PREMESSA

Descrizione parametrica di una popolazione

Sappiamo che un famiglia parametrica di funzioni densità di probabilità è definita da uno o più parametri $\Theta = \{\theta_1, \theta_2,\ldots, \theta_n\}$. Ad esempio, la d.d.p. di tipo esponenziale è completamente caratterizzata dal parametro θ e si interpreta come funzione di x con parametro θ fissato.

$$f(x;\theta) = \theta \exp(-\theta x)$$

La distribuzione normale è completamente specificata dai parametri $\theta_1 = \eta$ e $\sigma_2 = \sigma^2$:

$$f(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right]$$

Nei due casi la funzione x rappresenta la variabile misurabile e viene scelta sulla base dell'esperimento (es. pressione, frequenza cardiaca, ecc.). Il parametro Θ riporta informazioni sulla posizione esull'andamento della legge di distribuzione della variabile misurabile.

Come abbiamo già messo in evidenza nel capitolo relativo alle variabili aleatorie e ai rispettivi momenti, conoscere i parametri della d.d.p. significa essere in grado di caratterizzare una popolazione di dati in modo esatto.

Esempio: $\theta_1 = \eta$ valore medio della popolazione

$\theta_2 = \sigma$ deviazione standard della popolazione

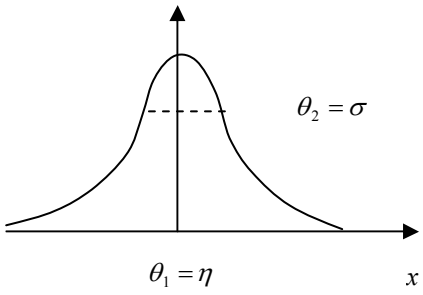


Fig. 1

Definizione e proprietà degli stimatori

In genere la legge si conosce o si suppone nota, mentre i parametri non sono noti e la loro stima è fatta sulla base dei risultati di un campione di *n* osservazioni di una variabile aleatoria misurabile $x = \{x_1, x_2,\ldots, x_n\}$.

Pertanto la ricerca della migliore stima del parametro incognito si può formulare come la ricerca della migliore funzione g derivata dal vettore osservazione x che sia la migliore stima di θ, che chiameremo $\hat{\theta}$. In particolare:

$$\hat{\theta} = g(x) = \{x_0, x_1,\ldots, x_n\}$$

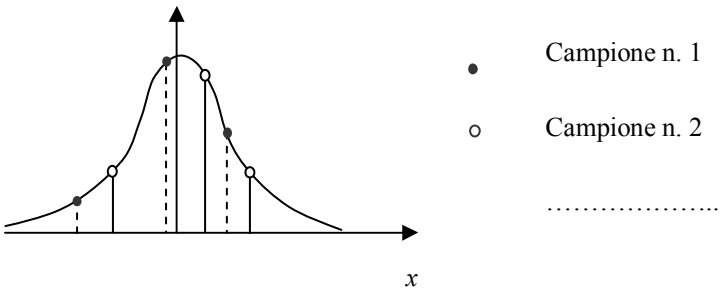


Fig. 2

Quindi si tratta di trovare la funzione 'g' che meglio stimi il parametro della popolazione. Un esempio di 'g' è la funzione che calcola il valore medio. Il valore medio campionario può essere scritto come:

$$\bar{x} = g\left(x_1, x_2, ..., x_n\right) = \frac{1}{n}\sum_i x_i$$

Un altro esempio di 'g' può essere la funzione che calcola la deviazione standard.

INFERENZA STATISTICA

Alla base della statistica c'è il concetto di stima, cioè sulla base dei risultati tratti da un campione si effettua una stima dei parametri della popolazione dalla quale il campione è stato estratto. La teoria dell'inferenza statistica riguarda prevalentemente due argomenti specifici:

a) la stima di intervallo, entro il quale si stabilisce che andrà a cadere il parametro della popolazione con un errore predefinito, sulla base di un dato campionario. Vedremo che per definire l'intervallo di confidenza è necessario conoscere: - la media campionaria, - la deviazione standard della popolazione (o in sua mancanza, si può stimare),- la numerosità del campione. La media della popolazione μ non è quindi richiesta per la stima dell'intervallo di confidenza: infatti l'intervallo di confidenza definisce gli estremi entro il quale si colloca la media della popolazione sulla base del dato campionario e assumendo una probabilità di errore α.
b) il test delle ipotesi, che è impiegato per verificare statisticamente la validità di una certa assunzione sulla base di un'evidenza campionaria. Per esempio, si può confrontare la media di una variabile rilevata su un campione con la media della popolazione per determinare se il campione può considerarsi estratto da quella popolazione. Si possono anche confrontare due medie campionarie per capire se vi sia una differenza dovuta al caso, oppure le medie delle popolazioni di provenienza siano realmente diverse.

Data una popolazione su cui è definita una variabile aleatoria X con ddp $f_X(x)$. Si effettui un campionamento casuale di *n* elementi $x = \{x_1, x_2, \ldots, x_n\}$. Se si estraggono più campioni dalla stessa popolazione si ottengono valori medi differenti per tutte le possibili estrazioni campionarie.
I valori delle medie, essendo funzione del campione, sono a loro volta una variabile aleatoria. Quindi si può ipotizzare che esista una distribuzione (campionaria) del valore medio, che avrà un proprio valore medio e una propria deviazione standard.
Notare che la deviazione standard indica la variabilità di una serie di misure. Quando si tratta della descrizione della variabilità di un valore statistico (es. la media ecc.) calcolato su un campione, si utilizza l'errore standard, definito come: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
Prima di procedere all'analisi dei dati campionari, è fondamentale che siano sempre verificati e soddisfatti alcuni assunti che riguardano la popolazione d'origine, dalla quale si presume che i dati campionari siano stati estratti, e il modo di estrarre i campioni.
Nel caso in cui anche uno solo dei presupposti non sia rispettato, neppure dopo appropriati tentativi di trasformazione dei dati che modificano la forma della distribuzione campionaria, possono ragionevolmente sorgere dubbi sulla validità delle analisi successive.

Il primo assunto da rispettare è l'indipendenza dei gruppi campionari: i campioni sottoposti ai differenti trattamenti dovrebbero essere generati per estrazione casuale da una popolazione, nella quale ogni soggetto abbia la stessa probabilità di essere incluso in un gruppo qualsiasi. In questo modo, i fattori aleatori o non controllati dovrebbero risultare casualmente distribuiti e non generare distorsioni od errori sistematici. E' una condizione che spesso è soddisfatta con facilità e che dipende quasi completamente dalla programmazione dell'esperimento.
Il secondo assunto, distintivo della statistica parametrica, riguarda la normalità delle distribuzioni.
Da essa deriva la relazione tra popolazione dei dati e medie dei campioni, secondo il teorema del limite centrale: se da una popolazione con media η e deviazione standard σ, i cui dati abbiano una forma di distribuzione non normale, si estraggono casualmente campioni di dimensione n sufficientemente grande ($n \geq 30$), le loro medie si distribuiranno normalmente con media generale η ed errore standard $\frac{\sigma}{\sqrt{n}}$.

La non-normalità della distribuzione delle medie è un indice serio di un'estrazione non casuale. La grande importanza pratica del teorema del limite centrale, che rende diffusamente applicabile la statistica parametrica, deriva dal fatto che gruppi di dati estratti da una popolazione distribuita in modo differente dalla normale, hanno medie che tendono a distribuirsi normalmente.
Se il campione estratto da una popolazione incognita è piccolo, si hanno dei risultati solo se il campione proviene da una distribuzione normale.
La distribuzione normale è la forma limite della distribuzione delle medie campionarie per n che tende all'infinito.
E' possibile comprendere il teorema del limite centrale in modo intuitivo, pensando come esempio al lancio dei dadi. Con un solo dado, i 6 numeri avranno la stessa probabilità e la distribuzione delle frequenze dei numeri ottenuti con i lanci ha forma rettangolare. Con due dadi, è possibile ottenere somme da 2 a 12 e tra esse quelle centrali sono più frequenti. All'aumentare del numero di dadi, la distribuzione delle somme o delle medie (la legge è valida per entrambe, poiché contengono la medesima informazione) è sempre meglio approssimata ad una distribuzione normale.

Il terzo assunto riguarda la omogeneità delle varianze: se sono formati per estrazione casuale dalla medesima popolazione i vari gruppi devono avere varianze eguali. Vedremo successivamente che nella statistica parametrica, è possibile verificare se esistono differenze significative tra medie campionarie, solamente quando i gruppi a confronto hanno la stessa varianza. Infatti, l'analisi statistica prevede un apposito test utilizzato per verificare l'uguaglianza statistica o meno tra le varianze di due campioni di dati.

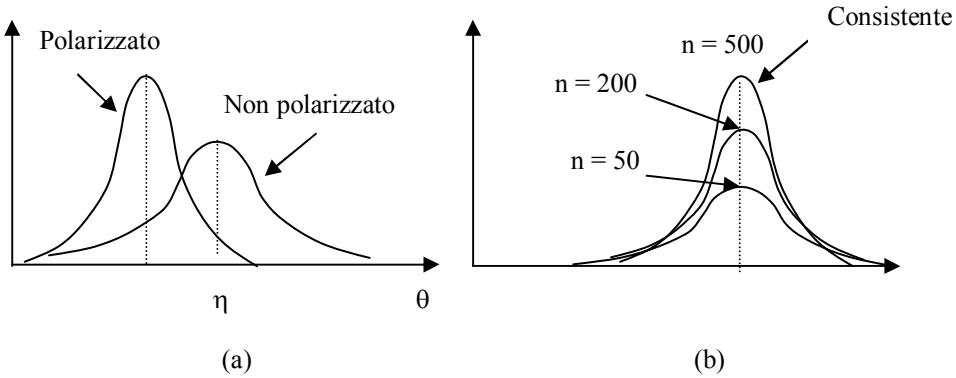
Lo stimatore della media è non polarizzato e consistente

Dopo aver trovato g(x), cerchiamo in generale la risposta a due quesiti:
a) quanto ottima è la stima effettuata con quel parametro;
b) se esistono stimatori migliori per il parametro incognito.

Il criterio che normalmente si usa è la minimizzazione dell'errore quadratico medio, ma è molto usato anche il criterio della massima verosimiglianza. In pratica si deve valutare la seguente probabilità valida per uno stimatore $\hat{\eta}$ del parametro η:

$$P\{|\hat{\eta} - \eta| \leq \varepsilon\} \geq 1 - \frac{Q}{\varepsilon^2}$$

dove ε è una quantità piccola a piacere e $Q = E[(\hat{\eta} - \eta)^2]$ è l'errore quadratico medio.
Si può dimostrare che uno stimatore il cui errore quadratico medio tende a zero, è non polarizzato (Fig. 3a) e consistente (Fig. 3b).



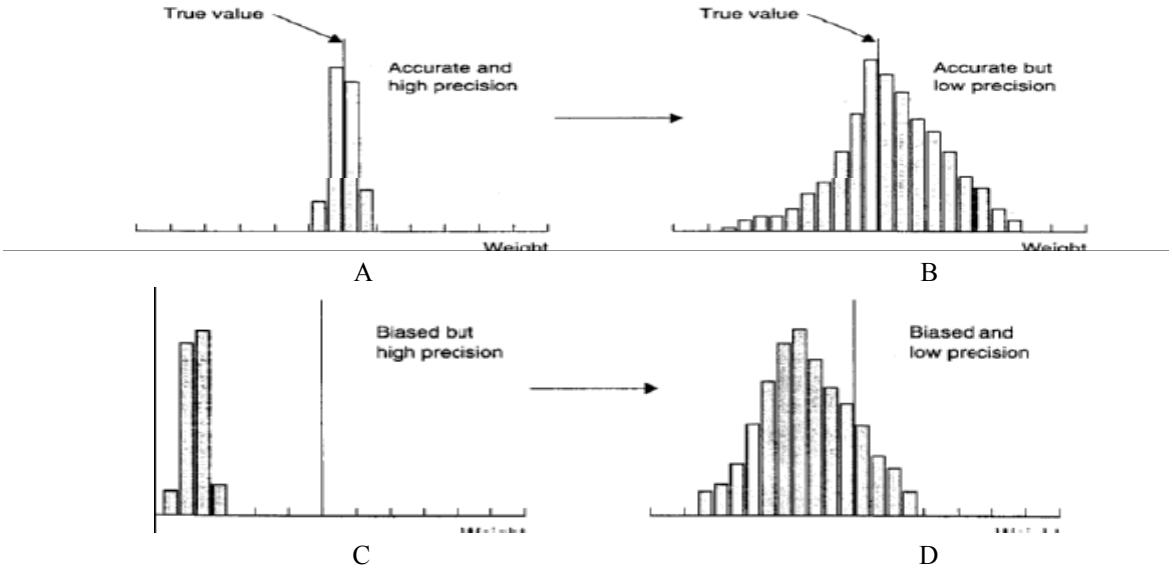
(a) (b)

Fig. 3

Uno stimatore è non polarizzato se il valore medio della distribuzione degli stimatori campionari tende al valore medio della popolazione. Uno stimatore è consistente se la varianza della distribuzione dello stimatore campionario diminuisce all'aumentare della numerosità del campione.
Sulla base di quanto sopra esposto, si può affermare che lo stimatore media campionaria $\bar{x}$ è non polarizzato e inoltre è consistente perché la varianza è $\frac{\sigma}{\sqrt{n}}$ e tende a zero all'aumentare di *n*.

Cos'è l'accuratezza e la precisione

Ambedue dipendono non dal tipo di stimatore scelto, ma dalla misura (strumento, tecnico, ecc).
L'accuratezza è la vicinanza di un valore misurato al suo valore reale e in buona parte dipende dallo strumento.
La precisione è la vicinanza di misure ripetute, al medesimo valore. Spesso dipende dalla capacità del tecnico di ripetere la misurazione con le stesse modalità e ha origine dalla sua esperienza o abilità.



Nella figura A le misure sono accurate, vicine al valore vero, e molto precise.
Nella figura B le misure sono accurate, ma poco precise, cioè differenti tra loro.
Nella figura C le misure sono non accurate, ma molto precise.
Nella figura D le misure sono non accurate e poco precise.

Variabile standardizzata Z

Abbiamo già avuto modo di dire che partendo da una popolazione con media η e deviazione standard σ, la variabile $\bar{x}$ segue una distribuzione approssimativamente normale con media ed errore standard dati da:

$$\eta_{\bar{x}} = \eta \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

A questo punto si definisce una nuova variabile detta variabile aleatoria standardizzata in base alla seguente trasformazione:

$$z = \frac{\bar{x} - \eta}{\sigma / \sqrt{n}}$$

Standardizzare una variabile significa riferirla ai parametri (media, deviazione standard) della popolazione dalla quale è stato estratto il dato campionario. Ciò fornisce una rappresentazione del dato campionario adeguata per effettuare un confronto tra risultati campionari ottenuti in laboratori differenti.

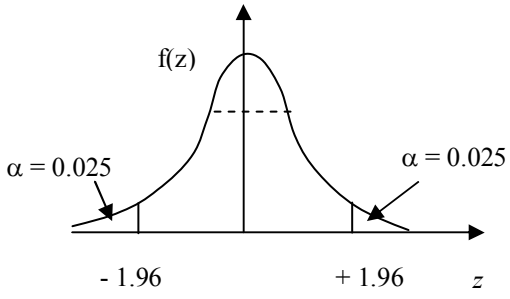


Fig. 4

La variabile *z* segue una distribuzione normale con media zero e deviazione standard unitaria. In questo modo è possibile stabilire l'area, e quindi la probabilità, che z sia compreso all'interno di un dato intervallo.

$$f(z) = e^{-z^2}$$

Quando si fissa una soglia, per esempio ai valori di z = ± 1.96, si ha la probabilità del 95% che il dato campionario cada all'interno di tale intervallo, e con un errore del 5% che il campione cada al di fuori di tale intervallo. Tenuto conto della definizione data per z, è possibile mettere in relazione il valore tabulato di z con il dato campionario

Per consolidata convenzione internazionale, i livelli di soglia delle probabilità α ai quali di norma si ricorre per effettuare i test statistici descritti nel paragrafo sull'inferenza statistica, sono tre: 0.05 (5%); 0.01 (1%); 0.001 (0.1%). Le tre probabilità e i valori critici più frequentemente utilizzati sono: 5% => 1.96; 1% => 2.58; 0.1% => 3.28.

Variabile standardizzata t di Student

Se il campione estratto da una popolazione incognita è piccolo e il campione proviene da una distribuzione normale di cui non si conosce la deviazione standard, si definisce una nuova variabile standardizzata T:

$$t_{n-1} = \frac{\bar{x} - \eta}{\hat{s}_{\bar{x}}} \qquad con \quad \hat{s}_{\bar{x}} = \frac{\hat{s}}{\sqrt{n}}$$

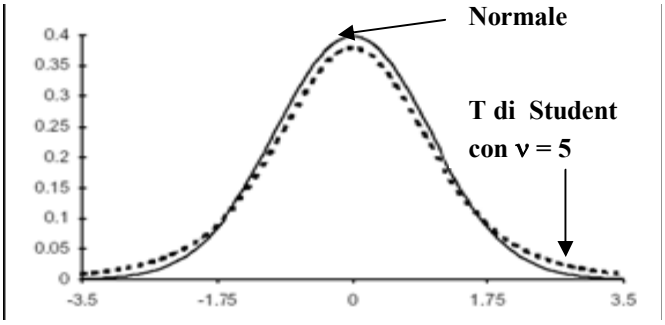
dove con $\bar{x}$ si è indicato il valore medio campionario e con $\hat{s}_{\bar{x}}$ la deviazione standard campionaria. In questo caso la deviazione standard $\hat{s}_{\bar{x}}$ non era nota a priori, ed è stata stimata direttamente dai dati.

La nuova variabile segue una distribuzione t di Student con (ν = n - 1) gradi di libertà, dove n è la dimensione del campione.

La distribuzione t di Student è decritta dalla seguente legge:

$$f(t) = (1 + \frac{t^2}{\nu})^{-(\nu+1)/2}$$

e dipende dal parametro ν. E' simile alla distribuzione di Gauss e per n molto grande potrebbe essere ben approssimata da essa.



Come per la variabile standardizzata z, così anche per la variabile standardizzata T è possibile leggere sulle tavole l'area corrispondente a determinati valori di T ed applicare i ragionamenti sviluppati per la variabile z.

Variabile standardizzata χ^2

Finora abbiamo discusso la distribuzione della variabile standardizzata relativa alla media campionaria, ma con ragionamento analogo si può studiare la distribuzione della varianza campionaria.

Supponiamo di estrarre da una popolazione normale con varianza σ^2 un certo numero di campioni casuali di numerosità n. Indicando con s^2 la varianza campionaria si ottiene la seguente variabile standardizzata:

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

è una variabile aleatoria con distribuzione χ^2 (chi quadro) di parametro ν = n – 1. Si dimostra che la distribuzione χ^2 ha media $\eta = \nu$ e $\sigma^2 = 2\nu$. Inoltre la distribuzione χ^2 è definita solo per valori positivi di x ed è simmetrica in particolare per bassi valori di ν.

Intervallo di confidenza per la media (varianza nota)

Come abbiamo già avuto modo di dire, la media campionaria è una buona stima delle media della popolazione, ma a causa dell'errore dovuta alla variabilità casuale del campione, la media campionaria coinciderà con la media della popolazione.
Ha pertanto più significato stimare un intervallo per la media, che ci dia informazioni sulla probabile entità della media stessa.

Abbiamo già detto che data una popolazione con media η e varianza σ^2, la variabile standardizzata $z = \frac{\bar{x} - \eta}{\sigma / \sqrt{n}}$ ha approssimativamente la distribuzione normale standardizzata. Apposite tavole forniscono l'area della distribuzione normale standardizzata in funzione dei valori di z. Pertanto, utilizzando i valori tabulati si possono determinare due valori c_1 e c_2 tali che:

$$p(c_1 \le z \le c_2) = 1 - \alpha$$

in genere tali intervalli sono scelti in modo simmetrico ($c_1 = c_2 = c$).
La precedente uguaglianza è equivalente a scrivere:

$$p(\bar{x} - c\frac{\sigma}{\sqrt{n}} \le \eta \le \bar{x} + c\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Fissato il livello di significatività $\alpha = 0.05$ (nella tabella α è l'area a sinistra di z = -c più quella a destra di z = c) si ottiene il seguente intervallo di confidenza:

$$p(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \eta \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95 .$$

La precedente equazione può essere interpretata nel modo seguente: se si ripete infinite volte l'esperimento di estrazione di *n* campioni e ad ogni esperimento si calcola la media $\bar{x}$ e l'intervallo $\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$ e $\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$, allora nel 95% dei casi, η è contenuto nell'intervallo calcolato.

Esempio.
Un campione di 200 cavie ha un peso medio 0.824 kg e una deviazione standard di 0.042. Determinare gli intervalli di confidenza al 95% e 99% sulla media η incognita.
Sappiamo che:

$$p(0.824 - 1.96\frac{0.042}{\sqrt{200}} \le \eta \le 0.824 + 1.96\frac{0.042}{\sqrt{200}}) = 0.95$$

dove il valore 1.96 è letto in corrispondenza delle tavole della variabile standardizzata z. Pertanto l'intervallo di confidenza per la media è: $0.8182 \le \eta \le 0.8298$, cioè il peso delle cavie è compreso nell'intervallo trovato con una probabilità del 95%.
Per $\alpha = 0.01$ si ha:

$$p(0.824 - 2.58\frac{0.042}{\sqrt{200}} \le \eta \le 0.824 + 2.58\frac{0.042}{\sqrt{200}}) = 0.99$$

e quindi l'intervallo di confidenza per la media è: $0.8153 \le \eta \le 0.8317$.

Si noti come all'aumentare del livello di confidenza (dal 95% al 99%) sia cresciuta anche l'ampiezza dell'intervallo. L'interpretazione è la seguente: se desidero conoscere l'intervallo di confidenza della media con una minore probabilità di commettere un errore, essendo i dati campionari invariati come numero, necessariamente l'intervallo di confidenza si deve allargare. Al contrario l'intervallo diminuisce al crescere della numerosità del campione, cioè al crescere dell'evidenza sperimentale.

Intervallo di confidenza per la media (varianza incognita)

Nel caso di varianza incognita, la variabile standardizzata $t_{n-1} = \dfrac{\overline{x} - \eta}{\dfrac{\hat{s}}{\sqrt{n}}}$ segue una distribuzione t di Student con (ν = n - 1) gradi di libertà. Fissata la numerosità del campione n, e stabilità la significatività α, mediante l'uso delle tavole si potranno allora determinare due valori c_1 e c_2 tali che:

$$p(c_1 \le v \le c_2) = 1 - \alpha$$

ipotizzando $c_1 = c_2 = c$, si ottiene:

$$p(\overline{x} - c\frac{\hat{s}}{\sqrt{n}} \le \eta \le \overline{x} + c\frac{\hat{s}}{\sqrt{n}}) = 1 - \alpha \qquad \text{dove: } \hat{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\overline{x}})^2}{(n-1)}}\,.$$

Per valori molto grandi di n, la distribuzione t di Student è praticamente uguale alla distribuzione normale e dunque si può usare z.

Esempio.
Un campione di 12 cavie ha avuto in un determinato periodo di tempo i seguenti incrementi di peso in grammi: 55,62,54,57,65,64,60,63,58,67,63,61. Qual'è l'intervallo di confidenza, al livello del 95%, dell'incremento di peso medio?
L'incremento di peso medio calcolato sulle 12 cavie risulta essere di gr. 60.75. Con 11 gradi di libertà e un livello di significatività $\alpha = 0.05$, sulla tabella si legge un valore del coefficiente c = 2,20. Pertanto l'intervallo di confidenza si ottiene risolvendo la seguente formula:

$$p(60.75 - 2.20\frac{\sqrt{16.38}}{\sqrt{12}} \le \eta \le 60.75 - 2.20\frac{\sqrt{16.38}}{\sqrt{12}}) = 1 - \alpha = 0.95$$

L'intervallo di confidenza risulta essere: $58.17 \le \eta \le 63.32$.

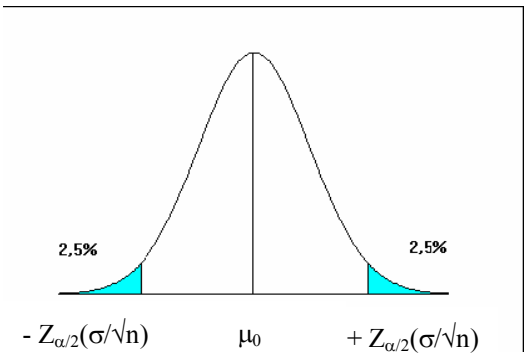
Test delle ipotesi

Un test statistico può essere definito come una procedura che, sulla base di dati campionari e con un certo grado di probabilità consente di decidere se è ragionevole respingere una certa ipotesi (ed accettare implicitamente l'ipotesi alternativa), oppure se non esistono elementi sufficienti per respingerla.
In pratica il test statistico è strutturato in:

- Ipotesi iniziale (H_0), detta anche ipotesi nulla, che solitamente sostiene la non diversità, cioè qualunque differenza, se riscontrata, è dovuta al caso. Essa afferma che gli effetti osservati nei campioni sono dovuti a fluttuazioni casuali, sempre possibili data la variabilità tra gli individui; si tratta di variazioni che sono tanto più marcate quanto più ridotto è il numero di osservazioni. Simbolicamente si scrive: $H_0 : \eta = \eta_0$, cioè non ci sono differenze tra le medie. Nel test delle ipotesi η e η_0 rappresentano le due evidenze da confrontare.
- Ipotesi alternativa H_a contrapposta all'ipotesi nulla H_0.
- Livello di significatività α per indicare l'errore massimo che accettiamo di commettere sostenendo che esiste una differenza. Il valore α viene detto errore di I specie.

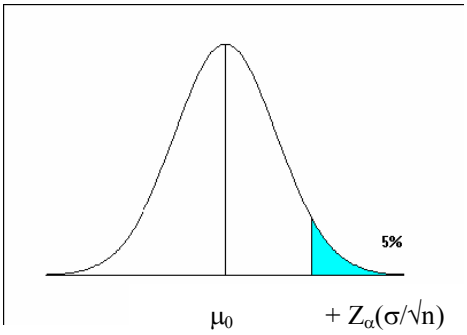
L'ipotesi alternativa, in rapporto al problema e al test utilizzato, può essere di tre tipi, tra loro mutuamente esclusivi:

1) bilaterale: $H_0 : \eta = \eta_0$ contro $H_a : \eta \neq \eta_0$. Per esempio, quando si confronta l'effetto di due sostanze A e B sull'accrescimento di due gruppi di animali, si è interessati a valutare quale abbia l'effetto maggiore e inoltre si ritiene che ambedue le risposte siano possibili. In questo caso l'ipotesi nulla stabilisce che le due diete sono comparabili, mentre l'ipotesi alternativa dice che una delle due risposte è prevalente.
I concetti espressi sul test bilaterale sono espressi nel seguente grafico:



Nel seguito ci si riferisce alla variabile standardizzata t, ma lo stesso ragionamento può essere applicato alla variabile z.
Il test bilaterale consiste nell'applicare la seguente formula: $p(-c \le t \le c) = 1 - \alpha$ per accettare l'ipotesi nulla, dove c viene letto sulle tavole di t di Student. Si fissa un valore per l'errore α, esempio 5%.
Poiché la distribuzione di T è simmetrica e dato che si usa il test bidirezionale, il valore di α da inserire nella relazione: $p(-c \le T \le c) = 1 - \alpha$ è la somma dei valori delle aree sottese dalle code della distribuzione.
Quando il valore calcolato di t è compreso tra ± c, si accetta l'ipotesi nulla, per cui $H_0 : \eta = \eta_0$ e quindi le eventuali differenze riscontrate tra i due valori sotto test sono dovute al caso (non c'è differenza significativa).
Quando il valore di t risulta esterno all'intervallo ± c, si rifiuta l'ipotesi nulla e si afferma che le differenze riscontrate sono significativamente diverse sulla base dell'errore prefissato.

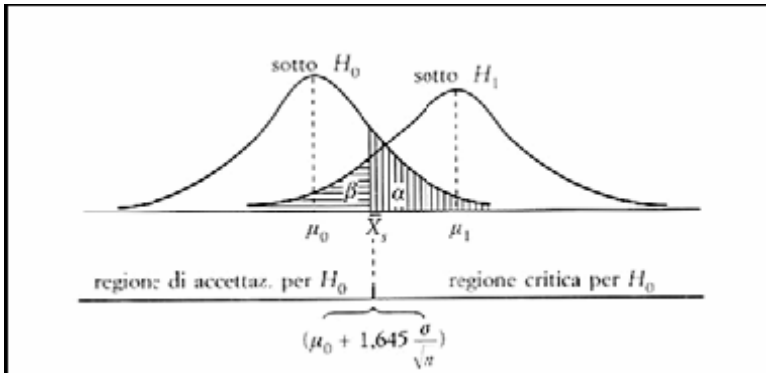
2) unilaterale: è unilaterale quando è possibile escludere a priori il fatto che la media di un campione possa essere minore o maggiore dell'altra. Per esempio quando si confrontano i risultati di un principio attivo (tossico) e di un placebo come dieta per cavie, non ci si può aspettare che gli animali ai quali è stato somministrato il tossico abbiano risultati migliori nella crescita rispetto a quelli ai quali è stato somministrato placebo.
Il test unilaterale si usa quando l'ipotesi nulla è: $H_0 : \eta = \eta_0$, mentre l'ipotesi alternativa è: $H_a : \eta < \eta_0$ (cioè: $p(t \le -c) = \alpha$), oppure $H_a : \eta > \eta_0$ ($p(t \ge c) = \alpha$).
Si rifiuta l'ipotesi nulla quando, quando $t < -c$, essendo l'ipotesi alternativa: $p(t \le -c) = \alpha$ per un prefissato valore di significatività α a cui corrisponde un valore c letto sulle tavole. Oppure si rifiuta l'ipotesi nulla, quando $t \ge c$, essendo l'ipotesi alternativa: $p(t \ge c) = \alpha$ per un prefissato valore di significatività α a cui corrisponde un valore c letto sulle tavole.
In caso contrario, cioè se $t > -c$ quando $H_a : \eta < \eta_0$, oppure $t < c$ quando $H_a : \eta > \eta_0$, allora si accetta l'ipotesi nulla.
I concetti espressi sul test unilaterale sono rappresentati nel seguente grafico:



La differenza tra test unilaterale e test bilaterale non è solamente una questione teorica: è una scelta con effetti pratici rilevanti, poiché è importante per la determinazione della zona di rifiuto dell'ipotesi nulla. In una distribuzione normale, prendendo come livello di significatività α = 5%,
- in un test ad una coda l'area di rifiuto dell'ipotesi nulla inizia dal valore critico Z_α = 1,645
- in un test a due code essa inizia dal valore critico $Z_{\alpha/2}$ = 1,96.

Errori di I e di II specie

L'insieme dei valori ottenibili con il test può essere diviso in due zone:
-una zona di rifiuto dell'ipotesi nulla, collocata oltre gli estremi della distribuzione secondo la direzione dell'ipotesi alternativa. Se il valore dell'indice statistico cade nella zona di rifiuto, si respinge l'ipotesi nulla. Si rifiuta l'ipotesi nulla quando il dato campionario appartiene ad una popolazione con distribuzione di dati diversa da quella su cui si basa l'ipotesi nulla.
- una zona di accettazione dell'ipotesi nulla, che comprende i restanti valori.
Con riferimento alla figura che segue, la distribuzione sulla quale si definisce l'ipotesi nulla H_0 può rappresentare la distribuzione dello stimatore su campioni prelevati da una popolazione di soggetti normali, mentre quella sulla destra può rappresentare la distribuzione dello stimatore su campioni che non sono normali, per esempio patologici, che rappresenta l'ipotesi alternativa H_a.
Quando il valore assunto dallo stimatore supera la soglia prefissata, si dice che le differenze tra i risultati campionari e il valore della popolazione normale non sono dovute al caso, ma il campione è parte di un'altra popolazione, per esempio i patologici. L'errore che posso commettere nel rifiutare l'ipotesi nulla è α (infatti, pur facendo parte della distribuzione dei normali, il valore dello stimatore cade alla destra della soglia).



La sovrapposizione delle due distribuzioni dice anche che posso commettere un errore anche quando accetto l'ipotesi nulla, ma essa è falsa, commettendo un errore β (pur facendo parte della distribuzione dei patologici, il valore dello stimatore cade alla sinistra della soglia).
Pertanto, nell'applicazione di un test statistico, si possono commettere due errori:
- rifiutare l'ipotesi nulla quando in realtà è vera: si parla di errore di prima specie (errore α) e corrisponde alla probabilità che il valore campionario cada nella zona di rifiuto, quando l'ipotesi nulla è vera;
- accettare l'ipotesi nulla quando in realtà è falsa: si parla di errore di seconda specie (errore β) e corrisponde alla probabilità che il valore campionario cada nella zona di accettazione (si accetta l'ipotesi nulla) quando in realtà essa è falsa.

Nella tabella che segue sono riportate le possibili situazioni.

	H_0 vera	**H_0 falsa**
Accetto H_0	**Scelta corretta $P = 1 - \alpha$**	**Errore II specie $P = \beta$**
Rifiuto H_0	**Errore I specie $P = \alpha$**	**Scelta corretta $P = 1- \beta$**

Il termine $P = 1- \beta$, il quale rappresenta l'area della distribuzione alla destra della soglia, è chiamato potenza del test. A parità di numerosità del campione, se diminuisco la probabilità di errore di I specie (α), aumento la probabilità di errore di II specie (β), cioè diminuisce la potenza del test.

Le fasi della verifica delle ipotesi

Le principali fasi richieste per mettere a punto un test d'ipotesi sono:
1 Definire l'ipotesi H_0
2 Definire l'ipotesi H_a

3 Specificare il livello di significatività α
4 Determinare la dimensione n del campione
5 Determinare la statistica del test
6 Fissare il valore (test unidirezionale) o i valori critici (test bidirezionale) che dividono le regioni di rifiuto e di accettazione
7 Calcolare il valore campionario della statistica
8 Confrontare il valore campionario della statistica con il/i valori critici
9 Prendere una decisione

Test delle ipotesi sulla media

Il test delle ipotesi può essere utilizzato per il confronto tra media campionaria e media della popolazione, il confronto tra dato campionario e media della popolazione, oppure tra medie campionarie. Così come per la stima d'intervallo, si utilizzano le statistiche di z o t a seconda che la varianza sia nota oppure sia necessario stimarla dai dati campionari.
Nel seguito prenderemo in esame alcuni esempi dell'applicazione di test parametrici.

Esempio:
Riprendiamo il caso dell'incremento del peso nelle 12 cavie (incremento di peso medio = 60.75 e varianza = 16.38). Inoltre, sapendo che cavie dello stesso tipo non sottoposte a dieta mostrano un incremento di peso medio di 65 grammi, ci si domanda se le risultanze campionarie siano tali da attribuire alla dieta le differenze riscontrate nell'incremento di peso oppure sia semplicemente dovuto al caso.

Si tratta di un test bidirezionale:
1) si fissa il livello di significatività: $\alpha = 0.05$
2) si specificano le due ipotesi: $H_0 : \eta = 65$; $H_a : \eta \neq 65$ (il test è bidirezionale, cioè l'area α relativa alla variabile standardizzata deve essere letta sulle due code, ciascuna di valore α/2).
3) Si individua la variabile aleatoria del test: $t = \dfrac{\bar{x} - 65}{\hat{s}_{\bar{x}}}$ Tale variabile descrive l'andamento dei risultati campionari sotto l'ipotesi nulla, cioè a condizione che le differenze siano dovute al caso (la dieta non ha effetto).
4) Si pone a confronto il valore assunto da t, calcolato sui dati campionari: $t = \dfrac{60.75 - 65}{\sqrt{16.38/12}} = -3.63$ con il valore critico c. Dato che T = - 3.63 < - 2.20 = -c, si rifiuta l'ipotesi nulla al livello di significatività $\alpha = 0.05$, si rifiuta cioè l'ipotesi che la differenza d = 60.75 – 65 sia dovuta al caso. Nel caso di test unidirezionale avremmo avuto: $H_0 : \eta = 65$; $H_a : \eta < 65$ (oppure $H_a : \eta > 65$). In questo caso il valore critico *c* soddisfa la relazione: $p(t \leq -c) = \alpha$ e c = 1.80. Pertanto: t = - 3.63 ≤ - 1.8 = - c e si rifiuta l'ipotesi nulla.

Esempio
Si supponga di avere a disposizione un gruppo di n_1 osservazioni campionarie relative ad una popolazione normale X con media η_x incognita e varianza nota σ_x^2, ed un secondo gruppo di n_1 osservazioni campionarie relative ad una popolazione normale Y con media η_y incognita e varianza nota σ_y^2. Supponiamo di voler stabilire se la differenza eventualmente riscontrata tra le due medie campionarie $\bar{x}$ e $\bar{y}$ sia da attribuire al caso o al fatto che le due medie η_x e η_y delle popolazioni di provenienza dei due campioni sono diverse. In altri termini si vuole decidere per l'eventuale significatività statistica della differenza: $d = \bar{x} - \bar{y}$.
Pertanto l'ipotesi nulla H_0 è:
H_0: $\bar{x} - \bar{y} = d$;
L'ipotesi alternativa è una delle seguenti:

H_1: $\bar{x} - \bar{y} > d$;
H_1: $\bar{x} - \bar{y} < d$;
H_1: $\bar{x} - \bar{y} \neq d$;

Nei primi due casi si fa un test unidirezionale, mentre nell'ultimo caso si fa un test bidirezionale.

In tal caso la variabile z è data da: $z = \frac{(\bar{x} - \bar{y}) - d}{\sqrt{\sigma_x^2 / n_1 + \sigma_y^2 / n_2}}$ ed ha, quando l'ipotesi nulla è vera, legge di distribuzione normale.

Si supponga ora di avere a disposizione due gruppi di n_1 e n_2 osservazioni campionarie relative a due popolazioni normali X e Y con rispettive media η_x e η_y incognite e varianze uguali e incognite. In questo caso si deve ricorrere ad una stima campionaria della varianza.
Tuttavia, per quanto riguarda il test, si procede con le ipotesi come nel caso visto in precedenza, tenendo conto che in questo caso il test da impiegare è quello basato sulla variabile T data da:

$$t = \frac{(\bar{x} - \bar{y}) - d}{\sqrt{s^2 (1/n_1 + 1/n_2)}} \text{ dove: } s^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}$$

Test F sull'uguaglianza di due varianze

Le assunzioni di validità di un test parametrico sul confronto tra due o più medie sono essenzialmente tre:
1 - l'indipendenza dei dati entro e tra campioni;
2 - l'omogeneità della varianza: il confronto tra due o più medie è valido se e solo se le popolazioni dalle quali i campioni sono estratti hanno varianze uguali;
3 - i dati o (detto ancor medio) gli scarti rispetto alla media sono distribuiti normalmente,

Con due campioni indipendenti, l'assunzione di validità più importante è quella dell'uguaglianza della varianza, perché rispetto ad essa il test t è meno robusto. La varianza è una stima della credibilità di una media: dati molto variabili, quindi con una varianza ampia, a parità del numero di osservazioni hanno medie meno credibili, appunto perché più variabili come i loro dati.
Per confrontare due medie, è quindi necessario che la loro credibilità sia simile, soprattutto quando i campioni hanno dimensioni molto differenti, per cui una varianza diversa determina la stima della probabilità α e del rischio β che possono essere sensibilmente differenti da quelli nominali o dichiarati.
Per l'applicazione del test t, l'omogeneità tra le varianze di due gruppi (A e B) è verificata con un test bilaterale, dove l'ipotesi nulla H_0 e l'ipotesi alternativa H_a sono:

$H_0 : \sigma_A = \sigma_B$

$H_a : \sigma_A \neq \sigma_B$

E' possibile anche un test unilaterale, quando una delle due varianze tende a essere sistematicamente maggiore o minore dell'altra. Ma, nella pratica della ricerca ambientale e biologica, in questo contesto di analisi preliminare per il confronto tra le medie di due campioni indipendenti, il test unilaterale è un caso più raro.

Tra i test parametrici più diffusi in letteratura e nelle pubblicazioni per verificare l'omogeneità della varianza bilaterale o unilaterale vi è sicuramente il test F o del rapporto tra le due varianze, definito come segue:

$$F_{(n_{\max}-1),(n_{\min x}-1)} = \frac{S_{\max}^2}{S_{\min}^2}$$

dove:

$S_{\max}^2$ è la varianza maggiore

$S_{\min}^2$ è la varianza minore

n_{max} è il numero di dati nel gruppo con varianza maggiore

n_{min} è il numero di dati nel gruppo con varianza minore

Fondato sull'ipotesi che le due varianze siano uguali (cioè che l'ipotesi nulla H0 sia vera), il rapporto tra esse dovrebbe essere uguale a 1.

I valori critici della distribuzione F dipendono dai gradi di libertà e cioè:
- $\nu_1 = n_1 - 1$ rappresenta i gradi di libertà del numeratore, riportati nella prima riga della tabella,
- $\nu_2 = n_2 - 1$ rappresenta i gradi di libertà del denominatore, riportati nella prima colonna della tabella.

Solo se si dimostra che l'ipotesi nulla è vera e pertanto che i due gruppi hanno varianze statisticamente uguali, è possibile usare il test t di Student per 2 campioni indipendenti.

Test di adattamento o del χ^2

In questo paragrafo ci occuperemo di un metodo statistico per stabilire se un campione di dati misurati si adatta ad una predeterminata distribuzione teorica. Un tale tipo di test si chiama test di adattamento.
Per applicare il test del chi quadro si deve disporre della frequenza osservata O_i e della frequenza attesa A_i. Per valutare la bontà dell'adattamento si usa la seguente variabile:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - A_i)^2}{A_i}$$

essendo K il numero delle classi su cui si desidera effettuare il confronto.
Si dimostra che per n sufficientemente grande, tale statistica ha una distribuzione approssimabile a quella della variabile χ^2 con ν = K-m-1 gradi di libertà, dove m è il numero dei parametri della distribuzione teorica stimati servendosi dei dati del campione.
Nel caso l'ipotesi nulla sia che i dati si adattino alla distribuzione teorica, allora si rifiuta l'ipotesi nulla H_0 se il valore di χ^2 calcolato dai dati è maggiore del valore critico χ^2_α, dove α è il livello di significatività.

Esempio.
In media in una classe di studenti del primo anno di un corso universitario l'età assume quattro valori come appare dalla seguente tabella:

Età	19	20	21	22
Percentuale	26%	32%	15%	27%

Effettuando una rilevazione in una certa classe di 350 studenti si ottengono i seguenti risultati:

Età	19	20	21	22
O_i	80	120	60	90

Si vuole sapere se, sulla base delle rilevazioni sperimentali, le due tabelle concordano statisticamente, oppure le differenze sono statisticamente rilevanti.
A questo punto, applicando la percentuale media al caso dei 350 studenti, si otterrebbero i seguenti risultati teorici:

Età	19	20	21	22
A_i	91	112	52.5	94.5
$\frac{(O_i - A_i)^2}{A_i}$	1.33	0.57	1.07	0.21

Sommando gli scarti su ciascuna classe si ottiene: $\chi^2 = 3.18$. I gradi di libertà ν = K-m-1 sono 3, in quanto nessun coefficiente è stato stimato dai dati (m = 0). Al livello di significatività del 5% si ottiene dalle tavole del χ^2 il seguente valore critico: $\chi^2_{0.05} = 7.81$. Poiché il valore di χ^2 calcolato dai dati è inferiore a quello ricavato dalle tavole, si accetta l'ipotesi nulla e cioè non esiste differenza significativa tra i dati campionari e i valori medi (popolazione).