

Test per confrontare un valore atteso con un numero

Situazione

Popolazione $\mathcal{N}(\mu, \sigma^2)$; varianza σ^2 nota.
 μ_0 numero reale fissato.

Test di livello α per μ

$$\text{Statistica: } Z_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

H_0	H_1	Rifiutiamo H_0 se	p-value
$\mu \leq \mu_0$	$\mu > \mu_0$	$Z_n > z_{1-\alpha}$	$\Phi(-z_n)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$Z_n < -z_{1-\alpha}$	$\Phi(z_n)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$ z_n > z_{1-\frac{\alpha}{2}}$	$2 - 2\Phi(z_n)$

dove z_n = valore di Z_n calcolato dal campione; z_β = quantile β della $\mathcal{N}(0, 1)$.

Perché funziona

Studiamo il test

H_0	H_1	Rifiutiamo H_0 se	p -value
$\mu = \mu_0$	$\mu \neq \mu_0$	$ z_n > z_{1-\frac{\alpha}{2}}$	$2 - 2\Phi(z_n)$

Osserviamo che se vale H_0 la v.a.

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

è una $\mathcal{N}(0, 1)$.

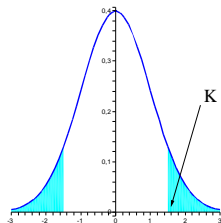
Volendo rifiutare H_0 è naturale farlo quando **la media aritmetica dei valori osservati è “molto diversa”** da μ_0 , cioè quando $|\bar{X}_n - \mu_0|$ è “grande”; **ovvero quando $|Z_n|$ è “grande”**.

Il livello α

Prendiamo un livello di significatività α : dovremo scegliere una regione di rifiuto del tipo:

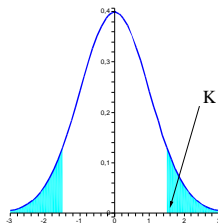
$|Z_n|$ “grande”, cioè $|Z_n| > K$, per qualche $K > 0$,

per cui **se** $H_0 : \mu = \mu_0$ **fosse vera la probabilità che** $|Z_n| > K$ **sia** (al massimo) α : dobbiamo determinare K in modo che l'area colorata valga α .



Come trovare K

K deve essere scelto in modo che l'area colorata valga α .

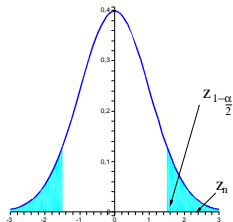


Ognuna delle 2 parti azzurre vale $\alpha/2$, quindi l'area a sinistra di K vale $1 - \alpha/2$.

Per definizione di quantile: $K = z_{1-\frac{\alpha}{2}}$.

Il p -value

Supponiamo che la statistica Z_n assuma il valore z_n e calcoliamo il p -value (il più piccolo livello a cui rifiutiamo H_0). Il test rifiuta H_0 se z_n cade sull'asse delle x nella zona azzurra, cioè se $|z_n| > z_{1-\frac{\alpha}{2}}$:



Più α è piccolo, più $z_{1-\frac{\alpha}{2}}$ va a destra. Quindi il più piccolo α per cui z_n cade nella zona azzurra è $\bar{\alpha}$ tale che $z_n = z_{1-\frac{\bar{\alpha}}{2}}$.

Il p -value

Quindi il p -value è $\bar{\alpha}$ tale che $z_n = z_{1-\frac{\bar{\alpha}}{2}}$.

Allora l'area a sinistra di z_n vale $1 - \frac{\bar{\alpha}}{2}$.

D'altra parte l'area a sinistra di z_n vale $\Phi(z_n)$, quindi da

$$\Phi(z_n) = 1 - \frac{\bar{\alpha}}{2}$$

ricaviamo

$$p\text{-value} = \bar{\alpha} = 2(1 - \Phi(z_n)).$$

Test per una media - varianza nota

Ricordiamo:

H_0	H_1	Rifiutiamo H_0 se	p-value
$\mu \leq \mu_0$	$\mu > \mu_0$	$Z_n > Z_{1-\alpha}$	$\Phi(-Z_n)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$Z_n < -Z_{1-\alpha}$	$\Phi(Z_n)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$ Z_n > Z_{1-\frac{\alpha}{2}}$	$2 - 2\Phi(Z_n)$

Analogamente a quanto appena visto si ricavano i due test per le altre due ipotesi (è un po' più difficile perché la legge di Z_n non è detto sia $\mathcal{N}(0, 1)$ e per α bisogna fare un sup sulle probabilità di rifiutare H_0).

Test più potenti.

Si potrebbe anche mostrare che **questo test** (come quelli che vedremo in seguito) è **il più potente test di livello α per la media** con campioni provenienti da popolazioni normali.

Test per una media - varianza ignota

Situazione

Popolazione $\mathcal{N}(\mu, \sigma^2)$; varianza σ^2 ignota.
 μ_0 numero reale fissato.

Test di livello α per μ

$$\text{Statistica: } T_n = \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}}.$$

H_0	H_1	Rifiutiamo H_0 se	$p\text{-value}=\bar{\alpha}$ tale che
$\mu \leq \mu_0$	$\mu > \mu_0$	$t_n > t_{1-\alpha}(n-1)$	$t_{1-\bar{\alpha}}(n-1) = t_n$
$\mu \geq \mu_0$	$\mu < \mu_0$	$t_n < -t_{1-\alpha}(n-1)$	$t_{1-\bar{\alpha}}(n-1) = -t_n$
$\mu = \mu_0$	$\mu \neq \mu_0$	$ t_n > t_{1-\frac{\alpha}{2}}(n-1)$	$t_{1-\bar{\alpha}/2}(n-1) = t_n $

dove t_n = valore di T_n calcolato dal campione; $t_{\beta}(n-1)$ = quantile β della $t(n-1)$.

Perché funziona

Se la popolazione è $\mathcal{N}(\mu, \sigma^2)$

abbiamo visto che

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

se σ^2 è incognita questo fatto è inutilizzabile, ma abbiamo visto (negli intervalli di confidenza) che

$$\frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}} \sim t(n-1),$$

dove $t(n-1)$ indica la v.a. di Student con $n-1$ gradi di libertà. Dopo di che si procede come nel test a varianza nota.

Test per la media di grandi campioni

I due test per il valore atteso di popolazioni normali (varianza nota o ignota) si possono utilizzare anche per popolazioni che seguono altri modelli statistici, purché il campione sia sufficientemente numeroso da consentire l'applicazione del Teorema del Limite Centrale.

Test su una frequenza (Bernoulli)

Un caso particolare è quello del test per il parametro p di una popolazione $\mathcal{B}(p)$. Si vuole confrontare p con un numero $p_0 \in (0, 1)$. Sia x_1, \dots, x_n un campione estratto da quella popolazione e sia $n\bar{x}_n \geq 5$ e $n(1 - \bar{x}_n) \geq 5$.

Test di livello α per p

$$\text{Statistica: } Z_n = \frac{\bar{X}_n - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

H_0	H_1	Rifiutiamo H_0 se	p -value
$p = p_0$	$p \neq p_0$	$ z_n > z_{1-\alpha/2}$	$2 - 2\Phi(z_n)$
$p \leq p_0$	$p > p_0$	$z_n > z_{1-\alpha}$	$\Phi(-z_n)$
$p \geq p_0$	$p < p_0$	$z_n < -z_{1-\alpha}$	$\Phi(z_n)$

dove z_n = valore di Z_n calcolato dal campione; z_β = quantile β della $\mathcal{N}(0, 1)$.

Test per confrontare due valori attesi

Vogliamo ora confrontare due medie: abbiamo un campione

$$X_1, \dots, X_n$$

da una popolazione con valore atteso (ignoto) μ_X ; e un altro campione

$$Y_1, \dots, Y_m$$

da una popolazione con valore atteso (ignoto) μ_Y .

Test per due medie

Il test che cerchiamo dovrà darci una regola di decisione per le seguenti situazioni (δ è un numero reale fissato, può valere anche 0):

H_0	H_1
$\mu_X = \mu_Y + \delta$	$\mu_X \neq \mu_Y + \delta$
$\mu_X \leq \mu_Y + \delta$	$\mu_X > \mu_Y + \delta$
$\mu_X \geq \mu_Y + \delta$	$\mu_X < \mu_Y + \delta$

In tutti i test sui valori attesi supponiamo che

- le popolazioni seguano la legge normale (quindi anche le medie campionarie sono v.a. normali);
- oppure i campioni siano abbastanza numerosi da poter applicare il Teorema del Limite Centrale (quindi le medie campionarie sono approssimabili con v.a. normali).

Campioni accoppiati o indipendenti

DEFINIZIONE di CAMPIONI ACCOPPIATI O INDIPENDENTI

Distinguiamo **due differenti situazioni**:

- 1 i campioni sono **accoppiati**, cioè si tratta di osservazioni diverse effettuate sugli stessi individui;
- 2 i campioni sono **indipendenti**, cioè si tratta di osservazioni provenienti da popolazioni diverse.

Esempi accoppiati

Esempi di campioni accoppiati:

- Si vuole stabilire se le rondini hanno in genere un'ala più sviluppata dell'altra: se ne prendono n e
 X_i =lunghezza dell'ala destra dell' i -esimo individuo;
 Y_i =lunghezza dell'ala sinistra dell' i -esimo individuo.
- Si vuole stabilire se un betabloccante è efficace: si prendono n pazienti e
 X_i =battiti cardiaci al minuto dell' i -esimo individuo prima della somministrazione del betabloccante;
 Y_i =battiti cardiaci al minuto dell' i -esimo individuo dopo la somministrazione del betabloccante.

Esempi indipendenti

Esempi di campioni indipendenti:

- Si vuole stabilire se una cura per l'otite è efficace: si prendono due gruppi di pazienti, uno lo si tratta con placebo, l'altro col farmaco:
 X_i =durata in giorni dell'otite per l' i -esimo individuo di un gruppo di pazienti trattati con placebo;
 Y_i =durata in giorni dell'otite per l' i -esimo individuo di un gruppo di pazienti trattati col farmaco.
- Si vuole stabilire se due sottospecie di Iris sono distinguibili misurando la lunghezza dei petali: si prende un gruppo di piante della prima sottospecie e un gruppo dell'altra sottospecie:
 X_i =lunghezza di un petalo di fiore dell' i -esimo individuo della prima sottospecie di Iris;
 Y_i =lunghezza di un petalo di fiore dell' i -esimo individuo della seconda sottospecie di Iris.

Test per due medie - campioni accoppiati

Situazione

Popolazione X con valore atteso μ_X , **accoppiata** con la popolazione Y con valore atteso μ_Y .

Idea: basta considerare il nuovo campione Z_1, \dots, Z_n definito come

$$Z_1 = X_1 - Y_1; \quad Z_2 = X_2 - Y_2; \quad \dots; \quad Z_n = X_n - Y_n.$$

Si tratta del campione delle differenze: **il test su due medie si riduce ad un test per una media!**

Ipotesi da 2 ad 1 media

Considerando $Z = X - Y$ le ipotesi si “traducono” in questo modo:

H_0 per 2 medie	H_0 per 1 media
$\mu_X = \mu_Y + \delta$	$\mu_Z = \delta$
$\mu_X \leq \mu_Y + \delta$	$\mu_Z \leq \delta$
$\mu_X \geq \mu_Y + \delta$	$\mu_Z \geq \delta$

Conclusione

Se si può supporre che la “popolazione differenza” Z segua la distribuzione normale, oppure se il campione è abbastanza numeroso da consentire l'applicazione del Teorema del Limite Centrale, si utilizzano i test per 1 media già visti.

Test per due medie - campioni indipendenti

Anzitutto osserviamo che se i campioni

$$X_1, \dots, X_n \quad \text{e} \quad Y_1, \dots, Y_m$$

sono indipendenti, può essere che $n \neq m$.

Situazione

Popolazione $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$;
popolazione $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

Allora

$$\bar{X}_n - \bar{Y}_m - \delta \sim \mathcal{N} \left(\mu_X - \mu_Y - \delta, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} \right),$$

Campioni indipendenti - varianze note

Test per due medie - campioni indipendenti, varianze note

H_0	H_1	Rifiuto H_0 se	p -value
$\mu_X - \mu_Y = \delta$	$\mu_X - \mu_Y \neq \delta$	$ z_{n,m} > z_{1-\alpha/2}$	$2 - 2\Phi(z_{n,m})$
$\mu_X - \mu_Y \leq \delta$	$\mu_X - \mu_Y > \delta$	$z_{n,m} > z_{1-\alpha}$	$\Phi(-z_{n,m})$
$\mu_X - \mu_Y \geq \delta$	$\mu_X - \mu_Y < \delta$	$z_{n,m} < -z_{1-\alpha}$	$\Phi(z_{n,m})$

$$\text{dove } z_{n,m} = \frac{\bar{x}_n - \bar{y}_m - \delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Campioni indipendenti - varianza incognita

Situazione

Popolazione $X \sim \mathcal{N}(\mu_X, \sigma^2)$;

popolazione $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$.

Varianza σ^2 incognita ma **uguale** per le due popolazioni.

Si potrebbe dimostrare che

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n + m - 2)$$

dove

$$S = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}.$$

La varianza combinata

Ricordando che S_X^2 e S_Y^2 sono rispettivamente la varianza campionaria del campione X e quella del campione Y ,

$$\begin{aligned} S^2 &= \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \\ &= \frac{1}{n+m-2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 \right) \end{aligned}$$

si interpreta come una *combinazione* delle due varianze campionarie.

DEFINIZIONE DI VARIANZA COMBINATA

S^2 è detta **varianza combinata** (*pooled variance* in inglese).

Campioni indipendenti - varianza ignota

Test per due medie - campioni indipendenti, varianze note

H_0	H_1	Rifiuto H_0 se	$p\text{-value}=\bar{\alpha}$ con
$\mu_X - \mu_Y = \delta$	$\mu_X - \mu_Y \neq \delta$	$ \hat{t} > t_{1-\alpha/2}(k)$	$t_{1-\bar{\alpha}/2}(k) = \hat{t} $
$\mu_X - \mu_Y \leq \delta$	$\mu_X - \mu_Y > \delta$	$\hat{t} > t_{1-\alpha}(k)$	$t_{1-\bar{\alpha}}(k) = \hat{t}$
$\mu_X - \mu_Y \geq \delta$	$\mu_X - \mu_Y < \delta$	$t < -t_{1-\alpha}(k)$	$t_{1-\bar{\alpha}}(k) = -t$

dove $k = n + m - 2$,

$$\hat{t} = \frac{\bar{x}_n - \bar{y}_m - \delta}{s \sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

e s è la varianza combinata osservata.

Confronto di due medie

I test che abbiamo visto per il confronto di due medie valgono se le popolazioni in gioco seguono la distribuzione normale (e si può dimostrare che sono i test più potenti possibile). Se però i campioni sono sufficientemente numerosi da consentire l'applicazione del Teorema del Limite Centrale, allora possiamo usare questi test anche per popolazioni non normali.

Ricordiamo inoltre che abbiamo visto il confronto di due medie solo nel caso in cui le due varianze o sono note oppure sono incognite ma uguali fra loro.

Esiste una formula anche per il caso in cui le due varianze sono incognite e diverse fra loro e anche altri test per campioni piccoli. Questi casi non li vediamo.

Confronto di due Bernoulli (grandi campioni)

Vediamo invece un'applicazione del test per confronto di medie al confronto di parametri di due popolazioni bernoulliane. Si usa il **test z per campioni indipendenti con varianze note**, supponendo che i campioni siano abbastanza numerosi da consentire l'applicazione del Teorema del Limite Centrale:

$$n\bar{x}_n \geq 5, \quad n(1 - \bar{x}_n) \geq 5, \quad m\bar{y}_m \geq 5, \quad m(1 - \bar{y}_m) \geq 5.$$

Inoltre si sostituisce

al posto di σ_X^2 la quantità $\bar{x}_n(1 - \bar{x}_n)$
al posto di σ_Y^2 la quantità $\bar{y}_m(1 - \bar{y}_m)$.

Test per due medie

Quello per due medie, campioni indipendenti e varianze note, ponendo $\delta = 0$, era:

H_0	H_1	Rifiuto H_0 se	p -value
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$ Z_{n,m} > Z_{1-\alpha/2}$	$2 - 2\Phi(Z_{n,m})$
$\mu_X \leq \mu_Y$	$\mu_X > \mu_Y$	$Z_{n,m} > Z_{1-\alpha}$	$\Phi(-Z_{n,m})$
$\mu_X \geq \mu_Y$	$\mu_X < \mu_Y$	$Z_{n,m} < -Z_{1-\alpha}$	$\Phi(Z_{n,m})$

dove

$$Z_{n,m} = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$

Test per due frequenze

Situazione

Abbiamo due popolazioni, la X_1, \dots, X_n sia $\mathcal{B}(p_1)$ e la Y_1, \dots, Y_m sia $\mathcal{B}(p_2)$, con

$$n\bar{x}_n \geq 5, \quad n(1 - \bar{x}_n) \geq 5, \quad m\bar{y}_m \geq 5, \quad m(1 - \bar{y}_m) \geq 5.$$

Test per due frequenze

H_0	H_1	Rifiuto H_0 se	p -value
$p_1 = p_2$	$p_1 \neq p_2$	$ z_n > z_{1-\alpha/2}$	$2 - 2\Phi(z_n)$
$p_1 \leq p_2$	$p_1 > p_2$	$z_n > z_{1-\alpha}$	$\Phi(-z_n)$
$p_1 \geq p_2$	$p_1 < p_2$	$z_n < -z_{1-\alpha}$	$\Phi(z_n)$

$$\text{dove } z_n = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n} + \frac{\bar{y}_m(1-\bar{y}_m)}{m}}}$$

Avvertenze

Il test per le due Bernoulli non è quello riportato sul libro di Bramanti.

Inoltre bisogna sempre far attenzione all'applicabilità dell'approssimazione normale, che richiede campioni piuttosto numerosi.