# Statistica applicata alle biotecnologie

- CdL Specialistica in Biotecnologie Alimentari e Vegetali.
- CdL Specialistica in Biotecnologie Industriali ed Ambientali.
- CdL Specialistica in Biotecnologie Mediche e Medicina Molecolare.

Dr. Nicola Ancona

www.ba.cnr.it/~iesina18

ancona@ba.issia.cnr.it

# Obiettivi del corso

1) Insegnare ad organizzare e sintetizzare i dati.

2) Insegnare a prendere decisioni su un gran numero di dati esaminando solo una piccola parte di essi.

Il primo obiettivo si raggiunge con i metodi ed i concetti della **statistica descrittiva**.

Il secondo con i metodi della **statistica inferenziale**.

# La statistica

La statistica è una disciplina scientifica concernente:

1) la raccolta, l'organizzazione, la sintesi e l'analisi dei dati;

2) l'inferenza su un corpo di dati, quando soltanto una parte di essi è osservata.

# Testi consigliati

- Wayne W. Daniel, Biostatistica: concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria, EdiSES s.r.l. – Napoli.
- Stanton A. Glantz, Statistica per discipline biomediche, 5° Edizione, Mc Graw Hill.
- M. Pagano, K. Gauvreau, Biostatistica, II Edizione, Idelson Gnocchi.

# La statistica descrittiva

# INDICE

- ✓Definizioni
- ✓Approccio statistico
- ✓Frequenze
- ✓Rappresentazioni grafiche
- ✓Sintesi dei dati
- ✓Variabilità dei dati

# Unità statistica

Una unità statistica è un singolo soggetto preso in considerazione nell'analisi.

Es.: se studiamo l'altezza media degli studenti di questa aula, ogni studente è una unità statistica.

# Caratteristica

Una caratteristica, detta anche feature, è un attributo o aspetto di una unità statistica.

Es.: se studiamo l’altezza media degli studenti di questa aula, l’altezza è una caratteristica.

# Variabile aleatoria o casuale

In generale una caratteristica (feature) è rappresentata da una variabile aleatoria che può assumere valori (modalità) diversi su unità statistiche differenti.

Es. pressione diastolica del sangue, la frequenza cardiaca, la statura di maschi adulti, il livello di espressione di un gene in un tessuto,…

# Variabile aleatoria

Indicheremo v.a. con lettere maiuscole $X$, $Y$ e con $x$, $y$ i valori che esse assumono.

Es.: sia $X$ l'altezza degli studenti di quest'aula. Allora $x_1$=1.74, $x_2$=1.65, $x_3$=1.82, $x_4$= 1.78, …

E' importante osservare che, in generale, una v.a. assume valori differenti.

# Variabile aleatoria discreta

Una variabile aleatoria si dice discreta quando assume un numero finito o numerabile di valori.

Es. numero di ricoveri giornalieri in un ospedale, numero di geni sovra-espressi nel colon,…

# Variabile aleatoria continua

Una variabile aleatoria si dice continua quando assume un numero infinito di valori in un determinato intervallo.

Es. statura di un individuo, livello di espressione di un gene in un tessuto,…

# Popolazione

Una popolazione è l'insieme di tutte le unità statistiche di interesse per un particolare studio.

oppure

Una popolazione è l'insieme dei valori che una variabile aleatoria può assumere.

Una popolazione può essere finita o infinita.

# Esempi di popolazioni

- Il peso dei bambini di una scuola elementare.
- Il livello di glicemia nei pazienti diabetici.
- Il livello di espressione del gene P53 in pazienti affetti da cancro al colon.

# Campione

Un campione è un insieme finito di unità statistiche estratte da una popolazione.

oppure

Un campione è un insieme finito di valori ottenuti da una variabile aleatoria.

# Proprietà di un campione

- Un campione è sempre finito.
- Un campione deve essere rappresentativo dell’intera popolazione.

# I dati

Il materiale di base della statistica è

costituito dai dati.

In generale i dati sono numeri che possono essere ottenuti da un processo di misura o di conteggio.

# Le fonti dei dati

- Le rilevazioni periodiche (es. pazienti dimessi al giorno).
- Le indagini (es. tipo di farmaco usato).
- Gli esperimenti (es. tipo di farmaco più efficace).
- Le fonti esterne.

# Approccio statistico

Osservazioni dei fenomeni

Formulazione delle leggi che regolano i fenomeni

Raccolta dei dati

Analisi dei dati ← **Statistica**

# Fonti di errore

- Raccolta di dati non corretta
- Presentazione inadeguata
- Analisi statistica inappropriata

} *impossibile la verifica e il confronto con altre ricerche*

# L'analisi statistica

Disegno sperimentale: è necessario per scegliere e programmare le osservazioni in natura e le ripetizioni in laboratorio, in funzione dell'OBIETTIVO della ricerca.

Campionamento: è necessario per raccogliere i dati; è limitato; DEVE essere rappresentativo della popolazione.

Descrizione dei dati: è necessaria per verificare l'adeguatezza del disegno sperimentale e del campionamento.

Inferenza: è necessaria per estendere le considerazioni dal campione alla popolazione.

Le conclusioni dell'analisi vanno sempre interpretate nell'ambito disciplinare in cui si conduce la ricerca.

# Approcci statistici

## *Statistica descrittiva*

Conoscere e rappresentare in maniera sintetica l'andamento di un fenomeno.

## *Statistica inferenziale*

- avanzare interpretazioni e previsioni sulla variabilità di un determinato fenomeno.
- prendere decisioni su un gran numero di dati esaminando soltanto una piccola parte (es: farmaco).

# Metodo di indagine

***La rilevazione statistica***

Definizione ed esplicitazione degli obiettivi (vincoli operativi e di costo)

Definizione della popolazione obiettivo (definendo in modo univoco le caratteristiche comuni che devono presentare le unità per essere incluse nella popolazione obiettivo e l'intervallo temporale dell'indagine)

Raccolta dei dati

Spoglio dei dati

Analisi/interpretazione dei dati

CHIAREZZA e SEMPLICITA'

## Strumenti di indagine

Questionari, interviste, MISURE

Tabelle (chiare, sintetiche, precise)

Rappresentazioni grafiche

## Strumenti di analisi

Frequenze, medie, indicatori di variabilità, *test statistici*

# Gli oggetti della statistica: DEFINIZIONI

***Unità statistiche***

Elementi o casi componenti il fenomeno collettivo e costituiscono l'oggetto diretto dell'osservazione

***Popolazione (o collettivo di studio)***

Insieme di tutte le unità statistiche accomunate da una o più caratteristiche

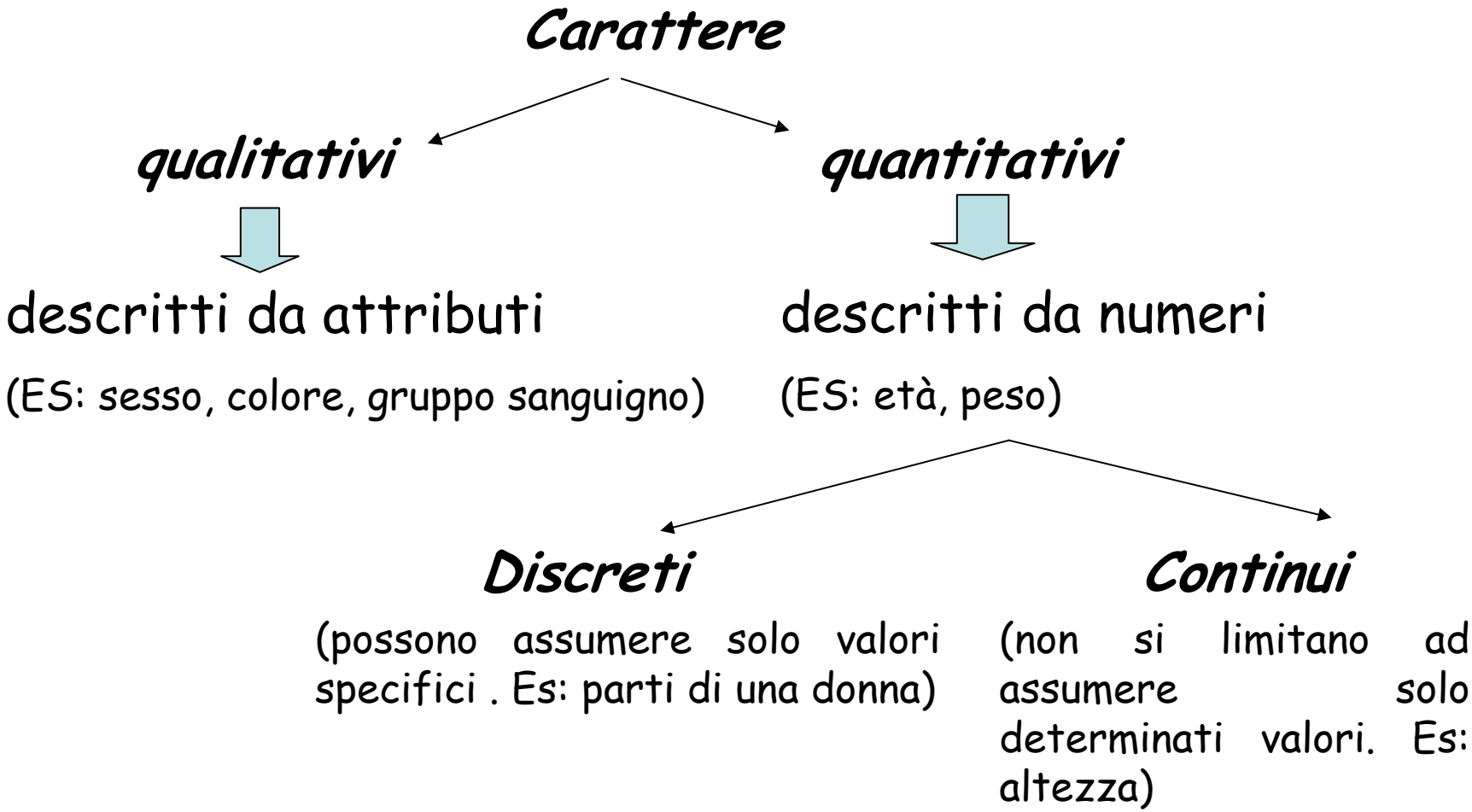
***Campione***

Parte rappresentativa della popolazione

***Carattere***

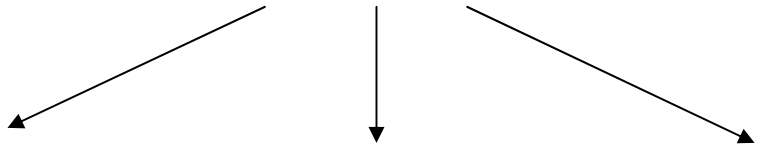
Aspetto, caratteristica, attributo dell'unità statistica

# Carattere


Carattere
qualitativi
quantitativi
descritti da attributi
(ES: sesso, colore, gruppo sanguigno)
descritti da numeri
(ES: età, peso)
Discreti
(possono assumere solo valori specifici . Es: parti di una donna)
Continui
(non si limitano ad assumere solo determinati valori. Es: altezza)

# Modalità

Ogni diversa presentazione del carattere osservato su ciascuna unità statistica



*Ordinate*
*Qualifica, professione…*

*Non ordinate*
*Sesso, professione…*

*Misurabili*
*Età, peso…*

| CARATTERE | |
|---|---|
| *qualitativo* | *quantitativo* |
| modalità non ordinate | |
| modalità ordinate | |
| | modalità misurabili |

si dirà misurato su SCALA

| |
|---|
| **nominale** |
| **ordinale** |
| **a intervalli** |

# Variabile statistica

***Variabile statistica***

Caratteristica che assume valori diversi su persone diverse. E' l'oggetto di studio.

ES: pressione diastolica, peso

*Quantitativa*

(possono essere misurate. Es: età, peso)

*Qualitativa*

(non possono essere misurate, ma solo divise in categorie, cioè classificate. Es: appartenenza ad un gruppo etnico)

**Statistica univariata:** per ogni individuo si raccolgono informazioni relative ad UNA variabile

**Statistica multivariata:** per ogni individuo si raccolgono informazioni relative ad PIU' variabili

# Esempio

Si vuole studiare il livello di glicemia in un gruppo di soggetti diabetici

- **Fenomeno collettivo** → livello di glicemia in soggetti
- **Unità statistica** → ogni paziente affetto dalla malattia
- **Popolazione (o collettivo statistico)** → l'insieme di tutti i diabetici
- **Carattere** → la glicemia
- **Modalità** → Valori diversi che assume la glicemia

Oltre al carattere glicemia ci sono altri caratteri:

- Sesso con modalità maschio-femmina
- Età con modalità età osservate

# Distribuzioni di frequenza

Tipica rappresentazione tabellare per variabili qualitative o per variabili quantitative discrete.

Nella tabella sono riportate:

- le **modalità** della variabile
- le **frequenze** associate a ciascuna modalità

# Distribuzioni di frequenza

**Frequenza assoluta**

Misura quante volte una certa modalità è stata osservata nel collettivo studiato. Solitamente si indica con il simbolo $n_i$

**Frequenza relativa**

Rappresenta la proporzione (talvolta in percentuale) di osservazioni che presentano una certa modalità della variabile analizzata

$$f_i = n_i / N$$

Per aver il valore percentuale: $n_i * 100$.

# Esempio

Su 50 soggetti è stato rilevato il gruppo sanguigno.

| **Gruppo** | **$n_i$** | **$f_i$** |
|---|---|---|
| A | 20 | 0.40 |
| B | 5 | 0.10 |
| AB | 2 | 0.04 |
| 0 | 23 | 0.46 |
| **TOT** | **50** | **1** |

PROPRIETA':

1. $\sum_i n_i = N$

2. $\sum_i f_i = 1$

# Frequenza cumulata (assoluta)

Si usa quando si intende stimare il numero totale di osservazioni inferiore (superiore) ad un valore prefissato

Si indica con $N_i$ la frequenza cumulata **assoluta** e con $F_i$ quella **relativa**.

Si può utilizzare solo quando il carattere è misurato almeno su scala ordinale.

Proprietà:

$$N_1 = n_1$$

$$N_N = N$$

| **Gruppo** | $\mathbf{n_i}$ | $\mathbf{f_i}$ | $\mathbf{N_i}$ | $\mathbf{F_i}$ |
|---|---|---|---|---|
| A | 20 | 0.40 | 20 | 0.40 |
| B | 5 | 0.10 | 25 | 0.50 |
| AB | 2 | 0.04 | 27 | 0.54 |
| 0 | 23 | 0.46 | 50 | 1.0 |
| **TOT** | **50** | **1** | | |

# Osservazioni

☺ Non si perde informazione rilevante

☹ Scarso potere di sintesi se le modalità sono numerose

☹ Non utilizzabile per le variabili continue

….non è del tutto vero….

# Distribuzioni di frequenza per variabili continue

Se siamo disposti a rinunciare ad alcune informazioni, la distribuzione di frequenza può essere costruita anche per variabili continue.

Generalmente si opera nel seguente modo:

- Si suddivide l'insieme dei valori che la variabile può assumere in intervalli, detti **classi**
- Si determina il numero di osservazioni che cadono all'interno di ciascuna classe

# Come costruire le classi?

Non esistono regole assolute per la costruzione , ma è buona norma:

- Evitare di costruire classi con frequenze basse
- Modulare l'ampiezza delle classi in funzione della disponibilità di informazione

| *Distribuzione dei medici in base all'età* | |
|---|---|
| **Classi di età** | **$n_i$** |
| [26,30[ | 40 |
| [30,34[ | 72 |
| [34,42[ | 120 |
| [42,52[ | 61 |
| [52,57[ | 15 |
| [57,65] | 10 |
| **TOT** | **318** |

# Esempio: classi e frequenze

Studiamo il livello di trigliceridi in 250 uomini normali.

| *Livelli di trigliceridi* | | | |
|---|---|---|---|
| **Trigliceridi (mg/100ml)** | **$n_i$** | **$N_i$** | **$f_i$** |
| [130,134] | 2 | 2 | 0.008 |
| [135,139] | 14 | 16 | 0.056 |
| [140,144] | 32 | 48 | 0.128 |
| [145,149] | 39 | … | … |
| [150,154] | 52 | … | … |
| [155,159] | 45 | | |
| [160,164] | 35 | | |
| [165,169] | 13 | | |
| [170,174] | 11 | | |
| [175,179] | 3 | | |
| [180,184] | 1 | | |
| [185,189] | 3 | | |
| **TOT** | **250** | | |

# Ampiezza di classe e densità di frequenza

Data la classe $[x_i, x_{i+1}[$ si chiama **ampiezza di classe**: $\alpha_i = x_{i+1} - x_i$

Si chiama **densità frequenza di classe** $d_i = \frac{n_i}{\alpha_i}$

Si chiama **distribuzione statistica** una successione di dati in cui alle modalità di un carattere sono associate le rispettive frequenze con cui dette modalità si presentano.

| *Distribuzione dei medici in base all'età* | | | |
|---|---|---|---|
| **Classi di età** | **$n_i$** | **$\alpha_i$** | **$d_i$** |
| [26,30[ | 40 | 4 | 10 |
| [30,34[ | 72 | 4 | 18 |
| [34,42[ | 120 | 8 | 15 |
| [42,52[ | 61 | 10 | … |
| [52,57[ | 15 | 5 | |
| [57,65] | 10 | 8 | |
| **TOT** | **318** | | |

# Rappresentazioni grafiche

Strumenti molto utili per visualizzare le caratteristiche di una variabile.

Ne esistono di vari tipi a seconda delle esigenze di analisi

- grafici a barre
- i grafici a settori circolari (grafici a torta)
- gli istogrammi
- i grafici a punti

***NOTA:***

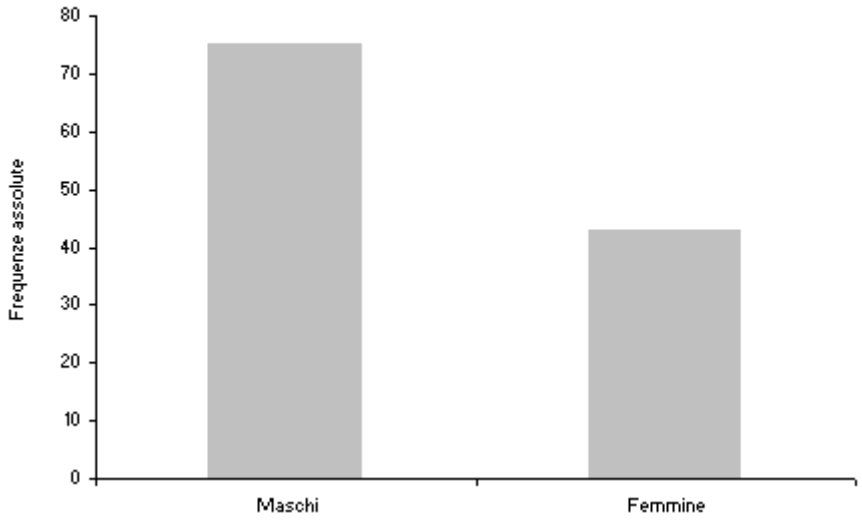
Alcuni riproducono le stesse informazioni di una distribuzione di frequenza, altri riassumono caratteristiche difficilmente rappresentabili con tabelle.

# Diagrammi a barre

Per rappresentare la **frequenza** con cui si presentano le modalità di un **carattere qualitativo** (sesso, religione praticata).

ES:

Rilevazione del carattere *Sesso* di 118 bambini di una scuola elementare. Sono risultati 75 bambini e 43 bambine.




**asse verticale:**

frequenza (assoluta o relativa) con cui le modalità si presentano

**asse orizzontale:**

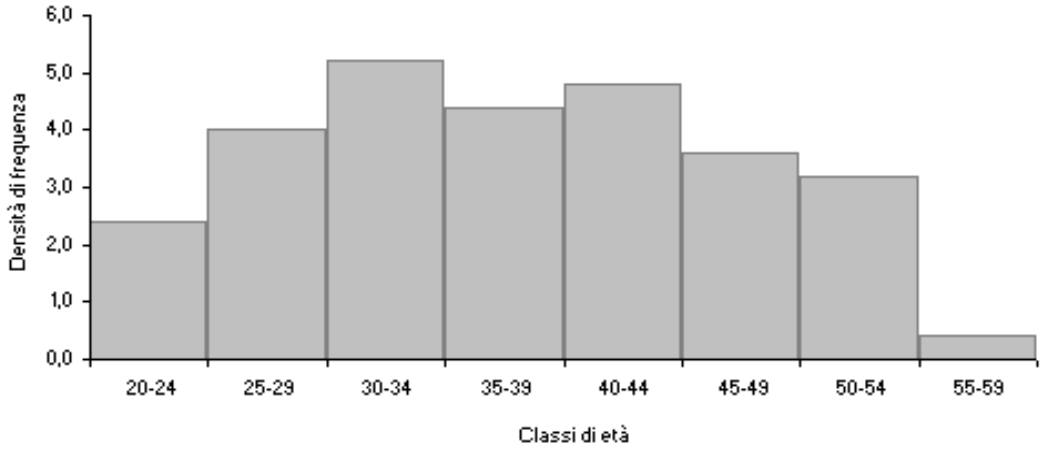
serve soltanto come base di appoggio dell'elemento grafico (le due barre)

# Istogrammi

Per rappresentare **dati quantitativi suddivisi in classi**. Ogni frequenza è rappresentata dall'area di un rettangolo, la cui base è uguale all'ampiezza della classe e l'altezza è pari alla *densità di frequenza.*

ES:

| CLASSI DI ETÀ | Operai Frequenza | Ampiezza della classe | Densità di frequenza |
|---|---|---|---|
| 20-24 | 12 | 5 | 2,4 |
| 25-29 | 20 | 5 | 4,0 |
| 30-34 | 26 | 5 | 5,2 |
| 35-39 | 22 | 5 | 4,4 |
| 40-44 | 24 | 5 | 4,8 |
| 45-49 | 18 | 5 | 3,6 |
| 50-54 | 16 | 5 | 3,2 |
| 55-59 | 2 | 5 | 0,4 |
| **Totale** | **140** | | |



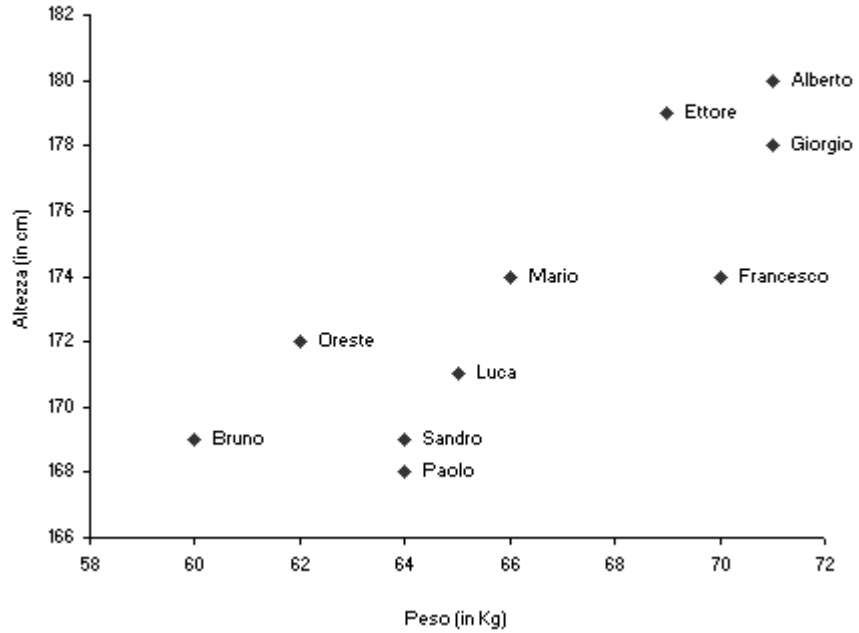
# I grafici a punti

Per rappresentare il valore assunto da due variabili su una stessa unità statistica (per esempio il peso e l'altezza di una persona, oppure l'età e il suo reddito mensile).

OSSERVAZIONE:

E' possibile verificare visivamente se le due variabili sono connesse, cioè se il comportamento di una è legato al comportamento dell'altra.

| ATLETI | Peso (X) | Altezza (Y) |
|---|---|---|
| Mario | 66 | 174 |
| Paolo | 64 | 168 |
| Luca | 65 | 171 |
| Giorgio | 71 | 178 |
| Sandro | 64 | 169 |
| Francesco | 70 | 174 |
| Alberto | 71 | 180 |
| Oreste | 62 | 172 |
| Bruno | 60 | 169 |
| Ettore | 69 | 179 |



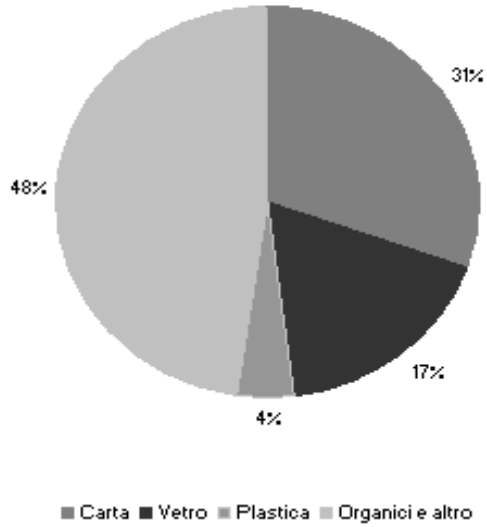
# Settori circolari

Il grafico a settori circolari calcolato sui valori percentuali. E' utilizzato per distribuzioni di variabili nominali al fine di evitare di stabilire un ordine.

Ogni settore del grafico rappresenta (in frequenza assoluta o nell'esempio proposto, percentuale) il peso assunto da ciascuna modalità.

**Tabella 14 - Raccolta di rifiuti urbani differenziata per tipo di rifiuto. Italia - Anno 2001** *(in tonnellate e composizioni percentuali)*

| TIPI DI RIFIUTO | Rifiuti urbani | Composizione % |
|---|---|---|
| Carta | 1.567.806 | 30,7 |
| Vetro | 874.921 | 17,1 |
| Plastica | 230.110 | 4,5 |
| Organici e altro | 2.441.958 | 47,7 |
| **Totale** | **5.114.795** | **100,0** |




$\alpha : 360° = n_i : N$

Evidenzia come sono distribuite le parti rispetto all'intero

# Vantaggi e svantaggi

☺ Conservano la maggior parte dell'informazione

☺ Sono di immediata comprensione

☹ Nonostante la semplicità, non sempre è chiaro quale sia la rappresentazione da utilizzare

# La sintesi dei dati

Si parte da una massa di dati relativi al fenomeno che si vuole analizzare, e si perviene a pochi numeri che descrivono le caratteristiche più rilevanti dei dati.

**Indici di centralità**: forniscono informazioni sui valori intorno ai quali sono prevalentemente concentrati i dati.

**Indici di dispersione**: misurano la dispersione dei dati intorno ai valori centrali.

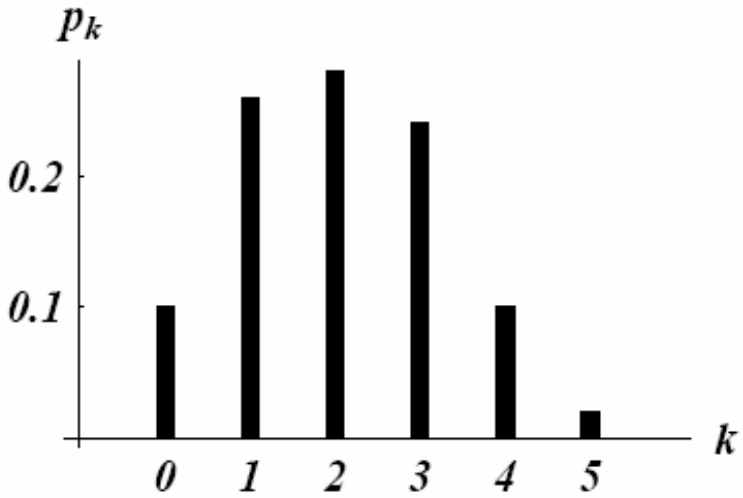
# La moda

Data la distribuzione di frequenze di un carattere discreto chiameremo **moda** il valore corrispondente alla frequenza più grande. Nel caso di un carattere continuo la moda è la classe (o il suo valore centrale) corrispondente al rettangolo più alto dell'istogramma.

Esempio: numero di figli maschi in 50 famiglie con 5 figli.

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 3 | 1 | 1 | 1 | 2 | 4 | 1 | 3 | 2 | 1 | 0 | 2 | 1 | 3 | 3 | 0 | 2 | 1 |
| 3 | 4 | 3 | 1 | 3 | 4 | 1 | 5 | 0 | 2 | 0 | 4 | 1 | 4 | 2 | 2 | 2 | 1 | 2 | 3 |
| 2 | 3 | 2 | 2 | 3 | 3 | 2 | 1 | 2 | 1 | | | | | | | | | | |

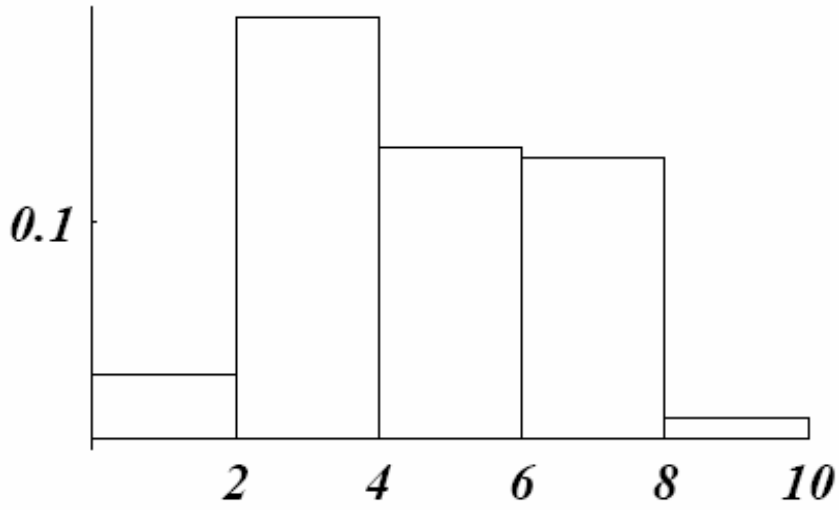
| $k$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $N_k$ | 5 | 13 | 14 | 12 | 5 | 1 |
| $F_k$ | 5 | 18 | 32 | 44 | 49 | 50 |
| $p_k$ | 0.10 | 0.26 | 0.28 | 0.24 | 0.10 | 0.02 |
| $f_k$ | 0.10 | 0.36 | 0.64 | 0.88 | 0.98 | 1.00 |

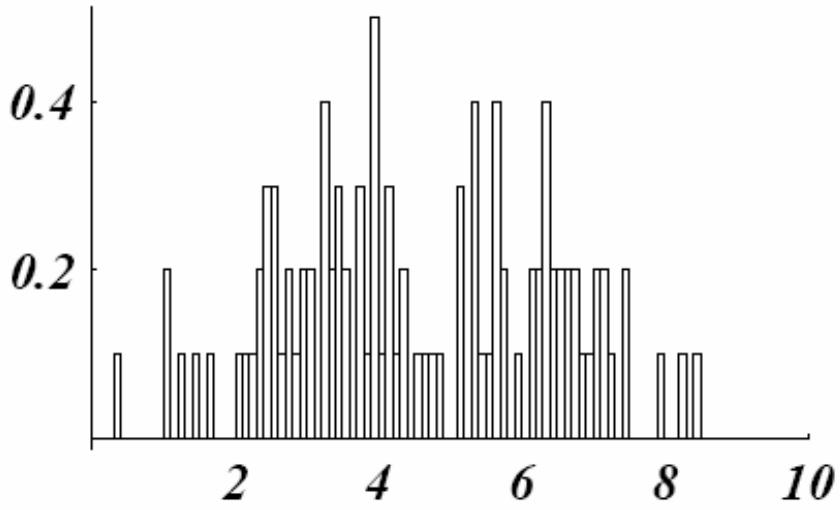


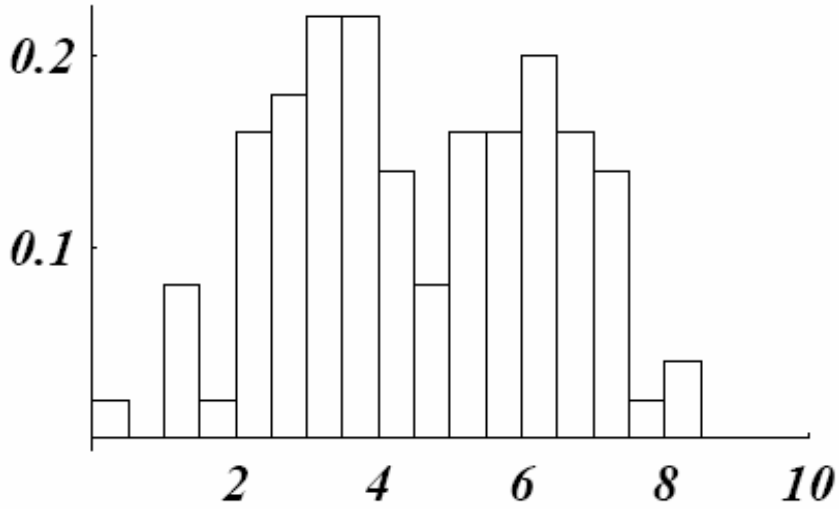
Esempio: livello di espressione di un gene in 100 soggetti analizzati.

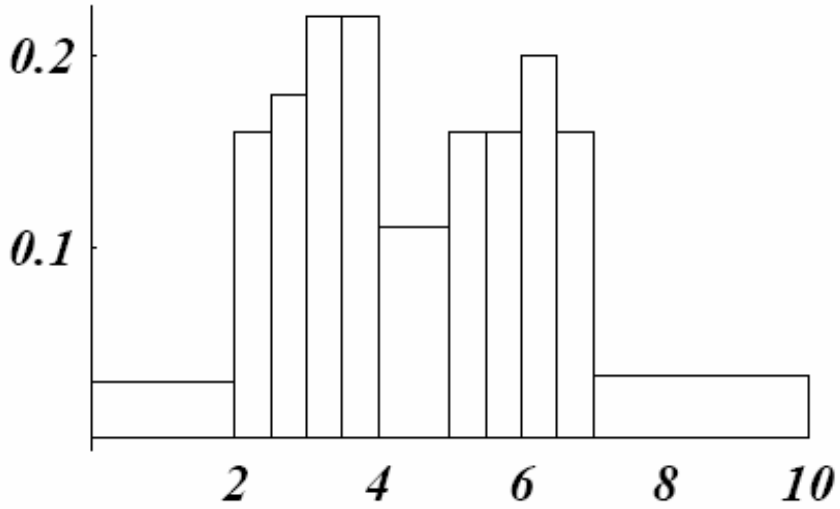
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.30 | 1.03 | 1.08 | 1.22 | 1.46 | 1.62 | 2.01 | 2.17 | 2.27 | 2.31 |
| 2.33 | 2.41 | 2.49 | 2.49 | 2.57 | 2.58 | 2.59 | 2.63 | 2.75 | 2.75 |
| 2.84 | 2.93 | 2.95 | 3.08 | 3.09 | 3.23 | 3.27 | 3.27 | 3.28 | 3.37 |
| 3.39 | 3.42 | 3.47 | 3.49 | 3.56 | 3.60 | 3.78 | 3.78 | 3.79 | 3.87 |
| 3.91 | 3.91 | 3.95 | 3.95 | 3.96 | 4.02 | 4.11 | 4.12 | 4.12 | 4.22 |
| 4.31 | 4.35 | 4.58 | 4.69 | 4.76 | 4.89 | 5.12 | 5.18 | 5.20 | 5.34 |
| 5.34 | 5.37 | 5.40 | 5.46 | 5.54 | 5.62 | 5.64 | 5.64 | 5.68 | 5.71 |
| 5.73 | 5.94 | 6.10 | 6.19 | 6.24 | 6.28 | 6.31 | 6.33 | 6.35 | 6.40 |
| 6.44 | 6.44 | 6.55 | 6.56 | 6.63 | 6.68 | 6.73 | 6.75 | 6.89 | 6.99 |
| 7.01 | 7.08 | 7.11 | 7.15 | 7.26 | 7.44 | 7.47 | 7.93 | 8.21 | 8.44 |

| $J_k$ | $N_k$ | $F_k$ | $p_k$ | $f_k$ |
|---|---|---|---|---|
| $[0.0, 2.0]$ | 6 | 6 | 0.06 | 0.06 |
| $[2.0, 4.0]$ | 39 | 45 | 0.39 | 0.45 |
| $[4.0, 6.0]$ | 27 | 72 | 0.27 | 0.72 |
| $[6.0, 8.0]$ | 26 | 98 | 0.26 | 0.98 |
| $[8.0, 10.0]$ | 2 | 100 | 0.02 | 1.00 |


0.1
2
4
6
8
10


0.4
0.2
2
4
6
8
10


0.2
0.1
2
4
6
8
10


0.2
0.1
2
4
6
8
10

# La media

Si definisce media del campione $x_1, x_2, \ldots, x_n$ del carattere $X$ la quantità:

$$m_X = \bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Esempio: $x_1 = 4$ ,$x_2 = 7$, $x_3 = 9$, $x_4 = 10$, $x_5 = 12$ $\Rightarrow$

$$\bar{x} = \frac{4+7+9+10+12}{5} = \frac{42}{5} = 8.4$$

# Proprietà

Se i dati $x_1, x_2, \ldots, x_n$ sono misure di un carattere discreto $X$ che assume valori $w_1, w_2, \ldots, w_M$ con frequenze relative $p_1, p_2, \ldots, p_M$, allora:

$$m_X = \bar{x} = \sum_{i=1}^{M} p_i w_i$$

Infatti, per definizione di frequenza relativa, si ha che $p_i = N_i/n$ dove $N_i$ è il numero di dati presenti nel campione con valore $w_i$. Allora

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{np_1w_1 + np_2w_2 + \ldots + np_Mw_M}{n} = \sum_{i=1}^{M} p_i w_i$$

## ESEMPIO

Consideriamo i punteggi ($X$) riportati ad un test attitudinale da un gruppo di 25 studenti:

2, 4, 6, 5, 7, 6, 7, 5, 2, 4, 3, 8, 4, 7, 8, 9, 6, 6, 5, 5, 6, 7, 7, 3, 4.

Allora X assume valori $w_i$ con frequenze relative $p_i$ date da:

| $w_i$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| $p_i$ | 2/25 | 2/25 | 4/25 | 4/25 | 5/25 | 5/25 | 2/25 | 1/25 |

$$\bar{x} = \frac{2}{25}2 + \frac{2}{25}3 + \frac{4}{25}4 + \frac{4}{25}5 + \frac{5}{25}6 + \frac{5}{25}7 + \frac{2}{25}8 + \frac{1}{25}9$$
$$= 5.44$$

# Osservazione

Se i dati $x_1, x_2, \ldots, x_n$ sono misure di un carattere discreto $X$ che assume valori $w_1, w_2, \ldots, w_M$ con <u>frequenze assolute</u> $N_1, N_2, \ldots, N_M$, allora:

$$m_X = \overline{x} = \sum_{i=1}^{M} \frac{N_1 w_1 + N_2 w_2 + \ldots + N_M w_M}{N_1 + N_2 + \ldots + N_M}$$

In questo caso dobbiamo normalizzare con la somma delle frequenze assolute.

Confronto con la formula che utilizza le frequenze relative.

Lo stesso procedimento si applica anche nel caso di caratteri continui o dati raggruppati in classi.

Esempio: distribuzione per età di un gruppo di pazienti.

| Anni $x_i$ | $n_i$ | Valore centrale $x'_i$ | $x'_i n_i$ | Risultato |
|---|---|---|---|---|
| [10,20[ | 5 | 15 | 15*5 | 75 |
| [20,30[ | 7 | 25 | 25*7 | 175 |
| [30,40[ | 5 | 35 | 35*5 | 175 |
| [40,50[ | 2 | 45 | 45*2 | 90 |
| [50,60[ | 3 | 55 | 55*3 | 165 |
| [60,70[ | 4 | 65 | 65*4 | 260 |
| TOT | 26 | | | 940 |

$$\bar{x} = \frac{x'_1 n_1 + x'_2 n_2 + .... + x'_k n_k}{n_1 + n_2 + .... + n_k} = \frac{\sum_{i=1}^{k} x'_i n_i}{\sum_{i=1}^{k} n_i}$$

$$\bar{x} = \frac{940}{26} = 36.15$$

# Proprietà

Dati due campioni $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_m$ con medie $\bar{x}$ e $\bar{y}$ e detto $z_1, z_2, \ldots, z_k$ il campione ottenuto unendo i primi due campioni con $k=n+m$, si ha:

$$\bar{z} = \frac{n\bar{x} + m\bar{y}}{k}$$

Infatti:

$$\bar{z} = \frac{1}{k}\sum_{i=1}^{k} z_i = \frac{1}{k}\left(\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i\right) = \frac{n\bar{x} + m\bar{y}}{k}$$

Osserva che $\bar{z}$ è la *media pesata* delle medie dei due campioni con pesi $n/k$ e $m/k$.

# ESEMPIO

Se l'età media di un gruppo di 15 donne ricoverate in clinica è di 45 anni e quella di un gruppo di 25 uomini è di 55 anni, l'età media di tutti i ricoverati è

$$\bar{x} = \frac{45*15+55*25}{40} = 51.25$$

# Media pesata

Assegnati i numeri $x_1,x_2,\ldots,x_n$ e i pesi $p_1,p_2,\ldots,p_n$ tali che:

1) $0 \leq p_i \leq 1$ per $i=1,2,\ldots,n$

2) $p_1+p_2+\ldots+p_n=1$

si chiama media pesata il numero:

$$\overline{x} = \sum_{i=1}^{n} p_i x_i$$

# La variabilità dei dati

MEDIE: 1solo numero ⇒ non mette in evidenza tutte le caratteristiche del fenomeno.

Valori di glicemia in 3 soggetti:

x1: 96 98 105 97 95

x2: 86 100 108 99 98

x3: 86 125 95 76 109

Andamento costante

Varia tra 86-108 intorno a media

Varia tra 76-125. E' molto disperso

MEDIA=98.2

Cioè tutti gli individui sono sani!!!!

# Varianza di un campione

Chiameremo varianza di un campione $x_1, x_2, \ldots, x_n$ di $X$ con media $\bar{x}$ la quantità

$$s_X^2 = \frac{1}{n} \sum_{i=1}^{n} \left( x_i - \bar{x} \right)^2$$

E scarto quadratico o standard deviation la radice quadrata $s_X$ della varianza.

# Osservazione

Quelle definite sono misure di dispersione del campione intorno alla media.

Grandi valori di $s^2_X$ indicano la presenza di dati molto lontani dalla media; piccoli valori di $s^2_X$ indicano dati concentrati intorno alla media.

Se $s^2_X=0$ allora tutti i dati sono uguali e coincidono con la media.

# Esempio

Livello di colesterolemia di cinque soggetti:

128 130 134 132 140

$$\bar{x} = \frac{128+130+134+132+140}{5} = 132.8$$

$$s_X^2 = \frac{1}{5}\left[(128-132.8)^2 + (130-132.8)^2 + \ldots + (140-132.8)^2\right]$$

$$= \frac{84.8}{5} = 16.96$$

$$s_X = \sqrt{16.96} = 4.12$$

# Proprietà

Se i dati $x_1,\ldots,x_n$ sono misure di un carattere discreto $X$ che assume valori $w_1,\ldots,w_M$ con frequenze relative $p1,\ldots,p_M$ allora:

$$s_X^2 = \sum_{i=1}^{M} p_i \left(w_i - \overline{x}\right)^2$$

Dimostrazione per esercizio.

# Proprietà

Dato un campione $x_1,\ldots,x_n$ con media $\bar{x}$ si ha:

$$s_X^2 = \overline{x^2} - \bar{x}^2$$

Infatti:

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2$$

$$= \overline{x^2} - 2\bar{x}\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2.$$

# Trasformazione lineare di un campione

Dato un campione $x_1,\ldots,x_n$ di $X$ con media $\bar{x}$ e varianza $s^2_X$ e due numeri $a$ e $b$, definito il nuovo campione $y_i=ax_i+b$ si ha:

$$\bar{y} = a\bar{x} + b \quad e \quad s_Y^2 = a^2 s_X^2$$

Infatti:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} ax_i + b = a\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} b.$$

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(ax_i + b - a\bar{x} - b\right)^2 =$$

$$= a^2\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = a^2 s_X^2 .$$

# Errore quadratico medio

Chiameremo errore quadratico medio di un campione $x_1, \ldots, x_n$ rispetto al numero $\alpha$ la quantità

$$\mathcal{E}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2$$

Osservazione. $\mathcal{E}(\alpha)$ misura l'errore che si commette se approssimiamo ogni elemento del campione con $\alpha$ .

Qual è il valore di $\alpha$ che minimizza $\mathcal{E}(\alpha)$ ?

# Proprietà

La media $\bar{x}$ di un campione $x_1, \ldots, x_n$ è il valore di $\alpha$ che minimizza l'errore quadratico medio del campione.

Infatti,

$$\frac{d\mathcal{E}(\alpha)}{d\alpha} = 0 \Rightarrow -\frac{1}{n} 2 \sum_{i=1}^{n} (x_i - \alpha) = 0 \Rightarrow$$

$$\sum_{i=1}^{n} (x_i - \alpha) = 0 \Rightarrow \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \alpha \Rightarrow \alpha = \frac{1}{n} \sum_{i=1}^{n} x_i$$

# Campione standardizzato

Diremo che x1, x2, …, xn è un campione standardizzato quando ha media nulla e varianza unitaria:

$$\overline{x} = 0 \quad e \quad s_X^2 = 1$$

# Standardizzazione di un campione

Dato il campione $x_1, \ldots, x_n$ con media $\bar{x}$ e varianza $s^2_X$, il campione

$$y_i = \frac{x_i - \bar{x}}{s_X} \qquad \mathrm{i} = 1,2,...,\mathrm{n}$$

è standardizzato.

Infatti se poniamo $a = \frac{1}{s_X}$ e $b = -\frac{\bar{x}}{s_X}$

allora $y_i = ax_i + b$. Calcoliamo:

$$\bar{y} = a\bar{x} + b \Rightarrow \bar{y} = \frac{1}{s_X}\bar{x} - \frac{\bar{x}}{s_X} \Rightarrow \bar{y} = 0.$$

$$s_Y^2 = a^2 s_X^2 \Rightarrow s_Y^2 = \frac{1}{s_X^2} s_X^2 \Rightarrow s_Y^2 = 1.$$

# Esempio

Consideriamo il campione:

$x_1$=1.3, $x_2$=2.4, $x_3$=-3.5, $x_4$=6.9, $x_5$=-5.2

con

$\bar{x}$=0.38, $s^2_x$=18.7256, $s_X$=4.3273

Il nuovo campione

$y_1$=0.2126, $y_2$=0.4668, $y_3$=-0.8966, $y_4$=1.5067,

$y_5$=-1.2895

è standardizzato, ossia $\bar{y}$=1 e $s^2_Y$=0.

# Coefficiente di variazione

Dato un campione $x_1, x_2, \ldots, x_n$ di $X$ con media $\bar{x}$ e varianza $s^2_X$, si chiama coefficiente di variazione il rapporto:

$$cv = \frac{s_X}{|\bar{x}|}$$

Il CV è una misura di variabilità relativa utile quando si vogliono confrontare le variabilità di due campioni relativi anche a caratteri eterogenei. Esso è il rapporto di due quantità con la stessa unità di misura, per cui il CV è adimensionato.

# Esempio

Si voglia confrontare la variabilità della diuresi nelle 24 ore e della pressione in cinque soggetti

| | Pressione (mmHg) | Urine(ml) |
|---|---|---|
| | 120<br>140<br>160<br>180<br>130 | 1250<br>1200<br>900<br>850<br>1080 |
| $\overline{x}$ | 146 | 1056 |
| $s$ | 21.5 | 158.5 |
| $100CV$ | **147.2** | **150.1** |

Media e deviazione standard mostrano molta variabilità!

Omogeneizzando le misure, si osserva che la variabilità è quasi uguale.

# Mediana e quantili

Consideriamo il campione $x_1, x_2, \ldots, x_n$ ed ordiniamo in ordine crescente i suoi elementi. Il campione ordinato sarà indicato con la notazione

$$x_{[1]}, x_{[2]}, \ldots, x_{[n]}$$

dove:

$$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$$

# Quantile

Chiameremo quantile di ordine $\alpha$ ($0<\alpha<1$) di un campione $x_1, x_2, \ldots, x_n$ un numero $q_\alpha$ maggiore o uguale di una frazione $\alpha$ degli elementi del campione ordinato $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ nel senso che il numero di $x_{[i]}$ che risulta minore o uguale di $q_\alpha$ non deve superare $\alpha(n+1)$.

# Procedura per il calcolo di $q_\alpha$

Si calcola il numero $\alpha(n+1)$. Poi:

- se $\alpha(n+1)$ è intero allora si considera l'indice $i=\alpha(n+1)$ e si pone $q_\alpha = x_{[i]}$ ;
- se $\alpha(n+1)$ non è intero allora si considera l'indice $i$ tale che $i<\alpha(n+1)<i+1$ e si pone

$$q_\alpha = \frac{x_{[j]} + x_{[j+1]}}{2}$$

# Definizioni

Il quantile di ordine $\alpha$ = ½ prende il nome di **mediana**.

I quantili di ordine $\alpha = k/4$, con $k$=1,2,3, si chiamano primo, secondo, terzo **quartile** rispettivamente.

I quantili di ordine $\alpha = k/10$, con $k$=1,2,…,9, si chiamano **decili**.

I quantili di ordine $\alpha = k/100$, con $k$=1,2,…,99, si chiamano **percentili**.

# Esempio

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.30 | 1.03 | 1.08 | 1.22 | 1.46 | 1.62 | 2.01 | 2.17 | 2.27 | 2.31 |
| 2.33 | 2.41 | 2.49 | 2.49 | 2.57 | 2.58 | 2.59 | 2.63 | 2.75 | 2.75 |
| 2.84 | 2.93 | 2.95 | 3.08 | 3.09 | 3.23 | 3.27 | 3.27 | 3.28 | 3.37 |
| 3.39 | 3.42 | 3.47 | 3.49 | 3.56 | 3.60 | 3.78 | 3.78 | 3.79 | 3.87 |
| 3.91 | 3.91 | 3.95 | 3.95 | 3.96 | 4.02 | 4.11 | 4.12 | 4.12 | 4.22 |
| 4.31 | 4.35 | 4.58 | 4.69 | 4.76 | 4.89 | 5.12 | 5.18 | 5.20 | 5.34 |
| 5.34 | 5.37 | 5.40 | 5.46 | 5.54 | 5.62 | 5.64 | 5.64 | 5.68 | 5.71 |
| 5.73 | 5.94 | 6.10 | 6.19 | 6.24 | 6.28 | 6.31 | 6.33 | 6.35 | 6.40 |
| 6.44 | 6.44 | 6.55 | 6.56 | 6.63 | 6.68 | 6.73 | 6.75 | 6.89 | 6.99 |
| 7.01 | 7.08 | 7.11 | 7.15 | 7.26 | 7.44 | 7.47 | 7.93 | 8.21 | 8.44 |

$n = 100.$ Mediana : $\alpha(n+1) = 50.5.$ Prenderemo $i = 50$ e

$$q_{\frac{1}{2}} = \frac{x_{[50]} + x_{[51]}}{2} = \frac{4.22 + 4.31}{2} = 4.265$$

Primo quartile : $\alpha(n+1) = 25.25.$ Prenderemo $i = 25$ e

$$q_{\frac{1}{4}} = \frac{x_{[25]} + x_{[26]}}{2} = \frac{3.09 + 3.23}{2} = 3.16$$

Prima di calcolare i quantili i dati devono essere ordinati.

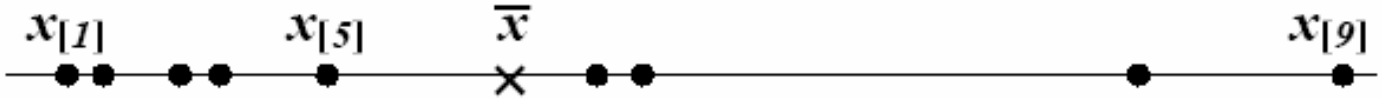
# Proprietà della mediana

La mediana, come la media e la moda, è un indice di centralità.

In generale la mediana è un indice più *robusto* della media nel senso che il suo valore è meno sensibile a variazioni o errori presenti nei dati.

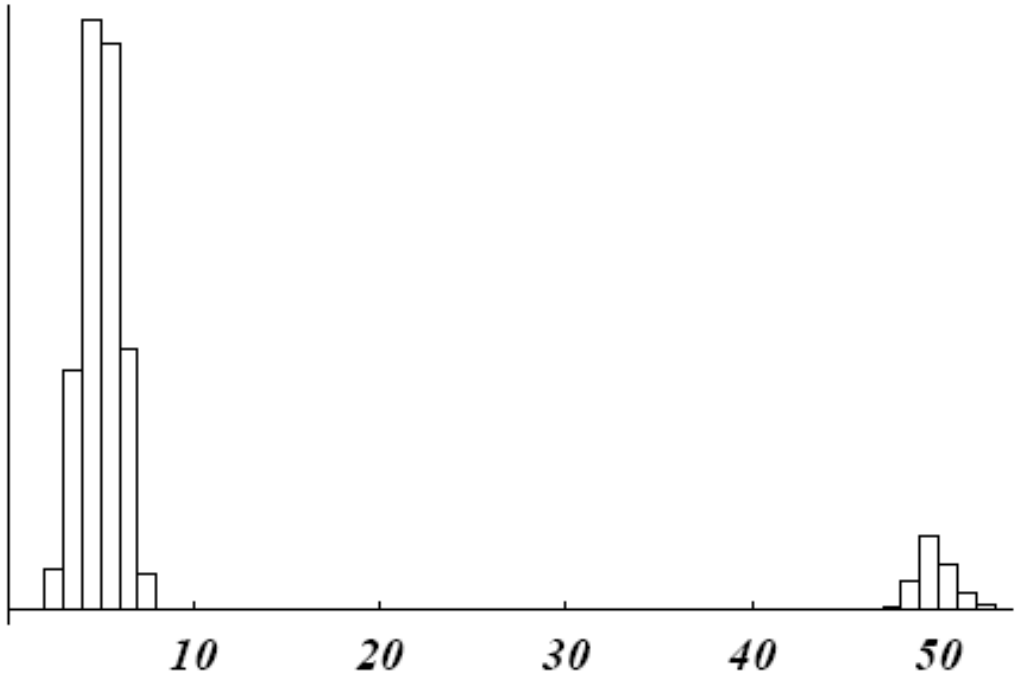
Vediamo perché.

# Esempio




In questo campione con n=9 la mediana coincide con $x_{[5]}$ e la media è indicata con $\overline{x}$. Consideriamo $x_{[9]}$. Ogni variazione anche piccola di questo punto influenza la media del campione. Queste variazioni, al contrario, non cambiano il valore della mediana del campione. Il valore della mediana rimane lo stesso finché il punto $x_{[9]}$ rimane alla destra di $x_{[5]}$.

# Esempio




I redditi di 1000 impiegati si distribuiscono intorno a 5, mentre quello di 100 dirigenti intorno a 50. Si ha che la mediana è 5.13 e la media 9.08. In questo caso la mediana è più rappresentativa del reddito tipico.

# Range e differenza interquartile

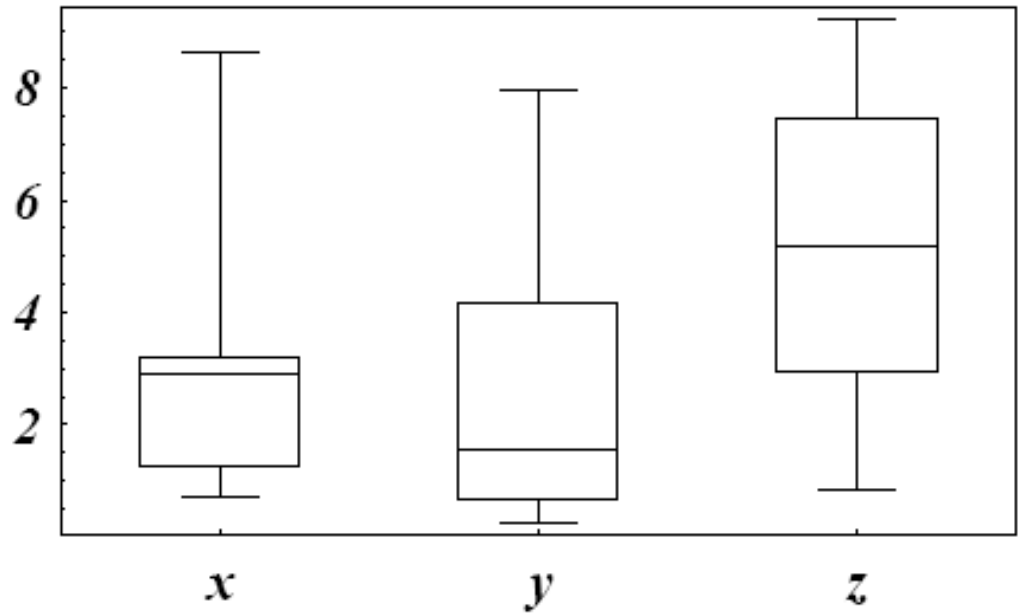
Chiameremo range del campione il numero $x_{[n]} - x_{[1]}$, ovvero l'ampiezza dell'intervallo che contiene tutti i dati.

Chiameremo differenza interquartile il numero $q_{3/4} - q_{1/4}$ , ossia l'ampiezza dell'intervallo delimitato dai due quartili.

Nota che range e differenza interquartile sono indici di dispersione.

# Box-plot

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x_{[i]}$ | 0.72 | 1.10 | 1.24 | 1.98 | 2.82 | 2.99 | 3.01 | 3.18 |
| | 3.31 | 8.64 | | | | | | |
| $y_{[j]}$ | 0.25 | 0.66 | 0.68 | 1.07 | 1.09 | 1.15 | 1.94 | 3.11 |
| | 4.18 | 4.79 | 6.18 | 7.94 | | | | |
| $z_{[k]}$ | 0.85 | 1.49 | 2.19 | 2.93 | 4.46 | 4.61 | 4.62 | 5.16 |
| | 5.67 | 6.41 | 6.46 | 7.45 | 7.66 | 8.65 | 9.22 | |



# Media geometrica

Sia $x_1,\ldots,x_n$ un campione tale che $x_i>0$, i=1,…,n, e consideriamo il campione: $y_i=\log(x_i)$. Calcoliamo:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} log(\, x_i \,)$$

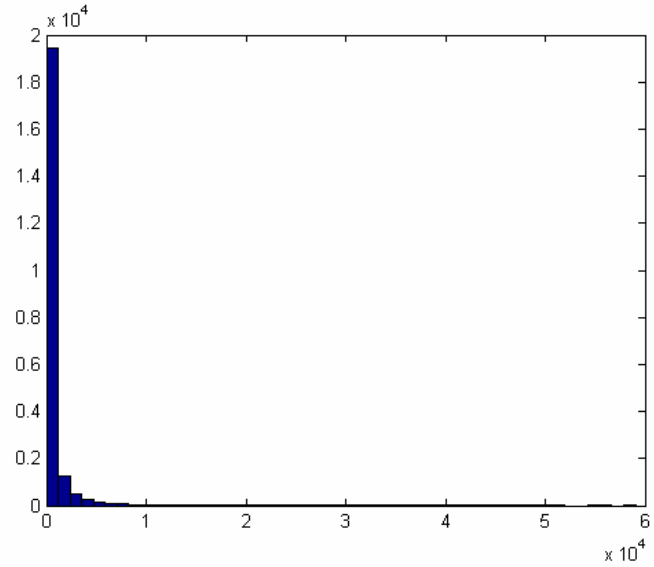
$$= \frac{1}{n} log\left(x_1 x_2 \cdot \ldots \cdot x_n\right) = log\left(x_1 x_2 \cdot \ldots \cdot x_n\right)^{\frac{1}{n}}$$

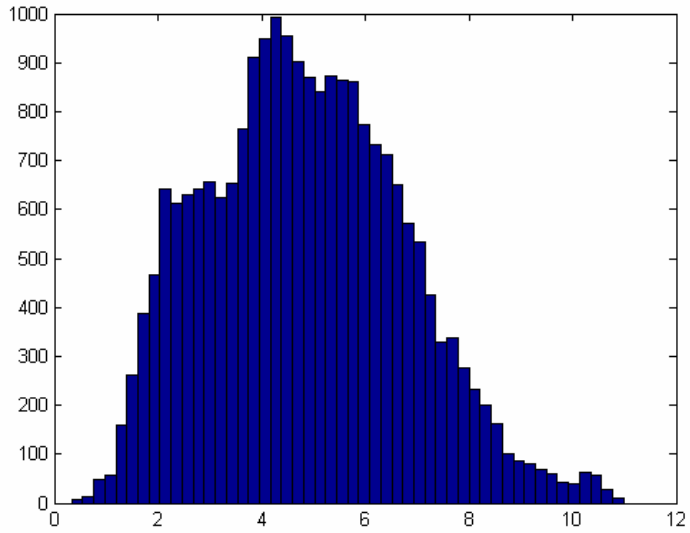
La quantità

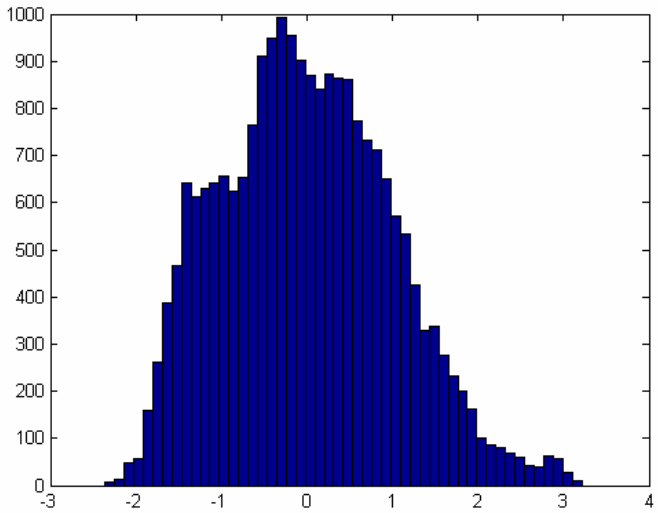
$$\left(x_1 x_2 \cdot \ldots \cdot x_n\right)^{\frac{1}{n}}$$

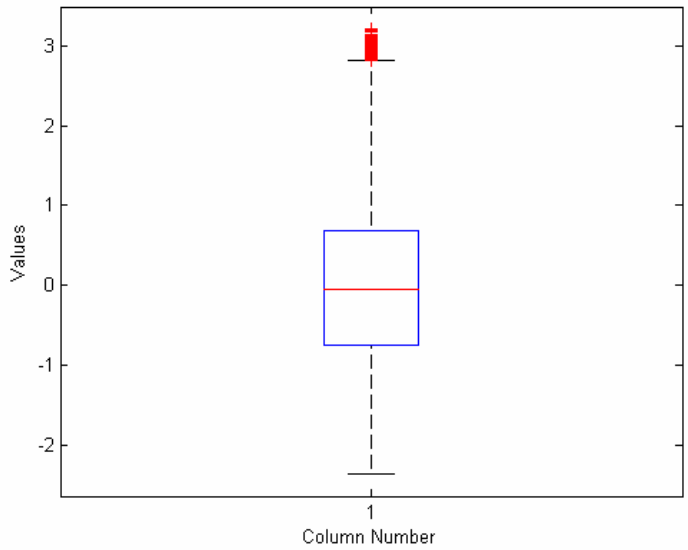
si chiama media geometrica del campione.

# Preprocessing di dati di microarray



# Media armonica

Supponiamo di conoscere il tempo di sopravvivenza di 5 rane sottoposte a trattamento digitalico. Calcoliamo il tempo di sopravvivenza medio del campione.

| Rane | Tempo di sopravvivenza (h) | Rane morte nell'unità di tempo |
|---|---|---|
| 1 | 10 | 1/10=0.1 |
| 2 | 16 | 1/16=0.062 |
| 3 | 21 | 1/21=0.047 |
| 4 | 32 | 1/32=0.031 |
| 5 | 48 | 1/48=0.02 |
| | 127 | 0.26 |

Quindi 0.26/5 è il numero medio di rane morte nell'unità di tempo per cui il tempo di sopravvivenza medio è 5/0.26=19.2 ore.

# Media armonica (cont.)

Per cui in generale dato un campione $x_1,\ldots,x_n$ la media armonica è data dalla seguente quantità:

$$\left( \frac{x_1^{-1} + \ldots + x_n^{-1}}{n} \right)^{-1}$$