Test su die popolazioni

Il test su due popolazioni serve per decidere se due appricci allo stesso problema hanno portato allo stesso risultato o no.

Confronto delle medie di due popolazioni normali, caso di varianze note

Siano $X_1, \dots, X_n$ e $Y_1, \dots, Y_m$ due campioni indipendenti estratti da due popolazioni normali con medie **incognite** μ_x e μ_y e varianze **note** σ_x^2 e σ_y^2 note.

Tipologie di test

H_0	H_1	ST	**Rifiuto H_0 se**
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert st \rvert > z_{\frac{\alpha}{2}}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$	$st > z_\alpha$
$\mu_x \geq \mu_y$	$\mu_x < \mu_y$		$st < -z_\alpha$

NB: le σ_x^2 e σ_y^2 sono note da esercizio. Quindi si devono calcolare solo le medie campionarie.

Esempio calcolato

- Prima delle vacanze: 5 corse, tempo medio 53.82 secondi
- Dopo le vacanze: 6 corse, tempo medio 54.41 secondi

Si vuole determinare se le vacanze abbiano influito negativamente sulla prestazione. Si supponga che la varianza $\sigma^2 = 0.1$ sia rimasta costante. Cosa è possibile concludere?

Dati:

- $\overline{X} = 53.82$ secondi
- $\overline{Y} = 54.41$ secondi
- $\sigma^2 = 0.1$ secondi
- $n = 5$
- $m = 6$
- $H_0 : \mu_x = \mu_y$ (*le vacanze non hanno influito sulla prestazione*)
- $H_1 : \mu_x < \mu_y$ (*le vacanze hanno influito negativamente sulla prestazione*)

Calcoli:

1. Calcoliamo il valore di ST:

$$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} = \frac{53.82 - 54.41}{\sqrt{\frac{0.1}{5} + \frac{0.1}{6}}} = -2.98$$

NB: il valore di α in questo caso non è dato, quindi possiamo calcolarlo tramite il p-value sapendo il tipo di test che stiamo facendo.
Dobbiamo trovare il valore di significatività di alpha tale per cui $-z_\alpha = -2.98$.
Quindi caloliamo il p-value:

```
> pnorm(-2.98)
[1] 0.0014
```

Conclusione:
Essendo che il p-value è molto basso **rifiutiamo l'ipotesi nulla** e concludiamo che le vacanze hanno influito negativamente sulla prestazione.

Confronto delle medie di due popolazioni normali, caso di varianze non note, MA UGUALI

Siano $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ due campioni indipendenti estratti da due popolazioni normali con medie μ_x e μ_y e varianze **incognite** σ_x^2 e σ_y^2 **uguali**.
Supponiamo che le varianze siano uguali, cioè $\sigma_x^2 = \sigma_y^2 = \sigma^2$ e usiamo lo stimatore dell **varianza combinata** (o pooled variance) S_p^2.

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Facciamo questo perchè lo consideriamo come unica popolazione, infatti la formula della varianza campionaria è:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Quindi combinando le due abbiamo:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Che serve per stimare la varianza comune σ^2.
Occhio: La formula calcola già il quadrato quindi nello stimatore lo dobbiamo tenere dentro la radice.

Quando assumere che le varianze delle due popolazioni sono uguali?

Possiamo assumere che le varianze delle due popolazioni siano uguali se abbiamo:

$$\frac{1}{2} < \frac{S_x^2}{S_y^2} < 2$$

Tipologie di test

H_0	H_1	ST	**Rifiuto H_0 se**
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert st \rvert > t_{\frac{\alpha}{2},n+m-2}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$st = \frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2(\frac{1}{n}+\frac{1}{m})}} \sim t_{n+m-2}$	$st > t_{\alpha,n+m-2}$
$\mu_x \geq \mu_y$	$\mu_x < \mu_y$		$st < -t_{\alpha,n+m-2}$

Esempio calcolato

Viene eseguito un esperimento ai fini di valutare l'usura di due diversi materia. Vengono testati 12 pezzi del materiale A e 10 pezzi del materiale B. L'usura media del materiale A è 85 unità con deviazione standard campionaria di 4 unità; il materiale B ha un'usura media di 81 unità con deviazione standard campionaria di 5 unità.
A livello di significatività del 5%, si può concludere che l'usura del materiale A supera quella del materiale B?
Si assuma che le due popolazioni abbiano distribuzione normale e varianze uguali.

Dati:

- $\overline{X} = 85$ unità
- $\overline{Y} = 81$ unità
- $S_x = 4$ unità
- $S_y = 5$ unità
- $n = 12$
- $m = 10$
- $H_0 : \mu_x \leq \mu_y$ (*l'usura del materiale A non supera quella del materiale B*)
- $H_1 : \mu_x > \mu_y$ (*l'usura del materiale A supera quella del materiale B*)
- $\alpha = 0.05$

```r
mean_x <- 85
n <- 12
mean_y <- 81
m <- 10
alpha <- 0.05
sd_x <- 4
sd_y <- 5

Sp <- (sd_x^2*(n-1) + (m-1)*sd_y^2)/ (n+m-2)
# Sp = 20.05
st <- (mean_x-mean_y)/(sqrt((Sp)*(11/60)))
# st = 2.08

st > qt(p=alpha, df=n+m-2, lower.tail = FALSE)
# rifiuto H_0 se TRUE, è true
# qt(p=alpha, df=n+m-2, lower.tail = FALSE) = 1.72

# Calcolo con il p-value
pt(q=st, df=n+m-2, lower.tail = FALSE)
# il valore del p-value è minore di alpha quindi rifiuto
```

1. Calcoliamo il valore di $\hat{S_p}^2$:

$$S_p^2 = \frac{11 \cdot 4^2 + 9 \cdot 5^2}{20} = 20.05$$

2. Calcoliamo il valore di ST:

$$st = \frac{85 - 81}{\sqrt{20.05 \cdot (\frac{1}{12} + \frac{1}{10})}} = 2.08$$

3. Calcoliamo il valore di $t_{\alpha,n+m-2}$:

$$t_{\alpha,n+m-2} = t_{0.05,20} = 1.72$$

Conclusione:
Essendo che $st > t_{\alpha,n+m-2}$, **rifiutiamo l'ipotesi nulla** e concludiamo che l'usura del materiale A supera quella del materiale B.

Confronto delle medie di due popolazioni normali, caso di varianze non note e DIVERSE

Siano $X_1, \dots, X_n$ e $Y_1, \dots, Y_m$ due campioni indipendenti estratti da due popolazioni normali con medie μ_x e μ_y e varianze **incognite** σ_x^2 e σ_y^2 **diverse**.
In questo caso non possiamo usare la varianza combinata, quindi dobbiamo stimare le varianze delle due popolazioni separatamente.

- Possiamo effettuare i testo solo su popolazioni numerose: $n, m \geq 30$

Tipologie di test

H_0	H_1	ST	**Rifiuto H_0 se**
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$		$\lvert st \rvert > z_{\frac{\alpha}{2}}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim N(0,1)$	$st > z_\alpha$
$\mu_x \geq \mu_y$	$\mu_x < \mu_y$		$st < -z_\alpha$

Con S_x^2 e S_y^2 varianze campionarie.

Confronto delle medie di due popolazioni normali nel caso di campioni accoppiati

Rispetto i metodi prima che facevano riferimento a due popolazioni indipendenti qui è diverso perchè ora facciamo riferimento a due coppie di osservazioni, con ciascuna coppia sullo stesso individuo.

	1	2	...	n
x_i	55	15		22
y_i	82	28		10

Quindi ad esempio pensiamo alla somministrazione di un farmaco che ha effetto sulla pressione, la misuriamo prima e dopo l'assunzione. Questo ripetuto su n persone diverse; quindi avremo $(x_1, y_1), \dots, (x_n, y_n)$ dove ogni coppia di valore è un'individuo.

Consideriamo un campione casuale $(X_1, Y_1), \dots, (X_n, Y_n)$ dove X e Y sono normali, e costruiamo le **differenz** $W_i = X_i - Y_i$.

Ora il campione che abbiamo calcolato $W_1, \dots, W_n$ possiamo considerarlo come un campione casuale estratto da una popolazione normale con media $\mu_w = \mu_x - \mu_y \in \mathbb{R}$ e varianza σ_w^2, entrambe incognite.

Quindi per confrontare le medie μ_x e μ_y ci riconduciamo ad effettuare un **test-t** sulla media μ_w con $\mu_0 = 0$.
test-t $\rightarrow$ *(test per la media di una popolazione normale, caso di varianza ignota)*.

Quindi i dati diventano:

- Media Campionaria: $\bar{w} = \frac{1}{n} \sum_{i=1}^{n} w_i$
- Deviazione standard: $\bar{s}_w = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (w_i - \bar{w})}$

I test e le ipotesi

Per effettuare i testo dobbiamo ricordare che $\mu_w = \mu_x - \mu_y$.
Quindi:

- $\mu_w = 0 \iff \mu_x = \mu_y$
- $\mu_w < 0 \iff \mu_x < \mu_y$
- $\mu_w > 0 \iff \mu_x > \mu_y$

Inoltre partiamo dall'ipotesi che $\mu_w = 0$.

Quindi ad esempio s evogliamo vedere l'efficacia di un farmaco contro il colesterolo (quindi che abbassa il valore), dobbiamo effettuare il seguente test:

- $H_0 : \mu_w = 0$ *il farmaco non ha avuto effetto perchè la media dei dati prima e dopo l'assunzione sono uguali.*, nello specifico: $\mu_x = \mu_y$
- $H_1 : \mu_w > 0$ *il farmaco ha avuto effetto perchè la media dopo l'assunzione è più bassa rispetto prima*, quindi: $\mu_x > \mu_y$ che porta la media finiale ad essere positiva.

Tabella di tutti i test

Test	H_0	H_1	ST	Rifiuto H_0 se
Confronto delle medie di due popolazioni normali, caso di varianze note	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$	$\lvert st \rvert > z_{\frac{\alpha}{2}}$
	$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$	$st > z_{\alpha}$
	$\mu_x \geq \mu_y$	$\mu_x < \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0,1)$	$st < -z_{\alpha}$
Confronto delle medie di due popolazioni normali, caso di varianze non note, MA UGUALI	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$	$\lvert st \rvert > t_{\frac{\alpha}{2}, n+m-2}$
	$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$	$st > t_{\alpha, n+m-2}$
	$\mu_x \geq \mu_y$	$\mu_x < \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$	$st < -t_{\alpha, n+m-2}$
Confronto delle medie di due popolazioni normali, caso di varianze non note e DIVERSE	$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$st = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim N(0,1)$	$\lvert st \rvert > z_{\frac{\alpha}{2}}$

Test	H_0	H_1	ST	Rifiuto H_0 se
	$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$st = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \sim N(0,1)$	$st > z_\alpha$
	$\mu_x \geq \mu_y$	$\mu_x < \mu_y$	$st = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \sim N(0,1)$	$st < -z_\alpha$
Confronto delle medie di due popolazioni normali nel caso di campioni accoppiati	$\mu_w = \mu_x - \mu_y$		i campioni diventano $w = x_i - y_i$ quindi si procede come una distribuzione sopra.	$st > t_{\alpha,n-1}$

Quando costruiamo i test per le popolazioni accoppiate **ricordiamo che**:

- $\mu_w = 0 \iff \mu_x = \mu_y$
- $\mu_w < 0 \iff \mu_x < \mu_y$
- $\mu_w > 0 \iff \mu_x > \mu_y$

Per le popolazioni accoppiate la media campionaria è:

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$$

La deviazione standard campionaria è:

$$\bar{s}_w = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(w_i - \bar{w})}$$

La varianza combinata campionaria si calcola come:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Nel caso le ampiezze dei campioni siano uguali si può usare la varianza campionaria comune:

$$S_p^2 = \frac{(S_x^2 + S_y^2)(n-1)}{2(n-1)}$$