Università degli Studi di Napoli “Parthenope”

Corso di Laurea in Statistica e Informatica per l'Azienda, la Finanza e le Assicurazioni (SIAFA)

STATISTICA
II MODULO

Sergio LONGOBARDI
longobardi@uniparthenope.it

Test d'ipotesi su campioni indipendenti provenienti da due popolazioni

Test d'ipotesi campioni indipendenti

Si è interessati a comparare dei parametri di due popolazioni sulla base dell'osservazione di due campioni indipendenti

TEST

- Medie
- Varianze

Popolazioni con distribuz. normale

- Proporzioni

Popolazione con distribuz. Bernoulliana

Test d'ipotesi campioni indipendenti

Test sulle medie μ_1 e μ_2 ➡ Si assume che:
$X_1 \sim (\mu_1;\sigma^2_1)$ e $X_2 \sim (\mu_2;\sigma^2_2)$

Ipotesi da verificare:

1. $H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 > \mu_2$
2. $H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 < \mu_2$
3. $H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$

Test d'ipotesi campioni indipendenti

Il test viene effettuato considerando la differenza tra le due medie di conseguenza le ipotesi da verificare diventano:

1. H_0: $\mu_1-\mu_2=0$ H_1: $\mu_1-\mu_2>0$
2. H_0: $\mu_1-\mu_2=0$ H_1: $\mu_1-\mu_2<0$
3. H_0: $\mu_1-\mu_2=0$ H_1: $\mu_1-\mu_2\neq0$

Si considera lo ***stimatore differenza*** dato dalla differenza tra le due medie campionarie:

$$\bar{X}_1-\bar{X}_2 \sim N\left(0, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

$$Z=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$$

Test d'ipotesi campioni indipendenti

Test per medie nel caso di Popolazioni normali e VARIANZE NOTE

Statistica test e distribuzione sotto l'ipotesi nulla: H_0: $\mu_1 = \mu_2$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \sim N(0,1)$$

Con σ^2_1 e σ^2_2 che indicano il valore delle varianze nelle due popolazioni e con n_1 e n_2 le dimensioni dei due campioni.

Ipotesi alternativa	*Regione di rifiuto*
$H_1: \ \mu_1 > \mu_2$	$Z \geq z_\alpha$
$H_1: \ \mu_1 < \mu_2$	$Z \leq -z_\alpha$
$H_1: \ \mu_1 \neq \mu_2$	$\lvert Z \rvert \geq z_{\alpha/2}$

Esempio

Test sulle medie di campioni indipendenti *varianze note*

Si vuole verificare l'efficacia di una campagna promozionale condotta da una catena alberghiera. A tal scopo si considerano due campioni di alberghi, con lo stesso numero di camere, il primo campione estratto tra gli alberghi in cui è stata effettuata la campagna promozionale ed il secondo estratto tra quelli in cui non è stata attuata nessuna promozione. Si vuole verificare (con un α=5%) se nel mese in cui è stata attuata la campagna promozionale si è verificato un incremento del numero dei clienti.

Test sulle medie di campioni indip. ***varianze note***

Il primo campione è composto da n_1=15 con una media di clienti pari a 1700, mentre il secondo campione di ampiezza n_2=12 ha registrato un numero medio di clienti paria a 1650. Si suppone che la varianza del numero di clienti è uguale per le due popolazioni ed è pari a 800. Il sistema d'ipotesi da verificare è il seguente:

$H_0: \mu_1 = \mu_2$	$\Rightarrow$	$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 > \mu_2$		$H_1: \mu_1 - \mu_2 > 0$

Si calcola la statistica test

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} = \frac{1700 - 1650}{\sqrt{800/15 + 800/12}} = 4{,}56$$

Per α=0,05 si ottiene che :$z_{0,05}$=1,645

e quindi poiché $Z \geq z_{\alpha}$ **(4,56>1,645)** si rifiuterà l'ipotesi nulla e si conclude che la campagna promozionale ha comportato un incremento dei clienti.

Test d'ipotesi campioni indipendenti

Test per medie nel caso di Popolazioni normali e VARIANZE IGNOTE E UGUALI

Statistica test e distribuzione sotto l'ipotesi nulla: H_0: $\mu_1 = \mu_2$

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2\left(1/n_1 + 1/n_2\right)}} \sim t\text{-}Student\left(n_1 + n_2 - 2\right)$$

Con $S^2{}_p$ che indica lo stimatore congiunto della varianza, ossia:

***Varianza pooled*, media ponderata rispetto ai gradi di libertà degli stimatori corretti**

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Ipotesi alternativa	*Regione di rifiuto*
$H_1: \ \mu_1 > \mu_2$	$T \geq t_\alpha$
$H_1: \ \mu_1 < \mu_2$	$T \leq -t_\alpha$
$H_1: \ \mu_1 \neq \mu_2$	$\lvert T \rvert \geq t_{\alpha/2}$

Test d'ipotesi sulle medie di campioni indipendenti

varianze ignote e uguali

Esercizio tratto da ***Borra, Di Ciaccio*** pag.376-377

Si vuole verificare l'efficacia di un trattamento per il controllo della pressione sanguigna, vengono esaminati due campioni di individui con pressione sanguigna elevata. Il primo gruppo è composto da **n_1=15** individui sottoposti al trattamento mentre il secondo gruppo di **n_2=12** individui è stato trattato con un placebo. Si vuole verificare l'ipotesi

H_0:μ_1=μ_2 → **H_0:μ_1-μ_2=0**

H_1: μ_1<μ_2 → **H_1: μ_1- μ_2 <0**

Sulla base dei seguenti dati campionari:

X_1= (155, 128, 130, 140, 10'6, 158, 130, 150, 145, 135, 128, 146, 122, 130, 135)

X_2= (146, 169, 167, 176, 169, 188, 158, 175, 180, 185, 178, 156)

Test d'ipotesi campioni indipendenti

Si suppone che la varianza del numero di clienti è uguale per le due popolazioni ma è incognita. Si calcolano quindi le medie campionarie e le varianze campionarie corrette:

X_1= (155, 128, 130, 140, 106, 158, 130, 150, 145, 135, 128, 146, 122, 130, 135)

X_2= (146, 169, 167, 176, 169, 188, 158, 175, 180, 185, 178, 156)

Le ***medie campionarie*** risultano 'pari a:

$$\bar{x}_1 = 135{,}87 \qquad \bar{x}_2 = 170{,}58$$

Le ***varianze campionarie corrette*** risultano 'pari a:

$$s_1^2 = 183{,}7 \qquad s_2^2 = 154{,}27$$

La ***stima congiunta della varianza s^2_p*** è pari a:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(15-1)183{,}7 + (12-1)154{,}27}{15+12-2} = 170{,}75$$

Test d'ipotesi sulla media di campioni indip. ***varianze ignote***

Avendo calcolato la varianza congiunta si calcola la statistica test t:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2\left(1/n_1 + 1/n_2\right)}} = \frac{135{,}87 - 170{,}58}{\sqrt{170{,}75\left(\frac{1}{15} + \frac{1}{12}\right)}} = -6{,}86$$

Fissato un α=0,01 con 25 gradi di liberta (g) si ottiene che $-t_{0,01}$=-2,485 e quindi poiché $T \leq -t_{\alpha}$ **(-6,86<-2,48)** si rifiuterà l'ipotesi nulla e si conclude che il trattamento ha comportato una diminuzione della pressione sanguigna.

Test d'ipotesi sulle varianze

La verifica d'ipotesi riguardante l'uguaglianza tra le varianze di popolazioni indipendenti è detta anche ipotesi di omoschedasticità. Analogamente al test sulle medie le ipotesi da verificare sono le seguenti:

Ipotesi da verificare:

1. $H_0: \sigma^2_1 = \sigma^2_2 \qquad H_1: \sigma^2_1 > \sigma^2_2$

2. $H_0: \sigma^2_1 = \sigma^2_2 \qquad H_1: \sigma^2_1 < \sigma^2_2$

3. $H_0: \sigma^2_1 = \sigma^2_2 \qquad H_1: \sigma^2_1 \neq \sigma^2_2$

Test d'ipotesi sulle varianze

Ponendo $\boldsymbol{\sigma^2_R=\sigma^2_1/\sigma^2_2}$ il sistema di ipotesi da testare diventa il seguente

Ipotesi da verificare:

1. H_0: $\sigma^2_R=1$	H_1: $\sigma^2_R>1$
2. H_0: $\sigma^2_R=1$	H_1: $\sigma^2_R< 1$
3. H_0: $\sigma^2_R=1$	H_1: $\sigma^2_R\neq 1$

La statistica test da utilizzare, con l'assunzione di normalità delle due popolazioni, è la seguente:

$$\frac{S_1^2}{S_2^2} \sim f\text{ - Fisher}\left(n_1-1, n_2-1\right)$$

Test d'ipotesi sulle varianze

Dopo aver calcolato la statistica test si procede utilizzando i valori della F di Fischer riportati sulle apposite tavole (Borra di Ciaccio pag.497-501)

Ipotesi alternativa	Regione di rifiuto
$H_1: \ \sigma_1^2 > \sigma_2^2$	$S_1^2 / S_2^2 \geq f_\alpha$
$H_1: \ \sigma_1^2 < \sigma_2^2$	$S_1^2 / S_2^2 \leq f_{1-\alpha}$
$H_1: \ \sigma_1^2 \neq \sigma_2^2$	$S_1^2 / S_2^2 \leq f_{1-\alpha/2}$ $S_1^2 / S_2^2 \geq f_{\alpha/2}$

Per valori della F non riportati sulle tavole si può utilizzare la seguente relazione:

$$f_{1-a/2} = \frac{1}{f^*_{a/2}}$$

Dove $f_{1\text{-}a/2}$ indica il valore critico inferiore di una F di fisher con n_1-1 e n_2-1 gradi di libertà mentre $f^*_{a/2}$ indica il valore critico superiore di una F di fisher con n_2-1 e n_1-1 gradi di libertà (si invertono i gradi di lib.)

Test d’ipotesi sulle varianze

Si vuole verificare la variabilità del prezzo di un genere alimentare in due diversi mercati. Nel primo mercato su un campione di $n_1=10$ prezzi si calcola una varianza campionaria corretta pari a $S^2_1=225{,}5$; nel secondo mercato su un campione di $n_2=8$ prezzi la varianza camp. corretta risulta pari a $S^2_2=180{,}4$. si vuole verificare il seguente sistema d’ipotesi:

$$H_0: \sigma^2_1 = \sigma^2_2$$

$$H_1: \sigma^2_1 \neq \sigma^2_2$$

Test d'ipotesi sulle varianze

Si calcola la statistica test:

$$\frac{S_1^2}{S_2^2} = \frac{225{,}5}{180{,}4} = 1{,}25$$

Essendo l'ipotesi alternativa bidirezionale, si confronta il valore della statistica test con i valori della F di Fisher $f_{1\text{-a/2}}$ e $f_{\text{a/2}}$.

Per α=5% il valore $f_{0,025}$ con gradi di libertà pari a 9 (n_1-1) e 7 (n_2-1) è uguale a 4,82, mentre il valore $f_{0,975}$ è pari a 0,238

Il valore $f_{0,975}$ non è presente sulle tavole e quindi si utilizza la trasformazione:

$$f_{1-a/2} = \frac{1}{f^*_{a/2}} \qquad f_{0,975} = \frac{1}{f^*_{0,025}} = \frac{1}{4{,}20} = 0{,}238$$

Poiché il valore della statistica test (1,25) risulta compreso tra $f_{1\text{-a/2}}$ e $f_{\text{a/2}}$ accetteremo l'ipotesi nulla che prevede l'uguaglianza delle varianze.

Test d’ipotesi sulle proporzioni

È possibile effettuare un test sulle proporzioni provenienti da due popolazioni bernoulliane. Il sistema d’ipotesi analogamente al test sulle medie può essere espresso come differenza tra le due proporzioni:

1. $H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 > 0$
2. $H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 < 0$
3. $H_0: \pi_1 - \pi_2 = 0$ $H_1: \pi_1 - \pi_2 \neq 0$

Test d'ipotesi sulle proporzioni

La statistica test al tendere di n_1 e n_2 a infinito si distribuisce come una Z sotto l'ipotesi nulla **H_0: $\pi_1 - \pi_2 = 0$** :

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}_p\left(1 - \bar{X}_p\right)\left(1/n_1 + 1/n_2\right)}} \sim N(0,1)$$

$\bar{X}_p$ rappresenta lo stimatore congiunto della proporzione:

$$\bar{X}_p = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

Ipotesi alternativa	*Regione di rifiuto*
$H_1: \ \pi_1 > \pi_2$	$Z \geq z_\alpha$
$H_1: \ \pi_1 < \pi_2$	$Z \leq -z_\alpha$
$H_1: \ \pi_1 \neq \pi_2$	$\lvert Z \rvert \geq z_{\alpha/2}$

Test d'ipotesi sulle proporzioni

Esempio

Si vuole confrontare il tasso di occupazione delle persone con laurea in legge (π_1) con quelle laureate in economia (π_2). L'ipotesi alternativa da testare è che i laureati in legge presentano un tasso di occupazione minore.

$$\boxed{\begin{aligned} H_0 &: \ \pi_1 = \pi_2 \\ H_1 &: \ \pi_1 < \pi_2 \end{aligned}}$$

Si estraggono due campioni casuali il primo, dei laureati in legge, è composto da n_1=80 laureati, il secondo è composto da n_2=120 laureati in economia. Nel primo campione (laurea in legge) 59 laureati sono occupati mentre nel secondo campione risultano occupati 93 laureati.

Test d'ipotesi sulle proporzioni

Esempio

La proporzione di occupati tra i laureati in legge è pari a 59/80=0,738, mentre quella dei laureati in economia è pari a 93/120=0,775. Lo stimatore congiunto di π è pari a :

$$\overline{X}_p = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1+n_2} = \frac{80\cdot 0{,}738 + 120\cdot 0{,}775}{80+120} = 0{,}76$$

La statistica test risulta pari a:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\overline{X}_p\left(1-\overline{X}_p\right)\left(1/n_1 + 1/n_2\right)}} = \frac{0{,}738-0{,}775}{\sqrt{0{,}76(1-0{,}76)\left(\frac{1}{80}+\frac{1}{120}\right)}} = -0{,}6$$

Al livello a=5% si ottiene $-z_{0,05}$=-1,645. Il valore della statistica test risulta maggiore di z_a (-0,6>-1,645) per cui si accetta H_0 (uguaglianza tra i due tassi).

Test d'indipendenza

Per verificare l'indipendenza tra due caratteri si può ricorrere ***all'indice di associazione Chi-quadrato di Pearson***

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

Si può dimostrare che questo indice si distribuisce asintoticamente come una chi quadrato con (k-1)(h-1) gradi di libertà.

Di conseguenza è possibile verificare se sussiste indipendenza tra i due caratteri impostando un test d'ipotesi come il seguente:

H_0: $p_{ij}=p_i p_j$ (indipendenza tra i caratteri)

H_1: $p_{ij} \neq p_i pp_{ij}$ (dipendenza tra i caratteri)

Test d'indipendenza

Per effettuare il test si calcola l'indice chi quadrato (statistica test) e si confronta con i valori di una chi quadrato con k-1 e h-1 gradi di libertà per un dato livello di α

Ipotesi alternativa	Regione di rifiuto
$H_1:\ p_{ij} \neq p_i p_j$	$\chi^2 \geq \chi^2_\alpha$

Dove χ^2_a indica il valore della chi quadrato con (K-1) x (H-1) gradi di libertà per il quale

$$P(\chi^2 \geq \chi^2_a) = \alpha$$

Test d'indipendenza

In un'indagine di marketing (139 interviste) si è chiesto di indicare la preferenza tra 3 alimenti liquidi e 3 alimenti solidi da consumare al mattino a colazione:

	Biscotti	Merendina	Fette biscottate	Totale
Caffelatte	45	8	5	58
Tè	7	5	31	43
Succo di frutta	5	27	6	38
Totale	57	40	42	139

La statistica test vale: $\chi^2 = 101{,}57$

Ponendo α=0,01 il valore critico per la distribuzione Chi-quadrato con (3-1)(3-1)=4 gradi di libertà è:

$$\chi^2_{0,01} = 13{,}227$$

Si rifiuta quindi l'ipotesi nulla di indipendenza tra i due caratteri.

P-value

Il p-value viene anche definito come ***livello di significatività osservato*** *in quanto rappresenta il valore più piccolo di a per il quale il campione porta a rifiutare l'ipotesi nulla.*

Quanto più il p-value è vicino a zero, tanto più è forte l'evidenza fornita dai dati ***contro*** *l'ipotesi nulla in direzione dell'ipotesi alternativa.*

Per calcolare il p-value, rispetto ad una statistica test Z, si può considerare:

$$\textit{p-value} = 1 - \Phi(Z)$$

P-value

Esempio con Test su media

Si ipotizza di testare la seguente ipotesi relativamente alla media di una popolazione la cui varianza è pari a 25

H_0: μ=30

H_1: μ>30

Si estrae un campione di ampiezza n=150 e la media campionaria risulta pari a 30,8.

$$\bar{X} \sim N(30, \frac{25}{150}) \rightarrow Z = \frac{\bar{X} - 30}{\sqrt{25/150}} = 1.96$$

La regione di rifiuto con un a=5% è costituita dai valori superiori a 1,645 e quindi si rifiuta l'ipotesi nulla (H_0: μ=30)

P-value

$$z=1{,}96 \begin{cases} z_{a=5\%}=1{,}645 \Rightarrow \text{Rifiuto } H_0 \\ z_{a=1\%}=2{,}33 \Rightarrow \text{Accetto } H_0 \end{cases}$$

*Riprendiamo la definizione: «Il p-value viene anche definito come **livello di significatività osservato** in quanto rappresenta il valore più piccolo di a per il quale il campione porta a rifiutare l'ipotesi nulla.»*

$$p\text{-}value=1-\Phi(1{,}96)=0{,}025=2{,}5\%$$

Di conseguenza osservare un p-value =2,5% vuol dire che si accetterà Ho se si effettua il test per valori di a<2,5% viceversa viene rifiutata Ho per qualsiasi a≥ 2,5%

P-value

$z_{a=5\%}=1,645 \Longrightarrow$ *Rifiuto* H_0

$z=1,96$

$z_{a=1\%}=2,33 \Longrightarrow$ *Accetto* H_0

*Riprendiamo la definizione: «Il p-value viene anche definito come **livello di significatività osservato** in quanto rappresenta il valore più piccolo di a per il quale il campione porta a rifiutare l'ipotesi nulla.»*

$$p\text{-}value=1-\Phi(1,96)=0,025=2,5\%$$

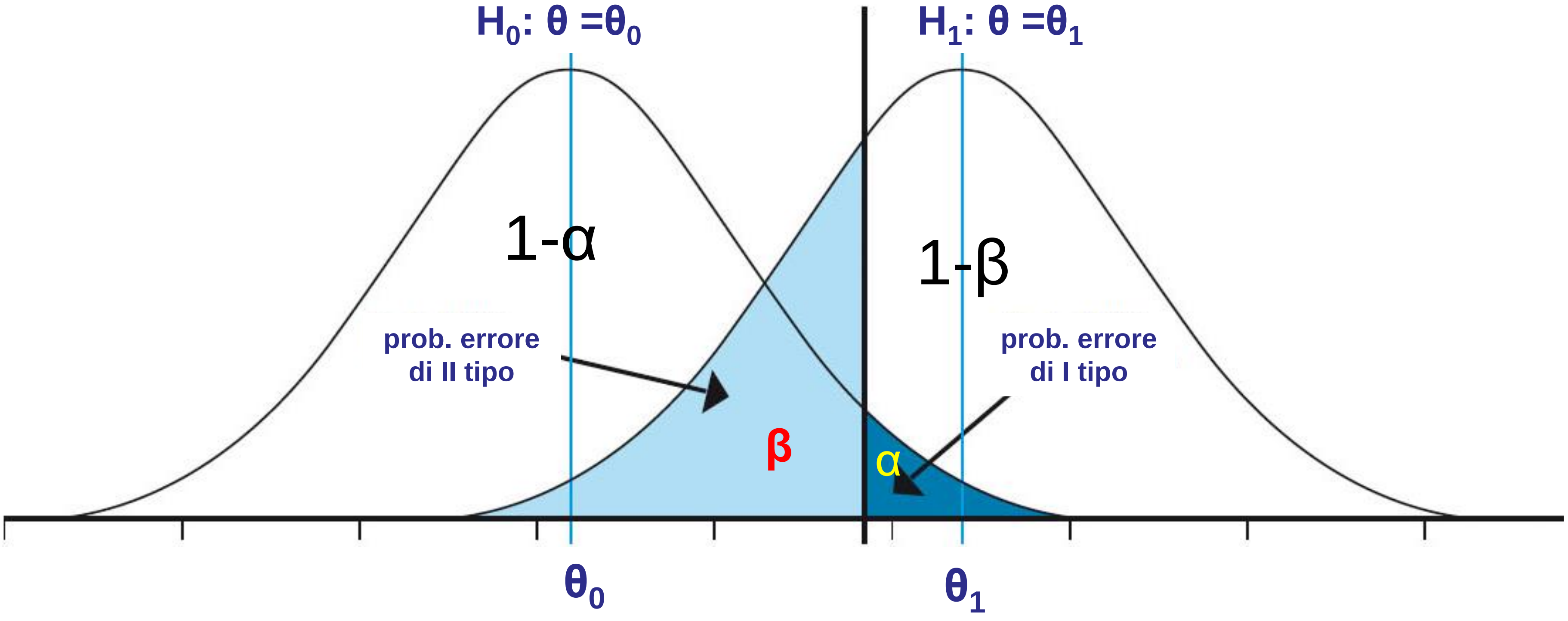
Di conseguenza osservare un p-value =2,5% vuol dire che si accetterà Ho se si effettua il test per valori di a<2,5% viceversa viene rifiutata Ho per qualsiasi a≥ 2,5%

Errore del I e II tipo

Si commette un ***errore del I tipo*** quando si rifiuta l'ipotesi nulla mentre questa è vera

Si commette un ***errore del II tipo*** quando si accetta l'ipotesi nulla mentre questa è falsa

Errore del I e II tipo



Errore del I e II tipo

	Decisione	
	Accetto H_0	**Rifiuto H_0**
H_0 è vera	Corretta 1-a	errore del I tipo a
H_0 è falsa	errore del II tipo β	Corretta 1-β