TEST DI IPOTESI (M.S. Bernabei)

Definizione. L' **ipotesi statistica** è una affermazione o una congettura intorno ad un parametro incognito ρ della popolazione, formulata sulla base di considerazioni teoriche o risultati sperimentali che riguarda un aspetto della popolazione studiata. L'ipotesi sottoposta a verifica viene in genere indicata con H_o e viene chiamata **ipotesi nulla**. Si chiama **test** il procedimento con cui si decide se, attraverso i dati del campione, accettare o rifiutare H_o. L'ipotesi H_o verrà rifiutata se dal confronto dei dati osservati con essa emergerà una "discrepanza considerevole". Il problema è "decidere" quanto deve essere grande la distanza tra il valore teorizzato e quello vero perché venga rifiutata l'ipotesi.

Esempio 1: L'industria A sostiene che le batterie elettriche da essa prodotte hanno una durata media di 36 mesi con una deviazione standard di 3 mesi. Un' industria automobilistica B è interessata al prodotto, ma prima di qualsiasi decisione di acquisto, intende controllare l' affermazione di A attraverso l'osservazione di un campione di batterie dalla popolazione composta dal numero indefinito di batterie che l'industria può potenzialmente produrre. L'affermazione di A, cioé μ=36, è un'ipotesi sulla media della popolazione, che si può approssimare con una distribuzione di Gauss se l'ampiezza del campione é sufficientemente ampia.

Data l'ipotesi nulla $H_0 : \rho = \rho_0$ allora l' **ipotesi alternativa** detta H_1 può essere
H_1 :

$\rho \neq \rho_0$

$\rho < \rho_0$

$\rho > \rho_0$

Relativamente all' Esempio precedente poiché B dubita dell'ipotesi H_0, penso che il valore medio dichiarato sia troppo alto allora si pone
H_1: μ < 36.

Il test può condurre a decisioni errate, infatti esso è eseguito su base probabilistica utilizzando i risultati ottenuti con un campione

Si confrontano quindi due possibili ipotesi:

- H_0: Ipotesi nulla
- H_1: Ipotesi alternativa

Il metodo per eseguire il test è analogo a quello usato per calcolare gli intervalli di confidenza. In generale nell'ipotesi nulla H_0 si suppone che un parametro ρ della popolazione abbia valore pari a ρ_0. Non essendo certi di tale supposizione si esegue un test di tale ipotesi mediante un'indagine campionaria. Consideriamo come parametro incognito la media μ di una popolazione.

TEST A DUE CODE PER UNA MEDIA

Si formulano le ipotesi:

- H_0: $\mu = \mu_0$ (ipotesi nulla: il valore del parametro μ è pari a μ_0)
- H_1: $\mu \neq \mu_0$ (ipotesi alternativa: il valore del parametro μ è diverso da μ_0)

Varianza nota: sia σ^2 la varianza (nota) della popolazione; la media campionaria $\overline{X}$ ha distribuzione normale con media μ_0 e varianza σ^2 / n (n dimensione dei campioni); quindi la variabile casuale

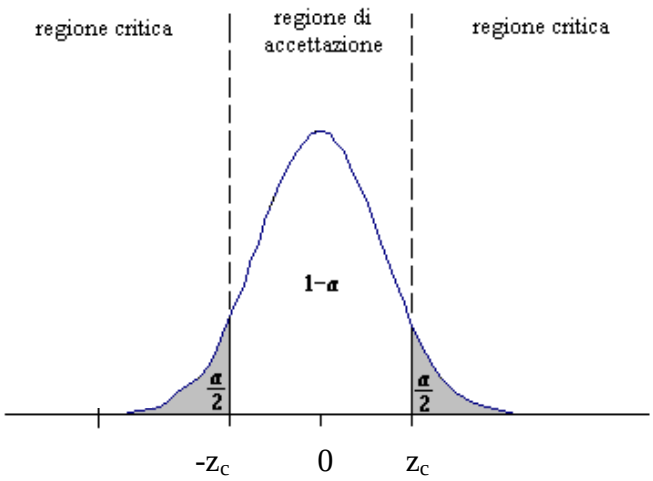
$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

ha distribuzione approssimativamente normale standard per n sufficientemente grande. Se l'ipotesi H_0 è vera, il valore Z è probabilmente vicino a *0*, anche se difficilmente sarà uguale. Se l'ipotesi H_0 è falsa, il valore Z è probabilmente abbastanza "lontano" da *0*. È opportuno quindi stabilire quanto deve discostarsi Z da 0 affinchè H_0 possa essere ritenuta falsa. Tale decisione si prende in termini probabilistici ed essa è in funzione di un grado di fiducia *1-a* già introdotto nell' ambito dell' intervallo di fiducia per una media, a cui corrisponde un intervallo $(-z_c, z_c)$ di valori critici. Allora

- se $Z < -z_c$ oppure $Z > z_c$ (**regione critica**): si deve rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a $a / 2$ di avere $Z < -z_c$ e si avrebbe una probabilità pari a $a / 2$ di avere $Z > z_c$; si accetta quindi H_1 e si dice che il test è stato **significativo**, in quanto ha cambiato l' ipotesi principale.
- se $-z_c \leq Z \leq z_c$ (**regione di accettazione**): non si può rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a *1-a* di avere $-z_c \leq Z \leq z_c$; si accetta quindi H_0 e si dice che il test **non** è stato **significativo**, in quanto non ha cambiato l' ipotesi principale.

Per mezzo delle tavole della distribuzione normale standard è possibile calcolare il valore critico z_c, tale che l'area della regione critica è pari ad *1-a*; se il valore della variabile casuale *Z* ottenuto dal campione cade nella regione di accettazione, H_0 si accetta; se il valore della variabile casuale *Z* ottenuto dal campione cade nella regione critica, H_1 si rifiuta.

Il test appena considerato è detto **Test a due code** perché l'ipotesi H_1 si accetta per valori *Z* sia maggiori di z_c che minori di $-z_c$. Questo tipo di test serve per verificare se Z è significativamente diverso da 0 o, equivalentemente, se il parametro μ è significativamente diverso da μ_0.



Test ad una coda per la media.

Varianza nota. A volte è necessario usare dei tests ad una coda in cui la regione critica è localizzata solo a destra di μ_0 o solo a sinistra di μ_0, o, se consideriamo la variabile standardizzata Z

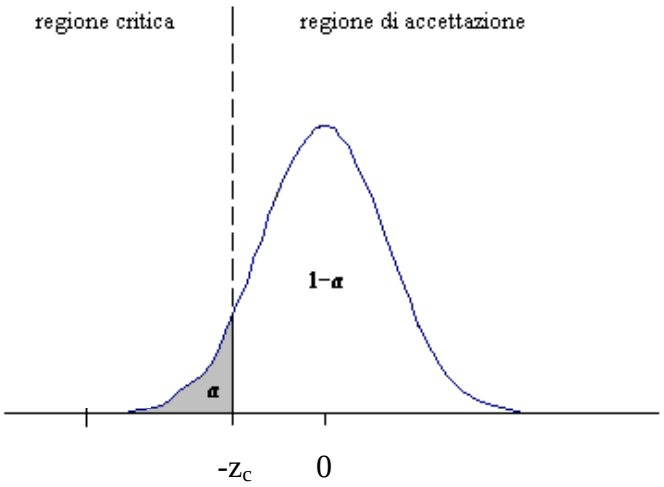
a destra di 0, o a sinistra di 0. Sia σ^2 la varianza (nota) della popolazione; la media campionaria $\overline{X}$ ha distribuzione normale con media μ_0 (quella ipotizzata nell' ipotesi nulla) e varianza σ^2 / n (n dimensione dei campioni); quindi la variabile casuale

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

ha distribuzione approssimativamente normale standard per n sufficientemente grande.

I caso: **coda a sinistra**

- H_0: $\mu = \mu_0$
- H_1: $\mu < \mu_0$

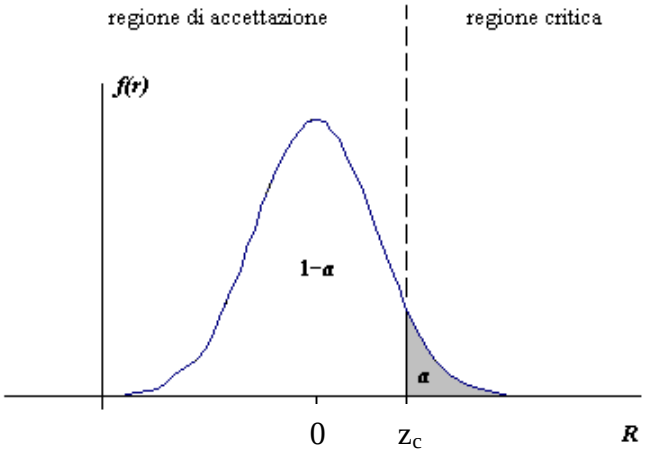


- se Z < $-z_c$ (**regione critica**) si deve rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a α di avere Z < $-z_c$; si accetta quindi H_1 e si dice che il test è stato **significativo**, in quanto ha cambiato l' ipotesi principale.
- se Z ≥$-z_c$ (**regione di accettazione**): non si può rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a $1-\alpha$ di avere Z ≥$-z_c$; si accetta quindi H_0 e si dice che il test **non** è stato **significativo**, in quanto non ha cambiato l' ipotesi principale.

Il test appena considerato è detto **Test ad una coda** perché l'ipotesi H_1 si accetta per valori Z maggiori di $-z_c$.

II caso: **coda a destra**

- H_0: $\mu = \mu_0$
- H_1: $\mu > \mu_0$



- se Z > z_c (**regione critica**) si deve rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a a di avere Z >z_c; si accetta quindi H_1 e si dice che il test è stato **significativo**, in quanto ha cambiato l' ipotesi principale.
- se Z ≤ z_c (**regione di accettazione**): non si può rifiutare l'ipotesi H_0 , in quanto se essa fosse vera si avrebbe una probabilità pari a 1-a di avere Z ≤ z_c; si accetta quindi H_0 e si dice che il test **non** è stato **significativo**, in quanto non ha cambiato l' ipotesi principale.

Il test appena considerato è detto **Test ad una coda** perché l'ipotesi H_1 si accetta per valori Z minori di z_c.

VARIANZA NON NOTA: sia σ^2 la varianza (non nota) della popolazione; la media campionaria $\overline{X}$ ha distribuzione normale con media μ_0 e varianza σ^2 / n (n dimensione dei campioni), quindi la variabile casuale:

$$T = \frac{\overline{X} - \mu_0}{s' / \sqrt{n}}$$

ha distribuzione T di Student con $\nu = n-1$ gradi di libertà; quindi per mezzo delle tavole della distribuzione di Student è possibile calcolare t_c, detto **valore critico**, tale che l'area della regione critica è pari ad $1-a$; se il valore della variabile casuale T ottenuto dal campione cade nella regione di accettazione, H_0 si accetta; se il valore della variabile casuale T ottenuto dal campione cade nella regione critica, H_0 si rifiuta.

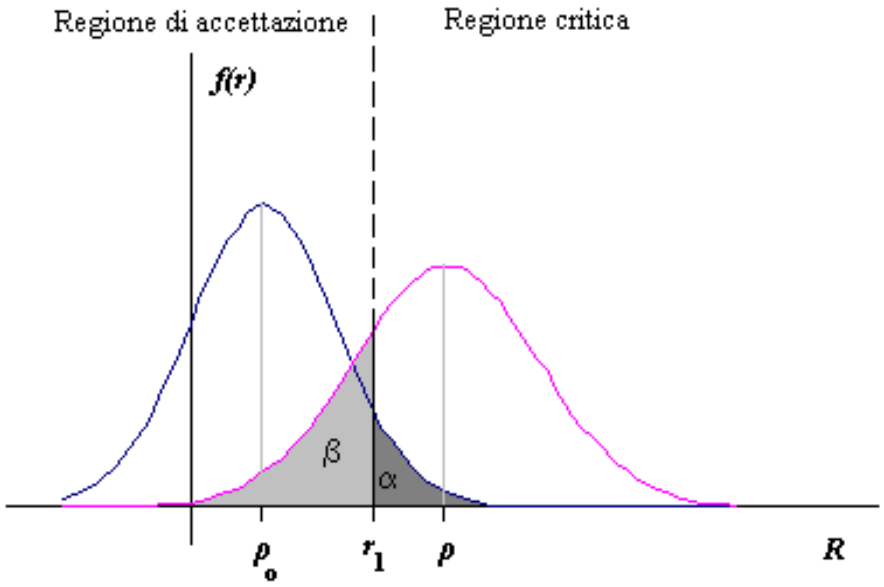
Osservazione: per grandi campioni ($n > 100$) la distribuzione T di Student con $\nu = n-1$ gradi di libertà è ben approssimata dalla distribuzione normale, quindi in tali ipotesi si può usare indifferentemente l'una o l'altra distribuzione; in particolare per grandi campioni ($n > 100$) la varianza della popolazione può essere considerata sempre nota perché s^2 fornisce una buona approssimazione di σ^2.

In tal caso vale il ragionamento fatto sopra per il test a due code e ad una coda, con T al posto di Z e t_c al posto di z_c.

ERRORI DI PRIMA E SECONDA SPECIE

La decisione di rifiutare o accettare H_0 è sempre presa su base probabilistica considerando dei risultati campionari, è quindi possibile commettere degli errori nel prendere tali decisioni:

- ERRORI DI PRIMA SPECIE: quando H_0 è vera, ma in base ai risultati campionari H_0 viene rifiutata
- ERRORI DI SECONDA SPECIE: quando H_0 è falsa, ma in base ai risultati campionari H_0 viene accettata



Si osserva che la probabilità di commettere un errore di prima specie è pari al livello di significatività *a*, mentre 1-*a* è la probabilità di accettare H_0 quando H_0 è vera. Sia β la probabilità di commettere un errore di seconda specie, mentre 1-β è la probabilità di accettare H_1 quando H_1 è vera. Il valore 1-β è generalmente detto POTENZA DEL TEST.

		Decisione	
		H_0	H_1
Realtà	H_0	DECISIONE ESATTA Prob.: $1 - \alpha$	ERRORI I SPECIE Prob.: α
	H_1	ERRORI II SPECIE Prob.: β	DECISIONE ESATTA Prob.: $1 - \beta$

ESEMPIO: il contenuto di nicotina delle sigarette di un certo tipo risulta normalmente distribuito con deviazione standard di 4mg. Se il contenuto medio di nicotina delle sigarette non deve superare 26mg e in un campione di 10 sigarette si sono ottenuti i seguenti valori di nicotina (in mg):

33 27 20 36 25 24 27 24 34 29

si può affermare, ad un livello di significatività pari a 0.05, che sia stato superato il livello massimo?
Risposta: si ha H_0: $\mu = 26$, H_1: $\mu > 26$ (test a una coda a destra). Dal livello di significatività $\alpha=0.05$ si ha $z_c=1.645$ (dalle tavole della distribuzione normale standard), inoltre dai dati campionari si ha $\bar{x}=27.9$ e quindi

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{27.9 - 26}{4/\sqrt{10}} = 1.502$$

cade nella regione di accettazione. Il campione non è statisticamente significativo si accetta l'ipotesi nulla H_0.
Osserviamo che quando il campione è troppo piccolo come in questo caso si deve fare l'ipotesi che la distribuzione della media campionaria è gaussiana, perchè non si può applicare il Teorema del Limite Centrale.

ESEMPIO: in base all'esperienza degli anni precedenti risulta che le votazioni, ad un certo esame scritto, riportate dagli studenti di un certo corso di laurea sono distribuite in maniera approssimativamente normale con media di 23 trentesimi. Se un gruppo di 50 studenti dell'anno in corso riporta una votazione media di 25 trentesimi con deviazione standard (con la correzione di Student) di 4 trentesimi, si può accettare l'ipotesi che tali studenti non differiscano da quelli degli anni precedenti al livello di significatività di 0.02
Risposta: si ha H_0: $\mu = 23$, H_1: $\mu \neq 23$ (test a due code). Dal livello di significatività $a=0.02$ si ha $t_c=2.423$ (dalle tavole della distribuzione t di Student con 49 gradi di libertà), inoltre dai dati campionari si ha $\bar{x}=25$, $s'=4$, quindi

$$T = \frac{\overline{X} - \mu_0}{S'/\sqrt{n}} = \frac{25 - 23}{4/\sqrt{50}} = 3.536$$

cade nella regione critica. Il campione è statisticamente significativo si rifiuta l'ipotesi nulla H_0.

TEST SULLE DIFFERENZE DI MEDIE

Si considera due popolazioni. Sia μ_1 la media della prima popolazione e sia μ_2 la media della seconda popolazione; si formula la seguente ipotesi nulla:

H_0: $\mu_1 = \mu_2$

mentre le ipotesi alternative possono essere:

H_1: $\mu_1 \neq \mu_2$ (test a due code)
H_1: $\mu_1 < \mu_2$ (test a una coda a sinistra)
H_1: $\mu_1 > \mu_2$ (test a una coda a destra)

Campioni indipendenti: non esiste alcuna relazione tra i risultati dei due campioni, i risultati di un campione non influenzano quelli dell'altro.

Varianza nota: siano σ_1^2, σ_2^2 le varianze (note) delle popolazioni; sia $\overline{X}_1$ la variabile casuale della media campionaria di campioni casuali di dimensione n_1 estratti dalla prima popolazione; sia $\overline{X}_2$ la variabile casuale della media campionaria di campioni casuali di dimensione n_2 estratti dalla seconda popolazione; allora la variabile casuale $\overline{X}_1 - \overline{X}_2$ ha distribuzione normale con media μ_1-μ_2 , che per H_0 è nulla, e varianza $\sigma_1^2/n_1 + \sigma_2^2/n_2$; quindi la variabile casuale:

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ha distribuzione normale standard, quindi per mezzo delle tavole della distribuzione normale standard è possibile calcolare z_c, tale che l'area della regione critica è pari ad a; se il valore della variabile Z ottenuto dal campione cade nella regione di accettazione, si accetta H_0; se il valore della variabile casuale Z ottenuto dal campione cade nella regione critica, si rifiuta H_0.

Varianze non note: si suppone per semplicità che $\sigma_1 = \sigma_2$ (**Test di Fisher**); siano $\overline{X}_1$ e S_1^2 rispettivamente la variabile casuale della media campionaria e la variabile casuale della varianza campionaria di campioni casuali di dimensione n_1 estratti dalla prima popolazione; siano $\overline{X}_2$ e S_2^2 rispettivamente la variabile casuale della media campionaria e la variabile casuale della varianza campionaria di campioni casuali di dimensione n_2 estratti dalla seconda popolazione; la stima di $\sigma = \sigma_1 = \sigma_2$ si può dare in termini di S_1^2 e S_2^2, cioè come media ponderata delle due varianze

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

inoltre la variabile casuale:

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S_p^2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S_p^2}}$$

ha distribuzione T di Student con $\nu = n_1 + n_2 - 2$ gradi di libertà; quindi per mezzo delle tavole della distribuzione t di Student è possibile calcolare t_c, tale che l'area della regione critica è pari ad a; se il valore della variabile casuale T ottenuto dal campione cade nella regione di accettazione, si accetta H_0; se il valore della variabile casuale T ottenuto dal campione cade nella regione critica, si rifiuta H_0.

Esempio: un campione di 40 capsule di analgesico è stato fabbricato da una macchina A, il peso medio è $\bar{x}$=330 mg, la deviazione standard è s=7 mg; una macchina B ha prodotto 50 capsule con peso medio $\bar{x}$=320 mg e deviazione standard s=6.5 mg. Sottoporre a test l'ipotesi che le due macchine producano capsule di stesso con un livello di significatività pari a 0.05
Risposta: si ha H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$ (test a due code). Si tratta di campioni indipendenti. Dal livello di significatività a=0.05 si ha t_c=1.98 (dalle tavole della distribuzione di Student), inoltre dai dati campionari si ha n_1=40, $\bar{x}_1$=330, s'_1=7, n_2=50, $\bar{x}_2$=320, s'_2=6.5 e quindi

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_p^2}} = \frac{330 - 320}{\sqrt{\left(\frac{1}{40} + \frac{1}{50}\right)\left(\frac{7^2(39) + 6.5^2(49)}{88}\right)}} = 7{,}009$$

cade nella regione critica. Il campione è statisticamente significativo si rifiuta l'ipotesi nulla H_0.

Confronto tra due medie con uguale varianze con Analisi dei dati di Excel

Nella finestra di Dialogo bisogna inserire il range della prima variabile, l'intervallo della seconda, la differenza che si ipotizza (0 se si vuole verificare che le due medie coincidano), Alfa, cioè *a* dove *1-a* è il grado di fiducia e dove visualizzare la Tabella.

Apparirà una tabella di output dove sono descritti le medie, le varianze, le numerosità dei due campioni, la media pesata delle varianze, la differenza delle medie ipotizzata, i gradi di libertà.
T-stat è il valore di T ottenuto dai due campioni che va confrontato con il valore **t critico a due code** se il test è a due code oppure con **t critico ad una coda** se il test è ad una coda.
Il valore **P(T<=t) una coda** è la probabilità di ottenere una differenza di medie campionarie maggiore o uguale a quella osservata, quindi dovrebbe essere **P(T>=t) una coda.**
P(T<=t) due code è la probabilità di ottenere una differenza di medie campionarie in modulo maggiore o uguale a quella osservata.