. Therefore, he has no idea about the answers. The exam is a multiple choice exam with 10 questions; each has 4 choices with only one correct answer. The student has decided to take his chances. The table below contains simulated data for 30 repetitions of a ten question exam. Answer the questions according to the table given below.

Trial #	Questions										No. of Cs
	1	2	3	4	5	6	7	8	9	10	
1	C	C	I	I	I	C	I	C	C	C	
2	I	C	I	C	I	I	C	I	I	I	
3	I	I	C	I	I	I	I	I	C	C	
4	C	I	I	I	C	I	I	I	I	I	
5	I	I	C	C	I	I	C	I	C	I	
6	I	I	I	I	I	C	I	I	I	I	
7	I	I	C	C	C	C	I	C	C	I	
8	I	I	C	I	I	I	I	I	I	I	
9	I	I	C	I	I	I	I	I	I	I	
10	I	I	I	I	I	I	I	I	I	I	
11	C	C	C	I	C	I	I	I	C	C	
12	I	C	C	I	I	I	I	C	I	I	
13	I	I	I	C	I	I	I	C	C	I	
14	I	I	I	I	I	I	I	I	I	I	
15	I	C	I	C	I	I	C	I	I	I	
16	I	I	C	I	I	I	I	I	C	I	
17	I	I	I	C	C	I	C	I	I	I	
18	I	C	I	C	I	I	C	I	I	I	
19	C	I	I	C	C	I	I	I	C	I	
20	C	I	I	I	I	I	I	I	C	I	
21	C	I	C	I	I	C	I	C	I	I	
22	C	C	I	I	I	I	I	I	C	I	
23	I	I	I	I	I	I	C	I	C	I	
24	I	I	I	I	I	C	I	I	I	I	
25	I	I	I	I	I	I	C	I	C	I	
26	I	I	I	I	C	C	I	I	I	C	
27	I	I	C	I	C	C	I	I	I	I	
28	I	I	C	C	I	I	C	I	I	C	
29	I	C	I	C	I	I	I	C	I	I	
30	I	I	C	I	C	I	C	I	C	I	
No. of Cs											

Table 12.6.7

a. Record the number of correct answers in each trial to the last column. How many of the trials did the student fail to score a 6 or higher?

b. Record the number of correct answers to each question to the last row. Look at the proportion of correct answers to each of the questions. Is it exactly 0.25?

	Question Number									
	1	2	3	4	5	6	7	8	9	10
No. of Cs										
Proportion										

Table 12.6.8

c. Using the last column of the table, make a frequency table for the number of correct answers.

No. of Cs	Frequency
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Table 12.6.9

d. What is the most frequent score on the exam? Is this number surprising?
e. What is the observed relative frequency that the student has 6 correct answers?
f. What is the observed relative frequency that the student has 0 correct answers?
g. Calculate the theoretical probability of getting 0 correct answers using the Multiplication Rule.
h. If it is known that the student has failed, what is the probability that he has 4 correct answers?

1. Simulate a Multiple Choice Exam
    a. Simulate his performance in MS Excel; make 30 simulation runs. Set the Excel sheet as in the snapshot. Write the formula in cell B3 as the Formula Bar and copy it to other cells. The IF command in Excel states that if the random number is less than or equal to 25, Excel will record a value of "C" for correct in the cell. Otherwise it records and "I" for incorrect.

Figure 12.6.2

    b. Why is there a ≤ 25 in the statement?
    c. He will pass the exam if he answers at least 6 of the questions correctly. According to this, in how many trials does he fail?

d. In any of the trials, does he have a chance to answer all the questions correctly?
e. Make a frequency table for the number of correct answers and identify which is the most frequent score.

12.6.3 X-Press School Newspaper

X-Press is a monthly school newspaper which started in September, 2008. It has 3 writers and 3 editors. According to the records from last year, it is estimated that the probability that each student will meet the deadline is 90%. Even though, this year has started out well for all the writers, the paper was delayed twice in 6 months due to lateness of two editors. Table 1 shows the data for the six months of 2009.

	Editors			Writers			
Months	**Tom**	**Bob**	**May**	**Ann**	**Lee**	**Eric**	**Late?**
1	on time	on time	on time	on time	on time	on time	No
2	on time	on time	on time	on time	on time	on time	No
3	on time	on time	**late**	on time	on time	on time	Yes
4	on time	on time	on time	on time	on time	on time	No
5	on time	on time	on time	on time	on time	on time	No
6	on time	**late**	**late**	on time	on time	on time	Yes
Number of late times per student	0	1	2	0	0	0	

Table 12.6.10: September 2008-February 2009 Performance Data

Table 12.6.10 shows that student writers showed 100% performance in the last six months.

1. Is this strong evidence of an improvement of the timeliness of the student writers?

2. The two times late in the last 6 months was due to two of the editors, Bob and/or May. Looking at the data, can we say that the editors have some problems?

3. May was late twice over the 6 months period. Do you think May has some problems?

Actually, the data in Table 12.6.10 is simulated. It is formed by generating random numbers for all the editors and writers as if all of them are meeting deadlines 90% of the time. It is just pure randomness that all the writers are showing 100% performance. There is no evidence that they will show the same performance in the next month.

4. If you assume that each the writer and the editor is 90% reliable, what is the probability of all of them meeting the deadline in any month?

5. What is the probability of not all of them meeting the deadline in any month?

6. Simulate the performance of the newspaper over the 6 months yourself. Create a table as Table 1 above in Excel. Write **=IF(RANDBETWEEN(1,10)=1,"late","on time")** in B3 and copy it to other cells until G8. In cell H3, write **=IF(COUNTIF(B3:G3,"late")>0,"Yes","No")** and copy it until H8. To count the number of lateness of a student write **=COUNTIF(B3:B8,"late")** in cell B9 and copy it until G9.

B3		f_x	=IF(RANDBETWEEN(1,10)=1,"late","on time")				

	A	B	C	D	E	F	G	H
1	`		Editors			Writers		
2	Months	Tom	Bob	May	Ann	Lee	Eric	Late?
3	1	on time						
4	2							
5	3							
6	4							
7	5							
8	6							
9	Number of late times/student							

Figure 12.6.3: Snapshot of the Simulation

7. Is there any editor or writer, whose performance is similar to May's in your experiment? Namely any editor or writer is late twice over the 6 months period?

8. Repeat the experiment in question 2a 10 times. Simulate the performance of the newspaper over the 6 months 10 times.

9. In how many of the runs, was an editor or writer late twice over the 6 month period?

10. How many times was the newspaper published on time for the entire 6 months?

11. The ten runs represent simulating 60 months. What percentage of times was the paper late?

12. Make a frequency table which shows the number of late times of the paper in 10 simulation runs. Calculate the relative frequencies.

13. Calculate the cumulative relative frequencies of the number of late times of the paper.

14. According to the cumulative relative frequencies you obtained in the previous question, what proportion of the time the paper is late 2 or more times over the 6 month period?

15. The newspaper has changed its policy. Now, it will publish the paper if just one editor or just one writer is late. Use the previously simulated data to evaluate this new policy.

16. Make a frequency table which shows the number of times the paper is late in 10 simulation runs. Calculate the relative frequencies and cumulative frequencies.

17. According to the cumulative relative frequencies you obtained in the previous question, what proportion of the time the paper is late?

18. Evaluate the new policy. Did it help the newspaper be published on time? (Hint: Compare the frequency tables from questions 2e and 2g)

12.6.4 High School Hockey—5 Game Series

1. Red Run High School hockey teams, the Quacks and the Tops, are playing a best-of-five series to determine the champion this year. The series ends when one team wins 3 games. We believe that both teams are equally likely to win in each game.
    a. Take a piece of paper, and cut it into 2 equal pieces. Write "Quacks" on one piece, and "Tops" on the other piece. Fold the papers, shuffle them, and pick one piece. Record what you have observed? Repeat this until you collect 3 of the same team's name. The total number of repetitions must not exceed 5.

    b. How many times did you fold-shuffle-pick until you collect 3 of the same team's name?

    c. Repeat the experiment in part (a) 10 times. How many trials resulted in the order of "Quacks – Quacks – Tops – Tops – Tops"?

    d. What is the probability of observing this specific sequence, "Quacks – Quacks – Tops – Tops – Tops"? Compare this probability with the relative frequency in your simulation. Are you surprised by the result?

2. Now, let's simulate the series by flipping a coin. Let heads be the Quacks and tails be the Tops. Repeat flipping the coin until you get either three heads or three tails. The total number of flips must not exceed 5. This will simulate one 5-game series.

    a. Do the same thing 20 times to complete the simulation runs and record the results in a table similar to the one below.

    b. How many times did Tops win the series in the 20 simulated experiments? Compare your results with your friends.

    c. What other physical experiments can you use to simulate the series? How?

3. Suppose 3 games have already been played, and the Quacks have won 2 out of 3 games.

    a. If we still assume that both teams are equally likely to win in each game, what is the probability that the Tops will come back and win the series?

    b. Historical data suggests that a team winning 2 out of the first 3 games will win the series 85% of the time and there is a 15% chance for the other team to win. This percentage is smaller than the value found in part a. According to this information, the Quacks and the Tops are not equally likely to win in each game. If the probability that the Tops will win two games in a row is 0.15, what is the probability of Tops winning each individual game? What is the probability of Quacks winning each individual game? Determine this value to three decimal places.

4. You will simulate the 4th game using the RAND (Math➜ Prb ➜ 1: Rand) function of your calculator. If the random number is less than or equal to the probability of the Tops winning that was found in 3b, then the Tops wins the game and the series continues. Otherwise the Quacks win the game and the series ends.

    a. Simulate this once. Who won the game?

b. Simulate the 4th and the 5th games 20 times. What percentage of time did the series end with the 4th game?

c. Explain why you cannot flip a coin to simulate the 4th game?

5. Simulate the 4th and the 5th games 100 times using MS Excel. Write the formulation in cell B2 and drag it to the bottom and to the right until you cover all the games and the simulation runs. The letter x corresponds the probability of the Tops winning that you have found earlier.

	A	B	C	D
1	Run Number	4th game	5th game	
2	1	=IF(RAND()<=x,"Tops","Quacks")		
3	2			
4	3			

Figure 12.6.4

a. How many times did Quacks win the series in 4 games?

b. How many times did Quacks win the series in 5 games?

c. How many times did Tops win the series?

12.6.5 2010 Bradley Cup—7 Game Series

1. The Finals of 2010 Bradley Cup Basketball Playoffs have begun. The Raleigh Elks and Lansing Moose are playing a best-of-7 series for the championship. Four games have already been played. The Elks won 3 out of the first 4 games and the Moose won only 1 game. Assuming that the Moose and the Elks are equally likely to win in each game, answer the 3 questions below.
    a. What is the probability that the Moose will win the series?

    b. If the Moose wins the 5th game, what is the probability that the Moose will win the series?

    c. What is the probability that the Elks will win the series if the Moose wins the 5th game?

Historical data indicates that a team winning 3 out of the first 4 games won the series 95% of the time. This data suggests that the two teams are not equally likely to win each game.
    d. Explain why the team that is ahead 3 to 1 may have a higher than 50% chance of winning each game.
    e. Perhaps the chance of winning in each game is only 40% for the Moose. What would be the probability that the Moose will win the series?
    f. Based on historical data, the chances of a team like Moose winning the series is only 0.05. What value raised to 3rd power yields 0.05? What does this number represent?

2. Simulate the 5th, 6th and 7th games 100 times using MS Excel. Use the value you found in part (f) above rounded to three decimal places to run your simulation.

    a. How many times did Elks win the series in 5 games?

    b. How many times did Elks win the series in 6 games?

c. How many times did Moose win the series?

d. Given the series did not end in five games, what proportion of time did the Elks eventually win the series?

12.6.6 Airline Overbooking

1. **Profits and Break-Even Analysis**
Great Lakes Airlines (GLA) operates out of Lansing, MI (LAN) and serves 20 cities in 7 states. They want to conduct a profit break-even analysis on Flight 425 which flies from Lansing (LAN) to in Milwaukee, WI (MKE). The average one-way airfare price is $300 and operating cost of the plane is $6,000 with an additional cost of $20 per customer. Flight 425 has 30-passenger seat capacity. Customers who cancel their flights or don't show up at the gate will pay GLA $ 75.

a. What is the revenue if the number of reserved passengers is 20 and all passengers show up?

b. What is the cost if the number of reservations is 20 and all passengers show up? What will be the profit or loss with 20 passengers?

c. The number of passengers where the cost equals revenue is called the break-even point. According to this, what is the break-even point for this case in which all reservations show up?

According to GLA flight records, 15% of the people who make reservation don't show up at the gate.

d. In Table 12.6.11, 5 customers out of 30 did not show up, so what is the profitability? (Remember to include the penalty cost for not showing up.) What would the profitability be if all the customers had shown up? What is the ratio of the first case to the second?

Customer Number	Show-up?	Customer Number	Show-up?
1	Yes	17	Yes
2	No	18	No
3	Yes	19	Yes
4	Yes	20	Yes
5	Yes	21	Yes
6	Yes	22	No
7	Yes	23	Yes
8	Yes	24	Yes
9	Yes	25	Yes
10	Yes	26	Yes
11	Yes	27	Yes
12	Yes	28	No
13	Yes	29	No
14	Yes	30	Yes
15	Yes	**Number of No-Shows**	5
16	Yes		

Table 12.6.11: Customer Show-ups with No Overbooking

e. Write a mathematical expression to calculate GLA's net profit as a function of the number of reservations and the number of no shows. Be careful to define all of your variables.

f. If it is known that 30 customers have already reserved their tickets, simulate the show-up behavior of customers in Excel Worksheet. Then calculate the net profit GLA earned on this flight. Use RANDBETWEEN(1,100) function to generate numbers between 1 and 100. If this random number is less than or equal to 15, it means the person will not show up. Count the number of no-shows using the function COUNTIF(B2:B31,"No").

	A	B	C	D
1	customer #	Show-up?		
2	1	Yes		
3	2	Yes		
4	3	Yes		
5	=IF(RANDBETWEEN(1,100)<=15,"No",			
6	"Yes")			
7	6	Yes		
8	7	Yes		
9	8	Yes		
10	9	Yes		
11	10	No		
12	11	Yes		

Figure 12.6.5: Simulation of Customer Show-ups at the Gate

g. Is it possible to fill all the seats without overbooking? What is the probability of that happening? If GLA has 250 flights per year, on average how many times will all 30 people show up?

2. Overbooking and Bumping

Overbooking is a common practice in airline companies and it is legal. Sometimes people cancel their flights or don't show up at the gate. This results in empty seats on the planes and, in other words, loss of revenue. By overbooking flights, the likelihood of empty seats (spoilage cost) because of those cancellations and no-shows is reduced.

If the number of ticketholders who arrive at the gate is more than the flight capacity, then some of the passengers will be bumped to other flights. They will receive some compensation for the disruption. They generally receive another ticket for the next available flight and compensation in the form of cash or a voucher for future travel. If the delay involves an overnight stay, they will be placed in a hotel free of charge. Compensation will vary depending upon whether the passenger volunteers to be bumped or is involuntarily bumped.

GLA wants to determine an optimal overbooking strategy to reduce the lost revenue due to cancellations and no shows on fully booked flights. They will apply the initial overbooking strategy on Flight 425. If a customer is denied boarding in case of overbooking, GLA will bump the customer to another flight and pay $250 compensation to that customer. GLA starts its overbooking strategy with 31 reservations on each flight.

a. Determine GLA's net profit if all 31 customers show up,

b. Determine GLA's net profit if only 30 customers show up,

c. Determine GLA's net profit if only 29 customers show up.

d. Write a mathematical expression to calculate GLA's net profit for the case of overbooking as a function of the number of reservations, the number of no shows and the number of people bumped. Be careful to define all of your variables.

e. Table 12.6.12 shows the frequency of the number of customers who did not show up over 100 days when GLA allows 32 reservations and overbooks 2 customers consistently. Fill in the blank cells under the number bumped and the profit/day column. Multiply the number of days by profit/day and write the solution under the 5th column. Take the average of the 5th column to find the average profit per day.

Number of No-Shows	Number Bumped	Profit/Day	Number of Days	Profit/Day * Number of Days
0			1	
1			1	
2			8	
3			17	
4			23	
5			15	
6			15	
7			7	
8			6	
9			5	
10			2	

Average Profit/Day ___

Table 12.6.12: Frequency table from 100 days

f. What number of no-shows was the most profitable? Why?

g. Table 12.6.13 shows the frequency of number of no-shows over 100 days when GLA allows 32 reservations. Fill in the blank cells under the number bumped column. Calculate the relative frequency for each number of no-shows and write it to the corresponding row in the 4th column. Write the cumulative relative frequencies in the 5th column.

Number of No-Shows	Number Bumped	Number of Days	Relative Frequency	Cumulative Relative Frequency
0		*1*		
1		*1*		
2		*8*		
3		*17*		

4		*23*		
5		*15*		
6		*15*		
7		*7*		
8		*6*		
9		*5*		
10		*2*		

Table 12.6.13: Frequency table from 100 days

Now, answer the questions below using the 5th column of Table 12.6.13.

h.   What percent of the time all 32 customers appear at the gate?

i.   What percent of the time was a customer bumped?

j.   What percent of the time did more than 4 customers not show up for their flight?

k.   When you multiply the daily profit of each number of occurrences with its relative frequency and sum the terms, you can find the average profit per day. Then, use the relative frequencies to calculate the average daily profit.

3. Optimal Overbooking Policy

Table 12.6.14 shows the 100-day simulation results of Flight 425 with different overbooking policies (35, 36, and 37 reservations). For example, according to this table, when GLA allowed 35 customers to reserve for Flight 425, 3 customers did not show up 8 times out of 100 days. In this case, 2 customers have to be bumped to other flights.

Number of No-Shows	Frequency of No-Shows for Different Policies		
	35	36	37
0	2	1	0
1	2	0	3
2	7	5	2
3	8	2	9
4	16	17	14
5	21	21	14
6	16	21	17
7	12	16	16
8	8	9	14
9	6	3	7
10	0	3	2
11	2	0	2
12	0	2	0
13	0	0	0
14	0	0	0

| 15 | 0 | 0 | 0 |

Table 12.6.14: No-Shows History of Flight 425

a. Using Table 12.6.14, calculate the daily average profits and find which policy is the most profitable?

b. Fill in the blank cells in Table 12.6.15 using the data in Table 12.6.14.

	Overbooking Policy		
	35	36	37
Probability of no bumping			
Probability of bumping 2 or more			
Average number of no-shows			
Average number of bumped passengers			
Average number of flying passengers			

Table 12.6.15: Evaluation of Overbooking Policies

c. Calculate the average daily profit using a specific reservation policy and the average number of flying passengers, no-shows, and bumped customers reported in Table 5. Compare this to the results in part a.

d. Do a 25-day simulation for a policy of 34 and a policy of 38 reservations. Compare your results with the previous policies. You will get extra credit if you do 100-day simulation!

e. Calculate the average daily profits of the 34 and 38 reservation policies, respectively.

f. What is the optimal overbooking policy according to your results?

Chapter 12 Summary

What have we learned?

The purpose of this chapter is to gain a better understanding of the nature of random behavior. Random behavior means that we do not know what will happen in the future. It is not that we do not know what is possible. Often we know one of a number of possibilities will happen. However, we do not know which of them will occur. One way to study situations that include multiple random events is to simulate them. We started by simulating two equally likely outcomes by flipping coins. When we have outcomes that are not equally likely we use a random number generator such as a calculator or computer. The purpose of the simulation is to try to discover would happen if a situation happened a large number of times. This can give us an idea of how likely a particular outcome is. We can use this information to make decisions.

We have also learned how to combine the probabilities of two or more events if we know the probability of each of them occurring. For independent events, events where the outcome of one event does not affect the outcome of the other event, we can multiply the probabilities together.

Finally, we learned how to use excel for simulations. Excel can be used for generating random numbers and can also do calculations to use to evaluate the results. When we flip coins or use our calculator to generate the results we have to write them down. Excel shows the result but we have to remember to use the paste special command to save the numbers generated to prevent them from changing the next time we enter a command.

Our intuition when it comes to random behavior is often wrong. People have a hard time getting rid of the belief that the future will correct for a surprising run of random behavior in the past. There is no "Law of Averages" that outcomes of a random event will "even out" within a small sample. The average of the results of a large number of trials will get closer to the expected value as an event is repeated. However, that is not because future results somehow make up for the past. It is simply that a small number of results have a small impact on the average of a large number of results.

Terms

Outcome	The result of an experiment or simulation involving uncertainty
Sequence	A list of outcomes when an event is repeated
String	A sequence of consecutive results with the same outcome
Mutually Exclusive	Two outcomes that cannot both happen at the same time
Complementary	If one outcome does not occur, the other outcome must occur
Collectively Exhaustive	The set of all possible outcomes
Success	The outcome that is being studied. It may or may not be an outcome that is desired
Simulate	To create a model for a situation with the same probabilities as the real world situation but that is easier to duplicate
Notation	The shortened way to write a mathematical statement
$P(A)$	The probability that the outcome will be A
$P(A^c)$	The probability that the outcome will be the complement of A
$P(A \text{ and } B)$	The probability that both A and B will occur
Independent Events	The outcome of one event does not affect the likelihood of any outcome of another event
Multiplication Principle	If two events are independent then $P(A \text{ and } B) = P(A) \cdot P(B)$

Chapter 12 (Basic Probability) Objectives

You should be able to:

- Identify the possible outcomes for an event.

- Determine the probability of each of the possible outcomes for an event.

- Determine whether or not two outcomes are complementary.

- Use the multiplication principle to determine the probability for outcomes of two or more independent events.

- Use "randInt" on the TI calculator and "randbetween" in excel to generate a random integer between two values

- Use "countif" in excel to find out how often a particular value appears in a list.

- Evaluate the results of a simulation to find the likelihood of a particular outcome.

Chapter 12 Study Guide

1. What is the difference between independent events and mutually exclusive events? Give an example of each.

2. What does it mean if two events are complementary? Give an example.

3. Give an example of two events that are mutually exclusive but not complementary?

4. If two events are independent and both have a 90% chance of occurring.

    a. What is the probability both will occur?

    b. What is the probability one will occur and one will not?

    c. What is the probability neither will occur?

    d. What is the probability if there are 3 events of them all occurring?

5. What does the notation $P(A)$ mean? What about the notation $P(A^c)$

6. If $P(A) = 0.75$ what is $P(A^c)$?

7. Are the results of a simulation always reliable? Explain

Section 13.0: Random Number of Successes and Probability Distributions

The previous chapter was designed to explore randomness in a number of decision contexts. We modeled various examples using random number simulations. Then we analyzed the output of the simulated data to determine the observed relative frequency of critical outcomes. We also applied the multiplication rule to estimate the probability of a specific outcome. In each of the examples, we started with a specific probability, p, of something happening.

In Koala Foods, p referred to the performance standard that there was 0.5 probability the phone would be answered at 8:02 a.m. For the high school newspaper, p referred to the 0.95 probability that a writer would submit an article on time. With regard to BT Auto, the p value of interest was the 0.1 likelihood that, on a particular day, one of the workers would be out sick due to the flu. In addition, we assumed that each repetition or individual was independent of every other and had the same p value.

Thus, whether or not the phone was answered on Monday at 8:02 a.m., did not affect the probability it was answered on Tuesday at 8:02 a.m. Whether writer X met or did not meet the deadline did not influence any other writer on the team. Similarly, worker A's absence or presence on a specific day did not change the probability for each and every one of the other workers on that day or any other.

The last common element of each problem was a value n, the number of times something was repeated. You flipped a coin 43 times. Phone calls were made to Koala Foods three days in a row. There were 10 writers, each of whom was to submit an article. In the case of BT Auto there were two sets of repetitions presented in rows and columns. Each row contained the simulated data for each of the 12 workers expected to show up for work that day. For each row n is 12. Each column had 50 repetitions for an individual worker. This corresponded to the days the worker was absent or present at work. For each column n is 50.

In summary, there were four common elements to the random event being analyzed. In each case:
1. There were only two possible outcomes.
2. The p value, the probability of success, was the same for each repetition.
3. The outcome of each repetition was assumed to be independent of every other repetition.
4. There were n identical repetitions of the random event.

13.0.1 The Number of Successes—Binomial Distribution

These four basic elements and assumptions are enough to enable us to build a probability model, a formula that can be used to calculate probabilities. This formula can complement and eventually replace the detailed simulations and associated data analysis presented in the previous chapter. The model we develop in this chapter is called the **Binomial Distribution**. It has two parameters, values that need to be specified in order to apply the distribution. These parameters are n and p.

Let: n = the number of times the random event is repeated
        p = the probability that a specific outcome, called a success, occurs.
        Then: $(1-p)$ is the probability of the specific outcome *not* occurring.

Mathematicians and statisticians use standard words. They call the outcome of interest a **success** and its complement a **failure**. The terms success and failure are not meant to have either positive or negative connotations. If we are interested in studying writers who do not meet deadlines, the event of interest is

"not meeting a deadline" and would be labeled a success in our example. However, if we are interested in counting the number of writers who do meet deadlines, the event of interest "meeting a deadline" would be labeled a success.

13.0.2 Random Variables

The variable X is called a **random variable**. Calling X a **variable** means that it can take on different values. Calling X a **random** variable means that the different values of X occur at random.

Let X= the number of successes.

When we repeat the random event n times, *X, the number of successes,* has a range from 0 to n. For example if we call Koala three days in a row, the number of times someone answered the phone is a number between 0 and 3. Similarly, the number of writers who meet the deadline for a specific edition of the paper can range from 0 to 10 writers. In this chapter, the teacher in charge of the school paper is considering publishing the paper even if only 8 writers have submitted their work. Thus he will be interested in knowing.

$P(X \geq 8)$

What is the probability that 8 or more writers have met the deadline? The Binomial Distribution has a formula that can be easily input into your calculator to answer this question. Similarly the manager of BT Auto is considering scheduling spare workers. He wants to know the probability that 2 workers will be absent on a given day. If X represents the number of workers absent, the manager wants to calculate $P(X = 2)$.

13.0.3 The Number of Attempts until the 1st Success—Geometric Distribution

In some situations we are interested in how long it will take to achieve a first success. This is also a random variable. Imagine you are in charge research and development for a drug company. Assume that each time your team studies a possible drug compound there is only a 1 and 500 chance that it will lead to a drug that can be sold. As a manager, you are interested in the random variable, the number of compounds tested before the team finds a saleable drug.

Imagine a different context. You are in charge of planning space shuttle flights. Your engineers have estimated that there is a 1 in 80 chance of a catastrophic failure. What is the probability that the first failure will occur on or before the 50th shuttle flight?

Now let's return to the examples from the previous chapter, Mr. Smith of Koala Foods is interested in the first day a call at 8:10 a.m. is not answered.

Let R = the number of times an event is repeated, until we observe a specific outcome for the first time.

The smallest possible value of R is just 1. You cannot have the first unanswered phone call without calling the first time. The largest possible value of R is infinity. Koala Foods may never miss a phone call at 8:10 a.m.

The formula used to calculate probabilities for this type of problem is called the **Geometric Distribution**. For example, the Geometric Distribution can be used to calculate the probability that the first day a call at

8:10 a.m. is not answered is on the fifth day, $P(R = 5)$. The Geometric Distribution is closely related to the concept of a Geometric Series that you studied in earlier courses. In the Geometric Distribution there is just one parameter, q, the probability of the outcome of interest.

13.0.4 Chapter Examples

In the previous chapter we explored the likelihood of no one answering the Koala Foods customer service line for several consecutive days. In this chapter we explore a wider range of possible outcomes. We introduce the Binomial Distribution as a standard for modeling situations like these. We also develop a second example based on *The Lancer* newspaper scenario. This scenario has two added features. We demonstrate how to use a cumulative distribution of the Binomial Distribution to answer questions. We also apply the concept of **conditional probability**.

The second set of examples develops the Geometric Distribution. We use two contexts. The first explores the probability distribution of the first day on which a call to the Koala Foods customer service line is answered at 8:02 am. The second example explores the time between space shuttle catastrophes.

13.0.5 Probability Distributions and Relative Frequency

There is a direct link between the simulations we carried out in the previous chapter and probability distributions. Recall that each of your simulated data sets was often very different from that of your neighbor. When you pooled the classroom data, you had a much larger sample. In general, as you increase the number of times you simulate a random event, the frequency of specific outcomes more and more closely approaches the probability values that are calculated by the Binomial Distribution formula. Thus, these formulas are short cuts to estimating probabilities. They replace the need to repeat the simulations over and over again in order to obtain reliable estimates of the relative frequency of the outcome being studied. At the close of the chapter we will offer opportunities to experience a large number of repeated repetitions of the same random experiment. This will be used to demonstrate how the relative frequency approaches the theoretical probabilities calculated by the probability distribution formula.

Section 13.1: Customer Service at Koala Foods—Binomial Distribution of Successes

In the previous chapter, Mr. Smith decided the data presented by AGB Company was inconclusive. Even if the 8:02 a.m. standard of 50% of calls answered was generally met, there was a one-in-eight chance that during a 3 day period no one answered the phone. He decided this was not enough evidence to take corrective action. Before seeking more data, Mr. Smith decided he wanted to better understand the likelihood of different outcomes from calling the help line three days in a row. There are four possibilities as to the number of times the call was answered: 0, 1, 2, and 3. He wondered, what is the probability of each of these outcomes assuming that the 50% standard is generally achieved over the course of a year?
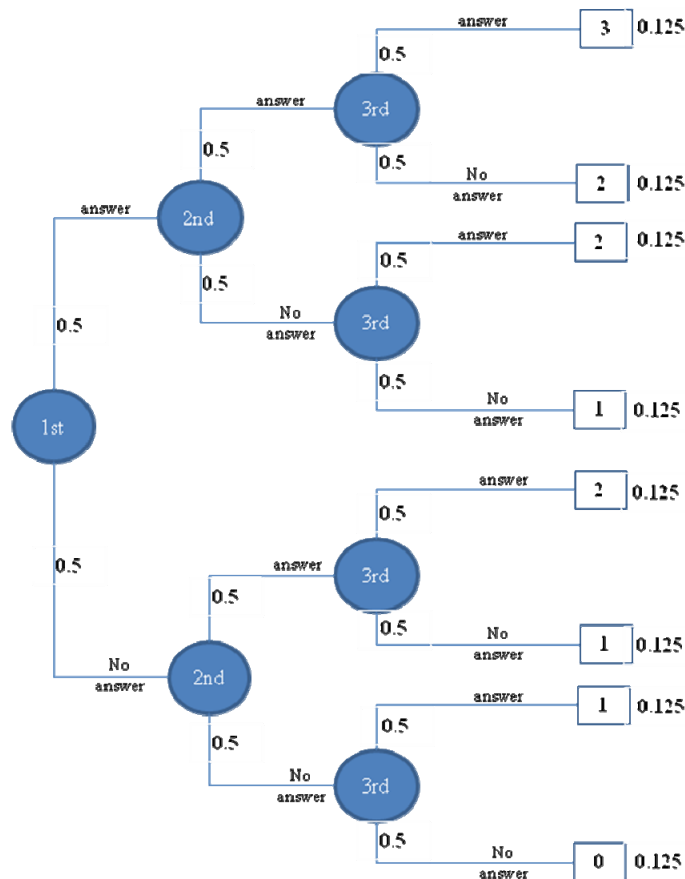

Figure 13.1.1: Probability tree—call three days in a row at 8:02 a.m.

Mr. Smith constructed a probability tree to lay out the sequence of events that might occur when calling three days in a row. Each day there are two possible outcomes; the call is answered or it is not answered. The standard set for answering calls at 8:02 a.m. was 50%, so these two possibilities are equally likely. To develop the tree, he multiplied the number of outcomes each day for three days and determined there was a total of 8 possible sequences. This tree is depicted in Figure 13.1.1. At the end of each path, he recorded the probability of that path or sequence of events. He also recorded in a box the number of calls that were answered on that path. The top path in the figure corresponds to the call being answered every day. The probability of that happening is 0.5 · 0.5 · 0.5 or 0.125. In this path, a total of three calls were answered. Similarly, at the bottom of the figure is the sequence of days with no answer each day. It too has a probability of occurrence of 0.125 with zero calls answered. In fact, the probability of each and every

sequence of three daily calls has a probability 0.125 because of the equal probability of an answer and no answer.

In thinking about the possible outcomes, Mr. Smith realized he was more interested in the number of times the phone was answered at 8:02 a.m. than in the specific sequence of what happened each day. The number of times the phone was answered is a random variable. To help with the analysis, he recorded at the end of each sequence the number of calls that were answered. He first analyzed the probability of the phone being answered exactly once. He identified the three distinct paths listed below.

Path	Day			Path Probability
	1	2	3	
1	answered	not answered	not answered	.0125
2	not answered	answered	not answered	.0125
3	not answered	not answered	answered	.0125

Table 13.1.1: One call answered out of three—mutually exclusive paths

Each of these paths has a probability of 0.125. These paths are considered mutually exclusive events even though pieces of the path overlap. They are mutually exclusive in that if one 3-day sequence occurred the other 3-day sequence did not occur. Because they are mutually exclusive, he can add probabilities together to determine the probability a call was answered on exactly 1 day:
$$P(X = 1) = 0.125 + 0.125 + 0.125 = 3(0.125) = 0.375$$

He next analyzed the probability of the phone being answered exactly twice. He again identified the three distinct paths listed below.

Path	Day		
	1	2	3
1	answered	answered	not answered
2	not answered	answered	answered
3	answered	not answered	answered

Table 13.1.2: Two calls answered out of three—mutually exclusive paths

Again, each of these mutually exclusive events has a probability of 0.125.
$$P(X=2) = 0.125 + 0.125 + 0.125 = 3(0.125) = 0.375$$

In summary, he noted that if none of the three calls were answered, there was cause for concern. However, he would need more convincing evidence before taking action. He also observed that the likelihood of 1 call or 2 calls being answered was exactly the same. Thus if only 1 out of 3 had been answered which is below the corporate standard, there would not yet be cause for concern.

13.1.1 Four Days at 8:02 a.m. with a Standard of 50%

Mr. Smith decided to use the same approach to explore a four-day series of calls. The extra day doubles the size of the tree, because now there are 16 possible sequences. They are depicted in Figure 13.1.2. Each of the 16 sequences is equally likely because the outcomes "answer" and "no answer" both have a probability 0.5. The top path corresponds to the call being answered every day. Its probability of occurrence is $(0.5)^4$ or 0.0625. At the other extreme, no calls answered also has a probability of 0.0625.
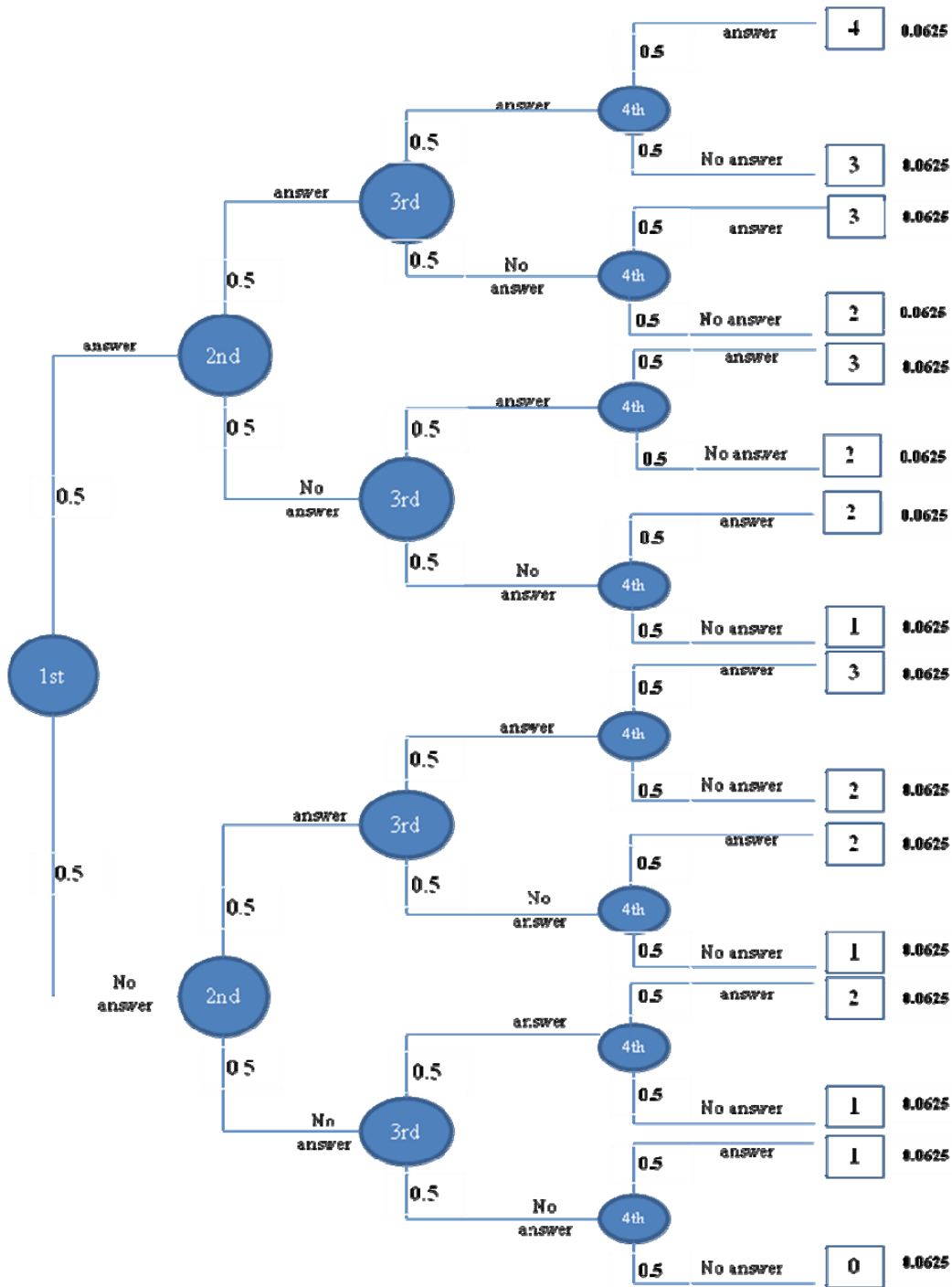
Figure 13.1.2: Probability tree—call four days at 8:02 a.m.

Mr. Smith thought about the possibility of four days in a row with no answer. If the 50% standard were generally being met, this would not likely occur. It has only a 0.0625 probability. If this were to happen, he was thinking that he would call in the help line manager and ask for an explanation.

There is no single criterion to determine if there is enough evidence that a problem exists. The alternative interpretation is the results are simply random fluctuations and no problem exists. Three common criteria

that are considered are 0.1, 0.05 and 0.01. The first criterion says that if there is less than a 10% chance that this poor outcome would occur randomly, the manager should take a closer look and take action to correct a problem. The 0.05 criterion is a higher standard of evidence. This says that if there is less than a 5% or 1-in-20 chance that this poor outcome would occur randomly, the manager should take a closer look and take action. Similarly, a 0.01 criterion involves a 1-in-100 chance of occurrence.

The decision as to what criterion to use involves balancing two points of view with regard to taking action or not taking action to correct a potential problem.
- What is involved in taking action to correct a possible problem? Is there a significant cost associated with taking action? Could taking action when no problem really exists cause harm?
- What are the consequences of waiting before taking action if there is, in fact, a problem? What is the cost of gathering more data before making a decision to act?

If taking action has major costs and consequences, such as firing someone, then the 0.01 criterion might be most appropriate.

Q1.   Describe another situation in which taking action to correct a perceived problem can cause harm if the problem does not really exist.

Q2.   Describe a situation in which delaying action to correct a problem can cause harm.

A scenario of no answers in four days falls between the 0.10 and 0.05 criteria. If Mr. Smith were to use a 5 day sample, the probability of no answer is $(0.5)^5$ which is 0.03. This would definitely cause him to act to correct a perceived problem of not answering the help line at 8:02 am.

Now let's look at some of the other possible outcomes of the phone check to see if call answerers are ready by 8:02 am. If the 50% standard was generally met, then 2 out of 4 calls answered would exactly match the standard. If only 1 out of 4 was answered, it would be below the standard. However, Mr. Smith was unsure of how strong an indication of a problem this would be. There were 4 different sequences that involved one answered call. Thus, the probability of this happening was simply

$$P(X = 1) = 4(.0625) = 0.25$$

A poor response that has a 1-in-4 chance of random occurrence should not produce any immediate concerns, nor cause him to act.

Mr. Smith looked back at the process he had been using to perform the calculations. He liked the pictorial representation but knew that when the number of days increased to 5 or more, it would not be possible to contain the picture on one page. He realized that his approach seemed to involve a standard three step process:
1. Determine the probability of a specific sequence of events
2. Count the number of different paths, each with the same probability that produce the same total number of answered calls.
3. Multiply the path probability by the number of paths with the same total of positive outcomes.

Mr. Smith already knew how to do step 1. It involved just multiplying the individual event probabilities. He wondered if there was a simple formula to help count the different paths. His assistant Sorel Ward recalled a combination formula she had learned in high school and thought it might help. She reasoned as follows.

"There are 4 paths that result in one answered call. These four paths correspond to the four
possible different days. In the language of selecting combinations this is equivalent to asking how
many ways are there of picking 1 (day) out of 4 (days). This is called 4 choose 1 and is
represented as $_4C_1$."

We have used the capital letter X, to represent the general random variable. We use the lower case x to
represent a specific value of the random variable in the next equation. The general formula for
combinations for choosing x items out of n possibilities is:

$$_nC_x = \frac{n!}{(n-x)!(x)!}$$

She then considered the case in which calls were answered 2 days out of the 4:

$$_4C_2 = \frac{4!}{(4-2)!(2)!} = \frac{24}{(2)(2)} = 6$$

The number of paths is 6 which is the same as the value determined applying the combinations formula.
Her reasoning seemed solid and the numbers matched.

13.1.2 Four Days at 8:05 a.m. with a Standard of 70%

Mr. Smith decided to use this same approach to determine the probability distribution of outcomes for
calls placed at 8:05 a.m. The corporate standard was to answer these calls 70% of the time. He started
with a probability tree as illustrated in Figure 13.1.3. The layout of sequences in this tree is exactly the
same as before. The number of paths that result in 1 call being answered is again 4. The number of paths
that result in 2 calls being answered is again 6. The differences in the tree are the probabilities associated
with each branch. As a result the end probability of each sequence of four days is different than in Figure
13.1.2. More interestingly, the probability of each sequence is not the same.

In the previous probability tree, the 50% standard means that answer and no answer are equally likely
random events. Now the probability of receiving an answer is 0.7 and of receiving no answer is 0.3. Let's
look at the two extreme paths of the tree. The top path corresponds to all four calls being answered. It has
a probability of occurrence that is $(0.7)^4$ or 0.2401. The bottom path corresponds to no calls being
answered. It has a probability of $(0.3)^4$ or 0.0081. If no calls were answered, Mr. Smith would be sure that
the call center was not generally meeting the 70% standard. He would act immediately to correct the
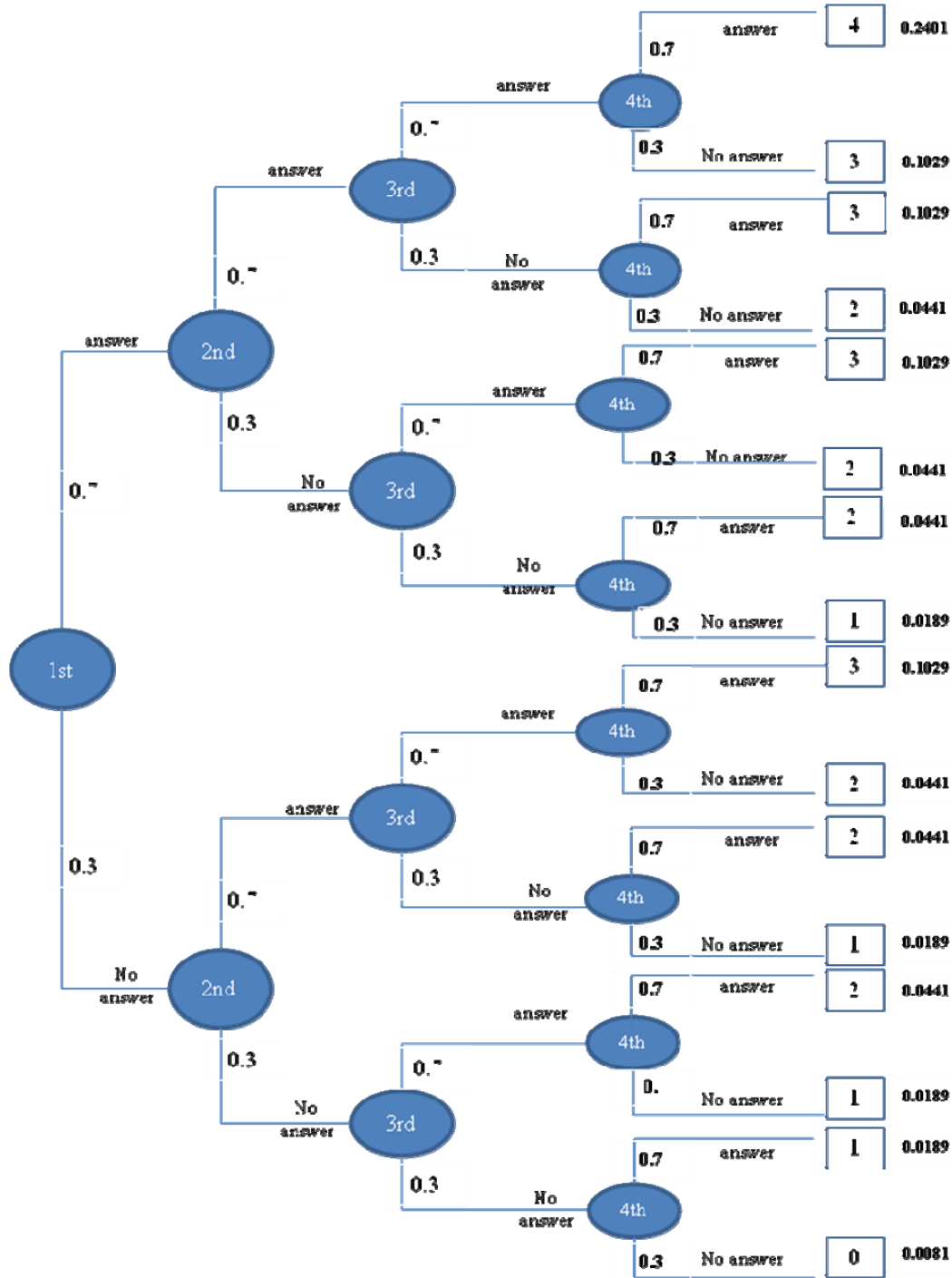problem.

Figure 13.1.3: Probability tree—call four days at 8:05 a.m.

Mr. Smith looked closely at the probability tree and noticed that specific probabilities occurred multiple times. For example, each sequence that resulted in 1 call being answered had the same probability of 0.0189. Because of the commutative law of multiplication, the probability of each path is the same. In each case there is one call answered and three calls not answered. The probability of one call being answered is just 0.7. The probability that three calls are not answered is 0.3 raised to the third power.

$(0.7)(0.3)(0.3)(0.3)$
$= (0.3)(0.7)(0.3)(0.3)$
$= (0.3)(0.3)(0.7)(0.3)$
$= (0.3)(0.3)(0.3)(0.7)$
$= (0.7)(0.3)^3$
$= 0.0189$

Now, because there are four different paths that result in exactly one of four calls being answered the total
probability is

$\quad$ $4(0.0189) = 0.0756$

This probability is below the 10% criterion for random fluctuation. Mr. Smith would, therefore, begin
looking more closely at the starting time performance of the call center.

This calculation paralleled the three step approach described earlier:
1. Determine the probability of a specific sequence of events.
2. Count the number of different paths each with the same probability that produce the same total
    number of answered calls.
3. Multiply the path probability by the number of paths with the same total of positive outcomes.

Mr. Smith repeated the process for all possible values of x, the number of calls answered. He organized
the results in Table 13.1.3. A total of 3 calls out of 4 answered is the most likely occurrence if the 70%
standard was generally being met. Answering 2 out 4 is below the standard. However, it would not be an
uncommon occurrence and Mr. Smith would not take action if that happened.

Number of Calls Answered	Number of Calls Not Answered	Path Probability	Number of equally likely paths	Total Probability
0	4	$(.7)^0(.3)^4 = .0081$	$_4C_0 = 1$	.0081
1	3	$(.7)^1(.3)^3 = .0189$	$_4C_1 = 4$	.0756
2	2	$(.7)^2(.3)^2 = .0441$	$_4C_2 = 6$	.2646
3	1	$(.7)^3(.3)^1 = .1029$	$_4C_3 = 4$	.4116
4	0	$(.7)^4(.3)^0 = .2401$	$_4C_0 = 1$	.2401
x	$4 - x$	$(.7)^x(.3)^{4-x}$	$_4C_x = \dfrac{4!}{(4-x)!(x)!}$	$_4C_x(.7)^x(.3)^{4-x}$

Table 13.1.3: Four days at 8:05 a.m.—probability distribution of the number of calls answered

Based on the above analysis, Mr. Smith, with Sorel Ward's help, believed he could write a general
formula to calculate P(x) as follows:
$$P(x) = {_4C_x} [(.7)^x (.3)^{4-x}]$$

Let's review this formula. The first part is the number of combinations (sequences) that lead to the same
exact number of answered calls, x. It is the combinations of 4 choose x. The second term represents the
probability of exactly x answered calls and $n - x$ unanswered calls. Because of the commutative product
law, this term is exactly the same irrespective of the daily sequence of answered and unanswered calls as
long as the total answered is the same x. In each case 0.7 is raised to the power of x, the number of
answered calls. And 0.3 is raised to the power of $(4 - x)$ the number of unanswered calls.

13.1.3 Four Days at 8:10 a.m. with a Standard of 90%

Mr. Smith is now interested in calculating the probabilities for four calls made at 8:10 a.m. when the corporate standard is 90%. This will guide his thinking as to when he should take action and when he should wait and gather more data. In looking at the above formula, the only changes that need to be made are to replace the 0.7 and its complement 0.3 with the new standard of 0.9 and its complement 0.1. He does not feel a need to build a tree to confirm his reasoning. He asked Sorel Ward to complete the table below:

Number of Calls Answered	Number of Calls Not Answered	Path Probability	Number of equally likely paths	Total Probability
0	4	$(.9)^0(.1)^4 = .0001$	$_4C_0 = 1$	.0001
1	3		$_4C_1 = 4$	
2	2		$_4C_2 = 6$	
3	1		$_4C_3 = 4$	
4	0		$_4C_0 = 1$	
x	$4 - x$	$(.9)^x(.1)^{4-x}$	$_4C_x = \dfrac{4!}{(4-x)!(x)!}$	$_4C_x(.9)^x(.1)^{4-x}$

Table 13.1.4: Four days at 8:10 a.m.—probability distribution of the number of calls answered

13.1.4 Five Days for Different Times and with Different Standards

Mr. Smith was thinking that all tests carried out by AGB should last five days. However, Koala has different standards for 8:02, 8:05, and 8:10 a.m. As a result, he plans to use different criterion for determining whether or not there is strong evidence that a particular standard is not being met. Mr. Smith asked Sorel Ward to apply the Binomial Distribution for each standard. She needed to generalize the equation that was developed earlier. She reasoned as follows:

1. Each time period has a different standard that can be represented by p, the probability the call is answered.
2. The probability that a call is not answered is just the complement of this probability. It equals (1-p).
3. In a five day test, the probability that exactly x calls are answered should therefore be:
   $P(x) = {_5C_x}[(p)^x(1-p)^{5-x}]$.

The first part is the number of combinations (sequences) that lead to the same exact number of answered calls, x. It is the combination of 5 choose x. The second term represents the probability of exactly x answered calls and n-x unanswered calls. Because of the commutative product law, this term is exactly the same for any value of p irrespective of the daily sequence of answered and unanswered calls as long as the total answered is the same x. In each case p is raised to the power of x, the number of answered calls. And (1-p) is raised to the power of (5-x) the number of unanswered calls. Sorel used this formula to complete Table 13.1.5.

Number of Calls Answered	5 Day Test of Calls Answered Time of Call and Koala Foods Standard		
	8:02 a.m. p = 0.5	8:05 a.m. p = 0.7	8:10 a.m. p = 0.9
x	P(x)		
0	0.031	0.002	0.000
1	0.156	0.028	0.000
2	0.313	0.132	0.008
3	0.313	0.309	0.073
4	0.156	0.360	0.328
5	0.031	0.168	0.590

Table 13.1.5: Five days—probability distribution of the number of calls answered

Mr. Smith reviewed the table and discussed with Sorel the circumstances in which he would take action.

Q3.   Assume AGB conducts a five day test at 8:02 a.m. Under what circumstances should Mr. Smith take corrective action?

Q4.   Assume AGB conducted a five day test at 8:02 a.m. Under what circumstances should Mr. Smith being seriously concerned but delay taking action until more data comes in?

Q5.   Assume AGB conducts a five day test at 8:05 a.m. Under what circumstances should Mr. Smith take corrective action?

Q6.   Assume AGB conducted a five day test at 8:05 a.m. Under what circumstances should Mr. Smith being seriously concerned but delay taking action until more data comes in?

Q7.   Assume AGB conducts a five day test at 8:10 a.m. Under what circumstances should Mr. Smith take corrective action?

Q8.   Assume AGB conducted a five day test at 8:10 a.m. Under what circumstances should Mr. Smith being seriously concerned but delay taking action until more data comes in?

Section 13.2: Binomial Distributions: Expected Value—What to Expect?

In the previous chapter, you experimented with flipping a fair coin numerous times. Now consider just two flips of a fair coin. The number of heads can take on the values 0, 1, and 2. Using the Binomial Distribution, the corresponding probabilities are 0.25 (TT), 0.50 (TH or HT), and 0.25 (HH). The expected value of any random variable is simply a weighted sum. First, you multiply each of the possible number of successes by the probability that it occurs. Then, you add those products together. Thus, the expected number of successes for two flips is:

$$E(X) = \sum x\, P(x) = 0\,(0.25) + 1\,(0.50) + 2\,(0.25) = 1.$$

We use the capital letter X, to represent the general random variable. We use the lower case x to represent a specific value of the random variable. Thus, on the left hand side of the above equation we use the capital X. When we take the weighted sum of the possible values, we use the lower case x.
There is a 0.5 likelihood that when flipping the coin twice, you will have one head out of two, which is the expected value. There is also a 0.5 probability of observing some other value. Now consider the case of five days of calling Koala at 8:02 a.m. We calculated the probabilities when $p = 0.5$ in the previous section.

$$P(X = 0) = 0.031$$
$$P(X = 1) = 0.156$$
$$P(X = 2) = 0.313$$
$$P(X = 3) = 0.313$$
$$P(X = 4) = 0.156$$
$$P(X = 5) = 0.031$$

In the above equations we use the capital X. For example, the term $P(X = 2)$ refers to the probability that the random variable X equals the specific value 2.

To determine the expected value, we again multiply each of the possible number of successes in a week by the probability that it occurs. Then, we add those products together. Thus, the expected number of calls answered per week is:

$$0(.031) + 1(.156) + 2(.313) + 3(.313) + 4(.156) + 5(.031) = 2.5$$

The expected number of answered calls is 2.5. Obviously, it is not possible to get this value since the number of times the call was answered must be an integer value. If that is the case, what does the term expected value of this random variable really suggest? The answer relates to understanding what happens as you repeat a random experiment over and over again. The calculated average of larger and larger samples will be very close to the expected value.

Above we calculated the expected value by using the basic weighted formula for expected value. The general formula for the expected for the Binomial Distribution is given below.

Let n = number or repetitions
        p = probability of success
$$E(X) = np$$

This formula makes intuitive sense. For example, if $n = 10$ and $p = 0.5$, the expected value is 5. In other words the long-term average will be five successes out of ten tries. This formula is not limited to p values of 0.5. For example, the Koala standard for 8:05 a.m. is that 70% of the calls should be answered. Using

this higher standard, the expected value of the number of calls answered for a five day week is just 5 times 0.7 or 3.5 calls.

Section 13.3: *The Lancer*—What If We Publish an Incomplete Paper?

In the previous chapter, we also explored the challenge of publishing the school newspaper, *The Lancer*, on time. Recall there were ten different writers, five editorial writers and five reporters. Even if each of the writers were 95% reliable in meeting the deadline, there was only a 60% chance the paper would be published on time. This was determined by using the multiplication rule for independent events: 0.95 raised to the 10th power is only 0.60.

Mr. Mitchell is committed to publishing the paper on time. He is considering publishing the paper even if all of the articles have not been submitted. He would like the paper to be on time at least 95% of the time. Mr. Mitchell is considering going to press even if only 9 articles have been submitted. He wonders if this will enable him to meet his goal of at least 95% of the time. Mr. Mitchell is considering increasing the number of writers above 10. He plans to evaluate the impact of changing *n*, the number of writers.

In the previous section we developed the Binomial Distribution for different numbers of repetitions. We considered 3 days, 4 days and 5 days. We now generalize the distribution for all values of *p*, the probability of success, and all values of *n*, the number of repetitions.

Let x = number of on time submissions,
        n = number of repetitions, and
        p = probability of success each time.
        $P(X = x) = {}_nC_x\,(p)^x\,(1–p)^{n-x}$

The first term is the number of different sequences that will result in exactly *x* successes and *n-x* failures. The rest of the equation is the probability any individual sequence has exactly *x* successes and *n-x* failures.

Now let's return to the specifics of Mr. Mitchell's decision. In terms of probability, we can express the probability of on time publication with a policy of publishing with only 9 articles as follows:
        $P(X \geq 9) = P(X = 9) + P(X = 10)$

This includes two terms. If 9 or 10 articles are submitted on time, the paper is published on time. Using the Binomial formula
        $P(X \geq 9) = P(X = 9) + P(X = 10)\ = {}_{10}C_9\,(0.95)^9\,(1 - 0.95)^1 + {}_{10}C_{10}\,(0.95)^{10}(1 - 0.95)^0$
        $P(X \geq 9) \approx 0.315 + 0.599 = 0.914$

This policy is close to his goal. If he were to use this policy, sometimes the paper would have ten articles and other times only nine. He is interested in knowing the proportion of editions published on time that are complete. This is an example of *conditional probability*: What is the probability that the issue is complete, *given* that it goes to press on time? The statement, "It goes to press on time," is a given *condition*. This is the source of the term *conditional* probability. To answer this question we write the following conditional probability statement:
        $P(X = 10$, given that the paper published on time$) = P(X = 10 \mid X \geq 9)$

In words this means that, given there were 9 or more articles submitted on, what is the probability that were in fact exactly ten articles submitted on time. The conditional probability formula is:
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

The denominator represents the probability of the given event, that the paper is published on time in this case.

$$P(X = 10 \mid X \geq 9) = \frac{P(X = 10 \cap X \geq 9)}{P(X \geq 9)}$$

The numerator $P(X = 10 \cap X \geq 9)$ contains the intersection of two events with regard to X. However, the only way that X can *both* equal 10 and be greater than or equal to 9 is if X equals 10. Therefore,

$$P(X = 10 \mid X \geq 9) = \frac{P(X = 10)}{P(X \geq 9)} = \frac{0.599}{0.914} = 0.655$$

When the paper is published on time, there is a little less than two-thirds likelihood that paper is complete with all ten articles.

13.3.1 Meeting the 95% Goal for Publishing on Time

The above policy would not enable Mr. Mitchell to meet his 95% on time goal. He therefore considers expanding the policy and allowing the paper to be published with as few as 8 articles. We now need also to find the probability that exactly eight articles are available on time.

$$P(X = 8) = {}_{10}C_8 \, (0.95)^8 (1-0.95)^2 = .075$$

Now, the probability of eight or more articles being available on time is

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) = 0.075 + 0.315 + 0.599 = 0.988$$

Thus if Mr. Mitchell go to press with just 8 articles, the paper will be published on time almost 99% of the time.

When the paper is published on time, it will have 8, 9, or 10 articles. He is now concerned how frequently the on time paper will have as few as 8 articles. This is another conditional probability question. To answer it, we write the following conditional probability statement:

$$P(X = 8 \mid \text{the paper is published on time}) = P(X = 8 \mid X \geq 8)$$

Applying the conditional probability formula, we have

$$P(X = 8 \mid X \geq 8) = \frac{P(X = 8 \cap X \geq 8)}{P(X \geq 8)} \,.$$

Now, the only way that X can be *both* greater than or equal to 8 and equal to 8 is if it is exactly equal to 8. In other words, $P(X = 8 \cap X \geq 8) = P(X = 8)$. Thus,

$$P(X = 8 \mid X \geq 8) = \frac{P(X = 8)}{P(X \geq 8)} \approx \frac{0.075}{0.988} \approx 0.076$$

Approximately, one out of every 13 on-time editions will have only eight articles, because $1/.076 \approx 13.2$

13.3.2 Twelve Writers—But Less Reliable

This year Mr. Mitchell had an unusually large group of students interested in working on *The Lancer*. He selected the best 12 of the group. However, this group seems to him to be much less reliable than past groups. He estimates, based on past experience, that each of the individuals in this group will meet the

paper deadline 80% of the time. Recall that the expected value of the Binomial Distribution is $E(X) = np$, where n is the number of possibilities (writers, in this case), and p is the probability of a success (meaning the writer meets the deadline in this case). So, the expected value of the number of writers meeting the deadline in this case is

$E(X) = 12(0.8) = 9.6$.

Thus on average there would be fewer than ten articles ready on time for publication. He once again is ready to consider publishing the paper with fewer than ten articles. To explore his options, he decides to use a cumulative table of the Binomial Distribution instead of performing calculations. The cumulative distribution sums all the probabilities up to and including a specific value of X. It is written as $P(X \leq x)$. For example if $x=8$, the cumulative probability is the probability that the random variable takes on any value from 0 up to 8. It involves summing up the probabilities as in the equation below.

$P(X \leq 8) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$
$\qquad + P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)$

$P(X < 8) = 0 + 0 + 0 + 0.0005 + 0.0033 + 0.0155 + 0.0532 + 0.1329 = 0.2054$

With calculators, it is not necessary to perform each step of this calculation. The cumulative distribution is a calculator function: binomcdf(. Table 13.3.1 lists $P(X \leq x)$ for every possible value of x from 0 to 12 for a range of values for p. For example, if $p = 0.8$, the probability that the number of successes will be 8 or fewer is 0.205. If $p = 0.8$, the probability that the number of successes will be 10 or fewer is 0.725.

Mr. Mitchell is, however, more interested in probability statements of the form $P(X \geq x)$. This can be determined from the table by using the principle of the complement. Thus, for example, the complement of nine or more is eight or fewer. Applying the principle of the complement, the probability of nine or more is equal to 1 minus the probability of eight or less.

$P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.205 = 0.795$

Mr. Mitchell found the table to be very helpful. He asked one of the students to calculate this probability by determining the probability of each individual outcome.

$P(X \geq 9) = P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$
$_{12}C_9 (0.8)^9 (1-.8)^3 + {}_{12}C_{10} (0.8)^{10}(1-.8)^2 + {}_{12}C_{11} (0.8)^{11} (1-.8)^1 + {}_{12}C_{10} (0.8)^{12}(1-.8)^0$

Q1.   What are the individual values and the total?

After watching the student go through each calculation, Mr. Mitchell was glad to have this simple lookup table.

In general, when X and x must be integers, $P(X \geq x+1) = 1 - P(X \leq x)$.

	p values			
	0.75	**0.8**	**0.85**	**0.9**
$X \leq 0$	0.000	0.000	0.000	0.000
$X \leq 1$	0.000	0.000	0.000	0.000
$X \leq 2$	0.000	0.000	0.000	0.000
$X \leq 3$	0.000	0.000	0.000	0.000
$X \leq 4$	0.003	0.001	0.000	0.000
$X \leq 5$	0.014	0.004	0.001	0.000
$X \leq 6$	0.054	0.019	0.005	0.001
$X \leq 7$	0.158	0.073	0.024	0.004
$X \leq 8$	0.351	0.205	0.092	0.026
$X \leq 9$	0.609	0.442	0.264	0.111
$X \leq 10$	0.842	0.725	0.557	0.341
$X \leq 11$	0.968	0.931	0.858	0.718
$X \leq 12$	1.000	1.000	1.000	1.000

Table 13.3.1: Binomial Cumulative Distribution for *n* =12

Thus, to explore his options, Mr. Mitchell is looking for the value of *x* such that
$$P(X \geq x+1) \geq 0.95.$$

Using the principle of the complement, we are looking for the largest value of *X* such that
$$P(X \leq (x–1) \leq 0.05.$$

Mr. Mitchell looked at the cumulative probabilities for both *X* less than or equal to seven and *X* less than or equal to 6. He found the following:
$$P(X \leq 7) = 0.073 \;\rightarrow\; P(X \geq 8) = 1- 0.073 = 0.927$$
$$P(X \leq 6) = 0.019 \;\rightarrow\; P(X \geq 7) = 1- 0.019 = 0.981$$

Thus if he wants the Lancer to be published on time at least 95% of the time, he will have to settle for a policy of publishing with seven or more articles.

Q2.   Do you think the Lancer should be published with 7 articles when the basic design was made for ten articles? What standard would you recommend and how often would you be able to meet that standard?

Mr. Mitchell was concerned that the paper might seem skimpy if there were only seven articles. He was thinking about a strategy that might help fill in the gap. One thought that came to mind is that sometimes there might be a surplus of articles for an on-time edition. There are twelve writers and only ten articles can be published in an issue. In that case, he could withhold one or more of these articles for a subsequent edition. Even though this year's writers were not as reliable as in the past, it was possible that 11 or even 12 of the writers might submit articles on time. He was therefore interested in determining the likelihood of this happening. He looked at the table for the value of *r* equal to 10 and applied the principle of the complement.
$$P(X \geq 11) = 1 – P(X \leq 10) = 1 - 0.725 = 0.275$$

He found that more than 27% of the time, he would have 11 or more articles. He could then publish a full-edition paper with ten articles. He would hold onto the one or two extra articles to use them later in the year if fewer than ten articles are submitted.

Mr. Mitchell estimated that the reliability of the writers was 0.8 based on experience with similar teams in the past. He had some concern about his estimate. He also knows that if he aggressively monitors their work habits, he may be able to increase their reliability.

Q3.   Assume the individual reliability is only 0.75. What should he use for the minimum number of articles so he can meet the 95% goal? What policy would you recommend?

If he supervises closely, he can increase the reliability to 85% after a couple of issues of the Lancer.

Q4.   Assume the individual on time reliability reaches 0.85. What standard should he use for the minimum number of articles so he can meet the 95% goal?

Mr. Mitchell believes that by the second semester, it may be possible to increase the reliability to 90%. He was wondering what would be the benefit of his efforts to increase reliability.

Q5.   Assume the individual on time reliability reaches 0.90. What standard should he use for the minimum number of articles so he can meet the 95% goal?

Section 13.4: Worker Absenteeism and Spare Workers at BT Auto Industries

BT Auto Industries is a small manufacturer based in Detroit, Michigan. It recently received several large orders and needs 25 workers on the job each day. Management wants to know how many spare workers to hire during the flu season. The last time they carried out a study, they simulated 10 weeks of flu season, 50 days. The process of simulation was time consuming. Kenneth Young is the chief industrial engineer. He knows that the Binomial Distribution can be used to perform the analysis much more quickly and effectively.

He knows if he does not have spare workers, most days he will be short workers. The likelihood that 25 workers out of 25 show up during flu season is easily calculated with the Binomial Probability Distribution formula
$$P(X = 25) = {}_{25}C_{25}(0.9)^{25}(1 - 0.9)^0 = 0.072$$

It is worse than he thought. There is only a 7.2% chance everyone will show up for work. That means on average 13 out of every 14 days he will be short workers. He next considers having just 1 spare worker available every day. This worker may also not show up if he has the flu. He will have enough workers if at least 25 of these workers and spares come to work.
$$P(X \geq 25) = P(X = 25) + P(X = 26)$$
$$= {}_{26}C_{25}(0.9)^{25}(1 - 0.9)^1 + {}_{26}C_{26}(0.9)^{25}(1 - 0.9)^1 = 0.187 + 0.065 = 0.252$$

This is a significant improvement but nowhere near the 95% goal. He also sees that the calculations will continue to get more complex as he adds spare workers. He will need to add together more and more terms. His trusty TI calculator can simplify this if he uses the concept of the complement and the Cumulative Distribution Function of the calculator. The complement of the event that 25 or more workers show up is that 24 or fewer come to work. Thus, the last calculation can be performed as follows.
$$P(X \geq 25) = 1 - P(X \leq 24)$$
$$= 1 - \text{binomcdf}(26, 0.9, 24) = 1 - 0.748 = 0.252$$

The first parameter within the parenthesis is n, the number of workers including spares. In this case it is 26. The second parameter is 0.9, the probability that a worker shows up. The third parameter is the value of interest, 24 workers or less. He was also interested in the average number of workers that would come in for each policy. Recall that the expected value is simply $n \cdot p$. With 25 workers, the average number that shows up is 22.5. Kenneth Young asked his assistant, Jay Trust, to continue to evaluate different numbers of spare workers by completing Table 13.4.1.

Spare Workers	Total number of workers	$P(X \leq 24)$	$P(X \geq 25)$	Average number that show up ($n \cdot p$)
0	25	0.928	0.072	22.5
1	26	0.749	0.251	23.4
2	27			
3	28			
4	29			
5	30			
6	31			
7	32			

Table 13.4.1: Spare workers and probability of having at least 25 workers

Q1.   How many spare workers are needed to reach the goal of 95% of the time there are at least 25 workers in the manufacturing plant?

Q2.   On average how many workers will show up for work under this policy? How many will not actually be used?

Q3.   Under this policy, what proportion of days will they still be short exactly 1 worker? On average how many days during 50 day flu season will they be short exactly 1 worker?

Q4.   Under this policy, what proportion of days will they still be short 2 or more workers?

Q5.   The company is considering a slightly lower standard than 95%? What would the standard have to be in order to reduce the number of spare workers by 1?

Q6.   The Center for Disease Control has been tracking the flu season. It has been worse than in past years. During the peak 3 weeks of the flu season, the CDC forecasts that 15% of workers will be out sick on any random day. How many spare workers should they schedule for the peak flu season to meet the 25 worker goal 95% of the time?

Section 13.5: First Time Something Happens: Geometric Distribution—Koala Foods 8:10 a.m.

Until now, the examples have explored the random number of total successes. In this section, we are interested in the number of times something is repeated until the first success or the first failure. For example, you might be interested in the first day that AGB finds a call is not answered at 8:10 a.m. In the case of a space shuttle, NASA was interested in the first time a Shuttle is destroyed in flight. The set of examples in this section develop the probability distribution for this class of problems. The formula for this class of problems is called the **Geometric Distribution**. It is closely related to the concept of a Geometric Series that you studied in earlier courses.

We use two contexts to develop and explore this distribution. The first explores the probability distribution of the first day on which a call to the Koala Foods customer service line is answered. The second example introduces a totally new context. It explores the probability distribution for space shuttle disasters.

We will again consider the quality control concern of Koala Foods. In the previous chapter, we investigated the probability that x successes occurred. Success was defined as the call reaching a customer service representative at a specified time. This gave rise to the Binomial Distribution. In this section, we will investigate the probability that the first time an 8:10 a.m. call is not answered occurs on the rth day that we call. This will help us develop an understanding of a new distribution that describes the first time an event occurs. This pattern is called the **Geometric Distribution**. It arises in numerous contexts. For example, in the pharmaceutical industry, managers are interested in the first compound of many they have tested that seems to address a specific disease. An outcome of interest can be a negative event. NASA managers were concerned with predicting probabilistically the first time they would have a catastrophic event during a shuttle mission.

After many weeks of working with the call center staff, Mr. Smith of Koala Foods is convinced that the 90% standard for 8:10 a.m. is being achieved. Nevertheless, he understands that there is still a 10% chance a call will not be answered. He wonders if AGB were to run a daily 8:10 a.m. test, what would be the first day a call was not answered? He begins by asking and answering a series of questions.

- What is the probability of not reaching a customer service representative at 8:10 a.m. on the first day? This he realizes is just 0.1, the likelihood that a call is not answered.

- What is the probability of reaching a customer service representative on the first day, and not reaching one on the second day? The likelihood that a call is answered the 1st day is 0.9. The chances of no answer the 2nd day is still, 0.1. The likelihood of this sequence is just the product of these two numbers, 0.9*0.1. This equals 0.09.

- Next, what is the probability of reaching a customer service representative at 8:10 a.m. on the 1st day and 2nd day, but not reaching one on the 3rd day? The likelihood of AN answer on the 1st day is 0.9. The same probability applies to the 2nd day. The chances of no answer the 3nd day is still 0.1. The likelihood of this sequence is just the product of these three numbers, 0.9*0.9 *0.1. This equals 0.081.

He reviews the questions and answers and is interested in developing a common pattern. Mr. Smith decides to approach the problem like he did with the Binomial Distribution. There he started by counting the number of answered calls over three days. For this problem, he first develops a table and then a

probability tree. In Table 13.5.1, Mr. Smith lays out each sequence discussed above. In Figure 13.5.1 he builds a probability tree but stops at 5 days of calls.

	Day		
Sequence	1	2	3
1	not answered		
2	answered	not answered	
3	answered	answered	not answered

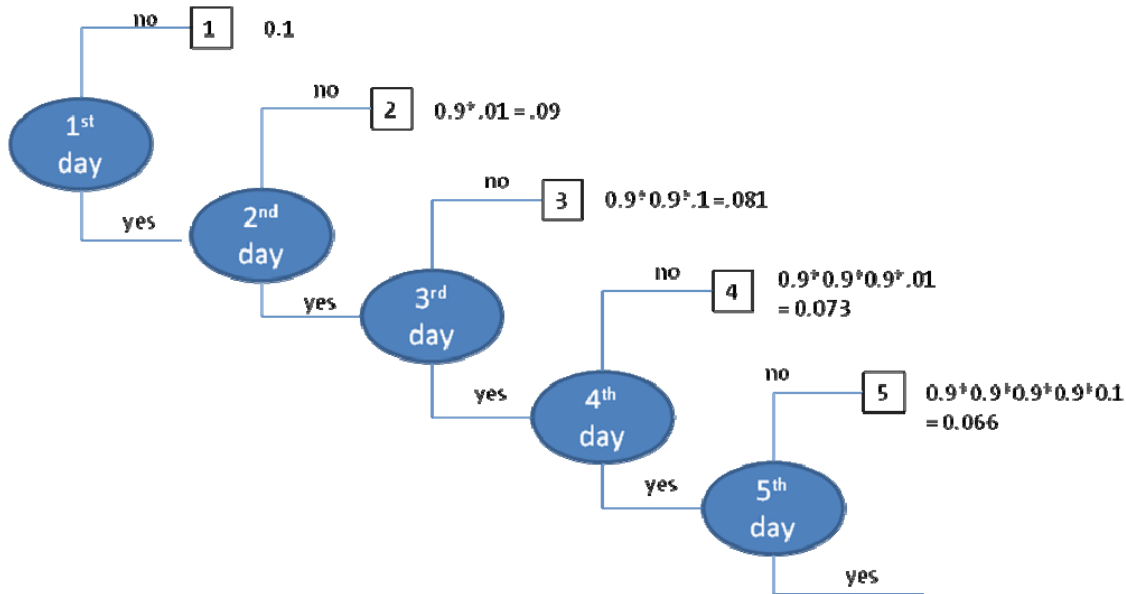**Table 13.5.1**: Sequence of days until call is not answered for 1st time

Figure 13.5.1: Probability tree for sequence of days until call is not answered for 1st time

Mr. Smith compared this table and figure to the earlier ones, Tables 13.1.1 and Figure 13.1.1, He is struck by how different they are. Here are his main observations about the table and figure.

- Looking at the table, he saw no sequences that looked the same. Each sequence was unique. Similarly, no number appeared more than once at the end of the tree.
- There is no 0 value in the tree or table.
- He stopped the probability tree at 5 days but there seems to be no limit to the number of days.
- The probabilities next to each sequence in the probability tree are all different.

After looking this over, Mr. Smith's initial reaction was "*oy vey.*" After closer inspection, he thought he saw a simple pattern. In each sequence there were consecutive days of answered calls followed by one unanswered call. For example, consider when the 1st unanswered call occurs on the 4th day. This means 3 days of answered calls followed by 1 day with no answer at 8:10 a.m. The probability is easily calculated by the multiplication rule of probability of independent events.

$$P(r = 4) = (0.9)(0.9)(0.9)(0.1) = (0.9)^3(0.1)$$

He also saw that he could simplify the formula as 0.9 raised to the 3rd power multiplied by 0.1. He repeated this for $r = 5$.

$$P(r = 5) = (0.9)(0.9)(0.9)(0.9)(0.1) = (0.9)^4(0.1)$$

He then reasoned that this could be generalized for any value *r*. In order for the first unanswered call to occur on day *r*, there must be *r*–1 days of answered calls followed by 1 day of no answer. This is the one and only possible sequence.

$$P(R = r) = (0.9)^{r-1}(0.1)$$

This is the formula for the Geometric Distribution with the probability of occurrence (an unanswered call) equal to 0.1. He decided to organize the probabilities in Table 13.5.2 and records the probabilities up to *r* = 5.

Number of the day that customer service is NOT reached for the first time: *r*	Probability customer service is reached for the first time on day *r*
0	
1	0.1
2	(0.9)(0.1) =0.09
3	$(0.9)^2(0.1)=0.081$
4	$(0.9)^3(0.1)=0.073$
5	$(0.9)^4(0.1)=0.066$
6	
7	
8	
9	
10	

Table 13.5.2: Probabilities of first recording occurring on day *r*

Q1.   What is happening to the probability as the number of days increases?

Q2.   If you know the probability, P(R=*r*), what calculation is required to find P(R=*r*+1)

Q3.   Why is there no probability value for *r*=0? Why is this different than the Binomial distribution?

Q4.   Use a calculator to complete the above table. What is the lowest value in the table?

Q5.   Can you find a number of days for which the calculator returns a probability of zero?

Q6.   What would a probability of zero mean? Do you believe that the probability is actually zero? Explain.

Mr. Smith was still somewhat puzzled by this distribution. It seems there is no largest value. It goes off to infinity. He is concerned because one of the fundamental concepts in probability theory is that the sum of all the probabilities must equal 1. He added up the first five values and the total was 0.411. This is equivalent to the following equation.

$$P(R \leq 5) = 0.410$$

He interpreted this to mean, that if he called at 8:10 a.m. every morning, there is a 0.410 probability that the 1st time the call goes unanswered will occur on or before the 5th day. Thus, there was less than a 50-50 chance that he would notice the first unanswered call before five days have passed.

He then added up the first 10 probabilities. The answer was 0.651. This was higher but still not very close to 1.

Q7.   What is the total for the first 20, 25 and 30 values? Are these close to 1?

Q8.   Determine for what value of R is the sum total greater than 0.99.

Mr. Smith asked the consultants at CBG if they could explain how the sum of the probabilities could add
to 1. Dr. Nat Finit reminded Mr. Smith about the concept of the sum of an infinite geometric series when
the common ratio is strictly less than 1. The probabilities of the Geometric Distribution form an infinite
geometric series.

Let a = value of 1st term
         p = the common ratio multiplier used to calculate each succeeding term

The formula for the infinite sum, S, is:

$$S = \frac{a}{(1-p)}$$

In this problem $a = 0.1$ and $p = 0.9$. The numerator a is the probability of no answer and the term p is its
complement.

$$S = \frac{0.1}{(1-0.9)} = \frac{0.1}{0.1} = 1$$

Thus the *infinite* sum does equal 1. Dr. Finit also explained that for the Geometric distribution a and p
will always be complements. Thus, the sum total will always be 1 for any pair of complementary values a
and p.

Mr. Smith found that calculating the cumulative probability was tedious.

        $P(R \le 5) = 0.1 + 0.09 + 0.081 + 0.073 + 0.066 = 0.410$ and
        $P(R \le 10) = 0.1 + 0.09 + 0.081 + 0.073 + 0.066 + 0.059 + 0.053 + 0.048 + 0.043 + 0.039 = 0.651$

He wondered if there was an easier way to find these probabilities. Dr. Nat Finit discussed with him a
simple trick for determining these probability sums. Dr. Finit described how to use the principle of the
complement to indirectly calculate these probabilities.

Let's take the first example $P(R \le 5)$. If R is less than or equal to five then the first unanswered occurred
on or before the 5th day. The complement of this event is that calls were answered on each and every day.
There was not one single day in the first five that a call was unanswered at 8:10 a.m.

        P(calls answered 5 days in a row) $= (0.9)(0.9)(0.9)(0.9)(0.9) = (0.9)^5 = 0.590$

Thus, $P(R \le 5) = 1 - $ P(calls answered 5 days in a row) $= 1 - 0.590 = 0.410$.

The same logic would apply to
        $P(R \le 10) = 1 - $ P(calls answered 10 days in a row) $= 1 - 0.9^{10} = 1 - 0.349 = 0.651$

In general, we can write that
        $P(R \le r) = 1 - $ P(calls answered r days in a row) $= 1 - (1 - 0.1)^r = 1 - (0.9)^r$

This formula was developed for an example in which the probability of an outcome was 0.1. This formula calculates the cumulative probability of the outcome being equal to or less than r. It is called the Cumulative Distribution Formula and abbreviated as cdf. The same logic would apply for any value of p, the probability of the outcome of interest.

$$P(R \leq r) = 1 - P(r \text{ failures in a row}) = 1 - (1 - p)^r$$

Q9.   Use this formula to determine the likelihood for Mr. Smith that the first time a call goes unanswered occurs on or before the 15th day.

Q10.   Determine for the smallest value of R such that the cumulative likelihood of finally having a call go unanswered is greater than 0.5.

Mr. Smith was interested in one last statistic, the expected value of the number of days until the 1st unanswered call at 8:10 a.m. He wrote down the general expression for expected value and the specific equation for the 8:10 a.m. case.

$$E(R) = \sum r \cdot P(r)$$
$$E(R) = 1 \cdot 0.1 + 2 \cdot (0.9 \cdot 0.1) + 3 \cdot (0.9^2 \cdot 0.1) + 4 \cdot (0.9^3 \cdot 0.1)$$
$$+ 5 \cdot (0.9^4 \cdot 0.1) + \ldots + r \cdot (0.9^{r-1} \cdot 0.1) + \ldots$$

Mr. Smith did not see an easy way to sum this infinite series. It was a weighted sum of the geometric series. Dr. Nat Finit chimed in. He said that with using derivatives learned in Calculus he could derive a simple formula for the expected value. The general formula for the Expected value of a Geometric Distribution with the probability p of an occurrence is just:

$$E(R) = 1/p$$

In this context the expected value is 10 days. On average the first time an 8:10 a.m. call is not answered is on the 10th day:

$$E(R) = 1/.1 = 10$$

However, the probability the 1st unanswered call occurs exactly on the 10th day is very small:

$$P(R = 10) = 0.9^9 \cdot 0.1 = 0.039$$

Section 13.6: Geometric Distribution—NASA Shuttle Catastrophic Failure

NASA and the astronauts all recognize and accept the fact that all manned space flight is inherently dangerous. In addition unlike automotive testing, it is not possible to actually test all of the systems under real-world conditions. Instead, they develop policies and procedures for trying to make the vehicles as safe as possible while being able to meet the space mission needs. Typically, they use engineering judgment to estimate the risks associated with space flight.

As work began on development of the space shuttle, the NASA team of engineers and managers collectively estimated that there was a 1 in 80 chance of a catastrophic failure. They anticipated flying the shuttle only 50 times. They were therefore optimistic that there would be no catastrophes. Based on this estimate, we will explore a number of related probabilities. We will also compare their initial estimates to the actual record of catastrophes to see if the experience to-date is in line with their original estimates.

By the end of 2010 there were 134 shuttle missions. NASA plans to retire the shuttle fleet in the near future. We will use the Binomial and Geometric Distributions to summarize the catastrophic risk probabilities for a total of 134 missions instead of the originally planned 50.

The probability that the first failure will be occur on any specific flight is very small. A one in eighty chance equals a probability of 0.0125. For example, the probability that the first failure occurs on the 25th flight is:

$$P(r = 25) = 0.0125(1 - 0.0125)^{24} = 0.0092$$

The likelihood that this first failure will occur on the 50th flight is even smaller.

$$P(r = 50) = 0.0125(1 - 0.0125)^{49} = 0.0067$$

However, the more appropriate question involves the cumulative probability. The likelihood that the first failure occurs on or before the 25th flight can be calculated by using the principle of the complement.

$$P(r \leq 25) = 1 - P(r > 25)$$

For the first failure to occur after the 25th flight, the first 25 flights must all be successful. Thus:

$$P(r \leq 25) = 1 - P(r > 25) = 1 - (1 - 0.0125)^{25} = 0.27$$

The first catastrophic failure occurred 73 seconds after liftoff and involved the Challenger. It was the 25th flight in the Shuttle series. Assume the estimate of one in eighty was accurate.

Q1.   What is the expected value of the first failure?

Q2.   Is it surprising that the first failure occurred on or before the 25th flight?

The Shuttle managers thought that 50 flights posed limited risk when the odds were one-in-eighty.

Q3.   What is the likelihood that the first failure would occur on or before 50 flights were completed? Should they have been more concerned?

The second catastrophe of NASA was the Columbia disaster which occurred on February 1, 2003 upon attempted reentry into the atmosphere. This shuttle launch was the 113th NASA shuttle and the first disaster happened during the 25th.

Q4.   How many flights were there from the first catastrophe until the second catastrophe? Is this consistent with a one-in-eighty probability?

The original estimate was 0.0125 probability of failure.

Q5.   If we were to use the actual Shuttle experience of 123 flights, what is the probability of a failure?

Q6.   Explain why it would be impossible to obtain a value of 0.0125 from the data for 134 flights.

In this section and the last, we developed Geometric Distribution. This distribution is closely linked to Binomial Distribution. Both distributions start with an identical process that involves repeating the same action over and over again. It could be flipping a coin, making a phone call, or launching a shuttle. The probability of success is the same for each independent repetition. The two distributions differ with regard to the random variable of interest. With the Binomial Distribution, the decision maker specifies n, the number of repetitions, coin flips or phone calls. He is interested in the random variable, X, the number of successes out of n repetitions. Consequently, the maximum number of successes is n and the minimum number is 0. With the Geometric Distribution, the decision maker is interested in the random variable, R, the number of repetitions until the 1st success. The minimum number is 1; there has to be at least one repetition to achieve the 1st success. However, the maximum is infinity. There may never be a success. Table 13.6.1 summarizes the differences.

Facet	Binomial	Geometric
Number of independent repetitions	n specified in the decision context	R random variable
Probability of success on each repetition	p	p
Number of successes	X random variable	1 stop at 1st success
Minimum value of random variable	0	1
Maximum value of random variable	n	infinity
Expected Value	np	1/p
Probability Distribution Formula (pdf)	$_nC_x(p)^x(1-p)^{n-x}$	$(1-p)^{r-1}*p$
TI Calculator Command (pdf)	binompdf(	geometpdf(
Cumulative Distribution Formula (cdf)	no formula	$1-(1-p)^r$
TI Calculator Command (cdf)	binomcdf(	geometcdf(

Table 13.6.1: Comparison of Binomial and Geometric Distributions

In the questions that follow, we demonstrate how the Binomial Distribution would be used to answer different types of questions about the space shuttle.

Q7.   What is the probability of zero catastrophes in 134 flights?

Q8.   What is the probability of exactly one catastrophe in 134 flights?

Q9.   What is the probability of exactly two catastrophes in 134 flights?

Q10.   What is the probability of two or more catastrophes in 134 flights?

Q11.   Assuming a 1 in 80 failure rate, would it be extremely rare to have 3 or more catastrophic failures in 123 flights?

Q12.   Use the above results to explain whether or not the original estimate was consistent with the actual pattern of catastrophes.

Section 13.7: Relative Frequency Approaches Theoretical Probabilities

The core concept of probability theory is that over the long range the observed frequency of an event will tend to approach the theoretical probability. We explore this idea by using simulated data. We compare the observed relative frequency to the theoretical values calculated with the Binomial Distribution function. We illustrate the concept of long-range by using different numbers of simulations. We ran the simulation 100, and 4000 times. In addition, we repeat each simulation twice to illustrate the point that each time we obtain somewhat different relative frequencies.

The charts below illustrate how simulated data approach the theoretical probabilities for the Binomial Distribution when $p = 0.5$ and $n = 5$. Each chart includes two sets of simulations of one week of calls. Five calls are made each week. Each chart and table compares the results to the theoretical probabilities. A capital N is used to specify the number of simulations. A small n refers to the number of days on which calls were made. In these examples n always equals 5 days of calls. In Figure 13.7.1 and Table 13.7.1, we use a p value of 0.5. In Figure 13.7.2 and Table 13.7.2, we use a p value of 0.7.

Let's review the first results for $N = 100$. When simulating 5 days, the number of calls answered will range from 0 to 5. The probability that exactly 0 calls are answered is 0.0313. In the first set of 100 simulations of 5 days, there was only one instance of 0 calls answered. The observed relative frequency was therefore 0.01. This is less than one-third the predicted value. In the second set of 100 simulations, there were 3 instances of calls answered. This time the observed frequency was 0.03, which is much closer to the theoretical value. For $N = 4000$, one set of simulations recorded 120 answered calls and the other recorded 134 answered calls. The corresponding relative frequencies were 0.03 and 0.034. Clearly, with N=4000, the relative frequencies were more consistently close to the theoretical values.
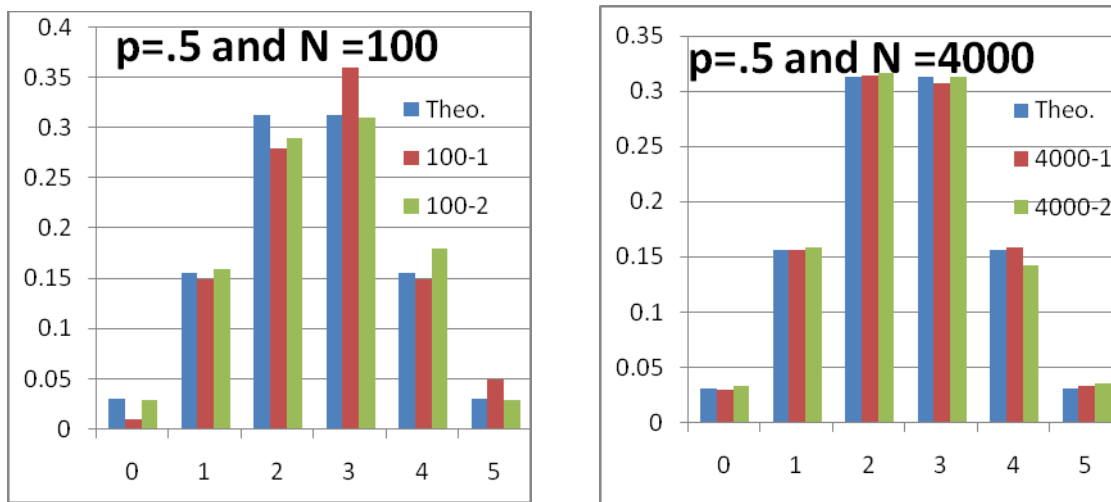

Figure 13.7.1: Frequency Histogram—simulated data and Binomial ($p = 0.5$ and $n = 5$)

		P = 0.5 and *n* = 5							
		N=100				N=4000			
X	Theoretical Probability	sample 100-1		sample 100-2		sample 4000-1		sample 4000-2	
		Number	Relative Frequency	Number	Relative Frequency	Number	Relative Frequency	Number	Relative Frequency
0	0.0313	1	0.0100	3	0.0300	120	0.030	134	0.034
1	0.1563	15	0.1500	16	0.1600	625	0.156	634	0.159
2	0.3125	28	0.2800	29	0.2900	1256	0.314	1265	0.316
3	0.3125	36	0.3600	31	0.3100	1228	0.307	1253	0.313
4	0.1563	15	0.1500	18	0.1800	637	0.159	572	0.143
5	0.0313	5	0.0500	3	0.0300	134	0.034	142	0.036
	Total	100	1	100	1	4000	1	4000	1
	Average	2.640		2.540		2.511		2.482	

Table 13.7.1: Frequency Distribution: Simulated Data and Binomial (*p*=0.5 and *n*=5)

Q1.   What was the biggest difference between sample 100-1 and 100-2?

Q2.   What was the biggest difference between a sample relative frequency and the theoretical probabilities?

Q3.   What was the smallest difference between sample 100-1 and 100-2?

Q4.   What was the smallest difference between a sample relative frequency and the theoretical probabilities?

Q5.   What was the biggest difference between sample 4000-1 and 4000-2?

Q6.   What was the biggest difference between a sample relative frequency for N=4000 and the theoretical probabilities?

The expected value for this example is 2.5.

Q7.   What was the biggest difference between the observed average and the expected value?

Q8.   Summarize your findings as to how well the relative frequencies match the theoretical probabilities for different values of N.

13.7.1 Another sample: *p*=0.7

In Figure 13.7.2 and Table 13.7.2, we illustrate how simulated data approach the theoretical probabilities for the Binomial Distribution when *p* = 0.7 and *n* = 5. Each chart includes two simulations of five repetitions and compares the results to the theoretical probabilities. A capital N is used to specify the number of simulations while a small *n* refers to the binomial repetitions. We present the results from 100 and 4000 simulations.
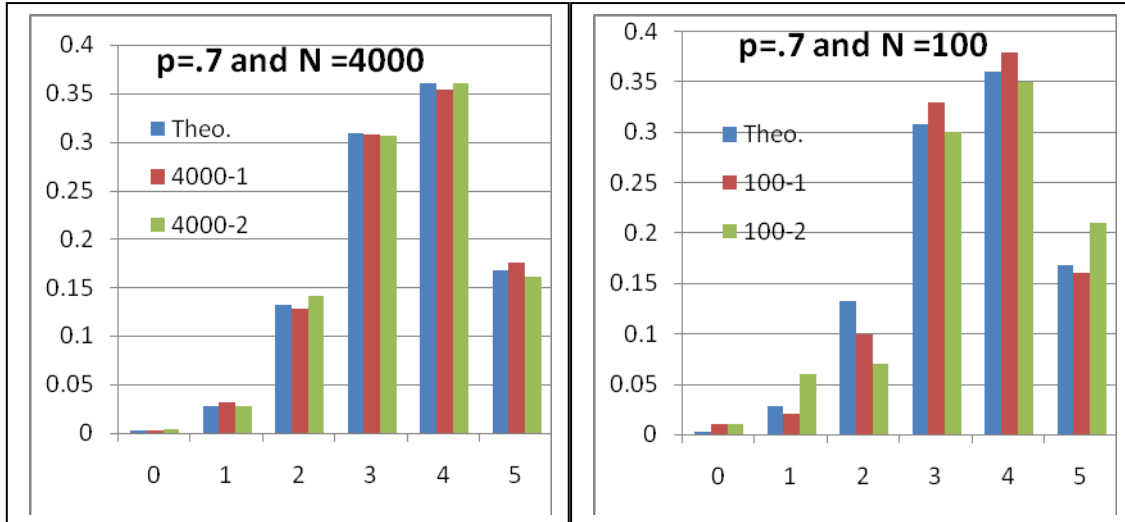
Figure 13.7.2: Frequency Histogram: Simulated Data and Binomial (*p*=0.7 and *n*=5)

		p=0.7 and n=5							
		N=100				N=4000			
	Theoretical Probability	sample 100-1		sample 100-2		sample 4000-1		sample 4000-2	
bin		Number	Relative Frequency	Number	Relative Frequency	Number	Relative Frequency	Number	Relative Frequency
0	0.0024	1	0.0100	1	0.0100	11	0.003	16	0.004
1	0.0284	2	0.0200	6	0.0600	126	0.032	109	0.027
2	0.1323	10	0.1000	7	0.0700	511	0.128	563	0.141
3	0.3087	33	0.3300	30	0.3000	1232	0.308	1225	0.306
4	0.3602	38	0.3800	35	0.3500	1416	0.354	1441	0.360
5	0.1681	16	0.1600	21	0.2100	704	0.176	646	0.162
	Total	100	1	100	1	500	1	500	1
	Average	**3.53**		**3.55**		**3.507**		**3.476**	

Table 13.7.2: Frequency Distribution: Simulated Data and Binomial (*p*=0.7 and *n*=5)

Q9.   What was the biggest difference between sample 100-1 and 100-2?

Q10.   What was the biggest difference between a sample relative frequency and the theoretical probabilities?

Q11.   What was the biggest difference between sample 4000-1 and 4000-2?

Q12.   What was the biggest difference between a sample relative frequency for N=4000 and the theoretical probabilities?

The expected value for this example is 3.5.

Q13.   What was the biggest difference between the observed average and the expected value?

Q14.   Summarize your findings as to how well the relative frequencies match the theoretical
          probabilities for different values of N.

Q15.   In general, for which value of p were the relative frequencies closer to the theoretical values?

Section 13.8: Chapter 13 (Binomial and Geometric Distributions) Homework Questions

1. Compute Combinations
   a) Find the value of $_6C_6$. Explain your answer
   b) Find the value of $_6C_0$. Explain your answer
   c) Find the value of $_6C_2$ - List every possible combination.
   d) Find the value of $_6C_4$ - List every possible combination.
   e) Compare and explain why the answers to c) and d) are the same

2. A bin of external hard disks contains four disks from four different manufacturers. In how many ways can you choose 2 disks from the bin?

3. Compute Combinations
   a) Find the value of $_{20}C_{20}$. Explain your answer
   b) Find the value of $_{20}C_4$. Provide a process for efficiently listing every possible combination.
   c) What combination is equivalent to part b)
   d) Find the value of $_{20}C_{12}$
   e) What combination is equivalent to part d)

4. A committee of 4 is to be formed from a group of 25 people. How many different committees are possible?

5. In a conference of 9 schools, how many intra-conference football games are played during the season if the teams all play each other exactly once?

6. A survey from Teenage Research Unlimited (Northbrook, Ill.) found that 40% of teenage consumers receive their spending money from part-time jobs. If nine teenagers are selected at random, find the probability that at least three of them will earn their spending money from part-time jobs.

7. The probability that a student is accepted to Kennedy College is 0.3. Assume 5 students from the same school apply. Use both the formula and the Binomial pdf or the Binomial cdf in your calculator to determine these probabilities.
   a) What is the probability that exactly 2 are accepted?
   b) What is the probability that 2 or fewer are accepted?
   c) The top universities in the US accept approximately one out of every ten applicants. Calculate a and b again if the admission probability is 0.1.

8. Courts pick the juries randomly from specific population that is often split evenly between males and females. In many jurisdictions, civil cases have juries of six people. If a jury has 6 members and the population contains half males and half females, calculate the following probabilities:
   a) What is the probability of having 3 males and 3 females on the jury?
   b) What is the probability of having all males?
   c) What is the probability of having all jurors of the same sex?
   d) What is the probability of having 5 or more females?

9. Now suppose, a jury has 12 members and the population have 25% minorities.
   a) What is the expected value of the number of minorities on a 12 person jury?
   b) What is the probability of not having any minorities on the jury?

c) What is the probability of having one or fewer minority members on the jury?
d) What is the probability of having 2 or fewer minorities on the jury?

10. If a student randomly guesses at ten multiple-choice questions. Each question has four possible choices.
   a) A score of six is passing. Find the probability of getting a six or more by using the Calculator function.
   b) Given the student passed, what is the probability the student passed with the minimum grade of 6?
   c) What is the expected value of the number questions the student will guess correctly?
   d) Which is more likely, 2 correct answers or 3 correct answers out of 10?
   e) A score of 9 or 10 earns an A grade. Find the probability of getting an A grade. DO this both with the formula and with binomial function for your calculator.
   f) What is the probability of getting no right answers? Compare the results to the simulation experiment in chapter 11.

11. A marketing manager of Mm Company has monthly meetings with his 5 consultants to review the marketing strategies of the company. The manager will run the meeting as long as 4 or more of his consultants attend the meeting.
   a) If each consultant is 85% reliable in attending the meeting, what is the probability of having the meeting each month?
   b) If the manager wants to have to cancel the meetings at most 5% of the time, what standard should the manager use for the minimum number of attendees?
   c) Generate five random numbers between 1 and 100 and assign each of them to one consultant (C1 to C5). If the assigned number is 85 or less then the consultant attends; if not, the consultant misses the meeting. Record your random numbers in the first row of the following table. Then count the number of attendees and record it in the last cell of first row.
   The table below will represent one year's worth of meetings

	C1	C2	C3	C4	C5	Number of attendees	Canceled Yes or No
Month 1							
Month 2							
Month 3							
Month 4							
Month 5							
Month 6							
Month 7							
Month 8							
Month 9							
Month 10							
Month 11							
Month 12							

Table 13.8.1

   d) Repeat it 11 more times and record the results.
   e) Based on your result, how many times was the meeting not canceled? What is the percentage of times the meeting was cancelled?
   f) Compare the percentage with the theoretical probability.

12. In a 2007 random survey of drivers on weekend nights, the police found that the proportion of drunk drivers (0.08 alcohol level) was .022. The state police are planning to establish some checkpoints to catch the drunk drivers.
    a) What is the probability the first drunk driver identified occurs on 5th pull-over?
    b) Each random stop takes approximately four minutes. The state police will stop 15 drivers on average each hour. What is the probability of catching the first drunk driver in the first hour of the check point?
    c) What is the probability that they will need to stop 30 or more drivers before finding the first drunk driver?
    d) On average how many drivers will be stopped before they find the first drunk driver?
    e) The checkpoint will be in place for four hours for a total of sixty random stops. On average how many drunk drivers will the state police find?
    f) What is the probability that they will find 4 or more drunk drivers during the four hours?

13. It is believed that on Saturday night between the hours of 1 am and 3 am that the percentage of drunk drivers is actually 5%.
    a) What is the probability that they will need to stop 15 or more drivers before they find the first drunk driver?
    b) On average how many drivers will be stopped before they find the first drunk driver?
    c) What is the probability of not catching any drunk driver in two hours in which 30 drivers are stopped?
    d) What is the probability that they will find 2 or more drunk drivers during the two hours?

14. NASA engineers as they began developing the Shuttle estimated based on engineering experience and judgment that there was a 1 in 80 chance of a catastrophic failure. They anticipated flying the shuttle only 50 times and were optimistic that there would therefore be no catastrophes. Based on this estimate answer the following question:
    a) 73 seconds after launching the 25th NASA shuttle, the shuttle (Challenger) exploded, killing the entire crew. What is the probability of having the first crash on the 25th flight?
    b) What is the probability of having the first crash on or before the 25th flight?
    c) What is the probability of having the first crash in first 50 flights?
    d) What is the probability of not having a crash in first 50 flights? How about first 80 flights? How about first 100 flights?
    e) On average how many flights will there be until the first crash occurs?
    f) The second catastrophe of NASA was the Columbia disaster which occurred on February 1, 2003. This shuttle launch was the 113th NASA shuttle and the first disaster happened in 25th. What is the probability of two disasters in 113 flights? What is the probability of two or more disasters in 113 flights?

15. A company wants to do a random survey of past purchasers of its product to find one who has experienced a particular problem. They know the problem exists with their product and wish to interview people to determine the impact on the customer when the problem occurs. They estimate that 1.5% of their products sold have this problem. Based on this information find the following probabilities:
    a) What is the probability of finding a first customer who experienced the problem by the 10th call?
    b) What is the probability of finding a first customer by the 50th call?
    c) What is the probability of not finding any customer with the problem in the first 50 calls?
    d) On average how many calls does an operator have to make to find a customer who experienced the problem?

e) If one of our operators has 100 customers to call, what is the probability of finding zero customers with damaged product?
f) What is the probability the operator will find 2 customers with damaged products by?
g) What is the probability the operator will find more than 2 customers with damaged products?

Airline Overbooking (Revisited from previous chapter)

GLA realized that simulation of the number of airline-customer who show-up is time-consuming. They recognize that probabilistic pattern of the number who show up fits the binomial distribution. They decided to use the Binomial distribution and its expected value to evaluate different overbooking strategies.

Here are a few things you need to remember from the previous chapter.
The average one-way airfare price is $300 and operating cost of the plane is $6,000 with an additional cost of $20 per customer. Flight 425 has 30-passenger seat capacity. GLA will payback $300 airfare price plus $250 compensation to the customers whose boarding is denied. Customers who cancel their flights or don't show up at the gate will pay GLA $ 75 out of the refund for their ticket. According to GLA flight records, on average 15% of the people who make reservation in advance don't show up at the gate.

16. No Airline Overbooking
    a) What is the expected number of arrivals if 30 customers have already reserved their tickets?
    b) What percent of the time will all 30 reserved customers appear at the gate?
    c) What is the most likely number of customers appearing at the gate? How often will this happen on average in 100 flights in which all 30 seats have bee reserved?

17. Airline Overbooking
    a) If the GLA management allows 35 customers to reserve seats, how many of them will show up on average?
    b) GLA is considering an overbooking policy such that there is only a 5% chance or less that someone will have to be bumped. What is the maximum number of reservations they can accept without exceeding this 5% standard? What is the probability that someone will be bumped with this policy?

	Number of Reserved Customers				
	31	32	33	34	35
P(X ≤ 30)	0.994	0.963	0.891	0.772	0.619

Table 13.8.2: Probability of less than or equal to 30 customers arriving at the gate

    c) GLA is considering an overbooking policy such that there is only a 10% chance or less that someone will have to be bumped. What is the maximum number of reservations they can accept without exceeding this 10% standard?
    d) Figure 1 shows the profit values for different overbooking strategies as the compensation is increased gradually from $150 to $300. Identify the optimal policy for each compensation level and on average how much money would GLA earn per overbooked flight. Which policies are never optimal? Explain why the profit curves are not strictly a monotonically increasing function of the number of reservations.
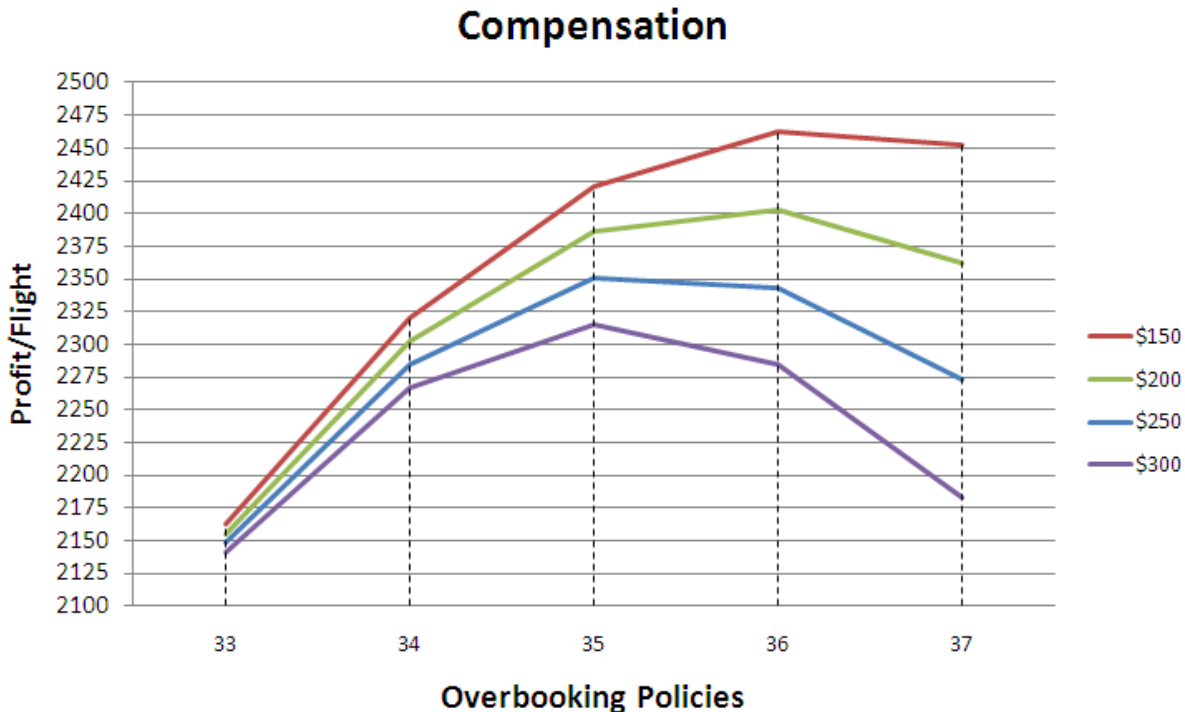
Compensation

Figure 13.8.1: Optimal overbooking for different compensation values

Bed &Breakfast Overbooking

18. B&B Profits and Break-Even Analysis
    Niagara House is a Bed & Breakfast hotel (B&B) located in Niagara Falls, NY. It offers 9 guest
    rooms, a full breakfast, and a private parking lot. Each room has a private bath, a queen size bed, and
    cable television. The rate for a room in Niagara House is $130 per night and customers have to pay a
    deposit for the first night in advance. If a customer wants to cancel a reservation, the customer has to
    cancel 24 hrs in advance. There is a $25 cancellation fee that is taken out of the deposit before the
    remainder is refunded. From the records, seven percent of the customers cancel their reservations.

    The hotel has some daily expenses. It has a $350 fixed cost that covers the cost of the property and
    pays an annual income to the owners. There is a $30 variable cost per room per day that includes the
    cost of breakfast for two people.

    a) When does the revenue of Niagara House cover its daily expenses including its fixed cost in the
       case where all reservations show up?
    b) How much profit does the B&B make on a day in which it fills all the rooms with all 9 reserved
       customers and all show up?
    c) If 9 customers already reserved their rooms for the day after tomorrow, what is the probability
       that all rooms will be occupied?
    d) If 9 customers already reserved their rooms for the day after tomorrow, what is the probability
       that exactly one customer will cancel his reservation 24 hrs in advance? How much profit will
       Niagara make in this instance?
    e) What is the probability that at least one customer will cancel his reservation out of 9 customers?
    f) What will be the profit if 2 of the 9 reserved customers cancel their reservation and what is the
       likelihood of this happening?

g) What is the Expected Profit if they have nine reservations for the day after tomorrow? Use the table below to help answer the question.

h) If 9 customers already reserved their rooms for the day after tomorrow, what is the probability that the hotel will not cover its expenses? Explain why this question does not match up with the answer provided in part a.

i) What is the probability that it will earn more than $450?

19. B&B Overbooking and Transfer
In order to cover its expenses and make profit, Niagara House should fill up as many rooms as possible. To ensure this, the hotel management uses an overbooking strategy. However, if the number of arrived customers is more than the number of available rooms, they have a standing mutual agreement to transfer them to an equivalent nearby B&B. They compensate the arrivals for their inconvenience with a gift card valued at $30.

a) If Niagara House allows 10 customers to reserve, what is the probability that exactly one customer will be transferred to a nearby hotel and given a gift card?

b) What will be the profit if all 10 reserved customers come to the hotel and one customer is transferred with a gift card?

c) If Niagara House lets 10 customers reserve rooms, what is the probability that two customers will cancel their reservations?

d) What will be the profit if 2 of the 10 reserved customers cancel their reservations?

e) Would you recommend Niagara B&B adopt this overbooking policy? Explain your answer.

20. B&B Optimal Overbooking Policy
a) What is the optimal number of reservations to accept if Niagara plans to allow for overbooking?

b) In order to avoid losing customer goodwill, Niagara House management considers increasing the compensation value to $40. Will the previously recommended overbooking strategy still be optimal? How much will the average profit per day change? Why is it not $10?

c) Without doing any more detailed calculations, what do you think would happen if the compensation was increased to $50?

d) Spring-Summer is the peak season for the Niagara House hotel. This period lasts for approximately 20 weeks. This offers an opportunity to make more profit. Because of high demand, Niagara House increases its rate for a room to $195 per night. The cancellation fee is $50 and the compensation for being transferred is increased to a $60 gift card. The daily expenses of the hotel did not change: it has a $350 fixed cost and a $30 variable cost per room. What is the optimal overbooking strategy during the peak season?

e) What will be the expected annual profit if Niagara House follows the 9-reservation policy throughout the year? Assume that seven percent of the customers cancel their reservations and the hotel is open for 50 weeks: 30 weeks are considered off-peak and 20 weeks are considered peak season. What percent of the total profit is earned during the peak season?

f) What will be the expected annual profit if Niagara House follows the 10-reservation policy throughout the year?

g) What will be the expected annual profit if Niagara House follows the 10-reservation policy during the off-peak season and 9-reservation policy during the peak season?

h) Because of the higher cancellation fee and extremely high demand in July, the cancellation probability might decrease to only 3%. What would be the optimal policy in the month of July?

Service Quality Assurance

21. The manager of Koala foods wants to track each call to assess the quality of service. He is trying to figure out whether his staff needs more training or not. Therefore, AGB Company decides to select

calls at random. An experienced customer service representative will listen to the conversation and categorize each as either "Satisfactory" or "Not unsatisfactory". For recent hires Mr. Smith wants AGB to ensure that by the end of a two-month training period a new operator achieves at least an 80 percent satisfactory rating. AGB picked 5 calls of an operator and here are the results:

Number	Quality
1	1
2	0
3	1
4	0
5	1

Table 13.8.3: "1" means quality is satisfying and "0" is not-satisfying

a) Is he/she doing the job satisfactorily?
b) Can we decide about this operator based on a sample of 5 calls?
c) What is the likelihood that an individual who regularly achieves the standard of 80% satisfaction will be below this standard in a sample of 5 calls? Use this to review your answer to part b.
d) What factors will affect how large a sample of calls to listen to and evaluate?
e) Assume AGB will review 20 calls of an operator who generally meets the 80% standard. What is the probability that in this sample of 20, the operator will meet or exceed the standard (at least 16 satisfactory calls a day)?
f) What is the probability that 60% or fewer of the calls will be judged satisfactory? 70% or fewer will be judged satisfactory?
g) The manager must make a decision, as to when to send the operator to training which will cost the company more than $500. She does not want to do this unless she is reasonably sure that the low performance is not just a function of natural variability. The manager will pick a number such that the chance of seeing that many bad calls is 5% or less if the operator generally meets the 80% standard. How many bad calls should she allow out of 20? What is the corresponding percentage?
h) Now consider the fact, AGB is going to review 50 calls of an operator. The manager will pick a number such that the chances of seeing that many bad calls is 5% or less if the operator is generally meeting the 80% standard. How many bad calls should she allow when the sample is 50? What is the corresponding percentage? Compare this answer to that of the previous problem and explain the difference.

22. Service Quality Sampling Strategy
a) Do you think that size of our samples should depend on the cost of evaluating each call?
b) What other factor(s) do you think should be considered when deciding on a sample size?
c) Do you think that picking calls from only one day is a good way for sampling? Why.
d) Do you think that the cost of sampling has any effect on the manager's decision in selecting number of the days for sampling?
e) Do you think that categorizing the calls in just two groups is a good idea?
f) What are the alternatives to categorize the calls?
g) Can we use Binomial distribution to calculate the probabilities if we have more than two groups?
h) Imagine a policy as follows for our goal of 80% satisfactory performance. If the number of bad calls is equal to or below a certain percentage X, the operator is sent to training. If it is equal to or above a certain percentage Y, the operator is certified as trained and does not need to be reviewed for another year. If the percentage is between X and Y, another sample of calls is selected and reviewed. Pick values of X and Y. Describe how you would apply this policy to an initial sample of 50 calls. Calculate the probability that an operator with a long-term 0.8 record might fall below

X in a sample of 50. Calculate the probability that an operator with long-term record of 0.75 might still score above Y.

Organ Harvesting

23. Organ donation and transplants save many lives that need vital organs e.g. heart, kidney, liver. Only a small percentage of deaths result in potential organ donation. In Jefferson County there are an estimated 60 deaths per year from which organs could be harvested. However, not everybody volunteers as a donor. Nationwide an estimated 25% of people sign organ donor cards. In addition, when a death occurs, hospitals try to convince family members to agree to donate organs. Some hospitals have special initiatives which have made them more successful than others in arranging family approved donations. Hospital staff efforts increase the total by anywhere from an additional 15% to 35% depending upon the hospital's effectiveness in reaching out to family members. In the County, there are currently 27 people in dire need of heart transplants within the next year.
    a) With just the pre-signed organ donor cards, what is the probability that there will be enough hearts available from 60 deaths to meet the needs of all 27 people?
    b) Now assume the Jefferson area hospitals have worked to reach out to family members but they are at the lower end of the range. Their efforts only add 15% to the overall total donation rate. Thus, 40% of eligible cadavers result in donated organs.
        i.   On average how many hearts will become available during the year?
        ii.   With the pre-signed organ donor cards and 15% added by the hospital, what is the probability that there will be enough hearts available from 60 deaths to meet the needs of all 27 people?
    c) Now assume the Jefferson area hospitals are in the middle range and add 25% to the overall total donation rate. Thus, 50% of eligible cadavers result in donated organs.
        i.   On average how many hearts will become available during the year?
        ii.   With the pre-signed organ donor cards and 25% added by the hospital, what is the probability that there will be enough hearts available from 60 deaths to meet the needs of all 27 people?
    d) The Jefferson area hospitals are striving to provide heart transplants to all patients on the waiting with at least 90% probability. What is the long-term average percentage of people that would have to agree to donate in order for the hospitals to have a 90% probability each year? Explain why the answer is not just 27/60 or 45%.

24. Organ Harvesting: Two neighboring counties of approximately equal have begun coordinating their efforts. Assume there are a total of 120 eligible deaths in the two counties and that there are 54 critical patients needing heart transplants.
    a) What is the long-term average percentage of people that would have to agree to donate from these two counties in order for the hospitals to have a 90% probability each year?
    b) A regional consortium is being developed that involves four counties. Assume there are a total of 240 eligible deaths in the two counties and that there are 108 critical patients needing heart transplants. What is the long-term average percentage of people that would have to agree to donate from these four counties in order for the hospitals to have a 90% probability each year?
    c) Discuss the impact of pooling efforts.

25. Organ Harvesting: There is an individual in Jefferson County who is extremely critical and is first on the list.
    a) On average how many eligible deaths would occur before the hospitals identify an individual with a signed organ donor card? Remember only 25% of individuals have signed cards.

b) If the hospital is successful in its efforts to increase the rate of donation to 60% of eligible deaths, on average how many eligible deaths would occur before the hospitals obtains a heart?

c) Consider a situation when the donation percentage is just 25 percent. There are supplementary efforts by local hospitals. What is the probability that the first individual that has an organ donor card is associated with
    i.   the 1st cadaver?
    ii.   the 2nd cadaver?
    iii.   the 3rd cadaver?
    iv.   the 4th or later?
    v.   Assume the deaths are spread equally throughout the year with an average of five per month. What is the probability that it will take longer than a month for a cadaver from an organ donor card to become available?

d) Assuming the hospitals' efforts increase the total percentage to 40%. What is the probability that the first individual or family to donate is associated with
    i.   the 1st cadaver?
    ii.   the 2nd cadaver?
    iii.   the 3rd cadaver?
    iv.   the 4th or later?
    v.   Assume the deaths are spread equally throughout the year with an average of five per month. What is the probability that it will take longer than a month for a cadaver from an organ donor card to become available?

Chapter 13 Summary

What have we learned?

In this chapter we have studied two related distributions, the Binomial Distribution and Geometric Distribution. The formulas for these distributions replace the need for the repeated simulations we did in the previous chapter. For both distribution there are 4 common elements to each random event being analyzed:

> There are 2 possible outcomes.
> The p value, probability of success, is the same for each repetition.
> The outcome of each repetition is assumed to be independent.
> There are n identical repetitions of the random event.

The Binomial Distribution has two parameters: p, the probability of success on any individual trial, and n, the number of trials that are performed. The formula to find the probability distribution function for a particular number of successes, x, is $_nC_x (p)^x (1-p)^{n-x}$. We can also use the TI calculator command binompdf(n,p,x). There is no formula to find the cumulative distribution function, which adds the probabilities from 0 to x. However, we can use the TI calculator the command binomcdf(n,p,x).

The Geometric Distribution has only one parameter: p, the probability of success on any individual trial. The probability that the first success occurs on the rth trial is $p(1-p)^{r-1}$. We can also use the TI calculator command geometpdf(p,r). The probability that the first success occurs in r or less days is $1-(1-p)^r$ and the TI calculator command is geometcdf(p,r).

Both these distributions can help us determine how likely a particular sequence of events would be. We can then use the results to make decisions. The three most common criteria for believing something is not random chance are: 0.1 (10%), 0.05 (5%), and 0.01 (1%).

Another important consideration is the expected value for a distribution. While the expected value is often not actually possible, it is the long run average if an event occurred many times. The best choice is the option that has the most favorable expected value.

A third consideration discussed in the chapter is the concept of conditional probabilities. A conditional probability asks the likelihood of an occurrence if we already know something that happened. The formula is $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ where A is the event we are interested in and B is the event we already know occurred.

Terms

Parameter	Values that need to be specified in order to apply a distribution
Success	The outcome we are interested in (not necessarily a good thing)
Failure	The complement of the outcome we are interested in
n	The number of times the random event is repeated
p	The probability that a success occurs (a number between 0 to 1)
$_nC_x$	The number of ways x items can be chosen out of n items $$\frac{n!}{(n-x)!(x)!}$$
Binomial Distribution	The probability distribution that is used to find the likelihood of a particular number of successes (x) for a given number of trials (n)
P(X = x)	The probability of x successes: $_nC_x(p)^x(1-p)^{n-x}$
binomcdf(*n*,*p*,*x*)	The TI calculator command to calculate the probability of 0 to x successes out of n trials if the probability of success is p
Geometric Distribution	The probability distribution that is used to find the likelihood of a particular trial (r) being the first success.
P(R = r)	The probability the rth trial will be the first success: $p(1-p)^{r-1}$
Expected Value	The long run average value over many repetitions
Binomial	$E(X) = np$
Geometric	$E(X) = 1/p$
Conditional Probability	The probability that something happens, *given* that something else happened

Chapter 13 (Binomial and Geometric Distributions) Objectives

You should be able to:

- Determine whether or not a situation has the 4 common elements needed to use binomial and geometric distributions

- Calculate combination values

- Determine whether a solution to a question requires binomial or geometric calculation

- Calculate binomial probability distribution functions

- Calculate binomial cumulative distribution functions

- Calculate geometric probability distribution functions

- Calculate geometric cumulative distribution functions

- Calculate conditional probabilities.

Chapter 13 Study Guide

1. What is the Binomial Distribution and for what is it used? Give a specific example.

2. What are the parameters needed for the Binomial Distribution?

3. How do you find the expected value for the Binomial Distribution and what does it represent?

4. What is the Geometric Distribution and for what is it used? Give a specific example.

5. What are the parameters needed for the Geometric Distribution?

6. How do you find the expected value for the Geometric Distribution and what does it represent?

7. What are the three criteria levels mentioned in the chapter?

    a. Give an example of when the highest would be more appropriate.

    b. Give an example of when the lowest would be more appropriate.

8. What does conditional probability mean?

    a. If the probability of B occurring is 40% and the probability of A and B both occurring is 20%, what is the probability of A occurring given that B occurs?

Section 14.0: Poisson and Exponential Distributions

In the previous chapter we developed two discrete distributions, the Binomial and the Geometric. Each distribution is based on simple assumptions about repeated independent events with the same probability of occurrence. From these assumptions we were able to derive a general formula for calculating individual probabilities. We also presented and explained the standard expressions for the expected value. In this chapter we introduce two new distributions, the Poisson and the Exponential. The Binomial Distribution has two parameters, *n* and *p*. The Poisson and Exponential Distributions need only one, the average rate. This is usually represented by the Greek lower case symbol, λ, lambda.

The Poisson Distribution is a natural outgrowth of a random sequence of events in which the occurrence of one event has no influence on the occurrence of another random event. For example, someone falling off a bike and being rushed to a hospital has no influence as to when another person will fall off a bike and be rushed to a hospital. Each event is independent of every other event. When a random variable distributes itself according to the Poisson Distribution there is a specific mathematical formula that is used to compute probabilities. This formula can determine the probability of any number of occurrences of the event, such as the number of calls to a 911 system. The entire random sequence of independent events that occur continuously is often called a Poisson Process. The number of accidents, emergencies, airplane crashes, and crimes are all examples of random variables that form a Poisson Process. It is also often used to describe the random arrival of people at a box office, or arrivals at a store, e-mails or phone calls to a help desk.

This distribution is named for the 19th century French mathematician Siméon-Denis Poisson who developed it in his "Research on the Probability of Criminal and Civil Verdicts." The formula was derived as an approximation to the Binomial Distribution when the number *n* of events is very large (i.e., in the hundreds or thousands) and *p* is small (i.e., less than 0.1). This derivation is beyond the scope of this text. Mathematical models of early telephone systems were amongst the first applications of the Poisson Process early in the twentieth century.

Anytime you observe a sequence of events, there are two ways to describe the situation. You can count the number of events in a minute, or hour, or day. Alternatively, you can measure the time between two sequential events. For example, if on average four calls come in during an hour, the average time between calls is one-quarter of an hour. The *number* of calls is obviously a discrete variable. The *time* between calls is a continuous variable. There is a fundamental theorem that links the Poisson and Exponential Distributions.

> If the number of random events in a time period is a random variable that behaves like the Poisson Distribution, then the time between calls behaves like the Exponential Distribution.

The Exponential Distribution is the first example in this text of modeling a continuous random variable. We will therefore need to introduce some basic principles about how probability theory is applied when the random variable is continuous.

Section 14.1: The CSI Team

In a suburban county outside of Detroit, Michigan, the sheriff's department only has one Crime Scene Investigation (CSI) Unit to assist the many small suburban police departments investigate major crime scenes. The County Sheriff's Department averages three major crime scene calls per day. Needless to say these calls come in totally at random. Except in the case of a rare crime spree, each incident is independent of every other incident. Therefore, these calls fit the description of the Poisson Process. Thus the number of calls each day can be modeled by the Poisson Distribution. The team is composed of three investigators who work together to analyze major crime scenes that occur in their jurisdiction. On average, the team can handle three investigations per day, each of which averages 2.5 hours. Thus the daily average workload is matched with the unit's average capability.

The CSI Unit is always on call and is available to respond to three crime scenes a day (i.e., 7.5 hours), five days a week. They are paid a salary, based on a pay rate of $32/hour. Even if there are not three calls on a given day, the investigators still earn a full day's pay. If the unit receives more than three calls on a given day, they are required to investigate all crime scenes in a timely manner. For more than 3 calls in a day, they are paid overtime at time-and-a-half their usual pay rate, or $48/hour. The Sheriff's Department has a decision to make: Should they maintain the CSI team as it is now. Should they just pay for overtime when necessary, or hire a new investigator and split the four investigators into two teams of two? If this happens, it would take about 3.5 hours for each team of two to investigate a crime scene. Each team could analyze two scenes per day. The Sheriff's Department wants to know if it is more cost effective to maintain the current CSI team structure or to hire one more investigator.

14.1.1 The Poisson Distribution

In order to decide if the Sheriff's Department should hire another investigator, we need to find some probabilities. In particular, we want to know the probability that the CSI team will get exactly x calls on a particular day, knowing that they average three calls per day. The average number of calls per day is called a parameter and is usually denoted with the lower case Greek letter lambda, λ.

Q1.   What is the value of λ?

Q2.   What values can x be?

The CSI probabilities can be modeled by the Poisson Distribution. Mathematicians have proven that the probability of exactly x events occurring where each event previously occurred with an average rate of λ is given by the function:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

This is called the probability density function for the Poisson Distribution. To see how this formula works, let's look at the probability of having exactly 2 crime scenes to investigate in a specific day.

$$P(x = 2) = \frac{3^2 e^{-3}}{2!}$$
$$= \frac{9}{2e^3}$$
$$\approx 0.224$$

There is 0.224 probability that there will be exactly two crime scenes to investigate on any given day. This probability is, on average, equivalent to exactly two calls once every four or five days. It also means that there is 0.776 (1 – 0.224) probability that there will either be fewer than two crime scenes or more than two crime scenes to investigate on any particular day.

Q3.   Use this function to find the probability that the CSI team will have zero calls on a particular day.

Even when using a calculator, this formula can sometimes be messy to use (see Figure 14.1.1).

Figure 14.1.1: Evaluating the probability function using a calculator

However, modern graphing calculators have the ability to directly evaluate probabilities based on the Poisson Distribution for different values of x and λ. The Poisson probability density function command is in the distribution menu, which is the ψfunction of the ☐ key. Within the distribution menu, scroll down until you find the poissonpdf(command and press ⊆. (See Figure 14.1.2) You need to enter the parameter, λ, which in this case is three, then a comma, followed by the value of x. For example, the probability that the CSI team will have zero calls on a particular day would look like this:
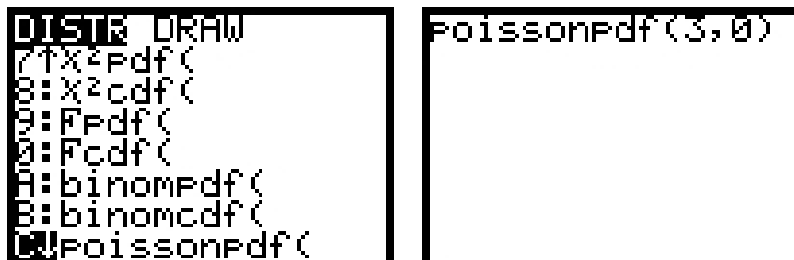
Figure 14.1.2: Accessing the Poisson probability density function on a graphing calculator

Q4.   Use the Poisson probability density function on your calculator to find the probability that the CSI team will have zero calls on a particular day. How does it compare to the value you found using the formula? Which method was easier?

Q5.   Use the Poisson probability density function on your calculator to find the probability that the CSI team will have exactly one call on a particular day.

Q6.   Find the probability that the CSI team will have exactly two calls on a particular day, and again for exactly three calls on a particular day.

Record your answers to questions 4, 5, and 6 in a table similar to Table 14.1.1. Round answers to three decimal places, and find and record the probabilities for x = 4, 5, 6, 7, 8, and 9.

x (calls on a particular day)	P(x)
0	0.050
1	0.149
2	0.224
3	0.224
4	0.168
5	0.101
6	0.050
7	0.022
8	0.008
9	0.003

Table 14.1.1: Probability of having *x* calls on a particular day if the average is three calls per day

Q7.   You should have found very small probabilities for *x* = 8 and *x* = 9. On average, about how many times in a year (i.e., 250 workdays) will exactly 8 calls occur? Exactly 9 calls?

Since the CSI team can handle three calls per day, we are really interested in the probability that they will have more than three calls on a particular day. Let's start by finding the probability that they will have *at most* three calls on a particular day. This is the same as asking for the probability of having exactly zero or one or two or three calls on a particular day. Since it is impossible to have two different numbers of crimes scenes to investigate on the same day, these events are said to be mutually exclusive. As such we can simply add all the individual probabilities to find the probability that the CSI team will have at most three calls on a particular day.

Q8.   Find the probability that the CSI team will have at most three calls on a particular day.

Q9.   If we know the probability of having at most three calls on a particular day, how can we find the probability of having more than three calls on a particular day?

Q10.   What is the probability that the CSI team will have more than three calls on a particular day?

Q11.   Using the same method, what is the probability that the CSI team will have more than four calls on a particular day?

14.1.2 Cost of Overtime

The Sheriff's Department's Deputy Chief, Dante Top, recently received the bi-weekly payroll statement. The bill for the two-week period was $2520 for overtime for the CSI Unit. He was puzzled since on an average day there would be no overtime. The first day of the ten-day period had four crime scene investigations. Each of the three investigators was paid for 2.5 hours at a rate of $48 per overtime hour. The total cost of overtime for that day was (2.5)($48)(3) = $360.

Day	Calls	Calls on Overtime	Cost of Overtime
Monday 1	4	1	1 · (2.5)($48)(3) = $360
Tuesday 1	5	2	2 · (2.5)($48)(3) = $720
Wednesday 1	2	0	0 · (2.5)($48)(3) = $0
Thursday 1	2	0	0 · (2.5)($48)(3) = $0
Friday 1	1	0	0 · (2.5)($48)(3) = $0
Monday 2	4	1	1 · (2.5)($48)(3) = $360
Tuesday 2	2	0	0 · (2.5)($48)(3) = $0
Wednesday 2	6	3	3 · (2.5)($48)(3) = $1,080
Thursday 2	3	0	0 · (2.5)($48)(3) = $0
Friday 2	3	0	0 · (2.5)($48)(3) = $0
Totals	32	7	$2,520

Table 14.1.2: CSI calls during one two-week period

Q12.   For how many of the ten days was there overtime?

Q13.   What was the busiest day? How many overtime calls were there? What was the cost of the overtime for the busiest day?

Q14.   How many overtime calls were there? What was the total cost of the overtime?

Q15.   What is the average number of calls they should have received in a week?

Q16.   What was the total number of calls in the first week? In the second week?

Q17.   One of the weeks was below average, yet there was overtime in that week. How could that happen?

x (calls on a particular day)	Number of Calls on Overtime	$P(x)$	Cost of Overtime $C(x)$	Contribution to Expected Overtime Cost/Day $P(x) \cdot C(x)$
0	0	0.050	0	
1	0	0.149	0	
2	0	0.224	0	
3	0	0.224	0	
4	1	0.168	$360	$60.48
5	2	0.101	$720	72.72
6	3	0.050	$1080	54.00
7	4	0.022	$1440	31.68
8	5	0.008	$1800	14.40
9	6	0.003	$2160	6.48
Totals		0.999		$239.76

Table 14.1.3: The cost of overtime

Using the values from Table 14.1.3, the expected cost of overtime per day is $239.76. On average, the cost of overtime for a bi-weekly pay period would be 10 times the expected cost per day, or $2,397.60.

Q18.   How does the expected cost for a biweekly pay period compare to the actual cost observed by the
        deputy chief? Explain the difference.

14.1.3 Exploring the Alternative Plan—Adding a Worker

Recall that the Sheriff's Department is considering adding a fourth CSI and forming two teams of 2 CSIs.
Each investigation will then require 3.5 hours. Therefore, each team can complete 2 investigations per
day. This increases the number of investigations the Department can do to 4 per day. This will clearly
reduce overtime costs, but there is the increased cost of the additional CSI.

When there were three investigators, the cost per biweekly pay period without overtime was 3($32/hr)(7.5
hrs/day)(10 days) = $7,200. If a fourth investigator is hired at the same rate of pay, the hours worked per
day by each investigator will be reduced to seven.

Q19.   Calculate the cost per biweekly pay period without overtime if a fourth investigator is hired. How
        does this compare to the present cost?

Q20.   In order for hiring a fourth investigator to be cost effective, by how much must the expected cost
        of overtime per biweekly pay period be reduced?

When performing analysis, we often first look at a specific example to make sure we understand the
respective costs. Thus, to compare the two plans, we will first look at the data for a particular two-week
period. We will then use the probability distribution to calculate the expected cost of each of the plans.

Table 14.1.4 shows the data from the same two-week period as Table 14.1.2. This time the cost of
overtime has been computed assuming that there were two teams of two CSI investigators. They are able
to cover four crime scenes per day.

Day	Calls	Calls on Overtime	Cost of Overtime
Monday 1	4	0	0 · (3.5)($48)(2) = $0
Tuesday 1	5	1	1 · (3.5)($48)(2) = $360
Wednesday 1	2	0	0 · (3.5)($48)(2) = $0
Thursday 1	2	0	0 · (3.5)($48)(2) = $0
Friday 1	1	0	0 · (3.5)($48)(2) = $0
Monday 2	4	0	0 · (3.5)($48)(2) = $0
Tuesday 2	2	0	0 · (3.5)($48)(2) = $0
Wednesday 2	6	2	2 · (3.5)($48)(2) = $720
Thursday 2	3	0	0 · (3.5)($48)(2) = $0
Friday 2	3	0	0 · (3.5)($48)(2) = $0
Totals	32	3	$1,008

Table 14.1.4: CSI Calls during the same two week period with two teams of two CSI investigators

Q21.   Include both base pay and overtime cost. Which structure, the present one with one CSI team or
        the proposed alternative with two CSI teams, would have been more cost effective for the two-
        week data in Table 14.1.2 and 14.1.4?

x (calls on a particular day)	Number of Calls on Overtime	P(x)	Cost of Overtime C(x)	Contribution to Expected Overtime Cost/Day P(x)*C(x)
0	0	0.050	0	
1	0	0.149	0	
2	0	0.224	0	
3	0	0.224	0	
4	0	0.168	0	
5	1	0.101	$336	$33.94
6	2	0.050	$672	33.60
7	3	0.022	$1008	22.18
8	4	0.008	$1344	10.75
9	5	0.003	$1680	5.04
Totals		0.999		$105.51

Table 14.1.5: The cost of overtime with two teams of two CSI investigators

Q22.   What is the expected savings on overtime costs per day if a fourth CSI were hired?

Q23.   On average, how much would be saved on overtime over a ten-day period?

Q24.   Do you think the deputy chief should hire another CSI? Why or why not?

14.1.4 Weekly Calculation of Overtime

The Deputy Chief reviewed the overtime. He noticed that in the first week there were a total of 14 crime scenes, one less than the weekly average of 15. Nevertheless, there were three cases handled on overtime on Monday and Tuesday. He was considering renegotiating the wage contract so that overtime was paid on a weekly basis rather than a daily basis. He would be willing to increase the base salary by 10% if the investigators would accept the new overtime rules. With the new rules, there would have been no overtime the first week. The second week, there were 18 calls, 3 more than the average. The investigators would have been paid overtime for three crime scenes in the second week. The pay raise would increase the base salary to $35.20 per hour (1.1 · $32). The overtime salary would now be $52.80 per hour (1.5 · $35.20).

He decides to first see whether or not this policy would have saved money during the most recent two week pay period. He will then attempt to calculate the new expected value for this policy and compare it to the current policy.

The base bi-weekly salary was $7,200. With the 10% pay raise, this increases by $720 dollars to $7,200 + $720 = $7,920.

The cost for overtime would have become.
Week 1 overtime = $0
Week 2 overtime = (3 cases)(3 people)(2.5 hours/case)($52.80 per person per hour) =$1,188

Thus the total cost would have been
        $7,920 + $1,188 = $9,108

The current overtime structure and salary cost $7,200+ $2,520 = $9,720

This would have been a savings of $612 for these two weeks. The results seemed encouraging so he moved ahead with the expected value calculation.

Q25.   What is the cost of investigating each overtime case under this policy?

One of the fundamental theorems of the Poisson Distribution is the addition theorem. If you add together Poisson random variables, the sum is still Poisson distributed. Thus if we add the total number of calls in five days, it will be Poisson distributed. If $\lambda=3$/day, then the five day average will be just $\lambda t = 3*5 = 15$. Now, when we apply the Poisson probability density function to a work-week rather than a single day, we use λt instead of just λ:

$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

$$= \frac{15^x e^{-15}}{x!}$$

In Table 14.1.6, we have tabulated the probability values for this Poisson distribution beginning with 16 crime scenes. For completeness we included the probability that the number of crimes is 15 or less.

Q26.   Why do we not need the specific probabilities for 15 or fewer crime scenes?

			Overtime	
x	Probability	Number of Overtime Cases	Cost of Overtime $C(x)$	Contribution to Expected Overtime Cost/Week $P(x) \cdot C(x)$
$x \le 15$	0.568	0	0	0
16	0.096	1	$396	$38.03
17	0.085	2	$792	$67.11
18	0.071	3	$1,188	$83.89
19	0.056	4	$1,584	$88.30
20	0.042	5	$1,980	$82.78
21	0.030	6	$2,376	$70.96
22	0.020	7	$2,772	$56.44
23	0.013	8	$3,168	$42.07
24	0.008	9	$3,564	$29.58
25	0.005	10	$3,960	$19.72
26	0.003	11	$4,356	$12.51
27	0.002	12	$4,752	$7.58
28	0.001	13	$5,148	$4.40
Totals	0.999			$603.39

Table 14.1.6: Cumulative Poisson Probability Table for $\lambda t = 15$

Use the table to answer the following questions.

Q27.   What is the probability that there will be 5 or fewer overtime cases?

Q28.   What is the probability that there will 10 or more overtime cases in a week?

Q29.   What is the probability that the cost of overtime for the week will be less than $1,000? Between
         $1,000 and $2,000? More than $3,000 for a week?

The expected value of *weekly* cost for overtime under this new policy is $603.39. Thus the expected total
bi-weekly cost will be:
         $7,920 + 2 · $603.39 = $7,920 + $1,206.78 = $9,126.78

Q30.   Is this more or less than the current expected total bi-weekly cost and by how much?

Q31.   If the new policy costs more, under what conditions would it cost less? If the new policy costs
         less, under what conditions would it cost more?

Q32.   What is the probability that there is no overtime ten days in a row?

14.1.5 Simulate Variability in Savings for Two Policies

The Deputy Chief could see that the new overtime policy would save on average more than $470.82 every
two weeks. However, he was concerned that in some bi-weekly pay periods the new policy might end up
costing more money. The Deputy wondered how often that might happen. He understood that if every day
in a week the number of crime scenes was three or less, the new policy of weekly overtime would save no
overtime costs. The higher base salary would therefore cost more money that week. He had trouble
imagining other ways that the policy might cost more. He decided to have an operations researcher
simulate 25 bi-weekly pay periods. This was equivalent to a year when you leave off two weeks for
vacations. The results are presented in Table 14.1.7

Week and Day	Pay Period Number												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Monday	0	2	2	4	2	5	5	4	5	2	4	2	5
1. Tuesday	3	3	3	4	3	1	4	2	5	3	3	3	8
1. Wednesday	3	4	2	1	4	2	1	1	5	1	5	0	4
1. Thursday	1	1	2	1	4	2	2	5	3	2	1	2	5
1. Friday	6	1	6	3	3	6	0	3	1	2	1	1	4
Week 1 Total	13	11	15	13	16	16	12	15	19	10	14	8	26
2. Monday	0	4	2	1	4	0	3	2	3	9	2	5	1
2. Tuesday	3	0	1	1	5	0	4	1	3	0	2	5	4
2. Wednesday	3	5	1	3	3	3	3	2	5	6	3	6	5
2. Thursday	0	2	4	2	2	3	4	1	2	1	3	4	2
2. Friday	4	4	3	3	2	3	2	2	2	4	1	3	7
Week 2 Total	10	15	11	10	16	9	16	8	15	20	11	23	19
Bi-weekly Total	23	26	26	23	32	25	28	23	34	30	25	31	45
Daily Average	2.3	2.6	2.6	2.3	3.2	2.5	2.8	2.3	3.4	3.0	2.5	3.1	4.5
Policy													
Daily Overtime	8640	9000	8640	7920	9000	9000	9000	8280	10080	10800	8280	10080	13680
Weekly Overtime	7920	7920	7920	7920	8712	8316	8316	7920	9504	9900	7920	11088	13860
Net Improvement	720	1080	720	0	288	684	684	360	576	900	360	-1008	-180

Week and Day	Pay Period Number												
	14	15	16	17	18	19	20	21	22	23	24	25	
1. Monday	4	5	0	5	2	0	3	5	8	1	2	1	
1. Tuesday	5	2	1	2	4	6	1	2	1	1	3	4	
1. Wednesday	2	3	3	4	5	1	3	2	4	4	3	2	
1. Thursday	3	2	2	1	4	4	2	5	5	2	1	3	
1. Friday	3	5	4	1	0	3	3	5	1	1	2	2	
Week 1 Total	17	17	10	13	15	14	12	19	19	9	11	12	
2. Monday	1	3	5	3	4	3	4	2	8	5	3	3	
2. Tuesday	0	3	2	3	3	3	5	2	6	3	4	5	
2. Wednesday	5	4	1	7	2	2	5	5	7	4	9	2	
2. Thursday	2	6	1	2	2	3	4	4	2	2	6	6	
2. Friday	4	3	4	1	3	6	4	2	4	2	2	6	
Week 2 Total	12	19	13	16	14	17	22	15	27	16	24	22	
Total	29	36	23	29	29	31	34	34	46	25	35	34	Avg.
Daily Average	2.9	3.6	2.3	2.9	2.9	3.1	3.4	3.4	4.6	2.5	3.5	3.4	3.02
Policy													
Daily Overtime	9360	10080	8640	9720	9000	9720	9720	10440	14760	8640	10800	10440	
Weekly Overtime	8712	10296	7920	8316	7920	8712	10692	9504	14256	8316	11484	10692	Avg.
Net Improvement	648	-216	720	1404	1080	1008	-972	936	504	324	-684	-252	$387

Table 14.1.7: Simulation of CSI in 250 Weekdays Grouped into 25 Pay Periods

The Table presents the randomly generated number of CSI each day for 250 weekdays. These data were generated according to the Poisson Distribution in an EXCEL spreadsheet.

Reading the Table
For example, in the week 1 of the first pay period there were 0 CSI on Monday, 3 on Tuesday, and a high of 6 on Friday. There were a total of 13 CSI in the first week and 10 CSI in the second week. The bi-weekly total was 23 CSI, for an average of 2.3 per day. Under the current policy, there would have been overtime paid on both the Friday of week 1 with 6 CSI and Friday of week 2 with 4 CSI. However, if overtime were paid only on a weekly basis there would have been no overtime costs. The total cost for the current policy is $8,640 and with weekly overtime is $7,920.

In building a mathematical model in a spreadsheet, there is always the possibility of a careless mistake. It is always a good idea to peruse the basic results to see if anything extreme stands out as a possible error. For example, the average daily CSI is 3.0. If the spreadsheet model was built correctly, it is extremely unlikely that the average for a simulated 250 weekdays would be below 2.5 or higher than 3.5. (This concept of a reasonable range for a sample average is developed in later chapters.)

Q33. What was the average for this simulated sample and does it seem reasonable?

He also wanted to look at the relative frequency distribution and compare it to the probability distribution. Table 14.1.8 has the relative frequency for the simulated data. Figure 14.1.3 presents the same graphically.

Q34. Are the relative frequency values close to the probabilities reported in Table 14.1.1? Do any values stand out as very different from the probabilities?

Bin	Absolute Frequency	Relative Frequency
0	13	0.052
1	38	0.152
2	56	0.224
3	52	0.208
4	40	0.160
5	31	0.124
6	12	0.048
7	3	0.012
8	3	0.012
9	2	0.008
10	0	0
More	0	0
Sum	250	1

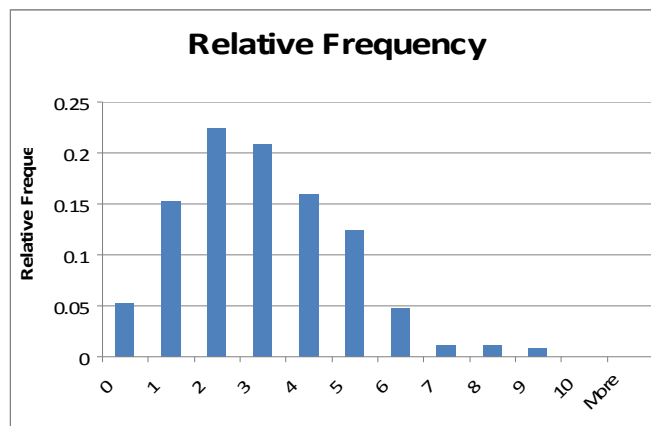**Table 14.1.8**: Relative frequency simulated data

Figure 14.1.3: Relative Frequency of Simulated Data

At the bottom of Table 14.1.7 are the calculated total costs. In this 1st pay period, there were 3 overtime CSI on Friday Week 1 and 1 overtime CSI on Friday in week 2. The total cost for this first pay period would be $7,200 for regular pay plus $360 for each overtime investigation.

$7,200 + 4($360) = $8,640

With the weekly overtime policy, there would be no overtime cost. Each week had fewer than 15 CSIs. Thus, the total cost would simply be the base pay of $7,920. The net savings for this 1st pay period is just

Net Savings = $8,640 - $7,920.

Q35. For which time period was the total cost the highest? Recreate the calculations of total cost for that pay period for both overtime policies.

Q36. For which time period was the total cost the lowest? Recreate the calculations of total cost for that pay period for both overtime policies.

As was mentioned, the Deputy Chief could envision a week or a bi-weekly pay period with no overtime that would lead to higher costs with the new policy. He wondered if this was likely to happen.

Q37.   What is the probability that there is no overtime five days in a row? Ten days in a row?

This probability was so small that he was not too concerned about this occurrence. He therefore was
interested in what new insights the simulation could provide about the relative costs of these two overtime
policies.

Q38.   What percent of the time did the new policy cost more than the old policy?

Q39.   What was the largest increased cost and during which pay period?

Q40.   How would you explain why the new policy cost so much more during those two weeks?

Q41.   Which other bi-weekly pay period had a similar pattern?

The Deputy Chief noticed that if every day in a week there were at least three CSIs, the new policy again
costs more.

Q42.   What is the probability that there would be three or more CSIs on a day? (Refer back to Table
        14.1.1.)

Q43.   What is the probability that there would be three or more CSIs on every day during a week?

Q44.   How many weeks experienced this pattern? Did these usually correspond to a net loss with the
        new policy? Is this consistent with the probability you calculated?

The Deputy Chief decided to take a closer look at the savings of the new policy.

Q45.   What was the largest savings and during which pay period?

Q46.   How many times did the new policy save more than $1,000 during a pay period?

Q47.   How did the average savings in this simulated example compare to the predicted average savings?

Section 14.2: Scheduled and Urgent Patients at a Health Care Clinic

Dr. Mahelia Johnson is the primary care physician at a small neighborhood clinic during the morning. She is on site from 8 AM until noon before going to the local hospital to supervise interns and see her patients who have been admitted. With the help of nurses and physician assistants she can see an average of four patients an hour. This is a total of 16 patients per morning. The majority of Dr. Johnson's time is spent consulting with patients that have been scheduled a week in advance. Besides scheduled patients, Dr. Johnson wishes to make time for patients with an urgent problem who call up early in the morning requesting an appointment that morning. On average there are two urgent requests per day. The actual number is a random variable that is Poisson distributed.

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$
$$= \frac{2^x e^{-2}}{x!}$$

Based on the average, Dr. Johnson is thinking of scheduling only 14 patients per day and leaving open the remaining two time slots to handle the average number of urgent patients.

Q1.   In what ways does using this average value not describe what might happen on any particular day?

For this policy to work perfectly, there would have to be exactly two urgent patient calls in a day.

Q2.   What is the probability of exactly two urgent patients in a morning?

Q3.   If fewer than two patients call in, there will be unused slots. What is the likelihood of this happening?

There is also the possibility that there will be more urgent patients than the slots she has made available.

Q4.   Use the answers to the above questions and the principle of complement to determine that likelihood of more than two urgent patients.

In an uncertain environment there is no one policy that will turn out to be best on every day. Dr. Johnson can never be sure that the number of slots reserved for urgent patients will be needed. Nor can she ever be sure that she will have enough slots for all of the urgent patients. Dr. Johnson is torn between her desire to see as many urgent patients as possible but also not wasting any of the time slots. Initially, her primary concern was with the urgent patients. If she could not treat a patient, Dr. Johnson would have to send him to an emergency room instead. This significantly increases the cost of medical care. It also usually means long waits in the emergency room until the patient is seen. She was considering leaving enough open slots, such that more than 90% of the time she would be able to see all of the urgent patients calling in that day.

Q5.   Use table 14.2.1 to determine how many slots Dr. Johnson would set aside to meet this 90% goal?

Q6.   With this policy how frequently would there be two or more unutilized slots? If four slots are set aside, there will be two or more unutilized slots if the number of urgent patients is 0, 1 or 2.

x	$P(x)$	Cumulative Probability
0	0.135	0.135
1	0.271	0.406
2	0.271	0.677
3	0.180	0.857
4	0.090	0.947
5	0.036	0.983
6	0.012	0.995
7	0.003	0.998
8	0.001	0.999

Table 14.2.1: Poisson Distribution with $\lambda = 2$

In order to determine the best policy, Dr. Johnson decided to perform an economic analysis. She recruited an industrial engineering undergraduate student, Marie Staats, at the local university. Ms. Staats determined that for each unused time slot, the clinic lost $138 in wasted resources. However, if an urgent patient was sent to an emergency room, the added cost for that visit was $285.

Table 14.2.2 presents the detailed information Ms. Staats will use to determine the expected value of the cost of each policy.

$$E(C(x)) = \sum [P(x) \cdot C(x)]$$

Under each policy and for each value of x, the table specifies the number of unused slots or the number of patients sent to the emergency room. For example, assume the policy is to reserve just one slot. Then if there are no urgent cases, there will be one unused slot, U 1. The cost is $138. However, if there were 4 urgent patients, then 3 would have to be sent to an emergency room, E 3. The associated cost would be 3*285 = $855. The expected cost for this policy is an average of $341 per day. A similar analysis for a policy that reserves 4 slots would cost on average $307 per day.

		Policy							
		Reserve 1		Reserve 2		Reserve 3		Reserve 4	
x	$P(x)$	U or E	$C(x)$	U or E	$C(x)$	U or E	$C(x)$	U or E	$C(x)$
0	0.135	U 1	138					U4	552
1	0.271	0	0					U3	414
2	0.271	E 1	285					U2	276
3	0.180	E 2	570					U 1	138
4	0.090	E 3	855					0	0
5	0.036	E 4	1140					E 1	285
6	0.012	E 5	1425					E 2	570
7	0.003	E 6	1710					E 3	855
8	0.001	E 7	1995					E 4	1140
Expected Total Cost			$341						$307

Table 14.2.2: Expected Value Analysis of Different Reserved Slot Policies

Q7.   Complete a similar table for policies that reserve two or three slots each day. What is the expected value for each of these policies?

Q8.   Given Dr. Johnson's concern for urgent patients and cost consciousness, what policy would you recommend and why?

After being presented the results, Dr. Johnson asked Marie to evaluate one additional policy. The doctor was considering reserving just two slots. However, if three or more urgent patients call in, she will squeeze in one more patient and come 15 minutes late to the hospital.

Q9.   How often would she come late to the hospital?

Q10.   For what values of x are there no added costs with this policy?

Q11.   Complete a table similar to Table 14.2.2 for this new policy. What is the expected value of the cost of this policy?

Managing the mix of scheduled and urgent patients is even more complex than we have described. All clinics have to deal with scheduled patients who do not show up for their appointments. The percentage of *no shows* can exceed 20% in some clinics. Thus the number of scheduled patients who actually come is a random variable. That random variable follows the Binomial Distribution, discussed in the previous chapter. The total number of patients, scheduled and urgent, is then the sum of two distinct random variables. To model this problem we would use random number generators to simulate the two groups of patients and determine the various costs. To reduce the likelihood of *no shows*, many health care providers pay staff to call all scheduled patients a day or two before their scheduled appointments. This practice can cut the rate of *no shows* by half. Modeling this problem, however, is beyond our scope.

Section 14.3: Exponential Distribution—Time Between Calls at a Dispatch Center

The City of Holdings police dispatch center receives on average four calls per hour. As is typical, the probability distribution of calls follows the Poisson Distribution. This also means that the average time between calls is one-quarter of an hour or 15 minutes. When using the Poisson probability density function, the key parameter is the average arrival rate multiplied by the time interval. Thus if we are interested in a two hour time period, we would use $\lambda t = 4*2 = 8$ as the parameter for the Poisson distribution. On average there are eight calls in this two-hour period. This principle also applies to time periods shorter than an hour.

For example, the average number of calls in a half hour is just $\lambda t = 4*(1/2) = 2$. However, the key parameter, the average, does not have to be an integer. If we were interested in a 20-minute interval, $\lambda t = 4*(1/3)$ or 4/3. There may not be a standard table for non-integer values but the Poisson probability density function can be used to calculate probabilities.

For example what is the probability that there will exactly one call during a twenty minute time interval?

$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

$$P(1) = \frac{\left(\frac{4}{3}\right)^1 e^{-\frac{4}{3}}}{1!}$$

$$P(1) \approx 0.351$$

In Holdings, the dispatch operator is allowed to take a five minute break each hour. During the break, his job is covered by a less experienced trainee. He usually takes a break immediately after completing his first call during an hour. What is the probability that the trainee will handle no calls? Five minutes is one-twelfth of an hour.

Thus, $\lambda t = 4*(1/12) = 1/3$.

$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

$$P(0) = \frac{\left(\frac{4}{12}\right)^0 e^{-\frac{4}{12}}}{0!}$$

$$P(0) = \frac{1e^{-\frac{1}{3}}}{1}$$

$$P(0) \approx 0.717$$

It is likely that the trainee will not have to handle any calls. If there are no calls during his five minute break, the time between the last call and the next call must be at least five minutes. The first statement implies the second statement.

0 calls in five minutes → the time between calls is more than five minutes

These two statements may seem redundant. However, one is phrased in terms of the number of calls, zero.
The other is phrased in terms of the time between calls, longer than five minutes. Let's see if the
implication goes in both directions. Does the following first statement imply the second?

The time between calls is more than five minutes → 0 calls in five minutes

It does. If we put these statements in probabilistic terms where t represents time before first call then
$$P(x = 0 \text{ in five minutes}) = P(t > 5 \text{ minutes})$$

$$P(x = 0 \text{ in five minutes}) = e^{-\frac{4}{12}} = P\left(t > \frac{1}{12} \text{ hour}\right)$$

From a mathematics perspective, there is nothing special about five minutes. Let's generalize our
statements in terms of "t" hours.

0 calls in t hours → the time between calls is more than t hours
$$P(x = 0 \text{ in } t \text{ hours}) = e^{-4t} = P(T > t \text{ hours})$$
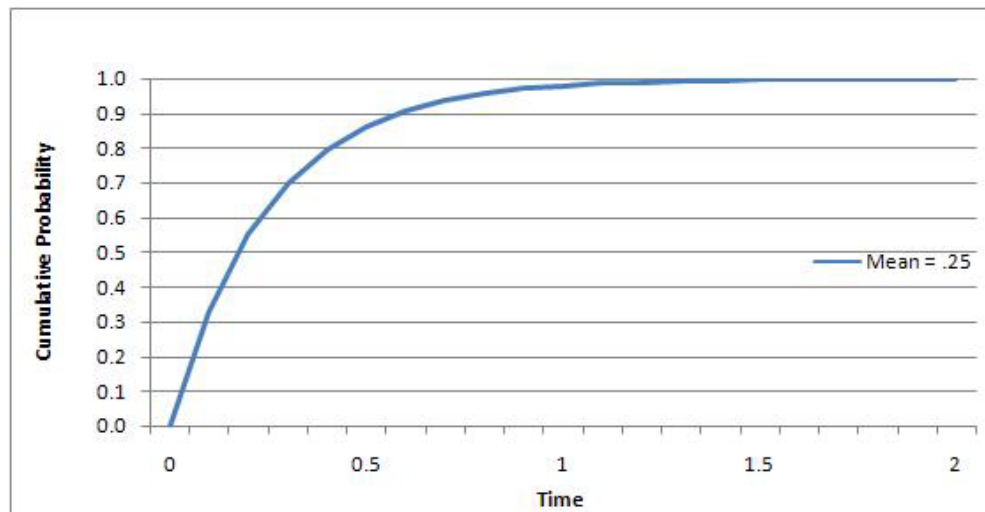

Figure 14.3.1: Exponential Cumulative Distribution: Mean = 0.25 hours

In our example $\lambda = 4$. We can generalize this formula for all values of lambda:
$$P(X = 0 \text{ in } t \text{ hours}) = e^{-\lambda t} = P(T > t \text{ hours})$$

We have just presented the formula for the continuous random variable, T, the time between calls. The
formula is presented in terms of calls longer than t. If we were interested in the likelihood of the interval
between calls being less than t, we would take the complement.
$$P(T < t \text{ hours}) = 1 - P(T > t \text{ hours})$$
$$= 1 - e^{-\lambda t}$$

This formula is the cumulative distribution formula for the Exponential Distribution.

As we noted earlier, the average time between calls is just the inverse of the arrival rate. If there are four calls per hour, the average time between calls is ¼ of an hour. In general.

$$E(T) = \frac{1}{\lambda}.$$

Returning to our original situation, we were interested in the probability that the time between two consecutive calls will be less than 5 minutes, 1/12 hour.

$$P\left(T < \tfrac{1}{12}\right) = 1 - e^{-\frac{4}{12}}$$

$$\approx 1 - 0.717$$

$$\approx 0.283$$

Once every four hours the dispatcher is allowed to take an extended 15 minute break, ¼ hour. When he goes on break, what is the probability that the next call will come in before his break ends?

$$P\left(T < \tfrac{1}{4}\right) = 1 - e^{-\frac{4}{4}}$$

$$= 1 - e^{-1}$$

$$\approx 0.632$$

Thus it is likely that the trainee will have to handle a call during the extended break.
Recall that since the average arrival rate is four per hour, the average time between calls is a quarter of an hour. Surprisingly the probability that the time between arrivals is less than the average is 0.632. The ratio of below average intervals to above average is 63:37 or 1.7:1.

We performed this calculation for λ=4. However, this 0.632 probability applies to the average for all exponential distributions.

$$P\left(T < \tfrac{1}{\lambda}\right) = 1 - e^{-\frac{\lambda}{\lambda}}$$

$$= 1 - e^{-1}$$

$$\approx 0.632$$

Many people confuse the average with the median and therefore believe that the chance of being below average is 0.5. This is true for symmetric distributions but is certainly not true for the exponential distribution.

The questions we raised with regard to *T* were framed as the probability of being more than some value or less than some other value. We never asked, for example, "what is the probability that the interval between two calls would be exactly .12513 hours? With a continuous random variable such as time, there is an infinitesimal likelihood that a specific value measured to many decimal places will occur. Instead probability statements are developed for a range. We can ask, what is the probability that the time between two consecutive calls will be more than 0.1 hours but less than 0.2 hours?

$$P(0.1 < T < 0.2)$$

A fundamental theorem of probability is that this can be determined by

$$P(0.1 < T < 0.2) = P(T < 0.2) - P(T < 0.1)$$

In general: $P(a < T < b) = P(T < b) - P(T < a)$.

The range for $T < 0.2$ includes all values from zero to 0.2 including those below 0.1. However, since we are interested only in values between 0.1 and 0.2, we need to subtract out the probability of $T < 0.1$.

$$P(T < 0.2) = 1 - e^{(-4)(0.2)}$$
$$= 1 - e^{-0.8}$$
$$\approx 0.551$$

$$P(T < 0.1) = 1 - e^{(-4)(0.1)}$$
$$= 1 - e^{-0.4}$$
$$\approx 0.330$$

$$P(0.1 < T < 0.2) = P(T < 0.2) - P(T < 0.1)$$
$$\approx 0.551 - 0.330$$
$$\approx 0.221$$

Thus there is a 0.221 probability that the time between two consecutive calls will be more than 0.1 hours, six minutes, and less than 0.2 hours or 12 minutes.

The dispatcher is allowed a half hour for lunch. Usually during lunch, he turns over the desk to a supervisor. However, when there is no supervisor available, he must eat his lunch at his desk. The department has agreed to pay him an extra half hour's salary, if his lunch break is interrupted by even one call. They will give him an extra hour's pay, if he is unable to complete even the first fifteen minutes without an interruption.

He will therefore receive no extra pay for $T > 0.5$ hours.

$$P(T > 0.5) = e^{(-4)(0.5)}$$
$$= e^{-2}$$
$$\approx 0.135$$

The dispatcher will receive an extra hour of pay for $T < 0.25$ hours.

$$P(T > 0.25) = 1 - e^{(-4)(0.25)}$$
$$= 1 - e^{-1}$$
$$\approx 1 - 0.368$$
$$\approx 0.632$$

He will receive an extra half hour of pay for $0.25 < T < 0.5$ hours.

$$P(0.25 < T < 0.5) = P(T < 0.5) - P(T < 0.25)$$
$$\approx (1 - 0.135) - (1 - 0.368)$$
$$\approx 0.233$$

On average how much extra hours of pay, H, will the dispatcher earn sitting at his desk while eating?

$$E(H) \approx (0)(0.135) + (0.5)(0.233) + (1)(0.632)$$
$$\approx 0.748 \text{ hours}$$

The dispatcher will earn zero hours with probability 0.135. He will earn a half hour of extra pay with 0.233 probability. Most likely, the dispatcher will earn one full hour of extra pay with probability 0.632. On average, he will therefore earn 0.75 hours of extra pay.

Section 14.4: When Will Firefighters Sleep?—Exponential Distribution

In the previous section, we explored interruptions during break times. We analyzed break times of different lengths: 5 minutes, 15 minutes, and a half hour. The issue of interruptions has important significance for firefighters. A common firefighter schedule involves working for 24 hours straight and then being off for 48 hours. The firefighters are on-duty and ready to respond far more than 40 hours per week, but they are only paid for daytime hours. One concern is the frequency with which their nightly sleep is interrupted by late night calls for service. Frequent sleep interruptions increase the stress of the job. In addition, firefighters awoken from deep sleep may initially be disoriented. In some locales firefighters are paid extra if they do not get, for example, at least five hours of uninterrupted rest.

The normal sleep cycle includes four stages of non-REM (rapid eye movement) sleep followed by a stage of REM sleep. Sleep cycles typically last between 90 and 110 minutes. During the non-REM sleep phase, the human body heals itself physically: repairing muscles and the immune system. During REM sleep, the body heals itself mentally. Most of the benefits from sleep are gained during the first two sleep cycles. Almost all of the benefits are achieved by the end of the third cycle. An individual completes four or five sleep cycles in a typical night of seven or eight hours of sleep. The last cycles have a much higher percentage of time spent in REM sleep. Stages 3 and 4 involve the deepest sleep cycles during non-REM sleep. It is hard to wake an individual during these stages; when woken it can take several minutes to become oriented and fully awake.

Each individual's sleep cycle is different. In addition the person's sleep cycle can vary from night to night. Table 14.3.1 contains a typical sequence of three sleep cycles.

Sleep Cycle 1		Sleep Cycle 2		Sleep Cycle 3	
Stage	Minutes	Stage	Minutes	Stage	Minutes
S1 & S2	30	S1 & S2	25	S1 & S2	25
S3, S4, S3	45	S3, S4, S3	35	S3	15
S2	15	S2	20	S2	15
REM	10	REM	20	REM	35
Total	100		100		90

Table 14.4.1: Stage Duration in a Series of Sleep Cycles

A city's fire call rate is generally much lower than its corresponding police call rate. The fire call rate of the City of Holdings is approximately 0.1 calls per hour between the hours of 10 PM and 7 AM. These hours are considered the firefighters' rest period.

Q1.   What is the probability that no fire call will interrupt the first two sleep cycles?

The first two sleep cycles last a total of 200 minutes, 3.33 hours. Therefore we are interested in

$$P\left(T > \tfrac{10}{3}\right) = e^{(-0.1)\left(\frac{10}{3}\right)}$$
$$\approx 0.717$$

Q2.   What is the probability that no fire call will interrupt the first three sleep cycles?

The first three sleep cycles last a total of 290 minutes, 4 and 5/6 hours. Therefore we are interested in

$$P\left(T > \tfrac{29}{6}\right) = e^{(-0.1)\left(\frac{29}{6}\right)}$$
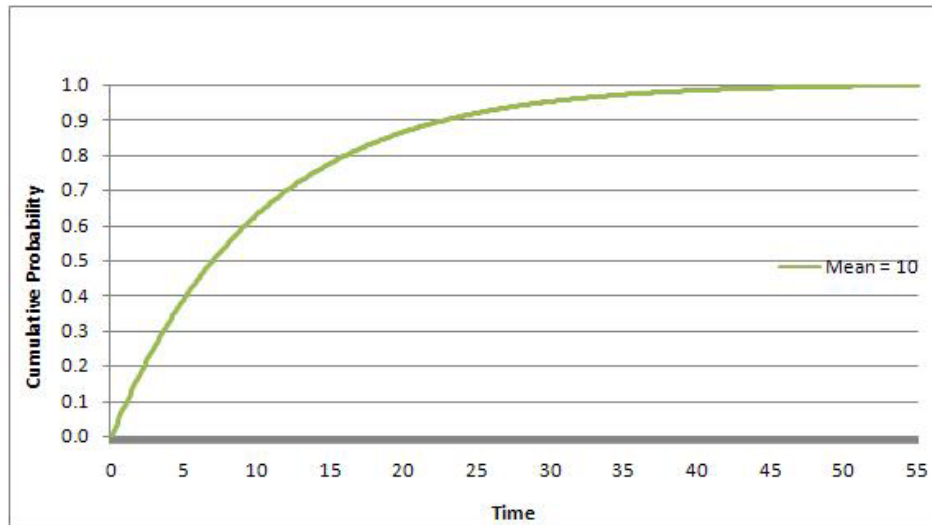
$$\approx 0.617$$

Figure 14.4.1: Exponential Cumulative Distribution: Mean = 10 hours

Q3.   What is the probability that there will be no fire call during the entire rest period of nine hours?

$$P(T > 9) = e^{(-0.1)(9)}$$

$$\approx 0.406$$

This is the same as having zero calls in a nine hour period.

$$P(X = 0) = \frac{\left[(0.1)(9)\right]^{0} e^{(-0.1)(9)}}{0!}$$

$$\approx 0.406$$

Of special concern are sleep interruptions that occur during deep sleep (i.e., stages three or four) of the first two cycles.

Q4.   What is the probability that the first fire call will occur during the deep sleep stage that occurs between 30 and 75 minutes of the rest period?

$$P(0.5 < T < 1.25) = P(T < 1.25) - P(T < 0.5)$$

$$= \left(1 - e^{(-0.1)(1.25)}\right) - \left(1 - e^{(-0.1)(0.5)}\right)$$

$$= e^{(-0.1)(0.5)} - e^{(-0.1)(1.25)}$$

$$\approx 0.951 - 0.882$$

$$\approx 0.069$$

This is approximately equivalent to an interruption once every 14 days.

$$\frac{1}{0.069} \approx 14$$

Q5.   What is the probability that the first fire call will occur during the deep sleep stage that occurs between 125 and 160 minutes of the rest period.

$$
\begin{aligned}
P(2.08 < T < 2.67) &= P(T < 2.67) - P(T < 2.08) \\
&= \left(1 - e^{(-.01)(2.67)}\right) - \left(1 - e^{(-0.1)(2.08)}\right) \\
&= e^{(-0.1)(2.08)} - e^{(-0.1)(2.67)} \\
&\approx 0.812 - 0.766 \\
&\approx 0.047
\end{aligned}
$$

This is approximately equivalent to an interruption once every 3 weeks.

$$\frac{1}{0.047} \approx 21$$

The analysis so far focused on the time from the start of the rest period until the first call. It is possible that a call could occur early in the rest period and yet later in the rest period they may still get five hours of continuous sleep. It is difficult to develop a simple formula to determine the probability that firefighters will get five hours of continuous sleep during the nine hour rest period. Instead in the optional Appendix at the end of this chapter we simulate the situation using a random number generator.

Section 14.5: Exponential Distribution—Reliability and Service Time

In the previous section we introduced the exponential distribution to describe the time between incidents in a Poisson Process. This distribution is also used to model other situations. It is used to model the lifetime of electronic components. One characteristic of these types of devices is that they do not wear out but rather fail at random. The exponential distribution is also used sometimes to describe the time to complete a task such as servicing a customer.

Suppose the lifetime distribution of a critical component is exponentially distributed with a mean of 2,000 hours. The component costs $150. There is a full money back warranty if the component fails in less than 500 hours. If it fails after 500 hours but before 1,000 hours, the refund is 50%. There is no warranty beyond 1,000 hours.

The expected time, $E(T)$, for the component to fail is the mean of the exponential distribution. It also equals the reciprocal of the parameter of the distribution. We can use this property to determine λ for this situation.

$$E(T) = \frac{1}{\lambda}$$

$$2000 = \frac{1}{\lambda}$$

$$\lambda = \frac{1}{2000}$$

Q1.   What is the probability the component will fail in less than 500 hours?

$$P(T < 500) = 1 - e^{\left(-\frac{1}{2000}\right)(500)}$$

$$= 1 - e^{-0.25}$$

$$\approx 0.221$$

Q2.   What is the probability the component fails after 500 hours but before 1,000 hours?

$$P(500 < T < 1000) = P(T < 1000) - P(T < 500)$$

$$= \left(1 - e^{\left(-\frac{1}{2000}\right)(1000)}\right) - \left(1 - e^{\left(-\frac{1}{2000}\right)(500)}\right)$$

$$= e^{\left(-\frac{1}{2000}\right)(500)} - e^{\left(-\frac{1}{2000}\right)(1000)}$$

$$= e^{-0.25} - e^{-0.5}$$

$$\approx 0.779 - 0.607$$

$$\approx 0.172$$

Q3.   What is the expected value of the warranty cost?

$$(0.221)(\$150) + (0.172)(\$75) + (0.607)(\$0) = \$45.75$$

Q4.   The company wants to set the full money back warranty at the 10th percentile. For what value of T
        is there a 10% failure rate?

$$P(T < t) = 1 - e^{(-t)\left(\frac{1}{2000}\right)}$$

$$0.10 = 1 - e^{-\frac{t}{2000}}$$

$$0.9 = e^{-\frac{t}{2000}}$$

$$\ln(0.9) = \ln\left(e^{-\frac{t}{2000}}\right)$$

$$-0.105 \approx -\frac{t}{2000}$$

$$t \approx 210 \text{ hours}$$

The time to service a bank customer is exponentially distributed with a mean of seven minutes.

Q5.   What proportion of bank customers are served in five minutes or less?

$$P(T < 5) = 1 - e^{-\frac{5}{7}}$$

$$\approx 0.510$$

Q6.   What proportion of bank customers take more than 10 minutes?

$$P(T > 10) = 1 - P(T < 10)$$

$$= 1 - \left(1 - e^{\left(-\frac{1}{7}\right)(10)}\right)$$

$$= e^{-\frac{10}{7}}$$

$$\approx 0.240$$

The company wants only 5% of the customers to take 10 minutes or longer. How much would they have
to decrease the average to achieve this standard?

$$P(T > t) = e^{-10\lambda} = 0.05$$

Take the natural logarithm of both sides of the equation.

$$\ln\left(e^{-10\lambda}\right) = \ln(0.05)$$

$$-10\lambda \approx -2.996$$

$$\lambda \approx 0.300$$

$$E(T) = \frac{1}{\lambda}$$

$$= \frac{1}{0.300}$$

$$\approx 3.33 \text{ minutes}$$

The bank would have to reduce the average service time to 3.33 minutes if they do want to exceed ten minutes 5% of the time.

Appendix: **Simulate the Likelihood of Five Hours Sleep during Nine Hours**

Simulation Process

1. Generate a random number between zero and one. This represents the probability that the time between calls will be less than t. In step 2 we find the corresponding value of t.
2. Invert an equation to convert that random number into a time value for the time interval between calls. This inversion is necessary because the random number generator gives numbers that are uniformly distributed, and our model requires the random numbers to be exponentially distributed. Convert decimal version to minutes.
3. Generate enough numbers until you have a call that starts after the rest period ends.
4. Place the sequence of calls on time clock
5. Assume a call will be completed a half hour after it begins.
6. Consider the last call that starts before the end of the rest period. Determine the end of the call; add a half hour to the start time. Calculate the time left after the call until the nine-hour rest period ends.
7. Determine whether or not there is a five-hour interval between the completion of a call and the start of the next call or before the end of the rest period.

Table 14.4.1 provides data for 28 days of simulating the call arrivals in the nine hour rest period. We will walk through the first two simulated rest periods.

Day 1

1. Random number = 0.492. This number will be used to find the cumulative probability for a corresponding value of T, $F(T)$. The cumulative distribution function for the exponential distribution is given below.

    $$F(T) = 1 - e^{-\lambda T}$$

2. Convert the random number to time by solving the following equation for T using the natural logarithm.

    $$F(T) = 1 - e^{-\lambda T}$$

    $$0.492 = 1 - e^{(-0.1)(T)}$$

    $$1 - 0.492 = e^{-0.1T}$$

    $$\ln(0.508) = \ln\left(e^{-0.1T}\right)$$

    $$-0.677 = -0.1T$$

    $$T = 6.77 \text{ hours} \rightarrow 6 \text{ hours} + (0.77 \cdot 60) \text{ minutes} \approx 6:46$$

Repeat steps 1 and 2 for the next call.

    Random number = 0.753

3.

$$F(T) = 1 - e^{-\lambda T}$$

$$0.753 = 1 - e^{(-0.1)(T)}$$

$$1 - 0.753 = e^{-0.1T}$$

$$\ln(0.247) = \ln\left(e^{-0.1T}\right)$$

$$-1.398 = -0.1T$$

$$T = 13.98 \text{ hours} \rightarrow 13 \text{ hours} + (0.98 \cdot 60) \text{ minutes} \approx 13:59$$

This second call would occur almost 14 hours after the first call. This is clearly after the conclusion of the 9 hour rest period.

4. 6:46 and then 20:45
5. The first call ends 6:46 + 0:30 = 7:16
6. The time from 7:16 until end of the nine-hour time period is 1:44.
7. Yes. They would have a continuous block of more than five hours at the beginning of the rest period.

Day 2
1. Random number = 0.238. This number will correspond to the cumulative probability for a corresponding value of T.
2. Convert to time by solving the following equation for T.

$$0.238 = 1 - e^{-0.1T}$$

$$T = 2.72 \rightarrow 2:43$$

Repeat steps 1 and 2 for the next call.
    Random number = 0.022
        $T = 0.22 \rightarrow 0:13$
    Random number = 0.256
        $T = 2.95 \rightarrow 2:57$
    Random number = 0.365
        $T = 4.53 \rightarrow 4:32$

3. If we add up the times, the fourth call begins after nine hours.
        2:43 + 0:13 + 2:57 + 4:32 = 10:27
4. The clock start times 2:43, 3:13, and 5:54. The second call's start time is delayed because it arrives during the 30 minute interval of the first call. Firefighters cannot start work on that call until 30 minutes after the first call.
5. The third call ends 5:54 + 0:30 = 6:24
6. The time from 6:24 until end of 9 hour time period is 2:35.
7. **No.** There was no continuous block of more than five hours during the rest period. The longest continuous sleep period was after the last call until the end of the rest period. It was only 2:35.

Exploring the Simulated Data
Q1.   How many rest periods were never interrupted by a call? What is the percentage? How does this compare to the calculated probability?

Q2.   How many rest periods were interrupted once? How many of these still had at least five hours of continuous sleep?

Q3.   How many rest periods were interrupted twice? How many of these still had at least five hours of
        continuous sleep?

Q4.   How many rest periods were interrupted three times? How many of these still had at least five
        hours of continuous sleep?

A total of 4 days out of 28 or 14% did not include a five-hour continuous sleep period.

Day	Random Number	Time Between Calls	Call Start Time	Call End Time	Intervals	Five continuous hours of sleep?
1	0.492	6:46	6:46	7:16	6:46	YES
	0.753	13:59	20:46	NA	1:44	
2	0.238	2:43	2:43	3:13		
	0.022	0:13	2:56 (3:13)	3:43		
	0.256	2:57	5:53	6:23		
	0.365	4:32	10:25	NA		
3	0.468	6:18	6:18	6:48	6:18	YES
	0.217	2:27	8:45	9:15	1:57	
4	0.221	2:29	2:29	2:59		
	0.298	3:31	6:01	6:31		
	0.282	3:19	9:20	NA		
5	0.112	1:11	1:11	1:41	1:11	
	0.849	18:53	20:04	NA	7:18	YES
6	0.001	0:00	0:00	0:30	0:00	
	0.053	0:32	0:33	1:03	0:02	
	0.731	13:07	13:40	14:10	7:56	YES
7	0.461	6:11	6:11	6:41	6:11	YES
	0.930	2:32	8:43	9:13	2:18	
8	0.517	7:16	7:16	7:46	7:16	YES
	0.825	17:25	0:41	1:11	1:13	
9	0.625	9:48	9:48	10:18	9:48	YES
10	0.772	14:46	14:46	15:16	9:00	YES
11	0.457	6:06	6:06	6:36	6:06	YES
	0.120	1:16	7:23	7:53	0:46	
	0.016	0:09	7:33	8:23 (8:03)	0:00	
	0.352	4:20	11:53	NA	0:37	
12	0.751	13:54	13:54	NA	9:00	YES
13	0.553	8:02	8:02	8:32	8:02	YES
	0.816	16:56	24:58	NA	0:27	
14	0.407	5:13	5:13	5:43	5:13	YES
	0.683	11:28	16:42	NA	3:16	
15	0.608	9:21	9:21	9:51	9:00	YES
16	0.392	4:58	4:58	5:28		
	0.125	1:20	6:18	6:48		
	0.617	9:35	15:53	NA		
17	0.554	8:04	8:04	8:34	8:04	YES
	0.485	6:38	14:42	NA	0:25	
18	0.838	18:12	18:12	NA	9:00	YES
19	0.325	3:55	3:55	4:25		
	0.981	15:40	19:36	NA		
20	0.783	15:15	15:15	NA	9:00	YES
21	0.448	5:56	5:56	6:26	5:56	YES

Day	Random Number	Time Between Calls	Call Start Time	Call End Time	Intervals	Five continuous hours of sleep?
	0.658	10:44	16:40	NA	2:33	
22	0.264	3:04	3:04	3:34	3:04	
	0.861	19:45	22:49	NA	5:25	YES
23	0.508	7:06	7:06	7:36	7:06	YES
	0.632	10:00	17:06	NA	1:24	
24	0.712	12:26	12:26	12:56	9:00	YES
25	0.395	5:01	5:01	5:31	5:01	YES
	0.498	6:53	11:55	12:25	3:28	
26	0.626	9:50	9:50	10:20	9:00	YES
27	0.519	7:18	7:18	7:48	7:18	
	0.643	10:18	17:36	18:06	1:11	YES
28	0.198	2:12	2:12	2:42	2:12	
	0.481	6:32	8:45	9:15	6:02	YES
	0.858	19:30	4:16	4:46	NA	

Table 14.Appendix.1: Simulated call times and sleep durations

Student Activity—Simulate Four days
- Generate 12 random numbers. This should be enough for four days.
- Convert each random number into a time for the next call.
- Use enough random numbers to create the first day's sequence.
- Determine the intervals for sleep including the interval at the end of the nine hours
- Determine if any interval was at least five hours.
- Use the remaining numbers to create three more days' sequence.
- Determine number of days that provided a five hour sleep interval.

Pool this information with the entire class to determine the percent of days with adequate sleep.

Section 14.6: Chapter 14 (Poisson and Exponential Distributions) Homework Questions

1. During tax season, the city's tax department receives on average three requests per day from homeowners for a review of their property tax assessment. The number of request is random and seems to follow the Poisson distribution. Find the probability that in a randomly selected office hour, the number of student arrivals is:
    a) 0
    b) 1
    c) 2
    d) 3
    e) 4 or more
    f) What is the probability that there will be two or more times the average number of requests on a random day?
    g) Provide an explanation for why you think the Poisson distribution would be appropriate?

2. Dr. Chelst teaches a large class with more than 100 students. He has found that on average five students arrive for every office hour he schedules. Because these students arrive independently of one another, he feels that random number of students who stop in follows the Poisson distribution. Find the probability that during an office hour, the number of student arrivals is:
    a) 0
    b) 1
    c) 2
    d) 3
    e) 4 or more
    f) What is the probability that there will be two or more times the average number of student visits in a random hour?
    g) What is the average time between arrivals?

3. Dr. Chelst generally schedules two hour blocks of office hours. Dr. Chelst expects to spend on average 10 minutes with each. He is therefore concerned as to the likelihood of too many students showing up and he will have to cut his time short to accommodate the extra students.
    a) What is the probability that six students will stop by in the first hour?
    b) What is the probability that six students will stop by in the first hour and an additional six students will stop by in the second hour?
    c) What is the probability that 12 students will stop by in the two hour block? How does this answer compare to the answer in part b? Explain the difference.
    d) What is the probability that more than six students will stop by in the first hour?
    e) What is the probability more than six students will stop by in the first hour and more than six students will stop by in the second hour?
    f) What is the probability that more than 12 will stop by in the two hour block? How does this answer compare to the answer in part e?
    g) Of the above questions, which probabilities do you think are of most interest to Dr. Chelst?

4. In a one-second period of time, a radioactive source emits a random number X of particles into a counter, where X has a Poisson distribution with mean 10.
    a) What is the probability that no particles are actually seen?
    b) What is the probability that fewer than 10 particles are actually seen?
    c) What is the probability that there will be twice the average or even more particles in a second?

5. Vehicles pass through a junction on a busy road based at an average rate of 320 per hour. The actual number in any hour is a random variable that can be described by the Poisson distribution
    a) Find the probability that none pass in a minute.
    b) Find the probability that none pass in a 5-minute period.
    c) What is the expected value of the number passing in two minutes?
    d) Find the probability that the closest integer number to this expected value actually pass through in a given two-minute period.
    e) In a two minute period, what is the probability that there will be two or more times the average number?
    f) In a five minute period, what is the probability that there will be two or more times the average number?
    g) In a ten minute period, what is the probability that there will be two or more times the average number?
    h) Summarize your observations about the likelihood that more than twice the average will pass during time intervals of various lengths.
    a) The higher the average, the smaller the probability of observing two or more times the average.

6. The coordinator of the local middle school soccer league regular gathers injury data. He found that there were a total of 72 injuries in last season's 60 games.
    a) Explain why we can use Poisson distribution with an average 1.2 injuries per game for this problem?
    b) Explain why a non-integer value is possible for the average?
    c) Find the probability that the team will have at most 1 injury in a game.
    d) Find the probability that the team will have one or more injuries in a game.
    e) Find the probability that the team will have exactly 2 injuries in this game.
    f) This month one team will play 6 games. What is the probability that in each and every game there will be at least one injury?

7. On average 6.5 traffic accidents occur on a particular stretch of road during a week. if the number of accidents follow a Poisson distribution:
    a) Find the probability that no accident will occur next week on this stretch of road.
    b) Find the probability that less than four accidents will occur next week.
    c) Find the probability that exactly five accidents will occur next week.
    d) What is the average time between accidents?

8. We continue with the previous problem in which on average 6.5 traffic accidents occur per week. We now focus on the number of accidents in a single day.
    a) What is the probability that there will be no accidents on a specific day?
    b) What is the probability of having four accidents or more on a specific day?
    c) On average how many days in a year will there be four or more accidents?
    d) What is the probability of having six or more accidents on a specific day?
    e) If six accidents occur, do you think this was simply a natural random fluctuation? What might explain what happened?

14.6.1 Homework Problems - Exponential Distribution (optional)

9. A deer has jumped into the road right after a car passed. It takes 15 seconds for it to cross the road. The number of cars passing through the road in an hour is a random number having a Poisson distribution with $\lambda = 35$. What is the probability of the deer crossing the road safely?

10. Dr. Chelst has found that on average five students arrive per hour during his office hours. The time between arrivals seems to be exponentially distributed.
    a) What is the probability that the time between two consecutive students is less than three minutes?
    b) What is the probability the time between two consecutive students is more than 20 minutes?

11. On average 6.5 traffic accidents occur on a particular stretch of road during a week. Assume number of accidents follow a Poisson distribution.
    a) What is the mean time between accidents?
    b) What is the probability that there will be an accident within the next 12 hours?
    c) What is the probability that it will be more than two days before the next accident occurs?

12. Mr. Sheriff is tracking the cars over the speed limit on I-40. He positions his vehicle right behind an overpass. He tries to ticket cars that exceed the speed limit by 10 miles per hour or more. The inter-arrival time between cars exceeding the speed limit by 10 miles per hour is exponentially distributed with a mean of 4 minutes. What is the probability of Mr. Sheriff ticketing a speeding car within the next 5 minutes?

13. Jack has just bought a used car. The previous owner told him that he renewed the battery a year ago. The lifespan of a battery of this brand is an exponentially distributed random number with a mean of 5 years. What is the probability of driving the car for the next 4 years without any battery problem?

14. Blind Man's Bluff is a kids' game, in which a person is chosen to be the blind man. The eyes of that person are blindfolded and he attempts to touch other players, who are scattered and avoiding him. If the blind man touches a person, that person assumes the role of the blind man. Six kids are playing the game, including the blind man. With this number of players, the time until the blind man touches a single person is exponentially distributed with a mean of 3 minutes.
    a) What is the probability that the blind man takes less than 1 minute to touch the first person?
    b) What is the probability that the blind man takes more than 5 minutes to touch the first person?

15. There is a water leak in Roberta's bathroom lavatory. She made a maintenance request and the apartment management assigned a plumber for this task. She wants to be at home while the maintenance is being done. The service time, including driving time, of the plumber is exponentially distributed with a mean of 30 minutes. It is 7:10 a.m. now, and the plumber has just called Roberta and said that he is leaving his office. Roberta has to leave home at 7:40 a.m. for work; otherwise, she will be late. (Her work starts at 8:00 a.m.)
    a) What is the probability that Roberta will wait at home more than 30 minutes?
    b) What is the probability that Roberta will leave home after 7:50 a.m.?
    c) It takes exactly 20 minutes drive from Roberta's home to her work. According to this, what is the probability that Roberta will be at least 1/2 hr late for her work?

16. A student sells Sunday Free Press newspaper on the weekends as a part time job at the intersection of Woodward Ave and MLK Ave. The number of vehicles passing by this intersection in an hour is a random number having a Poisson distribution with $\lambda=240$. Only 1% of the vehicles stop to buy a paper.
    a) What is average rate of vehicles stopping per hour to buy a newspaper?

b) Explain why the Poisson distribution would be appropriate to model this subset of drivers?

c) He still has 1 paper unsold and he has to leave in 15 minutes to catch a bus. What is the probability that he sells the last paper before he leaves?

17. The delay between the time people begin to feel symptoms of influenza and visit a doctor is exponentially distributed with a mean of 72 hours. Standard medical procedure is that patients are given an antiviral treatment if they seek medical help within 48 hours after seeing the symptoms. (Reference: *BMC Infectious Diseases* 2007, 7:76)

a) What is the proportion of the sick people who are given an antiviral intervention?

b) After an informative TV program regarding the dangers of the influenza pandemic, people have become more aware of influenza symptoms. Now they respond within 48 hours on average after seeing the onset symptoms of the disease. What proportion of sick people will be given an antiviral treatment?

c) What percent increase in treatment was caused by the TV programs?

18. Homeland Security has discovered a terrorist group that has stolen some secret files of the military. It has sent a search warrant to the Police Department in NYC where the terrorist member resides. The police team broke into the apartment where the terrorist hangs out and started searching. In the past, they have found that it takes on average a half hour for the team to find what they are looking for. It has been shown that the cumulative probability of locating a single target in a random search is exponentially distributed (Reference: *Theoretical Issues in Ergonomics Science* 2000, 1:4).

a) What is the probability that the team finds the files within 30 minutes?

b) A terrorist member is driving to the hangout and will be there within 40 minutes. What is the probability that the team will not have found the files before the terrorist arrives?

14.6.2 Expanded Homework Problems

I. Health Care Clinic—Reserving Slots for Random Urgent Patients

1. Dr. Johnson is the primary care physician at a clinic. The majority of Dr. Johnson's time is spent examining patients that have been scheduled in advance. Besides seeing patients who have scheduled appointments, Dr. Johnson must also see unscheduled patients who have an immediate urgent need to see him. Dr. Johnson wants to be able to reserve time for the urgent patients by limiting the number of scheduled patients in a day. The difficulty is deciding how many patients to schedule in advance so that he still has enough time to see any urgent patients that request time. The number of urgent patients each day is a random variable. If too many patients are scheduled in advance, then Dr. Johnson will not have enough time to see all of the urgent patients. As a result, the urgent patients will go instead to the emergency department. This is much more expensive for the patients and a poor use of emergency resources for the emergency room. However, if not enough patients are scheduled in advance, then if only a few urgent patients appear, Dr. Johnson ends up with extra time in his daily schedule. Regular patients seeking an appointment will have been delayed unnecessarily.

The data indicate that the number of urgent patients each day follows the Poisson distribution with a mean of 5 patients in a day. We explore Dr. Johnson's dilemma through a series of questions about the random pattern.

a. What is the probability of having 2 urgent patients or more in a day?
b. What is the probability of having no urgent patients in a day?
c. If the doctor keeps 4 slots available for urgent patients, what is the probability of not using all these 4 slots in a day? What is the probability that this will not be enough?

   d. For this policy of 4 reserved slots, calculate the probability of 0, 1, 2, 3 and 4 unused slots. Determine the expected value of the number of unused slots. Why is this number not simply the difference between the number of slots reserved and the average number of urgent patients?

   e. For this same policy, calculate the discrete probability of 0, 1, 2, … up to 10 urgent patients sent to the emergency room. Determine the expected value of the number urgent patients sent to the emergency room. Why is this number not simply the difference between the average number of urgent patients and number of slots reserved? Justify why we can stop at 10 or some other number when calculating the expected value.

   f. How many slots should be reserved for urgent patients to make sure that approximately three out of every four days all of the urgent patients will be accepted?

   g. With this policy, what is the probability that two or more or slots will be unused? What is the probability that two or more urgent patients will end up in the emergency room?

   h. For this policy, determine the expected value of the number of unused slots. Also for this policy, determine the expected value of the number urgent patients sent to the emergency room.

   i. Dr. Johnson recognizes that because of the uncertainty, there is no perfect strategy. She feels that as long as at the end of the day no more than 3 slots were unused, she can work during this spare time and catch up on a backlog of paperwork. She also does not want to frequently send more than 1 patient per day to the emergency room. Specify and analyze a policy we have not yet considered that you would recommend for reserving slots for urgent patients.

   j. Can you suggest other actions that Dr. Johnson can take to deal with the dilemma she faces?

II. EMS Dispatching Problem for Poisson Distribution

1. Six neighboring cities of Oakland County in Michigan use their own dispatching centers. Assume on average each city receives 2 emergency medical calls per hour:

   a. What is the probability a city receives 4 calls in a random hour?

   b. What is the probability a city receives less than 4 calls in a random hour?

   c. What is the probability a city receives 6 calls in a random hour?

   d. What is the probability a city receives less than 6 calls in a random hour?

   e. How many calls should the EMS units in one city be able to respond in an hour to make sure that at least 75 percent of the calls in that city will be responded by local units?

   f. How many calls should the EMS units in one city be able to respond in an hour to make sure that at least 95 percent of the calls in that city will be responded by local units?

   g. The EMS units in these cities are just first responders and the patients are transported to the hospital by a private ambulance system. If an EMS unit can handle on average two calls per hour, how many local EMS units are needed to meet the 75% and 95% standards?

   h. During Friday nights, the average increases to 4 calls per hour. How many calls should a city's EMS units be able to answer in an hour to make sure that at least 95 percent of the calls will be answered by local units?

2. (Optional) When the number of calls follows the Poisson distribution, the time between calls follows the exponential distribution.

   a. What is the probability of receiving the next call within 15 minutes?

   b. If an operator just finished answering a call, what is the probability that the operator would be able to take a 20-minute break before receiving the next call?

   c. Assume that answering a call takes 5 minutes of an operator's time. What is the probability of having another call come in while the dispatch operator is answering the first call?

3. Now, assume that those six main cities of Oakland County in Michigan use a shared EMS dispatching center. If on average there are on average 12 emergency calls per hour:

   a. What is the probability of receiving 12 calls in a random hour?

   b. What is the probability of receiving less than 12 calls in a random hour?

c.   What is the probability of receiving 24 calls in a random hour?
d.   How many calls should the EMS units be able to respond in an hour to make sure that at least 75 percent of the calls will be responded?
e.   How many calls should the center be able to respond in an hour to make sure that at least 95 percent of the calls will be responded?
f.   If the centers share the expenses of the dispatching unit, which policy is more economical for 95 percent case? Sharing or having 6 different centers?
g.   If during Friday nights the average increases to 15 calls, how many calls should the center be able to respond in an hour to make sure that at least 95 percent of the calls will be responded?

III. Tornados
<Wikipedia > "A tornado is a violent, dangerous, rotating column of air which is in contact with both the surface of the earth and a cumulonimbus cloud or, in rare cases, the base of a cumulus cloud. Scientists use some techniques to rate tornadoes based on their intensity. The Fujita scale rates tornadoes by damage caused. An EF0 tornado will probably damage trees but not substantial structures; whereas an EF5 tornado can rip buildings off their foundations leaving them bare and even deform large skyscrapers."
In the United States, 80% of tornadoes are EF0 and EF1 tornadoes. The rate of occurrence drops off quickly with increasing strength—less than 1% are violent tornadoes (EF4 or stronger).
Find more information about tornadoes in US and state of Oklahoma in the following website:
http://www.srh.noaa.gov/oun/tornadodata/ok/monthlyannual.php

1.   The state of Oklahoma prepares for tornadoes each year and allocates enough resources to repair and fix the damage of intense tornadoes. During the peak four months from April through July, the number of intense tornados in Oklahoma each year is a random variable that follows Poisson distribution with mean 7. On average, each intense tornado costs one million dollars for the state. At the beginning of each year the governor allocates the financial resources for handling the cost of intense tornados. If the number of intense tornados will be more than their prediction, they have to ask the Federal government for help.
a.   What is the probability of having 8 or more tornados?
b.   What is the probability of having 9 or more tornados?
c.   How much money should be allocated to make sure that they won't need to ask for money from the federal government with probability of 85 percent? How about 95 percent?
d.   Arkansas has a similar problem. Intense tornados during Tornado season in Arkansas follow a Poisson distribution with mean 6. How much money should be allocated to make sure that they won't ask for money from the Federal government with probability 85 percent? How about 95 Percent?
e.   Now suppose that Oklahoma and Arkansas signed a collaboration agreement to have a shared budget and resources for handling the tornados. Each state has to provide 50 percent of the money. How much money in total should be allocated to make sure that they won't ask for money from the Federal government with probability 85 percent? How about 95 Percent?
f.   Compare the allocated budget of Oklahoma with collaboration and without collaboration. Are they better off collaborating?
g.   Compare the allocated budget of Arkansas with collaboration and without collaboration. Are they better off collaborating?

2.   (Optional) The states are interested in better understanding the time between tornados.
d.   It is June 1st. What is the probability that the next tornado will hit in Oklahoma on or before June 10th?
e.   Write a general expression to determine the probability that Oklahoma will not experience a tornado within X days?
f.   It is June 1st. What is the probability that the next tornado will hit in Arkansas on or before June 10th?

g. Write a general expression to determine the probability that Arkansas will not experience a tornado within X days?

h. What is the probability that neither Oklahoma nor Arkansas experience a tornado on or before June 10th?

i. Write a general expression to determine the probability neither Oklahoma nor Arkansas will experience a tornado within X days? How does this expression compare to what was found in parts b and c?

j. Each state has one major team that helps during the first week after a major tornado strikes. What is the probability that while the Oklahoma team is working at a tornado site, another Tornado will strike? What is the probability that while the Arkansas team is working at a tornado site, another Tornado will strike?

Chapter 14 Summary

What have we learned?

In this chapter we have studied two related distributions, the Poisson Distribution and Exponential Distribution. The distributions have only one parameters, λ, the average rate of occurrence of a random event. Just as in the distributions we studied in other chapters, the occurrence of one random event does not influence the occurrence of another. For example, the occurrence of one medical emergency is independent of the next medical emergency.

The Poisson Distribution is used to study the number of times a random event occurs in a time period. The probability distribution function for a particular number, x, of events is $P(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$. We can also use the TI calculator command poissonpdf(λ,x). There is no formula to find the cumulative distribution function, which adds the probabilities from 0 to x occurrences. However, we can use the TI calculator the command poissoncdf(λ,x). Poisson Distributions can also be added and the sum will still follow a Poisson Distribution. Thus if λ is the average rate per day of a random event, 5λ will be the average for a 5 day workweek.

The Exponential Distribution is used to study the time between occurrences of a random event. If the number of random events in a time period is a random variable that behaves like the Poisson Distribution, then the time between calls behaves like the Exponential Distribution. The formula to find the probability that there is more than a particular time, t, between events is $P(T > t) = e^{-\lambda t}$. This distribution does not have a TI calculator command. The cumulative distribution function, that there is less than a particular time between events (from 0 to t) is $P(T < t)$ $1 - e^{-\lambda t}$. The probability that the time between events will be between two times, t_1 and t_2, is $P(t_1 < T < t_2) = e^{-\lambda t_1} - e^{-\lambda t_2}$.

Since λ is the average rate of occurrence per unit time, the expected value for the Poisson Distribution is λ for one unit in time and λt for an entire time period. As with the other distributions we have studied, the expected value for the Poisson Distribution may not actually be possible. The expected value for the exponential distribution is $1/\lambda$. Since the exponential distribution is an amount of time, the expected value for the Exponential Distribution is always possible. Interestingly, although $1/\lambda$ is the average time between events, the probability that the time between events is less than the average is 0.632. The ratio of below average intervals to above average is 63:37 or 1.7:1.

Terms

Parameter	Values that need to be specified in order to apply a distribution
Random Event	The occurrence of an event which happens at random times
Poisson Distribution	The probability distribution that is used to find the likelihood of a particular number of occurrences of a random event
Exponential Distribution	The probability distribution that is used to find the likelihood of the time between occurrences of a random event
λ **(lambda)**	The average rate of occurrence per unit time
Expected Value	The long run average value over many repetitions
Poisson	$E(X) = \lambda$ for one unit time and $E(X) = \lambda t$ for time period t
Exponential	$E(X) = 1/\lambda$

Chapter 14 (Decision Trees) Objectives

You should be able to:

- Determine whether or not the Poisson and Exponential distributions would be appropriate for a situation

- Determine whether a solution to a question requires use of the Poisson Distribution or the exponential distribution.

- Calculate Poisson probability distribution functions

- Calculate Poisson cumulative distribution functions

- Calculate Exponential probability distribution functions for more than a particular time

- Calculate Exponential cumulative distribution functions for less than a particular time

- Calculate Exponential probabilities for between two particular times.

Chapter 14 Study Guide

1. Compare and contrast the Poisson and Exponential Distributions.

    a. Give a specific example when you would use the Poisson Distribution.

    b. Give a specific example when you would use the Exponential Distribution.

2. What are the parameters needed for the Poisson and Exponential Distribution?

    a. Why do the distributions use the same parameter?

    b. If λ is 2 times per hour, what would λ be per day?

3. If λ is 4 times per day what would be the average time between occurrences?

    a. If you recorded when events occurred, would there be more intervals whose length was above this number, more intervals below this number, or the same number above and below?

    b. If λ increases, what happens to the average interval between occurrences of the random event?

4. Explain how to find the probability of 3 to 5 occurrences in an hour if λ is 4 times per hour?

    a. What is the probability?

5. Explain how to find the probability of between 15 and 45 minutes without an occurrence if λ is once per hour?

    a. What is the probability?

References

Finnerty, John D. (1977), How often will the firemen get their sleep? *Management Science* 23:1169-1173.

Kros, John, Dellana, Scott and West, David (2009), Overbooking Increases Patient Access at East
        Carolina University's Student Health Services Clinic, *Interfaces* 39:3, 271-287.

Section 15.0: The Normal Distribution

The Normal distribution is the most widely recognized of all probability distributions. It is a continuous distribution, which means its graph has no gaps. The shape of its graph is classically described as Bell-shaped. It is symmetric around the mean with long tails. It was originally discovered in the early part of the 18th century by the French mathematician, De Moivre and independently rediscovered in the 19th century by the German mathematician, Gauss. The distribution is described by a complex formula that cannot be used analytically to answer probability questions. Instead, all probability questions are answered by inserting values into a calculator or spreadsheet.

The normal distribution is the most important probability distribution that is used in statistics. This is due to a number of desirable mathematical properties that normal distributions have. Statisticians have been able to solve many problems in statistics only by assuming a normal distribution and using its mathematical properties. This is because there are many variables, such as intelligence or height, that when plotted appear similar to it. There might be an infinite number of variables of interest that have this characteristic.

However, the normal distribution is only a model. There are, in fact, no variables that completely match it. Furthermore, there are many other models, such as exponential growth (e.g., bacteria) or exponential decay (e.g., radioactivity), each also having possibly an infinite number of important variables that when plotted appear similar to them as well. The existence of other models has a major impact on the consequences of assuming the normal distribution if it is not reasonable to do so.

The normal distribution arises naturally when the random variable is the sum or mean of several of other independent random variables. The distribution has two independent parameters: the mean and the standard deviation. Sometimes the variance is used instead of the standard deviation. An important concept is that any normal distribution with mean of μ and standard deviation of σ can be transformed into an equivalent normal distribution with a mean of 0 and a standard deviation of 1. This new distribution is called the **standard normal** distribution. It can be used to quickly estimate with good accuracy probability questions about any normal distribution, sample mean, or sample proportion. With the advent of technology, calculators or computers with appropriate software are used to provide answers to probability questions based on normal distributions. For example, a graphing calculator determines the probability by finding the corresponding area under a normal distribution function. In this chapter, we use the graphing calculator as the primary tool to calculate probabilities.

Section 15.1: Cutting Fabric for Parachutes

Skydiving equipment has advanced considerably over the last several years. Round parachutes are seldom seen these days and have been replaced by modern, rectangular "ram-air" canopies that have better directional control and offer softer landings. Ram-air canopies are made of a series of inflatable tubes or "cells," connected side-by-side along their length. Each cell is designed to form the cross section of an airfoil, so when the parachute inflates, it forms a wing-shaped canopy, ready for flight. The front of each cell is open to the air, and the back is sewn closed. Once inflated, the ram-air canopy is a semi-rigid, rectangular plane, similar to an airplane wing. Figure 15.1.1 shows a ram-air canopy parachute with seven cells. (from http://www.uspa.org/AboutSkydiving/Equipment/tabid/128/Default.aspx)

Figure 15.1.1: A ram-air parachute (from http://ozcrw.tripod.com/crw_canopies.htm)

The fabric most often used to make parachutes is called rip stop nylon. It is commercially available in 20-meter rolls that are 150 centimeters wide. Depending on the final dimensions of the parachute, the size of each cell will vary as will the length of fabric that must be cut. Regardless of the size, the fabric must be cut with great precision. If the pieces are cut too small, the material cannot be used. If they are cut too large, the material can be used, but you will have wasted some of your fabric and may need to re-cut the fabric down to size. The margin of error for cutting the fabric is 5 millimeters.

Q1.   What is the relationship between meters and centimeters? between centimeters and millimeters?

15.1.1 Sampling the Cut Pieces

Production rate is a measure of productivity. The production rate represents the output after a given amount of time. Sky's the Limit is a small company that makes parachutes. Twin brothers Tom and Ken Speedy work in the cutting department at Sky's the Limit. They cut the large panels of fabric used in making the parachutes. In this case, their production rates would be the number of pieces each cuts per day. Tom and Ken Speedy are cutting pieces that are supposed to be 50 cm wide and 2 m long. On a particular day, the company decides to take random samples of 10 pieces of Tom's and Ken's work, to see if they are within specifications: 2 m long ± 5 mm and 50 cm wide ± 5 mm. The length and width of each piece of work that was sampled appear in Table 15.1.1. Ray Monitor, the production manager at Sky's the Limit, decides he needs to compute some statistics to help him make sense of the data. He decides to compare the *means* of the lengths of the two samples.

WORKER	Tom		Ken	
SAMPLE	Length (m)	Width (cm)	Length (m)	Width (cm)
1	2.004	50.3	1.999	49.9
2	1.999	50.3	2.001	50.1
3	2.002	50.2	1.999	50.3
4	1.997	50.3	2.000	49.9
5	2.003	49.7	1.999	50.0
6	2.000	49.8	2.002	49.7
7	2.000	49.7	1.999	50.1
8	1.998	50.2	1.999	49.9
9	1.996	49.8	2.002	50.1
10	2.001	49.7	2.000	50.0

Table 15.1.1: Lengths and widths of two samples measured to the nearest mm.

> **The *mean* of a sample is just the arithmetic average:**
> **Add each of the sampled items and divide by the number of items.**

> **If the elements of a data set are represented by $\underline{x}_i$, the**
> **usual notation for the mean of the data set is $\overline{x}$**

Q2.   Compute the mean of the lengths of each worker's sample. What do you observe?

The mean is one way to describe the center of a set of data. Ray was surprised that the two means were equal, because when he looked at the two sets of data, they somehow looked different to him.

Q3.   Do the two sets of lengths look different to you? If so, how are they different? If not, other than having the same mean, what similarities did you see?

Then Ray had an "Ah-ha" moment! He noticed that Tom's lengths varied between 1.997 and 2.003; while Ken's varied between 1.999 and 2.002. Ray can use a statistic called the *range* to express what he noticed mathematically. He just subtracted the smallest number from the largest number in each data set. The range is one way to measure how spread-out the individual data points are.

> **The *range* of a set of data is defined to be the difference between the**
> **largest and smallest numbers in the set of data.**

Q4.   Compute the range of the lengths for each worker's sample. What do you observe?

Now Ray wonders if he will observe the same things about the width data.

Q5.   Compute the mean and the range for each worker's sample of widths. What do you observe?

Ray finds the width data difficult to make sense of. Although the mean *and* range for each worker's width data are the same, he is still sure that he can see a difference.

Q6.   Do you see any difference in the two width data sets? If yes, how are they different? If not, besides having the same mean and range, what similarities do they have?

Ray decides that he needs another statistic to measure the spread of the data. He decides to use a statistic called the **standard deviation**. The standard deviation uses the difference, or deviation, of each data point from the mean to measure the spread of the data. So first, for each data point, x_i, the difference $\overline{x} - x_i$ is computed. Next, finding an average difference would make sense. However, some of the differences are going to be positive, and some negative, so if they were just added together, the positives and negatives would cancel each other. To avoid that, each of the differences is squared. Then the squared differences are added. If there are n elements in the data set, this can be written as:

$$\left(\overline{x} - x_1\right)^2 + \left(\overline{x} - x_2\right)^2 + \ldots + \left(\overline{x} - x_n\right)^2 = \sum_{i=1}^{n}\left(\overline{x} - x_i\right)^2$$

Q7.   Why does adding the squared differences avoid the problem of positives and negatives canceling each other?

To complete the process of finding a standard deviation, first we divide by $n-1$. Then we take the square root of the result, so that our units of measurement will be the same. The symbol for the standard deviation is S_x, and we can write:

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}\left(\overline{x} - x_i\right)^2}{n-1}}$$

Why divide by n–1 rather than n?

If we knew the means of the length and width for the entire population of nylon panels, we could just divide by n. However, we do not know those means, so we have to estimate them. The most convenient estimate to use is $\overline{x}$, the mean of the sample.

Now think about a sample of size 1. If there is only one data point, that data point is the mean of the sample. The total of the differences from all of the data points to the mean must be 0 because there is only one data point; it was used to calculate the mean. Thus the estimate of the standard deviation would be $0 \div 1 = 0$. That estimate says there is no variability in the population, which is extremely unlikely. No variability means that the entire population of nylon panels would have the same length and width.

In order to get around this difficulty, rather than dividing by n, we divide by $n – 1$. Now if our sample size is 1, our standard deviation would be $0 \div 0$, which is indeterminant. Its value could be any real number. This fits the situation where the sample size is 1. A sample of only 1 does not have enough information to estimate the standard deviation.

Q8.   Using this process, find the standard deviation for each of the worker's width data. What do you observe about the two standard deviations? Why do you think that happened?

Sometimes it is useful to work with a statistic called the **variance**. The variance is simply the square of the standard deviation. Therefore, it is denoted by S_x^2. Notice that the square root of the variance is the standard deviation, which is as it should be:

$$\boxed{\sqrt{S_x^2} = S_x}$$

Q9.   Calculate the variances of Tom's and Ken's samples.

Notice also that the variance is always computed as part of the computation of the standard deviation. Why, then, do we have two different statistics? The first reason is that the units of measure for the variance might not be meaningful, and even when they are meaningful, they are not appropriate to the problem context. For example, in our width problem above, the units in which the widths were measured are centimeters. However, the variance squares differences that are measured in centimeters; the units of the variance are *square* centimeters. From your study of geometry, you may recall that centimeters are used to measure distances, as in the width example, but square centimeters are used to measure area. It is not appropriate to use a unit of area to measure variation in a distance.

15.1.2 Putting Two Pieces Together

The second reason that we use two different statistics, the variance and the standard deviation, is that the standard deviation, which is the more meaningful statistic, lacks an important property that the less meaningful statistic, the variance, has. To see this, consider a third worker who takes a panel cut by Tom and another cut by Ken an stitches them together. Before he does so, he overlaps the panels by 1 cm. On average, how wide would the resulting panel be? How much variability should we expect?

The mean width of the samples taken from both Tom's and Ken's work is 50 cm. On average the width of two panels sewn together with a 1 cm overlap will be 99 cm. Now, how much variability will there be, on average?

The standard deviation and the variance are both measures of variability. Perhaps it seems that we could just add the two standard deviations or the two variances together, just like we did for the mean. Let's see what happens when we do.

SAMPLE	Tom's Width (cm)	Ken's Width (cm)	Width of doubled panel
1	50.3	49.9	50.3 + 49.9 – 1 = 99.2
2	50.3	50.1	99.4
3	50.2	50.3	99.5
4	50.3	49.9	99.2
5	49.7	50.0	98.7
6	49.8	49.7	98.5
7	49.7	50.1	98.8
8	50.2	49.9	99.1
9	49.8	50.1	98.9
10	49.7	50.0	98.7

Table 15.1.2: Widths of double-panels made from two workers samples.

Table 15.1.2 shows the widths of a double panel when one of the panels from Tom's sample and one of the panels from Ken's sample are overlapped and sewn together. You can use a graphing calculator to compute the mean, standard deviation, and variance of the widths of the doubled panels. To begin, in the mode screen, choose to display two decimal places. Then use STAT-EDIT to access the data lists and enter the "Width of doubled panel" data into List 1, as shown in Figure 15.1.2. Next, Use STAT-CALC to access the 1-Var Stats feature to calculate the one- variable statistics, as shown in Figure 15.1.3. The calculator displays only the mean, $\bar{x}$, and the standard deviation, S_x. From the screen capture, we can see

that $\bar{x} = 99$ and $S_x = 0.33$. We must use S_x to calculate the variance, S_x^2. The screens in Figure 15.1.4

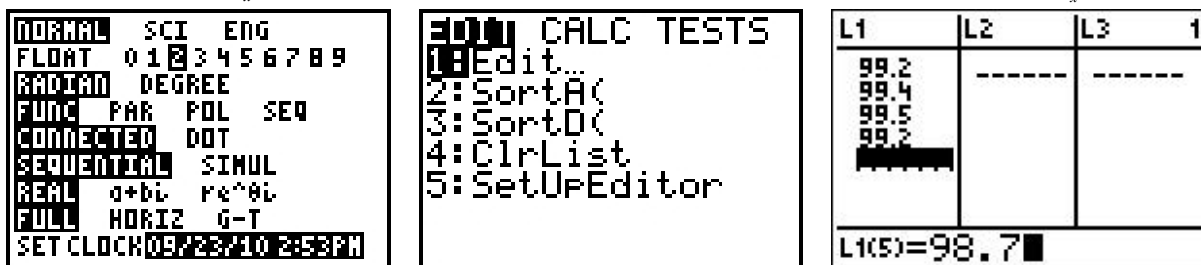show how to access S_x from the VARS-Statistics menu, and in the last screen, we see that $S_x^2 = 0.11$.

Figure 15.1.2: Entering the data into a calculator
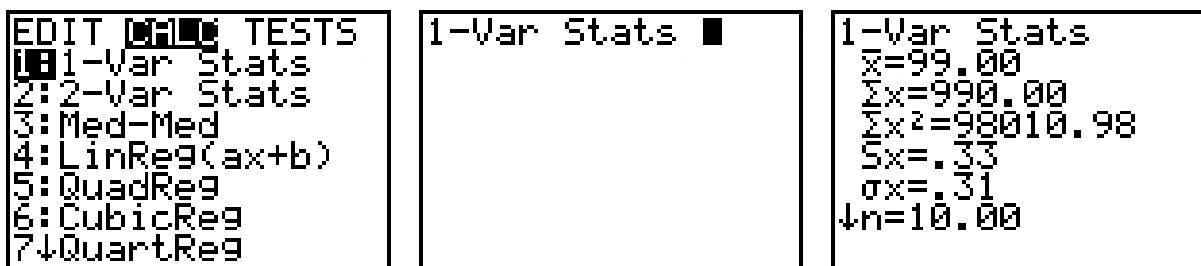

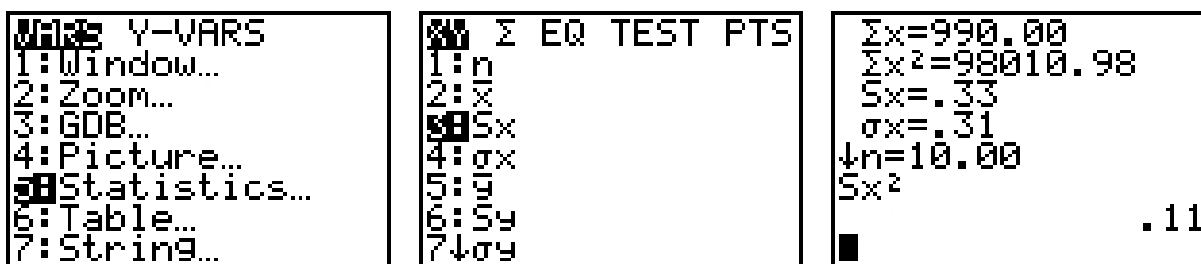**Figure 15.1.3**: Using the calculator to compute statistics

Figure 15.1.4: Using the standard deviation to find the variance

Q10.   Add the means of Ken's and Tom's samples that you calculated earlier and subtract the 1 cm overlap. Compare your answer to the mean of the doubled panels. Observation?

Q11.   Add the variances of Ken's and Tom's samples that you calculated earlier. Compare your answer to the variance of the doubled panels. Observation?

Q12.   Add the standard deviations of Ken's and Tom's samples that you calculated earlier and compare your answer to the standard deviation of the doubled panels. Observation?

You should have found that when the elements from two data sets are added, the means and variances can also be added to find the mean and variance of the new data set. However, the standard deviations cannot be added.

Why can't the standard deviations be added? You might recall from algebra that the square root of a sum is not the same as the sum of the square roots. For example, the following two equations are not the same.

$$\sqrt{4+9} = \sqrt{13} \approx 3.61$$
$$\sqrt{4} + \sqrt{9} = 2 + 3 = 5$$

So, when two variances are added, you get the variance of the data set formed by adding the two sets of data. However, the standard deviation is the square root of the variance. The square root of a sum of two variances does not equal the sum of the square root of each variance.

15.1.3 Putting the Whole Canopy Together

Recall that the mean and range of Tom's and Ken's samples of parachute panels were the same, but that Ken's sample had the smaller standard deviation. From past experience, Ray Monitor, their supervisor, knows that when something is cut to a specification, the cuts are approximately normally distributed about the mean.

To see what that means, let's graph a normal distribution based on Tom's width data. We cannot know the mean, μ, nor the standard deviation, σ, of the widths of all of the panels Tom has ever cut, but we do have the mean, $\overline{x}$, and the standard deviation, S_x, of the sample that he cut for Ray Monitor. We can use those statistics, $\overline{x}$ and S_x, as estimates of the actual parameters, μ and σ, of Tom's widths. Figure 15.1.5 shows the process of graphing a normal distribution of widths that have a mean of 50 cm and a standard deviation of 0.28 cm. Notice the bell shape. The vertical line in the fourth graph goes through $x = 50$, which is the mean of the distribution.
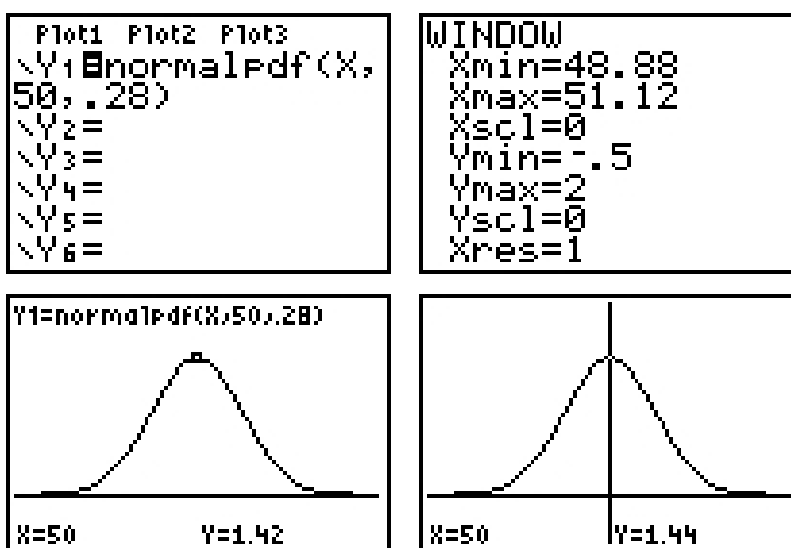
Figure 15.1.5: Graph of a normal distribution with $\mu = 50$ and $\sigma = 0.28$

Q13.   What proportion of the area under the bell-shaped curve and above the *x*-axis appears to be to the left of the vertical line? What proportion appears to be to the right?

Q14.   What do you think is the probability that a value of *x* is greater than 50? … less than 50?

Q15.   Use the mean and standard deviation of Ken's width data to generate a similar graph. What proportion of the area of your graph appears to be on either side of a vertical line through the mean?

Q16.   What do you think is the probability that a value of *x* on your graph is greater than 50 or less than 50?

In order to use a normal distribution, Ray Monitor needs to know the mean and standard deviation for the widths of all of the panels that *each* of his workers cuts, not just Ken and Tom. That information would not be available unless every panel had been measured after it was cut. That would have been very costly. Instead, Ray decides to use the mean and standard deviations from the samples that were measured. He reasons that these statistics are probably good estimates of the actual mean and standard deviation. Recall that the mean of each of their samples was 50 cm, but the standard deviation of Tom's sample was 0.28 cm, while the standard deviation of Ken's was 0.16.

When data are normally distributed with mean, μ, and standard deviation, σ, the distribution features of a calculator can be used to find the probability that a particular value falls within a given range. For example, the range of interest in this problem is widths between 49.5 cm and 50.5 cm, because widths in this range are within specifications. So, a key question that Ray asks himself is, "What is the probability that the panels Tom cuts meet the width specifications?" In other words, what is the probability that Tom cuts a panel with a width between 49.5 cm and 50.5 cm?

We can answer this question using the normal cumulative density function (cdf), as shown in Figure 15.1.5. First we use the mode key to set the calculator to display 3 decimal places. The key parameters are the mean, 50, standard deviation, 0.28, and the endpoints of the interval, 49.5 and 50.5 The syntax is normalpdf(49.5,50.5,50,.28). From the figure, we can see that the probability Tom cuts a panel with its width within specifications is 0.926, or 92.6%.
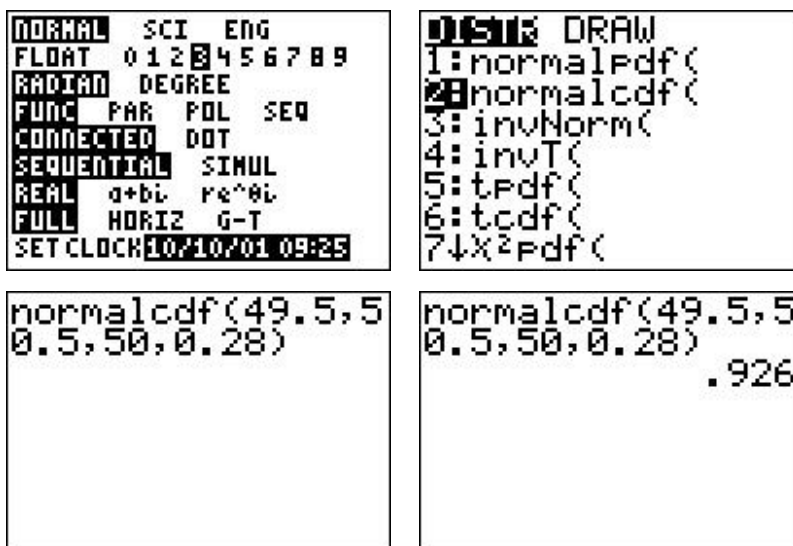
Figure 15.1.5b Setting up a normal cdf on a graphing calculator

Table 15.1.2 presents a sample of 10 measurements from Tom's work.

Q17.   What percentage of this sample meets the specifications? Is this consistent with what the normal distribution predicts?

Q18.   Using the mean and standard deviation of Ken's sample, find the probability that Ken cuts a panel with its width within specifications.

Ray Monitor knows that Ken and Tom are two of his best workers. He decides to collect a random sample of twenty-five cut panels from each of his 50 workers. When he computes the statistics from the large sample, he finds that the mean is 50.05 cm and the standard deviation is 0.3 cm

Q19.   Assuming a normal distribution and using the mean and standard deviation of the large sample,
        find the probability that the width of a panel cut by any one of Ray's workers meets
        specifications.

Q20.   In a production run of 7,000 panels, on average, how many will meet the width specifications?
        How many will not?

Each panel cut to the correct width specifications is worth $6.00. Each panel that does not meet the width
specifications can be used for other products, but is then worth only $1.50.

Q21.   Find the average value of a production run of 7,000 panels.

Seven panels are stitched together to form a parachute canopy. Remember that the panels are overlapped
1 cm for each seam and that there will be 6 seams in a rectangular parachute as shown in Figure 15.1.1.
The seven panels are chosen randomly from all of the panels cut by all of the workers. Now, suppose that
the specifications for the entire parachute canopy indicate that the width should be 344 cm ± 2 cm. Ray
Monitor would now like to know the probability that a completed parachute canopy meets this width
specification.

First, he needs to find the mean and the standard deviation of the width of an entire parachute canopy
after it has been sewn together. Recall that the mean of a sum is the same as the sum of the means. He
reasons that, on average, the width of a finished canopy will be 7×50.05 cm less the total of 6 cm of
overlap. So, the average width is going to be 344.35 cm.

Next, he needs to calculate the standard deviation for the width of the seven panels that are sewn together
to form the canopy. However, he remembers that he cannot just add the standard deviations, because the
square root of a sum is not the same as the sum of the square roots. He must square the standard deviation
to find the variance, add the variances together, and then take the square root. The standard deviation of
the large sample was 0.3 cm and the variance is 0.09 cm^2. The total variance of the seven pieces is seven
times the individual variance.

$$7 \cdot 0.09 = 0.06 \text{ and } \sqrt{0.63} \approx 0.794$$

So, the standard deviation for the width of 7 panels sewn together is 0.794. Now Ray is ready the use the
normalcdf(function to compute the probability that a finished parachute canopy will meet the width
specifications. Figure 15.1.6 displays Ray's result. Almost 98% of the parachutes will meet the width
specifications.

```
normalcdf(344-2.
0,344+2.0,344.35
,.794)
        .9796095557
```

Figure 15.1.6: Computation of the probability of meeting the width specification

Q22.   Why were 344 – 2 and 344 + 2 used in setting up the normalcdf?

15.1.4 Waste vs. Production—An Economic Analysis

The two brothers, Tom and Ken Speedy are both up for raises. Unfortunately, their supervisor, Ray Monitor, knows that the company can only afford a small raise for one of them. Ray's dilemma is deciding which brother is more deserving of the raise.

Ray knows that on an average day, Tom produces 180 cut panels, while Ken produces 171. Ray also remembers that although the mean widths of their production samples were equal, Ken's sample had a smaller standard. Ray decides to see if this matters. Recall from the previous section that 92.6% of Tom's cut panels met specifications, while 99.8% of Ken's did.

Ray decides to calculate the value per day of each worker to the company. Tom produces more cut panels per day, but only 92.6% of them are usable for parachute canopies. The other 7.4% are unusable for parachute canopies and must be used for other products. Ray also knows that the value to the company of each usable panel is $6.00, while the value to the company of an unusable panel is $1.50. Ray computes Tom's per day value to the company this way:
$$(0.926)(180)(\$6.00) + (0.074)(180)(\$1.50) = \$1,020.06.$$

Q23.   Explain each part of the formula above that Ray used to compute Tom's per day value to the company.

Q24.   Using a similar formula, compute Ken's per day value to the company.

Q25.   Based on the per day value to the company of these two workers, which one should get the raise?

Section 15.2: Automobile Battery Warranties

Alex Murphy is a senior at Chesterfield High School in Chesterfield Township, Michigan. In a case of what he likes to call Murphy's Law, the battery in his recently purchased used car had just died and he needed to replace it as soon as possible. He intended to keep this car through his senior year of high school as well as through his four years of college, so a reliable car battery with a high-quality warranty was essential. Alex decided to purchase a car battery that costs \$110 and includes a 36-month full replacement warranty with an additional 6-month pro-rated warranty. Let T represent the time, in months, from the purchase of the car battery until it fails. Table 15.2.1 shows the replacement values for the warranty period.

Time, in months, from the purchase of the car battery that it fails	Replacement Value (% of original purchase price)
$T \leq 36$	100
$36 < T < 39$	75
$39 < T < 42$	50
$T \geq 42$	0

Table 15.2.1: Replacement values for battery failure within warranty period

15.2.1 Using a Normal Distribution to Find the Probability of Battery Failure

After doing some additional research, Alex found an old article in an automotive magazine with some data about battery life for the same battery that he was considering purchasing. The magazine collected data on a sample of 50 such batteries. Alex put the data into his graphing calculator and used it to compute some statistics. He also used his calculator to draw a histogram of the data. The set-up of Alex's calculator appears in Figure 15.2.1, and his results are displayed in Figure 15.2.2.

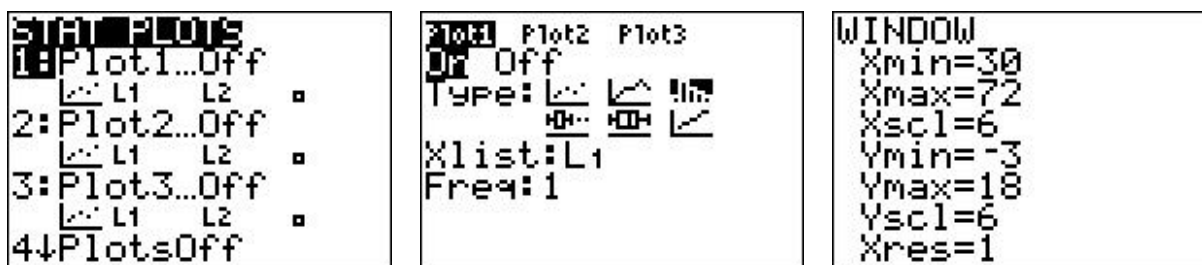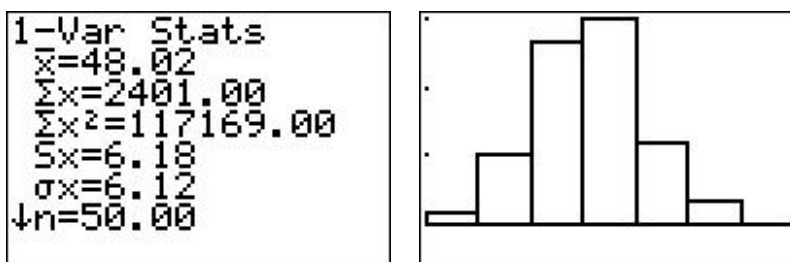**Figure 15.2.1**: Alex's calculator set-up

Figure 15.2.2: Alex's battery statistics and histogram

Alex thought he remembered seeing histograms with the same shape as his battery histogram when he learned about normal distributions in his math class. Looking at the histogram and the statistics, Alex noticed that the mean battery life was a little more than 48 months and the standard deviation was close to

6 months. He decided to assume that the battery life is normally distributed with a mean of 48 months and a standard deviation of 6 months. Alex decided to look at the graph of this normal distribution and consider whether it seemed to be a good representation of the data in his histogram. He did this on his graphing calculator by (a) opening the Y= menu, (b) pressing 2nd VARS to view the distributions menu and selecting normalpdf(, and finally (c) specifying the variable, the mean, and the standard deviation. Figure 15.2.3 shows the calculator screens for these steps.
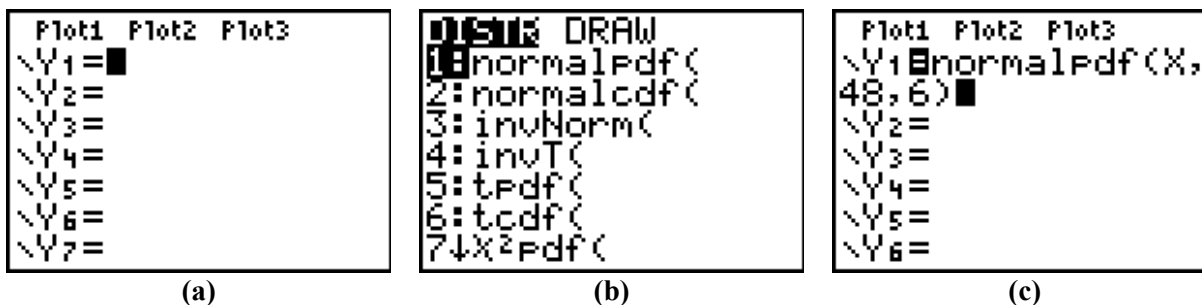

(a)   **(b)**   **(c)**

Figure 15.2.3: Steps to graphing a normal distribution on a graphing calculator

The most challenging step in graphing the normal distribution is setting up an appropriate viewing window. In order to set the Xmin and Xmax, Alex reasoned that the mean should be in the middle, half-way between the Xmin and Xmax,. He also thought that it made sense to scale the *x*-axis using the standard deviation; which is 6. His real challenge came in setting-up the *y*-axis. At first, he couldn't see anything, but then he remembered learning that the area between the graph of a normal distribution and the *x*-axis is always 1. Since his *x*-values were fairly large, he decided to use fairly small values for *y*. The viewing window Alex used and his graph of this normal distribution are given in Figure 15.2.4.
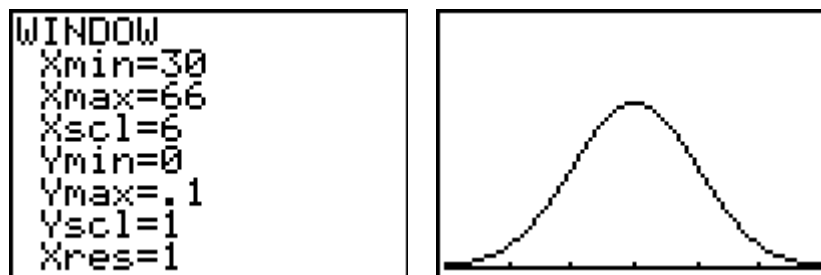
Figure 15.2.4: Alex's viewing window and graph of the normalpdf with $\mu = 48$ and $\sigma = 6$

Q1.   What are some features that you notice about the graph of this normal distribution?

Alex now wanted to find the probability that the battery he has chosen will fail within the ranges given in the terms of the battery warranty. At first he was not concerned about the probability of the battery failing within the first 36 months. In that case he would get 100% of his money back. Instead, he considered the first range in the pro-rated warranty period. If the car battery fails between 36 and 39 months of his purchase, Alex would be refunded 75% of the original purchase price.

Q2.   How much money would Alex get back if the battery fails between 36 and 39 months?

One way to find the probability is to use the normalcdf(feature on the graphing calculator. Recall that for this particular car battery, Alex assumed that $\mu = 48$ and $\sigma = 6$. He wanted to know the probability that the battery would fail between 36 and 39 months of purchase. Figure 15.2.5 shows the calculation. The output tells us the probability that the battery would fail between 36 and 39 months of the purchase is approximately 4.4%.

```
normalcdf(36,39,
48,6)
        .0440571667
```

Figure 15.2.5: Calculation of the Normal Cumulative Density Function

Q3.   On average, out of every 100 of these batteries that are made, how many fail between 36 and 39
        months after purchase? How many out of every 1,000?

Alex likes to visualize things whenever possible. He remembered a second way to use his calculator to
find this probability. This time, he used the graph he created earlier and option 7 under the CALC menu,
which looks like this: $\int f(x)\,dx$. To use this feature of the calculator, while looking at the graph, Alex

        (a) pressed 2nd TRACE to call up the CALC (calculate) menu and pressed number 7,
        (b) entered the lower limit,36, for the range of months he was interested in, then
        (c) entered the upper limit, 39. Figure 15.2.6 shows Alex's calculator screens for these steps.

```
CALCULATE
1:value
2:zero
3:minimum
4:maximum
5:intersect
6:dy/dx
7:∫f(x)dx
```

```
Y1=normalpdf(X,48,6)




Lower Limit?
X=36█
```

```
Y1=normalpdf(X,48,6)




Upper Limit?
X=39
```

      (a)   **(b)**   **(c)**
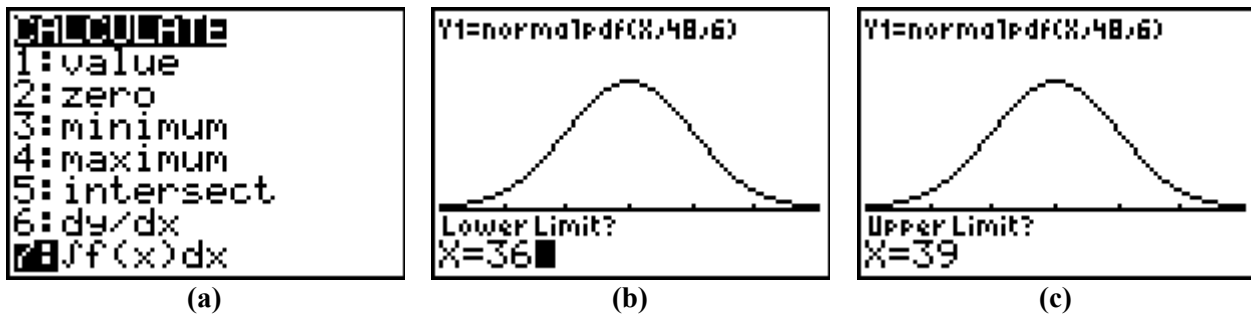
Figure 15.2.6: Alex's calculator steps to find a probability graphically

Once Alex hit enter, his calculator did something dynamic and then showed the screen depicted in Figure
15.2.7 Notice that both of Alex's methods determined that the probability of the battery failing between
36 and 39 months is about 4.4%. Given this graphical representation, Alex also interpreted that value as
the area of the shaded region in Figure 15.2.7.

```



∫f(x)dx=.04405707
```
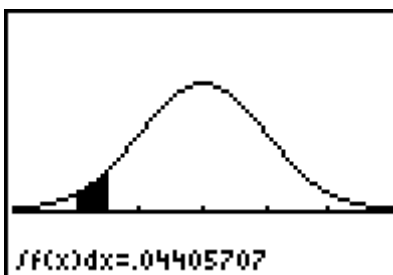
Figure 15.2.7: The probability calculated graphically

Next, Alex decided to calculate the probability that the battery fails beyond the warranty period. Since he
wouldn't get any money back if this happens, he was very interested in the likelihood that the battery will

fail after 42 months. He wanted to use the normalcdf command, but he needed to know both the lower and *upper* limits of that range. The lower limit is clearly 42 months.

Q4.   What is the upper limit?

Alex realized that, in theory, the upper limit for the range of months when he wouldn't get any money back is infinite. Since there is no infinity button on the calculator, he wondered how to tell the calculator that the upper limit is without bound? The answer Alex tried is to use the largest number he can possibly tell the calculator. In scientific notation, the largest number the calculator can interpret is 1×10^{99}, which is usually written as 1E99. Alex accessed the EE command by pressing 2nd, comma. The command he used and the calculator's result are shown in Figure 15.2.8.

```
normalcdf(42,1E9
9,48,6)
          .8413447404
```

Figure 15.2.8: Alex's use of scientific notation to represent infinity

Q5.   What is the probability that the car battery will fail beyond the warranty period?

Alternatively, you do not need to use an absurdly large number. We know that with any normal distribution, there is less than a one in ten thousand chance of observing a value that is more than four standard deviations above the mean.

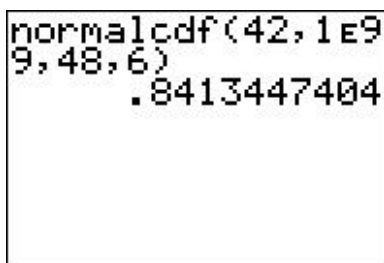Q6.   What value do you obtain when you multiply the standard deviation by 4 and add it to the mean?
         Input this value as the upper limit and compare your answer to that of the previous question.

Option 7 on the CALC menu gives us a nice way to visualize the area under the curve. Let's use that feature to find the same probability. Before repeating this process, we want to clear the shading that we created. While looking at the graph, press 2nd PRGM, and then press option 1. This will clear the shaded region, but also redraw the normal curve. Now that your graph is displayed, we can try to calculate the probability: press 2nd TRACE, option 7, and then enter your lower limit and upper limit.

Q7.   Did you get the same answer? Why or why not?

Option 7 under the CALC menu will only allow you to enter a lower and upper limit within the viewing window of your graph. Try to use this technique to find the probability that the car will fail between 42 months and the maximum *x*-value in your window.

Q8.   How does that compare with the value you found between 42 months and infinity when you used
         the normalcdf command?

Q9.   Can you explain why the probabilities are so close?

Use either technique (normalcdf or option 7) to find the probabilities for all the ranges in the warranty period. Record your answers in a table similar to Table 15.2.8.

Q10.   Do your values make sense?

Q11.   What happens when you add all the probabilities together? Why?

Probability of battery failure at time T
$P(T < 36) =$
$P(36 < T < 39) = 0.044$
$P(39 < T < 42) =$
$P(T > 42) =$

Table 15.2.8: Recording the probabilities of battery failure at time T

Q12.   Since Alex intended to keep his car for 5 years, what is the probability that the battery will fail before 60 months has passed?

Q13.   Do you think he will need to buy another car battery before then? Explain your reasoning.

15.2.2 Expected Value of the Warranty

Once we know the probabilities for each range, we can calculate the expected value of the warranty. Recall that the expected value of a random variable is the long-term average value. So, the expected value of a battery warranty is the long-term average amount of money that a customer gets back when the battery fails. It is calculated by multiplying each probability by the amount of money you would get back and then adding them all together.

Q14.   What is the expected value of the car battery warranty?

Q15.   From Alex's perspective, what does that value mean?

A consumer will only purchase one car battery, in which case the notion of expected value referring to the long run average doesn't really make sense. In the long run, Alex still only has the one car battery that either fails during the warranty period or fails outside the warranty period. On the other hand, the company that sells the battery would be interested in the expected value simply because of the large number of car batteries they have sold. From the company's perspective, they would want to have the expected value of the warranty as small as possible. The warranty requires them to pay in the event of their battery's failure.

Q16.   Suggest changes to the warranty policy to reduce the expected cost of the warranty to less than $10?

Q17.   What happens to the average warranty costs if the company can improve the life of the battery so it lasts, on average 49 months? 50 months?

Since the average life of the car battery is 48 months and Alex intended to keep this car for 5 years, he was concerned about what may happen outside of the warranty period.

Q18.   What is the probability that the battery will last to the average, that is, find $P(T > 48)$?

Q19.   What is the probability that it will last 54 months, or $P(T > 54)$?

Q20.   What is the probability that it will last 5 years or 60 months?

Alex's sister Josephine just returned home for Thanksgiving break during her senior year in college. Her car has the same battery that Alex researched and considered buying. Her car battery is 48 months old. Based on what Alex showed her, she was concerned that her battery would not make it through her final six months of school. Alex applied the formula for conditional probability to determine the probability that a 48 month old battery will last at least another six months.

$$P(T > 54 \mid T > 48) = \frac{P(T > 54)}{P(T > 48)}$$

Q21.   What is this probability? Does Josephine need to be concerned?

15.2.3 A Special Offer

The company that makes the batteries is considering offering a special coupon for owners of really old batteries that are approaching their end of life. They plan to offer this coupon to owners with the 5% oldest batteries. They therefore need to know the lifetime value for which there is only a 5% chance of surpassing. Conversely, 95% of the batteries will have failed before reaching this age. When given the probability, you can use your calculator to find the corresponding **value** on the normal curve. To do this, we will once again call up the distribution menu from the calculator. However, this time we will be inverting the normal distribution. The command is option 3, invNorm(, and the parameters are the percentage, the mean, and the standard deviation, respectively. Figure 15.29 shows the calculator screens.
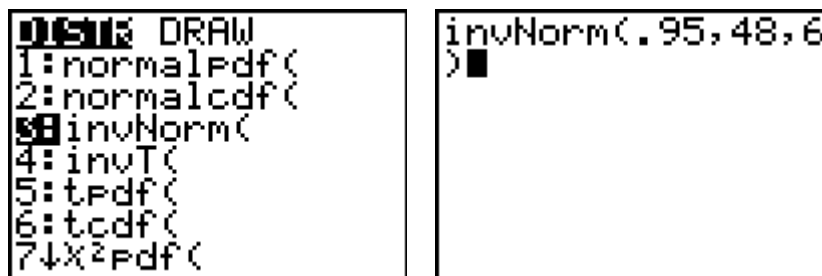
Figure 15.2.9: Using the invNorm command

Q22.   There a 95% chance that the battery will fail before how many months?

The company is considering limiting the coupon to just the 1% oldest batteries.

Q23.   There a 99% chance that the battery will fail before how many months?

Q24.   What if government regulations forced the company to offer the coupon to the 10% oldest batteries. There is a 90% chance the battery will fail before how many months?

Section 15.3: The Standard Normal Distribution

One of the important features of any normal distribution is that it can be rescaled to the **standard normal distribution**. In the standard normal distribution, the mean is 0 and the standard deviation is 1. The standard normal distribution allows you to make quick probability estimates, because the probabilities of being ±1, ±2, ±3 or ±4 standard deviations from the mean are well-known. Those probabilities are given in Table 15.3.1. Thus, 0.68 is a good estimate of the probability of being within 1 standard deviation from the mean, 0.95 is a good estimate of the probability of being within 2 standard deviations from the mean, and 0.99 is a good estimate of the probability being within 3 standard deviations. The table also shows that there is very little chance of being 4 standard deviations or more from the mean. We can also use the symmetry of a normal distribution together with these probabilities to determine that the probability of x being between the mean and 1 standard deviation above the mean. This range has a probability of about 0.34.

Domain of x	Probability
$\mu - \sigma \leq x \leq \mu + \sigma$	68.26%
$\mu - 2\sigma \leq x \leq \mu + 2\sigma$	95.45%
$\mu - 3\sigma \leq x \leq \mu + 3\sigma$	99.73%
$\mu - 4\sigma \leq x \leq \mu + 4\sigma$	99.99%

Table 15.3.1: Probabilities Based on Standard Deviations in a Normal Distribution

Q1.   What is the approximate probability of x being between the mean and 2 standard deviations above the mean?

Q2.   What is the approximate probability of x being between 1 and 2 standard deviations above the mean?

Q3.   What would the probabilities for questions 1 and 2 be if the same questions were asked about values below the mean?

15.3.1 Converting Values to the Standard Normal

We will use the car battery example from the previous section to demonstrate how to use the standard normal distribution. Recall that Alex assumed that battery life is normally distributed with $\mu = 48$ and $\sigma = 6$. Suppose he wanted to find the probability that the car battery would fail in the first 36 months. The first step in using the standard normal distribution is to find what is known as the z-score for 36 months. The z-score of any piece of data, x, is simply the number of standard deviations it is away from the mean. A z-score is found using the following formula:

$$z = \frac{x - \mu}{\sigma}$$

Thus, the z-score for the car battery failing in the first 36 months, is:

$$z = \frac{36 - 48}{6} = -\frac{12}{6} = -2$$

This means that 36 months is 2 standard deviations below the mean.

Q4.   How does the z-score tell you the value is *below* the mean?

We saw in the previous section that the probability was also the area under the normal curve.

Q5.   What is the largest value a probability can take on?

Q6.   What do you think the total area under the curve is? Explain your thinking.

We are looking for the probability that the car battery will fail in the first 36 months. This is the same as asking for the probability that a value is less than 2 standard deviations below the mean.

Once you have rescaled the data to a z-score, you can now use the standard normal distribution to estimate the probability that the battery will fail in the first 36 months. Our standard normal estimate that x lies within 2 standard deviations of the mean is 0.95. Our estimate of the probability that z lies further than 2 standard deviations from the mean is 0.05. But there are two ways that the value of z could be further than 2 standard deviations from the mean. The value of z could be greater than the mean (z is positive), or it could be less than the mean (z is negative). Using the symmetry of the normal distribution, the 0.05 probability of being further than 2 standard deviations from the mean needs to be divided in half. Thus, the probability of being more than 2 standard deviations *above* the mean is 0.025; while the probability of being more than 2 standard deviations *below* the mean is also 0.025. Applying this reasoning to Alex's question, there is only about a $2\frac{1}{2}\%$ chance that the car battery would fail within the first 36 months.

Q7.   How could you estimate the probability that the battery fails after 36 months?

The probability we found contains one of the tails of the distribution. There is slightly more work to be done if we want to estimate a probability which does not contain a tail of the distribution. For example, suppose Alex wanted to find the probability that the battery will fail between 48 and 60 months. First, he would need to convert these data points to z-scores.

Q8.   Find the z-scores for 48 and 60 months.

Recall that the standard normal z-score is the mean of the distribution. Using the symmetry of the normal distribution, the probability that $z < 0 = 0.5$, and the probability that $z > 0 = 0.5$. Now, the probability estimate that z is between -2 and 2 is 0.95. But using symmetry again, one-half of that probability is for being negative, and the other half positive. So, the probability estimate that z is between 0 and 2 is $\frac{1}{2}$ of 0.95, which is 0.475. So, there is almost 1 chance in 2 that the battery will fail between 48 and 60 months.

Q9.   Use a calculator to sketch the standard normal distribution and shade the area that corresponds to falling between the z-scores representing battery failure between 48 and 60 months.

Q10.   Estimate the probability that $z < 1$?

Q11.   Estimate the probability that $z < -2$?

Q12.   Estimate the probability that the battery will fail between 36 and 54 months?

A common error in this situation involves subtracting in the incorrect order. But there is an easy way to tell if the order is wrong. If you use the previous values and subtract them in the wrong order, your result is negative.

Q13.   How can you tell that a negative number cannot be the correct answer to a question regarding probability (or area)?

Section 15.4: *Rapping Skoop Dogg*—Seasonal Demand

The Pineapple Technology Corp. has developed a new item just in time for the holiday season. *Rappin'*
Skoop Dogg is an upright stuffed hound dog, dressed in a leather-like jacket and pants costume. It sings
and dances while standing on a replica of a stage and holding a microphone. There are ten different rap
songs that *Skoop Dogg* can sing as well as dance to. Additional songs can be downloaded from the
Pineapple Technology Corp. web site. There are three settings for activating the dog. *Skoop Dogg* has a
motion detector and can be set to activate when anyone approaches. It can also be activated by clapping
your hands or pushing a button. Commercials for *Rappin' Skoop Dogg* can be seen on the VideoTube
website.

The Chillioz Toy store chain is planning on placing a large order in anticipation of the holiday
Seasonseason. The Pineapple Technology Corp. charges $45 per item for purchases of more than 10,000
units. Chillioz Toys plans to price *Skoop Dogg* at $99. However, items left over after the holiday season
will be steeply discounted to $30. There is an inventory carrying cost of $2 per item left over after the
holiday season ends. This data is summarized below:

P = Purchase Price	$45
S = Selling Price	$99
D = After Christmas sale price	$30
H = Inventory Carry Over Cost	$2
Profit on regular sale	$99 – 45 = $54
Net Cost at end of season for leftovers	$30 – 45 – 2 = -$17

Based on past experience with similar products, executives at Chillioz Toy know there is high potential,
but also a great deal of uncertainty. They forecast that the demand for this item is normally distributed
with mean of 21,000 and standard deviation of 4200 (one-fifth of the mean).

Roy Charles, VP for marketing, is interested in analyzing the impact of different ordering policies. He
starts by wondering how much to order so that there is only a 40% chance that the demand will be less
than this number. He therefore wants to determine the value of *x* such that:
   P(X < *x*) = 0.40

To determine this value, VP Charles uses the **inverse** of the **normal distribution**. invNorm(is one of the
functions that can be found in the DISTR menu of a graphing calculator. The syntax is invNorm(*x*,μ,σ).
Figure 15.4.1 contains a calculator screen for this calculation. From this calculation, VP Charles can see
that there is about a 40% chance that the demand for *Rappin' Skoop Dog* will be less than 19,936.

```
invNorm(.4,21000
,4200)
            19935.94
```
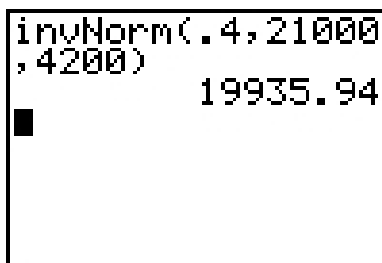
Figure 15.4.1: Using the inverse of the normal distribution

VP Charles noticed that the profit margin was much larger than the potential loss if a product is not sold before the end of the Holiday season. He is considering ordering more than the average demand. This time, he wants to find out how many to order so that there is only a 40% chance that the demand will be *more* than this number.

$$P(X > x) = 0.40$$

However, the calculator only provides the left hand tail of the Normal Distribution. He must first apply the principle of the complement.

$$P(X > x) = 1 - P(X < x)$$

If the area to right of the x is going to be 0.4, the area to the left will have to equal 0.6.

Q1.   Use the inverse normal function to determine how many toys to order.

VP Charles wants to look at the decision from a different point of view. He is hoping to sell as many units as possible and does not want to run out of stock. He considers placing an order that is 1,000 units more than the average. He wishes to know the probability that the demand will be more than that amount.

   a) Select "y = ", then define Y1=Normalpdf(X,21000,4200)

   b) Adjust the window setting before displaying the graph. Set Window: Xmin=21000–4×4200, Xmax=21000+4×4200, Xscl=4200

   c) ZOOM->ZoomFit will adjust Ymin and Ymax

   d) Press ENTER.

   e) The Numerical Integral function will determine the area/probability from the lower limit to the higher limit. CALC->∫f(x)dx. Input the *Lower limit* and *Upper limit*. What are the *Lower* and *Upper* limits? Since we are determining the probability <u>above</u> 22000, the *Lower limit* will be 22000 and use Xmax for the *Upper limit.*

Q2.   What is the probability?

Until now VP Charles focused just on the uncertainty of the demand and not on profits. He considers two scenarios.

        Scenario I: He orders only 21000 units but the demand was actually 22000

In this case Cheelioz would sell out. The total profit is just 21000 · $54= $1,134,000.
However, the company lost 1000 potential sales and missed out on $54,000 extra profits.

        Scenario II: He orders 22000 units but the demand was actually 21000.

In this case Cheelioz would not sell out. They would have to discount 1000 units after the season. The total profit is just 21000 · $54 – 1000 · $17 = $1,117,000

15.4.1 Two Policies—Simulated Demand

VP Charles is not sure how to analyze the decision. How does he balance the risk of having too few against the risk of having too many? He decides to ask two assistant vice presidents, Tom Kawtion and Daniella Egr, to make recommendations. Tom suggests a conservative policy of ordering 22,000 *Rappin' Skoop Dogg* toys. Daniella is more optimistic and suggests an order of 26,000. Roy Charles attempts to evaluate these two policies by simulating ten random seasons of demand. Table 15.4.1 presents an analysis of Tom's policy with the randomly generated demand listed in the second column. The third column records how many items would have been sold. In the first example the demand was only 14092. This was less than the amount ordered. As a result only 14092 were sold. The rest are marked down at the end of season. In the last column the number sold is multiplied by $54. From this is subtracted the number of marked down items multiplied by the net loss of $17. In the second simulation, the demand was 23444. This is greater than the amount ordered. Thus the number that could be sold is just 22000, the number ordered.

Tom's Policy – Order 22,000				
Simulate #	Simulated Demand	Sold	Discounted	Profit
1	14092	14092	$22000 - 14092 = 7908$	$14092 \cdot 54 - 7908 \cdot 17$ $= \$626,532$
2	23447	22000	0	$22,000 \cdot 54$ $= \$1,118,00$
3	20152	20152	$22000 - 20152 = 1848$	$20152 \cdot 54 - 1848 \cdot 17$ $= \$1,056,792$
4	18917	18917	$22000 - 18917 = 3083$	$18917 \cdot 54 - 3083 \cdot 17$ $= \$969,107$
5	28899	22000	0	$1,188,000
6	13100	13100	8900	$556,100
7	25176	22000	0	$1,188,000
8	22283	22000	0	$1,188,000
9	17613	17613	4387	$876,523
10	23450	22000	0	$1,188,000
Average	20713	19387	2613	$1,002,505

Table 15.4.1: Simulated Demand and a Conservative Order Policy

Q3.   How many times did the demand exceed the order policy? In each case what value is recorded in the column marked Discounted and why?

Q4.   What was the average demand for the simulated data? How close is it to the forecasted average? Are you surprised that the number is several hundred units below the forecasted average?

Q5.   What were the least and most profits earned?

Q6.   Why does the exact same total profit of $1,188,000 appear several times?

Table 15.4.2 is set to repeat the above analysis for Daniella's order policy. You are asked to help Roy Charles complete the calculations for simulations 2 thru 10.

Daniella's Policy – Order 26,000				
Simulate #	Simulated Demand	Sold	Discounted	Profit
1	14092	14092	26000-14092 = 11908	$14092 \cdot 54 - 11908 \cdot 17$ $= \$558,532$
2	23447			
3	20152			
4	18917			
5	28899			
6	13100			
7	25176			
8	22283			
9	17613			
10	23450			
Average	20713			

Table 15.4.2: Simulated Demand and an Optimistic Order Policy

Q7.   How many times did the demand exceed the order policy?

Q8.   What were the least and most profits earned?

Q9.   How do these profits compare to the results for Tom's order policy.

Q10.   Why is the maximum total profit not repeated several times?

Q11.   What is the average profit under Daniella's Policy?

Q12.   Which policy would you recommend to VP Roy Charles?

Q13.   Which student policy earned the most profit?

Q14.   What was the average profit?

15.4.2 Class Simulation Demand

VP Roy Charles volunteered to give a lecture on demand and inventory management in a business school course. He decides to use this opportunity to explore other possible order strategies. He also wants to use a larger random sample. In teams of two, he asks every pair of students to come up with their recommended order policy in units of 100. They record that value in a copy of the following table. Next, he is going to simulate on his calculator a random number based on the normal distribution with a mean of 21000 and a standard deviation of 4200. After he calls out the demand and records it on the board, each team is to calculate its profit for that simulation. At the end of the twenty simulations, the team with the highest average profit will receive a free *Rappin' Skoop Dogg*.

Student's Policy – Order _____				
Simulation #	Simulated Demand	Sold	Discounted	Profit
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
Average				

Table 15.4.3: Simulated Demand and Student Order Policy

Q15.   Which student policy earned the most profit?

Q16.   What was the average profit?

15.4.3 A Formula for Ordering

After class, the instructor Dr. Regina Smarty whispered to the VP that operations researchers have developed a mathematical formula to determine the optimal order policy. This policy is optimal on average. The formula uses the selling price (*S*), purchase price (*P*), discounted price (*D*), and inventory carrying cost (*H*).

According the formula the optimal value of *x* is such that

$$P(X < x) = \frac{S - P}{S - D + H} = \frac{99 - 45}{99 - 30 + 2} = \frac{54}{71} \approx 0.761$$

In this case, the optimal policy is the value of *x* such that the probability that demand is less than *x* is 0.761. If no integer value of X exactly matches this probability, the formula states to pick the next largest integer.

Q17.   Using the invNorm(function, what is the optimal number to order?

Q18.   How does this value compare to the best policy found by the students?

VP Charles did not have time to check out the claims of Dr. Smarty. On his way back to the office, he was thinking of calculating the average profit for this formula derived policy. He was going to use the class-simulated data in Table 15.2.3.

Q19.   Why might the best student's order quantity have produced a higher average profit than the order quantity based on the formula?

Section 15.5: Chapter 5 (Normal Distribution) Homework Questions

1. An expert witness in a paternity suit testifies that the length (in days) of human gestation is approximately normally distributed with parameters μ = 270 and σ = 10. The defendant in the suit is able to prove that he was out of the country during a period that began 290 days before the birth of the child and ended 240 days before the birth. If the defendant was, in fact, the father of the child, what is the probability that the mother could have had the very long or very short gestation indicated by the testimony? (This problem is adapted from "A first course in probability" book by Ross)

2. Light bulbs manufactured by Acme Corporation last 1000 hours with a standard deviation of 100 hours.
    a) Assuming that the life of a bulb is normally distributed, what is the probability that an Acme light bulb will last at least 950 hours?

    b) The amount of time in a year that the closet light is turned on has a mean of 950 hours. What is probability that light bulb will be on more than 1000 hours in a random year? (Assume 200 for the standard deviation.)

    c) Find the probability of replacing the light bulb in one year. (Hint: The probability of replacing in one year is failure probability less than lifetime. Difference between two normal distributions is a normal distribution with mean of subtracting two means and variance of summation of two variances)

3. Assume that the life of a laptop computer has a normal distribution with a mean of 38 months and a standard deviation of 8 months.
    a) Find the probability that a randomly selected laptop will need to be replaced in less than 42 months.

    b) What should the warranty period be to replace a malfunctioning laptop if the company does not want to replace more than 5% of all laptops sold?

4. Factory tests of the life of tires indicated that tire life is normally distributed with a mean mileage of 50,000 and a standard deviation 3,000 miles.

    a) What is the probability that the life of a random tire will exceed 55,000 miles?

    b) What is the probability that the life of a random tire will not reach 47,000 miles?

    c) The factory wants to set the guaranteed mileage so that no more than 10% of the tires will have to be replaced. What guaranteed mileage should the factory announce?

5. Graduate Management Aptitude Test (GMAT) scores are widely used by graduate schools of business as an entrance requirement. Suppose that in one particular year, the mean score for the GMAT was 476, with a standard deviation of 107. Assuming that the GMAT scores are normally distributed, answer the following questions:
    a) What is the probability that a randomly selected score from this GMAT falls between 476 and 650.

    b) What is the probability of receiving a score greater than 750 on a GMAT test?

    c) What is the probability of receiving a score of 540 or less on a GMAT test?

    d) What is the probability of receiving a score between 330 and 440?

    e) If a business school wants to consider the top 5 percent students based on their GMAT scores, which score should be considered as a minimum score?

6. After each surgery, a surgeon has appointments with her patients to make sure that the healing process is proceeding well. The duration of each appointment is assumed to be normally distributed with a mean of 10 minutes and a standard deviation of 3 minutes.
    a) Find the probability of a randomly selected appointment taking less than 15 minutes.

    b) Assuming independent appointment times, what is the probability that at least one of the doctor's 20 appointments will be longer than 15 minutes?

    c) If the surgeon works 210 minutes before his lunch, what is the probability of visiting 20 patients before his lunch?

7. The target value for the thickness of a machined cylinder is 8cm. The upper specification limit is 8.2cm and the lower specification limit is 7.9cm. A machine produces cylinders that have a mean thickness of 8.1cm and a standard deviation of 0.1cm. (adapted from http://stat.wvu.edu/srs/modules/Normal/diameters.html)
    a) What is the probability of producing a cylinder within the limits?

    b) What is the probability of exceeding the upper limit?

    c) If the thickness of a cylinder is too thin and it is less than the lower limit, there is no way to use the product and it will be wasted. But in case the cylinder produced is too thick, the cylinder can be used with some rework. Write a formula for the expected value of the profit based on the production and rework costs.

8. Suppose you must establish regulations concerning the maximum number of people who can occupy an elevator. You decide to assume that the weight of one person chosen at random follows a normal distribution with a mean of 70kg and a standard deviation of 20kg. What's the probability that the total weight of 8 people exceeds 600kg?

9. An assembly line consists of 6 stations in order to produce a special type of machine. The task at each station takes 70 minutes on average and has a 20 minute standard deviation.

    a) What is the probability that a machine can be finished in 8 hours of work?

    b) Assume we add a worker to each station in order to reduce the average time of each station to an hour. The standard deviation is assumed not to change. Now what is the probability of finishing a machine in 8 hours of work?

10. Workers A and B are responsible for fixing products returned by customers. Suppose they start their day with 13 products. If the time they spend to fix each product follows normal distribution with following parameters

	Mean(minute)	Standard Deviation(minute)
A	30	10
B	40	5

Table 15.5.1

a) Which employee has more variability in his work?

b) Which one is preferred by a manager who favors consistency?

c) What is the probability of worker A finishing all the products in 7 hours?

d) What is the probability of worker B finishing all the products in 7 hours?

e) What is the probability of overtime for worker A? How about for worker B?

11. Ms. Kuo has had significant stomach pain for the last two days. She scheduled an appointment at the Dreyer Urgent Care Center for today. When she goes to the hospital, she has to first register at the reception desk. When the appointment time comes, a nurse will call her and do pre-screening before the doctor starts examining her.

The duration of each of the tasks is normally distributed. The mean duration and standard deviation for each task is given in Table 15.5.2. Since Ms. Kuo is in significant pain, she is eligible for the fast track. She does not have to wait between these tasks.

Tasks	Mean (minutes)	Standard Deviation (minutes)
1) Patient registration	3.45	0.85
2) Nurse primary screening (triage)	6.40	2.00
3) Medical Doctor thorough examination	15.90	5.60

Table 15.5.2: Tasks to be performed at the Urgent Care Center

a. What is the probability that the registration will take more than 3 minutes?

b. Her appointment is at 9:00 a.m. Since the registration takes some time, she has to arrive at the registration desk earlier than the appointment time. To be 90% confident that she will be on time for her appointment, how early should she arrive at the registration desk?

c. What is the probability that the triage will take less than 4 minutes?

d. What is the probability that the registration and the triage will take less than 8 minutes?

e. Calculate the average stay in the Urgent Care, including the registration, triage, and examination.

12. After the examination, the doctor may ask her to have a lab test or an X-ray scan, or both. If he does that, he gets the test results soon afterwards. The doctor meets with the patient to review the test results.

Tasks	Mean	Standard Deviation
1) Lab test	13.67	5.3

2) X-ray room	20.70	6.5
3) Doctor after testing meeting	7.58	2.5

Table 15.5.3: Possible tasks after doctor's thorough examination

a) The doctor has sent Ms. Kuo to the lab. What is the probability that she will be back within 7 minutes?

b) The doctor thinks he can thoroughly examine one patient until Ms. Kuo comes back from the lab. The doctor will see only one patient during this time even if Ms. Kuo is not yet back from the lab. Assume Ms. Kuo's lab tests took 16 minutes to complete and she is returning to the doctor's office. What is the probability that Ms. Kuo will wait at most one minute for the doctor?

c) The doctor has started the thorough examination of that next patient. What is the probability that after finishing this next patient, he will wait for Ms. Kuo more than 2 minutes to return from the lab?

d) Ms. Kuo has a friend, Hatice Myer, who works fifteen minutes from the clinic. She has agreed to leave work and pick her up after the examination. Ms. Kuo is trying to decide what time she should tell her to come. Describe the basic tradeoff that Ms. Kuo must take into account in specifying the time? What are the uncertainties to take into account?

e) Ms. Kuo talked to the receptionist as to how frequently patients require additional tests. The receptionist's estimates are provided in Table 15.5.4.

Alternative	Possible Tests	Chances
1	No test	40%
2	Lab test	30%
3	X-ray	15%
4	Lab test and X-ray	15%

Table 15.5.4: Likelihood of additional tests

For each of the four alternatives, calculate the mean total time spent in the Urgent Care and its standard deviation. Do not forget that the doctor will meet with Ms. Kuo a second time if tests are required.

Possible Tests	Total Time Spent at Urgent Care	
	Mean	**Standard Deviation**
No test		
Lab test only		
X-ray only		
Lab test and X-ray		

Table 15.5.5: Parameters of Time Spent in Urgent Care with Additional Tests after Thorough Exam

f) Ms. Kuo is considering asking her friend to come by at 9:45 a.m. Calculate the probability of departing by 9:45 a.m. for each of the four alternatives: No test, Lab test, X-ray, and Lab test and X-ray.

Possible Tests	Probability
No test	
Lab test	
X-ray	
Lab test and X-ray	

Table 15.5.6

g) If Ms. Kuo does not require any test, she will have to wait to be picked up. What is the probability that she will have to wait at least 10 minutes if she has no tests?

h) Now consider all of the possible situations. If her friend comes exactly at 9:45 a.m., how likely will Hatice end up waiting more than 10 minutes? (Hint: You will need to determine the weighted sum of the individual alternatives. The weights are the likelihood of each alternative.)

Possible Tests	Chances	Probability of departure later than 9:55 a.m. in each event	Weights
No test	40%		
Lab test	30%		
X-ray	15%		
Lab test and X-ray	15%		

Table 15.5.7

13. The total cholesterol is assumed to be normally distributed with a mean 5.21 mmol/L and a standard deviation of 1.03 mmol/L. The medical interns are observing the total cholesterol levels of patients in a research study. The Hyperlipidaemia Society classifies the total cholesterol level as,
   - desirable (<5.2 mmol/L)
   - borderline (5.2 to 6.4 mmol/L)
   - abnormal (6.4 to 7.8 mmol/L)
   - high (>7.8mmol/L)
   a) What percent of the patients have cholesterol that is in the desirable range?

   b) What percent of the patients have cholesterol levels that are borderline?

   c) What percent of the patients have abnormal cholesterol levels?

   d) What percent of the patients have high cholesterol?

14. A pharmacist is selling two types of cholesterol lowering drugs, Drug A and Drug B, and he is considering the stocking levels of each drug. From his records, the demand for Drug A is 20 boxes with a standard deviation of 5 and the demand for Drug B is 10 boxes with a standard deviation of 3 boxes a month and both are independent and normally distributed.

   a) How many boxes of each drug should he have at the beginning of each month if he wants to fill 90% of each demand?

   b) The combined monthly number of patients with high cholesterol in the area hospital is normally distributed with a mean of 30 and a standard deviation of 5.8. The hospital is considering using

Drug C for lowering cholesterol. It is a substitute for both Drug A and Drug B. How many boxes of Drug C should be kept at the beginning of each month if the hospital wants to fulfill 90% of the patient requirements?

c) Compare the number of Drug C boxes to the total number of boxes of Drug A and B. Explain why the hospital needs less or more boxes of Drug C to satisfy 90% of the demand?

d) The prices for Drug A, B, and C are is $44, $40, and $47, respectively. If the hospital wants to fulfill 90% of the patient requirements, which option is more cost effective, stockpiling both Drug A and Drug B, or stockpiling Drug C?

e) At what price for Drug C would a manager be indifferent between stocking just Drug C or keeping both Drug A and B?

15. Recent reports have shown that there is an inverse relationship between calcium intake and body weight. Thus, changing the calcium intake may reduce the prevalence of obesity. From an experiment, done on middle-aged women, it is known that if the ratio of calcium intake to protein intake is 9 (mg:g), then the average yearly weight change will be $\mu_1 = 0.43 \ kg$ with a standard deviation of $\sigma_1 = 0.61 \ kg$ $\sigma_1 = 0.605 \ kg$. The weight gain is assumed to be normally distributed.

a) What is the probability that a middle-aged woman will lose weight in a year if her calcium to protein intake ratio is 9?

b) What is the probability that she will gain more than 1 kg/year if she follows the suggestion?

   If the ratio of calcium to protein intake is 20 (mg:g), then the average weight change will be $\mu_2 = -0.01 \ kg/year$ with a standard deviation of $\sigma_2 = 0.57 \ kg/year$.

c) According to this given information, how likely a middle-aged woman will lose weight in a year if her calcium to protein intake ratio is 20?

d) What is the probability that she will gain more than 1 kg/year if her calcium to protein intake is 20?

e) Which calcium to protein ratio is more helpful to a person who wants to lose weight and keep fit?

16. There are various intervention methods to reduce the impact of an influenza pandemic. Among those, vaccination is the most effective one. It reduces both morbidity and mortality. However, a vaccine might not be available in the initial stages of the epidemic. It takes on average 45 days to develop a vaccine and it takes another 15 days until it is readily available. Therefore, there is value in slowing the spread of the disease until the vaccine is available.

The peak of the pandemic can be delayed by other effective intervention methods such as use of antiviral drugs, self isolation, and contact reduction. Very sick people should be isolated from other people. Contact reduction is ensured by closure of schools and day care centers. Antiviral treatment is given to 10% of the patients, who are very sick.

The spread of the pandemic after its emergence is assumed to be normally distributed. Table 15.5.7 provides estimates of the mean number of days and standard deviation until the pandemic reaches its peak.

	μ	σ
No Intervention	39	13
Isolation	45	15
Contact reduction	52	17
Antiviral Treatment	53	18
All interventions combined	77	26

Table 15.5.8: Peak of the epidemic (in days) with different intervention methods

a. In how many days, does the pandemic reaches its peak on average if no intervention is used?

b. In how many days, will the vaccine be readily available?

c. What is the probability that the pandemic reaches its peak before the vaccine is readily available if no intervention is used?

d. What is the probability that the pandemic reaches its peak before the vaccine is readily available if antiviral treatment is used?

e. Even though the antiviral treatment is the second most effective tool for delaying the peak, why do you think other intervention methods might be preferred?

f. How many days does self isolation delay the peak of the epidemic on average compared to no intervention?

g. Do you think contact reduction considerably slows down the spread of the epidemic?

h. Why is the impact of all interventions combined not the sum of other interventions?

i. What is the probability that the pandemic reaches its peak before the vaccine is readily available if all interventions are used simultaneously?

Chapter 15 Summary

What have we learned?

Terms

Chapter 15 (Normal Distribution) Objectives

You should be able to:

Chapter 15 Study Guide

Section 16.0: Introduction

What do a Boeing jet, a chicken breast, and a cell phone have in common? We utilize the products and eat the food without ever considering the products may not work or the food may be contaminated. When we board an airplane, we assume the plane will get us to our destination. When we eat a chicken breast, we assume that the meat will not make us sick. When we dial a phone number, we assume we will hear ringing. In fact, we expect *all* Boeing jets, *all* chicken breasts, and *all* cell phones to perform as they should. Every day we use complex things and, in call cases, we assume things are going to function correctly.

Not only do we assume these complex items will perform as they should, but we assume that things that look the same are the same. However, we know that upon close examination, no two things are alike. If every item and every activity is different from that which precedes it or succeeds it, then how can such a complex random world exist? How can things operate with all of this randomness?

Someone or some group must be creating consistency out of randomness. If so, then how do they do this? People conducting **quality control** procedures do all of this. Quality control is what makes our world possible and protects our way of life. In this chapter, we will learn how to create consistency out of randomness by managing the randomness using quality control.

www.flightglobal.com
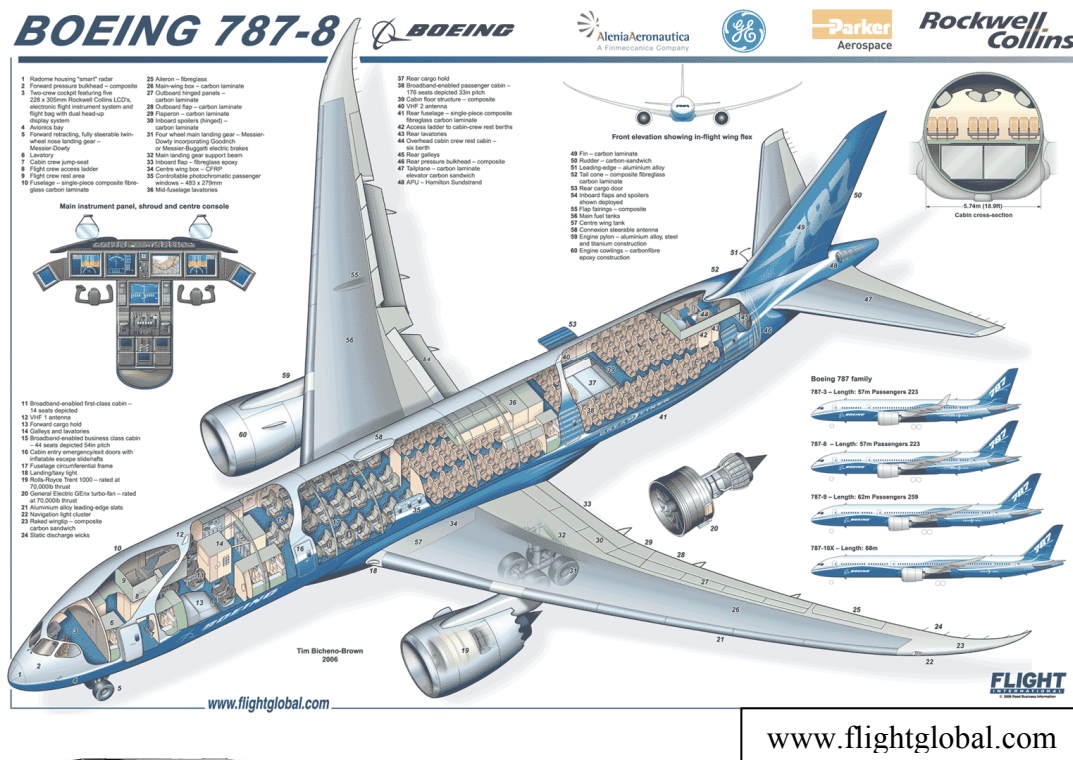
www.apple.com

www.fosterfarms.com

The essence of quality control is to manage the variability during processes such as the assembly of planes, the packing of chicken, or the moderating of phone calls. This *process variability* appears as either "natural" or "unnatural."

Natural variability is a regular part of the process and is the inherent variation of the process. In this chapter you will use the average of averages and the variance of averages to characterize natural variability.

Unnatural variability is the occurrence of uncommon things and suggests the underlying process has changed. These uncommon things might be ongoing, intermittent or a one-time event. The ongoing, intermittent uncommon things might happen in a pattern or in no identifiable pattern. In this chapter you will learn tests for unnatural variability.

The amazing thing is that engineers and scientists utilize *quality control* to determine when uncommon things begin. Thus, they can detect problems before they occur as well as identify problems when they occur. They do both by determining when a process begins to change (detection <u>before</u> a problem occurs) and when it has changed (detection <u>when</u> a problem occurs) from having natural variability to unnatural variability. We use the word "control" because the technique allows us to keep the natural variation within acceptable limits. We use the word "quality" because keeping the natural variation within acceptable limits creates products and processes that function the way they should. Thus, **quality control** is a technique that informs us whether the products and processes we are producing are within the statistical measures we have put in place to manage imprecision.

An example of natural and unnatural variability is:
- *Pens*: The inherent variation in the length of the pen or the cap can be considered "natural" variation; however, a crack in the pen or cap of some pens can be considered an "unnatural" variation.
- *Cookies*: The inherent variation in the diameter of the cookie can be considered "natural" variation; however, broken cookies can be considered an "unnatural" variation.

This change in variability is how scientists determine when a system's performance changes. They notice that the system's behavior changes to such a degree that the natural variation no longer accounts for this shift. Scientists consider this shift to be an "unnatural variation". They use quality control techniques to detect this shift. These are the same techniques that allow airplanes, chicken, cell phones, and all devices to be produced consistently so they function as they should.

The term *quality control* refers to statistical control. **Statistical control** is the stability and predictability of a process over time. A process is *in statistical control* if it is stable over time and exhibits natural variability, meaning its regular random behavior remains unchanged and consistent.

A process is *out of statistical control* if it changes from its regular random behavior, or its norm, such that the process exhibits unnatural variability. It is usually quite difficult to determine intuitively whether natural or unnatural variation exists in a process. Therefore, it is difficult to determine intuitively whether intervention is necessary to correct the unnatural variation (Benneyan, 1998). In this chapter we will explore quality control techniques.

Section 16.1: Assembling the Fuselage of a Boeing 787

We can use quality control to manage the randomness in the manufacturing of individual parts or the assembly of these parts into complex assemblies. An airplane is an excellent example of where quality control is crucial to the manufacturing and assembly of parts. Figure 16.1.1 is a picture of the main assembly pieces of the Boeing 787.
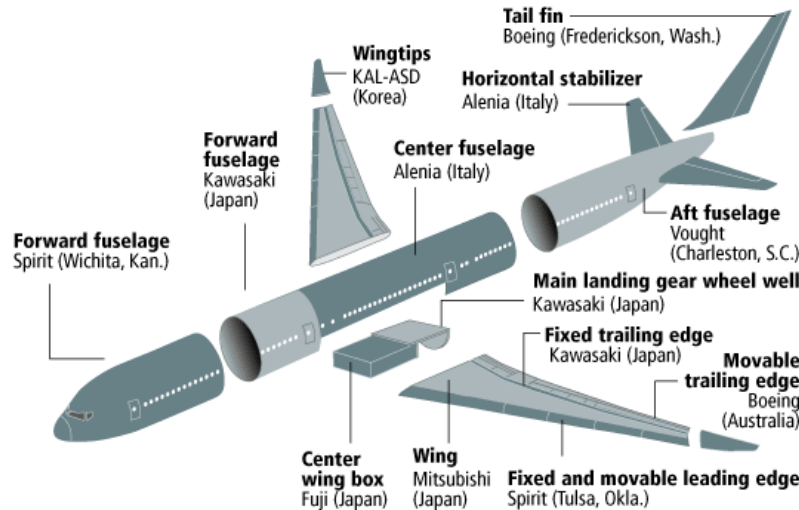
Figure 16.1.1: Main assembly pieces of the Boeing 787 airplane

Bob and Molly are two structural engineers who work for Boeing in Seattle, Washington. Together, they oversee the assembly of the fuselage sections of the Boeing 787. When looking at Figure 16.1.1, the middle sections of the airplane are the fuselage sections, the part of the plane where the passengers sit. Each fuselage section is constructed at a different international location and shipped to the plant in Seattle, Washington where Bob and Molly supervise a large team that joins sections together.

While there are four distinct fuselage sections in Figure 16.1.1, three fuselage sections arrive in Seattle because the *small forward fuselage (Japan)* and *center fuselage (Italy)* merge into a single *center fuselage section* before reaching Seattle. Thus, Bob and Molly's team are in charge of merging the three main fuselage sections which are the *forward fuselage (Wichita, Kansas),* the *center fuselage section which is the small forward fuselage (Japan)* and *center fuselage (Italy) combined*, and the *aft fuselage (Charleston, SC)*. Although the task of assembling the fuselage seems easy, Bob and Molly's team have a very difficult task due to the massive size of these three sections and the necessary precision needed to create a sturdy and sound passenger cabin.

One of the ways that Bob and Molly ensure that the complete fuselage meets the specifications needed to continue assembling a Boeing 787 plane is by measuring the length of the fuselage. While it is near impossible to repeatedly build fuselages that are an exact length, Bob and Molly have a good idea of what variation is acceptable for the assembly process. If the complete fuselage length is randomly within a range of established values, Bob and Molly consider the assembly process of the three sections to be in *statistical control* since there is *natural variability* in the assembly process.

Because Bob and Molly's team complete many fuselages at a time, Bob and Molly need a way to monitor the assembly process to ensure it stays in statistical control. Moreover, they need a way to be able to anticipate when the assembly process starts to falter before they assemble too many faulty completed fuselages.

Q1.   What are possible implications if a completed fuselage is faulty?

For these reasons, Bob and Molly employ quality control procedures using $\overline{x}$ -Charts (pronounced "*x*-bar charts") and *R*-charts to ensure that the assembly of the three sections of the fuselage stays in statistical control. These procedures will be explained in subsequent sections.

We need to understand the assembly process in the Seattle, Washington. Bob and Molly's team will put together five fuselages at a time, and then take a break to evaluate their work and make sure the completed fuselages meet their standards. Bob and Molly use statistical language when referring to this assembly process. They call each set of five completed fuselages a sample, and then call each of the five fuselages a trial in that sample.

Because Bob and Molly want to anticipate when the assembly process starts to become out of statistical control, Bob and Molly will study the natural variations of the completed fuselages. As stated earlier, they use the length of the complete fuselage as one measure of variability. While there are many types of variability, Bob and Molly decided to assess only two types of variability due to time and cost constraints. The first type of variability they will evaluate is the variability *between* the samples (i.e. the variability between the average length of sets of five completed fuselages) to ensure that the samples are consistent. The second type of variability they will evaluate is the variability *within* a sample (i.e. the variability in length between the five completed fuselages in a set) to make sure there is not a wide range of lengths within a set.

Bob and Molly use $\overline{x}$ -Charts to evaluate the variability of length between samples of completed fuselages and *R*-Charts to evaluate the variability of length within samples.

16.1.1 Class Activity—Modeling the Assembly and Measuring the Complete Fuselages from a Boeing 787

As a class, you will assemble fuselages for the Boeing 787 from three main fuselage subsections. You will then measure the assembled fuselages, grouping the measurements into sets of samples and computing the average for each sample set.

First, each of the three sections of the fuselage needs to be created.
- Obtain a sheet with five Boeing 787 fuselages, each divided into three sections as in Figure 16.1.2.
- Cut out each fuselage so that you have five strips of paper.
- Notice that there are two breaks along the horizontal sides of the fuselage. Conduct an "eyeball" cut on the strips in the middle of the breaks so that you have three fuselage sections, a forward fuselage section, a center fuselage section, and an aft fuselage section. When done, you should have fifteen fuselage sections (5 fuselages × 3 sections).
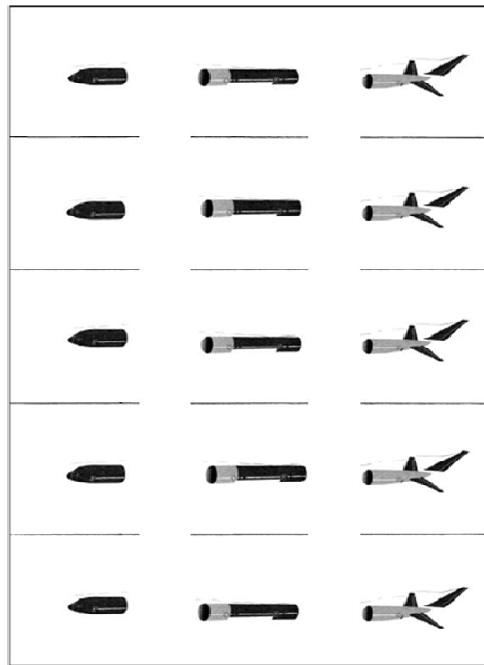
Figure 16.1.2: Boeing 787 fuselage without cutting lines (not to scale, see Worksheet 1)

Second, you need to model the shipping process from the three different locations to Seattle, Washington.
- Gather all the pieces of paper for the front fuselage section, place them into a box, and shake them up. Note, if you are doing this in a group, then the entire group should place their front fuselage pieces into the box together.
- If you are in a group, take five pieces for the front fuselage section from the shaken box and give to each person. By shaking up the box in the step before, everyone should receive a random sampling of each section of the fuselage.
- Repeat the last two bullet points for the center and aft sections of the fuselage.

Third, you need to model the assembly process.
- Place the three fuselage pieces together, end to end, to form a strip. *When putting the three sections together, you do not want any overlap but you want to minimize the gaps between the sections.* Tape the 3 pieces at the seams to assemble a complete fuselage of a Boeing 787. See Figure 16.1.3 for an example of an assembled fuselage. Each participant will complete five Boeing 787 fuselage assemblies. The taping together of the sections represents the assembly operation.

Figure 16.1.3: Example of an assembled Boeing 787 fuselage. Note, the dotted lines represent exaggerated edges that have been cut with scissors

Lastly, you need to measure each completed fuselage for quality control purposes.
- Measure each fuselage using the tool provided in Figure 16.1.4 or a ruler and record their lengths in Table 16.1.1.
- Pool the rest of the class's measurements in Table 16.1.1 (See Worksheet 3)
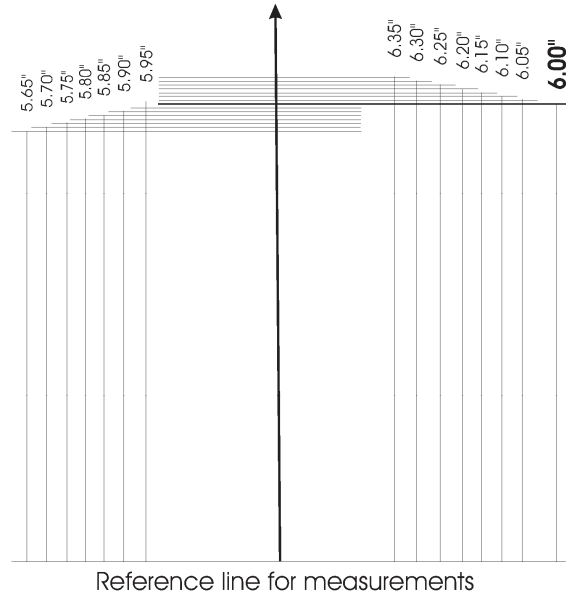
Figure 16.1.4: Measuring tool for completed Boeing 787 fuselages (not to scale, see Worksheet 2)

With this data, you can model the process Bob and Molly undertake to determine if the assembly of fuselage sections is in statistical control.

Section 16.2: Monitoring the Quality of Boeing 787 Fuselage Assembly

Recall that Bob and Molly employ quality control procedures to ensure that the assembling of the three sections of the fuselage stay in statistical control. Moreover, they want to be able to anticipate when the assembly process begins to become out of statistical control.

16.2.2 Using an $\bar{x}$-Chart

One measure Bob and Molly use to gauge the quality of complete fuselage is the length of the fuselage. If the complete fuselage length is randomly within a range of values, Bob and Molly consider the assembly process of the three sections to be in *statistical control* since there is *natural variability* in the assembly process. One way to study the natural variations of the completed fuselage lengths is to look at the variability *between* the samples (i.e. the variability between the average length of sets of five completed fuselages) to ensure that the samples are consistent. We use an $\bar{x}$-Chart (pronounced "x-bar chart") to examine this type of variability.
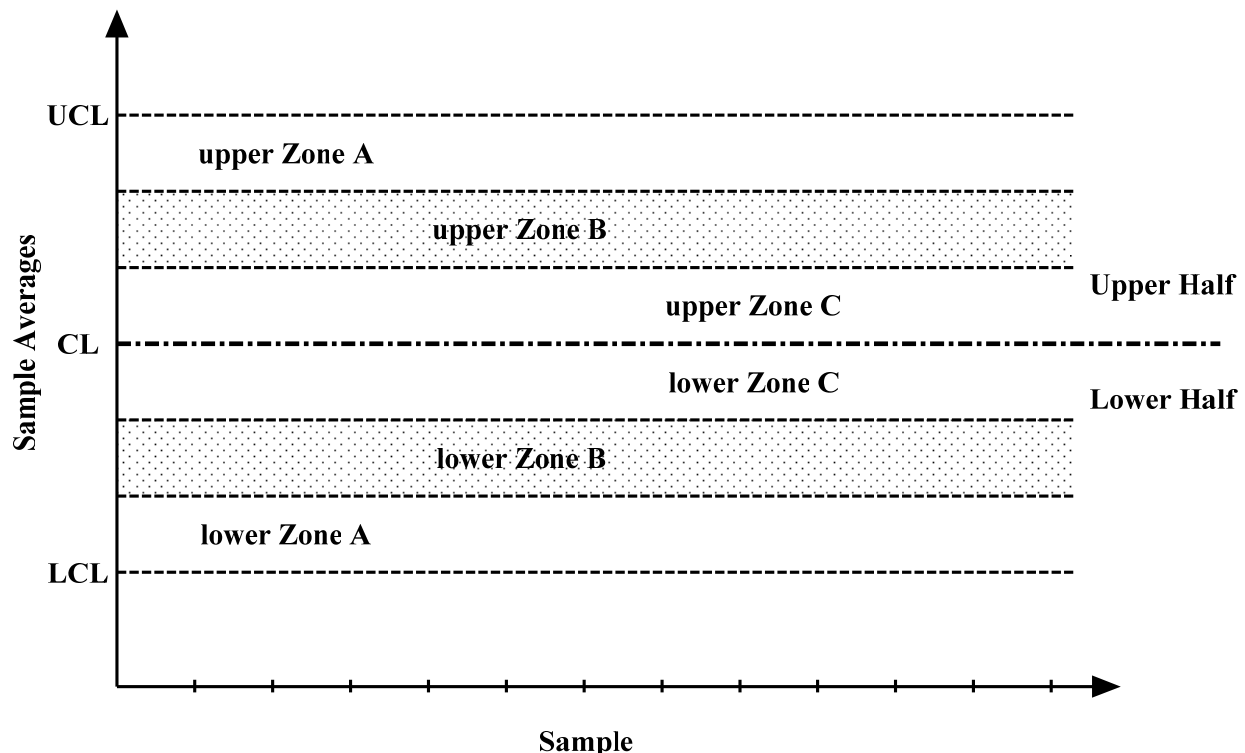
Figure 16.2.1: A general example of a $\bar{x}$-Chart

Figure 16.2.1 is an example of a $\bar{x}$-Chart similar to the one Bob and Molly use to evaluate the average sample length of the completed fuselages. Notice there is six zones between the LCL and the UCL. LCL stands for Lower Control Limit; UCL stands for the Upper Control Limit; and CL stands for Center Line. You will learn how to interpret the six zones below and you will learn how to calculate the LCL, CL, UCL, and the six zones in a subsequent section.

How do Bob and Molly use this chart? For each sample, they calculate the average length of the five completed fuselages and then plot the value on the $\bar{x}$-Chart. Depending on where the data point lands on the $\bar{x}$-Chart, Bob and Molly will be able to judge whether the assembly of the fuselage is in statistical control with regards to length.

When Bob and Molly plot sample averages on their $\overline{x}$-Chart, they are looking for particular characteristics that indicate there is natural variability in the fuselage assembly process.

Pattern Characteristics of Natural Variability between Sample Averages
Characteristic 1: Most of the points are near the center or Center Line (CL).
Characteristic 2: The number of points on one side of the Center Line should be about equal to the number of points on the other side of the Center Line.
Characteristic 3: There are few points near the Lower Control Limit (LCL) or near the Upper Control Limit (UCL).
Characteristic 4: Rarely or never is a point beyond the LCL or the UCL.

Let's return to the Class Activity to see how Bob and Molly use the $\overline{x}$-Chart. Recall, you cut out the sections for five fuselages and randomly reassembled the fuselage. You then measured the completed fuselage and recorded the class results in Table 16.2.1.

In order to use the $\overline{x}$-Chart, you must first calculate the average for each sample (remember a sample is a set of trials, in this case, it is a set of measurements from five complete fuselages). This average is the "$\overline{x}$" in $\overline{x}$-Charts. In statistics, "$\overline{x}$" means the average or mean of the data.

Q1.   Find the average for each sample from your class's pooled data. Record these averages in Table 16.2.1.

Table 16.2.2 is an example of a class set of data.

Sample	Trial number 1	2	3	4	5	Average	Sample	Trial Number 1	2	3	4	5	Average
1	5.95	5.9	6	5.9	5.85	5.92	19	5.95	6.1	6.05	5.95	5.9	5.99
2	6.05	5.95	5.8	6.1	6.25	6.03	20	5.95	6	5.95	6.15	5.95	6
3	6	6.05	6.1	5.9	6.15	6.04	21	6.05	6	5.95	6.15	6.1	6.05
4	6.2	6	6	6.05	6.05	6.06	22	6.1	6.2	5.9	6	5.9	6.02
5	5.8	6	6.1	6.05	6	5.99	23	6	6	5.95	6.05	6	6
6	6.05	6	5.9	6.1	6	6.01	24	6.05	6.05	6.05	6.05	6.1	6.06
7	6	6	6.05	6.05	6.05	6.03	25	6.2	6	6	6.1	6.05	6.07
8	5.9	5.85	5.85	6.05	6.1	5.95	26	6	5.85	5.95	6	6.1	5.98
9	6	6.1	5.95	6.05	6.05	6.03	27	6	6.05	5.95	6.05	6.05	6.02
10	6	6	6	5.95	5.95	5.98	28	6.1	6	6	6.05	6.05	6.04
11	5.9	5.8	6	6.05	5.8	5.91	29	6.05	6	6.2	6.05	6.1	6.08
12	5.95	6	6.1	6.1	6.15	6.06	30	6	6	6.05	6.05	5.95	6.01
13	6.05	5.9	5.95	6.05	6	5.99	31	6.05	6.05	6	5.95	5.95	6
14	5.9	6	6.05	6.1	6	6.01	32	6.05	6	6.05	6	6.05	6.03
15	6	6	6	5.9	6.05	5.99	33	6.1	6.05	5.95	6.05	6.1	6.05
16	5.95	6.1	5.95	6.05	6	6.01	34	6	6	6	6.05	6.05	6.02
17	5.95	6	6.1	6.1	6.1	6.05	35	6	6.1	6.05	6.05	6.05	6.05
18	5.9	6.05	6	6.2	6.05	6.04							

Table 16.2.2: Example pooled class data for complete Boeing 787 fuselages (in inches) with sample averages

Just as Bob and Molly have an $\bar{x}$–Chart to evaluate the variability between samples, there is a $\bar{x}$-Chart for the modeling of completed fuselage assembly. Notice in Figure 16.2.2, there are six zones as well as the Upper Control Limit (UCL), Center Line (CL), and the Lower Control Limit (LCL).

x-Bar Chart for Completed Fuselage Lengths
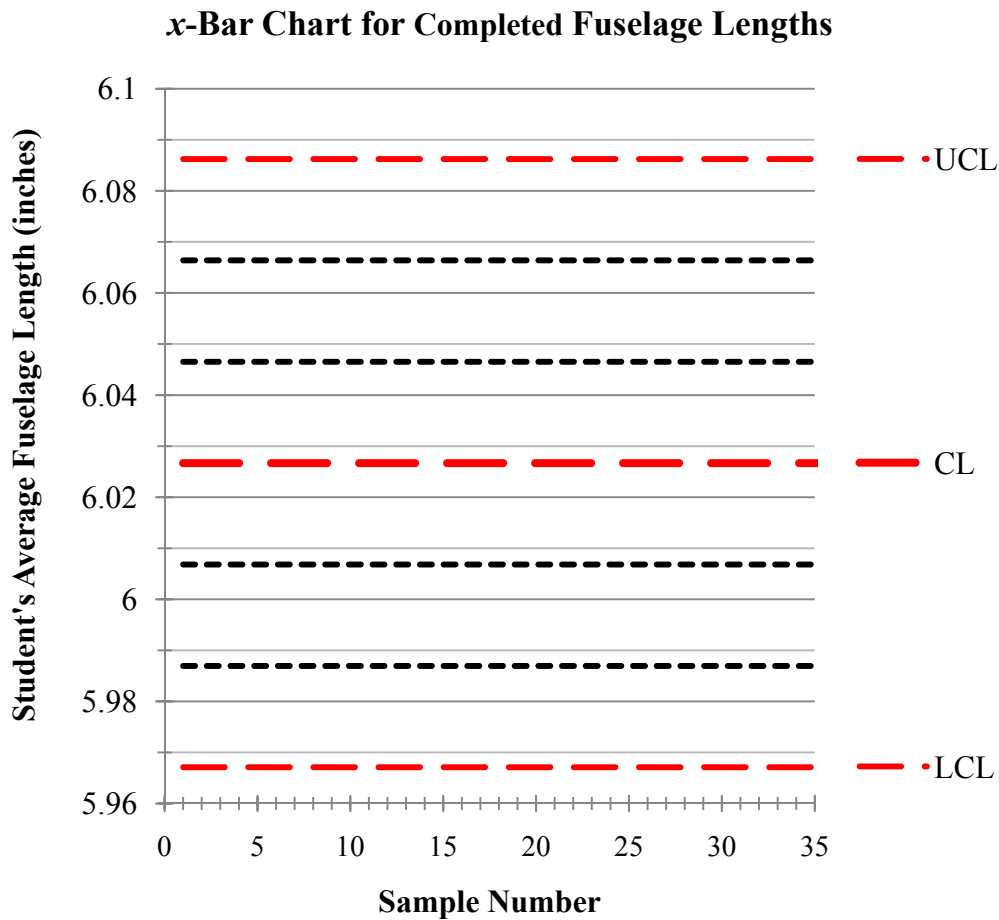
Figure 16.2.2: The $\overline{x}$ –Chart for the average sample lengths of completed fuselages from the Boeing 787 model activity

As Bob and Molly monitor the quality of the fuselages, they plot one average sample length at a time onto the $\overline{x}$ –Chart. For each new data point, they go through a series of tests, see below, to anticipate if the fuselage assembly process is out of statistical control.

Tests for Unnatural Patterns in Variability between Average Samples using Figure 16.2.1
Test 1: A single point falls beyond Zone A.
Test 2: Two out of three consecutive points are on one side of the Center Line (CL) and fall in Zone A or beyond.
Test 3: Four out of five consecutive points are on one side of the Center Line and fall in Zone B or beyond.
Test 4: Eight successive points are on one side of the Center Line and fall in Zone C or beyond.

Figure 16.2.3 shows how the tests for unnatural patterns applies to the general example of a $\overline{x}$ -Chart.
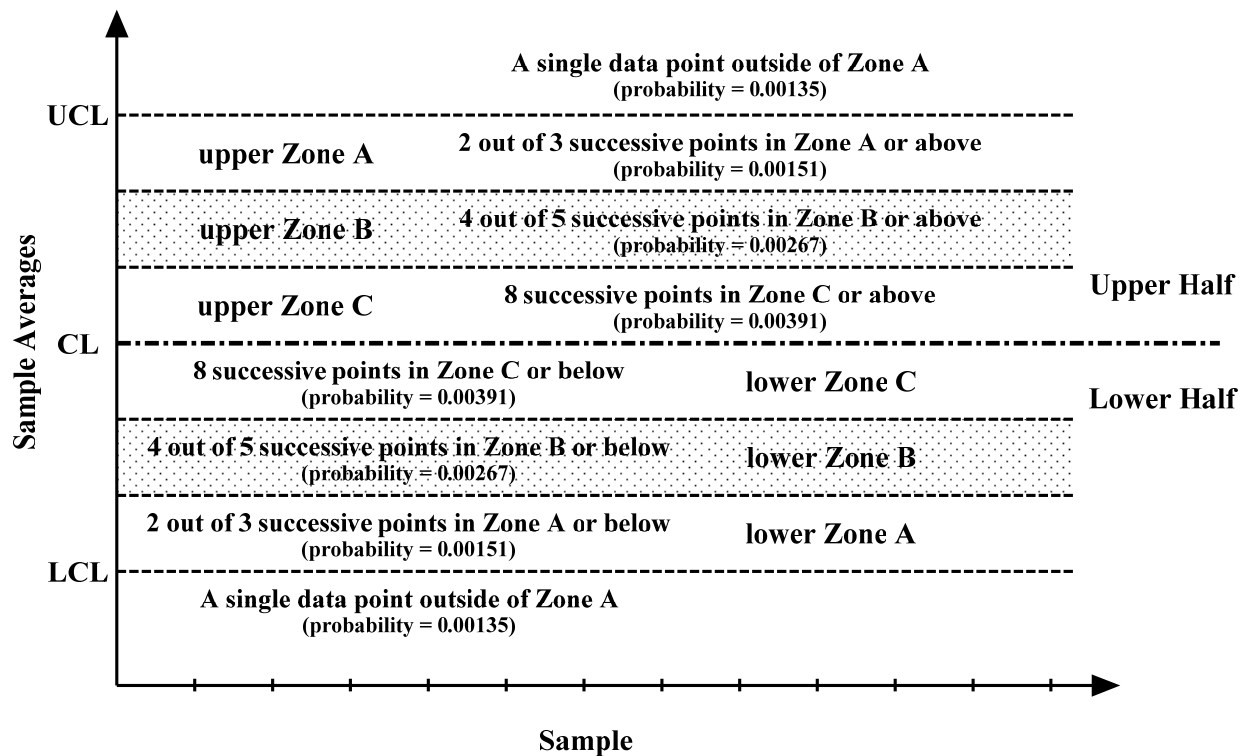
Figure 16.2.3: Summary of tests for unnatural patterns in a general example of a $\overline{x}$ –Chart

The *order* from point to point is not random. The points are ordered sequentially in time.

Test 1 means if we are at one data point then the probability of the next data point being beyond Zone A is extremely low at 0.00135 or 0.135%. If this happens then we need to determine if it is just a random event or if something is fundamentally wrong with the process.

Test 2 means that when we are at one data point then the probability of two out of the next three consecutive data points on one side of the Center Liner and falling in Zone A or beyond is extremely low at 0.001510778 or 0.151%. If this happens then it is unlikely to have come from a process performing correctly and it is likely something is fundamentally wrong with the process.

Test 3 means that when we are at one data point then the probability of four out of five of the next five consecutive points on one side of the Center Line and falling in Zone B or beyond is extremely low at 0.002668262 or 0.267%. If this happens then it is unlikely to have come from a process performing correctly and it is likely something is fundamentally wrong with the process.

Test 4 means that when we are at one data point then the probability the next eight consecutive points on one side of the Center Line and falling in Zone C or beyond is extremely low at 0.00390625 or 0.391%. If this happens then it is unlikely to have come from a process performing correctly and it is likely something is fundamentally wrong with the process.

Figure 16.2.4 is the example pooled class data from for complete Boeing 787 fuselages from Table 16.2.2 plotted onto the $\overline{x}$ –Chart for the average sample lengths of completed fuselages from the Boeing 787 model activity (Figure 16.2.1).

x-Bar Chart for Complete Fuselage Lengths
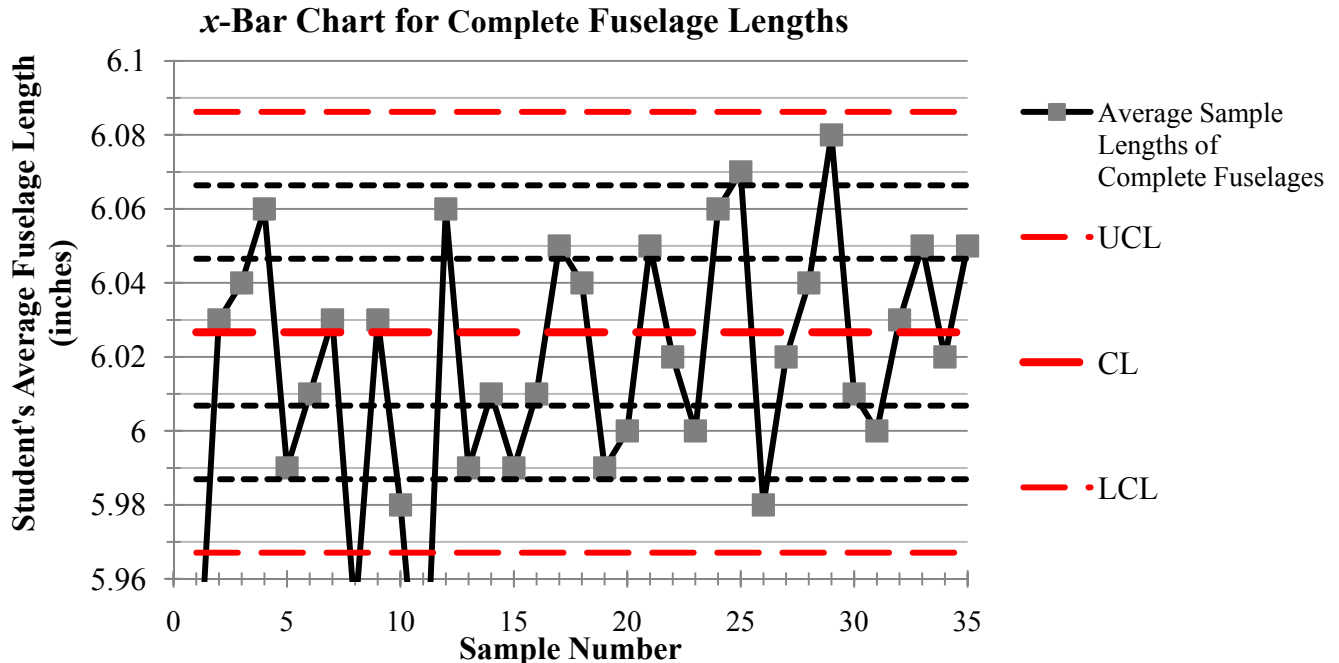
Figure 16.2.4: Example pooled class data plotted onto the $\overline{x}$–Chart for the average sample lengths of completed fuselages from the Boeing 787 model activity

Notice from the start, Sample 1 average fuselage length (5.92 in.) is outside of Zone A. Thus, the sample satisfies the condition for Test 1 of unnatural patterns. What does this mean about the assembly process? If the process is in statistical control, then there is a very low probability ($P = 0.00135$) that a data point will randomly be outside of Zone A. Because this probability is so low, it is not likely that this bad sample average happened by chance. Thus, a more likely explanation is that the fuselage assembly process is not in statistical control; however, Test 1 does not guarantee that this is the case. Instead, it is an indicator that something might be awry and that the class (or Bob and Molly) should stop the assembly and discern if there is an issue.

Sample 8 average length and Sample 11 average length (5.95 and 5.91 in., respectively) also satisfies the condition for Test 1 of unnatural patterns. Thus, after these samples, the class (or Bob and Molly) should also take time to evaluate the assembly process. While it is possible that these average lengths occurred randomly during the assembly process, a more likely explanation is that their large variability in length compared to the other samples is caused by something going wrong during the assembly. It is in these moments that Bob and Molly can anticipate when the assembly process might become out of statistical control before they make more faulty fuselages (and cost the company, time, money, and wasted resources).

At no other data points in Figure 16.2.4 does there appear to be unnatural patterns (the data fails Tests 2 – 4). Consequently, there is no cause of alarm that the assembly process is out of statistical control for these data points.

Now apply the tests for unnatural patterns to your class data. The greater the system instability, the more points will be identifies by these tests.

Q2.   Using the $\overline{x}$–Chart in Figure 16.2.2 (see Worksheet 4), plot the class averages of the sample lengths of completed fuselages one at a time and determine if the class's assembly process is ever out of statistical control. If it is, at what sample in the assembly process does this occur? Why do

you think the assembly process is out of statistical control? What are the implications for being out of statistical control?

The tests for unnatural patterns in variability between average samples are the four basic tests recommended for industry use to identify problems in the process. In the next section we will look at some other indicators of unnatural variation.

16.2.2 Using an *R*-Chart

Recall that Bob and Molly decided to assess two types of variability concerning the length of the complete Boeing 787 fuselages. They use an $\bar{x}$ –Chart to evaluate the variability *between* the samples (i.e. the variability between the average length of sets of five completed fuselages) to ensure that the samples are consistent. They use an *R*-chart to evaluate is the variability *within* a sample (i.e. the variability in length between the five completed fuselages in a set) to make sure there is not a wide range of lengths within a set. In other words, the key characteristic for the *R*-chart is to confirm that the variability in the process is stable.
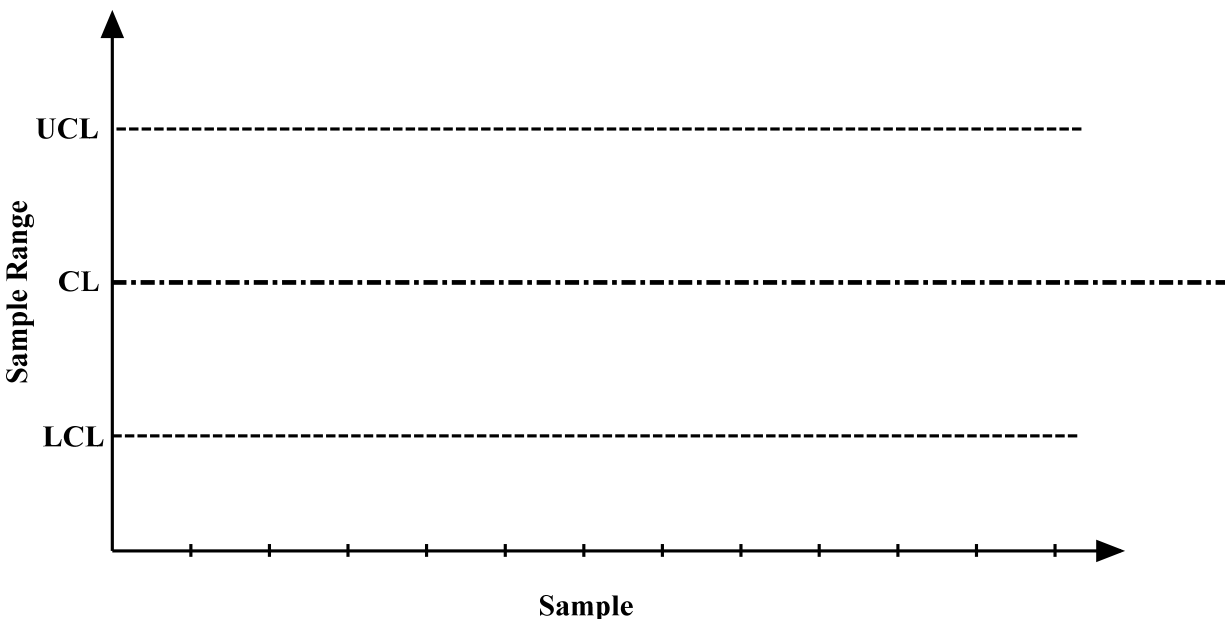
Figure 16.2.5: A general example of an *R*-Chart

Figure 16.2.5 is an example of an *R*-Chart similar to the one Bob and Molly use to evaluate the range of trails lengths of the completed fuselages. Notice that the *R*-chart looks similar to the $\bar{x}$ -Chart. There is a Lower Control Limit (LCL), a Center Line (CL), and an Upper Control Limit (UCL). However, these values will not be the same as the LCL, CL, and UCL for an $\bar{x}$ -Chart, you use different values and a different calculation processes for an *R*-Chart.

How do Bob and Molly use this chart? For each sample, they calculate the range of the five trials. Range is the measure of the spread of the trials in each sample. Range is calculated by taking the difference between the largest value and the smallest value in a sample. Mathematically, the smallest and largest values are the minimum value and the maximum value, respectively. Another way to express this is:

Range of sample = (maximum trial value) – (minimum trial value)

In context of the Boeing 787 fuselages, the range of a sample is the difference between the maximum complete fuselage length and the minimum complete fuselage length for each set of five fuselages. Depending on where the data point lands on the *R*-Chart, Bob and Molly will be able to judge whether the process variability is stable and the assembly process is in statistical control.

Let's return to the Class Activity to see how Bob and Molly use the *R*-Chart. Recall, you cut out the sections for five fuselages and randomly reassembled the fuselage. You then measured the completed fuselage and recorded the class results in Table 16.2.1.

In order to use the *R*-Chart, you must first calculate the range for each sample. This range is explains the "*R*" in *R*-Charts.

Q3.   Find the range for each sample from your class's pooled data. Record these averages in Table 16.2.1.

Table 16.2.3 is an example of a class set of data with the sample ranges calculated. Note, the trial values are the same as is Table 16.2.2.

Sample	Trial number					Range	Sample	Trial Number					Range
	1	2	3	4	5			1	2	3	4	5	
1	5.95	5.9	6	5.9	5.85	0.15	19	5.95	6.1	6.05	5.95	5.9	0.2
2	6.05	5.95	5.8	6.1	6.25	0.45	20	5.95	6	5.95	6.15	5.95	0.2
3	6	6.05	6.1	5.9	6.15	0.25	21	6.05	6	5.95	6.15	6.1	0.2
4	6.2	6	6	6.05	6.05	0.2	22	6.1	6.2	5.9	6	5.9	0.3
5	5.8	6	6.1	6.05	6	0.3	23	6	6	5.95	6.05	6	0.1
6	6.05	6	5.9	6.1	6	0.2	24	6.05	6.05	6.05	6.05	6.1	0.05
7	6	6	6.05	6.05	6.05	0.05	25	6.2	6	6	6.1	6.05	0.2
8	5.9	5.85	5.85	6.05	6.1	0.25	26	6	5.85	5.95	6	6.1	0.25
9	6	6.1	5.95	6.05	6.05	0.15	27	6	6.05	5.95	6.05	6.05	0.1
10	6	6	6	5.95	5.95	0.05	28	6.1	6	6	6.05	6.05	0.1
11	5.9	5.8	6	6.05	5.8	0.25	29	6.05	6	6.2	6.05	6.1	0.2
12	5.95	6	6.1	6.1	6.15	0.2	30	6	6	6.05	6.05	5.95	0.1
13	6.05	5.9	5.95	6.05	6	0.15	31	6.05	6.05	6	5.95	5.95	0.1
14	5.9	6	6.05	6.1	6	0.2	32	6.05	6	6.05	6	6.05	0.05
15	6	6	6	5.9	6.05	0.15	33	6.1	6.05	5.95	6.05	6.1	0.15
16	5.95	6.1	5.95	6.05	6	0.15	34	6	6	6	6.05	6.05	0.05
17	5.95	6	6.1	6.1	6.1	0.15	35	6	6.1	6.05	6.05	6.05	0.1
18	5.9	6.05	6	6.2	6.05	0.3							

Table 16.2.3: Example pooled class data for complete Boeing 787 fuselages (in inches) with sample ranges

Just as Bob and Molly have an *R*-Chart to evaluate the variability within samples, there is an *R*-Chart for the modeling of completed fuselage assembly; See Figure 16.2.5.

R-Chart for Complete Fuselage Lengths
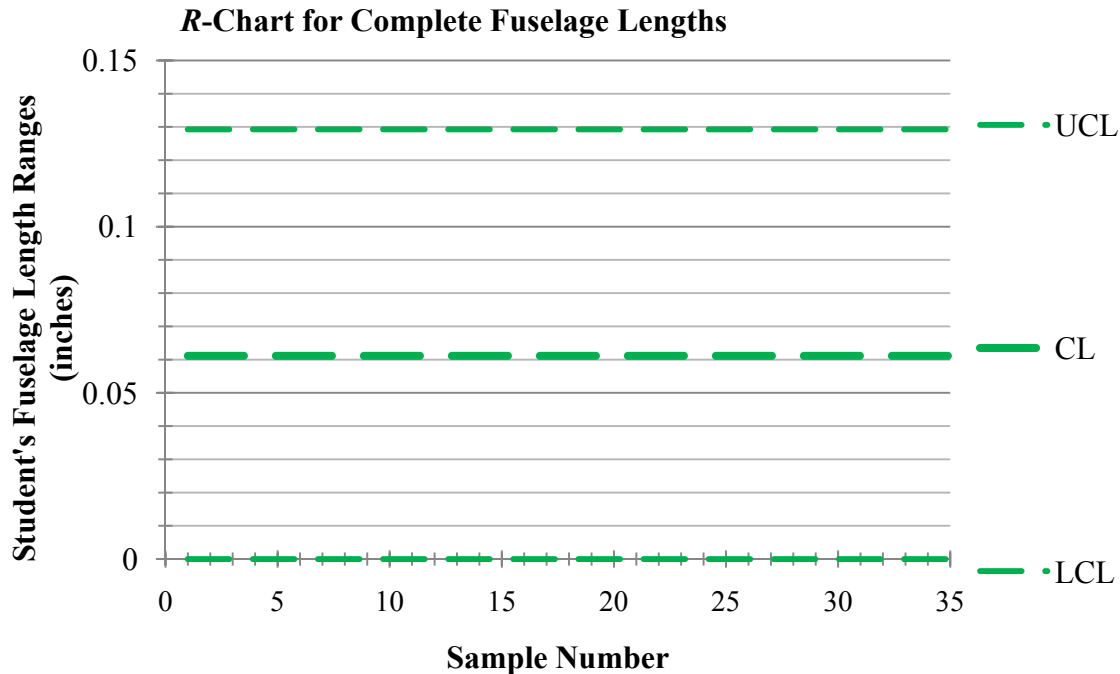
Figure 16.2.6: The *R*–Chart for the sample range of lengths of completed fuselages from the Boeing 787 model activity

Notice the values of the Upper Control Limit (UCL), Center Line (CL), and the Lower Control Limit (LCL) in Figure 16.2.6; these are difference values from the UCL, CL, and LCL of the $\bar{x}$ –Chart (see Figure 16.2.1).

A point to consider with the *R*-Chart is that there will never be negative ranges.

Q4.   Why will there never be negative data points on a *R*-Chart?

When Bob and Molly plot sample ranges on their *R*-Chart, they are looking for particular characteristics that indicate there is natural variability in the fuselage assembly process. There are not specific tests for the *R*-Chart as there are for the $\bar{x}$ –Chart, but Bob and Molly visually check for *trends*, *shifts in level*, and *cycles* and *seasonality*.

A *trend* is a continuous movement of data points up or down in one direction. This is a long series of consecutive points with a continuous movement either up or down, but always in one direction.

A *shift in level* is a long series of data points in which all points are either above or below the Center Line (CL) and are not a trend (i.e. moving in one direction). Data points can move up or down, but remain above or below the CL.

A *cycle* is a short trend in the data occurring in repeating patterns. Similarly, *seasonality* is a longer tendency in the data occurring in a repeated pattern. It is called "seasonal" because often these longer tendencies are attributes to changes in the season that result in changes in the environment. An example of environment change is cold, dry climate in the winter and hot, humid climate in the summer. It is the repeated patterns of *cycles* and *seasonality* that violates natural variations' primary characteristic of having no discernable pattern. Thus, arguing whether a repeated pattern is a cycle or a seasonality is not important since there terms are often used interchangeably. Instead, the key is that a repeated pattern

indicates the process has moved from a state of natural variation (i.e. in statistical control) to a state of unnatural variation (i.e. out of statistical control) and that some corrective action may need to be taken.

As Bob and Molly monitor the stability of the process variation, they plot one sample range at a time onto the *R*-Chart. For each new data point, they visually check for *trends*, *shifts in level*, and *cycles and seasonality* to anticipate if the fuselage assembly process is out of statistical control.

Figure 16.2.6 is the example pooled class data from for complete Boeing 787 fuselages from Table 16.2.3 plotted onto the *R*-Chart for the sample range of lengths of completed fuselages from the Boeing 787 model activity (Figure 16.2.5).
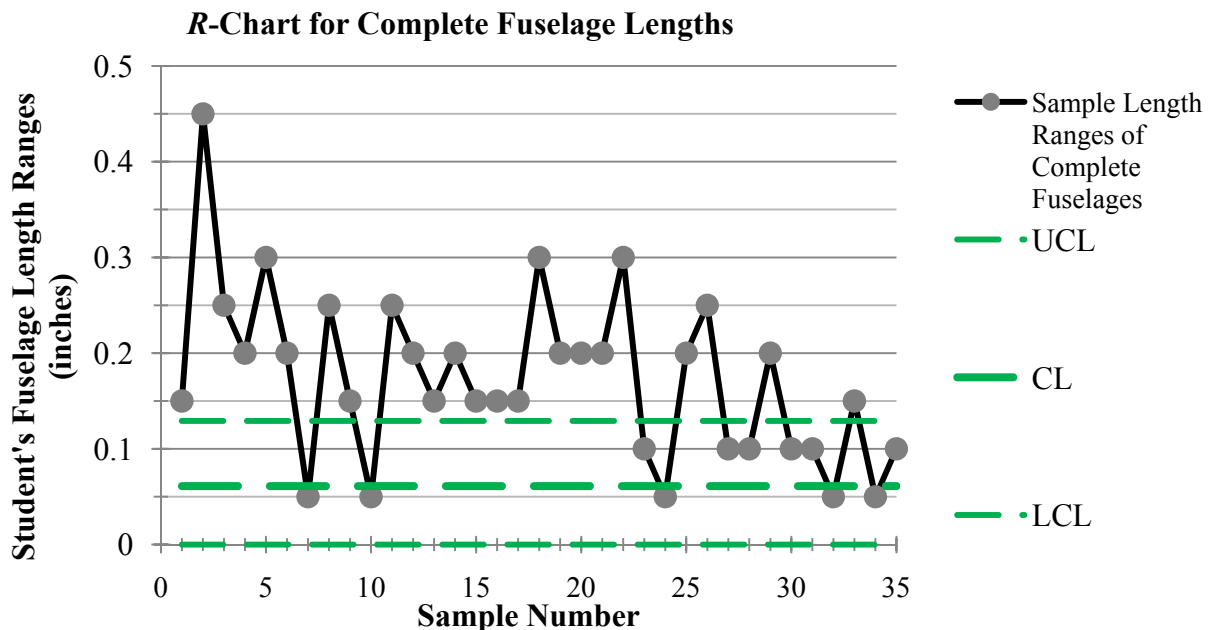
Figure 16.2.7: Example pooled class data plotted onto the *R*-Chart for the sample length ranges of completed fuselages from the Boeing 787 model activity

There is no clear trend in the data points, meaning at no point is there continuous movement in one direction. Moreover, there is no repeated pattern in the data; consequentially, there is no cycle or seasonality in the data. However, there is a shift in level. All the samples except for Samples 7, 10, 24, 32, and 34 have ranges above the Center Line (CL).

What does this mean about the example class data's assembly process? There is a wide variability between trials in a sample. Therefore, the process variability is not stable and the assembly process is out of statistical control. Molly and Bob probably would have stopped the fuselage assembly to evaluate the process sometime around Sample 3 or Sample 4. The probability of being below the LCL or above the UCL is small (1% in each case). In Figure 16.2.6 we see that there are a lot of samples above the UCL even though the probability of being above the UCL is 0.01 (1%) there is a lot of variability between trials within the first few samples indicating and that there may be issues with the assembly process.

Now apply these visual indicators (*trends*, *shifts in level*, and *cycles and seasonality*) to your class data. The greater the system instability, the more points will be identified by these tests.

Q5.   Using the *R*-Chart in Figure 16.2.5 (see Worksheet 5), plot the class ranges of the sample lengths of completed fuselages one at a time and determine if the class's assembly process is ever out of statistical control. If it is, at what sample in the assembly process does this occur? Why do you

think the assembly process is out of statistical control? What are the implications for being out of statistical control?

Now that you have evaluated both types of variability for the assembly process for Boeing 787 fuselages, you need to look at the quality control system as a whole.

Q6.   What can you say about the fuselage assembly process? Does your class have natural or unnatural variability as you joined fuselage sections together? Why?

Q7.   If we have unnatural variability, what are some suggestions for tweaking the assembly process so that the process is in statistical control? What are the pros and cons of these suggestions?

Q8.   Now that you have a firmer grasp on $\bar{x}$-Charts and R-Charts, are the data points in $\bar{x}$-Charts correlated to the data points in R-Charts? For example, if Sample 1 has a high $\bar{x}$-value, then will Sample 1 have a high R-value or vice-versa (or the inverse)?

In the last two sections, we learned how Bob and Molly use two types of variability (variability between samples and variability within a sample) to determine the quality of the assembly process as their team constructed complete Boeing 787 fuselages. They take the appropriate data points (sample length averages and sample length rages, respectively) and plot them onto quality control charts ($\bar{x}$–Chart and R-Chart, respectively) to determine if the assembly process is in statistical control. But where do the charts come from? How do Bob and Molly know what the Upper Control Limit (UCL), Center Line (CL), and Lower Control Limit (LCL) should be for each chart? The next three sections will explain how to make these calculations, but first we need data.

Section 16.3: Compiling Data to Create Quality Control Charts for the Assembly of Boeing 787 Fuselages

When Bob and Molly's team assemble the fuselages, they do so under normal manufacturing conditions. However, in order to create quality control charts, their team must take some time to assemble fuselages in optimal conditions. It is in these tightly controlled conditions that Bob and Molly are able to establish what variation is acceptable and what variation is not acceptable.

As a class, you will assemble fuselages for the Boeing 787 from three main fuselage subsections under controlled conditions. You will then measure the assembled fuselages, grouping the measurements into sets of samples and computing the average for each sample set.

First, each of the three sections of the fuselage needs to be created to the best of your ability.
- Obtain a sheet with five Boeing 787 fuselages, each divided into three sections as in Figure 16.3.1.
- Cut out each fuselage so that you have five strips of paper.
- *Precisely* cut strips along each line so you have three fuselage sections, a forward fuselage section, a center fuselage section, and an aft fuselage section. When done, you should have fifteen fuselage sections (5 fuselages × 3 sections).
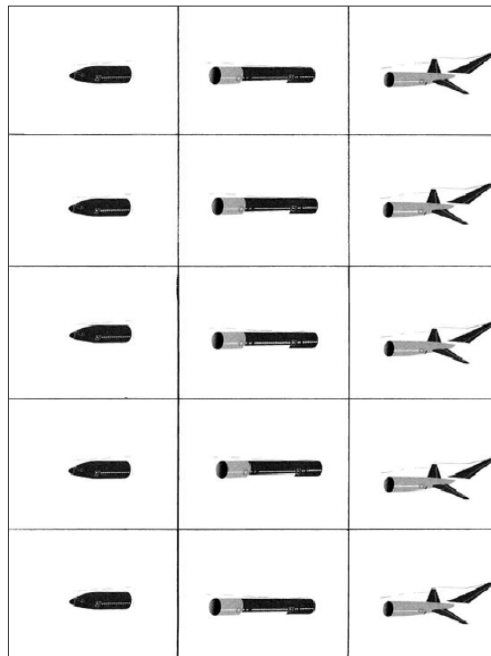
Figure 16.3.1: Boeing 787 fuselage with cutting lines (not to scale, see Worksheet 6).

Second, you need to model the shipping process from the three different locations to Seattle, Washington.
- Gather all the pieces of paper for the front fuselage section, place them into a box, and shake them up. Note, if you are doing this in a group, then the entire group should place their front fuselage pieces into the box together.
- If you are in a group, take five pieces for the front fuselage section from the shaken box and give to each person. By shaking up the box in the step before, everyone should receive a random sampling of each section of the fuselage.
- Repeat the last two bullet points for the center and aft sections of the fuselage.

Third, you need to model the assembly process.

- Place the three fuselage pieces together, end to end, to form a strip. *When putting the three sections together, you want do not want any overlap but you want to minimize the gaps between the sections* (just as before). Tape the 3 pieces at the seams to assemble a complete fuselage of a Boeing 787. Be as precise as possible. Each participant will complete five Boeing 787 fuselage assemblies. The taping together of the sections represents the assembly operation.

Lastly, you need to measure each completed fuselage for quality control purposes.
- Measure each fuselage using the tool provided in Figure 16.2.4 (see Worksheet 2) or a ruler and record their lengths in Table 16.2.4. Be as precise as possible.
- Pool the rest of the class's measurements in Table 16.2.4 (See Worksheet 7)

With this data, you can model the process Bob and Molly undertake to determine the standards for statistical control.

16.3.1 Creating an $\bar{x}$-Chart for the Assembly of Boeing 787 Fuselages

Now that you have sample length averages created from the best possible conditions, it is possible to determine the boundaries of statistical control and create a $\bar{x}$–Chart. Below is a general $\bar{x}$–Chart for reference.
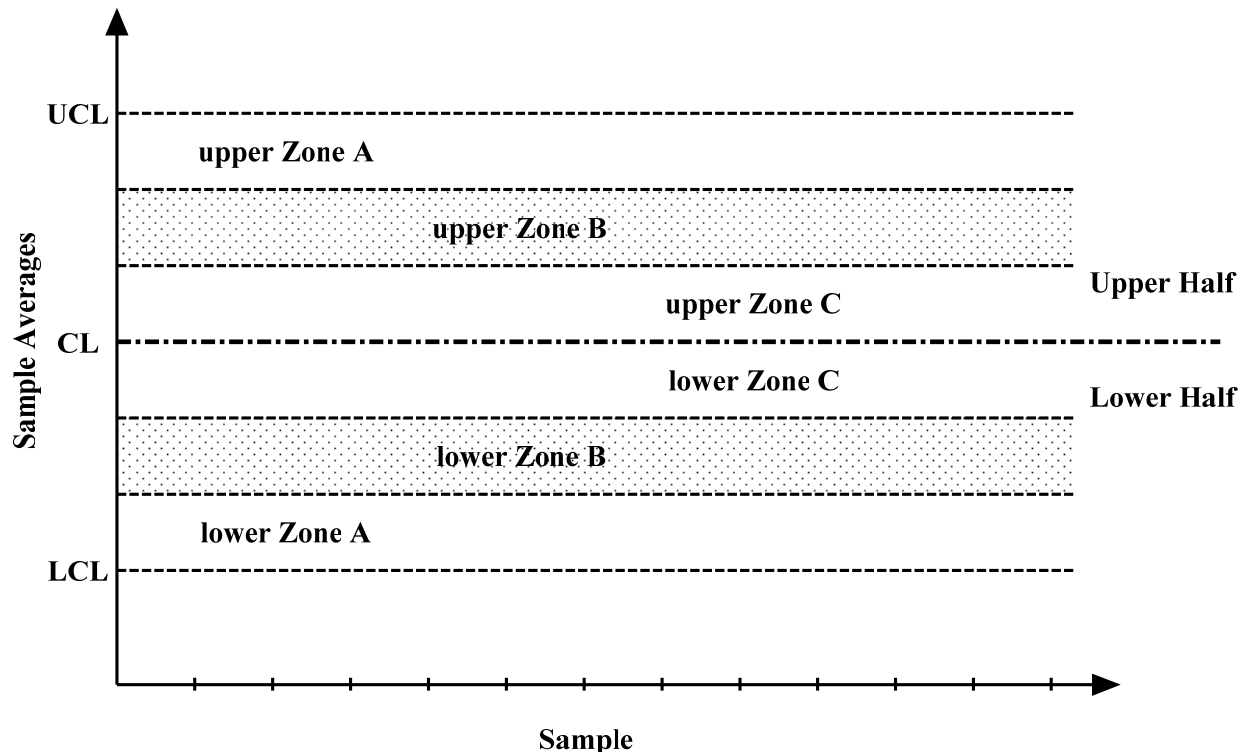
Figure 16.3.2: A general example of a $\bar{x}$-Chart

First, we need to find the average of the sample average lengths ($\bar{\bar{x}}$) and the estimated standard deviation (*s*) of the sample average lengths of the Boeing 787 fuselage. You can calculate $\bar{\bar{x}}$ and *s* by hand or you can utilize Excel and have the computer calculate the values for you. If this is the case, you would use the "=AVERAGE(" formula to calculate the average of averages and the "=STDEV(" formula to calculate the estimated standard deviation of the averages.

Q1.   Find (a) the average length for each sample ($\overline{x}$), (b) the average of sample average lengths ($\overline{\overline{x}}$), and (c) the estimated standard deviation (s) of the sample averages lengths. Record these values in Table 16.2.4 (see Worksheet 7).

Q2.   Is the average of averages the same value as the average of all the data combined (meaning taking the average of all five trials for all samples)?

Table 16.3.1 is an example of a class set of data collected under optimal conditions. The sample averages ($\overline{x}$) are calculated along with the average of the sample average lengths ($\overline{\overline{x}}$) and the estimated standard deviation (s).

Sample	Trial number 1	2	3	4	5	Average	Sample	Trial Number 1	2	3	4	5	Average
1	6	6	6	6	6.05	6.01	20	6	6	6.05	6.05	6.1	6.04
2	6	6	6	6.05	6.1	6.03	21	6	6.05	6	6.05	6.05	6.03
3	6	6.05	6	6.05	6	6.02	22	6.05	6	6.05	6.05	6	6.03
4	6.05	6.05	6.05	6.05	6	6.04	23	6.05	6	6	6.05	6.05	6.03
5	6.05	6.05	6.05	6	6	6.03	24	6	6.05	6.05	6.05	6.05	6.04
6	6.05	6.05	6.05	6	6	6.03	25	6	6	6.05	6.05	6.05	6.03
7	6.05	6.05	5.95	5.95	6	6	26	6	6.05	6.05	6.1	6.1	6.06
8	6	6	6	5.9	6.05	5.99	27	6	6.05	6.05	6.05	6	6.03
9	6	6	6	5.95	6.05	6	28	6.05	6	6	6	6.05	6.02
10	6.05	6.05	6.05	6	6	6.03	29	6	6.05	6.05	6	6	6.02
11	6.05	6.05	6	6	6	6.02	30	6	6	6	6	6	6
12	6	6	6	6	6.05	6.01	31	6	6.05	6.05	6	6.05	6.03
13	6	6	6	6.05	6.05	6.02	32	6.05	6.05	6.05	6.05	6.05	6.05
14	6.05	6.05	6	6	6	6.02	33	6.05	6.05	6.05	6.05	6.05	6.05
15	6.05	6.05	6.05	6.1	6.15	6.08	34	6	6.05	6.1	6.05	6	6.04
16	6	6	6	6	5.95	5.99	35	6.05	6	6.05	6.05	6.05	6.04
17	6	6	6	6	5.95	5.99	36	6.05	6	6.05	6.1	6.05	6.05
18	6	6	6.05	6.05	6.1	6.04	Average of Averages, $\overline{\overline{x}}$						6.02667
19	6	6	6	6.05	6.05	6.02	Estimated Standard Deviation, s						0.01986

Table 16.3.1: Example pooled class data under optimal conditions for complete Boeing 787 fuselages (in inches) with sample length averages

With the average of averages ($\overline{\overline{x}}$) and the estimated standard deviation (s) of data collected under the best possible conditions, we are able to calculate the necessary values to establish the standards for natural variability between samples.

The Center Line (CL) is determined by the average of sample averages ($\overline{\overline{x}}$); specifically, the Center Line is the horizontal line $y = \overline{\overline{x}}$. The Upper Control Limit (UCL) and the Lower Control Limit (LCL) are

three standard deviations away from the Center Line (CL), meaning the UCL is the horizontal line $y = \bar{\bar{x}} + 3s$ and the LCL is the horizontal line $y = \bar{\bar{x}} - 3s$.

In creating a $\bar{x}$ -Chart, the Upper and Lower Control Limits are nearly always set to three estimated standard deviations of the averages (3s). Why? Recall from Normal Distribution chapter (Chapter 15) that for a normal distribution, a bin with a lower bin value of -3s and an upper bin value of +3s will include 99.7% of all the values. This means that for a process with only natural variation, sample averages have a 0.997 chance of being between the UCL and LCL with a probability of 0.997. In other words, we would anticipate seeing sample average lengths with natural variability (i.e. a process in statistical control) between the UCL and LCL 99.7% of the time. With known variation within defined limits (Upper Control Limit and Lower Control Limit), Bob and Molly can engineer an assembly process that allows them to manage the imprecision of putting together Boeing 787 fuselage sections.

- Why can we assume that the sample average fuselage lengths are normally distributed?

  Four powerful words - the Central Limit Theorem. The Central Limit Theorem states if we use averages of the trails for our data analysis, then the sample averages follow a normal curve distribution.

  For the $\bar{x}$ -Chart, we plot data points that are averages of sets of five trials. Because our data points are average values, we can invoke the Central Limit Theorem when analyzing our data.

- Why did Bob and Molly choose to look at five trials (i.e. completed fuselages) at a time? How many trials should be in each sample?

  The minimum number of trials needed in each sample is dependent upon the practical situation in which the Central Limit Theorem is used. Since we will use it to manage imprecision in a very specific way when using quality control charts, we know the practical situation and then, the minimum number of trials. When the Central Limit Theorem is applied to quality control, *the minimum number of trials needed in a sample is five*.

Figure 16.3.2 illustrates the Center Line ($\bar{\bar{x}}$), the Upper Control Limit ($\bar{\bar{x}} + 3s$) and the Lower Control Limit ($\bar{\bar{x}} - 3s$). So how are the six zones formed? By calculating $\bar{\bar{x}} \pm 1s$ and $\bar{\bar{x}} \pm 2s$.
In this way,

       upper <u>Zone A</u> is the area from $\bar{\bar{x}} + 2s$ to $\bar{\bar{x}} + 3s$.
       upper <u>Zone B</u> is the area from $\bar{\bar{x}} + 1s$ to $\bar{\bar{x}} + 2s$.
       upper <u>Zone C</u> is the area from $\bar{\bar{x}}$ to $\bar{\bar{x}} + 1s$.
       lower <u>Zone C</u> is the area from $\bar{\bar{x}}$ to $\bar{\bar{x}} - 1s$.
       lower <u>Zone B</u> is the area from $\bar{\bar{x}} - 1s$ to $\bar{\bar{x}} - 2s$.
       lower <u>Zone A</u> is the area from to $\bar{\bar{x}} - 3s$.

With this information, we can add more information to Figure 16.3.2. Figure 16.3.3 is a more detailed illustration of a general a $\bar{x}$ -Chart
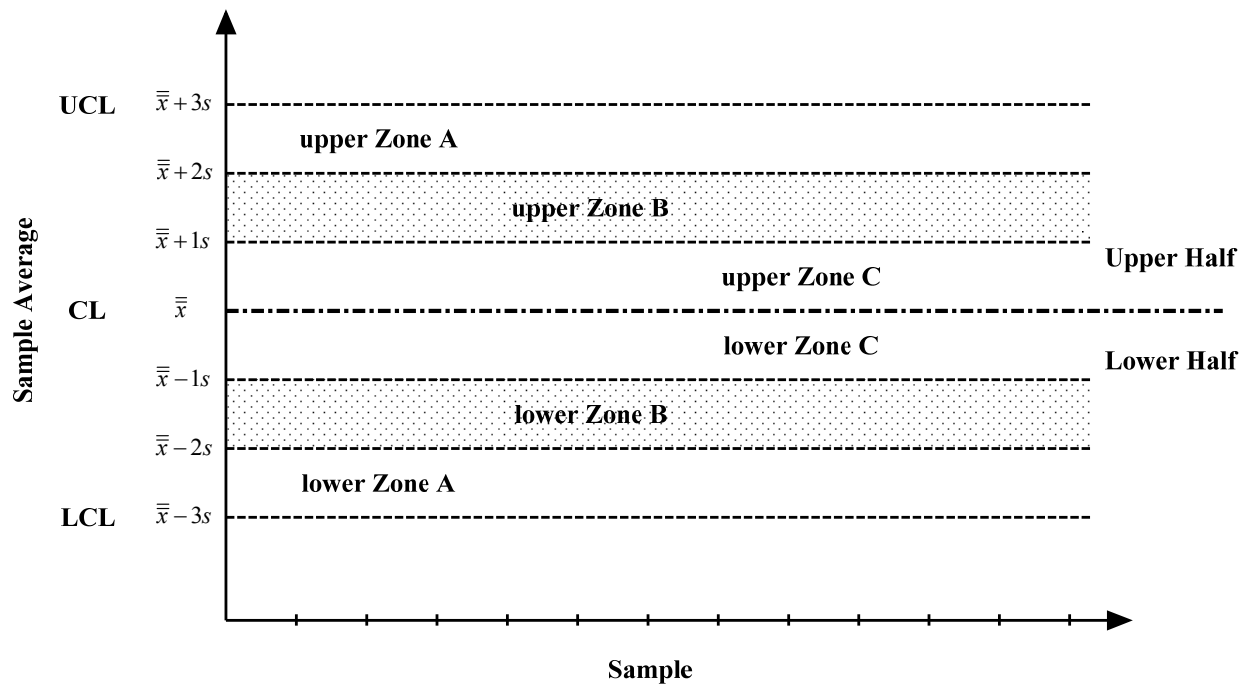
Figure 16.3.3: A general example of a $\overline{x}$-Chart with the formulas for values

As stated above, the sample average lengths follow a Normal Distribution function. Thus, we know that the probability that a data point will be between the Upper Control Limit (UCL) and the Lower Control Limit (LCL) is 0.997.

Q3.   What is the probability that a data point will be in Zone C (meaning both upper and lower Zone C)?

Q4.   What is the probability that a data point will be in Zone B?

Q5.   What is the probability that a data point will be in Zone A?

Q6.   Look at Figure 16.2.7: Summary of tests for unnatural patterns in a general example of a $\overline{x}$ – Chart. Notice that the probability of a single data point being outside upper Zone A has natural variability is 0.00135. How is this value calculated?

Let's apply this knowledge to the example pooled class data collective under optimal conditions

Table 16.3.1 is an example of a class set of data collected under optimal conditions. We already have a good idea of what the values are for the Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL) as well as $\overline{\overline{x}} \pm 1s$ and $\overline{\overline{x}} \pm 2s$ from Figure 16.3.3. Recall this $\overline{x}$ –Chart was given earlier with very little explanation as to why these values set standards to measure natural variability between sample averages.

x-Bar Chart for Completed Fuselage Lengths
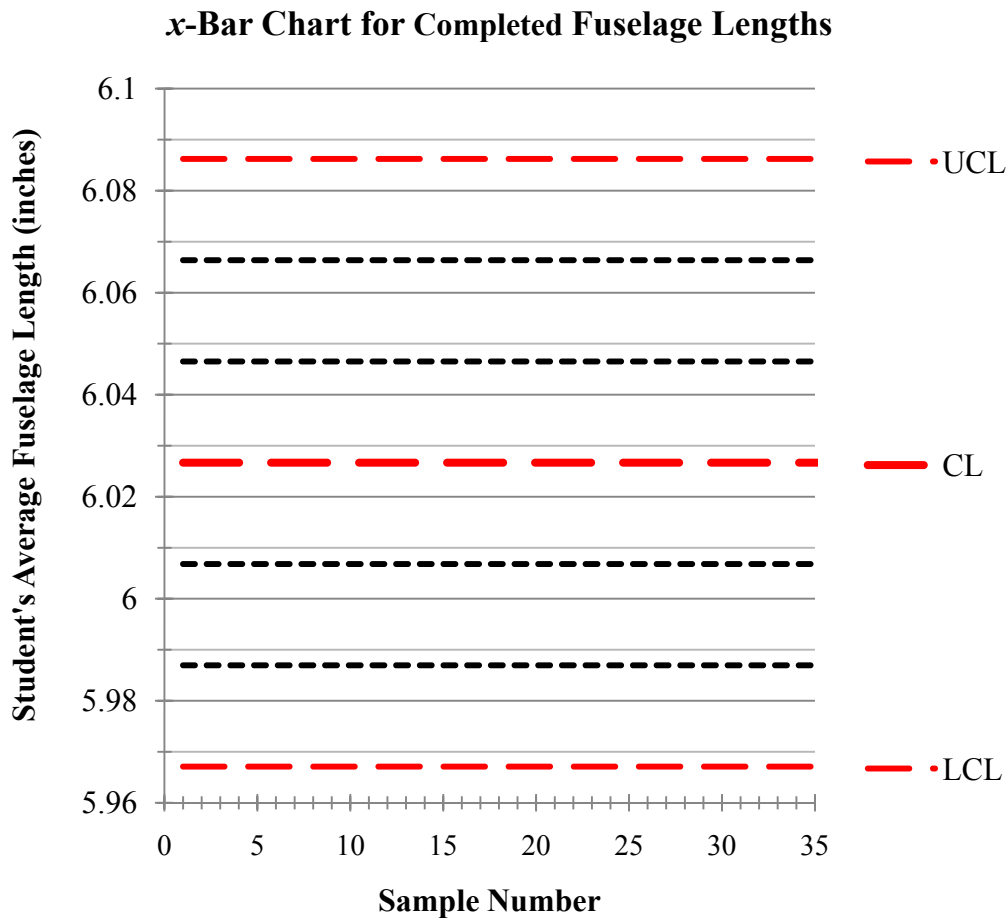
Figure 16.3.4: The $\overline{x}$ -Chart for the average sample lengths of completed fuselages from the Boeing 787 model activity

We can now calculate the values of the horizontal lines.

Q7.   Using the data from Table 16.3.1, calculate the exact Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL) as well as $\overline{\overline{x}} \pm 1s$ and $\overline{\overline{x}} \pm 2s$ by hand.

To summarize, Bob and Molly create their $\overline{x}$ -Chart based upon data collected under optimal conditions. These conditions establish the natural variation in the fuselage assembly process. By taking the average of the trial lengths for each sample and invoking the Central Limit Theorem and the Normal Distribution, Bob and Molly are able to calculate the Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL) as well as $\overline{\overline{x}} \pm 1s$ and $\overline{\overline{x}} \pm 2s$. With the six zones established, Bob and Molly can now use their chart to gauge the assembly process of the Boeing 787 fuselage under normal manufacturing conditions.

16.3.2 Creating an *R*-Chart for the Assembly of Boeing 787 Fuselages

Like $\overline{x}$ -Charts, *R*-Charts have an Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL). However, because the *R*-Chart is made up of the range of the sample and not the average of the sample, the UCL, CL, and LCL take on different values through different calculation processes.

Below is a general *R*-Chart for your reference.
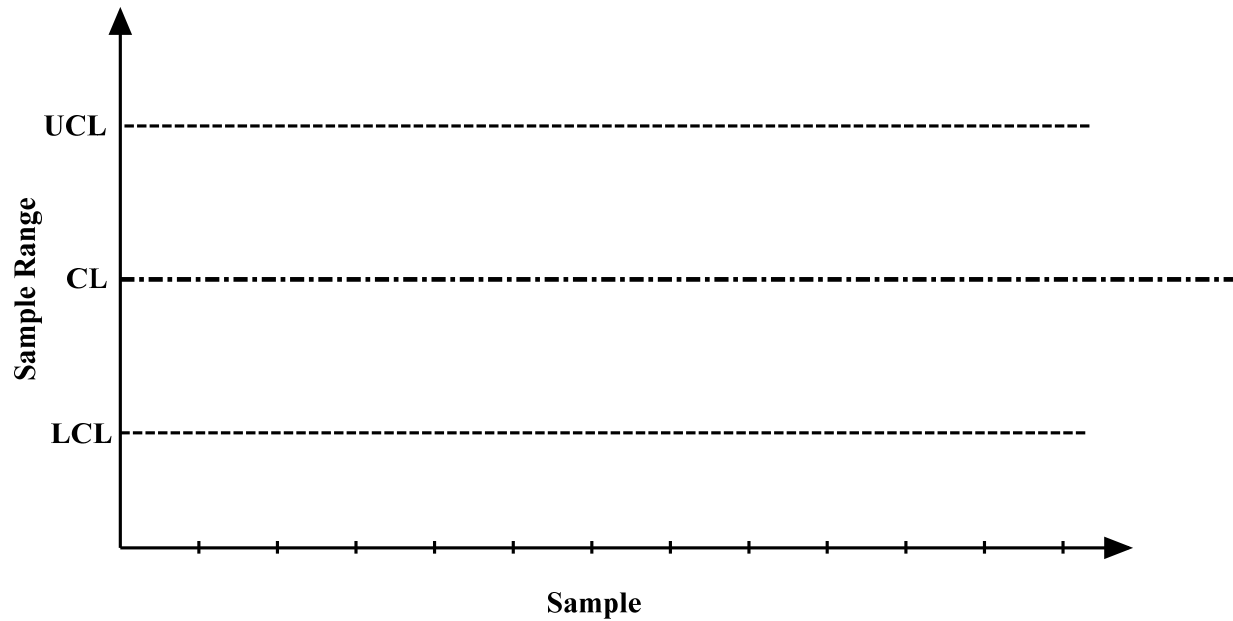
Figure 16.3.5: A general example of a *R*-Chart

First, we need to find the average of the ranges of the sample lengths ($\bar{R}$). You can calculate $\bar{R}$ by hand or you can utilize Excel and have the computer calculate the values for you. If this is the case, you would use the "=AVERAGE(" formula to calculate the average of the ranges.

Q8.   Find (a) the range for each complete fuselage sample (*R*), and (b) the average of the ranges of the sample lengths ($\bar{R}$). Record these values in Table 16.2.4 (see Worksheet 7).

Table 16.3.2 is an example of a class set of data collected under optimal conditions. The sample ranges (*R*) are calculated along with the average of the ranges of the sample lengths ($\bar{R}$).

| | **Trial number** | | | | | | | **Trial Number** | | | | | |
Sample	1	2	3	4	5	Range	Sample	1	2	3	4	5	Range
1	6	6	6	6	6.05	0.05	20	6	6	6.05	6.05	6.1	0.1
2	6	6	6	6.05	6.1	0.1	21	6	6.05	6	6.05	6.05	0.05
3	6	6.05	6	6.05	6	0.05	22	6.05	6	6.05	6.05	6	0.05
4	6.05	6.05	6.05	6.05	6	0.05	23	6.05	6	6	6.05	6.05	0.05
5	6.05	6.05	6.05	6	6	0.05	24	6	6.05	6.05	6.05	6.05	0.05
6	6.05	6.05	6.05	6	6	0.05	25	6	6	6.05	6.05	6.05	0.05
7	6.05	6.05	5.95	5.95	6	0.1	26	6	6.05	6.05	6.1	6.1	0.1
8	6	6	6	5.9	6.05	0.15	27	6	6.05	6.05	6.05	6	0.05
9	6	6	6	5.95	6.05	0.1	28	6.05	6	6	6	6.05	0.05
10	6.05	6.05	6.05	6	6	0.05	29	6	6.05	6.05	6	6	0.05
11	6.05	6.05	6	6	6	0.05	30	6	6	6	6	6	0
12	6	6	6	6	6.05	0.05	31	6	6.05	6.05	6	6.05	0.05
13	6	6	6	6.05	6.05	0.05	32	6.05	6.05	6.05	6.05	6.05	0
14	6.05	6.05	6	6	6	0.05	33	6.05	6.05	6.05	6.05	6.05	0
15	6.05	6.05	6.05	6.1	6.15	0.1	34	6	6.05	6.1	6.05	6	0.1
16	6	6	6	6	5.95	0.05	35	6.05	6	6.05	6.05	6.05	0.05
17	6	6	6	6	5.95	0.05	36	6.05	6	6.05	6.1	6.05	0.1
18	6	6	6.05	6.05	6.1	0.1	Average of Ranges, $\overline{R}$						0.06111
19	6	6	6	6.05	6.05	0.05							

Table 16.3.2: Example pooled class data under optimal conditions for complete Boeing 787 fuselages (in inches) with sample length ranges

With the average of the ranges of the sample lengths ($\overline{R}$) from data collected under the best possible conditions, we are able to calculate the necessary values to establish the standards for natural variability within samples.

Similar to $\overline{x}$-Charts, the Center Line (CL) of a R-Chart is determine by an average. However, instead of the CL being the average of averages ($\overline{\overline{x}}$), the CL of an R-Chart is the average of the ranges of the sample lengths ($\overline{R}$). Thus, the Center Line is the horizontal line $y = \overline{R}$.

Unlike $\overline{x}$-Charts, the Upper Control Limit (UCL) and the Lower Control Limit (LCL) are not as easy to calculate because the data points on an R-Chart are not normally distributed.

Q9.   Why are the data points on an R-Chart not normally distributed?

Because the distribution of the range (R) is not normally distributed, it is difficult to compute directly the Upper Control Limit (UCL) and Lower Control Limit (LCL). This difficult is cause by the fact that the type of distribution used to estimate the standard deviation changes with the number of trials within each sample. This means that the distribution for samples with 2 trials is different than the distribution for samples with 5 trials and different still than the distribution for samples with 10 trials. To make the R-

chart accessible to industry people, statisticians and engineers have developed table look-up methods to closely approximate the Upper and Lower Control Limits.

To obtain the Upper Control Limit (UCL) and the Lower Control Limit (LCL), we first compute the average range ($\bar{R}$) and multiply that value by the factor D1 for the Lower Control Limit (LCL) and a factor D2 for the Upper Control Limit (UCL).

$$\text{LCL} = \bar{R} \times \text{D1}$$
$$\text{UCL} = \bar{R} \times \text{D2}$$

You can find the values for D1 and D2 in Table 16.3.3.

Number of Trials	D1 (for LCL)	D2 (for UCL)
1	no range because you need a minimum of 2 trials	
2	0.000	3.267
3	0.000	2.574
4	0.000	2.282
5	0.000	2.115
6	0.000	2.004
7	0.076	1.924
8	0.136	1.864
9	0.184	1.816
10	0.223	1.777
11	0.256	1.744
12	0.283	1.717
13	0.307	1.693
14	0.328	1.672

Table 16.3.3: The LCL and UCL factors for *R*-Charts

With these factors (D1 and D2), the probability is approximately 0.98 or 98% that a single range value (*R*-value) is within the Lower Control Limit (LCL) and the Upper Control Limit (UCL) for a process in statistical control. Conversely, there is an approximate of probability of 0.01 that a single range value (*R*-value) is outside either the Lower Control Limit (LCL) or the Upper Control Limit (UCL) for a process in statistical control.

With this information, we can add more information to Figure 16.2.4. Figure 16.3.6 is a more detailed illustration of a general a *R*-Chart
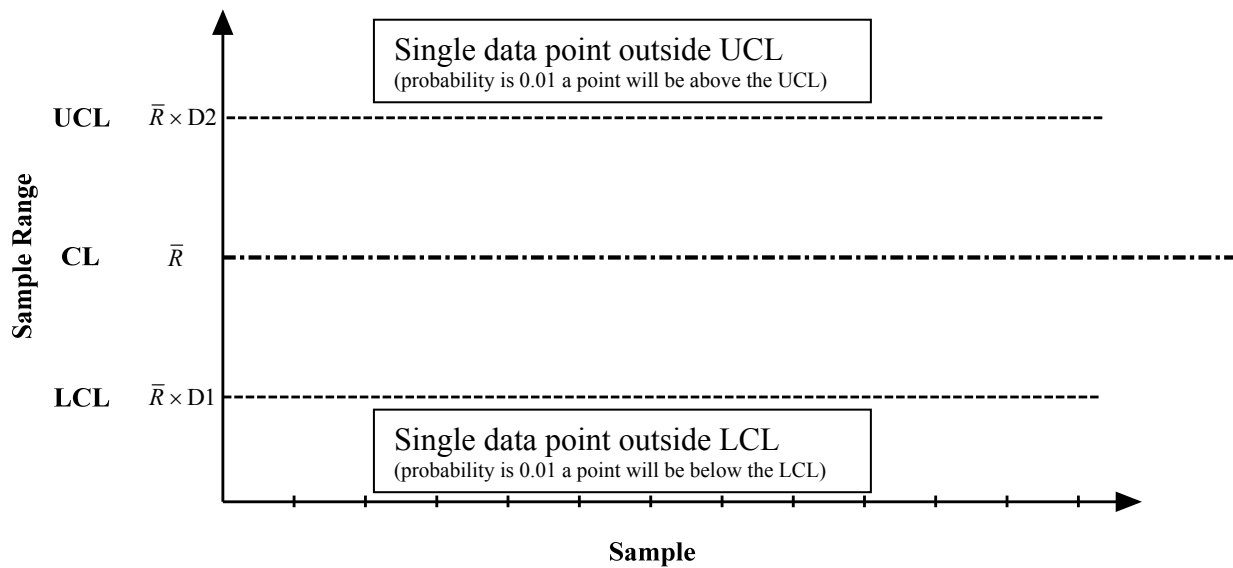
Figure 16.3.6: A general example of a *R*-Chart with the formulas for values

Let's apply this knowledge to the example pooled class data collective under optimal conditions

Table 16.3.1 is an example of a class set of data collected under optimal conditions. We already have a good idea of what the values are for the Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL) from Figure 16.3.7. Recall this *R*-Chart was given earlier with very little explanation as to why these values set standards to measure natural variability within the samples.
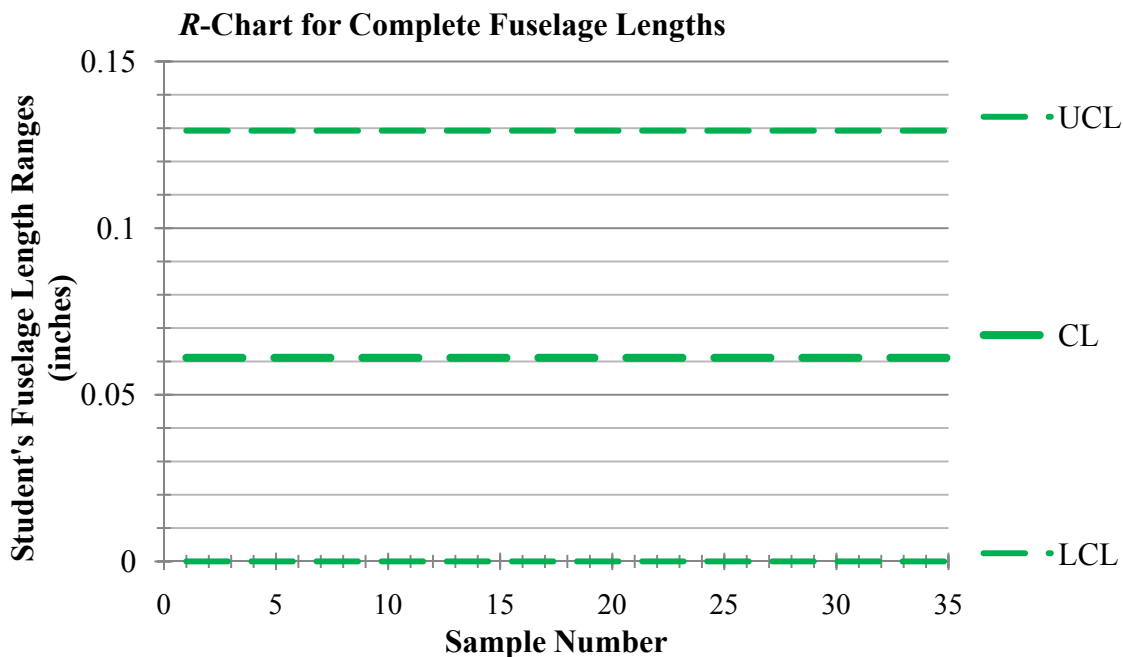
Figure 16.3.7: The *R*–Chart for the sample range of lengths of completed fuselages from the Boeing 787 model activity

We can now calculate the values of the horizontal lines.

Q10.   Using the data from Table 16.3.2 and the UCL and LCL factors in Table 16.3.3, calculate the exact Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL) by hand.

Q11.   What is the probability of being below the LCL if the LCL is zero?

To summarize, Bob and Molly create their *R*-Chart based upon data collected under optimal conditions. These conditions establish what can be considered natural variation in the fuselage assembly process. By taking the range of the trial lengths for each sample and utilizing Table 16.3.3, Bob and Molly are able to calculate the Upper Control Limit (UCL), the Center Line (CL), and the Lower Control Limit (LCL). With these standards established, Bob and Molly can now use their chart to gauge the assembly process of the Boeing 787 fuselage under normal manufacturing conditions.

Quality Control techniques like the one you studied in this section are used to manage imprecision in almost every industry in the world (remember the jet, the chicken breast, and the cellphone?). Because we can tell when a system begins to exhibit unnatural variation, we know when it goes "out of control" and can takes steps to correct whatever is causing the unnatural variation. It is this ability to control a process even though it has randomness that allows industry to make reliable and quality produces for us to use and consume.

Summary
Quality Control techniques like the one you studied in this chapter are used to manage imprecision in almost every industry in the world (remember the jet, the chicken breast, and the cellphone?). Because we can tell when a system begins to exhibit unnatural variation, we know when it goes "out of control" and can takes steps to correct whatever is causing the unnatural variation. It is this ability to control a process even though it has randomness that allows industry to make reliable and quality produces for us to use and consume.

Section 16.4: Chapter 16 (Quality Control) Homework Questions

1. The first step in the making steel is smelting iron ore in a furnace with coal and air.   A special type of coal, known as coke is used during the steel making process. Coke is produced by heating coal in the absence of air to eliminate many of the impurities in the coal.   The bulk density of coke is a critical parameter that will ultimately affect the fuel efficiency and productivity of the steel making process. The bulk density of the coke can change depending on how the coal is handled and the moisture content of the coal. To test the bulk density of coke, the coke in place in a standard sized container and weighed. The bulk density is equal to the sample mass divided by the volume that the coke occupies.   The Iron Harbor steel facility has determined the follow from when the process was in control:

x-bar chart	
$\bar{\bar{x}}$	48.7 lb/cu ft
UCL	48.8 lb/cu ft
LCL	48.5/cu ft

R-chart	
R	0.3 lb/cu ft
UCL	0.7 lb/cu ft
LCL	0 lb/cu ft

The follow data was collected over the past sixteen days. (see Excel spreadsheet as well)

Date	sample number	bulk density observations				
		1	2	3	4	5
1-Dec	1	48.7	48.9	49.0	48.7	48.7
1-Dec	2	48.7	48.9	48.8	48.7	48.7
2-Dec	3	48.6	48.6	48.6	48.7	48.7
2-Dec	4	48.7	48.9	48.8	48.8	48.7
3-Dec	5	48.5	48.5	48.5	48.7	48.6
3-Dec	6	48.9	48.7	48.8	48.7	48.5
4-Dec	7	48.5	48.7	48.6	48.6	48.7
4-Dec	8	48.7	48.6	48.7	48.9	49.0
5-Dec	9	48.5	48.6	48.4	48.5	48.9
5-Dec	10	48.6	48.2	48.6	48.4	48.7
6-Dec	11	48.7	48.2	48.4	48.8	48.6
6-Dec	12	48.7	48.8	48.6	48.6	48.5
7-Dec	13	48.7	48.8	48.8	48.9	48.8
7-Dec	14	48.7	48.6	48.5	48.6	48.7
8-Dec	15	48.9	48.4	48.9	48.8	48.9
8-Dec	16	48.7	48.8	48.5	48.8	48.8
9-Dec	17	48.5	48.6	48.8	48.6	48.6
9-Dec	18	48.5	48.8	48.6	48.5	48.6

10-Dec	19	48.6	48.5	48.6	48.5	48.6
10-Dec	20	48.6	48.6	48.5	48.7	48.7
11-Dec	21	48.6	48.8	48.5	48.6	48.2
12-Dec	22	48.1	48.5	48.5	48.6	48.6
12-Dec	23	48.5	48.6	48.8	48.9	48.5
13-Dec	24	48.7	48.6	48.6	48.6	48.7
13-Dec	25	48.8	48.7	48.8	48.7	49.0
14-Dec	26	48.8	48.9	48.7	48.7	48.6
14-Dec	27	48.5	48.6	48.6	48.7	48.7
15-Dec	28	48.7	48.7	48.6	48.6	48.4
15-Dec	29	48.6	48.6	48.5	48.8	48.9
16-Dec	30	48.8	48.7	48.7	48.9	48.9
16-Dec	31	48.8	48.8	48.4	48.4	48.6

a. How many samples are collected in one day?

b. What is the average and range for the first reading on December 7th?

c. What is the range of all the samples taken?

d. What day has the minimum sample density and what is the density?

e. What is the maximum sample density and what day and sample does it occur?

f. Do the maximum and minimum density measurements have any effect on the sample mean? Explain.

2. Refer to the data from Iron Harbor in Problem 1.
   a.   The upper and lower control limits are based on 3 standard deviations from the mean, what is the probability of a sample value of the average bulk density will fall between the UCL and the LCL?

   b.   Use Excel to create the x-bar and R charts.

   c.   Plot the $\bar{\bar{x}}$ and R data for Dec 1 – Dec 16 on your charts create in part a.

   d.   Do you notice any days the process was not in control? How do you know the process is not in control? Did this show on the x-bar or the R-chart?

   e.   How do these control charts help employees monitor the process?

3, This problem is based on the article by Ted Miller and Burton Balch, "Statistical Process Control in Food Processing," ISA Transactions, vol. 20, iss. 1, p 35-37. (available in the teacher notes)
J.M. Smucker plant in Lexington, KY produces Jif™ peanut butter (Balch, 1991). A critical variable is the peanut butter process is the salt content; if the salt content is too large or too small the peanut butter will not taste good. The plant has been using SPC to monitor salt content since the late 1980's. Below is a table of the salt content measurements taken when the process was believed to be in control.

Day	salt content observations				
	1	2	3	4	5
6/1	110.731	112.777	112.039	112.531	112.706
6/2	112.260	112.818	111.813	112.280	112.777
6/3	112.675	112.519	111.914	111.767	112.039
6/4	111.068	112.254	111.905	112.637	112.531
6/5	112.237	112.840	111.805	112.307	112.706
6/6	112.696	112.326	112.491	112.175	113.071
6/7	112.373	112.131	112.251	112.008	112.507
6/8	112.363	112.633	112.487	112.829	112.248
6/9	111.986	111.975	112.678	112.663	112.694
6/10	111.749	112.274	112.036	112.532	113.531
6/11	112.100	111.722	111.713	111.731	111.695
6/12	111.690	112.621	112.641	112.650	112.600
6/13	112.518	112.055	111.430	112.477	112.090
6/14	111.247	111.157	112.229	112.217	112.242
6/15	111.988	113.240	112.404	111.944	111.087
6/16	112.330	112.255	111.474	111.115	111.640
6/17	112.259	112.870	112.504	111.534	112.028
6/18	111.384	112.926	112.186	112.027	112.344
6/19	111.959	111.785	111.877	111.676	112.035
6/20	111.966	112.797	112.329	111.715	111.310
6/21	111.285	111.990	112.441	113.337	112.050

6/22	112.747	112.896	112.061	112.132	111.988
6/23	112.192	112.298	111.297	111.610	113.032
6/24	112.388	111.812	112.115	110.900	112.742
6/25	112.002	112.143	112.841	111.015	112.606
6/26	112.061	112.485	111.391	112.044	112.503
6/27	112.178	111.746	112.682	112.240	111.643
6/28	111.496	111.477	112.297	112.290	112.304
6/29	112.221	112.860	111.800	112.329	112.661
6/30	112.167	111.829	112.721	112.336	111.546

a.   How often were these samples taken?
b.   What is the average and range for June 12th?
c.   All sample averages fall between what two numbers?
d.   Plot the data in Excel
e.   What would be the $\bar{\bar{x}}$ and the s value for the salt content data?
f.   Graph the data with the Center Line (CL) and the LCL and UCL line
g.   Create a histogram of the sample averages in Excel using bin sizes equal using 6 zones, with each bin equal to 1s. What is the value of the zone lines between upper Zone A and upper Zone B?
h.   Based on your histogram created in (g) what is the how many of the averages fall into the upper and lower Zone C? What percentage is this? What would you expect?

4. New salt content data from the Lexington Jif™ plant (see Problem 3) is given in Excel spreadsheet and shown below.

Day	salt content observations				
	1	2	3	4	5
15-Jul	110.742	111.786	112.900	112.176	111.700
16-Jul	111.494	112.668	112.136	111.997	112.636
17-Jul	111.685	112.753	111.525	111.315	112.448
18-Jul	112.407	111.749	111.734	112.108	112.546
19-Jul	112.507	112.635	112.086	111.480	111.842
20-Jul	111.444	112.591	111.514	111.629	112.488
21-Jul	112.358	112.809	112.297	111.903	112.557
22-Jul	112.081	111.551	112.814	111.811	112.433
23-Jul	112.536	111.953	112.110	112.667	112.092
24-Jul	112.481	112.868	112.413	111.139	112.397
25-Jul	112.217	112.295	111.579	112.966	113.011
26-Jul	112.297	112.486	112.446	112.441	111.784
27-Jul	112.789	112.072	113.044	111.721	112.536
28-Jul	111.831	112.985	112.056	112.424	112.030
29-Jul	112.774	111.997	111.770	112.439	111.646
30-Jul	112.045	111.978	112.222	111.427	112.848
31-Jul	109.378	109.878	110.335	111.452	112.070

1-Aug	112.237	113.177	113.558	113.228	112.847
2-Aug	113.136	112.427	111.442	111.948	111.374
3-Aug	112.645	112.732	111.942	113.398	111.466
4-Aug	111.774	112.678	111.502	112.665	112.576
5-Aug	112.820	112.355	112.478	112.228	112.358
6-Aug	112.788	112.103	112.055	112.390	111.637
7-Aug	111.791	112.931	112.735	112.126	111.594
8-Aug	112.488	112.807	112.049	112.221	112.806
9-Aug	111.456	111.914	113.015	112.564	111.661
10-Aug	111.827	112.555	111.045	112.208	112.084
11-Aug	112.413	111.593	111.951	111.239	112.414
12-Aug	112.493	111.588	112.048	112.113	112.404
13-Aug	112.143	111.924	112.359	111.399	111.567
14-Aug	112.245	111.829	112.192	111.789	112.135
15-Aug	112.069	112.310	111.803	111.955	112.735
16-Aug	112.787	112.608	112.532	112.094	112.181
17-Aug	112.236	112.472	112.279	112.031	112.478
18-Aug	111.892	112.521	112.431	113.085	112.739
19-Aug	111.921	111.280	110.897	111.481	112.338
20-Aug	111.714	112.007	112.550	112.275	112.695
21-Aug	111.732	112.171	112.381	112.120	111.744
22-Aug	111.471	113.236	111.630	112.487	112.109
23-Aug	112.395	111.213	112.965	111.583	112.036
24-Aug	111.830	111.912	111.350	111.525	111.745
25-Aug	111.457	112.822	112.217	111.790	112.147
26-Aug	111.869	111.564	112.433	112.317	112.048
27-Aug	112.245	112.475	111.746	112.272	111.838
28-Aug	112.515	111.958	112.590	112.882	113.143
29-Aug	112.583	112.933	112.522	111.436	112.428
30-Aug	111.236	111.560	112.891	112.342	112.686
31-Aug	110.029	110.166	110.528	111.195	112.086
1-Sep	112.089	113.448	113.541	113.168	112.364
2-Sep	112.504	113.055	112.532	112.312	111.183
3-Sep	112.274	112.647	112.655	112.749	112.980
4-Sep	112.085	112.077	111.854	112.450	112.338
5-Sep	111.454	112.324	111.696	112.891	111.744
6-Sep	112.458	112.837	112.212	111.628	111.669
7-Sep	111.998	112.896	112.060	112.532	112.572
8-Sep	112.423	112.538	112.280	111.968	112.962
9-Sep	112.194	112.815	111.646	112.428	111.378
10-Sep	113.298	111.978	112.039	113.020	112.607
11-Sep	112.965	112.700	112.245	112.150	112.167
12-Sep	111.665	111.944	112.979	111.909	111.726

13-Sep	111.933	112.106	111.966	112.884	112.137
14-Sep	112.957	112.270	112.894	111.941	112.635
15-Sep	112.316	110.921	113.194	112.070	112.485
16-Sep	112.571	112.668	112.527	110.972	112.774
17-Sep	112.414	112.205	112.419	112.433	113.490
18-Sep	112.932	111.838	111.872	111.562	111.683
19-Sep	112.625	111.922	112.164	112.312	113.169
20-Sep	112.876	112.082	112.507	112.353	111.026
21-Sep	111.858	111.495	111.713	111.044	111.593
22-Sep	112.265	112.221	112.472	112.291	111.574

a. All of the measurements are given to 3 decimal points. What does this tell you about the precision of the measuring equipment? How many decimal points should your chart, averages, UCL and LCL have? Explain.

b. Plot the new July 15th to September 22nd data on the control chart created in Problem 1 (f).

c. Do you notice any days that the process was out of control?

d. The peanut butter plant instituted a new cleaning procedure. At the end of every month, the plant had a scheduled shutdown during which the salt chute was thoroughly cleaned. When the process was restarted, some of the salt would build up in the chute. The salt was feeding at the correct rate, however, not all the salt was being transferred to the peanut butter, which led to low salt readings. The operator would then try to correct the process by increasing the salt feed rate. However, the salt build up would by then reach its steady state, and the salt readings would increase. In turn, the operator would lower the salt rate. Do you see evidence of this in your control chart? Explain. What suggestions would you make to help with this problem?

5. Shipping of medical devices requires the packaging to maintain its integrity throughout the inventory, shipping and delivery systems. A common test done by medical device manufacturers is seal strength of packaging components. This is a destructive test (product is destroyed during the testing process and cannot be re-used) in which the package is pulled apart on a test machine and the amount of force required to tear the package is measured. A medical device company manufactures and packages prosthetic heart valves. The packaging must keep the valve sterile, as well as protect it from handling issues in the distribution and storage system. The company has determined from past experience that an in control packaging process will have average seal strength $\bar{x}$ of 8.96 N with an s^2 value of 0.03 N.

a. Plot a x-bar control chart including a the center line, and upper and lower control limits of $\bar{\bar{x}} \pm 3s$.

If the seal is broken, the hospital cannot use the device. In addition, the FDA regulates seal strength of medical packaging and testing procedures are outline in ASTM F88/F88M Standard Test Method for Seal Strength of Flexible Barrier Materials. The company is very concerned about the seal strength and tests samples 6 times per day. The quality department would like to further divide the LCL and UCL into A,B and C zones.

b. Plot a control chart showing upper and lower Zones A, B, and C.

The company has collected data over the past 2 months.

c. Plot the data on the control chart created in part (a).

| | | | sample of packaging | | | |
Date	1	2	3	4	5	6
1-Feb	8.38	8.79	9.24	8.95	8.76	8.68
2-Feb	9.15	8.93	8.88	9.13	8.75	9.18
3-Feb	8.69	8.61	9.06	9.04	8.78	8.77
4-Feb	8.92	9.10	9.08	9.13	8.91	8.67
5-Feb	8.82	8.66	9.12	8.69	8.73	9.08
6-Feb	9.02	9.20	9.00	8.84	9.10	8.91
7-Feb	8.70	9.21	8.80	9.05	9.09	8.86
8-Feb	8.92	9.15	8.92	9.07	9.23	9.05
9-Feb	8.54	9.04	8.97	9.00	8.71	9.27
10-Feb	9.28	9.00	9.07	9.06	9.06	8.79
11-Feb	9.20	8.91	9.30	8.77	9.09	8.81
12-Feb	9.27	8.90	9.05	8.89	9.19	8.88
13-Feb	8.79	9.06	8.74	8.90	8.87	8.97
14-Feb	8.65	9.22	8.63	9.15	9.01	8.66
15-Feb	8.91	8.97	8.95	9.26	8.69	8.82
16-Feb	9.33	9.05	8.66	8.86	8.63	9.14
17-Feb	9.17	8.86	9.44	8.67	8.79	9.15
18-Feb	8.68	9.15	9.11	9.21	9.02	9.07
19-Feb	8.97	9.02	9.20	8.92	8.90	9.04
20-Feb	8.73	8.80	9.25	9.17	8.93	8.72
21-Feb	9.08	9.20	8.90	8.97	9.20	8.66
22-Feb	8.85	9.29	9.11	8.74	8.81	9.10
23-Feb	8.50	8.96	8.91	9.05	8.72	8.86
24-Feb	8.58	9.05	9.08	8.72	8.90	8.93
25-Feb	9.04	8.94	8.85	9.02	8.64	8.71
26-Feb	8.98	8.81	8.96	8.80	8.93	8.91
27-Feb	9.00	8.80	8.86	9.17	9.19	9.12
28-Feb	9.09	8.92	8.95	8.97	9.07	8.99
1-Mar	8.89	9.07	8.84	9.09	9.05	9.31
2-Mar	9.18	8.85	8.59	8.44	8.67	9.02
3-Mar	8.77	8.88	9.10	8.99	9.16	8.77
4-Mar	8.95	9.03	8.93	8.78	8.67	9.37
5-Mar	8.73	9.07	8.92	9.04	8.57	9.27
6-Mar	8.71	8.89	8.81	8.84	8.62	8.69
7-Mar	8.78	8.66	9.21	8.97	8.80	8.94
8-Mar	8.83	8.71	9.05	9.01	8.90	8.98
9-Mar	9.07	8.78	8.99	8.82	9.09	8.86
10-Mar	9.12	9.23	9.34	9.11	9.25	9.09
11-Mar	8.65	9.05	8.57	8.70	9.24	9.02
12-Mar	9.15	8.97	8.63	8.81	8.88	9.11
13-Mar	8.91	8.90	8.90	9.07	8.63	9.08

14-Mar	9.30	9.09	9.00	8.55	8.99	9.14
15-Mar	9.14	9.18	9.27	8.91	8.91	8.82
16-Mar	9.06	9.02	8.66	9.01	8.76	9.24
17-Mar	8.78	9.06	9.21	8.96	8.73	8.75
18-Mar	8.88	9.24	8.90	9.09	9.11	9.05
19-Mar	9.10	8.99	8.87	9.26	8.96	9.21
20-Mar	8.74	9.05	8.63	9.40	8.87	8.90
21-Mar	9.29	9.12	9.27	9.16	8.98	8.94
22-Mar	8.95	8.75	8.86	9.27	8.84	8.77
23-Mar	8.85	8.92	8.87	9.23	8.93	9.26
24-Mar	8.99	9.24	8.86	9.13	9.01	8.45
25-Mar	9.36	8.91	9.07	9.11	9.15	9.09
26-Mar	8.47	9.19	9.05	8.96	9.05	9.05
27-Mar	9.48	9.25	8.82	8.83	8.70	8.75
28-Mar	9.13	8.85	8.95	9.00	9.35	9.23
29-Mar	8.91	9.08	9.02	8.49	8.82	8.68
30-Mar	8.77	8.50	8.72	8.99	8.97	9.07
31-Mar	9.00	8.71	8.87	8.74	8.74	9.18

d. Based on your plot in part (c) have there been any unnatural patterns? Explain your answer based on the Tests for Unnatural Patterns (Western Electric, 1956).

e. The company is interested in testing a new material for the packaging. Before implementing, the company produced prototypes and tested the material in the same way as the current material. The table below shows the seal strength data of the new packaging material. Based on this information, do you think the company should change to the new material? Explain.

Day	Observation					
	1	2	3	4	5	6
1	7.53	8.37	9.26	8.68	8.3	8.14
2	9.07	8.65	8.54	9.05	8.29	9.14
3	8.16	7.99	8.9	8.87	8.34	8.33
4	8.63	8.98	8.95	9.05	8.61	8.12
5	8.41	8.1	9.01	8.15	8.24	8.93
6	8.83	9.19	8.78	8.46	8.99	8.6
7	8.18	9.19	8.39	8.89	8.97	8.5
8	8.63	9.07	8.61	8.93	9.23	8.87
9	7.85	8.86	8.71	8.78	8.2	9.31
10	9.35	8.78	8.93	8.9	8.89	8.37
11	9.17	8.6	9.38	8.32	8.97	8.4
12	9.33	8.59	8.88	8.56	9.16	8.54
13	8.36	8.89	8.26	8.58	8.52	8.72
14	8.08	9.22	8.04	9.08	8.81	8.1

15	8.6	8.73	8.68	9.31	8.16	8.42
16	9.45	8.88	8.09	8.5	8.04	9.06
17	9.13	8.49	9.66	8.11	8.36	9.08
18	8.14	9.07	9	9.2	8.82	8.92
19	8.72	8.83	9.17	8.62	8.58	8.85
20	8.25	8.37	9.28	9.13	8.64	8.22

  f. One of your employee's suggests increase the sample size from 6 to 20, and instead of using the R-chart look at the standard deviations. The testing is a destructive test, therefore tested product must be scrapped. It costs the company $834 to produce and package each heart valve. What would be the daily cost to increase the sample size from 6 to 20? Do think the increase in the accuracy of the measurement would be worth the increase in cost?

6. In the manufacture of cell phones, the circuit board provides the "brains" of the phone. They can process millions of calculations per second, switching FM channels as the phone moves between cell stations, compressing and decompressing the voice stream. The components of the circuit board must fit into precise locations on the board, and therefore the pin alignment is critical (see diagrams below). The manufacture of one of components has implemented a quality control program to monitor pin-to-pin distance, a critical parameter. While the assembly process was in control the follow data points were obtained.
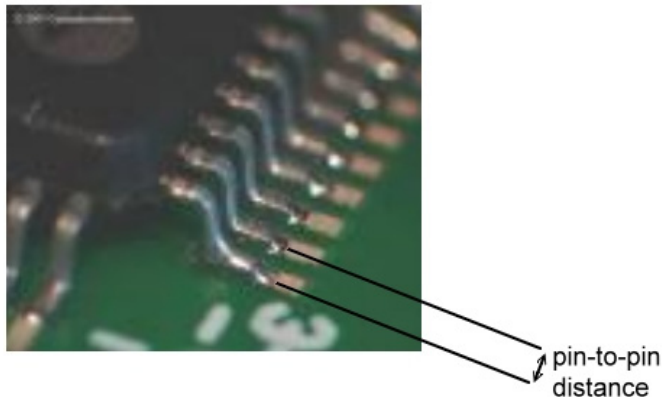
Example of a cell phone circuit board from *HowStuffWorks.com*. Component pins must line up with the circuit board during the assembly of the cell phone.

Parts of a cell phone: *http://electronics.howstuffworks.com/cell-phone6.htm*

The microprocessor on the circuit board in the cell phone. *http://electronics.howstuffworks.com/cell-phone6.htm*

Close up view of the surface mount pins on the circuit board
http://hackedgadgets.com/2008/01/14/surface-mount-soldering-tutorial/

   a. Calculate the Upper Control Limit, Center Line and Lower Control Limit for the $\bar{x}$ chart and
       graph in Excel based on the data that was taken while the process was considered in control.

			pin-pin dist mm				
Day	Hour	Data pt	1	2	3	4	5
3-Oct	1	1	2.39	2.60	2.52	2.57	2.59
3-Oct	2	2	2.63	2.57	2.54	2.59	2.67
3-Oct	3	3	2.49	2.59	2.53	2.54	2.43
3-Oct	4	4	2.48	2.52	2.55	2.52	2.45
3-Oct	5	5	2.52	2.52	2.62	2.59	2.58
3-Oct	6	6	2.57	2.48	2.55	2.59	2.47
3-Oct	7	7	2.54	2.51	2.49	2.54	2.50
3-Oct	8	8	2.53	2.45	2.54	2.47	2.53
3-Oct	9	9	2.58	2.58	2.52	2.53	2.59
3-Oct	10	10	2.50	2.59	2.52	2.53	2.63

4-Oct	1	11	2.63	2.56	2.54	2.52	2.47
4-Oct	2	12	2.55	2.59	2.55	2.61	2.58
4-Oct	3	13	2.49	2.57	2.48	2.49	2.53
4-Oct	4	14	2.55	2.53	2.63	2.51	2.53
4-Oct	5	15	2.54	2.44	2.59	2.65	2.59
4-Oct	6	16	2.53	2.62	2.53	2.50	2.52
4-Oct	7	17	2.63	2.55	2.50	2.49	2.59
4-Oct	8	18	2.51	2.53	2.55	2.57	2.57
4-Oct	9	19	2.47	2.52	2.58	2.52	2.56
4-Oct	10	20	2.55	2.53	2.53	2.48	2.63
5-Oct	1	21	2.54	2.55	2.55	2.55	2.52
5-Oct	2	22	2.57	2.52	2.52	2.57	2.54
5-Oct	3	23	2.53	2.58	2.53	2.54	2.57
5-Oct	4	24	2.56	2.54	2.50	2.53	2.51
5-Oct	5	25	2.53	2.54	2.53	2.56	2.52
5-Oct	6	26	2.53	2.55	2.55	2.53	2.55
5-Oct	7	27	2.52	2.57	2.53	2.58	2.54
5-Oct	8	28	2.56	2.55	2.56	2.52	2.53
5-Oct	9	29	2.52	2.53	2.51	2.53	2.52
5-Oct	10	30	2.55	2.56	2.53	2.54	2.52

b. Calculate the Upper Control Limit, Center Line and Lower Control Limit for the R-chart and graph in Excel from the data that was taken while the process was considered in control.   Are the Upper and Lower Control limit lines equal distance from the Center Line? Explain.

c. Data has been collected over the last week. Plot this data on your control charts from parts (a) and (b). Would you consider this process in control? Why or why not?

d. When the tooling begins to wear the pin-to-pin distance will begin to steadily decrease, can you recommend some guidelines for the operator as to when the tooling needs to be changed based on the quality control chart.

7. The textile industry in the United States has been faced with increasing competition from overseas. In order to remain competitive the US companies are relying on superior quality, innovative products and quick market response to customer demands. New technology in nanoscience, electrotextiles, nonwovens, medical textiles and geotextiles have increased the quality levels and production control techniques used in the textile industry (Clapp, 2001).   The textile industry for many years relied on inspection of final goods to achieve their quality requirement. However, this method resulted in defective product that could not be used, a waste of both raw material and labor. Inspection methods today include electronic monitoring of production process in order to make adjustments to the process before defective material is produced (Clapp, 2001). Synthetic material such as polyester, nylon and polypropylene can be textured to give the yarn the look and feel of natural fibers. During the texturing process, the yarn is fed at high speeds into a machine that simultaneously twists and draws the yarn. During this process the yarn

is heated to a critical temperature, if set too high the yarn will become brittle, and if set too low the yarn will not have the appropriate twist (Hearle, 2001).

Temperature data from ABS Yarn, Incorporated's yarn texturing area is given. The data was obtained when the processes was considered in control.

Yarn dye process						
Temperature Data (in F)						
Date	1	2	3	4	5	6
1-Aug	105.2	105.0	101.0	102.9	99.8	91.7
2-Aug	101.2	106.0	104.4	98.7	99.9	91.2
3-Aug	103.2	98.7	99.3	100.8	106.0	98.2
4-Aug	100.2	102.5	99.3	111.0	104.9	108.1
5-Aug	108.8	98.4	102.2	106.8	99.1	101.9
6-Aug	106.0	104.7	105.8	108.0	104.2	109.4
7-Aug	98.8	97.5	100.5	90.8	95.4	102.5
8-Aug	99.2	105.1	96.4	101.9	103.2	102.9
9-Aug	100.6	99.7	102.8	87.5	97.8	100.7
10-Aug	100.3	108.8	102.3	98.3	105.2	103.5
11-Aug	106.1	93.4	105.0	101.5	105.1	106.0
12-Aug	102.5	102.8	108.9	106.4	104.7	103.5
13-Aug	100.5	98.0	98.5	101.1	109.9	98.1
14-Aug	109.2	100.4	108.6	106.6	104.6	104.9
15-Aug	101.5	105.1	103.6	104.7	102.0	95.9
16-Aug	108.2	104.3	99.3	104.1	103.1	99.6
17-Aug	105.4	98.0	107.5	104.3	103.7	107.1
18-Aug	100.8	99.4	104.6	105.9	102.0	99.9
19-Aug	102.5	104.1	104.7	109.4	93.5	103.9
20-Aug	104.6	97.3	105.7	98.7	103.9	99.8
21-Aug	107.1	94.5	105.7	111.2	109.2	101.1
22-Aug	100.7	101.0	104.2	106.0	104.2	113.3
23-Aug	103.0	103.4	103.1	101.4	111.7	96.0
24-Aug	99.7	92.7	98.5	105.1	103.1	102.0
25-Aug	106.1	102.4	100.9	94.5	96.7	100.6
26-Aug	107.3	102.7	101.1	99.7	112.5	106.0
27-Aug	101.9	106.7	102.7	104.4	108.8	102.7
28-Aug	103.5	101.9	98.9	107.2	111.4	100.6
29-Aug	105.4	105.2	105.2	102.1	104.0	105.8
30-Aug	104.6	104.3	94.3	99.4	101.7	102.7

a. Create an x-bar and R chart for the yarn texturing temperature process from the data given showing the center line, upper control limit and lower control limit based on three standard deviations.

b. Data is given for the last week, plot this on the control charts created in (a).

c.   It was found that operators often over compensated when the processes was not within the control limits. In order to refine the temperature adjustment, it was determined that incorporating zones on the control process would allow operators to monitor and adjust the process with more precision and control. Add the three zones (A,B,C) to x-bar chart created in (a).

d.   Temperature data is given for the next week. Plot this on the x-bar chart created in (c) and the R-chart created in (a). Based on this data, do you think the zone procedures improved the process?

8. In the summer of 2002, the USDA completed a lot by lot final inspection of 116 million pounds of ground beef delivered by 17 different meat companies (Higgins, 2003). Results were disappointing, and a bit scary.   About 8.6 million pounds of the beef failed microbial tests for coliforms, salmonella, and other pathogens (Higgins, 2003). One million pounds of beef was found to be contaminated with E. coli. These results led to a dramatic shift from a final pass/fail inspection of the product to a process based control system (Higgins, 2003). The USDA as well as the UK Food Safety Regulations now mandates Hazard Analysis Critical Control Points (HACCP) in the seafood, meat, poultry and juice food industries (Morris, 2003). HACCP requires these food manufacturers to identify, control and monitor potential food hazards. They are required to identify critical control points in the production process where contamination could occur, establish targets and control limits, and monitor the process (Morris, 2003). Not only are these regulations mandatory, but food recalls can significantly impact a company, not only the direct financial cost of recovery, disposal and liability claim, but also the long term cost of the negative publicity.

Food borne bacteria can grow quickly if left at room temperature. Refrigeration temperatures are critical. At temperatures above 40°F some bacteria can double their numbers in 20 minutes and in 2 hours there will be sufficient numbers to cause illness if ingested (USDA Food Safety and Inspection Service, 1999).   CoolCargo, Inc. is a refrigerated trucking company that transports ground beef to area schools. It is imperative that the truck remains below the critical 40°F point. They have installed five temperature probes in various locations around the cargo area. The system monitors the temperature hourly and adjusts the temperature based on the average reading. The CoolCargo would like to maintain the temperature at an average of 37.5°F with upper and lower control limits of ±2.2°F. They would like the range to average 1.5°F, with an upper limit of 2.5 °F. They have collected reading from the five temperature probes 3 times per day, shown below.

Date	reading	data pt	Temperatures				
			1	2	3	4	5
1-Sep	1	1	37.8	37.2	38.3	35.3	38.0
1-Sep	2	2	36.4	39.1	38.0	39.4	35.0
1-Sep	3	3	41.5	35.8	34.1	40.3	40.0
2-Sep	1	4	36.6	35.3	37.7	35.3	37.9
2-Sep	2	5	35.6	39.1	35.6	44.3	36.3
2-Sep	3	6	39.0	39.0	40.0	35.0	37.0
3-Sep	1	7	34.8	35.9	35.8	40.2	39.5
3-Sep	2	8	34.4	41.7	38.9	32.3	41.7
3-Sep	3	9	36.1	36.2	42.4	33.4	38.1
4-Sep	1	10	39.9	34.6	41.1	39.0	38.6

4-Sep	2	11	35.3	35.4	38.6	36.9	35.0
4-Sep	3	12	38.4	43.1	38.5	34.7	38.9
5-Sep	1	13	40.1	36.8	38.1	37.5	32.6
5-Sep	2	14	38.1	35.2	41.4	33.1	37.0
5-Sep	3	15	40.9	37.4	33.4	33.8	34.3
8-Sep	1	16	35.1	34.2	34.2	40.8	34.2
8-Sep	2	17	38.0	32.7	39.8	36.3	37.0
8-Sep	3	18	34.4	32.0	37.7	37.1	40.4
9-Sep	1	19	37.9	41.3	38.9	36.8	31.4
9-Sep	2	20	38.3	37.6	39.3	38.4	42.0
9-Sep	3	21	36.3	35.8	34.7	41.0	45.0

    a.   Create an x-bar chart with the centerline, UCL_x, LCL_x at 3-standard deviations, and plot the data on the chart. Do you notice any points out of control? Based on your x-bar chart can CoolCargo, Inc. safely transport the ground beef?

Based on this x-bar chart CoolCargo can safely transport the ground beef. Average temperature stays below 40°F which is the critical temperataure.

    b.   Create an R-chart, and plot the data on the chart. Do you notice any points out of control? Based on this data can CoolCargo safely transport the ground beef?

    c.   Based on this information, is this truck capable of safely transporting ground beef? What type of process control charting would you recommend to CoolCargo? Explain.

    d.   What would you recommend CoolCargo do to improve the process?

9. Hospitals must deal with different types of waste in their operations. Infected waste (waste that has possibly been contaminated) cannot be thrown in a normal trash can, but must be disposed of in a special biohazard container. It is estimated that 6% of a hospital's waste is considered infectious or hazardous. The annual cost to dispose of hazardous waste to an average hospital is about $2 million per year, roughly the same cost of a new CAT scanner. Due to the high potential for savings, the local hospital is trying to reduce the amount of infectious waste. The amount of hazardous waste varies depending on the number of people in the hospital, the procedures performs and the personnel working. The hospital team conducted a study in order to create an x-bar chart of hospital waste.

Mean	7 lbs/day
Upper control limit	8.2 lbs/day
Lower control limit	5.8 lbs/day

Using this as a baseline the improvement team looked for ways to reduce the amount of hazardous waste. They noticed that people were throwing regular trash such as soft drinks and paper in the hazardous waste containers. The team developed an employee training program to educate the staff on hazardous and non-hazardous waste disposal. The data was again collected for the next 30 weeks and averaged by week. The data is reported below:

    a.   Calculate the mean and standard deviation for the new data.

    b.   What are the new limits for a 3 standard deviation x-bar chart?

    c.   How does the new average and standard deviation of the x-bar chart compare with the old (pre-training) mean and standard deviation? Higher or lower?

    d.   Based on this data do you think the training was effective?

    e.   If disposing of hazardous waste costs $100/lb, what is the expected annual savings from this project?

10. Reference: "Correcting package weights puts giveaway on ice", Packaging Digest, Vol. 46, Issue 7, p. 40-42, available in teacher notes.

Companies must carefully monitor package weights. Over-filled packages amount to product give-aways. There are legal requirements and fair trade regulations that protect the consumer from negligent under-filled packages. In addition, under-fills can be detrimental to customer satisfaction and will result in a loss of repeat business.

Oregon Freeze Dry (OFD) is the world's largest producer of freeze dried foods. They make private label freeze dried foods under the private label, Mountain House (http://www.mountainhouse.com/), supply other food manufacturers and are the largest supplier to the US military. The Instructions for rehydrating freeze dried food is based on the net freeze dried food content weight. If the freeze dried weight is not consistent, when water is added the food will not have the desired quality, taste, texture and/or flavor. OFD has installed a real-time SPC system that allows for operators to immediately adjust the equipment to bring the weight back into control. Packaging operators take samples every hour.

One top selling product is Beef Stroganoff with Noodles. It is packed in a horizontal form/fill/seal machine. This machine forms the pouch, adds the product and seals the pouch. The pouch should be packaged at 136 grams. OFD would like a probability of 99.7% that the net contents are at 136 ± 3 grams. The following data was collected while the equipment was running in control

   a. What are the Upper Control Limit, Lower Control Limit and centerline for the x-bar and R charts?

   b. Why do you think the OFD has the average packaging weight higher than the requirement?

   c. Engineering want to spend $800 on improvements to the machine. The think this will reduce the range to an upper control limit of 2 grams with an average of 1 gram. This will also allow them to decrease the average fill per package to 136 grams. How many packages of Beef Stroganoff with Noodles would they have to produce to break even if it costs them $0.058 per gram for the food? Do you think customer service would suffer due to an increase in under-filled pouches?

Acknowledgment

The authors gratefully acknowledge Ms. Becca Purser, age 13, and Ms. Katlyn Purser, age 10, for working out the procedure for assembling the Boeing 787 fuselage. They did the assembly operation and discovered a flaw in the initial assembly instructions. They analyzed the final assemblies as well as the process and determined the problem. They then hypothesized corrective action, tested their hypothesis in a second assembly operation. This proved their corrective actions were correct. Finally, Becca and Katlyn made recommendations to us, which we incorporated into the instructions. Without their assistance, the assembly problem would not be working correctly.

References

Benneyan, J.C. 1998. "Use and interpretation of statistical quality control charts," *International Journal for Quality in Health Care,* vol. 10, no. 1, pp. 69-73.

NIST 2010, United States National Institute of Standards and Technology, March 2010, http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc321.htm#Factors%20for%20Calculating%20Limits

Shewhart, W.A. 1986. *Statistical Method from the viewpoint of Quality Control*, Dover Publications, Mineola, NY. This is an unabridged edition of the original 1939 work with a new foreword written by Edwards Deming.

Western Electric 1956. *Statistical Quality Control Handbook,* Western Electric Co, 5th printing, Mack Printing Company, Easton, PA, Nov. 1977.
Note: Western Electric was the manufacturing subsidiary of AT&T when it was a government sanction telephone monopoly within the U.S. It was broken up in the 1980s. This was their guide for applying statistical quality control principles to the company's manufacturing and other operations. It is considered the gold standard for applied quality control.
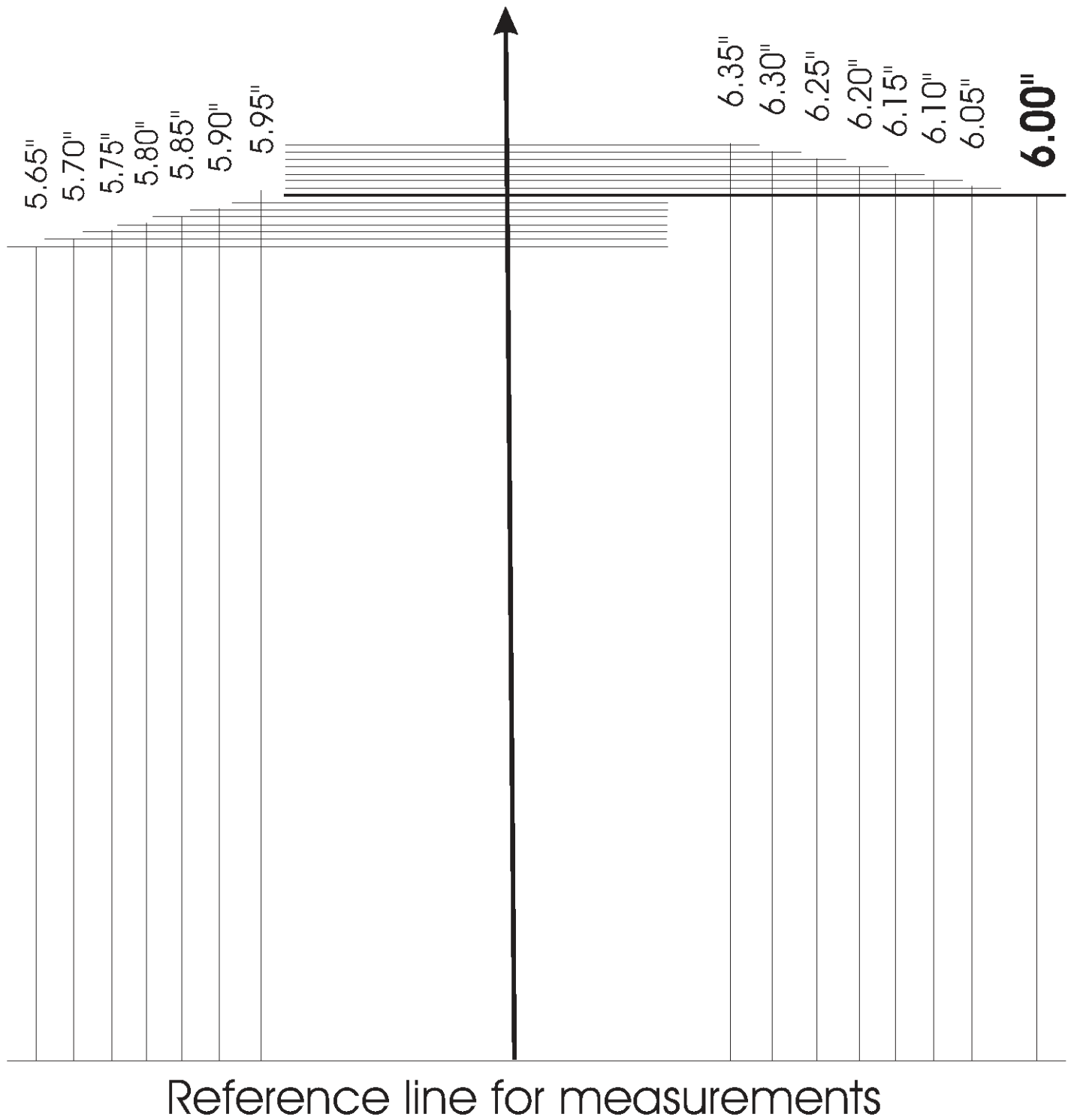
http://www.seattlepi.com/dayart/20070629/787assemblers.gif

Worksheet 1
Figure 16.2.2: Boeing 787 fuselage without cutting lines.
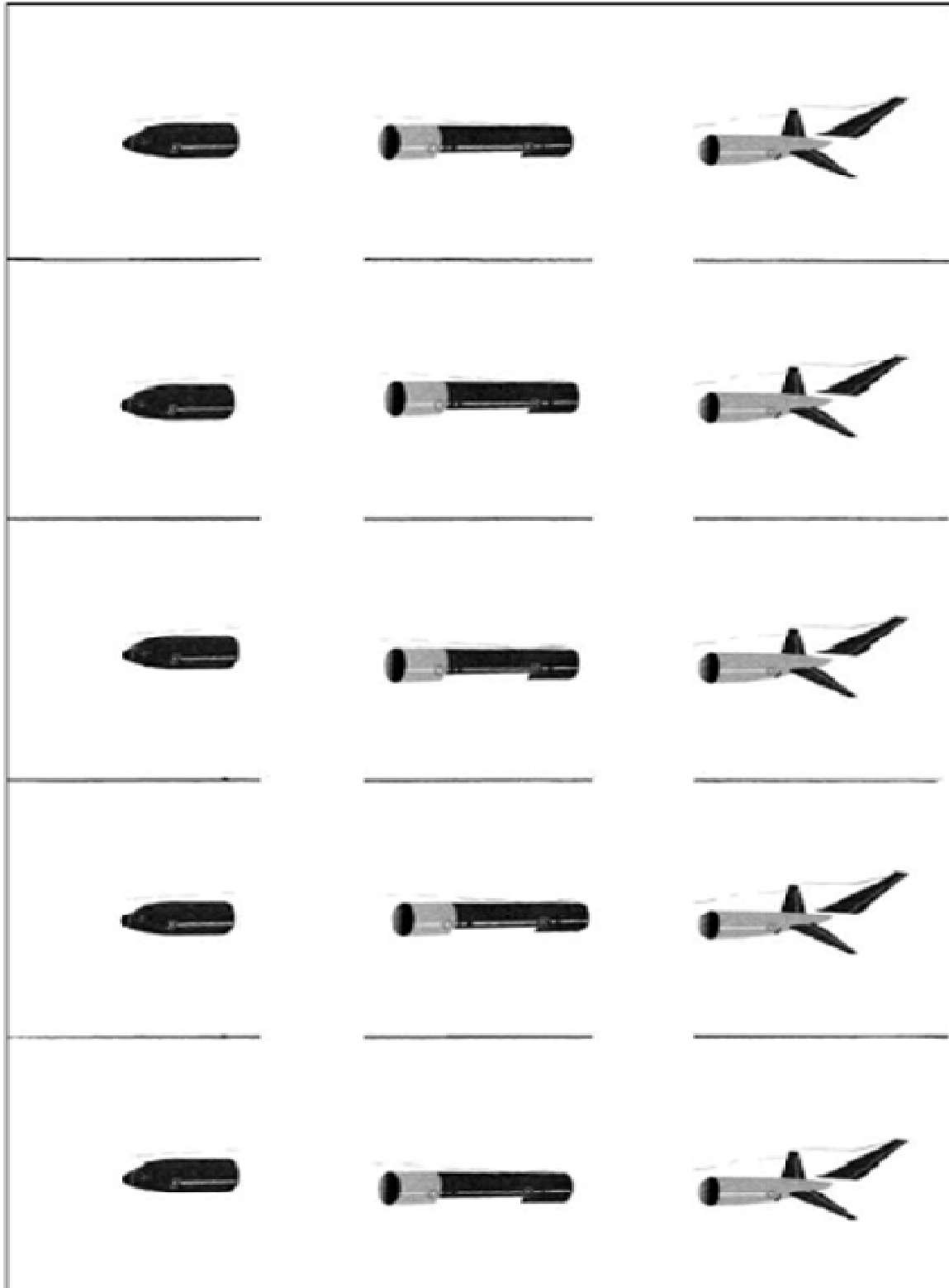
Worksheet 2
**Figure 16.2.4:** Measuring tool for completed Boeing 787 fuselages

Reference line for measurements

Worksheet 3
Table 16.2.1: Pooled data measurement for completed fuselages

Sample	Trial number					Average	Range
	1	2	3	4	5		
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							

Worksheet 4

Figure 16.2.6: The $\bar{x}$ -Chart for the average sample lengths of completed fuselages from the Boeing 787 model activity
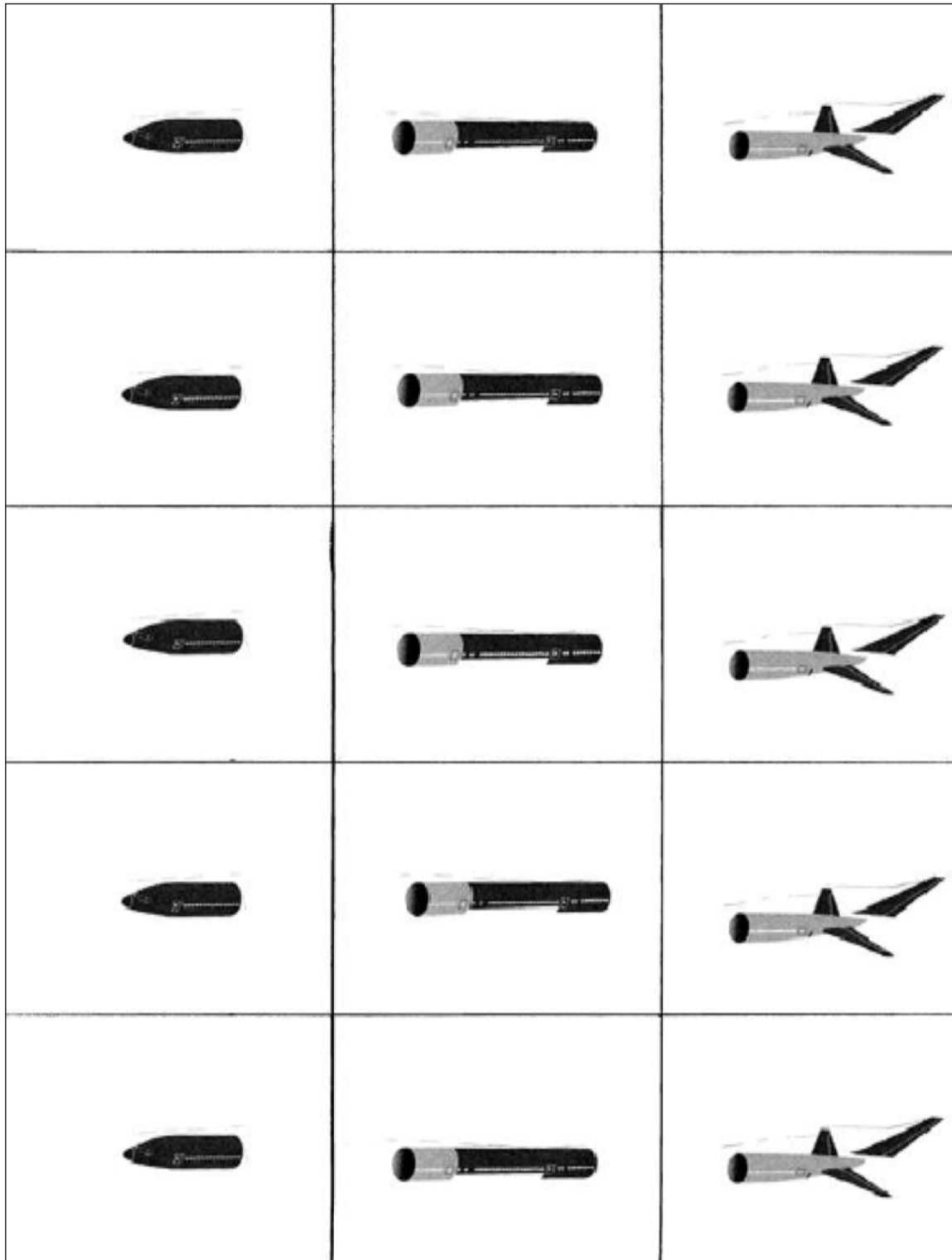

x-Bar Chart for Completed Fuselage Lengths
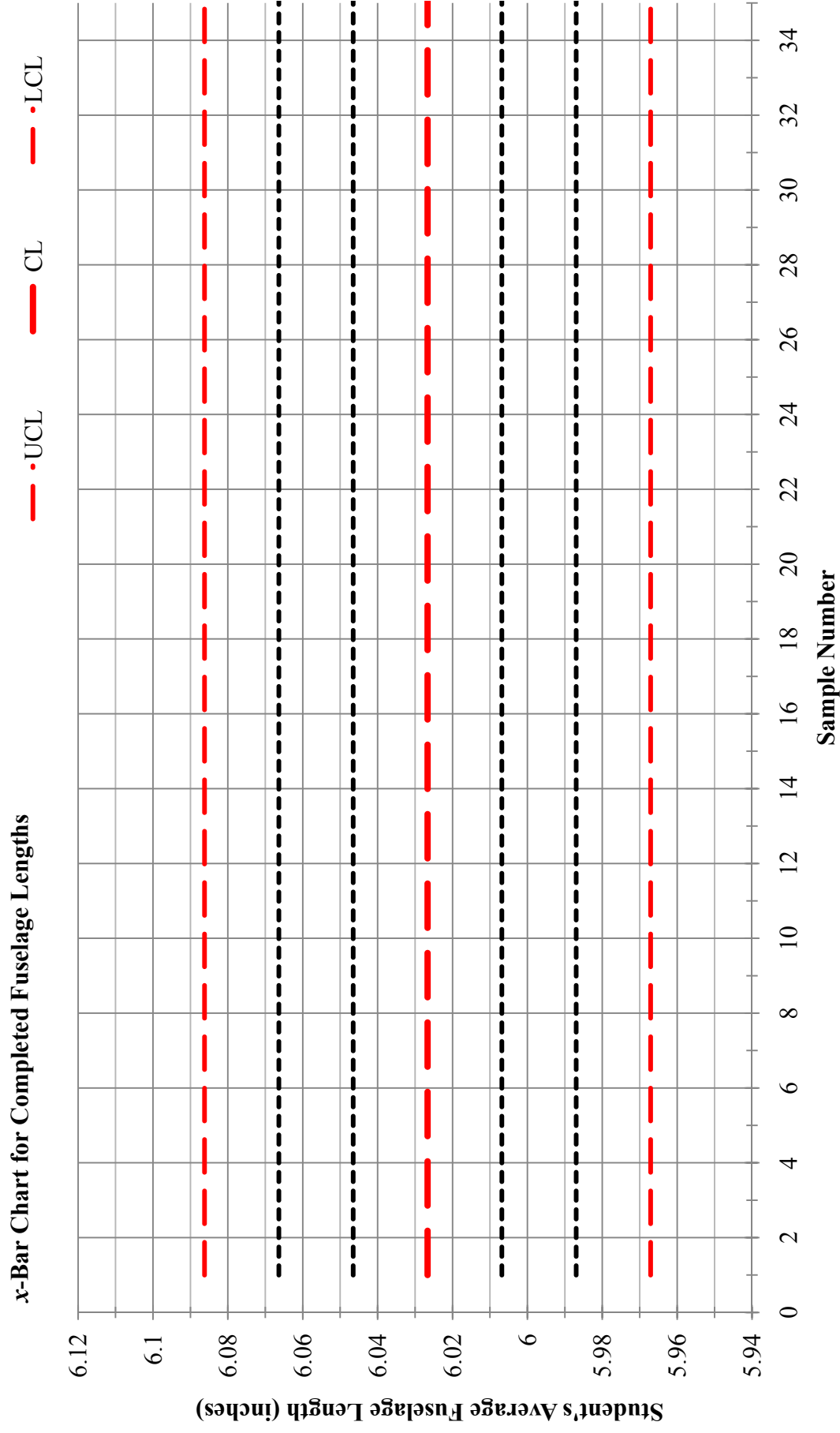
Worksheet 5

Figure 16.2.11: The *R*-Chart for the sample length ranges of completed fuselages from the Boeing 787 model activity

_R_-Chart for Complete Fuselage Lengths

- – – UCL
- – – CL
- – – LCL

y-axis: **Student's Fuselage Length Ranges (inches)** — 0, 0.1, 0.2, 0.3, 0.4, 0.5

x-axis: **Sample Number** — 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34
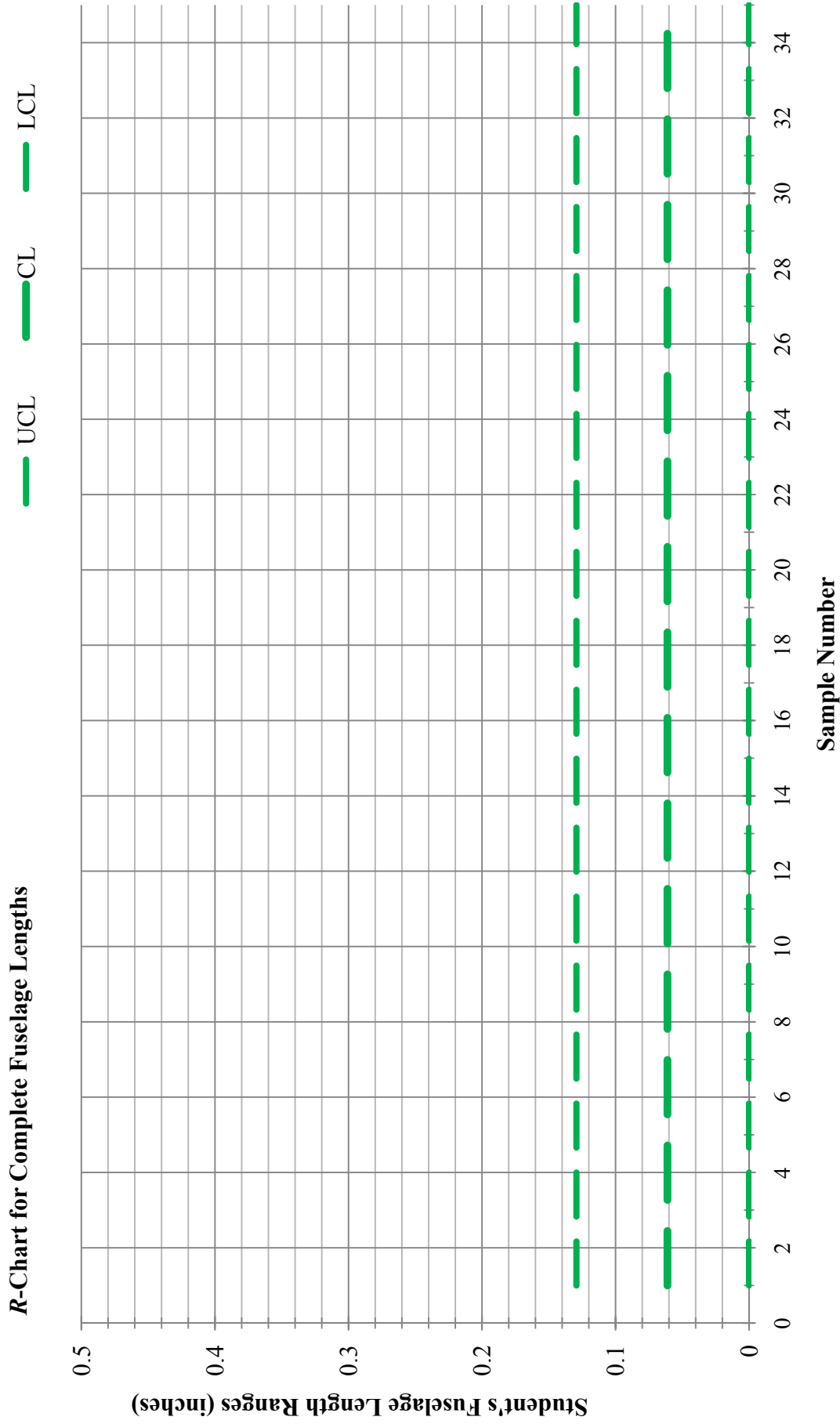
Worksheet 6
Figure 16.2.12: Boeing 787 fuselage with cutting lines

Worksheet 7
Table 16.2.4: Pooled data measurement for completed fuselages (with cutting lines)

Sample	Trial number					Average	Range
	1	2	3	4	5		
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							
Average of Averages, $\bar{\bar{x}}$							
Estimated Standard Deviation, s							
Average of Ranges, $\bar{R}$							

Section 17.1: Spam Email

People all around the world use medical testing to find occurrences of diseases. However, tests are flawed: they sometimes find things that don't exist or they miss things that do exist. In this chapter, we will explore the probability of the occurrence of false positives and false negatives, not only in medical testing but also when dealing with email spam filters and credit card fraud.

17.1.1 False Positives and Negatives in Medicine

A **false positive result** occurs when a test detected something that did not exist. A **false negative result** means a test failed to detect something that did exist. The most common discussion of false positives and negatives takes place within the field of medicine. One reason for this is that false positives and negatives are quite prevalent in medical testing. Another reason for this is that making a mistake as in missing a positive for a test can be life threatening. Tens of millions of people in the U.S. undergo medical tests each year and even a small percentage of false positives raise concerns.

When considering medical testing, a false positive may occur because the tests look at surrogate or early indicators of the disease; they do not look for the actual disease. For example, mammograms look for a mass in the breast; they do not look for cancer. If a false positive result is given, this could lead to extra costly procedures and unnecessary anguish for the patient.

On the other hand, a false negative may occur because the tests are not sensitive enough to see something, or perhaps the patient does not have the disease indicator the test is looking for. If a false negative result is given, this is a missed diagnosis, which may lead to a more serious illness.

17.1.2 Spam Email Filters

The discussion of false positives and false negatives usually occurs within the medical field. However, we can also investigate the idea of false positives and negatives when discussing spam email filters. That is, a filter may mark an email as spam when it is in fact wanted (or "ham") email (this is a false positive). Other the other hand, a filter may send an email to your inbox when it is spam (this is a false negative).

Q1.   What are some of the consequences of an email that is not spam, being filtered as spam?

Suppose an email filter is used to detect spam email. When testing for spam, there are four possible results. These possibilities are summarized in the table below:

	Spam	Not Spam
Filter recognized as spam	True Positive	False Positive
Filter did not recognize as spam	False Negative	True Negative

Table 17.1.1: Possible results for spam email filter

If we know the probability that an email is filtered as spam given it was indeed spam, could we find the probability that an email is spam given that it was filtered as spam?

For example, suppose we know that the probability an email was filtered into the spam folder given that the email is spam. The notation used for this is P(filtered | spam). We can estimate this probability by counting the number of spam emails in the spam filter, then dividing by all the spam emails we have received over a long period of time.

Q2.   Given this information, what do you think P(spam | filtered) is?

We will develop how to find the second probability in the next section. But first, we will explore this example further.

Suppose that over a long period of time the average per cent of spam emails is 0.20 or 20%. This is known as the **prior probability**. The prior probability is an estimate of the probability of an event occurring before any new information is obtained.

We are given P(filtered | spam) = 0.90 and P(not filtered | not spam) = 0.95. That is, we know that the probability of an email being filtered given than it was spam (i.e., a true positive) is 0.9 or 90%; the probability of an email not being filtered given that it was not spam (i.e., a true negative) is 0.95 or 95%. Table 17.1.2 illustrates this given information. Figure 17.1.1 shows the information in a tree diagram.

Prior Probability	P(spam) = 0.20	P(not spam) = 0.80
	Spam	**Not Spam**
Filter recognized as spam	P(filtered | spam) = 0.90	P(filtered | not spam) = ?
Filter did not recognize as spam	P(not filtered | spam) = ?	P(not filtered | not spam) = 0.95

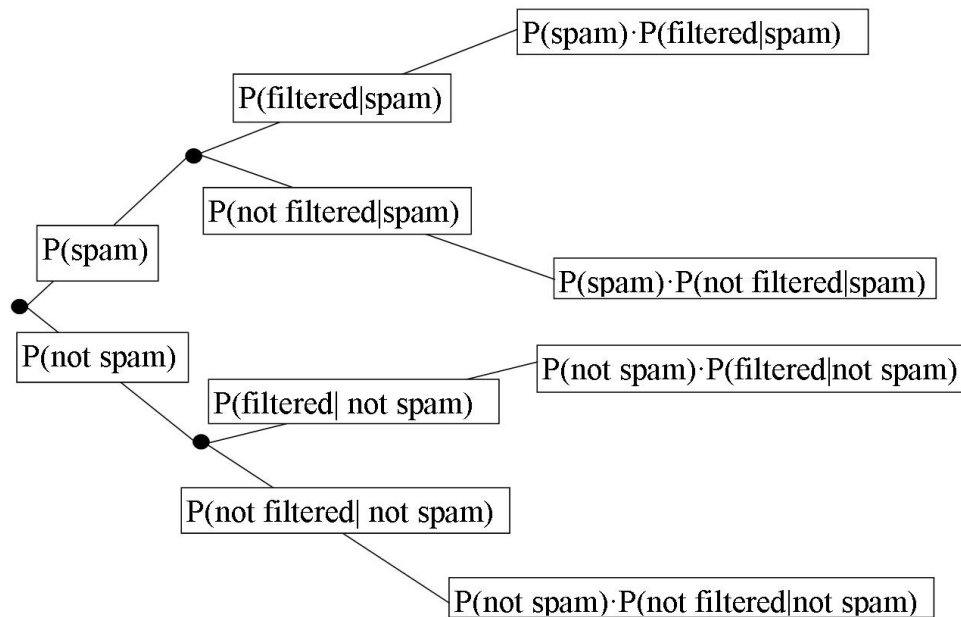**Table 17.1.2**: Possible results for spam email filter

Figure 17.1.1: Probability of spam emails represented as a tree diagram.

Remember that in a probability tree diagram, the probability for any branch of the tree is the product of the probabilities of each part of the branch. That is why you see the far right notations on the tree.

Q3.   Which pair of events are complements of one another? How do you know?

Q4.   Given the information in Table 17.1.2 and your response to the previous question, what is P(not filtered | spam)?

Q5.   What is P(filtered | not spam)?

Q6.   Complete Table 17.1.3 with the information obtained in questions 3 and 4.

Prior Probability	P(spam) = 0.20	P(not spam) = 0.80
	Spam	**Not Spam**
Filter recognized as spam	P(filtered | spam) = 0.90	P(filtered | not spam) = ___
Filter did not recognize as spam	P(not filtered | spam) = ___	P(not filtered | not spam) = 0.95

Table 17.1.3: Possible results for spam email filter

Next, consider this information in another way. Imagine you received 100 emails in one day. Since the prior probability of spam email is 0.20, then on average 20 emails out of the 100 emails are spam and 80 are not spam. This is represented graphically in Figure 17.1.2, the twenty spam emails are shown shaded. In this figure, prior probability P(spam) is represented by the shaded circles, and prior probability P(not spam) is represented by the un-shaded circles. Shaded rectangles show the same information in a tree diagram.
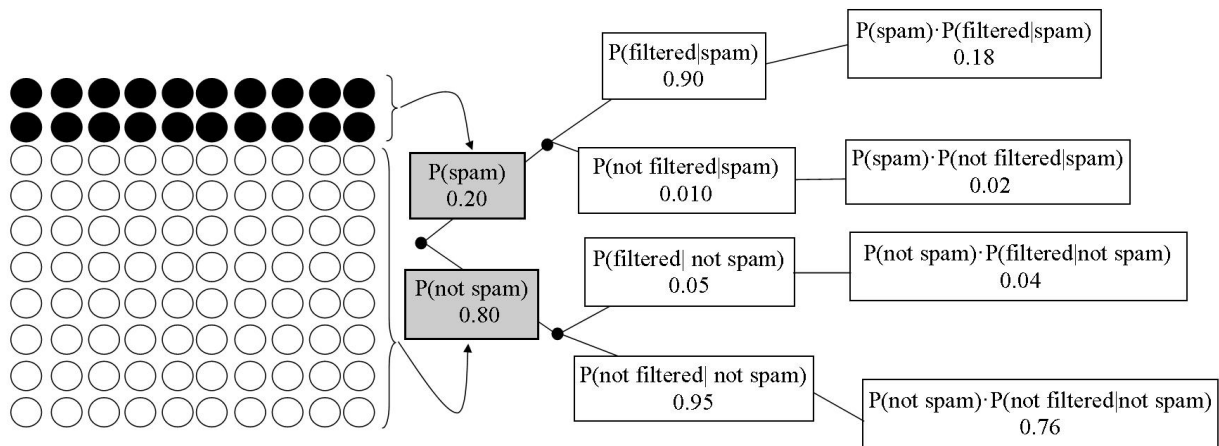

Figure 17.1.2: Pictorial representation of the spam email problem

Of the 20 spam emails, there is 90% probability that each will be identified as spam. Thus, on average, 18 of these 20 spam emails will be identified as spam and sent to the spam folder.

In Figure 17.1.3, a box is drawn around the spam emails that are filtered. The shaded rectangle in the tree diagram represents spam emails that are filtered as spam.
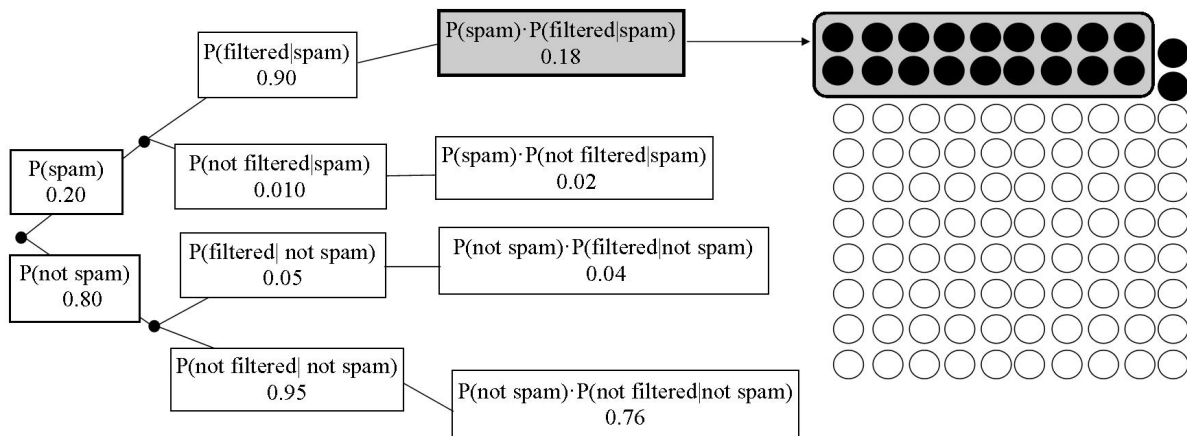
Figure 17.1.3: Emails that are spam and are filtered

In Figure 17.1.4, look at the 80 emails that are not spam, these are represented by the un-shaded circles. These are real emails from friends, family, school and coupons from your favorite stores that you want to open and read. The email filter will correctly identify 95% of these.

Q7.   How many of the 80 emails that are *not spam* will be filtered as spam?

Figure 17.1.4 shows shaded boxes around all of the emails that filtered as spam. The solid box shows the spam emails that also filtered as spam (true positives), and the dashed box shows the emails that are NOT spam, and are filtered as spam (false positives). The arrows highlight the relationship to the tree diagram.
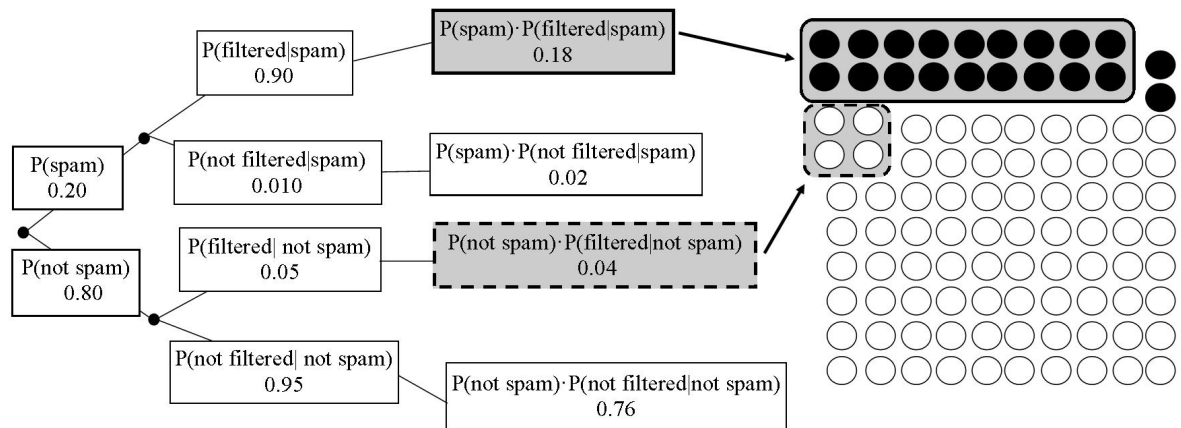

Figure 17.1.4: All emails that are filtered as spam

Q8.   How many of the total 100 emails are filtered as spam?

Q9.   How many of the total 100 emails are *not* filtered as spam?

Q10.   We have 4 groups of emails. Fill in the quantity for each group in Table 17.1.4.

Group	Quantity	Probability
Spam emails that are filtered as spam		
Spam emails that are not filtered as spam		
Not spam emails that are filtered as spam		
Not spam emails that are not filtered as spam		
Total		

Table 17.1.4: Groups of emails

Figure 17.1.5 shows only the emails that are filtered as spam, where the shaded circles represent the spam emails and the un-shaded circles represent non-spam emails.
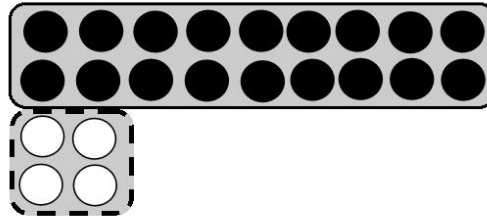
Figure 17.1.5: Emails that are filtered as spam

The prior probability of a spam email was 20%. From the table and diagram we can easily see that the spam emails still equal 20%:

$$\frac{18+2}{100} = 0.20$$

Recall that the relative frequency of an event corresponds to the probability of the event:

$$probability = \frac{frequency\ of\ desired\ event}{total\ frequency}$$

Use this information and Figure 17.1.5, we can calculate *P*(spam | filtered), that is, the probability of spam given the email was filtered as spam. The *desired event* is the spam emails that are filtered, which is represented by the shaded circles in the diagram, and equals 18. The *all possibilities* is the total number of filtered emails, both spam and not spam, which is (18 + 4) = 22. Therefore, we can now calculate P(spam | filtered):

$$P\left(spam \mid filtered\right) = \frac{frequency\ of\ desired\ event}{total\ frequency}$$

$$\frac{0.18}{0.18+0.04} = 0.82 \qquad\qquad \frac{18}{18+4} = 0.82$$
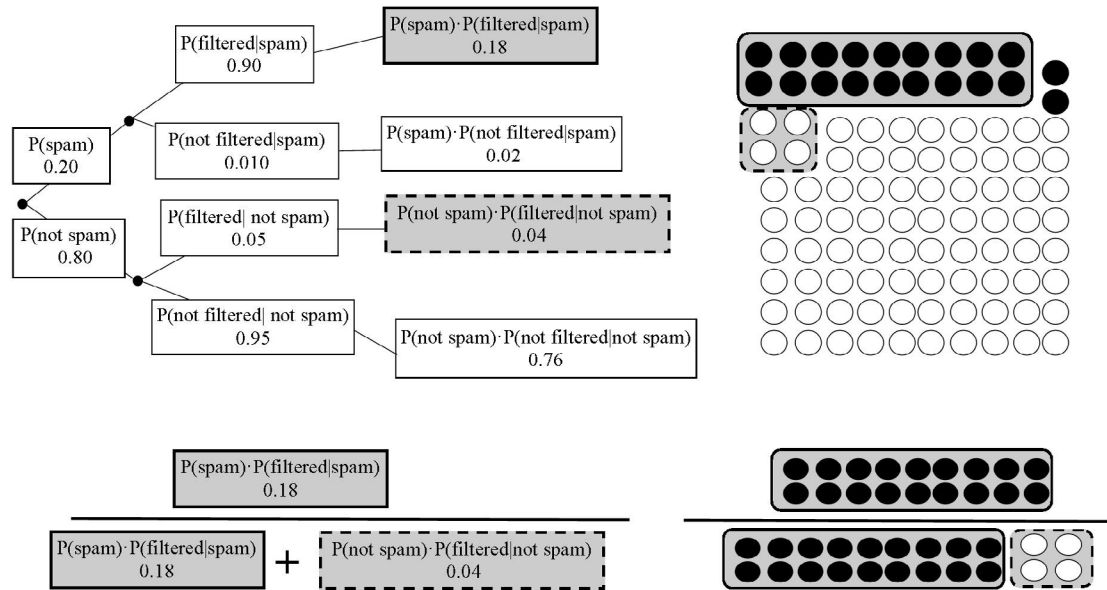
Figure 17.1.6 shows this calculation pictorially.

Figure 17.1.6: Calculating *P*(spam | filtered)

Q11.   Use this information to find the *P*(not spam | filtered).

Q12.   What is the probability and consequence of *P*(spam | not filtered)?

Q13.   List two other examples of where false positive or negative results may occur (other than the field
          of medicine and spam email). In each example, answer the following:
          a.   Why do you think false positive results occur?
          b.   What are some consequences of a false positive result?
          c.   Why do you think false negative results occur?
          d.   What are some consequences of a false negative result?

Section 17.2: Lyme Disease

Lyme disease is disease spread through infected blacklegged tick bites, see Figure 17.2.1. Typical symptoms of Lyme disease include fever, fatigue, headache and a characteristic ring-shaped rash, see Figure 17.2.2. Rash is observed in approximately 70-80% of infected individuals.

Figure 17.2.1: Blacklegged (deer) tick

Figure 17.2.2: Characteristic ring-shaped rash cause by Lyme disease

If left untreated Lyme disease can lead to serious complications such as severe joint pain and swelling, severe headaches, heart palpitations, arthritis, and neurological problems. Most cases of Lyme disease, if caught early, can be cured with antibiotics taken for a few weeks. It is important to treat early in order to prevent the more serious complications which can often last for years. For more information about Lyme disease, visit the CDC website on Lyme disease (http://www.cdc.gov/ncidod/dvbid/lyme/index.htm).

The diagnosis of Lyme disease is based on the symptoms found during a physical exam – especially the characteristic rash, and a possible exposure to infected ticks. In addition, there are several laboratory tests available. The CDC recommends a two-step process. The first step is uses an ELISA test; this is an enzyme-linked immunoassay which detects antibodies in the blood. If this test is positive, the CDC recommends an immunoblot, such as a Western blot test, which is used to detect specific proteins in the blood. The Western Blot test can use the same sample of blood but takes approximately one week to complete. The CDC does not recommend performing the immunoblot before the ELISA test because of the potential for a higher number of false positives.

17.2.1 False Positive and Negative Results in Lyme Disease Testing

Dr. Tyler has a patient, Tommy, with symptoms of fatigue, chills, fever, headache and swollen lymph nodes. Tommy does recall while camping two weeks ago removing several ticks, however, Dr. Tyler does not find any circular rash. Dr. Tyler knows that approximately 70-80% of people with Lyme disease will

have a rash, however, because of the possible exposure to tick bites, Dr. Tyler decides to perform the ELISA test.

Q1.   What are the consequences of a false negative in Lyme Disease?

Q2.   What are the consequences of a false positive?

Medical tests often are characterized by their sensitivity and specificity. **Sensitivity** refers to the likelihood of a *positive test result* in patients *with the disease* (i.e., a true-positive) or *P*(Positive test | has the disease). The sensitivity of a test is how well a test detects the disease. A test with a lower sensitivity is more likely to fail to identify many patients with the disease. This is the complement of a false-negative.

Specificity refers to the likelihood of a *negative test result* in patients *without the disease* (i.e., a true negative) or *P*(Negative test | does not have the disease). A test with a lower specificity will diagnoses many patients without the disease as having the disease. This is the complement of a false-positive.

> **"Sensitivity" in Medical Testing versus**
> **"Sensitivity Analysis" in Excel**
>
> Medical testing sensitivity should not be confused with the Sensitivity Report used in Excel. The Excel Sensitivity Report showed us how much the profitability must be increased in order to bring a particular decision variable that was outside the optimal solution into the optimal solution. In this way it helps to determine how "sensitive" the optimal solution is the changes in data values. The Excel Sensitivity Report shows how changes to the situation will affect the optimal solution.

The ELISA test has a sensitivity of 85% for Lyme disease. In other words, the probability that the test results are positive given that the patient has Lyme disease is 85% (i.e., the probability of a true positive is 85%). Furthermore, the specificity of the ELISA test is given at 88%. Therefore the probability that the test results are negative given that the patient does not have Lyme disease is 88% (i.e., the probability of a true negative is 88%). This information is detailed in Tables 17.2.1a and 17.2.1b.

	Lyme Disease	**No Disease**
Positive Test Result	True Positive	False Positive
Negative Test Result	False Negative	True Negative

Table 17.2.1a: Possible results for Lyme disease testing

	Lyme Disease	**No Disease**
Positive Test Result	*P*(positive result | disease) = 0.85	*P*(positive result | no disease) = 0.12
Negative Test Result	*P*(negative result | disease) = 0.15	*P*(negative result | no disease) = 0.88

Table 17.2.1b: Possible results for Lyme disease testing, with probabilities

Suppose we would like to determine the probability that a patient has the disease given that the test result was positive. We can use a probability tree to determine *P*(disease | positive result).

First, the prior probability of having Lyme disease is needed. In this case, Dr. Tyler, based on her experience with many patients who have potential exposure to an infected tick and with these symptoms, estimates that there is a 20% probability of Lyme disease.

For any given patient, either the test result was positive or it was not, and either the patient has the disease or the patient does not. These possibilities are illustrated in Figure 17.2.3, where *PR* stands for "positive result," *NR* stands for "negative result," *D* stands for "disease," and *D'* stands for "no disease."

Q3.   Fill in the tree diagram in Figure 17.2.3.
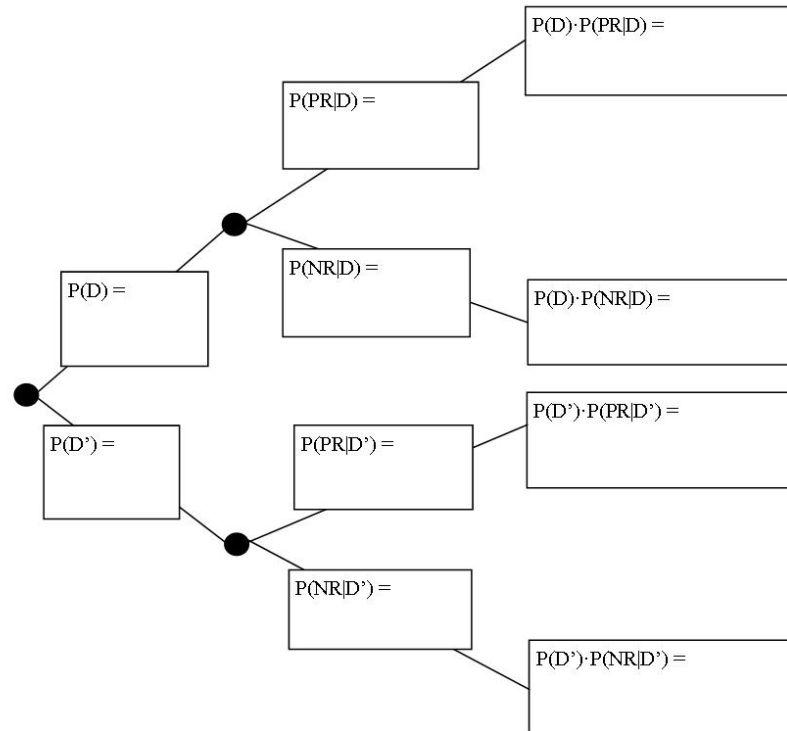

**Figure 17.2.3**: Tree diagram for Lyme Disease

The information can also be represented in a table, as shown below is Table 17.2.2.

Q4.   Fill in Table 17.2.2

Prior Probability	$P(D) = 0.20$	$P(D') = 0.80$
	Lyme Disease	**No Lyme Disease**
Positive Test Result	$P(D) \cdot P(PR \mid D) =$ ___	$P(D') \cdot P(PR \mid D') =$ ___
Negative Test Result	$P(D) \cdot P(NR \mid D) =$ ___	$P(D') \cdot P(NR \mid D') =$ ___

Table 17.2.2: Lyme Disease information

Now, suppose for Tommy that the ELISA test comes back positive. Dr. Tyler would like to know the probability the patient, Tommy, does *not* have Lyme disease: P(No Lyme Disease | Positive Test Result). In this case, the event of interest is "does not have Lyme disease and a positive test result", and all possibilities would be "all possibilities of having a positive test result". This is shown in the tree diagram in Figure 17.2.4.
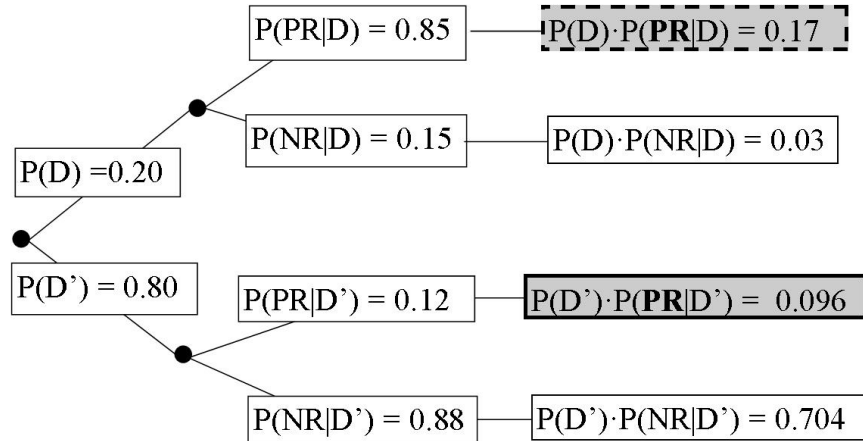
$$P(D'|PR) = \frac{P(D')\cdot P(\mathbf{PR}|D') = 0.096}{P(D)\cdot P(\mathbf{PR}|D) = 0.17 \ + \ P(D')\cdot P(\mathbf{PR}|D') = 0.096} = \mathbf{0.361}$$

Figure 17.2.4: Probability of not having Lyme disease given a positive ELISA test result

The same information can also be displayed as a table, as shown in Table 17.2.3. Note that the Lyme disease column sums to equal the prior probability of having Lyme disease as stated earlier. This is the same for the No Lyme disease column.

	Lyme disease	**No Lyme disease**	**Total**
Positive test result	$P(D) \cdot P(PR\|D)$ = 0.20 · 0.85 = 0.170	$P(D') \cdot P(PR\|D')$ = 0.80 · 0.12 = 0.096	0.170 + 0.096 = 0.266
Negative test result	$P(D) \cdot P(NR\|D)$ = 0.20 · 0.15 = 0.030	$P(D') \cdot P(NR\|D')$ = 0.80 · 0.88 = 0.704	.030 + 0.704 = 0.734
Total	0.170 + 0.030 = 0.20	0.096 + 0.704 = 0.80	1.00

Table 17.2.3: Probability of not having Lyme disease given a positive ELISA test result

Q5.　Suppose the ELISA test comes back positive. What is the probability the patient, Tommy, does have Lyme disease?

Q6.　If you were Dr. Tyler, and the ELISA test came back positive for Lyme disease, what would you do? What would be a consequence of your action?

Dr. Tyler decides to follow the CDC recommended guideline and follow up with a Western blot. The combined ELISA and Western blot test have a sensitivity of 0.90 and a specificity of 0.99.

Q7.　Using this information, fill in Table 17.2.4.

	Lyme Disease	No Lyme Disease
Positive Test Result	P(Positive Result | Disease) = ____	P(Positive Result | No Disease) = ____
Negative Test Result	P(Negative Result | Disease) = ____	P(Negative Result | No Disease) = ____

Table 17.2.4: Probabilities for combined ELISA and Western blot test for Lyme's disease

Q8.   If the combined test comes back positive for Lyme disease, would you expect the probability that the patient has Lyme disease to be higher or lower than with just the ELISA test? Explain.

Q9.   Using the same prior probabilities as before, fill in the Table 17.2.5.

	Lyme disease	No Lyme disease	Total
Positive test result	$P(D) \cdot P(PR\|D)$ =	$P(D') \cdot P(PR\|D')$ =	
Negative test result	$P(D) \cdot P(NR\|D)$ =	$P(D') \cdot P(NR\|D')$ =	
Total			

Table 17.2.5: Conditional probabilities for combined ELISA and Western blot test

Q10.   The combined tests come back negative for Lyme disease, what is the probability that Tommy has Lyme disease? What would you do if you were Dr. Tyler?

Q11.   The combined tests come back positive for Lyme disease, what is the probability that Tommy has Lyme disease? What would you do if you were Dr. Tyler?

17.2.2 A New Prior Probability

Dr. Tyler just went to an educational training class sponsored by the American Lyme Disease Foundation. It was emphasized that there is an ongoing problem of the inappropriate use of serology to test patients with a low pre-test probability of Lyme disease. After consideration, Dr. Tyler is not too sure of her original estimate of a 20% probability of Lyme disease. She thinks it might be closer to 10%.

Q12.   Re-calculate the probability of Tommy having Lyme disease given a positive test result using the combined ELISA and Western blot tests using a prior probability of 10%. Do you think this would change Dr. Tyler's decision?

Q13.   Dr. Tyler sees that Tommy has not been on antibiotics in a couple of years. She would still give Tommy 2-weeks of antibiotics based on his symptoms, if the probability of Lyme disease given a positive test result was 0.60 or higher. What is the lowest pre-test probability of Lyme disease she could estimate to obtain a $P(D \mid PR) = 0.60$?

Q14.   Why is it important that Dr. Tyler estimate the Lyme disease accurately?

Section 17.3: Identifying Credit Card Fraud

Fraud is a deception deliberately practiced in order to secure unfair or unlawful gain. According to the law, this deception can be a federal crime. If caught, a citizen can be charged with a felony and be sentenced to jail. Currently, Bernard "Bernie" Madoff is one of the most notorious fraudsters. He was convicted of eleven federal offenses on March 12, 2009 and has been sentenced to 150 years in federal prison.

There are many different types of fraud. Mr. Madoff was convicted of securities fraud, wire fraud, mail fraud, money laundering, making false statements, perjury, theft from an employee benefit plan, and making false filings with the SEC. While Mr. Madoff is an extreme case of different types of fraud, there are four common types of fraud: internal fraud, insurance fraud, telecommunications fraud, and credit card fraud.

Internal fraud is when employees of a business "cook the books" meaning somehow they misrepresent financial reports (typically for their own financial gain). This could happen at any level in the company.

Insurance fraud can be found in four areas of insurance: home insurance, crop insurance, automobile insurance, and medical insurance. For all subgroups, typically a customer files a faulty claim in order to collect money.

Telecommunications fraud is the theft of a telecommunication service (telephones, cell phones, computers etc.) or the use of telecommunication service to commit other forms of fraud.

Finally, **credit card fraud** refers to falsified credit card applications or fraudulent credit card purchases. Many citizens have been or will be victims of credit card fraud. This experience can range from a false purchase charged on a monthly credit card bill to identity theft. Credit card companies work around the clock to protect their customers' accounts. While this is a very costly endeavor, the alternatives (like losing customers or paying for fraudulent purchases) are even more costly.

In order for major credit card companies to detect fraudulent credit card transactions, they use sophisticated data-mining programs that, after examining numerous variables and purchasing patterns, will flag prospective fraudulent purchases. Stated another way, credit card companies use programs that analyze large amounts of data in order to reliably detect unusual behaviors that could be indicators of fraudulent purchases. Ideally, these programs would identify (i.e. flag) all the transactions that are fraudulent a hundred percent of the time without incorrectly flagging legitimate transactions. However, the reality is that these data-mining software programs are not perfect. They sometimes falsely identify transactions as fraudulent (saying transactions are fraudulent when they are not or false positive) or the programs do not flag transactions that are fraudulent (letting fraudulent transactions through as though they are legitimate or false negative).

Q1. What are some of the consequences to the credit card company of falsely identifying transactions as fraudulent? What are some of the consequences to the credit card company of letting fraudulent transactions through as though they were legitimate?

Q2. What are some of the consequences to the credit card *holder* of falsely identifying transactions as fraudulent? What are some of the consequences to the credit card *holder* of letting fraudulent transactions through as though they were legitimate?

There are multiple computer programs available for credit card companies to use. These programs score the transactions on a scale from 1 to 10. Transactions with a score of 8 or above are considered to have a high

risk of being fraudulent. In these cases, the credit card company denies the transaction and puts a freeze on the account until the customer can be reached to verify the transactions. For a score greater than 4, but less than 8, there is a medium risk that the charge is fraudulent. For these medium-scored transactions, the credit card company will further investigate the transaction by contacting the merchant; often these are smaller charges and are not discovered until after the transaction has processed. Scores below 4 are considered to have a low probability of risk of being fraudulent so the transactions proceeds with no further action form the credit card company. See Table 17.3.1 for a summary.

Score	Risk level	Action
Score >8	High	Deny transaction, freeze account, contact card holder
4 ≤ score ≤ 8	Medium	Allow transactions, call merchant, further internal investigation
Score < 4	Low	Allow transaction, no further investigation

Table 17.3.1: Summary of credit card transaction scores

One credit card company, Leader Card, is looking to invest in a data analysis software program for identifying fraudulent credit card transactions. Leader Card hires a research team, Dietzstein, to investigate which data analysis software program they should invest in.

The Dietzstein team did some background reading. They learned that when they compare the quality of data analysis software programs, they should use the same data for all of the programs. Otherwise, the results from the software programs may be inconsistent.

The Dietzstein team obtained 500,000 records of transactions from Leader Card that contained at least 30 different information fields from a database. The Dietzstein team cannot reveal what the 30 fields represent due to privacy agreements, but examples would include date and time of transaction, amount, geographic location of transaction, name and type of company, transactional history of card holder, payment history, etc. These 500,000 records were collected over a 12-month period and 100,000 of these records (20%) have already been identified as fraudulent transactions.

The Dietzstein team selects the three most popular data analysis software programs to test using their data pool: TRIPPER, WAGON, and PIN3.

The following tables report the resulting probabilities from running the data through each program.

TRIPPER Software Program	Fraudulent Transaction	Legitimate Transaction
Program scores transaction as *high* risk	0.67	0.03
Program scores transaction *medium* risk	0.13	0.20
Program scores transaction as *low* risk	0.2	0.77

Table 17.3.1: The resulting probabilities for the TRIPPER software program given 20% of the data records are fraudulent transactions and 80% of the data records are legitimate transactions.

WAGON Software Program	Fraudulent Transaction	Legitimate Transaction
Program scores transaction as *high* risk	0.76	0.05
Program scores transaction *medium* risk	0.18	0.11
Program scores transaction as *low* risk	0.06	0.84

Table 17.3.2: The resulting probabilities for the WAGON software program given 20% of the data records are fraudulent transactions and 80% of the data records are legitimate transactions.

PIN3 Software Program	Fraudulent Transaction	Legitimate Transaction
Program scores transaction as *high* risk	0.74	0.15

| Program scores transaction *medium* risk | 0.22 | 0.20 |
| Program scores transaction as *low* risk | 0.04 | 0.65 |

Table 17.3.3: The resulting probabilities for the PIN3 software program given 20% of the data records are fraudulent transactions and 80% of the data records are legitimate transactions.

Q3.   For each computer program, calculate the following (to three decimal places). You could use a calculator or a spreadsheet for these calculations. You may want to create a table or probability tree to help you.

    a.   Calculate the probability a transaction is *fraudulent* given that is was scored as a *high*: P(fraud | high).

    b.   Calculate the probability a transaction is *fraudulent* given that is was scored as a *medium*: P(fraud | medium).

    c.   Calculate the probability a transaction is *fraudulent* given that is was scored as a *low*: P(fraud | low).

    d.   Calculate the probability a transaction is *legitimate* given that was scored as a *high*: P(legitimate | high).

    e.   Calculate the probability a transaction is *legitimate* given that is was scored as a *medium*: P(legitimate | medium).

    f.   Calculate the probability a transaction is *legitimate* given is was scored as a *low*: P(legitimate | low).

Q4.   Given your results in Q1 and Q3, which computer program do you think the Dietzstein research team should recommend to the Leader Card company? What criteria did you use?

Even with the Dietzstein research team's investigation, Leader Card company is still nervous about choosing the best data-mining software to detect fraudulent credit card transactions. The software is very expensive and it will cost Leader Card a lot of time and money to train employees on the software.

The Leader Card company decides to utilize in-house accounts to calculate the extra cost associated with fraudulent transactions. The Leader Card accountants determine:

- The average extra cost of a fraudulent transaction being given a low risk score and therefore being processed is $500 since the credit card company (not the consumer) will have to absorb the cost of the transaction.
- The average extra cost of a legitimate transaction receiving a high risk score is $100 due to the credit card company having to freeze the account and call the customer unnecessarily.
- The average extra cost of a legitimate transaction receiving a medium risk score is $60 due to the cost of investigating the transaction.
- The average extra cost of fraudulent change receiving a high or medium risk score is $0.
- The average cost of a legitimate charge receiving a low risk score is $0.

Q5.   Calculate and interpret the expected extra cost for each of the software programs.

Q6.   Given this information, which software program should Leader Card credit card company purchase? Why?

Q7.   In what ways could the Leader Card company off-set the expected extra cost?

Terms

Credit Card Fraud

Insurance Fraud

Internal Fraud

False Negative Result

False Positive Result

Prior Probability

Sensitivity

Specificity

Telecommunications Fraud

References

Bernard Madoff - Wikipedia, the free encyclopedia. (2011). . Retrieved May 6, 2011, from
     https://secure.wikimedia.org/wikipedia/en/wiki/Bernie_Madoff

Division of Vector Borne Infectious Diseases: Rash Photos | CDC Lyme Disease. (2008). Retrieved May 6,
     2011, from http://www.cdc.gov/ncidod/dvbid/lyme/ld_LymeDiseaseRashPhotos.htm

Division of Vector Borne Infectious Diseases: Transmission | CDC Lyme Disease. (2007). Retrieved May
     6, 2011, from http://www.cdc.gov/ncidod/dvbid/lyme/ld_transmission.htm

Health Grades Inc. (2010). Prevalence and Incidence of Lyme disease - WrongDiagnosis.com. Retrieved
     February 25, 2010, from http://www.wrongdiagnosis.com/l/lyme_disease/prevalence.htm.

Johnson, B. J. B., Robbins, K. E., Bailey, R. E., Cao, B., Sviat, S. L., Craven, R. B., et al. (1996).
     Serodiagnosis of Lyme disease: Accuracy of a two-step approach using a flagella-based elisa and
     immunoblotting. *The Journal of Infectious Diseases*, *174*(2), 346-353.

The Laboratory Diagnosis of Lyme Disease II The Banbury Center, Cold Spring Harbor 9-12 September
     2007 Emerging Views on Serodiagnosis, http://www.aldf.com/2nd_Banbury_Conference.shtml

Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). A comprehensive survey of data mining-based fraud
     detection research. *Artificial Intelligence Review*, 1-14.
     http://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf

Scherokman, B. (1997). Selecting and interpreting diagnostic tests. *The Permanente Journal*, *1*(2), 4-7.

Stolfo, S. J., Fan, D. W., Lee, W., Prodromidis, A. L., & Chan, P. K. (1997). *Credit card fraud detection
     using meta-learning: Issues and initial results.* Paper presented at the AAAI-97 Workshop on Fraud
     Detection and Risk Management. http://www.aaai.org/Papers/Workshops/1997/WS-97-07/WS97-
     07-015.pdf

Tape, T. G. (n.d.). Likelihood Ratios: University of Nebraska Medical Center. Retrieved March 4, 2010,
     from http://darwin.unmc.edu/dxtests/Default.htm.

Suozzi, T. R., & Lawrence, J. H. (2010). *Nassau County SPIN: Telecommunications Fraud*. Retrieved from
     http://www.wantagh.li/spin/telecommunications_fraud.pdf

WebMD, LLC. (2005). Lyme Disease Test. Retrieved February 25, 2010, from
     http://arthritis.webmd.com/lyme-disease-test.

Section 18.0: Introduction

Queues are waiting lines, and we have all experienced them. It is estimated that Americans spend over a billion hours per year waiting in queues. Queueing theory is the mathematical study of queues that operations researchers and industrial engineers use to increase the efficiency of queueing systems. The term **queueing system** refers to the people currently waiting in the queue and the person(s) being served. Inefficient queues not only hurt customer satisfaction, but they also have an impact on a nation's economy. If the time we wasted waiting in queues could instead be spent productively, it would amount to the equivalent of half a million additional workers.
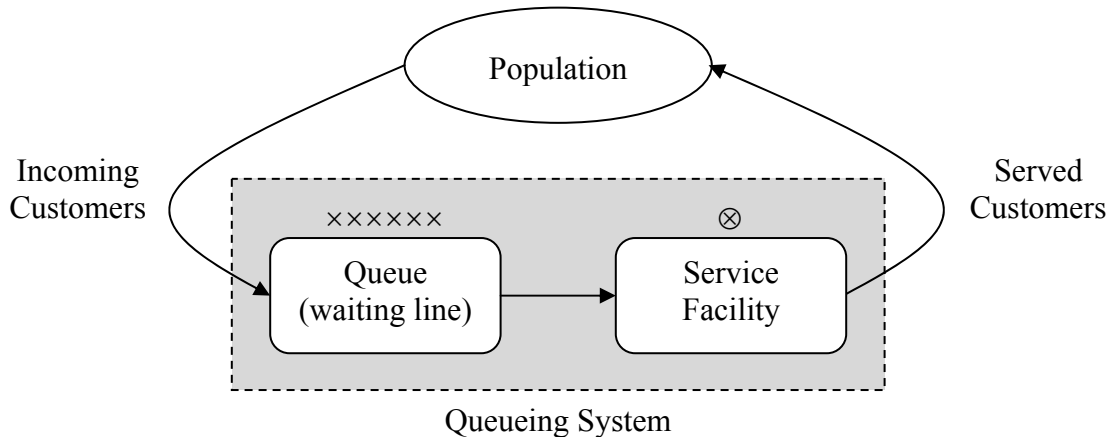
Figure 18.0.1: Basic queueing model

Q1.   List three queues in which you have spent time waiting. Indicate whether each queue you listed is either physical or nonphysical.

Q2.   Estimate how much time you spend waiting in queues during a typical week.

Q3.   Reflecting on the time you spend in queues, propose three strategies for reducing the amount of time customers spend waiting.

Q4.   From a manager's perspective, what are the downsides of your suggestions?

Not all queues are created equal. Some queues are designed to include distractions to make the time you spend waiting more pleasant. Other queues are quite boring and may lead you to frustration as you think about ways you could be using your time more productively.

Q5.   Propose three strategies for making the time spent in a queue more pleasant.

Q6.   Comparing your responses to numbers 3 and 5, which approach seems more cost effective?

Section 18.1: Post Office in Britton, MI—Single Server

The village of Britton, in southeast Michigan, has a population of 4500 people. Due to its small size, there is only one server at a time working at Britton's post office. The USPS has decided to cut back service at the counter to four hours on Saturday mornings. Historical data indicate that on average eight customers per hour arrive at the post office for service between 9 am and 1 pm. The single server in the post office is able to, on average, serve one customer every six minutes.

Q1.   What is the average rate of arrivals per unit of time? Include appropriate units in your answer.

Q2.   What is the average interarrival time (i.e., amount of time between arrivals)?

Q3.   While the server is continuously busy, what is the average rate of service per unit of time?

Q4.   What is the relationship between the average rate of customer arrivals and the average time between customer arrivals? Be sure to use the same unit of time to make the comparison.

In order to study waiting time in the post office in Britton, Dave Goldsman, a USPS supervisor, decided to collect data on one typical Saturday morning. Table 18.1.1 shows the data he collected for the first 30 customers who arrived at the post office on that Saturday morning. For each customer he recorded the time of arrival and the service time. From those two statistics, he was able to calculate the time service began, how long the customer waited in queue to begin service and the time in the system.

Dave Goldsman was specifically interested in two measures. The first measure was the waiting time until a customer begins service. This is called W_q. The second measure is the total time a customer spends in the post office. The sum of the time in queue plus the time it takes to be served and is called W. For example, let's look at the first two customers in the table.

The first customer arrived at 9:00:06, shortly after the post office opened at 9 a.m. and had no waiting time. The customer's service time was 9 minutes and 39 seconds. Before the customer completed service, another customer arrived at 9:04:41. The second customer began service immediately after the first customer left at 9:09:45. This customer waited in line 5 minutes and 4 seconds and was served in 6 minutes and 10 seconds. The total time this customer spent in the system was 11 minutes and 14 seconds.

Dave Goldsman summarized the data for the first 15 customers. These 15 arrived in the first 2 hour and 17 minutes. The average hourly arrival rate is calculated as

$$\frac{15}{2 + \frac{17}{60}} = \frac{15}{2.28\overline{3}} \approx 6.6 \text{ customers per hour}$$

Q5.   Explain why this does not match the 8 customers per hour mentioned in the beginning of this example.

Mr. Goldsman noticed there was significant variability in the time between arrivals.

Q6.   What were the shortest and longest time intervals between arrivals? Do these seem reasonable?

He averaged the service times and found the average was 5 minutes and fifty seconds

Q7.   How does this compare with the estimated average?

Q8.   What were the shortest and longest service times? Why do you think there might this amount of variability in the time it takes to service a customer?

Dave found the average waiting time in queue, *Wq,* to be 4 minutes and 16 seconds. However, he noticed that there was a great deal of variability in the waiting. Many customers did not wait at all.

Q9.   What percentage of customers did not have to wait?

Q10.   What was the longest wait? What events contributed to this long wait?

The average time in the system, *W*, was 10 minutes and 6 seconds.

Q11.   What two numbers combine to make up this total of 10 minutes and 6 seconds?

Dave Goldsman was also interested in how busy the postal worker was.

Q12.   What was the longest continuous time that the worker was busy with the first 15 customers?

Q13.   How many minutes did he spend serving customers between 9 am and 11 am? Include the customer who is the midst of being served at 11am.

Q14.   What percent of the time was the worker busy during his first two hours?

Mr. Goldsman did all of his analysis for the first 15 customers.

Q15.   Complete Table 18.1.1 for the next fifteen customers by calculating and recording the missing values.

Q16.   For this second group of 15 customers, what was their average arrival rate and average service time?

Q17.   For customers 16-30, what was the average waiting time in queue? What was their average time in the system?

Q18.   Which group of customers had the longer average waiting time in queue? Why do you think the two groups differ on this average?

Q19.   What was the overall average waiting time in queue for all 30 customers?

Table 18.1.1 contains a lot of numerical data and is a little unwieldy. It is difficult to gain much insight into the queueing system by staring at rows and columns of numbers. Calculating averages is one way to help make sense of all the data. Creating graphic displays allows for a more visual interpretation of the queueing system and its performance. In Figure 18.1.1 we see a snapshot of the state of the system. We can see if a customer is being served and how many are waiting in the queue.

Customer	Arrival Time	Time Service Begins	Time in Queue	Service Time	Departure Time	Time in System
1	9:00:06	9:00:06	0:00:00	0:09:39	9:09:45	0:09:39
2	9:04:41	9:09:45	0:05:04	0:06:10	9:15:55	0:11:14
3	9:33:17	9:33:17	0:00:00	0:04:24	9:37:41	0:04:24
4	9:42:49	9:42:49	0:00:00	0:02:12	9:45:01	0:02:12
5	10:00:42	10:00:42	0:00:00	0:02:41	10:03:23	0:02:41
6	10:03:17	10:03:23	0:00:06	0:04:34	10:07:57	0:04:40
7	10:27:55	10:27:55	0:00:00	0:05:34	10:33:29	0:05:34
8	10:38:40	10:38:40	0:00:00	0:10:11	10:48:51	0:10:11
9	10:41:09	10:48:51	0:07:42	0:09:08	10:57:59	0:16:50
10	10:47:46	10:57:59	0:10:13	0:08:49	11:06:48	0:19:02
11	10:54:36	11:06:48	0:12:12	0:08:32	11:15:20	0:20:44
12	10:58:57	11:15:20	0:16:23	0:01:40	11:17:00	0:18:03
13	11:10:11	11:17:00	0:06:49	0:01:13	11:18:13	0:08:02
14	11:16:50	11:18:13	0:01:23	0:02:56	11:21:09	0:04:19
15	11:16:55	11:21:09	0:04:14	0:09:47	11:30:56	0:14:01
		Averages	4:16	5:50		10:06

Customer	Arrival Time	Time Service Begins	Time in Queue	Service Time	Departure Time	Time in System
16	11:23:28			0:04:07		
17	11:28:42			0:05:50		
18	11:28:56			0:11:46		
19	11:35:37			0:11:24		
20	11:39:09			0:00:18		
21	11:49:36			0:07:06		
22	11:54:55			0:04:11		
23	12:13:59			0:00:27		
24	12:16:30			0:04:22		
25	12:22:10			0:06:02		
26	12:24:58			0:02:46		
27	12:29:33			0:03:03		
28	12:32:21			0:04:14		
29	12:39:44			0:07:35		
30	12:55:42			0:14:53		
		Averages				

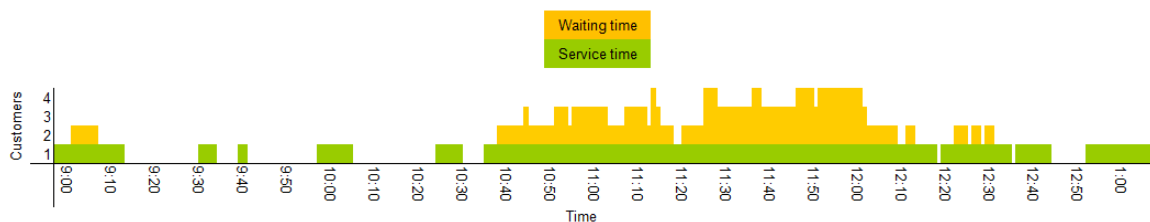**Table 18.1.1**: Collected data for post office queueing system

Figure 18.1.1: Number of customers present over time

Q20.   What is the greatest number of customers waiting at any one time during the interval shown?

Q21.   The clerk is allowed to take one ten minute break. When would have been a good time for him to have taken the break on this Saturday morning?

Q22.   Whose perspective did you use in answering the previous question?

Q23.   When was the clerk continuously busy?

Q24.   When was there at least one person waiting in line continuously?

Q25.   When was the longest continuous time when there were 3 people waiting?

Q26.   How might the number of customers waiting impact people who stopped by the post office?

18.1.1 Second Saturday

Dave Goldsman understood that each day was just one representation of what might happen. He therefore gathered a second set of data as presented in Table 18.1.2. On this Saturday there were 34 customers. This was slightly more than the average of eight per hour. The average service time was 7 minutes and 22 seconds as compared to forecasted average of 6 minutes. The average customer experience was very different from the first Saturday. The average time waiting was more than a half hour.

Q27.   How many customers came in each hour? How many hour periods were below average and how many were above average?

Assume the arrival pattern follows the Poisson distribution with an average of 8 per hour.

Q28.   What is the probability that there will 12 or more arrivals in an hour?

Q29.   What is the probability that there will be 5 or fewer in an hour?

Q30.   Based on these probabilities would it be surprising to have as many as 12 customers arrive in an hour or as few as 5 customers?

Q31.   Which customer waited the longest and how long was that wait to begin service?

Q32.   What percent of the customers waited more than 30 minutes for service?

Q33.   What was the average time a customer spent in the post office: waiting time plus service time?

Q34.   How did this arrival pattern contribute to the long waits?

Customer	Times					
	Arrival	Service Begins	In Queue	Service	Departure	In System
1	9:04:57	9:04:57	0:00:00	0:03:19	9:08:16	0:03:19
2	9:05:31	9:08:16	0:02:45	0:09:34	9:17:50	0:12:19
3	9:13:11	9:17:50	0:04:39	0:11:23	9:29:13	0:16:02
4	9:18:47	9:29:13	0:10:26	0:00:28	9:29:41	0:10:54
5	9:21:32	9:29:41	0:08:09	0:22:38	9:52:19	0:30:47
6	9:26:06	9:52:19	0:26:13	0:02:39	9:54:58	0:28:52
7	9:29:55	9:54:58	0:25:03	0:19:51	10:14:49	0:44:54
8	9:36:03	10:14:49	0:38:46	0:01:06	10:15:55	0:39:52
9	9:37:09	10:15:55	0:38:46	0:01:03	10:16:58	0:39:49
10	9:41:18	10:16:58	0:35:40	0:13:55	10:30:53	0:49:35
11	9:48:55	10:30:53	0:41:58	0:04:35	10:35:28	0:46:33
12	9:56:14	10:35:28	0:39:14	0:14:56	10:50:24	0:54:10
13	10:00:51	10:50:24	0:49:33	0:01:24	10:51:48	0:50:57
14	10:08:27	10:51:48	0:43:21	0:20:46	11:12:34	1:04:07
15	10:24:10	11:12:34	0:48:24	0:01:01	11:13:35	0:49:25
16	10:25:34	11:13:35	0:48:01	0:08:52	11:22:27	0:56:53
17	10:28:24	11:22:27	0:54:03	0:02:38	11:25:05	0:56:41
18	10:31:51	11:25:05	0:53:14	0:13:21	11:38:26	1:06:35
19	10:34:23	11:38:26	1:04:03	0:01:16	11:39:42	1:05:19
20	10:43:12	11:39:42	0:56:30	0:01:16	11:40:58	0:57:46
21	10:53:26	11:40:58	0:47:32	0:01:16	11:42:14	0:48:48
22	10:56:57	11:42:14	0:45:17	0:07:35	11:49:49	0:52:52
23	11:02:04	11:49:49	0:47:45	0:01:33	11:51:22	0:49:18
24	11:05:51	11:51:22	0:45:31	0:03:35	11:54:57	0:49:06
25	11:09:05	11:54:57	0:45:52	0:02:13	11:57:10	0:48:05
26	11:27:41	11:57:10	0:29:29	0:06:59	12:04:09	0:36:28
27	11:44:39	12:04:09	0:19:30	0:09:05	12:13:14	0:28:35
28	12:08:04	12:13:14	0:05:10	0:11:15	12:24:29	0:16:25
29	12:25:33	12:25:33	0:00:00	0:14:40	12:40:13	0:14:40
30	12:35:03	12:40:13	0:05:10	0:03:27	12:43:40	0:08:37
31	12:38:03	12:43:40	0:05:37	0:06:50	12:50:30	0:12:27
32	12:44:29	12:50:30	0:06:01	0:11:50	13:02:20	0:17:51
33	12:45:32	13:02:20	0:16:48	0:04:51	13:07:11	0:21:39
34	12:51:40	13:07:11	0:15:31	0:09:24	13:16:35	0:24:55
		Averages	**0:30:07**	**0:07:22**		

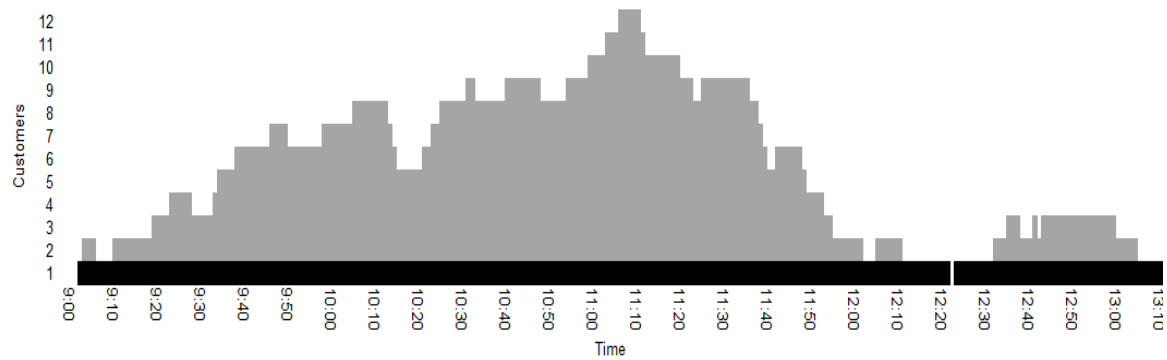**Table 18.1.2**: Second Saturday—collected data for post office queueing system

Figure 18.1.2: Number of customers present over time on the second Saturday

Q35.   How many people were waiting at 10:00, 11:00, 12:00, and 1 pm?

Q36.   What was the average of these four times? Why would you not average in the number at 9:00 am?

Q37.   What is the greatest number of customers waiting at any one time during the interval shown?

Q38.   The post office was set to close at 1 pm. Although no new arrivals were allowed in, the clerk could not close up. Why? How long was the clerk delayed in closing the post office?

Q39.   Describe a similar experience in which a service worker had to continue working even though the service was closed to new customers.

18.1.2 Building the Model

Dave Goldsman understood that the data he had collected represented two very different Saturdays. He also noted that on one day the average service time was below average and the number of customers was also below average. The opposite was true on the second Saturday. He decided to build a mathematical model to estimate average performance. He was also interested in knowing how long the waits would be during a busier time of the year when the average rate increases to 9 customers per hour and sometimes to 9.5 customers per hour. He decided to use queueing theory to determine the estimated average waiting time in queue and total time spent in the system. He understands that queueing theory models might not be appropriate for the first hour during the start-up of operations in the post office. It takes time before a queue can build up after a facility opens. He believes the models would be helpful in predicting average waiting times in the last few hours of the morning.

Rate Diagrams
Being an expert on queueing theory, Dave Goldsman first constructs a **rate diagram** of the queueing system to show how it makes transitions between various **states**. The state of a queueing system indicates the number of people in the queueing system (those waiting and the one being served) at any particular time. Only one person may arrive at the queue or depart from the system at a time. Thus, the only possible transitions in the state of the queueing system are to either increase by one or decrease by one.
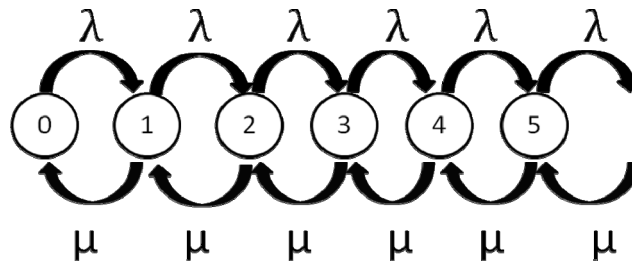
Figure 18.1.3: General rate diagram for a one-server queueing system

Figure 18.1.3 shows the general case using the variables λ and μ to represent the average arrival and service rates respectively. Figure 18.1.4 shows the specific case of the Britton, MI post office using the estimated arrival and service rates.
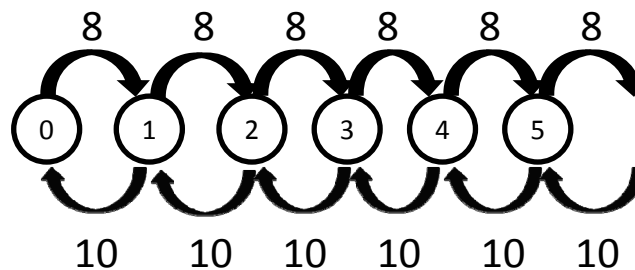

Figure 18.1.4: Specific rate diagram for one-server Post Office queueing system

Each node in the rate diagram represents the state of the queueing system. The number inside the node shows the number of people in the queueing system. The figure explicitly displays up to 5 people in the system. However, in theory, the number of potential customers waiting could be very large. The arrows continue so as to indicate there is no theoretical limit to that number. The mathematical assumption that the model is based on is that this maximum number approaches infinity.

Recall that when the state of the system is three, there is one customer being served and two more are waiting in the queue. Each directed arc represents the rate of transition between the states it connects. An arc leading to a higher-numbered state represents a customer entering the queueing system. An arc that leads to a lower-numbered state represents a customer, having been served, leaving the system.

To develop some understanding of the rate diagram, consider state 0 in Figure 18.1.4. There is only one arc leading out of state 0, 8 customers per hour. This represents the rate of customers arriving to the empty queueing system. They will start service without waiting. There is also only one arc leading into state 0, 10 customers per hour. This represents the rate of the only customer currently in the system being served and leaving the system.

Q40.   Describe the possible transitions *from* state 1.

Q41.   Describe the possible transitions *into* state 4.

In general the typical, long run average behavior of any queueing system is such that the average rate of people entering a queueing system has to be the same as the average rate of people leaving the system.

Q42.   If this general principle were not true of queueing systems, describe the two situations that could occur and their implications.

This is related to the *steady-state* concept discussed in the Markov Chain chapter. Given enough time, queueing systems approach a steady state. More specifically, when a queueing system is in steady-state, the rate at which the queueing system makes the transition from state i to state $(i + 1)$ must be equal to the rate at which the system makes the transition from state $(i + 1)$ to state i.

To have a practical analogy for this concept, consider a large department store whose only connections between floors are escalators. This store's lack of elevators makes it like a queueing system in terms of possible transitions. The only floors directly accessible to customers are the one immediately above and the one immediately below their current floor. Each day the store is open, the number of customers who took an escalator from the second floor to the third floor must be equal to the number of customers who took an escalator from the third floor to the second floor. During any particular hour, the rates of customers going from the second to the third floors and from the third to the second floors are not necessarily equal. It is only when considering the long term averages that the rates up and down between adjacent floors are equal.

Balance Equations—Rate Diagrams
The equation that represents this principle of long term averages is called the **balance equation**. There is a separate balance equation for each possible state of a queueing system.

We must not lose sight of our purpose. Dr. Goldsman is analyzing the queueing system to determine some measures of performance: average number of customers in the system and average length of time in the system. The critical values he needs to calculate are the probabilities of being in any particular state. We use P_n to denote the probability that the queueing system is in state n. Recall that n represents the number of people in the system. There are two ways to interpret P_n, and they are both important to understand.

One way to conceive of P_n is that it represents the proportion of time that the system is in state n. The other is that P_n is the probability of the system being in state n during any random observation or snapshot of the system.

Q43.   Interpret the meaning of the equation $P_0 = 0.20$ two ways.

Let's consider the balance equation for the transitions between states 0 and 1 using the specific rate diagram, Figure 18.1.3. The average rate from state 0 to state 1 is found by multiplying the probability of being in state 0, P_0, by the rate of departures from state 0 which is 8. The average rate from state 1 to state 0 is found by multiplying the probability of being in state 1, P_1, by the rate of arrivals to state 0 from state 1, which is 10. Therefore, the balance equation for the transitions between states 0 and 1 is as follows.
$$8P_0 = 10P_1$$

Table 18.1.3 below shows all the possible state transitions using the information from the rate diagram in Figure 18.1.4.

State i	State $(i+1)$	Rate from i to $(i+1)$	Rate from $(i+1)$ to i	Balance Equation
0	1	8	10	$8P_0=10P_1$
1	2	8	10	$8P_1=10P_2$
2	3	8	10	$8P_2=10P_3$
3	4	8	10	$8P_3=10P_4$
4	5	8	10	$8P_4=10P_5$
		8	10	
n	$n+1$	8	10	$8P_n=10P_{n+1}$
		8	10	

Table 18.1.3: Deriving balance equations

Q44.   Using λ and μ, write the general case of the balance equation between states 0 and 1.

Table 18.1.3 shows a system of equations with unknown probabilities. The first five equations with six unknowns are explicitly listed. There is also a general equation for states n and $n+1$.

Some algebraic manipulation will allow us to find these unknown probabilities. Consider the balance equation for the transition from state 0 to state 1. If we solve this equation for P_1, the result will be useful moving forward, as we shall see.

$$8P_0 = 10P_1$$
$$\left(\tfrac{8}{10}\right)P_0 = P_1$$

This means that P_1 will be 0.8 times the currently unknown value of P_0. Similarly we will need to isolate one of the probabilities in each of the balance equations.

State i	State $(i+1)$	Balance Equation	Solved for P_{i+1}	Substitute with P_0	General pattern
0	1	$8P_0 = 10P_1$	$\left(\frac{8}{10}\right)P_0 = P_1$		$\left(\frac{8}{10}\right)P_0 = P_1$
1	2	$8P_1 = 10P_2$	$\left(\frac{8}{10}\right)P_1 = P_2$	$\left(\frac{8}{10}\right)\left[\left(\frac{8}{10}\right)P_0\right] = P_2$	$\left(\frac{8}{10}\right)^2 P_0 = P_2$
2	3	$8P_2 = 10P_3$	$\left(\frac{8}{10}\right)P_2 = P_3$	$\left(\frac{8}{10}\right)\left[\left(\frac{8}{10}\right)^2 P_0\right] = P_3$	$\left(\frac{8}{10}\right)^3 P_0 = P_3$
3	4	$8P_3 = 10P_4$	$\left(\frac{8}{10}\right)P_3 = P_4$	$\left(\frac{8}{10}\right)\left[\left(\frac{8}{10}\right)^3 P_0\right] = P_4$	$\left(\frac{8}{10}\right)^4 P_0 = P_4$
4	5	$8P_4 = 10P_5$	$\left(\frac{8}{10}\right)P_4 = P_5$	$\left(\frac{8}{10}\right)\left[\left(\frac{8}{10}\right)^4 P_0\right] = P_5$	$\left(\frac{8}{10}\right)^5 P_0 = P_5$
⋀	⋀	⋀	⋀	⋀	⋀
$n-1$	n	$8P_{n-1} = 10P_n$	$\left(\frac{8}{10}\right)P_{n-1} = P_n$	$\left(\frac{8}{10}\right)\left[\left(\frac{8}{10}\right)^{n-1} P_0\right] = P_n$	$\left(\frac{8}{10}\right)^n P_0 = P_n$

Table 18.1.4: Algebraically manipulated balance equations

Making a series of algebraic substitutions, now we can write each balance equation in terms of P_0. Notice that the first such balance equation in Table 18.1.4 already shows the probability of being in state 1 in terms of the probability of being in state 0.

The process for deriving the expressions for the other probabilities is illustrated below for the balance equation for the transitions between states one and two. Earlier, we solved for P_1 in terms of P_0. All that is required is to substitute the earlier result into the following equation. For convenience, we use 0.8 in place of $\frac{8}{10}$ in each case.

> $(0.8)P_1 = P_2$, and $(0.8)P_0 = P_1$, so
> $(0.8)[(0.8)P_0] = P_2$
> $(0.8)^2 P_0 = P_2$

A similar substitution is used to determine P_3.

> $(0.8)P_2 = P_3$
> $(0.8)[(0.8)^2 P_0] = P_3$
> $(0.8)^3 P_0 = P_3$

In general for any value of n, a similar substitution is used to determine P_n.

> $(0.8)P_{n-1} = P_n$
> $(0.8)[(0.8)^{n-1} P_0] = P_n$
> $(0.8)^n P_0 = P_n$

Solve for P_0

As a result of these substitutions, we can write every state probability as a function of P_0. We need an equation that can be used to find P_0. These probabilities represent all of the possible states of the system and therefore must sum to one:

> $P_0 + P_1 + P_2 + P_3 + P_4 + P_5 + \ldots + P_n + \ldots = 1$

Using the results from table 18.1.4, we can substitute and rewrite the above equation such that every term is written as a function of P_0:

> $P_0 + 0.8^1 P_0 + 0.8^2 P_0 + 0.8^3 P_0 + 0.8^4 P_0 + 0.8^5 P_0 + \ldots + 0.8^n P_0 + \ldots = 1$
> $(1 + 0.8^1 + 0.8^2 + 0.8^3 + 0.8^4 + 0.8^5 + \ldots + 0.8^n + \ldots)P_0 = 1$

This equation contains an infinite number of terms. The terms inside the parentheses form an infinite geometric series, because any two consecutive terms have the same ratio. Recall that the sum of an infinite geometric series is:

$$S = \frac{a}{1-r}$$

In this equation, a is the first term in the sum and r is the common ratio used to find each succeeding term. In this case, the first term is 1 and the common ratio is 0.8.

$$S = \frac{1}{1-0.8} = \frac{1}{0.2} = 5$$

Now, substituting 5 for the infinite series, we have

$$5P_0 = 1 \rightarrow P_0 = 0.2$$

Average Server Utilization: ρ
This means that that the probability that an arriving customer will not have to wait is 0.2. This also means that the server is not busy 20% of the time. Let's see why this answer makes sense. What is the workload of the server? On average 8 customers arrive per hour. On average each customer requires 6 minutes of service. Thus on average the server will be busy 48 minutes per hour. Thus, the server is busy 80% of the time and therefore not busy on average 20% of the time. Notice that 0.8 is the ratio of the arrival rate to the service rate. This ratio is called the average server utilization and represented by Greek letter ρ. This equals the fraction of time the server is busy.

$$\rho = \tfrac{8}{10}$$

For the general case,

$$\rho = \frac{\lambda}{\mu} \text{ and } P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho.$$

The specific probability for each state can then be determined:
$P_0 = 0.2$
$P_1 = 0.8\, P_0 = 0.8(0.2) = 0.16$
$P_2 = 0.8^2 P_0 = 0.8^2(0.2) = 0.128$
$P_3 = 0.8^3 P_0 = 0.8^3(0.2) = 0.102$
$P_4 = 0.8^4 P_0 = 0.8^4(0.2) = 0.082$
$P_5 = 0.8^5 P_0 = 0.8^5(0.2) = 0.066$

n	P_n	Cumulative P_n	nP_n	Cumulative Sum
0	0.200	0.2	0.000	0.000
1	0.160	0.36	0.160	0.160
2	0.128	0.488	0.256	0.416
3	0.102	0.59	0.307	0.723
4	0.082	0.672	0.328	1.051
5	0.066	0.738	0.328	1.379
6	0.052	0.79	0.315	1.693
7	0.042	0.832	0.294	1.987
8	0.034	0.866	0.268	2.255
9	0.027	0.893	0.242	2.497
10	0.021	0.914	0.215	2.712
11	0.017	0.931	0.189	2.900
12	0.014	0.945	0.165	3.065
13	0.011	0.956	0.143	3.208
14	0.009	0.965	0.123	3.331
15	0.007	0.972	0.106	3.437
16	0.006	0.978	0.090	3.527
17	0.005	0.983	0.077	3.604
18	0.004	0.987	0.065	3.669
19	0.003	0.99	0.055	3.723
20	0.002	0.992	0.046	3.769
21	0.002	0.994	0.039	3.808
22	0.001	0.995	0.032	3.841
23	0.001	0.996	0.027	3.868
24	0.001	0.997	0.023	3.890
25	0.001	0.998	0.019	3.909
26	0.001	0.999	0.016	3.925
27	0.000		0.013	3.938
28	0.000		0.011	3.949
29	0.000		0.009	3.958
30	0.000		0.007	3.965
31	0.000		0.006	3.971
32	0.000		0.005	3.977
33	0.000		0.004	3.981
34	0.000		0.003	3.984
35	0.000		0.003	3.987
36	0.000		0.002	3.989
37	0.000		0.002	3.991
38	0.000		0.002	3.993
39	0.000		0.001	3.994
40	0.000		0.001	3.995

Table 18.1.5: Calculation of expected queue length

18.1.3 Number of People in System and Waiting

No one will be waiting to begin service whenever the system is in either state 0 or state 1. This probability is $0.2 + 0.16$ for a total probability 0.36. On average more than one-third of the time, no one will be waiting for service.

In the data that Dave collected, the maximum number of people in the post office was 4. Dave wants to know the long-term probability that the number of people in the post office will exceed 4.

Q45.   What is the probability that there will be four or fewer people in the post office?

Q46.   What is the probability that there will be five or more people in the post office?

More importantly Dave Goldsman is interested in the average number of people in the post office and the average time they spend in the post office. The average length of the line is represented by L. The average time spent in the system is W. Recall that the long-term average of a random variable is its expected value. The basic definition of expected value is the probabilistic weighted sum of the observed values. In this problem the random variable is the number of people, n, in the post office.

$$L = \Sigma n\, P_n = 0P_0 + 1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 + \ldots + nP_n$$

$$L = 0P_0 + 1(0.8^1 P_0) + 2(0.8^2 P_0) + 3(0.8^3 P_0) + 4(0.8^4 P_0) + 5(0.8^5)P_0 + \ldots + n(0.8^n P_0) + \ldots$$

$$L = [0 + 1(0.8^1) + 2(0.8^2) + 3(0.8^3) + 4(0.8^4) + 5(0.8^5) + \ldots + n(0.8^n) + \ldots]\, P_0$$

In the general case the traffic intensity is represented as ρ and P_0 equals $(1 - \rho)$. Thus,

$$L = [0 + 1(\rho^1) + 2(\rho^2) + 3(\rho^3) + 4(\rho^4) + 5(\rho^5) + \ldots + n(\rho^n) + \ldots]\,(1 - \rho)$$

This equation requires significant algebraic manipulation to determine its sum. We provide the results in the equation below.

$$L = \left(1 - \rho\right)\left(\frac{\rho}{\left(1 - \rho\right)^2}\right) = \frac{\rho}{1 - \rho}$$

The general equation for the average number of people in line reduces mathematically to just the ratio of the fraction of time the server is busy, ρ, divided by the fraction of time the server is not busy, $(1 - \rho)$. In our example, $\rho = 0.8$, and

$$L = \frac{0.8}{1 - 0.8} = \frac{0.8}{0.2} = 4$$

Thus the long-term average exceeds the observed average in the one Saturday morning sample.

John D.C. Little developed a general formula that applies to every queueing system. The formula, known as Little's Law, relates the W, the average time in the system, to L, the average number of people in the queue:

$$L = \lambda W$$

As a way to understand this relationship, imagine a customer who has just arrived and joined the queueing system. On average the time that individual spends in the system will be W. Now imagine the instant that individual completes service. How many customers will have arrived during the individual's time in the system, W? On average, since customers arrive at a rate of λ per hour, while spending W hours in the system, the average total number of customers arriving is λW. Thus, as the individual leaves and looks back over the shoulder to see L, the number of people in the system, there should be on average λW people.

Similarly if we know L, we can determine W by dividing both sides of the equation by λ:

$$W = \frac{L}{\lambda}.$$

In this case, customers arrive at a rate of 8 per hour and the average number of customer in the system is 4. Thus,

$$W = \frac{4}{8} = 0.5 \text{ hours}$$

On average customers will spend a half hour when visiting the post office on a Saturday morning.

W represents the total average time a person spends in the system. That time includes two components, W_q, the average time spent waiting in line, and the average service time. Thus, the average time in the system, W, equals the average spent waiting to be served, W_q, plus the average time in service, $\frac{1}{\mu}$.

$$W = W_q + \text{average service time} = W_q + \frac{1}{\mu}$$

Similarly, this equation may be manipulated to obtain a formula for W_q:

$$W_q = W - \frac{1}{\mu}$$

In this instance the average service time is 6 minutes or 0.1 hours.

$$W_q = W - \frac{1}{\mu} = 0.5 - 0.1 = 0.4 \text{ hours or 24 minutes}$$

Dave Goldsman was very concerned. He considered an average of 24 minutes waiting to begin service as too long. He also wondered what would happen when the average arrival rate was higher. He was also concerned that at certain times in the year, people come in with more stuff and the average time to service a customer increase.

Q47.   Complete the following table.

λ	μ	L	W	W_q
8	10	4	0.5	0.4
More Arrivals				
8.5	10			
9	10			
Slower Service				
8	9			
8	8.5			

Table 18.1.6: Calculate queue and waiting times

18.1.4 Options to Improve Performance

Dave is considering the options he has to improve service. The three options he is considering are:

A. A 40-hour retraining program designed to improve the efficiency of the postal worker. Typically, participants can reduce their average service time by 1 minute after completing the program. That will cost the postal service $2500.

B. A redesign of the workplace and the addition of some specialty equipment will reduce the average service time by 1.5 minutes. The cost of these changes is estimated to be $5,000.

C. He is considering increasing the hours of operation by one more hour to five hours. Currently on average 8 customers arrive per hour for a total of 32 per Saturday. Dave believes that by keeping the office open an extra hour, the total will increase to 35 customers over the five hour period. Keeping the office open an extra hour each Saturday, will cost an average of $65 per Saturday. The office is expected to be open 50 Saturdays per year.

Dave carries out the following analysis.

Option A → The average service is reduced from 6 minutes to 5 minutes.

The average service rate is now 12 customers per hour, $\rho = \frac{8}{12} = 0.667$, and $L = \frac{\rho}{1-\rho} = \frac{0.667}{0.333} = 2$.

Q48. Use Little's Law to calculate the average number of customers in the system, and find the average number of customers waiting in queue under Option A.

Option B → The average service is reduced from 6 minutes to 4.5 minutes. The average service rate is now 13.3 customers per hour. $\rho = \frac{8}{13.33} = 0.6$ and $L = \frac{\rho}{1-\rho} = \frac{0.6}{0.4} = 1.5$

Q49. Use Little's Law to calculate the average number of customers in the system, and find the average number of customers waiting in queue under Option B.

Option C → The average arrival is reduced from 8 customers per hour to 7 per hour. $\rho = \frac{7}{10} = 0.7$ and $L = \frac{\rho}{1-\rho} = \frac{0.7}{0.3} = 2.33$

Q50. Use Little's Law to calculate the average number of customers in the system, and find the average number of customers waiting in queue under Option C.

Q51. Consider the cost and the impact on waiting. Which would you recommend?

Q52.   What justification can be provided for considering option C?

Option	Added Cost	W	W_q
Current	0	30	24
A	$2500		
B	$5000		
C	$3250		

Table 18.1.6: Evaluate post office options

18.1.5 Merge Two Office—2 Servers

Dave decided to consider a broader view of the services provided in the region. There is another small post office in Macon Township, a ten minute drive from Britton. It services approximately the same number of people. It too provides only limited hours on Saturdays. He is considering recommending that on Saturdays the Macon office close and that the worker from Macon Township join the worker at the Britton office. This will save the post office an estimated $45 per Saturday in reduced utility and cleanup expenses. There will be no labor savings as the worker will simply work in a different location but the same number of hours.

As a result, there will now be two servers in the Britton office. For simplicity sake, we will assume that all of the customers who would otherwise use the Macon office on Saturday will make the short drive to Britton instead.

The first task is to create the new rate diagram. The new average arrival rate is 16 customers per hour.

Q53.   Why does an arrival rate of 16 describe the new situation?

Notice that the rate diagram in Figure 18.1.4 shows an average service rate of 10 customers per hour for the Post Offices when the system is in state 1. However, it is 20 customers per hour whenever there are two or more customers in the system

Q54.   Why does this representation of the service rate describe the new situation?
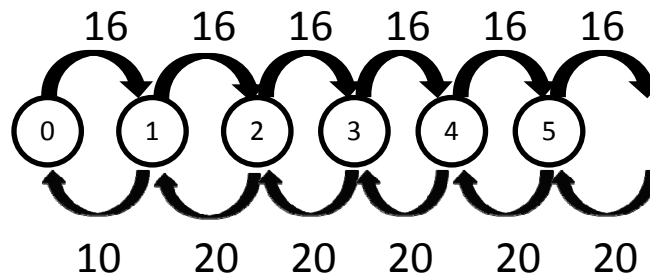

Figure 18.1.4: Specific rate diagram for two-server post office queueing system

In the table below we present the balance equations for each pair of neighboring states. In the General Pattern column, the number, "2" appears as a multiplier on the left side of the equation.

Q55.   What is the justification for introducing a "2"?

State i	State $(i+1)$	Balance Equation	Solved for P_{i+1}	Substitute with P_0	General pattern
0	1	$16P_0=10P_1$	$\frac{16}{10}P_0 = P_1$		
1	2	$16P_1=20P_2$	$\frac{16}{20}P_1 = P_2$	$\left(\frac{16}{20}\right)\left(\frac{16}{10}\right)P_0 = P_2$	$2\left(\frac{16}{20}\right)^2 P_0 = P_2$
2	3	$16P_2=20P_3$	$\frac{16}{20}P_2 = P_3$	$\left(\frac{16}{20}\right)\left[2\left(\frac{16}{20}\right)^2 P_0\right] = P_3$	$2\left(\frac{16}{20}\right)^3 P_0 = P_3$
3	4	$16P_3=20P_4$	$\frac{16}{20}P_3 = P_4$	$\left(\frac{16}{20}\right)\left[2\left(\frac{16}{20}\right)^3 P_0\right] = P_4$	$2\left(\frac{16}{20}\right)^4 P_0 = P_4$
4	5	$16P_4=20P_5$	$\frac{16}{20}P_4 = P_5$	$\left(\frac{16}{20}\right)\left[2\left(\frac{16}{20}\right)^4 P_0\right] = P_5$	$2\left(\frac{16}{20}\right)^5 P_0 = P_5$
⋮	⋮	⋮	⋮	⋮	⋮
n–1	n	$16P_{n-1} = 20P_n$	$\frac{16}{20}P_{n-1} = P_n$	$\left(\frac{16}{20}\right)\left[2\left(\frac{16}{20}\right)^{n-1} P_0\right] = P_n$	$2\left(\frac{16}{20}\right)^n P_0 = P_n$

Table 18.1.7: Two-server algebraically-manipulated balance equations

Q56.   How does the introduction of "2" simplify the general pattern representation?

As noted previously, the random variable is the number of people, n, in the post office. The definition of L is unchanged

$$L = \Sigma n\, P_n = 0P_0 + 1P_1 + 2P_2 + 3P_3 + 4P_4 + 5P_5 + \ldots + nP_n$$

$$L = 0P_0 + 1(2)(0.8^1 P_0) + 2(2)(0.8^2 P_0) + 3(2)(0.8^3 P_0) + 4(2)(0.8^4 P_0) \\ + 5(2)(0.8^5)P_0 + \ldots + n(2)(0.8^n P_0) + \ldots$$

$$L = [0 + 1(2)(0.8^1) + 2(2)(0.8^2) + 3(2)(0.8^3) + 4(2)(0.8^4) \\ + 5(2)(0.8^5) + \ldots + n(2)(0.8^n) + \ldots]\, P_0$$

Figure 18.1.5 is the general rate diagram for a 2 server queueing systems. We have labeled the arrival rate λ_c so as to reflect that this arrival rate is the combined arrival rates of the two post offices.
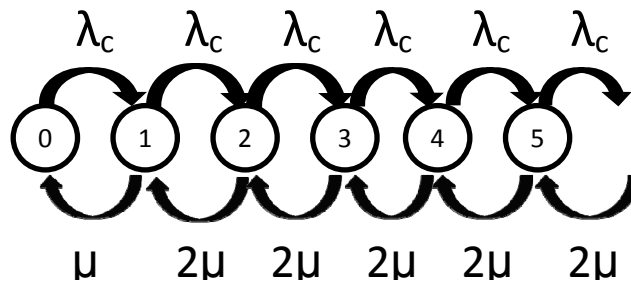

Figure 18.1.5: General rate diagram for Two-server Post Office queueing system

The average utilization of the two servers will be

$$\rho = \frac{\lambda_c}{2\mu}$$

In this particular case,

$$\rho = \frac{16}{2 \cdot 10} = \frac{16}{20} = 0.8$$

Q57.   Why is this utilization rate for the two servers the same as for the individual post offices in the previous section?

For the general case the equation for the average number of people in the system is

$$L = [0 + 1(2\rho^1) + 2(2\rho^2) + 3(2\rho^3) + 4(2\rho^4) + 5(2\rho^5) + \ldots + n(2\rho^n) + \ldots] P_0$$

This equation requires significant algebraic manipulation to determine its sum. We provide the results in the equation below.

$$L = \frac{2\rho}{1-\rho^2}$$

In our post office example the average utilization is 0.8 and the average number of people in the combined post office is now

$$L = \frac{2(0.8)}{1-(0.8)^2} = \frac{1.6}{0.36} = 4.44$$

Q58.   Is this better or worse than the two separate post offices? Explain.

Little's formula for relating L and W applies to all queueing systems. We can determine W with
$$L = \lambda W$$

$$W = \frac{L}{\lambda}$$

With the combined post offices, customers arrive at a rate of 16 per hour. Thus,

$$W = \frac{4.4}{16} = 0.28 \text{ hrs} = 16.7 \text{ min}$$

The average time waiting to begin service is now

$$W_q = W - \frac{1}{\mu} = 0.28 - \frac{1}{10} = 0.28 - 0.10 = 0.18 \text{ hrs} = 10.7 \text{ min}$$

Option	Cost Change	W	W_q
Current	0	30	24
A: Training	+$2500	15	10
B: Redesign	+$5000	11	6.8
C: More Hours	+$3250	20	14
Merged	-$2250	16.7	10.7

Table 18.1.8: Merged Post Office Option

Let's think a little more deeply about the situation from the perspective of people in Britton and those near Macon who would on average have to travel ten minutes further.

Q59.   If you lived in Macon, would you prefer the current situation or a Saturday merger?

Q60.   How would your answer change if you regularly shopped in Britton on Saturdays?

Consider the cost and the impact on both groups.

Q61.   What would you recommend?

Dave reexamined the options and realized that combining two options might even be better. He considered merging the two offices on Saturday and sending both postal workers for retraining. The savings from the merger would pay in part for the extra training for two workers.

Q62.   What is the net cost of this combination of options?

Recall that retraining reduces the average service time to 5 minutes. As a result the average utilization is

$$\rho = \frac{16}{2 \cdot 12} = \frac{16}{24} = 0.67$$

$$L = \frac{2\rho}{1-\rho^2} = \frac{2(0.67)}{1-(0.67)^2} = 2.4$$

With the combined post offices, customers arrive at a rate of 16 per hour. Thus,

$$W = \frac{2.4}{16} = 0.15 \text{ hrs} = 9 \text{ min} .$$

The average time waiting to begin service is now

$$W_q = W - \frac{1}{\mu} = 9 - 5 = 4 \text{ min}$$

Consider the cost and the impact on both groups.

Q63.   Now, what would you recommend?

Section 18.2: Airport Security Screening

The Transportation Security Agency (TSA) is concerned about long back-ups in the security screening area at Charlene York International Airport (CYA). The area has 7 x-ray screening devices. TSA can vary the number of staff to operate these devices up to a maximum of seven. In the morning, the number of screeners can be adjusted every hour. Any unneeded screeners are assigned other tasks.

As customers go through the screening area, they move through a series of security screening steps. First, their boarding pass and id are checked. Then, they wait to join one of the screening lines. There they wait to approach a long counter where they begin unloading their personal items into bins to be passed through an x-ray detection device. As their personal items are screened, they pass through a metal detection device. The primary backlog is created by the bins passing through the x-ray equipment. Each bin is looked at carefully. There is significant variability as the number of bins and contents vary from customer to customer. On occasion people have to be specially checked or pass through the metal detector a second time.

Ralph Waldo is in charge of personnel staffing for CYA. Ralph Waldo's data collectors found that the average time for this x-ray and personal screening was 45 seconds. However, at times of heightened security, this average could increase to 52 seconds per customer.

Average Screening Service Time = 45 seconds					
Time Interval		Customer Arrivals		Number of Screeners	
Begin	End	Total	Per Minute	Minimum	Min. plus 1
6:00 AM	7:00 AM	315	5.25	4	5
7:00 AM	8:00 AM	375	6.25	5	6
8:00 AM	9:00 AM	470	7.83		
9:00 AM	10:00 AM	500	8.33		
10:00 AM	11:00 AM	430	7.17		
11:00 AM	Noon	390	6.50		

Table 18.2.1: Airport arrivals and number of screeners

His staff has collected data on the total number of arrivals in hour intervals from 6 am until noon. For example between 6 am and 7 am on average 315 customers arrive daily. The average arrival rate per minute is therefore 5.25. The average service time per customer is 45 seconds. The total screening workload per minute is 5.3 multiplied by 45 seconds. This equals 3.94 minutes of work each minute. To complete this work will require a minimum of 4 screeners available each minute. Similarly, 7 am and 8 am on average 375 customers arrive daily. The average arrival rate per minute is therefore 6.25. The average service time per customer is 45 seconds. The total screening workload per minute is 6.25 multiplied by 45 seconds. This equals 4.69 minutes of work each minute. To complete this work will require a minimum of 5 screeners available each minute.

Q1.   Complete the above table to determine the minimum number of screeners required each hour.

Ralph Waldo is concerned that staffing to the minimum could result in long waits to be screened. He therefore, asks Dr. N. Queue to determine the wait in queue for the minimum number of screeners and one more than the minimum number.

Little's law relates the average number in the queueing system to the average waiting time. It relates L_q and W_q as well as L and W. Ralph Waldo is interested in W_q. The relevant formula is

$L_q = \lambda W_q$

Dr. N. Queue applied the appropriate formula for L_q as a function of ρ, the average utilization. This utilization is determined by dividing the arrival rate by the total service capacity. This service capacity is the rate of service by one server multiplied by the number of servers, c. Thus the average utilization is

$$\rho = \frac{\lambda}{c\mu}$$

The average utilization with 4 screeners is determined as follows. Since the arrival rate is customers per minute, we need to define the service rate in the same units of measurement. The average service time, 45 seconds, equals 0.75 minutes. This is equal to $1/\mu$. Thus the service rate is 4/3 customers per minute. Then,

$$\rho = \frac{5.25}{4 \cdot \frac{4}{3}} = 0.984.$$

If there were 4 screeners, they would be busy on average 98.4% of the time. If there were 5 servers, the corresponding value is 78.8% of the time:

$$\rho = \frac{5.25}{5 \cdot \frac{4}{3}} = 0.788.$$

Q2.   Help Dr. N. Queue complete the table below to determine the average utilization for each hour for the different number of screeners.

Time Interval		Minimum			Minimum Plus 1		
Begin	End	Screeners	ρ	W_q	Screeners	ρ	W_q
6:00 AM	7:00 AM	4	0.984	11.6	5	0.788	0.4
7:00 AM	8:00 AM	5	0.938		6	0.782	
8:00 AM	9:00 AM	6			7		
9:00 AM	10:00 AM	7			8		
10:00 AM	11:00 AM	6			7		
11:00 AM	Noon	5			6		

Figure 18.2.2: Airport screening and waiting in line

Dr. N. Queue found the formula for L_q with 4 servers in *Modeling Random Processes for Engineers and Managers* by James J. Solberg.

$$L_q = \frac{32\rho^5}{3 + 6\rho + 3\rho^2 - 4\rho^3 - 8\rho^4}$$

She applied this formula to the first time period, 6 am to 7 am.

$\rho = (0.984)$

$$L_q = \frac{32(0.984)^5}{3 + 6(0.984) + 3(0.984)^2 - 4(0.984)^3 - 8(0.984)^4}$$

$L_q = 60.83$ customers

She then calculated:

$$W_q = \frac{L_q}{\lambda} = \frac{60.83}{5.25} = 11.6 \text{ minutes}$$

In summary, assume there are four screening machines fully staffed and operating this first hour. If John Q. Public arrived at 6:47 am, he would see on average more than 60 customers waiting in line ahead of him. His average time in queue would be less than 12 minutes. This did not seem unreasonably long for Ralph Waldo.

Q3.   Explain why it will take on average less than 12 minutes to service 60 customers in line.

Ralph Waldo was also interested in determining the impact of staffing one more x-ray machine. Dr. N. Queue found the formula for L_q with 5 servers:

$$L_q = \frac{625\rho^6}{24 + 72\rho + 84\rho^2 + 20\rho^3 - 75\rho^4 - 125\rho^5}$$

It was obvious to her that the formulas were getting more and more complicated. Dr. Queue wondered if there was some approximate formula that would be easier to use. She found the following approximation in Solberg's book.

$$L_q = \frac{\rho^{\sqrt{2(c+1)}}}{1-\rho}$$

Dr. Queue decided to see how good the approximation was by first applying it to the 4 server example. Here are her calculations using the approximation.

$$L_q = \frac{(0.984)^{\sqrt{2(4+1)}}}{1-(0.984)} = 60.89 \text{ customers}$$

$$W_q = \frac{L_q}{\lambda} = \frac{60.89}{5.25} = 11.6 \text{ minutes}$$

Dr. Queue was excited to find that the result was less than 0.1% different from the exact value. The nice thing about the formula is it does not get more complicated as the number of servers increases. Before proceeding to use the formula for all of the analysis, she decided to check out the approximation for 5 servers between 7 am and 8 am.

$$\rho = \frac{5.25}{5\left(\frac{4}{3}\right)} = 0.788$$

The exact formula resulted is applied below.

$$L_q = \frac{625(0.788)^6}{24 + 72(0.788) + 84(0.788)^2 + 20(0.788)^3 - 75(0.788)^4 - 125(0.788)^5}$$

$$L_q = 1.97$$

The average number of customers in queue drops dramatically to an average of less than 2 customers. The average time in queue would now be less than 1 minute.

$$W_q = \frac{L_q}{\lambda} = \frac{1.97}{5.25} = 0.4 \text{ minutes}$$

The approximate formula yielded an estimate of 2.06 customers. This was 4.7% higher. Dr. N. Queue felt this number was close enough to justify using the approximate formula for the rest of the analysis.

$$L_q = \frac{(0.788)^{\sqrt{2(5+1)}}}{1-(0.788)} = 2.06$$

Ralph Waldo needed to make a tough decision. The cost of staffing and operating one addition x-ray machine was $45 per hour. Was it worth spending that much money to almost eliminate any waiting time for customers arriving between 6 am and 7 am.

Q4.   What would you recommend and why?

Dr. N. Queue hired a local high school student, Matt Wannabee, to complete the table. Matt was also curious about the approximation. He therefore decided to check it out for himself using the 7am to 8 am data and 5 servers. Matt was also going to time how long it took him to do the exact and approximate calculations and obtain the correct answer. (With complex calculations such as these, mistakes are common and the calculations may need to be redone. If the answers differ by more than 5%, you have made a mistake. Determine the total time until you obtain the correct answer.)

Q5.   Determine L_q by using the exact formula for 5 servers and $\rho = 0.938$. How long did it take to obtain this number?

Q6.   Determine L_q by using the approximate formula for 5 servers and $\rho = 0.938$. What is the percentage difference? How long did it take?

Q7.   Use the approximate formula to complete Table 18.2.2

The completed results were presented to Ralph Waldo. Now he had to decide how many x-ray machines to staff each hour. Should he staff the minimum each hour or should he sometimes add one more screener at a cost of $45 per hour.

Q8.   He reviewed his budget and decided he could go above the minimum required staffing for just one hour. What hour would you recommend and why?

18.2.1 Heightened Security Screening

At least once a month, CYA is ordered on heightened security. As a result the average time to screen a customer increases to 52 seconds. Although this does not seem like much of an increase, Ralph Waldo is concerned about the possibility of long back-ups. Dr. Queue has told him the waiting time is a steep function of the average utilization and is not a straight line.

Q9.   He asked Matt Wannabee to start with the basic calculations to determine the Minimum number of screeners. Complete the Minimum column in Table 18.2.3.

Q10.   Mr. Waldo planned to use the minimum number of screeners. How many more would he need each hour when compared to the minimum for a regular day?

Q11.   Ralph took a closer look and noticed CYA was going to have a problem. What problem did Ralph
        Waldo find?

Average Screening Service Time = 52 seconds					
Time Interval		Arrivals Per Minute	Minimum Screeners	ρ	W_q
Begin	End				
6:00 AM	7:00 AM	5.25			
7:00 AM	8:00 AM	6.25			
8:00 AM	9:00 AM	7.83			
9:00 AM	10:00 AM	8.33			
10:00 AM	11:00 AM	7.17			
11:00 AM	Noon	6.5			

Figure 18.2.3: Heightened security and waiting in line

Whenever, possible Ralph Waldo planned to use the minimum number of screeners. He asked Matt to
complete the table.

Q12.   Using the approximation, Matt Wannabee proceeded to determine L_q and then W_q for each hour.
        Complete Table 18.2.3.

Ralph Waldo is considering hiring some low-paid part-time aids to help during heightened security. These
aids would assist passengers and help ensure that they moved efficiently through the screening area. It is
projected these aids would reduce the average service time to 50 seconds.

Q13.   Would these extra personnel reduce the required minimum number of screeners?

Q14.   Would they be used every hour or for only some of the hours? Explain.

Section 18.3: Customer Complaints about Front and Center Ticket Sales

Amit Juan is the chief operating officer for Front and Center Ticket Sales (FACTS). FACTS sells tickets to a wide variety of events at area venues. Each FACTS retail outlet consists of a single sales counter inside a music store within a shopping mall. Their competitors are national Internet-based ticket sellers. Amit is concerned about recent customer complaints about long lines at their retail locations

Q1.   How could complaints about long lines translate into a decrease in profit for FACTS?

To address the complaints, Amit hires a queueing consultant, Dr. Jack Small. Most of the complaints are from customers seeing long lines and waiting in queues at malls on Friday afternoons, so that is where Dr. Small begins.

Before making any recommendations, Dr. Small needs to observe the queueing system and collect some data.

Q2.   What data would be useful for Dr. Small?

In addition to spending several Friday afternoons making observations and collecting data, Dr. Small also makes two important assumptions about the behavior of the queueing system:
* individual customers arrive to purchase tickets randomly, and
* the time to complete a purchase is also random.
The first assumption seems reasonable, so Amit accepts it, but he is skeptical about the second one.

Q3.   List three reasons Dr. Small might cite to explain why the length of time needed to complete a ticket sale transaction might vary.

In addition to the arrival and service times being random, they are distributed exponentially. An exponential distribution is related to the Poison distribution in Chapter 16.

18.3.1 Data and Variables

Dr. Small observed that all of FACTS's retail outlets consist of a single server working behind a ticket window. The results of Dr. Small's data collection were that, on average, 12 customers arrived per hour and each customer spent, on average, four minutes being served. Dr. Small also notices that many potential customers left either before joining the queue or shortly after joining it. She also noted that the largest number of customers she observed in line including the one in service was 4. She, therefore, decides to assume that this is a finite capacity queue with a maximum of 4 customers in the queue.

Q4.   Given these findings, is it surprising that customers were complaining about spending too long waiting in the queue?

Q5.   If 20 customers arrive per hour, on average, how much time elapses between successive arrivals?

Q6.   If λ customers arrive per hour, on average, how much time elapses between successive arrivals?

Q7.   If it takes the server four minutes to serve one customer, on average, how many customers can be served per hour?

Q8.   If it takes the server $\frac{1}{\mu}$ hours to serve one customer, on average, how many customers can be

served per hour?

Just as in the infinite capacity case, the parameter ρ is defined as $\rho = \frac{\lambda}{\mu}$. This ratio is the relative rate of

arrivals as compared to the service rate. However, in the finite capacity system, this is not the *utilization rate*. In a finite capacity the rate of arrivals can exceed the service rate because when the system is at capacity, no more arrivals enter the system. In this situation, we can use an alternative term for ρ, the *traffic intensity*.

In this example, when a customer sees four customers already in the system, Dr. Small found that the typical customer chooses not to buy tickets now.

Q9.   With four customers already in line, how long would a new arrival have to wait on average to
        reach the ticket agent? Would you wait this long to buy tickets?

Obtaining a value for ρ is vital to developing a mathematical model of a queue. For the value of ρ to have any meaning, it is necessary for λ and μ to be expressed in the same units.

Q10.   What is the average arrival rate, λ, of the queue at FACTS on Friday afternoons?

Q11.   What is the average service rate, μ, of the queue at FACTS on Friday afternoons?

Q12.   What is the *traffic intensity*, ρ, of the queue at FACTS on Friday afternoons?

Q13.   What are the units of the *traffic intensity*?

18.3.2 Building the Model

Q14.   Describe the possible transitions from state one.

Q15.   Describe the possible transitions from state two.

Q16.   Describe the possible transitions from state four.

Q17.   In which state is it appropriate to consider potential customer whose business was lost?

Dr. Small uses an important principle of queueing as the foundation of his model: when the queueing system reaches a **steady-state** (the typical, long-run average behavior of the system), the average rate of people entering a queueing system has to be the same as the average rate of people leaving the system. The general rate transition diagram in Figure 18.3.1 is the same as Figure 18.1.2 except that it now stops at state 4. The equation that represents this principle is called the **balance equation**. There is a separate balance equation for each possible state of a queueing system.
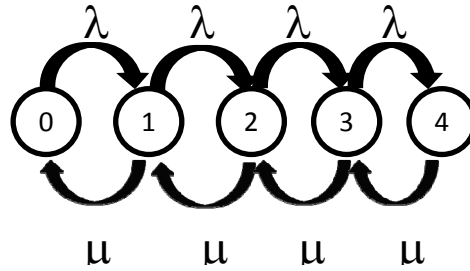
Figure 18.3.1: General four state finite capacity rate diagram

The balance equations for this diagram are the same as depicted in Table 18.1.3 except that there is no state past state 4. The balance equation for the transitions between states zero and one is given below. We can solve for P_1 in terms of P_0. We can reorganize the equation and substitute ρ for the ratio, λ/μ.

$$\lambda P_0 = \mu P_1$$

$$\frac{\lambda P_0}{\mu} = \frac{\mu P_1}{\mu}$$

$$\frac{\lambda}{\mu} P_0 = P_1$$

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$P_1 = \rho P_0$$

State i	State $(i+1)$	Balance Equations			
		Balance Equation	**Solved for P_{i+1}**	**General Pattern in Terms of P_0**	**Function of ρ**
0	1	$\lambda P_0 = \mu P_1$	$P_1 = \dfrac{\lambda}{\mu} P_0$	$P_1 = \dfrac{\lambda}{\mu} P_0$	$P_1 = \rho P_0$
1	2				
2	3				
3	4				

Table 18.3.1: Finite capacity balance equations and general pattern

Complete the list of balance equations in the column marked Equality. Then write each term P_i as a function of P_0. Lastly, substitute ρ for the ratio, λ/μ.

Q18.   What is the general formula for P_n, the probability of the queueing system being in state n?

As a result of the data Dr. Small collected over the course of her observations of FACTS's queueing system on Friday afternoons, we know the values of λ and μ (see questions 10 and 11). Therefore, the unknown quantities in the balance equations are the probabilities of being in any particular state. This is

not a directly observable feature, so Dr. Small must calculate it somehow. A property of probability is the key.

Q19.   What is the probability that the queueing system is in one of its possible states;
        that is, $P_0 + P_1 + P_2 + P_3 + P_4 = ?$

Combining your answers to questions 18 and 19, we have the following equation, which can be solved for P_0.

$$P_0 + \rho^1 P_0 + \rho^2 P_0 + \rho^3 P_0 + \rho^4 P_0 = [1 + \rho^1 + \rho^2 + \rho^3 + \rho^4] P_0 = 1$$

$$P_0 = \frac{1}{1 + \rho + \rho^2 + \rho^3 + \rho^4}$$

We know the value of ρ, $\rho = \frac{20}{15} \approx 1.333$, and we can determine the probability of being in each of the states 0 through 4.

Q20.   Calculate the probability of the queueing system being in each of its possible states.

State (*n*)	P_n
0	0.104
1	0.139
2	0.185
3	0.246
4	0.327

Q21.   Confirm that $\sum_{i=0}^{4} P_i = 1$

Customers who arrive when the system is in state 4 will balk; that is, they will not join the queue in order to buy tickets.

Q22.   What proportion of customers chooses not to buy tickets at this ticket counter?

In the earlier infinite capacity example, the proportion of time the server was busy was simply ρ or equivalently $1 - P_0$. This is not true in the finite capacity example. For that we need to consider when the server is busy and not busy.

Q23.   In which state(s) will the server be busy and in which state(s) will the server not be busy?

Q24.   What proportion of time is the server busy?

18.3.3 Some Results

Two of the key measures of the performance of a queueing system are the expected number of customers in the queue, L, and the expected waiting time for a customer in the queueing system, W. These are the factors that Amit has to take into consideration in assessing the significance of the customer complaints he has received. If potential customers see a long line, they may not join the queue to purchase tickets. If they experience long waits while in the queueing system, they may not return as customers. Amit needs to

decide if the potential loss in business is large enough to warrant opening a second service window at FACTS's retail locations.

To calculate L for a queueing system with capacity n, Dr. Small uses the formula below.

$$L = \sum_{i=0}^{n} i \cdot P_i$$

Q25.   What is the expected number of customers in the queueing system at FACTS retail locations on Friday afternoons?

To calculate W, Dr. Small uses Little's Law.

$L = \bar{\lambda} W$, where $\bar{\lambda}$ is the average rate of arrivals for each possible state

The difference between λ and $\bar{\lambda}$ has to do with the fact that there are no arrivals when the system has reached its maximum capacity. The formula for calculating $\bar{\lambda}$ for a queueing system of with capacity n is shown below.

$$\bar{\lambda} = \lambda\left(1 - P_n\right)$$

Q26.   What is the value of $\bar{\lambda}$ for the queueing system at FACTS?

Q27.   What is a customer's expected waiting time in the queueing system (include units in your answer)?

This represents the total in the system. Part of that total time is the average time to process the individual's ticket request. That was found to be an average of 4 minutes.

Q28.   What is W_q, a customer's expected time waiting in the queue until (s)he reaches the ticket counter?

Q29.   From a customer's perspective, do the expected number of people in the system and expected waiting time in the system seem reasonable?

18.3.4 Second Server at FACTS

Amit is concerned enough about the queueing system's performance and customer complaints to ask Dr. Small to determine the impact of opening a second service window during the Friday afternoon rush. The parameters of the system, λ and μ, will not be affected by the introduction of a second server. However, with two servers each serving, on average, one customer every four minutes, the customers will move through the queueing system faster. When working simultaneously, they can process an average of two customers every four minutes or one every two minutes. Dr. Small begins her analysis by constructing a new rate diagram for the two-server model. Because of the extra serve, she believes that customers would be willing to wait even if there are more than four people are in the system. Dr. Small sets the limit at 7 customers before assuming that customers will go elsewhere. Figure 18.3.2 is a finite capacity modification of the earlier rate diagram Figure 18.1.5 for the infinite capacity case.
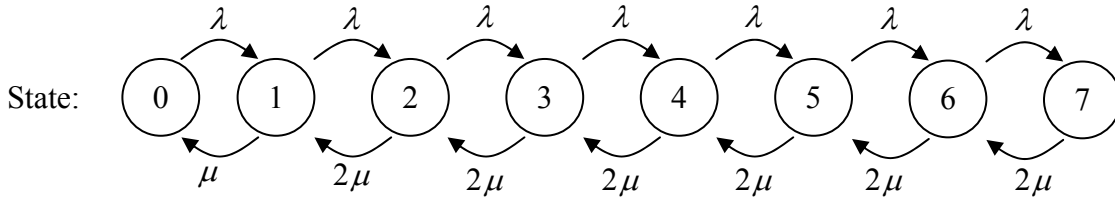
Figure 18.3.2: Rate diagram for a two-server model queueing system at FACTS

Q30.   Explain why most of the rates on the service arcs are 2μ.

Q31.   Explain why the rate on the leftmost service arc is μ.

Because Dr. Small already has the data for the queueing system, it is merely a matter of recalculating the measures of performance.

Q32.   Complete Table 18.3.2 below to show all the balance equations in a useful form. This parallels the infinite capacity example in Table 18.1.7

State i	State $(i+1)$	Rate from i to $(i+1)$	Rate from $(i+1)$ to i	Balance Equation	Balance Equation Solved for P_{i+1}	Balance Equation in Terms of P_0
0	1	λ	μ	$\lambda P_0 = \mu P_1$	$P_1 = \dfrac{\lambda}{\mu} P_0$	$P_1 = \dfrac{\lambda}{\mu} P_0$
1	2	λ	2μ	$\lambda P_1 = 2\mu P_2$	$P_2 = \dfrac{1}{2}\left(\dfrac{\lambda}{\mu}\right) P_1$	$P_2 = \dfrac{1}{2}\left(\dfrac{\lambda}{\mu}\right)^2 P_0$
2	3	λ	2μ	$\lambda P_2 = 2\mu P_3$	$P_3 = \dfrac{1}{2}\left(\dfrac{\lambda}{\mu}\right) P_2$	$P_3 = \left(\dfrac{1}{2}\right)^2\left(\dfrac{\lambda}{\mu}\right)^3 P_0$
3	4					
4	5					
5	6					
6	7					

Table 18.3.2: Balance equations for two server scenario

Q33.   What is the general formula for the probability of a two-server queueing system being in state n, P_n?

Q34.   Explain why the equation below is true.

$$\frac{\lambda}{\mu}P_0 + \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 P_0 + \left(\tfrac{1}{2}\right)^2\left(\frac{\lambda}{\mu}\right)^3 P_0 + \left(\tfrac{1}{2}\right)^3\left(\frac{\lambda}{\mu}\right)^4 P_0 + \left(\tfrac{1}{2}\right)^4\left(\frac{\lambda}{\mu}\right)^5 P_0 + \left(\tfrac{1}{2}\right)^5\left(\frac{\lambda}{\mu}\right)^6 P_0 + \left(\tfrac{1}{2}\right)^6\left(\frac{\lambda}{\mu}\right)^7 P_0 = 1$$

Q35.   Substitute the known values of λ and μ to solve the above equation for P_0.

Q36.   Use the value of P_0 to calculate the probability of the two-server queueing system being in each
of its possible states.

State (n)	P_n
0	
1	
2	
3	
4	
5	
6	
7	

Q37.   Use the formula below to calculate the expected number of customers in the queueing system, L.

$$L = \sum_{i=0}^{n} i \cdot P_i$$

Q38.   Use the formula below to calculate the average rate of arrivals for each possible state, $\overline{\lambda}$.
$$\overline{\lambda} = \lambda\left(1 - P_n\right)$$

Q39.   Use Little's formula, shown below, to determine a customer's expected waiting time in the
queueing system.

Q40.   What is W_q, the average time until reaching the ticket counter?

In the earlier infinite capacity example, the proportion of time the servers were busy was simply ρ or
equivalently $1 - P_0$. This is not true in the finite capacity example. For that we again need to consider
when the servers are busy and not busy.

Q41.   In which states will both servers be busy?

Q42.   In which state will both servers not be busy?

Q43.   In which state will one server be busy and the other not be busy?

Let's assume that the two servers share all of the work equally.

Q44.   What proportion of time is each server busy and conversely not busy?

Q45.   What would you recommend the servers do with this spare time without affecting the service they
provide?

18.3.5 Economic Analysis—Diminishing Returns

Just looking at the expected waiting times, Amit notices that under the current one-server model, a customer is likely to wait 7.4 minutes to begin processing a ticket request. By adding a second server, that number is reduced to 2.21 minutes. More importantly the percentage of customers who refuse to join the queue drops from 32.4% to 2.4%. He has to decide if adding a second server will increase profit enough to warrant the additional expenses.

The expenses FACTS will incur to change from a one-server to a two-server ticket outlet include about $2,000 per location to purchase additional computers and other hardware. The major expense will be the $15 per hour that will have to be paid to the new employees. Each FACTS location is open from 10:00 a.m. to 9:00 p.m. on Monday through Saturday, and 10:00 a.m. to 6:00 p.m. on Sunday. The Friday afternoon rush lasts from 2:00 p.m. until 9:00 p.m. Instead of doubling the staffing at the FACTS retail locations full time, Amit is considering using a second server just for Friday afternoons.

Q46.   What is the added operating cost per location per week of hiring an additional part-time server?

Recall from question 21 that the only time customers are lost from the system is when the system has reached its capacity. For the Friday afternoon rush under the one-server model, P_4 is about 32.4%. This means that an average of about 1 in 3 customers is lost. With an additional server P_7 was reduced to only about 2.4%. That means on average only about 1 in 40 customers is lost.

Q47.   Why did Dr. Small originally consider the capacity of the queueing system to be four?

Q48.   Is that reasoning still valid under the two-server model? Why or why not?

In trying to decide what to do, Amit assumes that 32.4% of customers will be lost if he stays with a single-server model, and that 2.4% customers will be lost if he upgrades to a two-server model. His sales data suggest that each location averages about 1,400 ticket orders per week and the average ticket order generates $6 of profit. The average number of ticket orders during the Friday afternoon rush is 85. These ticket sales are typically for greater numbers of tickets, so they are also more profitable, generating an average of $8 per order.

Q49.   Calculate lost profit per location under the assumptions of the single-server model for one Friday afternoon.

Q50.   Should Amit decide to switch to a part-time two-server model?

Section 18.3: Chapter 18 (Queueing Theory) Homework Questions

Under development

Chapter 18 Summary

What have we learned?

We have learned about several types of queueing systems and how to gauge their performance. When studying a queueing system, you first need to determine some important features.

- How many servers does the queueing system have?

- What is the capacity of the system?

- What is the average arrival rate of the customers?

- What is the average service rate of the server(s)?

To analytically study a queueing system, several steps must be taken.

- Collect data to determine the parameters (λ and μ) of the system.

- Construct a mathematical model of the system.

    o Draw a rate diagram

    o Write balance equations

Once you have some basic information about a queueing system, you need to perform some calculations to determine its performance.

- Write and manipulate balance equations.

- Calculate P_0, L, W_q, and other performance measures.

After analyzing a queueing system, you must make decisions, weighing customer satisfaction against cost.

- Are the waiting times acceptable?

- Do there need to be more servers?

- Are there other ways to improve customers' experiences in the system?

Terms

Queue	A queue is a waiting line
Queueing Theory	Queueing theory is the mathematical study of queues that operations researchers and industrial engineers use to increase the efficiency of queueing systems
Queueing System	A queueing system is made up of all the customers waiting in the queue and all the customers being served
Customer	In queueing theory, a customer typically refers to a person either waiting in the queue or being served. Taken more broadly, anything that waits for some kind of service (e.g., packages being shipped, airplanes on a runway, emails being routed, parts being manufactured) can be considered customers
Server	In the typical sense, a server is a person who provides customers with the service for which they are waiting. Servers could also be machines, computer programs, or anything else that provides service to customer
State	The state of a queueing system refers to the number of customers in the system at any particular time
Rate Diagram	The rate diagram of a queueing system shows the possible states of the system and the rates of transition between adjacent states
Balance Equation	A balance equation shows that, in the long-run, the average rate of customers leaving a particular state must equal the average rate of customers entering that state
Average Server Utilization	The average server utilization is the percentage of time the server will be busy in infinite-capacity queueing systems. In finite-capacity systems, this parameter (ρ) is referred to as the *traffic intensity*

Notation

W	Average waiting time in the system
W_q	Average waiting time in the queue
L	Average number of customers in the system
L_q	Average number of customers in the queue
n	State of the system
P_n	Probability the system is in state n
λ **(lambda)**	Average arrival rate
$\dfrac{1}{\lambda}$	Average time between arrivals
λ	Average arrival rate for each possible state (only used in finite-capacity queueing systems)
μ **(mu)**	Average service rate
$\dfrac{1}{\mu}$	Average service time
ρ **(rho)**	Average traffic intensity/average server utilization
c	Number of servers in a queueing system

Formulas

Formula	Single-server, infinite-capacity	c-servers, infinite-capacity	Single-server, n-capacity
	Type of system		
$\lambda P_m = \mu P_{m+1}$	✓		✓ $(m < n)$
$\lambda P_m = c \cdot \mu P_{m+1}$		✓ $(m > c)$	
$\rho = \dfrac{\lambda}{\mu}$	✓		✓
$\rho = \dfrac{\lambda}{c\mu}$		✓	
$P_n = \rho^n P_0$	✓		✓
$P_n = c \cdot \rho^n \cdot P_0$		✓	
$P_0 = 1 - \rho$	✓		
$P_0 = \dfrac{1}{\sum_{i=0}^{n} \rho^i}$			✓
$L = \dfrac{\rho}{1-\rho}$	✓		
$L = \dfrac{2\rho}{1-\rho^2}$		✓ (only for $c = 2$)	
$L_q = \dfrac{\rho^{\sqrt{2(c+1)}}}{1-\rho}$		✓ (approximation)	
$L = \sum_{i=0}^{n} i \cdot P_i$			✓
$L = \lambda W$	✓	✓	
$L_q = \lambda W_q$	✓	✓	
$L = \bar{\lambda} W$			✓
$L_q = \bar{\lambda} W_q$			✓
$\bar{\lambda} = \lambda(1 - P_n)$			✓
$W_q = W - \dfrac{1}{\mu}$	✓	✓	✓

Chapter 18 (Queueing Theory) Objectives

You should be able to:

- Determine the number of servers in and the capacity of a queueing system

- Design a data collection plan for studying a queueing system

- Know the meanings of all notation used in queueing theory

- Construct a rate diagram for a queueing system

- Write and manipulate balance equations

- Apply appropriate formulas to determine the performance measures for a queueing system

- Propose and evaluate various methods for improving a queueing system

Chapter 18 Study Guide

1. Explain the difference between a *queue* and a *queueing system*.

2. A *parameter* is defined as "a quantity that describes a statistical population." In this chapter, we have referred to average arrival rate (λ), average service rate (μ), and average server utilization (ρ) as parameters of queueing systems. Explain the importance of these values and why they are referred to as *parameters*.

3. To study a queueing system, what type of data would need to be collected?

4. From a manager's perspective, what are the issues that need to be balanced when considering if and how to improve a queueing system?

5. A portion of a queueing system's rate diagram and the related balance equation are shown below. Explain the principle behind the balance equation.

| Rate Diagram | Balance Equation |

$$18P_2 = 21P_3$$
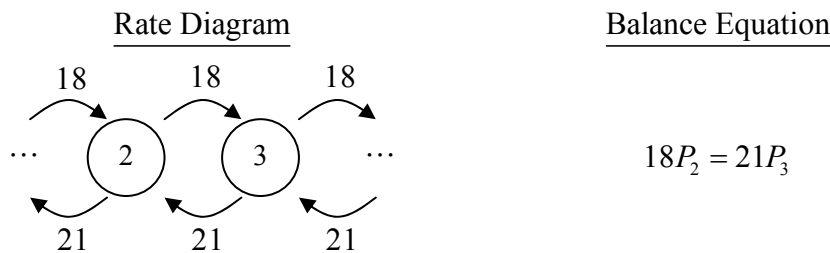
6. If W is the amount of time (in hours) that a customer spends in the queueing system, W_q is the amount of time (in hours) the customer spends waiting in the queue, and μ is the average service rate (in customers per hour), explain why the following formula makes sense. Include in your explanation the units involved.

$$W_q = W - \frac{1}{\mu}$$

7. As a queueing system consultant, list four ways a queueing system could be improved, whether it would improve the performance measures (as opposed to improving customers' experiences in the system) of the system, and whether the cost of implementing the improvement would be high or low.

| Improvement	Improve performance?	Cost

8. Discuss the pros and cons of using the formula $L_q = \dfrac{\rho^{\sqrt{2(c+1)}}}{1-\rho}$.

9. The only possible values of the *average server utilization* parameter in a queueing system are $0 < \rho < 1$. Discuss the implications if ρ took on a value outside this interval.

If $\rho < 0$, then

If $\rho = 0$, then

If $\rho = 1$, then

If $\rho > 1$, then

Section 19.0 Introduction

In Chapter 10, a tool for managing projects called the Critical Path Method (CPM) was shown. In this chapter a new tool is introduced called **Program Evaluation Review Technique (PERT)**.

A Little History

The Special Projects Office in the Department of the Navy created PERT in 1958. It was designed to manage the development of the Polaris Missile and Submarine program. As a result of this innovative project management method, the Polaris Missile and Submarine program was completed two years ahead of schedule.
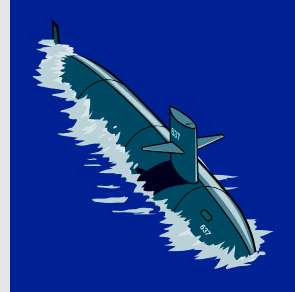
The Critical Path Method (CPM) and the Program Evaluation Review Technique (PERT) share many characteristics. For example, they both utilize activity dependence, critical paths, and slack time.

However, these two techniques were developed independently of each other. In fact, PERT was developed before CPM. PERT is generally used for the research and development of projects with a large amount of uncertainty in activity completion times. For example, industries with new or rapidly changing technology use PERT. In comparison, CPM is more appropriate for projects that have been completed repeatedly. For example, companies that build pre-fabricated houses or manufacture popular car models use CPM.

In this chapter, the ideas from the CPM chapter are built upon to demonstrate how one can plan, monitor, and control a project even if he/she does not know exact completion times of activities.

Section 19.1 Preparing a Taco Dinner

Recall the "Preparing a Taco Dinner" problem from Chapter 10. In the original problem, Maya and her friends are given exact completion times for each activity based upon the average time it takes for different chefs to prepare this meal (see Figure 19.1.1). Maya uses CPM to determine the start and finish times of each activity, and she finds the critical path of activities needed to get the dinner cooked and served. The steps of CPM are reviewed below.

> **CPM Algorithm**
>
> **Step 1**: Compute the earliest start time (EST) and the earliest finish time (EFT) for each activity in the order-requirement digraph.
>
> **Step 2**: Compute the latest start time (LST) and the latest finish time (LFT) for each activity in the order-requirement digraph.
>
> **Step 3**: Compute the slack for each activity in the order-requirement digraph.
>
> **Step 4**: Identify the critical path and find the slack along each path.

Activity Code	Activity Description	Activity Dependence	Completion Time (minutes)
A	Find recipe for taco salad on Internet	—	5
B	Find recipe for dessert in Mom's recipe box	—	6
C	Make graham cracker crust	B	4
D	Clean lettuce	A	3
E	Find cheese in refrigerator	A	1
F	Shred cheese	E	4
G	Brown hamburger and assemble tacos (including cheese)	D, F	14
H	Prepare the pudding	B	8
I	Assemble the pudding pie	C, H	4
J	Put out all the food for dinner	G, I	2

Table 19.1.1: Activity chart for making a taco dinner

Using the CPM algorithm, Maya discovers that it takes 26 minutes to complete the dinner and the critical path of activities is *A-E-F-G-J.* This means that the activities that determine how much time is needed to prepare dinner and the order in which they should be accomplished are: *Find recipe for taco salad on Internet*, *Find cheese in the refrigerator*, *Shred cheese*, *Brown hamburger and assemble tacos*, and *Put out all the food for dinner*.

The other activities are done simultaneously, but these activities will not keep Maya and her friends from putting dinner on the table.

This problem context can also be used to introduce the Program Evaluation Review Technique (PERT), another type of project planning algorithm.

Unless Maya and her friends have cooked this exact dinner many times, they really do not know exactly how much time each activity will take. While professional chefs are a good gauge of completion time,

chances are the time it takes the chefs to prepare, cook, and serve this dinner is different from the time it takes Maya and her friends.

PERT allows Maya to answer the question "How much time will it take us to prepare a taco dinner?" in a way that allows for flexibility in completion times.

For each activity, there are three completion times that Maya needs to take into consideration: the shortest possible time (or the optimistic time), the longest possible time (or the pessimistic time), and the most probable time.

Maya assumes that the completion times from Chapter 10 are the most probable times (m) to complete the activities of preparing a taco dinner in Table 19.1.1 since these times are an average of different chefs' activity completion times. However, she also needs to consider a range of possible completion times. Table 19.1.2 is the same chart as Table 19.1.1 but with columns added for shortest and longest possible preparation times.

Activity Code	Activity Description	Activity Dependence	Completion Time (minutes)			
			Shortest Possible Time (a)	Most Probable Time (m)	Longest Possible Time (b)	Expected Time (t_E)
A	Find recipe for taco salad on Internet	—		5		
B	Find recipe for dessert in Mom's recipe box	—		6		
C	Make graham cracker crust	B		4		
D	Clean lettuce	A		3		
E	Find cheese in refrigerator	A		1		
F	Shred cheese	E		4		
G	Brown hamburger and assemble tacos (including cheese)	D, F		14		
H	Prepare the pudding	B		8		
I	Assemble the pudding pie	C, H		4		
J	Put out all the food for dinner	G, I		2		

Table 19.1.2: Activity chart for making a taco dinner for PERT

Q1.   Work with a partner and fill in the columns for shortest possible time (a) and longest possible time (b) in Table 19.1.2 (ignore the Expected Time column (t_E) for now). You may use whole numbers or fractions; however, the numbers must be greater than zero. To make the problem more interesting, try not to make the numbers on either side of the Most Probable Time (m) be the exact same distance from that middle value. Note: you and your partner's times may differ from your classmates' times.

Now that the shortest and longest possible completion times have been determined, Maya must calculate the **expected completion time** (t_E) for each activity. Expected time refers to time Maya *expects* to finish by when she accounts for a range of possible times. This may or may not be the same as the most probable time. For instance, if it takes Maya and her friends much longer to complete an activity, the expected time will be longer than the most probable time; if Maya and her friends finish an activity quickly, the expected time will be shorter than the most probable time.

The expected completion time is a weighted average, where the most probable time gets more weight in the calculation. Specifically, the most probable time has a weight of 4 while the shortest and longest times each have weights of 1. These weights are based on the Beta distribution, which is discussed briefly in the following section.

Formula for Expected Completion Time:

$$t_E = \frac{a + 4m + b}{6},$$

where
- a = the shortest completion time,
- b = the longest completion time, and
- m = the most probable completion time.

For example, consider activity A (find a recipe for taco salad on the Internet). Suppose the shortest time Maya and her friends could complete activity A is 2 minutes ($a = 2$) and the longest it would take is 6 minutes ($b = 6$). Then the expected completion time for activity A is

$$t_E = \frac{(2) + 4(5) + (6)}{6} = \frac{28}{6} \approx 4.67 \text{ minutes}.$$

Therefore, Maya and her friends can expect that it will take approximately 4.67 minutes to find the recipe for taco salad on the Internet. This is slightly shorter than the most probable time.

Q2.   Using the numbers you filled in for shortest and longest possible activity completion times, calculate the expected activity completion time (t_E) for each activity and fill in the last column of Table 19.1.2.

Q3.   The expected activity completion time (t_E) should be slightly different values than the most probable time (m). When will these two values be the same?

Q4.   Based on your expected completion times, do you think it will take Maya and her friends more or less than 26 minutes (the length of the critical path) to complete the taco dinner?

Next, Maya creates an order-requirement digraph modeling the activity dependence of preparing a taco dinner project, as shown in Figure 19.1.1. She uses the t_E time (rather than the most probable time) as the weight of each node. From there, she can find the critical path of the project and the estimated project completion time.
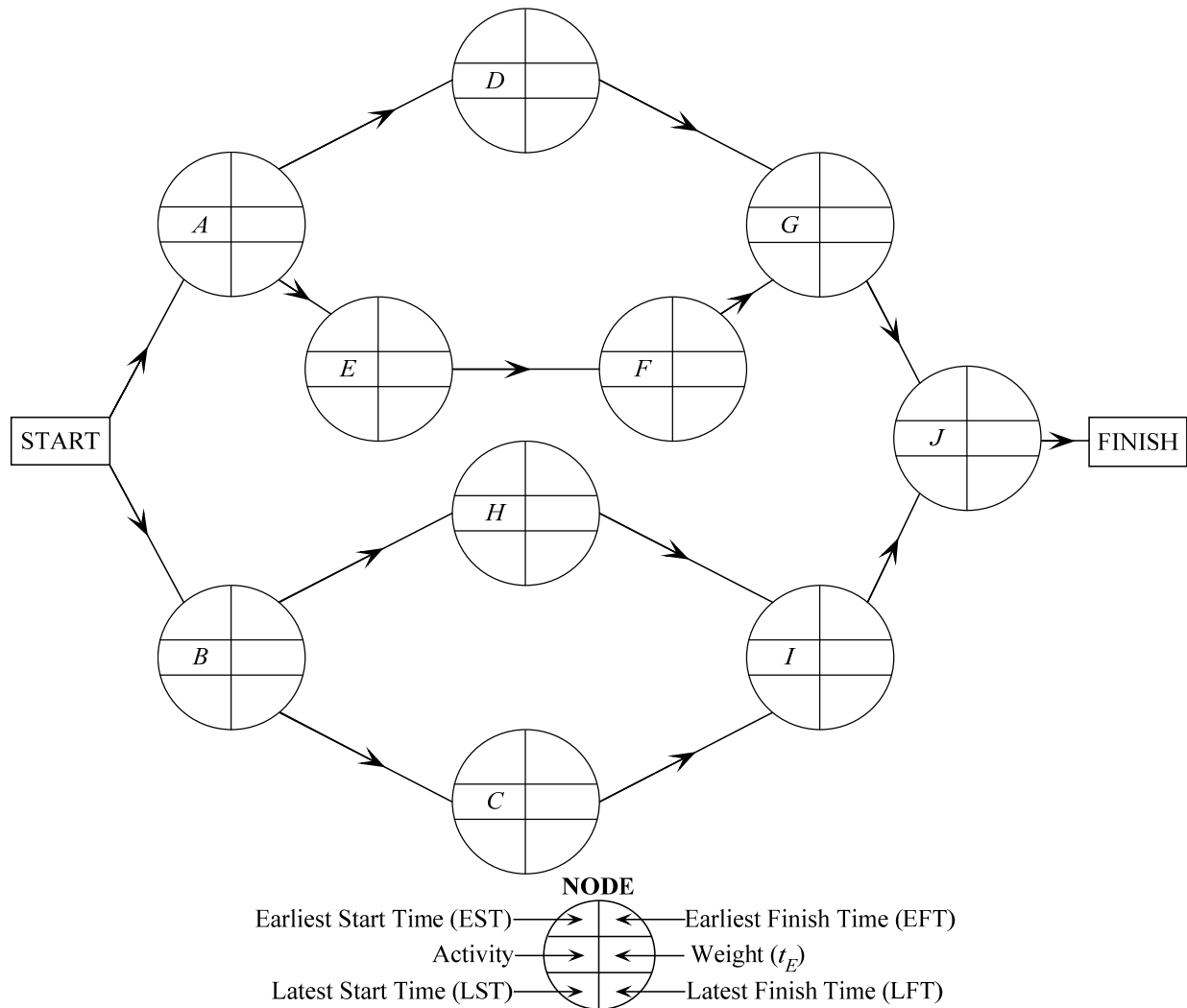
NODE

Earliest Start Time (EST)———→ ←———Earliest Finish Time (EFT)
Activity———→ ←——— Weight (t_E)
Latest Start Time (LST)———→ ←———Latest Finish Time (LFT)

Figure 19.1.1: Order-requirement digraph representing preparing a taco dinner

Q5.   Using Figure 19.1.1, fill in the expected times (t_E) as the weight of the order-requirement digraph.

Q6.   Calculate the earliest start time (EST), earliest finish time (EFT), latest start time (LST), and latest finish time (LFT) for each activity using the same algorithm as for Critical Path Method (CPM) in Chapter 10.

Q7.   Determine the critical path for preparing a taco dinner. How long will Maya and her friends to complete the project?

Q8.   How is your answer different from the answer from Chapter 10 [*A-E-F-G-J*, 26 minutes]? How did your choice of shortest and longest completion times change your answer? Why did it change?

Q9.   Thinking about the process, what would have happened if each of your shortest and longest times were the same distance from the most probable values in your activity table?

Thus, PERT has some of the same features as CPM from Chapter 10, but it goes one-step further because it considers a range of completion times for each activity. In the next section, a more formal example of PERT is presented.

Section 19.2 Relocation of CHEMPACK containers for the Super Bowl

The Super Bowl is a popular sporting event in the U.S. More than 70,000 people sit in the stadium on the Sunday game day in addition to thousands of fans gathered nearby to experience the atmosphere. With so many people crowded together in one place, the Super Bowl could be a potential target for a bioterrorist attack.

The public health department needs to be prepared in case such a horrible event should occur. They will plan for a worst-case scenario and determine the best way to distribute needed supplies if such an event did occur.

A CHEMPACK container is one type of response item in a bioterrorist attack. One CHEMPACK container stores enough doses of antidote to help about 1,000 people exposed to various types of nerve agents or chemical weapons. They are currently stored in hospital emergency departments all over the country. Because the antidotes are only effective right after exposure, the public health department needs large quantities of CHEMPACK containers nearby so that all exposed people can get the antidote quickly if there is an attack on Super Bowl Sunday. The public health department must decide how to move CHEMPACK containers from various parts of the state and country to facilities near the Super Bowl.

Angel Meluso, who works for the public health department, wants to know when to start planning to move the CHEMPACK containers so that they are in location for game day. He does not want to move them too early because he does not want to leave other areas of the state and country vulnerable. However, he definitely needs the CHEMPACK containers positioned in time for game day. He also wants to know which activities of the positioning are critical to his timetable. A problem that arises in planning to move CHEMPACK containers is that Angel does not know how long each part of the plan will take to implement. Therefore, he gives the following time estimates for each activity in the move:
- *a* represents an optimistic time (shortest possible time)
- *b* represents a pessimistic time (longest possible time)
- *m* represents a most likely time (most probable time)

19.2.1 Using the Critical Path Method

Table 19.2.1 lists the activities and possible times (in days) needed to complete the activities.

Activity Code	Activity Description	Activity Dependence	Completion Time (days)		
			a	*m*	*b*
A	Decide the number of CHEMPACK containers to move		0.5	1	2
B	Decide where the CHEMPACK containers will be moved to	A	1	1.5	3
C	Decide where the CHEMPACK containers will be moved from	A	1	2	3
D	Get approval from State and Federal Authorities	B, C	0.5	1	3
E	Discuss strategy with current location(s)	D	0.5	1	1.5
F	Discuss strategy with receiving location(s)	D	1	2	2.5
G	Set up transportation of CHEMPACK containers	E, F	0.5	1	2
H	Prepare receiving site(s)	F	2	3	5
I	Prepare CHEMPACK containers for relocation	E	1	1.5	2
J	Load and transport CHEMPACK containers	G, I	1	1.5	3
K	Unload CHEMPACK containers at receiving site(s)	H, J	0.5	1	2

Table 19.2.1: Activity chart for CHEMPACK container relocation

Angel will use PERT (Program Evaluation Review Technique) to determine how long it will take to plan and move the CHEMPACK containers to the appropriate locations for Super Bowl Sunday.

Angel begins by adding a column to Table 19.2.1 that lists the expected time for each activity (See Table 19.2.2). He finds the expected times using the formula mentioned in section 19.1.1. He then draws an order-requirement digraph and determines the critical path of the network, a process very similar to the Critical Path Method found in Chapter 10.

Activity Code	Activity Description	Activity Dependence	Completion Time (days)			
			a	m	b	t_E
A	Decide the number of CHEMPACK containers to move		0.5	1	2	1.08
B	Decide where the CHEMPACK containers will be moved to	A	1	1.5	3	1.67
C	Decide where the CHEMPACK containers will be moved from	A	1	2	3	2.00
D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25
E	Discuss strategy with current location(s)	D	0.5	1	1.5	1.00
F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.92
G	Set up transportation of CHEMPACK containers	E, F	0.5	1	2	1.08
H	Prepare receiving site(s)	F	2	3	5	3.17
I	Prepare CHEMPACK containers for relocation	E	1	1.5	2	1.50
J	Load and transport CHEMPACK containers	G, I	1	1.5	3	1.67
K	Unload CHEMPACK containers at receiving site(s)	H, J	0.5	1	2	1.08

Table 19.2.2: Activity chart for CHEMPACK container relocation including the expected time

Recall that the expected time (t_E) is found in each row by using the formula

$$t_E = \frac{a + 4m + b}{6}.$$

For example, activity D (Get approval from State and Federal Authorities) could optimistically be completed in 0.5 days; it could pessimistically be completed in 3 days; or it could most likely be completed in 1 day. Thus, the expected time to complete activity D is:

$$t_E = \frac{(0.5) + 4(1) + (3)}{6} = \frac{7.5}{6} = 1.25 \text{ days}.$$

This weighted average is an approximation of the expected value of a probability distribution called the Beta Probability Distribution (BPD). When the values for a, b, and m are known, the BPD is used by operations researchers to find t_E. That is, the BDP can be used to model events which are constrained to take place within a certain time interval. For example, Figure 19.2.1 shows an illustration of the BPD for certain values of a, b, and m.
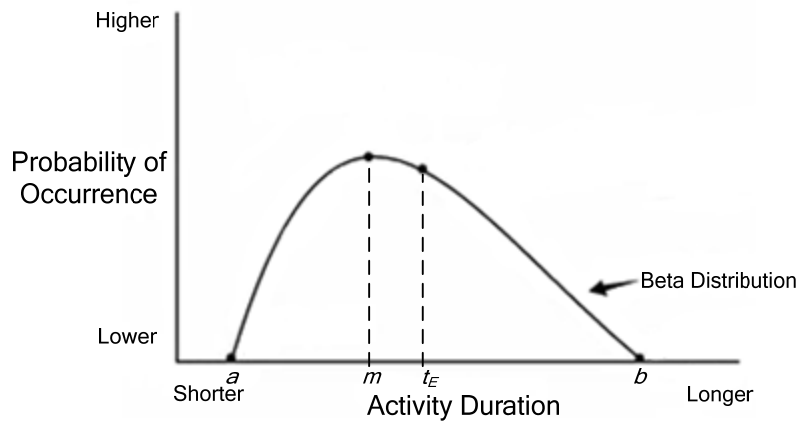
Figure 19.2.1: An illustration of the Beta Probability Distribution

In earlier chapters, several probability distributions were discussed as well as the expected value for each of them. Rather than devote an entire chapter to the BPD, we only mention that the BPD underlies PERT probabilities, and the expected completion time (based on the BPD) is approximated by the formula

$$t_E = \frac{a + 4m + b}{6}.$$

Now, Angel uses the activities and expected time from Table 19.2.2 to create the network shown in Figure 19.2.2

Figure 19.2.2: The completed order-requirement digraph for CHEMPACK container relocation

Figure 19.2.2 illustrates the flow of activities from Table 19.2.2. Notice that the weight of the order-requirement digraph is the expected time (t_E) for each activity. The earliest start time (EST), earliest finish time (EFT), latest start time (LST), and latest finish time (LFT) are calculat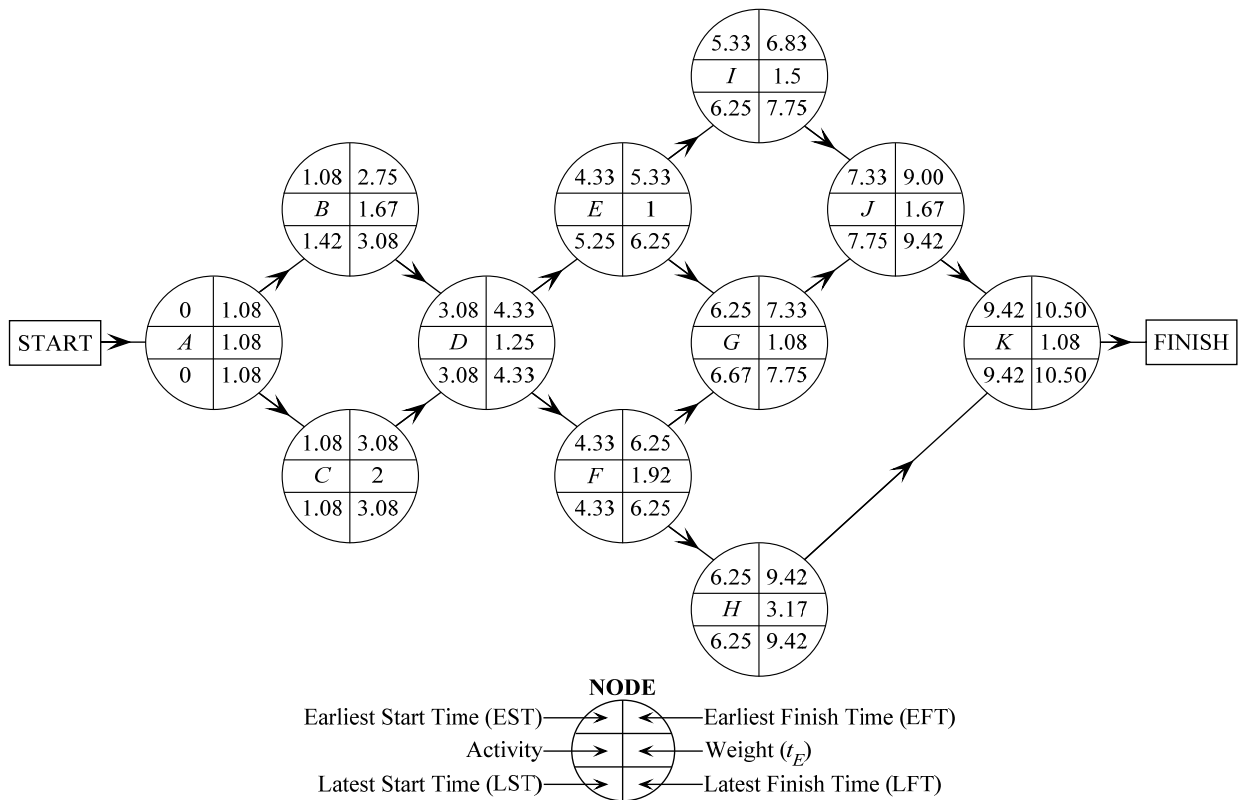ed using the same techniques as for Critical Path Method (CPM) in Chapter 10. (Refer to Chapter 10 to review how to do draw an order-requirement digraph and how to calculate forward and backward passes.)

Angel uses Figure 19.2.2 to find the critical path. Recall the critical path is the longest path through the network because this path determines the overall duration of the project.

In this problem, Angel determines that the critical path is *A-C-D-F-H-K* (Decide the number of CHEMPACK containers to move; Decide where the CHEMPACK containers will be moved from; Get approval from State and Federal Authorities; Discuss strategy with receiving location(s); Prepare receiving site(s); Unload CHEMPACK containers at receiving site(s)).

Angel notes that the last activity in the critical path (activity *K*: Unload CHEMPACK containers at receiving site(s)) finishes at 10.5 days. Therefore, the expected time to relocate the needed CHEMPACK containers is 10.5 days. Another way of stating this is that the *total expected time of the project* (T_E) is 10.5 days. This means that Angel can expect, after taking into account the shortest, probable, and longest possible times for the activities, to take about 10.5 days to do the CHEMPACK container preparation and relocation for the Super Bowl.

Next, Angel creates a chart and lists the slack times for the activities similar. (Refer to Chapter 10 for a review of slack times) Note that a critical activity shows a slack of zero in this chart. Angel could also create a Gantt chart to display slack, as was done in Chapter 10, but for PERT he uses a chart like Table 19.2.3.

Activity	EST	LST	EFT	LFT	Slack (days)		Critical Activity
					LST – EST	LFT – EFT	
A	0.00	0.00	1.08	1.08	0.00	0.00	Yes
B	1.08	1.42	2.75	3.08	0.33	0.33	
C	1.08	1.08	3.08	3.08	0.00	0.00	Yes
D	3.08	3.08	4.33	4.33	0.00	0.00	Yes
E	4.33	5.25	5.33	6.25	0.92	0.92	
F	4.33	4.33	6.25	6.25	0.00	0.00	Yes
G	6.25	6.67	7.33	7.75	0.42	0.42	
H	6.25	6.25	9.42	9.42	0.00	0.00	Yes
I	5.33	6.25	6.83	7.75	0.92	0.92	
J	7.33	7.75	9.00	9.42	0.42	0.42	
K	9.42	9.42	10.50	10.50	0.00	0.00	Yes

Table 19.2.3: Application of CPM algorithm to determine the critical path

Again, Angel can see that the critical path is *A-C-D-F-H-K* and the estimated project completion time is 10.5 days.

Up to this point—other than calculating and using the expected time as the weight of an activity rather than the exact time to determine the critical path—the problem has looked the same as the CPM algorithm. In the next section, the PERT method takes project planning a step further.

19.2.2 Determining the Probability of Finishing on Time

Recall that PERT is best used for projects that have never been planned before. Consequently, there is uncertainty in the completion times of the activities as well as in the total expected time of the project. In

fact, how can Angel be sure that they will finish the project in exactly 10.5 days? He cannot be sure, assuming the uncertainty of all the other activity times. However, a range of completion times may help him feel more confident in his projection.

For example, if Angel knew there was a high probability that the project could be completed sometime between 10 days and 11 days, he might feel more comfortable planning on a 10.5 day completion time for relocating CHEMPACK containers. If he knew that the project could be completed sometime between 5 days and 16 days, he might feel uncomfortable planning on 10.5 days. Therefore, he needs to consider more than just the expected times because the spread of the data affects the probabilities of completion.

Because PERT is built upon the Beta Probability Distribution (BPD), Angel uses the BPD to estimate the variance and standard deviation of the probable finishing times. From there, he will use the normal distribution to understand and predict the probabilities of finishing times at and around the expected times.

The formula to find the estimated variance for an activity (based on the BPD) is:

$$s^2 = \left(\frac{\text{largest} - \text{smallest}}{6} \right)^2 = \left(\frac{b-a}{6} \right)^2.$$

Angel uses this formula to find the estimated variance for each activity. Table 19.2.4 lists the already calculated variances for each activity of the Relocation of CHEMPACK containers project.

Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
A	Decide the number of CHEMPACK containers to move		0.5	1	2	1.08	Yes	0.06
B	Decide where the CHEMPACK containers will be moved to	A	1	1.5	3	1.67		0.11
C	Decide where the CHEMPACK containers will be moved from	A	1	2	3	2.00	Yes	0.11
D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25	Yes	0.17
E	Discuss strategy with current location(s)	D	0.5	1	1.5	1.00		0.03
F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.92	Yes	0.06
G	Set up transportation of CHEMPACK containers	E, F	0.5	1	2	1.08		0.06
H	Prepare receiving site(s)	F	2	3	5	3.17	Yes	0.25
I	Prepare CHEMPACK containers for relocation	E	1	1.5	2	1.50		0.03
J	Load and transport CHEMPACK containers	G, I	1	1.5	3	1.67		0.11
K	Unload CHEMPACK containers at receiving site(s)	H, J	0.5	1	2	1.08	Yes	0.06
							Total Project Variance	**0.72**

Table 19.2.4: Activity chart for CHEMPACK container relocation including the variance

Angel knows that *the estimated variance for the entire project is the sum of the variances of the activities on the critical path*. That is, the total estimated variance is *not* the sum of the variances of all the activities. Only the variances of the activities on the critical path are included.

Thus, Angel calculates the total estimated variance by summing the variances of the critical activities A, C, D, F, H, K:

$$0.06 + 0.11 + 0.17 + 0.06 + 0.25 + 0.06 = 0.72.$$

Therefore, 0.72 is the estimated variance of the whole project, as shown in Table 19.2.4.

Statisticians and engineers use the estimated standard deviation (s) instead of the estimated variance to talk about how values are distributed because the units of variance are nonsensical. Angel finds the estimated standard deviation by taking the square root of the variance:

$$s = \sqrt{0.72} \approx 0.85 \,.$$

This means that the expected time for finishing the CHEMPACK container project has an estimated standard deviation of almost one day. Angel can use the estimated standard deviation to determine the probability that the project finishes in 10.33 days, 9 days, 12 days, etc. instead of the projected time of 10.5 days.

PERT uses the idea of a normal distribution (see Chapter 15) to determine these probabilities.

> **Why is the normal distribution used even though PERT is based upon the Beta Probability Distribution (BPD)?**
>
> The BPD is used to estimate the expected completion times and variances for each activity. This distribution is appropriate for these calculations because it emphasizes the most probable value more than the optimistic or pessimistic values. However, once the expected completion times and variances are calculated, the normal distribution is more appropriate.
>
> The project completion time is estimated by adding the expected activity times along the critical path. Since these are assumed to be random and independent the *Central Limit Theorem* can be invoked (which implies that the sum of a large number of independent random variables will be approximately normally distributed, regardless of the distribution of the individual random variables).
>
> Because of these properties, the inventors of PERT (mathematicians and statisticians) realized the normal distribution is the best model of project networks for the management of complex projects. It is very useful from a statistical point of view that the normal distribution can be used for the project completion time since (a) this distribution is symmetric with a single peak value and (b) this distribution is described completely by its mean and standard deviation. Therefore, two normal curves with the same mean and standard deviation are always identical.

Suppose Angel is interested in the probability that the project will be completed in the 10 days before the Super Bowl if the expected time is 10.5 days. To find this, he uses the normal distribution to do the following calculation:

$$P(T_E < 10)$$

where T_E is the total time it takes to complete the project within 10 days given the estimated mean ($\bar{x}$) is 10.5 days and the estimated standard deviation (s) is 0.85. Figure 19.2.3 presents this graphically.
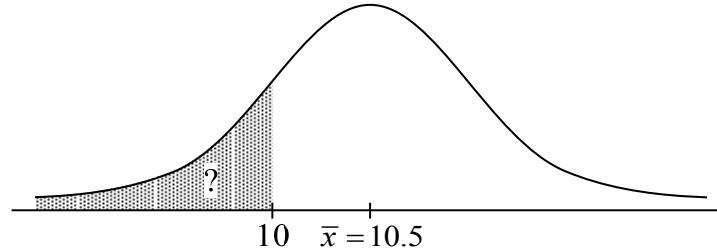

Figure 19.2.3: A graphical display of $P(T_E < 10)$ with $\bar{x} = 10.5$ and $s = 0.85$.

To calculate the probability (i.e., the area under the curve), there are many applications could be used, including Excel, graphing calculators, or online Java applets. Angel decides to use a graphing calculator. (See the Appendix for an explanation on how to use Excel to calculate normal distribution probabilities.)

Under the DISTR menu in the graphing calculator, Angel selects normalcdf(and presses enter. The syntax for the normalcdf(function is:

normalcdf(start, end, mean, standard deviation).

Angel wants to the know the probability that the project will be completed in 10 days if the expected time is 10.5 days and the standard deviation is 0.85. Therefore, he inputs:

normalcdf(0, 10, 10.5, 0.85).

Figure 19.2.4 shows what this looks like in the graphing calculator.

Figure 19.2.4: A graphing calculator screen of $P(T_E < 10)$ with $\bar{x} = 10.5$ and $s = 0.85$

Thus, Angel has approximately a 28% chance of relocating the CHEMPACK containers in time for the Super Bowl if he starts 10 days prior, given that project takes approximately 10.5 days to complete.

Note the key to this answer is that the expected total project completion time is 10.5 days. The calculation $P(T_E < 10)$ would change if the expected total project completion time change.

For example, the critical path time is 9 days, instead of 10.5 days. Then the probability that the project will be completed in 10 days would no longer be 28%. See Figure 19.2.5 for an illustration of this example.
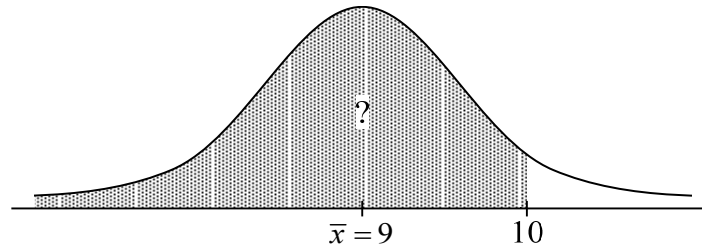
Figure 19.2.5: A graphical display of $P(T_E < 10)$ with $\bar{x} = 9$ and $s = 0.85$.

Using the graphing calculator, Angel inputs normalcdf(0, 10, 9, 0.85) and discovers that $P(T_E < 10) = 0.880$. So instead of approximately a 28% chance of completing the project on time, Angel now has an 88% chance of relocating the CHEMPACK containers on time given the fact he estimates that the project would take 9 days to complete. The probability of completing the project in the 10 days before the Super Bowl changes drastically depending on our expected project completion time.

Q1.   Determine the probability of relocating the CHEMPACK containers in time for the Super Bowl if Angel starts 10 days prior, given that project takes approximately:
        a.   11 days to complete
        b.   8 days to complete

Angel returns back to the original calculation of expected project completion time: $T_E = 10.5$ days. But now, instead of stating the time within which he wants to complete the project, he decides to investigate how long it would take to be 95% certain the CHEMPACK containers would be in position in time.

In this case, his is calculating $P(T_E < ?) = 0.95$. Figure 19.2.6 shows this graphically.
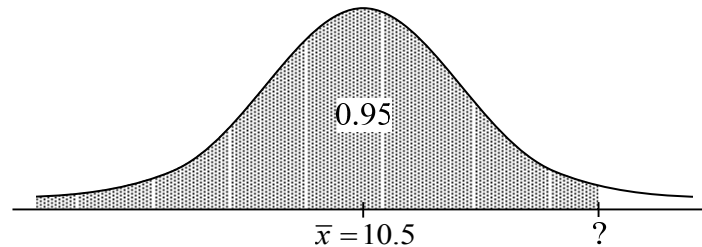
Figure 19.2.6: A graphical display of $P(T_E < ?) = 0.95$ with $\bar{x} = 10.5$ and $s = 0.85$.
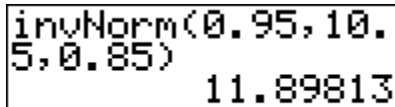
To solve for the value in his graphing calculator, Angel selects invNorm(under the DISTR menu presses enter. The syntax for the invNorm(function is:

invNorm(probability, mean, standard deviation).

To determine how long it will take to be 95% certain the CHEMPACK containers will be in position in time, Angel inputs:

invNorm(0.95, 10.5, 0.85).

Figure 19.2.7 shows what this looks like in the graphing calculator.

```
invNorm(0.95,10.
5,0.85)
                11.89813
```

Figure 19.2.7: A graphing calculator screen of $P(T_E < ?) = 0.95$ with $\bar{x} = 10.5$ and $s = 0.85$.

Therefore, in order to be 95% sure the CHEMPACK containers will be in place for Super Bowl Sunday, Angel needs to begin the process at least 11.89 days before the Super Bowl.

Q2.   Why is this time greater than 10.5 days (the estimated completion time)?

Next, Angel wonders what the probability of completing the relocation of the CHEMPACK containers within 12 full days is. He knows it should be a little more than 95%, since the 0.95 is the probability for 11.89 days. Calculating this, he gets:
$$P(T_E < 12) = 0.961$$

As he predicted, this is a little bigger than the 0.95 probability.

Q3.   Should Angel plan for 12 days to complete the project given we estimate an expected total project completed of 10.5 days? What are some advantages of having that extra 1.5 days? What are some disadvantages?

Next, suppose Angel finds out that activity *H* (prepare CHEMPACK container receiving site(s)) will take exactly 3 days to complete. He wants to know how this changes when he must start in order to be 95% sure of the relocation is finished by Super Bowl Sunday. To do this, he has to recalculate the whole network including earliest start times, earliest finish times, latest start times, latest finish times, and project completion.
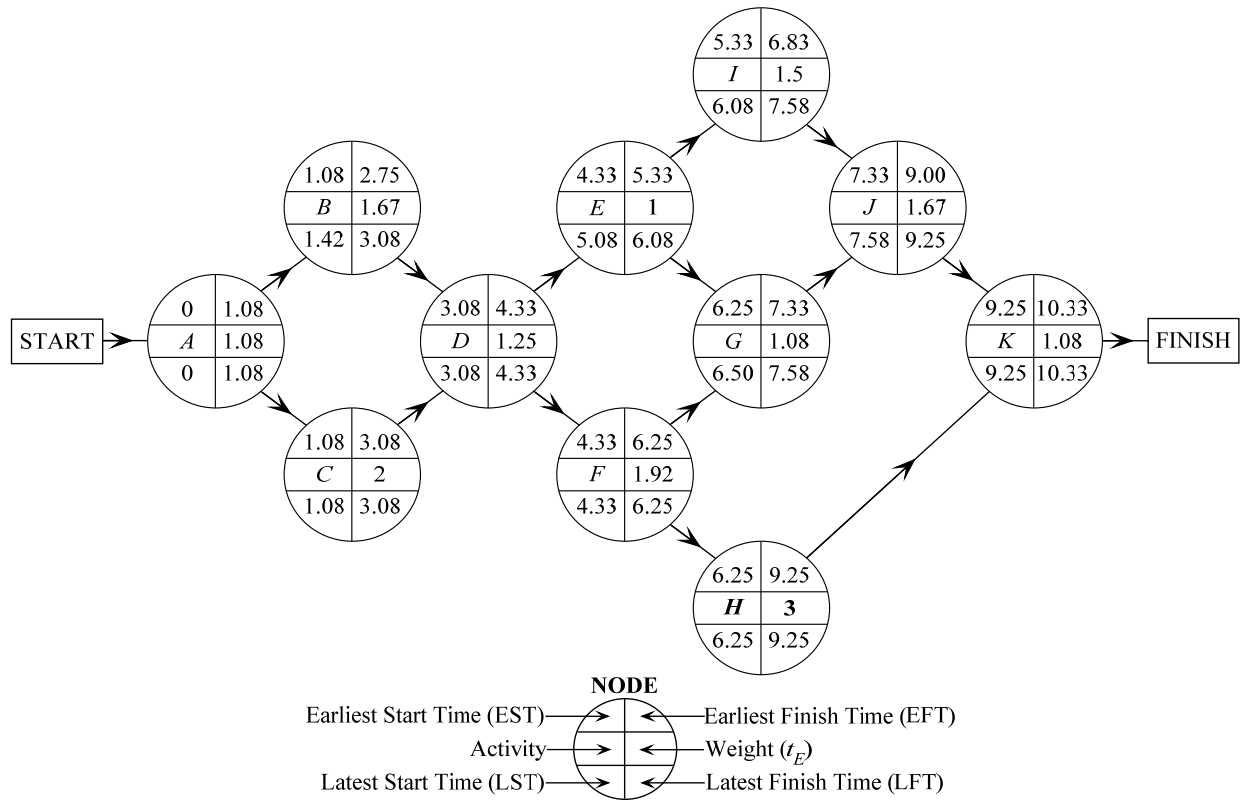
Figure 19.2.8: The completed order-requirement digraph for CHEMPACK container relocation with activity *H* taking exactly 3 days to complete

Even with the altered time for activity *H*, the critical path remains *A-C-D-F-H-K*. But now the total time for the project decreases from 10.5 days to 10.33 days.

Angel notices that activity *H* has the largest estimated variance ($s^2 = 0.25$) in Table 19.2.4 of the original problem. However, because he now knows exactly how long it takes to prepare the CHEMPACK container receiving site(s), this value will not vary. Thus, activity *H*'s variance become zero ($s^2 = 0$). Table 19.2.5 reflects these changes for activity *H* and the estimated variance for the whole project.

Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
A	Decide the number of CHEMPACK containers to move		0.5	1	2	1.08	Yes	0.06
B	Decide where the CHEMPACK containers will be moved to	A	1	1.5	3	1.67		0.11
C	Decide where the CHEMPACK containers will be moved from	A	1	2	3	2.00	Yes	0.11
D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25	Yes	0.17
E	Discuss strategy with current location(s)	D	0.5	1	1.5	1.00		0.03
F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.92	Yes	0.06
G	Set up transportation of CHEMPACK containers	E, F	0.5	1	2	1.08		0.06
H	**Prepare receiving site(s)**	**F**	**3**	**3**	**3**	**3**	**Yes**	**0**
I	Prepare CHEMPACK containers for relocation	E	1	1.5	2	1.50		0.03
J	Load and transport CHEMPACK containers	G, I	1	1.5	3	1.67		0.11
K	Unload CHEMPACK containers at receiving site(s)	H, J	0.5	1	2	1.08	Yes	0.06
							Total Project Variance	**0.47**

Table 19.2.5: Activity chart for CHEMPACK container relocation including the estimated variance for the Super Bowl knowing that Activity H takes exactly 3 days to complete

The overall project estimated variance reduces from 0.72 to 0.47. This means the estimated standard deviation reduces from 0.85 to $s = \sqrt{0.47} \approx 0.69$.

Repeating the calculations from above, Angel wants to know how long it would take if he wanted to be 95% certain the CHEMPACK containers would be in position in time, given the expected project completion time is 10.33 days (the new value of T_E).

In this case, he is calculating $P(T_E < ?) = 0.95$. Graphically, this looks like

0.95

$\bar{x} = 10.33$?

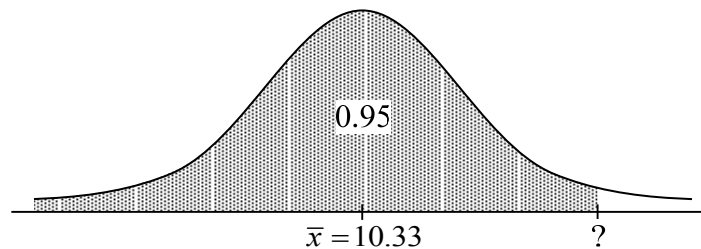**Figure 19.2.9**: A graphical display of $P(T_E < ?) = 0.95$ with $\bar{x} = 10.33$ and $s = 0.69$.

Using the graphing calculator [invNorm(0.95,10.33,0.69], he gets $(T_E < \mathbf{11.46}) = 0.95$.

Therefore, in order to be 95% sure he has the CHEMPACK containers in place for Super Bowl Sunday given the expected completion time is 10.33 days, Angel needs to begin the process 11.46 days (or more) before the Super Bowl.

Taking these calculations further, Angel determines the probability of completing relocation within 12 full days:

$$P(T_E < 12) = 0.992$$

As seen in the calculation above, Angel is 99.2% certain that he can relocate the CHEMPACK containers within 12 days of Super Bowl Sunday given the expected completion time is 10.33 days.

The project management tool PERT allows people, like Angel, to make decisions about when to begin projects that have uncertain activity times. In this chapter, examples of projects were given in which the activities and activity dependence were already determined. In the homework, you will be asked to create these tables on your own.

19.2.3 Steps of the PERT Algorithm

The following are the steps of the PERT process.
1. Create a list of activities necessary to complete a project of uncertain duration.
2. Decide on the dependencies of each activity, meaning each activity determines what other activities must be finished in order for the activity to begin.
3. Create a table that lists the activities and activity dependencies as well as shortest possible completion time (a), most probable completion time (m), and longest possible completion time (b) for each activity.
4. Find the expected time for each activity by using the Beta Probability Distribution (BPD) formula of $t_E = \dfrac{a + 4m + b}{6}$ and fill in the chart.
5. Use the expected times to create and complete a network. Calculate the earliest start time (EST), earliest finish time (EFT), latest start time (LST), and latest finish time (LFT) for each activity.
6. Find the critical path through the activity network and the total time for that path. This is the expected finish time; however, there is a lot of uncertainty in this time since the activity times are based upon estimated times as well.
7. Find the estimated variance (s^2) for each activity using BPD formula $s^2 = \left(\dfrac{b-a}{6}\right)^2$. Then sum the estimated variances of the critical activities to find the estimated variance for the whole project.
8. Use the normal distribution with the expected time for the mean and the estimated standard deviation (s, the square root of the variance calculated in step 7) to make decisions about the probability of finishing on time given the expected project completion time is $\bar{x}$. You use the following formula to calculate the probability:

    $$P(T_E \leq \text{some finish time}) = \text{probability}.$$

    This formula can be used to find either a percent change of finishing within a certain time or a certain time necessary to guarantee a specific value for probability.

This chapter concludes with a final more complex example and asks you to use the information to make decisions about this new product.

Section 19.3 Construction of a TV Tower

In this section, a more complex example of where PERT might be very helpful is presented. Keene General Contractors (KGC) is currently preparing a bid for the construction of a new 225-foot television antenna tower and the construction of a building adjacent to the tower that will be used to house transmission and electrical equipment in Lansing, Michigan. This construction has become necessary because of all the new HDTV signals that are being transmitted; more towers are needed to keep up with demand.

KGC is bidding only on the construction of (1) the tower and its electrical equipment, (2) the building, (3) the connecting cable between the tower and building, and (4) site preparation. Transmission and other equipment to be housed in the building are not included in the bid and will be obtained separately by the local television station in Lansing. The site for the tower is at the top of a hill to minimize the required height of the tower. The building site is at a slightly lower elevation than the tower base and near a main road. There will be a crushed gravel service road and an underground cable between the tower and building. A fuel tank will be installed above ground on a concrete slab adjacent to the building.

In addition to preparing a cost estimate for the bid package, an estimate of the time it would take to complete this work is also required. The TV station management is very concerned about the time to complete this contract and therefore requested bids be prepared for an expected project completion date. They also requested that a 95% probability goal be established for the expected completion date.

Due to the novelty of this contract, the length of time each activity will take to complete is unknown. However, R. Scott, the general manager of KGC, can estimate the optimistic (a), the pessimistic (b), and the most likely (m) time for each activity. Table 19.3.1 lists each activity and the three activity time estimates. The dependency of the activities and the order of construction work were developed from the following logic:

- Survey the work and procurement of the structural steel and electrical equipment so that the tower work can start as soon as the contract is signed.
- Grading of the tower site and building site can begin when the survey is complete.
- After tower site is graded, footings and anchors can be poured.
- After building site is graded and basement excavated, building footings can be poured.
- Septic tanks can be installed when grading and excavating of the building site is done.
- Construction of connecting road can start as soon as the survey is completed.
- Exterior and interior basement walls can be poured as soon as the footings are in.
- Basement floor and fuel tank slab should go in after basement walls.
- Floor beams can go in after the basement walls and basement floor.
- Main floor slab and concrete block walls go in after floor beams.
- Roof slab can go on after block walls are up.
- Interior can be completed as soon as roof slab is on.
- Install fuel tank any time after slab is in.
- Drain tile and storm drain for building to in after septic tank.
- As soon as tower footings and anchors are in and tower steel and equipment are available, tower can be erected.
- Connecting cable in tower site, drain tile, and storm drain can be put in as soon as the tower is up.
- Main cable between the building and tower goes in after connecting cable at tower site is in and basement walls are up.
- Tower site can be backfilled and graded as soon as storm drain, connecting cable, and main cable are in.
- Clean up of tower site is done after backfilling and grading is done.

- Backfill around building and grade after main cable is in and after storm drain is in.
- Clean up of building site after backfilling and grading is done.
- Obtain job acceptance.

From this logic, R. Scott developed Table 19.3.1.

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)			
			a	*m*	*b*	t_E
A	Sign contract and complete negotiations		4	5	6	
B	Survey site	A	4	6	8	
C	Grade building site and excavate	B	6	8	10	
D	Grade tower site	B	21	30	39	
E	Procure structural steel	A	85	85	85	
F	Procure electrical equipment	A	120	120	120	
G	Procure concrete for footings	D	25	42	59	
H	Erect tower and install E. E.	E, F, G	25	38	51	
I	Install connecting cable	H	4	8	12	
J	Install drain tile and storm drain	H	18	35	52	
K	Backfill and grade tower site	J, V	4	8	12	
L	Pour building footings	C	21	29	37	
M	Pour basement and tank slabs	N, O	11	14	17	
N	Pour outside basement walls	L	30	34	38	
O	Pour walls for basement rooms	L	7	9	11	
P	Pour concrete floor beams	M	10	11	12	
Q	Pour main floor slab	P	10	12	14	
R	Pour roof slab	Q	13	15	17	
S	Complete interior framing	R	30	42	54	
T	Lay roof	S	2	3	4	
U	Paint building interior	T	13	19	25	
V	Install main cable-site to building	I	25	35	45	
W	Install fuel tank	M	2	3	4	
X	Install septic tank	C	8	12	16	
Y	Install drain tile and storm drain	X	10	15	20	
Z	Backfill around building	Y	7	9	11	
AA	Lay base for connecting road	B	13	15	17	
BB	Complete grading and surface connecting road	AA	5	8	11	
CC	Clean up tower site	K	3	5	7	
DD	Clean up building site	Z	2	3	4	
EE	Contract acceptance obtained and sign-off	U, W, BB, CC, DD	5	5	5	

Table 19.3.1: Activity chart for the construction of a TV tower

R. Scott takes a minute to study this chart and learn a little about the TV tower. He notices that there are more than 26 activities, so the activities are named *AA*, *BB*, etc. when the 26 alphabet letters are exhausted. He also notices that the exact order of the activities is not necessarily in the alphabetical order. For example activity *K* (Backfill and grade tower site) actually depends on activity *V* (Install main cable-site to building) as well as activity *J* (Install drain tile and storm drain). Another example is activity *M* depends on activity *N* and activity *O*. This is common for activity dependence charts since the list of activities is often created first and then the dependencies are developed later.

Now, R. Scotts calculates the expected finish time (t_E) for each activity from shortest possible completion time (a), most probable completion time (m), and longest possible complete time (b), using the formula:

$$t_E = \frac{a + 4m + b}{6}.$$

Table 19.3.2 lists these values.

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)			
			a	m	b	t_E
A	Sign contract and complete negotiations		4	5	6	5
B	Survey site	A	4	6	8	6
C	Grade building site and excavate	B	6	8	10	8
D	Grade tower site	B	21	30	39	30
E	Procure structural steel	A	85	85	85	85
F	Procure electrical equipment	A	120	120	120	120
G	Procure concrete for footings	D	25	42	59	42
H	Erect tower and install E. E.	E, F, G	25	38	51	38
I	Install connecting cable	H	4	8	12	8
J	Install drain tile and storm drain	H	18	35	52	35
K	Backfill and grade tower site	J, V	4	8	12	8
L	Pour building footings	C	21	29	37	29
M	Pour basement and tank slabs	N, O	11	14	17	14
N	Pour outside basement walls	L	30	34	38	34
O	Pour walls for basement rooms	L	7	9	11	9
P	Pour concrete floor beams	M	10	11	12	11
Q	Pour main floor slab	P	10	12	14	12
R	Pour roof slab	Q	13	15	17	15
S	Complete interior framing	R	30	42	54	42
T	Lay roof	S	2	3	4	3
U	Paint building interior	T	13	19	25	19
V	Install main cable-site to building	I	25	35	45	35
W	Install fuel tank	M	2	3	4	3
X	Install septic tank	C	8	12	16	12
Y	Install drain tile and storm drain	X	10	15	20	15
Z	Backfill around building	Y	7	9	11	9
AA	Lay base for connecting road	B	13	15	17	15
BB	Complete grading and surface connecting road	AA	5	8	11	8
CC	Clean up tower site	K	3	5	7	5
DD	Clean up building site	Z	2	3	4	3
EE	Contract acceptance obtained and sign-off	U, W, BB, CC, DD	5	5	5	5

Table 19.3.2: Activity chart for the construction of a TV Tower

Figure 19.3.1 applies these expected times to the order-requirement digraph and calculates the earliest start time (EST), earliest finish time (EFT), latest start time (LST), and latest finish time (LFT) for each activity for the TV tower.

26	34	
BB	8	
211	219	

31	46	
Y	15	
192	207	

46	55	
Z	9	
207	216	

55	58	
DD	3	
216	219	

96	107	
P	11	
117	128	

107	119	
Q	12	
128	140	

119	134	
R	15	
140	155	

134	176	
S	42	
155	197	

176	179	
T	3	
197	200	

179	198	
U	19	
200	219	

219	224	
EE	5	
219	224	

FINISH

96	99	
W	3	
216	219	

NODE

Earliest Start Time (EST) —— →← —— Earliest Finish Time (EFT)
Activity —— →← —— Weight (T_E)
Latest Start Time (LST) —— →← —— Latest Finish Time (LFT)

206	214	
K	8	
206	214	

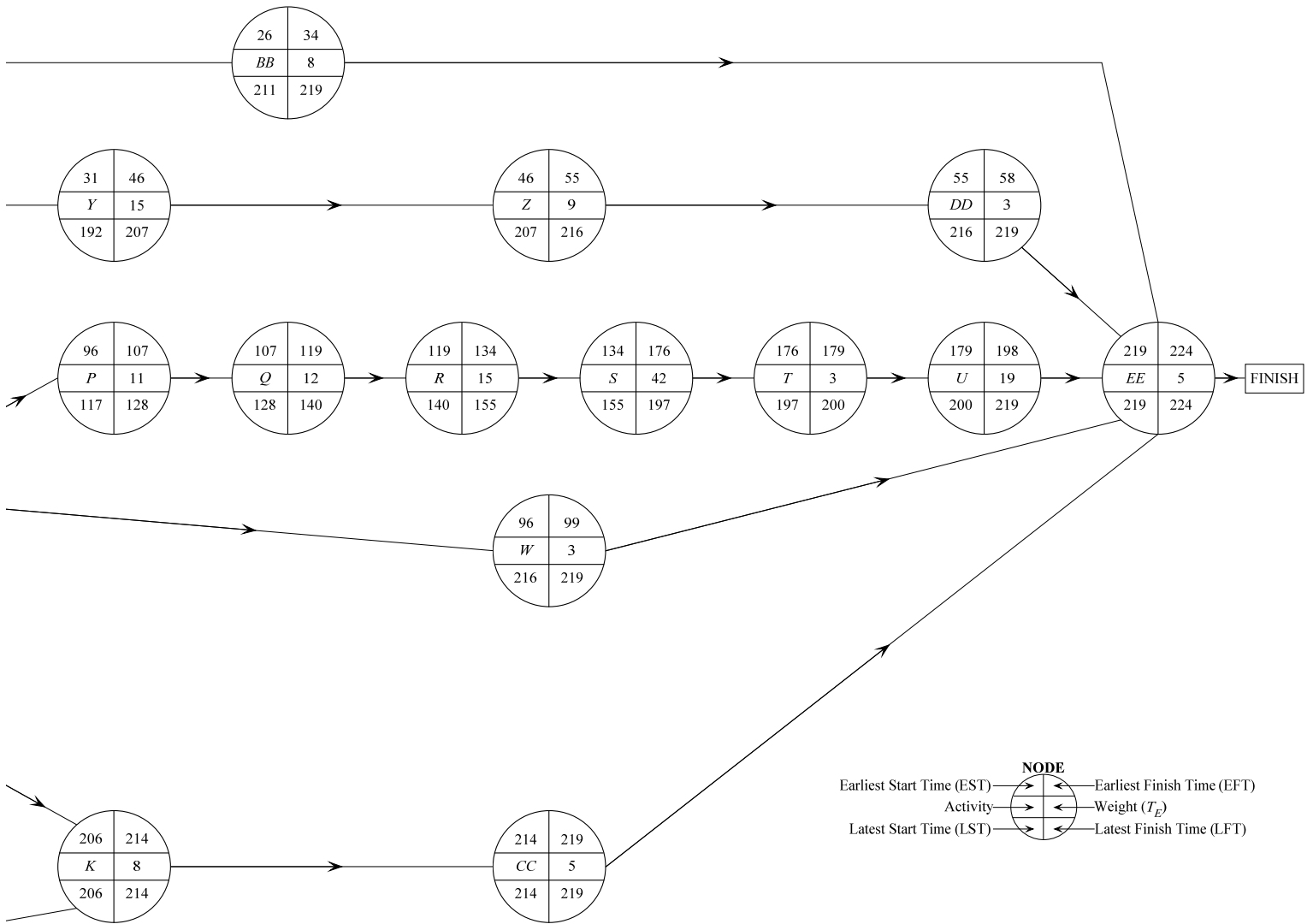|		
214	219	
CC	5	
214	219	
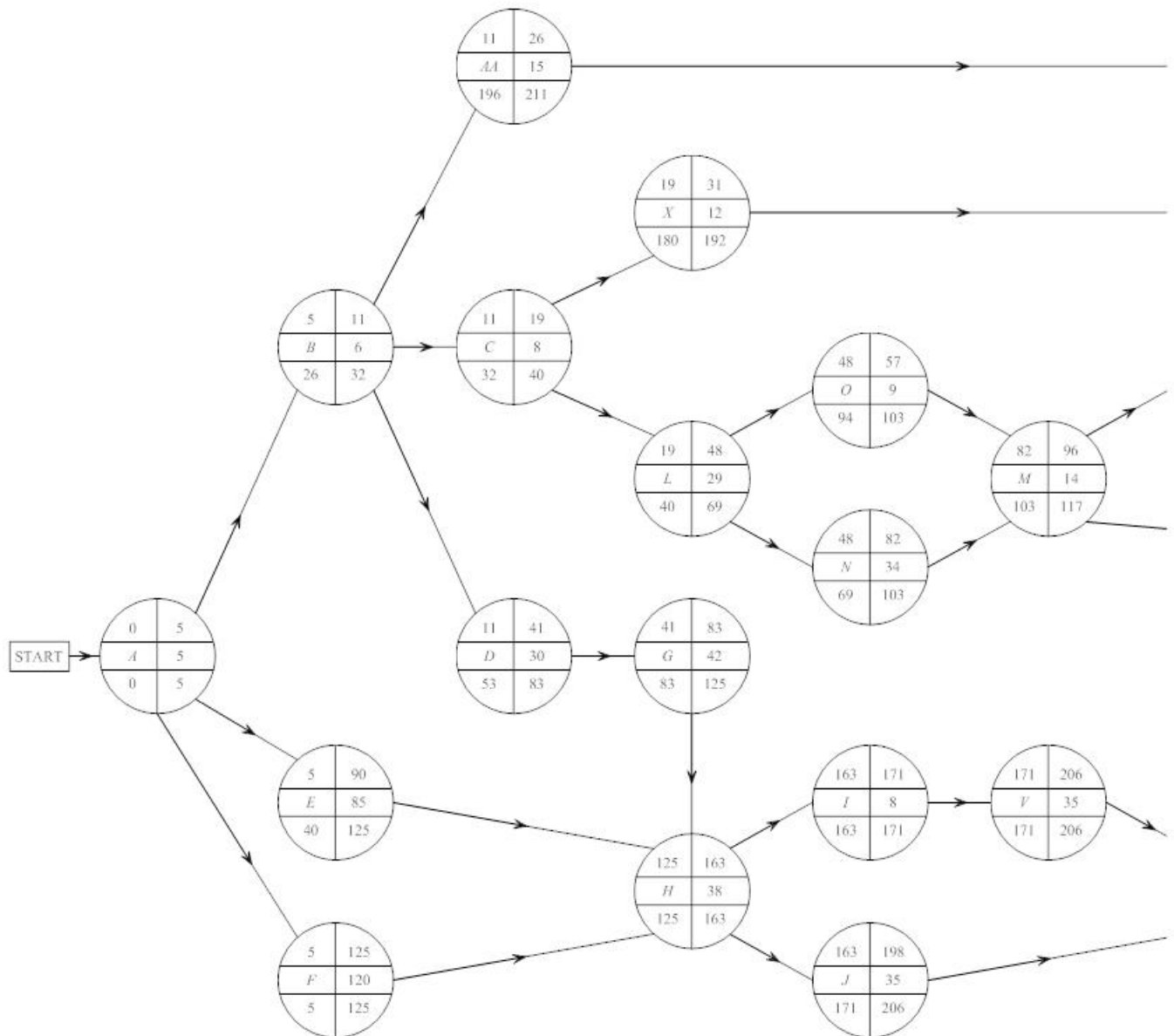
Figure 19.3.1: The completed order-requirement digraph for the construction of a TV Tower

Q1.   What are the critical activities of the TV Tower Project?

Q2.   What is the expected time (T_E) to complete the TV Tower Project?

Q3.   What is the estimated standard deviation (s) for the TV Tower project? Use Table 19.3.3 to calculate the answer.

Q4.   We estimate 2-4 weeks to lay the roof of this building. If the building will have 10,000 square feet of roof, use the Internet to see if this value is an appropriate estimate. If not, change the Activity Chart to reflect what you found and see if that changes the critical path time.

Q5.   One question R. Scott needs to answer in the bid is how many weeks are needed to be 95% certain the project is completed on time. Using the normal distribution, determine how many weeks we need to plan for in order to be 95% certain we will complete the TV Tower Project on time.

Q6.   Using the normal distribution, find the probability that the project can be completed in under 200 weeks.

Q7.   Assume we know that Activity O, Pour walls for basement rooms, will take exactly 9 weeks to complete. How does this change your answers in Question (12) and Question (13)?

Q8.   Look carefully at the activity dependence in Table 19.3.1. Choose one activity you think is not going to be done in the time suggested and change the time accordingly. How does your change affect the critical path?

Q9.   What if R. Scott asked you to find the "worst possible scenario". Use PERT and the most pessimistic times to find the critical path and completion time.

Q10.   You have been asked to help R. Scott document the work you did on the construction project planning with PERT. Write a document addressing the requirements of the original bid:
       a.   An estimation of the time it would take to complete this work.
       b.   Establish a 95% probability goal for the expected completion date.
       Additionally, construct a table that lists the probabilities of completing the project within the following weeks: 208 (four years), 216, 224, 232, 240, 248, 254, and 260 (5 years).

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	*m*	*b*	*t_E*		
A	Sign contract and complete negotiations		4	5	6	5		0.11
B	Survey site	A	4	6	8	6		0.44
C	Grade building site and excavate	B	6	8	10	8		0.44
D	Grade tower site	B	21	30	39	30		9
E	Procure structural steel	A	85	85	85	85		0
F	Procure electrical equipment	A	120	120	120	120		0
G	Procure concrete for footings	D	25	42	59	42		32.11
H	Erect tower and install E. E.	E, F, G	25	38	51	38		18.78
I	Install connecting cable	H	4	8	12	8		1.78
J	Install drain tile and storm drain	H	18	35	52	35		32.11
K	Backfill and grade tower site	J, V	4	8	12	8		1.78
L	Pour building footings	C	21	29	37	29		7.11
M	Pour basement and tank slabs	N, O	11	14	17	14		1
N	Pour outside basement walls	L	30	34	38	34		1.78
O	Pour walls for basement rooms	L	7	9	11	9		0.44
P	Pour concrete floor beams	M	10	11	12	11		0.11
Q	Pour main floor slab	P	10	12	14	12		0.44
R	Pour roof slab	Q	13	15	17	15		0.44
S	Complete interior framing	R	30	42	54	42		16
T	Lay roof	S	2	3	4	3		0.11
U	Paint building interior	T	13	19	25	19		4
V	Install main cable-site to building	I	25	35	45	35		11.11
W	Install fuel tank	M	2	3	4	3		0.11
X	Install septic tank	C	8	12	16	12		1.78
Y	Install drain tile and storm drain	X	10	15	20	15		2.78
Z	Backfill around building	Y	7	9	11	9		0.44
AA	Lay base for connecting road	B	13	15	17	15		0.44
BB	Complete grading and surface connecting road	AA	5	8	11	8		1
CC	Clean up tower site	K	3	5	7	5		0.44
DD	Clean up building site	Z	2	3	4	3		0.11
EE	Contract acceptance obtained and sign-off	U, W, BB, CC, DD	5	5	5	5		0
						Total Project Variance		

Table 19.3.3: Activity chart for the construction of a TV Tower including the estimated variance

Appendix: **Using Excel in PERT**

Many of the calculations for CPM are not mathematically difficult. Almost all the calculations are the sum or difference of two activity times. Even though the process requires a lot of logic to determine which numbers are added or subtracted, the actual arithmetic is not hard. Because of this, CPM calculations do not need to require Excel to do the math comfortably.

PERT is a little different since the calculations are more substantial. Additionally, there is a need to use charts, calculators or computers to find probabilities of the normal distribution. Therefore, in this appendix, we discuss how using Excel with PERT can simplify and speed up the calculations as well an easily manipulate solutions for analysis.

Referring to the relocation of CHEMPACK containers for the Super Bowl, we will describe how you can use Excel to keep track of the data and calculate what you need to find. The charts will look similar to the charts that you created in CPM, but you will do the calculations in Excel.

First, we create a new Excel spreadsheet by entering in Table 19.2.1, the activity chart, into Excel. You enter this information by hand. Here is an example:

	A	B	C	D	E	F	G
	Activity Code	**Activity Description**	**Activity Dependence**	**Completion Time (days)**			
1 2				a	m	b	t_E
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	
6	D	Get approval from State and Federal Authorities	B, C	0.5	1	3	
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	
9	G	Set up transportation of CHEM-PACKs	E, F	0.5	1	2	
10	H	Prepare receiving site(s)	F	2	3	5	
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	
12	J	Load and transport CHEM-PACKs	G, I	1	1.5	3	
13	K	Unload CHEM-PACKs at receiving site(s)	H, J	0.5	1	2	

Figure 19.Appendix.1: An example of the activity chart for CHEMPACK container relocation in Excel

Next, we will find the expected time (t_E) for each activity using the Beta Probability Distribution formula, $t_E = \dfrac{a + 4m + b}{6}$. Look at activity A. We type the formula,

$$= (D3+4*E3+F3)/6$$

into the formula bar of G3. To avoid having to type in the formula for every activity, use the "fill-handle" tool and drag the formula down the column for each row with an activity.

	A	B	C	D	E	F	G
1	**Activity Code**	**Activity Description**	**Activity Dependence**	**Completion Time (days)**			
2				a	m	b	t_E
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	1.08333
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2
6	D	Get approval from State and Federal Authorities	B,C	0.5	1	3	1.25
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667
9	G	Set up transportation of CHEM-PACKs	E,F	0.5	1	2	1.08333
10	H	Prepare receiving site(s)	F	2	3	5	3.16667
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5
12	J	Load and transport CHEM-PACKs	G,I	1	1.5	3	1.66667
13	K	Unload CHEM-PACKs at receiving site(s)	H,J	0.5	1	2	1.08333

Figure 19.Appendix.2: An example of the activity chart for CHEMPACK container relocation in Excel with the expected times

As you can see, these numbers for t_E are the same numbers you found by hand in section 19.2 (see Table 19.2.2), but now Excel has done the work for you.

At this point, even with Excel, you might create the order-requirement digraph and find the critical path. However, in this appendix we will use Excel to find the earliest start time (EST), earliest finish time (EFT), latest start time (LST), and latest finish time (LFT) for each activity. In order to use Excel to calculate the EST, EFT, LST, and LFT, you will need to be very familiar with the problem.

Keep in mind, doing this chart *after* you draw the order-requirement digraph may be the easiest way. In this case, you would just enter the numbers found on the nodes into the chart.

However, for this appendix we briefly describe a way you can use Excel that bypasses drawing a order-requirement digraph and filling out the nodes. While the process will remain the same no matter the project planning problem (CPM or PERT), each spreadsheet is individualized and cannot be "copied", "pasted", and used for a different project plan. The spreadsheets cannot be replicated because each project has different activity dependence.

Below is a blank table similar to Table 19.2.3. The table is placed to the right of the dependence chart (see Figure 19.Appendix.2); consequently, the activities on each row match up.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Variance (s^2)	EST	LST	EFT	LFT	Slack (days)		Critical Activity
2				a	m	b	t_E						LST – EST	LFT – EFT	
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	1.08333								
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667								
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2								
6	D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25								
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1								
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667								
9	G	Set up transportation of CHEM-PACKs	E, F	0.5	1	2	1.08333								
10	H	Prepare receiving site(s)	F	2	3	5	3.16667								
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5								
12	J	Load and transport CHEM-PACKs	G, I	1	1.5	3	1.66667								
13	K	Unload CHEM-PACKs at receiving site(s)	H, J	0.5	1	2	1.08333								

Figure 19.Appendix.3: An example of a blank EST/EFT/LST/LFT chart for CHEMPACK container relocation in Excel

Just as with an order-requirement digraph, we will make a version of the "forward pass" and then the "backward pass" in Excel in order to calculate the EST, LST, EFT, and LFT. First, we fill in the EST and EFT columns for the forward pass. Activity *A* is the earliest activity, so place a zero in the earliest start time (EST), which is cell I3 in our example spreadsheet. To calculate the earliest finish time (EFT), add the expected activity time (t_E) to the EST.

Note, you do not want to just type in the value 1.08 into cell K3, activity *A*'s EFT. Instead, you create a formula that allows Excel to add the expected activity time (t_E) to the earliest start time (EST). For our spreadsheet, the formula for cell K3 would look like:

$$=I3 + G3$$

We use a formula in Excel because we can utilize the "fill-handle" tool and quickly find all earliest finish times (EFT) by filling down in the K column since EFTs are always the corresponding EST plus t_E.

Now we have calculated all EFTs for the Excel spreadsheet, but we still need to find the ESTs. Each activity's EST is based upon on the EFTs of the activities it is dependent on. Therefore, each EST is found by entering a specific formula separately.

If an activity is dependent on only one activity, then its EST is equal to the EFT of the activity it is dependent on.

For example, activity *F* is only dependent on activity *D*. Therefore, activity *F*'s earliest start time (EST) is equal to activity *D*'s earliest finish time (EFT) which is 4.33 days. In Excel, using the example spreadsheet, the formula for I8, activity *F*'s EST, would be

$$=K6$$

or activity *D*'s EFT.

If an activity is dependent on two or more activities, then the activity's EST is the larger EFT of the previous activities. Why? Because an activity cannot start until all of the activities it is dependent on finish. So the last activity to finish is when our activity can start.

How do we capture this logic in Excel? We use an Excel formula called the "MAX" function.

Recall, if an activity is dependent on two or more activities, then the activity's earliest start time (EST) is the *larger* earliest finish time (EFT) of the previous activities. We want the maximum EFT of the activities depended on.

For example, activity *G* is dependent on activities *E* and *F*. Therefore, activity *G* cannot start until both activity *E* and activity *F* have finished. Activity *G*'s EST is 6.25 days since activity *G* must wait for activity *F*'s EFT of 6.25 days. In Excel, using the example spreadsheet, the formula for I9, activity *G*'s EST, would be

$$=MAX(K7, K8)$$

or the maximum value between activity *E*'s EFT and activity *E*'s EFT.

Note the syntax for the MAX function in Excel is MAX(number1, number2, ...) meaning the function should spit out the maximum value in a list of specified numbers.

In order to calculate EST column, you must be patient and methodical as you enter in specific formulas for each activity's EST. The "fill-handle" tool will not help you in this case. Because activity dependence varies, each EST cell formula will vary as well. This is why we explained earlier that project planning spreadsheets are project specific and do not transfer between projects.

Once you complete the EST column, you have completed the equivalent of the forward pass through the order-requirement digraph (See Figure 19.Appendix.4). Now we will do a backward pass using Excel.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Variance (s^2)	EST	LST	EFT	LFT	Slack (days)		Critical Activity
2				a	m	b	t_E						LST – EST	LFT – EFT	
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	1.08333		0.00		1.08				
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667		1.08		2.75				
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2		1.08		3.08				
6	D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25		3.08		4.33				
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1		4.33		5.33				
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667		4.33		6.25				
9	G	Set up transportation of CHEM-PACKs	E, F	0.5	1	2	1.08333		6.25		7.33				
10	H	Prepare receiving site(s)	F	2	3	5	3.16667		6.25		9.42				
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5		5.33		6.83				
12	J	Load and transport CHEM-PACKs	G, I	1	1.5	3	1.66667		7.33		9.00				
13	K	Unload CHEM-PACKs at receiving site(s)	H, J	0.5	1	2	1.08333		9.42		10.50				

Figure 19.Appendix.4: An example of a complete EST/EFT chart for CHEMPACK container relocation in Excel

With the earliest start and finish times (EST/EFT) completed, we begin calculating the latest start and finish times (LST/LFT) for each activity in Excel. For the relocation of CHEMPACK containers for the Super Bowl, it takes an estimated 10.5 days to complete the project. Therefore, the latest activity *K* (the final node) can finish is 10.5 days. Another way to state this is the latest finish time (LFT) for activity *K* is 10.5 days.

A quick way to figure out what the final node's LFT should be is to scan the EFT column and look for the largest time value. Recall, the largest EFT is going to represent the finishing time of the final node on the

critical path (i.e., the total project completion time). To find the final node's LFT, which is the largest time value in the EFT column, we are going to evoke the MAX() function again. For CHEMPACK containers, our final node is activity K. Therefore, instead of just typing 10.5 days into cell L13, activity K's LFT, we will type in the formula

$$=MAX(K3:K13).$$

> **How come we do not want to type in the value 10.5 in cell L13?**
>
> The power of the spreadsheet is that you can change one number, such as an activity time, and the entire spreadsheet will recalculate values based upon the change.
>
> If you input 10.5 into cell L13, then you guarantee that the LST/LFT of the project will always be based upon 10.5 days instead of the expected project completion time (whatever that value may be).

K3:K13 represents cells K3 through K13, or in this case the EFTs for all activities. MAX() tells Excel to find the largest value out of these cells.

By using the MAX() function, we ensure that activity K, the final activity, will have an LFT equal to the largest EFT. Remember, the final activity is not the last activity on the activity chart, but the activity that is the last one to finish in the project.

To calculate the latest start time (LST), subtract the expected activity time (t_E) from the LFT. For our example, the LST for activity K is:

$$LFT - t_E = 10.5 - 1.08 = 9.42 \text{ days.}$$

Note the process of finding the LST is similar to calculating EFT. Before, we added t_E to the start time to find the finish time. Now, because we are moving backwards (from right to left through the order-requirement digraph), we subtract t_E from the finish time.

Just as before, you do not want to just type in the value 9.42 into cell J13 for the LST of activity K. Instead, you create a formula that allows Excel to subtract the expected activity time (t_E) from the latest finish time (LFT). For our spreadsheet, the formula for activity K's LST would look like:

$$=L13 - G13$$

We use a formula in Excel because we can utilize the "fill-handle" tool and quickly find all latest start times (LST) since LSTs are always the corresponding LFT minus t_E.

We have figured out how to calculate all the LSTs for the Excel spreadsheet, but we still need to find the LFTs. Each activity's LFT is based upon on the LSTs of the activities dependent on it. Therefore, each LFT is found by entering a specific formula separately. This seems a little backwards, but that is the key. We are moving backwards through the order-requirement digraph, so an activity finishes when its dependent activity starts.

If an activity has only one dependent activity, then its LFT is equal to the LST of the dependent activity. For example, activity G only has activity J dependent on it. Therefore, activity G's latest finish time (LFT) is equal to activity J's latest start time (LST) which is 7.75 days, meaning activity G finishes when activity J starts. In Excel, using the example spreadsheet, the formula for L9, activity G's LFT, would be

$$=J12$$

or activity J's LST.

If an activity has two or more dependent activities, then the activity's LFT is the smaller LST of the succeeding activities. Why? It is because an activity finishes when the first dependent activity starts. So the first dependent activity's start time is the latest our activity finishes.

How do we capture this logic in Excel? We use an Excel formula called the "MIN" function. Recall, if an activity has two or more activities dependent on it, then the activity's LFT is the *smaller* LST of the succeeding activities. We want the minimum LST of the dependent activities.

For example, activity D has two activities depending on it, activities E and F. Therefore, activity D must be finished before activity E and activity F start. Activity E's LST is 5.25 days and activity F's LST is 4.33 days. Since activity F starts first, the latest activity D can finish is 4.33 days in order for activity F to start. In Excel, using the example spreadsheet, the formula for L6, activity D's LFT, would be

$$=MIN(J7, J8)$$

or the minimum value between activity E's LST and activity F's LST.

Note the syntax for the MIN function in Excel is MIN(number1, number2, ...) meaning the function should spit out the minimum value in a list of numbers.

In order to calculate LFT column, you must be patient and methodical as you enter in specific formulas for each activity's LFT. The "fill-handle" tool will not help you in this case. Because activity dependence varies, each LFT cell formula will vary as well.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Variance (s^2)	EST	LST	EFT	LFT	Slack (days)		Critical Activity
2				a	m	b	t_E						LST – EST	LFT – EFT	
3	A	Decide the number of CHEM-PACKS to move		0.5	1	2	1.08333		0.00	0.00	1.08	1.08			
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667		1.08	1.42	2.75	3.08			
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2		1.08	1.08	3.08	3.08			
6	D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25		3.08	3.08	4.33	4.33			
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1		4.33	5.25	5.33	6.25			
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667		4.33	4.33	6.25	6.25			
9	G	Set up transportation of CHEM-PACKs	E, F	0.5	1	2	1.08333		6.25	6.67	7.33	7.75			
10	H	Prepare receiving site(s)	F	2	3	5	3.16667		6.25	6.25	9.42	9.42			
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5		5.33	6.25	6.83	7.75			
12	J	Load and transport CHEM-PACKs	G, I	1	1.5	3	1.66667		7.33	7.75	9.00	9.42			
13	K	Unload CHEM-PACKs at receiving site(s)	H, J	0.5	1	2	1.08333		9.42	9.42	10.50	10.50			

Figure 19.Appendix.5: An example of a complete EST/EFT, LST/LFT chart for CHEMPACK container relocation in Excel

Once we calculate the EST/LST and LST/LFT in the spreadsheet, we can finish the remainder of the chart quickly. The slack is calculated just as the column heading indicates, using LST – EST or LFT – EFT. For the LST – EST column in our example spreadsheet, the formula for cell M3, activity A's slack time is

$$=J3 – I3$$

and use the "fill-handle" tool and drag the formula down column M for the rest of the activities. Repeat the same steps for LFT – EFT expect use the formula

$$=L3 – K3$$

in cell N3, activity A's slack time.

The two columns should be equal since the slack can be calculated either way. The critical path includes the critical activities, which are the activities with no slack time. Your chart should look similar to Figure 19.Appendix.6.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Variance (s^2)	EST	LST	EFT	LFT	Slack (days)		Critical Activity
2				a	m	b	t_E						LST – EST	LFT – EFT	
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	1.08333	0.0625	0.00	0.00	1.08	1.08	0.00	0.00	
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667	0.111111111	1.08	1.42	2.75	3.08	0.33	0.33	
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2	0.111111111	1.08	1.08	3.08	3.08	0.00	0.00	
6	D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25	0.173611111	3.08	3.08	4.33	4.33	0.00	0.00	
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1	0.027777778	4.33	5.25	5.33	6.25	0.92	0.92	
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667	0.0625	4.33	4.33	6.25	6.25	0.00	0.00	
9	G	Set up transportation of CHEM-PACKs	E, F	0.5	1	2	1.08333	0.0625	6.25	6.67	7.33	7.75	0.42	0.42	
10	H	Prepare receiving site(s)	F	2	3	5	3.16667	0.25	6.25	6.25	9.42	9.42	0.00	0.00	
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5	0.027777778	5.33	6.25	6.83	7.75	0.92	0.92	
12	J	Load and transport CHEM-PACKs	G, I	1	1.5	3	1.66667	0.111111111	7.33	7.75	9.00	9.42	0.42	0.42	
13	K	Unload CHEM-PACKs at receiving site(s)	H, J	0.5	1	2	1.08333	0.0625	9.42	9.42	10.50	10.50	0.00	0.00	

Figure 19.Appendix.6: An example of a complete EST/EFT, LST/LFT, and Slack chart for CHEMPACK container relocation in Excel

So far, we have learned how to calculate for each activity the expected times (t_E), EST/EFT, LST/LFT, and slack time as well as calculate the critical path for a project. While these calculations provide us a lot of information about the project, using Excel allows us to easily make changes to the project without needing to recalculate values.

We can also use Excel to calculate the probabilities of finishing the project within certain times.

First, we need the estimated variance (s^2) from the expected value of the finish times of the activities. The equation for the estimated variance for a Beta Probability Distribution (BPD) is

$$s^2 = \left(\frac{b-a}{6}\right)^2 .$$

We add a column to the activity chart and label it variance (s^2). In the first activity cell enter the formula in and use the "fill-handle" tool to calculate the other activities' estimated variances. For our example spreadsheet, the formula for H3, the estimated variance for activity A, is
=((F3 – D3)/6)^2.

Second, we need to calculate the total project estimated variance. Recall, this total is the sum of the estimate variances of the critical activities $A, C, D, F, H,$ and K. For our example spreadsheet, the formula for H15, the project estimated variance is
=H3 + H5 + H6 + H8 + H10 + H13.

Another way to calculate the total project variance is to use the SUMIF function in Excel. The syntax for this function is SUMIF(rang, criteria, sum range). This means if a range of cells meets a certain criteria, then sum up specified values.

For total project variance, we only want to add the variances of critical activities, not the whole column. Therefore, if the critical activity column has cells with 'Yes', then we want to sum up the corresponding cells in the variance column. For our example spreadsheet, the formula for H15, the total project variance, is

=SUMIF(O3:O13,"Yes",H3:H13).

The advantage of using SUMIF is that if your critical activities change, you do not need to add up new cells. Excel will automatically recalculate the total project variance for you based on the new set of critical activities.

Third, we need to calculate the estimated standard deviation for the project. The estimated standard deviation is just the square root of the corresponding variance $\sqrt{s^2} = s$. For our example spreadsheet, the formula for H16, the project estimated standard deviation is
=SQRT(H15).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Variance (s^2)	EST	LST	EFT	LFT	Slack (days)		Critical Activity
2				a	m	b	t_E						LST – EST	LFT – EFT	
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	1.08333	0.0625	0.00	0.00	1.08	1.08	0.00	0.00	Yes
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667	0.111111111	1.08	1.42	2.75	3.08	0.33	0.33	
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2	0.111111111	1.08	1.08	3.08	3.08	0.00	0.00	Yes
6	D	Get approval from State and Federal Authorities	B, C	0.5	1	3	1.25	0.173611111	3.08	3.08	4.33	4.33	0.00	0.00	Yes
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1	0.027777778	4.33	5.25	5.33	6.25	0.92	0.92	
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667	0.0625	4.33	4.33	6.25	6.25	0.00	0.00	Yes
9	G	Set up transportation of CHEM-PACKs	E, F	0.5	1	2	1.08333	0.0625	6.25	6.67	7.33	7.75	0.42	0.42	
10	H	Prepare receiving site(s)	F	2	3	5	3.16667	0.25	6.25	6.25	9.42	9.42	0.00	0.00	Yes
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5	0.027777778	5.33	6.25	6.83	7.75	0.92	0.92	
12	J	Load and transport CHEM-PACKs	G, I	1	1.5	3	1.66667	0.111111111	7.33	7.75	9.00	9.42	0.42	0.42	
13	K	Unload CHEM-PACKs at receiving site(s)	H, J	0.5	1	2	1.08333	0.0625	9.42	9.42	10.50	10.50	0.00	0.00	Yes
14															
15							TOTAL Project Variance	0.722222222							
16							s	0.849836586							

Figure 19.Appendix.7: An example of a complete chart for CHEMPACK container relocation in Excel

We finally have the needed information to manipulate project completion time probabilities. In order to determine probabilities, we need use the NORMSDIST() function in Excel.

Beneath our activity chart, we are going to set up two values: *P* and *Days* (see Figure 19.Appendix.8). The *P* is for the probability of an event, and the *Days* is for the number of days before Super Bowl Sunday it takes to complete the project.

In our spreadsheet example, we want to calculate the probability that the project will be completed within 10 days given that the estimated project completion time is 10.5 days.

To calculate the probability, we use the Excel function NORMDIST(). The syntax for NORMDIST() is
NORMDIST(*x*, mean, standard_deviation, culmulative).

The *x* represents the value we are trying to calculate the probability for. In our example, *x* = 10. The mean represents the mean value of the specific normal distribution. In our example, the mean is 10.5, the expected project completion time. The standard_deviation represents the standard deviation (*s*) of the specific normal distribution. In our example, *s* = 0.85. Finally, cumulative represents whether or not the distribution is a cumulative distribution function.   For this entry, you type either TRUE or FALSE. In our example (as with all PERT problems), we use a cumulative distribution function; therefore, we will type TRUE.

Putting this all together, our entry should look like:
=NORMDIST(10,10.5,0.85,TRUE)

However, you want to be as general as possible in Excel so that if you decide to change a value somewhere in your spreadsheet, all other values will also update. Therefore, to calculate the probability that the CHEMPACK containers will be relocated in the 10 days before Super Bowl Sunday, we would input into cell G18
=NORMDIST(I18,L13,H16,TRUE)

Where cell I18 is the 10 day time frame, cell L13 is the project completion time, cell H16 is the standard deviation, and TRUE is the indicator for a cumulative distribution.

The Excel function tells us we have a 27.8% chance of relocating CHEMPACK containers in time if we start 10 days before Super Bowl Sunday.

Lastly, we want to determine how many days we need in order to be 95% certain the project will be completed by Super Bowl Sunday. Therefore, we input 0.95 into cell G19, the specified probability. To calculate the number of days, we use the Excel function NORMINV(). The syntax for NORMINV is
=NORMINV (probability, mean, standard_deviation).

The probability represents the probability we are trying to reach. In our example, the probability is 0.95. The mean represents the mean value of the specific normal distribution. In our example, the mean is 10.5, the expected project completion time. The standard_deviation represents the standard deviation (*s*) of the specific normal distribution. In our example, *s* = 0.85.

Putting this all together, our entry should look like:
=NORMINV (0.95, 10.5, 0.85).
More generally, we would input into cell I19
=NORMINV(G19,L13,H16)

The Excel function tells us we need to plan for 11.9 days in order to be 95% certain the relocation of CHEMPACK containers is completed in time for Super Bowl Sunday.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Activity	Activity Description	Activity	Completion Time (days)				Variance (s^2)	EST	LST	EFT	LFT	Slack (days)		Critical
2	Code		Dependence	a	m	b	t_E						LST – EST	LFT – EFT	Activity
3	A	Decide the number of CHEM-PACKs to move		0.5	1	2	1.08333	0.0625	0.00	0.00	1.08	1.08	0.00	0.00	Yes
4	B	Decide where the CHEM-PACKs will be moved to	A	1	1.5	3	1.66667	0.111111111	1.08	1.42	2.75	3.08	0.33	0.33	
5	C	Decide where the CHEM-PACKs will be moved from	A	1	2	3	2	0.111111111	1.08	1.08	3.08	3.08	0.00	0.00	Yes
6	D	Get approval from State and Federal Authorities	B , C	0.5	1	3	1.25	0.173611111	3.08	3.08	4.33	4.33	0.00	0.00	Yes
7	E	Discuss strategy with current location(s)	D	0.5	1	1.5	1	0.027777778	4.33	5.25	5.33	6.25	0.92	0.92	
8	F	Discuss strategy with receiving location(s)	D	1	2	2.5	1.91667	0.0625	4.33	4.33	6.25	6.25	0.00	0.00	Yes
9	G	Set up transportation of CHEM-PACKs	E , F	0.5	1	2	1.08333	0.0625	6.25	6.67	7.33	7.75	0.42	0.42	
10	H	Prepare receiving site(s)	F	2	3	5	3.16667	0.25	6.25	6.25	9.42	9.42	0.00	0.00	Yes
11	I	Prepare CHEM-PACKs for relocation	E	1	1.5	2	1.5	0.027777778	5.33	6.25	6.83	7.75	0.92	0.92	
12	J	Load and transport CHEM-PACKs	G , I	1	1.5	3	1.66667	0.111111111	7.33	7.75	9.00	9.42	0.42	0.42	
13	K	Unload CHEM-PACKs at receiving site(s)	H , J	0.5	1	2	1.08333	0.0625	9.42	9.42	10.50	10.50	0.00	0.00	Yes
14															
15							TOTAL Project Variance	0.722222222							
16							s	0.849836586							
17															
18		Probability of project completion within 10 days					P =	0.27815	Days	10					
19		Number of days to be 95% certain of project completion					P =	0.95	Days	11.89785679					

Figure 19.Appendix.8: An example of the activity chart for CHEMPACK container relocation in Excel with possible probabilities

In this Appendix, we demonstrate how you can use Microsoft Excel to do PERT project management. We first calculated the expected times (t_E), EST/EFT, LST/LFT, and slack time for each activity. We then determined the critical path and calculated the total project estimated variance based upon the critical activities' estimated variances. Finally, we utilized built in normal distribution functions in Excel to calculate various probabilities based upon our project completion time. We emphasize each spreadsheet is a unique PERT project and you cannot transfer one PERT spreadsheet to another PERT spreadsheet. However, once you understand the procedure, it is not too difficult to recreate an Excel spreadsheet for each project you are planning.
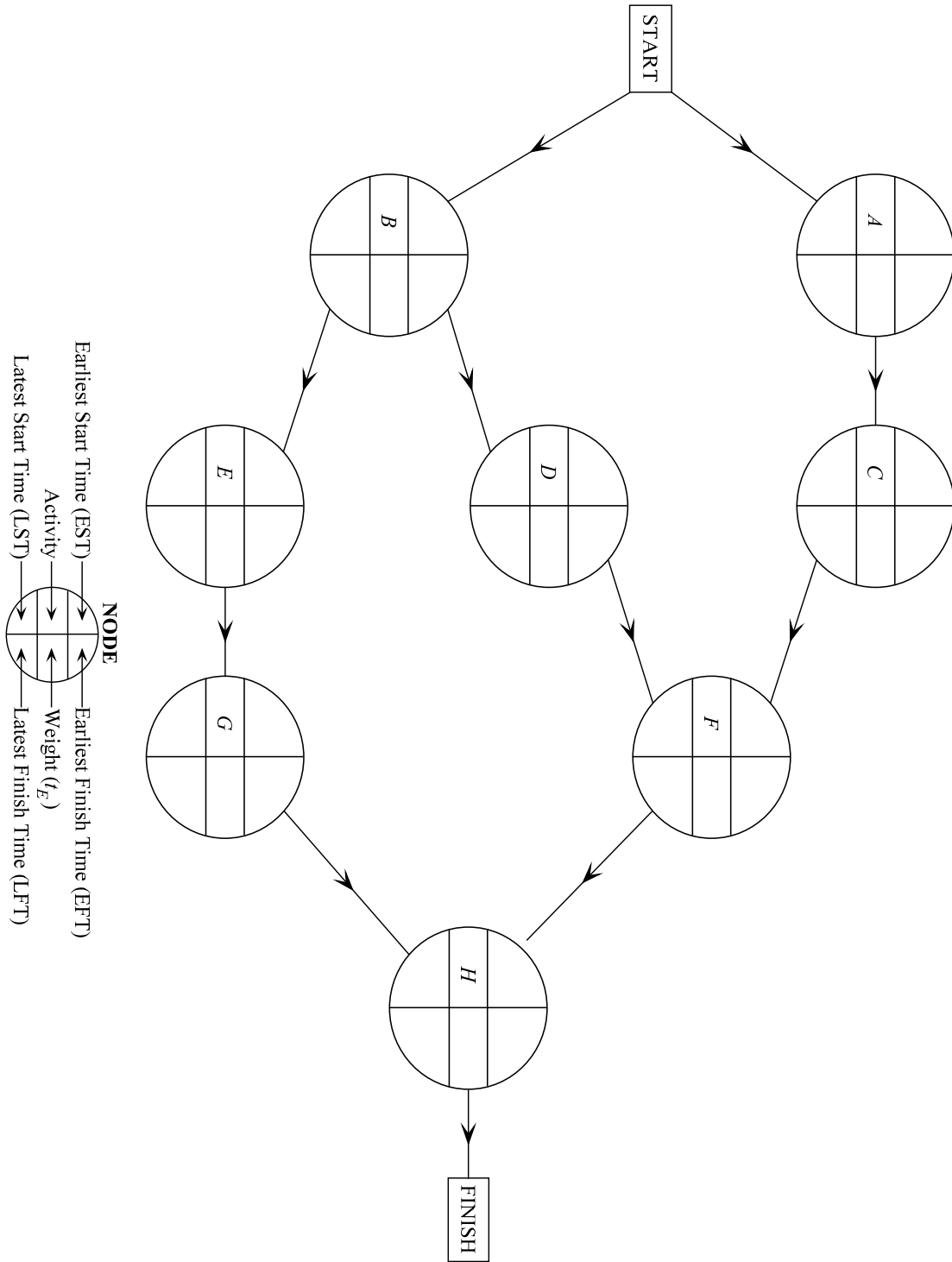
Section 19.4: Chapter 19 (PERT) Homework Questions

1. Given the information in the table below, answer the following:
   a. Compute the expected time for each activity (t_E).
   b. Identify the critical path.
   c. Compute the expected time of the project (T_E).
   d. Compute the variance for each activity (s^2).
   e. Compute the probability of completing the project within 74 days.

PROJECT

Activity Code	Activity Dependence	Completion Time (days)				Critical Activity	Variance (s^2)
		a	m	b	t_E		
A	-----	6	12	24			
B	-----	2	5	8			
C	A	20	35	50			
D	B	1	1	1			
E	B	17	29	47			
F	C, D	13	16	19			
G	E	16	19	28			
H	F, G	4	7	10			

Total Project Variance

PROJECT

Activity	EST	LST	EFT	LFT	Slack (days)		Critical Activity
					LST – EST	LFT – EFT	
A							
B							
C							
D							
E							
F							
G							
H							

Critical Path: _____

Project Completion Time is _____ days.

NODE

Earliest Start Time (EST) —— Earliest Finish Time (EFT)
Activity —— Weight (t_E)
Latest Start Time (LST) —— Latest Finish Time (LFT)

START

A

B

C

D

E

F

G

H

FINISH

2. Given the network information below for the Corporate Consolidation, answer the following:
   a. Compute the expected time for each activity (t_E).
   b. Draw an order-requirement digraph illustrating the network of this project.
   c. Identify the critical path.
   d. Compute the expected time of the project (T_E).
   e. Compute the variance for each activity (s^2).
   f. Compute the probability of completing the project within 112 days; within 116 days.
   g. Compute the probability of completing Activity K, Negotiate with Unions, by day 90.

CORPORATE CONSOLIDATION

Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
A	Codify accounts	-----	16	19	28			
B	Files articles of unification	-----	30	30	30			
C	Unify price and credit policy	-----	60	72	90			
D	Unify personnel policies	-----	18	27	30			
E	Unify data processing	A	17	29	47			
F	Train accounting staff	A	4	7	10			
G	Pilot run data processing	E	12	15	18			
H	Calculate P&L and balance sheet	F, G	6	12	24			
I	Transfer real property	B	18	27	30			
J	Train sales force	C	20	35	50			
K	Negotiate with unions	D	40	55	100			
L	Determine capital needs	H	11	20	29			
M	Explain new personnel policies	K	14	23	26			
N	Secure line of credit	I, L	13	16	19			

Total Project Variance

CORPORATE CONSOLIDATION

| Activity | EST | LST | EFT | LFT | Slack (days) | | Critical Activity |
					LST – EST	LFT – EFT	
A							
B							
C							
D							
E							
F							
G							
H							
I							
J							
K							
L							
M							
N							

Critical Path: _____

Project Completion Time is _____ days.

3. The Vlaamse Maatschappij voor Watervooziening (VMW) is a Flemish water distribution company that covers approximately 50% of Flanders in the northern region of Belgium. The VMW services 2.5 million customers with a pipeline network of 27,000 km (~16.7 miles) and a yearly production of 140 billion liters (~37 million gallons) of water.

    VMW is remodeling the water production center (WPC) of Kluizen, which VMW owns. This center currently produces and delivers water by transforming surface water into drinkable water and distributing it toward customers. Therefore, surface water is taken from the area around the WPC and is stored in two open reservoirs with a total capacity of 11,000,000 m^3 (~2,905,892,576 gallons).

    VMW wants to expand the storage capacity of the treated water by building two extra reservoirs for the treated water. Each new reservoir will have a capacity of 15,000m^3 (~3,962,581 gallons).

    A list of the detailed activities of these steps is depicted in the table below.

    a. Compute the expected time for each activity (t_E).
    b. Draw an order-requirement digraph illustrating the network of this project.
    c. Identify the critical path.
    d. Compute the expected time of the project (T_E).
    e. Compute the variance for each activity (s^2).
    f. Compute the probability of completing the project on schedule. Does this probability make sense? Why?
    g. How many weeks are needed to be 95% certain the project is completed on time?

BUILDING WATER RESERVOIRS

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
A	Obtain building license	-----	4	24	36			
B	Obtain environmental license	-----	13	18	25			
C	Find a contractor	A, B	8	24	25			
D	Design reservoirs	-----	29	30	41			
E	Determine the specifications of the reservoirs	D	12	27	39			
F	Call for material bids	E	4	9	12			
G	Steel fabrication	F	15	47	73			
H	Negotiations with power distribution company	E	36	40	45			
I	Pour Concrete	F	12	26	30			
J	Install electrical work	H, I	5	14	21			
K	Getting HSC operating	G, J	24	45	49			
L	Update existing high-pressure room and pumps	------	46	57	75			
M	Construct pipes between reservoirs	J, L	11	14	20			

Total Project Variance

BUILDING WATER RESERVOIRS

Activity	EST	LST	EFT	LFT	Slack (weeks)		Critical Activity
					LST – EST	LFT – EFT	
A							
B							
C							
D							
E							
F							
G							
H							
I							
J							
K							
L							
M							

Critical Path: _____

Project Completion Time is _____ Weeks.

4. NASA's goal in 2050 is to have humans living on the planet Mars. In order to build colonies for people to inhabit Mars, NASA needs to provide shelter for relocated families. Your company, Wolfpack, is submitting a bid to build shelters. You envision a pod-like unit that robots can assemble on Mars before a family arrives. You anticipate that the assembly of these shelter units will consist of 16 separate tasks, described below.

   a. Draw an order-requirement digraph illustrating the network of this project.
   b. Identify the critical path.
   c. Compute the expected time of the project (T_E).
   d. Compute the probability of completing the project within 2 years (104 weeks); within 3 years (156 weeks).
   e. Because you are submitting a bid to NASA, you want to be fairly comfortable with the estimated time you give in your proposal. How many weeks are needed for you to be 98% certain the project is completed on time? What are some drawbacks to giving NASA this estimated time?
   f.
      i. Assume we know that Activity O, Pressurize Dome, will take exactly 3 weeks to complete. How does this change your answers in Question (18.5.4b) and Question (18.5.4c)?
      ii. Assume we know that Activity O, Pressurize Dome, will take exactly 5 weeks to complete. How does this change your answers in Question (18.5.4b) and Question (18.5.4c)?
   g. What if NASA asked you to find the "worst possible scenario". Use PERT and the most pessimistic times to find the critical path and completion time.

ASSEMBLY OF MARS SHELTER UNITS

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	*m*	*b*	t_E		
A	Assemble pad	-----	14	18	37	20.5		14.69
B	Assemble flooring	-----	15	34	42	32.17		20.25
C	Assemble wall units	-----	9	36	45	33		36
D	Assemble dome frame	-----	37	50	55	48.67		9
E	Install floors	A, B	4	12	21	12.17		8.03
F	Install interior walls	C, E	22	29	34	28.67		4
G	Install dome frame	D, F	7	13	15	12.33		1.78
H	Install plumbing	F	30	40	56	41		18.78
I	Install atomic power plant	F	12	19	45	22.17		30.25
J	Install pressurized unit	G, I	32	34	46	35.67		5.44
K	Install heating/cooling units	I	15	25	28	23.83		4.69
L	Install satellite	G, I	2	3	5	3.17		0.25
M	Install commode and sink	H, K	1	2	4	2.17		0.25
N	Complete interior finish work	M	6	10	17	10.5		3.36
O	Pressurize dome	J, K	2	3	8	3.67		1
P	Install entertainment unit	N	2	4	5	3.83		0.25

Total Project Variance

ASSEMBLY OF MARS SHELTER UNITS

Activity	EST	LST	EFT	LFT	Slack (weeks)		Critical Activity
					LST – EST	LFT – EFT	
A							
B							
C							
D							
E							
F							
G							
H							
I							
J							
K							
L							
M							
N							
O							
P							

Critical Path: _____

Project Completion Time is _____ weeks.

5. Bayside High is considering switching from a hardcover yearbook to a digital yearbook on a CD-ROM. Zack, the editor of the yearbook and originator of the idea, is putting together a plan for the yearbook committee to execute. Below is a list of activities that need to be accomplished.

    a. Interview your yearbook staff and faculty advisor to find out the optimistic, pessimistic, and most probable times for each activity.
    b. Compute the expected time (t_E) and variance (s^2) for each activity.
    c. Identify the critical path.
    d. Compute the expected time of the project (T_E).
    e. Compute the probability of completing the project within the school year.
    f. Can you make any recommend ways in which the yearbook staff could shorted their total project completion time (T_E)?

CREATING A DIGITAL YEARBOOK

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
A	Get a cost estimate for burning the CD-ROMs	-----						
B	Create a budget	A						
C	Decide on a name and theme	-----						
D	Send out order forms with the price included	B						
E	Obtain photographs of the student body	-----						
F	Obtain photographs of the faculty and staff	-----						
G	Advertise for stories and photographs from the student body	-----						
H	Collect stories and photographs from both student body and yearbook staff	G						
Table continues on the next page.								

CREATING A DIGITAL YEARBOOK (continued)

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
I	Select and edit material	H						
J	Design a layout	C						
K	Develop a design for the CD-ROM label and the case	C						
L	Obtain pictures and member list of all clubs and sports teams	-----						
M	Collect orders and money	D						
N	Input photographs and material into the layout design	E, F, I, J, L						
O	Final edit	N						
P	Create a master copy	K, O						
Q	Send out master copy to be mass produced	P						
R	Disperse the yearbook to students who placed an order	M, Q						
						Total Project Variance		

CREATING A DIGITAL YEARBOOK

Activity	EST	LST	EFT	LFT	Slack (weeks)		Critical Activity
					LST – EST	LFT – EFT	
A							
B							
C							
D							
E							
F							
G							
H							
Table continues on the next page.							

CREATING A DIGITAL YEARBOOK (continued)

Activity	EST	LST	EFT	LFT	Slack (weeks)		Critical Activity
					LST – EST	LFT – EFT	
I							
J							
K							
L							
M							
N							
O							
P							
Q							
R							

Critical Path: _____

Project Completion Time is _____ weeks.

NODE

Earliest Start Time (EST)
Activity
Latest Start Time (LST)

Earli
Weig
Lates

Chapter 19 Summary

What have we learned?

Terms

Program Evaluation Review Technique (PERT)

Expected Completion Time

Chapter 19 (Location Problems) Objectives

You should be able to:

- Given a contextual problem, estimate the optimistic/pessimistic/most probable activity completion times and calculate the expected activity completion time.

- Given a contextual problem, find the critical path of a project using the critical path method (via order-requirement digraph or Excel spreadsheet).

- Investigate expected total project completion time using the normal distribution.

- **Analyze and interpret results; make decisions based on results.**

Chapter 19 Study Guide

1.

References

Ciano, B. (2011). Using PERT for estimating tasks. Retrieved March 15, 2011, from
   http://www.brunociano.com/2007/08/using-pert-for-estimating-tasks.html

Structured Data, LLC. (2011). RiskAMP: The beta-PERT Distribution. Retrieved May 3, 2011, from
   http://www.riskamp.com/library/pertdistribution.php

Tannenbaum, P. (2007). *Excursions in modern mathematics with mini-excursions (6th ed.).* Upper Saddle
   River, NJ: Pearson Education. p. 280.

Vanhoucke, M., & Demeulemeester, E. (2003). The application of project scheduling techniques in a real-
   life environment. *Project Management Journal, 34*(1), 30-42.

Worksheet 1
Table 19.1.2: Activity chart for making a taco dinner for PERT

PREPARING A TACO DINNER

Activity Code	Activity Description	Activity Dependence	Completion Time (minutes)			
			Shortest Possible Time (*a*)	Most Probable Time (*m*)	Longest Possible Time (*b*)	Expected Time (*t_E*)
A	Find recipe for taco salad on Internet	—		5		
B	Find recipe for dessert in Mom's recipe box	—		6		
C	Make graham cracker crust	B		4		
D	Clean lettuce	A		3		
E	Find cheese in refrigerator	A		1		
F	Shred cheese	E		4		
G	Brown hamburger and assemble tacos (including cheese)	D, F		14		
H	Prepare the pudding	B		8		
I	Assemble the pudding pie	C, H		4		
J	Put out all the food for dinner	G, I		2		

Worksheet 2
Figure 19.1.1: Order-requirement digraph representing the problem of preparing a taco dinner
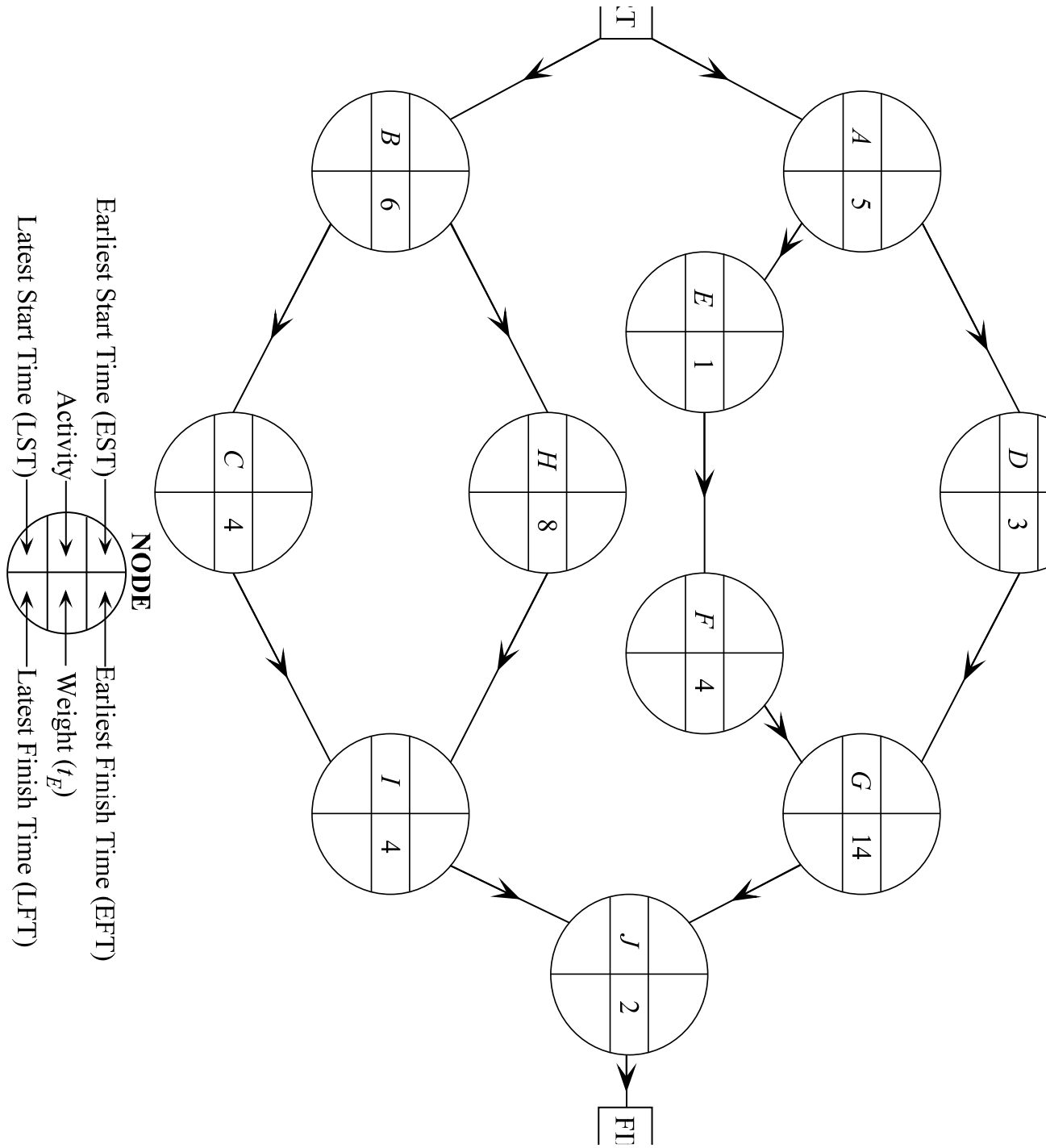
Worksheet 3
Table 19.2.4: Activity chart for CHEMPACK container relocation including the expected time and variance

RELOCATION OF CHEMPACK CONTAINERS FOR THE SUPER BOWL

Activity Code	Activity Description	Activity Dependence	Completion Time (days)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
A	Decide the number of CHEMPACK containers to move		0.5	1	2			
B	Decide where the CHEMPACK containers will be moved to	A	1	1.5	3			
C	Decide where the CHEMPACK containers will be moved from	A	1	2	3			
D	Get approval from State and Federal Authorities	B, C	0.5	1	3			
E	Discuss strategy with current location(s)	D	0.5	1	1.5			
F	Discuss strategy with receiving location(s)	D	1	2	2.5			
G	Set up transportation of CHEMPACK containers	E, F	0.5	1	2			
H	Prepare receiving site(s)	F	2	3	5			
I	Prepare CHEMPACK containers for relocation	E	1	1.5	2			
J	Load and transport CHEMPACK containers	G, I	1	1.5	3			
K	Unload CHEMPACK containers at receiving site(s)	H, J	0.5	1	2			
							Total Project Variance	

Worksheet 4
Figure 19.2.1: The order-requirement digraph of the relocation of CHEMPACK containers for the Super Bowl
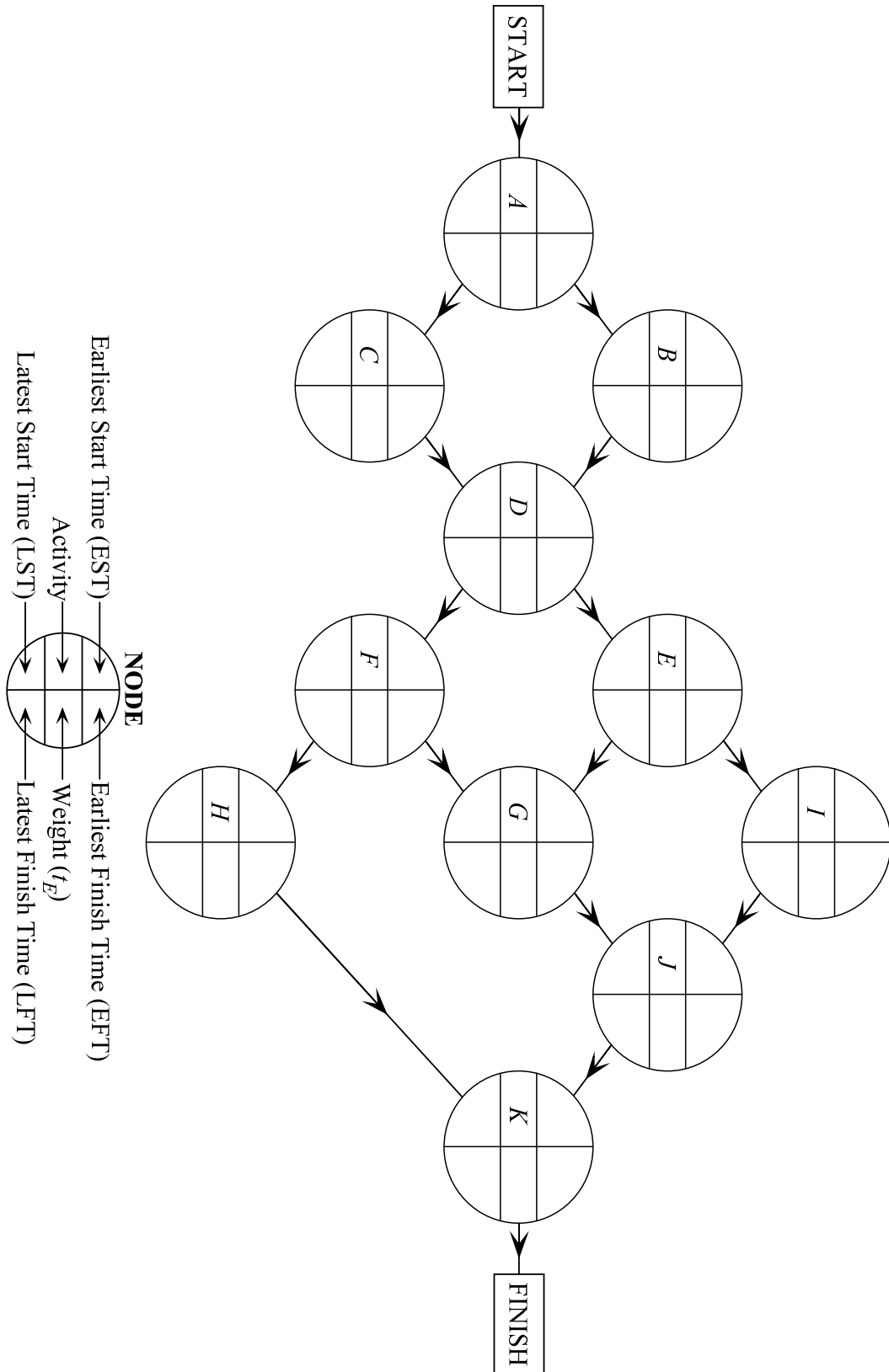
Worksheet 5
Table 19.2.3: Chart for EST, EFT, LST, LFT, slack time, and critical path of the Relocation of CHEMPACK containers for the Super Bowl

RELOCATION OF CHEMPACK CONTAINERS FOR THE SUPER BOWL

Activity	EST	LST	EFT	LFT	Slack (days)		Critical Activity
					LST – EST	LFT – EFT	
A							
B							
C							
D							
E							
F							
G							
H							
I							
J							
K							

Critical Path: _____

Project Completion Time is _____ days.

Worksheet 6

Figure 19.3.1: The order-requirement digraph for the construction of a TV Tower

NODE

Earliest Start Time (EST) ——————→ ←—————— Earliest Finish Time (EFT)
Activity ——————→ ←—————— Weight (t_E)
Latest Start Time (LST) ——————→ ←—————— Latest Finish Time (LFT)
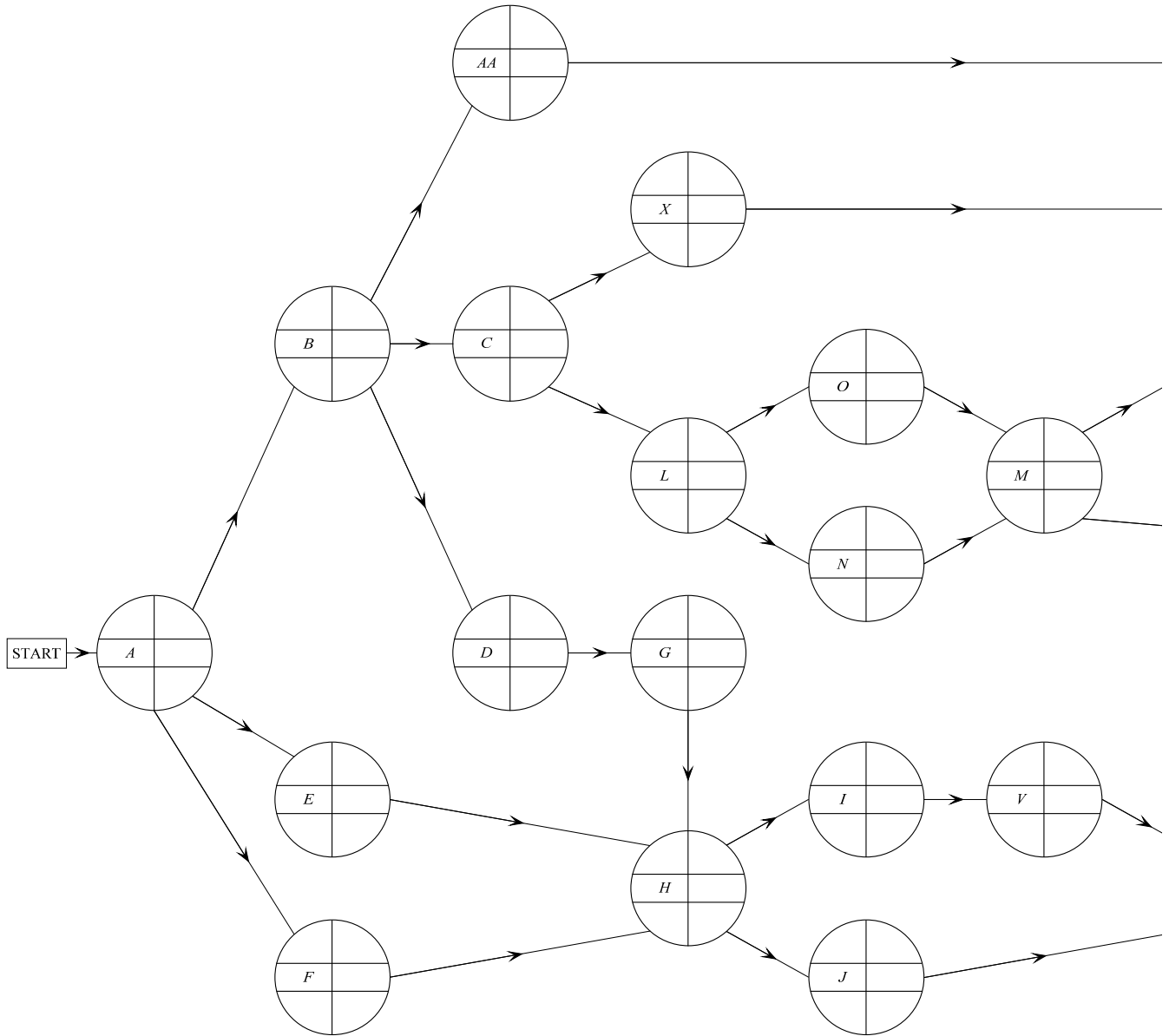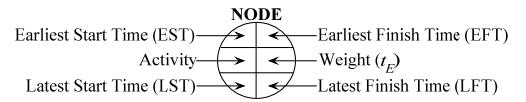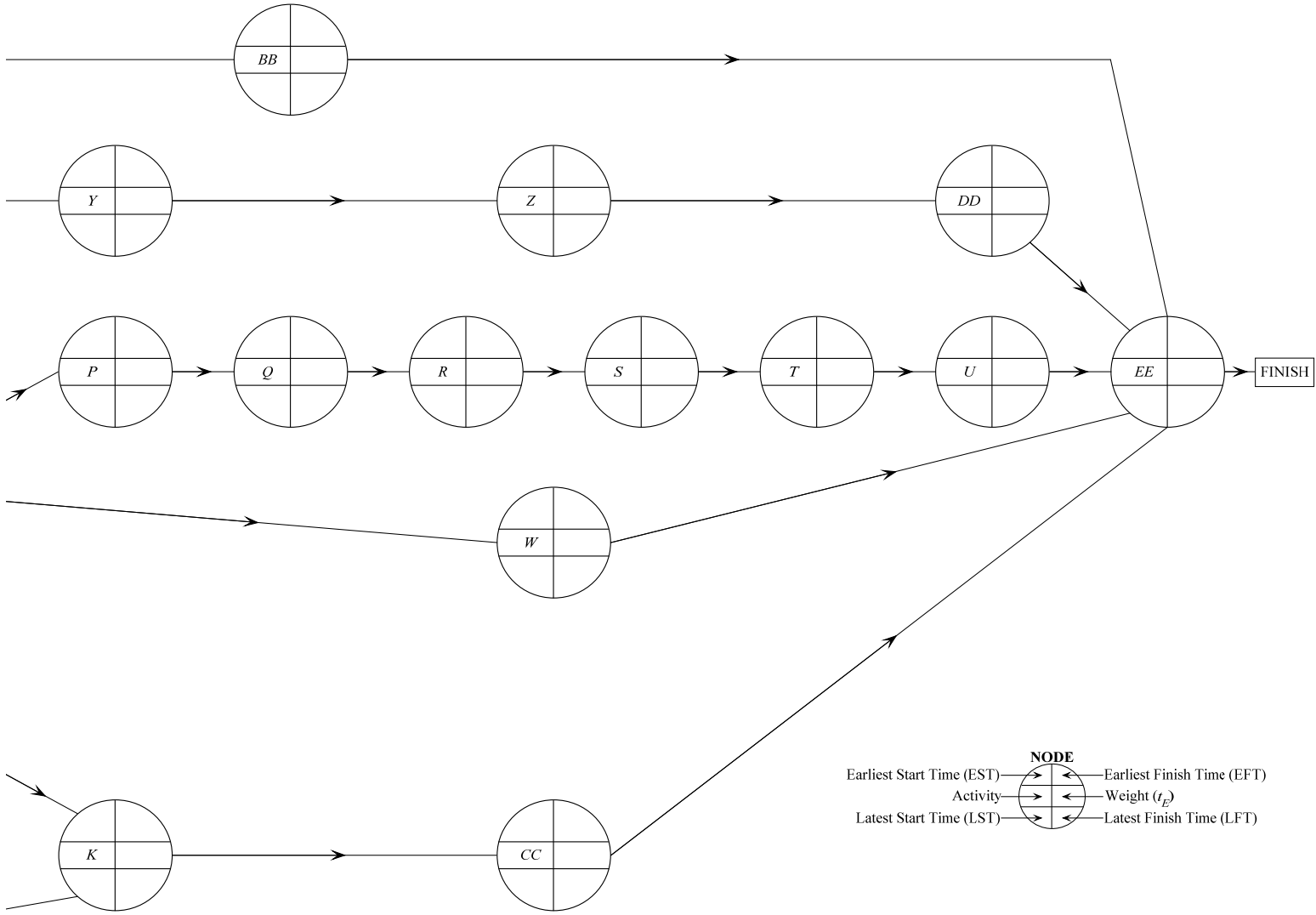
Worksheet 7
Table 19.3.3: Activity chart for the construction of a TV Tower

CONSTRUCTION OF A TV TOWER

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	*m*	*b*	t_E		
A	Sign contract and complete negotiations		4	5	6			
B	Survey site	A	4	6	8			
C	Grade building site and excavate	B	6	8	10			
D	Grade tower site	B	21	30	39			
E	Procure structural steel	A	85	85	85			
F	Procure electrical equipment	A	120	120	120			
G	Procure concrete for footings	D	25	42	59			
H	Erect tower and install E. E.	E, F, G	25	38	51			
I	Install connecting cable	H	4	8	12			
J	Install drain tile and storm drain	H	18	35	52			
K	Backfill and grade tower site	J, V	4	8	12			
L	Pour building footings	C	21	29	37			
M	Pour basement and tank slabs	N, O	11	14	17			
N	Pour outside basement walls	L	30	34	38			
O	Pour walls for basement rooms	L	7	9	11			
P	Pour concrete floor beams	M	10	11	12			

Worksheet 7 (Continued)
Table 19.3.3: Activity chart for the construction of a TV Tower

CONSTRUCTION OF A TV TOWER

Activity Code	Activity Description	Activity Dependence	Completion Time (weeks)				Critical Activity	Variance (s^2)
			a	m	b	t_E		
Q	Pour main floor slab	P	10	12	14			
R	Pour roof slab	Q	13	15	17			
S	Complete interior framing	R	30	42	54			
T	Lay roof	S	2	3	4			
U	Paint building interior	T	13	19	25			
V	Install main cable-site to building	I	25	35	45			
W	Install fuel tank	M	2	3	4			
X	Install septic tank	C	8	12	16			
Y	Install drain tile and storm drain	X	10	15	20			
Z	Backfill around building	Y	7	9	11			
AA	Lay base for connecting road	B	13	15	17			
BB	Complete grading and surface connecting road	AA	5	8	11			
CC	Clean up tower site	K	3	5	7			
DD	Clean up building site	Z	2	3	4			
EE	Contract acceptance obtained and sign-off	U, W, BB, CC, DD	5	5	5			
							Total Project Variance	

Section 20.1: Using Markov Chains to Make Predictions about Total Cholesterol

You probably have heard your parents or other adults talk about high cholesterol. Cholesterol is a soft, fat-like, waxy substance found in the bloodstream and in all your body's cells. Everyone needs cholesterol, but a person with high total cholesterol (HTC) has an increased risk of having a heart attack or stroke. Different people use different numbers, but generally a count of over 240 mg/dL is considered high. Researchers in the health industry, such as pharmaceutical companies, often study the chances of people having high cholesterol. In particular, some researchers have considered how total cholesterol (TC) changes as people grow older or how TC levels change over time. In the first section of this chapter, we explore this situation by following the changes in a patient's Total Cholesterol (TC) at discrete five year intervals. In this model we choose to study a patient's progression between ATC and HTC at five year intervals in order to make it easier to look at a longer period of time.

20.1.1 Total Cholesterol

Let us consider a 40-year-old patient who currently has appropriate total cholesterol (ATC). We would like to model the probabilities of the patient continuing to have ATC and of the patient's TC increasing to High Total Cholesterol (HTC). The patient's TC is currently at an acceptable level (ATC), but factors such as poor diet and lack of exercise could force his/her TC to increase or it might stay the same. Similarly, a patient's TC may be High (HTC), and it may stay high (HTC) or it may become ATC.

Q1.   What are some factors that may decrease a person's TC from HTC to ATC?

According to a data collection done by a researcher who looked at several hundred patients and examined what happened after five years, we approximate the following probabilities concerning a 40-year-old patient's TC in 5 year intervals.

> P(ATC stays ATC) = 0.88
> P(ATC becomes HTC) = 0.12
> P(HTC becomes ATC) = 0.16
> P(HTC stays HTC) = 0.84

In each case you would interpret that the probability that the first state of cholesterol level moves to the second state with the probability given. For example, P(ATC stays ATC) = 0.88 means that the probability that someone with an acceptable total cholesterol will still have acceptable cholesterol after five years is 0.88 or 88%.

Q2.   What do you notice about the first two and the second two probabilities? Why is this true?

Q3.   If a patient has HTC, what is the probability that he still has HTC after 5 years? Write your answer as a reduced fraction.

Now, we can look at long-term probabilities of Total Cholesterol in patients. For example, what is the probability that a patient with ATC will still have ATC in 10 or 15 years?

For simplicity, we make an *important assumption* that the probabilities are independent, meaning that the probability of each five year period does not depend on the probability before. We cannot be sure of long term results, but we can use these numbers as approximations over time. Additionally, in the next section,

we will consider what happens if there is intervention that affects the probabilities, but for now, we will assume that the probabilities are independent.

Q4.   Is it reasonable to assume that the probabilities will be the same for every five-year interval or would you expect them to change based on starting age? Explain.

Q5.   What is the probability that a person with ATC now will continue to have ATC for 15 years, that is have it all the way through? We find this by multiplying the probabilities because we assume that the events are independent. (See Chapter 11 for a review of this idea of multiplying probabilities if needed.)

Q6.   What is the probability that a person with HTC will continue to have HTC for 20 years?

Q7.   What is the probability that a person with ATC will have acceptable cholesterol for ten years, and then get HTC in the following 5 years?

We could continue to do this many times by hand or with a calculator for each question and patient we have over long periods of time. However, matrices offer a way to make the calculations more efficient and allow us to do significant mathematical work without having to do individual calculations. When one uses matrices with these probabilities, they use what is called a **discrete Markov chain process**. Markov chains have the **Markov property**, which means that each future state depends only on the present state and is independent of past states (also known as "Memorylessness"), and Markov chains are a type of **stochastic process**, which means that all state transitions are probabilistic (also known as "random process"). A **Markov chain** is a stochastic process with the Markov property.

To use the discrete Markov chain process, we must build a matrix and then use it to predict the future for all patients in all situations. We use the matrix P (called the **transition matrix**) to consider the probabilities for patient cholesterol levels in 5 years.

$$P = \begin{bmatrix} P(\text{ATC} \to \text{ATC}) & P(\text{ATC} \to \text{HTC}) \\ P(\text{HTC} \to \text{ATC}) & P(\text{HTC} \to \text{HTC}) \end{bmatrix} = \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix}.$$

Q8.   Explain where each of these numbers came from.

In addition to the matrix, the Markov process model can be visually displayed as in Figure 1. We saw above that the same probability matrix is used to move from each five year period to the next one. This fact is represented by the four numbers on the arrows between the circles in Column 1 and Column 2 and between Column 2 and Column 3.
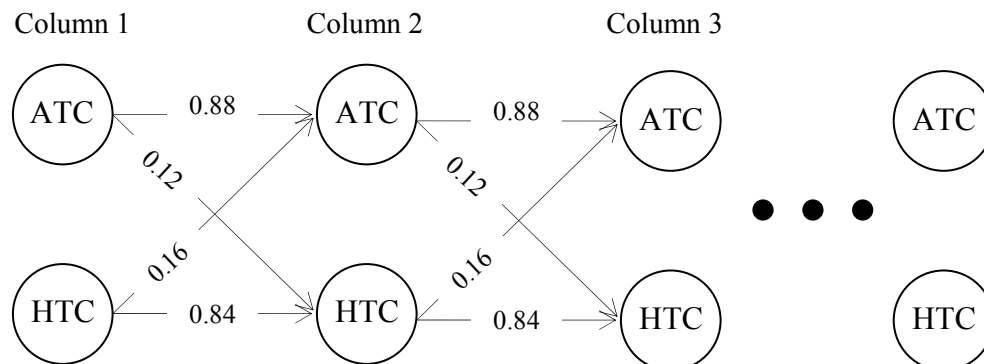
Figure 20.1.1: Diagram illustrating the Markov process model to examine the progression of a patient's Total Cholesterol over time, in five-year intervals

We can use the diagram to examine the progression of a patient's Total Cholesterol. For example, if a patient has HTC, we can find the probability he will reach a certain state after a few periods of five years by multiplying the probabilities on the arrows. To find the probability that a patient with ATC at 40 has HTC at 50, we would look at the two paths from ATC in Column 1 to HTC in Column 3. The patient could continue to have ATC in the first five years but then have HTC after the second five years (i.e., $0.88 \cdot 0.12$). Alternatively, the patient could have HTC after the first five years and then continue to have HTC after the second five years (i.e., $0.12 \cdot 0.84$). Thus, we calculate $0.88 \cdot 0.12 + 0.12 \cdot 0.84$ as the two paths to get from ATC to HTC after 10 years and conclude that there is a 0.2064 probability that a patient with ATC at 40 has HTC at 50.

Q9.   What is the probability that a patient at age 40 with HTC will have ATC at 50?

Q10.   What is the probability that a patient at age 40 with HTC will have ATC at 60?

In this case, we had 8 possible paths to look at in the diagram to go from age 40 to age 60. At this point, we may notice that using the diagram to determine probabilities is time-consuming and as we continue to look at more values and more years, it will get even worse. We can use matrices instead. Recall that our transition matrix (defined as the matrix that allows us to look at probabilities for transition in a process) is:

$$P = \begin{array}{c} \\ ATC \\ HTC \end{array} \begin{array}{cc} ATC & HTC \\ \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} \end{array}.$$

To determine the probability that a patient with ATC at 40 has HTC at 50, we needed to calculate $0.88 \cdot 0.12 + 0.12 \cdot 0.84 = 0.2064$. Notice that we could obtain this same calculation if we multiplied the transition matrix P by itself:

$$P \cdot P = \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} \cdot \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix}$$

$$= \begin{bmatrix} 0.88 \cdot 0.88 + 0.12 \cdot 0.16 & \boxed{0.88 \cdot 0.12 + 0.12 \cdot 0.84} \\ 0.16 \cdot 0.88 + 0.84 \cdot 0.16 & 0.16 \cdot 0.12 + 0.84 \cdot 0.84 \end{bmatrix}$$

$$= \begin{bmatrix} 0.7936 & 0.2064 \\ 0.2752 & 0.7248 \end{bmatrix}$$

Recall from algebra: when you multiply a 2×2 matrix by a 2×2 matrix, you get a 2×2 matrix. If you need a review of multiplication of matrices, you may look in an algebra book or on the internet.

Here, we see that the $P(\text{ATC} \rightarrow \text{HTC}) = 0.2604$, which is exactly what we calculated by hand.

Q11.   Using the matrix, what is the probability that a patient at age 40 with HTC will have ATC at age 50? Compare your answer to your response for (8).

Next, to calculate the probability that a patient at age 40 with HTC will have ATC at age 60, we can again use the transition matrix. This time, we would need to multiply P by itself four times because there are four five-year intervals in this time period.

$$P \cdot P \cdot P \cdot P = \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} \cdot \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} \cdot \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} \cdot \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} = \begin{bmatrix} 0.6866 & 0.3134 \\ 0.4179 & 0.5821 \end{bmatrix}$$

To make this calculation a bit easier, we can instead write and enter into our calculator:

$$P^4 = \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix}^4 = \begin{bmatrix} 0.6866 & 0.3134 \\ 0.4179 & 0.5821 \end{bmatrix}$$

Q12.   Using the matrix, what is the probability that a patient at age 40 with HTC will have ATC at 60? Compare your answer to your response for (9).

20.1.2 Matrix Multiplication on the Graphing Calculator

We can do the above example on a TI calculator (TI-83 or TI-84).

First, press the MATRIX key.
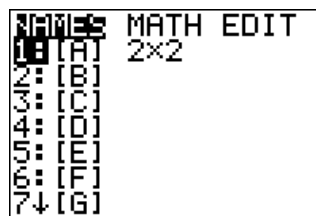
Select EDIT 1: [›] [›] [ENTER] .

Choose a 2×2 matrix, and enter the probabilities into [A].

Go back to the main screen: [2nd] [MODE] .

Go back to the matrix screen: [2nd] [x^{-1}], and press [ENTER] to get to matrix [A].

Raise [A] to the 4th power: [^] [4] [ENTER] .

Use the right and left arrows to see the entire matrix.

```
[A]^4
[[.68660224 .31…
 [.41786368 .58…
```
```
[A]^4
…224 .31339776]
…368 .58213632]]
```

20.1.3 1x2 Matrix[1]

You can also use the matrix to predict the cholesterol levels for individual patients for years to come. We let $\begin{bmatrix} p_t & q_t \end{bmatrix}$ represent the probabilities a patient has ATC (p) and HTC (q) at time t. In this text, we call $\begin{bmatrix} p_t & q_t \end{bmatrix}$ a 1×2 matrix. If, for example, a patient has HTC at age 60, then $p_{60} = 0$ (i.e., the probability that this person has ATC is 0) and $q_{60} = 1$ (i.e., the probability that this person has HTC is 1) so $\begin{bmatrix} p_{60} & q_{60} \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix}$, When we are considering a single individual at a specific time, the probability matrix will always be $\begin{bmatrix} 0 & 1 \end{bmatrix}$ or $\begin{bmatrix} 1 & 0 \end{bmatrix}$ because they either have HTC or have ATC. Later, we will look at whole populations, rather than a single individual, and consider the probability matrices of the entire population.

If we look at another patient, we could write $\begin{bmatrix} p_{40} & q_{40} \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and conclude that this patient has ATC at age 40 (i.e., they initially have ATC with probability 1 and HTC with probability 0). Using the matrix P, we can find the probability the patient will have ATC at age 45: $\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} = \begin{bmatrix} 0.88 & 0.12 \end{bmatrix}$.

Therefore, the patient will have ATC with probability 0.88.

> Recall from algebra: when you multiply a 1×2 matrix by a 2×2 matrix, you get a 1×2 matrix. In this case, it is just the top row of the second matrix.

Q13.   Continuing to look at this answer vector, what is the probability that this same patient will have HTC?

These problems may seem a little trivial, but we can use matrix multiplication to determine many more probabilities.

Q14.   What are these two probabilities (p_{80} and q_{80}) for the patient at age 80?

Q15.   Find the two probabilities (p_{80} and q_{80}) for the patient at age 80 if we know for certain that they have HTC at age 40.

Q16.   Would a patient with ATC at age 40 be more or less likely to have a heart attack or stroke at age 80 than a patient with HTC at age 40? Explain your reasoning.

Next, we will consider an entire population. If in a given population, 60% have ATC and 40% have HTC, what are the two probabilities for this population in five years?

In this case, our 1x2 matrix will represent the population, rather than an individual: $\begin{bmatrix} p & q \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$.

To determine the probabilities of this population in five years, we will multiply the matrix $\begin{bmatrix} p & q \end{bmatrix}$ by matrix P:

$$\begin{bmatrix} 0.6 & 0.4 \end{bmatrix} \cdot \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} = \begin{bmatrix} 0.592 & 0.408 \end{bmatrix}.$$

Therefore, based on our model, in five years, 59.2% of the population will have ATC and 40.8% of the population will have HTC.

Q17.   Consider another population where 45% have ATC and 55% have HTC. Predict the cholesterol levels in ten years for this population.

Q18.   Does your answer make sense? Why or why not?

We can perform similar operations for the number of people, rather than the percentages. For example, if in a population of 1500 people, 900 have ATC and 600 have HTC, how many people will have ATC after five years?

$$\begin{bmatrix} 900 & 600 \end{bmatrix} \cdot \begin{bmatrix} 0.88 & 0.12 \\ 0.16 & 0.84 \end{bmatrix} = \begin{bmatrix} 888 & 612 \end{bmatrix}$$

Therefore, according to our model, 888 people will have ATC after five years. (Notice that we still have 888 + 612 = 1500 people in our population.)

Q19.   Consider another population where 800 people have ATC and 800 people have HTC. How many people will have ATC after ten years?

20.1.4 The Effect of Medicine

Now let us look at an issue that is important to doctors. What if the patients take medicine? How will that affect their chances of having ATC or HTC?

People with HTC sometimes take HMG Co-A reductase inhibitors (statins) as part of their treatment plan to reduce their risk of heart attack and stroke. Sometimes, factors, such as diabetes, excess weight, or high blood pressure, may make it necessary for a doctor to prescribe statins to reduce cholesterol. We consider the specific example of a person with type 2 diabetes being prescribed statins (again, a researcher collected data to calculate these probabilities, which are specific to a person with type 2 diabetes taking statins). If a patient begins taking statins at age 60, the probabilities change in our model and we have a different matrix to calculate the probability of patients having ATC and HTC after they take statins. Here is the matrix:

$$P_{statins} = \begin{array}{c} \\ ATC \\ HTC \end{array} \begin{array}{cc} ATC & HTC \\ \begin{bmatrix} 0.92 & 0.08 \\ 0.27 & 0.73 \end{bmatrix} \end{array}.$$

When a patient takes statins, their TC levels should generally decrease. The probabilities in this transition matrix reflect that if a patient on statins has ATC, they will continue to have ATC five years later 92% of the time, or they will have HTC five years later 8% of the time. In addition, there are a greater percentage of patients who start out with HTC and move to ATC five years later than in the original probability matrix. We again notice that each of the rows sums to 1.

Q20.   Calculate the probability of a patient having ATC at age 65 if he starts out at age 40 with HTC and *does not* initiate statins.

Q21.   Calculate the probability of a patient having ATC at age 65 if he starts out at age 40 with HTC and starts statins at age 45.

Q22.   Comparing your answers from problems 20 and 21, is the probability of the patient having ATC at age 65 smaller or larger if the patient starts statins treatment at age 45? Why or why not?

Q23.   Suppose a group of people start statins at age 40. Initially, 450 people have ATC and 500 people have HTC. Predict how many people will have ATC after fifteen years.

Section 20.2: Using Markov Chains to Predict about Medication Adherence

In this section, we introduce the more formal idea of Markov processes. A **Markov process** is one type of stochastic process. A **stochastic process** is defined to be a set of random events in which each event, as it happens in order, is based on random probabilities. **Markov processes** are simply sets of these ordered random events identified by probabilities and that have the quality that each step in the process has probabilities independent of the ordered events before and after.

20.2.1 Adherence Levels

To examine Markov processes, we will consider patients' adherence to the treatment plans prescribed by their doctors. Not all patients take their medicine as often as is prescribed by their physician. Patients have different levels of adherence to their treatment plans for diseases such as diabetes. Medical adherence is the extent to which a patient follows the suggestions of their health care providers. It can greatly influence the effectiveness of medications.

Here are some questions to think about:
Q1. How would you describe medical adherence to a non-health professional?

Q2. Do you think most people take their medicine as prescribed?

Q3. What are some reasons that people might not take their medication?

Q4. What could we do to encourage more people to take their medications?

Let us consider a group of patients that has just initiated statins. We learned that statins are used by patients with high cholesterol to reduce their risk of heart attacks and strokes. We want to analyze their adherence to statins over time. Their adherence can be divided into four different levels:

Adherence State	NON	LOW	MED	HIGH
Percentage of time statins is taken	0-10%	10-40%	40-80%	80-100%

To read this table, we see that MED adherence means that a person takes his or her medications appropriately 40-80% of the time. Make sure you can interpret NON, LOW, and HIGH Adherence as well.

After studying behavior of a large cohort of patients, we determined the percentages for patients' adherence levels during their first year of treatment. A patient starting statins will begin the process according to this initial probability distribution. This matches the table above for the adherence level.

$$q_0 = \begin{bmatrix} P(\text{NON}) & P(\text{LOW}) & P(\text{MED}) & P(\text{HIGH}) \end{bmatrix} = \begin{bmatrix} 0.091 & 0.165 & 0.257 & 0.487 \end{bmatrix}$$

This means that of all of the patients who start taking statins, about 48.7% will have high adherence during the first year, 25.7% will have medium adherence during the first year, 16.5% will have low adherence during the first year, and 9.1% will be nonadherent during the first year.

A researcher (one from first section) also calculated the probabilities for patients moving between different adherence states from year to year. We use the following *yearly* transition matrix Q to determine the probability a patient is in one of the given adherence states after a certain number of years.

$$Q = \begin{bmatrix} P(\text{NON} \to \text{NON}) & P(\text{NON} \to \text{LOW}) & P(\text{NON} \to \text{MED}) & P(\text{NON} \to \text{HIGH}) \\ P(\text{LOW} \to \text{NON}) & P(\text{LOW} \to \text{LOW}) & P(\text{LOW} \to \text{MED}) & P(\text{LOW} \to \text{HIGH}) \\ P(\text{MED} \to \text{NON}) & P(\text{MED} \to \text{LOW}) & P(\text{MED} \to \text{MED}) & P(\text{MED} \to \text{HIGH}) \\ P(\text{HIGH} \to \text{NON}) & P(\text{HIGH} \to \text{LOW}) & P(\text{HIGH} \to \text{MED}) & P(\text{HIGH} \to \text{HIGH}) \end{bmatrix}$$

$$= \begin{matrix} & \text{NON} & \text{LOW} & \text{MED} & \text{HIGH} \\ \text{NON} \\ \text{LOW} \\ \text{MED} \\ \text{HIGH} \end{matrix} \begin{bmatrix} 0.787 & 0.106 & 0.082 & 0.025 \\ 0.498 & 0.205 & 0.213 & 0.084 \\ 0.199 & 0.154 & 0.390 & 0.257 \\ 0.028 & 0.046 & 0.189 & 0.737 \end{bmatrix}$$

One way that might help us see this a little better is to do another diagram like was done in the first example. In this case, we need four circles and will have a much more complicated diagram.
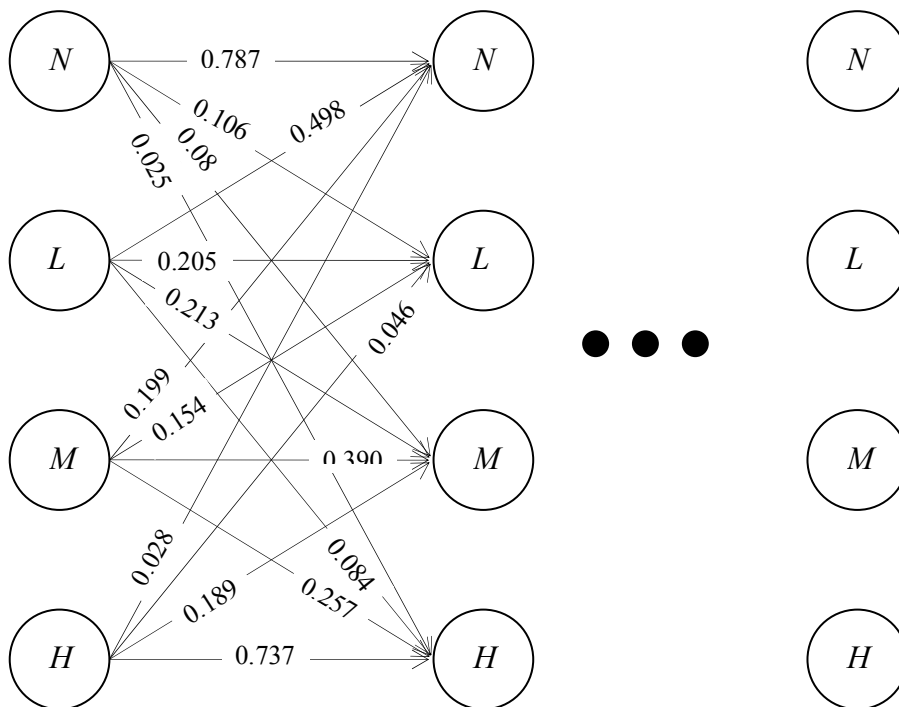

Figure 20.2.1: Diagram illustrating the Markov process model to examine the progression of a patient's adherence level to statins over time, in one-year intervals

For example, we see the probability of going from high to low adherence from one year to the next is 4.6%, and the probability of continuing to have high adherence from one year to the next is 73.7%.

Q5. What is the probability that after a year, a patient will move from medium to low adherence?

Q6. What is the probability that after a year, a patient will move from Low to High Adherence in a year?

Q7.   Patients initiate statins according to the initial probability distribution q_0, meaning that when they are first introduced to statins, they will have the probabilities given in q_0 for adherence. Find the probabilities of four different adherence states 3 years after their initial year of statin treatment (i.e., find q_3).

Q8.   Using your answer to 7, is the percentage of patients that is nonadherent in 3 years after their initial year of statin treatment greater or less than the percentage that is nonadherent in the first year of statin initiation? What about the other adherence states?

Q9.   Let us assume that we know with certainty that a given patient will have high adherence to statins in the first year.

Q10.   Write the patient's initial probability distribution.

Q11.   What is the probability she will be highly adherent in the second year as well (i.e., what is q_1)?

Q12.   What is the probability she will have high adherence in the third year after starting the statin treatment (i.e., what is q_2)?

Q13.   What do you notice about the percentages in the matrix q_2 compared to the percentages in q_1?

Q14.   What is the probability she will have high adherence through the third year and then fall to medium adherence in the fourth year? (Hint: use Figure 20.2.1 and extend it forward)

Q15.   Consider a population of 1,000 people, where 400 people have high adherence, 200 people have medium adherence, 250 people have low adherence, and 150 people are non-adherent. Predict the adherence levels for this group five years after starting the statin treatment.

Q16.   Consider a population of 1,000 people, where 250 people have high adherence, 250 people have medium adherence, 250 people have low adherence, and 250 people are non-adherent. Predict the adherence levels for this group 15 years after starting the statin treatment. Analyze your results.

20.2.2 Steady State Matrix

Now we would like to determine the long term adherence behavior of a large group of patients. We will use the transition matrix and see what happens to the probabilities in the matrix over a period of time.

Q17.   First calculate Q^2.

Q18.   Now calculate Q^4.

Q19.   Finally, calculate Q^{32}. Approximately what percentage of patients will have medium adherence in the long term?

Did you notice something unusual about the Q^{32} matrix? Just to be sure, use your calculator to find Q^{50}. What has happened? All four of the rows of the matrix are identical and not changing. When this happens, the final matrix is called the **steady state matrix**. It is important that we know that often (but not always) transition matrices when raised to high enough powers become steady. The steady state matrix in this example is

$$
Q^{32} = \begin{array}{c} \\ \text{NON} \\ \text{LOW} \\ \text{MED} \\ \text{HIGH} \end{array}
\begin{array}{cccc}
\text{NON} & \text{LOW} & \text{MED} & \text{HIGH} \\
\left[\begin{array}{cccc}
0.458 & 0.110 & 0.179 & 0.253 \\
0.458 & 0.110 & 0.179 & 0.253 \\
0.458 & 0.110 & 0.179 & 0.253 \\
0.458 & 0.110 & 0.179 & 0.253
\end{array}\right]
\end{array}
$$

Q20.   What percentage of patients will be nonadherent in the long term? Describe what this means.

Q21.   Now calculate Q^{100}.

Q22.   What will happen in general if we try to calculate Q raised to powers greater than 32? What does this mean?

One of the really interesting ideas here is that the final probabilities of the four states do not depend on the patient's initial state. To see this, choose two different vectors that might represent the beginning percentages of adherent patients. Then multiply each of them by Q^{32}. You will get the same answer with either beginning patient vector!! This is a very powerful notion for these Markov process steady state systems.

Section 20.3: Video Game

Casey just purchase a video game, *Markov Unchained*. The main character in the game is Markov, and he must beat all the levels of the game to "unchanged" the world. To do so, Markov needs to collect gold coins, and once he has collected enough gold coins he beats one level and goes to the next level of the game. On each turn, a player can choose whether Markov is to enter a contest or to solve a puzzle. To go to the next level of the game, the player has to collect 6 or more gold coins. The game is over if the player loses all the gold golden coins. There are no "lives" in this game, so once the gold coins are gone, it is game over.

If you choose to have Markov play the contest, he can either win or lose up to 3 golden coins on his turn. The contest choice for Markov is called "high risk and high gain" as you can lose a lot, but you can also win a lot when you play it.

x (Number of coins won in contest)	$P(x)$
-3	0.1
-2	0.15
-1	0.25
0	0
1	0.25
2	0.15
3	0.1

Fortunately, we can use the Markov chain technique to investigate this problem further. Here is a general transition matrix that we will develop for this problem of what might happen in a turn for Markov Unchained.

$$Q = \begin{bmatrix} P(0 \text{ to } 0) & P(0 \text{ to } 1) & P(0 \text{ to } 2) & P(0 \text{ to } 3) & P(0 \text{ to } 4) & P(0 \text{ to } 5) & P(0 \text{ to } 6) \\ P(1 \text{ to } 0) & P(1 \text{ to } 1) & P(1 \text{ to } 2) & P(1 \text{ to } 3) & P(1 \text{ to } 4) & P(1 \text{ to } 5) & P(1 \text{ to } 6) \\ P(2 \text{ to } 0) & P(2 \text{ to } 1) & P(2 \text{ to } 2) & P(2 \text{ to } 3) & P(2 \text{ to } 4) & P(2 \text{ to } 5) & P(2 \text{ to } 6) \\ P(3 \text{ to } 0) & P(3 \text{ to } 1) & P(3 \text{ to } 2) & P(3 \text{ to } 3) & P(3 \text{ to } 4) & P(3 \text{ to } 5) & P(3 \text{ to } 6) \\ P(4 \text{ to } 0) & P(4 \text{ to } 1) & P(4 \text{ to } 2) & P(4 \text{ to } 3) & P(4 \text{ to } 4) & P(4 \text{ to } 5) & P(4 \text{ to } 6) \\ P(5 \text{ to } 0) & P(5 \text{ to } 1) & P(5 \text{ to } 2) & P(5 \text{ to } 3) & P(5 \text{ to } 4) & P(5 \text{ to } 5) & P(5 \text{ to } 6) \\ P(6 \text{ to } 0) & P(6 \text{ to } 1) & P(6 \text{ to } 2) & P(6 \text{ to } 3) & P(6 \text{ to } 4) & P(6 \text{ to } 5) & P(6 \text{ to } 6) \end{bmatrix}$$

Just to review, answer the following questions about the game:

Q1.   What happens when Markov has zero coins?

Q2.   What happens when Markov has six or more coins?

So, as soon as a place has zero gold coins, the game is over, the player can never leave this state (he won't be given more coins. As soon as the player gets at least 6 gold coins the player goes to the next level and the game is reset. Therefore, the states 0 and 6 are called **absorbing states**, because it is impossible to leave these states once they are reached, the $P(0,0)$ is 1 and the $P(6,6)$ is also 1. Because it is possible to get to an absorbing state from every state (not necessarily in one step) this is called an absorbing Markov chain. For example, if a player started with one gold coin, they could get to either absorbing state 0 or

absorbing state 6. To get to state 6 from state 1, would take at least 2 turns, but it is still possible. In an absorbing Markov chain, a state that is not absorbing is called a **transient state**.

In the following table, we see the probabilities that you would win or lose to a certain number of coins on one turn.

Number of coins at end of turn / Number of gold coins at beginning of turn	≤ 0 *lose*	1	2	3	4	5	≥ 6 *Go to next level*
≤ 0 *Already lost*	1	0	0	0	0	0	0
1	0.5	0	0.25	0.15	0.10	0	0
2	0.25	0.25	0	0.25	0.15	0.10	0
3	0.10	0.15	0.25	0	0.25	0.15	0.10
4	0	0.10	0.15	0.25	0	0.25	0.25
5	0	0	0.10	0.15	0.25	0	0.50
≥ 6 *At next level*	0	0	0	0	0	0	1

From this chart, we fill in the Q transition Matrix as shown above with the actual probabilities:

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ .5 & 0 & 0.25 & 0.15 & 0.10 & 0 & 0 \\ .25 & 0.25 & 0 & 0.25 & 0.15 & 0.10 & 0 \\ .10 & 0.15 & 0.25 & 0 & 0.25 & 0.15 & 0.10 \\ 0 & 0.10 & 0.15 & 0.25 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0.10 & 0.15 & 0.25 & 0 & 0.50 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As an example to help you understand this transition matrix, we do the following: In one turn, the probability of starting with 2 coins and ending up with ≤ 0 coins is 0.25. You find this by looking at the intersection of the third row and the second column. The reason for that value is that the probability of losing either 2 or 3 coins (See Table 1) is 0.1 and 0.15 which sums to 0.25.

Q3.   Why are 0 and 6 absorbing states? Explain in terms of the game.

Q4.   Why is the $P(1,1) = 0.5$?

Q5.   Why is the $P(5,1) = 0$?

Q6.   Choose any other non-zero entry in the matrix and explain why it is that value.

Q7.   Using the calculator, find the matrix that will show the probabilities of changes in coin numbers after 4 turns. (Hint: $Q^{4)}$

Q8.   If the game starts with Markov having 3 golden coins *this is also randomly determined at the beginning of the level*, what is the probability he will proceed to the next level after 8 turns? 16 turns? 64 turns? (Again, use your calculator).

Q9.   If the game starts with Markov having only 1 gold coin, what is the probability he will have lost after 8 turns? 16 turns? 64 turns?

A player can also choose for Markov to solve a puzzle. The puzzle is either solved or not solved in a turn and Markov gains one gold coin if he solves the puzzle and loses one gold coin if he does not solve the puzzle. The puzzle choice for Markov is called "low risk and low gain" as you can lose little, but you can only win a little when you play it.

The probabilities for winning and losing a coin are in the table below.

x (Number of coins won in contest)	$P(x)$
-1	.45
0	0
1	0.55

Notice that each puzzle, you can only win or lose 1 coin. There will be many zeros in the table below.

Q10.   Fill in the transition table for the puzzle

Number of gold coins	≤ 0 Lose	1	2	3	4	5	≥ 6 Next level
≤ 0 lose							
1							
2							
3							
4							
5							
≥ 6 Next level							

Q11.   Then write the transition matrix from the table you made:

R=

$$\begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{pmatrix}$$

Q12.   If you have Markov play the puzzle choice only, and you start with 2 coins, what is your probability of winning after 8 turns?

Q13.   How many puzzles would Markov have to play if he starts with 4 coins to have a 90% chance of winning?

Q14.   What do you think is the better strategy: always chose the contest or the always chose the puzzle? Does it matter how many turns you take? Justify your answer.

Q15.   What is the better strategy if you are only going to play for 4 turns? What about for 16 turns?

Q16.   What is the probability of NOT losing (staying in the game) for each of the strategies after 4 turns?

Q17.   Do these matrices reach a steady-state? What does that mean in terms of the game?

Q18.   What is your probability of winning if you start with 3 coins, and alternate puzzle, contest, puzzle, contest?

20.3.1 Markov Chain Improvement

Have you ever played a video game against someone that plays all the time while you have never played before? Have you ever noticed that you improve over time (We hope!) when you play video games? There is a learning curve to a new game, meaning that after playing a game for a while you learn some new tricks or you get better at the controller so you can quickly go through the first several levels. In Markov Unchained, let's say that every time you play the game for 4 turns, you see an improvement in just the contest probabilities for winning gold coins based on the following table:

These are the original probabilities

This is the change every four turns

Sum the two columns to left

x (number of gold coins)	Initial P1(x) Turns(1,2,3,4)	Increase in Probability after 4 turns	P2(x) after 4 turns During turns (5,6,7,8)	P3(x) after 8 turns During turns (9, 10, 11,12)
-3	0.1	-0.01	0.09	
-2	0.15	-0.05	0.10	
-1	0.25	-0.1	0.15	
0	0	0	0	0
1	0.25	0.1	0.35	
2	0.15	0.15	0.20	
3	0.1	0.1	0.11	

Q19.   Fill in the rest of the table.

Q20.   What is the probability of losing 2 coins on a turn after 10 turns? Does this make sense?

Q21.   What happens to the probability of winning a level over time? Why?

Q22.   Why would this chart not make sense if you go on for many turns?

Q23.   Create the transition matrix for winning and losing coins during turns 5 to 8. Call it Q2. Then multiply $Q^4 * Q2^4$. What does this matrix tell you specifically?

Q24.   Do you think that you really can get better that fast? If not, estimate how many turns from your experience that you might need to see an improvement.

# Section 20.4: Chapter 20 (Markov Chains) Homework Question

1. Suppose a mouse starts in room 1 and moves from room to room by choosing from the available doors with equal chance (see figure below). Note: the mouse must move; that is, it cannot stay in the same place from one transition to the next.
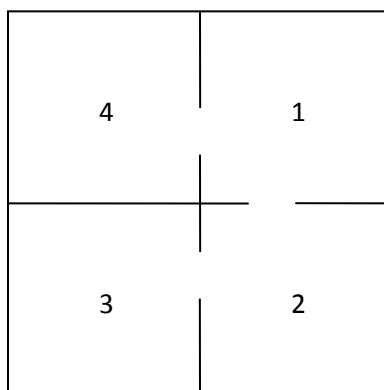
   a. Determine the probabilities of the mouse being in room 1, 2, 3, or 4 next.
   b. Suppose the mouse starts in room 3 instead. Determine the probabilities of the mouse being in room 1, 2, 3, or 4 next.
   c. Continuing in this way, find the transition matrix for this problem.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{cccc} 1 & 2 & 3 & 4 \\ \left[\begin{array}{cccc} \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \\ \_ & \_ & \_ & \_ \end{array}\right] \end{array}$$

   d. Determine whether or not this matrix reaches a steady state. If it does, interpret the meaning of the steady-state matrix in terms of the problem. If not, what are some patterns you notice over time?
   e. Suppose the available doors change so that there are now two ways to go between rooms 3 and 4 (see figure below) and it is equally likely that the mouse will choose either of these doors. Complete (a) through (d) for this new problem.
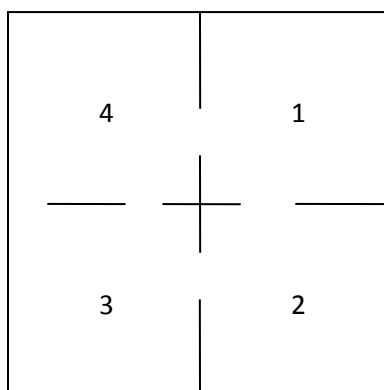
2. Consumers are often loyal to a particular brand. For example, of frozen orange juice consumers, 25 percent are loyal to a leading brand-name orange juice while 14 people are loyal to a private label orange juice. In other words, if a person purchases the leading brand-name frozen orange juice, there is a 25% chance that their next purchase will be the same brand. If a person purchased the private label frozen orange juice, there is a 14% chance that their next purchase will also be the private label orange juice. For the purposes of this problem, we will assume that the only product options for the consumer are the leading brand-name and the private label.
    a. Create a transition matrix to represent the purchases of frozen orange juice over time.
    b. If a person purchases the brand-name frozen orange juice, what is the probability that they will purchase the private label two purchases from now?
    c. If a person purchases the private label frozen orange juice, what is the probability that they will purchase the private label three purchases from now?
    d. Does your transition matrix reach a steady state? If it does, interpret the meaning of the steady-state matrix in terms of the problem. If not, what are some patterns you notice over time?

3. Annabelle heard that Bill was going to ask her to prom. Annabelle told Christy her answer to Bill's request. Christy relays the message to Dahlia, who in turns relays the news to Eugene, and so forth, always to some new person. Suppose there is a probability of 0.1 that a person will change the answer from "yes" to "no" when transmitting it to the next person and a probability of 0.15 that a person will change the answer from "no" to "yes".
    a. Create a diagram to illustrate the progression of Annabelle's response over time.
    b. Create a transition matrix with the probabilities.
    c. Assume that Annabelle says "yes" to Bill's request. What is the probability that the fourth person heard a "no" instead?
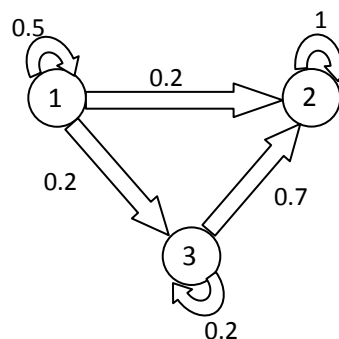
4. Suppose there is a world that is only 2-dimensions. In this world, there are three types of people: squares, rectangles, and triangles. Assume that, of the sons of the squares, 80 percent are squares, 10 percent are rectangles, and 10 percent are triangles. Of the sons of the rectangles, 70 percent are rectangles, 20 percent are squares, and 10 percent are triangles. Of the sons of the triangles, 50 percent are triangles, 25 percent are squares, and 25 percent are rectangles. Assume that every person has at least one son.
    a. Create a transition matrix with the probabilities.
    b. Find the probability that a randomly chose grandson of a square is a triangle.
    c. Determine whether or not this matrix reaches a steady state. If it does, interpret the meaning of the steady-state matrix in terms of the problem. If not, what are some patterns you notice over time?

5. A Markov chain is *absorbing* if and only if the following conditions hold:
    *   one of the probabilities is equal to 1 ($p_{ii} = 1$)
    *   it is possible to go from any other state to the state with probability 1

    For example, the first transition matrix and diagram below represent an absorbing Markov chain and the second matrix and diagram does not.
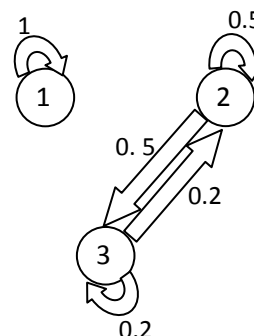
Example 1:

$$\begin{array}{c c c c} & 1 & 2 & 3 \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \left[\begin{array}{ccc} 0.5 & 0.25 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0.75 & 0.25 \end{array}\right] \end{array}$$

Example 2:

$$\begin{array}{c c c c} & 1 & 2 & 3 \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.75 & 0.25 \end{array}\right] \end{array}$$

a. Explain why Example 1 represents an absorbing Markov Chain while Example 2 does not.
b. Create a transition matrix that represents an absorbing Markov Chain and draw the corresponding diagram.

6. At a four-year university, a student has a 15% chance of flunking out during a given year, a 5% chance of having to repeat the year, and an 80% chance of finishing the year. Use the states below.

State	Meaning
1	Freshman
2	Sophomore
3	Junior
4	Senior
5	Has flunked out
6	Has graduated

a. Create a transition matrix for this Markov chain.
b. Find the probability that a freshman will graduate after four years.
c. Find the probability that a freshman will flunk out after four years.
d. Is this Markov chain absorbing? (Use the definition of *absorbing* from question (5).) If so, what does that mean in terms of this problem?

7. Suppose a town has two pizzerias, Carlo's and Dante's. Carlo did some research and determined that there is a probability of *x* that a customer of Carlo's will buy his next pizza at Carlo's and a probability of *y* that a customer of Dante's will switch and buy his next pizza at Carlo's.
a. Create a transition matrix for this Markov chain.
b. Find the probability that a customer of Carlo's will be a customer of Carlo's after two pizza purchases.

c. Find the probability that a customer of Carlo's will be a customer of Dante's after two pizza purchases.
d. Choose probabilities for x and y. Then complete (b) and (c) with your probabilities.

References

Grinstead, C. M., & Snell, J. L. (1997). *Introduction to probability* (2nd ed.). American Mathematical Society.

Huang, Rui; Perloff, Jeffrey M.; and Villas-Boas, Sofia B. (2006) "Effects of Sales on Brand Loyalty," *Journal of Agricultural & Food Industrial Organization*: Vol. 4: Iss. 1, Article 5.

Lial, M. L., Greenwell, R. N., & Ritchey, N. P. (2004). *Finite mathematics* (8th ed.). Addison Wesley.

---

[1] The authors would like to thank Catherine Sitek for her ideas and suggestions for the content in this chapter.

A product of Project MINDSET www.mindsetproject.org

May 2011