Statistica I corso A-E

Lezione XXXII

Statistica Inferenziale

Prova d'ipotesi

Noi utilizzeremo nel calcolo degli intervalli di confidenza e nelle prove di ipotesi le seguenti statistiche test

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

σ è nota

$\leftarrow$ statistica Z

$$t_{n-1} = \frac{\bar{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim t\text{-Student}$$

S deviazione campion[aria]

S^2 = Varianza campionaria

$$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2} \text{ - Chi quadro}$$

Cos'è un'ipotesi?

- Un'ipotesi è una affermazione (assunzione) circa il parametro della popolazione:
  - media della popolazione

    Esempio: In questa città, il costo medio della bolletta mensile per il cellulare è μ = $42

  - proporzione della popolazione

    Esempio: In questa città, la proporzione di adulti con il cellulare è p = .68

L'ipotesi Nulla, H_0

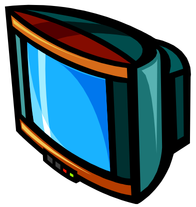
- Rappresenta l'ipotesi (numerica) che deve essere verificata

> Esempio: Il numero medio di TV nelle case americane è uguale a tre $(H_0 : \mu = 3)$

- Si riferisce sempre al parametro della popolazione, non alla statistica campionaria

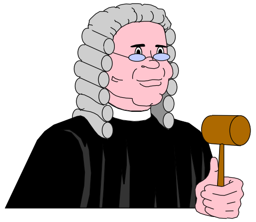
$$H_0 : \mu = 3$$

$$\cancel{H_0 : \overline{X} = 3}$$



L'ipotesi Nulla, H_0

- Iniziamo con l'assunzione che l'ipotesi nulla sia vera
  - Simile alla nozione di innocenza a meno che venga dimostrata la colpevolezza



- **Si riferisce allo status quo**
- Contiene sempre "=" , "≤" o "≥"
- Può o può non essere rifiutata

L’Ipotesi Alternativa, H_1

- È l’opposto dell’ipotesi nulla
  - e.g., Il numero medio di TV nelle case americane non è uguale a 3 (H_1: $\mu \neq 3$)

- **Sfida lo status quo**

- Può o può non essere sostenuta

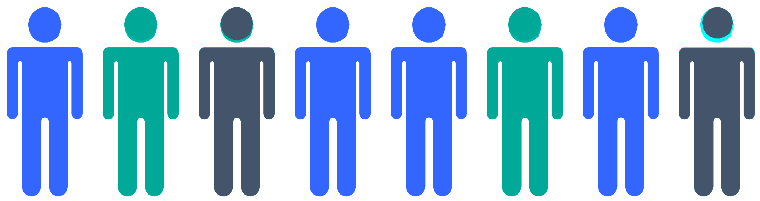
- È generalmente l’ipotesi che il ricercatore cerca di dimostrare contro lo status quo

Processo della Verifica di Ipotesi

Affermazione: l'età media della popolazione è 60.

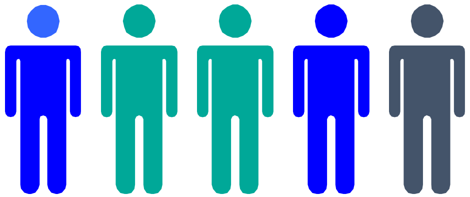
(Ipotesi nulla:

H_0: μ = 60)



Popolazione

Adesso selezioniamo un campione casuale



Campione

È $\overline{x_n}$=35 probabile se μ = 60?

Supponiamo l'età media del campione sia 20: $\overline{X_n}$ = 35

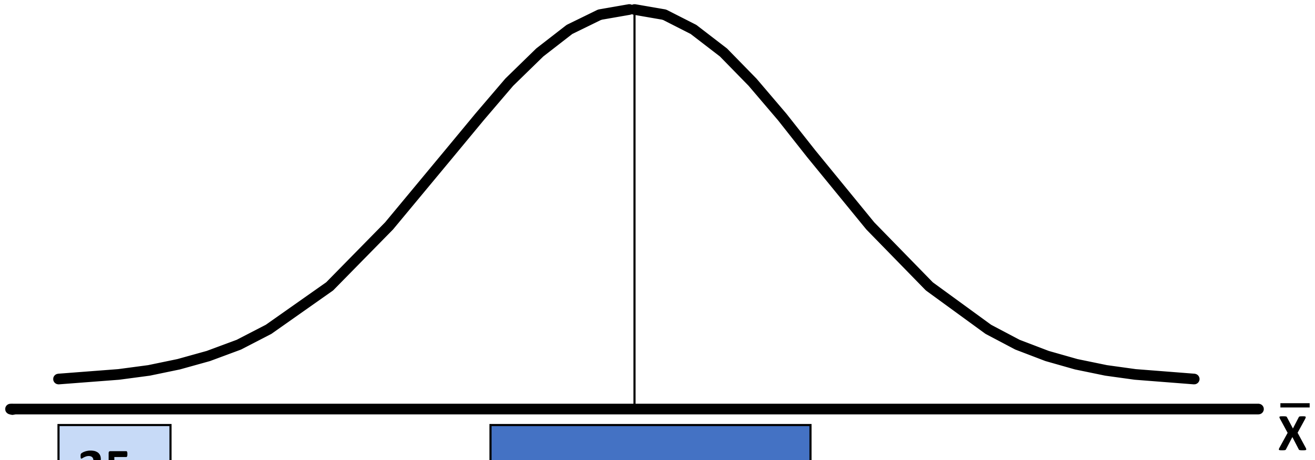
Se non è probabile,

RIFIUTA

Ipotesi Nulla

Motivazione per Rifiutare H_0

Distribuzione Campionaria di $\bar{X}$



35

$\mu = 60$

Se H_0 è vera

Se non è probabile che si possa ottenere questo valore per la media campionaria...

... se infatti questa è la media della popolazione...

... allora rifiutiamo l'ipotesi nulla che

$\mu = 60$.

Livello di Significatività e la Regione di Rifiuto

Livello di significatività = α

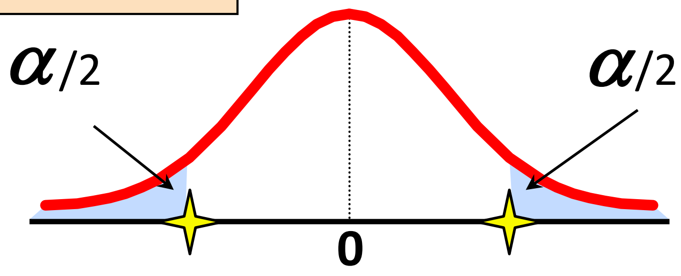
Representa i valori critici

Regione di rifiuto è ombreggiata

H_0: $\mu = 3$

H_1: $\mu \neq 3$

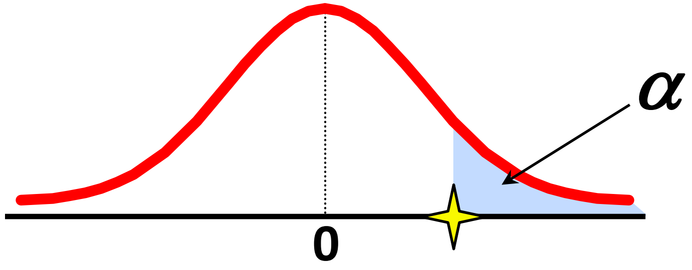
Test a due code



H_0: $\mu \leq 3$

H_1: $\mu > 3$

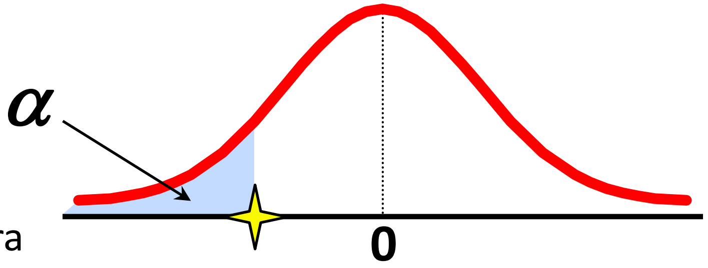
Test coda destra



H_0: $\mu \geq 3$

H_1: $\mu < 3$

Test coda sinistra



Livello di Significatività, α

- **Definisce i valori della statistica campionaria che sono improbabili se l'ipotesi nulla è vera**
  - Definisce la regione di rifiuto della distribuzione campionaria
- È indicato con α (livello di significatività)
  - Valori comuni sono 0.01, 0.05, or 0.10
- Fissato a priori dal ricercatore
- Fornisce il valore critico del test

Errori nel Processo Decisionale

- **Errore di Primo Tipo o prima specie**
  - Rifiutare un'ipotesi nulla quando quest'ultima è vera
  - Considerato un tipo di errore molto serio

La probabilità dell'errore di prima specie è α

- α **è anche detto** livello di significatività del test
- α **è scelto a priori** dal ricercatore

Errori nel Processo di Decisione

- **Errore di Secondo Tipo o seconda specie**
  - Non rifiutare l'ipotesi nulla quando quest'ultima è falsa

La probabilità dell'errore di secondo tipo è β

La quantità 1- β è detta potenza del test

Il termine potenza del test è dovuto al fatto che

1- β è la probabilità di rigettare l'ipotesi nulla quando questa non è vera

Risultati e Probabilità

Legenda:
Risultato
(Probabilità)

Possibili Risultati Verifica di Ipotesi

	Stato di Natura	
Decisione	H_0 Vera	H_0 Falsa
Non Rifiutare H_0	**No errore** **$(1 - \alpha)$**	**Errore Secondo Tipo** **(β)**
Rifiutare H_0	**Errore Primo tipo** **(α)** Livello di significatività del test	**No Errore** **$(1 - \beta)$** Potenza del test

Riassumendo:

Errore di prima specie: è l'errore che si commette rifiutando l'ipotesi nulla H_0 quando H_0 è vera.

Il livello di significatività è **la probabilità di rifiutare $\boldsymbol{H_0}$ quando $\boldsymbol{H_0}$ è vera**:

$\mathrm{P}(\text{ rifiuto } \mathrm{H}_0 \mid \mathrm{H}_0) = \alpha$ (α è il livello di significatività);

Errore di seconda specie: è l'errore che si commette non rifiutando l'ipotesi nulla quando questa non è vera.

$$\mathrm{P}\,(\,non\,rifiuto\;\mathrm{H}_0 \mid \mathrm{H}_1) = \beta$$

La probabilità di commettere un errore di seconda specie si indica con β.

N.B: Come traduciamo in un evento l'affermazione « rifiuto H_0 » e «$non\,rifiuto\;\mathrm{H}_0$»? Cosa implica nella pratica « $\mid \mathrm{H}_0$»(dato H0) oppure «$\mid \mathrm{H}_1$» (dato H1)?

Domanda: Come traduciamo in un evento l'affermazione « rifiuto H_0 » e «$non\ rifiuto$ H_0»? Cosa implica nella pratica « | H_0»(dato H0) oppure «| H_1» (dato H1)? Supponiamo che H_0 sia $\theta = \theta_0$ e H_1 sia $\theta = \theta_1$

- Errore di prima specie: rifiutare H_0 quando H_0 <u>è vera</u> $\mathrm{P}(\text{ rifiutare } \mathrm{H}_0 \mid \mathrm{H}_0) = \alpha$ (livello di significatività)
- Errore di seconda specie: non rifiutare H_0 quando H_0 <u>**non è vera**</u> $\mathrm{P}\left(\text{ non rifiutare } \mathrm{H}_0 \mid \mathrm{H}_1\right) = \beta$

La prova di ipotesi ha come obiettivo minimizzare entrambi gli errori. Il test che minimizza α (alfa) e β (beta) fu formulato da **Neyman e Pearson** come segue: la regione di rifiuto deve avere la forma:

$$R : \frac{L(\theta_0)}{L(\theta_1)} < k$$

La costante k è scelta in modo che la probabilità che il rapporto tra la funzione di verosimiglianza calcolata in θ_0 (teta zero) e la funzione di verosimiglianza calcolata in θ_1 (teta uno) minore sia uguale al livelllo di significatività α (alfa): $\mathrm{P}\left(\frac{\mathrm{L}(\theta_0)}{\mathrm{L}(\theta_1)} < \mathrm{k} \,\middle|\, \boldsymbol{H_0}\right) = \alpha$

Oss: Ricordiamo che la funzione di verosimiglianza è la probabilità congiunta di osservare le osservazioni del campione fissato il parametro. Se $\boldsymbol{H_0}$ vera dovrei avere $\mathrm{L}(\theta_0)$ maggiore di $\mathrm{L}(\theta_1)$

Come applicare la regola di decisione in pratica? Prova di ipotesi della media di una popolazione normale quando la varianza è nota

Il test di Neyman-Pearson nel caso di prova di ipotesi per la media di una popolazione normale, $X \sim N(\mu, \sigma^2)$,

permette di dimostrare che la regione di rifiuto è una diseguaglianza sulla media campionaria.

Esempio Regione Rifiuto: $\overline{X}_n < x_c$ **oppure**, $\overline{X}_n > x_c$, **oppure** $\overline{X}_n < x_L \vee \overline{X}_n > x_U$, dove x_c è detto valore critico e così pure x_L e x_U

Supponiamo che la varianza sia nota mentre la media sia incognita.

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

$$H_1 \ ---> \mu = \mu_1$$

In generale si fissa un livello di significatività "α" e si cerca di capire come è fatta la regione di rifiuto. Il tipo di diseguaglianza della regione di rifiuto dipende dal tipo di diseguaglianza tra le due ipotesi

PRIMO CASO: TEST UNILATERALE DESTRO

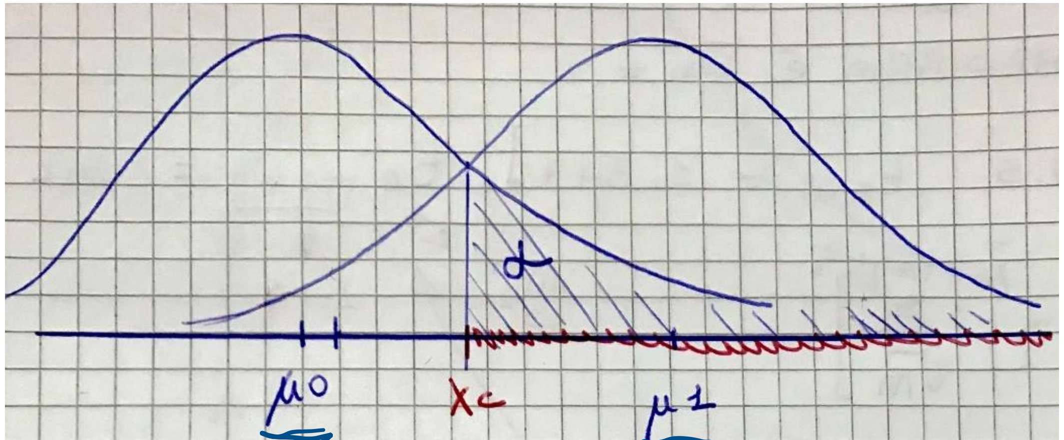
Immaginiamo ora che $\mu_0 < \mu_1$, quindi di questo tipo:

La regione di rifiuto si ha quando la media campionaria è maggiore di un certo valore x_c .

$$R: \overline{X}_n > x_c$$ chi è

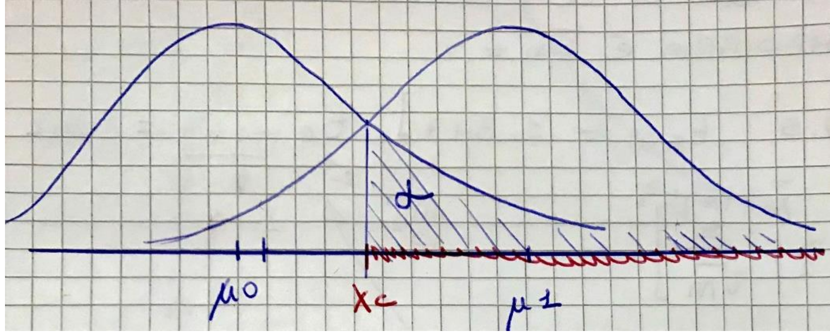
La zona di rifiuto si trova nella parte che contiene l'ipotesi alternativa ovvero quando vado a finire troppo a destra di μ_0. Osserviamo che la regione di rifiuto deve contenere l'ipotesi alternativa.

Oss: x_c è fissato e una volta fissato rifiuteremo l'ipotesi nulla se il valore osservato della media campionaria cade nella zona di rifiuto.
Il valore critico $\mathbf{x}_c$ viene fissato in base al livello di significatività α.



la zona di rifiuto è quella rossa ed è nella direzione dove si trova l'ipotesi alternativa

COME CALCOLIAMO IL VALORE CRITICO? $R: \bar{X}_n > X_c$



Il valore critico x_c deve essere determinato in modo che:

$$P(R\, H_0 | H_0) = P\left(\bar{X}_n > x_c \mid H_0\right) = \alpha$$

Sappiamo che la media campionaria è una normale quindi possiamo standardizzarla ma rispetto a quale media della popolazione? Dobbiamo usare

$\mu = \mu_0$ (ipotesi H_0) oppure $\mu = \mu_1$(ipotesi H_1)? La probabilità da calcolare è condizionata alla ipotesi nulla «$\mid H_0$» cioé dobbiamo assumere che l'ipotesi nulla sia vera quindi la media di $\bar{X}_n$ è μ_0 . Avremo dunque $\bar{X}_n \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$.

Usiamo la statistica Z: $\boldsymbol{Z_n = \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}} \sim N(0,1)$.

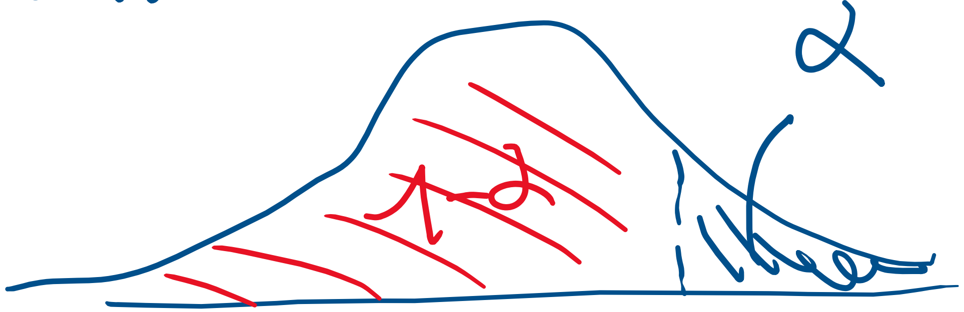
Procedura: dato il livello di significatività α troviamo il valore critico della statistica Z $\boldsymbol{z_c}$ che soddisfa $P(\boldsymbol{Z_n} > z_c) = \alpha$. Dal disegno si comprende che $\boldsymbol{z_c} = \boldsymbol{z_{1-\alpha}}$ quindi troviamo il valore critico $\boldsymbol{x_c}$ risolvendo $\frac{x_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = z_{1-\alpha}$ cioé $\boldsymbol{x_c = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}}$

$$P(\bar{X}_n > x_c \mid H_0) = \alpha$$

devo trovare x_c?
dato α

$$\alpha = P(\bar{X}_n > x_c \mid H_0) = P\Big(\underbrace{\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}}_{Z \sim N(0,1)} > z_c\Big) = \alpha$$

dato α trovo $z_c = z_{1-\alpha}$ usando le tavole



$z_c = z_{1-\alpha}$

$$\frac{x_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = z_{1-\alpha}$$

$x_c - \mu_0 = \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$

$$\boxed{x_c = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}}$$

$R: \bar{X}_n > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$

Come determino la potenza del test dato $\boldsymbol{x_c}$?

La **potenza del test è** 1- β dove β è la probabilità dell'errore di seconda specie

$$P\,(\,non\,rifiuto\;H_0|\,H_1)\;=\beta$$

Nel caso di TEST UNILATERALE DESTRO la **regione di rifiuto è $\bar{X}_n > x_c$** mentre quella di accettazione (non rifiuto) è $\bar{X}_n \le x_c$

$$P\,(\,non\,rifiuto\;H_0|\,H_1)\;=\beta = P\,(\bar{X}_n \le\; x_c\,|\,H_1) = P\big(Z \le z_q\big) = P(Z \le z_\beta) = F(z_\beta)$$

$$1-\beta = P\,(\bar{X}_n >\; x_c\,|\,H_1) = P(Z > z_q)$$

Dove il valore critico di Z è calcolato supponendo che la media sia $\boldsymbol{z_q = \frac{x_c-\mu_1}{\frac{\sigma}{\sqrt{n}}}}$

ATTENZIONE il pedice di z_q è β perché β è l'area lasciata alla sinistra di z_q

Esercizio 6

Da una popolazione distribuita in modo normale avente media (μ) incognita e varianza $\sigma^2=16$ viene estratto un campione casuale di $n=64$ elementi. Date le seguenti due ipotesi:

$$H_0 : \mu=20 \qquad\qquad H_1 : \mu = 22$$

a) Definire la probabilità di commettere un errore di prima specie e quindi determinare il punto critico e la regione di rifiuto per l'ipotesi nulla al livello di significatività $\alpha=0.05$;

b) Definire la potenza del test e quindi calcolarne il valore nella situazione considerata.

(Si tenga conto che alcuni quantili della distribuzione Normale standardizzata Z sono i seguenti: $z_{0.0091}=-2.36$, $z_{0.950}=1.640$, $z_{0.965}=1.812$, $z_{0.975}=1.96$, $z_{0.985}=2.17$, $z_{0.99}=2.326$, $z_{0.995}=2.576$).

Dati problema n = 64 campione casuale iid

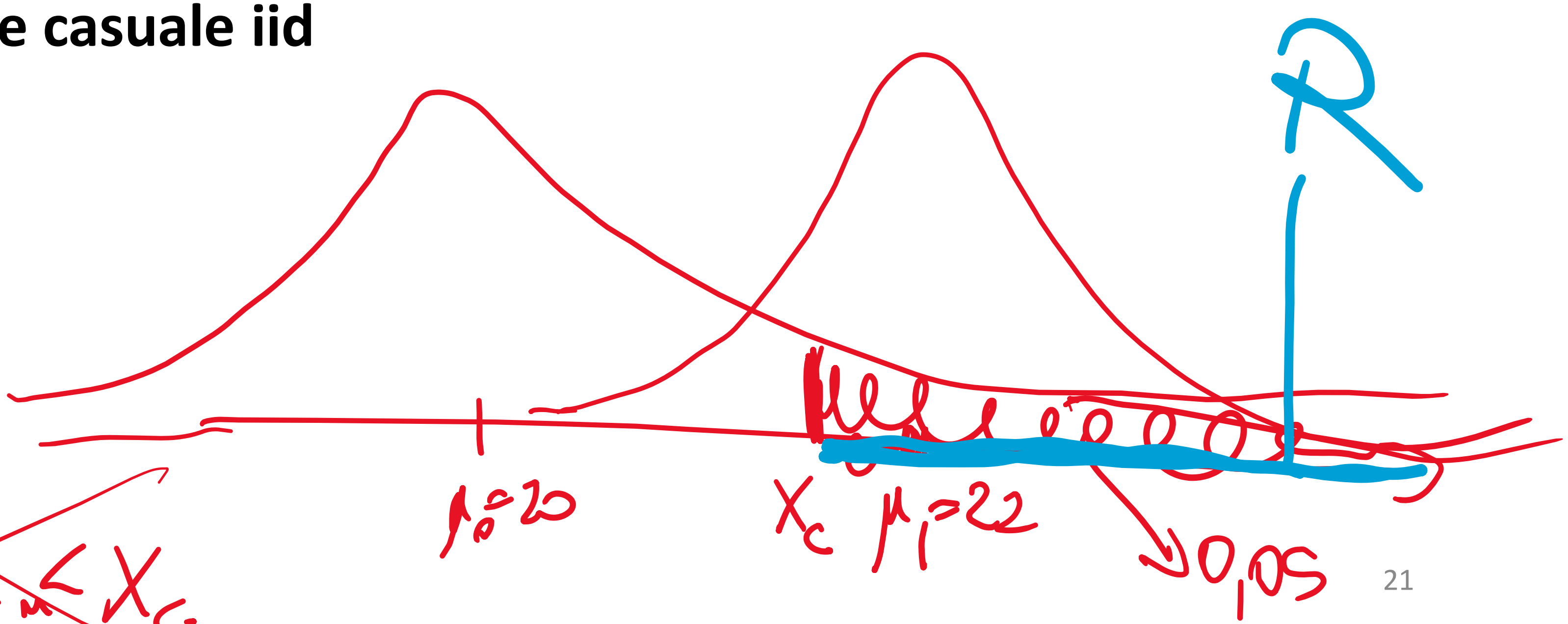
$X \sim N(\mu, \sigma^2) \qquad \sigma^2 = 16$

Ipotesi

$H_0 : \mu_0 = 20$

$H_1 : \mu_1 = 22$

$\mu_1 > \mu_0$



R $\quad \bar{X}_n > X_c \quad$ opp. $\cancel{\bar{X}_n < X_c}$

$P(\text{RIFIUTARE } H_0 \mid H_0) = \alpha$ PROBABILITÀ DI RIFIUTARE H_0 QUANDO È VERA.

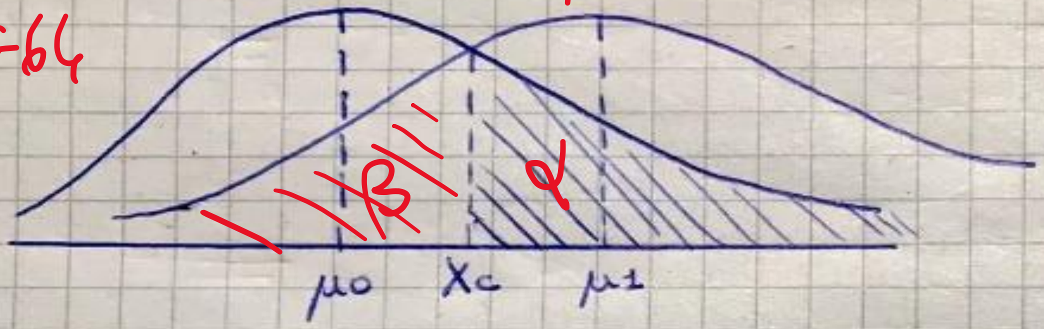
$\mu_1 > \mu_0$

$\mu_0 = 20$

$\mu_1 = 22$

$\sigma^2 = 16 \quad \sigma = 4$

$n = 64$

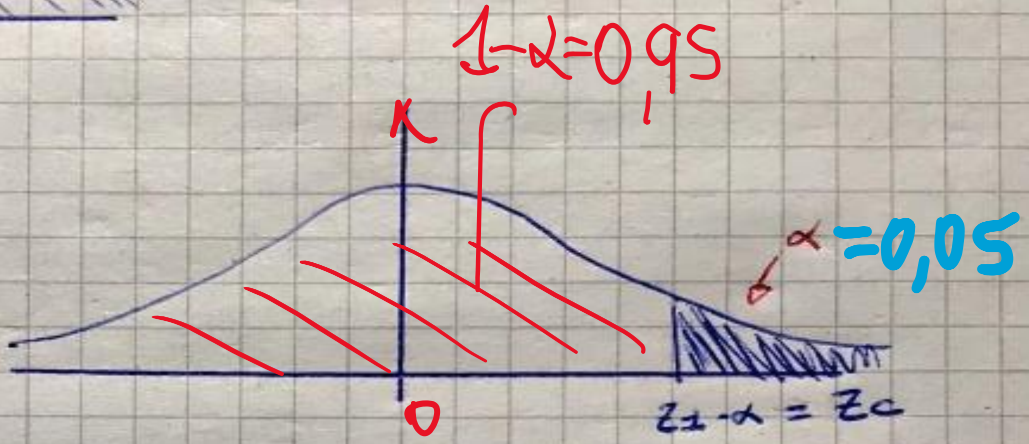


$R: \bar{X}_n > X_c$

$P(\bar{X}_n > X_c) = \alpha = 0{,}05$

NORMALIZZAZIONE

$$\frac{X_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = z_{1-\alpha}$$



$z_{1-\alpha} = 1 - 0{,}05$

$$\frac{X_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{X_c - 20}{\frac{4}{8}} = z_{0,95} = 1{,}64$$

$\frac{\sigma}{\sqrt{n}} = \frac{4}{8}$

$X_c = 1{,}64 \cdot \frac{1}{2} + 20 = 20{,}22$

$R: \bar{X}_n > 20{,}22$

La probabilità di commettere un errore di prima specie è la probabilità di rifiutare H_0 quando H_0 è vero

$$P(\text{Rifiutare } H_0 \mid H_0) = \alpha$$

La potenza del test è la probabilità di rifiutare H_0 quando H_0 è falsa $P(\text{Rifiuto } H_0 \mid H_1) = 1-\beta$

Modo alternativo la potenza del Test è
$1-\beta$ dove β è la probabilità dell'errore
di 2ª specie ovvero la probabilità di
non rifiutare H_0 quando H_0 è falsa

$$\beta = P(\text{Non rifiuto } H_0 \mid H_1)$$

$x_c = 20{,}22$

$$\beta = P(\bar{X}_n \le x_c \mid H_1) = P(Z \le z_q) = P(Z \le z_\beta) = F(z_\beta)$$

$$z_q = \frac{x_c - \mu_1}{\sigma/\sqrt{n}} = \frac{20{,}22 - 22}{4/8} = -2.36$$

L'esercizio fornisce $z_{0.0091} = -2.36 \longrightarrow \beta = 0.0091 \longrightarrow$

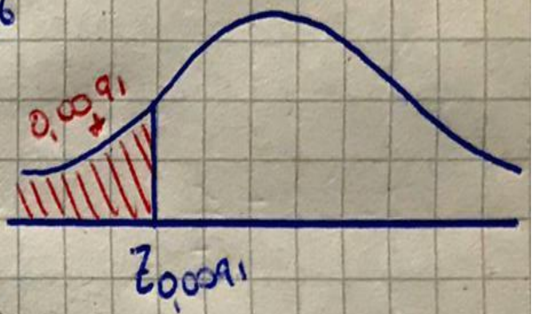
$$Potenza\ test = 1 - \beta = 1 - 0.0091 = 0.9909$$

$1 - \beta = P(\text{Rifiutare } H_0 \mid H_1) = P(\bar{X}_m > \overset{x_c}{20{,}22} \mid H_1) = P(Z > \bar{Z})$

$\bar{Z} = \dfrac{20{,}22 - \mu_1}{\frac{6}{\sqrt{m}}} = \dfrac{20{,}22 - 22}{\frac{4}{8}} = -2{,}36$

$z_{0{,}0091} = -2{,}36$

$P(Z > z_{0{,}0091}) = 1 - \beta = 1 - 0{,}0091 = 0{,}9909$



Secondo CASO: TEST UNILATERALE SINISTRO

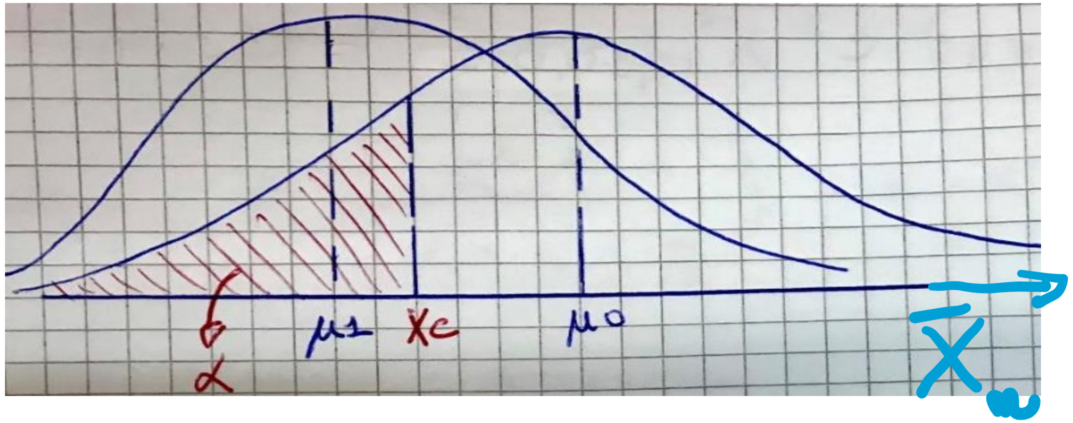
Immaginiamo ora che $\mu_0 > \mu_1$, quindi di questo tipo:

La regione di rifiuto si ha quando la media campionaria è minore di un certo valore X_c .

$$R: \overline{X}_n < x_c$$

La zona di rifiuto si trova nella parte che contiene l'ipotesi alternativa ovvero quando vado a finire troppo a sinistra di μ_0. Osserviamo che la regione di rifiuto deve dalla parte dell'ipotesi alternativa.

Oss: x_c è fissato e una volta fissato rifiuteremo l'ipotesi nulla se il valore osservato della media campionaria cade nella zona di rifiuto.
Il valore critico x_c viene fissato in base al livello di significatività α.



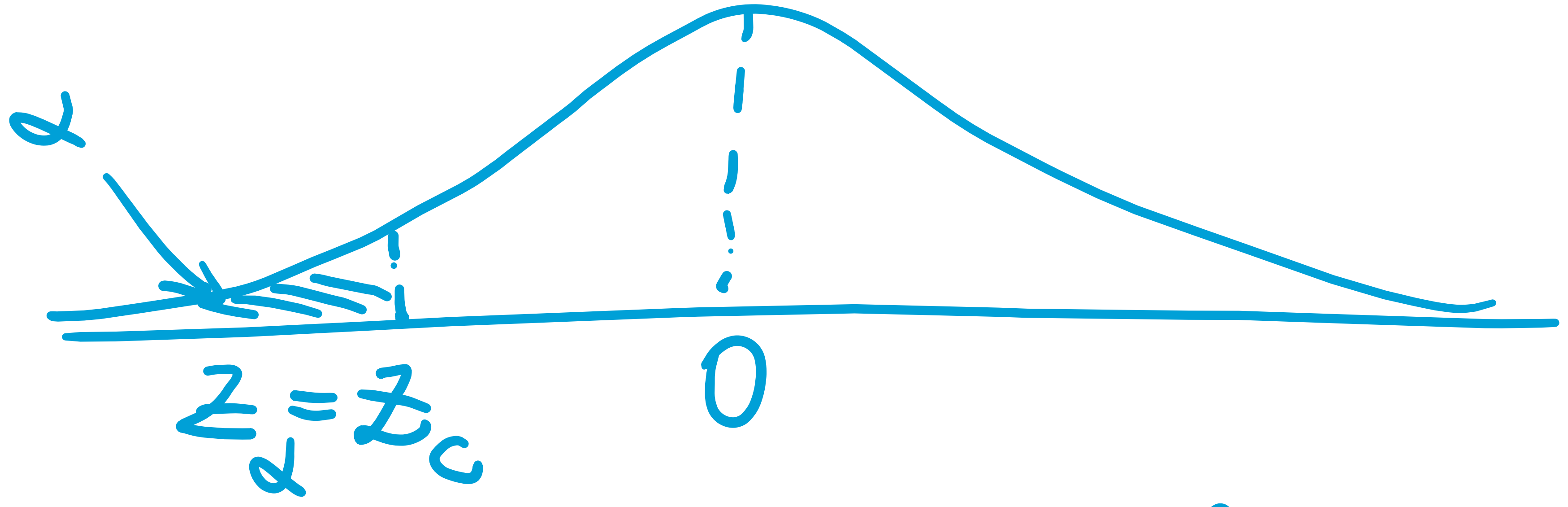
$$\mathbf{R}: \overline{X}_n < x_c$$

$$\mathbf{P}\,(\overline{X}_n < x_c \mid H_0) = \mathbf{P}(\mathbf{Z} < z_c) = \alpha$$

$$z_c = \frac{x_c - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

come determino lo z critico?

Nel caso di Test Unilaterale sinistro



Dunque lo Z_c è proprio il quantile di ordine α

$$Z_\alpha = \frac{X_c - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$\boxed{X_c = \mu_0 + \frac{\sigma}{\sqrt{n}} Z_\alpha} \qquad \boxed{Z_\alpha = -Z_{1-\alpha}} \qquad \boxed{X_c = \mu_0 - \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}}$$

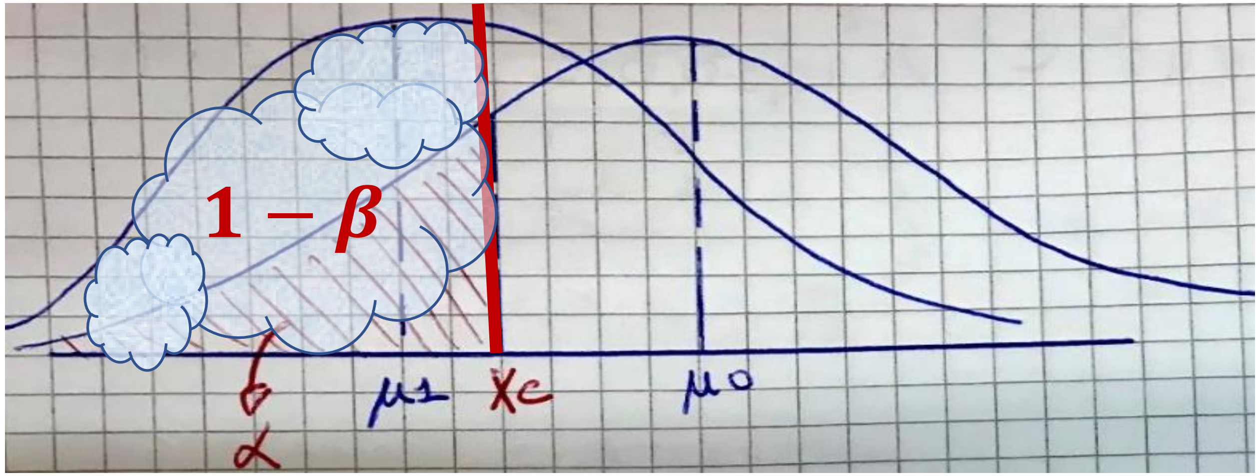
Secondo CASO: TEST UNILATERALE SINISTRO potenza del test

Ricordiamo $\mu_0 > \mu_1$, quindi la zona di rifiuto è $\mathrm{R}: \overline{\mathrm{X}}_n < x_c$

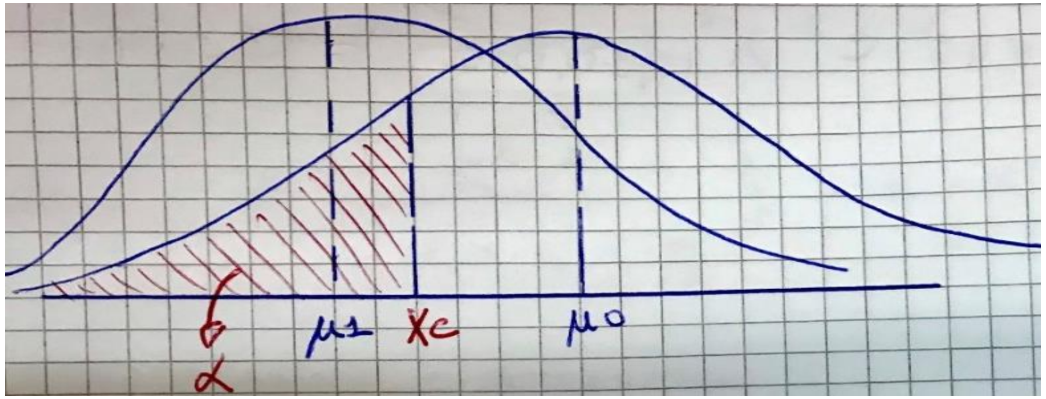
Potenza del test

$$1-\beta = \mathrm{P}\,(\boldsymbol{Rifiuto}\ \boldsymbol{H_0} \mid \boldsymbol{H_1}) = \mathrm{P}\,(\overline{\boldsymbol{X}}_{\boldsymbol{n}} < \boldsymbol{x_c} \mid \boldsymbol{H_1}) = \mathrm{P}(\mathrm{Z} < \boldsymbol{z_q}) = \mathrm{P}(\mathrm{Z} < \boldsymbol{z_{1-\beta}})$$

$$\boldsymbol{z_q} = \boldsymbol{z_{1-\beta}} = \frac{\boldsymbol{x_c} - \boldsymbol{\mu_1}}{\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}}$$



Secondo CASO: TEST UNILATERALE SINISTRO



$$\mu_0 > \mu_1$$

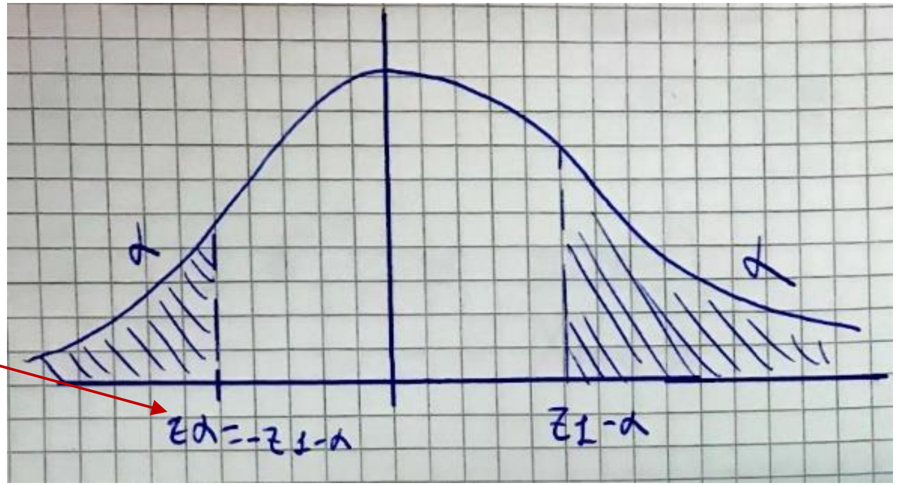
$$R: \bar{X}_n < x_c$$

La zona di rifiuto si trova nella parte che contiene l'ipotesi alternativa ovvero quando vado a finire troppo a sinistra di μ_0

Il valore critico x_c viene fissato in base al livello di significatività α usando la statistica Z.

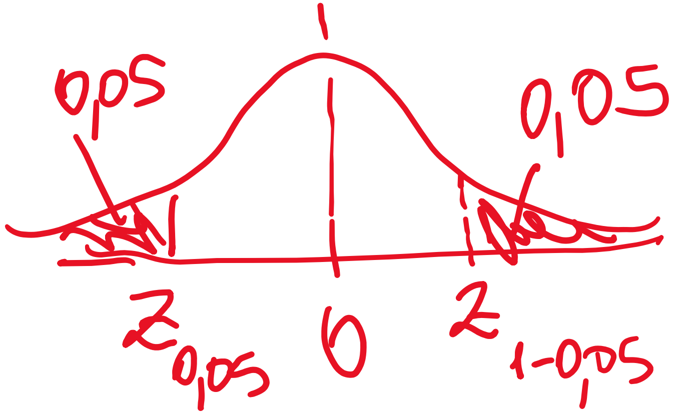
$$P\,(\bar{X}_n < x_c \mid H_0) = P(Z < z_c) = \alpha$$

$$z_c = \frac{X_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -\,z_{1-\alpha}$$



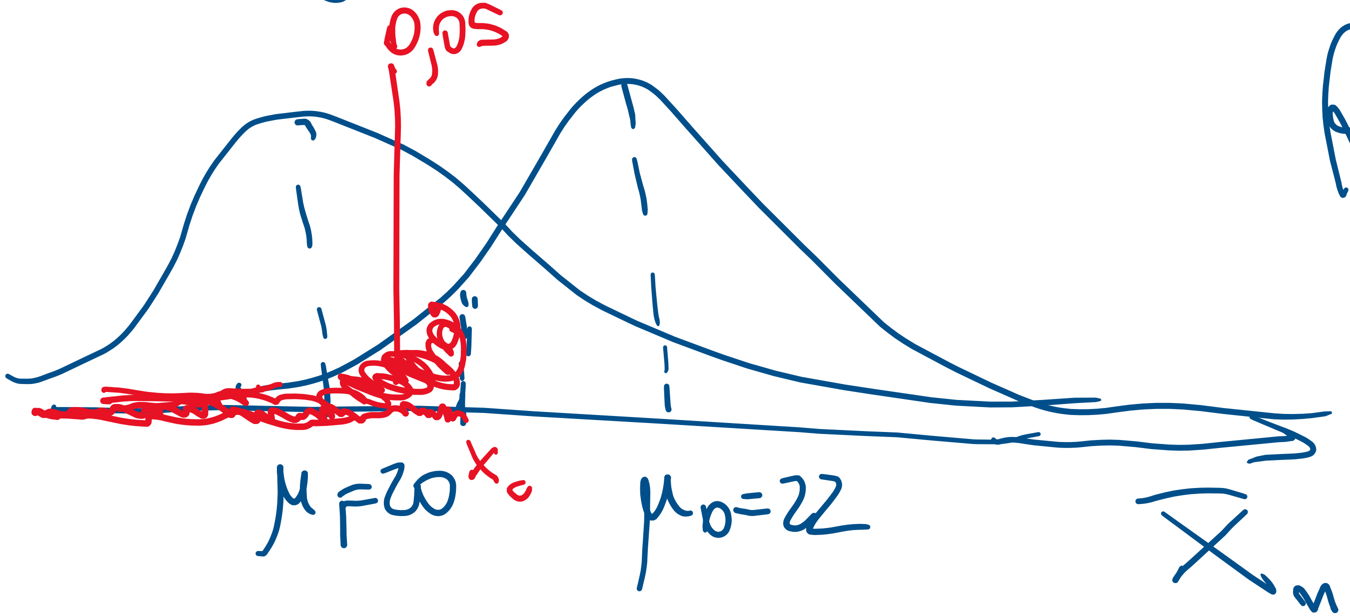
NOTA BENE: $-\,z_{1-\alpha}$

Esercizio 6 Scambiando le ipotesi



$\mu_0 = 22$ $\mu_1 = 20$

1) Determinare la regione di rifiuto a livello di significatività $\alpha = 0,05$



$R: \bar{X}_n < x_c$

$z_c = z_{0,05} = -z_{1-0,05} = -z_{0,95}$

$z_c = -1.64$

$$\frac{X_c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = Z_c \qquad \mu_0 = 22 \quad \sigma = \sqrt{16} \quad n = 64 \qquad \frac{X_c - 22}{\frac{4}{8}} = -1{,}64$$

$$X_c = 22 - 1{,}64 \cdot \frac{1}{2} = 22 - 0{,}82 = 21{,}18$$

Potenza del test

$1 - \beta = 1 - 0{,}0091$

$$Z_q = \frac{X_c - \mu_1}{\frac{\sigma}{\sqrt{n}}} = \frac{21{,}18 - 20}{\frac{4}{8}} = 1{,}18 \cdot 2 = 2{,}36$$

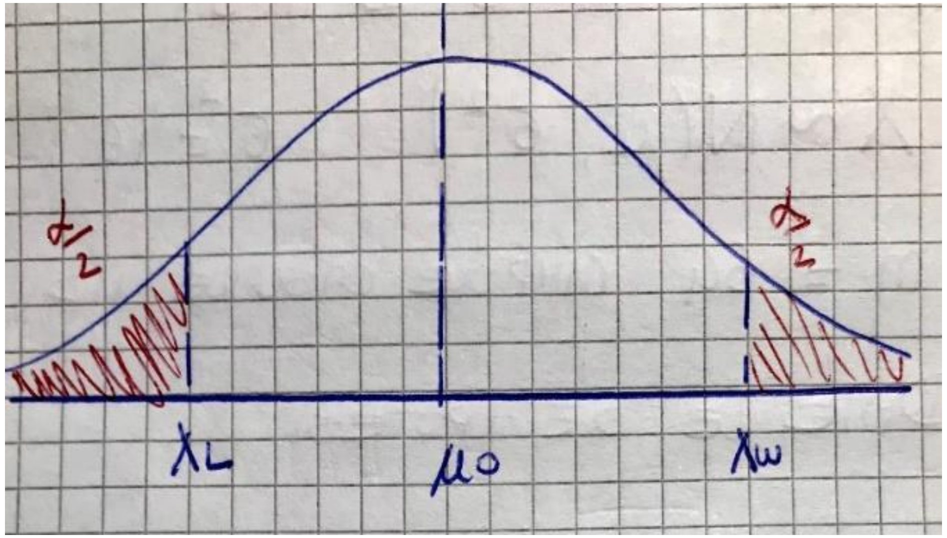
$Z_{0,0091} = -2{,}36$

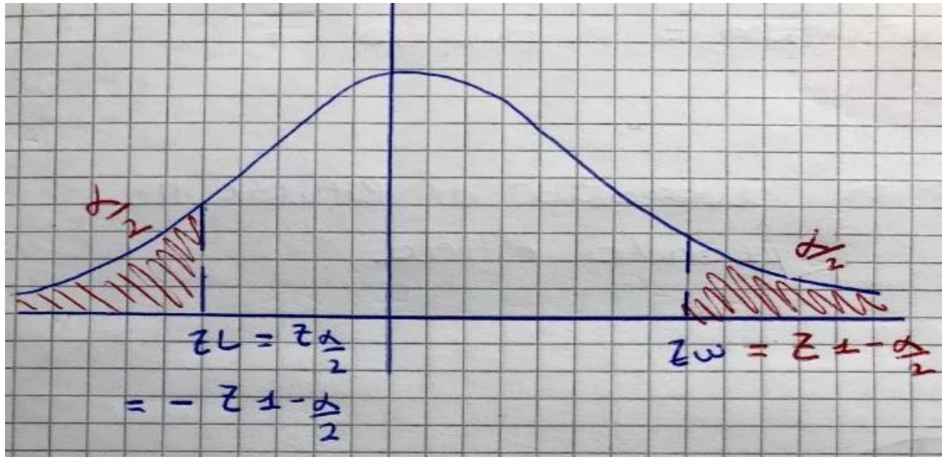
$2{,}36 = Z_{1-0{,}0091}$

TERZO CASO: TEST BILATERALE

Ipotesi $\boldsymbol{H_0}$ $\mu = \mu_0$ $\boldsymbol{H_1}$ $\mu \neq \mu_0$

La regione di rifiuto si ha quando la media campionaria è maggiore di un certo valore x_U e minore di un certo valore x_L





$$R: \overline{X}_n < x_L \quad \vee \quad \overline{X}_n > x_U$$

I valori critici x_L e x_u sono fissati in base al livello di significatività α.

$$P(\overline{X}_n < x_L \vee \overline{X}_n > x_u) = \alpha \quad P(Z_n < -z_c \vee Z_n > z_c) = \alpha$$

$$z_c = z_{1-\frac{\alpha}{2}} \qquad \frac{x_L - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -z_c = -z_{1-\frac{\alpha}{2}}$$

$$\frac{x_u - \mu_0}{\frac{\sigma}{\sqrt{n}}} = z_c = z_{1-\frac{\alpha}{2}}$$