

Inferenza Statistica

- Consiste nel trarre delle conclusioni (*fare inferenza*) relativamente a una certa *popolazione* sulla base di un *campione*.
- *Popolazione*: totalità dei casi (unità statistiche) sui quali si manifesta il fenomeno oggetto di studio.
- *Campione*: sottoinsieme delle unità che compongono la popolazione estratte in modo casuale. L'estrazione può essere:
 - *con ripetizione* (o *con reimmisione*): una volta estratta, un'unità viene *reinserita* nella popolazione e quindi può essere selezionata più volte;
 - *senza ripetizione* (o *senza reimmisione*): una volta estratta, una unità *non* viene reinserita nella popolazione e quindi può essere selezionata al più una volta.
- *Vantaggi*: in termini di tempo e costi rispetto a una rilevazione completa (consenso).

Esempio

- Si supponga che la *popolazione* sia composta di $N = 100$ soggetti che indichiamo con le *etichette* $1, 2, \dots, 100$ e che siamo interessati a *stimare* l'altezza media dei soggetti che compongono questa popolazione sulla base di un *campione* di dimensione $n = 5$.
- *Estrazione con ripetizione*: per 5 volte si seleziona un numero casuale tra 1 e 100; il campione sarà formato dalle unità con le etichette corrispondenti:
37, 25, 80, 89, 80
- *Estrazione senza ripetizione*: per 5 volte si seleziona un numero casuale tra 1 e 100 togliendo via via quelli già selezionati (se il primo numero selezionato è 8, questo non potrà essere più selezionato).
8, 100, 88, 39, 29

- Per l'*estrazione* si può far ricorso a programmi informatici o alle tavole dei numeri casuali.
- L'*altezza media* della popolazione può essere stimata tramite l'altezza media dei soggetti che compongono il campione.
- Nel caso di *campionamento con ripetizione*:

Unità	Altezza
37	170,75
25	186,14
80	173,39
89	185,12
80	173,39
Media	177,76

Formalizzazione in termini statistici

- Si rappresenta il fenomeno di interesse tramite una *variabile aleatoria* (X), discreta o continua, che ha una certa distribuzione nella popolazione che viene indicata, in generale, con $f(x)$.
- Per semplicità si considera solo il campionamento con ripetizione.
- *Campione casuale*: variabile aleatoria multipla di dimensione n
 $(X_1, X_2, \dots, X_n)$
le cui componenti sono indipendenti e identicamente distribuite ognuna con distribuzione $f(x)$.
- *Campione osservato*: realizzazione di $(X_1, X_2, \dots, X_n)$
 $(x_1, x_2, \dots, x_n)$

- *Spazio campionario* (Ω): insieme di tutti i possibili campioni. Può essere discreto o continuo a seconda della natura (discreta o continua) della variabile X .
- *Distribuzione del campione*: funzione di densità o di probabilità del campione ($X_1, X_2, \dots, X_n$)

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \cdots f(x_n) = \prod_{i=1}^n f(x_i)$$

Esempio

- Si assuma che il *numero di figli per famiglia* (variabile discreta) abbia la seguente distribuzione nella popolazione

x	$f(x)$
0	0,1
1	0,3
2	0,6
1	

- Ci sono 9 campioni distinti di dimensione $n = 2$. Questi formano lo *spazio campionario* (Ω).

x_1	x_2	$f(x_1)$	$f(x_2)$	$f(x_1, x_2)$
0	0	0,1	0,1	0,01
0	1	0,1	0,3	0,03
0	2	0,1	0,6	0,06
...	...	...	...	...
2	1	0,6	0,3	0,18
2	2	0,6	0,6	0,36
1				1

Esempio

- Si assuma che il numero di addetti per impresa (variabile discreta) abbia nella popolazione distribuzione di Poisson con parametro $\lambda = 3,5$.

$$X \sim Po(3,5)$$

- Lo spazio campionario (Ω) ha infiniti elementi.
- La probabilità del campione osservato

2, 4, 3, 3

è pari a

$$\begin{aligned} f(2,4,3,3) &= f(2) \cdot f(4) \cdot f(3) \cdot f(3) = \\ &= \frac{3,5^2}{2!} e^{-3,5} \cdot \frac{3,5^4}{4!} e^{-3,5} \cdot \frac{3,5^3}{3!} e^{-3,5} \cdot \frac{3,5^3}{3!} e^{-3,5} = 0,00163 \end{aligned}$$

- In pratica, la distribuzione della variabile aleatoria X non è completamente nota, ma nota a meno di uno o più *parametri* (θ) che la caratterizzano:

$$X \sim f(x; \theta)$$

- *Spazio dei parametri* (Θ): insieme di tutti i possibili valori che può assumere il parametro θ .
- Dato che il parametro θ non è noto, si vogliono trarre delle conclusioni su di esso sulla base di un campione estratto dalla popolazione. Cioè si vuole *fare inferenza* su θ .

- *Esempio 1*: si assume che l'*altezza* (variabile continua) abbia nella popolazione distribuzione normale con media μ e varianza σ^2 non noti:

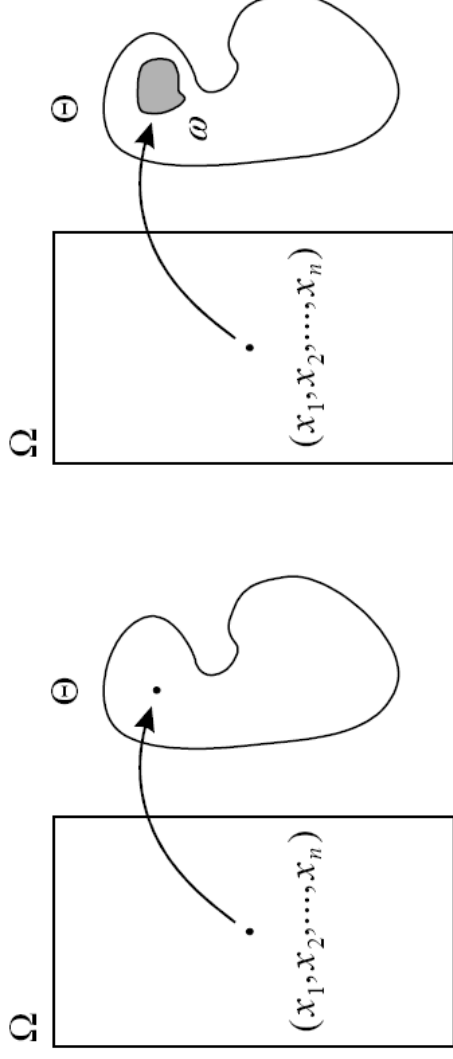
$$X \sim N(\mu, \sigma^2), \quad \mu \in (-\infty, \infty), \quad \sigma^2 \in (0, \infty),$$

- *Esempio 2*: si assume che il *numero di addetti per impresa* (variabile discreta) abbia nella popolazione distribuzione di Poisson con parametro λ non noto:

$$X \sim Po(\lambda), \quad \lambda \in (0, \infty)$$

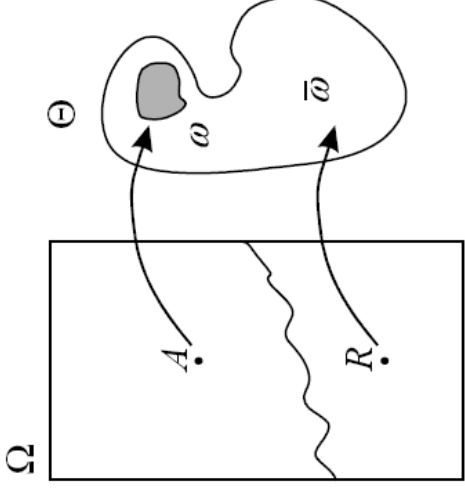
Metodi di inferenza

- **Stima dei parametri:** sulla base del campione si assegna al parametro di interesse (θ) un valore (stima puntuale) o un insieme di valori (stima per intervallo).



- *Esempio 1:* sulla base di un campione di soggetti si stima che l'altezza media degli italiani è pari a 175,36cm o è compresa nell'intervallo (173; 178).
- *Esempio 2:* sulla base di un campione di famiglie si stima che il numero medio di figli per famiglia in Italia è 1,8 o che tale media è compresa nell'intervallo (1,5; 2,3).

- **Verifica delle ipotesi:** si formula un'ipotesi, o congettura, sul parametro di interesse (θ) e si verifica, sulla base del campione, se tale ipotesi è o meno plausibile.



- *Esempio 1:* si formula l'ipotesi che l'altezza media degli italiani sia pari a 175cm e sulla base di 10 soggetti estratti casualmente si decide se tale ipotesi è plausibile o meno.
- *Esempio 2:* si formula l'ipotesi che il numero medio di figli delle famiglie lombarde sia inferiore di quello delle famiglie calabresi. Sulla base di due campioni di 50 famiglie ognuno, di cui il primo estratto in Lombardia e il secondo in Calabria, si decide se tale ipotesi è plausibile o meno.

Diversi approcci all'inferenza

- **Classico** (Fisher, Neyman, Pearson): concezione frequentista della probabilità; fa uso unicamente delle informazioni contenute nel campione.
- **Bayesiano** (Lindley, Savage, de Finetti): concezione soggettivista della probabilità; fa uso anche di informazioni a priori sui parametri espresse tramite distribuzioni di probabilità.
- **Teoria delle decisioni** (Wald): tiene conto delle conseguenze di decisioni alternative espresse tramite funzioni di perdita; può fare uso anche di informazioni a priori.

Statistiche campionarie

- Servono a sintetizzare l'informazione contenuta in un campione.
- Una **statistica campionaria** è una qualsiasi funzione del campione

$$t(X_1, X_2, \dots, X_n)$$

- *Statistiche campionarie più comuni:*

➤ Media campionaria:
$$\bar{X} = \sum_{i=1}^n X_i / n$$

➤ Varianza campionaria:
$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$$

➤ Statistiche d'ordine:
$$Y_1 \leq Y_2 \leq \dots \leq Y_n$$

- *In particolare* saremo interessati alla media campionaria ($\bar{X}$) e alla varianza campionaria (S^2).

Esempio

- Si consideri nuovamente il campione di dimensione $n = 5$ soggetti per ognuno dei quali si riporta l'altezza (in cm.)

170,75; 186,14; 173,39; 185,12; 173,39

- La *media campionaria* risulta pari a

$$\bar{x} = \frac{170,75 + 186,14 + 173,39 + 185,12 + 173,39}{5} = 177,76$$

- La *varianza campionaria* risulta pari a

$$s^2 = \frac{(170,75 - 177,76)^2 + (186,14 - 177,76)^2 + \dots + (173,39 - 177,76)^2}{4} = 52,93$$

Distribuzioni campionarie

- Ogni statistica campionaria ha una sua distribuzione, detta **distribuzione campionaria**.
- Esempio: si riprendano i 9 campioni distinti di dimensione $n = 2$ sui quali si è osservato il numero di figli

x_1	x_2	$f(x_1, x_2)$	$\bar{X}$	$\bar{X}$	$f(\bar{X})$
0	0	0,01	0	0	0,01
0	1	0,03	0,5	0,5	0,06
0	2	0,06	1	1	0,21
1	0	0,03	0,5	1,5	0,36
1	1	0,09	1	2	0,36
1	2	0,18	1,5		
2	0	0,06	1		
2	1	0,18	1,5		
2	2	0,36	2		
			1		1

Proprietà della distribuzione della media campionaria

- Per qualsiasi distribuzione $f(x)$ nella popolazione, il valore atteso e la varianza della *media campionaria* sono pari a:

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

dove $\mu = E(X)$ è la media di X nella popolazione e $\sigma^2 = V(X)$ è la sua varianza.

Proprietà della distribuzione della varianza campionaria

- Per qualsiasi distribuzione $f(x)$ nella popolazione, il valore atteso della *varianza campionaria* è:

$$E(S^2) = \sigma^2$$

- Esempio: si riprenda la variabile *numero di figli per famiglia*

X	f(x)	X·f(x)	(X-μ)²·f(x)
0	0,1	0	0,225
1	0,3	0,3	0,075
2	0,6	1,2	0,15
	1	1,5	0,45

da cui la *media* e la *varianza* di X sono $\mu = 1,5$ e $\sigma^2 = 0,45$.

- *Media e varianza campionaria* per ognuno dei 9 campioni

X₁	X₂	f(x₁,x₂)	$\bar{X}$	S^2
0	0	0,01	0	0
0	1	0,03	0,5	0,5
0	2	0,06	1	2
1	0	0,03	0,5	0,5
1	1	0,09	1	0
1	2	0,18	1,5	0,5
2	0	0,06	1	2
2	1	0,18	1,5	0,5
2	2	0,36	2	0
		1		

- Distribuzione della *media campionaria*:

$\bar{X}$	$f(\bar{x})$	$\bar{X} \cdot f(\bar{x})$	$[\bar{X} - E(\bar{X})]^2 \cdot f(\bar{x})$
0	0,01	0	0,0225
0,5	0,06	0,03	0,06
1	0,21	0,21	0,0525
1,5	0,36	0,54	0
2	0,36	0,72	0,09
	1	1,5	0,225

- Si ha quindi

$$E(\bar{X}) = 1,5 = \mu \quad e \quad Var(\bar{X}) = 0,225 = \frac{\sigma^2}{n}$$

- Distribuzione della *varianza campionaria*:

S^2	$f(s^2)$	$S^2 \cdot f(s^2)$
0	0,46	0
0,5	0,42	0,21
2	0,12	0,24
	1	0,45

- Si ha quindi

$$E(S^2) = 0,45 = \sigma^2$$

Distribuzione della media campionaria in alcuni casi particolari

- Se X ha distribuzione *Bernoulliana*, $X \sim \text{Bin}(1, p)$,

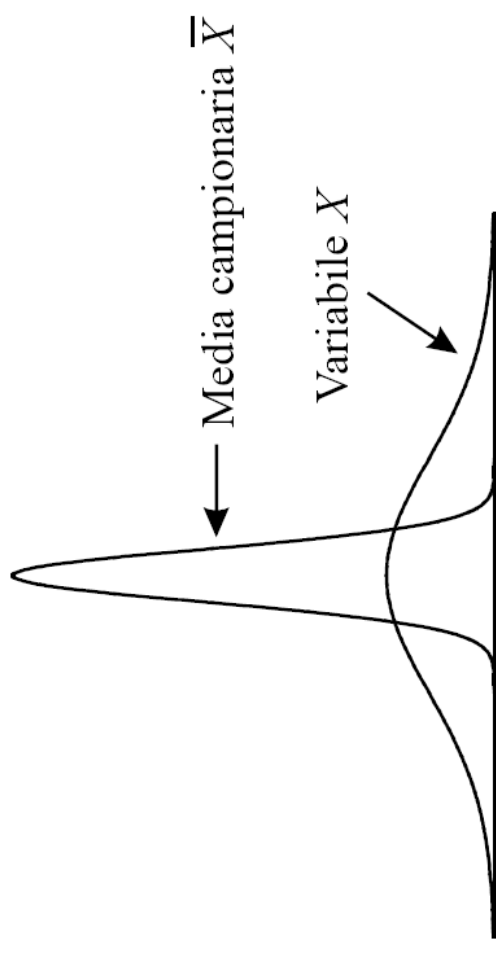
$$P(\bar{X} = \bar{x}) = P(n\bar{X} = n\bar{x}) = \binom{n}{n\bar{x}} p^{n\bar{x}} (1-p)^{n-n\bar{x}}$$

- Se X ha distribuzione *Normale*, $X \sim N(\mu, \sigma^2)$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Nel caso di *grandi campioni* ($n \geq 30$), a prescindere dalla distribuzione di X , approssimativamente

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$



Esempio

- Supponiamo che X abbia distribuzione Bernoulliana, $X \sim \text{Bin}(1, p)$, e che da tale popolazione si estraggano campioni di dimensione $n = 5$.

- Se p fosse pari a 0,3, la distribuzione di $\bar{X}$ sarebbe

$\bar{x}$	$n\bar{x}$	$f(\bar{x})$
0,0	0	0,168
0,2	1	0,360
0,4	2	0,309
0,6	3	0,132
0,8	4	0,028
1,0	5	0,002
		1,000

$$P(\bar{X} = 0) = P(5\bar{X} = 0) = \binom{5}{0} 0,3^0 0,7^5 = 0,168$$

$$P(\bar{X} = 0,2) = P(5\bar{X} = 1) = \binom{5}{1} 0,3^1 0,7^4 = 0,360$$

Distribuzione della varianza campionaria in alcuni casi particolari

- Se X ha distribuzione *Normale*, $X \sim N(\mu, \sigma^2)$

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

- La distribuzione $\chi^2(r)$ è chiamata *chi-quadro* con r gradi di libertà (parametro della distribuzione).
- La *media* e la *varianza* di una v.a. $Y \sim \chi^2(r)$ sono pari a

$$E(Y) = r \quad e \quad V(Y) = 2r$$

- Per il calcolo dei *centili* si fa ricorso ad opportune tavole che riportano i valori χ^2_α tali che

$$P(Y \geq \chi^2_\alpha) = \alpha$$

Esempio

- Si supponga che X abbia nella popolazione distribuzione *Normale*

$$X \sim N(10, 20)$$

e che da tale popolazione vengano estratti campioni di dimensione $n = 25$.

- La *media campionaria* ha distribuzione

$$\bar{X} \sim N(10, 0,8)$$

- Per la *varianza campionaria* si ha

$$\frac{24 \cdot S^2}{20} \sim \chi^2(24)$$

da cui il centile corrispondente a $\alpha=0,05$ è pari a 36,42, cioè

$$P\left(\frac{24 \cdot S^2}{20} \geq 36,42\right) = 0,05;$$

il centile corrispondente a $\alpha=0,99$ è pari a 10,86, cioè

$$P\left(\frac{24 \cdot S^2}{20} \geq 10,86\right) = 0,99.$$

Media campionaria standardizzata

- In alcuni casi si farà uso della *media campionaria standardizzata*

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$$

- Se X ha distribuzione *Normale*, $X \sim N(\mu, \sigma^2)$, si ha

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \quad (\text{normale standardizzata})$$

- Usualmente la varianza della popolazione (σ^2) non è nota e quindi si utilizza al suo posto la varianza campionaria (S^2)

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

- Se X ha distribuzione *Normale*, $X \sim N(\mu, \sigma^2)$, si ha

$$T \approx t(n-1)$$

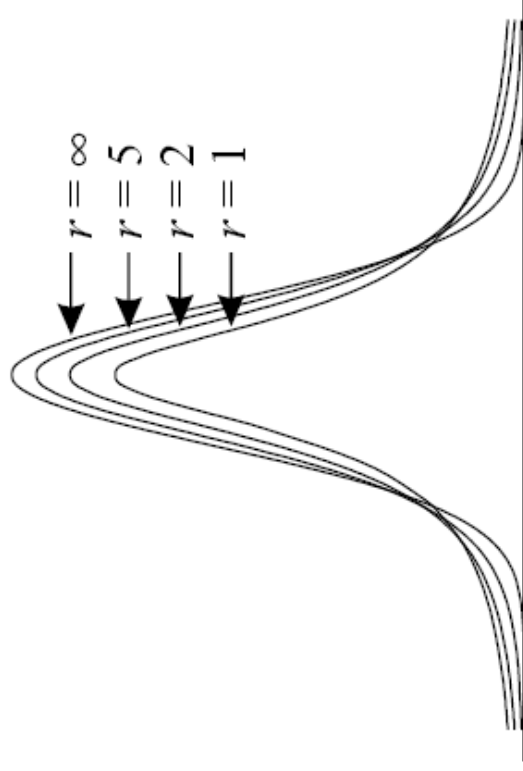
- La distribuzione $t(r)$ è chiamata *t di Student* con r gradi di libertà.

- La distribuzione *t di Student* è simmetrica intorno allo 0. La *media* e la *varianza* di una v.a. $Y \sim t(r)$ sono pari a
- Per il calcolo dei *centili* si fa ricorso ad opportune tavole che riportano i valori t_α tali che
- Quando r è elevato (≥ 100) la distribuzione *t di Student* è approssimabile con la Normale standardizzata.

$$E(Y) = 0 \quad V(Y) = r/(r-2)$$

$$P(Y \geq t_\alpha) = P(Y \leq -t_\alpha) = \alpha$$

Funzione di densità



Esempio

- Si supponga che X abbia nella popolazione distribuzione *Normale*

$$X \sim N(10, 20)$$

e che da tale popolazione vengano estratti campioni di dimensione $n = 25$.

- La *media campionaria* standardizzata (con S^2) ha distribuzione

$$T \sim t(24)$$

- Il centile corrispondente a $\alpha = 0,05$ è $t_{\alpha} = 1,711$, cioè

$$P(T \geq 1,711) = 0,05,$$

mentre quello della normale standardizzata è 1,645.

- Se il campione ha dimensione $n = 121$,

$$T \sim t(120);$$

in questo caso il centile corrispondente a $\alpha = 0,05$ è $t_{\alpha} = 1,658$ che è molto più vicino a quello della Normale standardizzata.

Stima puntuale

- **Stimare:** attribuire un valore plausibile a un parametro incognito.
- **Stimatore del parametro** θ : ogni statistica
$$T = t(X_1, X_2, \dots, X_n)$$
utilizzata per stimare θ . Come ogni altra statistica, lo stimatore è una variabile aleatoria in quanto funzione della variabile aleatoria multipla $(X_1, X_2, \dots, X_n)$ che rappresenta il campione.
- **Stima:** la singola determinazione dello stimatore,

$$t = t(x_1, x_2, \dots, x_n)$$

- cioè il valore che lo stimatore assume in corrispondenza di un campione osservato $(x_1, x_2, \dots, x_n)$.
- Per stimare un parametro θ possiamo utilizzare *diversi stimatori* (es. per stimare la media della popolazione potremmo usare la mediana, anziché la media, campionaria). E' allora importante trovare dei *criteri per scegliere* lo stimatore più appropriato.

Criteri di valutazione di uno stimatore

- Uno dei criteri principali è l'**Errore quadratico medio** (MSE) che per lo stimatore

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

di θ è definito come

$$MSE(T) = E[(T - \theta)^2]$$

- *Esempio:* si assuma che $X \sim \text{Bin}(1, p)$ e che i campioni hanno dimensione $n = 5$. Se $p = 0,3$ si ha

$\bar{x}$	$f(\bar{x})$	$(\bar{x} - 0,3)^2$	$(\bar{x} - 0,3)^2 f(\bar{x})$
0,0	0,168	0,09	0,0151
0,2	0,360	0,01	0,0036
0,4	0,309	0,01	0,0031
0,6	0,132	0,09	0,0119
0,8	0,028	0,25	0,0071
1,0	0,002	0,49	0,0012
1,000			0,0420

$MSE(\bar{X}) = 0,042$

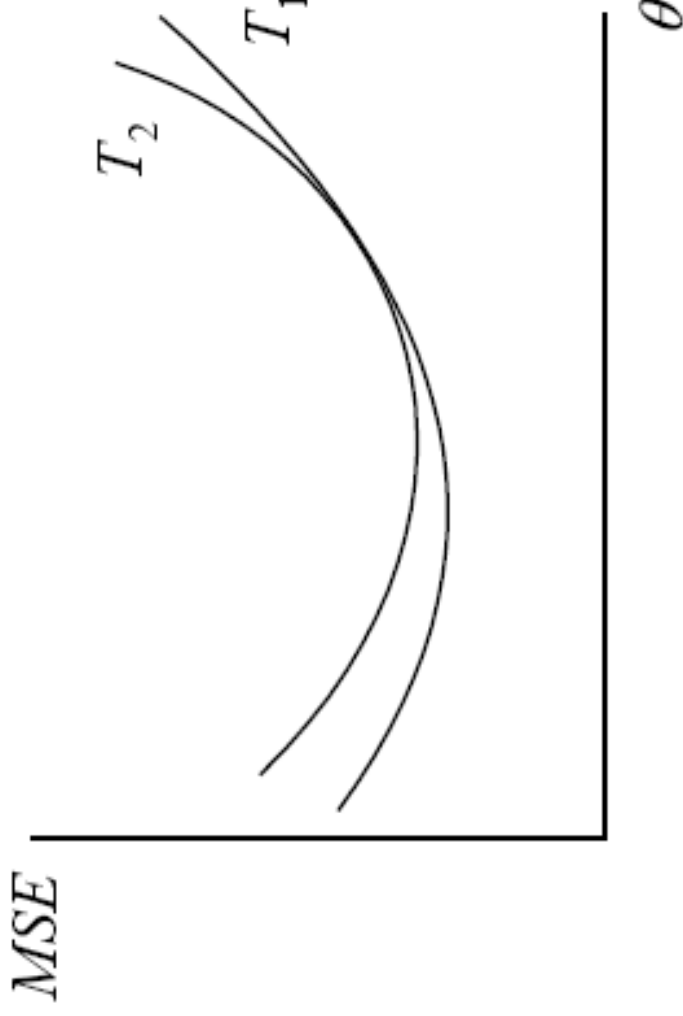
- MSE è una funzione di θ che misura l'errore che si commette in media utilizzando T per stimare θ . Tra due stimatori è preferibile quello che ha *uniformemente* (per ogni θ) l'MSE più piccolo.

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

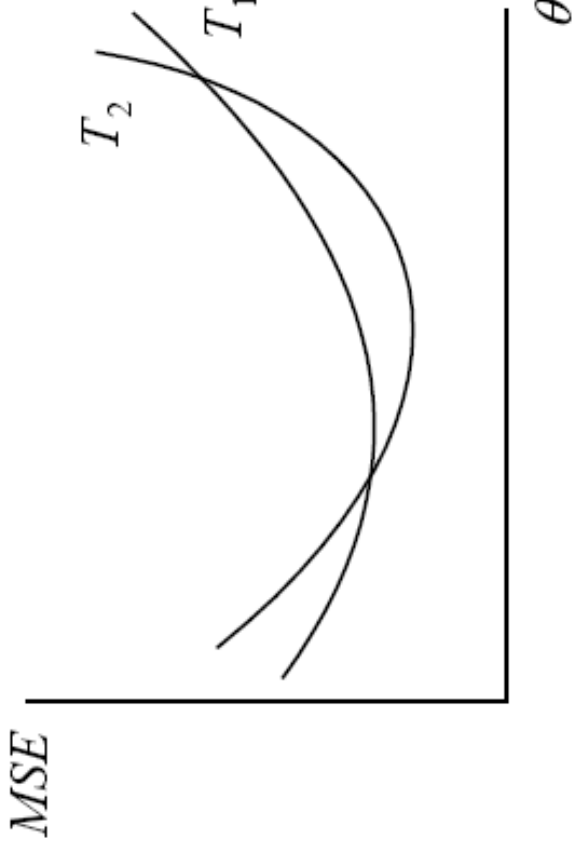
- Dati due stimatori T_1 e T_2 di θ , T_1 è **più efficiente** di T_2 se

$$MSE_{\theta}(T_1) \leq MSE_{\theta}(T_2), \quad \forall \theta$$

e la disuguaglianza vale in senso stretto per almeno un valore di θ .



- In generale non esiste uno stimatore con MSE *uniformemente minimo* quindi usualmente ci si restringe alla classe degli stimatori non distorti.



- Lo stimatore $T = t(X_1, X_2, \dots, X_n)$ di θ si dice **non distorto** se

$$E_{\theta}(T) = \theta, \quad \forall \theta$$
- Si chiama **distorsione** (o *bias*) di uno stimatore T di θ :

$$B_{\theta}(T) = E_{\theta}(T) - \theta$$
- La distorsione può essere positiva (in media T sovrastima θ) o negativa (in media T sottostima θ).

- L'errore quadratico medio può essere decomposto come segue:

$$MSE(T) = Var(T) + [B(T)]^2$$

- Se T è uno stimatore non distorto, si ha

$$MSE(T) = Var(T)$$

- Dati due stimatori non distorti T_1 e T_2 di θ , T_1 è **più efficiente** di T_2 se

$$Var_{\theta}(T_1) \leq Var_{\theta}(T_2), \quad \forall \theta$$

e la disuguaglianza vale in senso stretto per almeno un valore di θ .

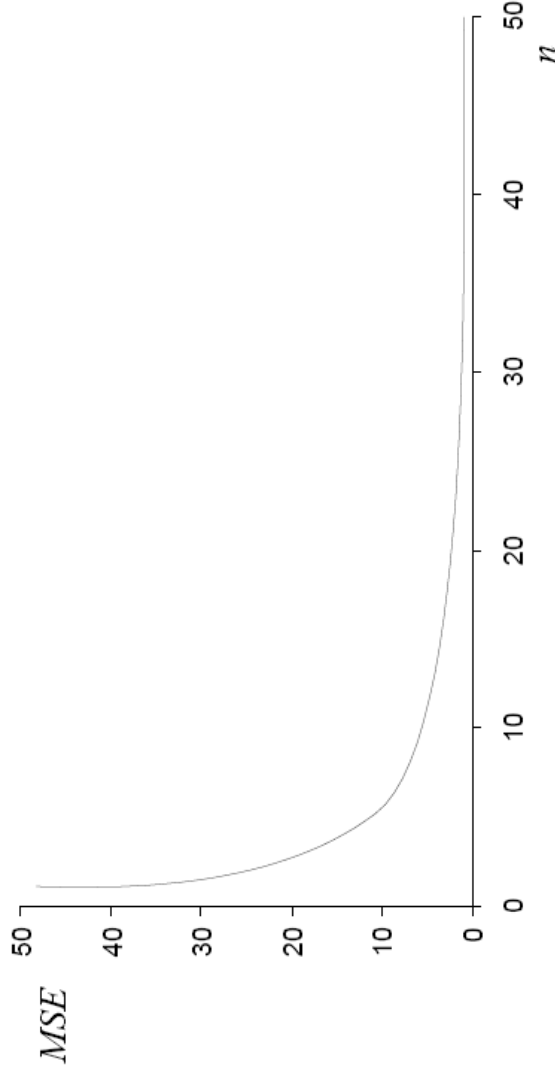
Esempio

- A prescindere dalla popolazione di riferimento, la *media campionaria* è uno stimatore non distorto della media della popolazione:
- Di conseguenza l'errore quadratico medio coincide con la varianza:

$$E(\bar{X}) = \mu$$
$$MSE(\bar{X}) = Var(\bar{X}) = \frac{\sigma^2}{n}$$

e quindi diminuisce (la sua precisione aumenta) al crescere della numerosità del campione.

- Ad esempio, se la varianza della popolazione è $\sigma^2 = 48,5$ si ha



Esempio

- Si consideri lo *stimatore della varianza*
- Questo stimatore è proporzionale alla varianza campionaria

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

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2$$

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

e la sua distorsione è pari a

$$B(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} < 0$$

Lo stimatore *sottostima* la varianza della popolazione. La distorsione tende a 0 al crescere della dimensione campionaria.

- L'errore quadratico medio di questo stimatore non coincide con la sua varianza.

Proprietà asintotiche di uno stimatore

- Le *proprietà asintotiche* riguardano il comportamento di uno stimatore T quando la dimensione del campione tende a infinito.
- In questo caso si usa la notazione T_n per indicare che lo stimatore è applicato a un campione di dimensione n .
- Uno stimatore T_n di θ è **consistente** se per ogni $\varepsilon > 0$
$$\lim_{n \rightarrow \infty} P(|T_n - \theta| \leq \varepsilon) = 1, \quad \forall \theta$$
- Uno stimatore T_n di θ è **consistente in media quadratica** se

$$\lim_{n \rightarrow \infty} E[(T_n - \theta)^2] = 0, \quad \forall \theta$$

La consistenza in media quadratica implica quella semplice.

- Uno stimatore T_n di θ è **asintoticamente non distorto** se

$$\lim_{n \rightarrow \infty} B(T_n) = 0, \quad \forall \theta$$

Metodi per la determinazione di uno stimatore

- Metodo della massima verosimiglianza
- Metodo dei momenti
- Metodo dei minimi quadrati

Metodo della massima verosimiglianza

- Dato un campione $(x_1, x_2, \dots, x_n)$ proveniente da $X \sim f(x; \theta)$, si definisce **funzione di verosimiglianza** la probabilità (densità) del campione interpretata come funzione del parametro θ ,

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \dots \cdot f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

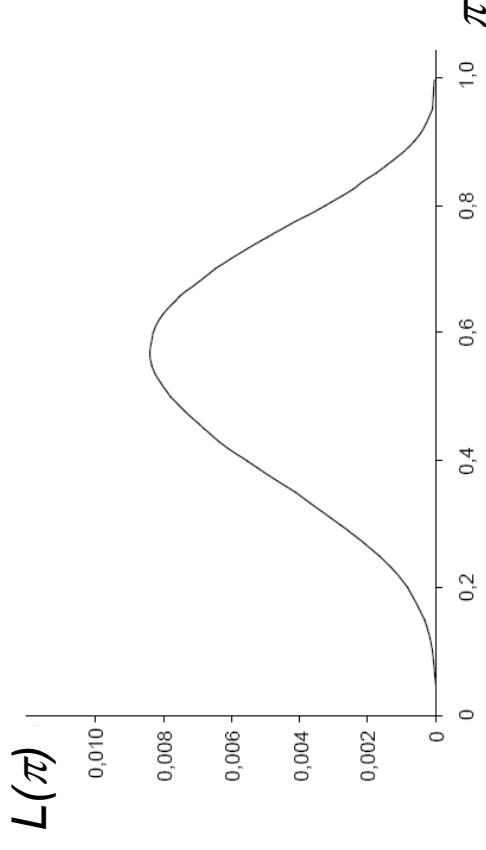
- Quanto più è elevata la funzione di verosimiglianza per un certo valore di θ , tanto più questo valore è plausibile. Ciò permette di stabilire qual è il valore più plausibile del parametro.

Esempio

- Si consideri un campione $(x_1, x_2, \dots, x_n)$ estratto da una popolazione *Bernoulliana* con parametro π , cioè $X \sim \text{Bin}(1, \pi)$.
- La *funzione di verosimiglianza* è pari a

$$L(\pi) = \prod_{i=1}^n \pi^{x_i} (1 - \pi)^{1-x_i} = \pi^{\sum_{i=1}^n x_i} (1 - \pi)^{\sum_{i=1}^n (1-x_i)} = \pi^s (1 - \pi)^{n-s}$$

- dove $s = \sum_{i=1}^n x_i$ è il numero di successi e $n-s$ di insuccessi
- Se il campione osservato è $(1, 0, 1, 0, 0, 1, 1)$ si ha $n = 7$, $s = 4$ e $n-s = 3$ da cui $L(\pi) = \pi^4(1 - \pi)^3$.



- La *stima di massima verosimiglianza* è il valore di θ che massimizza la funzione di verosimiglianza, cioè

$$\hat{\theta} = t(x_1, x_2, \dots, x_n) \quad \text{tale che} \quad L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

- Lo *stimateur di massima verosimiglianza*, come ogni altro *stimateur* è una variabile casuale, ed è definito come:

$$t(X_1, X_2, \dots, X_n)$$

- Generalmente risulta conveniente massimizzare la *log-verosimiglianza*

$$l(\theta) = \log[L(\theta)]$$

risolvendo l'equazione

$$\frac{d l(\theta)}{d \theta} = 0$$

e controllando che in corrispondenza del valore trovato di θ la derivata seconda di $l(\theta)$ sia negativa.

Esempio

- Nel caso della popolazione Bernoulliana $X \sim \text{Bin}(1, \pi)$, la funzione di log-verosimiglianza è pari a

$$l(\pi) = \log[L(\pi)] = s \log(\pi) + (n - s) \log(1 - \pi)$$

- La derivata prima di $l(\pi)$ rispetto a π è

$$\frac{d l(\pi)}{d \pi} = \frac{s}{\pi} - \frac{n - s}{1 - \pi}$$

che è pari a 0 quando $\pi = \hat{\pi}$, dove

$$\hat{\pi} = \frac{s}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

è la stima di massima verosimiglianza di π . Si noti che questa coincide con la media campionaria. Quindi lo *stimatore di massima verosimiglianza* di π è:

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

- Per il campione osservato (1,0,1,0,0,1,1), si ha $\hat{\pi} = 4/7 = 0,571$

Esempio

- Si assume che il campione è estratto da una *popolazione Normale* con media e varianza incognite, $X \sim N(\mu, \sigma^2)$.
- Lo *stimatore di massima verosimiglianza* della media, μ , coincide con la media campionaria

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

- Lo *stimatore di massima verosimiglianza* della varianza, σ^2 , è dato da

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

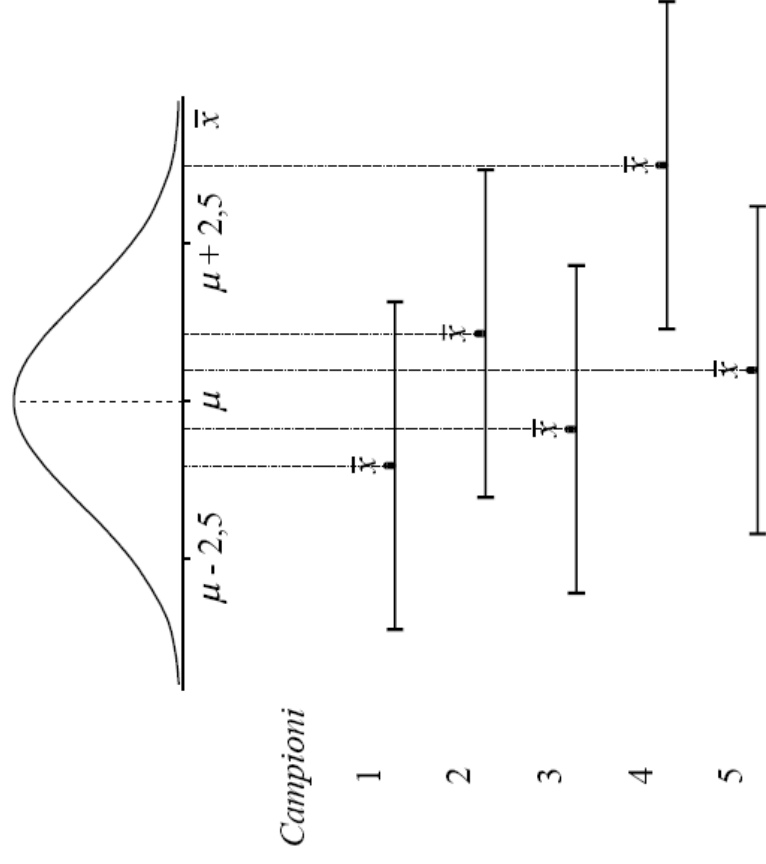
Proprietà di uno stimatore di massima verosimiglianza

- Se $\hat{\theta}$ è lo stimatore di massima verosimiglianza di θ , allora $g(\hat{\theta})$ è lo stimatore di massima verosimiglianza di $g(\theta)$. Ad esempio lo stimatore di σ nel caso di popolazione Normale è $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.
- In genere permette di ottenere *stimatori efficienti* nella classe di quelli non distorti.
- Sotto certe condizioni di regolarità, lo stimatore di massima verosimiglianza gode delle seguenti *proprietà asintotiche*:
 - è *consistente*;
 - è *asintoticamente efficiente*;
 - ha *distribuzione asintotica normale*.

Stima per intervallo

- In ogni stima è ovviamente insito un certo *margin di errore*, ad esempio la probabilità $P(\bar{X} = \mu)$ è nulla quando X è continua.
- Nell'ambito della *stima per intervallo* un parametro viene stimato tramite un *intervallo di confidenza* anziché con un singolo valore come nell'ambito della *stima puntuale*.

• Esempio: Se $X \sim N(\mu, 44)$, $n = 27$ e si usano intervalli del tipo $(\bar{X} - 2,5; \bar{X} + 2,5)$ anziché $\bar{X}$ per stimare μ , si ottiene



Intervallo di confidenza

- Dato un campione $(X_1, X_2, \dots, X_n)$ estratto da una popolazione di cui interessa stimare il parametro θ , e date due statistiche campionarie $L_1 = L_1(X_1, X_2, \dots, X_n)$ e $L_2 = L_2(X_1, X_2, \dots, X_n)$ con $L_1 \leq L_2$ per ogni possibile campione, l'intervallo (L_1, L_2) è un *intervallo di confidenza* per θ al $(1-\alpha)100\%$ se

$$P(L_1 \leq \theta \leq L_2) = 1 - \alpha$$

- Il *coefficiente fiduciario* $1-\alpha$ è solitamente pari al 90%, 95% o 99%.
- L'*ampiezza di un intervallo di confidenza* è data da

$$A = L_2 - L_1$$

Intervalli di confidenza per la media

(popolazione normale, varianza nota)

- Se la popolazione ha *distribuzione normale*, $X \sim N(\mu, \sigma^2)$, si ha

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

- Di conseguenza

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

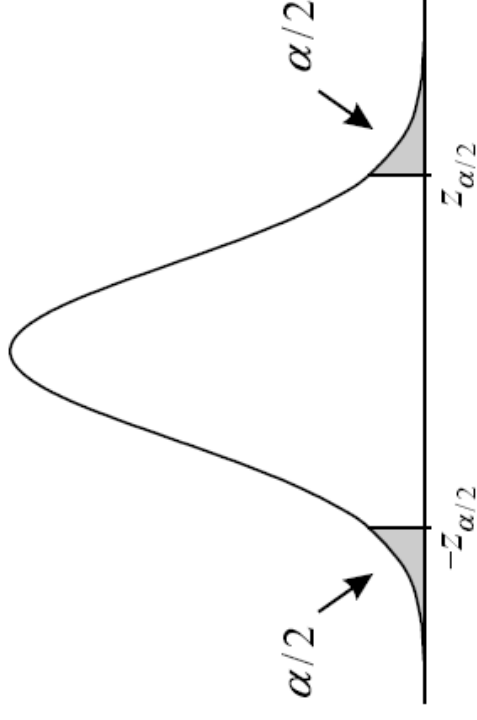
dove $z_{\alpha/2}$ è il quantile tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$ e $Z \sim N(0, 1)$

- L'*intervallo di confidenza* per μ al $(1-\alpha)100\%$ è dato da

$$\left(\bar{X} - z_{\alpha/2} \sqrt{\sigma^2/n}, \bar{X} + z_{\alpha/2} \sqrt{\sigma^2/n}\right)$$

la cui ampiezza è

$$A = 2z_{\alpha/2} \sqrt{\sigma^2/n}$$



Esempio

- Si supponga di voler costruire l'*intervallo di confidenza* al 95% per μ sulla base del campione

170,75	186,14	173,39	185,12	173,39
--------	--------	--------	--------	--------

sotto l'assunzione $X \sim N(\mu, \sigma^2)$, con varianza $\sigma^2 = 50$.

- La *media campionaria* è pari a $\bar{x} = 177,758$
- Dato che $1-\alpha = 0,95$, si ha che $\alpha = 0,05$ e il corrispondente *quantile* della distribuzione Normale standard è pari a $z_{\alpha/2} = z_{0,025} = 1,96$. Si noti che $\Phi(z_{0,025}) = 0,975 = 1-\alpha/2$.
- L'*intervallo di confidenza* è quindi

$$(177,758 - 1,96\sqrt{50/5}, 177,758 + 1,96\sqrt{50/5}) = (171,560, 183,956)$$

che ha ampiezza pari a

$$A = 2 \cdot 1,96\sqrt{50/5} = 12,396$$

Intervalli di confidenza per la media

(popolazione normale, varianza non nota)

- Se $X \sim N(\mu, \sigma^2)$ ma σ^2 non è nota, si utilizza il risultato

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t(n-1)$$

- Di conseguenza

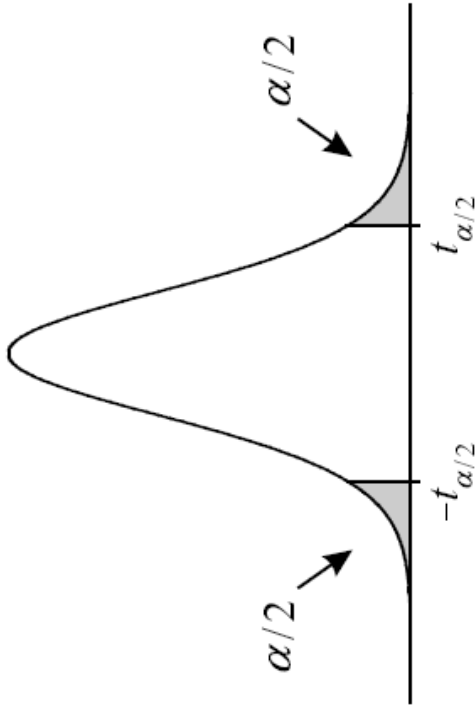
$$P\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

dove $t_{\alpha/2}$ è il quantile tale che $P(T \geq t_{\alpha/2}) = \alpha/2$ e $T \sim t(n-1)$

- L'intervallo di confidenza per μ al $(1-\alpha)100\%$ è dato da

$$\left(\bar{X} - t_{\alpha/2} \sqrt{S^2/n}, \bar{X} + t_{\alpha/2} \sqrt{S^2/n}\right)$$

la cui ampiezza è $A = 2t_{\alpha/2} \sqrt{S^2/n}$



Esempio

- Si supponga di voler costruire l'*intervallo di confidenza* al 95% per μ sulla base del campione

170,75	186,14	173,39	185,12	173,39
--------	--------	--------	--------	--------

sotto l'assunzione $X \sim N(\mu, \sigma^2)$, con varianza σ^2 non nota.

- La *media campionaria* e la *varianza campionaria* sono pari a

$$\bar{x} = 177,758 \quad e \quad s^2 = 52,932$$

- Dato che $1-\alpha = 0,95$, si ha che $\alpha = 0,05$ e il corrispondente *quantile* della distribuzione $t(4)$ è pari a $t_{\alpha/2} = t_{0,025} = 2,776$.

- L'*intervallo di confidenza* è quindi

$$(177,758 - 2,776\sqrt{52,932/5}, 177,758 + 2,776\sqrt{52,932/5}) = (168,726, 186,790)$$

che ha ampiezza pari a $A = 2 \cdot 2,776\sqrt{52,932/5} = 18,064$
maggiore di quella calcolata per il caso della varianza nota.

Intervalli di confidenza per la media

(popolazione qualsiasi, varianza non nota, grandi campioni)

- Per una *popolazione qualsiasi*, nel caso di *grandi campioni* ($n \geq 30$) si ha che

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

ha approssimativamente distribuzione Normale standard.

- Di conseguenza

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

dove $z_{\alpha/2}$ è il quantile tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$ e $Z \sim N(0, 1)$

- L'*intervallo di confidenza* per μ al $(1-\alpha)100\%$ è dato da

$$\left(\bar{X} - z_{\alpha/2} \sqrt{S^2/n}, \bar{X} + z_{\alpha/2} \sqrt{S^2/n}\right)$$

la cui ampiezza è $A = 2z_{\alpha/2} \sqrt{S^2/n}$

Esempio

- Si supponga di voler costruire l'*intervallo di confidenza* al 90% per μ sulla base di un campione di dimensione $n = 100$ che ha media e varianza campionarie pari rispettivamente a

$$\bar{x} = 15,52 \quad e \quad s^2 = 24,47$$

- Dato che $1-\alpha = 0,90$, si ha che $\alpha = 0,10$ e il corrispondente *quantile* della distribuzione Normale standard è pari a $z_{\alpha/2} = z_{0,05} = 1,645$. Si noti che $\Phi(z_{0,05}) = 0,95 = 1-\alpha/2$.
- L'*intervallo di confidenza* è quindi

$$(15,52 - 1,645\sqrt{24,47/100}, 15,52 + 1,645\sqrt{24,47/100}) = (14,706, 16,334)$$

che ha ampiezza pari a

$$A = 2 \cdot 1,645\sqrt{24,47/100} = 1,627$$

Intervalli di confidenza per la media

(popolazione Bernoulliana, grandi campioni)

- Per una *popolazione Bernoulliana*, nel caso di *grandi campioni* ($n \geq 30$) si ha che

$$\frac{\bar{X} - \pi}{\sqrt{\bar{X}(1 - \bar{X})}/n}$$

ha approssimativamente distribuzione Normale standard.

- Di conseguenza

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \pi}{\sqrt{\bar{X}(1 - \bar{X})}/n} \leq z_{\alpha/2}\right) = 1 - \alpha$$

dove $z_{\alpha/2}$ è il quantile tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$ e $Z \sim N(0, 1)$

- L'*intervallo di confidenza* per π al $(1 - \alpha)100\%$ è dato da

$$\left(\bar{X} - z_{\alpha/2} \sqrt{\bar{X}(1 - \bar{X})}/n, \bar{X} + z_{\alpha/2} \sqrt{\bar{X}(1 - \bar{X})}/n\right)$$

la cui ampiezza è

$$A = 2z_{\alpha/2} \sqrt{\bar{X}(1 - \bar{X})}/n$$

Esempio

- Si supponga di voler costruire l'*intervallo di confidenza* al 99% per π sulla base di un campione di dimensione $n = 50$ in cui ci sono $\sum_{i=1}^n x_i = 32$ successi, da cui

$$\bar{x} = 32/50 = 0,64 \quad e \quad \bar{x}(1 - \bar{x}) = 0,64 \cdot (1 - 0,64) = 0,23$$

- Dato che $1 - \alpha = 0,99$, si ha che $\alpha = 0,01$ e il corrispondente *quantile* della distribuzione Normale standard è pari a $z_{\alpha/2} = z_{0,005} = 2,576$. Si noti che $\Phi(z_{0,005}) = 0,995 = 1 - \alpha/2$.
- L'*intervallo di confidenza* è quindi

$$(0,64 - 2,576\sqrt{0,23/50}, 0,64 + 2,576\sqrt{0,23/50}) = (0,465, 0,815)$$

che ha ampiezza pari a

$$A = 2 \cdot 2,576\sqrt{0,23/50} = 0,350$$

Intervalli di confidenza per la varianza

(popolazione normale, media e varianza non note)

- Se $X \sim N(\mu, \sigma^2)$ con μ non noto, si ha

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2(n-1)$$

- Di conseguenza

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

dove $\chi^2_{1-\alpha/2}$ e $\chi^2_{\alpha/2}$ sono i quantili della distribuzione $\chi^2(n-1)$ tali che

$$P(\chi^2 \geq \chi^2_{1-\alpha/2}) = 1 - \alpha/2 \quad \text{e} \quad P(\chi^2 \geq \chi^2_{\alpha/2}) = \alpha/2$$

- L'intervallo di confidenza per σ^2 al $(1-\alpha)100\%$ è dato da

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

Esempio

- Si supponga di voler costruire l'*intervallo di confidenza* al 90% per σ^2 sulla base del campione (già visto in precedenza)

170,75	186,14	173,39	185,12	173,39
--------	--------	--------	--------	--------

sotto l'assunzione $X \sim N(\mu, \sigma^2)$ con μ non noto.

- La *varianza campionaria* è pari a

$$s^2 = 52,932;$$

dato che $1-\alpha = 0,90$, si ha che $\alpha = 0,10$ e i corrispondenti *quantili* della distribuzione $\chi^2(4)$ sono pari a

$$\chi^2_{1-\alpha/2} = \chi^2_{0,95} = 0,71 \quad e \quad \chi^2_{\alpha/2} = \chi^2_{0,05} = 9,49$$

- L'*intervallo di confidenza* è quindi

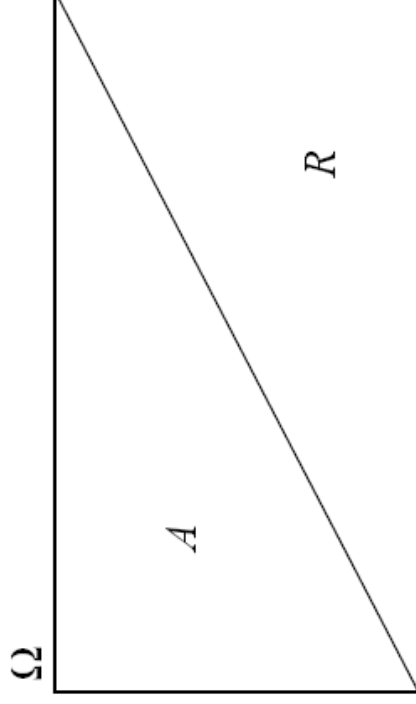
$$(4/9,49 \cdot 52,932, 4/0,71 \cdot 52,932) = (22,311, 298,208)$$

che ha ampiezza pari a $A = 298,208 - 22,311 = 275,897$

Verifica delle ipotesi

- Un'*ipotesi statistica* è un'affermazione o una congettura riguardante un parametro θ che caratterizza il modello descrittivo della popolazione, $f(x;\theta)$, con $\theta \in \Theta$, dove Θ è lo spazio parametrico.
- Solitamente, oltre a un'*ipotesi nulla*,
$$H_0 : \theta \in \Theta_0 \subset \Theta$$
si formula un'*ipotesi alternativa* (contrapposta a quella nulla),
$$H_1 : \theta \in \overline{\Theta_0} \subset \Theta$$
- Un'*ipotesi* si dice *semplice* se specifica completamente la distribuzione, cioè $\theta = \theta_0$; altrimenti si dice *composta*.

- La **verifica delle ipotesi** (**test statistico**) è un procedimento con il quale si decide, alla luce di un campione, se rifiutare o meno H_0 in favore di H_1 . Si tratta di dividere lo spazio campionario (Ω) in due sottoinsiemi: *zona di accettazione* (A) e *zona di rifiuto* (R).



Esempio

- Un'impresa vuole verificare, sulla base di un campione casuale di 50 unità, se il peso medio dei prodotti che le sono stati consegnati è effettivamente pari a 1,2Kg., come dichiarato dal fornitore.
- Sotto l'assunzione che il peso dei prodotti abbia distribuzione $X \sim N(\mu, \sigma^2)$, l'*ipotesi nulla* (che in questo caso è *semplice*) è

$$H_0 : \mu = 1,2$$

Siccome l'impresa viene danneggiata solo se il peso è minore di quanto dichiarato, si pone come *ipotesi alternativa* (*composta*)

$$H_1 : \mu < 1,2$$

- Un modo naturale di condurre la *verifica delle ipotesi* è rifiutare H_0 in favore di H_1 quando la media campionaria è significativamente più piccola della media dichiarata, cioè quando

$$\bar{x} \leq 1,2 - \varepsilon, \quad \varepsilon > 0$$

La zona di accettazione e la zona di rifiuto sono pari a:

$$A = \{\mathbf{x} : \bar{x} > 1,2 - \varepsilon\}, \quad R = \{\mathbf{x} : \bar{x} \leq 1,2 - \varepsilon\}$$

- Se l'*ipotesi alternativa (composta)* è $H_1 : \mu \neq 1,2$, è sensato rifiutare quando

$$|\bar{x} - 1,2| \geq \varepsilon, \quad \varepsilon > 0$$

La zona di accettazione e la zona di rifiuto, in questo caso sono pari a:

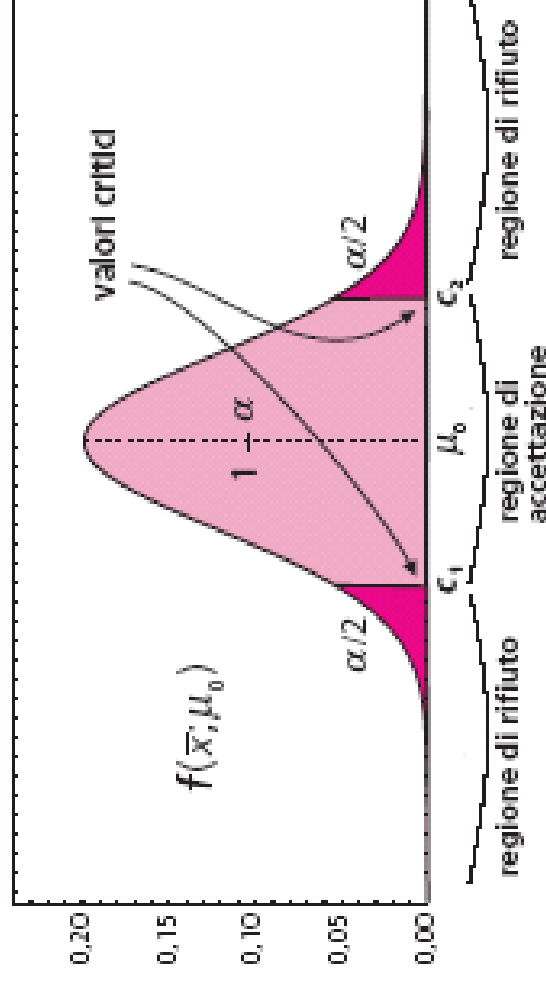
$$A = \{\mathbf{x} : |\bar{x} - 1,2| < \varepsilon\}, \quad R = \{\mathbf{x} : |\bar{x} - 1,2| \geq \varepsilon\}$$

Esempio

Supponiamo che la popolazione sia Normale con media μ incognita e varianza σ^2 nota. Si vuole verificare:

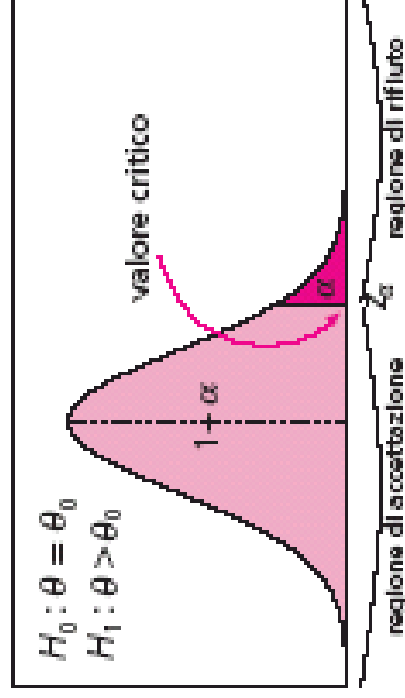
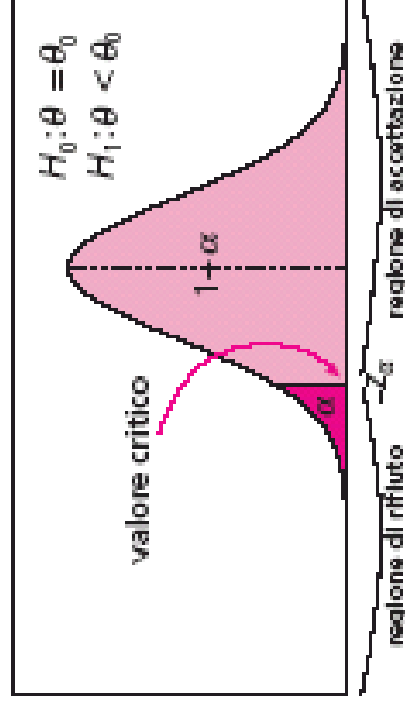
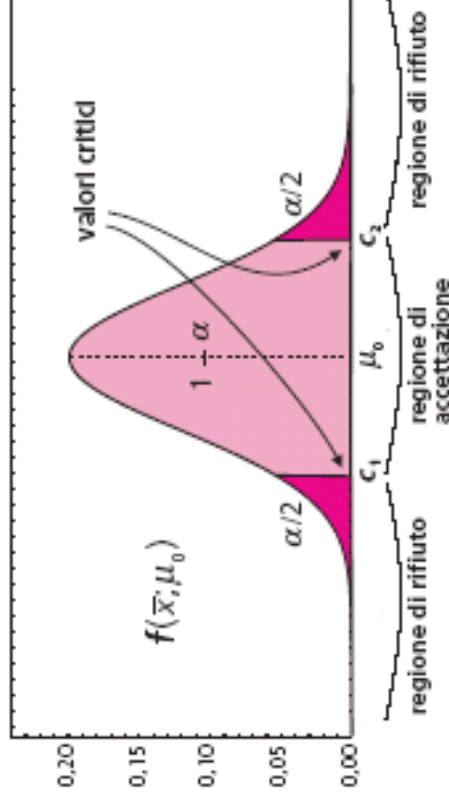
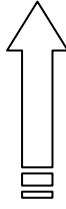
$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

Considerando come statistica test la media campionaria $\bar{X}$ sappiamo che sotto l'ipotesi nulla questa si distribuisce come una Normale con media $\mu = \mu_0$ e varianza σ^2/n .



In corrispondenza dell'ipotesi alternativa si possono configurare diverse regioni di rifiuto:

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$$



Criteri di valutazione di un test statistico

- Nel condurre un test possiamo incorrere in due tipi di errore:
 - *errore di I tipo*: si rifiuta H_0 quando è vera
 - *errore di II tipo*: si accetta H_0 quando è falsa
- Questi errori possono essere commessi con *probabilità*:
 - $\alpha(\theta) = P_\theta(\mathbf{X} \in R), \quad \theta \in \Theta_0$
 - $\beta(\theta) = P_\theta(\mathbf{X} \in A), \quad \theta \in \overline{\Theta}_0$
- La *potenza del test* è la probabilità di rifiutare H_0 quando è falsa,

$$\pi(\theta) = 1 - \beta(\theta) = P_\theta(\mathbf{X} \in R), \quad \theta \in \overline{\Theta}_0$$

	Decisione	
	Accetto H_0	Rifiuto H_0
H_0 è vera	Corretta $1 - \alpha$	Errore del I tipo α
H_0 è falsa	Errore del II tipo β	Corretta $1 - \beta$

- La potenza cresce al crescere della dimensione campionaria e della distanza tra il valore vero del parametro e l'ipotesi nulla.

Esempio

- In relazione all'esempio precedente, si supponga che per verificare

$$\begin{cases} H_0 : \mu = 1,2 \\ H_1 : \mu < 1,2 \end{cases}$$
 si rifiuti quando $\bar{X} \leq 1$.

- Sotto l'assunzione $X \sim N(\mu, \sigma^2)$, la probabilità di rifiutare quando il valore vero della media è μ è data da

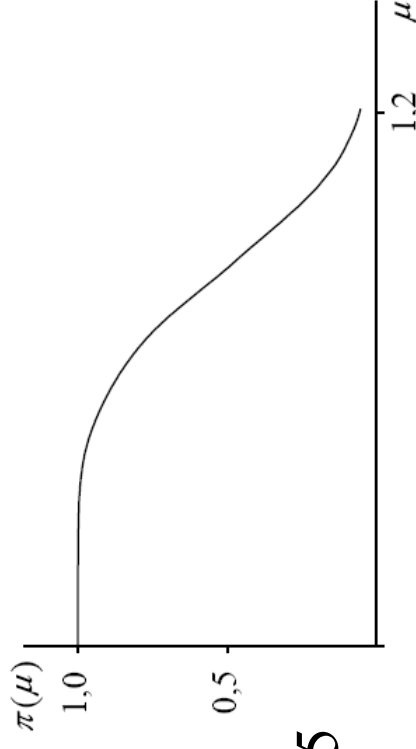
$$P_{\mu}(\bar{X} \leq 1) = P_{\mu} \left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq \frac{1 - \mu}{\sqrt{\sigma^2/n}} \right) = \Phi \left(\frac{1 - \mu}{\sqrt{\sigma^2/n}} \right)$$

- Se $\sigma^2 = 0,5$ e con $n = 50$,
la *probabilità dell'errore di I tipo* è

$$P_{1,2}(\bar{X} \leq 1) = \Phi(-2) = 1 - 0,9772 = 0,0228$$

- Se la media vera fosse $\mu = 1,05$,
la *potenza del test* sarebbe

$$P_{1,05}(\bar{X} \leq 1) = \Phi(-0,5) = 1 - 0,6915 = 0,3085$$



Test uniformemente più potenti e test corretti

- Siccome minimizzare entrambe le probabilità di errore è usualmente impossibile, si usa fissare un livello massimo per quella di I tipo, chiamata *livello di significatività* α (0,1, 0,05, 0,01), e si cerca di minimizzare quella di II tipo (o massimizzare la potenza).
- Un test con funzione di potenza $\pi^*(\theta)$ è uniformemente più potente al livello α se
$$\pi^*(\theta) \geq \pi(\theta), \quad \theta \in \overline{\Theta}_0$$
dove $\pi(\theta)$ è la funzione di potenza di un qualsiasi altro test al livello α .
- In alcune situazioni, il test uniformemente più potente *non esiste* e quindi si ricorre alla definizione di *non distorsione*.
- Un test al livello α è *non distorto* se

$$\pi(\theta) \geq \alpha, \quad \forall \theta \in \overline{\Theta}_0$$

- Un metodo che usualmente porta a test con queste proprietà è quello basato sul *rapporto di verosimiglianze* in cui si rifiuta H_0 se

$$\lambda = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} = \frac{\max_{\theta \in \Theta_0} L(\theta)}{L(\hat{\theta})} \leq k$$

dove $0 < k < 1$ è scelto in base al livello di significatività voluto.

Verifica di ipotesi sulla media

(popolazione normale, varianza nota)

- Si assume che $X \sim N(\mu, \sigma^2)$, con σ^2 noto.
- Si considera l'ipotesi nulla $H_0: \mu = \mu_0$ e quindi la *statistica test*

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1) \quad (\text{sotto } H_0)$$

- Se $H_1: \mu > \mu_0$, si rifiuta H_0 se

$$z \geq z_\alpha \quad \text{oppure se} \quad \bar{x} \geq \mu_0 + z_\alpha \sigma / \sqrt{n}$$

dove z_α è il quantile della distribuzione Normale standard tale che $P(Z \geq z_\alpha) = \alpha$ con $Z \sim N(0, 1)$.

- Se $H_1: \mu < \mu_0$, si rifiuta H_0 se

$$z \leq -z_\alpha \quad \text{oppure se} \quad \bar{x} \leq \mu_0 - z_\alpha \sigma / \sqrt{n}$$

- Se $H_1: \mu \neq \mu_0$, si rifiuta H_0 se

$$|z| \geq z_{\alpha/2} \quad \text{oppure se} \quad \bar{x} \leq \mu_0 - z_{\alpha/2} \sigma / \sqrt{n} \quad \text{o} \quad \bar{x} \geq \mu_0 + z_{\alpha/2} \sigma / \sqrt{n}$$

Un altro modo per evidenziare il risultato del test è quello di calcolare il **p-value**.

p-value : probabilità di osservare un valore della statistica test uguale o più estremo del valore ottenuto dal campione, sotto l'ipotesi nulla.

E' una quantità che misura l'evidenza fornita dai dati contro l'ipotesi nulla: minore è il valore del p-value, più è forte l'evidenza contro l'ipotesi nulla.

In pratica

se $\text{p-value} < \alpha$  rifiuto H_0

Esempio

- Si supponga di voler verificare l'ipotesi che l'*altezza media degli italiani* è 172cm sulla base del campione

170,75	186,14	173,39	185,12	173,39
--------	--------	--------	--------	--------

sotto l'assunzione $X \sim N(\mu, \sigma^2)$, con varianza $\sigma^2 = 50$.

- L'ipotesi nulla è H_0 : $\mu = 172$ e quindi, dato che $\bar{x} = 177,758$, la statistica test Z assume il valore

$$z = \frac{177,758 - 172}{\sqrt{50/5}} = \frac{5,758}{3,162} = 1,821$$

- Se H_1 : $\mu > \mu_0$ e $\alpha = 0,05$, si rifiuta H_0 in quanto
$$1,821 = z \geq z_\alpha = 1,645$$
- Se H_1 : $\mu \neq \mu_0$ e $\alpha = 0,05$, non si rifiuta H_0 in quanto
$$1,821 = |z| < z_{\alpha/2} = 1,96$$

Verifica di ipotesi sulla media

(popolazione normale, varianza non nota)

- Se $X \sim N(\mu, \sigma^2)$, con σ^2 non noto, e l'ipotesi nulla è $H_0: \mu = \mu_0$, si utilizza la statistica test

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim t(n-1) \quad (\text{sotto } H_0)$$

- Se $H_1: \mu > \mu_0$, si rifiuta H_0 se
$$t \geq t_\alpha$$
- dove t_α è il quantile della distribuzione t di Student tale che $P(T \geq t_\alpha) = \alpha$ con $T \sim t(n-1)$.
- Se $H_1: \mu < \mu_0$, si rifiuta H_0 se
$$t \leq -t_\alpha$$
- Se $H_1: \mu \neq \mu_0$, si rifiuta H_0 se

$$|t| \geq t_{\alpha/2}$$

Esempio

- Nel caso della verifica dell'ipotesi $H_0: \mu = 172$ considerata in precedenza, si supponga di non conoscere la varianza.
- La *media campionaria* e la *varianza campionaria* sono pari a

$$\bar{x} = 177,758 \quad e \quad s^2 = 52,932$$

da cui la statistica test T assume il valore

$$t = \frac{177,758 - 172}{\sqrt{52,932/5}} = \frac{5,758}{3,254} = 1,770$$

- Se $H_1: \mu > \mu_0$ e $\alpha = 0,01$, non si rifiuta H_0 in quanto
$$1,770 = t < t_{\alpha} = 3,747$$
- Se $H_1: \mu \neq \mu_0$ e $\alpha = 0,05$, non si rifiuta H_0 in quanto

$$1,770 = |t| < t_{\alpha/2} = 2,776$$

Verifica di ipotesi sulla media

(popolazione qualsiasi, grandi campioni)

- Per una popolazione qualsiasi, nel caso di grandi campioni ($n \geq 30$), se l'ipotesi nulla è $H_0: \mu = \mu_0$ si può utilizzare la statistica test

$$Z = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim N(0, 1) \quad (\text{sotto } H_0)$$

- Se $H_1: \mu > \mu_0$, si rifiuta H_0 se

$$z \geq z_\alpha \quad \text{oppure se} \quad \bar{x} \geq \mu_0 + z_\alpha S/\sqrt{n}$$

- Se $H_1: \mu < \mu_0$, si rifiuta H_0 se

$$z \leq -z_\alpha \quad \text{oppure se} \quad \bar{x} \leq \mu_0 - z_\alpha S/\sqrt{n}$$

- Se $H_1: \mu \neq \mu_0$, si rifiuta H_0 se

$$|z| \geq z_{\alpha/2} \quad \text{oppure se} \quad \bar{x} \leq \mu_0 - z_{\alpha/2} S/\sqrt{n} \quad \text{o} \quad \bar{x} \geq \mu_0 + z_{\alpha/2} S/\sqrt{n}$$

Esempio

- Si consideri l'ipotesi $H_0: \mu = 1,2$ contro $H_1: \mu < 1,2$
- Si supponga che il campione di $n = 50$ prodotti che è stato estratto ha media e varianza pari rispettivamente a

$$\bar{x} = 1,059 \quad e \quad s^2 = 0,158$$

e quindi la *statistica test* assume il valore

$$z = \frac{1,059 - 1,22}{\sqrt{0,158/50}} = -\frac{0,141}{0,0562} = -2,509$$

- Al livello di significatività $\alpha = 0,01$, si rifiuta H_0 in quanto
 $-2,509 = z \leq -z_\alpha = -2,326$

Verifica di ipotesi sulla media

(determinazione della numerosità campionaria)

Ipotizziamo una dimensione ampia del campione tale da garantire l'applicazione dell'approssimazione alla Normale.

La procedura segue i seguenti passi:

1. specificare il livello di significatività α
2. specificare il valore di μ_1 e il corrispondente valore di β
3. selezionare una stima iniziale di σ
4. calcolare la numerosità campionaria

Sia

z_α il valore per cui $P(Z \geq z_\alpha) = \alpha$

z_β il valore per cui $P(Z < z_\beta) = 1 - \beta$

allora:

$$H_1: \mu \neq \mu_0 \quad n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_1 - \mu_0} \right]^2$$

$$\begin{array}{ll} H_1: \mu > \mu_0 & n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_1 - \mu_0} \right]^2 \\ H_1: \mu < \mu_0 & \end{array}$$

Verifica di ipotesi sulla probabilità di successo (popolazione Bernoulliana, grandi campioni)

- Si assume che $X \sim \text{Bin}(1, \pi)$ (popolazione Bernoulliana).
- Si considera l'ipotesi nulla $H_0: \pi = \pi_0$ e quindi la statistica test

$$Z = \frac{\bar{X} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}/n}$$

che per grandi campioni ($n \geq 30$), ha approssimativamente distribuzione Normale standard $N(0, 1)$, sotto H_0 .

- Se $H_1: \pi > \pi_0$, si rifiuta H_0 se

$$z \geq z_\alpha \quad \text{oppure se} \quad \bar{x} \geq \pi_0 + z_\alpha \sqrt{\pi_0(1 - \pi_0)}/n$$

- Se $H_1: \pi < \pi_0$, si rifiuta H_0 se

$$z \leq -z_\alpha \quad \text{oppure se} \quad \bar{x} \geq \pi_0 - z_\alpha \sqrt{\pi_0(1 - \pi_0)}/n$$

- Se $H_1: \mu \neq \mu_0$, si rifiuta H_0 se

$$|z| \geq z_{\alpha/2} \quad \text{oppure se} \quad \bar{x} \leq \pi_0 - z_{\alpha/2} \sqrt{\pi_0(1 - \pi_0)}/n \quad \text{o} \quad \bar{x} \geq \pi_0 + z_{\alpha/2} \sqrt{\pi_0(1 - \pi_0)}/n$$

Esempio

- Si supponga di voler verificare l'ipotesi $H_0: \pi = 0,52$ contro $H_1: \pi > 0,52$ al livello $\alpha = 0,10$ sulla base di un campione di dimensione $n = 50$ in cui ci sono 32 successi.
- La media campionaria è

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{32}{50} = 0,64$$

e quindi la *statistica test* assume il valore

$$z = \frac{0,64 - 0,52}{\sqrt{0,52(1 - 0,52)/50}} = \frac{0,12}{0,0707} = 1,698$$

- Si rifiuta H_0 in quanto

$$1,698 = z \geq z_{\alpha} = 1,282$$

Verifica di ipotesi sulla varianza

(popolazione normale, media non nota)

- Se $X \sim N(\mu, \sigma^2)$ con μ non noto, e l'ipotesi nulla è $H_0: \sigma^2 = \sigma_0^2$ si utilizza la statistica test

$$C = \frac{(n-1)S^2}{\sigma^2} \approx \chi^2(n-1) \quad (\text{sotto } H_0)$$

- Se $H_1: \sigma^2 > \sigma_0^2$, si rifiuta H_0 se

$$c \geq \chi_\alpha^2 \quad \text{oppure se} \quad s^2 \geq \chi_\alpha^2 \sigma_0^2 / (n-1)$$

- dove χ_α^2 è il quantile della distribuzione $\chi^2(n-1)$ tale che $P(\chi^2 > \chi_\alpha^2) = \alpha$.

- Se $H_1: \sigma^2 < \sigma_0^2$, si rifiuta H_0 se

$$c \leq \chi_{1-\alpha}^2 \quad \text{oppure se} \quad s^2 \leq \chi_{1-\alpha}^2 \sigma_0^2 / (n-1)$$

- Se $H_1: \sigma^2 \neq \sigma_0^2$, si rifiuta H_0 se

$$\{c \leq \chi_{1-\alpha/2}^2 \quad \text{o} \quad c \geq \chi_{\alpha/2}^2\} \quad \text{oppure} \quad \{s^2 \leq \chi_{1-\alpha/2}^2 \sigma_0^2 / (n-1) \quad \text{o} \quad s^2 \geq \chi_{\alpha/2}^2 \sigma_0^2 / (n-1)\}$$

Esempio

- Si supponga di voler verificare l'ipotesi $H_0: \sigma^2 = 75$ contro $H_1: \sigma^2 \neq 75$ al livello $\alpha = 0,10$ sulla base di un campione (già considerato in precedenza)

170,75	186,14	173,39	185,12	173,39
--------	--------	--------	--------	--------

sotto l'assunzione $X \sim N(\mu, \sigma^2)$, con media μ non nota.

- La varianza campionaria è

$$s^2 = 52,932$$

e quindi la *statistica test* assume il valore

$$c = \frac{4 \cdot 52,932}{75} = 2,832$$

- Non si rifiuta H_0 in quanto

$$0,71 = \chi^2_{1-\alpha/2} \leq c \leq \chi^2_{\alpha/2} = 9,49$$