Università di Catania

Metodi Matematici e Statistici

Dr. Giovanni Nastasi
giovanni.nastasi@unict.it

Dipartimento di Matematica e Informatica
Università degli Studi di Catania, Italy

CdS in Informatica
A.A. 2023-2024

Statistica descrittiva

Introduzione

Con il termine statistica si intende l'insieme dei metodi scientifici per raccogliere, riassumere e analizzare i dati relativi ad una popolazione, intesa come una collezione di eventi osservabili (altezze di un gruppo di persone, pezzi di una produzione industriale, insieme di misure relative al carico di rottura di una trave).

In particolare la **statistica descrittiva** studia il modo di rappresentare i dati e di fissare parametri che forniscono indicazioni sintetiche sulla popolazione statistica in esame.

Esempio. Due Paesi con più o meno la stessa popolazione sono colpiti da una grave epidemia. Dopo un mese si effettua una rilevazione statistica sul tasso di mortalità. In entrambi i Paesi sono state ospedalizzate 10.000 persone. Nel Paese X ne sono morte 310, nel paese Y ne sono morte 270.

La popolazione di X si chiede: abbiamo un sistema sanitario peggiore di quella di Y?

Introduzione

Esempio (continuo). Il risultato è particolarmente deludente in quanto X è un ricco Paese europeo, mentre Y è un Paese più povero gestito da un regime autoritario.

Si esegue una indagine statistica per fascia d'età: giovani, adulti, anziani.

Sono stati ospedalizzati:
in X 1000 giovani, 3000 adulti, 6000 anziani (non ce l'hanno fatta 10 giovani, 60 adulti, 240 anziani);
in Y 5000 giovani, 4000 adulti, 1000 anziani (non ce l'hanno fatta 100 giovani, 120 adulti, 50 anziani).

Questo corrisponde ai tassi di mortalità:
in X 1% giovani, 2% adulti, 4% anziani;
in Y 2% giovani, 3% adulti, 5% anziani.

Si scopre che i tassi di mortalità per fasce di età sono in realtà tutti superiori in Y!

Introduzione

Esempio (continuo). La conclusione sul fatto che X abbia un sistema sanitario peggiore di Y sarebbe stata affrettata. La causa va ricercata ad esempio nel fatto che X ha una popolazione con un'aspettativa di vita maggiore (in quanto Paese ricco con qualità della vita maggiore), mentre in Y si muore di meno perché quelli che dovrebbero essere anziani sono in realtà già morti.

La recente pandemia di COVID-19 ci ha abituati ad analisi di questo tipo. Inoltre i media ci hanno sommersi di dati statistici e questo ha messo in evidenza quanto sia importante comprendere analisi di questo tipo.

Oggi per essere cittadini consapevoli è importante avere competenze statistiche tanto quanto la capacità di leggere e scrivere.

Tipologie di dati

I dati che non sono stati organizzati, sintetizzati o elaborati sono chiamati **dati grezzi** e difficilmente forniscono di per sé qualche informazione.

Un'indagine statistica si effettua su una o più caratteristiche della popolazione.
Si parla di

- **caratteri qualitativi** quando essi sono dei dati di natura non numerica
  (es. indagine sul colore delle automobili);
- **caratteri quantitativi** quando essi sono delle grandezze numeriche.
  I caratteri di tipo quantitativo si distinguono in
  - **discreti** se assumono un numero limitato di valori
    (es. indagine sul voto conseguito ad un certo esame universitario);
  - **continui** quando assumono qualsiasi valore reale in un certo intervallo
    (es. indagine sull'altezza di un gruppo di persone).

I dati, se di tipo numerico, vengono raggruppati o per **singoli valori** o per **classi di valori**.

Raggruppamento per singoli valori

Solitamente i dati si ordinano in una sequenza crescente $x_1, x_2, \ldots, x_n$.

Alcuni valori possono presentarsi più volte. La molteplicità con cui un valore si presenta ne costituisce la **frequenza assoluta**. La differenza tra il valore maggiore e quello minore rappresenta la **variabilità** o **range** dei dati.

Indichiamo con f_i la frequenza di x_i e con N la numerosità dell'insieme dei dati.
Allora si ha che $\sum_{i=1}^{n} f_i = N$.

Si definiscono inoltre **frequenze relative** o **probabilità empiriche** le quantità $p_i = f_i/N$.
Allora si ha che $\sum_{i=1}^{n} p_i = 1$.

Si definiscono infine **frequenze relative cumulate** le quantità $F_i = \sum_{k=1}^{i} p_k$.
Queste ultime costituiscono una approssimazione costante a tratti della funzione di ripartizione.

Raggruppamento per singoli valori

Le quantità prima introdotte si riassumo in tabelle del tipo

x_i	f_i	p_i	F_i
x_1	f_1	p_1	F_1
x_2	f_2	p_2	F_2
.	.	.	.
.	.	.	.
.	.	.	.
x_{n-1}	f_{n-1}	p_{n-1}	F_{n-1}
x_n	f_n	p_n	F_n

Definizione. Dicesi **media** della popolazione la quantità

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = \sum_{i=1}^{n} x_i p_i.$$

Definizione. Dicesi **mediana** m_e il valore che è preceduto e seguito da un eguale numero di valori x_i ordinati in modo non decrescente.

Raggruppamento per singoli valori

Proprietà. La mediana rende minima la somma dei valori assoluti degli scarti, cioè la quantità

$$S(a) = \sum_{i=i}^{N} |x_i - a| \qquad \text{è minima se } a = m_e.$$

Osservazione. La mediana suddivide i dati in due gruppi di uguale numerosità. I due gruppi potrebbero non essere distinti ma avere al più un elemento in comune, cioè la stessa mediana nel caso di un numero dispari di valori.

Ciascuno dei due gruppi può, a sua volta, essere diviso in due gruppi di uguale numerosità, in pratica valutandone, come delineato sopra, la mediana. I tre elementi divisori in questo caso si dicono **quartili** e dividono la popolazione in quattro gruppi (non necessariamente distinti) di uguale numerosità.

Ovviamente il secondo quartile coincide con la mediana. La differenza tra il terzo e il primo quartile prende il nome di distanza interquartile. Spesso si usa la semidistanza interquartile, ottenuta dividendo per due la distanza interquartile.

Più in generale si definiscono **quantili** di ordine k i $k-1$ valori che dividono la popolazione in k gruppi di uguale numerosità.

Esempio: distribuzione dei voti di 10 studenti

Consideriamo la distribuzione di voti di 10 studenti riportati in tabella.
Determiniamone la mediana e i quartili.

studente	voto
1	25
2	22
3	28
4	30
5	15
6	18
7	13
8	17
9	25
10	26

Ordiniamo i valori in ordine crescente

13, 15, 17, 18, 22, 25, 25, 26, 28, 30

I due valori centrali sono 22 e 25 e quindi la mediana è data dalla loro media aritmetica, cioè $m_e = 23.5$.

Per determinare il primo quartile consideriamo il gruppo formato dai primi cinque elementi. La mediana di questo gruppo è il valore centrale 17, che è quindi il primo quartile. Similmente si trova che il terzo quartile è 26 considerando il gruppo dei rimanenti 5 elementi.

In realtà quando si hanno valori ripetuti il calcolo dei quantili non è sempre immediato e bisogna ricorrere alla procedura che verrà delineata per il caso di dati raggruppati.

Raggruppamento per singoli valori

Definizione. I valori di frequenza massima costituiscono la **moda**. Se l'elemento di massima frequenza è unico si ha un insieme di dati **unimodale** altrimenti l'insieme è **multimodale**.

Definizione. Dicesi **varianza empirica** la quantità

$$\bar{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 p_i.$$

La quantità $\bar{\sigma}$ rappresenta la **deviazione standard empirica**.

Definizione. Dicesi **momento empirico centrato** di ordine r con $r > 1$ la quantità

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^r f_i = \sum_{i=1}^{n}(x_i - \bar{x})^r p_i.$$

Osservazione. Spesso si vuole vedere se i dati si possono considerare ottenuti da una distribuzione teorica di probabilità.

Definizione. Si definiscono i seguenti coefficienti

$$\beta_1 = \frac{(\mu_3)^2}{(\mu_2)^3} \qquad \text{(coefficiente di asimmetria o skewness)},$$
$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} \qquad \text{(kurtosi)}.$$

Osservazione. Si può calcolare che per una distribuzione normale si ha che $\beta_1 = 0$ e $\beta_2 = 3$. Pertanto affinché i dati seguano una distribuzione normale occorre necessariamente che $|\beta_1| \ll 1$ e che $\beta_2 \approx 3$.

Raggruppamento per classi di valori

Nel raggruppamento per classi si considera una suddivisione dei valori in sottointervalli. Ciascun sottoinsieme della suddivisione costituisce una **classe**.

Gli estremi e il punto medio di ciascun sottointervallo costituiscono gli **estremi** e il **valore centrale** della classe.

In generale non si richiede che le classi siano contigue. In tale caso si definiscono pure i **confini** delle classi considerando il valore medio tra l'estremo superiore di una classe e quello inferiore della classe successiva.

Raggruppamento per classi di valori

Esempio. Nella tabella seguente sono stati raccolti i dati relativi al peso di 100 studenti. Si hanno 5 classi.

peso (kg)	f_i	p_i	F_i
60-62	5	0.05	0.05
63-65	18	0.18	0.23
66-68	42	0.42	0.65
69-71	27	0.27	0.92
72-74	8	0.08	1.0

I valori 60 e 62 sono i limiti inferiori e superiori della prima classe, 63 e 65 quelli della seconda classe e via dicendo. I confini della prima classe sono 59.5 e 62.5, quelli della seconda classe 62.5 e 65.5, etc. Si osservi che nel caso della prima classe il valore 59.5 in realtà è arbitrario e qui è stato scelto solo per ottenere delle classi di eguale ampiezza.

Funzione di ripartizione empirica

All'interno di ciascuna classe si assume una approssimazione lineare della funzione di ripartizione ottenendo una poligonale come segue: nella prima classe $F(x)$ si considera variabile tra zero e F_1, nella seconda classe tra F_1 e F_2, ed infine nell'ultima classe tra F_{n-1} e 1.

Tale ricostruzione empirica della funzione di ripartizione consente di definire i quantili q_α di ordine $\alpha \in [0,1]$ come nel caso teorico, cioè q_α soddisfa $\alpha = F(q_\alpha)$.

Illustriamo il procedimento per determinare il quantile di ordine α.

1. Determiniamo l'indice j tale che $F_{j-1} < \alpha \leq F_j$.
2. Nell'intervallo $[x_{j-1}, x_j]$ si ha

$$F(x) = F_{j-1} + \frac{F_j - F_{j-1}}{x_j - x_{j-1}}(x - x_{j-1}), \qquad \forall x \in [x_{j-1}, x_j].$$

3. Sostituendo $F(x) = \alpha$ e $x = q_\alpha$ si ha

$$q_\alpha = x_{j-1} + (\alpha - F_{j-1})\frac{x_j - x_{j-1}}{F_j - F_{j-1}}.$$

Rappresentazioni grafiche

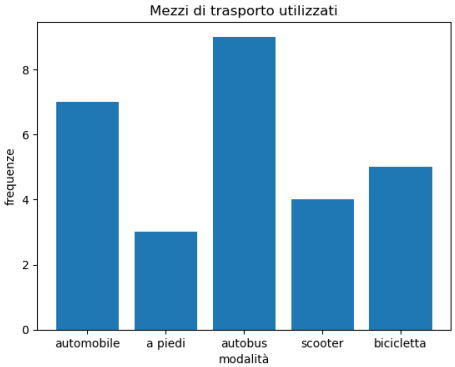
Istogrammi

A partire dalle distribuzioni di frequenza possiamo costruire grafici espressivi e ben leggibili che risultano utili in fase di comunicazione.

Istogramma. Si riportano sull'asse verticale le frequenze e su quello orizzontale si rappresentano tanti segmenti quante sono le modalità.

Esempio: Mezzi di trasporto utilizzati

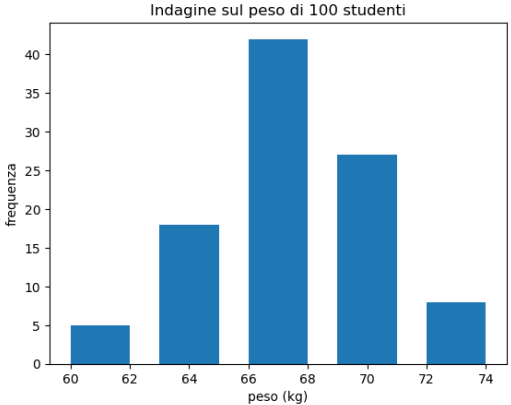
Modalità	Frequenza
automobile	7
a piedi	3
autobus	9
scooter	4
bicicletta	5
totale	28



Istogrammi

Esempio: Indagine sul peso di 100 studenti.

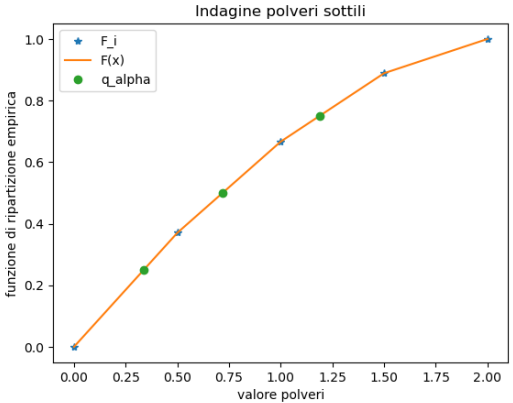
peso (kg)	f_i	p_i	F_i
60-62	5	0.05	0.05
63-65	18	0.18	0.23
66-68	42	0.42	0.65
69-71	27	0.27	0.92
72-74	8	0.08	1.0



Istogrammi

Esercizio. Un monitoraggio sulla densità di polveri sottili nell'aria ha condotto alle frequenze riportate nella tabella sotto, in opportune unità di misura. Si calcolino i quartili empirici.

classi	frequenze
$[0, 0.5[$	10
$[0.5, 1[$	8
$[1, 1.5[$	6
$[1.5, 2]$	3



Istogrammi

Esercizio. L'altezza di 2000 individui di una popolazione è riportata nel file 'Data_altezze.dat'.

1. Calcolare media e deviazione standard della popolazione.
2. Costruire un istogramma a 20 barre.
3. È possibile adattare ai dati una distribuzione normale?

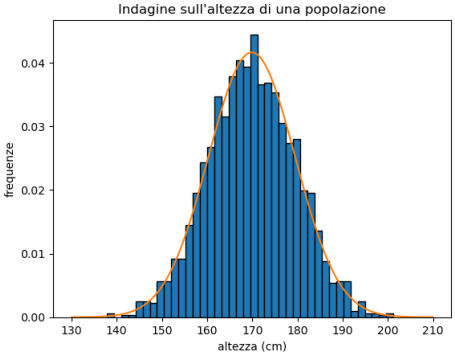
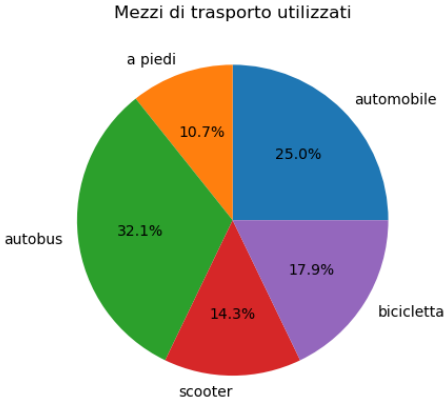


Grafico a torta (pie chart)

Un grafico a torta o aerogrmma è costruito a partire da un cerchio suddiviso in settori circolari. Ciascun settore rappresenta una classe. L'ampiezza di ciascun settore è proporzionale alla relativa frequenza percentuale.

Esempio: Mezzi di trasporto utilizzati

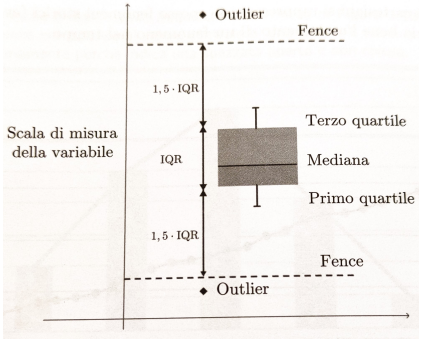
Modalità	Frequenza
automobile	7
a piedi	3
autobus	9
scooter	4
bicicletta	5
totale	28



Box-plot

Il **box-plot**, il cui nome completo è box and whiskers plot (diagramma a scatola e baffi), è composto da un rettangolo (box) e da due tratti a T (baffi).

Il rettangolo è compreso tra il primo e il terzo quartile e mostra l'ampiezza della metà centrale della distribuzione. L'altezza della scatola è il range interquartile (IQR), cioè la distanza tra il primo e il terzo quartile. La linea all'interno della scatola rappresenta la mediana.

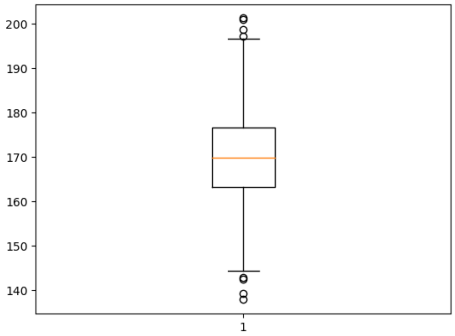


I valori anomali (outliers) sono determinati come quei valori che si trovano al di fuori delle barriere (fence), cioè al di fuori del range $[q_{0.25} - 1.5 \cdot IQR, q_{0.75} + 1.5 \cdot IQR]$.

I due tratti a T sono detti baffi (whiskers). La fine del baffo è determinata dal più piccolo (risp. grande) valore che non sia anomalo.

Box-plot

Esempio. L'altezza di 2000 individui di una popolazione è riportata nel file 'Data_altezze.dat'. Costruiamo il box-plot.



I cerchietti rappresentano gli outliers.

Per l'implementazione in Python si veda: matplotlib.pyplot.boxplot

Grafico di probabilità normale

Il **grafico di probabilità normale** è una tecnica grafica per stabilire se un insieme di dati segue approssimativamente una distribuzione normale o meno.

Si ordinano i dati in modo crescente e per ogni osservazione ordinata si trova il valore che ci aspetteremmo per l'osservazione se i dati seguissero una distribuzione normale standard.

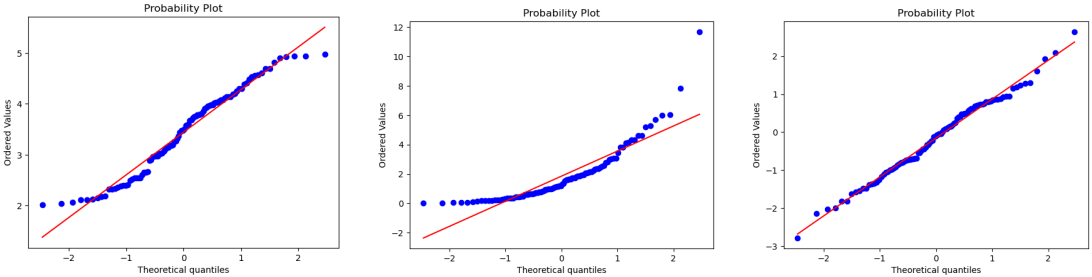
Questo calcolo si effettua in modo approssimato.
Per ogni dato x_i si determina z_i tale che

$$\Phi(z_i) = \begin{cases} 1 - 0.5^{1/n} & \text{se } i = 1 \\ \dfrac{i - 0.3175}{n + 0.365} & \text{se } i = 2, 3, \ldots, n-1 \\ 0.5^{1/n} & \text{se } i = n \end{cases}$$

ove n è il numero di osservazioni.

Per l'implementazione in Python si veda: scipy.stats.probplot

Grafico di probabilità normale



Le figure mostrano i grafici di probabilità normale per tre differenti dataset.
Dal grafico a sinistra e da quello centrale si evince che i dati non seguono una distribuzione normale, nel grafico a destra si osserva invece un buon adattamento dei dati con la distribuzione normale.

Grafico quantile-quantile (Q-Q plot)

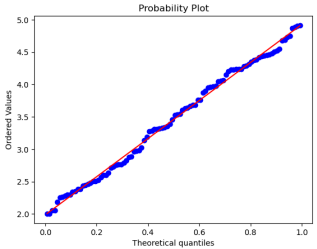
Più in generale, si parla di **grafico quantile-quantile** o **Q-Q plot**.

Esso è una tecnica grafica per stabilire se un insieme di dati segue approssimativamente una data distribuzione o meno.

Si esegue confrontando i quantili empirici, calcolati a partire dal dataset, con quelli teorici della distribuzione di cui si vuole verificare l'adattamento.

Per l'implementazione in Python si veda: scipy.stats.probplot

In questo caso occorre specificare la distribuzione target.



La figura mostra il *Q-Q plot* del dataset di sinistra della precedente slide, confrontato con i quantili della distribuzione uniforme.
Si osserva un buon adattamento.

Statistica inferenziale

Statistica inferenziale

La **statistica inferenziale** è l'insieme delle metodologie atte a trarre conclusioni circa una popolazione statistica dalla analisi di **campioni** estratti dalla popolazione in esame.

Esempio. L'analisi dei pezzi difettosi di una produzione. Spesso non è possibile controllare tutti i pezzi per l'eccessiva numerosità ed inoltre la produzione potrebbe ancora continuare nel futuro. Quindi se si vuole stimare la percentuale di pezzi guasti, in molte circostanze l'unica possibilità pratica è di analizzare sottoinsiemi dell'intera popolazione.

Definizione. Un insieme di v.a. indipendenti $X_1, X_2, \ldots, X_n$ di medesima legge si dice **campione** di rango n.

Osservazione. Se la legge è gaussiana il campione si dice di tipo gaussiano, se la legge è quella esponenziale il campione si dice di tipo esponenziale, eccetera. Assunta la tipologia di legge, occorre stimarne i parametri che la identificano.

Stimatori puntuali

Definizione. Dato un campione di rango n, $X_1, X_2, \ldots, X_n$, dicesi **statistica** una v.a. T del tipo

$$T = t(X_1, X_2, \ldots, X_n)$$

ove t è una generica funzione del campione.

Definizione. Sia θ un parametro incognito e $\psi(\theta)$ una sua funzione. Si dice **stimatore** (puntuale) di $\psi(\theta)$ una qualunque statistica T a valori nell'insieme delle immagini di $\psi(\theta)$.

Esempio. In una catena di produzione, si considerano n pezzi e per ciascuno si verifica se è difettoso o meno, assegnando al pezzo il valore uno se è guasto, zero altrimenti. Assumendo che il guasto di ogni pezzo sia indipendente dagli altri, si ottiene un campione di rango n di leggi di Bernoulli $B(1, p)$ il cui parametro p è incognito e si dovrà stimare dal campione in esame. Si consideri lo stimatore

$$\hat{p} = t(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

che per la legge dei grandi numeri tende in probabilià al valore teorico p per $n \to \infty$.

Stimatori puntuali

Definizione. Uno stimatore T di $\psi(\theta)$ dicesi **non distorto** se

$$E^{\theta}[T] = \psi(\theta),$$

ove $E^{\theta}[T]$ è da intendersi come la media di T rispetto alla densità $f^{\theta}(x)$ con θ incognito.

Osservazione. L'essere non distorto è un proprietà desiderabile per uno stimatore in quanto se analizziamo un singolo campione il valore di T in generale non coincide con $\psi(\theta)$. Però, in linea teorica, raccogliendo un numero sempre più grande di campioni, la media dei corrispondenti valori di T tenderà al valore teorico $\psi(\theta)$.

Definizione. Dato un campione $X_1, X_2, \ldots, X_n$ si chiama **media campionaria**

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Stimatori puntuali

Proprietà. Dato un campione $X_1, X_2, \ldots, X_n$, la media campionaria $\bar{X}_n$ è uno stimatore non distorto della media teorica μ.

Dimostrazione. Dalle proprietà della media si ha

$$E^\theta[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E^\theta[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu.$$

Per la varianza occorre fare una distinzione: il caso in cui è nota la media teorica e il caso in cui non lo è.

Proprietà. Dato un campione $X_1, X_2, \ldots, X_n$, di cui è nota la media teorica μ, uno stimatore non distorto della varianza teorica σ^2 è dato da

$$\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2.$$

Dimostrazione. Dalle proprietà della media e della varianza si ha

$$E^\theta[\bar{\sigma}_n^2] = \frac{1}{n}\sum_{i=1}^{n} E^\theta[X_i^2] - 2E^\theta[X_i]\mu + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} E^\theta[X_i^2] - E^\theta[X_i]^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma^2 = \sigma^2.$$

Stimatori puntuali

Osservazione. Se non è nota la media teorica, la varianza empirica

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

costituisce uno stimatore che non gode della proprietà di essere non distorto.

Proprietà. Dato un campione $X_1, X_2, \ldots, X_n$, lo stimatore

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

costituisce uno stimatore non distorto della varianza teorica σ^2 delle X_i.
Dimostrazione. Osserviamo intanto che

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

$$= \left(\sum_{i=1}^{n}X_i^2\right) - 2n\bar{X}_n^2 + n\bar{X}_n^2 = \left(\sum_{i=1}^{n}X_i^2\right) - n\bar{X}_n^2.$$

Stimatori puntuali

Dimostrazione (continuo). Dalla proprietà della varianza $\mathrm{VAR}(X) = E[X^2] - (E[X])^2$, si ha

$$E^\theta[X_i^2] = \mathrm{VAR}^\theta(X_i) + (E^\theta[X_i])^2 = \sigma^2 + \mu^2$$

ove μ è la media teorica delle X_i e

$$E^\theta[\bar{X}_n^2] = \mathrm{VAR}^\theta(\bar{X}_n) + (E^\theta[\bar{X}_n])^2 = \left(\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{VAR}^\theta(X_i)\right) + \mu^2 = \frac{\sigma^2}{n} + \mu^2$$

Allora, si ha

$$E^\theta[S^2] = \frac{1}{n-1}\left[\left(\sum_{i=1}^{n}E^\theta[X_i^2]\right) - nE^\theta[\bar{X}_n^2]\right] = \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]$$

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

Stimatori puntuali

Osservazione. Per $n \gg 1$, $\bar{\sigma}^2$ e S^2 forniscono valori vicini, ma per piccoli campioni si può avere una differenza significativa.

Esempio. Se lanciando una moneta 10 volte esce testa 8 volte, si può stimare la probabilità di successo in un singolo lancio tramite $\hat{p} = 8/10$ e la varianza tramite

$$S^2 = \frac{1}{10-1}\sum_{i=1}^{10}\left(X_i - \frac{8}{10}\right)^2 = \frac{1}{9}\left[2\left(0 - \frac{8}{10}\right)^2 + 8\left(1 - \frac{8}{10}\right)^2\right] = \frac{1}{9}\left[2\left(\frac{8}{10}\right)^2 + 8\left(\frac{2}{10}\right)^2\right]$$
$$\approx 0.1778$$

Proprietà. Nel caso di campioni dati da coppie (X_i, Y_i), $i = 1, 2, \ldots, n$ di v.a. indipendenti di medesima legge (congiunta), uno stimatore del coefficiente di correlazione tra X_i e Y_i è dato da

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}$$

Proprietà degli stimatori

Proprietà. Siano X e Y due v.a. indipendenti con $X \sim N(0,1)$ e $Y \sim \chi^2(n)$. Allora $T = \frac{X}{\sqrt{Y}}\sqrt{n} \sim t(n)$.

Proprietà. Supponiamo che $Z_1, Z_2, \ldots, Z_n$ sia un campione di leggi gaussiane di media μ e varianza σ^2. Allora gli stimatori

$$Z = \frac{\bar{Z}_n - \mu}{\sigma}\sqrt{n} \qquad \text{e} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2,$$

con $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$, sono v.a. indipendenti e

$$W = \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1).$$

Proprietà degli stimatori

Proprietà. Supponiamo che $Z_1, Z_2, \ldots, Z_n$ sia un campione di leggi gaussiane di media μ e varianza σ^2. Allora

$$T = \frac{Z}{\sqrt{W}}\sqrt{n-1} = \frac{\bar{Z}_n - \mu}{S}\sqrt{n} \sim t(n-1).$$

Dimostrazione. Segue dalle due proprietà precedenti. Infatti

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

e

$$W = \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1).$$

Per la prima proprietà si ha

$$T = \frac{Z}{\sqrt{W}}\sqrt{n-1} = \frac{\bar{Z}_n - \mu}{\sigma}\sqrt{n}\frac{\sigma}{S\sqrt{n-1}}\sqrt{n-1} = \frac{\bar{Z}_n - \mu}{S}\sqrt{n} \sim t(n-1).$$

Intervalli di confidenza

Gli **intervalli di confidenza** non danno un singolo valore per la quantità da stimare ma un intervallo che viene dedotto dai valori del campione in esame e che quindi ha carattere aleatorio.

Sostanzialmente si garantisce, entro un errore non superiore ad α, che il valore vero dello stimatore stia nell'intervallo considerato. Si noti, però, che prendere α eccessivamente piccolo, se da un lato diminuisce la probabilità di sbagliare dall'altro rende l'intervallo troppo grande e praticamente inutile.

Di solito nelle applicazioni si assume $\alpha = 0.05$ oppure $\alpha = 0.01$ ovvero, in termini percentuali, si stabiliscono intervalli di confidenza con un livelli di fiduciosità del 95% o del 99%.

Definizione. Dato un campione $X_1, X_2, \ldots, X_n$ e assegnato $\alpha \in [0, 1]$, dicesi **intervallo di confidenza** di livello $1 - \alpha$ per $\psi(\theta)$ l'intervallo (aleatorio) $I_X = [T_1, T_2]$ tale che

$$P^\theta(\psi(\theta) \in I_X) = 1 - \alpha$$

ove $T_1 = t_1(X_1, X_2, \ldots, X_n)$ e $T_2 = t_2(X_1, X_2, \ldots, X_n)$ sono due statistiche con $T_1 < T_2$.

Intervallo di confidenza per la media

Supponiamo che $X_1, X_2, \ldots, X_n$ sia un campione di leggi gaussiane oppure che sia valida l'approssimazione normale. Sia μ la media di ciascuna X_i e sia $\alpha \in [0,1]$ fissato.

Bisogna determinare un intervallo $I = [a,b]$ tale che

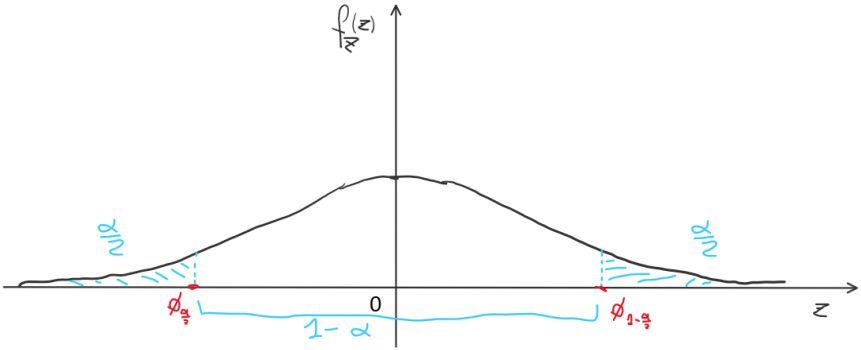
$$P(a \leq \mu \leq b) = 1 - \alpha.$$

Distinguiamo due casi.

1. Supponiamo che la varianza σ^2 delle X_i sia nota.
In tale caso per il Teorema del limite centrale si ha, approssimativamente, che

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

Intervallo di confidenza per la media



Si osserva che

$$P(\phi_{\alpha/2} \leq Z \leq \phi_{1-\alpha/2}) = 1 - \alpha.$$

Inoltre si ha che $\phi_{1-\alpha/2} = -\phi_{\alpha/2}$, quindi

$$P(-\phi_{1-\alpha/2} \leq Z \leq \phi_{1-\alpha/2}) = 1 - \alpha.$$

Intervallo di confidenza per la media

Abbiamo quindi che

$$-\phi_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \leq \phi_{1-\alpha/2}.$$

Da cui si ottiene

$$\bar{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2}.$$

Quindi l'intervallo di confidenza per la media nel caso in cui la varianza sia nota è

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2},\, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\phi_{1-\alpha/2}\right].$$

Esercizio

Si vuole testare un dispositivo con uno strumento che fornice delle misure di voltaggio. Si eseguono 9 misurazioni registrando i valori in volt

$$11,\ 13.2,\ 12.3,\ 10.9,\ 13,\ 10.5,\ 12.3,\ 13,\ 13.15.$$

È nota la precisione dello strumento e si ha $\sigma = 1$ V.

1. Si determinino gli intervalli di confidenza al 95% e al 99%.
2. Determinare gli stessi intervalli di confidenza nel caso in cui si avesse $\sigma = 1.4$ V.
3. Sempre con precisione $\sigma = 1$ V, determinare gli stessi intervalli con la stessa media delle misure ma supponendo che essa provenga da un campione di 20 misurazioni.

Intervallo di confidenza per la media

2. Supponiamo che la varianza σ^2 delle X_i NON sia nota.

Da una proprietà enunciata in precedenza e usando il Teorema del limite centrale si ha, approssimativamente, che

$$T = \frac{\bar{X}_n - \mu}{S}\sqrt{n} \sim t(n-1),$$

essendo S^2 lo stimatore della varianza.
Di conseguenza, osserviamo che che

$$P(-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2}) = 1 - \alpha.$$

Da cui si ottiene che

$$-t_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{S}\sqrt{n} \leq t_{1-\alpha/2}.$$

Quindi l'intervallo di confidenza per la media nel caso in cui la varianza non sia nota è

$$\left[\bar{X}_n - \frac{S}{\sqrt{n}}t_{1-\alpha/2},\, \bar{X}_n + \frac{S}{\sqrt{n}}t_{1-\alpha/2}\right].$$

Esercizio

Viene effettuato un test di rottura di un certo materiale ottenendo i seguenti valori in megapascal (MPa).

19.8	10.1	14.9	7.5	15.4	15.4
15.4	18.5	7.9	12.7	11.9	11.4
11.4	14.1	17.6	16.7	15.8	
19.5	8.8	13.6	11.9	11.4	

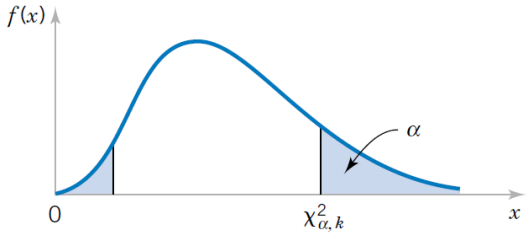
Dopo aver verificato graficamente che il campione proviene da una popolazione distribuita approssimativamente in modo normale, determinare l'intervallo di confidenza al 95% per la media.

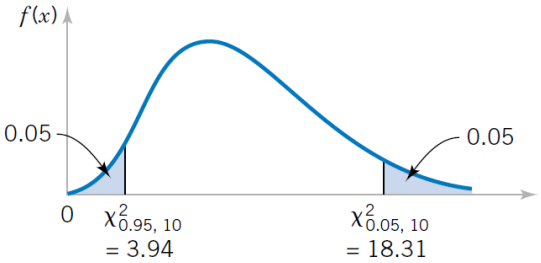
Intervallo di confidenza per la varianza

Supponiamo che $X_1, X_2, \ldots, X_n$ sia un campione di leggi gaussiane oppure che sia valida l'approssimazione normale. Sia σ^2 la varianza di ciascuna X_i e sia $\alpha \in [0, 1]$ fissato.

Da una proprietà enunciata in precedenza e usando il Teorema del limite centrale si ha, approssimativamente, che

$$W = \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1).$$





Intervallo di confidenza per la varianza

Si osservi che

$$P\left(\chi^2_{\alpha/2}(n-1) \leq W \leq \chi^2_{1-\alpha/2}(n-1)\right) = 1-\alpha.$$

Da cui si ottiene che

$$\chi^2_{\alpha/2}(n-1) \leq \frac{S^2}{\sigma^2}(n-1) \leq \chi^2_{1-\alpha/2}(n-1).$$

Di conseguenza

$$\frac{1}{\chi^2_{1-\alpha/2}(n-1)} \leq \frac{\sigma^2}{S^2(n-1)} \leq \frac{1}{\chi^2_{\alpha/2}(n-1)}$$

e infine

$$\frac{S^2(n-1)}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{S^2(n-1)}{\chi^2_{\alpha/2}(n-1)}.$$

Quindi l'intervallo di confidenza per la varianza è

$$\left[\frac{S^2(n-1)}{\chi^2_{1-\alpha/2}(n-1)}, \frac{S^2(n-1)}{\chi^2_{\alpha/2}(n-1)}\right].$$

Intervallo di confidenza per la varianza

È possibile anche determinare intervalli di confidenza unilateri.
Cioè osservando che

$$P\left(\chi^2_\alpha(n-1) \leq W\right) = 1-\alpha$$

si trova così che

$$\sigma^2 \geq \frac{S^2(n-1)}{\chi^2_\alpha(n-1)},$$

che rappresenta l'intervallo di confidenza per il limite inferiore della varianza.

Analogamente, si può osservare che

$$P\left(W \leq \chi^2_{1-\alpha}(n-1)\right) = 1-\alpha.$$

Si trova così che

$$\sigma^2 \leq \frac{S^2(n-1)}{\chi^2_{1-\alpha}(n-1)},$$

che rappresenta l'intervallo di confidenza per il limite superiore della varianza.

Esercizio

Un macchinario riempie automaticamente delle bottiglie. Da un campione di 20 misurazioni si ottengono i seguenti valori (in litri)

2.05, 2.04, 1.98, 1.96, 2.03, 2.01, 1.97, 1.99, 2.01, 2.05
1.96, 1.95, 2.04, 2.01, 1.97, 1.96, 2.02, 2.04, 1.98, 1.94

Se la varianza fosse troppo grande, la proporzione di bottiglie sotto o sovrariempite sarebbe non accettabile.

Calcolare l'intervallo di confidenza al 95% per il limite superiore per la deviazione standard.

Intervallo di confidenza per la proporzione

Supponiamo che $X_1, X_2, \ldots, X_n$ sia un campione di leggi di Bernoulli, cioè $X_i \sim B(1,p)$ per ogni $i = 1, 2, \ldots, n$ e sia $\alpha \in [0,1]$ fissato.

Utilizzando l'approssimazione normale

$$Z = \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} = \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}\sqrt{n}$$

segue approssimativamente una legge $N(0,1)$.

Osservando che $P(|Z| \leq \phi_{1-\alpha/2}) = 1 - \alpha$ si ottiene la relazione

$$\bar{X}_n - \phi_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X}_n + \phi_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Esplicitando rispetto a p, per $n \gg 1$ sviluppando in serie di $1/n$ e arrestandosi al primo ordine si ottiene l'intervallo di confidenza approssimato per la proporzione

$$\left[\bar{X}_n - \phi_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}},\ \bar{X}_n + \phi_{1-\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right].$$

Intervallo di confidenza per la proporzione

Ad esempio, supponiamo che un campione di ampiezza n è estratto da una popolazione e supponiamo di avere due classi di interesse e che X osservazioni di questo campione appartengono ad una classe. Quindi $p = X/n$ è uno stimatore della proporzione della popolazione che appartiene a questa classe.

Esempio. Un macchinario produce un componente di un motore per automobili. Si considera un campione di 85 pezzi e si osserva che 10 hanno una rugosità superficiale maggiore delle specifiche consentite.
Quindi in questo caso $\bar{X}_n = 10/85 = 0.12$.
Applicando la formula della slide precedente, un intervallo di confidenza al 95% per la proporzione è dato da

$$0.12 - 1.96\sqrt{\frac{0.12 \cdot 0.88}{85}} \leq p \leq 0.12 + 1.96\sqrt{\frac{0.12 \cdot 0.88}{85}}$$

ovvero

$$0.05 \leq p \leq 0.19$$

Riferimenti bibliografici

1. V. Romano, Metodi matematici per i corsi di ingegneria, Città Studi, 2018.
2. P. Baldi, Calcolo delle probabilità e statistica, Mc Graw-Hill, Milano, 1992.
3. R. Scozzafava, Incertezza e probabilità, Zanichelli, 2001.