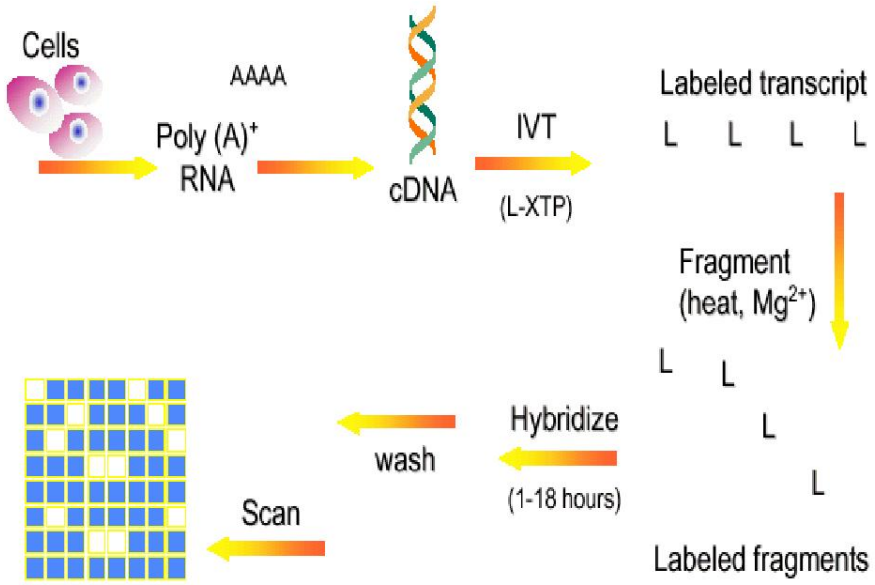
# Statistics for functional bioinformatics - 1

Giorgio Valentini
(Adapted from a lecture by David Gifford)

*valenti@disi.unige.it*

# Starting point

- The experimental setup [affymetrics slide]



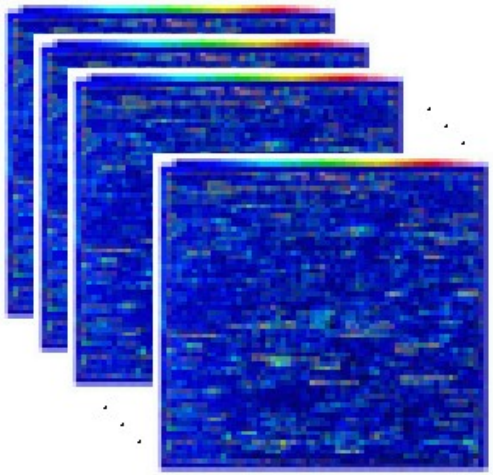

- Variation in the measurements comes from
  - "nuisance" variation in repeated experiments
  - "interesting" variation across different experiments
- Statistical methods are required to characterize either type of variation

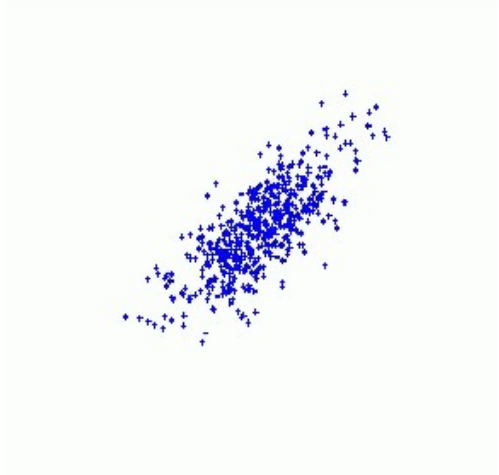
# Topics from statistics

- Elementary concepts, methods
  - population, observation, random variable, random sample
  - statistics, variance, covariance, correlation
  - model, likelihood, likelihood principle, max likelihood
  - exponential family of distributions, examples
  - central limit theorem, implications
  - data transformations
- Measures of confidence
  - confidence intervals
- Significance testing
  - statistical tests, test statistics
  - p-values, power of a test

# Elementary concepts

- Population
  - the set of items we are interested in studying





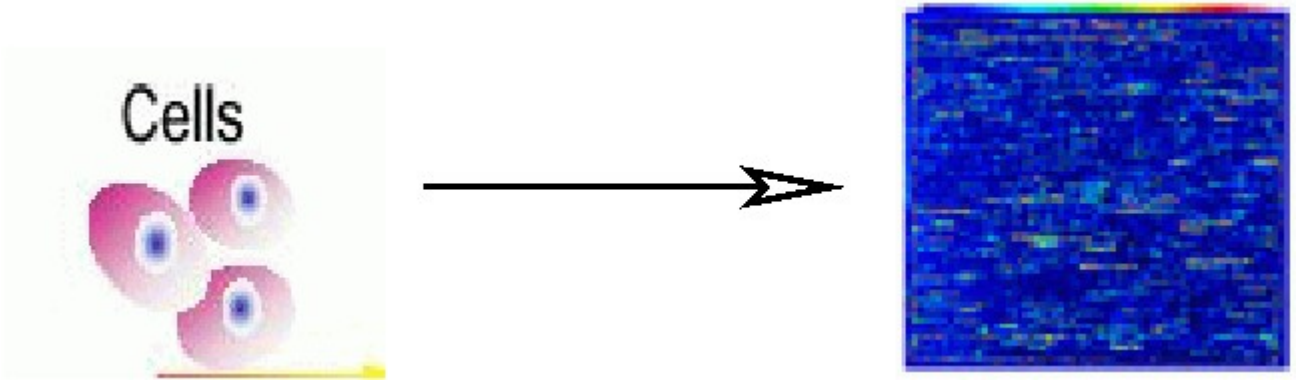
  - (a large number of) repetitions of the same experiment
  - collection of different experiments (nutrient content/type, temperature, cell-cycle)

Elements in the population in these cases correspond to individual experiments

# Elementary concepts

- Observations
  - interpreted, coded

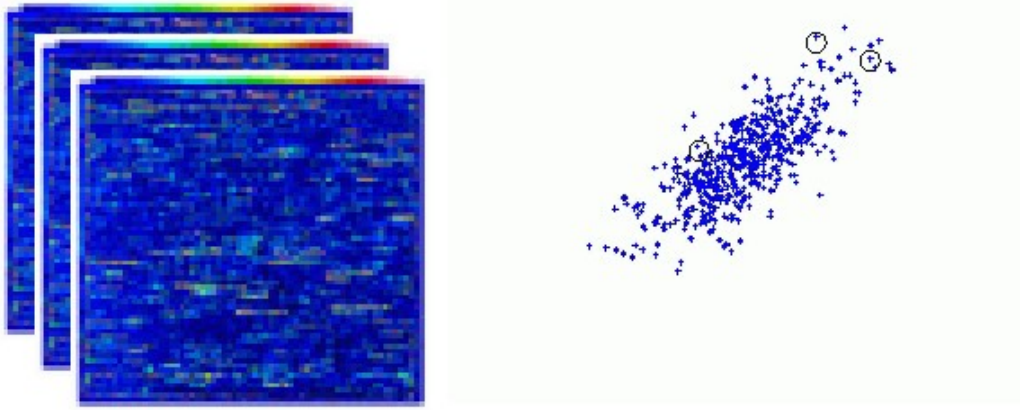
  For example, we almost never directly observe the quantities of interest



# Elementary concepts

- Random sample
  - a set of random draws from the population (with replacement)

  For example, cell cycle measurements at three time points



  Are these ever random draws?

# Elementary concepts

- Random variable
  - a mapping from (experimental) outcomes to numerical values

  Example: $X_1$ is a random variable corresponding to the expression level of gene 1

  $x_1^{(2)}$ is a realization of $X_1$ in experiment 2

| | Experiment 1 | Experiment 2 | ... |
|---|---|---|---|
| Gene 1 | 181 | 1 | 137 |
| Gene 2 | 499 | 229 | 218 |
| Gene 3 | 167 | 147 | 120 |
| ... | 296 | 110 | 380 |

  Note: $P(X_1 = 181)$ is a statement about the population, not about the observed data

# Elementary concepts

- Statistics
  - any function computed from the observed data (random sample)

  For example, mean expression level of gene 1

$$\bar{x}_1 = \frac{1}{n} \sum_{t=1}^{n} x_1^{(t)} \tag{1}$$

  where $x_1^{(t)}$ is the observed value of the random variable $X_1$ in experiment $t$.

# Elementary concepts

- Correlation
  - measures linear relations between variables

Sample correlation between two genes (1 and 2) across $n$ experiments

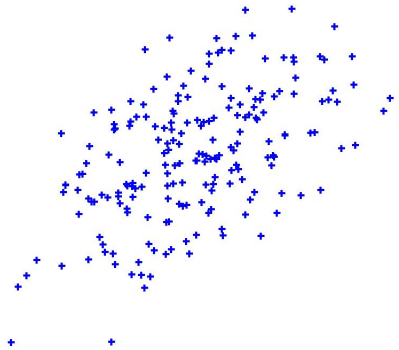
$$\widehat{C}_{12} = \frac{\overbrace{\frac{1}{n}\sum_{t=1}^{n}(x_1^{(t)} - \bar{x}_1)(x_2^{(t)} - \bar{x}_2)}^{\text{Sample covariance } \widehat{\Sigma}_{12}}}{\sqrt{\hat{\sigma}_1^2}\sqrt{\hat{\sigma}_2^2}} \tag{2}$$

where $\hat{\sigma}_i^2$, $i = 1, 2$ are sample variances

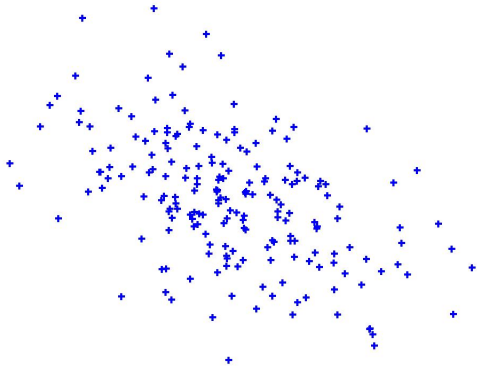
$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n}(x_1^{(t)} - \bar{x}_1)^2 \tag{3}$$

# Elementary concepts

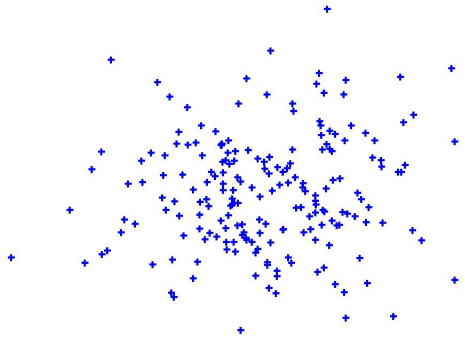
- Scatter plots of (hypothetical) genes



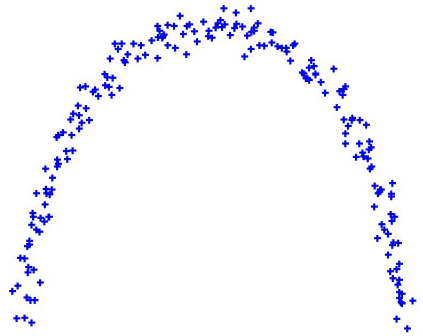
positive correlation



negative correlation



zero correlation



zero correlation

# Statistical models

- Statistical models attempt to characterize the population of interest
- A generative model aims to be able to recreate the observed data (or population of interest)
- A multivariate Gaussian model

$$
\begin{aligned}
Z_i &\sim N(0,1) && (4)\\
X &= AZ + \mu && (5)\\
\Sigma &= E[(X-\mu)(X-\mu)^T] && (6)\\
&= E[(AZ)(AZ)^T] && (7)\\
&= E[AZZ^TA^T] && (8)\\
&= AE[ZZ^T]A^T && (9)\\
&= AA^T && (10)
\end{aligned}
$$

- A multivariate Gaussian model

$$
\begin{aligned}
p(x|\theta) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \} && (11)\\
X &\sim N(\mu,\Sigma) && (12)
\end{aligned}
$$

where $\mu$ is the mean vector and $\Sigma$ is the covariance matrix

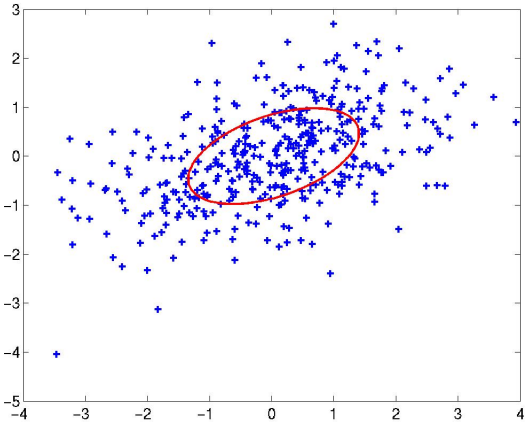
# Statistical models

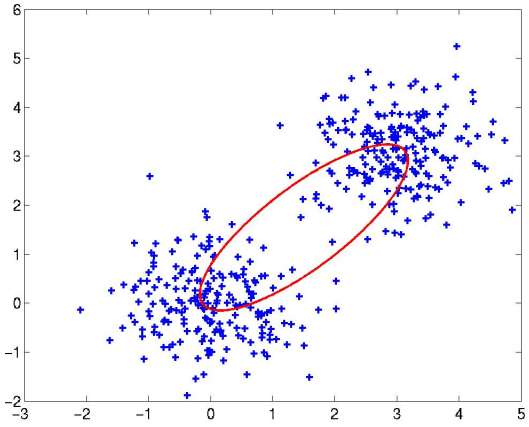
- Statistical models attempt to characterize the population of interest
- A generative model aims to be able to recreate the observed data (or population of interest)
- A multivariate Gaussian model

$$p(x|\theta) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \} \quad (15)$$

$$X \sim N(\mu, \Sigma) \quad (16)$$

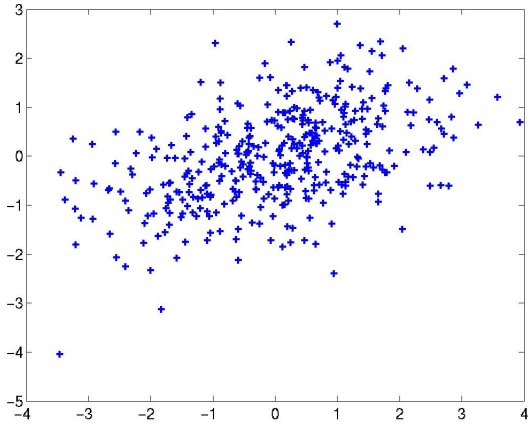
where $\mu$ is the mean vector and $\Sigma$ is the covariance matrix



# Likelihood functions

- Assume we have a probability model $p(x|\theta)$ with parameter $\theta$ ($\theta$ can be a vector of parameters)
- Given observed data $D = \{x^{(1)}, \ldots, x^{(n)}\}$ we wish to find an appropriate setting of the parameters $\theta$ so that the model "best" accounts for the observed data




- A likelihood function is the likelihood of the observed data as a function of $\theta$ (the parameters)

$$L(x^{(1)}, \ldots, x^{(n)}|\theta) = \prod_{t=1}^{n} p(x^{(t)}|\theta) \tag{17}$$

and is sufficient for adjusting the parameters $\theta$.

# Maximum likelihood principle: Binomial

- Maximum likelihood principle: we find the parameter $\widehat{\theta}$ that maximize the likelihood of the observed data

$$\widehat{\theta} = \arg\max_{\theta} L(x^{(1)}, \ldots, x^{(n)}|\theta) \tag{18}$$

The Maximum likelihood estimate (MLE) for the Binomial PMF is

$$P(k_N|\theta) = \binom{N}{k}\theta^k(1-\theta)^{(N-k)} \tag{19}$$

$$logP(k_N|\theta) = log\binom{N}{k} + klog\theta + (N-k)log(1-\theta) \tag{20}$$

$$\frac{\mathrm{d}\, P(k_N|\theta)}{\mathrm{d}\theta} = \frac{k}{\theta} - \frac{N-k}{1-\theta} \tag{21}$$

$$0 = \frac{k}{\theta} - \frac{N-k}{1-\theta} \tag{22}$$

$$\widehat{\theta} = k/N \tag{23}$$

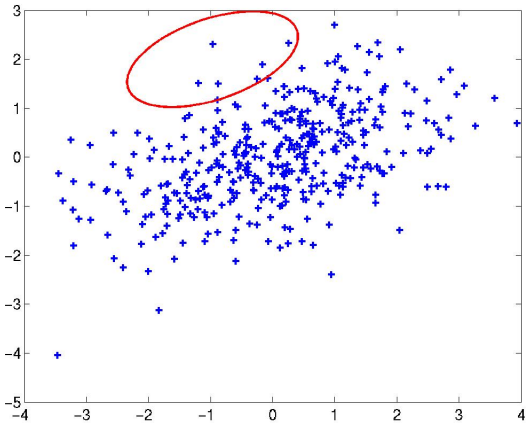
# Maximum likelihood principle: Gaussian

- All the information is in the likelihood function

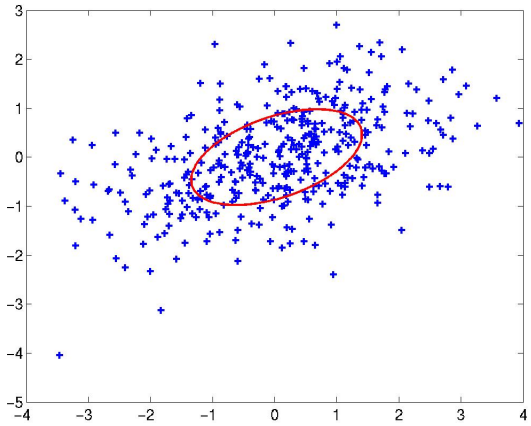
$$L(x^{(1)}, \ldots, x^{(n)}|\theta) = \prod_{t=1}^{n} p(x^{(t)}|\theta) \tag{8}$$

- Maximum likelihood principle: we find the parameters $\hat{\theta}$ (mean and covariance) that maximize the likelihood of the observed data

$$\hat{\theta} = \arg\max_{\theta} L(x^{(1)}, \ldots, x^{(n)}|\theta) \tag{9}$$



bad setting of parameters
(low likelihood)



good setting
(high likelihood)

# Maximum likelihood estimation

- A multivariate Gaussian model

$$p(x|\theta) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \} \quad (10)$$

- Given observed data $D = \{x^{(1)}, \ldots, x^{(n)}\}$, the maximum likelihood estimates of the parameters are:

  1. Sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x^{(t)} \quad (11)$$

  2. Sample covariance

$$\hat{\Sigma}_{ij} = \frac{1}{n}\sum_{t=1}^{n} (x_i^{(t)} - \hat{\mu}_i)(x_j^{(t)} - \hat{\mu}_j) \quad (12)$$

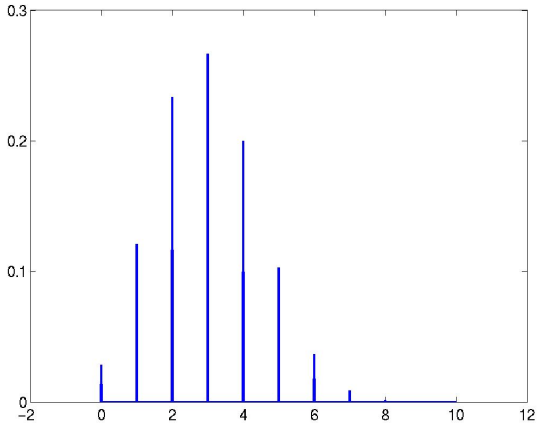
# Exponential family of distributions

- Binomial, multinomial
- Poisson
- Gaussian
- Exponential
- Gamma

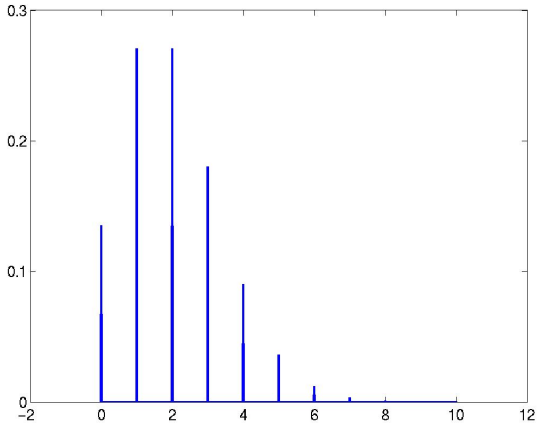
  ...

  - For exponential distributions, sample statistics (mean, variance, covariance) are the maximum likelihood estiates for the model parameters
  - Thus, for all sufficient statistics, simply calculate the statistic from the sample to fit the distribution

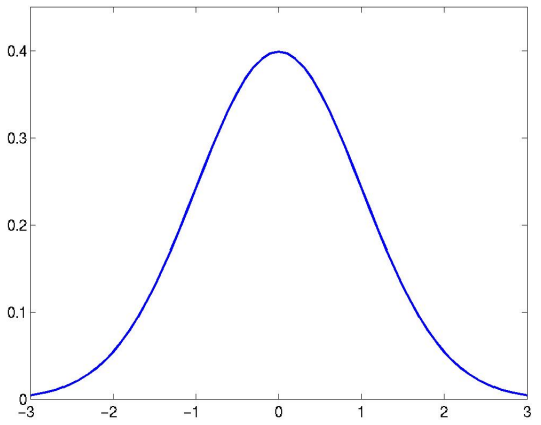
# Exponential family of distributions



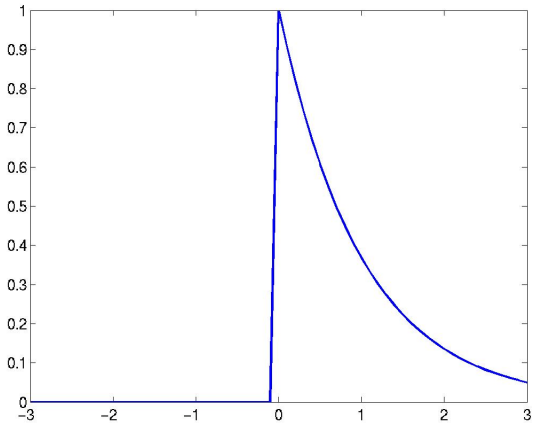

Binomial




Poisson




Gaussian (normal)




Exponential

# Central limit theorem

Let $X^{(1)}, \ldots, X^{(n)}$ be independent (vector valued) random variables corresponding to any distribution with mean $\mu$ and covariance $\Sigma$, then for large $n$,

$$\sqrt{n}(\bar{X} - \mu) \sim N(0, \Sigma) \tag{13}$$

where $\bar{X}$ is the mean

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X^{(t)} \tag{14}$$

# Statistical tests

- Possible things that we might want to test:

  1. whether a gene is cell cycle related
  2. if a gene has a differential response to a pathogen

  etc.

- For the purposes of illustration, we try to test whether the observed correlation between two genes is significant

# Statistical tests

- Testing involves several steps:
  1. Select the hypotheses such as
     $H_0$ two genes are uncorrelated
     $H_1$ they have a non-zero correlation
  2. Choose a test statistic $T(X)$
     – need to define how we will measure differences between the hypothesis
  3. Observe a random sample $D = \{x^{(1)}, \ldots, x^{(n)}\}$
  4. Compute the observed value for the test statistic

$$T_{obs} = T(x^{(1)}, \ldots, x^{(n)}) \tag{18}$$

  5. Compute the significance level (P-value) for rejecting the null hypothesis $H_0$

$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \tag{19}$$

  6. The P-value is the probability we reject $H_0$ when $H_0$ is true

# Statistical tests: example

- Defining the hypothesis:

  Let $X_1$ and $X_2$ are the random variables corresponding to the expression levels of the two genes

  The null hypothesis $H_0$: $X_1$ and $X_2$ are uncorrelated:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right) \tag{21}$$

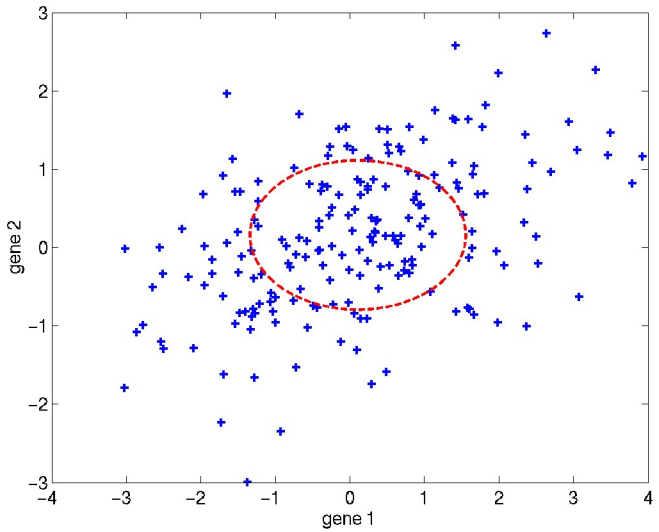
  The alternative hypothesis $H_1$: $X_1$ and $X_2$ can be correlated:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{22}$$

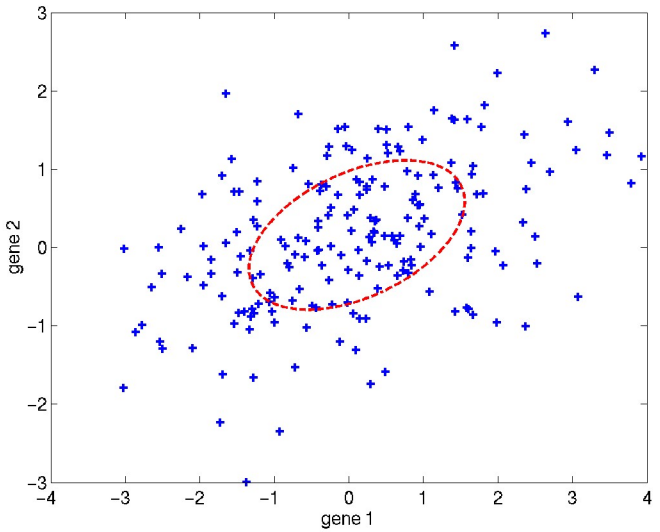
  where $\Sigma_{ij}$ is the covariance between $X_i$ and $X_j$ ($\sigma_i^2 = \Sigma_{ii}$)

# Statistical tests: example

- The alternative hypothesis $H_1$ is more expressive in terms of explaining the observed data




null hypothesis




alternative hypothesis

- We need to find a way of testing whether this difference is significant

# Test statistic

- Likelihood ratio statistic

$$T(X^{(1)}, \ldots, X^{(n)}) = 2 \log \frac{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_1)}{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_0)} \tag{23}$$

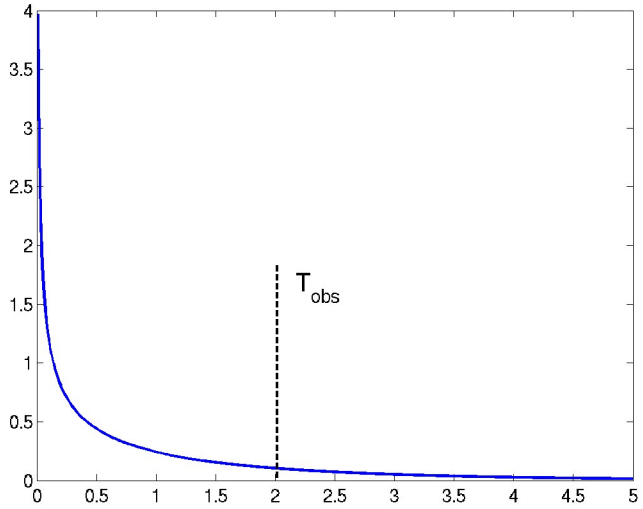
  Larger values of $T$ imply that the model corresponding to the null hypothesis $H_0$ is much less able to account for the observed data

- To evaluate the P-value, we also need to know the sampling distribution for the test statistic

  In other words, we need to know how the test statistic $T(X^{(1)}, \ldots, X^{(n)})$ varies if the null hypothesis $H_0$ is correct

# Test statistic cont'd

- For the likelihood ratio statistic, the sampling distribution is $\chi^2$ with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses




- Once we know the sampling distribution, we can compute the P-value

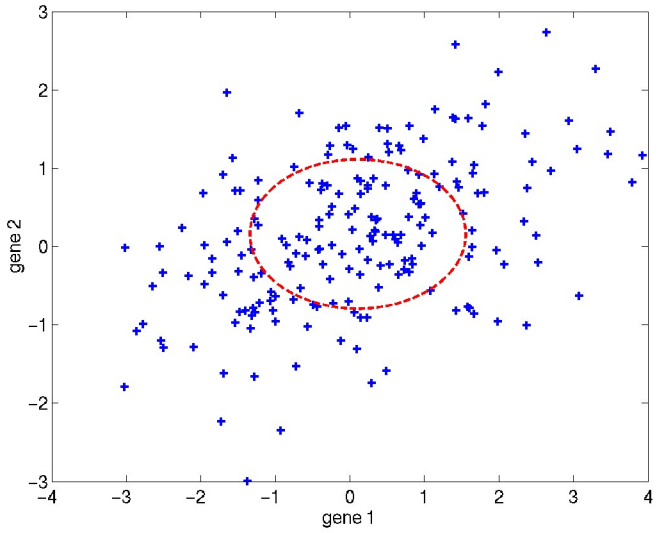
$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \tag{24}$$

# Degrees of freedom

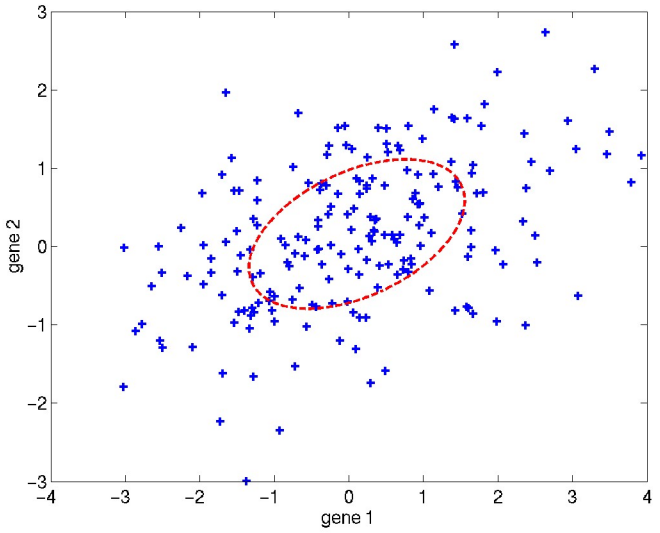
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$




$H_0$




$H_1$

- The observed data overwhelmingly supports $H_1$

# Maximum a Posterior Estimators (MAP)

- Assume that we know something about a coin before we observe $N$ trials
- Prior knowledge can take on many forms
  - Assumptions (mRNA levels are never negative)
  - Data (other experiments suggests that protein A regulates gene B)
  - Estimates (our best estimate of the parameters so far)
- How do we express this knowledge so that it can be used in a principled way?
- Represent this knowledge as a distribution over model parameters
  - In the case of a coin, as a distribution over $\theta$

# Bayes' Rule

- Key to Bayesian analysis is Bayes' Rule

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \tag{31}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{32}$$

# Bayesian Inference

- If we believe that Gene A can be in low, medium, or high state of expression, and it influences Gene B as follows, and the prior on A is as given:
  - $P(B|A_L) = 0.2$ and $P(A_L) = 0.4$
  - $P(B|A_M) = 0.4$ and $P(A_M) = 0.4$
  - $P(B|A_H) = 0.8$ and $P(A_H) = 0.2$
- Given that gene B is turned on, what is the probability that gene A is in the high state?

$$P(A_H|B) = \frac{P(B|A_H)P(A_H)}{P(B)} \tag{33}$$

$$P(A_H|B) = \frac{P(B|A_H)P(A_H)}{P(B|A_L)P(A_L) + P(B|A_M)P(A_M) + P(B|A_H)P(A_H)} \tag{34}$$

$$P(A_H|B) = \frac{0.8 \times 0.2}{0.2 \times 0.4 + 0.4 \times 0.4 + 0.8 \times 0.2} \tag{35}$$

$$P(A_H|B) = 0.4 \tag{36}$$

# Maximum a Posterior Estimators (MAP)

- Bayesians use prior knowledge when analyzing data
  - This can lead to different conclusions from the same data, depending on your prior
- Frequentists believe that conclusions from data should always be the same
- Using Bayes' Rule in our Binomial example:

$$P(\theta|k_N) = \frac{P(k_N|\theta)P(\theta)}{P(k_N)} \tag{37}$$

- Let's represent $P(\theta)$ as:

$$P(\theta) = C(\alpha)\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \tag{38}$$

$$\alpha_1 = pS+1 \tag{39}$$

$$\alpha_2 = (1-p)S+1 \tag{40}$$

# Dirichlet Distributions

- $P(\theta)$ is a Dirichlet distribution, and is a conjugate distribution to the Binomial distribution:

$$
\begin{aligned}
P(\theta) &= C(\alpha)\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} && (41)\\
\alpha_1 &= pS+1 && (42)\\
\alpha_2 &= (1-p)S+1 && (43)
\end{aligned}
$$

- This binomial form of the Dirichlet distribution is called the Beta distribution.
- Now:

$$
\begin{aligned}
P(\theta|k_N) &= \frac{\binom{N}{k}C(\alpha)\theta^{k+pS}(1-\theta)^{(N-k)+(1-p)S}}{P(k_N)} && (44)\\
\frac{\mathrm{d}\,P(\theta|k_N)}{\mathrm{d}\theta} &= \frac{k+pS}{\theta}-\frac{(N-k)+(1-p)S}{1-\theta} && (45)\\
\widehat{\theta_{MAP}} &= \frac{k+pS}{N+S} && (46)
\end{aligned}
$$