

Formulario di Statistica

- Formulario di Statistica
 - Statistica descrittiva
 - I dati
 - Unità statistiche e popolazione
 - Tipi di Statistica
 - Statistica descrittiva
 - Analisi esplorativa
 - Variabili statistiche
 - Distribuzioni di frequenza
 - Frequenza assoluta
 - Frequenza relativa
 - Frequenze cumulative
 - Altri tipi di distribuzioni di frequenza
 - Rappresentazioni grafiche
 - Diagrammi circolari
 - Diagrammi a Barre
 - Diagrammi a bastoncini
 - Istogrammi
 - Poligoni di frequenza
 - Stima della densità
 - Funzione di ripartizione empirica
 - Diagrammi di dispersione
 - Indici sintetici
 - Indici di posizione: media aritmetica
 - Indice di posizione: mediana
 - Quantili
 - Boxplot
 - Indici di posizione: moda
 - Indici di variabilità
 - Indici di variabilità: campo di variazione
 - Indici di variabilità: scarto interquartilico
 - Indici di variabilità: varianza e scarto quadratico medio
 - Indici di variabilità: coefficiente di variazione
 - Simmetria e asimmetria
 - Indice di simmetria
 - Curtosi e indice di curtosi

- Calcolo delle probabilità
- Inferenza statistica
 - Utilizzo
 - Campionamento
 - Modelli statistici parametrici
 - Verifica del modello
 - Procedure inferenziali
 - Statistiche campionarie
 - Somma e media campionaria
 - Varianza campionaria
 - Stima puntuale - Stime
- Formule e Esempi
 - Frequenza relativa
 - Frequenza cumulata
 - Stima della densità
 - Funzione di ripartizione empirica
 - Indici di posizione: media aritmetica
 - Indice di posizione: mediana
 - Indici di posizione: moda
 - Indici di variabilità: campo di variazione
 - Indici di variabilità: scarto interquartilico
 - Indici di variabilità: varianza e scarto quadratico medio
 - Indici di variabilità: coefficiente di variazione
 - Indice di simmetria
 - Curtosi e indice di curtosi
 - Somma e media campionaria
- Comandi in R

Statistica descrittiva

I dati

Unità statistiche e popolazione

I dati rappresentano informazioni di una **popolazione**, ovvero l'intera collezione di unità statistiche sulle quali si cerca l'informazione.

Esistono due tipi di popolazioni:

- **popolazione reale**: unità che hanno una esistenza fisica, sono popolazioni effettive e quindi finite, possono essere osservate in modo completo (**censimento**) o parziale (**campionamento**)
- **popolazione virtuale**: hanno un'esistenza concettuale e sono derivate dalla potenziale replicabilità a piacere della sperimentazione; sono potenzialmente *infinite* e quindi esaminabili solo in modo parziale

censimento: si esaminano *tutte* le unità di una popolazione reale, con riferimento a determinate caratteristiche di interesse

Per popolazioni reali i censimenti sono raramente effettuati, si opta più spesso per un campione.

campionamento: si esamina un *sottoinsieme* finito di unità statistiche, appartenenti ad una *popolazione reale o virtuale*, selezionate mediante l'esperimento di campionamento.

L'**esperimento di campionamento** è un particolare esperimento, assimilabile all'estrazione casuale di alcuni elementi da un'urna.

È un **esperimento casuale (aleatorio)** dal momento che risultano possibili una pluralità di esiti (campioni osservati) e prima di effettuare il campionamento non è possibile individuare con certezza quale potenziale campione verrà selezionato (**variabilità campionaria**).

Affinché il campione porti informazioni sull'intera popolazione, la sua estrazione deve essere casuale.

- il campione va scelto in modo che rifletta le **caratteristiche della popolazione**
- Esistono vari **piani di campionamento**, il più semplice è il **campionamento casuale semplice**, assimilabile all'estrazione casuale con reinserimento di elementi da un'urna.

Tipi di Statistica

I metodi statistici si possono dividere in due grandi classi.

- **Statistica descrittiva:** metodi per la descrizione, la presentazione e la sintesi dei dati disponibili, al fine di individuarne la struttura essenziale. Le finalità sono principalmente di tipo descrittivo, poichè si sintetizzano le informazioni disponibili, che riguardano la totalità della popolazione. anche quando i dati disponibili rappresentano un campione estratto da una popolazione, nella statistica descrittiva non se ne tiene conto.
- **Statistica inferenziale:** sono metodo per ricavare dai dati campionari informazioni sulla popolazione di riferimento e per quantificare la fiducia da accordare a tali informazioni. Si utilizza metodi del **calcolo delle probabilità**

Statistica descrittiva

Gli elementi più rilevanti:

- Metodi grafici e numerici per descrivere e sintetizzare i dati osservati
- Distinzione fra tecniche di **analisi univariata** (relative ad una singola caratteristica) e tecniche di **analisi multivariata**, ovvero per lo studio congiunto di due o più caratteristiche 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.

Analisi esplorativa

Un'analisi esplorativa dei dati ha 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 dataset
- codificare e riorganizzare i dati nella forma più conveniente per l'analisi
- utilizzare metodo grafici e numerici per ricavare alcune informazioni preliminari sui dati osservati

Si suppone che i dati siano già stati acquisiti e siano disponibili nella forma di **matrice dei dati** (anche detto **data frame**), questi sono cosiddetti **dati grezzi** ad esempio:

| unità | Genere | età | Livello istruzione | Dis |
|---------|--------|-----|--------------------|-----|
| Andrea | M | 28 | 3 | 5.0 |
| claudio | M | 17 | 2 | 7.5 |
| Lucia | F | 20 | 3 | NA |
| ... | ... | ... | ... | ... |

con **NA** si intende un dato mancante.

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

- Genere: Maschio (M) o Femmina (F) - variabile qualitativa nominale
- Età: età in anni - variabile quantitativa discreta
- Livello di istruzione: 1 = nessuna istruzione, 2 = scuola dell'obbligo, 3 = diploma, 4 = laurea - variabile qualitativa ordinale
- Dis: distanza dal luogo di lavoro in km - variabile quantitativa continua

Ogni **riga** corrisponde a un'unità statistica e contiene i valori su essa relativi delle caratteristiche dmoltelcresce.

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. Sono, in genere, aggettivi, valori numerici o espressioni verbali.

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

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** (normali): non è possibile individuare un ordinamento naturale delle modalità (es. genere, colore degli occhi)
 - **variabili qualitative ordinali**: è possibile invece individuare un ordinamento naturale delle modalità (es. livello di istruzione)

- **Variabile dicotomica** (binaria): se una variabile qualitativa ha soltanto due modalità
- **Variabili quantitative** (numeriche): se le modalità sono espresse in forma numerica (diverse dalle codifiche numeriche). in particolare si individuano:
 - **variabili quantitative discrete**: se Y è un insieme finito o al più numerabile (es. numero di figli, età)
 - **variabili quantitative continue**: se Y è un insieme continuo (ad esempio distanza, altezza, reddito). si noti che la continuità ca intensa come una potenziale continuità co come opportuno **riferimento semplificativo**

Una variabile quantitativa può essere con una **scala di intervalli**, se non esiste uno zero naturale e non arbitrario (ad esempio "temperatura", perchè lo 0 è convenzionale e non ha senso dire che 30° è il doppio di 15°). Una variabile quantitativa è con una **scala di rapporti** se invece esiste uno zero con tali caratteristiche (es. peso, altezza, reddito).

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) costruita 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.

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.

Il valore che occupa la posizione i -esima, $y_{(i)}$, si dice avere **rango i**

Con **supporto** si intende un insieme delle modalità di Y effettivamente osservate nella popolazione, indicata con S_Y . il numero di elementi all'interno di S_Y sarà sempre minore o uguale a n .

Distribuzioni di frequenza

Frequenza assoluta

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.

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

un esempio:

| Genere | Frequenza |
|--------|-----------|
| M | 5 |
| F | 7 |
| Totale | 12 |

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.

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

Le classi vanno definite in modo che: non siano ne troppe ne troppo poche. Una regola indicativa è di utilizzare circa \sqrt{n} classi.

Una **tabella (distribuzione) di frequenza** con modalità raggruppate in classi:

| Classi | $y_0 \dashv y_1$ | \dots | $y_{i-1} \dashv y_i$ | \dots | $y_{J-1} \dashv y_J$ | Totale |
|--------|------------------|---------|----------------------|---------|----------------------|--------------------|
| Freq | f_1 | \dots | f_i | \dots | f_J | $\sum_{i=1}^J f_i$ |

nota che il simbolo \dashv indica che l'estremo sinistro è incluso nella classe, mentre quello destro no, in notazione matematica standard: $[y_{i-1}, y_i)$.

Frequenza relativa

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:

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

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

Qualora si abbia un solo dato, la variabile statistica è detta **degenere**

Frequenze cumulative

Quando si hanno **variabili con modalità ordinabili**, 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 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:

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

Altri tipi di distribuzioni di frequenza

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 |

Rappresentazioni grafiche

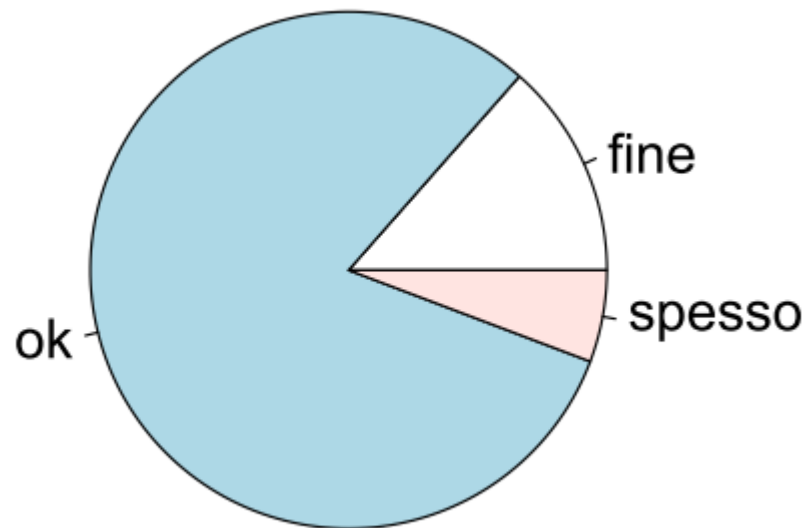
Oltre alle tabelle di frequenza, risulta utile introdurre alcune rappresentazioni grafiche, dette **diagramma statistici** (grafici)

- Per dati categoriali si possono utilizzare, ad esempio:
 - **diagrammi circolari** (o a torta)
 - **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**
 - **box-plot**

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.

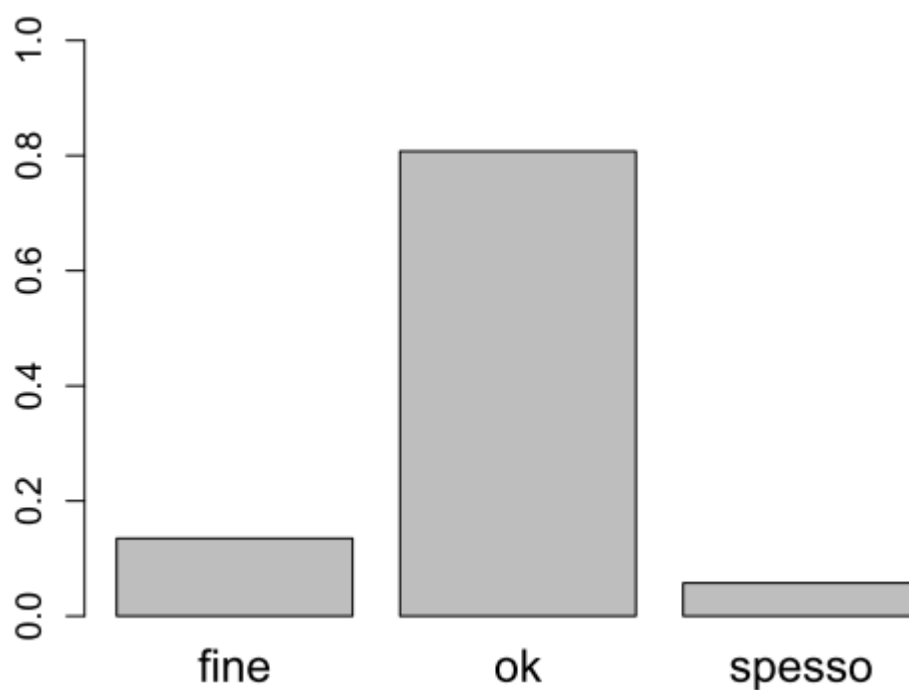
Esempio. Perni (continua). Considerando i dati riferiti alla produzione dei perni, le diverse modalità relative al diametro sono rappresentate dagli spicchi della torta, la cui dimensione è proporzionale alla corrispondente frequenza.



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.

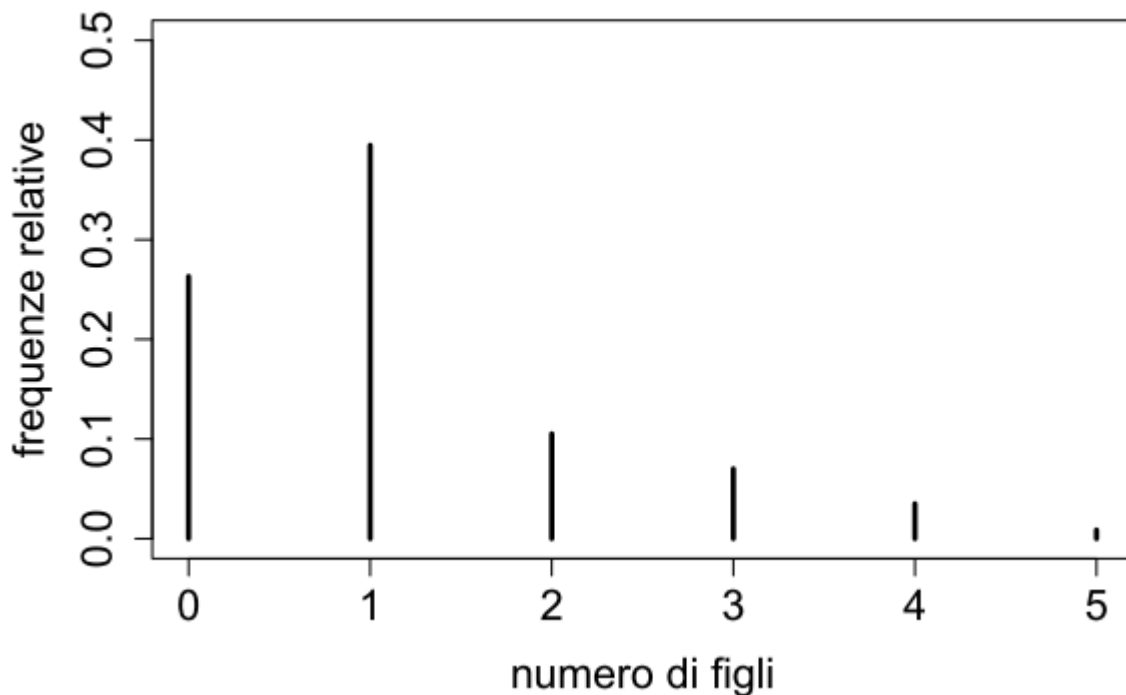


I rettangoli hanno la base uguale e sono separati per non implicare alcuna continuità. Le altezze sono proporzionali alle frequenze delle modalità; si considerano le frequenze relative affinché il confronto abbia senso.

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 L 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 valore indicato è rappresentato **dall'area**.

Il rettangolo corrispondente alla i -esima classe e ha come intervallo $[y_{j-1}, y_j]$ e ha:

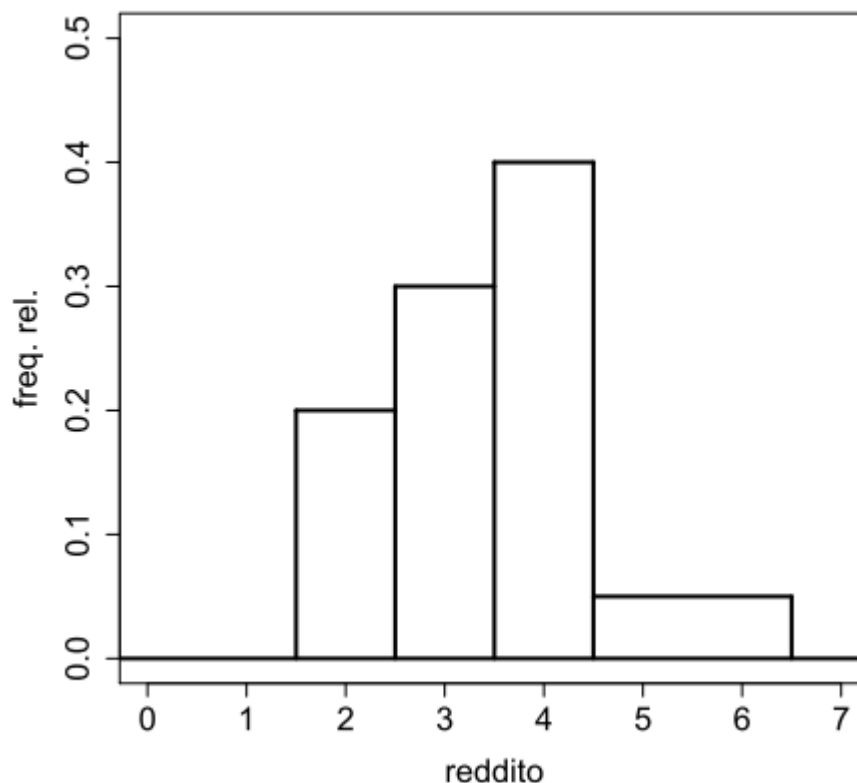
- altezza (e quindi area) proporzionale a, oppure pari a, $f_j / (y_j - y_{j-1})$: **istogramma delle frequenze assolute**
- altezza (e quindi area) proporzionale a, oppure pari a, $p_j / (y_j - y_{j-1})$: **istogramma delle frequenze relative**

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

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. | 0.2 | 0.3 | 0.4 | 0.1 | 1 |

L'associato istogramma della frequenze relative corrisponde a

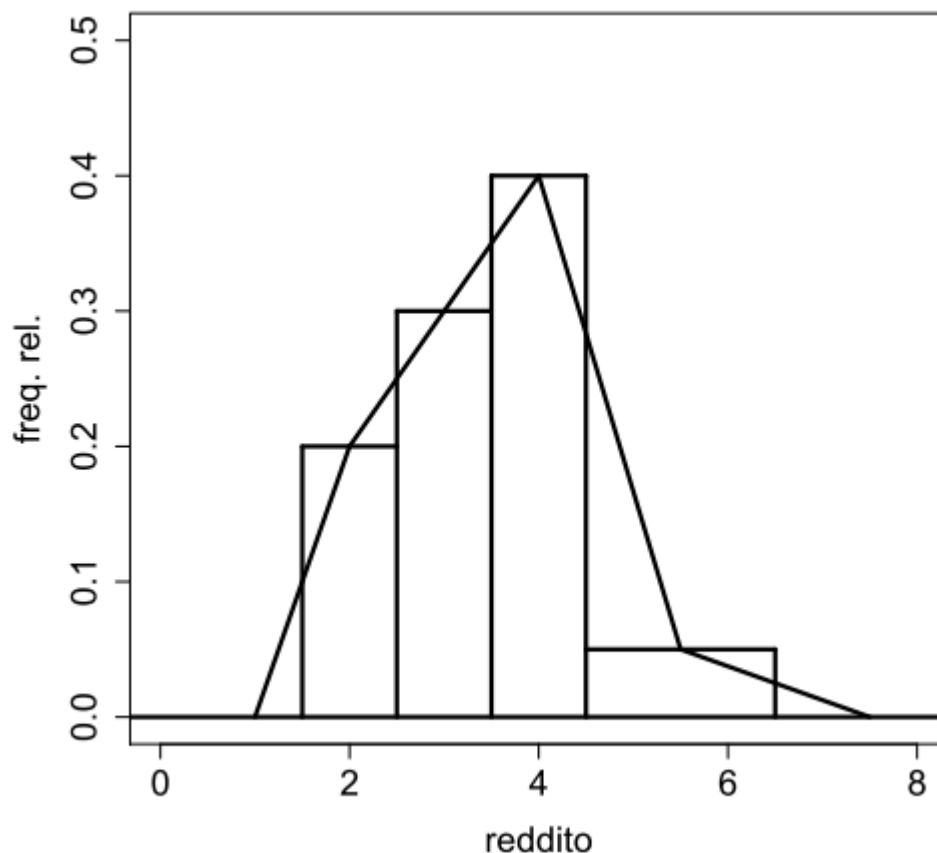


Poligoni di frequenza

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

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

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:

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

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

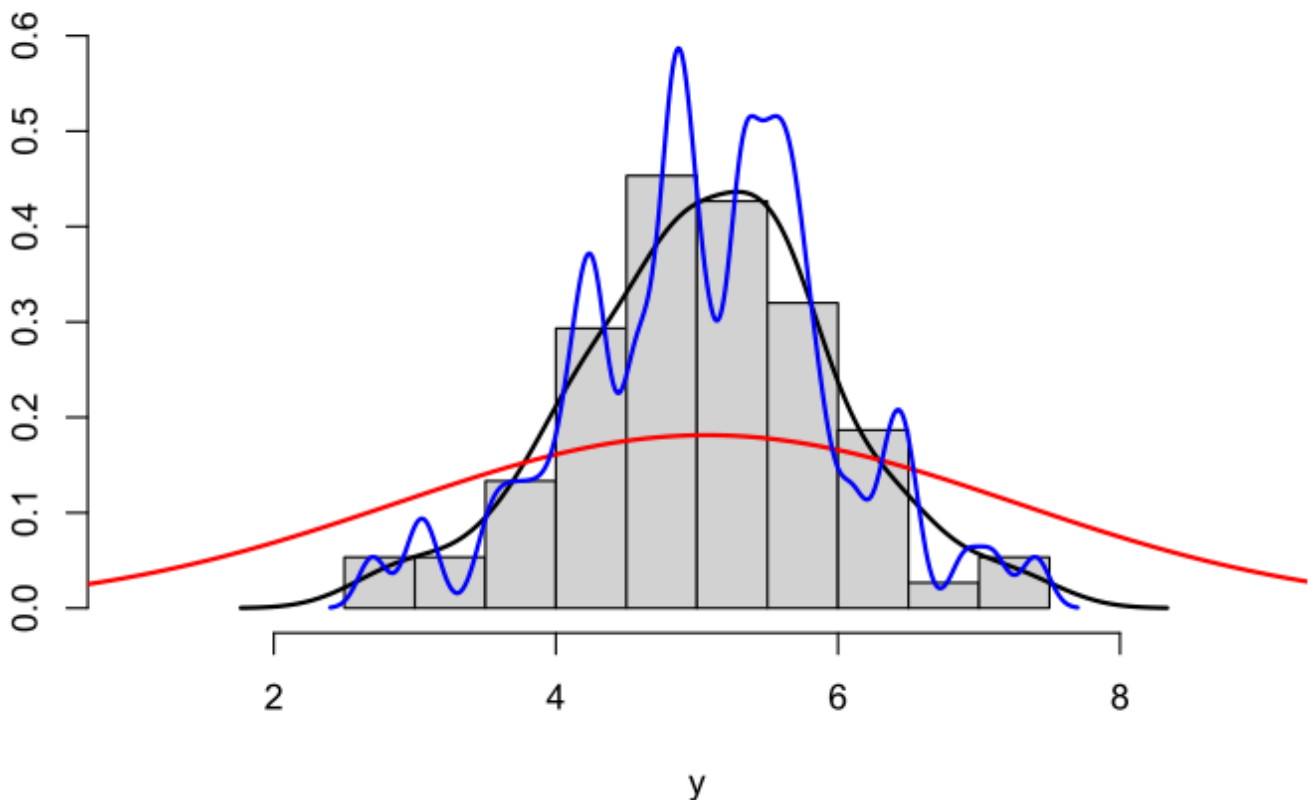
Ad ogni dato y_i si sovrappone non un rettangolo ma una curva che risulta essere più smussata. La sua altezza è proporzionale alla frequenza dei punti e la sua ampiezza dipende dalla banda b .

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 $y \in \mathbb{R}$ corrisponde al rapporto tra il numero di osservazioni minori o uguali a y e il numero totale di osservazioni.

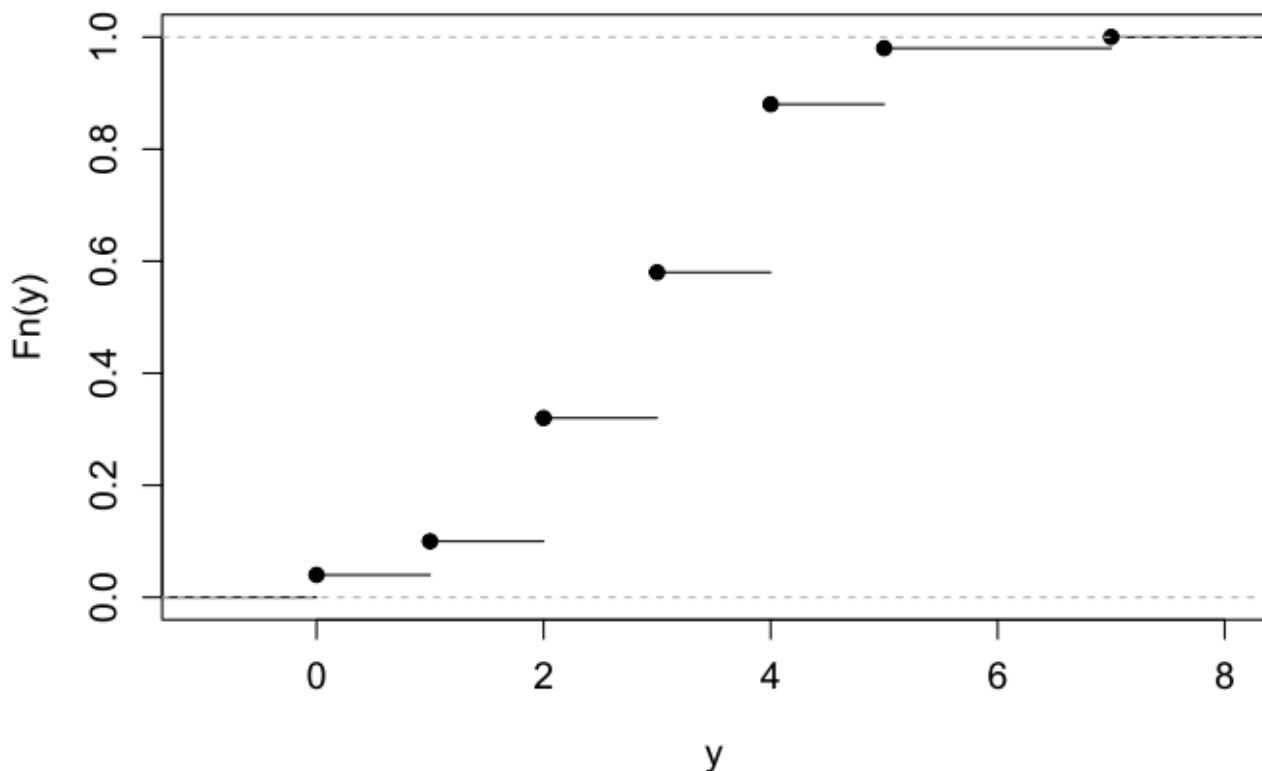
$$F_n(y) = \frac{\text{no. oss.} \leq y}{\text{no. totale oss}}, \quad y \in \mathbb{R}$$

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 funzioni 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 | Totale |
|-----------------|------|------|------|------|------|------|------|--------|
| 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 | 1 |



Diagrammi di dispersione

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

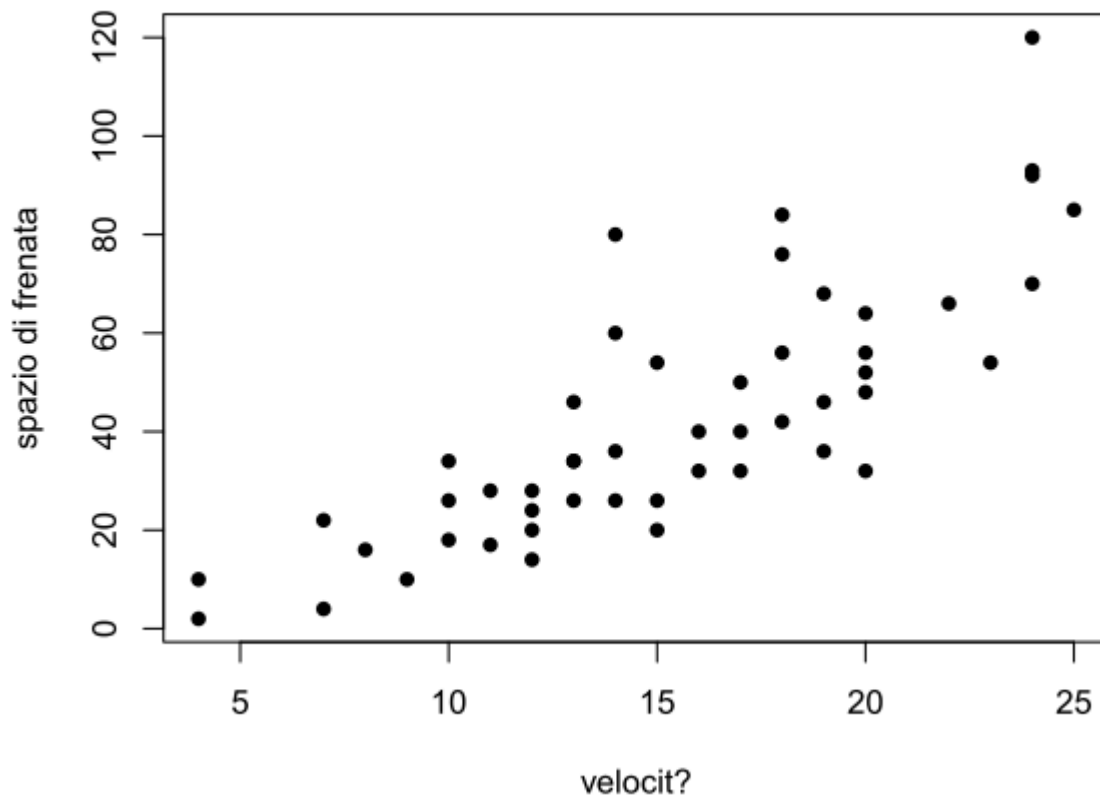
Nel caso di due **variabili quantitative**, per una prima analisi della relazione tra le variabili si possono usare i **diagrammi di dispersione** (scatter plot).

se x_i, y_i 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

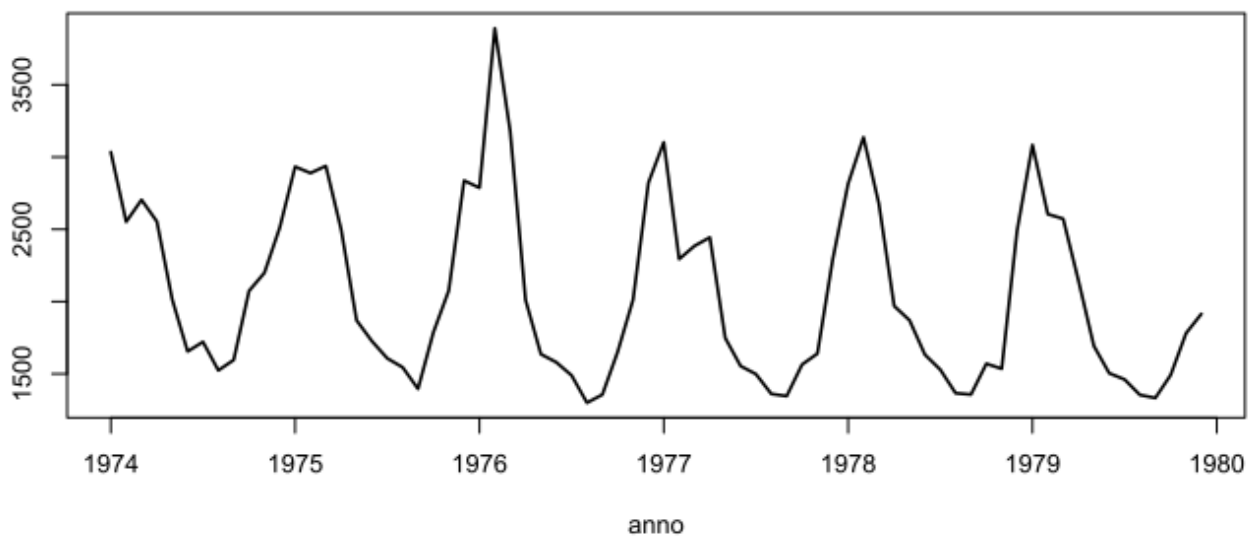
dispersione.



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.

Indici sintetici

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

Nel caso in cui i dati derivate 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 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 μ_Y o semplicemente con μ .

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:

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

Qualora si abbiano delle classi di modalità, si può approssimare la media aritmetica sostituendo ad ogni classe il punto medio di essa $m_j = (y_{j-1} + y_j)/2$:

La media aritmetica risente della presenza di osservazioni anomale o estreme (outliers).

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.

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 $Y - E(Y)$, si ha:

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

3. **proprietà di linearità.** sia $aY + b$, $a, b \in \mathbb{R}$, una trasformata lineare della variabile Y , allora:

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

Indice di posizione: mediana

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

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 nelle 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 $y_{0.5} = (y_{(n/2)} \text{ e } y_{(n/2+1)})/2$

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.

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 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.

Quantili

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

Data una variabile qualitativa ordinale o quantitativa Y , y_α 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.

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.

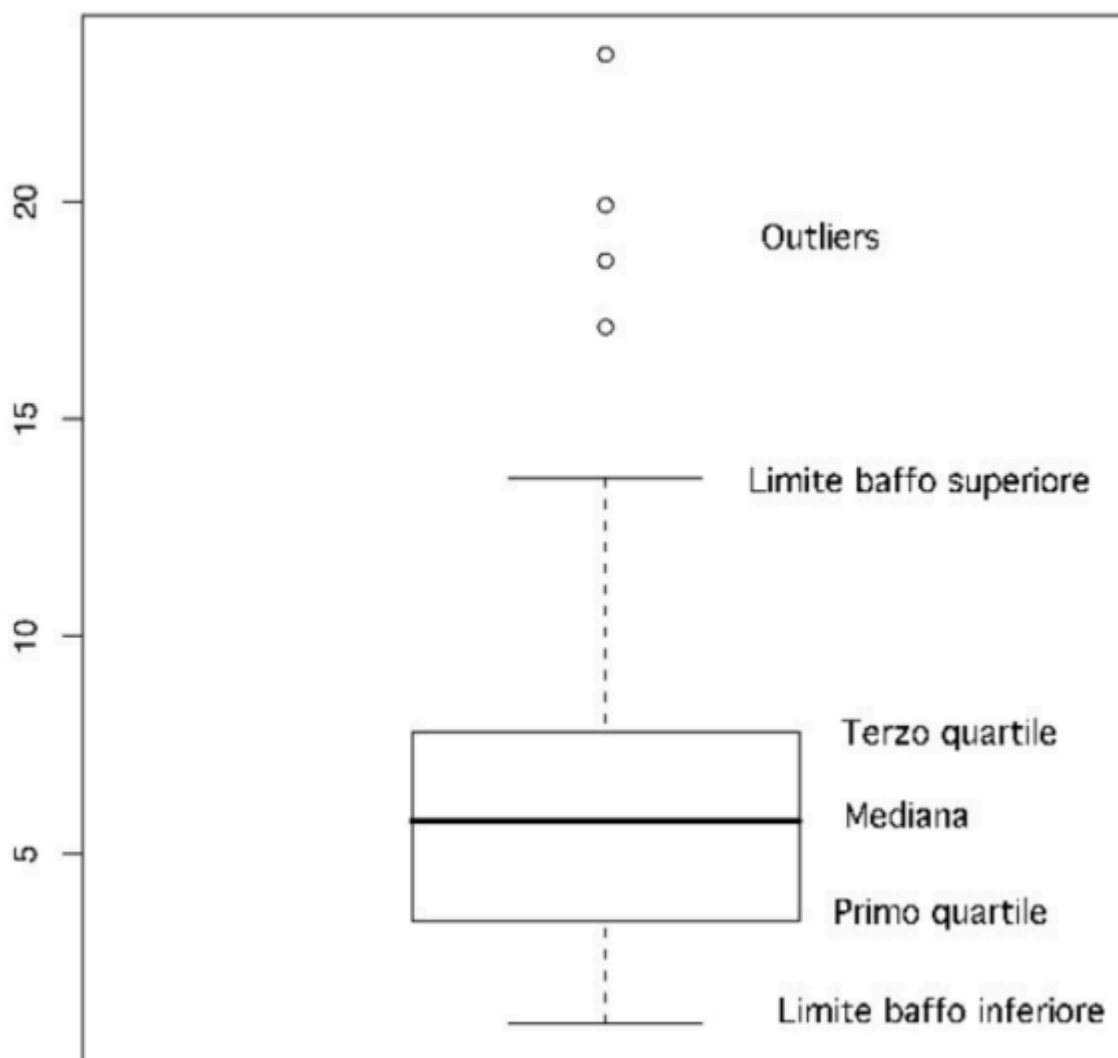
I quantili di livello $\alpha = 0.10, 0.20, \dots, 0.90$ vengono chiamati **decili** e dividono le osservazioni ordinate in 10 parti uguali.

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

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**.

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.

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.

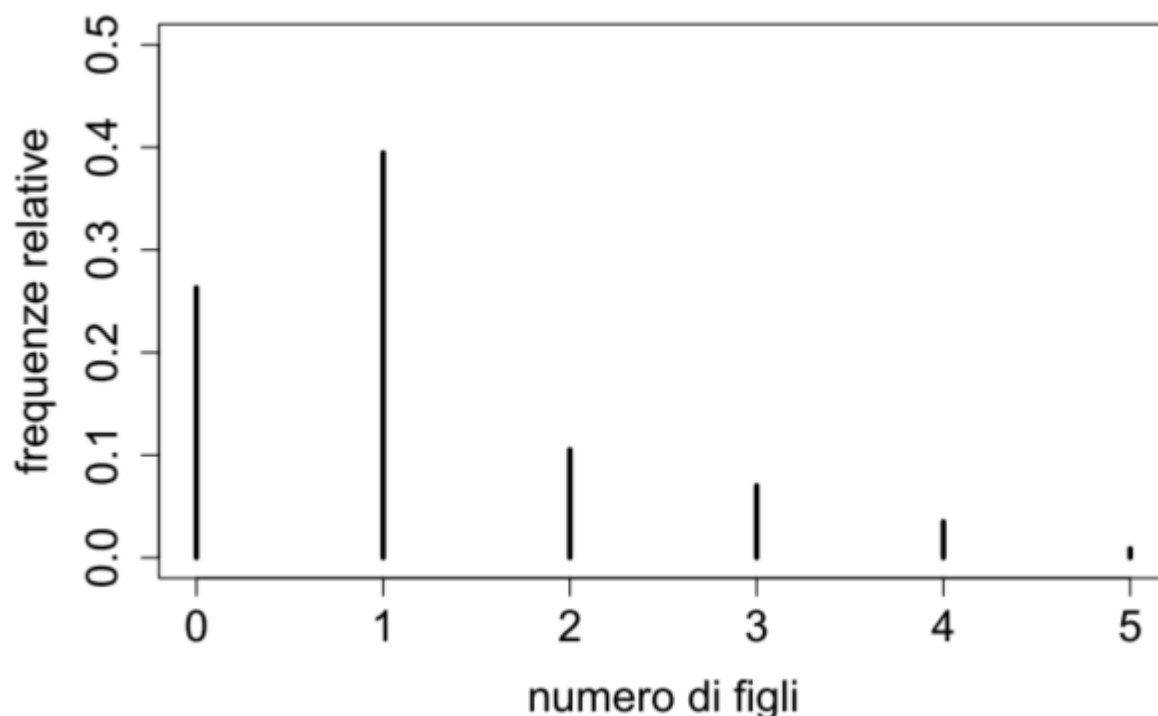
La moda di una variabile statistica Y corrisponde al valore y_{mo} del supporto S_Y a cui è associata la frequenza, relativa o assoluta, più alta.

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.

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.

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

- **campo di variazione**
- **scarto interquantilico**
- **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.

Indici di variabilità: campo di variazione

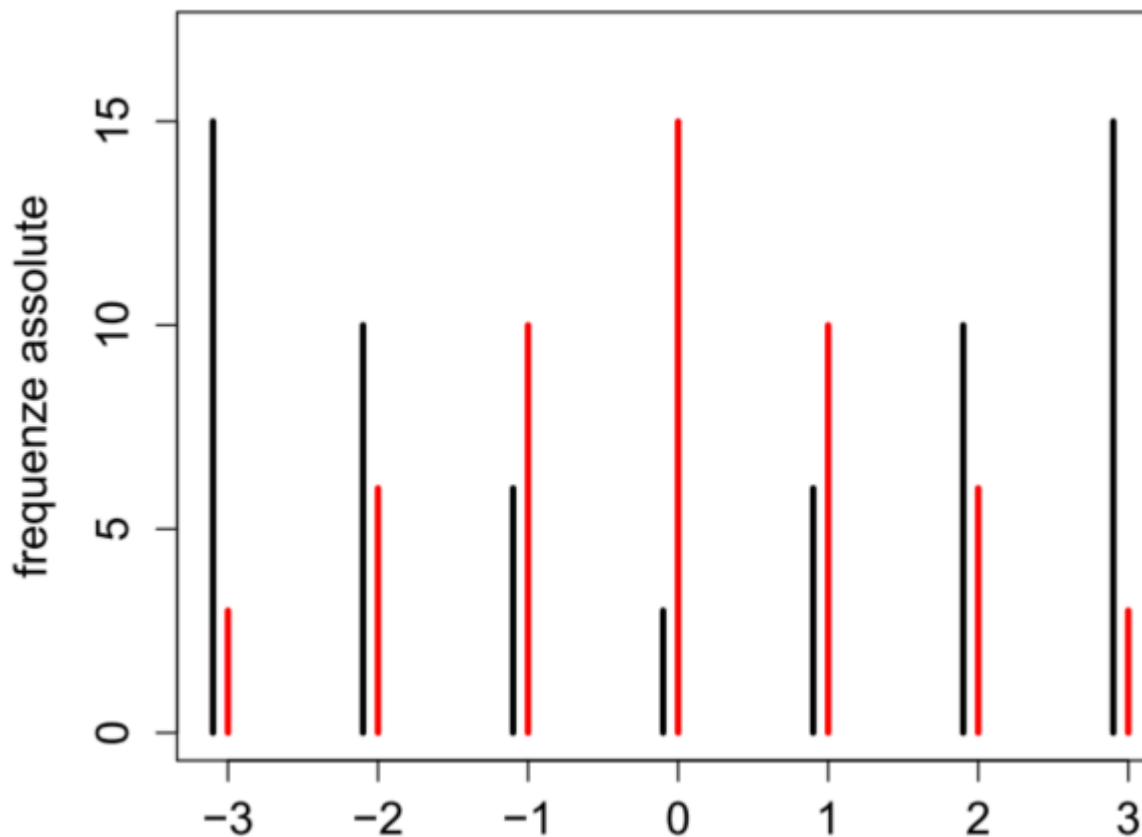
Sia Y una variabile statistica quantitativa. Il **campo di variazione** (range) corrisponde a

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

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

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

R_Y è sensibile alla presenza di valori anomali. Inoltre è un indice piuttosto povero, come dimostra il seguente graco 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 $Y = (1, 2, 2, 2, 2, 2, 5)$, poichè $y_{0.75} = y_{0.25} = 2$. In presenza di poche osservazioni nomale ha proprietà di robustezza.

Indici di variabilità: varianza e scarto quadratico medio

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

Lo scarto quadratico medio di Y , indicato con σ_Y o con σ , 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**

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

La varianza soddisfa le seguenti proprietà:

1. **Proprietà di non negatività:** $V(Y) \geq 0$ e $V(Y) = 0$ se e solo se Y è degenere.
2. **Formula per il calcolo:**

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

3. **Proprietà di invarianza per traslazione:**

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

4. **Proprietà di omogeneità di secondo grado:**

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

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**.

$$CV_Y = \frac{\sigma_Y}{E(Y)}$$

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.

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

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

Simmetria e asimmetria

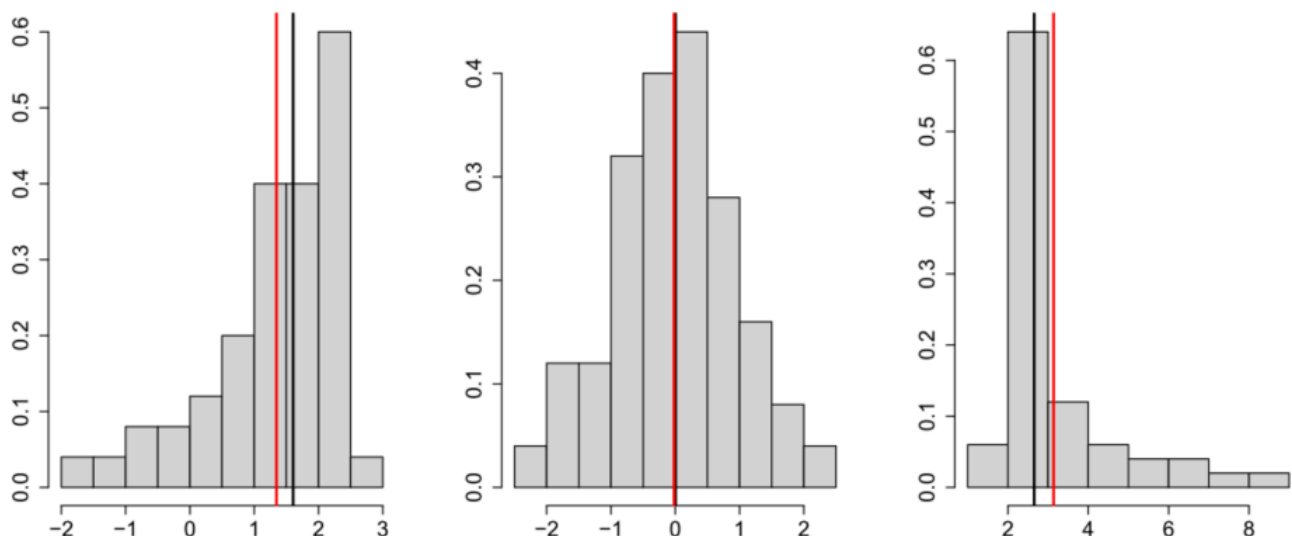
Una **distribuzione di frequenza** (ad esempio rappresentata con un istogramma o un diagramma a bastoncini) è **simmetrica** se la sua metà di destra si sovrappone alla sua metà di sinistra (dove la metà è identificata dalla mediana).

si noti che:

- se l'assimmetria è **positiva** media > mediana
- se la distribuzione è **simmetrica** media = mediana
- se l'assimmetria è **negativa** media < mediana

Per una distribuzione di frequenza unimodale e simmetrica si ha che: media \approx mediana \approx 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 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 σ_Y , allora l'indice di simmetria corrisponde a

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

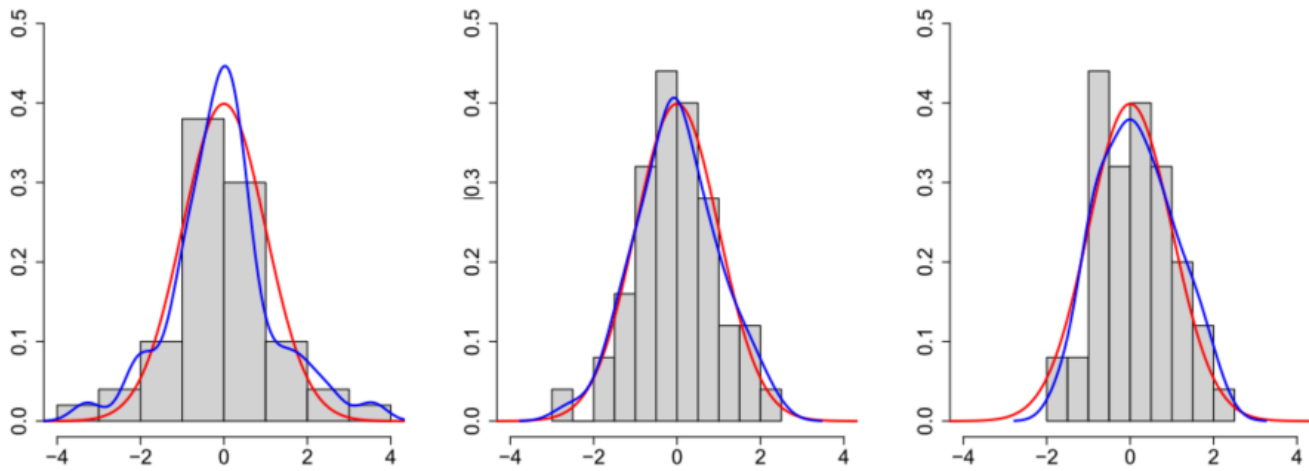
- Se la distribuzione di frequenza è **simmetrica**, $\gamma_Y \approx 0$;
- se c'è **asimmetria negativa**, $\gamma_Y < 0$;
- se c'è **asimmetria positiva**, $\gamma_Y > 0$.

Curtosi e indice di 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).



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

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

Se si dispone dei dati grezzi $Y = (y_1, \dots, y_n)$, e si sono preventivamente calcolati $E(Y)$ e σ_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, $\beta_Y < 3$.

Calcolo delle probabilità

da fare

Inferenza statistica

Utilizzo

Viene utilizzata per studiare una popolazione in maniera parziale.

Viene usata per:

- ricavare dai dati campionari informazioni sulla popolazione di interesse
- Quantificare la fiducia da assegnare a tali informazioni

L'obiettivo è quello di arrivare a conclusioni valide per tutta la popolazione, basandosi su un campione.

Campionamento

Si effettuano indagini campionarie quando:

- presenza di vincoli di tempo e/o problemi di costo
- la popolazione di interesse può essere infinita e virtuale
- la rilevazione potrebbe distruggere le unità statistiche o essere potenzialmente dannosa
- la precisione dei risultati nelle rilevazioni censuarie potrebbe non essere adeguata

È essenziale che i dati campionari possano essere interpretati come risultato di un esperimento aleatorio, perché altrimenti verrebbe meno la rappresentatività del campione e la possibilità di ricavare informazioni utili sulla popolazione (fenomeno) di interesse (inferenza).

Campioni casuali semplici

Sono formati da n realizzazioni indipendenti (con $n \geq 1$) di un esperimento base, nelle medesime condizioni.

Le unità vengono selezionate dalla popolazione di riferimento in modo che ognuna abbia la stessa probabilità di essere scelta (con popolazioni numerose un campionamento con reinserimento o meno non cambia sostanzialmente il risultato).

I dati osservati sono riferiti ad una caratteristica di interesse, rilevata sulle n unità statistiche che costituiscono il campione.

L'**ipotesi fondamentale** su cui poggia l'inferenza statistica è che i dati campionari x sono una realizzazione di un vettore di variabili casuali $X = (X_1, X_2, \dots, X_n)$

Nell'inferenza statistica c'è un rovesciamento di punto di vista: non si influenzano i dati (il fenomeno) ma il modello probabilistico. Il processo in questione è, in definitiva, la popolazione (fenomeno) oggetto di indagine.

Nel **campionamento casuale semplice**, le variabili casuali X_1, \dots, X_n sono **indipendenti e identicamente distribuite**, cioè con lo stesso modello probabilistico e tali da non influenzarsi a vicenda.

Modelli statistici parametrici

Da un campione casuale semplice $X = (X_1, X_2, \dots, X_n)$, la **distribuzione di probabilità** delle singole variabili casuali dipende dalla natura dei dati e del fenomeno oggetto di indagine.

La distribuzione assunta per le variabili casuali del campione dipende da costanti ignote dette **parametri**, ad esempio, $p, \mu, \sigma^2, \lambda, \dots$ o più parametri, indicati generalmente come $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, con $k \geq 1$. Nell'*inferenza statistica parametrica* si assume che la distribuzione delle variabili casuali del campione sia nota a meno dei valori dei parametri, che corrispondono tipicamente agli aspetti di interesse dell'analisi.

Vengono usate le seguenti assunzioni, che definiscono un **modello statistico parametrico** per i dati di un campione casuale semplice:

- le variabili casuali X_1, X_2, \dots, X_n sono indipendenti
- tutte le X_i hanno la stessa distribuzione di probabilità
- tale distribuzione è nota a meno dei valori di un numero d di parametri, indicati generalmente come $\theta = (\theta_1, \theta_2, \dots, \theta_d), d \geq 1$

Lo scopo dell'inferenza statistica parametrica è utilizzare i dati osservati x_1, \dots, x_n per ottenere informazioni su θ , i cui possibili valori appartengono ad un certo insieme Θ detto **spazio parametrico**.

La **scelta del modello** è molto importante, poiché le conclusioni inferenziali dipendono fortemente dalle assunzioni fatte.

Nella specificazione del modello statistico parametrico, è importante considerare:

- la natura dei dati (qualitativi o quantitativi, discreti o continui, ecc.) $\times \dots \times \{0, 1\}^n$, cioè l'insieme di tutti i possibili vettori x
- gli aspetti notevoli presenti nei dati come posizione, variabilità, simmetria, curtosi, ecc.

- tutte le informazioni sul meccanismo generatore dei dati.

Esistono anche modelli per **dati dipendenti** e/o **non identicamente distribuiti** (ad esempio per serie storiche e spaziali), e modelli che prescindono dalla forma della distribuzione di probabilità delle variabili casuali del campione (modelli non parametrici)

Lo spazio parametrico è $\Theta = (0, 1)$ e lo spazio campionario è

$S_X = 0, 1X \dots X0, 1 = 0, 1^n$, cioè l'insieme di tutti i possibili vettori n -dimensionali costituiti da 0 e 1.

Se le n osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, è ragionevole ipotizzare che $X_i, i = 1, \dots, n$ siano indipendenti con distribuzione $N(\mu, \sigma^2)$.

Si può verificare graficamente l'ipotesi di normalità considerando istogrammi e q-q plot, μ è la misura vera dell'oggetto in esame e σ^2 è riconducibile alla precisione dello strumento di misura.

Verifica del modello

Talvolta in casi complessi si necessita un **controllo empirico del modello**

alcuni strumenti possono essere:

- l'**istogramma delle frequenze relative** e la **stima della densità**
- i grafici dei quantili (**q-q plot**)

L'istogramma e la stima della densità basate sui dati campionari possono essere interpretate, in ambito inferenziale, come stime della funzione di densità

Procedure inferenziali

Si possono individuare *tre classi* generali di procedure che affrontano i problemi inferenziali, con riferimento al parametro di interesse θ :

- la **stima puntuale**: si vuole ottenere, sulla base dell'osservazione x , una congettura puntuale su θ
- la **stima intervallare** o **regione di confidenza**: si vuole ottenere, sulla base dell'osservazione x , un sottoinsieme (intervallo) di Θ in cui è plausibilmente incluso θ
- **verifica di ipotesi**: data una congettura o un'ipotesi su θ , si vuole verificare, sulla base dell'osservazione x , se essa è accettabile (cioè in accordo con i dati x).

Statistiche campionarie

La statistica inferenziale è caratterizzata da una componente di incertezza, poiché i dati campionari x sono interpretati come realizzazione di un vettore casuale X , ripetendo l'esperimento si ottengono dati diversi dal primo.

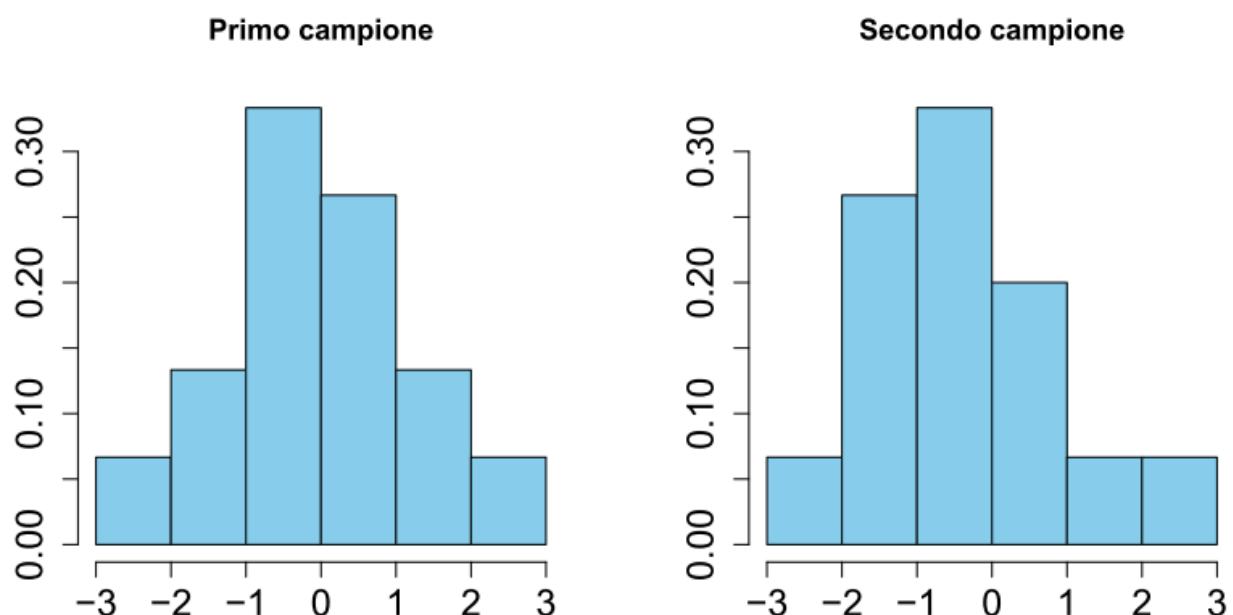
Si chiama **statistica campionaria** ogni trasformata $T = t(X_1, \dots, X_n)$ che sintetizza opportunamente il campione $X = (X_1, \dots, X_n)$.

La distribuzione di probabilità T , che è una funzione di $X = (X_1, \dots, X_n)$, dipende dal parametro incognito θ . Quindi, va intesa **sotto** θ , cioè nell'ipotesi che θ sia il vero valore del parametro, qualunque esso sia.

Dato un campione casuale $X = (X_1, \dots, X_n)$, sono esempi di statistiche campionarie:

- la **somma campionaria** $S_n = \sum_{i=1}^n X_i$, la media campionaria $\bar{X} = n^{-1} S_n$, la **varianza campionaria** $S^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- le **statistiche ordinate** $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, dove $X_{(1)}$ è la variabile casuale che occupa l' i -esima posizione del campione.
- la **variabile casuale minimo** $X_{(1)} = \min X_1, \dots, X_n$ e la **variabile casuale massimo** $X_{(n)} = \max X_1, \dots, X_n$
- la **mediana campionaria**, definita da $X_{((n+1)/2)}$ se n è dispari, e da $(X_{(n/2)} + X_{(n/2+1)})/2$ se n è pari
- i **momenti campionar**i, centrati e non centrati: $n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^r$ e $n^{-1} \sum_{i=1}^n X_i^r$, con $r \in \mathbb{N}^+$

anche se due rilevazioni campionare hanno due risultati diversi, si possono ricondurre al medesimo esperimento osservando gli istogrammi prodotti dai due campioni.



Somma e media campionaria

Sia X_1, \dots, X_n un campione casuale semplice tratto da una determinata popolazione. Si definiscono, rispettivamente, la **somma campionaria** (somma del campione) e la **media campionaria** (media del campione) le variabili casuali $S_n = \sum_{i=1}^n X_i$, $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = \frac{S_n}{n}$

Poiché X_1, \dots, X_n sono **indipendenti e identicamente distribuite** (quindi anche la stessa media μ e la stessa varianza σ^2), allora:

$$E(S_n) = \sum_{i=1}^n E(X_i) = n\mu \quad V(S_n) = \sum_{i=1}^n V(X_i) = n\sigma^2$$

$$E(\overline{X}_n) = \frac{E(S_n)}{n} = \mu \quad V(\overline{X}_n) = \frac{V(S_n)}{n^2} = \frac{\sigma^2}{n}$$

Se le variabili casuali X_1, \dots, X_n sono **gaussiane** $N(\mu, \sigma^2)$, allora anche somma e media campionaria sono variabili casuali gaussiane, più precisamente:

$$S_n \sim N(n\mu, n\sigma^2), \quad \overline{X}_n \sim N(\mu, \sigma^2/n)$$

Valgono, inoltre, i seguenti risultati con riferimento a variabili casuali X_1, \dots, X_n indipendenti:

- se $X_i \sim Bi(K_i, p)$, $i = 1, \dots, n$, allora $S_n \sim Bi(\sum_{i=1}^n K_i, p)$
- se $X_i \sim P(\lambda_i)$, $i = 1, \dots, n$, allora $S_n \sim P(\sum_{i=1}^n \lambda_i)$
- se $X_i \sim X^2(r_i)$, $i = 1, \dots, n$, allora $S_n \sim X^2(\sum_{i=1}^n r_i)$

La media campionaria \overline{X}_n è utile in un ambito inferenziale quando si vuole fare inferenza su μ (media della popolazione).

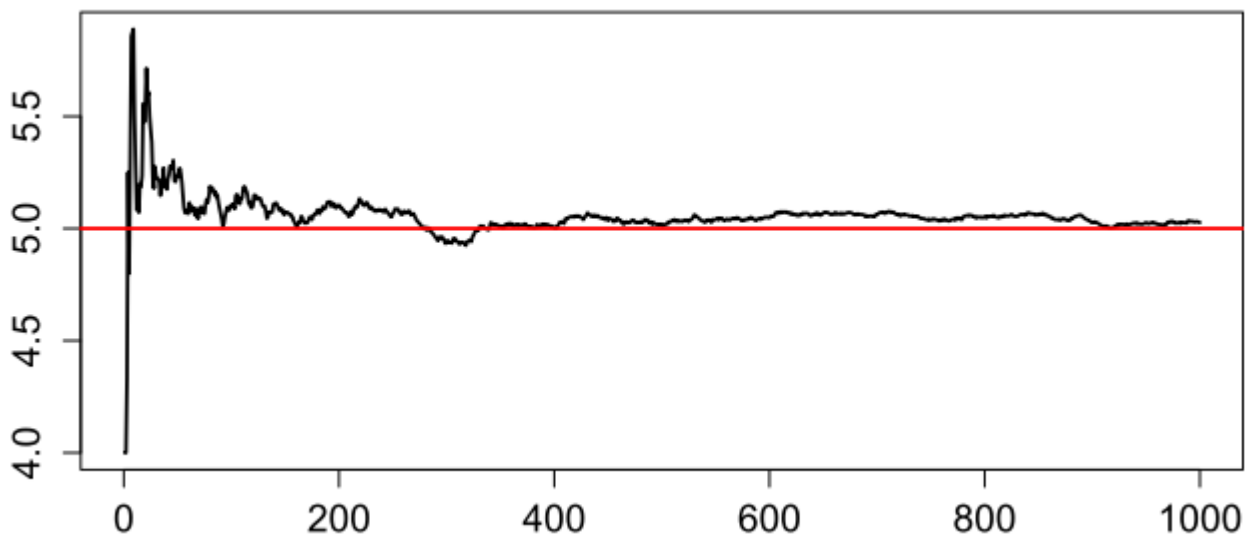
Quindi, la media campionaria \overline{X}_n definisce uno **Stimatore** per μ e il suo valore osservato \overline{x}_n che corrisponde alla media calcolata sul campione, viene utilizzato come **Stima** per μ .

Al crescere di n , la variabile casuale media campionaria ha una distribuzione di probabilità sempre più concentrata attorno alla media della popolazione μ .

Formalmente, si afferma che vale la **legge debole dei grandi numeri**, cioè che, nelle condizioni poste in precedenza, se $n \rightarrow \infty$:

$$\overline{X}_n \xrightarrow{p} \mu$$

Con la scrittura \xrightarrow{p} si intende la **convergenza in probabilità**, una notazione di convergenza probabilistica illustrata nel seguente esempio



Al crescere di n , la distribuzione di probabilità della media campionaria è sempre più concentrata attorno alla media della popolazione μ .

Dice che, sotto condizioni generali, la distribuzione di probabilità della media campionaria tende ad una distribuzione normale al crescere di n , indipendentemente dalla distribuzione delle singole variabili casuali del campione

Per le variabili casuali **somma** e **media campionaria** vale un importante risultato: il **teorema limite centrale**.

Data una successione di variabili casuali $X_i, i \geq 1$, indipendenti e identicamente distribuite, con media μ e varianza $\sigma^2 \neq 0$ finite, allora la **Somma standardizzata** e la **media campionaria standardizzata** coincidono e sono tali che, per $n \rightarrow \infty$:

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} Z \sim N(0,1)$$

La scrittura \xrightarrow{d} indica la **convergenza in distribuzione**: al crescere di n la distribuzione di probabilità è sempre più simile a quella di Z .

Per n fissato sufficientemente elevato (almeno $n > 30$), valgono le seguenti utili approssimazioni:

$$\bar{X}_n \dot{\sim} N(\mu, \sigma^2/n), \quad S_n \dot{\sim} N(n\mu, n\sigma^2)$$

dove $\dot{\sim}$ indica la distribuzione approssimata.

Per il teorema limite centrale, se n è sufficientemente elevato, si possono ancora utilizzare le **distribuzioni gaussiane** (approssimate), valgono le seguenti approssimazioni:

$$P(a < \bar{X}_n \leq b) \doteq \Phi\left(\frac{b - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right)$$

$$P(a < S_n \leq b) \doteq \Phi\left(\frac{b - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - n\mu}{\sigma\sqrt{n}}\right)$$

Varianza campionaria

Sia X_1, \dots, X_n un campione casuale semplice tratto da una determinata popolazione, si definisce **varianza campionaria** la variabile casuale

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

La varianza campionaria può venire calcolata utilizzando la seguente regola di calcolo:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \bar{X}_n^2$$

Poiché $X_i, i = 1, \dots, n$ sono **indipendenti e identicamente distribuite** (quindi anche la stessa media μ e la stessa varianza σ^2), allora:

$$E(S^2) = \frac{n-1}{n} \sigma^2 = \sigma^2 - \frac{1}{n} \sigma^2$$

La **varianza campionaria** S^2 è utile in ambito inferenziale quando, utilizzando i dati campionari, si vuole fare inferenza su σ^2 (varianza della popolazione).

Nel caso si abbia un n piccolo si usa la **varianza campionaria corretta**:

$$S_c^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Da cui:

$$E(S_c^2) = \frac{n}{n-1} E(S^2) = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2$$

Qualora le variabili casuali X_1, \dots, X_n siano **gaussiane** $N(\mu, \sigma^2)$, allora la varianza campionaria e la varianza campionaria corretta hanno una distribuzione di probabilità legata al modello χ^2 :

$$\frac{n}{\sigma^2} S^2 = \frac{n-1}{\sigma^2} S_c^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2(n-1)$$

Dove $X^2(n-1)$ indica un modello chi-quadrato di parametro (gradi di libertà) $n-1$.

La variabile casuale media si può standardizzare con:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Se al posto di σ si considera $S_c = \sqrt{S_c^2}$ si ha la variabile casuale chiamata **media campionaria studentizzata**, e si ha che:

$$\frac{\bar{X}_n - \mu}{S_c/\sqrt{n}} \sim t(n-1)$$

dove $t(n-1)$ indica una variabile casuale t Student con $n-1$ gradi di libertà.

Date due variabili casuali con distribuzione $N(\mu_X, \sigma_X^2)$ e $N(\mu_Y, \sigma_Y^2)$, le associate varianze campionarie (indipendenti), si può verificare che:

$$\frac{[nS_X^2/\sigma_X^2]/(n-1)}{[nS_Y^2/\sigma_Y^2]/(m-1)} \sim F(n-1, m-1)$$

Dove $F(n-1, m-1)$ indica una variabile casuale F di Fisher con $(n-1)$ e $(m-1)$ gradi di libertà.

Stima puntuale - Stime

Formule e Esempi

Frequenza relativa

Si ottiene dividendo la frequenza assoluta per il numero totale di osservazioni.

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

Esempio: se in un campione di 100 persone 30 sono di genere femminile, la frequenza relativa del genere femminile è:

$$p_F = \frac{30}{100} = 0.3$$

Frequenza cumulata

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

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 |

Stima della densità

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

Funzione di ripartizione empirica

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

Indici di posizione: media aritmetica

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:

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

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.

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 $Y - E(Y)$, si ha:

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

3. **proprietà di linearità.** sia $aY + b$, $a, b \in \mathbb{R}$, una trasformata lineare della variabile Y , allora:

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

Indice di posizione: mediana

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

- se n è **dispari**: alla posizione $(n+1)/2$, cioè $y_{0.5} = y_{(n+1)/2}$
- se n è **pari**: alla media delle posizioni $n/2$ e $(n/2)+1$, cioè $y_{0.5} = \frac{y_{(n/2)} + y_{(n/2)+1}}{2}$

Indici di posizione: moda

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.

Indici di variabilità: campo di variazione

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

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

Dispositivo A: $R = 15.64 - 14.40 = 1.20$

Dispositivo B: $R = 15.80 - 13.83 = 1.97$

Indici di variabilità: scarto interquartilico

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

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 e scarto quadratico medio

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

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

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**

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

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

| y_i | 0 | 1 | 2 | 3 | 4 | Totale |
|-------|-----|----|----|---|---|--------|
| f_i | 109 | 65 | 22 | 3 | 1 | 200 |

$$V(Y) = \frac{1}{200} \sum_{j=0}^4 (y_j - 0.61)^2 f_j = \frac{1}{200} [(0 - 0.61)^2 \cdot 109 + (1 - 0.61)^2 \cdot 65 \dots]$$

inoltre: $\sigma_Y = \sqrt{V(Y)} = 0.780$ La varianza soddisfa le seguenti proprietà:

1. **Proprietà di non negatività:** $V(Y) \geq 0$ e $V(Y) = 0$ se e solo se Y è degenere.
2. **Formula per il calcolo:**

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

3. **Proprietà di invarianza per traslazione:**

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

4. **Proprietà di omogeneità di secondo grado:**

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

Indici di variabilità: coefficiente di variazione

$$CV_Y = \frac{\sigma_Y}{E(Y)}$$

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 |

| Regione | Merchi (migliaia di tonnellate) | Passeggeri (migliaia) |
|-----------|----------------------------------|-----------------------|
| 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). Poichè $E(X) = 15554.75$, $V(X) = 53376636$, $E(Y) = 139.75$, $V(Y) = 13978.19$, si ha che

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

Indice di simmetria

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

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 σ_Y , allora l'indice di simmetria corrisponde a

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

Curtosi e indice di curtosi

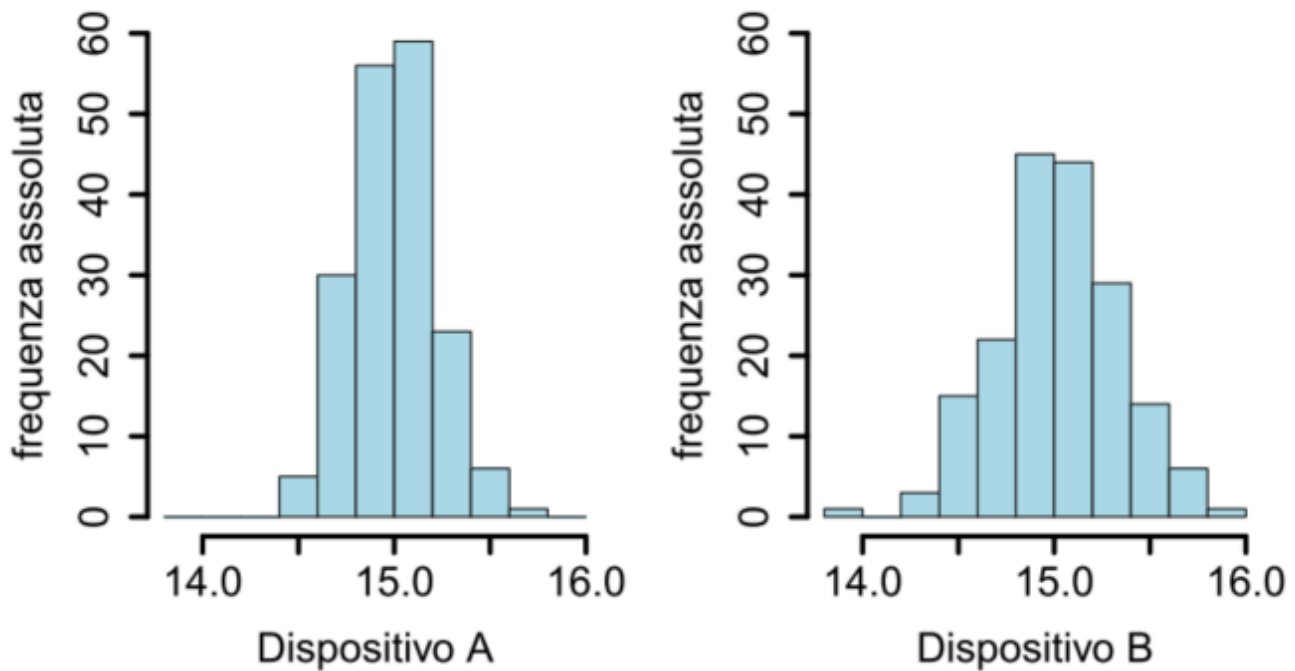
$$\beta_Y = \frac{E[(Y - E(Y))^4]}{\sigma_Y^4}$$

Se si dispone dei dati grezzi $Y = (y_1, \dots, y_n)$, e si sono preventivamente calcolati $E(Y)$ e σ_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, poichè 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.

Somma e media campionaria

Esempio: Procedura di controllo. Si è verificato un inconveniente su una linea di produzione che determina la presenza di 1/10 di pezzi difettosi. La procedura di controllo della qualità prevede che, se si individuano almeno 5 pezzi difettosi su $n \geq 1$ scelti a caso, il processo viene posto in revisione. Sia S_n la somma di $n \geq 1$ variabili casuali $\text{Ber}(1/10)$ indipendenti. Si cerca il valore di n tale che ci sia una probabilità pari a 0.9 di porre il processo in revisione. Quindi, $n \geq 1$ deve essere tale che

$$P(S_n \geq 5) = P\left(\frac{S_n - (n/10)}{\sqrt{n9/100}} \geq \frac{5 - (n/10)}{\sqrt{n9/100}}\right) \doteq P\left(Z \geq \frac{5 - (n/10)}{\sqrt{n9/100}}\right)$$

Sia 0.9 con $Z \sim N(0, 1)$. Poiché il valore critico $z_{0.9} = -1.282$, si cerca n tale che $[5 - (n/10)]\sqrt{n9/100} \doteq -1.282$, con $n \geq 50$.

Si ottiene come soluzione il valore 85.58, quindi $n = 86$, può essere una scelta ragionevole.

Comandi in R

```
print("Prossimamente anche in 32 bit!")
```