

# Modelli descrittivi, statistica e simulazione

Master per Smart Logistics specialist

Roberto Cordone  
([roberto.cordone@unimi.it](mailto:roberto.cordone@unimi.it))

La **statistica descrittiva** si occupa delle situazioni in cui

- si compiono delle analisi (qualitative o quantitative);
- il risultato dell'analisi non è sempre lo stesso.

Esempi: alle casse di un supermercato

- i clienti pagano in contanti, con bancomat o con carta?
- i clienti hanno il talloncino per il parcheggio?
- quanto impiega la cassiera a smaltire un cliente?
- ogni quanto tempo arriva un cliente alla cassa?
- ...

Lo scopo della statistica descrittiva è

- organizzare, sintetizzare e presentare i risultati delle analisi
- in modo da facilitarne l'elaborazione per prendere decisioni utili

**Popolazione** è l'insieme degli elementi che sono oggetto dell'analisi

*Per es., i clienti*

**Unità** è il singolo elemento che compone la popolazione

*Per es., il singolo cliente*

**Carattere** è il fenomeno analizzato

*Per es., il modo di pagamento, il possesso del talloncino, il tempo di servizio o la frequenza di arrivo alla cassa*

In generale, ogni unità ha diversi caratteri

Le popolazioni possono essere

- **reali**, cioè **analizzate completamente** unità per unità  
*(i clienti di giovedì 21 dalle 9.00 alle 13.00)*
- **virtuali**, cioè **definite rigorosamente, ma non analizzate per intero**  
*(i clienti del 2016)*

La definizione di una popolazione fa sempre riferimento a un preciso spazio (posizione o regione) e tempo (istante o intervallo)

Un carattere si può esprimere in diverse **modalità**,  
che sono i **valori specifici assunti in ciascuna unità**

Un carattere si classifica in base al tipo di modalità che ammette:

- **binario** (o **dicotomico**), quando è **presente in alcune unità e non in altre**  
(Per es., il possesso del talloncino per il parcheggio)
- **qualitativo** (o **categoriale**) quando è **espresso a parole**  
(Per es., il mezzo di pagamento)
- **quantitativo** (o **numerico**) quando è **espresso da un numero**  
(Per es., il tempo di servizio o la frequenza di arrivo)

Un **carattere qualitativo** si definisce anche **mutabile**

Una mutabile può essere

- **sconnessa** (o **nominale**) quando  
non esiste nessun ordinamento naturale fra le sue possibili modalità  
(Per es., il modo di pagamento)
- **ordinale** (o **ordinata**, o **ordinabile**) quando  
esiste un ordinamento naturale fra le sue possibili modalità
  - **lineare**  
(Per es., la qualità del servizio: sufficiente, discreto, buono, ottimo)
  - **ciclico**  
(Per es., il giorno della settimana dedicato alla spesa: lunedì, . . .)

Un **carattere quantitativo** si definisce anche **variabile**

Una variabile può essere

- **discreta** (o **intera**, o **enumerabile**) quando i suoi possibili valori possono essere elencati, in quanto costituiscono un insieme finito o infinito numerabile  
(*Per es., il numero di articoli acquistati, la lunghezza della coda*)
- **continua** (o **reale**, o **misurabile**) quando può assumere un insieme continuo di valori, tipicamente un intervallo  
(*Per es., il tempo di servizio, la frequenza di arrivo alla cassa*)

La distinzione tra discreto e continuo può dipendere dalla precisione degli strumenti di misura (ad es., l'altezza in cm)

I **dati** sono l'**insieme delle misure fatte sulla popolazione**

Se le misure riguardano solo un sottoinsieme di unità, tale sottoinsieme si indica con il termine **campione**

Per esempio, uno studio medico può avere

- come popolazione l'insieme di tutti i pazienti, presenti e futuri
- come unità i singoli pazienti
- come caratteri le caratteristiche rilevate sui pazienti
- come modalità i valori rilevati
- come campione i pazienti oggetto della rilevazione

La statistica descrittiva

- rileva alcuni caratteri di alcune o tutte le unità della popolazione
- ne ricava informazioni sulla popolazione
- allo scopo di dedurre altre informazioni utili o di prendere decisioni

*Per cominciare, supporremo che le misure riguardino tutta la popolazione*

**Distribuzione unitaria** è l'insieme dettagliato di tutte le rilevazioni fatte

È una lista di  $(n + 1)$ -uple composte da:

- unità
- carattere 1
- ...
- carattere  $n$

Questa descrizione è già sufficientemente significativa quando

- le unità non sono troppe
- esiste una rappresentazione efficace

Casi tipici sono

- **serie territoriali**: l'unità è un punto o una regione
- **serie temporali** o **storiche**: l'unità è un istante o intervallo di tempo



# Rappresentazioni grafiche

Per le serie territoriali, la rappresentazione più comune è il **cartogramma** o **mappa tematica**



Per le serie temporali, la rappresentazione più comune è il **grafico a linee**



Vedi Esercizio 1-1

# Distribuzioni di frequenza

Di solito, però i dati sono troppo dettagliati per essere comprensibili

Occorre un procedimento di **aggregazione** basato sulla frequenza delle diverse modalità di un carattere nella popolazione

- **frequenza assoluta**  $n_i$  è il **numero di unità della popolazione in cui il carattere  $x$  assume la modalità  $x_i$**  per ogni  $i \in M$
- **frequenza relativa**

$$f_i = \frac{n_i}{\sum_{i \in M} n_i} \text{ per ogni } i \in M$$

dove  $\sum_{i \in M} n_i = N$  è il **numero totale di unità nella popolazione**

- **frequenza percentuale**

$$p_i = 100 \cdot f_i \text{ per ogni } i \in M$$

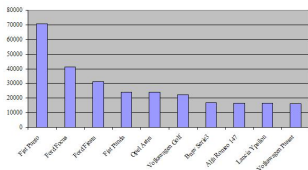
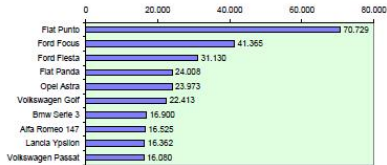
Ovviamente valgono le seguenti proprietà

$$\sum_{i \in M} f_i = 1 \quad \sum_{i \in M} p_i = 100 \text{ per ogni } i \in M$$

# Rappresentazioni grafiche delle distribuzioni

Anziché le singole rilevazioni, si rappresentano le frequenze delle rilevazioni che presentano la stessa modalità

Il **grafo a barre** e l'**istogramma** sono le rappresentazioni più comuni per la **frequenza assoluta**



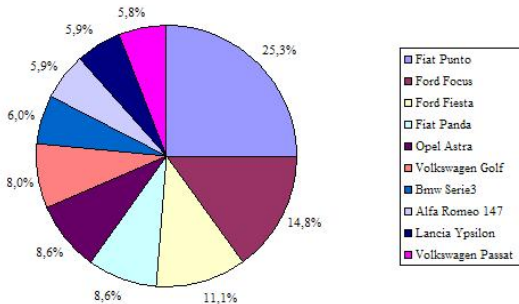
Se sono caratteri nominali, si introduce un ordinamento arbitrario, spesso dettato dai dati (*per sottolineare quali sono più frequenti*)

*Vedi Esercizio 1-2*

Per i caratteri continui, occorre una divisione in **classi** o **intervalli**

*Ne riparlamo fra poco*

Il **diagramma a torta** si usa invece per la **frequenza relativa** e **percentuale**



*Vedi Esercizio 1-3*

# Distribuzioni di frequenza cumulata

I caratteri ordinali e quelli quantitativi hanno anche le frequenze cumulate

- **frequenza assoluta cumulata** è il numero di unità della popolazione in cui il carattere  $x$  assume modalità  $\leq x_i$

$$N_i = \sum_{j: x_j \leq x_i} n_j$$

- **frequenza relativa cumulata**

$$F_i = \sum_{j: x_j \leq x_i} f_j = \frac{N_i}{N}$$

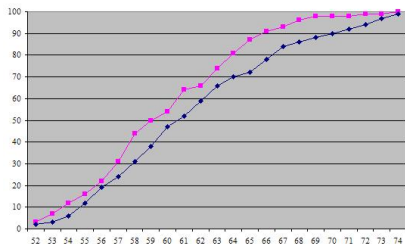
dove  $\sum_i n_i = N$  è il numero totale di unità nella popolazione

- **frequenza percentuale cumulata**

$$P_i = \sum_{j: x_j \leq x_i} p_j = 100F_i$$

Sono tutte funzioni crescenti da 0 a, rispettivamente  $N$ , 1 e 100

La rappresentazione tipica per le distribuzioni cumulate è il *grafico a linee*



Consente di ricavare informazioni dal confronto fra popolazioni diverse

- se il grafico della prima popolazione sta tutto a sinistra di quello della seconda (cioè sotto), la prima popolazione ha valori più bassi del carattere rilevato  
(*se sono costi o consumi, è meglio; se guadagni o qualità è peggio*)

# Definizione delle classi per caratteri continui

I diagrammi delle distribuzioni richiedono un numero finito di casi

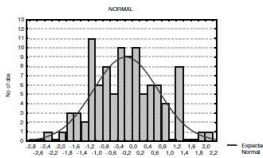
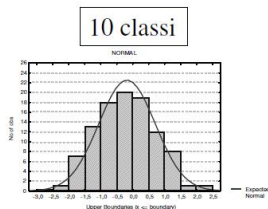
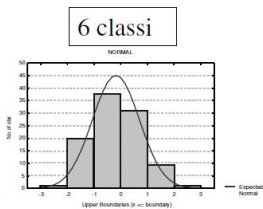
Per i caratteri continui, occorre una divisione in intervalli che rispetti le seguenti condizioni (via via meno importanti)

- 1 **mutuamente esclusivi**, per evitare che alcune unità contino più volte;
- 2 **complessivamente inclusivi**, per evitare che alcune unità non contino;
- 3 **di ampiezza costante** se possibile (*a volte non lo è*);
- 4 **senza classi vuote** se possibile (*senza cancellarle, se ci sono*);
- 5 **senza classi aperte**: si definiscano **un minimo e un massimo corretti**;
- 6 **con estremi interi e di ampiezza dispari** per avere un centro intero.

# Definizione delle classi per caratteri continui

Un numero errato di classi rende i dati illeggibili

- troppo poche classi cancellano le oscillazioni di frequenza;
- troppe classi producono oscillazioni spurie di frequenza.



Una buona regola euristica è

$$1 + \log_2 N \leq C \leq \sqrt{N}$$

Vedi Esercizio 1-4 ed Esercizio 1-5