Corso di laurea in ECONOMIA AZIENDALE

STATISTICA I

Inferenza statistica

13° settimana

A.A. 2008/2009

Carla Rampichini

Test Unilaterali

- In molti casi, l'ipotesi alternativa si concentra su una particolare direzione

H_0: μ ≤ 3
H_1: μ > 3

⇒ Questo è un test sulla coda destra perchè l'ipotesi alternativa è focalizzata sulla coda destra, al di sopra della media 3

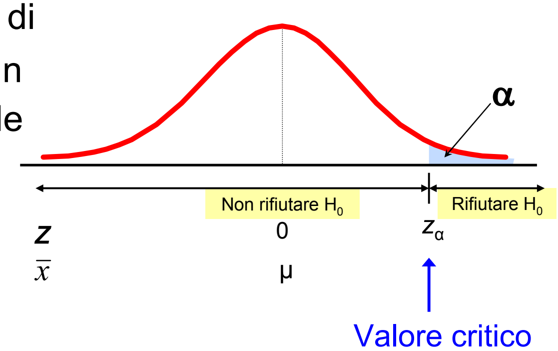
H_0: μ ≥ 3
H_1: μ < 3

⇒ Questo è un test sulla coda sinistra perchè l'ipotesi alternativa è focalizzata sulla coda sinistra, al di sotto della media 3

Test sulla coda destra

- C'è solo un valore critico, perchè l'area di rifiuto è solo in una delle code

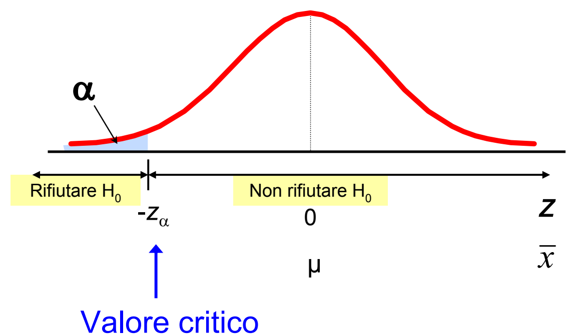
H_0: μ ≤ 3
H_1: μ > 3



Test sulla coda di sinistra

- C'è solo un valore critico, perchè l'area di rifiuto è solo in una delle code

H_0: μ ≥ 3
H_1: μ < 3

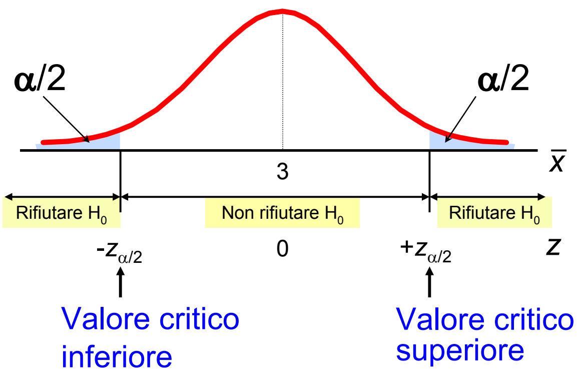


Test bilaterali

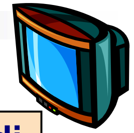
- In alcune situazioni, l'ipotesi alternativa non specifica un'unica direzione

H_0: μ = 3
H_1: μ ≠ 3

- Ci sono due valori critici, che definiscono i due intervalli della regione di rifiuto



Esempio verifica di ipotesi



Verificare l'ipotesi che il numero medio di TV nelle case italiane sia uguale a 3. (assumiamo X~N(μ, σ = 0.8)

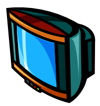
- Fornire le appropriate ipotesi nulla e alternativa
  - H_0: μ = 3 , H_1: μ ≠ 3 (test bilaterale)
- Specificare il livello di significatività desiderato
  - viene scelto α = .05 per questo test
- Scegliere la dimensione del campione
  - viene selezionato un campione di dimensione n = 100

Esempio verifica di ipotesi *(continuazione)*

- Determinare la tecnica appropriata
  - σ è nota quindi utilizziamo la statistica test Z
- Calcola i valori critici
  - per α = .05 i valori critici z sono ±1.96
- Raccogli i dati e calcola la statistica test
  - Supponiamo che i risultati campionari siano

$$n = 100,\ \bar{X} = 2.84 \quad (\sigma = 0.8,\ \text{si assume nota})$$

Quindi il valore della statistica test è:

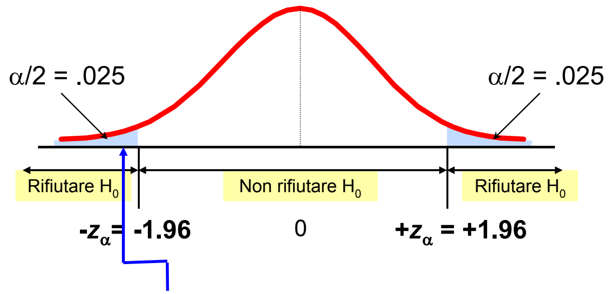


$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-0.16}{0.08} = -2.0$$

Esempio verifica di ipotesi *(continuazione)*

- La statistica test calcolata sul campione cade nella regione di rifiuto?

Rifiutare H_0 se z < -1.96 o z > 1.96; altrimenti non rifiutare H_0

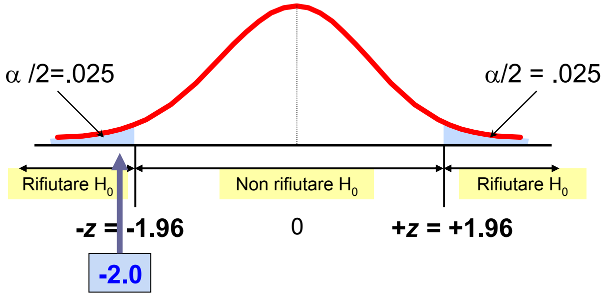


Qui, z = -2.0 < -1.96, quindi il valore della statistica test cade nella regione di rifiuto

Esempio verifica di ipotesi

(continuazione)

- Prendere una decisione ed interpretare il risultato



Siccome z = -2.0 < -1.96, rifiutiamo l'ipotesi nulla e concludiamo che ci sono sufficienti evidenze che il numero medio di TV nelle case italiane non sia uguale a 3

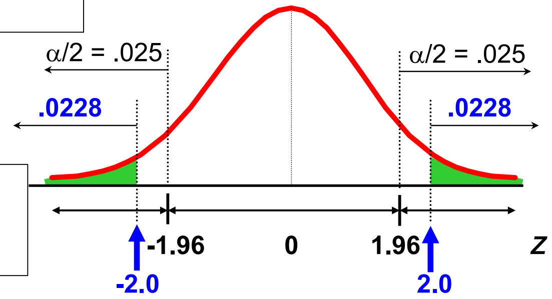
Esempio: *p*-value

- Qual è la probabilità di osservare una media campionaria di 2.84 (o un valore più lontano dalla media, in entrambe le direzioni) se la vera media è μ = 3?

$\bar{x}$ = 2.84 viene standardizzato
➔ z = -2.0

$$P(z < -2.0) = .0228$$
$$P(z > 2.0) = .0228$$

***p*-value**
= 2×.0228 = .0456



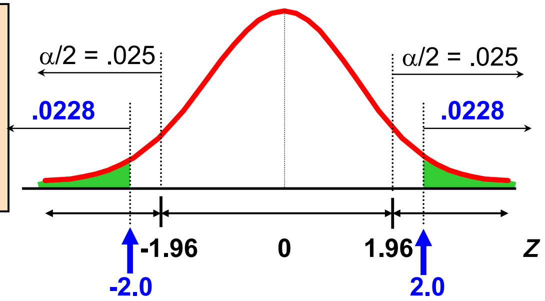
Esempio: *p*-value

(continuazione)

- Confrontare il p-value con α
  - Se p-value < α , rifiutare H_0
  - Se p-value ≥ α , non rifiutare H_0

qui: p-value = .0456
α = .05

poichè .0456 < .05, rifiutiamo l'ipotesi nulla



p-value verso livello di siginificatività

- La procedura classica dei test di significatività può essere di 2 tipi
  1. si calcola una probabilità – il *p*-value – che indica quanto sia *inusuale* ottenere il risultato osservato se l'ipotesi nulla è vera e quindi usa il *p*-value per valutare l'evidenza dei dati CONTRO l'ipotesi nulla. La decisione se rifiutare o meno l'ipotesi è lasciata al ricercatore, che esprime un giudizio in base al contesto in cui deve decidere.
  2. si fissa un limite – il livello di significatività – per questa probabilità e quindi si rifiuta automaticamente l'ipotesi nulla se il *p*-value supera questo limite.
- Di solito si preferisce il secondo approccio perchè:
  - consente al ricercatore di studiare la potenza del test nel rifiutare l'ipotesi nulla al variare dell'ipotesi alternativa e di decidere l'ampiezza campionaria di conseguenza.
  - esiste una corrispondenza biunivoca tra l'intervallo di confidenza al livello 1-α per la media e un test a due code per quella media: valori plausibili prodotti dall'intervallo di confidenza sono esattamente quelli che non sarebbero rifiutati in un test a due code con significatività α

P-value vs livello di significatività: pro e contro

- Uno svantaggio dell'approccio con livello di siginificatività fissato a priori è che il processo decisionale è fissato prima di vedere i dati e questo rende più rigido il giudizio del ricercatore:
- Un p-value grande implica un'evidenza debole, mentre un p-value piccolo implica un'evidenza forte
- Il livello di significatività può essere visto come il massimo p-value per il quale si rifiuta l'ipotesi nulla
- Storicamente, l'approccio del p-value fu sviluppato e diffuso da R. A. Fisher, scienziato brillante ed esperto analista di dati, mentre l'approccio basato sul ivello di significatività è stato sviluppato dai due statistici-matematici J. Neyman and E. Pearson. Neyman sviluppò anche la nozione di intervallo di confidenza
- La discussione su quale di questi due approcci sia il migliore è accesa ancora oggi, ma molti statistici applicati sono favorevoli ad un approccio 'combinato': fissano il livello di significatività a priori per pianificare l'esperimento, una volta osservato ilcampione, procedono alla stima per intervalli e utilizzano il p-value per valutare l'evidenza empirica in fase di test delle ipotesi.

Test sulla media, varianza incognita

- Consideriamo il carattere *X*~N(μ,σ2), con media e varianza incognite
- Si dispone di un campione casuale di ampiezza *n*
- Vogliamo sottoporre a verifica empirica un'ipotesi riguardante la media della popolazione μ, in una delle forme

  H0: μ ≤ μ 0 vs H1: μ > μ 0

  H0: μ ≥ μ 0 vs H1: μ < μ 0

  H0: μ = μ 0 vs H1: μ ≠ μ 0

- La migliore statistica test è la media campionaria; però σ non è nota ➔dovremo sostituire σ con la deviazione standard campionaria *S*

Test sulla media di una distribuzione normale con σ incognita

- statistica test *T*

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

T ha distribuzione *t* di Student con *n*−1 gdl

Consideriamo il test

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

La regola di decisione è:

$$\text{rifiutare } H_0 \text{ se } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1\,,\,\alpha}$$

Test T sulla Media (σ non nota) *(continuazione)*

- Per un test bilaterale:

Consideriamo il test

$$H_0: \mu = \mu_0 \qquad H_1: \mu \neq \mu_0$$

(Assumiamo che la popolazione abbia distribuzione normale, e la varianza della popolazione non sia nota)

La regola di decisione è rifiutare H_0 se:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1\,,\,\alpha/2} \quad \text{o} \quad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1\,,\,\alpha/2}$$

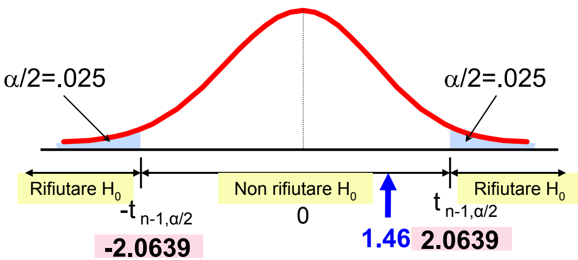
Esempio: test bilaterale sulla media di una distribuzione normale (σ non nota)

- Il costo medio di una camera doppia in hotel a Roma si dice che sia 168 € per notte.
- Un campione casuale di 25 hotel ha fornito $M(\overline{x})$= 172.50 € e s =15.40 €.
- Verifica l'ipotesi ad un livello di significatività α = 0.05.

H_0: μ = 168
H_1: $\mu \neq$ 168

Soluzione esempio: test T bilaterale

H_0: μ = 168
H_1: $\mu \neq$ 168



- **α = 0.05**
- **n = 25**
- **σ è incognita, quindi usiamo la statistica T**
- **Valore Critico:** **$t_{24\,,\,.025}$ = ± 2.0639**

$$\rightarrow t_{n-1} = \frac{\overline{x} - \mu}{s / \sqrt{n}} = \frac{172.50 - 168}{15.40 / \sqrt{25}} = 1.46$$

Non rifiutare H_0:

non c'è evidenza sufficiente che il costo medio differisca da 168 €

Test sulla media di una distribuzione normale con varianza incognita

$X_1, \ldots, X_n$, con $X \sim N(\mu, \sigma^2)$, σ^2 incognita.

- $H_0 : \mu = \mu_0$, di dimensione α: $P(x \in R \mid H_0) = \alpha$
- Statistica test $\overline{X}$, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.
- La regione critica R varia in base all'ipotesi alternativa:
  - $H_1 : \mu > \mu_0$, $R = \{\mathbf{x} : \overline{X} \geq \mu_0 + t_{(\alpha,g)} \frac{S}{\sqrt{n}}\}$
  - $H_1 : \mu < \mu_0$, $R = \{\mathbf{x} : \overline{X} \leq \mu_0 - t_{(\alpha,g)} \frac{S}{\sqrt{n}}\}$
  - $H_1 : \mu \neq \mu_0$, $R = \{\mathbf{x} : \overline{X} \leq \mu_0 - t_{(\alpha/2,g)} \frac{S}{\sqrt{n}}, \overline{X} \geq \mu_0 + t_{(\alpha/2,g)} \frac{S}{\sqrt{n}}\}$
- I quantili $t_{(\alpha,g)}$ sono relativi ad una v.c. $\mathcal{T}$ di Student con $g = n - 1$ gradi di libertà.
- Per n grande ($n > 100$) la distribuzione della v.c. $\mathcal{T}$ di Student è approssimata dalla $N(\mu, \sigma^2) \Rightarrow$ i quantili $t_{(\alpha,g)}$ possono essere sostituiti dai quantili z_α.

Test sulla Proporzione della Popolazione

- Riguarda variabili categoriche
- Due possibili risultati
  - "Successo" (una certa caratteristica è presente)
  - "Insuccesso" (la caratteristica non è presente)
- La frazione o proporzione della popolazione nella categoria dei "successi" è indicata con π
- Assumiamo che il campione sia grande

Proporzioni

(continuazione)

- La proporzione campionaria di successi viene indicata con p
  - $$p = \frac{X}{n} = \frac{\text{numero di successi nel campione}}{\text{dimensione del campione}}$$
- Quando $n\pi(1-\pi) > 9$, la distribuzione di p può essere approssimata con una distribuzione normale con media e deviazione standard:

$$E(p) = \pi \qquad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Verifica di ipotesi su proporzioni

***np*(1 – *p*) > 9**

- La distribuzione campionaria di p è approssimativamente normale, quindi usiamo la statistica test Z:

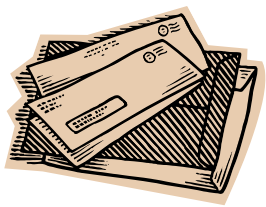
$$Z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

***np*(1 – *p*) < 9** — Non discusso in questo corso

Esempio: test Z su proporzioni

Una società di marketing afferma che il suo tasso di risposta alla mailing è dell'8%. Per verificare questa ipotesi, si considera un campione aleatorio di 500 clienti e si ottengono 25 risposte.

Verificare l'ipotesi ad un livello α = .05.



Verifica approssimazione normale: la nostra approssimazione per π è

p = 25/500 = .05

$np(1 - p) = (500)(.05)(.95) = 23.75 > 9$ ✓

Test Z sulla proporzione: soluzione

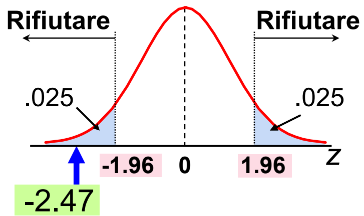
H_0: $\pi = .08$

H_1: $\pi \neq .08$

α = .05

n = 500, p = .05

Valori Critici: ± 1.96



Statistica Test:

$$z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{.05-.08}{\sqrt{\frac{.08(1-.08)}{500}}} = -2.47$$

Decisione:

Rifiutare H_0 con α = .05

Conclusione:

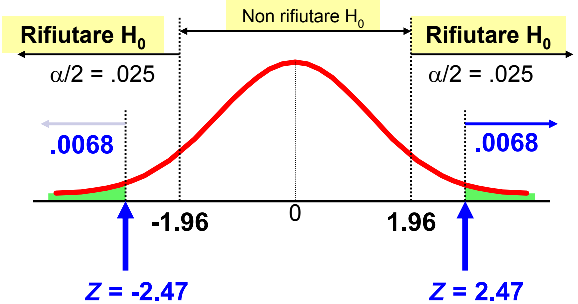
C'è evidenza empirica sufficiente per rifiutare l'ipotesi che il tasso di risposta sia dell'8%.

Soluzione p-value

(continuazione)

Calcolare il p-value e confrontare con α

(Per un test bilaterale il p-value è sempre a due code)



p-value = .0136:

$$P(Z \le -2.47) + P(Z \ge 2.47)$$
$$= 2(.0068) = 0.0136$$

Rifiutare H_0 perchè p-value = .0136 < α = .05

Potenza del Test

- Ricordare i possibili risultati della verifica di ipotesi:

Decisione	Stato di Natura: H_0 Vera	Stato di Natura: H_0 Falsa
Non Rifiutare H_0	**No errore** ($1 - \alpha$)	**Errore di II Tipo** (β)
Rifiutare H_0	**Errore di I Tipo** (α)	**No Errore** ($1 - \beta$)

Legenda: Risultato (Probabilità)

- β rappresenta la probabilità dell'errore di II tipo
- $1 - \beta$= P($x \notin R$| H_1) è la potenza del test sotto H_1
  ➔probabilità di rifiutare l'ipotesi nulla quando è falsa

Potenza = P($x \notin R$) probabilità di rifiutare l'ipotesi nulla

Errori e potenza del test

- Il livello di significatività α è la probabilità di rifiutare l'ipotesi nulla quando è vera, cioè la probabilità di commettere un errore di I tipo.
- Un errore di II tipo consiste nel NON rifiutare l'ipotesi nulla quando è FALSA. La probabilità di commetter errore di II tipo si indica con β e in generale non è nota.
- La potenza del test è la probabilità di rifiutare l'ipotesi nulla. Sotto H_1 la potenza del test è pari a $1-\beta$, mentre sotto H_0 è pari al livello di significatività α, che è anche la probabilità di errore di I tipo.
- Il calcolo della potenza del test richede la conoscenza del valore di μ sotto H_1.Spesso l'ipotesi alternativa è composta e a volte anche bidirezionale, quindi non c'è un unico valore di μ sotto H_1.
- Di solito si calcola la potenza del test per tutti i valori di interesse e per vari valori della dimensione campionaria, così da scegliere il valore di n appropriato per garantire una potenza del test adeguata per alternative di particolare interesse per lo studio che si sta conducendo. NB: La potenza è una proprietà del test (regola decisionale) ➔ i dati campionari non servono!

Errore di secondo tipo

Assumiamo che la popolazione abbia distribuzione normale e la varianza della popolazione sia nota. Consideriamo il test

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

La regola di decisione è:

$$\text{Rifiutare } H_0 \text{ se } z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$$

$$\text{Rifiutare } H_0 \text{ se } \bar{x} = \bar{x}_c > \mu_0 + z_\alpha \sigma / \sqrt{n}$$

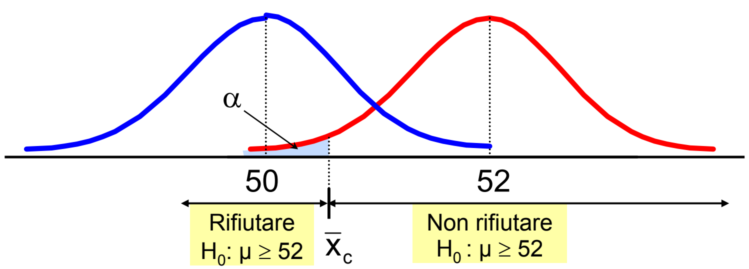
Se l'ipotesi nulla è falsa e la vera media è μ*, allora la probabilità dell'errore di secondo tipo è

$$\beta = P(\bar{X} < \bar{x}_c \mid \mu = \mu^*) = P\left(Z < \frac{\bar{x}_c - \mu^*}{\sigma / \sqrt{n}} \right)$$

Esempio errore di II tipo

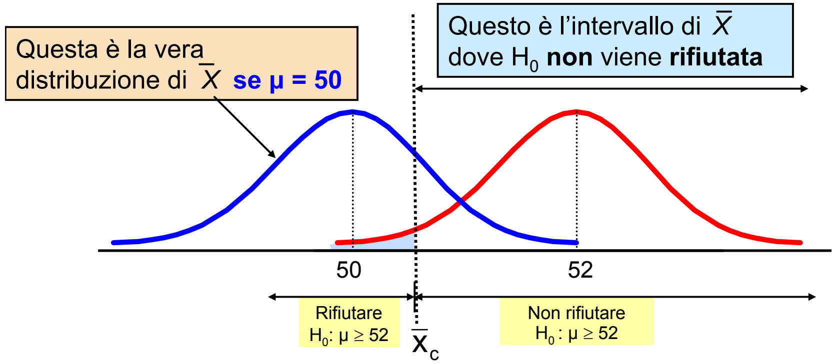
- Errore di II tipo corrisponde alla probabilità di NON rifiutare un'ipotesi nulla falsa

Supponiamo che **H_0: μ ≥ 52** non venga rifiutata quando infatti la vera media è **μ* = 50**



Esempio errore di II tipo *(continuazione)*

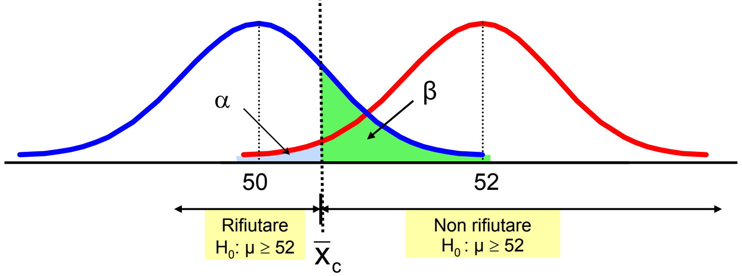
- Supponiamo che H_0: μ ≥ 52 non venga rifiutata quando infatti la vera media è μ* = 50



Esempio Errore di Secondo Tipo *(continuazione)*

- Supponiamo che H_0: μ ≥ 52 non venga rifiutata quando infatti la vera media è μ* = 50

Qui, $\beta = P(\bar{X} \geq \bar{x}_c | \mu^* = 50)$



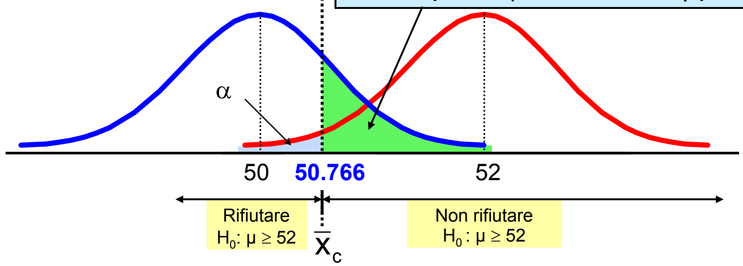
Calcolo di β

- Supponiamo n = 64 , σ = 6 , e **α** = .05

$$\bar{x}_c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = 52 - 1.645 \frac{6}{\sqrt{64}} = \boxed{50.766}$$

(per H_0 : μ ≥ 52)

Quindi β = P($\bar{X} \geq$ 50.766 | μ*=50)

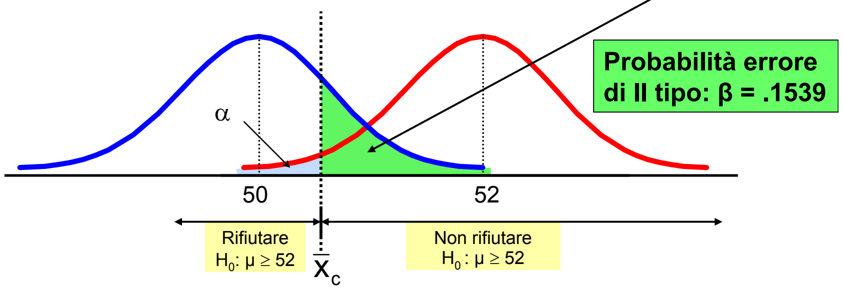


Calcolo di β

(continuazione)

- Supponiamo *n* = 64 , *σ* = 6 , e α = .05

$$P(\overline{X} \geq 50.766 | \mu^* = 50) = P\left(Z \geq \frac{50.766 - 50}{6/\sqrt{64}} \right) = P(Z \geq 1.02) = \boxed{0.1539}$$



Esempio potenza del test

Se vale H_1: μ* = 50:

- Probabilità dell'errore di II tipo: β = 0.1539
- Potenza del test : 1 – β = 1 – 0.1539 = 0.8461

Legenda: **Risultato** **(Probabilità)**

	Stato di Natura	
Decisione	H_0 Vera	H_0 Falsa
Non Rifiutare H_0	**No errore** **1 - α = 0.95**	**Errore di II tipo** **β = 0.1539**
Rifiutare H_0	**Errore di I tipo** **α = 0.05**	**No errore** **1 - β = 0.8461**

(Il valore di β e la potenza saranno diversi per diversi valori di μ*)

Test su due campioni

Test su due campioni

- Medie delle Popolazioni, Campioni Dipendenti
- Medie delle Popolazioni, Campioni Indipendenti
- ~~Proporzioni delle Popolazioni~~
- ~~Varianze delle Popolazioni~~

Esempi:

- Stesso gruppo prima vs. dopo trattamento
- Gruppo 1 vs. indipendente Gruppo 2
- ~~Proporzione 1 vs. Proporzione 2~~
- ~~Varianza 1 vs. Varianza 2~~

Confronto tra le medie di due popolazioni

Consideriamo un carattere rilevato su due popolazioni (*X* sulla prima, *Y* sulla seconda)

Il confronto tra due medie si può esprimere in termini di **differenza**

$\begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$ equivale a $\begin{cases} H_0 : \mu_X - \mu_Y \geq 0 \\ H_1 : \mu_X - \mu_Y < 0 \end{cases}$ (test coda sinistra)	$\begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$ equivale a $\begin{cases} H_0 : \mu_X - \mu_Y \leq 0 \\ H_1 : \mu_X - \mu_Y > 0 \end{cases}$ (test coda destra)	$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$ equivale a $\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$ (test bidirezionale)

Differenza tra medie di campioni dipendenti

- Confrontiamo le medie di 2 popolazioni appaiate
  - Misure ripetute (prima/dopo)
  - Dati appaiati

	Popolazione 1	Popolazione 2
Media popolazione	μ_X	μ_Y
Dev. Std. popolazione	σ_X	σ_Y
Ampiezza campione	n	n

***n* coppie**

***2n* osservazioni**

- Assunzioni:
  - Entrambe le popolazioni hanno distribuzione Normale
  - Campione casuale di coppie

Differenza tra due medie (campioni dipendenti)

- Idea: calcolare le differenze … $D_i = X_i - Y_i$
- … e poi calcolare la differenza media $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$
- Infatti, le differenze D_i sono n v.a. indipendenti e identicamente distribuite con valore atteso

$$E(D_i) = E(X_i - Y_i) = E(X_i) - E(Y_i) = \mu_X - \mu_Y$$

- Dalle proprietà della media campionaria segue che

$$E(\bar{D}) = E(D_i) = \mu_X - \mu_Y$$

$$\bar{D} \sim Normale$$

Questa è la differenza oggetto di ipotesi

Differenza tra due medie (campioni dipendenti)

- Per sottoporre a test l'ipotesi nulla che la differenza tra le medie sia d_0 si usa la statistica test

$$T_0 = \frac{\bar{D} - d_0}{S_d / \sqrt{n}} \overset{H_0}{\sim} t_{n-1}$$

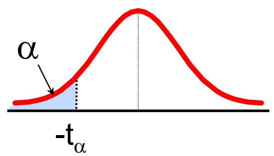
Essendo passati dalle 2*n* osservazioni alle *n* differenze si ha un test *t* come per la media di una singola popolazione

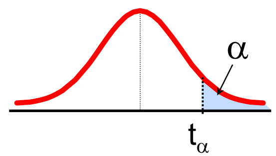
- d_0 = differenza tra medie sotto H_0, p.e H_0: $\mu_X - \mu_Y = d_0$
- S_d = deviazione standard campionaria delle differenze
- n = dimensione del campione (n. di coppie = n. di differenze)

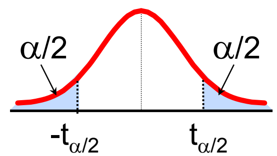
Differenza tra due medie(campioni dipendenti)

Campioni dipendenti: test per verificare se la differenza è nulla ($d_0 = 0$)

$\begin{cases} H_0: \mu_X - \mu_Y \geq 0 \\ H_1: \mu_X - \mu_Y < 0 \end{cases}$ (test coda sinistra)	$\begin{cases} H_0: \mu_X - \mu_Y \leq 0 \\ H_1: \mu_X - \mu_Y > 0 \end{cases}$ (test coda destra)	$\begin{cases} H_0: \mu_X - \mu_Y = 0 \\ H_1: \mu_X - \mu_Y \neq 0 \end{cases}$ (test bidirezionale)







Rifiutare H_0 se $t_0 < -t_{n-1,\alpha}$

Rifiutare H_0 se $t_0 > t_{n-1,\alpha}$

Rifiutare H_0 se $t_0 < -t_{n-1,\alpha/2}$ o $t_0 > t_{n-1,\alpha/2}$

dove $t_0 = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$

Campioni dipendenti: esempio

I rappresentati della vostra azienda partecipano ad un workshop sul “servizio ai clienti”. La partecipazione ha o non ha prodotto un cambiamento nel numero di reclami?

Si dispone dei seguenti dati:

Rappresent.	Numero di Reclami: Prima (1)	Dopo (2)	(2) - (1) Differenze, d_i
C.B.	6	4	- 2
T.F.	20	6	-14
M.H.	3	2	- 1
R.K.	0	0	0
M.O.	4	0	- 4
			-21

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = -4.2$$

$$s_d = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}} = 5.67$$

Campioni dipendenti: soluzione

- La partecipazione al workshop ha prodotto un cambiamento nel numero di reclami (ad $\alpha = 0.01$)?

$$\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$$

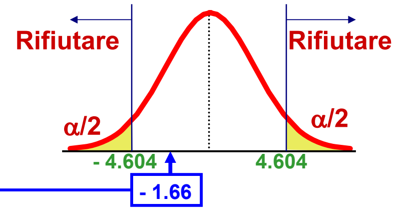
$\alpha = .01$ $\quad \bar{d} = -4.2$

Valore Critico= ± 4.604

g.d.l. = n - 1 = 4

Statistica Test:

$$t_0 = \frac{\bar{d} - d_0}{s_d / \sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{5}} = \boxed{-1.66}$$



Decisione: non rifiutare H_0

(la statistica T non è nella regione di rifiuto)

Conclusione: Non c’è un cambiamento significativo nel numero di reclami.

Differenza tra due medie (campioni indipendenti)

Medie delle popolazioni, campioni indipendenti

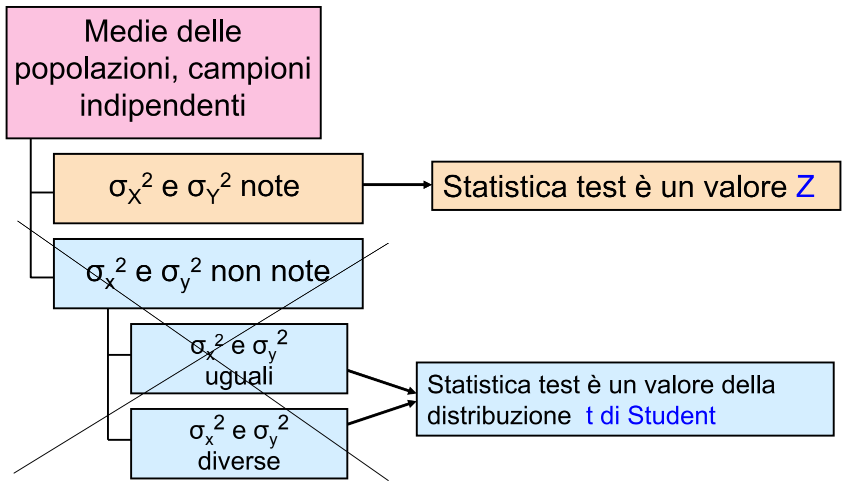
Obiettivo: Sottoporre a test un’ipotesi riguardante la differenza fra le medie di due popolazioni, es.

$$H_0: \mu_X - \mu_Y = d_0$$

	Popolazione 1	Popolazione 2
Media popolazione	μ_X	μ_Y
Dev. Std. popolazione	σ_X	σ_Y
Ampiezza campione	n_X	n_Y

Campioni indipendenti: i valori di un campione non influenzano i valori dell’altro → si hanno n_X+n_Y variabili aleatorie mutuamente indipendenti

Differenza tra due medie (campioni indipendenti)



Differenza tra due medie (campioni indipendenti)

Medie delle popolazioni, campioni indipendenti

- σ_X^2 e σ_Y^2 note *
- σ_X^2 e σ_Y^2 non note

Assunzioni:

1. Due campioni casuali indipendenti (in generale i due campioni hanno dimensione diversa, ma possono anche avere identica dimensione)
2. Entrambe le popolazioni hanno distribuzione Normale
3. Le varianze delle popolazioni sono note a priori

Quando entrambi i campioni sono grandi (diciamo >50) il test è approssimativamente valido anche senza le assunzioni 2 e 3

E' molto raro che le entrambe le varianze siano note a priori, per cui in pratica questo test è utile nelle situazioni in cui si dispone di grandi campioni

Differenza tra due medie (campioni indipendenti e grandi)

- Se i due campioni sono indipendenti e σ_X^2 e σ_Y^2 sono note, la varianza della differenza è

$$\sigma^2_{\overline{X}-\overline{Y}} = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

- Se entrambe le popolazioni sono Normali, la variabile aleatoria

$$Z = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}}$$

ha distribuzione Normale standard

Differenza tra due medie (campioni indipendenti e grandi)

La statistica test per $\mu_X - \mu_Y$ è :

$$Z_0 = \frac{\left(\overline{X}-\overline{Y}\right)-d_0}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \overset{H_0}{\sim} N(0,1)$$

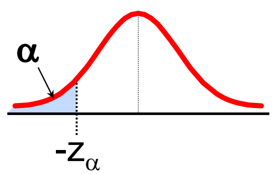
d_0 = differenza ipotizzata, es. H_0: $\mu_X - \mu_Y = d_0$ o H_0: $\mu_X - \mu_Y \geq d_0$

Quando le varianze NON sono note a priori vanno sostituite con le varianze campionarie: se entrambi i campioni sono grandi la statistica test ha ancora, con buona approssimazione, una distribuzione Normale standard

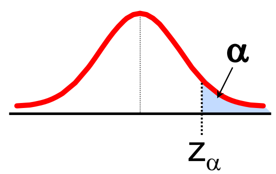
Differenza tra due medie (campioni indipendenti e grandi)

Campioni indipendenti: test per verificare se la differenza è nulla ($d_0 = 0$)

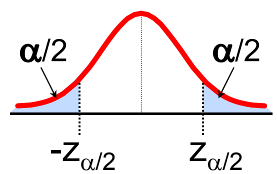
$\begin{cases} H_0: \mu_X - \mu_Y \geq 0 \\ H_1: \mu_X - \mu_Y < 0 \end{cases}$ (test coda sinistra)	$\begin{cases} H_0: \mu_X - \mu_Y \leq 0 \\ H_1: \mu_X - \mu_Y > 0 \end{cases}$ (test coda destra)	$\begin{cases} H_0: \mu_X - \mu_Y = 0 \\ H_1: \mu_X - \mu_Y \neq 0 \end{cases}$ (test bidirezionale)



Rifiutare H_0 se $z_0 < -z_\alpha$



Rifiutare H_0 se $z_0 > z_\alpha$



Rifiutare H_0 se $z_0 < -z_{\alpha/2}$ o $z_0 > z_{\alpha/2}$

Campioni indipendenti: esempio

- Il direttore di una catena di negozi riceve lamentele sui tempi di consegna del negozio *X*; a tal fine decide di effettuare un confronto con i tempi di consegna del negozio *Y* per appurare se vi è abbastanza evidenza empirica per affermare che *X* ha tempi più lunghi di *Y*

$$\begin{cases} H_0 : \mu_X - \mu_Y \le 0 \\ H_1 : \mu_X - \mu_Y > 0 \end{cases}$$

- I dati sono due campioni casuali di tempi di consegna, espressi in giorni lavorativi

	Negozio *X*	Negozio *Y*
Ampiezza campione	80	60
Media campionaria	12.56	9.41
Varianza campionaria	5.22	2.87

Campioni indipendenti: soluzione

- I campioni sono indipendenti ed entrambi grandi → anche senza specificare la distribuzione di *X* e *Y* e anche se le varianze non sono note a priori si può usare la statistica test z_0 per campioni indipendenti

$$z_0 = \frac{(12.56 - 9.41) - 0}{\sqrt{\frac{5.22}{80} + \frac{2.87}{60}}} = 9.3672$$

- Al livello di significatività dell'1% la regione di rifiuto è a destra con $z_\alpha = 2.33$
- Poiché 9.37>2.33 si rifiuta l'ipotesi nulla: i tempi di consegna del negozio *X* sono significativamente maggiori di quelli del negozio *Y*

Intervalli di confidenza e test delle ipotesi

- Il test delle ipotesi ha la forma

$$\text{statistica test} = \frac{\text{statistica} - \text{parametro}}{\text{dev.std.}(\text{statistica})}$$

- L'intervallo di confidenza ha la forma:

$$\text{statistica} \pm \text{valore critico} * \text{dev.std.}(\text{statistica})$$