Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

Analisi e valutazione del benessere della società: caso studio sugli indici di vita

Carmela Pia Senatore
Università degli studi di Salerno

1. Introduzione

L'analisi statistica coinvolge sempre un gruppo di elementi, che possono essere individui, oggetti o altro, in cui si manifesta il fenomeno di interesse. Questo gruppo è noto come "popolazione" o "universo" e può comprendere un numero limitato o infinito di unità. Nel primo caso, si parla di una "popolazione finita", mentre nel secondo caso, si tratta di una "popolazione infinita". Per comprendere le caratteristiche di una popolazione finita, è possibile osservare l'intero gruppo di elementi o un sottoinsieme estratto da esso, che chiamiamo "campione". Nel caso di una popolazione infinita, invece, l'analisi statistica può essere condotta solo attraverso l'analisi di un campione estratto da essa.

L'analisi dei fenomeni coinvolge l'impiego dell'inferenza statistica, un ramo della statistica che si differenzia dalla probabilità in un modo peculiare:

- Nella probabilità, si affrota un processo che genera dati e ci si chiede quale sia la probabilità di un certo evento.
- Nell'inferenza statistica, invece, si parte dai risultati osservati e ci si pone la domanda opposta: cosa possiamo dedurre sul processo che ha originato tali dati?

2. Variabili casuali

Una variabile casuale è una regola che associa ad ogni evento un unico numero reale. Formalmente, una **variabile casuale** è una funzione misurabile a valori reali definita su uno spazio. L'insieme dei valori che una variabile casuale può assumere con probabilità positiva si chiamerà **supporto** della variabile casuale.

$$x : A- > x \in R$$

Crea una corrispondenza tra il dominio degli eventi e il dominio R dell'insieme dei numeri. Il supporto della variabile casuale può far distinguere la v.c. in:

- DISCRETA: se il supporto è un insieme finito o numerabile di numeri reali.

- CONTINUA: se il suo supporto è un intervallo limitato o illimitato di R

2.1. Variabili casuali Continue

Una variabile casuale **continua** può assumere tutti i valori compresi in un intervallo misurabile reale. Formalmente , una v.c. x continua è una funzione misurabile e a valori reali che assegna a ogni evento E di uno spazio di probabilità continuo un numero reale $x \in R$. Le v.c. continue presentano una complessità aggiunta analiticamente poichè per queste non è possibile elencare tutti i valori assunti dalla v.c. essendo un infinità non numerabile. A questo tipo di variabile casuale viene associata una **funzione di densità**. Una v.c. continua X è ben definita se, per ogni x_0 reale e prefissato, è nota la probabilità che tale v.c. assuma un valore in un intervallo di ampiezza infinitesimo rispetto a x_0 mediante la relazione seguente:

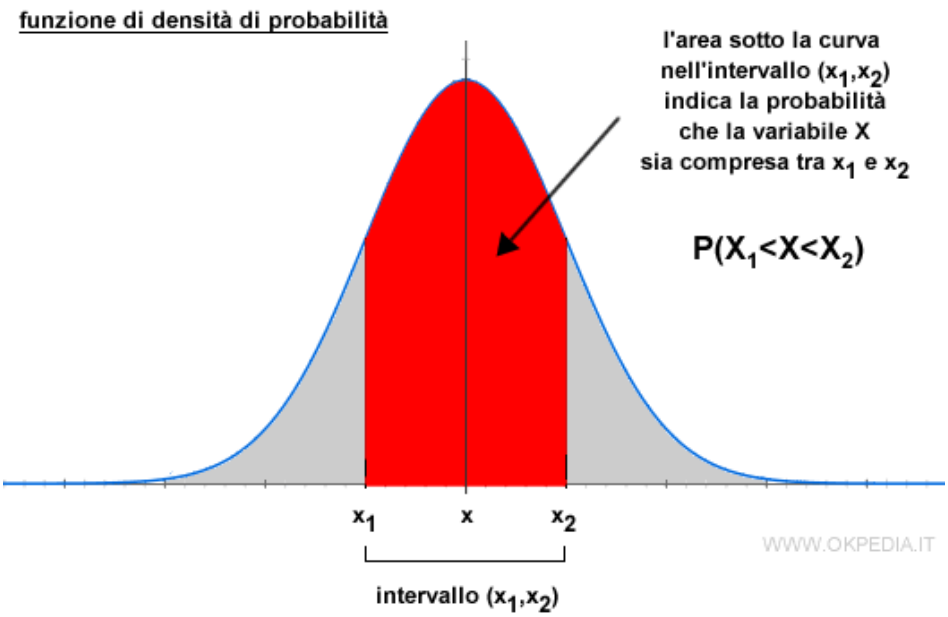


Figure 1: Funzione di probabilità

$$P_r(x_0 < X \leq x_0 + d_x) = f(x_0)dx$$

Dal punto di vista grafico la probabilità corrisponde all'aria sottesa alla curva di densità nell'intervallo

$$[x_0; x_1]$$

. E' pressocchè impossibile calcolare la probabilità che

$$x = x_0$$

per il semplice fatto che questo integrale avrà ampiezza nulla. E' infatti un evento possibile ma con probabilità nulla.

La funzione

$$F(x)$$

è detta **funzione di ripartizione** ed è una funzione matematica che descrive la probabilità che la variabile casuale assume un valore minore o uguale a un certo valore, calcolata nel punto x_0:

$$F(x_0) = Pr(X \leq x_0) = \int_{-\infty}^{x_0} f(w)d_w$$

In R tramite dei prefissi da inserire prima della v.c. a cui si vuole far riferimento permettono di conoscere le principali informazioni:

- Il prefisso r permette di simulare in maniera casuale un campione che simula la distribuzione della variabile casuale a cui si vuole far riferimento;

- il prefisso p permette di calcolare la funzione di distribuzione in un punto o in un insieme di punti di una determinata variabile casuale;
- il prefisso q permette di calcolare i quantili della v.c.;
- il prefisso d permette di calcolare la funzione di probabilità (densità) in un punto o in un insieme di punti di una determinata variabile casuale.

2.2. Variabile casuale Normale

La distribuzione normale gioca un ruolo fondamentale nel campo del calcolo delle probabilità e della statistica, essendo considerata la più importante. Il suo ruolo centrale è così significativo che è difficile immaginare lo sviluppo di tali discipline senza riconoscerle un ruolo di primaria importanza. Il termine "normale" nel suo nome suggerisce che questa distribuzione è la norma, indicando la sua presenza quasi regolare nelle applicazioni e nei sviluppi teorici.

L'introduzione della distribuzione normale nella storia della probabilità è collegata alla ricerca della distribuzione degli errori casuali. Sin dal 1632, Galileo descrisse con precisione le caratteristiche essenziali che una distribuzione degli errori doveva possedere nel "Dialogo dei massimi sistemi". Tuttavia, non si preoccupò di derivarne la forma analitica in quel contesto. La prima formulazione analitica della funzione di densità della variabile casuale normale apparve nel 1733 grazie a De Moivre, che la utilizzò come approssimazione alla somma di variabili casuali binomiali. Successivamente, negli anni 1770-71, Daniel Bernoulli fornì la prima tavola della funzione di densità, mentre Laplace, a partire dal 1810, la incluse nei suoi Teoremi Limite Centrale.

La variabile casuale normale non solo approssima la distribuzione empirica di molti fenomeni reali, ma funge anche da punto di riferimento per stabilire confronti e dedurre risultati asintotici. La sua presenza nelle applicazioni e la sua importanza nei confronti teorici la rendono una componente essenziale nel panorama del calcolo delle probabilità e della statistica.

Le proprietà della v.c. Normale:

- 1. Simmetria: La distribuzione normale è simmetrica rispetto alla sua media, il che significa che la metà dei dati si trova sopra la media e l'altra metà sotto la media. La curva a campana della distribuzione normale è simmetrica rispetto al suo valore centrale.

- 2. Media e mediana coincidono: Nella distribuzione normale, la media e la mediana hanno lo stesso valore, il che la rende particolarmente utile per rappresentare dati centrali.

- 3. Concentrazione dei dati: La maggior parte dei dati in una distribuzione normale si concentra vicino alla media, con code più sottili che si estendono verso i valori più alti o più bassi. Questo comportamento è utile per comprendere la tendenza centrale dei dati.

- 4. Parametri ben definiti: La distribuzione normale è completamente descritta dalla sua media (valore atteso) e dalla deviazione standard (misura della dispersione). Questi due parametri sono sufficienti per caratterizzare completamente la distribuzione.

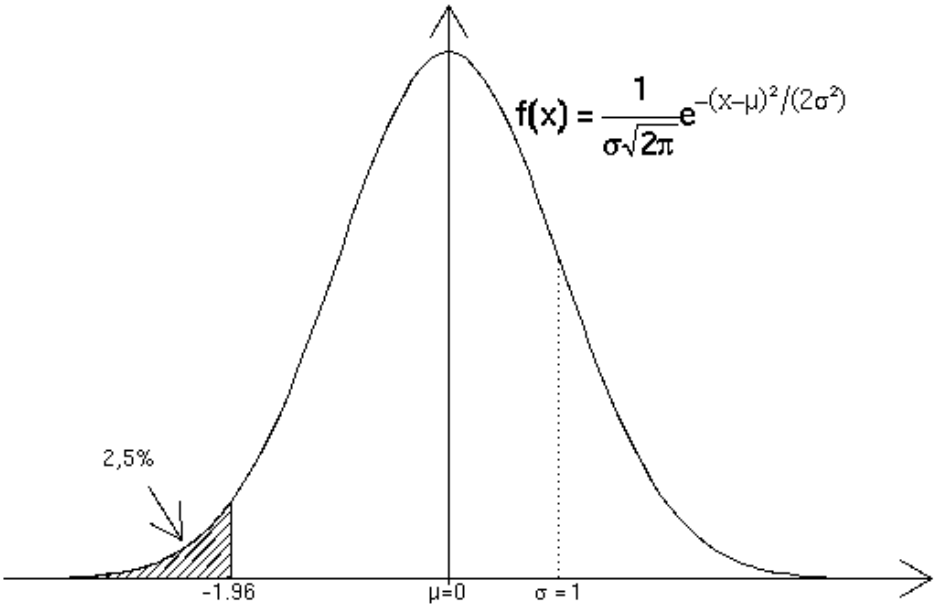


Figure 2: Funzione di densità della v.c. Normale

- 5. Teorema del limite centrale: Questo teorema afferma che la somma di un grande numero di variabili casuali indipendenti, ciascuna con una distribuzione qualsiasi, avrà una distribuzione approssimativamente normale. Questo rende la distribuzione normale fondamentale in molte applicazioni, poiché molte variabili casuali reali tendono a comportarsi in modo approssimativamente normale quando si sommano.

- 6. Facilità di calcolo: La distribuzione normale è matematicamente ben definita, il che rende più facile effettuare calcoli e condurre analisi statistiche. Esistono tavole standard e software statistici che semplificano notevolmente il calcolo delle probabilità e dei punteggi z (z-scores) associati alla distribuzione normale.

- 7. Legge dei grandi numeri: La distribuzione normale è correlata alle leggi dei grandi numeri, che stabiliscono che con un numero sufficientemente grande di campioni, le medie campionarie si avvicineranno alla media della popolazione e seguiranno una distribuzione normale.

Una v.c. X continua si dice v.c. Normale (oppure v.c. Gaussiana) con parametri μ ,per indicare la media, e σ^2 , per indicare la varianza, la si indica con $X \sim N(\mu, \sigma^2)$, se è definita su tutto l'asse reale inoltre la funzione di densità di probabilità è:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})}$$

con

$$-\infty < x < \infty$$

Poiché il parametro μ può assumere qualsiasi valore reale mentre il parametro σ^2 può solo essere non negativo, lo spazio parametrico di questa v.c. è il semipiano positivo, cioè:

$$\Omega(\mu, \sigma^2) = (\mu, \sigma^2) : -\infty < \mu < +\infty; 0 \leq \sigma^2 < +\infty$$

Lo studio analitico della funzione di densità di una v.c. Normale mostra che essa ha una forma campanulare simmetrica rispetto al suo valore medio (punto di ascissa $x = \mu$), in corrispondenza del quale si presenta il massimo ovvero $f(x) = \frac{1}{\sigma\sqrt{2\pi}}$. Quindi, il parametro μ è la moda, la mediana e il valore medio della v.c. X. Utilizzando le derivate seconde, si dimostra che $f(x)$ presenta due flessi equidistanti dal punto $x = \mu$ in corrispondenza delle ascisse $x = \mu \pm \sigma$. Inoltre, $f(x) - > 0$ per $x - > \pm\infty$, cioè l'asse x è un asintoto orizzontale per tale funzione.

Se si modifica il valore medio (μ) a parità di varianza (σ^2) nella variabile casuale normale, la funzione di densità subisce una traslazione lungo l'asse x.

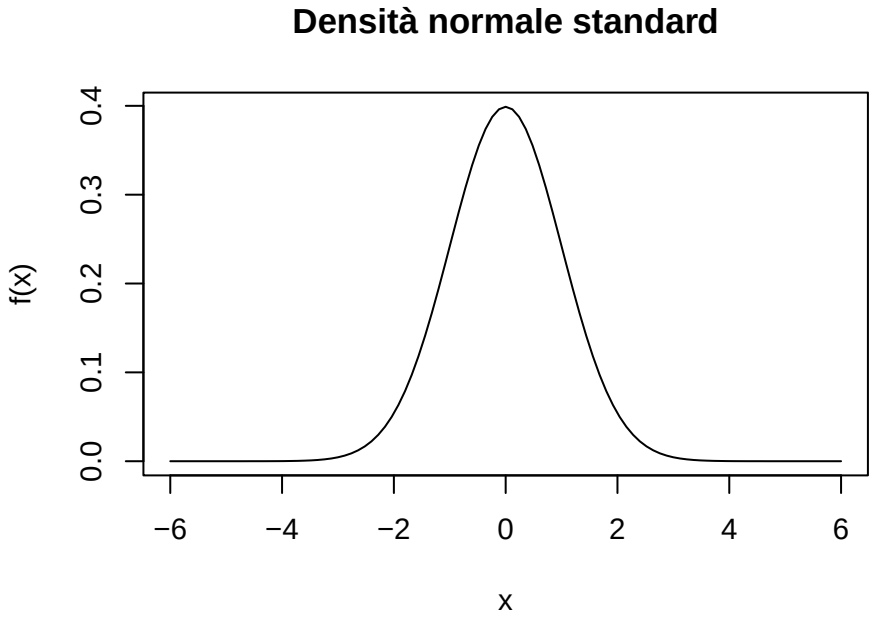
Al contrario, aumentando la varianza (σ^2) a parità di valore medio (μ), i flessi della distribuzione si allontanano dalla media (μ), e di conseguenza, la funzione di densità assegna una maggiore probabilità alla variabile casuale di assumere valori lontani dal valore medio.

Infine, quando la varianza (σ^2) tende a zero, la distribuzione della variabile casuale normale diventa degenere, poiché con probabilità 1 assume valori infinitamente vicini a $x = \mu$.

I momenti caratteristici della v.c. Normale risultano uguali a:

$$E(X) = \mu; Var(X) = \sigma^2; Asym(X) = 0; Kurt(X) = 3$$

Funzione di probabilità La funzione di densità viene così calcolata:



Quartili I quartili sono misure di posizione utilizzate per suddividere una distribuzione di dati in quattro parti uguali, o quarti. La variabile casuale normale ha una distribuzione a forma di campana simmetrica e può essere completamente descritta da due parametri: la media (μ) e la deviazione standard (σ). I quartili di una distribuzione normale sono definiti in modo che ciascun quartile contenga una proporzione specifica dei dati. In particolare:

1. Primo Quartile (Q1): Il **primo quartile** suddivide il 25% inferiore dei dati, quindi il 25% dei dati si trova al di sotto di Q1. Per una distribuzione normale, Q1 corrisponde al valore di -0,6745 deviazioni standard sotto la media.

2. Secondo Quartile (Q2): Il **secondo quartile** è anche noto come mediana ed è il valore che divide il 50% dei dati. Per una distribuzione normale, la mediana coincide con la media, quindi Q2 è uguale a μ che è uguale a 0.

3. Terzo Quartile (Q3): Il **terzo quartile** suddivide il 25% superiore dei dati, quindi il 25% dei dati si trova al di sopra di Q3. Per una distribuzione normale, Q3 corrisponde al valore di +0,6745 deviazioni standard sopra la media.

È importante notare che, per una variabile casuale normale, i quartili sono disposti in modo equidistante lungo la distribuzione. Ciò significa che l'intervallo tra Q1 e Q2 è uguale all'intervallo tra Q2 e Q3. Inoltre, la distribuzione normale è simmetrica rispetto alla sua media, quindi Q1 è equidistante da Q2 e la media μ, così come Q3 è equidistante da Q2 e μ.

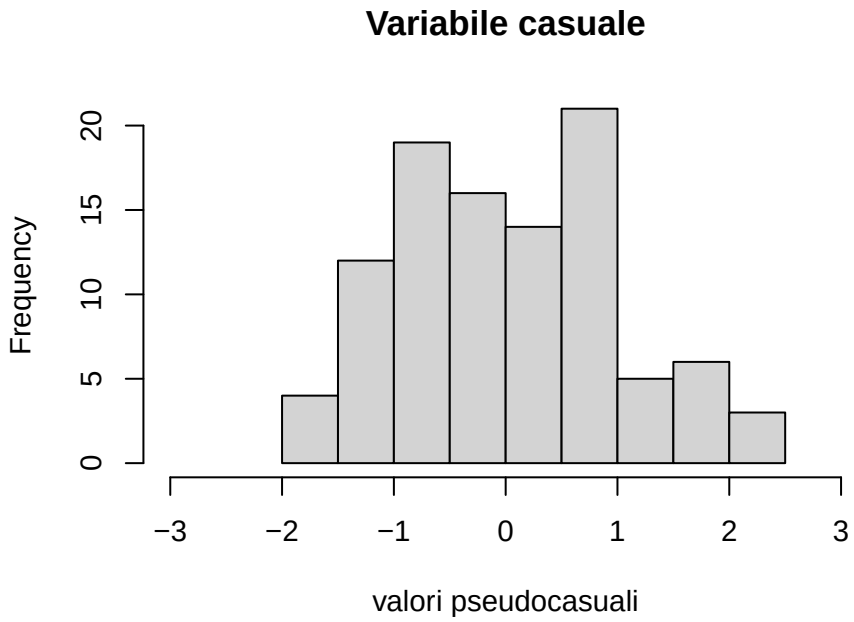
```
[1]       -Inf -0.6744898  0.0000000  0.6744898        Inf
```

Variabile casuale normale

E' quindi possibile generare una sequenza di numeri casuali dalla v.c. Normale fornendo il numero n della lunghezza del campione e

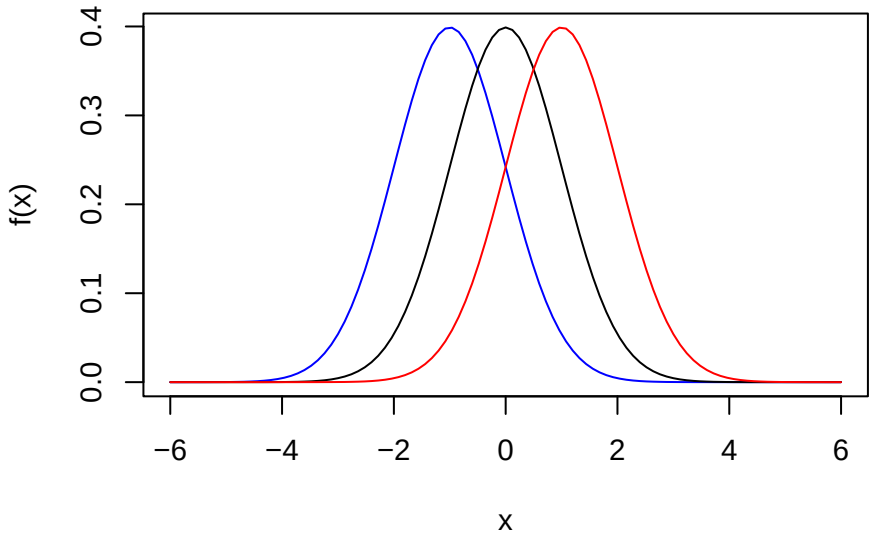
$$\mu, \sigma$$

la media e la deviazione standard della densità normale.



Confronto tra valori medi differenti Quando il parametro μ varia, la curva della distribuzione subisce traslazioni lungo l'asse delle ascisse senza che la sua forma cambi.

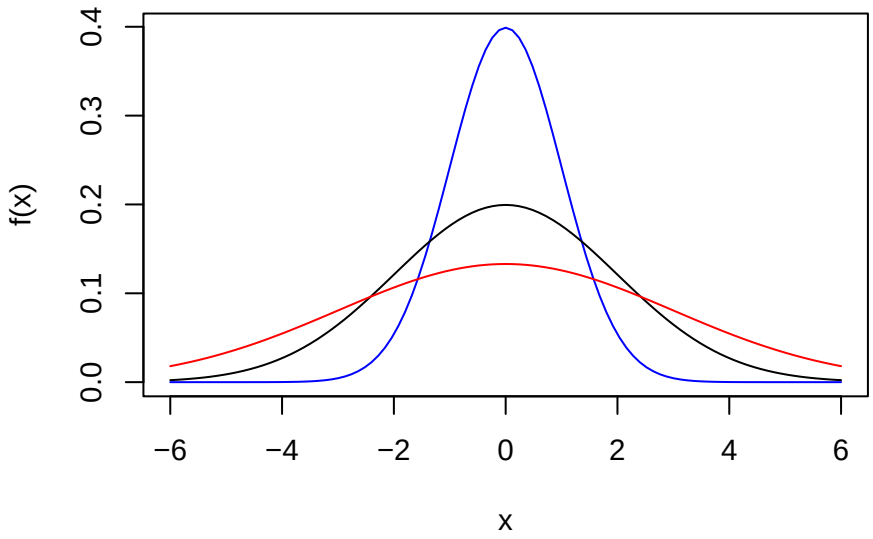
- Traslazioni sulla retta delle ascisse: Aumentando il valore di μ, la curva della distribuzione normale si sposta verso destra sull'asse delle ascisse. Per cui il picco della curva si sposterà nella direzione positiva dell'asse X. Al contrario, diminuendo il valore di μ, la curva si sposterà verso sinistra sull'asse delle ascisse. Questo cambiamento della posizione della curva è noto come traslazione.
- Forma invariata: La forma della curva di una distribuzione normale rimane sempre una campana simmetrica, indipendentemente dalla variazione del parametro μ. La simmetria è mantenuta, e la distribuzione resta identica in forma.
- Centro della distribuzione: Il parametro μ rappresenta il centro o la posizione della distribuzione normale. Aumentando μ, si sposta il centro della distribuzione verso destra, mentre diminuendolo lo sposta verso sinistra.
- Punto di Massima Probabilità: Il valore di μ coincide con il punto di massima probabilità (il picco) della distribuzione normale. Quindi, variando μ, modifichiamo la posizione del valore più probabile della variabile casuale.



Confronto tra deviazioni standard differenti La deviazione standard è cruciale per descrivere quanto i dati nella distribuzione si allontanano dalla media.

- Misura della Dispersione: La deviazione standard, σ, misura quanto i dati nella distribuzione sono dispersi intorno alla media. Un valore di σ maggiore indica una maggiore dispersione, mentre un valore minore indica una minore dispersione.
- Larghezza della Distribuzione: La deviazione standard influenza direttamente la larghezza della distribuzione normale. Maggiore è il valore di σ, più ampia sarà la distribuzione e viceversa. Se σ è piccolo, la distribuzione sarà concentrata intorno alla media, mentre se σ è grande, la distribuzione sarà più "allargata."
- Variabilità: Un valore di σ elevato indica una maggiore variabilità nei dati, il che significa che i punti dati sono più distanti dalla media. Al contrario, un σ basso indica una minore variabilità e punti dati più vicini alla media.
- Scarti Standard: La deviazione standard è utilizzata per calcolare gli scarti standard, che ci permettono di valutare quanto un dato specifico si discosta dalla media. Gli scarti standard sono spesso espressi in unità di σ.

In figura, all'aumentare di σ la curva tende ad essere più piatta mentre al diminuire dello stesso parametro la curva si allungherà verso l'alto restringendosi contemporaneamente ai lati.



Variabile casuale Normale standardizzata

Sia

$$X \sim N(\mu, \sigma^2)$$

,la trasformazione lineare

$$Z = (X - \mu)/\sigma$$

definisce la **variabile casuale Normale Standardizzata**

$$Z \sim N(0, 1)$$

, le cui funzioni di densità e ripartizione sono universalmente indicate con i simboli

$$\phi(z)$$

e

$$\Phi(z)$$

, rispettivamente. $\Phi(z)$:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} d_y$$

La funzione di distribuzione della variabile casuale standard Z non può essere espressa in una forma esplicita. Pertanto, sono state create tavole che forniscono il valore dell'integrale $\phi(z)$ per valori specifici di z con una precisione sufficiente.

REGOLA 68-95-99,7

In statistica, la regola 68-95-99,7, nota anche come regola empirica, è un'abbreviazione utilizzata per ricordare la percentuale di valori che si trovano all'interno di una banda attorno alla media in una distribuzione normale con un'ampiezza di due, quattro e sei deviazioni standard, rispettivamente; più precisamente, il 68,27%, il 95,45% e il 99,73% dei valori si trovano rispettivamente all'interno di una, due e tre deviazioni standard della media. Nella nozione matematica,questo può essere espresso come segue, X è un'osservazione da una variabile casuale con distribuzione normale,

$$\mu$$

è la media della distribuzione e

$$\sigma$$

è la sua deviazione standard.

La regola dei tre sigma più debole può essere derivata dalla disuguaglianza di Chebyshev, affermando che anche per variabili non distribuite normalmente, almeno l'88,8% dei casi dovrebbe rientrare in intervalli di tre σ correttamente calcolati. Per le distribuzioni unimodali, la probabilità di essere all'interno dell'intervallo è almeno del 95% secondo la disuguaglianza di Vysochanskij-Petunin. In altre parole, se si assume che una variabile casuale segua una distribuzione normale, si può aspettare che la maggior parte dei dati si concentri intorno alla media, e che l'ampiezza della distribuzione diminuisca man mano che ci si allontana dalla media. Il valore di sigma rappresenta la dispersione dei dati attorno alla media, e una maggiore dispersione indica una distribuzione più ampia La regola del 3 sigma può essere utile per la valutazione della precisione di una misura o per la rilevazione di eventuali valori anomali nella distribuzione dei dati. Ad esempio, se si osserva un valore che si trova a più di tre sigma dalla media, potrebbe essere considerato come un valore anomalo o "outlier" che richiede ulteriori indagini.

Per una qualsiasi variabile aleatoria normale $X \sim N(\mu, \sigma)$ risulta:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < \frac{X - \mu}{\sigma} < 3) = P(-3 < Z < 3) = 0.99$$

Quindi la probabilità che una variabile aleatoria $X \sim N(\mu, \sigma)$ assuma valori in un intervallo avente come centro μ e semiampiezza 3σ è prossima all'unità. Questa proprietà delle variabili aleatorie normali è nota come regola del 3σ.

Per un ulteriore dimostrazione si sostituiscono i valori dell'espressione con $\mu = 0$ e $\sigma = 1$:

```
[1] 0.9973002
```

3. Stima puntuale

3.1. Campioni casuali e stimatori

Si definisce **popolazione** l'insieme delle informazioni statistiche che esauriscono il problema oggetto dello studio. Popolazione è sinonimo di v.c. X e la conoscenza della popolazione X coinciderà strettamente con la conoscenza della funzione di ripartizione. Dalla popolazione viene , quindi, estratto prescelto o individuato un sottoinsieme di n unità statistiche e la

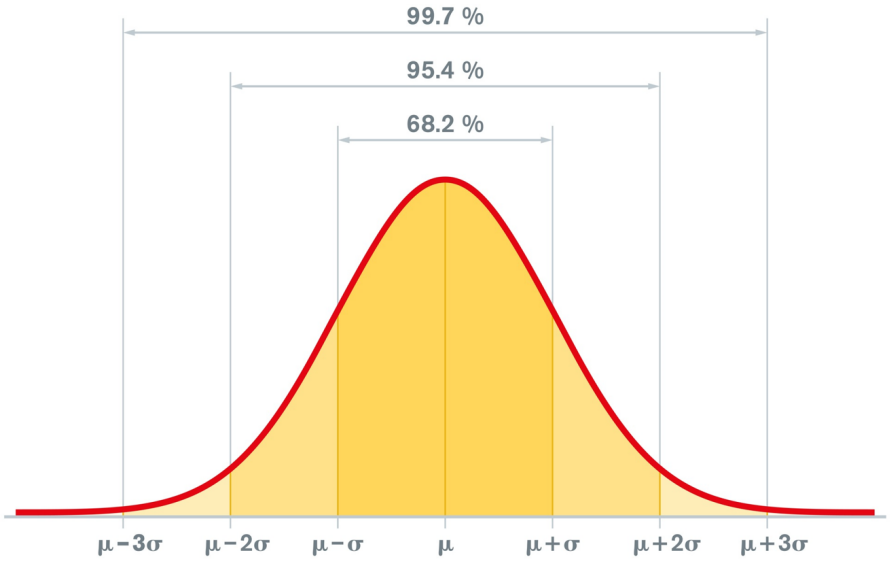


Figure 3: Regola 68-95-99,7

procedura di selezione genera una $n-upla$ di v.c. la cui determinazione numerica specifica una $n-upla$ di numeri reali detti *campione osservato*. Ogni numero reale è una effettiva realizzazione di una i-esima variabile casuale X_i. Una collezione di v.c. $X = (X_1, X_2, ..., X_N)$ ottenuta con un procedimento di estrazione dalla v.c. $X \sim f(x;\theta)$ è un **campione casuale** se le v.c. $X_1, X_2, ..., X_n$ sono mutuamente indipendenti r la distribuzione di probabilità marginale di ciascuna X coincide con la distribuzione di X. Per cui i campioni devono essere **IID** (indipendente e identicamente distribuita). Ciò significa che le osservazioni nel campione sono selezionate in modo casuale e non influenzate da alcuna caratteristica della popolazione madre.
L'obiettivo principale dell'utilizzo di campioni casuali è quello di ottenere informazioni sulla popolazione di riferimento, in modo da poter fare inferenza e prendere decisioni sulla base di queste informazioni.

Attraverso il campione si assegna un valore numerico ai parametri incogniti della popolazione. Per cui la definizione vera e propria di **stimatore** è una statistica ottenuta in funzione del campione casuale che caratterizza la popolazione mentre la stima è il suo corrispondente valore numerico calcolato sulla base del campione osservato.

- La popolazione oggetto di indagine viene assunta ''fissa", sebbene non nota, e tali sono anche la sua media
$$\mu_x = E(x)$$
, la sua varianza
$$\sigma_x^2 = Var(X)$$
, il suo indice di asimmetria, ecc...
  - Quindi, tutti gli indici sintetici relativi alla popolazione sono fissi ma incogniti, (assumendo che esistano, a volte non esistono), vengono indicati in simboli con $\theta_1, \theta_2, \ldots, \mu_x, \sigma_x^2$ e rappresentano i parametri non noti della popolazione.

– Avendo come unica informazione della popolazione il campione, il problema diventa ora individuare delle funzioni del campione (ovvero delle statistiche campionarie!) che riescano a derivare una 'ragionevole approssimazione' dei parametri ignoti.

Data una popolazione $X \sim f(\theta)$, sia {X1, X2, . . . , Xn} un campione casuale. Uno **stimatore** (parametrico puntuale) per θ è una statistica campionaria

$$\hat{\theta} = T(X1, X2, ..., Xn)$$

utilizzata per dedurre l'informazione su θ contenuta nel campione. La stima è il valore osservato dello stimatore, cioè il valore calcolato sui dati osservati,

$$\hat{\theta} = T(X1, X2, ..., Xn)$$

.

Tipiche statistiche sono media campionaria e varianza campionaria:

Media campionaria è:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Varianza campionaria è:

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

3.2. Proprietà desiderabili di uno stimatore

Per la valutazione dell'accettabilità della stima si confrontano e valutano le proprietà degli stimatori. Se il valore stima viene da uno stimatore con proprietà ottimali allora è plausibile, altrimenti no.

Esistono in generale più stimatori, $\hat{\theta}_1$, $\hat{\theta}_2$,. . . , di uno stesso parametro θ. Tra questi, va scelto il 'migliore'. Consideriamo il caso generale di uno stimatore $\hat{\theta}$ usato per inferire il valore θ, dove θ è un parametro che governa la popolazione $X \sim f(x; \theta)$. Sono proprietà desiderabili per lo stimatore:

- sufficienza
- non distorsione (o correttezza)
- efficienza
- consistenza

Proprietà di sufficienza di uno stimatore

Il concetto di sufficienza fu esplicitato da Fisher nel 1920 e si trova in ogni decisione statistica perché concerne l'utilizzo essenziale delle informazioni campionarie riguardanti il parametro θ mediante una sintesi T_n. La sufficienza come proprietà statistica deriva da questa constatazione: **quando da un campione osservato si cerca di produrre una valutazione**

numerica che riguarda il parametro θ si opera una sintesi, cioè una riduzione ad un sottospazio di dimensioni generalmente molto inferiori. Ebbene, tale sintesi dovrà essere ricercata in modo da non disperdere quelle caratteristiche riguardanti θ che sono contenute nel campione, in breve la statistica efficiente dovrà preservare l'essenziale riguardante θ. *Definizione*: Si dice che T_n è sufficiente per θ se la distribuzione condizionata di $(X_1, ..., X_n)$ dato che T_n ha assunto un valore t_0 non dipende da θ.

Se T_n è sufficiente per θ, tutte le informazioni riguardanti θ, che pure esistono nel campione casuale, vengono trasferite nello stimatore T_n . Infatti, una volta osservato un valore t_0, non vi è più alcuna informazione riguardante θ nella distribuzione condizionata del campione casuale.

- È una proprietà difficile da formalizzare
- è il minimo sindacale che possiamo chiedere ad uno stimatore, poiché qualsiasi stimatore non sufficiente sarebbe inutilizzabile

La sufficienza,tuttavia, non permette di pervenire alla scelta di uno stimatore.

Proprietà finite di uno stimatore

Stimatore corretto Uno **stimatore** $\hat{\Theta} = t(X_1, X_2, ..., X_n)$ del parametro non noto θ della popolazione è detto **corretto** (non distorto) se e solo se per ogni $\theta \in \Theta$ si ha:

$$E(\hat{\Theta}) = \theta$$

ossia se il valore medio dello stimatore $\hat{\Theta}$ è uguale al corrispondente parametro non noto della popolazione. La *distorsione* (=bias) di uno stimatore T_n è definita da

$$b(T_n) = E(T_n) - \theta$$

. Si parla allora di distorsione positiva se $E(T_n) > 0$, negativa se

$$E(T_n) < 0$$

. In maniera evidente uno stimatore non distorto presenta distorsione praticamente nulla.

$$b(T_N) = 0 => E(T_n) = \theta$$

. La distorsione è una proprietà desiderabile di uno stimatore percheè, pur non asserendo sulla singola stima, richiede che la procedura inferenziale prescelta per la stima non produca deviazioni rispetto al parametro θ.

Stimatore efficiente Uno stimatore è efficiente se soddisfa la seguente uguaglianza:

$$MSE(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2$$

Se lo stimatore è corretto, l'errore quadratico medio può scriversi come:

$$MSE(\hat{\Theta}) = E[\hat{\Theta} - E(\hat{\Theta})]^2 = Var(\hat{\Theta})$$

, coincidendo con la varianza dello stimatore. L'MSE misura la varianza di $\hat{\theta}$ intorno a θ. Per confrontare l'efficienza di due stimatori , dati due stimatori $\hat{\theta}_1, \hat{\theta}_2$, si dice che $\hat{\theta}_1$ è più efficiente di $\hat{\theta}_2$ se $MSE(\hat{\theta}_1)<MSE(\hat{\theta}_2)$. L'**efficienza relativa** risolve il problema del confronto tra due stimatori ma ciò non esclude che possano esservi altri stimatori la cui variabilità sia minore di quella dei due stimatori considerati. Occorre allora chiedersi se esiste un limite inferiore per la variabilità di uno stimatore di un certo parametro. Chiameremo **stimatore non distorto con varianza uniformemente minima** (Uniformly Minimum Variance Unbiased Estimator (UMVUE)) uno stimatore $\hat{\theta}$ di θ non distorto e che, tra tutti gli stimatori non distorti, possiede la varianza più piccola (quindi è il più efficiente). Quindi significa che: 1. E' non distorto , per cui bias=0 e MSE=varianza 2. Ha varianza minima

Diseguaglianza di Cramer-Rao Sia $\hat{\Theta} = t(X_1, X_2, ..., X_n)$ uno stimatore corretto del parametro non noto θ di una popolazione caratterizzata da funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x;\theta)$. Se sono soddisfatte le seguenti ipotesi:

$$\frac{\partial}{\partial\theta} log f(x;\theta) \; esiste \; per \; ogni \; x \; e \; per \; ogni \; \theta \in \Theta$$

$$E\{[\frac{\partial}{\partial\theta} log f(X;\theta)]^2\} \; esiste \; finito \; per \; ogni \; \theta \in \Theta$$

la varianza dello stimatore $\hat{\Theta}$ soddisfa la disuguaglianza:

$$Var(\hat{\Theta}) \geq \frac{1}{nE\{[\frac{\partial}{\partial\theta} log f(X;\theta)]^2\}}$$

Si noti che la disuguaglianza di Cramér–Rao individua l'estremo inferiore della varianza di uno stimatore corretto, ma non implica che esista sempre uno stimatore con varianza uguale al suo estremo.

Dunque se:

$$Var(\hat{\Theta}) = \frac{1}{nE\{[\frac{\partial}{\partial\theta} log f(X;\theta)]^2\}}$$

allora $\hat{\Theta}$ è uno stimatore corretto con varianza uniformemente minima per il parametro θ.

Stimatore corretto con varianza uniformemente minima per una popolazione normale Si desidera verificare che X sia uno stimatore corretto con varianza uniformemente minima del valore medio $E(X) = \mu$ di una popolazione normale descritta da una variabile aleatoria $X \sim N(\mu, \sigma)$ avente varianza nota σ^2.

La densità di probabilità che caratterizza la popolazione è:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

con $x \in R, \mu \in R, \sigma > 0$

Poiché $E(X) = \mu$, il parametro da stimare è $\theta = \mu$. Osserviamo che:

$$logf(x;\mu) = -log(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}$$

e quindi:

$$\frac{\partial}{\partial\mu} logf(x;\mu) = \frac{x-\mu}{\sigma^2}$$

Essendo $Var(X) = \sigma^2$ risulta:

$$E[\frac{\partial}{\partial\mu} logf(X;\mu)]^2 = E[(\frac{x-\mu}{\sigma^2})^2] = \frac{1}{\sigma^4}E[(X-\mu)^2] = \frac{Var(X)}{\sigma^4} = \frac{1}{\sigma^2}$$

e quindi:
$Var(\overline{X}) = \frac{\sigma^2}{n}$, $\frac{1}{n\frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$

segue quindi che $\overline{X}$ è uno stimatore corretto con varianza uniformemente minima del valore medio μ di una popolazione normale con varianza nota σ^2.

Occorre sottolineare che la media campionaria $\overline{X}$ non è sempre uno stimatore corretto con varianza uniformemente minima del valore medio di una specifica popolazione.

Per campioni di grande ampiezza alcune proprietà asintotica di uno stimatore sono la correttezza asintotica e la consistenza.

Proprietà asintotiche di uno stimatore

E' ragionevole richiedere che le proprietà statistiche di uno stimatore migliorino al crescere della numerosità campionaria e tale aspettativa viene trattata in questa parte con delle proprietà definite asintotiche perché valide quando n->infinito. Questa ragionevolezza non viene solo dal desiderio di rendere il campione rappresentativo per la popolazione ma anche dal fatto che uno stimatore che possiede proprietà di tipo asintotiche utilizza nella direzione giusta ogni nuovo dato disponibile. Per questo, lo studio delle proprietà asintotiche è importante anche quando occorre lavorare su campioni di numerosità finita.

Stimatore asintoticamente corretto

Uno stimatore $\hat{\Theta}_n = t(X_1, X_2, ..., X_n)$ del parametro non noto θ della popolazione è detto **asintoticamente corretto** (asintoticamente non distorto) se e solo se per ogni $\theta \in \Theta$ si ha:

$$\lim_{n->+\infty} E(\hat{\Theta}_n) = \theta$$

ossia se il valore medio dello stimatore Θ_n tende al crescere dell'ampiezza del campione casuale al corrispondente parametro non noto della popolazione.

Quindi, uno stimatore asintoticamente non distorto è uno stimatore eventualmente distorto per n finito, ma la cui distorsione tende a zero al crescere della numerosità campionaria.

Stimatore asintoticamente corretto della varianza per una popolazione normale

Si desidera verificare che:

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

è uno stimatore asintoticamente corretto della varianza σ^2 di una popolazione.
Ricordando che $E(S^2) = \sigma^2$, si ottiene immediatamente:

$$\lim_{n \to \infty} E(\hat{\Theta}_n) = \lim_{n \to \infty} \frac{n-1}{n} E(S^2) = \sigma^2$$

Dunque, per una popolazione normale lo stimatore $(n-1)S^2/n$ della varianza σ^2, individuato sia con il metodo dei momenti che con il metodo della massima verosimiglianza, è asintoticamente corretto.

Proprietà di consistenza di uno stimatore

Stimatore consistente Uno stimatore $\hat{\Theta}_n = t(X_1, X_2, ..., X_n)$ del parametro non noto θ della popolazione è detto **consistente** se e solo se per ogni $\in > 0$ si ha:

$$\lim_{n \to +\infty} P(|\hat{\Theta}_n - \theta| < \in) = 1$$

ossia se e solo se $\hat{\Theta}_n$ converge in probabilità a θ.

Infatti, se si verifica che all'aumentare di n cresce la probabilità che il parametro stimato coincida con quello della popolazione di riferimento, si dice che lo stimatore è consistente (o coerente).

In particolare se:

$$\lim_{n \to \infty} E(\hat{\Theta}_n) = \theta$$

$$\lim_{n \to \infty} Var(\hat{\Theta}_n) = 0$$

Quindi, una condizione sufficiente affinché lo stimatore sia consistente è che sia asintoticamente corretto e la sua varianza tende a zero al crescere del campione; da notare che uno stimatore può essere consistente senza essere asintoticamente corretto.

Un metodo molto utilizzato per la scelta di uno stimatore è invece il **Best Linear Unbiased Estimator (BLUE)**, che consiste nella scelta nella classe degli stimatori lineari corretti di θ di quello che presenta la varianza campionaria minima.

3.3. Stima dei parametri

La stima dei parametri in statistica è un processo che mira a determinare i valori sconosciuti di una popolazione, come la media o la varianza, utilizzando le informazioni tratte da un campione rappresentativo prelevato dalla popolazione. Questa stima può essere fatta in due modi principali: stime puntuali e stime per intervallo.

1. **Stima Puntuale:** La stima puntuale è un metodo in cui calcoliamo un singolo valore numerico per rappresentare la nostra migliore previsione del valore del parametro di interesse. Ad esempio, se vogliamo stimare la media di una popolazione, calcoliamo la media campionaria e la utilizziamo come stima puntuale della media della popolazione. Questo si basa sull'idea che se il campione è rappresentativo, le caratteristiche del campione dovrebbero riflettere quelle della popolazione. Tuttavia, la stima puntuale non fornisce informazioni sulla precisione o l'accuratezza della stima. Non dice quanto il valore stimato sia vicino al vero valore del parametro, e questo può portare ad errori.

2. **Intervallo di Confidenza:** Per superare le limitazioni della stima puntuale, spesso viene utilizzato un intervallo di confidenza. Un intervallo di confidenza è una gamma di valori che indica con quale grado di sicurezza (noto come grado di fiducia) il vero valore del parametro si trovi all'interno di quell'intervallo. In altre parole, fornisce un margine di errore attorno alla stima puntuale. Ad esempio, potremmo dire che siamo al 95% sicuri che il valore del parametro sia compreso tra due limiti specifici calcolati a partire dal campione. L'intervallo di confidenza fornisce una misura della precisione della stima e tiene conto della variabilità nei dati campionari. Un intervallo di confidenza più stretto indica una stima più precisa, mentre un intervallo più ampio indica una stima meno precisa.

Metodi per la costruzione degli stimatori

Ogni metodo di stima dà luogo a una soluzione numerica(=stima), ma per giudicare la qualità statistica di un metodo occorre fare riferimento alla corrispondente v.c. (=stimatore) generata dal campionamento casuale

Metodo dei momenti

Il metodo dei momenti conduce a stimatori naturali per i parametri e richiede due sole condizioni: • esistenza dei momenti delle v.c. in numero uguale a quello dei parametri da stimare; • conoscenza delle relazioni tra momenti e parametri che caratterizzano la popolazione(il parametro deve entrare nella definizione di momento).

Sia $X \sim f(x;\theta)$ una v.c. caratterizzata da un vettore θ di $m >= 1$ parametri che possiede momento $m - esimo$ assoluto: $E(|X|^m) < \infty$ e momenti $E(X^r) = \mu_r$. Si indica con

$$M_(r,n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

con $(r = 1, 2, \dots)$ i momenti campionari generati al campione casuale $(x_1, x_2, ..., x_n)$. Poichè i momenti della popolazione sono funzioni del vettore θ, cioè $\mu_r = \mu_r(\theta)$, il **metodo dei momenti** consiste nel risolvere rispetto a θ il sistema delle prime *mequazioni*:

$$\mu_r(\theta) = M_{r,n} \qquad r = 1, \dots, m$$

ottenendo gli stimatori dei momenti $\hat{\theta}_1, \dots, \hat{\theta}_m$

Ottenendo così per i parametri $(\theta_1, \theta_2, ..., \theta_n)$ i corrispondenti stimatori che dipendono dal campione osservato. In breve consiste nell' eguagliare il momento r-esimo del campione al momento r-esimo della popolazione. Verrà fuori un sistema di m equazioni tante quante sono i parametri da stimare in r incognite. Affinché il metodo dei momenti sia utilizzabile occorre che il sistema ammetta un'unica soluzione. Alcune volte per ottenere tali stimatori è necessario utilizzare un numero maggiore di equazioni rispetto al numero dei parametri non noti da stimare. Una volta calcolati i momenti campionari, si possono utilizzare le proprietà dei momenti per stimare i parametri della distribuzione di una popolazione. Ad esempio, per una distribuzione normale, la media campionaria è un buon stimatore della media della popolazione e la varianza campionaria è un buon stimatore della varianza della popolazione. Il metodo dei momenti è semplice da utilizzare e richiede solo un numero limitato di dati. Tuttavia, gli stimatori ottenuti possono essere meno precisi rispetto ad altri metodi, come il metodo dei minimi quadrati, sopratutto quando la distribuzione dei dati non è nota.

Metodo dei momenti per una popolazione normale Si è interessati a determinare con il metodo dei momenti gli stimatori dei parametri μ e σ^2 di una popolazione normale descritta da una variabile aleatoria $X \sim N(\mu, \sigma)$ di densità di probabilità:

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

con $(x \in R, \mu \in R, \sigma > 0)$.
Occorre quindi stimare due parametri μ e σ^2. Poiché $E(X) = \mu$ e $E(X^2) = \sigma^2 + \mu^2$, si ha:
$\hat{\mu} = \frac{x_1, x_2, ..., x_n}{n}$, $\hat{\sigma^2} + \hat{\mu^2} = \frac{x_1^2 + x_2^2 + ... + x_n^2}{n}$ da cui si ricava:

$$\hat{\sigma^2} = \frac{x_1^2 + x_2^2 + ... + x_n^2}{n} - \frac{(x_1 + x_2 + ... + x_n)^2}{n^2} = \frac{1}{n}[\sum_{i=1}^{n} x_1^2 - \overline{x} \sum_{i=1}^{n} x_i] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{(n-1)s^2}{n}$$

Applicato al caso specifico del dataset ho scelto di analizzare la variabile relativa al guadagno personale la cui ditribuzione potrebbe essere approssimativamente normale, prima di proseguire ho infatti effettuato un test di shapiro wilks per verificare se il campione di dati segue o meno una distribuzione normale. L'obiettivo principale del test è determinare se c'è evidenza sufficiente per rifiutare l'ipotesi nulla che i dati siano distribuiti normalmente.

```
    Shapiro-Wilk normality test

data:  bl_total$‘Personal earnings‘
W = 0.95562, p-value = 0.1364
```

Definito un livello di significatività $alpha = 0.05$, il p-value del test prossimo allo 0.13 permette di non rifiutare l'ipotesi nulla suggerendo che non c'è sufficiente evidenza per rifiutare l'ipotesi nulla che i dati seguano una distribuzione normale.Per questo motivo si può assumere che i dati seguono approssimativamente una distribuzione normale.

```
R> # Media della variabile Personal Earnings -> guadagno personale
R> stimamu = mean(bl_total$‘Personal earnings‘)
R>
R> # Varianza della variabile ‘Personal earnings‘
R> stimasigma2 = (length(bl_total$‘Personal earnings‘) - 1)*var(bl_total$‘Personal earning
```

La stima del parametro μ con il metodo dei momenti è $\mu = 37435.95$ e la stima del parametro σ^2 con il metodo dei momenti è $\sigma^2 = 1979126$.

Gli stimatori derivati con il metodo dei momenti:

- sono **consistenti** (perchè le varianze dei momenti campionari tendono a 0 per n -> ∞)
- sono **asintoticamente non distorti ed asintoticamente Normali**
- non sempre sono efficienti, neppure asintoticamente
- possono essere distorti per numerosità finite
- possono non essere coerenti e, per campioni di dimensione moderata, privi di senso

Metodo della massimaverosimiglianza

Il metodo della verosimiglianza deriva da un principio elementare: «tra i possibili valori del parametro θ, si preferisce quello che corrisponde alla massima probabilità di generare i dati osservati». La ragionevolezza del metodo deriva dal seguente ragionamento: la funzione di verosimiglianza rappresenta la probabilità di osservare, prima dell'esperimento, quel particolare campione che si è effettivamente verificato.

Sia $X_1, X_2, ..., X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\theta_1, \theta_2, ..., \theta_k) = L(\theta_1, \theta_2, ..., \theta_k; x_1, x_2, ..., x_n)$ del campione osservato $(x_1, x_2, ..., x_n)$ è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale $X_1, X_2, ..., X_n$, ossia:

$$L(\theta_1, \theta_2, ..., \theta_k) = L(\theta_1, \theta_2, ..., \theta_k; x1, x2, ..., xn) = f(x_1; \theta_1, \theta_2, ..., \theta_k) * f(x_2; \theta_1, \theta_2, ..., \theta_k) ... f(x_n; \theta_1, \theta_2, ..., \theta_k)$$

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\theta_1, \theta_2, \ldots, \theta_k$. Tale metodo cerca quindi di determinare da quale funzione di probabilità congiunta (nel caso di popolazione discreta) oppure di densità di probabilità congiunta (nel caso di popolazione assolutamente continua) è più verosimile (è più plausibile) che provenga il campione osservato $(x_1, x_2, ..., x_n)$.

Pertanto si cercano di determinare i valori $\theta_1, \theta_2, \ldots, \theta_k$ che rendono massima la funzione di verosimiglianza e che quindi offrano, in un certo senso, la migliore spiegazione del campione osservato $(x_1, x_2, ..., x_n)$.

I valori di $\theta_1, \theta_2, \ldots, \theta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$; essi costituiscono le stime di massima verosimiglianza dei parametri non noti $\theta_1, \theta_2, \ldots, \theta_k$ della popolazione. Tali stime dipendono dal campione osservato $(x_1, x_2, ..., x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori di massima verosimiglianza $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$ dei parametri non noti $\theta_1, \theta_2, \ldots, \theta_k$ della popolazione, detti stimatori di massima verosimiglianza.

Il MLE è un metodo molto utilizzato perché è semplice da utilizzare e perché fornisce stimatori efficaci per molti tipi di distribuzioni. Inoltre, gli stimatori ottenuti con il MLE sono spesso insensibili a piccole perturbazioni nei dati e sono asintoticamente efficienti, cioè tendono ad avere varianze più piccole di qualsiasi altro stimatore per la stessa distribuzione.

Tuttavia, il MLE richiede che si conosca la forma esatta della distribuzione dei dati, e può essere sensibile a anomalie nei dati come outlier. Inoltre, gli stimatori ottenuti con il MLE possono essere meno precisi per campioni di dimensioni ridotte.

Come per qualsiasi metodo di stima, è importante valutare sempre la validità e l'accuratezza degli stimatori ottenuti attraverso il metodo della massima verosimiglianza confrontandoli con altri metodi o con i dati di riferimento.

Metodo della massimaverosimiglianza per una popolazione normale

Si desidera determinare lo stimatore di massima verosimiglianza dei parametri μ e σ^2 di una popolazione normale caratterizzata da funzione densità di probabilità:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

con $(x \in R, \mu \in R, \sigma > 0)$, si ha:

$$L(\mu, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

con $(\mu \in R, \sigma > 0)$
dove le $x_i \in R$. Si nota che:

$$logL(\mu, \sigma^2) = -\frac{n}{2} log\sigma^2 - \frac{n}{2} log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

con $(\mu \in R, \sigma > 0)$ e quindi si ha:
$\frac{\partial logL(\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = \frac{n}{\sigma^2} (\frac{1}{n} \sum_{i=1}^{n} x_i - \mu)$
$\frac{\partial logL(\mu,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^4} (\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2)$
Lo stimatore di massima verosimiglianza e dei momenti del valore medio μ è la media campionaria $\overline{X}$. Invece lo stimatore di massima verosimiglianza e dei momenti della varianza σ^2 è $(n-1)S^2/n$.

4. Intervalli di confidenza

Nella teoria della stima si è visto come, data una popolazione $X \sim f(x;\theta)$, con θ incognito, sia possibile costruire una stima per θ. In particolare se X_1, X_2, X_N è un campione casuale estratto da X, uno stimatore per θ (parametrico puntuale: parametrico perché si conosce la variabile casuale che dipende da teta; puntuale: quando si costruisce uno stimatore si ha un solo valore numerico) è una statistica campionaria $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$ utilizzata per dedurre l'informazione su θ contenuta nel campione. La stima è il valore osservato dello stimatore, cioè il valore calcolato sui dati osservati. I metodi di stima puntuali anche se

corredati di tutte le proprietà giudicate desiderabili e ottimali, difficilmente potranno fornire delle stime che coincidono con il parametro incognito, poiché ci si dovrà sempre attendere un **certo errore di campionamento**. Nasce , quindi, l'esigenza di associare allo stimatore una misura dell'errore di stima commesso, in modo tale da **valutare quanto la stima sia da considerarsi 'vicina' al parametro incognito**. Definito il **grado di plausibilità* si potrà dividere lo spazio parametrico in due sottoinsiemi: uno di valori 'possibili' per θ secondo il grado di plausibilità fissato e un altro di valori poco 'possibili' per θ. (POSSIBILI e non PROBABILI, perché θ è un parametro e non una v.c.) Così invece di stimare un unico valore per θ, si stimerà un insieme di valori possibili a cui verrà associato il grado di plausibilità scelto il quale deve essere interpretato come livello di confidenza per l'insieme. Sia $X_1, X_2, ..., X_n$ un campione casuale di ampiezza n estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso continuo), $f(x;\theta)$, dove θ denota il parametro non noto della popolazione.

Denotiamo con $Cn = g_1(X_1, X_2, ..., X_n)$ e con $\overline{C}n = g_2(X_1, X_2, ..., X_n)$ due statistiche (funzioni osservabili del campione casuale) che soddisfino la condizione $Cn < \overline{C}n$, cioè che godono della proprietà che per ogni possibile fissato campione osservato $x = (x_1, x_2, ..., x_n)$ risulti $g_1(x) < g_2(x)$.

Fissato un coefficiente di confidenza $1 - \alpha \quad (0 < \alpha < 1)$, è utile scegliere le statistiche Cn e $\overline{C}n$ in modo tale che

$$P(Cn < \theta < \overline{C}n) = 1 - \alpha$$

allora si dice che $(Cn, \overline{C}n)$ è un intervallo di confidenza (intervallo di fiducia) di grado $1 - \alpha$ per θ. Inoltre, le statistiche Cn e $\overline{C}n$ sono dette limite inferiore e superiore dell'intervallo di confidenza.

Se $g_1(x)$ e $g_2(x)$ sono i valori assunti dalle statistiche Cn e $\overline{C}n$ per il campione osservato $x = (x_1, x_2, ..., x_n)$, allora l'intervallo $(g_1(x), g_2(x))$ è detto stima dell'intervallo di confidenza di grado $1 - \alpha$ per θ ed i punti finali $g_1(x)$ e $g_2(x)$ di tale intervallo sono detti rispettivamente stima del limite inferiore e stima del limite superiore dell'intervallo di confidenza.

4.1. Caso studio: Popolazione normale

Consideriamo una variabile casuale X con distribuzione normale $N(\mu, \sigma)$, dove μ rappresenta il valore medio e σ la deviazione standard.

Ecco alcuni problemi che possono essere esaminati:

1. **Intervallo di Confidenza per μ con Varianza Conosciuta:** Determinare un intervallo di confidenza di livello $1 - \alpha$ per il valore medio μ quando la varianza σ^2 della popolazione normale è nota.

2. **Intervallo di Confidenza per μ con Varianza Sconosciuta:** Determinare un intervallo di confidenza di livello $1 - \alpha$ per il valore medio μ quando la varianza della popolazione normale è sconosciuta.

3. **Intervallo di Confidenza per σ^2 con μ Conosciuto:** Determinare un intervallo di confidenza di livello $1 - \alpha$ per la varianza σ^2 quando il valore medio μ della popolazione normale è noto.

4. **Intervallo di Confidenza per σ^2 con μ Sconosciuto:** Determinare un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 quando il valore medio della popolazione normale è sconosciuto.

Intervallo di confidenza per media con varianza nota

Si considera la statistica test che si distribuisce come una normale standard, in particola modo si sceglie questa quantità in quanto rappresenta una **quantità pivotale**, perchè:

- è funzione del campione casuale e del parametro incognito;
- ha distribuzione indipendente dal parametro incognito.

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

Fissato $\alpha_1 = -z_{\alpha/2}$ e $\alpha_2 = z_{\alpha/2}$ si ha:

$$P(-z_{\alpha/2} < Z_n < z_{\alpha/2}) = 1 - \alpha.$$

Effettuando le sostituzioni si ottiene:

$$P(X_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < X_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha.$$

Ponendo $C_n = X_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ e $\overline{C}_n = X_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ si ha una stima dell'intervallo di confidenza di grado $1-\alpha$ per il valore medio μ :

$$C_n < \mu < \overline{C}_n$$

con $C_n = \overline{C}_n = \overline{x}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ dove $\overline{x}_n$ è la media campionaria del campione. Dopo aver osservato il campione è quindi possibile determinare l'intervallo di confidenza per la media.

Nell'esempio seguente vogliamo determinare l'guadagno medio mondiale di vita nel dataset *bl_total*

```
[1] 37435.95
```

Calcoliamo l'intervallo di confidenza di grado $1-\alpha = 0.95$ nel quale dovrebbe essere compreso tale valore, supponendo che la popolazione da cui proviene il campione sia normale con deviazione standard nota $\sigma = 14256.99$ *dollari*:

```
[1] 32902.96
```

```
[1] 41968.93
```

La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ è (32902.96, 41968.93), notando che la media campionaria è compresa in tale intervallo.

Intervallo di confidenza per media con varianza non nota

Si considera la statistica test che si distribuisce come una T-student con n-1 gradi di libertà:

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim T(n-1)$$

dove la varianza campionaria è:

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

Fissando $\alpha_1 = -t_{\alpha/2,n-1}$ e $\alpha_2 = t_{\alpha/2,n-1}$ si ha:

$$P(-t_{\alpha/2,n-1} < T_n < t_{\alpha/2,n-1})$$

Possiamo dunque dedurre che la stima dell'intervallo di confidenza di grado $1-\alpha$ per il valore medio μ è:

$$\overline{x}_n - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} < \mu < \overline{x}_n + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}$$

dove $\overline{x}_n$ è la media campionaria del campione e s_n è la deviazione standard. Dopo aver osservato il campione è quindi possibile costruire l'intervallo di confiedenza per la media con la varianza incognita.

Nell'esempio seguente vogliamo determinare il guadagno medio personale nel dataset *bl_total*

Calcoliamo l'intervallo di confidenza di grado $1 - alpha = 0.90$ nel quale dovrebbe essere compreso tale valore, supponendo che la popolazione da cui proviene il campione sia normale:

```
[1] 33534.06
```

```
[1] 41337.84
```

La stima dell'intervallo di confidenza di grado $1-\alpha = 0.90$ è (33534.06, 41337.84), chiaramente la media campionaria è compresa in essa.

Intervallo di confidenza per varianza con media nota

Si considera la variabile aleatoria che si distribuisce come una chi-quadrato con n-1 gradi di libertà:

$$V_n = \sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma})^2 \sim X_{n-1}^2$$

Fissando $\alpha_1 = \chi^2_{1-\alpha/2,n}$ e $\alpha_2 = \chi^2_{\alpha/2,n}$ si ha:

$$P(\chi^2_{1-\alpha/2,n} < V_n < \chi^2_{\alpha/2,n}) = 1 - \alpha$$

Pertanto possiamo affermare che la stima dell'intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 è:

$$\frac{(n-1)s_n^2 + n(\overline{x}_n - \mu)^2}{\chi^2_{\alpha/2,n}} < \sigma^2 < \frac{(n-1)s_n^2 + n(\overline{x}_n - \mu)^2}{\chi^2_{1-\alpha/2,n}}$$

dove $\overline{x}_n$ è la media campionaria e s_n^2 è la varianza campionaria.

Nell'esempio seguente vogliamo stimare l'intervallo di confidenza per l'guadagno personale nel dataset *bl_total*

Supponiamo che sia distribuito normalmente con valore medio $\mu = 79$ e varianza non nota σ^2, determiniamo una stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ per la varianza σ^2.

```
[1] 132184062
```

```
[1] 328722897
```

La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza della popolazione normale è quindi (132184062, 328722897). I risultati sono molto elevati ma comprensibili visto che la variabile è relativa alla percezione di guadagno annuale medio in migliaia di dollari.

Intervallo di confidenza per varianza con media non nota

Consideriamo la statistica test Q_n che si distribuisce come una chi-quadrato con n-1 gradi di libertà:

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

Fissando $\alpha_1 = \chi^2_{1-\alpha/2,n-1}$ e $\alpha_2 = \chi^2_{\alpha/2,n-1}$ si ha:

$$P(\chi^2_{1-\alpha/2,n-1} < Q_n < \chi^2_{\alpha/2,n-1}) = 1 - \alpha$$

Osservando le precedenti disuguaglianze possiamo affermare che la stima dell'intervallo di confidenza di grado $1 - \alpha$ per la varianza σ^2 è:

$$\frac{(n-1)s_n^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s_n^2}{\chi^2_{1-\alpha/2,n-1}}$$

dove s_n^2 è la varianza campionaria.

Nell'esempio seguente vogliamo stimare l'intervallo di confidenza per l'guadagno personale nel dataset *bl_total:*

Determiniamo una stima dell'intervallo di confidenza di grado $1-alpha = 0.95$ per la varianza di una popolazione supposta normale:

```
[1] 135098883
```

```
[1] 340215681
```

La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per la varianza della popolazione normale è (135098883, 340215681).

5. Test delle ipotesi

Il *test statistico* è una decisione operativa presa sulla base di risultati sperimentali, tenendo conto di considerazioni probabilistiche. La problematica del test può essere suddivisa in tre fasi:

- Formulare ipotesi sulla v.c. X;
- Osservare il campione casuale;
- In base ai risalutati campionari decidere se accettare o rifiutare l'ipotesi fatta.

IPOTESI STATISTICA: Un'ipotesi statistica è una affermazione sulla distribuzione di una o più variabili casuali. In particolare, si può derivare un test delle ipotesi:

- Supponendo nota la distribuzione di probabilità di X per cui l'inferenza si riferisce ai soli parametri che la specificano (test parametrico);
- Oppure senza fare assunzioni circa la firma analitica della distribuzione di probabilità di X per cui l'inferenza riguarda sia la forma della distribuzione che i suoi parametri.

Si definisce ipotesi statistica **semplice** una ipotesi statistica che specifica completamente la distribuzione della v.c.; in caso contrario viene chiamata ipotesi statistica **composta**. Le ipotesi vengono indicate con la lettera H. Esempio: Data una v.c. $X \sim N(\mu, 9)$ l'ipotesi H : $\mu = 15$ è semplice perché specifica completamente la distribuzione della v.c. X; l'ipotesi H : $\mu > 15$ è composta.

Il dilemma della verifica delle ipotesi coinvolge la creazione di un test ψ che suddivide l'insieme di possibili campioni, rappresentati dalle n-tuple $(x_1, x_2, \ldots, x_n)$ generati dal vettore aleatorio $X_1, X_2, \ldots, X_n$, in due sottoinsiemi distinti: una regione di accettazione A per l'ipotesi nulla e una regione di rifiuto R per l'ipotesi nulla. Il test ψ stabilisce che l'ipotesi nulla è accettata come valida se il campione osservato $(x_1, x_2, \ldots, x_n) \in A$ e viene rifiutata se $(x_1, x_2, \ldots, x_n) \in R$. Nel caso in cui l'ipotesi nulla sia erronea, l'ipotesi alternativa risulta vera e viceversa.

Da queste premesse derivano due possibili errori:

1. **Errore di Tipo I:** Rifiutare l'ipotesi nulla H_0 quando questa è effettivamente vera. Questo errore è indicato con la probabilità $\alpha(\theta) = P(\text{rifiutare } H_0|\theta),\ \theta \in \Theta_0$.
2. **Errore di Tipo II:** Accettare l'ipotesi nulla H_0 quando questa è falsa. Questo errore è indicato con la probabilità $\beta(\theta) = P(\text{accettare } H_0|\theta),\ \theta \in \Theta_1$.

Nella seguente tabella sono riassunti tutti i possibili casi:

	Rifiutare H_0	Accettare H_0
H_0 vera	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
H_0 falsa	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

Nel processo di creazione del test, è vantaggioso stabilire in anticipo la probabilità di commettere un errore di tipo I e successivamente cercare un test ψ che minimizzi la probabilità di commettere un errore di tipo II. *La ragione alla base della scelta di fissare una probabilità di errore di tipo I relativamente bassa è motivata dal fatto che spesso le ipotesi sono formulate in modo tale che l'errore di tipo I risulti più critico, e di conseguenza, il decision maker desidera limitare al massimo la probabilità di commettere tale errore.*

Solitamente la probabilità di commettere un errore di tipo I si sceglie uguale a 0.05, 0.01, 0.001 ed il test viene rispettivamente detto statisticamente significativo, statisticamente molto significativo e statisticamente estremamente significativo.

Sia θ_0 un sottoinsieme dello spazio parametrico θ (l'insieme dei possibili valori). Si vuole verificare l'ipotesi nulla :

$$H_0 : \theta \in \theta_0$$

contro l'ipotesi alternativa:

$$H_1 : \theta \in \theta_1$$

Dove :

$$\theta_0 \bigcup \theta_1 = \theta e \theta_0 \bigcap \theta_1 = \bigotimes$$

L'ipotesi statistica riguardante la v.c. X e, quindi il parametro θ, implica una bipartizione dello spazio parametrico θ in due regioni: θ_0 e θ_1 di cui una rappresenta l'ipotesi nulla H_0 e l'altra l'ipotesi alternativa H_1.

Se H_0 è un'ipotesi statistica semplice allora θ_0 consiste in un solo punto θ_0 il quale determina completamente la distribuzione di X. Se l'ipotesi statistica composta per il parametro θ include valori reali in una sola direzione ($\theta > \theta_0$) l'ipotesi si dirà **unidirezionale, altrimenti bidirezionale**.

Il test bilaterale è il seguente:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

mentre il test unilaterale sinistro è:

$$H_0 : \theta \leq \theta_0$$

$$H_1 : \theta > \theta_0$$

il test unilaterale destro è:

$$H_0 : \theta \geq \theta_0$$

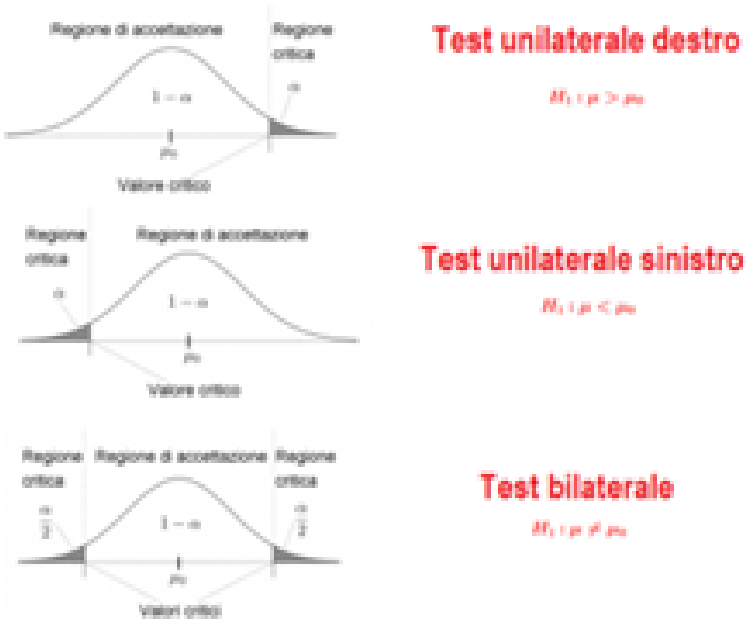


Figure 4: Test delle ipotesi

$$H_1 : \theta < \theta_0$$

avendo fissato a priori un livello di significatività α.

Criterio del p-value

Il *p-value* rappresenta la probabilità di ottenere risultati altrettanto o meno compatibili con quelli osservati durante il test, considerando un'ipotesi presumibilmente vera. In altre parole, fornisce una stima di quanto sia probabile osservare dati simili a quelli del campione in analisi quando l'ipotesi nulla è valida. Questa probabilità, essendo compresa tra 0 e 1, assume valori numerici che indicano quanto i dati siano in accordo con l'ipotesi in questione.

L'utilità del p-value emerge nel processo di decisione riguardo all'accettazione o al rifiuto dell'ipotesi nulla. Nello specifico, si compara il *p-value* con il livello di significatività α, e in base al confronto, si prende una decisione:

- Se $p \leq \alpha$, si opta per il rifiuto dell'ipotesi nulla.
- Se $p > \alpha$, si decide di non rifiutare l'ipotesi nulla.

Ad esempio, se $\alpha = 0.05$, allora $p - value \geq 0,05$ implica che il test non è statisticamente significativo (cioè, può trattarsi di un effetto casuale del campionamento) e l'ipotesi è accettata, mentre *p-value < 0,05* implica che l'ipotesi è rifiutata ed il test è, più in particolare:

- statisticamente significativo se $0.01 \leq p - value < 0.05$;

- molto significativo se $0.01 \leq p - value < 0.01$;
- estremamente significativo se $p - value < 0.001$.

In generale, più è basso il p-value, maggiore è la significatività statistica della differenza osservata.

5.1. Test delle ipotesi su Popolazione normale

(i) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota;

(ii) Verifica di ipotesi sul valore medio μ nel caso in cui la varianza della popolazione normale è non nota;

(iii) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;

(iv) Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio della popolazione normale è non noto.

Test su media con varianza nota

Test bilaterale

Sia $X_1, X_2, ..., X_n$ un campione casuale estratto da una popolazione normale descritta da una variabile aleatoria $X \sim N(\mu, \sigma)$ con varianza nota σ^2.

Si considerino le ipotesi:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

La statistica test è:

$$Z_n = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

Pertanto il test bilaterale ψ di misura α è il seguente:

- Si accetti H_0 se: $-z_{\alpha/2} < \frac{\overline{X}_n - \mu_0}{\sigma\sqrt{n}} < z_{\alpha/2}$
- Si rifiuti H_0 se: $\frac{\overline{X}_n - \mu_0}{\sigma\sqrt{n}} < -z_{\alpha/2}$ oppure $\frac{\overline{X}_n - \mu_0}{\sigma\sqrt{n}} > z_{\alpha/2}$

Denotando con $z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}$ la stima osservata della statistica test possiamo calcolare il p-value per il test bilaterale considerato:

$$pvalue = P(Z_n < -|z_{os}|) + P(Z_n > |z_{os}|) = 2P(Z_n > |z_{os}|) = 2[1 - P(Z_n \leq |z_{os}|)]$$

dall'uguaglianza osserviamo che il *p-value* corrisponde alla probabilità, supposta vera l'ipotesi nulla, che la statistica del test Z_n assuma un valore uguale o più estremo di quello effettivamente osservato z_{os}.

Si propone un esempio:

Un gruppo di persone discute sul guadagno medio mondiale (tenendo conto dei paesi inclusi). Alcuni sostengono che il guadagno medio mondiale sia 40000 *dollari* altri sostengono che sia di \$37435.95 dollari \$. Pertanto, si costruisce un test delle ipotesi con misura $\alpha = 0.05$ per verificare l'ipotesi nulla $H_0 : \mu = 37435.95$ contro l'ipotesi alternativa $H_1 : \mu \neq 37435.95$

```
[1] 1.959964
```

```
[1] 1.10864
```

```
[1] 0.2675855
```

Si nota che $z_{\alpha/2} = 1.959964$ e $z_{os} = 1.10$ cade all'interno della regione di accettazione; occorre quindi accettare l'ipotesi nulla con un livello di significatività del 5%. Si nota anche che $pvalue > \alpha$ e quindi anche il criterio del p-value consiglia di accettare l'ipotesi nulla.

Test su media con varianza non nota

Test bilaterale

Sia $X_1, X_2, ..., X_n$ un campione casuale estratto da una popolazione normale con varianza non nota σ^2.

Si considerino le ipotesi:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

La statistica test è:

$$T_n = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}$$

Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente:

- si accetti H_0 se: $-t_{\alpha/2,n-1} < \frac{\overline{x}_n - \mu_0}{s_n\sqrt{n}} < t_{\alpha/2,n-1}$
- si rifiuti H_0 se: $\frac{\overline{x}_n - \mu_0}{s_n\sqrt{n}} < -t_{\alpha/2,n-1}$ oppure $\frac{\overline{x}_n - \mu_0}{s_n\sqrt{n}} > t_{\alpha/2,n-1}$

Denotando la stima della statistica test con $t_{os} = \frac{\overline{x}_n - \mu_0}{s_n\sqrt{n}}$ calcoliamo ora il p-value per il test bilaterale considerato:

$$pvalue = P(T_n < -|t_{os}|) + P(T_n > |t_{os}|) = 2P(T_n > |t_{os}|) = 2[1 - P(T_n \leq |t_{os}|)]$$

Facciamo un esempio di quanto descritto:

Un gruppo di persone discute sul guadagno medio personale mondiale (tenendo conto dei paesi inclusi). Alcuni sostengono che questo sia prossimo in media a $20 miladollari$ altri sostengono che sia di $37 miladollari$. Pertanto, si costruisce un test delle ipotesi con misura $\alpha = 0.01$ per verificare l'ipotesi nulla $H_0 : \mu = 20mila$ contro l'ipotesi alternativa $H_1 : \mu \neq 20mila$

```
[1] 2.715409
```

```
[1] 5.327813
```

```
[1] 5.103831e-06
```

Si nota che $z_{\alpha/2} = 2.71$ e $z_{os} = 5.32$ cade al di fuori della regione di accettazione; occorre quindi rifiutare l'ipotesi nulla con un livello di significatività del 1%. Si nota anche che $pvalue < \alpha$ e quindi anche il criterio del p-value consiglia di rifiutare l'ipotesi nulla.

Test su varianza con valore medio noto

Test bilaterale

Sia $X_1, X_2, ..., X_n$ un campione casuale estratto da una popolazione normale con valore medio noto μ.

Si considerino le ipotesi:

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

La statistica test è:

$$V_n = \sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma_0})^2 = \frac{(n-1)S_n^2}{\sigma_0^2} + (\frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}})^2$$

Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente:

- si accetti H_0 se: $\chi^2_{1-\alpha/2,n} < \sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma_0}) < \chi^2_{\alpha/2,n}$
- si rifiuti H_0 se: $\sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma_0}) < \chi^2_{\alpha/2,n}$ oppure $\sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma_0}) > \chi^2_{\alpha/2,n}$

Facciamo un esempio di quanto descritto:

Alcuni sostengono che la varianza del guadagno medio mondiale sia 203261649 altri sostengono che sia di 303261649. Pertanto, si costruisce un test delle ipotesi con misura $\alpha = 0.05$ per verificare l'ipotesi nulla $H_0 : \sigma^2 = 203261649$ contro l'ipotesi alternativa $H_1 : \sigma^2 \neq 203261649$.

```
[1] 22.87848
```

```
[1] 56.89552
```

```
[1] 55.20314
```

Si nota che $\chi^2_{\alpha/2} = 22$, $\chi^2_{1-\alpha/2} = 56$. Poichè il valore osservato $\chi^2 = 55.20$ è compreso nella regione di accettazione, si accetta l'ipotesi nulla con un livello di significatività del 5%.

Test su varianza con valore medio non noto

Test bilaterale

Sia $X_1, X_2, ..., X_n$ un campione casuale estratto da una popolazione normale con valore medio noto μ.

Si considerino le ipotesi:

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

La statistica test è:

$$Q_n = \frac{(n-1)S_n^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- si accetti H_0 se: $\chi^2_{1-\alpha/2,n-1} < (\frac{(n-1)s_n^2}{\sigma_0^2}) < \chi^2_{\alpha/2,n-1}$
- si rifiuti H_0 se: $\frac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{\alpha/2,n-1}$ oppure $\frac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{\alpha/2,n-1}$

Facciamo un esempio di quanto descritto:

Alcune persone sostengono che la varianza dello stipendio percepito sia di \$203261649 \$ altri invece sostengono che sia \$300000000 \$, pertanto costruiamo un test delle ipotesi di misura $\alpha = 0.05$ per verificare l'ipotesi nulla $H_0 : \sigma^2 = 300000000$ contro l'ipotesi alternativa $H_1 : \sigma^2 \neq 300000000$:

```
[1] 22.10563
```

```
[1] 55.66797
```

```
[1] 20.82933
```

Si nota che $\chi^2_{\alpha/2} = 22$, $\chi^2_{1-\alpha/2} = 55.66$. Poichè il valore osservato $\chi^2 = 20.82$ è compreso nella regione di rifiuto, si rifiuta l'ipotesi nulla con un livello di significatività del 5%.

5.2. Criterio del chi-quadrato

Test chi-quadrato bilaterale

In molte situazioni pratiche, è comune voler verificare se un campione osservato provenga da una popolazione con una specifica variabile aleatoria X e la corrispondente funzione di distribuzione $F_X(x)$.

Il criterio chi-quadrato utilizza la seguente statistica:

$$Q = \sum_{i=1}^{r} \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)^2$$

Per campioni sufficientemente grandi di dimensione n, il test chi-quadrato bilaterale, con livello di significatività α, si esplicita come segue:

- Si accetta l'ipotesi H_0 se:

$$\chi^2_{1-\frac{\alpha}{2},r-k-1} < \chi^2 < \chi^2_{\frac{\alpha}{2},r-k-1}$$

- Si rifiuta l'ipotesi H_0 se:

$$\chi^2 < \chi^2_{1-\frac{\alpha}{2},r-k-1}$$

o

$$\chi^2 > \chi^2_{\frac{\alpha}{2},r-k-1}$$

Dove $\chi^2_{\frac{\alpha}{2},r-k-1}$ e $\chi^2_{1-\frac{\alpha}{2},r-k-1}$ sono le soluzioni delle seguenti equazioni:

$$P(Q < \chi^2_{1-\frac{\alpha}{2},r-k-1}) = \frac{\alpha}{2} \quad \text{e} \quad P(Q < \chi^2_{\frac{\alpha}{2},r-k-1}) = 1 - \frac{\alpha}{2}$$

Popolazione normale

Nonostante sia stato effettuato il test di shapiro-wilks precedentemente, si è interessati a verficare ulteriormente la normalità del guadagno medio mondiale.

Attraverso l'impiego del test chi-quadrato con un livello di significatività $\alpha = 0.05$, l'obiettivo è esaminare se la popolazione di provenienza del campione possa essere adeguatamente descritta da una variabile aleatoria X con una densità normale.

Immaginiamo di suddividere l'insieme dei valori potenziali di questa variabile aleatoria normale X in $r = 5$ sottoinsiemi. Attraverso l'utilizzo dei quantili della distribuzione normale, possiamo delineare tali sottoinsiemi:

```
[1] 25436.97 33823.98 41047.91 49434.93
```

Occorre ora determinare il numero di elementi del campione che cadono negli intervalli:

```
[1] 12  3  6  7 10
```

Possiamo dedurre che: $n_1 = 12,\ n_2 = 3,\ n_3 = 6,\ n_4 = 7,\ n_5 = 10$.
Calcoliamo adesso la statistica test χ^2:

```
[1] 6.473684
```

$\chi^2 = 6.47$.
La distribuzione normale ha due parametri non noti (μ, σ^2) e quindi $k = 2$. Pertanto, la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con $r - k - 1 = 2$ gradi di libertà. Occorre quindi calcolare $\chi^2_{\alpha/2,2}$ e $\chi^2_{1-\alpha/2,2}$ con $\alpha = 0.05$.

```
[1] 0.05063562
```

```
[1] 7.377759
```

da cui segue che $\chi^2_{1-\alpha/2,r-k-1} = 0.05$ e $\chi^2_{\alpha/2,r-k-1} = 7.37$, dato che $\chi^2 = 6.46$ è evidente che l'ipotesi nulla di popolazione normale può essere accettata.

Affiliation:
Carmela Pia Senatore
Università degli studi di Salerno
Matricola: 0522501721
E-mail: c.senatore50@studenti.unisa.it