Statistica test per la verifica di ipotesi sulla media μ

	Distribuzione popolazione			
	$X \sim$ Normale		$X \sim f(X)$ qualunque	
	Varianza popolazione		**Varianza popolazione**	
Dimens. campion	**nota**	**non nota (s^2 stima campionaria)**	**nota**	**non nota (s^2 stima campionaria)**
n piccolo	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim T_{n-1}$	?	?
n grande	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$

Test per la media di una popolazione (varianza della popolazione nota)

Il manager delle poste di Napoli è interessato a stabilire se il tempo medio di attesa dei clienti allo sportello è rimasto invariato oppure è cambiato nell'ultimo anno rispetto al precedente (quando era pari a 30 minuti)

Assume un livello di significatività del 5% (α=0,05)
Su un campione di n=150 clienti osserva che il tempo medio di attesa è pari a 30,2 minuti $(\bar{x} = 30{,}2)$

Si ipotizza che il manager conosca anche che la deviazione standard dei tempi di attesa della popolazione è pari a 5,38 minuti (σ=5,38; σ^2=29)

Cosa può concludere il manager sulla base dell'informazione contenuta nel campione?

Ingredienti del test di ipotesi

L'ipotesi riguarda la media μ

$\mu=30$ è l'ipotesi H_0; $\mu\neq30$ è l'ipotesi H_1

Tempo di attesa nella pop. $X \sim f(X;\mu,\sigma^2)$

σ^2 è nota ($\sigma^2=29$)

n è grande ($n=150$) ⇒ $\overline{X} \sim N\left(\mu;\frac{\sigma^2}{n}\right)$ per il TLC

α livello di significatività = Prob. di errore di I tipo =0,05

tempo medio di attesa osservato sul campione

$\overline{x} = 30{,}2$

Test per la media di una popolazione (varianza della popolazione nota, n grande)

$$H_0 : \mu = 30$$
$$H_1 : \mu \neq 30$$

Sotto H_0, $\quad \overline{X} \sim N\left(30, \frac{29}{150}\right)$

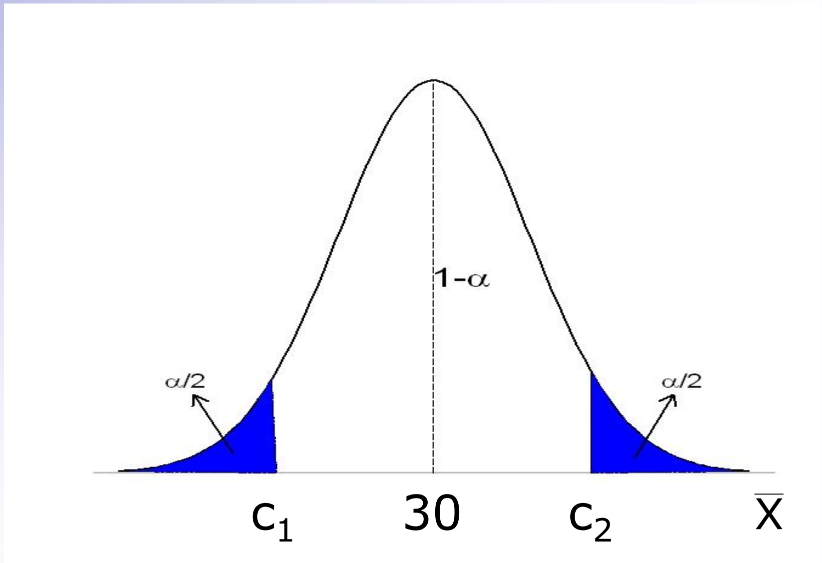
per il teor. del lim.centrale

Regione di rifuto R

$$R = \{\overline{X} < c_1\} \cup \{\overline{X} > c_2\}$$

$$P(\overline{X} \in R \ \text{ sotto } H_0) = \alpha$$

Osservazione:
per trovare c_1 e c_2 dovrei risolvere un integrale
In alternativa ragiono in termini di normale standardizzata



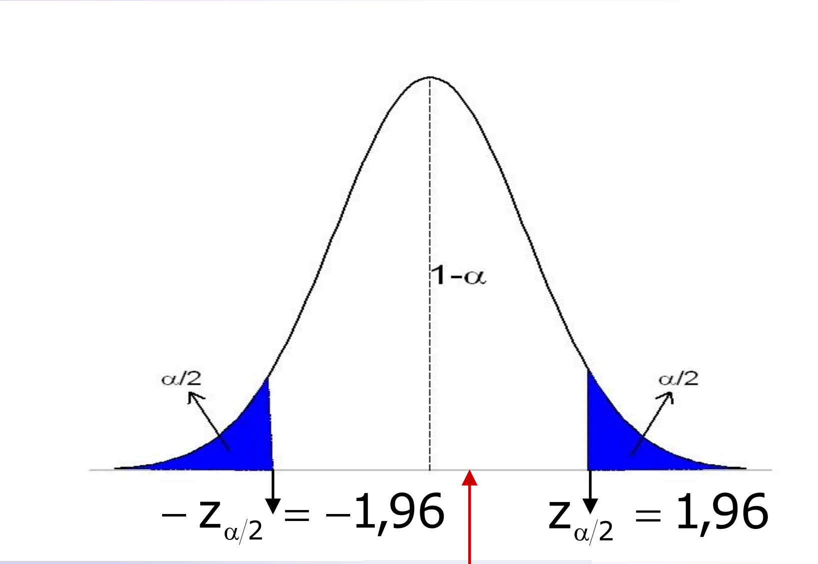
Test per la media di una popolazione (varianza della popolazione nota, n grande)

Sotto H_0, $\dfrac{\overline{X}-30}{\sqrt{29/150}} \sim N(0,1)$

$\alpha = 0{,}05;\ \ \alpha/2 = 0{,}025$

$z_{\alpha/2} = z_{0,025} = 1{,}96$

perché $\Phi(1{,}96) = P(Z < 1{,}96) = 0{,}975$



z=0,45

$R = \{Z < -1{,}96\} \cup \{Z > 1{,}96\}$

Valore osservato della statistica test $\quad z = \dfrac{(\overline{x}-30)}{\sqrt{29/150}} = \dfrac{(30{,}2-30)}{\sqrt{29/150}} = 0{,}45$

$-1{,}96 < 0{,}45 < 1{,}96$ il manager non può rifiutare l'ipotesi nulla

Non c'è sufficiente evidenza empirica per sostenere che il tempo medio di attesa sia cambiato

Con l'approccio del p-value

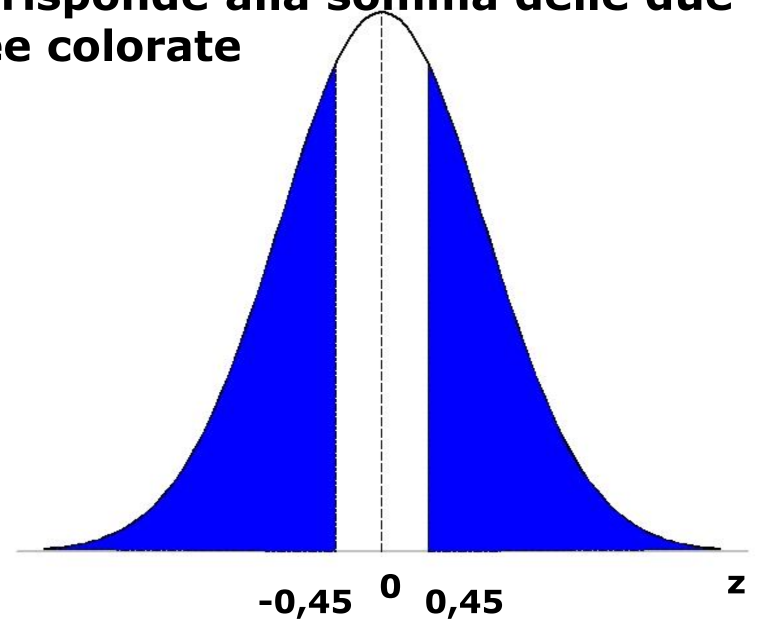
Valore osservato della statistica test z=0,45

$$p-\text{value}=P(Z<-0{,}45)+P(Z>0{,}45)=2P(Z>0{,}45)=$$
$$=2[1-\Phi(0{,}45)]=0{,}65$$

Per qualunque livello di significatività α che si ritenga appropriato,
p-value > α

⇨ non c'è sufficiente evidenza nei dati per rifiutare l'ipotesi nulla

Graficamente il p-value corrisponde alla somma delle due aree colorate



Ruolo della dimensione del campione

Lasciando immutati i termini del problema inferenziale, cosa cambierebbe se il manager ottenesse lo stesso risultato campionario da un campione di n=15 clienti?

Non si può dire niente sulla distribuzione della media campionaria, a meno che non si faccia l'ulteriore ipotesi che la distribuzione della popolazione è normale

Aggiungiamo quindi l'ipotesi di normalità della popolazione

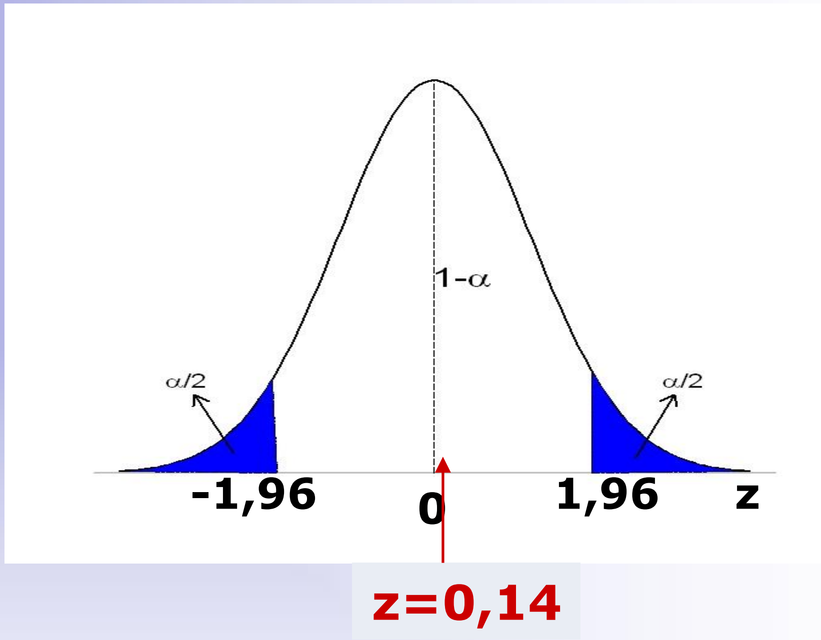
⇨ Anche per piccoli campioni $\overline{X} \sim N\left(30, \frac{29}{15}\right)$

Test per la media di una popolazione (varianza della popolazione nota, n piccolo)

$$H_0 : \mu = 30$$

$$H_1 : \mu \neq 30$$

$$z = \frac{(30{,}2 - 30)}{\sqrt{\frac{29}{15}}} = 0{,}14$$



$-1{,}96 < 0{,}14 < 1{,}96$

⇨ il manager non può rifiutare l'ipotesi nulla al livello di significatività del 5%

Conclude che il tempo medio di attesa non si è significativamente modificato

Se la varianza della popolazione non è nota

Riprendiamo l'esempio precedente con
$X \sim N(\mu,\sigma^2)$
Supponiamo che il manager non conosca la varianza σ^2 della popolazione che deve quindi essere stimata dal campione utilizzando la varianza campionaria corretta
Ipotizziamo che $S^2 = 32{,}8$ da un campione di $n=15$ clienti
Che cosa cambia nella verifica dell'ipotesi?

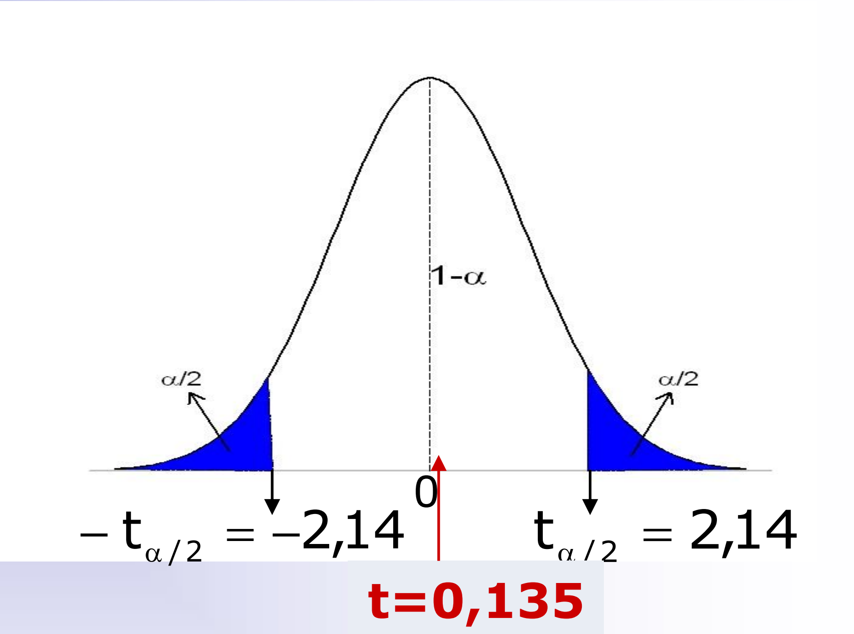
Test per la media di una popolazione (varianza della popolazione non nota, n piccolo)

$$H_0 : \mu = 30$$

$$H_1 : \mu \neq 30$$

$$T_{n-1} = \frac{(\overline{X} - \mu_0)}{\sqrt{\frac{S^2}{n}}}$$ è la statistica test

n-1=14 g.d.l.



Valore osservato della statistica test

$$t = \frac{(30,2 - 30)}{\sqrt{\frac{32,8}{15}}} = 0,135$$

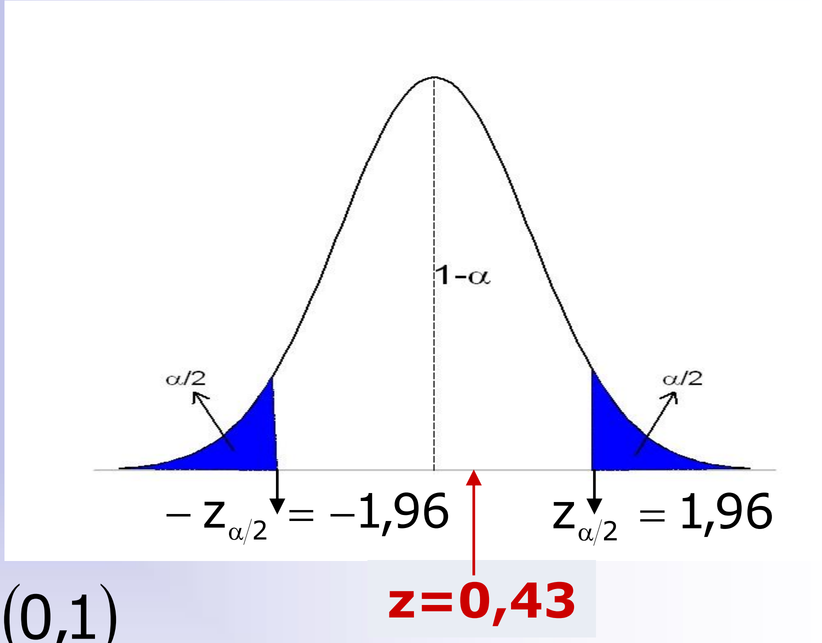
$$P(T_{14} > 2,14) = \frac{\alpha}{2} = 0,025$$

$-2,14 < 0,135 < 2,14$

⇨ l'ipotesi nulla non può essere rifiutata al livello di significatività del 5%

Test per la media di una popolazione (varianza della popolazione non nota, n grande)

Sempre considerando la stima campionaria della varianza,
se n=150 invece di n=15



Sotto H_0, per TLC $\dfrac{\overline{X}-30}{\sqrt{32{,}8/150}} \sim N(0,1)$

Valore osservato della statistica test $z = \dfrac{(\overline{x}-30)}{\sqrt{32{,}8/150}} = \dfrac{(30{,}2-30)}{\sqrt{32{,}8/150}} = 0{,}43$

$-1{,}96 < 0{,}43 < 1{,}96$ il manager non può rifiutare l'ipotesi nulla al livello di significatività del 5%

Osservazione:

Abbiamo introdotto i test sulla media di una popolazione utilizzando un test bilaterale, ma tutti i risultati descritti sono applicabili a test unilaterali per la media della popolazione

Esempio: test unilaterale a sinistra sulla media della popolazione

Un'industria automobilistica acquista un lotto di batterie per autovetture della durata media di 4000 ore, secondo quanto ha dichiarato il costruttore

Gli acquirenti vogliono verificare al livello di significatività del 5% sulla base di un campione di 30 autovetture che le batterie abbiano almeno una durata di 4000 ore

Si conosce che la deviazione standard σ della popolazione di batterie è pari a 141,42 ore (σ=141,42; σ^2=20000), che la popolazione è normale e che la media campionaria riscontrata è pari a 3985 ore.

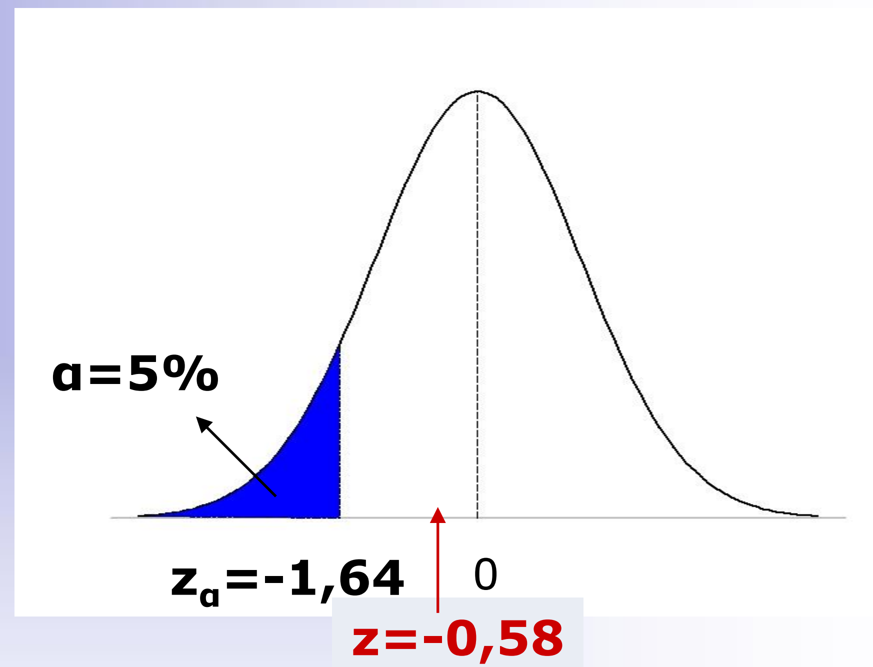
Esempio: test unilaterale a sinistra sulla media della popolazione

$$H_0 : \mu = 4000$$

$$H_1 : \mu < 4000$$

$$\overline{X} \sim N\left(4000, \frac{20000}{30}\right)$$

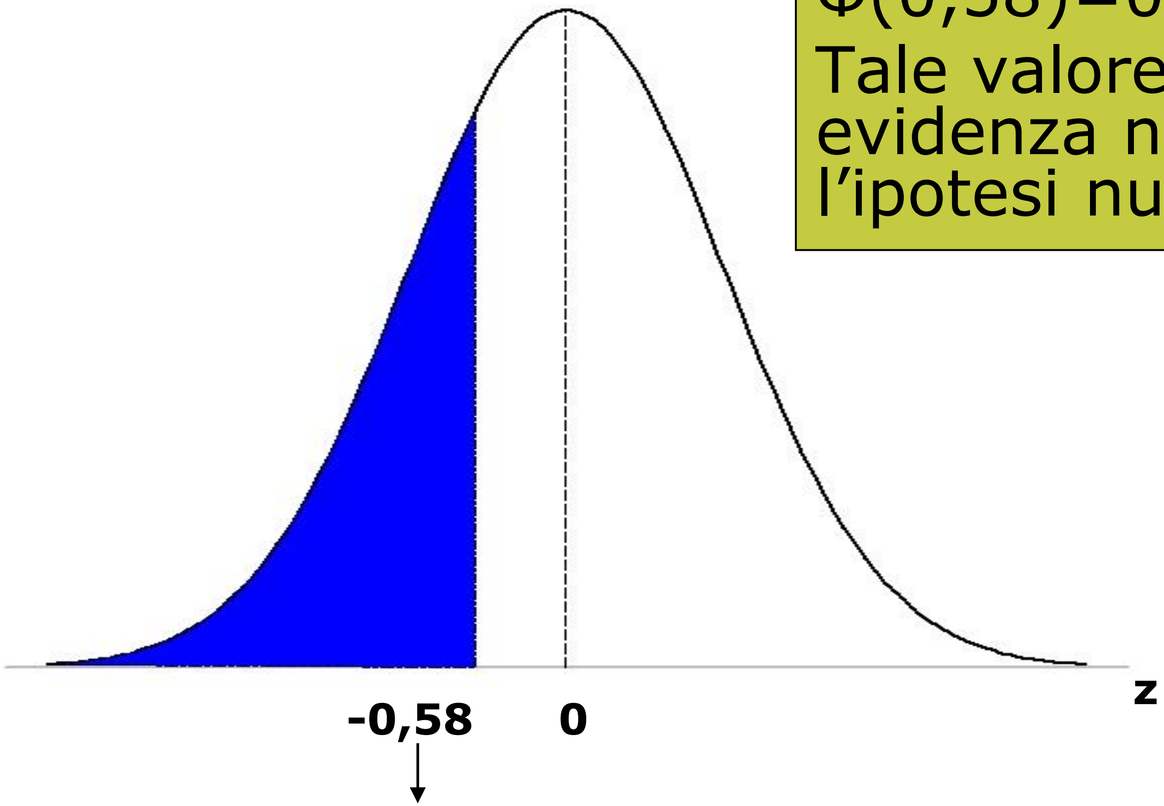
$$z = \frac{(3985 - 4000)}{\sqrt{\frac{20000}{30}}} = -0{,}58$$



Poiché -0,58> -1,64 allora il manager non può rifiutare l'ipotesi nulla al livello di significatività del 5%

⇨ la durata media delle batterie si può ritenere pari almeno a 4000 ore, come dichiarato dal produttore

Approccio del p-value



valore osservato z della statistica test

p-value=P(z<-0,58)=1-Φ(0,58)=0,29

Tale valore non mostra evidenza nei dati contro l'ipotesi nulla

15

Esempio: test unilaterale a destra sulla media della popolazione

Un produttore di vernici assicura che il tempo medio di asciugatura non è superiore a 15 minuti

La ditta acquirente prima di acquistare il prodotto prova il prodotto su 200 pezzi per verificare l'affermazione del produttore ad un livello di significatività dell'1% e riscontra un tempo medio di asciugatura pari a 15,8 minuti

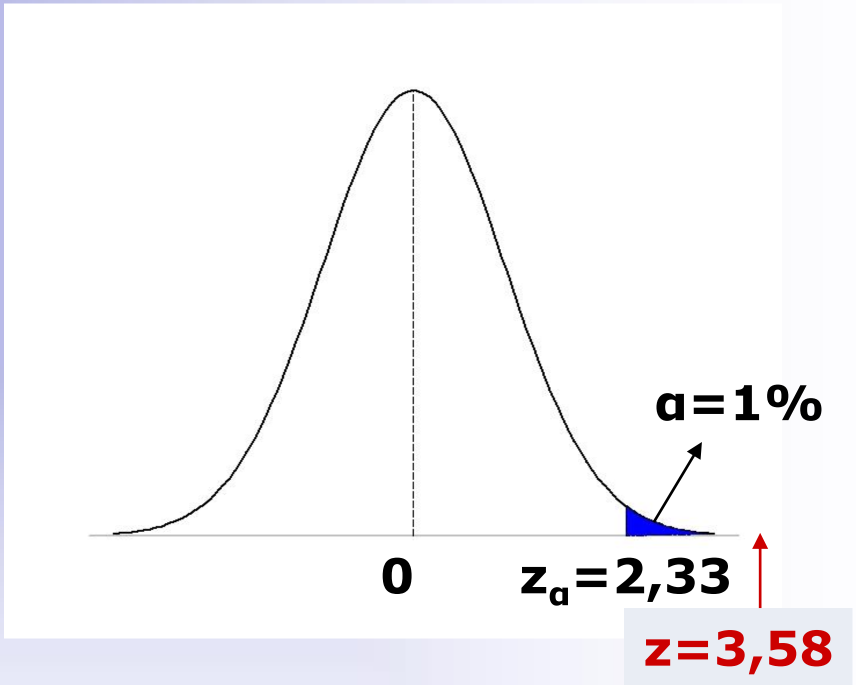
E' noto che la distribuzione dei tempi di asciugatura è normale con una varianza pari a 10

Esempio: test unilaterale a destra sulla media della popolazione

$$H_0 : \mu = 15$$
$$H_1 : \mu > 15$$

$$\overline{X} \sim N\left(15, \frac{10}{200}\right)$$

$$z = \frac{(15{,}8 - 15)}{\sqrt{\frac{10}{200}}} = 3{,}58$$



Poiché 3,58> 2,33 allora il manager rifiuta l'ipotesi nulla ad un livello di significatività dell'1%

⇨ i dati campionari smentiscono l'affermazione del produttore

Il tempo medio di asciugatura è significativamente superiore a 15 minuti

Approccio del p-value

p-value=$P(z>3{,}58)=1-\Phi(3{,}58)=0{,}00017$

Tale valore mostra una netta evidenza dei dati contro l'ipotesi nulla

Test per una proporzione

X è una v.c. dicotomica che può assumere i valori 0 e 1 la cui distribuzione di probabilità è una Bernoulli di parametro π

I sistemi di ipotesi possono essere:

$$H_0 : \pi = \pi_0 \qquad H_0 : \pi = \pi_0 \qquad H_0 : \pi = \pi_0$$

$$H_1 : \pi \neq \pi_0 \qquad H_1 : \pi > \pi_0 \qquad H_1 : \pi < \pi_0$$

Per n grande la statistica test sotto H_0 è una normale standardizzata

(in tal caso

la media campionaria=proporzione campionaria)

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

Esempio: test bilaterale sulla proporzione

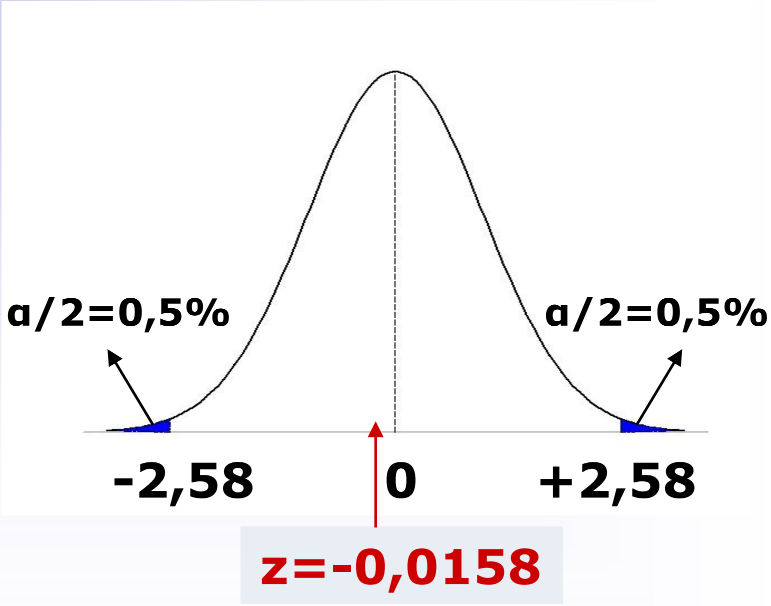
Si sono effettuati 1000 lanci di una moneta e si è ottenuto 498 volte testa

Decidere se la moneta è truccata oppure no ad un livello di significatività pari a 1%

$$H_0 : \pi = 0,5$$

$$H_1 : \pi \neq 0,5$$

$$z = \frac{0,498 - 0,5}{\sqrt{\frac{0,5(1-0,5)}{1000}}} = -0,0158$$



Si può concludere che ad un livello di significatività del 1% la moneta non sia truccata

Esempio: test unilaterale sulla proporzione

Un’industria farmaceutica asserisce che un farmaco è efficace nel 95% dei casi

Su un campione di 300 persone il farmaco è stato efficace su 230 unità

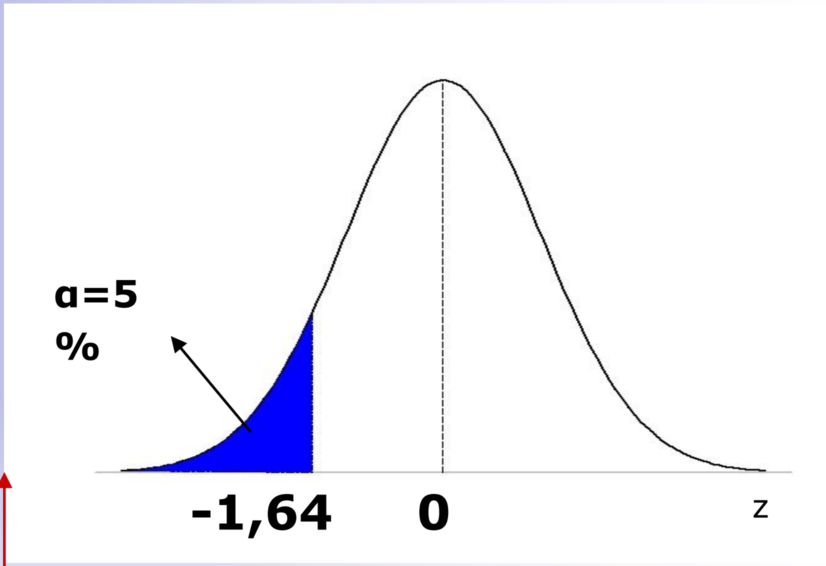
Si può concludere ad un livello di significatività del 5% che l’affermazione sia legittima?

Esempio: test unilaterale sulla proporzione

$$H_0 : \pi = 0,95$$
$$H_1 : \pi < 0,95$$

$$\bar{x} = \frac{230}{300} = 0,76$$

$$z \frac{0,76 - 0,95}{\sqrt{\frac{0,95(1 - 0,95)}{300}}} = -14,3$$



z=-14,3

Poiché -14,3 < -1,64 si rifiuta l'ipotesi nulla ad un livello di significatività del 5%

i dati campionari smentiscono l'affermazione dell'industria

La proporzione di pazienti sui quali il farmaco è efficace è significativamente inferiore al 95%

Connessione tra intervallo di confidenza e verifica di ipotesi

Tra gli intervalli di confidenza e la verifica di ipotesi esiste un legame che permette di costruire test di ipotesi bidirezionali sulla base dell'intervallo di confidenza corrispondente

Se voglio verificare un'ipotesi sul parametro θ (H_0:$\theta=\theta_0$ contro H_1:$\theta\neq\theta_0$) ad un livello di significatività α, posso costruire l'intervallo di confidenza per lo stesso parametro al livello di confidenza 1-α

Se l'intervallo contiene il valore di θ_0, allora posso accettare l'ipotesi nulla

Esempio

Un produttore asserisce che il contenuto di bibita delle lattine che produce ha una media pari a 300 ml. Si estrae un campione di 200 lattine e si riscontra una media di 305ml con varianza pari a 4000. Si può dire al livello di significatività del 5% che il contenuto medio delle lattine sia diverso da quello dichiarato dal produttore?

$$H_0 : \mu_1 = 300$$
$$H_1 : \mu_1 \neq 300$$

Per verificare questo sistema di ipotesi al livello di significatività del 5%, costruisco la stima per intervallo al livello di confidenza del 95%

$$P\left(305 - 1{,}96\frac{63{,}25}{\sqrt{200}} \leq \mu \leq 305 + 1{,}96\frac{63{,}25}{\sqrt{200}}\right) = 0{,}95$$

$$P(296{,}23 \leq \mu \leq 313{,}76) = 0{,}95$$

Il valore 300 è compreso nell'intervallo

⇨ L'ipotesi H_0 non si può rifiutare al livello di significatività del 5%