

# Statistica e Laboratorio

## 2. Statistica descrittiva: analisi univariate

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche

Università di Udine

via Tomadini 30/a - Udine

[paolo.vidoni@uniud.it](mailto:paolo.vidoni@uniud.it)

<https://elearning.uniud.it/>

# Sommario

- 1 **Sommario e introduzione**
- 2 Variabili statistiche
- 3 Distribuzioni di frequenza
- 4 Rappresentazioni grafiche
- 5 Indici sintetici

# Sommario

- **Introduzione**
- **Variabili statistiche**
- **Distribuzioni di frequenza**
- **Rappresentazioni grafiche**
- **Indici sintetici**

# Introduzione alla statistica descrittiva

- Metodi grafici e numerici per descrivere e sintetizzare i dati osservati.
- Distinzione fra tecniche di **analisi univariata**, cioè relative ad una singola caratteristica (variabile) di interesse, e tecniche di **analisi multivariata**, ovvero per lo studio congiunto di due o più caratteristiche (variabili) di interesse.
- Alcune nozioni di base sono utili anche per la statistica inferenziale.
- Come premessa ad una analisi inferenziale, è sempre opportuno effettuare uno studio descrittivo con riferimento al particolare campione osservato.

In particolare, l'uso di procedure grafiche e numeriche premette una analisi esplorativa dei dati (*Exploratory Data Analysis*) che si rileva fondamentale per la definizione di successive analisi più avanzate e per la presentazione finale dei risultati.

# Analisi esplorativa

Usualmente un'analisi statistica inizia con una prima esplorazione del data set, con l'obiettivo di:

- capire come i dati sono stati raccolti e se sono di natura osservazionale o sperimentale;
- individuare le unità statistiche, discutere la presenza di dati mancanti ed, eventualmente, *ripulire* il data set;
- codificare e riorganizzare i dati nella forma più conveniente per l'analisi;
- utilizzare metodi grafici e numerici per ricavare alcune informazioni preliminari sui dati osservati (analisi esplorativa dei dati).

“Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst.”

(D.C. Hoaglin, F. Mosteller and J.M. Tukey)

# Sommario

- 1 Sommario e introduzione
- 2 Variabili statistiche**
- 3 Distribuzioni di frequenza
- 4 Rappresentazioni grafiche
- 5 Indici sintetici

## Matrice dei dati

Si suppone che i dati siano già stati acquisiti e che siano disponibili nella forma di **matrice dei dati** (un data frame in R). Questi sono i cosiddetti **dati grezzi**.

Unità	Genere	Età	Livistr	Dist
Andrea	M	28	3	5.0
Claudio	M	17	2	7.5
Lucia	F	20	3	12.0
Giuseppe	M	32	4	3.2
Mara	F	16	2	NA
Luca	M	34	4	12.3
Aldo	M	18	3	25.0
Arianna	F	25	3	7.7

NA: dato mancante.

La matrice dei dati fornisce informazioni sulla popolazione in esame con riferimento a:

- Genere: Maschio (M), Femmina (F).
- Età (in anni compiuti):  $0, 1, 2, \dots$ ;
- Livello di istruzione (con codificazione numerica): 1 Analfabeta, 2 Scuola dell'obbligo, 3 Diploma, 4 Laurea;
- Distanza dal luogo di lavoro: numeri reali non negativi.

Ogni **riga** corrisponde ad una unità statistica e contiene i valori su essa rilevati delle caratteristiche di interesse.

Ogni **colonna** corrisponde ad una caratteristica di interesse e contiene i valori di tale caratteristica rilevati sulle varie unità statistiche.




# Variabili statistiche

Una **variabile** è una caratteristica delle unità statistiche che, al variare dell'unità, può assumere una pluralità di valori.

Le **modalità** di una variabile sono i valori che essa può assumere (e si presumono noti preliminarmente). Sono, in genere, aggettivi, valori numerici, espressioni verbali.

Le variabili si indicano con le lettere maiuscole, ad esempio  $Y$ , mentre una generica modalità si indica con  $y$ . L'insieme  $\mathcal{Y}$  è l'insieme di tutte le possibili modalità di  $Y$ .

**Esempio.** Con riferimento alla matrice dei dati presentata in precedenza, si hanno le variabili  $Y_1$ ="Genere", con  $\mathcal{Y}_1 = \{M, F\}$ ;  $Y_2$ ="Età", in anni compiuti, con  $\mathcal{Y}_2 = \{0, 1, 2, \dots\}$ ;  $Y_3$ ="Livello di istruzione", con  $\mathcal{Y}_3 = \{1, 2, 3, 4\}$ , avendo scelto la codifica della tabella precedente;  $Y_4$ ="Distanza", con  $\mathcal{Y}_4 = \mathbf{R}^+$ , anche se si può pensare ad una classificazione su intervalli prefissati.  $\mathbf{R}^+$  indica l'insieme dei numeri reali non negativi. 

Le variabili si possono classificare nel seguente modo.

- **Variabili qualitative (categoriali)**, se le modalità sono espresse in forma verbale. In particolare, si individuano:
  - ▶ **variabili qualitative sconnesse (nominali)**, per le quali non è possibile individuare un ordinamento naturale delle modalità (ad esempio, “Genere”, “Colore degli occhi”, “Religione professata”);
  - ▶ **variabili qualitative ordinali**, per le quali è invece possibile individuare un ordinamento naturale delle modalità (ad esempio, “Livello di istruzione”).
- **Variabili quantitative (numeriche)**, se le modalità sono espresse in forma numerica (da non confondere con le codifiche numeriche). In particolare, si individuano:
  - ▶ **variabili quantitative discrete**, se  $\mathcal{Y}$  è un insieme finito o al più numerabile (ad esempio, “Età” in a.c., “Numero di figli”);
  - ▶ **variabili quantitative continue**, se  $\mathcal{Y}$  è un insieme continuo (ad esempio, “Distanza”, “Altezza”, “Reddito”). Si noti che la continuità va intesa come *potenziale continuità* o come opportuno *riferimento semplificativo*.

Se le modalità di una variabile qualitativa sono solo due, si parla di **variabile dicotomica (binaria)**.

Una variabile quantitativa può essere con **scala di intervalli**, se non esiste uno zero naturale e non arbitrario. Una variabile quantitativa è con **scala di rapporti** se invece esiste uno zero con tali caratteristiche.

Ad esempio, la variabile “Temperatura”, in gradi centigradi, è su scala di intervalli poiché lo zero è convenzionale. Quindi, non ha senso affermare che la temperatura di  $30^{\circ}$  è due volte più calda della temperatura di  $15^{\circ}$ .

La variabile “Reddito” invece è su scala di rapporti. In questo caso ha senso affermare che un reddito di 20000 euro è il doppio di un reddito di 10000 euro.

- Esistono analisi statistiche adatte per lo studio dei diversi tipi di dati.
- Tra le varie tipologie di variabili esiste implicitamente una gerarchia.

Le variabili quantitative continue possono essere discretizzate, le variabili quantitative discrete possono essere tradotte in variabili qualitative ordinali, quelle ordinali possono essere considerate nominali.

Le analisi statistiche sono via via più ricche, man mano che si ascende la gerarchia.

- Le analisi univariate considerano una sola variabile rilevata sulle unità statistiche.
- Nello studio congiunto di due o più variabili si parla di analisi statistica bivariata o, in generale, multivariata.

La variabile  $Y$  viene rilevata su una popolazione (campione) costituita da  $n$  unità e si ottiene una successione di modalità osservate  $(y_1, \dots, y_i, \dots, y_n)$ , dove  $y_i$ ,  $i = 1, \dots, n$ , è il valore assunto da  $Y$  con riferimento all'unità  $i$ -esima.

È utile distinguere tra variabile e risultato della sua rilevazione sulla popolazione (campione).

Si definisce **variabile statistica** la rilevazione  $(y_1, \dots, y_i, \dots, y_n)$  di una certa variabile  $Y$  su una determinata popolazione (campione). È una colonna della matrice dei dati.

La stessa variabile rilevata su popolazioni (campioni) diverse dà luogo, in genere, a variabili statistiche differenti. Si usa il simbolo  $Y$  per indicare anche la variabile statistica.

**Esempio.** Con riferimento alla matrice dei dati vista in precedenza, alla variabile  $Y = \text{“Età”}$  corrisponde la variabile statistica

$Y = (28, 17, 20, 32, 16, 34, 18, 25)$ , con  $n = 8$ .



Nel caso di dati (campioni)  $y_1, \dots, y_i, \dots, y_n$  di tipo numerico può essere utile considerare l'**insieme dei dati (campione) ordinato**

$y_{(1)}, \dots, y_{(i)}, \dots, y_{(n)}$ , ottenuto disponendo le osservazioni in ordine non decrescente.

Il valore che occupa la posizione  $i$ -esima,  $y_{(i)}$ , si dice avere **rango**  $i$ ,  $i = 1, \dots, n$ . Si noti che il **minimo** e il **massimo** corrispondono rispettivamente a  $y_{(1)} = \min(y_1, \dots, y_n)$  e  $y_{(n)} = \max(y_1, \dots, y_n)$ .

Non tutte le modalità potenzialmente assumibili dalla variabile  $Y$  possono venire effettivamente rilevate in una popolazione (campione).

Il **supporto** di una variabile statistica  $Y$ , indicato con  $S_Y$ , è l'insieme delle modalità di  $Y$  effettivamente osservate nella popolazione (campione);  $S_Y = \{y_1, \dots, y_j, \dots, y_J\}$ . Si noti che  $J \leq n$ .

Le modalità osservate, che concorrono a costituire  $S_Y$ , sono tra loro distinte, cioè vanno prese una volta sola anche se ripetute.

Nel caso di variabili qualitative ordinali e quantitative si suppone che le modalità appartenenti al supporto vengano ordinate secondo un ordine crescente:  $y_1 < y_2 < \dots < y_J$ .

**Esempio.** Con riferimento alla variabile  $Y$  = “Età”, considerata in precedenza, il supporto è  $S_Y = \{16, 17, 18, 20, 25, 28, 32, 34\}$ , mentre  $\mathcal{Y} = \{0, 1, 2, \dots\}$ . ◇

# Sommario

- 1 Sommario e introduzione
- 2 Variabili statistiche
- 3 Distribuzioni di frequenza**
- 4 Rappresentazioni grafiche
- 5 Indici sintetici



## Frequenze assolute

I dati grezzi (la variabile statistica), pur rappresentando pienamente il contenuto dell'osservazione, usualmente non permettono di cogliere in modo chiaro le caratteristiche del fenomeno in esame.

È utile passare dai dati in forma grezza ad una **tabella di frequenza** che fornisca una sintesi dei dati in un formato facile da capire.

**Esempio.** *Colesterolo*. Si è misurato il livello di colesterolo sierico a  $n = 2294$  soggetti maschi, discriminando i pazienti in due classi di età: 25-34 anni e 55-64 anni. Ci si chiede se il livello di colesterolo sia maggiore negli adulti piuttosto che nei giovani.

“Livello di colesterolo” (pazienti età 25-34 anni): 80, 83, 86, 90, 93, 96, 109, 112, 116, 119, 120, ..., 250, 251, 251, 251, 252, 252, 252, 253, ...

“Livello di colesterolo” (pazienti età 55-64 anni): 84, 85, 96, 97, 97, 101, 119, 122, 133, 138, 140, ..., 266, 266, 268, 270, 272, 272, 284, 290, ...

I dati rilevati sono troppi per cercare di ricavare informazioni utili solamente guardandoli. È necessario operare una sintesi.

Definire una tabella dove si considerano le frequenze con cui le diverse modalità, o classi di modalità, sono state osservate. ◇

Se  $y_j \in S_Y$ ,  $j = 1, \dots, J$ , è una delle modalità osservate di  $Y$ , si dice **frequenza assoluta** di  $y_j$  il numero di volte che  $y_j$  risulta osservata. Si indica con  $f_j$ . Evidentemente,  $f_j > 0$ ,  $j = 1, \dots, J$ , e  $\sum_{j=1}^J f_j = n$ .

La lista delle modalità osservate accompagnate dalle rispettive frequenze assolute è detta **distribuzione di frequenza assoluta** e si rappresenta con una **tabella di frequenza** del tipo

Modalità	$y_1$	$\cdots$	$y_j$	$\cdots$	$y_J$	Totale
Frequenza	$f_1$	$\cdots$	$f_j$	$\cdots$	$f_J$	$\sum_{j=1}^J f_j$

**Esempio.** Con riferimento alla matrice dei dati vista in precedenza, si ricavano le seguenti tabelle di frequenza (assoluta) associate alle variabili “Genere” e “Livello di istruzione”:

Genere	frequenza
M	5
F	3
Totale	8

Liv. di istruz.	frequenza
<i>Scuola obbligo</i>	2
<i>Diploma superiore</i>	4
<i>Laurea o superiore</i>	2
Totale	8



Una tabella di frequenza riferita ad una **variabile statistica qualitativa** è detta **serie statistica**.

Se la **variabile statistica** è **quantitativa continua**, si osservano, a meno di effetti di arrotondamento, tante modalità distinte quante sono le unità statistiche:  $S_Y$  corrisponde all'insieme dei dati grezzi e  $f_j = 1$ ,  $j = 1, \dots, J$ .

Questo può accadere, in alcuni casi, anche con variabili statistiche quantitative discrete.

È conveniente definire **classi di modalità** contigue e contare le unità che appartengono a ciascuna classe.

Le classi vanno definite di modo che: non siano né troppe né troppo poche; siano disgiunte; comprendano tutte le modalità osservate.

La regola di individuare un numero di classi (di uguale ampiezza) pari a  $\sqrt{n}$  in molti casi va bene. Talvolta è necessario fare qualche aggiustamento o utilizzare regole più sofisticate.

Le classi non hanno necessariamente un'ampiezza costante.

Si ottiene una **tabella (distribuzione) di frequenza assoluta con modalità raggruppate in classi**

Classi	$y_0 \vdash y_1$	$\cdots$	$y_{j-1} \vdash y_j$	$\cdots$	$y_{J-1} \vdash y_J$	Totale
Freq.	$f_1$	$\cdots$	$f_j$	$\cdots$	$f_J$	$\sum_{j=1}^J f_j$

dove  $f_j$  è la frequenza assoluta associata alla classe  $y_{j-1} \vdash y_j$ , che corrisponde all'intervallo  $(y_{j-1}, y_j]$ . Analogamente,  $y_{j-1} \vdash y_j$  corrisponde all'intervallo  $[y_{j-1}, y_j)$  e  $y_J -$  indica  $(y_J, +\infty)$ .

Una tabella di frequenza così ottenuta è detta **seriazione statistica**.

**Esempio.** Con riferimento alla matrice dei dati vista in precedenza, si ricava la seguente seriazione statistica associata alla variabile “Distanza”:

Dist	frequenza
0 ÷ 5	2
5 ÷ 15	4
15–	1
Totale	7



## Frequenze relative

La **frequenza relativa** di una modalità  $y_j$ , o di una classe di modalità  $y_{j-1} \dashv y_j$ , è la proporzione  $p_j$  di unità statistiche portatrici di tale modalità o classe di modalità. Corrisponde a

$$p_j = \frac{f_j}{\sum_{j=1}^J f_j} = \frac{f_j}{n}, \quad j = 1, \dots, J.$$

Evidentemente,  $p_j > 0$ ,  $j = 1, \dots, J$ , e  $\sum_{j=1}^J p_j = 1$ .

Si possono definire anche le **frequenze relative percentuali**, definite come  $p_j 100$ ,  $j = 1, \dots, J$ .

Le frequenze relative sono utili per percepire il peso delle varie modalità e per operare confronti tra diverse popolazioni.

Se  $S_Y = \{y_1\}$ , allora  $J = 1$ ,  $f_1 = n$ ,  $p_1 = 1$  e la **variabile statistica**  $Y$  è detta **degenere**.

**Esempio.** Si consideri la tabella che fornisce la distribuzione per sesso della popolazione residente in Italia (confini attuali) ricavata dai censimenti del 1861 e del 1981; i dati sono espressi in migliaia.

		Freq. ass.	Freq. rel.	Freq. rel. %
1861	M	13399	0.5089	50.89
	F	12929	0.4911	49.11
	Totale	26328	1	100
1981	M	27506	0.4863	48.63
	F	29051	0.5137	51.37
	Totale	56557	1	100

Analizzando le frequenze relative si ha una rappresentazione immediata di come si è modificata la struttura della popolazione italiana. ◇



**Esempio.** *Colesterolo* (continua). Considerando i dati sul livello di colesterolo sierico, si ottiene la seguente tabella, dove le modalità sono state raggruppate in classi di ampiezza 40

Liv. colesterolo (mg/100 ml)	$f_j$ (età 25-34)	$f_j$ (età 55-64)	$p_j$ (età 25-34)	$p_j$ (età 55-64)
80 ┤ 120	13	5	0.012	0.004
120 ┤ 160	150	48	0.141	0.039
160 ┤ 200	442	265	0.414	0.216
200 ┤ 240	299	458	0.280	0.373
240 ┤ 280	115	281	0.108	0.229
280 ┤ 320	34	128	0.032	0.104
320 ┤ 360	9	35	0.008	0.029
360 ┤ 400	5	7	0.005	0.006
Totale	1067	1227	1	1

**Commento:** i soggetti più giovani hanno una porzione più elevata di osservazioni inferiori a 200 mg/100 ml, mentre i più anziani presentano una porzione più elevata al di sopra di questo valore. ◇

**Esempio. Perni.** In uno stabilimento industriale ci sono tre macchinari per la produzione di perni di acciaio, che devono rispettare le specifiche di diametro.

Per valutarne l'efficacia del procedimento produttivo si analizzano  $n = 400$  perni che vengono classificati, con riferimento agli standard richiesti per il diametro, in:

- fine: il diametro è troppo fine rispetto alle specifiche richieste;
- ok: il diametro soddisfa le specifiche richieste;
- spesso: il diametro è troppo spesso rispetto alle specifiche richieste.

Ogni perno poi viene classificato a seconda del macchinario che lo ha prodotto.

I dati grezzi prendono la forma di una lunga tabella che risulta di difficile analisi

Perno	Macchinario	Diametro
1	1	ok
2	3	ok
3	1	fine
...	...	...
400	2	spesso

I dati grezzi vengono sintetizzati determinando la distribuzione delle frequenze assolute e relative, suddividendo i perni in base al macchinario che li ha prodotti.

## Tabella delle frequenze assolute.

	Fine	Ok	Spesso	Totale
Macchinario 1	10	102	8	120
Macchinario 2	34	161	5	200
Macchinario 3	10	60	10	80

Suddividendo ogni riga per il suo totale si ottiene la tabella delle frequenze relative. L'utilizzazione delle frequenze relative è utile per confrontare la bontà dei tre macchinari, dal momento che la dimensione dei gruppi di perni classificati sulla base del macchinario di produzione è diversa.

	Fine	Ok	Spesso	Totale
Macchinario 1	0.083	0.850	0.067	1
Macchinario 2	0.170	0.805	0.025	1
Macchinario 3	0.125	0.750	0.125	1



## Frequenze cumulate

Quando si hanno **variabili con modalità ordinabili** (qualitative ordinali o quantitative), può essere utile considerare la frequenza con cui si presentano modalità di ordine inferiore o uguale ad un certo valore.

La **frequenza assoluta cumulata**  $F_j$  o, analogamente, la **frequenza relativa cumulata**  $P_j$  definiscono la frequenza assoluta o relativa di modalità o classi di modalità non superiori alla  $j$ -esima,  $j = 1, \dots, J$ .

Si ottengono cumulando progressivamente le frequenze, più precisamente

$$F_j = \sum_{i=1}^j f_i, \quad P_j = \sum_{i=1}^j p_i, \quad j = 1, \dots, J.$$

Evidentemente,  $F_1 = f_1$ ,  $F_J = n$ ,  $P_1 = p_1$ ,  $P_J = 1$ .

**Esempio.** *Colesterolo* (continua). Considerando i dati sul livello di colesterolo sierico

Liv. colesterolo (mg/100 ml)	$F_j$ (età 25-34)	$F_j$ (età 55-64)	$P_j$ (età 25-34)	$P_j$ (età 55-64)
80 ┤ 120	13	5	0.012	0.004
120 ┤ 160	163	53	0.153	0.043
160 ┤ 200	605	318	0.567	0.259
200 ┤ 240	904	776	0.847	0.632
240 ┤ 280	1019	1057	0.955	0.861
280 ┤ 320	1053	1185	0.987	0.965
320 ┤ 360	1062	1220	0.995	0.994
360 ┤ 400	1067	1227	1	1

**Commento:** i soggetti più anziani tendono ad avere livelli di colesterolo più elevati rispetto ai soggetti più giovani. ◇

## Serie storica

Quando si misura un fenomeno nel **tempo**, si ottiene una distribuzione di frequenza che prende il nome di **serie storica** (**temporale**).

**Esempio.** Si considera il numero di occupati in Italia dal 1997 al 2001.

Anno	No. occupati (in migliaia)
1997	20207
1998	20435
1999	20692
2000	21080
2001	21514



## Serie spaziale

Quando si misura un fenomeno nello **spazio**, si ottiene una distribuzione di frequenza che prende il nome di **serie spaziale (territoriale)**.

**Esempio.** Si considera il numero di occupati in Italia nel 2002, suddivisi per ripartizione territoriale.

Ripartizione territoriale	No. occupati (in migliaia)
Nord	11461
Centro	4513
Sud e Isole	6286





# Sommario

- 1 Sommario e introduzione
- 2 Variabili statistiche
- 3 Distribuzioni di frequenza
- 4 Rappresentazioni grafiche**
- 5 Indici sintetici

# Rappresentazioni grafiche

Oltre alle tabelle di frequenza, risulta utile introdurre alcune rappresentazioni grafiche, dette **diagrammi statistici (grafici)**.

L'osservazione di un buon grafico fornisce informazioni interessanti su un insieme di dati con una semplice "occhiata".

Un grafico è di solito di più facile e immediata consultazione rispetto ad una tabella.

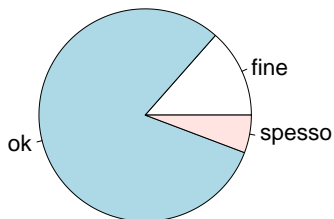
La scelta del grafico dipende dalla natura dei dati. Si utilizzano rappresentazioni grafiche diverse per dati discreti, continui, serie storiche, ecc.

- Per dati categoriali si possono utilizzare, ad esempio,
  - ▶ **diagrammi circolari**
  - ▶ **diagrammi a rettangoli**
  - ▶ **diagrammi a rettangoli multipli**
- Per dati numerici si possono utilizzare, ad esempio,
  - ▶ **diagrammi a bastoncini**
  - ▶ **istogrammi**
  - ▶ **poligoni di frequenza**
  - ▶ **stima della densità**
  - ▶ **funzione di ripartizione empirica**
  - ▶ **diagrammi di dispersione**
  - ▶ **boxplot** (che verranno presentati in seguito)

# Diagrammi circolari

I **diagrammi circolari (a torta)** sono utili per rappresentare **serie statistiche sconnesse**, riferite a dati qualitativi nominali o eventualmente ordinali (dati categoriali). L'area del settore circolare deve essere proporzionale alla frequenza della modalità corrispondente.

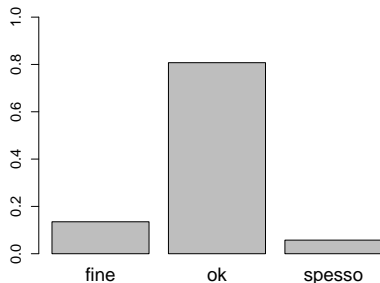
**Esempio.** *Perni* (continua). Considerando i dati riferiti alla produzione dei perni, le diverse modalità relative al diametro sono rappresentate dagli spicchi della torta, la cui dimensione è proporzionale alla corrispondente frequenza.



## Diagrammi a barre

I **diagrammi a rettangoli (a barre)** sono utili per rappresentare **serie statistiche sconnesse**, riferite a dati qualitativi nominali o eventualmente ordinali (dati categoriali). Le altezze dei rettangoli sono proporzionali alle frequenze delle modalità. Le basi hanno la stessa dimensione e sono separate per non implicare alcuna continuità.

**Esempio.** *Perni* (continua). Considerando i dati riferiti alla produzione dei perni, si ottiene il seguente diagramma a rettangoli.



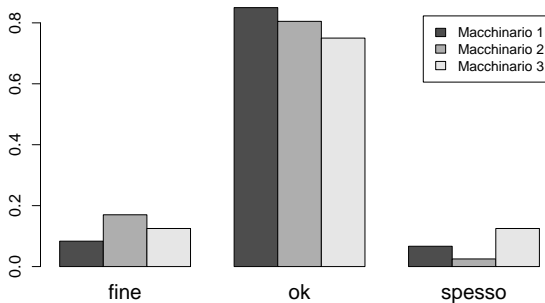
I **diagrammi a rettangoli (a barre) multipli** sono utili per rappresentare **serie statistiche sconnesse**, riferite a dati qualitativi nominali o eventualmente ordinali (dati categoriali), nel caso in cui la distribuzione di frequenza risulta suddivisa secondo un determinato criterio di classificazione.

I rettangoli hanno la base uguale e sono separati per non implicare alcuna continuità. Le altezze sono proporzionali alla frequenze delle modalità; si considerano le **frequenze relative** affinché il confronto abbia senso.

I rettangoli vengono raggruppati tenendo conto del criterio di classificazione. È opportuno evitare di disporre un numero elevato di rettangoli a confronto.

**Esempio.** *Perni* (continua). Considerando i dati riferiti alla produzione dei perni, si ottiene il seguente diagramma a rettangoli multipli.

Le diverse modalità relative alla dimensione dei perni vengono rappresentate considerando in modo distinto i tre macchinari di produzione .

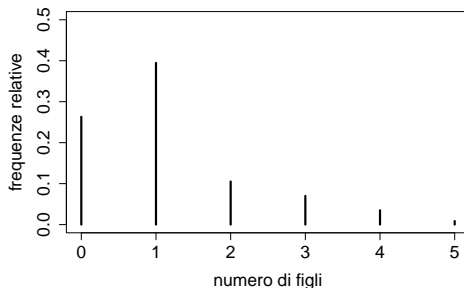


## Diagrammi a bastoncini

I **diagrammi a bastoncini** servono per rappresentare **distribuzioni di frequenza assoluta o relativa**, riferite a **dati qualitativi discreti**.

L'altezza dei bastoncini è proporzionale o pari alla frequenza, assoluta o relativa, della modalità.

**Esempio. Figli.** Si considera il numero di figli con riferimento alle famiglie residenti in un determinato territorio. La distribuzione di frequenza relativa è rappresentata con il seguente diagramma.





# Istogrammi

Gli **istogrammi** si utilizzano per rappresentare **distribuzioni di frequenza assoluta o relativa** con modalità raggruppate in classi, riferite usualmente a **dati quantitativi continui**.

L'istogramma è un insieme di rettangoli adiacenti, ognuno rappresentativo di una classe, posti su un piano cartesiano.

Il rettangolo corrispondente alla classe  $j$ -esima  $y_{j-1} \preceq y_j$ ,  $j = 1, \dots, J$ , ha come base l'intervallo  $[y_{j-1}, y_j]$  e

- altezza (e quindi area) proporzionale a, oppure pari a,  $f_j / (y_j - y_{j-1})$ : **istogramma delle frequenze assolute**;
- altezza (e quindi area) proporzionale a, oppure pari a,  $p_j / (y_j - y_{j-1})$ : **istogramma delle frequenze relative**.

Se i rettangoli hanno la stessa base, allora l'altezza è proporzionale a  $f_j$  o a  $p_j$ .

Se le classi estreme sono aperte, ad esempio  $-y_1$  e  $y_{J-1}-$ , vanno chiuse scegliendo opportunamente gli estremi  $y_0$  e  $y_J$ .

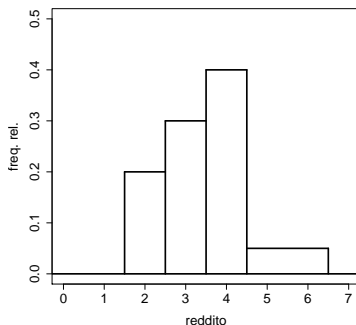
Se si considerano altezze pari a  $p_j/(y_j - y_{j-1})$ , la somma delle aree dei rettangoli è pari a 1.

L'istogramma può essere utilizzato anche per descrivere distribuzioni di frequenza associate a **variabili statistiche quantitative discrete**, quando si hanno molte modalità osservate distinte.

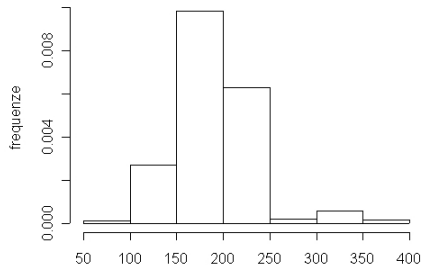
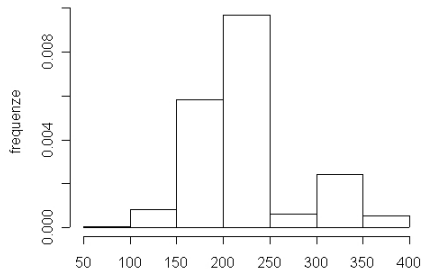
**Esempio.** *Reddito*. Si consideri la seguente seriazione riferita alla variabile reddito (lordo mensile in migliaia di euro)

Reddito	1.5-2.5	2.5-3.5	3.5-4.5	4.5-6.5	Tot.
freq. rel.	0.2	0.3	0.4	0.1	1

L'associato istogramma della frequenza relative corrisponde a



**Esempio.** *Colesterolo* (continua). Considerando i dati sul livello di colesterolo sierico, si ottengono i seguenti istogrammi riferiti, rispettivamente, ai pazienti di età 25-34 anni e di età 55-64 anni.

**eta' 25-34****eta' 55-64**

I grafici sono basati sulle **frequenze relative** e suggeriscono le stesse conclusioni fatte sulla base delle tabelle di frequenza. La distribuzione di frequenza relativa dei pazienti più anziani è spostata più a destra rispetto a quella dei pazienti giovani. ◇

## Poligoni di frequenza

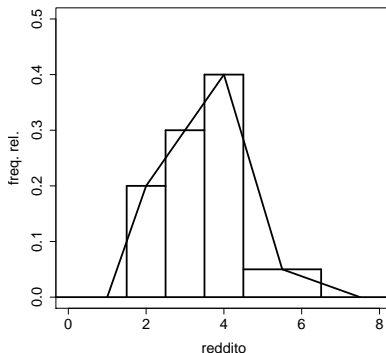
Un **poligono di frequenza** è uno smussamento locale dell'istogramma.

Per costruirlo si introducono due classi adiacenti alle classi esterne  $y_0 \dashv y_1$  e  $y_{J-1} \dashv y_J$ , ognuna con ampiezza uguale alla classe vicina e frequenza assoluta pari a zero.

Il poligono si ottiene unendo i punti di mezzo dei lati superiori dei rettangoli dell'istogramma con una linea spezzata.

Solo se i rettangoli hanno la stessa base, l'area sottesa dalla linea spezzata coincide con la somma dell'area dei rettangoli

**Esempio.** *Reddito* (continua). Si consideri la seriazione riferita alla variabile reddito. Partendo dall'associato istogramma si ottiene il corrispondente poligono di frequenza.



## Stima della densità

In alternativa all'istogramma, è possibile definire una stima della distribuzione delle frequenze tramite una curva che risulti essere più smussata (in questo modo si tiene conto che la variabile è continua).

Date le osservazioni  $y_1, \dots, y_i, \dots, y_n$ , la funzione che rappresenta la **stima della densità con il metodo del nucleo** è definita come

$$f_n(y) = \frac{1}{nb} \sum_{i=1}^n K\left(\frac{y - y_i}{b}\right), \quad y \in \mathbf{R},$$

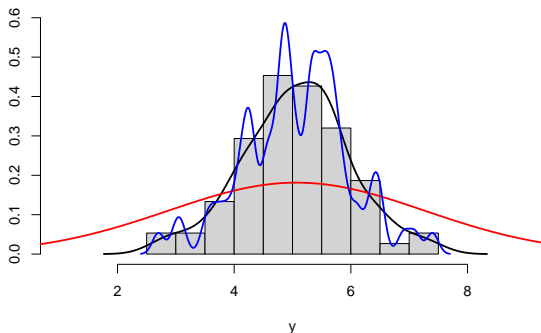
dove  $K(\cdot)$  è detto **nucleo (kernel)**,  $b > 0$  è la **banda** e  $\mathbf{R}$  è l'insieme dei numeri reali.

Ad ogni dato  $y_i$  si sovrappone non un rettangolo ma una curva che risulta essere più smussata. La sua altezza è proporzionale alla frequenza dei punti e la sua ampiezza dipende dalla banda  $b$ .

Si possono scegliere diversi nuclei  $K(\cdot)$ , che devono soddisfare ad alcune proprietà; in particolare,  $K(u) \geq 0$  e  $\int u^2 K(u) du = 1$ .

È importante scegliere la banda  $b$  in modo opportuno (in genere i software operano una scelta ottimale): se  $b$  è troppo grande il grafico risulta appiattito, mentre se  $b$  è troppo piccolo il grafico si avvicina ad un grafico a bastoncini.

Dato un insieme di osservazioni numeriche, si costruisce l'istogramma delle frequenze relative, la stima della densità con scelta ottimale per  $b$  (**nero**), con  $b$  troppo grande (**rosso**) e troppo piccolo (**blu**).





# Funzione di ripartizione empirica

Un'ulteriore rappresentazione grafica per **dati quantitativi**, e che risulta in molti casi particolarmente efficace, è fornita dalla **funzione di ripartizione empirica**.

La funzione di ripartizione empirica è una funzione il cui valore nel punto  $y \in \mathbf{R}$  corrisponde al rapporto tra il numero di osservazioni minori o uguali a  $y$  e il numero totale di osservazioni

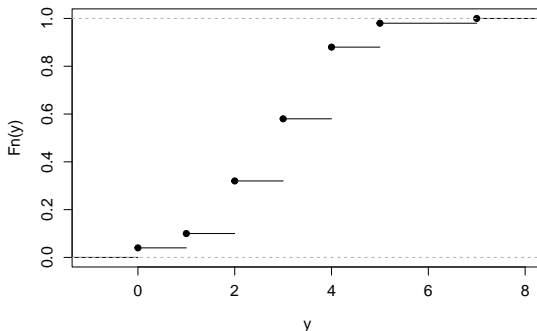
$$F_n(y) = \frac{\text{no. oss. } \leq y}{\text{no. totale oss.}}, \quad y \in \mathbf{R}.$$

Al variare di  $y$ , fornisce la proporzione cumulata di unità statistiche che presentano modalità minori o uguali a  $y$  ed è quindi una funzione a gradini.

La nozione di funzione di ripartizione empirica è utile per una rappresentazione grafica delle frequenze relative cumulate, in particolare per variabili quantitative discrete.

**Esempio.** Si consideri la seguente tabella delle frequenze relative e relative cumulate riferita ad una variabile quantitativa discreta. Di seguito viene rappresentata la corrispondente funzione di ripartizione empirica.

$Y$	0	1	2	3	4	5	7	Tot.
freq. rel.	0.04	0.06	0.22	0.26	0.30	0.10	0.02	1
freq. rel. cum.	0.04	0.10	0.32	0.58	0.88	0.98	1.00	

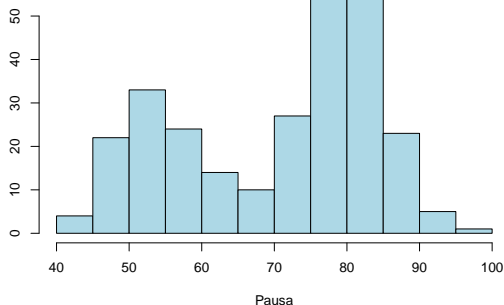


**Esempio.** *Geyser Old Faithful*. Si dispone di dati riferiti alle durate delle pause (in minuti) e alla tipologia delle eruzioni che precedono le pause (lunga o corta), con riferimento al geyser Old Faithful che si trova nel parco nazionale di Yellowstone, Wyoming, USA.

Si hanno le seguenti  $n = 272$  osservazioni

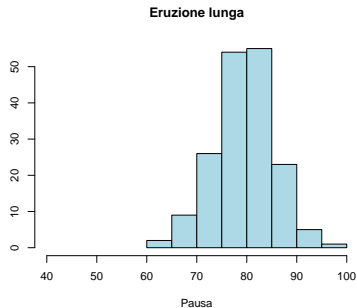
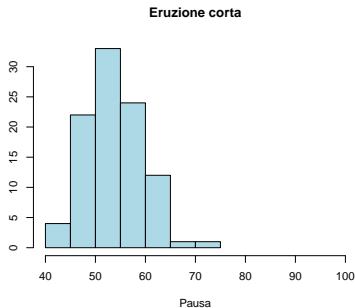
Pausa	Eruzione
79	Lunga
54	Corta
74	Lunga
62	Corta
85	Lunga
55	Corta
⋮	⋮
90	Lunga
46	Corta
73	Lunga

Considerando tutti i dati disponibili si può calcolare il seguente istogramma. Si individuano classi di ampiezza costante pari a 5 minuti.



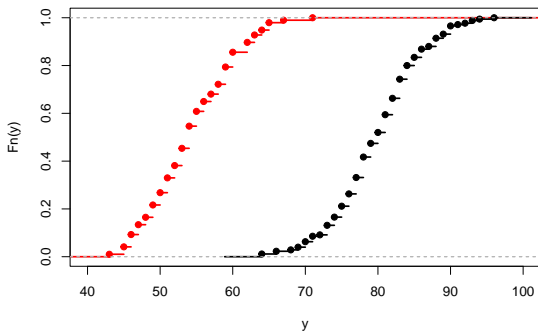
La forma dell'istogramma è dovuta al fatto che si sono considerati congiuntamente i dati sulla durata delle pause, senza distinguere tra durata corta o lunga dell'eruzione precedente.

Se si considerano le durate delle pause, raggruppando i dati in base alla tipologia dell'eruzione precedente, si ottengono i seguenti istogrammi



Le pause successive ad un'eruzione corta tendono ad essere più corte di quelle che seguono un'eruzione lunga.

Considerando i due insiemi di dati distinti, si possono determinare le associate funzioni di ripartizione empiriche (**rosso**=eruzione corta, **nero**=eruzione lunga), che confermano questa affermazione.



# Diagrammi di dispersione

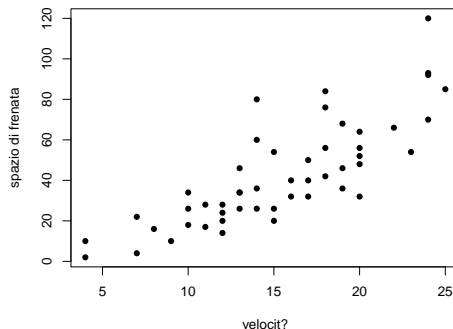
In molti casi, per ogni unità statistica vengono raccolti dati di più variabili.

Nel caso di due **variabili quantitative**, per una prima analisi della relazione tra le variabili si possono usare i **diagrammi di dispersione (scatterplot)**.

Se  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , sono i valori delle due variabili, il diagramma di dispersione si ottiene rappresentando i punti in un piano cartesiano.

Per più di due variabili, si possono ottenere i diagrammi di dispersione per ogni coppia di variabili.

**Esempio. Velocità.** Si dispone di dati riferiti alla velocità, in miglia orarie, e allo spazio di frenata, in piedi, per  $n = 50$  automobili degli anni '20. Si costruisce il seguente diagramma di dispersione.



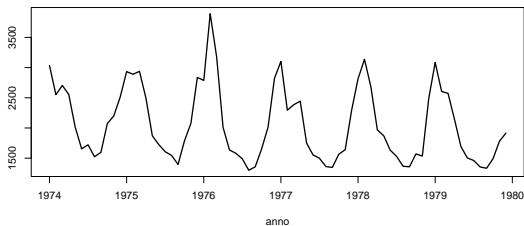
Lo spazio di frenata aumenta al crescere della velocità, con una relazione che si discosta leggermente da quella lineare.





Se i dati bivariati presentano una **componente temporale**, si ha una serie storica. In questo caso, si possono rappresentare i dati in funzione del tempo utilizzando un particolare diagramma di dispersione con i punti uniti da una linea spezzata.

**Esempio.** *Patologie polmonari.* Si considerano i dati riferiti al numero di decessi mensili per patologie polmonari (bronchiti, asma, enfisema) rilevati nel Regno Unito dal 1974 al 1979.



I decessi, come prevedibile, aumentano nei mesi invernali inoltre, tra il 1975 e il 1976, c'è stato un evidente aumento del numero dei decessi.



# Sommario

- 1 Sommario e introduzione
- 2 Variabili statistiche
- 3 Distribuzioni di frequenza
- 4 Rappresentazioni grafiche
- 5 Indici sintetici**

# Indici sintetici

Fino ad ora si sono utilizzate tabelle di frequenza e diagrammi statistici per descrivere e presentare dati ottenuti da rilevazioni statistiche.

Questi strumenti possono non essere adeguati per descrivere in modo efficace alcuni aspetti rilevanti dell'insieme dei dati e della associata distribuzione di frequenza.

È interessante indagare i seguenti aspetti dei dati:

- la **posizione**, cioè il *centro* dei dati;
- la **variabilità**, cioè la *dispersione* dei dati;
- la forma della distribuzione di frequenza, considerando in particolare la **simmetria** e la **curtosi** (pesantezza delle code).

Si presenteranno alcuni **indici sintetici** che descrivono la posizione, la variabilità, la simmetria e la curtosi di una variabile statistica.

Nel caso in cui i dati derivino da un indagine campionaria, gli indici vengono detti indici campionari.


## Indici di posizione: media aritmetica

Un aspetto rilevante dei dati è rappresentato dal suo **centro**, cioè dal punto attorno al quale le modalità osservate si dispongono.

Un **indice di posizione** è espresso nell'ordine di grandezza di  $Y$  e individua tale centro, che costituisce, in alcuni casi, il baricentro della associata distribuzione di frequenza.

La **media aritmetica**, che è l'indice di posizione più noto, si può calcolare per una **variabile quantitativa**  $Y$  e si indica con  $E(Y)$ , con  $\mu_Y$  o semplicemente con  $\mu$ .

**Esempio.** Sia  $Y = (27, 30, 30)$  la variabile statistica che descrive i voti riportati in tre esami da uno studente. La media aritmetica dei voti è  $\mu_Y = (27 + 30 + 30)/3 = 29$ . Si noti che 29 non corrisponde a nessuno dei voti ottenuti.

Se  $Y = (28, 30, 30)$ , allora  $\mu_Y = (28 + 30 + 30)/3 = 29.3$ , che non corrisponde a nessuna potenziale modalità per  $Y$ . In entrambi i casi la media sintetizza i valori osservati indicandone un centro. 

Se si dispone dei **dati grezzi**  $y_1, \dots, y_n$ , allora

$$E(Y) = \frac{1}{n} \sum_{i=1}^n y_i.$$

Se, con riferimento ad una *variabile quantitativa discreta*  $Y$ , si dispone della **tabella di frequenza assoluta o relativa**, allora

$$E(Y) = \frac{1}{n} \sum_{j=1}^J y_j f_j = \sum_{j=1}^J y_j p_j.$$

Se, con riferimento ad una *variabile quantitativa continua*  $Y$ , si dispone della **tabella di frequenza assoluta o relativa con modalità raggruppate in classi** (ad esempio  $y_{j-1} \dashv y_j$ ,  $j = 1, \dots, J$ ), si calcola il punto centrale  $y_j^c = (y_{j-1} + y_j)/2$ ,  $j = 1, \dots, J$ , delle singole classi e

$$E(Y) = \frac{1}{n} \sum_{j=1}^J y_j^c f_j = \sum_{j=1}^J y_j^c p_j.$$

Se ci sono classi aperte, il punto centrale viene individuato dopo aver convenientemente “chiuso la classe”.

Questa procedura approssimata per il calcolo di  $E(Y)$  è equivalente a quella che si definisce quando si dispone dei dati grezzi se viene soddisfatta una delle seguenti ipotesi:

- le osservazioni che cadono in una classe coincidono con il punto centrale della classe;
- le osservazioni sono distribuite in modo uniforme nella classe di appartenenza.

Non è detto che  $E(Y)$  coincida con una delle modalità osservate o osservabili. Può essere vista anche come il valore di equiripartizione sulle unità statistiche del totale delle osservazioni.

La media aritmetica risente della presenza di osservazioni anomale o estreme (non è un indice robusto).

Esistono altre tipologie di medie, che non vengono considerate in questa sede.

**Esempio.** Si consideri la seguente tabella di frequenza

$y_j$	0	1	2	3	4	Totale
$f_j$	109	65	22	3	1	200

È immediato concludere che  $E(Y) = 122/200 = 0.61$ .



**Esempio.** Si consideri la seguente tabella di frequenza con modalità raggruppate in classi

Classe	0 + 10	10 + 15	15 + 20	Totale
freq. rel.	0.30	0.52	0.18	1

I valori centrali delle classi sono, rispettivamente,  $y_1^c = 5$ ,  $y_2^c = 12.5$  e  $y_3^c = 17.5$ , da cui si conclude che

$$E(Y) = 5 \cdot 0.30 + 12.5 \cdot 0.52 + 17.5 \cdot 0.18 = 11.15.$$



**Esempio.** Un lavoratore può raggiungere il luogo di lavoro in bicicletta o in automobile. Vorrebbe scegliere il mezzo di trasporto che gli consente il maggiore risparmio di tempo.

Con questo obiettivo, va a lavorare per 12 giorni in automobile e per 12 giorni in bicicletta e registra il tempo impiegato (in minuti):

Automobile  $X = (23, 32, 44, 21, 36, 30, 28, 33, 45, 34, 29, 31)$

Bicicletta  $Y = (22, 24, 22, 33, 26, 31, 24, 28, 32, 31, 37, 24)$

Il tempo minore in assoluto si ha in auto (21 minuti), ma sempre in auto si ha anche il tempo maggiore in assoluto (45 minuti).

Calcolando le medie aritmetiche si ha  $E(X) = 32.17$  e  $E(Y) = 27.83$ , quindi per raggiungere il posto di lavoro, in media, si impiega meno tempo in bicicletta.





**Esempio. Polveri sottili.** Si vuole studiare l'emissione di polveri sottili (PM), in grammi per 5 litri, per  $n = 13$  veicoli a gasolio.

I dati grezzi vengono riportati nella seguente tabella, dove si individuano anche i veicoli con un alto chilometraggio (A) e un basso chilometraggio (B)

Veicolo	1	2	3	4	5	6	7	8
Km	B	A	A	B	B	B	A	B
PM	2.30	2.15	3.50	2.60	2.75	2.82	4.05	2.25

Veicolo	9	10	11	12	13
Km	B	A	A	B	B
PM	2.68	3.00	4.02	2.85	3.38

La media aritmetica risulta essere  $E(Y) = 38.35/13 = 2.95$  gr/5 lt.

Se al posto dell'osservazione  $y_{11} = 4.02$  si avesse  $y_{11} = 40.2$ , ad esempio per effetto di un errore nella rilevazione, la media risulterebbe pari a  $E(Y) = 74.53/13 = 5.73$  gr/5 lt, che è quasi il doppio del valore precedente.

**Commento:** la media aritmetica è sensibile alla presenza di valori anomali, che possono essere legati a errori oppure essere valori estremi non dovuti ad errori. ◇

La media aritmetica non si calcola per dati categoriali. Un'eccezione si ha con **variabili dicotomiche**, le cui modalità si possono codificare come 0 e 1. In questo caso la media fornisce la proporzione di esiti 1 sul totale delle osservazioni.

**Esempio.** *Polveri sottili* (continua). Nel caso dello studio sulle emissioni di polveri sottili, se si considera la variabile dicotomica  $X$ , con modalità  $1 = B$  e  $0 = A$ , si ha che  $E(X) = 8/13 = 0.615$ .

Quindi, il 61.5% dei veicoli hanno un basso chilometraggio. ◇

La media aritmetica soddisfa le seguenti **proprietà**.

**1) Proprietà di Cauchy.** Sia  $S_Y = \{y_1, \dots, y_J\}$ , con  $y_1 < \dots < y_J$ , allora

$$y_1 \leq E(Y) \leq y_J.$$

La media è compresa tra il più piccolo e il più grande valore osservato. Infatti, per ogni  $j = 1, \dots, J$

$$\begin{aligned} y_1 \leq y_j \leq y_J &\Rightarrow y_1 p_j \leq y_j p_j \leq y_J p_j \Rightarrow \\ \sum_{j=1}^J y_1 p_j &\leq \sum_{j=1}^J y_j p_j \leq \sum_{j=1}^J y_J p_j \Rightarrow \\ y_1 \sum_{j=1}^J p_j &\leq \sum_{j=1}^J y_j p_j \leq y_J \sum_{j=1}^J p_j, \end{aligned}$$

da cui si ottiene la tesi, poiché  $\sum_{j=1}^J p_j = 1$ .

**2) Proprietà di baricentro.** Sia  $Y - E(Y)$  la variabile scarto di  $Y$  dalla sua media  $E(Y)$ , allora

$$E(Y - E(Y)) = 0.$$

Infatti, considerando i dati grezzi e le modalità osservate  $y_i - E(Y)$ ,  $j = 1, \dots, J$ , della variabile  $Y - E(Y)$ ,

$$\begin{aligned} E(Y - E(Y)) &= \frac{1}{n} \sum_{i=1}^n (y_i - E(Y)) = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n E(Y) \\ &= E(Y) - \frac{1}{n} n E(Y) = 0. \end{aligned}$$

**3) Proprietà di linearità.** Sia  $aY + b$ ,  $a, b \in \mathbf{R}$ , una trasformata lineare della variabile  $Y$ , allora

$$E(aY + b) = aE(Y) + b.$$

Infatti, considerando i dati grezzi e le modalità osservate  $ay_i + b$ ,  $j = 1, \dots, J$ , della variabile  $aY + b$ ,

$$\begin{aligned} E(aY + b) &= \frac{1}{n} \sum_{i=1}^n (ay_i + b) = \frac{1}{n} \sum_{i=1}^n ay_i + \frac{1}{n} \sum_{i=1}^n b \\ &= a \frac{1}{n} \sum_{i=1}^n y_i + \frac{1}{n} nb = aE(Y) + b. \end{aligned}$$

## Indici di posizione: mediana

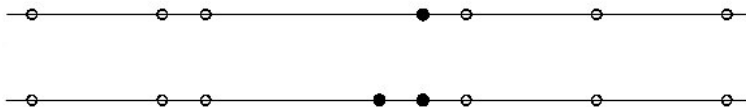
La **mediana** si può calcolare per una **variabile qualitativa ordinale** o **quantitativa**  $Y$  e si indica con  $y_{0.5}$ . È quel valore che, rispetto all'ordinamento non decrescente delle osservazioni, le divide in due parti uguali. È il valore centrale.

La mediana corrisponde a ogni valore  $y_{0.5}$  tale che:

- almeno il 50% delle unità statistiche presenta modalità inferiori o pari a  $y_{0.5}$ ;
- almeno il 50% delle unità statistiche presenta modalità superiori o pari a  $y_{0.5}$ .

Se si dispone dei **dati grezzi**  $y_1, \dots, y_n$ , ordinati secondo un *ordinamento non decrescente*, allora la mediana di  $y_{0.5}$  corrisponde

- alla modalità che si trova nella posizione  $(n+1)/2$ , se  $n$  è **dispari**, cioè  $y_{0.5} = y_{(n+1)/2}$ ;
- alle modalità che si trovano nelle posizioni  $n/2$  e  $(n/2) + 1$ , se  $n$  è **pari**, cioè  $y_{0.5} = y_{n/2}$  e  $y_{0.5} = y_{(n/2)+1}$ .



Se  $y_{n/2}$  e  $y_{(n/2)+1}$  non coincidono, la mediana può non essere unica.

Nel caso di variabili quantitative con  $n$  pari, si può avere anche un intervallo di valori  $[y_{n/2}, y_{(n/2)+1}]$  che soddisfano alla definizione di mediana. In questo caso si può prendere il punto di mezzo come **mediana convenzionale**.

**Esempio. Voti.** Si consideri la variabile statistica qualitativa ordinale  $Y$  che descrive il voto di  $n = 5$  studenti,

$$Y = (\text{sufficiente}, \text{sufficiente}, \text{buono}, \text{buono}, \text{ottimo}).$$

Poiché  $n$  è dispari,  $y_{0.5} = y_{(n+1)/2} = y_3 = \text{buono}$ .

Se invece

$$Y = (\textit{sufficiente}, \textit{sufficiente}, \textit{sufficiente}, \textit{buono}, \textit{buono}, \textit{ottimo}),$$

$n$  è pari, quindi  $y_{0.5} = y_{n/2} = y_3 = \textit{sufficiente}$  e  $y_{0.5} = y_{(n/2)+1} = y_4 = \textit{buono}$ .

Infine, se

$$Y = (\textit{sufficiente}, \textit{sufficiente}, \textit{buono}, \textit{buono}, \textit{ottimo}, \textit{ottimo}),$$

$n$  è pari, ma  $y_{0.5} = y_{n/2} = y_3 = \textit{buono}$  e  $y_{0.5} = y_{(n/2)+1} = y_4 = \textit{buono}$ ;  
quindi, *buono* è l'unica mediana.  $\diamond$



**Esempio. Serie TV.** Si consideri la variabile statistica quantitativa discreta  $Y$  che descrive il numero di puntate, di una serie televisiva, viste da  $n = 8$  famiglie

$$Y = (0, 1, 3, 3, 4, 6, 6, 6).$$

Poiché  $n$  è pari,  $y_{0.5} = y_{n/2} = y_4 = 3$  e  $y_{0.5} = y_{(n/2)+1} = y_5 = 4$ . In questo caso, sia 3 che 4 sono valori mediani e, in generale, ogni punto dell'intervallo  $[3, 4]$  è un valore mediano; la mediana convenzionale è 3.5.

Se invece

$$Y = (0, 1, 3, 3, 4, 6, 6),$$

$n$  è dispari e c'è un'unica mediana  $y_{0.5} = y_{(n+1)/2} = y_4 = 3$ .



Se si dispone soltanto della **distribuzione di frequenza relativa o assoluta**, si può operare nel seguente modo.

Si suppone che le modalità del supporto  $S_Y = \{y_1, \dots, y_J\}$  siano ordinate in senso crescente.

Se sono note le **frequenze assolute**  $f_j$ ,  $j = 1, \dots, J$ , e quindi la dimensione  $n$  della popolazione, la mediana corrisponde,

- se  $n$  è **dispari**, alla modalità  $y_j$  che presenta la frequenza assoluta cumulata  $F_j$  più piccola tale che  $F_j \geq (n + 1)/2$ ;
- se  $n$  è **pari**, alla modalità  $y_j$  che presenta la frequenza assoluta cumulata  $F_j$  più piccola tale che  $F_j \geq n/2$  e alla modalità  $y_j$  che presenta la frequenza assoluta cumulata  $F_j$  più piccola tale che  $F_j \geq (n/2) + 1$ .

Nel caso con  $n$  pari si possono avere due valori mediani distinti o più di due, se si considerano variabili quantitative.

Se sono note solo le **frequenze relative**  $p_j$ ,  $j = 1, \dots, J$ , e quindi la dimensione  $n$  della popolazione non risulta nota, allora la mediana corrisponde alla modalità  $y_j$  che presenta frequenza relativa cumulata più piccola tale che  $P_j \geq 0.5$ .

Se esiste una modalità  $y_j$  tale che  $P_j = 0.5$ , allora sia  $y_j$  che  $y_{j+1}$  soddisfano la definizione di mediana.

Per l'individuazione della mediana è utile, in questo contesto, l'analisi della funzione di ripartizione empirica, con particolare riferimento al livello 0.5.

La mediana è un indice di posizione robusto rispetto a valori anomali dei dati.

Se si dispone soltanto della **distribuzione di frequenza relativa o assoluta con modalità raggruppate in classi**, si può operare allo stesso modo.

Quindi, si individuerà una **classe mediana**.

**Esempio.** Sia  $Y$  la variabile quantitativa discreta che descrive il numero di componenti delle famiglie residenti in Liguria alla data del Censimento 1981. Si riporta la associata tabella di frequenza.

No. componenti	$f$	$F$	$p$	$P$
1	197906	197906	0.272	0.272
2	203709	401615	0.281	0.553
3	168536	570151	0.232	0.785
4	117509	687660	0.162	0.947
5	29727	717387	0.041	0.988
6	6577	723964	0.009	0.997
7	1707	725671	0.002	0.999
8 o più	906	726577	0.001	1
Totale	726577		1	

Poiché  $n = 726577$  è dispari, la mediana è unica e corrisponde alla modalità della famiglia che si trova nella posizione  $(n + 1)/2 = 363289$ , dopo avere ordinato le famiglie secondo il numero crescente di componenti. Quindi  $y_{0.5} = 2$ .

Si noti che a 2 corrisponde la frequenza assoluta cumulata più piccola che risulta maggiore o uguale a  $(n + 1)/2 = 363289$ .

Se si prendono in esame le frequenze relative cumulate,  $y_{0.5} = 2$  è la modalità che presenta frequenza relativa cumulata più piccola tale che  $P_j \geq 0.5$

Si consideri la tabella di frequenza riferita alla regione Campania

No. componenti	$f$	$F$	$p$	$P$
1	225641	225641	0.144	0.144
2	304325	529966	0.194	0.338
3	278879	808845	0.178	0.516
4	355488	1164333	0.226	0.742
5	228494	1392827	0.146	0.888
6	98924	1491751	0.063	0.951
7	42894	1534645	0.027	0.978
8 o più	34999	1569644	0.022	1
Totale	1569644		1	

Poiché  $n = 1569644$  è pari, la mediana corrisponde alla modalità della famiglia che si trova nella posizione  $n/2 = 784822$  e alla modalità della famiglia che si trova nella posizione  $(n/2) + 1 = 784823$ , dopo avere ordinato le famiglie secondo il numero crescente di componenti.

Tali famiglie presentano la stessa modalità 3, quindi  $y_{0.5} = 3$ .

Si noti che a 3 corrisponde la frequenza assoluta cumulata più piccola che risulta maggiore o uguale sia a  $n/2 = 784822$  che a  $(n/2) + 1 = 784823$ .

Se si considerano le frequenze relative cumulate,  $y_{0.5} = 3$  è la modalità che presenta frequenza relativa cumulata più piccola tale che  $P_j \geq 0.5$  ◇

**Esempio.** *Polveri sottili* (continua). Nel caso dello studio sulle emissioni di polveri sottili, le osservazioni ordinate in senso crescente sono

2.15 2.25 2.30 2.60 2.68 2.75 **2.82** 2.85 3.00 3.38 3.50 4.02 4.05

La mediana è l'osservazione in posizione  $(n + 1)/2 = 7$ , ovvero  $y_{0.5} = 2.82$ .

Se al posto dell'osservazione  $y_{11} = 4.02$  si avesse l'osservazione  $y_{11} = 40.2$ , l'ordine dei dati cambierebbe di poco e la mediana rimarrebbe  $y_{0.5} = 2.82$ .



La mediana è un indice di posizione robusto, cioè poco sensibile ai valori anomali.

Data una variabile statistica, è utile e conveniente calcolare, se possibile, sia la media che la mediana.

# Quantili

La mediana può venire interpretata come un particolare **quantile di livello**  $\alpha$ , con  $\alpha \in (0, 1)$ , indicato con la scrittura  $y_\alpha$ .

Data una **variabile qualitativa ordinale o quantitativa**  $Y$ ,  $y_\alpha$  è quel valore che, rispetto all'ordinamento non decrescente delle osservazioni, risulta preceduto da  $\alpha 100\%$  osservazioni e seguito da  $(1 - \alpha) 100\%$  osservazioni, a meno degli effetti di discretezza.

Il quantile di livello  $\alpha$ , con  $\alpha \in (0, 1)$ , di una variabile statistica  $Y$  corrisponde a ogni valore  $y_\alpha$  tale che:

- almeno  $\alpha 100\%$  unità statistiche presenta modalità inferiori o pari a  $y_\alpha$ ;
- almeno  $(1 - \alpha) 100\%$  unità statistiche presenta modalità superiori o pari a  $y_\alpha$ .

È evidente che, se  $\alpha = 0.5$ , si ottiene la definizione di mediana.



I quantili di livello  $\alpha = 0.25, 0.5, 0.75$  vengono chiamati **quartili** e dividono le osservazioni ordinate in 4 parti uguali.

I quantili di livello  $\alpha = 0.10, 0.20, \dots, 0.90$  vengono chiamati **decili** e dividono le osservazioni ordinate in 10 parti uguali.

I quantili di livello  $\alpha = 0.01, 0.02, \dots, 0.99$  vengono chiamati **percentili** e dividono le osservazioni ordinate in 100 parti uguali.

Se si dispone dei **dati grezzi**  $y_1, \dots, y_n$ , ordinati secondo un ordinamento non decrescente, per individuare  $y_\alpha$  si calcola  $\alpha(n+1)$ .

- Se si ottiene un valore intero,  $y_\alpha$  è la modalità che si trova nella posizione  $\alpha(n+1)$ , cioè  $y_\alpha = y_{\alpha(n+1)}$ .
- Se si ottiene un valore non intero,  $y_\alpha$  corrisponde alle modalità che hanno come posizioni i valori interi che seguono e precedono immediatamente  $\alpha(n+1)$ .

Se  $\alpha = 0.5$ , la procedura è equivalente a quella utilizzata per il calcolo della mediana in presenza di dati grezzi.

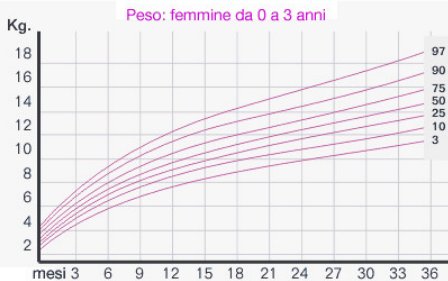
Se tali valori non coincidono, il quantile può non essere unico.

Nel caso di variabili quantitative si può avere anche un intervallo di valori che soddisfano alla definizione di quantile. In questo caso si può prendere il punto di mezzo come **quantile convenzionale**.

Con riferimento alla nozione di quantile si possono fare considerazioni analoghe a quelle introdotte per la mediana, per quanto riguarda il calcolo nel caso in cui si disponga solo della **distribuzione di frequenza**, assoluta o relativa, e la situazione con modalità raggruppate in classi.

Il livello di riferimento non sarà più 0.5 ma, in generale, il valore  $\alpha$ .

Una applicazione dei percentili riguarda la definizione di curve di crescita, come evidenziato nel grafico seguente.



**Esempio.** *Asfalto.* Si considerano i dati relativi ai valori di resistenza alla rottura di  $n = 24$  misture di asfalto (in megapascal)

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

La media aritmetica corrisponde a  $E(Y) = 195.4$ . Per il calcolo dei quantili di livello  $\alpha$  si segue la procedura indicata in precedenza, che richiede in via preliminare il calcolo di  $\alpha(n + 1)$ .

Avendo ordinato le osservazioni in senso crescente, si ottengono i quartili (convenzionali)

$$0.25(24 + 1) = 6.25 \Rightarrow y_{0.25} = (y_6 + y_7)/2 = 115.5$$

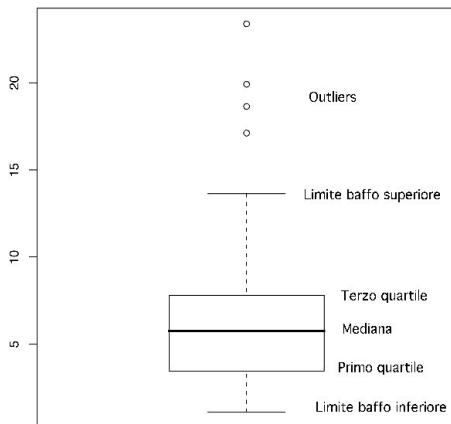
$$0.5(24 + 1) = 12.5 \Rightarrow y_{0.5} = (y_{12} + y_{13})/2 = 207$$

$$0.75(24 + 1) = 18.75 \Rightarrow y_{0.75} = (y_{18} + y_{19})/2 = 243.5$$



# Boxplot

Il **diagramma a scatola e baffi** (**box and whiskers plot**), abbreviato spesso in **boxplot**, fornisce una sintesi grafica efficace dell'insieme di dati basata sui quantili.



La **scatola** contiene il 50% centrale della distribuzione di frequenza ed è delimitata dal primo quartile  $y_{0.25}$  e dal terzo quartile  $y_{0.75}$ .

In corrispondenza della mediana  $y_{0.5}$  viene tracciata una **linea**.

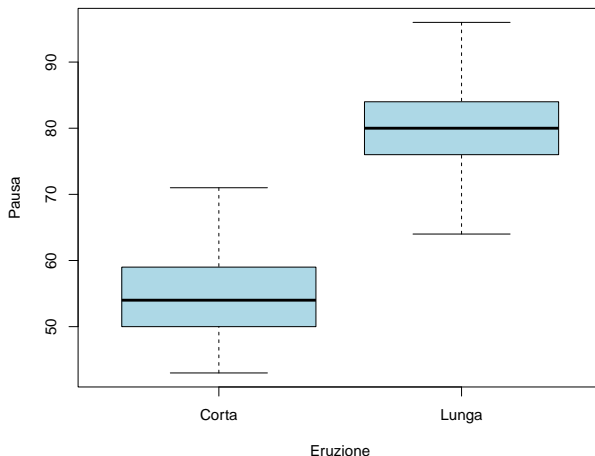
I **baffi** si prolungano fino al valore minimo e massimo osservati o fino ai percentili  $y_{0.01}$  e  $y_{0.99}$ . La lunghezza dei baffi può venire specificata in modo alternativo con l'obiettivo di far emergere potenziali valori anomali (**outliers**).

Esistono diverse varianti a seconda del software utilizzato.

Il boxplot è un grafico molto efficace, soprattutto quando si vogliono fare comparazioni tra due o più insiemi di dati.

**Esempio.** *Geyser Old Faithful* (continua). Si considerano i dati riferiti alle durate delle eruzioni del geyser Old Faithful.

La comparazione dei boxplot, riferiti alle due tipologie di eruzioni prima delle pause, confermano le affermazioni fatte in precedenza.



## Indici di posizione: moda

La **moda** si può calcolare per **variabili qualitative o quantitative**, si indica con  $y_{mo}$  e corrisponde alla modalità che si verifica con maggior frequenza.

La moda di una variabile statistica  $Y$  corrisponde al valore  $y_{mo}$  del supporto  $S_Y$  a cui è associata la frequenza, relativa o assoluta, più alta.

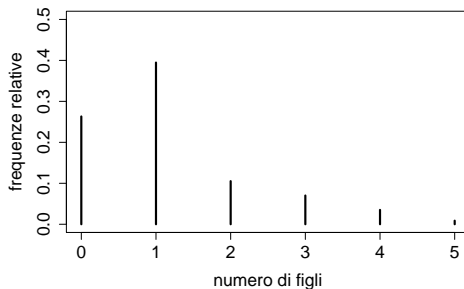
La moda è la modalità più comune e non è detto che sia unica.

La moda è un indice di posizione molto grezzo. Può non individuare il centro dei dati. Ci possono essere distribuzioni **unimodali**, **bimodali** o **multimodali**.

Nel caso in cui si abbia una tabella di frequenza con modalità raggruppate in classi, si può individuare la **classe modale**, soltanto se le classi hanno tutte la stessa ampiezza.



**Esempio. Figli** (continua). Si considera il numero di figli con riferimento alle famiglie residenti in un determinato territorio.



Dalla analisi del diagramma a bastoncini si conclude facilmente che  $y_{mo} = 1$  e la distribuzione è unimodale.



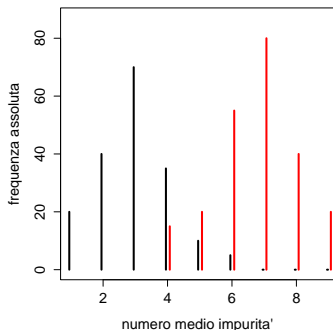
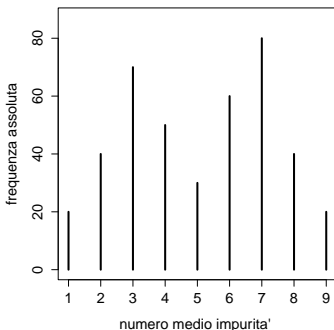
**Esempio. Creta.** Si vuole valutare la qualità della creta proveniente da due diverse cave. A tale scopo si rileva il numero medio di impurità per  $\text{cm}^2$  su 410 vasi, 180 costruiti con la creta della prima cava e 210 con la creta della seconda.

Si considera la distribuzione di frequenza assoluta.

No. medio impurità (per $\text{cm}^2$ )	$f_j$ (cava 1)	$f_j$ (cava 2)	$f_j$ (totale)
1	20	0	20
2	40	0	40
3	70	0	70
4	35	15	50
5	10	20	30
6	5	55	60
7	0	80	80
8	0	40	40
9	0	20	20
Totale	180	230	410

La distribuzione di frequenza, considerando tutti i 410 vasi, risulta bimodale con picchi in 3 e 7, come risulta dal grafico di sinistra.

Dall'analisi del grafico di destra, con le distribuzioni di frequenza distinte per cava di provenienza della creta (**nero**=cava 1, **rosso**=cava 2), si conclude che la bimodalità deriva dal fatto che il materiale delle due cave non ha la stessa qualità.



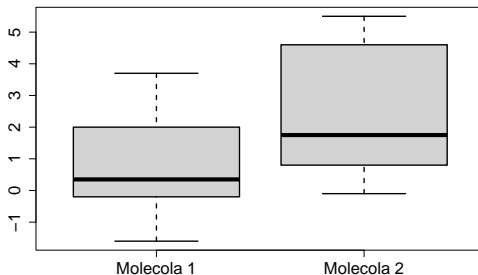
**Esempio. Sonnifero.** Per valutare e confrontare l'effetto come sonnifero di due distinte molecole si sono considerati  $n = 10$  volontari, senza storia pregressa di insonnia, ai quali è stato somministrato in una notte un placebo e in un'altra il sonnifero.

Per ogni soggetto si considera la differenza tra ore di sonno con il sonnifero e con il placebo, distinguendo tra le due molecole.

Soggetto	Ore sonno extra (Molecola 1)	Ore sonno extra (Molecola 2)
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0	4.6
10	2.0	4.6

L'effetto della molecola 1 non è molto chiaro poiché 4 soggetti su 10 hanno dormito di meno, ma 2 su 10 hanno manifestato un aumento di sonno superiore alle 3 ore. La media e la mediana sono, rispettivamente, 0.75 e 0.35 e indicano una moderata efficacia del sonnifero. Il primo e terzo quartile corrispondono a -0.7 e 1.4.

L'effetto della molecola 2 è più deciso. La media e la mediana sono, rispettivamente, 2.45 e 1.75. Il primo e terzo quartile corrispondono a 0.45 e 4.5. I boxplot confermano l'analisi.



## Indici di variabilità

L'individuazione del centro di un insieme di dati, tramite opportuni indici di posizione, può non essere sufficiente per descrivere in modo completo la associata distribuzione di frequenza.

**Esempio.** *Inquinamento.* Per confrontare l'efficacia di due diversi dispositivi per contenere l'inquinamento atmosferico si sono analizzati i fumi prodotti da una certa industria.

Si sono considerati 360 campioni di fumo e si è misurata la quantità di pulviscolo inquinante in g/min. In 180 si è utilizzato il dispositivo anti-inquinante A, mentre sui campioni rimanenti si è utilizzato il dispositivo anti-inquinante B. Si hanno i seguenti dati:

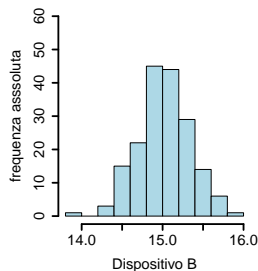
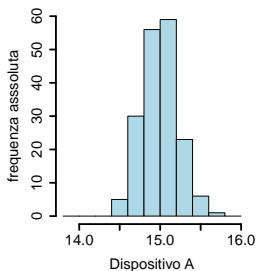
Dispositivo A: 14.98654, 15.14828, 15.15741, ..., 14.92036, 14.93270

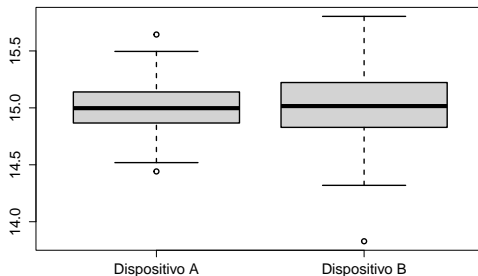
Dispositivo B: 14.62067, 15.26097, 14.87602, ..., 14.74481, 15.13047

Si calcolano alcuni valori di sintesi

Dispositivo	$y_{(1)}$	$y_{0.25}$	$y_{0.5}$	$\mu$	$y_{0.75}$	$y_{(n)}$
A	14.44	14.87	15.00	15.00	15.14	15.64
B	13.83	14.83	15.02	15.02	15.22	15.80

Entrambi i dispositivi sembrano tarati allo stesso modo, poiché forniscono valori centrati sul valore di 15 g/min.





Dalla analisi dei grafici si nota che gli scostamenti dal valore centrale sembrano essere più elevati per il dispositivo B.

I dati riferiti a B sono più dispersi attorno al valore centrale, cioè mostrano una variabilità più accentuata.






Si presentano i seguenti **indici di variabilità** utili per **variabili quantitative**:

- **campo di variazione**;
- **scarto interquartilico**;
- **varianza (e scarto quadratico medio)**;
- **coefficiente di variazione**.

Non si considerano gli indici di variabilità per **variabili qualitative**, detti anche **indici di mutabilità**.

La variabilità di una variabile statistica si traduce nella diversificazione delle modalità osservate. Se  $Y$  è quantitativa, tale diversificazione si intende sia come **diversità** di valori osservati sia come **distanza** fra essi.

**Esempio.** Se  $Y$  è una variabile statistica degenere,  $S_Y = \{y_1\}$  e la sua variabilità è nulla. Se  $Y_1 = (1, 1, 1, 2, 2)$  e  $Y_2 = (1, 1, 1, 10, 10)$ , i due supporti corrispondenti contengono due modalità, ma la variabilità di  $Y_2$  è più accentuata di quella di  $Y_1$ . 

## Indici di variabilità: campo di variazione

Sia  $Y$  una variabile statistica quantitativa. Il **campo di variazione** (**range**) corrisponde a

$$R_Y = y_{(n)} - y_{(1)},$$

ossia alla differenza tra l'osservazione più grande e l'osservazione più piccola.

Se  $Y$  è degenere,  $R_Y = 0$ ; altrimenti  $R_Y > 0$ .

**Esempio.** *Inquinamento* (continua). Con riferimento ai dati sulla misurazione dell'inquinamento da fumi si ottiene che

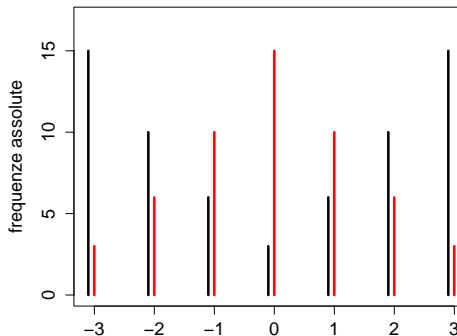
$$\text{Dispositivo A: } R = 15.64 - 14.40 = 1.20$$

$$\text{Dispositivo B: } R = 15.80 - 13.83 = 1.97$$

Questi valori indicano una maggiore variabilità nelle misurazioni riferite alla metodica B.



$R_Y$  è sensibile alla presenza di valori anomali. Inoltre è un indice piuttosto povero, come dimostra il seguente grafico dove si considerano le distribuzioni di frequenza di due diverse variabili statistiche (rappresentate in **nero** e in **rosso**) con lo stesso campo di variazione ma diversa variabilità.



## Indici di variabilità: scarto interquartilico

Lo **scarto interquartilico** di una variabile statistica quantitativa  $Y$  corrisponde a

$$SI_Y = y_{0.75} - y_{0.25},$$

ossia alla differenza tra il terzo e il primo quartile.

$SI_Y$  esprime la lunghezza dell'intervallo dove cade il 50% centrale della distribuzione di frequenza. Nel boxplot corrisponde all'ampiezza della scatola

L'indice  $SI_Y$  può essere nullo anche per variabili non degeneri; ad esempio, si annulla per  $Y = (1, 2, 2, 2, 2, 2, 5)$ , poiché  $y_{0.75} = y_{0.25} = 2$ . In presenza di poche osservazioni anomale ha proprietà di robustezza.

**Esempio.** *Inquinamento* (continua). Con riferimento ai dati sulla misurazione dell'inquinamento da fumi, si ottiene che, con il dispositivo A,  $SI = 0.27$ , mentre, con il dispositivo B,  $SI = 0.39$ . Questi valori indicano ancora una maggiore variabilità nelle misurazioni riferite al dispositivo B.



## Indici di variabilità: varianza

Il più importante indice di variabilità per variabili quantitative è la **varianza**, che si indica con  $V(Y)$ , con  $\sigma_Y^2$  o semplicemente con  $\sigma^2$ .

Data una variabile statistica  $Y$  con media aritmetica  $E(Y)$ , si ha

$$V(Y) = E[(Y - E(Y))^2].$$

La varianza è la media aritmetica della variabile scarto  $Y - E(Y)$  elevata al quadrato e misura la dispersione dei dati attorno alla media. L'unità di misura è pari a quella dei dati elevata al quadrato.

Per il calcolo di  $V(Y)$ , si può riprendere quanto detto con riferimento alla media aritmetica.

Lo **scarto quadratico medio** di  $Y$ , indicato con  $\sigma_Y$  o con  $\sigma$ , è la radice quadrata aritmetica (l'unica positiva) della varianza

$$\sigma_Y = \sqrt{V(Y)};$$

è nella stessa unità di misura di  $Y$ .

Se si dispone dei **dati grezzi**  $Y = (y_1, \dots, y_n)$ , e si è preventivamente calcolata  $E(Y)$ , allora la varianza corrisponde a

$$V(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - E(Y))^2.$$

Se si dispone della **distribuzione di frequenza assoluta o relativa**,

$$V(Y) = \frac{1}{n} \sum_{j=1}^J (y_j - E(Y))^2 f_j = \sum_{j=1}^J (y_j - E(Y))^2 p_j.$$

Se si dispone della **distribuzione di frequenza assoluta o relativa con modalità raggruppate in classi** (ad esempio  $y_{j-1} \vdash y_j$ ,  $j = 1, \dots, J$ ), è necessario calcolare il *punto centrale*  $y_j^c = (y_{j-1} + y_j)/2$ ,  $j = 1, \dots, J$ , delle singole classi.

In questo caso

$$V(Y) = \frac{1}{n} \sum_{j=1}^J (y_j^c - E(Y))^2 f_j = \sum_{j=1}^J (y_j^c - E(Y))^2 p_j.$$

Questa procedura approssimata per il calcolo di  $V(Y)$  è equivalente a quella che si definisce quando si dispone dei dati grezzi se le osservazioni che cadono in una classe coincidono con il punto centrale della classe.

**Esempio.** *Inquinamento* (continua). Con riferimento ai dati sulla misurazione dell'inquinamento da fumi, si ottiene che, con il dispositivo A,  $\sigma^2 = 0.046$ , mentre, con il dispositivo B,  $\sigma^2 = 0.099$ . Questi valori indicano ancora una maggiore variabilità nelle misurazioni riferite al dispositivo B. ◇

**Esempio.** Si consideri la seguente tabella di frequenza dalla quale si ricava che  $E(Y) = 0.61$

$y_j$	0	1	2	3	4	Totale
$f_j$	109	65	22	3	1	200

È immediato concludere che

$$V(Y) = \frac{1}{200} [(0 - 0.61)^2 \cdot 109 + (1 - 0.61)^2 \cdot 65 + (2 - 0.61)^2 \cdot 22 + (3 - 0.61)^2 \cdot 3 + (4 - 0.61)^2 \cdot 1] = 0.608.$$

Inoltre,  $\sigma_Y = \sqrt{V(Y)} = 0.780$ .





**Esempio.** Si consideri la seguente tabella di frequenza con modalità raggruppate in classi

Classe	0 + 10	10 + 15	15 + 20	Totale
freq. rel.	0.30	0.52	0.18	1

I valori centrali delle classi sono, rispettivamente,  $y_1^c = 5$ ,  $y_2^c = 12.5$  e  $y_3^c = 17.5$ , da cui si conclude che

$$\begin{aligned}
 V(Y) &= (5 - 11.15)^2 \cdot 0.30 + (12.5 - 11.15)^2 \cdot 0.52 \\
 &\quad + (17.5 - 11.15)^2 \cdot 0.18 = 19.55.
 \end{aligned}$$



La varianza soddisfa le seguenti **proprietà**.

1) **Proprietà di non negatività.**  $V(Y) \geq 0$ , con  $V(Y) = 0$  se e solo se  $Y$  è degenere.

2) **Formula per il calcolo.**

$$V(Y) = E(Y^2) - (E(Y))^2.$$

Infatti, sfruttando la proprietà di linearità della media aritmetica,

$$\begin{aligned} V(Y) &= E\{(Y - E(Y))^2\} = E\{Y^2 + (E(Y))^2 - 2YE(Y)\} \\ &= E(Y^2) + (E(Y))^2 - 2E(Y)E(Y) = E(Y^2) - (E(Y))^2. \end{aligned}$$

### 3) Proprietà di invarianza per traslazioni.

$$V(Y + b) = V(Y), \quad b \in \mathbf{R}.$$

Infatti, sfruttando la proprietà di linearità della media aritmetica,

$$\begin{aligned} V(Y + b) &= E\{(Y + b - E(Y + b))^2\} = E\{(Y + b - E(Y) - b)^2\} \\ &= E\{(Y - E(Y))^2\} = V(Y). \end{aligned}$$

### 4) Proprietà di omogeneità di secondo grado.

$$V(aY) = a^2 V(Y), \quad a \in \mathbf{R}.$$

Infatti, sfruttando la proprietà di linearità della media aritmetica,

$$\begin{aligned} V(aY) &= E\{(aY - E(aY))^2\} = E\{(aY - aE(Y))^2\} \\ &= E\{a^2(Y - E(Y))^2\} = a^2 E\{(Y - E(Y))^2\} = a^2 V(Y). \end{aligned}$$

Dalla 2) discende che, se  $E(Y) = 0$  allora  $V(Y) = E(Y^2)$ .

Dalla 3) e dalla 4) discende che  $V(aY + b) = a^2V(Y)$ .

**Esempio.** Si consideri la tabella di frequenza vista in precedenza, dalla quale si è ricavato che  $E(Y) = 0.61$

$y_j$	0	1	2	3	4	Totale
$f_j$	109	65	22	3	1	200

È immediato concludere che

$$E(Y^2) = \frac{1}{200} [0^2 \cdot 109 + 1^2 \cdot 65 + 2^2 \cdot 22 + 3^2 \cdot 3 + 4^2 \cdot 1] = 0.98.$$

Usando la formula per il calcolo, si ottiene che

$$V(Y) = E(Y^2) - (E(Y))^2 = 0.98 - 0.61^2 = 0.608,$$

che coincide con il valore ottenuto con la definizione di varianza.  $\diamond$

## Indici di variabilità: coefficiente di variazione


Con riferimento a **variabili statistiche che assumono solo valori positivi** si può introdurre un indice adimensionale di variabilità detto **coefficiente di variazione**

$$CV_Y = \frac{\sigma_Y}{\mu_Y}.$$

È un indice di variabilità relativa, nel senso che misura la variabilità dei dati tenendo conto dell'ordine di grandezza del fenomeno.

Essendo un numero puro, permette il confronto tra insiemi di dati diversi, ad esempio, con unità di misura diverse o con valori medi molto distanti.

Per variabili non necessariamente positive si considera il valore assoluto della media e quindi  $CV_Y = \sigma_Y / |\mu_Y|$ .

**Esempio.** *Inquinamento* (continua). Con riferimento ai dati sulla misurazione dell'inquinamento, si ottiene che con il dispositivo A  $CV = \sqrt{0.046}/15 = 0.014$ , mentre con il dispositivo B si ottiene, come prevedibile, un valore più elevato  $CV = \sqrt{0.099}/15.02 = 0.021$ . 

**Esempio.** Si consideri la seguente tabella di frequenza che riporta le merci e i passeggeri sbarcati, con riferimento agli scali portuali di alcune regioni italiane nel 1988.

Regione	Merci (migliaia di tonnellate)	Passeggeri (migliaia)
Friuli V.-G.	22806	42
Veneto	21849	248
Emilia-R.	12627	3
Marche	4937	266

Ci si chiede se è più variabile lo sbarco di merci (variabile  $X$ ) o lo sbarco di passeggeri (variabile  $Y$ ). Poiché  $E(X) = 15554.75$ ,  $V(X) = 53376636$ ,  $E(Y) = 139.75$ ,  $V(Y) = 13978.19$ , si ha che

$$CV_X = 0.47, \quad CV_Y = 0.85.$$

Nonostante la varianza di  $X$  sia più elevata, risulta maggiore, in termini relativi, la variabilità del numero di passeggeri.  $\diamond$

Una **variabile statistica** con **media nulla** e **varianza unitaria** è detta **standardizzata**.

Dalle proprietà della media e della varianza si conclude che

- data una variabile  $Y$ , allora  $Z = (Y - \mu_Y)/\sigma_Y$  è una variabile standardizzata;
- data una variabile  $Z$  standardizzata, allora la variabile  $Y = \sigma Z + \mu$  è tale che  $E(Y) = \mu$  e  $V(Y) = \sigma^2$ .

A volte è utile trasformare i dati  $y_1, \dots, y_n$  di modo tale che i dati trasformati  $z_1, \dots, z_n$  abbiano media nulla e varianza unitaria. La trasformazione appropriata è  $z_i = (y_i - \mu_Y)/\sigma_Y$ ,  $i = 1, \dots, n$ .

Con la standardizzazione dei dati possono emergere più chiaramente alcune ulteriori caratteristiche della distribuzione di frequenza, oltre la posizione e la variabilità.

La standardizzazione è utile anche per confrontare e analizzare più insiemi di dati, tenendo conto che possono avere con unità di misura e livello medio diversi.

# Simmetria e asimmetria

Una **distribuzione di frequenza** (ad esempio rappresentata con un istogramma o un diagramma a bastoncini) è **simmetrica** se la sua metà di destra si sovrappone alla sua metà di sinistra (dove la metà è identificata dalla mediana).

Un istogramma asimmetrico presenta una coda più lunga dell'altra. Se la coda destra è più lunga, si parla di **asimmetria positiva**, se la coda sinistra è più lunga si ha **asimmetria negativa**.

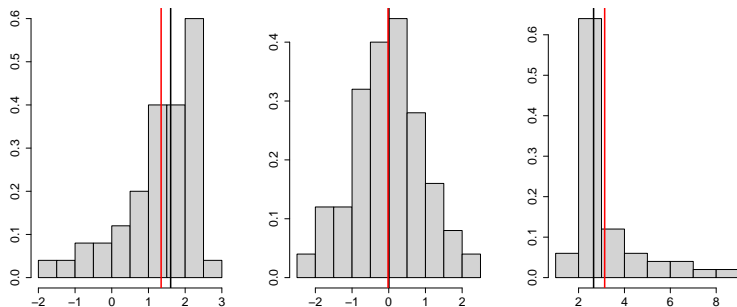
Si noti che:

- se l'asimmetria è positiva:  $\text{media} > \text{mediana}$ ;
- se c'è simmetria:  $\text{media} \approx \text{mediana}$ ;
- se l'asimmetria è negativa:  $\text{media} < \text{mediana}$ .

Per una distribuzione di frequenza unimodale e simmetrica si ha che:  
 $\text{media} \approx \text{mediana} \approx \text{moda}$ .



Nel grafici sottostanti la **linea nera** indica la mediana e la **linea rossa** indica la media.



Nel primo grafico si ha una distribuzione di frequenza con asimmetria negativa, nel terzo una distribuzione di frequenza con asimmetria positiva, mentre nel secondo c'è una sostanziale simmetria.

## Indice di simmetria

Data una variabile statistica quantitativa  $Y$ , con media aritmetica  $E(Y)$ , l'indice **indice di simmetria** più utilizzato è

$$\gamma_Y = \frac{E[(Y - E(Y))^3]}{\sigma_Y^3},$$

dove  $\sigma_Y = \sqrt{V(Y)}$  è lo scarto quadratico medio di  $Y$ .

Se si dispone dei **dati grezzi**  $Y = (y_1, \dots, y_n)$ , e si sono preventivamente calcolati  $E(Y)$  e  $\sigma_Y$ , allora l'indice di simmetria corrisponde a

$$\gamma_Y = \frac{(1/n) \sum_{i=1}^n (y_i - E(Y))^3}{\sigma_Y^3}.$$

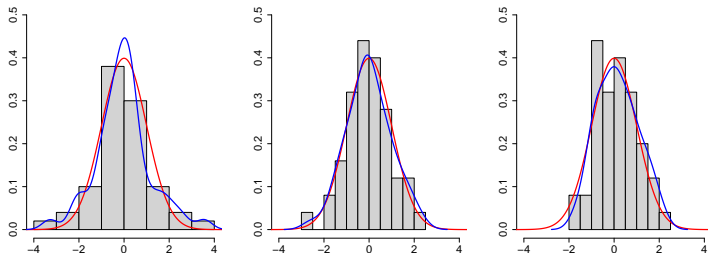
Se la distribuzione di frequenza è **simmetrica**,  $\gamma_Y \approx 0$ ; se c'è **asimmetria negativa**,  $\gamma_Y < 0$ ; se c'è **asimmetria positiva**,  $\gamma_Y > 0$ .

# Curtosi

La **curtosi** corrisponde ad un allontanamento dalla distribuzione di frequenza normale (o gaussiana), che viene considerata come riferimento.

Una **distribuzione platicurtica (ipornormale)** presenta un maggiore appiattimento e *code leggere*, mentre una **distribuzione leptocurtica (ipernormale)** manifesta un maggiore allungamento e *code pesanti*.

**Istogramma**, **densità normale** e **stima della densità** nel caso di distribuzione leptocurtica (sinistra), distribuzione normocurtica (centro) e distribuzione platicurtica (destra).



## Indice di curtosi

Data una variabile statistica quantitativa  $Y$ , con media aritmetica  $E(Y)$ , l'indice **indice di curtosi** più utilizzato è

$$\beta_Y = \frac{E[(Y - E(Y))^4]}{\sigma_Y^4},$$

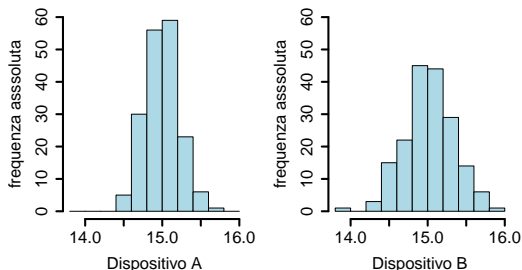
dove  $\sigma_Y = \sqrt{V(Y)}$  è lo scarto quadratico medio di  $Y$ .

Se si dispone dei **dati grezzi**  $Y = (y_1, \dots, y_n)$ , e si sono preventivamente calcolati  $E(Y)$  e  $\sigma_Y$ , allora l'indice di curtosi corrisponde a

$$\beta_Y = \frac{(1/n) \sum_{i=1}^n (y_i - E(Y))^4}{\sigma_Y^4}.$$

Se la distribuzione di frequenza è **normocurtica**,  $\beta_Y \approx 3$ ; se è **leptocurtica**,  $\beta_Y > 3$ ; se è **platicurtica**,  $\beta_Y < 3$ .

**Esempio.** *Inquinamento* (continua). Con riferimento ai dati sui due dispositivi anti-inquinamento, poiché media e mediana coincidono, si conclude che le distribuzioni di frequenza sono simmetriche; i valori dell'indice di simmetria sono pari a 0.095 e  $-0.225$ , rispettivamente.



Dalla analisi dell'istogramma, oltre alla conferma della sostanziale simmetria, si deduce che la seconda distribuzione presenta code più pesanti della prima. Infatti, i valori dell'indice di curtosi sono pari a 2.937 e 3.398, rispettivamente. ◇