Elementi di Statistica Descrittiva

Vlacci Fabio

A.A. 2023-24

Terminologia

- In un *esperimento* ogni risultato delle caratteristiche *osservabili* dell'esperimento si dice *esito*.

Terminologia

- In un *esperimento* ogni risultato delle caratteristiche *osservabili* dell'esperimento si dice *esito*.
- Se un esperimento può essere ripetuto si dice *prova* ogni singola esecuzione dell'esperimento.

Terminologia

- In un *esperimento* ogni risultato delle caratteristiche *osservabili* dell'esperimento si dice *esito*.
- Se un esperimento può essere ripetuto si dice *prova* ogni singola esecuzione dell'esperimento.
- Un *evento* è un insieme di esiti.

Terminologia

- In un *esperimento* ogni risultato delle caratteristiche *osservabili* dell'esperimento si dice *esito*.
- Se un esperimento può essere ripetuto si dice *prova* ogni singola esecuzione dell'esperimento.
- Un *evento* è un insieme di esiti. I singoli esiti sono anche detti *eventi elementari* e, in questo senso, un evento è anche detto *evento composto* (da eventi elementari)).

Terminologia

- In un *esperimento* ogni risultato delle caratteristiche *osservabili* dell'esperimento si dice *esito*.
- Se un esperimento può essere ripetuto si dice *prova* ogni singola esecuzione dell'esperimento.
- Un *evento* è un insieme di esiti. I singoli esiti sono anche detti *eventi elementari* e, in questo senso, un evento è anche detto *evento composto* (da eventi elementari)).
- Un esito o un evento di un esperimento si dicono *casuali* o *aleatori* se non è possibile prevederne il verificarsi a priori in modo certo.
- La totalità degli eventi elementari associati ad un esperimento è lo *spazio campionario* dell'esperimento.

Terminologia

- In un *esperimento* ogni risultato delle caratteristiche *osservabili* dell'esperimento si dice *esito*.
- Se un esperimento può essere ripetuto si dice *prova* ogni singola esecuzione dell'esperimento.
- Un *evento* è un insieme di esiti. I singoli esiti sono anche detti *eventi elementari* e, in questo senso, un evento è anche detto *evento composto* (da eventi elementari)).
- Un esito o un evento di un esperimento si dicono *casuali* o *aleatori* se non è possibile prevederne il verificarsi a priori in modo certo.
- La totalità degli eventi elementari associati ad un esperimento è lo *spazio campionario* dell'esperimento.

evento $\mathcal{A} \to A \subset S$ spazio campionario

Dati campionari

In genere in una raccolta di dati campionari di una popolazione o *rilevazione campionaria*, oltre agli esiti ovvero alle caratteristiche della popolazione che si intendono registrare, vengono indicati anche le frequenze (relative e/o assolute) per tali dati.

Dati campionari

In genere in una raccolta di dati campionari di una popolazione o *rilevazione campionaria*, oltre agli esiti ovvero alle caratteristiche della popolazione che si intendono registrare, vengono indicati anche le frequenze (relative e/o assolute) per tali dati.

Inoltre *prima* di procedere alla rilevazione campionaria andrebbe appositamente definito una procedura per il *piano di campionamento*

Dati campionari

In genere in una raccolta di dati campionari di una popolazione o *rilevazione campionaria*, oltre agli esiti ovvero alle caratteristiche della popolazione che si intendono registrare, vengono indicati anche le frequenze (relative e/o assolute) per tali dati.

Inoltre *prima* di procedere alla rilevazione campionaria andrebbe appositamente definito una procedura per il *piano di campionamento* Ad esempio, se bisogna campionare un'area territoriale (abbastanza estesa), sarà cura del ricercatore stabilire se (per ragioni pratiche o teoriche) sia meglio seguire uno schema di rilevamento casuale o con geometria regolare (ad esempio uniforme o a grappoli).

Rappresentazione dei dati statistici

Le frequenze ma anche le serie storiche dei dati possono essere visualizzati usando dei diagrammi; in particolare potranno essere usati dei *diagrammi a punti*, dei *diagrammi a linea* o *istogrammi* e degli *areogrammi* come i diagrammi a torta o pie charts.

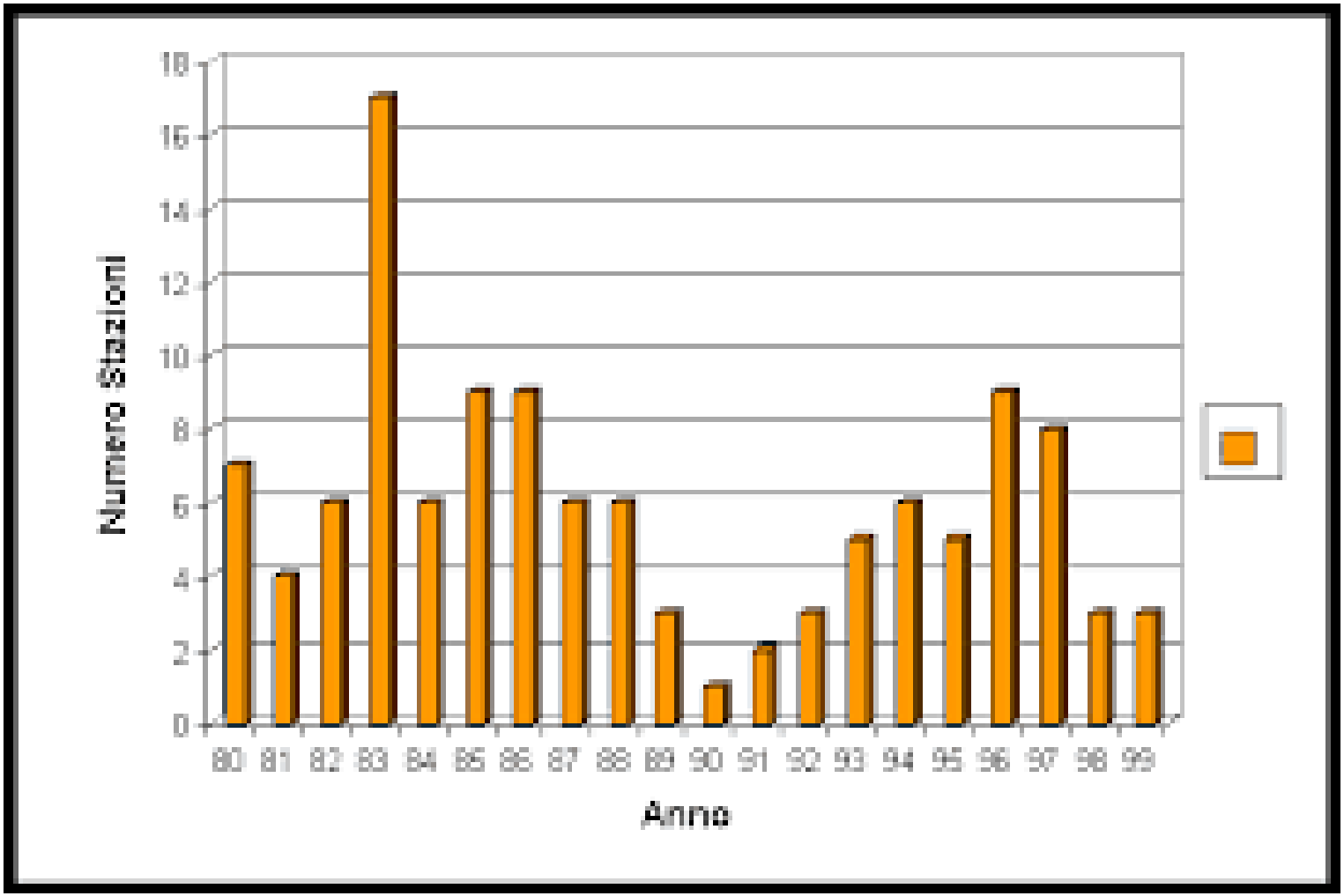
Rappresentazione dei dati statistici

Le frequenze ma anche le serie storiche dei dati possono essere visualizzati usando dei diagrammi; in particolare potranno essere usati dei *diagrammi a punti*, dei *diagrammi a linea* o *istogrammi* e degli *areogrammi* come i diagrammi a torta o pie charts. Quando il numero di osservazioni è molto grande, è bene considerare una rappresentazione a istogrammi per intervalli o *classi* di valori.

Rappresentazione dei dati statistici

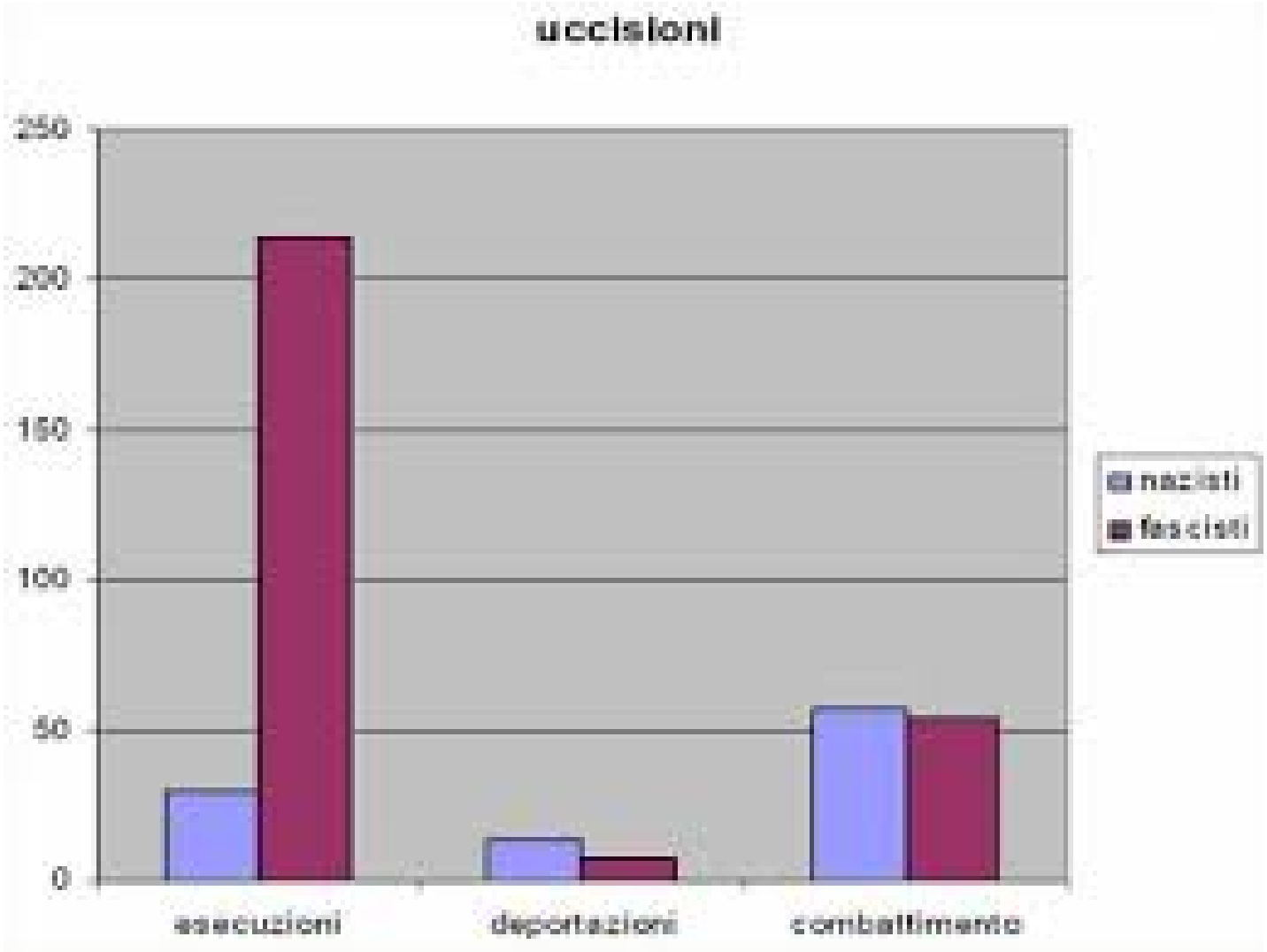
Le frequenze ma anche le serie storiche dei dati possono essere visualizzati usando dei diagrammi; in particolare potranno essere usati dei *diagrammi a punti*, dei *diagrammi a linea* o *istogrammi* e degli *areogrammi* come i diagrammi a torta o pie charts. Quando il numero di osservazioni è molto grande, è bene considerare una rappresentazione a istogrammi per intervalli o *classi* di valori. La scelta del numero di intervalli è una questione delicata ed *influenza* l'aspetto dell'istogramma.

Esempio di Istogramma

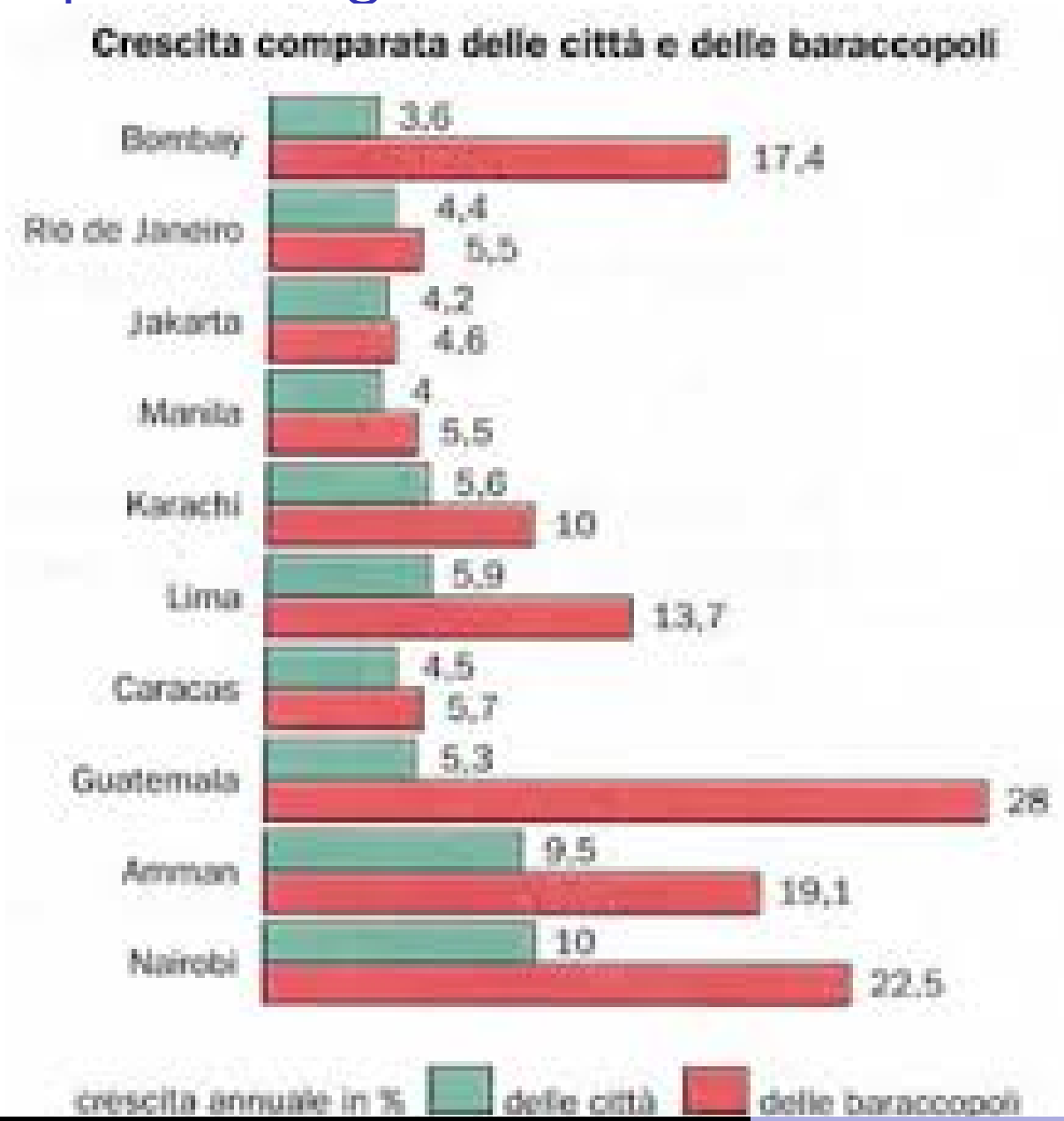


Esempio di Istogramma

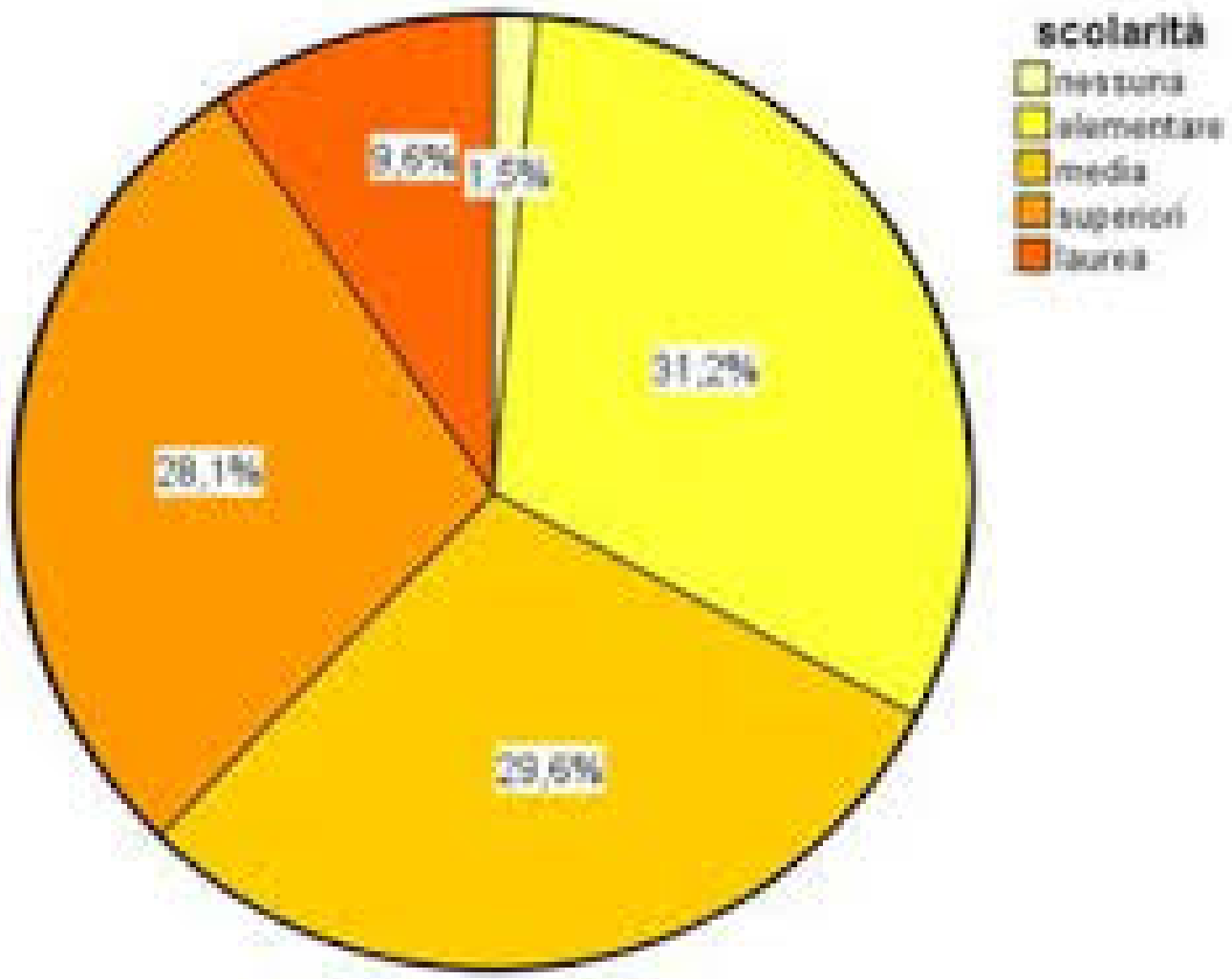
Esempio di Istogramma



Esempio di Istogramma



Pie chart o Diagramma a torta



Indicatori statistici di dati numerici

Per ottenere una descrizione sommaria di un insieme di dati numerici e per confrontare fra loro insiemi diversi di dati numerici campionari è opportuno ricavare dei termini riassuntivi detti *indicatori statistici*.

Indicatori statistici di dati numerici

Per ottenere una descrizione sommaria di un insieme di dati numerici e per confrontare fra loro insiemi diversi di dati numerici campionari è opportuno ricavare dei termini riassuntivi detti *indicatori statistici*.

In particolare tra gli indicatori di *tendenza centrale* si evidenziano

- media campionaria

Indicatori statistici di dati numerici

Per ottenere una descrizione sommaria di un insieme di dati numerici e per confrontare fra loro insiemi diversi di dati numerici campionari è opportuno ricavare dei termini riassuntivi detti *indicatori statistici*.

In particolare tra gli indicatori di *tendenza centrale* si evidenziano

- media campionaria
- mediana

Media campionaria o Media aritmetica dei dati

Se $x_1,\ x_2, , \ldots,\ x_n$ sono i dati numerici rilevati, la media campionaria è il numero

$$\overline{x} := \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Media campionaria o Media aritmetica dei dati

Se $x_1,\ x_2, , \ldots,\ x_n$ sono i dati numerici rilevati, la media campionaria è il numero

$$\overline{x} := \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- La media campionaria $\overline{x}$ NON è necessariamente uno dei dati;

Media campionaria o Media aritmetica dei dati

Se $x_1,\ x_2, , \ldots,\ x_n$ sono i dati numerici rilevati, la media campionaria è il numero

$$\overline{x} := \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- La media campionaria $\overline{x}$ NON è necessariamente uno dei dati;
- tuttavia $\min\{x_i\}_{i=1,\ldots,n} \leq \overline{x} \leq \max\{x_i\}_{i=1,\ldots,n}$.
- Esistono anche ALTRE medie (geometrica, armonica, ecc.), ma quella campionaria o aritmetica può essere presa come *baricentro* dei dati.

Osservazione importante

Poichè la somma di numeri reali è commutativa e associativa, allora

$$\overline{x} := \frac{\overbrace{x_1 + \ldots + x_1}^{n_1}}{n} + \frac{\overbrace{x_2 + \ldots + x_2}^{n_2}}{n} + + \cdots + \frac{\overbrace{x_n + \ldots + x_n}^{n_n}}{n}$$

con $n_1 + n_2 + \ldots + n_n = n$.

Osservazione importante

Poichè la somma di numeri reali è commutativa e associativa, allora

$$\overline{x} := \frac{\overbrace{x_1 + \ldots + x_1}^{n_1}}{n} + \frac{\overbrace{x_2 + \ldots + x_2}^{n_2}}{n} + + \cdots + \frac{\overbrace{x_n + \ldots + x_n}^{n_n}}{n}$$

con $n_1 + n_2 + \ldots + n_n = n$.
Si noti che

$$\overline{x} = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n$$

ove $p_j = \frac{n_j}{n}$ è la frequenza relativa di x_j.

Mediana

Un altro indicatore statistico di tendenza centrale è la *mediana campionaria*.

Mediana

Un altro indicatore statistico di tendenza centrale è la *mediana campionaria*.
Esso è il valore che divide in due parti uguali i dati, quando questi sono ordinati in senso crescente o decrescente.

Mediana

Un altro indicatore statistico di tendenza centrale è la *mediana campionaria*.
Esso è il valore che divide in due parti uguali i dati, quando questi sono ordinati in senso crescente o decrescente. Più precisamente, se il numero n dei dati è dispari, allora il valore della mediana è dato dal dato (detto *mediano*) di posizione $(n+1)/2$ nell'elenco ordinato di dati. Se invece n è pari vi saranno due dati mediani (quello di posizione $n/2$ e il successivo) e allora andrà considerata come mediana la media capionaria dei due dati mediani.

Quartili e percentili

Se il numero di osservazioni/dati è elevato, può essere utile suddividere ulteriormente l'insieme dei dati procedendo al calcolo della mediana della prima e della seconda metà dei dati una volta ordinati in senso crescente o decrescente.

Quartili e percentili

Se il numero di osservazioni/dati è elevato, può essere utile suddividere ulteriormente l'insieme dei dati procedendo al calcolo della mediana della prima e della seconda metà dei dati una volta ordinati in senso crescente o decrescente.
In questo modo i dati sono suddivisi in 4 classi dette *quartili*.

Quartili e percentili

Se il numero di osservazioni/dati è elevato, può essere utile suddividere ulteriormente l'insieme dei dati procedendo al calcolo della mediana della prima e della seconda metà dei dati una volta ordinati in senso crescente o decrescente.
In questo modo i dati sono suddivisi in 4 classi dette *quartili*.
In generale, se i dati possono essere divisi in 100 classi, dette *percentili*, il primo quartile corrisponde al 25-simo percentili mentre il 50-esimo percentile coincide con la mediana o secondo quartile.

Quartili e percentili

Se il numero di osservazioni/dati è elevato, può essere utile suddividere ulteriormente l'insieme dei dati procedendo al calcolo della mediana della prima e della seconda metà dei dati una volta ordinati in senso crescente o decrescente.
In questo modo i dati sono suddivisi in 4 classi dette *quartili*.
In generale, se i dati possono essere divisi in 100 classi, dette *percentili*, il primo quartile corrisponde al 25-simo percentili mentre il 50-esimo percentile coincide con la mediana o secondo quartile.
Per un valore p ($0 \leq p \leq 100$) si parla di p-esimo percentile.

Moda

La *moda campionaria* (detta anche *valore normale* o *norma*) è il dato (o i dati) che si presenta(no) con frequenza massima.

Moda

La *moda campionaria* (detta anche *valore normale* o *norma*) è il dato (o i dati) che si presenta(no) con frequenza massima. Di conseguenza si parla di *distribuzione dei dati* di tipo *unimodale* o *bimodale* e in generale *plurimodale*.

Moda

La *moda campionaria* (detta anche *valore normale* o *norma*) è il dato (o i dati) che si presenta(no) con frequenza massima. Di conseguenza si parla di *distribuzione dei dati* di tipo *unimodale* o *bimodale* e in generale *plurimodale*.
La moda è di facile identificazione in un istogramma;

Moda

La *moda campionaria* (detta anche *valore normale* o *norma*) è il dato (o i dati) che si presenta(no) con frequenza massima. Di conseguenza si parla di *distribuzione dei dati* di tipo *unimodale* o *bimodale* e in generale *plurimodale*.
La moda è di facile identificazione in un istogramma; se i dati sono raggruppati in classi, allora la classe con massima frequenza è detta *classe modale*.

Indicatore di dispersione dei dati

Se la media campionaria $\overline{x}$ ha il ruolo di *baricentro* dei dati $\{x_1, \ldots, x_n\}$, allora la dispersione dei dati x_j da $\overline{x}$ può essere calcolata considerando gli *scarti*

$$d_i := x_i - \overline{x}$$

oppure

$$|d_i| = |x_i - \overline{x}| \geq 0.$$

Indicatore di dispersione dei dati

Se la media campionaria $\overline{x}$ ha il ruolo di *baricentro* dei dati $\{x_1, \ldots, x_n\}$, allora la dispersione dei dati x_j da $\overline{x}$ può essere calcolata considerando gli *scarti*

$$d_i := x_i - \overline{x}$$

oppure

$$|d_i| = |x_i - \overline{x}| \geq 0.$$

Tuttavia, risulta

$$\sum_{i=1}^{n} d_i = 0$$

Indicatore di dispersione dei dati

Se la media campionaria $\overline{x}$ ha il ruolo di *baricentro* dei dati $\{x_1, \ldots, x_n\}$, allora la dispersione dei dati x_j da $\overline{x}$ può essere calcolata considerando gli *scarti*

$$d_i := x_i - \overline{x}$$

oppure

$$|d_i| = |x_i - \overline{x}| \geq 0.$$

Tuttavia, risulta

$$\sum_{i=1}^{n} d_i = 0$$

in quanto

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \overline{x}) = n\overline{x} - n\overline{x} = 0.$$

Varianza campionaria o stimata dei dati

Si dice *varianza campionaria* o *varianza stimata* dei dati $\{x_1, \ldots, x_n\}$ di media campionaria $\overline{x}$ il numero (non negativo)

$$s^2 := \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \sum_{i=1}^{n} \frac{d_i^2}{n-1}$$

Varianza campionaria o stimata dei dati

Si dice *varianza campionaria* o *varianza stimata* dei dati $\{x_1, \ldots, x_n\}$ di media campionaria $\overline{x}$ il numero (non negativo)

$$s^2 := \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \sum_{i=1}^{n} \frac{d_i^2}{n-1}$$

Si dice infine che

$$s := \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

è la *deviazione standard stimata* o *scarto quadratico medio stimato* dei dati $\{x_1, \ldots, x_n\}$.

Alcune osservazioni

Si parla di *Varianza* (Var) e *Scarto quadratico* o *Deviazione standard* (σ) se nelle formule precedenti si considera n invece di $n-1$, vale a dire

$$Var := \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \sum_{i=1}^{n} \frac{d_i^2}{n}$$

$$\sigma = \sqrt{Var}$$

Alcune osservazioni

Si parla di *Varianza* (Var) e *Scarto quadratico* o *Deviazione standard* (σ) se nelle formule precedenti si considera n invece di $n-1$, vale a dire

$$Var := \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \sum_{i=1}^{n} \frac{d_i^2}{n}$$

$$\sigma = \sqrt{Var}$$

La differenza fra la varianza e la varianza campionaria o stimata (e quindi fra la deviazione standard e quella stimata) risulta rilevante solo per n piccoli; per n grandi tali differenze sono trascurabili.

Alcune osservazioni

Si parla di *Varianza* (Var) e *Scarto quadratico* o *Deviazione standard* (σ) se nelle formule precedenti si considera n invece di $n-1$, vale a dire

$$Var := \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \sum_{i=1}^{n} \frac{d_i^2}{n}$$

$$\sigma = \sqrt{Var}$$

La differenza fra la varianza e la varianza campionaria o stimata (e quindi fra la deviazione standard e quella stimata) risulta rilevante solo per n piccoli; per n grandi tali differenze sono trascurabili.
Si osservi inoltre che la deviazione standard e la deviazione standard stimata hanno la stessa unità di misura dei dati.

Una formula alternativa....

Da $(x_i - \overline{x})^2 = x_i^2 - 2x_i\overline{x} + \overline{x}^2$ e dal fatto che $n\overline{x} = x_1 + \ldots + x_n$ si ricava anche

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