

Metodi quantitativi per le scienze sociali

Servizio Sociale e Scienze Criminologiche (L-39)

Andrea Priulla

Statistiche Descrittive

Il codice della lezione

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 Descrivere la variabilità dei dati
- 4 Misure di posizione

Distribuzioni di frequenza

Dalla scorsa lezione sappiamo che:

- la **statistica descrittiva** sintetizza le informazioni raccolte in formati semplici limitando la perdita di informazioni e le distorsioni

Distribuzioni di frequenza

Dalla scorsa lezione sappiamo che:

- la **statistica descrittiva** sintetizza le informazioni raccolte in formati semplici limitando la perdita di informazioni e le distorsioni

Le variabili si dividono in:

- **quantitative**: numeri $\rightarrow$ discrete e continue
- **qualitative**: categorie $\rightarrow$ nominali e ordinali

Distribuzioni di frequenza

Dalla scorsa lezione sappiamo che:

- la **statistica descrittiva** sintetizza le informazioni raccolte in formati semplici limitando la perdita di informazioni e le distorsioni

Le variabili si dividono in:

- **quantitative**: numeri $\rightarrow$ discrete e continue
- **qualitative**: categorie $\rightarrow$ nominali e ordinali

Tra le metodologie di base della **statistica descrittiva** che possiamo usare in entrambi i casi troviamo:

- **tabelle**
- **grafici**

La lezione

1 Distribuzioni di frequenza

- Dati categoriali
- Dati quantitativi
- Distribuzioni di popolazioni e distribuzioni di dati campionari
- La forma della distribuzione

2 Descrivere il centro dei dati

3 Descrivere la variabilità dei dati

4 Misure di posizione

Dati categoriali: le frequenze

Nel caso dei **dati categoriali**:

- la rappresentazione **tabellare** prevede un elenco con due **colonne**:
 - le **categorie della variabile**
 - le **frequenze assolute**
- è possibile utilizzare anche le frequenze **relative** e **percentuali**

Dati categoriali: le frequenze

Nel caso dei **dati categoriali**:

- la rappresentazione **tabellare** prevede un elenco con due **colonne**:
 - le **categorie della variabile**
 - le **frequenze assolute**
- è possibile utilizzare anche le frequenze **relative** e **percentuali**

Cosa sono le **frequenze**?

- **assolute**: numero di osservazioni che ricadono nella categoria considerata
- **relative**: proporzione di osservazioni che ricadono nella categoria
- **percentuali**: percentuale di osservazioni che ricadono nella categoria

Dati categoriali: la proporzione

Frequenza relativa: **proporzione** di osservazioni che ricadono in una categoria:

- **proporzione:** **numero di osservazioni** in una categoria **diviso** il **numero totale**
 - è **sempre compresa** tra 0 e 1
 - es. su 100 student*, 30 sono fuori sede
 - **proporzione** $\rightarrow 30/100 = 0.3$
 - se tutt* fossero fuori sede $\rightarrow 100/100 = 1$
 - nessun* è fuori sede $\rightarrow 0/100 = 0$

Dati categoriali: la proporzione

Frequenza relativa: **proporzione** di osservazioni che ricadono in una categoria:

- **proporzione:** **numero di osservazioni** in una categoria **diviso** il **numero totale**
 - è **sempre compresa** tra 0 e 1
 - es. su 100 student*, 30 sono fuori sede
 - **proporzione** $\rightarrow 30/100 = 0.3$
 - se tutt* fossero fuori sede $\rightarrow 100/100 = 1$
 - nessun* è fuori sede $\rightarrow 0/100 = 0$

Frequenza percentuale: **proporzione** moltiplicata per 100:

- è **sempre compresa** tra 0 e 100
- **percentuale** $\rightarrow 30/100 \times 100 = 0.3 \times 100 = 30\%$
- se tutt* fossero fuori sede $\rightarrow 100/100 \times 100 = 100\%$
- nessun* è fuori sede $\rightarrow 0/100 \times 100 = 0\%$

Esempio

Riprendiamo l'esempio della scorsa lezione.

Nome	Età	Sesso	Tit. studio	Attività	Peso (kg)	Punteggio
Mario	32	M	Laurea	Occupato	72	65
Alessia	39	F	Laurea	Occupato	55	53
Giovanni	46	M	Diploma	Disoccupato	79	53
Francesco	28	M	Diploma	Studente	63	78
Claudia	51	F	Diploma	Casalinga	64	30
Flavia	60	F	Diploma	Casalinga	62	21
Fabio	35	M	Dottorato	Occupato	71	85
Elisa	38	F	Diploma	Occupata	64	45
Clara	41	F	Diploma	Casalinga	65	28

Esempio

Riprendiamo l'esempio della scorsa lezione.

Nome	Età	Sesso	Tit. studio	Attività	Peso (kg)	Punteggio
Mario	32	M	Laurea	Occupato	72	65
Alessia	39	F	Laurea	Occupato	55	53
Giovanni	46	M	Diploma	Disoccupato	79	53
Francesco	28	M	Diploma	Studente	63	78
Claudia	51	F	Diploma	Casalinga	64	30
Flavia	60	F	Diploma	Casalinga	62	21
Fabio	35	M	Dottorato	Occupato	71	85
Elisa	38	F	Diploma	Occupata	64	45
Clara	41	F	Diploma	Casalinga	65	28

Vogliamo valutare la **distribuzione** di tre variabili: Attività, età e titolo di studio

Dati categoriali: distribuzioni di frequenze

Viene chiamata **distribuzione di frequenze** una tabella composta da:

- una **lista** di tutte le categorie di una variabile
- un **numero** che rappresenta **quante volte** quella categoria viene osservata nei dati

Possiamo avere:

- **distribuzione di frequenze assolute** → a ciascuna categoria associo il corrispondente **conteggio**
- **distribuzione di frequenze relative** → a ciascuna categoria associo la **proporzione** o la **percentuale**

Distribuzioni di frequenza

Table: Notazione distribuzione di frequenze

X	Frequenze		
	Assoluta	Relativa	Percentuale
X_1	n_1	f_1	p_1
X_2	n_2	f_2	p_2
...	...	...	...
X_j	n_j	f_j	p_j
...	...	...	...
X_K	n_K	f_K	p_K
Totale	n	1	100

Date n unità statistiche di cui n_j presentano la j -esima modalità, si definisce:

- $f_j = n_j/n$ la frequenza **relativa**
- $p_j = f_j \times 100$ la frequenza **percentuale**

Table: Distribuzione frequenze delle variabile **attività**

Attività	Frequenze		
	Assoluta	Relativa	Percentuale
Occupato/a	4	0.44	44.4
Disoccupato/a	1	0.11	11.1
Casalingo/a	3	0.33	33.3
Studente	1	0.11	11.1
Totale	9	1	100

La tabella ci indica:

- **Attività**: le **categorie** della variabile **attività**
- **Frequenze**: le **frequenze** associate a ciascuna **categoria**
- Es. il 44.4% degli individui risulta occupato.

Table: Distribuzione frequenze delle variabile
Titolo di studio

Titolo di studio	Frequenze		
	Assoluta	Relativa	Percentuale
Diploma	6	0.66	66.6
Laurea	2	0.22	22.2
Dottorato	1	0.11	11.1
Totale	9	1	100

- La variabile titolo di studio è **ordinale**: le categorie vanno ordinate.

Nel caso di una variabile ordinale, ci può interessare valutare la frequenza con cui si presentano modalità di ordine **inferiori o uguale** a una certa modalità. Si calcolano quindi le **frequenze cumulate**

Table: Distribuzione frequenze **cumulate** della variabile **Titolo di studio**

Titolo di studio	Frequenze Cumulate		
	Assoluta	Relativa	Percentuale
Diploma	6	0.66	66.6
Laurea	8	0.88	88.8
Dottorato	9	1	100
Totale	9	1	100

La tabella ci indica:

- **Frequenze percentuali:** il 66% degli individui non supera il diploma.

Dati categoriali: grafici a barre

Un altro modo di rappresentare le **informazioni quantitative** su una **distribuzione di frequenza** è utilizzare il **grafico a barre**:

- ogni **categoria** è rappresentata da una **barra**
- l'**altezza** della barra indica la **frequenza relativa**
- le barre sono **separate nettamente** per indicare che:
 - la variabile è **categoriale**
 - si può utilizzare con variabile categoriali ordinate e non ordinate.

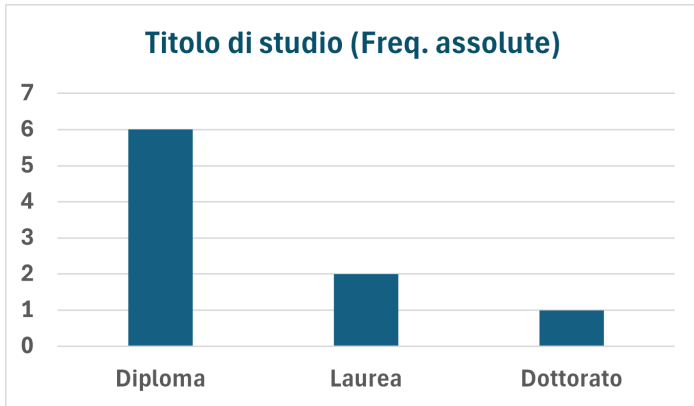


Figure: Grafico a barre della variabile titolo di studio

Dati categoriali: il grafico a torta

Un altro tipo di rappresentazione molto comune è il **grafico a torta**:

- il grafico si presenta come un **cerchio suddiviso in spicchi**
- ogni **categoria** è rappresentata da uno **spicchio**
- l'**ampiezza** dello spicchio indica la **frequenza relativa**

Criticità:

- utile se le categorie sono **poche** e presentano frequenze relative **molto diverse**
- **meno efficace** del grafico a barre quando le frequenze relative sono **simili**
- l'**ampiezza** dello spicchio non è direttamente quantificabile

La lezione

1 Distribuzioni di frequenza

- Dati categoriali
- **Dati quantitativi**
- Distribuzioni di popolazioni e distribuzioni di dati campionari
- La forma della distribuzione

2 Descrivere il centro dei dati

3 Descrivere la variabilità dei dati

4 Misure di posizione

Dati quantitativi: distribuzioni di frequenze

Anche nel caso dei **dati quantitativi** possiamo usare **tabelle** e **grafici**:

- le **tabelle** prevedono un elenco con due **colonne**:
 - gli **intervalli** in cui vengono divise le osservazioni (es. da 10 a 100 euro)
 - le **frequenze** associate a ogni **intervallo**
- è possibile utilizzare anche le frequenze **relative** e **percentuali**

Vediamo un **esempio** . . .

Esempio

Possiamo, anche in questo caso, utilizzare le **distribuzioni di frequenze** nel modo seguente:

Esempio

Possiamo, anche in questo caso, utilizzare le **distribuzioni di frequenze** nel modo seguente:

- **dividere** la scala di misurazione della variabile in un **insieme di intervalli**

Esempio

Possiamo, anche in questo caso, utilizzare le **distribuzioni di frequenze** nel modo seguente:

- **dividere** la scala di misurazione della variabile in un **insieme di intervalli**
- **contare** quante osservazioni sono presenti all'interno di ogni intervallo (**frequenze assolute!**)

Consideriamo ora la variabile **età**.

- Essendo una variabile **quantitativa**, conviene fare una suddivisione **in classi**. Consideriamo ad esempio 3 classi di età.

Nome	Età	Classe di età
Mario	32	28-40
Alessia	39	28-40
Giovanni	46	41-50
Francesco	28	28-40
Claudia	51	51-60
Flavia	60	51-60
Fabio	35	28-40
Elisa	38	28-40
Clara	41	41-50

Table: Distribuzione frequenze delle variabile **Classe di età**

Classe di età	Frequenze		
	Assoluta	Relativa	Percentuale
28-42	6	0.66	66.6
43-52	2	0.22	22.2
53-60	1	0.11	11.1
Totale	9	1	100

- La suddivisione in classi aiuta a **sintetizzare**.
- Si perde tuttavia informazione sulla **variabilità** all'interno della classe. Es. è possibile che, nella classe 28-40 ci siano 4 persone con la stessa età!
- Nel caso di classi vuote, conviene comunque riportarle nella tabella ... rimane sempre un'informazione!

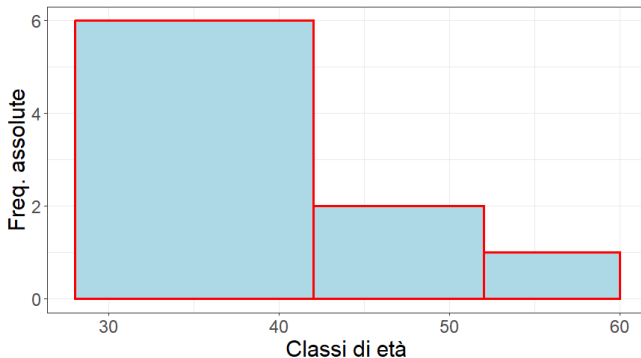
Dati quantitativi: l'istogramma

Queste informazioni possono essere rappresentate utilizzando
l'**istogramma**:

- simile al **grafico a barre**
- nell'asse delle ascisse si rappresentano **gli intervalli** in cui è stata suddivisa la variabile
- in questo caso le 'barre' **non sono separate**
- gli **intervalli**:
 - includono **tutti** i possibili valori della variabile
 - sono **mutualmente esclusivi**: ogni valore è incluso in **un solo intervallo**
- l'altezza della barra indica la **frequenza relativa**

Dati quantitativi: l'istogramma

- La **larghezza** della barra **dipende dall'ampiezza** della classe.



- La classe 28-42 ha ampiezza 14; La classe 43-52 ha ampiezza 9; La classe 53-60 ha ampiezza 7.

Istogrammi: la scelta del numero di intervalli

Come scegliamo il **numero di intervalli**?

Istogrammi: la scelta del numero di intervalli

Come scegliamo il **numero di intervalli**?

- **intervalli piccoli:**

- si mantengono più dettagli, il numero di barre cresce
- più difficile da interpretare
- potrebbe essere complesso risalire alla **forma della distribuzione**

Istogrammi: la scelta del numero di intervalli

Come scegliamo il **numero di intervalli**?

- **intervalli piccoli:**

- si mantengono più dettagli, il numero di barre cresce
- più difficile da interpretare
- potrebbe essere complesso risalire alla **forma della distribuzione**

- **intervalli grandi:**

- meno dettagli, il numero di barre diminuisce
- più facile da interpretare
- potremmo perdere le **informazioni** che ci servono

Istogrammi: la scelta del numero di intervalli

Come scegliamo il **numero di intervalli**?

- **intervalli piccoli:**

- si mantengono più dettagli, il numero di barre cresce
- più difficile da interpretare
- potrebbe essere complesso risalire alla **forma della distribuzione**

- **intervalli grandi:**

- meno dettagli, il numero di barre diminuisce
- più facile da interpretare
- potremmo perdere le **informazioni** che ci servono

- **principio generale:** includo due osservazioni nello stesso intervallo se le considero simili dal punto di vista pratico

- es. divido tra 0-100 euro e 100-200 euro perché penso che siano due fasce di prezzo diverse che attirano consumatori diversi

La lezione

1 Distribuzioni di frequenza

- Dati categoriali
- Dati quantitativi
- **Distribuzioni di popolazioni e distribuzioni di dati campionari**
- La forma della distribuzione

2 Descrivere il centro dei dati

3 Descrivere la variabilità dei dati

4 Misure di posizione

Distribuzioni di popolazioni e distribuzioni di dati campionari

A seconda della tipologia di dati che abbiamo a disposizione possiamo avere:

- **distribuzioni di dati di popolazione** → utilizzo i dati su **tutta** la popolazione
- **distribuzioni di dati campionari** → utilizzo i dati del **campione**

La distribuzione di dati campionari **approssima** quella di popolazione:

- quando la dimensione campionaria **aumenta**:
 - le **proporzioni campionarie** in ciascun intervallo si approssimano alle vere **proporzioni della popolazione**
 - la distribuzione dei dati campionari diventa **più simile** alla distribuzione di popolazione

Le distribuzioni di dati: esempio

Ad esempio, consideriamo il caso seguente:

- analizziamo le osservazioni di una **variabile continua**
- aumentiamo indefinitamente l'**ampiezza campionaria**
- aumentiamo il numero di **intervalli** in cui abbiamo suddiviso la distribuzione
 - gli intervalli quindi diventano più piccoli

Le distribuzioni di dati: esempio

Ad esempio, consideriamo il caso seguente:

- analizziamo le osservazioni di una **variabile continua**
- aumentiamo indefinitamente l'**ampiezza campionaria**
- aumentiamo il numero di **intervalli** in cui abbiamo suddiviso la distribuzione
 - gli intervalli quindi diventano più piccoli

Cosa succede alla **forma dell'istogramma**?

- più aumentiamo l'ampiezza campionaria e il numero di intervalli più la distribuzione viene approssimata da una curva liscia
- con poche osservazioni o pochi intervalli, invece, abbiamo dei 'gradoni'

Le distribuzioni di dati: esempio

Ad esempio, consideriamo il caso seguente:

- analizziamo le osservazioni di una **variabile continua**
- aumentiamo indefinitamente l'**ampiezza campionaria**
- aumentiamo il numero di **intervalli** in cui abbiamo suddiviso la distribuzione
 - gli intervalli quindi diventano più piccoli

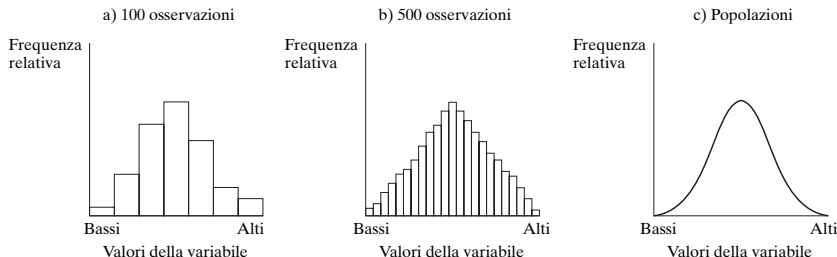
Cosa succede alla **forma dell'istogramma**?

- più aumentiamo l'ampiezza campionaria e il numero di intervalli più la distribuzione viene approssimata da una curva liscia
- con poche osservazioni o pochi intervalli, invece, abbiamo dei 'gradoni'

Vediamo un esempio...

Le distribuzioni di dati: esempio

La figura mostra due istogrammi di dati campionari, uno costruito su una numerosità di 100 e l'altro di 500 osservazioni:



- il terzo grafico indica la distribuzione nella popolazione
- anche se abbiamo una variabile discreta la sua distribuzione a livello di popolazione può essere adeguatamente approssimata da una curva liscia

La lezione

1 Distribuzioni di frequenza

- Dati categoriali
- Dati quantitativi
- Distribuzioni di popolazioni e distribuzioni di dati campionari
- **La forma della distribuzione**

2 Descrivere il centro dei dati

3 Descrivere la variabilità dei dati

4 Misure di posizione

La forma della distribuzione

Perché ci **interessa** la **forma della distribuzione**?

- ci permette di avere **informazioni** sulle caratteristiche della distribuzione

La forma della distribuzione

Perché ci **interessa** la **forma della distribuzione**?

- ci permette di avere **informazioni** sulle caratteristiche della distribuzione

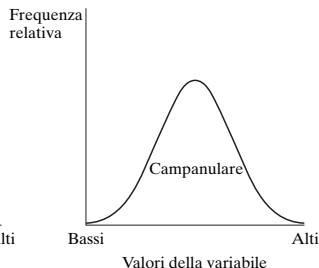
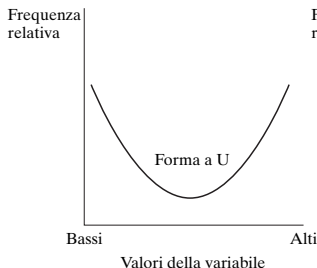
Prendiamo ad esempio due forme:

La forma della distribuzione

Perché ci **interessa** la **forma della distribuzione**?

- ci permette di avere **informazioni** sulle caratteristiche della distribuzione

Prendiamo ad esempio due forme:



La forma della distribuzione: esempio

Confrontiamo le due distribuzioni:

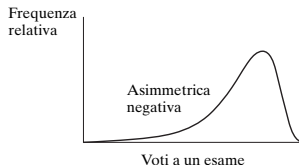
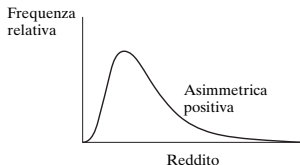
- **distribuzione a U**: le frequenze **più elevate** sono riferite ai valori **più bassi** e **più alti** assunti dalle osservazioni
 - indica una **polarizzazione** della variabile rispetto a due insieme di soggetti
- **forma campanulare**: le frequenze **più elevate** sono riferite ai valori **centrali**
 - indica che la **maggior parte** dei soggetti tende ad assumere valori prossimi al valore centrale

In entrambi i casi le distribuzioni sono **simmetriche**:

- il lato della distribuzione a sinistra del valore centrale è l'immagine speculare del lato a destra dello stesso valore

La forma della distribuzione: esempio

La **maggior parte** delle distribuzioni che vengono osservate **non sono** simmetriche:



- le parti della figura in cui le frequenze sono ridotte sono chiamate **code della distribuzione**
- le distribuzioni asimmetriche possono essere di due tipi:
 - **asimmetriche positive**: una lunga coda a destra
 - **asimmetriche negative**: una lunga coda a sinistra

E' importante?

- **reddito**: un **numero elevato** di persone con **redditi bassi** e un **numero ridotto** di persone con **redditi elevati**

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
 - La media
 - La mediana
 - La moda
- 3 Descrivere la variabilità dei dati
- 4 Misure di posizione

Descrivere il centro dei dati

Le **tabelle** e i **grafici** ci danno informazioni su:

- come si distribuisce una variabile?
- che forma ha la distribuzione di frequenze?
- che tipo di asimmetria caratterizza la distribuzione di frequenze?

Descrivere il centro dei dati

Le **tabelle** e i **grafici** ci danno informazioni su:

- come si distribuisce una variabile?
- che forma ha la distribuzione di frequenze?
- che tipo di asimmetria caratterizza la distribuzione di frequenze?

Altri tipi di **statistiche descrittive** ci permettono di:

- sintetizzare le informazioni sulla variabile tramite **sintesi numeriche**
- confrontare diversi gruppi
 - es. le famiglie nel sud Italia fanno più figli di quelle al nord
 - es. il tasso di criminalità in USA è maggiore di quello in Canada

Descrivere il centro dei dati

Le **tabelle** e i **grafici** ci danno informazioni su:

- come si distribuisce una variabile?
- che forma ha la distribuzione di frequenze?
- che tipo di asimmetria caratterizza la distribuzione di frequenze?

Altri tipi di **statistiche descrittive** ci permettono di:

- sintetizzare le informazioni sulla variabile tramite **sintesi numeriche**
- confrontare diversi gruppi
 - es. le famiglie nel sud Italia fanno più figli di quelle al nord
 - es. il tasso di criminalità in USA è maggiore di quello in Canada

Queste statistiche si chiamano **misure di tendenza centrale**:

- noi vedremo: media, mediana e moda

L'operatore sommatoria

Dati quattro valori (x_1, x_2, x_3, x_4) , l'**operatore sommatoria** ($\sum$) esprime la **somma dei $k = 4$ elementi**

$$x_1 + x_2 + x_3 + x_4 = \sum_{k=1}^4 x_k$$

Ad esempio, in una distribuzione di frequenza **per modalità**:

$$\sum_{k=1}^K n_k = N$$

- n_k : frequenza assoluta della k -esima modalità della variabile X
- K : numero di modalità di cui sommare le rispettive n_k osservazioni
- N : totale delle osservazioni

L'operatore sommatoria

L'operatore sommatoria $\sum$ gode delle seguenti proprietà

- $\sum_{k=1}^K (x_k + y_k) = \sum_{k=1}^K x_k + \sum_{k=1}^K y_k$
- $\sum_{k=1}^K (x_k + c) = \sum_{k=1}^K x_k + \sum_{k=1}^K c$
- $\sum_{k=1}^K (x_k * c) = c * \sum_{k=1}^K x_k$
- $\sum_{k=1}^K c = Kc$

L'operatore sommatoria

Esempio:

- $x_1 = 4; x_2 = -1; x_3 = 2, x_4 = 7$
- $K = 4$

L'operatore sommatoria

Esempio:

- $x_1 = 4; x_2 = -1; x_3 = 2, x_4 = 7$
- $K = 4$

a. $\sum_{k=1}^4 x_i = x_1 + x_2 + x_3 + x_4 = 4 + (-1) + 2 + 7 = 12$

b. $\sum_{k=1}^4 x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 4^2 + (-1)^2 + 2^2 + 7^2 = 70$

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
 - La media
 - La mediana
 - La moda
- 3 Descrivere la variabilità dei dati
- 4 Misure di posizione

La media

Le **misure di tendenza centrale**:

- ci aiutano a descrivere la **tipicità della distribuzione**
 - cioè quali valori **caratterizzano tipicamente** la distribuzione che stiamo considerando
- ci permettono, quindi, di descrivere **il centro** di una distribuzione di frequenze di una **variabile quantitativa**

Tra le **misure di tendenza centrale** troviamo **la media**:

- **la media**: somma dei valori assunti dalle osservazioni divisa per il totale delle osservazioni

Partiamo da un esempio sul **tasso di attività femminile**

Esempio: tasso di attività femminile

Table: Tasso di attività femminile in Sud America e Est Europa

Sud America		Est Europa	
Nazione	Occupazione	Nazione	Occupazione
Argentina	48	Rep. Ceca	83
Bolivia	58	Estonia	82
Brasile	52	Ungheria	72
Cile	50	Lettonia	80
Colombia	62	Lituania	80
Ecuador	40	Polonia	81
Guyana	51	Rep. Slovacca	84
Paraguay	44	Slovenia	81
Peru	45		
Uruguay	68		
Venezuela	55		

Occupazione femminile espressa come % dell'occupazione maschile.

Fonte: *H. Dev. Report 2005*, UN Development Programme

Esempio: tasso di attività femminile

La tabella ci mostra il tasso di occupazione femminile come percentuale dell'occupazione maschile:

- Es: in Argentina il numero di donne nella forza lavoro era il 48% del numero dei uomini, per ogni 100 lavoratori ci sono 48 lavoratrici

Calcoliamo **la media** per l'Est Europa:

- prima faccio la **somma** dei valori osservati:

$$83 + 82 + 72 + 80 + 80 + 81 + 84 + 81 = 643$$

- poi **divido** per il numero delle osservazioni (il numero di stati):

$$\text{Est Europa} \rightarrow 643/8 = 80.4$$

$$\text{Sud America} \rightarrow 573/11 = 52.1$$

- possiamo concludere quindi che l'occupazione femminile in Est Europa era considerevolmente più alta di quella in Sud America

La media: la notazione

Un po' di notazione generale:

- la numerosità campionaria è indicata con n
- per una **generica variabile** y , le osservazioni possono essere indicate come:

$$y_1, y_2, y_3, \dots, y_n$$

- dove y_1 è la prima osservazione mentre y_n è l' n -esima osservazione
 - con i dati dell'Est Europa: $y_1 = 83, y_2 = 82, \dots, y_n = 81$

Utilizzando questa **notazione** la **media campionaria** viene indicata come:

$$\bar{y}$$

dove $\bar{y}$ viene letto come "y-barrato"

- la barra viene usata per indicare la **media campionaria**. Quindi $\bar{x}$ è la media campionaria della variabile x

La media: la notazione

Utilizzando la notazione precedente possiamo scrivere:

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

Per scrivere una formula più compatta possiamo utilizzare il simbolo Σ :

- Σ indica una **sommatoria**, cioè una **somma**
- $\Sigma y_i \rightarrow$ somma delle osservazioni y_i (i indica la singola osservazione)

Nell'esempio sull'Est Europa:

$$\Sigma y_i = y_1 + y_2 + \dots + y_8 = 83 + 82 + \dots + 81 = 643$$

Quindi possiamo scrivere la media campionaria come:

$$\bar{y} = \frac{\Sigma y_i}{n}$$

Le proprietà della media

La media è caratterizzata da queste **proprietà**:

- il calcolo della media è **appropriato solo** per variabili **quantitative**:
 - cosa indica il "sesso medio" o il "frutto medio"?

Le proprietà della media

La media è caratterizzata da queste **proprietà**:

- il calcolo della media è **appropriato solo** per variabili **quantitative**:
 - cosa indica il "sesso medio" o il "frutto medio"?
- la media può essere influenzata notevolmente dagli **outlier**
 - **outlier** → "stanno fuori" → **valore anomalo**
 - **outlier** → osservazioni che assumono valori **molto al di sotto** o **al di sopra** di quello assunto dalla maggioranza delle osservazioni
 - esempio nella prossima slide!

Le proprietà della media

La media è caratterizzata da queste **proprietà**:

- il calcolo della media è **appropriato solo** per variabili **quantitative**:
 - cosa indica il "sesso medio" o il "frutto medio"?
- la media può essere influenzata notevolmente dagli **outlier**
 - **outlier** → "stanno fuori" → **valore anomalo**
 - **outlier** → osservazioni che assumono valori **molto al di sotto** o **al di sopra** di quello assunto dalla maggioranza delle osservazioni
 - esempio nella prossima slide!
- la media viene "**tirata**" verso le code
 - se una distribuzione è asimmetrica **positiva** la media avrà un valore spostato più a **destra** nel grafico
 - se una distribuzione è asimmetrica **negativa** la media avrà un valore spostato più a **sinistra** nel grafico

Esempio: effetti degli outlier nel calcolo del reddito

Immaginiamo che la gelateria *da Ciccio* abbia 7 dipendenti con un **reddito medio annuo** pari a 42900\$

- Di fatto però il reddito dei dipendenti è il seguente:

10000\$, 11300\$, 11500\$, 12000\$, 12500\$, 15000\$, **228000\$**

- il reddito **228000\$** è chiaramente un'osservazione **outlier**
- il reddito medio annuo calcolato **senza** questa osservazione, infatti, è pari a **12050\$**
- un valore molto **distante** dalla media di 42900\$ ottenuto in precedenza
- l'outlier ci farebbe pensare che nella gelateria tutti siano abbastanza ricchi, mentre invece hanno uno stipendio abbastanza ridotto

Esempio: effetti degli outlier nel calcolo del reddito

L'esempio precedente ci mostra che:

- la media può essere un valore che non rappresenta la **tipicità** delle osservazioni
- infatti la **maggior parte** delle osservazioni era minore di 15000\$ ma la media era nettamente più elevata (42900\$)
- in altre parole: il reddito era **tipicamente** inferiore a 15000\$!

Esempio: effetti degli outlier nel calcolo del reddito

L'esempio precedente ci mostra che:

- la media può essere un valore che non rappresenta la **tipicità** delle osservazioni
- infatti la **maggior parte** delle osservazioni era minore di 15000\$ ma la media era nettamente più elevata (42900\$)
- in altre parole: il reddito era **tipicamente** inferiore a 15000\$!

Questo accade quando:

- **campioni piccoli**: poche osservazioni e quindi anche una osservazione anomala influenza fortemente i risultati
- in presenza di distribuzioni **fortemente asimmetriche**
 - la media tende a spostarsi verso la coda lunga → in questo caso verso il valore anomalo!

La media ponderata

La **media ponderata** è una media **pesata**:

- utilizzata per calcolare la media di osservazioni con **pesi diversi**
- utilizzata per fare la media di **due o più** insiemi di dati

La media ponderata

La **media ponderata** è una media **pesata**:

- utilizzata per calcolare la media di osservazioni con **pesi diversi**
- utilizzata per fare la media di **due o più** insiemi di dati

Ipotezziamo di avere due insiemi di dati:

- **insieme 1**: n_1 osservazioni con media campionaria $\bar{y}_1$
- **insieme 2**: n_2 osservazioni con media campionaria $\bar{y}_2$

La media ponderata

La **media ponderata** è una media **pesata**:

- utilizzata per calcolare la media di osservazioni con **pesi diversi**
- utilizzata per fare la media di **due o più** insiemi di dati

Ipotizziamo di avere due insiemi di dati:

- **insieme 1**: n_1 osservazioni con media campionaria $\bar{y}_1$
- **insieme 2**: n_2 osservazioni con media campionaria $\bar{y}_2$

La media il **nuovo insieme** che unisce 1 e 2 è data da:

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2}$$

dove:

- il numeratore è la somma di tutte le osservazioni del nuovo insieme
 - infatti $n\bar{y} = \sum y_i$
- il denominatore è l'ampiezza campionaria totale ($n = n_1 + n_2$)

La media ponderata

Consideriamo i dati sull'occupazione femminile:

- **Sud America:** $n_1 = 11$; $\bar{y}_1 = 52.1$
- **Est Europa:** $n_2 = 8$; $\bar{y}_2 = 80.4$

La media considerando tutte le nazioni sarà:

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \frac{11 \times 52.1 + 8 \times 80.4}{11 + 8} = \frac{573 + 643}{19} = \frac{1216}{19} \simeq 64$$

La media ponderata

Consideriamo i dati sull'occupazione femminile:

- **Sud America:** $n_1 = 11$; $\bar{y}_1 = 52.1$
- **Est Europa:** $n_2 = 8$; $\bar{y}_2 = 80.4$

La media considerando tutte le nazioni sarà:

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \frac{11 \times 52.1 + 8 \times 80.4}{11 + 8} = \frac{573 + 643}{19} = \frac{1216}{19} \simeq 64$$

- **Metodi quantitativi per le scienze sociali:** voto 28; 6 CFU
- **Psicologia del lavoro e delle organizzazioni:** voto 22; 9 CFU

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n_1 + n_2} = \frac{6 \times 28 + 9 \times 22}{6 + 9} = \frac{168 + 198}{15} = \frac{366}{15} \simeq 24.4$$

Morale: la media è più vicina al gruppo più "pesante"

La lezione

- 1 Distribuzioni di frequenza
- 2 **Descrivere il centro dei dati**
 - La media
 - **La mediana**
 - La moda
- 3 Descrivere la variabilità dei dati
- 4 Misure di posizione

La mediana

Tra le misure di **tendenza centrale** troviamo la **mediana**:

- misura il centro dei dati come la media
- a seconda del caso potrebbe essere **più appropriata** della media

La mediana

Tra le misure di **tendenza centrale** troviamo la **mediana**:

- misura il centro dei dati come la media
- a seconda del caso potrebbe essere **più appropriata** della media

Cosa è la **mediana**?

- è il valore della variabile assunto dall'**osservazione centrale** del campione **ordinato in maniera crescente**
- divide in **due parti** il campione ordinato in maniera crescente; ciascuna parte contiene lo **stesso** numero di osservazioni (il 50%)

La mediana

Tra le misure di **tendenza centrale** troviamo la **mediana**:

- misura il centro dei dati come la media
- a seconda del caso potrebbe essere **più appropriata** della media

Cosa è la **mediana**?

- è il valore della variabile assunto dall'**osservazione centrale** del campione **ordinato in maniera crescente**
- divide in **due parti** il campione ordinato in maniera crescente; ciascuna parte contiene lo **stesso** numero di osservazioni (il 50%)

Due casi:

- **numerosità dispari**: esiste solo un valore centrale
 - es. 1, 2, 3 $\rightarrow$ la mediana è 2
- **numerosità pari**: due valori centrali, la mediana è la media tra i due
 - es 1, 2, 3, 4 $\rightarrow$ 2 e 3 sono centrali $\rightarrow$ la mediana è 2.5

La mediana: esempi

Consideriamo il caso del reddito dei dipendenti della gelateria:

12500\$, 10000\$, 11500\$, 228000\$, 11300\$, 15000\$, 12000\$

- Qual è il **valore mediano**?

La mediana: esempi

Consideriamo il caso del reddito dei dipendenti della gelateria:

12500\$, 10000\$, 11500\$, 228000\$, 11300\$, 15000\$, 12000\$

- Qual è il **valore mediano**?
- ... prima vanno ordinate!

10000\$, 11300\$, 11500\$, 12000\$, 12500\$, 15000\$, 228000\$

- abbiamo 7 osservazioni **ordinate in modo crescente**

La mediana: esempi

Consideriamo il caso del reddito dei dipendenti della gelateria:

12500\$, 10000\$, 11500\$, 228000\$, 11300\$, 15000\$, 12000\$

- Qual è il **valore mediano**?
- ... prima vanno ordinate!

10000\$, 11300\$, 11500\$, 12000\$, 12500\$, 15000\$, 228000\$

- abbiamo 7 osservazioni **ordinate in modo crescente**
- la mediana è il valore centrale → **12000\$**
- la mediana presenta quindi un valore **più tipico** della media in questo caso (42900\$)

La mediana

Tassi di attività femminile:

72, 80, 80, 81, 81, 82, 83, 84

- abbiamo 8 osservazioni, **la numerosità è pari**
- i valori centrali sono 81 e 81
- la mediana è $(81 + 81)/2 = 81$

La mediana

Esiste una **formula**?

- l'osservazione centrale si trova nella **posizione** $(n + 1)/2$
- se $n = 7$ allora la posizione mediana è $(7 + 1)/2 = 4$
- se $n = 8$ allora la posizione mediana è $(8 + 1)/2 = 4.5$
 - quindi sarà **la media** tra il valore delle osservazioni in posizione 4 e 5

La mediana

Esiste una **formula**?

- l'osservazione centrale si trova nella **posizione** $(n + 1)/2$
- se $n = 7$ allora la posizione mediana è $(7 + 1)/2 = 4$
- se $n = 8$ allora la posizione mediana è $(8 + 1)/2 = 4.5$
 - quindi sarà **la media** tra il valore delle osservazioni in posizione 4 e 5

Attenzione, errori molto comuni:

- occorre sempre **ordinare le osservazioni**
- la formula non ci da il valore della mediana ma la sua **posizione**

La mediana, i dati ordinali

La **mediana** può essere utilizzata anche per variabili **ordinali**:

- la tabella mostra la distribuzione della variabile livello di istruzione rilevata nell'indagine Aspetti della Vita Quotidiana

T. studio	Frequenza	%
Senza titolo	1742	9.06%
Scuola elementare	3538	18.41%
Scuola media	5650	29.40%
Diploma	6216	32.34%
Laurea	2010	10.46%
Dottorato ricerca	61	0.33%

(Fonte: *Indagine Multiscopo, Aspetti della vita quotidiana*, anno 2013, Istat)

La mediana, i dati ordinali

Come si procede:

- l'ampiezza del collettivo è $n = 19217$

La mediana, i dati ordinali

Come si procede:

- l'ampiezza del collettivo è $n = 19217$
- la posizione mediana è data da $(n + 1)/2 = 19218/2 = 9609$

La mediana, i dati ordinali

Come si procede:

- l'ampiezza del collettivo è $n = 19217$
- la posizione mediana è data da $(n + 1)/2 = 19218/2 = 9609$
- la prima categoria occupa le posizioni da 0 a 1742

La mediana, i dati ordinali

Come si procede:

- l'ampiezza del collettivo è $n = 19217$
- la posizione mediana è data da $(n + 1)/2 = 19218/2 = 9609$
- la prima categoria occupa le posizioni da 0 a 1742
- la seconda occupa le posizioni da 1742 a $1742 + 3538 = 5280$

La mediana, i dati ordinali

Come si procede:

- l'ampiezza del collettivo è $n = 19217$
- la posizione mediana è data da $(n + 1)/2 = 19218/2 = 9609$
- la prima categoria occupa le posizioni da 0 a 1742
- la seconda occupa le posizioni da 1742 a $1742 + 3538 = 5280$
- la terza occupa le posizioni da 5280 a $5280 + 5650 = 10930$

La mediana, i dati ordinali

Come si procede:

- l'ampiezza del collettivo è $n = 19217$
- la posizione mediana è data da $(n + 1)/2 = 19218/2 = 9609$
- la prima categoria occupa le posizioni da 0 a 1742
- la seconda occupa le posizioni da 1742 a $1742 + 3538 = 5280$
- la terza occupa le posizioni da 5280 a $5280 + 5650 = 10930$

La posizione 9609 la troviamo nella **terza categoria**, quindi la mediana è la categoria **scuola media**

Le proprietà della mediana

La mediana è caratterizzata da queste **proprietà**:

- per il suo calcolo richiede che le osservazioni siano **ordinabili**:
 - va bene per variabili **quantitative** o **ordinali**
 - **non** va bene per variabili in scala **nominale**

Le proprietà della mediana

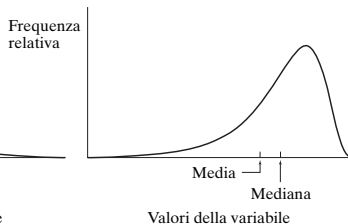
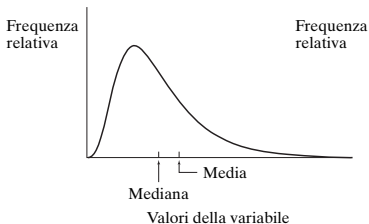
La mediana è caratterizzata da queste **proprietà**:

- per il suo calcolo richiede che le osservazioni siano **ordinabili**:
 - va bene per variabili **quantitative** o **ordinali**
 - **non** va bene per variabili in scala **nominale**
- la mediana è uguale alla media se la distribuzione è **simmetrica**
 - considerate i numeri 4, 5, 7, 9, 10, la mediana è la media sono pari a 7

Le proprietà della mediana

La mediana è caratterizzata da queste **proprietà**:

- per il suo calcolo richiede che le osservazioni siano **ordinabili**:
 - va bene per variabili **quantitative** o **ordinali**
 - **non** va bene per variabili in scala **nominale**
- la mediana è uguale alla media se la distribuzione è **simmetrica**
 - considerate i numeri 4, 5, 7, 9, 10, la mediana è la media sono pari a 7
- se è **asimmetrica** la media è **spostata** verso la coda più lunga:



- la mediana non è sensibile alla **distanza** tra i valori della distribuzione e la loro media:
 - tiene conto solo delle caratteristiche ordinali dei dati
 - considerate i seguenti gruppi, hanno tutti mediana pari a 10!

Gruppo 1: 8, 9, 10, 11, 12

Gruppo 2: 8, 9, 10, 11, 100

Gruppo 3: 0, 9, 10, 10, 10

Gruppo 4: 8, 9, 10, 100, 100

- la mediana non è sensibile alla **distanza** tra i valori della distribuzione e la loro media:
 - tiene conto solo delle caratteristiche ordinali dei dati
 - considerate i seguenti gruppi, hanno tutti mediana pari a 10!

Gruppo 1: 8, 9, 10, 11, 12

Gruppo 2: 8, 9, 10, 11, 100

Gruppo 3: 0, 9, 10, 10, 10

Gruppo 4: 8, 9, 10, 100, 100

- la mediana **non è influenzata** dagli outlier, questi tre gruppi hanno la stessa mediana:

Gruppo 1: 8, 9, 10, 11, 12

Gruppo 2: -80000, 9, 10, 11, 100

Gruppo 3: 0, 9, 10, 10, 1000000

Confronto tra media e mediana

Confrontiamo la **media** e la **mediana**:

- la **mediana** risulta più appropriata della **media**
 - distribuzioni fortemente asimmetriche
 - presenza di outlier
 - variabili categoriali ordinali

Confronto tra media e mediana

Confrontiamo la **media** e la **mediana**:

- la **mediana** risulta più appropriata della **media**
 - distribuzioni fortemente asimmetriche
 - presenza di outlier
 - variabili categoriali ordinali
- la **mediana** presenta però alcuni svantaggi:
 - per dati discreti, che assumono un numero limitato di valori, la mediana può essere **uguale** anche per distribuzioni **molto diverse**
 - prendiamo un caso come esempio

La tabella riassume le risposte date da 365 persone alla domanda 'Quanti partner hai avuto negli ultimi 12 mesi?'

Risposta	Frequenza	Percentuale	% Cumulata
0	102	27.9	27.9
1	233	63.8	91.7
2	18	4.9	96.6
3	9	2.5	99.1
4	2	0.6	99.7
5	1	0.3	100

Dalla tabella sappiamo che:

- la **mediana** è pari a 1 (il 50% delle osservazioni ha avuto 1 o 0)

La tabella riassume le risposte date da 365 persone alla domanda 'Quanti partner hai avuto negli ultimi 12 mesi?'

Risposta	Frequenza	Percentuale	% Cumulata
0	102	27.9	27.9
1	233	63.8	91.7
2	18	4.9	96.6
3	9	2.5	99.1
4	2	0.6	99.7
5	1	0.3	100

Dalla tabella sappiamo che:

- la **mediana** è pari a 1 (il 50% delle osservazioni ha avuto 1 o 0)
- la **media** campionaria è pari a :

$$\bar{y} = \frac{\sum y_i}{n} = \frac{102(0) + 233(1) + 18(2) + 9(3) + 2(4) + 1(5)}{365} = 0.85$$

Spostiamo 102 osservazioni da 0 a 5:

Risposta	Frequenza	Percentuale	% Cumulata
0	0	0	0
1	233	63.8	63.8
2	18	4.9	68.7
3	9	2.5	71.2
4	2	0.6	71.8
5	103	28.2	100

Dalla tabella sappiamo che:

Spostiamo 102 osservazioni da 0 a 5:

Risposta	Frequenza	Percentuale	% Cumulata
0	0	0	0
1	233	63.8	63.8
2	18	4.9	68.7
3	9	2.5	71.2
4	2	0.6	71.8
5	103	28.2	100

Dalla tabella sappiamo che:

- la **mediana** è pari a 1 (il 50% delle osservazioni ha avuto 1 o 0)

Spostiamo 102 osservazioni da 0 a 5:

Risposta	Frequenza	Percentuale	% Cumulata
0	0	0	0
1	233	63.8	63.8
2	18	4.9	68.7
3	9	2.5	71.2
4	2	0.6	71.8
5	103	28.2	100

Dalla tabella sappiamo che:

- la **mediana** è pari a 1 (il 50% delle osservazioni ha avuto 1 o 0)
- la **media** invece sale fino a **2.2**!
- la **media**, infatti, tiene conto anche del **valore numerico** delle osservazioni e non solo del loro ordinamento

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
 - La media
 - La mediana
 - La moda
- 3 Descrivere la variabilità dei dati
- 4 Misure di posizione

La moda

L'ultima misura di tendenza centrale che vediamo è la **moda**:

- **Moda**: il valore osservato nella distribuzione con maggiore frequenza

Quando si utilizza?

- se occorre sintetizzare distribuzioni **estremamente** discrete come nel caso delle variabili **categoriali**
 - se 10 scelgono arance e 2 scelgono pesche allora la moda sono le arance
- **possiamo** applicare la moda a dati misurati su **qualunque scala**
 - sia per variabili quantitative che categoriali!

La moda

L'ultima misura di tendenza centrale che vediamo è la **moda**:

- **Moda**: il valore osservato nella distribuzione con maggiore frequenza

Quando si utilizza?

- se occorre sintetizzare distribuzioni **estremamente** discrete come nel caso delle variabili **categoriali**
 - se 10 scelgono arance e 2 scelgono pesche allora la moda sono le arance
- **possiamo** applicare la moda a dati misurati su **qualunque scala**
 - sia per variabili quantitative che categoriali!

Dettagli:

- le distribuzioni si dicono **bi-modali** quando si osservano due picchi nella distribuzione
 - quindi anche **tri-modali** ecc
- la media, la mediana e la moda sono **identiche** se la distribuzione è **unimodale** e **simmetrica**
 - le distribuzioni campanulari hanno questa caratteristica

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 **Descrivere la variabilità dei dati**
 - Cosa si intende con variabilità?
 - Il campo di variazione (range)
 - La deviazione standard
- 4 Misure di posizione

Descrivere la variabilità dei dati

Le misure di tendenza centrale:

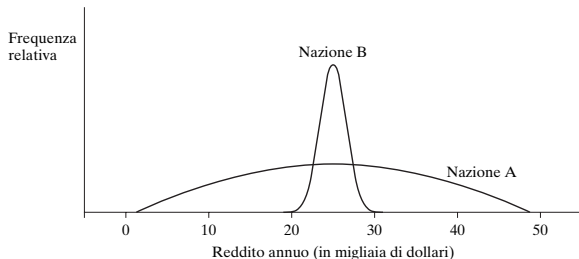
- ci danno **informazioni** sul valore **tipico** di una distribuzione
- **non** ci danno **informazioni** sulla **dispersione** delle osservazioni intorno al valore tipico

Descrivere la variabilità dei dati

Le misure di tendenza centrale:

- ci danno **informazioni** sul valore **tipico** di una distribuzione
- **non** ci danno **informazioni** sulla **dispersione** delle osservazioni intorno al valore tipico

Per esempio, i paesi A e B che hanno lo **stesso** reddito annuo **medio** (25000\$) ma presentano queste due **distribuzioni**:



I redditi sono molto più **variabili** in A, come lo **quantifichiamo**?

Cosa si intende con variabilità?

Poiché la statistica si occupa di analizzare come si manifesta un determinato fenomeno, bisogna porsi il problema della **variabilità**

- La **variabilità** di una distribuzione esprime la tendenza delle unità di un collettivo ad assumere diverse modalità del carattere.

Cosa si intende con variabilità?

Poiché la statistica si occupa di analizzare come si manifesta un determinato fenomeno, bisogna porsi il problema della **variabilità**

- La **variabilità** di una distribuzione esprime la tendenza delle unità di un collettivo ad assumere diverse modalità del carattere.

Per misurare la variabilità si utilizzano degli **indici** che sintetizzano la diversità delle modalità assunte **rispetto ad un valore di riferimento** (la media, il primo quartile, etc..)

- La variabilità è **nulla** nel caso in cui tutte le unità presentino la stessa modalità (es. tutti hanno la stessa età)
- Aumenta all'aumentare della diversità tra le modalità assunte dalle varie unità.

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 **Descrivere la variabilità dei dati**
 - Cosa si intende con variabilità?
 - **Il campo di variazione (range)**
 - La deviazione standard
- 4 Misure di posizione

Il campo di variazione (range)

La prima **misura di variabilità** che vediamo è:

- **Campo di variazione (range)**: differenza tra il valore più **alto** e il valore più **basso** della distribuzione

Il campo di variazione (range)

La prima **misura di variabilità** che vediamo è:

- **Campo di variazione (range)**: differenza tra il valore più **alto** e il valore più **basso** della distribuzione

Il **campo di variazione** è il modo più semplice per descrivere la variabilità di una variabile quantitativa:

- la nazione A ha un campo di variazione pari a

$$50000\$ - 0 = 50000\$$$

Il campo di variazione (range)

La prima **misura di variabilità** che vediamo è:

- **Campo di variazione (range)**: differenza tra il valore più **alto** e il valore più **basso** della distribuzione

Il **campo di variazione** è il modo più semplice per descrivere la variabilità di una variabile quantitativa:

- la nazione A ha un campo di variazione pari a

$$50000\$ - 0 = 50000\$$$

- la nazione B ha un campo di variazione pari a

$$30000\$ - 20000 = 10000\$$$

Il campo di variazione (range)

La prima **misura di variabilità** che vediamo è:

- **Campo di variazione (range)**: differenza tra il valore più **alto** e il valore più **basso** della distribuzione

Il **campo di variazione** è il modo più semplice per descrivere la variabilità di una variabile quantitativa:

- la nazione A ha un campo di variazione pari a

$$50000\$ - 0 = 50000\$$$

- la nazione B ha un campo di variazione pari a

$$30000\$ - 20000 = 10000\$$$

- si osserva una **maggiore variabilità** nella **distribuzione** dei redditi della nazione A

Il campo di variazione (range)

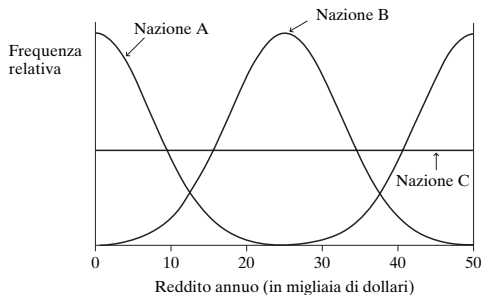
Il **campo di variazione** non ci da informazioni sulle altre caratteristiche della variabile

- ci dice solo la differenza tra valore **minimo** e **massimo**

Il campo di variazione (range)

Il **campo di variazione** non ci da informazioni sulle altre caratteristiche della variabile

- ci dice solo la differenza tra valore **minimo** e **massimo**
- ad esempio:



- le tre distribuzioni hanno la **stessa media** (25000\$) e lo **stesso campo di variazione** (50000\$)

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 Descrivere la variabilità dei dati**
 - Cosa si intende con variabilità?
 - Il campo di variazione (range)
 - La deviazione standard
- 4 Misure di posizione

La deviazione

Altre misure di variabilità si basano sul concetto di **deviazione**:

- la **deviazione** di un'osservazione y_i della media campionaria $\bar{y}$ è data dalla differenza tra i due valori

$$(y_i - \bar{y})$$

La deviazione

Altre misure di variabilità si basano sul concetto di **deviazione**:

- la **deviazione** di un'osservazione y_i della media campionaria $\bar{y}$ è data dalla differenza tra i due valori

$$(y_i - \bar{y})$$

Ogni **osservazione** presenta una deviazione dalla media:

- deviazione **positiva**: valore al di sopra della media

La deviazione

Altre misure di variabilità si basano sul concetto di **deviazione**:

- la **deviazione** di un'osservazione y_i della media campionaria $\bar{y}$ è data dalla differenza tra i due valori

$$(y_i - \bar{y})$$

Ogni **osservazione** presenta una deviazione dalla media:

- deviazione **positiva**: valore al di sopra della media
- deviazione **negativa**: valore al di sotto della media

La deviazione

Altre misure di variabilità si basano sul concetto di **deviazione**:

- la **deviazione** di un'osservazione y_i della media campionaria $\bar{y}$ è data dalla differenza tra i due valori

$$(y_i - \bar{y})$$

Ogni **osservazione** presenta una deviazione dalla media:

- deviazione **positiva**: valore al di sopra della media
- deviazione **negativa**: valore al di sotto della media
- la **somma** delle deviazioni è pari a 0

$$\sum (y_i - \bar{y}) = 0$$

La deviazione: esempio

Utilizziamo i dati sull'occupazione femminile in Est Europa

- le osservazioni:

83, 82, 72, 80, 80, 81, 84, 81

La deviazione: esempio

Utilizziamo i dati sull'occupazione femminile in Est Europa

- le osservazioni:

83, 82, 72, 80, 80, 81, 84, 81

- la media è pari a 80.375

La deviazione: esempio

Utilizziamo i dati sull'occupazione femminile in Est Europa

- le osservazioni:

83, 82, 72, 80, 80, 81, 84, 81

- la media è pari a 80.375
- faccio la somma delle deviazioni:

$$\begin{aligned}\sum (y_i - \bar{y}) &= (83 - 80.375) + \dots + (81 - 80.375) = \\ &= 2.625 + 1.625 - 8.375 - 0.375 - 0.375 + 0.625 + 3.625 + 0.625 = 0\end{aligned}$$

La deviazione: esempio

Utilizziamo i dati sull'occupazione femminile in Est Europa

- le osservazioni:

83, 82, 72, 80, 80, 81, 84, 81

- la media è pari a 80.375
- faccio la somma delle deviazioni:

$$\begin{aligned}\sum (y_i - \bar{y}) &= (83 - 80.375) + \dots + (81 - 80.375) = \\ &= 2.625 + 1.625 - 8.375 - 0.375 - 0.375 + 0.625 + 3.625 + 0.625 = 0\end{aligned}$$

Conseguenze:

- anche la **media** delle deviazioni è pari a 0!

La deviazione: esempio

Utilizziamo i dati sull'occupazione femminile in Est Europa

- le osservazioni:

83, 82, 72, 80, 80, 81, 84, 81

- la media è pari a 80.375
- faccio la somma delle deviazioni:

$$\begin{aligned}\sum (y_i - \bar{y}) &= (83 - 80.375) + \dots + (81 - 80.375) = \\ &= 2.625 + 1.625 - 8.375 - 0.375 - 0.375 + 0.625 + 3.625 + 0.625 = 0\end{aligned}$$

Conseguenze:

- anche la **media** delle deviazioni è pari a 0!
- per avere una misura della **variabilità media** usiamo le deviazioni in **valore assoluto** o le deviazioni **al quadrato**

La varianza

- Per misurare la **variabilità** si calcola la **varianza**:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\text{somma delle deviazioni al quadrato}}{\text{numerosità del campione}}$$

La varianza

- Per misurare la **variabilità** si calcola la **varianza**:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\text{somma delle deviazioni al quadrato}}{\text{numerosità del campione}}$$

La **varianza**:

- è approssimativamente la **media** delle **deviazioni al quadrato**
- il numeratore della varianza $\sum (y_i - \bar{y})^2$ viene chiamato **devianza** o **somma dei quadrati**
- All'aumentare di n , diminuisce la varianza.

La deviazione standard

Perché non usiamo direttamente la **varianza**?

- l'unità di misura della varianza è il **quadrato** dell'unità di misura originaria dei dati
 - cosa vuol dire la frase: *una deviazione media di 100\$ al quadrato*?
- la varianza quindi è di **difficile interpretabilità**

La deviazione standard

Perché non usiamo direttamente la **varianza**?

- l'unità di misura della varianza è il **quadrato** dell'unità di misura originaria dei dati
 - cosa vuol dire la frase: *una deviazione media di 100\$ al quadrato?*
- la varianza quindi è di **difficile interpretabilità**

Per questo motivo:

- utilizziamo la radice quadrata della varianza → la **deviazione standard** σ
- la **deviazione standard** σ è facile da interpretare:
 - σ è una sorta di **distanza tipica** delle osservazioni dalla media
 - σ ha la **stessa unità di misura** della variabile originaria

Esempio: la variabilità fra i punteggi dei quiz

Prendiamo per esempio due **campioni**:

- **Campione 1:** 0, 4, 4, 5, 7, 10
- **Campione 2:** 0, 0, 1, 9, 10, 10

Esempio: la variabilità fra i punteggi dei quiz

Prendiamo per esempio due **campioni**:

- **Campione 1**: 0, 4, 4, 5, 7, 10
- **Campione 2**: 0, 0, 1, 9, 10, 10

Per entrambi i campioni possiamo calcolare:

- la **media**: pari a 5 in entrambi i casi
- il **campo di variazione**: pari a 10 in entrambi i casi

Esempio: la variabilità fra i punteggi dei quiz

Prendiamo per esempio due **campioni**:

- **Campione 1**: 0, 4, 4, 5, 7, 10
- **Campione 2**: 0, 0, 1, 9, 10, 10

Per entrambi i campioni possiamo calcolare:

- la **media**: pari a 5 in entrambi i casi
- il **campo di variazione**: pari a 10 in entrambi i casi

Cosa possiamo dire sulla **variabilità**?

- il campione 1 è meno variabile attorno alla media
- la maggior parte dei punteggi nel campione 1 sono **prossimi** al valore medio
- i valori del campione 2 sono **più lontani** dal valore medio

Calcoliamo la **deviazione standard** per il primo campione (s_1)

- calcolo la **devianza**:

$$\sum (y_i - \bar{y})^2 = (0-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (7-5)^2 + (10-5)^2$$

$$\sum (y_i - \bar{y})^2 = 56$$

Calcoliamo la **deviazione standard** per il primo campione (s_1)

- calcolo la **devianza**:

$$\sum (y_i - \bar{y})^2 = (0-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (7-5)^2 + (10-5)^2$$

$$\sum (y_i - \bar{y})^2 = 56$$

- calcolo la **varianza**:

$$\sigma_1^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{56}{6} \simeq 9.3$$

Calcoliamo la **deviazione standard** per il primo campione (s_1)

- calcolo la **devianza**:

$$\sum (y_i - \bar{y})^2 = (0-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (7-5)^2 + (10-5)^2$$

$$\sum (y_i - \bar{y})^2 = 56$$

- calcolo la **varianza**:

$$\sigma_1^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{56}{6} \simeq 9.3$$

- calcolo la **deviazione standard**:

$$\sigma_1 = \sqrt{\sigma_1^2} = \sqrt{11.2} = 3.3$$

- per il campione 2 otteniamo $\sigma_2 = \sqrt{26.4} = 5.1$

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

- è sempre **maggiore o uguale a 0**: $\sigma \geq 0$
 - qualsiasi numero elevato al quadrato è pari a un numero positivo
 - la radice quadrata di un numero positivo è sempre un numero positivo

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

- è sempre **maggiore o uguale a 0**: $\sigma \geq 0$
 - qualsiasi numero elevato al quadrato è pari a un numero positivo
 - la radice quadrata di un numero positivo è sempre un numero positivo
- $\sigma = 0$ solo se **tutte** le osservazioni hanno lo **stesso valore**

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

- è sempre **maggiore o uguale a 0**: $\sigma \geq 0$
 - qualsiasi numero elevato al quadrato è pari a un numero positivo
 - la radice quadrata di un numero positivo è sempre un numero positivo
- $\sigma = 0$ solo se **tutte** le osservazioni hanno lo **stesso valore**
- **maggiore** è la variabilità intorno alla media **maggiore** è il valore di σ

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

- è sempre **maggiore o uguale a 0**: $\sigma \geq 0$
 - qualsiasi numero elevato al quadrato è pari a un numero positivo
 - la radice quadrata di un numero positivo è sempre un numero positivo
- $\sigma = 0$ solo se **tutte** le osservazioni hanno lo **stesso valore**
- **maggiore** è la variabilità intorno alla media **maggiore** è il valore di σ
- se sui dati si effettua una **trasformazione di scala** anche le deviazioni standard vengono trasformate
 - se misuro in *km* invece che in *m* anche la deviazione standard sarà misurata in *km*

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

- la ragione per cui si utilizza $(n - 1)$ piuttosto che n al denominatore dipende dall'**inferenza dei parametri**
 - dobbiamo tenere conto che stiamo usando la **media campionaria** (che ha una sua variabilità) e non la **media della popolazione**
 - la **varianza campionaria** tende a essere più piccola di quella della **popolazione**

Le proprietà della deviazione standard

La **deviazione standard campionaria** σ :

- la ragione per cui si utilizza $(n - 1)$ piuttosto che n al denominatore dipende dall'**inferenza dei parametri**
 - dobbiamo tenere conto che stiamo usando la **media campionaria** (che ha una sua variabilità) e non la **media della popolazione**
 - la **varianza campionaria** tende a essere più piccola di quella della **popolazione**

Se utilizziamo i dati dell'intera **popolazione** abbiamo che:

- utilizziamo n al posto di $(n - 1)$
- la **varianza** è la **media delle deviazioni al quadrato**

Interpretare la grandezza di s

Come si **interpreta** la grandezza di σ :

Interpretare la grandezza di s

Come si **interpreta** la grandezza di σ :

- utilizziamo i dati dell'esempio:
 - campione 1: $\sigma_1 = 3.3$
 - campione 2: $\sigma_2 = 5.1$

Interpretare la grandezza di s

Come si **interpreta** la grandezza di σ :

- utilizziamo i dati dell'esempio:
 - campione 1: $\sigma_1 = 3.3$
 - campione 2: $\sigma_2 = 5.1$
- possiamo dire che il campione 2 presenta una **variabilità maggiore** del campione 1
 - assumendo che i dati siano misurati con la stessa unità di misura!

Interpretare la grandezza di s

Come si **interpreta** la grandezza di σ :

- utilizziamo i dati dell'esempio:
 - campione 1: $\sigma_1 = 3.3$
 - campione 2: $\sigma_2 = 5.1$
- possiamo dire che il campione 2 presenta una **variabilità maggiore** del campione 1
 - assumendo che i dati siano misurati con la stessa unità di misura!

La varianza

Supponiamo di avere una distribuzione di frequenze di una variabile X con K modalità

X	n
-3	5
-1	7
0	4
2	3
5	2
6	4

Come si calcola la **varianza**?

La varianza

In questo caso, la varianza si calcola come

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^K (x_j - \bar{x})^2 n_j$$

Dove n_j è la numerosità corrispondente alla j -esima modalità della variabile X .

La varianza

In questo caso, la varianza si calcola come

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^K (x_j - \bar{x})^2 n_j$$

Dove n_j è la numerosità corrispondente alla j -esima modalità della variabile X .

Dato che n_j/n sono le frequenze relative f_j

$$\sigma^2 = \sum_{j=1}^K (x_j - \bar{x})^2 f_j$$

Formule semplificate

Esistono delle formule **semplificate** per il calcolo della varianza

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^n x_j^2 - \bar{x}^2$$

Nel caso in cui si abbia la distribuzione di frequenze della X

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^K x_j^2 n_j - \bar{x}^2 = \sum_{j=1}^K x_j^2 f_j - \bar{x}^2$$

Esempio: varianza con distribuzione di frequenze

Riprendiamo l'esempio precedente

X	n
-3	5
-1	7
0	4
2	3
5	2
6	4

- Utilizziamo la formula semplificata
- Ci serve calcolare x^2

Esempio: varianza con distribuzione di frequenze

Riprendiamo l'esempio precedente

X	n	x^2
-3	5	9
-1	7	1
0	4	0
2	3	4
5	2	25
6	4	36

- Utilizziamo la formula semplificata
- Ci serve calcolare x^2
- Ci serve calcolare le frequenze relative f

Esempio: varianza con distribuzione di frequenze

Riprendiamo l'esempio precedente

X	n	x^2	f	$x^2 f$
-3	5	9	0.20	1.80
-1	7	1	0.28	0.28
0	4	0	0.16	0
2	3	4	0.12	0.48
5	2	25	0.08	2
6	4	36	0.16	5.76

La formula semplificata è

$$\sigma^2 = \sum_{j=1}^K x_j^2 f_j - \bar{x}^2 = 10,32 - 0,72^2 = 9,8016$$

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 Descrivere la variabilità dei dati
- 4 **Misure di posizione**
 - Quartili e altri percentili
 - Misure di variabilità: lo scarto interquartile
 - Box Plot: rappresentazione grafica di 5 misure di posizione

Le misure di posizione

Le **misure di posizione**:

- ci forniscono informazioni sulla **posizione** dell'osservazione nella distribuzione
- ci forniscono informazioni sul **valore** al di sotto del quale ricade una determinata **percentuale** delle osservazioni
 - esempio: il 25% degli studenti prende un voto inferiore a 25

Le misure di posizione

Le **misure di posizione**:

- ci forniscono informazioni sulla **posizione** dell'osservazione nella distribuzione
- ci forniscono informazioni sul **valore** al di sotto del quale ricade una determinata **percentuale** delle osservazioni
 - esempio: il 25% degli studenti prende un voto inferiore a 25

Abbiamo già incontrato alcune misure di posizione:

- per calcolare il **campo di variazione** utilizziamo:
 - il **minimo**: nessuna osservazione (0%) ricade al di sotto del minimo
 - il **massimo**: tutte le osservazioni (100%) ricadono al di sotto del massimo
- la **mediana** è una **misura di posizione**: il 50% delle osservazioni ricade al di sotto della mediana

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 Descrivere la variabilità dei dati
- 4 **Misure di posizione**
 - **Quartili e altri percentili**
 - Misure di variabilità: lo scarto interquartile
 - Box Plot: rappresentazione grafica di 5 misure di posizione

Quartili e altri percentili

Tra le **misure di posizione** troviamo i **percentili**:

- Il percentile divide la distribuzione in 100 parti di uguale **numerosità**
- il **p-esimo percentile** è il valore nella distribuzione al di sotto del quale ricade il **p%** delle osservazioni e al di sopra del quale ricade il **(100-p) %** delle osservazioni

Quartili e altri percentili

Tra le **misure di posizione** troviamo i **percentili**:

- Il percentile divide la distribuzione in 100 parti di uguale **numerosità**
- il **p-esimo percentile** è il valore nella distribuzione al di sotto del quale ricade il **p%** delle osservazioni e al di sopra del quale ricade il **(100-p) %** delle osservazioni

Quartili e altri percentili

Facciamo alcuni **esempi**:

- la **mediana** è il **50-esimo percentile**:
 - $p = 50 \rightarrow$ la mediana è più grande del 50% delle osservazioni e più piccola del $(100 - 50)\% = 50\%$ delle osservazioni

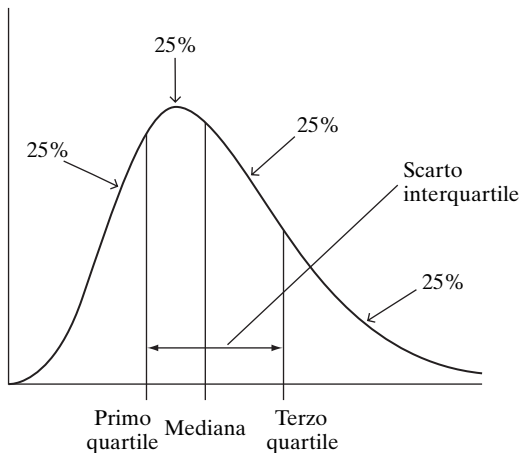
Quartili e altri percentili

Facciamo alcuni **esempi**:

- la **mediana** è il **50-esimo percentile**:
 - $p = 50 \rightarrow$ la mediana è più grande del 50% delle osservazioni e più piccola del $(100 - 50)\% = 50\%$ delle osservazioni
- i **quartili**: dividono la distribuzione in 4 parti che contengono un **quarto** delle osservazioni:
 - **primo quartile** $\rightarrow$ **25-esimo** percentile
 - un **quarto** delle osservazioni ricade al di sotto
 - **secondo quartile** $\rightarrow$ **mediana** $\rightarrow$ **50-esimo** percentile
 - **terzo quartile** $\rightarrow$ **75-esimo** percentile
 - un **quarto** delle osservazioni ricade al di sopra

Quartili e altri percentili

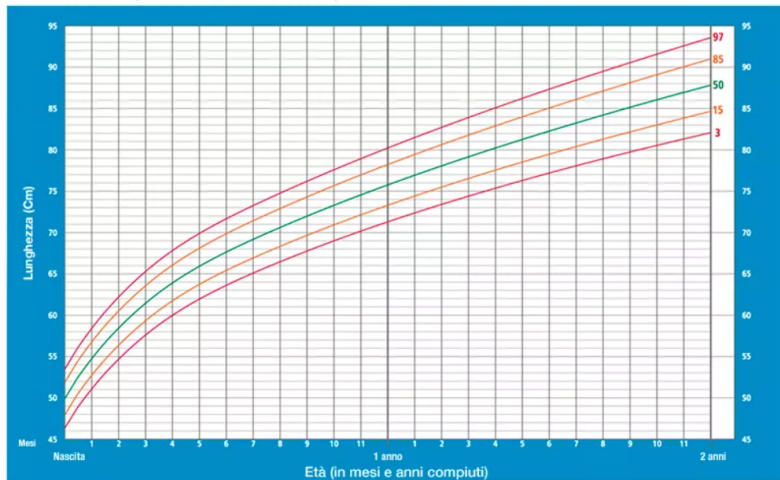
I **quartili** dividono la distribuzione in 4 parti, ciascuna contiene il 25% delle osservazioni:



Esempio: le curve di crescita

Lunghezza per età dei maschi

Percentili (dalla nascita ai 2 anni di età)



Curve di crescita dei bambini dell'OMS

Figure: Curve di crescita dei bambini maschi

Esempio: le curve di crescita

Questi grafici aiutano a **confrontare** le misurazioni individuali di un bambino con quelle di altri coetanei dello stesso sesso ed età, permettendo di valutare se la crescita procede in modo armonico.

- Si raccolgono dati da un ampio campione rappresentativo di bambini **della stessa età e sesso**. Questi dati vengono poi ordinati e **suddivisi in 100 gruppi**, ciascuno rappresentante **un percentile**.
- La curva del 50° percentile indica il **valore mediano**: se l'altezza di un bambino di 2 anni si colloca su questa curva, significa che **metà dei bambini sono più alti e metà più bassi di lui**.
- Variazioni nel tempo tra i percentili possono richiedere ulteriori approfondimenti da parte del pediatra.

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 Descrivere la variabilità dei dati
- 4 **Misure di posizione**
 - Quartili e altri percentili
 - **Misure di variabilità: lo scarto interquartile**
 - Box Plot: rappresentazione grafica di 5 misure di posizione

Lo scarto interquartile

I **quartili** possono essere utilizzati per calcolare lo **scarto interquartile (IQR, inter quartile range)**:

- **scarto interquartile**: differenza fra il **primo** e il **terzo** quartile

Lo scarto interquartile

I **quartili** possono essere utilizzati per calcolare lo **scarto interquartile (IQR, inter quartile range)**:

- **scarto interquartile**: differenza fra il **primo** e il **terzo** quartile
- rappresenta il **campio di variazione** per il 50% delle unità centrali

Lo scarto interquartile

I **quartili** possono essere utilizzati per calcolare lo **scarto interquartile (IQR, inter quartile range)**:

- **scarto interquartile**: differenza fra il **primo** e il **terzo** quartile
- rappresenta il **campo di variazione** per il 50% delle unità centrali
- come il **campo di variazione** e la **deviazione standard** aumenta quando aumenta la **variabilità**

Lo scarto interquartile

I **quartili** possono essere utilizzati per calcolare lo **scarto interquartile (IQR, inter quartile range)**:

- **scarto interquartile**: differenza fra il **primo** e il **terzo** quartile
- rappresenta il **campo di variazione** per il 50% delle unità centrali
- come il **campo di variazione** e la **deviazione standard** aumenta quando aumenta la **variabilità**
- a differenza del **campo di variazione** non è influenzato dalla presenza di **outlier**

La lezione

- 1 Distribuzioni di frequenza
- 2 Descrivere il centro dei dati
- 3 Descrivere la variabilità dei dati
- 4 Misure di posizione**
 - Quartili e altri percentili
 - Misure di variabilità: lo scarto interquartile
 - **Box Plot: rappresentazione grafica di 5 misure di posizione**

La sintesi a 5 numeri

Le misure che abbiamo visto fino a ora possono essere utilizzate per la **sintesi a 5 numeri**:

- utilizziamo 5 **misure di posizione**: minimo, primo quartile, mediana, terzo quartile, massimo
- ci permette di descrivere congiuntamente la **centralità** e la **variabilità** di una distribuzione

Esempio: output di un software statistico sulla distribuzione dei tassi di criminalità in USA

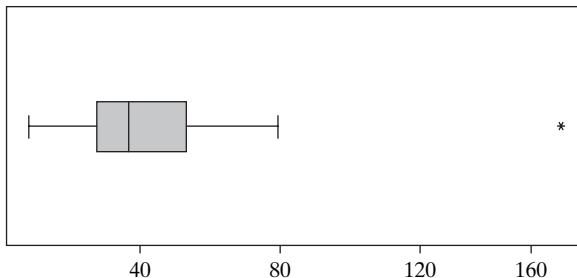
100%	Max	79.0
75%	Q3	51.0
50%	Med	36.5
25%	Q1	27.0
0%	Min	8.0

La **sintesi a 5 numeri** può essere utilizzata per costruire il **box plot**

Il box plot

Il **box plot** è un tipo di grafico che presenta questi elementi:

- il **box** (scatola) che contiene il **50%** centrale della distribuzione (dal **primo** al **terzo quartile**)
- la **mediana** è rappresentata da una linea che attraversa verticalmente il box
- le **linee** vengono chiamate **whiskers** (i baffi del gatto) e si estendono fino al **minimo** o **massimo** se **non** sono presenti **outlier**



Tasso di criminalità

Il box plot

I boxplot sono utili per **confrontare** due o più distribuzioni

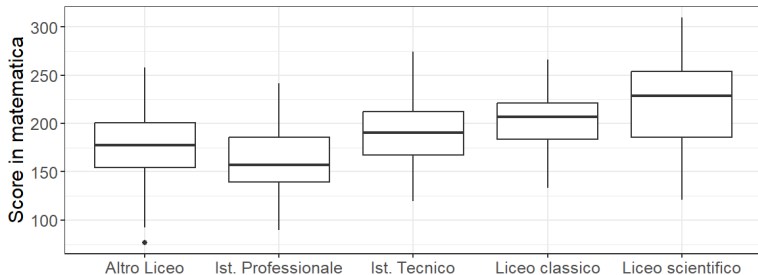


Figure: Boxplot score in matematica per tipo di diploma

Il box plot

I boxplot sono utili per **confrontare** due o più distribuzioni

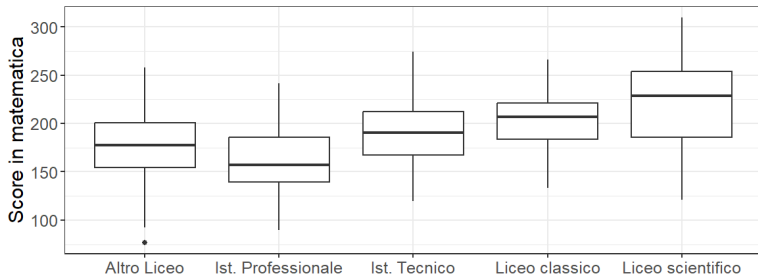


Figure: Boxplot score in matematica per tipo di diploma

- I punteggi sono più bassi negli istituti professionali
- Gli studenti dello scientifico mostrano una **maggiore variabilità**

Outlier

I **box plot** possono essere utilizzati per individuare gli **outlier**:

- un'osservazione viene definita **outlier** se ricade a più di $1.5 \times IQR$ al di sopra del **terzo quartile** o a più di $1.5 \times IQR$ al di sotto del **primo quartile**
 - definizione **formale** di outlier

Come segnaliamo gli **outlier** nel **box plot**:

- se **non ci sono** outlier allora i **whisker** del box si estendono fino al minimo e al massimo
 - non esistono osservazioni che sono oltre i $1.5 \times IQR$ i quartili
- se **ci sono** outlier allora i **whisker** del box si estendono fino all'osservazione che è compresa entro $1.5 \times IQR$
 - gli outlier vengono **indicati** separatamente, solitamente con dei pallini

Attenzione: bisogna sempre considerare gli outlier come *potenziali*

- potrebbero segnalare qualcosa (errori, osservazioni particolari, ...)