Test d’ipotesi per due campioni

Giorgio Corani - (IDSIA, SUPSI)

Statistica Applicata (G2A)

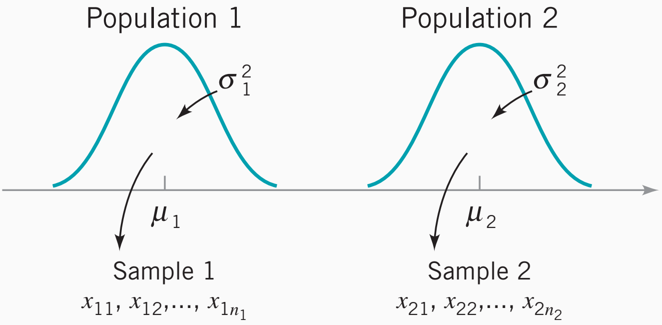
Testo di riferimento

- Douglas P. Montgomery, *Introduction to Statistical Process Control*, 6th Edition, Wiley.

Confrontare due popolazioni

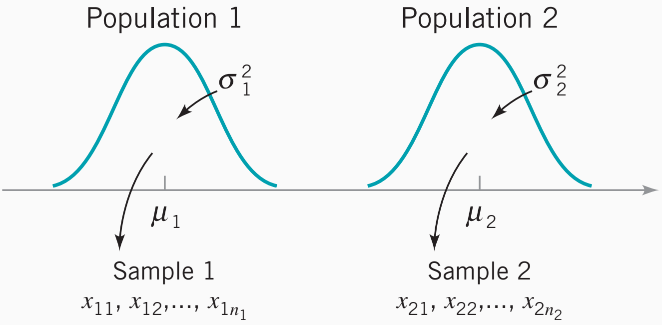
- Sinora abbiamo studiato test d'ipotesi e intervalli di confidenza riguardanti il parametro (μ o π) di una singola popolazione.
- Adesso studiamo come confrontare i parametri di due popolazioni.

Confrontare due popolazioni



- La prima popolazione ha media μ_1 e varianza σ_1^2.
- La seconda popolazione ha media μ_2 e varianza σ_2^2.

Confrontare due popolazioni



- I due campioni hanno dimensione n_1 e n_2.
- Assumiamo che i campioni siano estratti in modo *indipendente* dalle due popolazioni.
- Più avanti vedremo il caso dei campioni non indipendenti (*appaiati*).

Uguaglianza delle varianze

- Assumiamo $\sigma_1^2 = \sigma_2^2$.
- Questa assunzione ci permette di stimare σ^2 facendo una media pesata di s_1^2 e s_2^2. Questo è generalmente più accurato rispetto a stimare in modo indipendente le due varianze.

Confrontare la media di due popolazioni

- Il test a due code è:

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 \\
H_1 &: \mu_1 \neq \mu_2
\end{aligned}
$$

Sui due campioni misuriamo:

- $\bar{x}_1$ e $\bar{x}_2$: media del primo e del secondo campione.
- s_1^2 e s_2^2: varianza del primo e del secondo campione.

Distribuzione campionaria di $\bar{x}_1 - \bar{x}_2$

- Supponiamo di estrarre molte volte due campioni di dimensioni n_1 e n_2 e di misurare ogni volta $\bar{x}_1 - \bar{x}_2$.
- Assumendo:
  - che le due popolazioni abbiano la stessa varianza σ^2.
  - n_1 e n_2 >15-20 (per avere la normalità di $\bar{x}_1$ e $\bar{x}_2$):

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Statistica con σ nota

- Dato:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

sotto H_0 abbiamo:

$$\frac{\bar{x}_1 - \bar{x}_2 - \overbrace{(\mu_1 - \mu_2)}^{\text{ipotizzato 0 in } H_0}}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

- Ma σ è ignota e quindi non possiamo usare questa statistica.

- La statistica del t test è invece:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- che si distribuisce come una t con $(n_1 + n_2 - 2)$ gradi di libertà.
- s_p sostituisce σ ed è spiegato nella prossima slide.

Varianza pesata

- Per stimare σ^2 usiamo la media pesata di s_1^2 e s_2^2:

$$s_P^2 = \frac{(n_1 - 1)}{n_1 + n_2 - 2} \cdot s_1^2 + \frac{(n_2 - 1)}{n_1 + n_2 - 2} \cdot s_2^2$$
$$s_P = \sqrt{s_P^2}$$

- s_P^2: varianza pesata (*pooled*)
  - i due pesi sono proporzionali ai gradi di libertà dei due campioni.
  - se i campioni hanno uguale ampiezza, s_p^2 è la media di s_1^2 e s_2^2.

La statistica del test

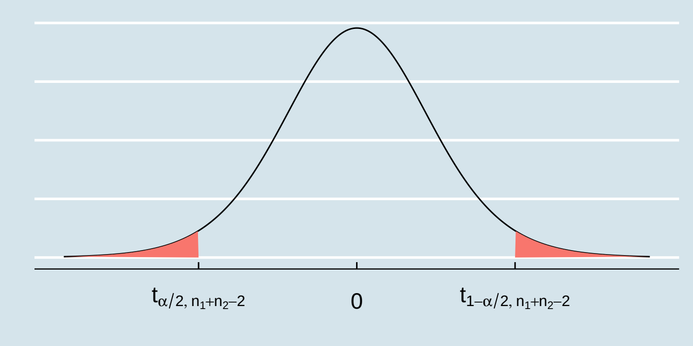
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- $\bar{x}_1 - \bar{x}_2$ è la stima di $\mu_1 - \mu_2$ in base ai dati del campione.
- $s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ è una stima della deviazione standard di $\bar{x}_1 - \bar{x}_2$, che misura quanto $\bar{x}_1 - \bar{x}_2$ è disperso attorno a $\mu_1 - \mu_2$

Regioni di rifiuto

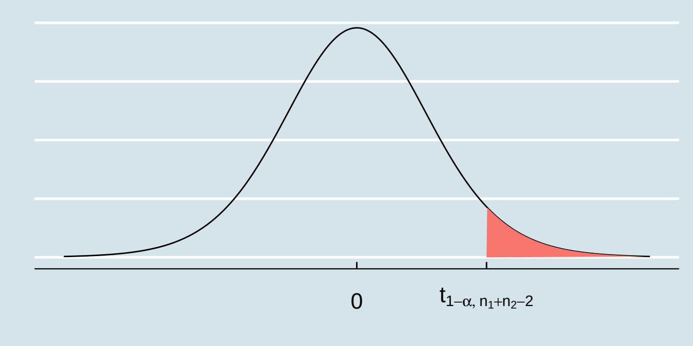
Test a due code

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$
  - Regione di rifiuto: $t_0 < t_{n_1+n_2-2,\alpha/2}$ e $t_0 > t_{n_1+n_2-2,1-\alpha/2}$
  - Ogni coda contiene probabilità $\alpha/2$



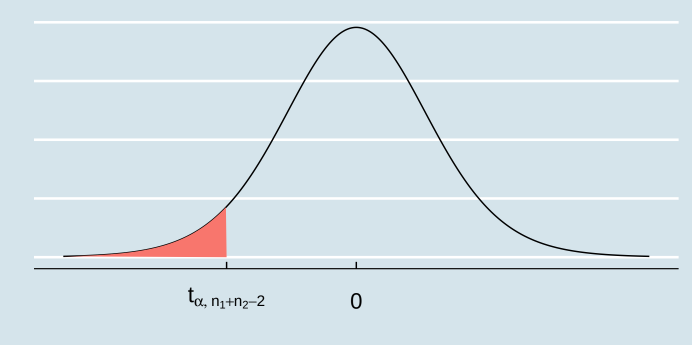
Test con coda destra

- $H_0 : \mu_1 \leq \mu_2$
- $H_1 : \mu_1 > \mu_2$
  - Valori *positivi* della statistica tendono a supportare H_1
  - Regione di rifiuto: $t_0 > t_{n_1+n_2-2,\alpha}$ (contiene probabilità α).



Test a una coda (sinistra)

- $H_0 : \mu_1 \geq \mu_2$
- $H_1 : \mu_1 < \mu_2$
  - Valori *negativi* della statistica supportano H_1
  - Regione di rifiuto: $t_0 < -t_{n_1+n_2-2,\alpha}$ (contiene probabilità α)



Esempio: resa di due catalizzatori

- Si confronta la resa di un processo chimico gestito dalla stessa azienda in due diversi impianti.
- Vogliamo testare se i due impianti hanno la stessa resa media.
- Facciamo quindi il test a due code:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Esempio: resa di due catalizzatori

I dati ($n_1 = n_2 = 8$) sono:

- $\bar{x}_1 = 92.25$, $s_1 = 2.39$
- $\bar{x}_2 = 92.73$, $s_2 = 2.98$

- Svolgiamo il test con $\alpha = 0.05$.

Esempio: resa di due catalizzatori

- Siccome n_1=n_2, s_p^2 è la media di s_1^2 e s_2^2:

$$s_p^2 = \frac{7}{14}s_1^2 + \frac{7}{14}s_2^2 = \frac{2.39^2 + 2.98^2}{2} = 7.3$$
$$s_p = \sqrt{s_p^2} = \sqrt{7.3} = 2.7$$

Statistica e valori critici

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

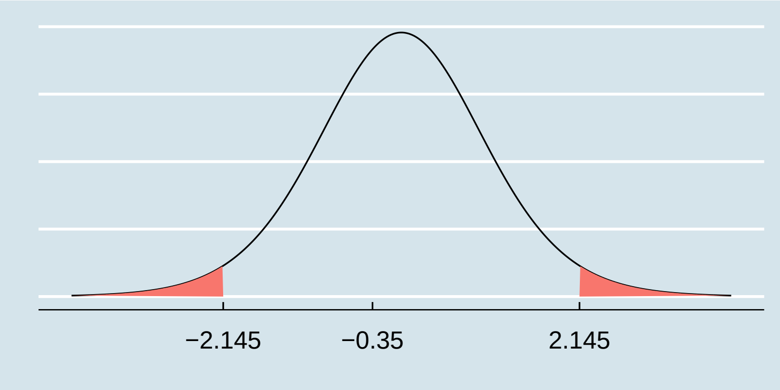
$$= \frac{92.25 - 92.73}{2.7\sqrt{\frac{1}{8} + \frac{1}{8}}}$$

$$= -0.35$$

- I valori critici sono $\pm t_{.975,14} = \pm 2.145$.

Decisione

- La statistica è in regione di **non rifiuto**: non c'è evidenza che la resa dei due impianti sia diversa.



Intervallo di confidenza di $\mu_1 - \mu_2$

- L'intervallo contiene i valori plausibili della differenza $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm t_{1-\alpha/2, n_1+n_2-2} \cdot s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- $t_{1-\alpha/2, n_1+n_2-2}$: quantile $(1 - \alpha/2)$ della t con $(n_1 + n_2 - 2)$ gradi di libertà; corrisponde al valore critico del test.

- $s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ è il denominatore della statistica e rappresenta l'errore standard di $(\bar{x}_1 - \bar{x}_2)$

Intervallo di confidenza vs test d’ipotesi

- Se l’ipotesi $\mu_1 = \mu_2$ è plausibile alla luce dei dati disponibili:
  - il test a due code non rifiuta H_0
  - il CI contiene 0.
- Se l’ipotesi $\mu_1 = \mu_2$ non è plausibile:
  - il test a due code rifiuta H_0
  - il CI non contiene 0.

Intervallo di confidenza (confidence interval, CI)

- I gradi di libertà sono 8-1+8-1 = 14

$$\bar{x}_1 - \bar{x}_2 \pm t_{.975,14} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(92.25 - 92.73) \pm 2.145 \cdot 2.7 \sqrt{\frac{1}{8} + \frac{1}{8}} = (-3.38, 2.42)$$

- Lo 0 è uno dei valori plausibili di $\mu_1 - \mu_2$, in quanto contenuto dal CI.
- Questo è coerente con l'esito del test, che non rifiuta H_0.

Esempio di test a una coda

- Uno studio riporta le percentuali di calcio misurate in cemento standard e cemento addizionato con piombo (*doped*), al termine di uno stress test.
- Maggiori percentuali di calcio indicano maggiore resistenza all'infiltrazione dell'acqua.
- Siccome il cemento doped è più costoso, si richiede forte evidenza che contenga una maggiore percentuale di calcio rispetto a quello standard.
- Svolgere il test usando α=0.05.

Formulazione del cemento

- L'ipotesi alternativa è ciò che vogliamo dimostrare, cioè che il cemento *doped* contiene una maggiore percentuale di calcio.
- Il test quindi è:

$$H_0 : \mu_{\text{standard}} \geq \mu_{\text{doped}}$$
$$H_1 : \mu_{\text{standard}} < \mu_{\text{doped}}$$

Formulazione del cemento: dati

$$n_{\text{standard}} = 10$$
$$\bar{x}_{\text{standard}} = 87.0$$
$$s_{\text{standard}} = 5.0$$

$$n_{\text{doped}} = 15$$
$$\bar{x}_{\text{doped}} = 90.0$$
$$s_{\text{doped}} = 4.0$$

Formulazione del cemento: s_P

La varianza pooled è:

$$s_P^2 = \frac{9 \cdot (5)^2 + 14 \cdot (4)^2}{10 + 15 - 2} = 19.52$$
$$s_P = \sqrt{19.52} = 4.4$$

La statistica è:

$$t_0 = \frac{\bar{x}_{\text{standard}} - \bar{x}_{\text{doped}}}{s_p \sqrt{\frac{1}{n_{\text{standard}}} + \frac{1}{n_{\text{doped}}}}} = \frac{87 - 90}{4.4\sqrt{\frac{1}{10} + \frac{1}{15}}} = -1.67$$

Regione di rifiuto

- Se $\bar{x}_{\text{doped}} > \bar{x}_{\text{standard}}$, la statistica è negativa e tende supportare H_1.
- In particolare, rifiutiamo H_0 se $t_0 < t_{0.05,23} = -1.71$.
- La statistica (-1.67) è in regione di non-rifiuto: non abbiamo forte evidenza che il cemento *doped* aumenti il contenuto di calcio.

Esercizio: p - value

- La statistica è comunque vicina alla regione di rifiuto.
- Usando le tavole, provate a stimare approssimativamente il valore del p-value.

Esercizio: confronto fra metodi di vendita

- Si confrontano i pezzi venduti settimanalmente di un certo prodotto in due gruppi di supermercati.
- I supermercati del primo adottano la collocazione a scaffale, mentre quelli del secondo utilizzano uno spazio dedicato.
- Con confidenza 95%., possiamo concludere che le vendite medie nel caso dello spazio dedicato siano significativamente superiori a quelle dello scaffale?

Dati

$$n_{\text{scaffale}} = 10$$
$$n_{\text{dedicato}} = 10$$

$$\bar{x}_{\text{scaffale}} = 50.3$$
$$\bar{x}_{\text{dedicato}} = 72$$

$$s^2_{\text{scaffale}} = 350$$
$$s^2_{\text{dedicato}} = 157$$

Campioni appaiati

Campioni non indipendenti (appaiati)

- A volte le osservazioni delle due popolazioni sono appaiate (*paired*)
- L'osservazione i-esima è presa in condizioni omogenee per entrambi i campioni, ma la condizione cambia ad ogni osservazione.

Esempio di campioni appaiati

- Vogliamo comparare le misure di durezza del metallo svolte da due tipi di punte.
- La macchina spinge la punta nel metallo con una forza prestabilita. La durezza del metallo si deduce dalla profondità del foro.
- Potremmo testare le due punte su pezzi di metallo diversi tratti dalla stessa produzione e poi applicare il t-test per campioni indipendenti.
- Ma le differenze di misura sarebbero dovute sia al tipo di punta sia alle piccole differenze tra i pezzi.

Campioni appaiati

- È meglio usare entrambe le punte su ogni pezzo di metallo.
- I pezzi di metallo sono il fattore appaiante dei due campioni.
- Misuriamo quindi la differenza di misura su ogni pezzo di metallo.

t-test appaiato

- Consideriamo due campioni costituiti da n osservazioni appaiate:
- Calcoliamo la differenza, in ogni osservazione, tra il valore del primo e del secondo campione.
- Otteniamo così il campione delle differenze, che ha lunghezza n, media $\bar{d}$ e deviazione standard s_d.

t-test appaiato

- Denotiamo $\mu_d = \mu_1 - \mu_2$ e facciamo questo test sul campione delle differenze:

$$
\begin{aligned}
H_0 &: \mu_d = 0 \\
H_1 &: \mu_d \neq 0
\end{aligned}
$$

- La statistica è:

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

- la cui distribuzione campionaria è una t con n-1 gradi di libertà.

Intervallo di confidenza

L'intervallo di confidenza è simmetrico attorno a $\bar{d}$:

$$\bar{d} \pm t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$$

Misure di durezza svolte dalle due punte.

metallo	punta A	punta B	diff
1	10.0	9.9	0.1
2	9.9	9.9	0.0
3	9.8	9.9	-0.1
4	10.0	9.8	0.2
5	9.9	9.9	0.0
6	10.0	9.8	0.2
7	9.9	10.0	-0.1
8	10.1	9.9	0.2

Le misure medie delle due punte sono significativamente diverse?

$$H_0 \ : \mu_d = 0$$
$$H_1 \ : \mu_d \neq 0$$

- $\bar{d}$ = 0.06
- s_d = 0.13
- t_0 = 0.06/ (0.13 / sqrt(8)) = 1.31

- I valori critici sono: $t_{\alpha/2,7}, t_{1-\alpha/2,7}$, cioè $\pm$ 2.31

- La statistica è in regione di *non rifiuto*. Non c'è differenza sistematica tra le misure medie dei due strumenti, anche se le misure su ogni singolo pezzo sono differenti.

Intervallo di confidenza

- L'intervallo di confidenza è:

$$\bar{d} \pm t_{n-1,1-\alpha/2} \frac{s_D}{\sqrt{n}}$$

- i cui estremi sono [0.17, -0.05]
- Questo intervallo contiene lo 0, coerentemente con l'esito del test precedente (test a due code e CI svolti con lo stesso α danno esiti coerenti).

Esercizio

- Uno studio ha chiesto a 14 guidatori di parcheggiare due auto identiche, ma con volante di diversa dimensione.
- I tempi di parcheggio (espressi in secondi) mostrano una differenza media $\bar{d}$=1.21 ed una deviazione standard della differenza s_D=12.68.
- Con confidenza 90%, possiamo dire che il tempo medio di parcheggio delle due auto è significativamente diverso?

Confronto fra due proporzioni

Confronto fra due proporzioni

- Vogliamo confrontare i parametri π_1 e π_2 che caratterizzano due distribuzioni binomiali.
- Il test si basa su una approssimazione per campioni ampi per i quali $p_1 = \frac{X_1}{n_1}$ e $p_1 = \frac{X_2}{n_2}$ sono normalmente distribuite.
- Entrambi i campioni devono contenere almeno 5 successi e 5 insuccessi.
- Questo permette di applicare l'approssimazione normale alla binomiale per entrambe le popolazioni.

Confronto fra due proporzioni

- Abbiamo estratto due campioni di dimensione n_1 e n_2 dalle due popolazioni, il cui numero di successi è rispettivamente X_1 e X_2.
- Il test di uguaglianza a due code è:

$$\begin{aligned} H_0 &: \pi_1 = \pi_2 \\ H_1 &: \pi_1 \neq \pi_2 \end{aligned}$$

Statistica del test

$$Z = \frac{(p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

dove :

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

- Sotto H_0 la statistica si distribuisce come una $N(0, 1)$.
- Nel caso di test a una coda, la statistica rimane identica ma cambia la regione di rifiuto.

Riassunto regioni di rifiuto

H_1	Regione di rifiuto	p-value
$\pi_1 \neq \pi_2$	$z < z_{\alpha/2}$ e $z > z_{1-\alpha/2}$	$2(1 - \Phi(\lvert z \rvert))$
$\pi_1 > \pi_2$	$z > z_{1-\alpha}$	$1 - \Phi(z)$
$\pi_1 < \pi_2$	$z < z_{\alpha}$	$\Phi(z)$

Intervallo di confidenza di $\pi_1 - \pi_2$

I valori plausibili di $\pi_1 - \pi_2$ sono contenuti nell'intervallo:

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- In questo caso non c'è una perfetta corrispondenza tra CI e test a due code, perchè il test ed il CI stimano l'errore standard di $\pi_1 - \pi_2$ in modo (leggermente) diverso.

Esempio: valutare l’efficacia di un farmaco

- Per valutare l’efficacia di un farmaco si svolge un *randomized trial*.
- In modo casuale ad alcuni pazienti viene somministrato il farmaco; ad altri il placebo.
- Alla fine del periodo di cura, è necessario analizzare se c’è una differenza statisticamente significativa fra i due gruppi.

Esempio: valutare l'efficacia di un farmaco

- Il gruppo farmaco contiene 227 pazienti, di cui 163 guariti al termine del periodo.
- Il gruppo placebo contiene 262 pazienti, di cui 154 guariti al termine del periodo.
- Il farmaco è significativamente più efficace del placebo?

Valutare l'efficacia di un farmaco

- Entrambi i gruppi contengono almeno 5 successi e 5 insuccessi; possiamo quindi fare il test che usa l'approssimazione per campioni larghi.

- Vogliamo provare a dimostrare l'efficacia del farmaco, quindi facciamo il test:

$$
\begin{aligned}
H_0 &: \pi_{\text{farmaco}} \leq \pi_{\text{placebo}} \\
H_1 &: \pi_{\text{farmaco}} > \pi_{\text{placebo}}
\end{aligned}
$$

- Vogliamo forte evidenza che il farmaco sia efficace, quindi usiamo $\alpha = 0.01$.

Valutare l'efficacia di un farmaco

- La regione di rifiuto del test contiene valori *positivi* di $p_{\text{farmaco}} - p_{\text{placebo}}$ e quindi della statistica.
- Regione di rifiuto: $Z_0 > \Phi^{-1}(.99)$ = 2.33

Valutare l'efficacia di un farmaco

$$
\begin{aligned}
p_{\text{farmaco}} &= 163/227 = 0.72 \\
p_{\text{placebo}} &= 154/262 = 0.59 \\
\bar{p} &= (163 + 154)/(227 + 262) = 0.65 \\
Z &= \frac{p_{\text{farmaco}} - p_{\text{placebo}}}{\sqrt{\bar{p} \cdot (1 - \bar{p}) \cdot 1/n}} = 3.01 > 2.33 \\
\text{p-value} &= 1 - \Phi(3.01) = 0.0013
\end{aligned}
$$

- La statistica è in regione di rifiuto.
- Il p-value è un ordine di grandezza più piccolo di α

Esercizio

- Un processo produce cuscinetti per l'albero motore.
- Si preleva un campione di 85 cuscinetti, che risulta contenere 12 non-conformi.
- Il processo produttivo viene quindi rivisto. Si preleva un nuovo campione di 85 cuscinetti, che risulta contenere 8 non-conformi.
- Possiamo concludere con confidenza del 95% che la frazione di non-conformi è significativamente decresciuta?

Soluzioni

Esercizio: p - value

- Il valore critico (quinto percentile) è $t_{0.05,23} = -1.71$
- Per 23 gradi di libertà, da tabella troviamo il decimo percentile $t_{0.1,23} = -t_{0.9,23} = -1.319$.
- La statistica (-1.67) è compresa fra il 5 ed il 10 percentile.
- Il p-value è calcolato integrando la distribuzione di $-\infty$ a -1.67. Concludiamo che $0.05 <$ p-value < 0.1

Vendite a scaffale vs spazio dedicato

$$H_0 \;: \mu_{\text{dedicato}} \leq \mu_{\text{scaffale}}$$
$$H_1 \;: \mu_{\text{dedicato}} > \mu_{\text{scaffale}}$$

- s_p = $\sqrt{(350 + 157)/2}$ = 15.92
- statistica t = $\frac{72 - 50.3}{s_p\sqrt{(1/10+1/10)}}$ = 3.05
- valore critico: $t_{0.95,18}$=1.73
- Rifiutiamo H_0: le vendite medie con spazio dedicato sono significativamente superiori a quelle dello scaffale.

Tempi di parcheggio

- L'intervallo di confidenza per il tempo medio di parcheggio è:

$$1.21 \pm \frac{12.68}{\sqrt{14}} \cdot t_{0.95,13} = [-4.79, 7.21]$$

- e quindi la differenza nei tempi di parcheggio non risulta statisticamente significativa.

- Una conclusione analoga si può ottenere calcolando la statistica (0.35) e verificando che ricade all'interno dei valori critici ($\pm$ 1.77).

- Testiamo che dopo l'intervento il processo sia diventato meno difettoso:

$$
\begin{aligned}
H_0 &: \pi_1 \leq \pi_2 \\
H_1 &: \pi_1 > \pi_2
\end{aligned}
$$

Produzione di cuscinetti a sfera

$$\bar{p} = \frac{8+12}{85+85} = 0.118$$

$$Z = \frac{12/85 - 8/85}{\sqrt{.118(1-.118)\cdot(\frac{1}{85}+\frac{1}{85})}} = 0.95$$

valore critico: $\Phi^{-1}(1-\alpha) = \Phi^{-1}(0.95) = 1.64$

p-value :$1 - \Phi(Z) = 1 - \Phi(0.95) = 0.18$

- Il test non rifiuta H_0