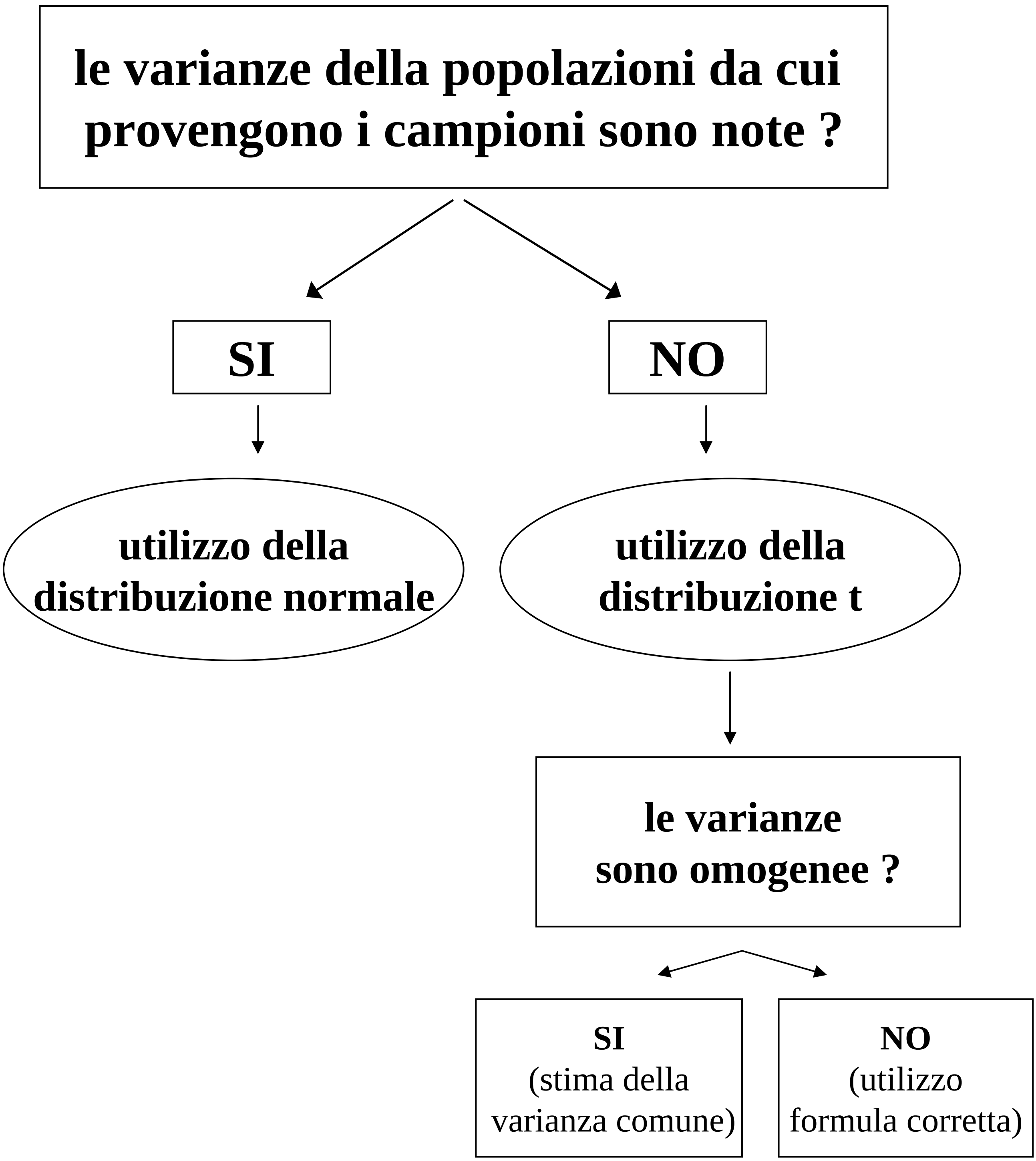
CONFRONTO TRA LA MEDIE DI DUE CAMPIONI INDIPENDENTI

ipotesi sul confronto tra le medie di due campioni indipendenti

Obiettivo: decidere, attraverso il confronto tra le medie dei due campioni indipendenti, se tali campioni provengono da due popolazioni diverse o meno.

informazioni sulla varianza del fenomeno



Confronto tra medie in due campioni indipendenti (σ è noto)

ASSUNZIONI

1) che entrambi i campioni siano distribuiti normalmente;

2) che siano tra loro indipendenti;

3) che le popolazioni da cui derivano abbiano varianze omogenee.

la distribuzione della differenza tra medie

quando σ *è noto*, la distribuzione campionaria della differenza tra le due medie ha le seguenti caratteristiche:

a) si distribuisce in forma normale

b) $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 = 0$

c) $\sigma^2_{\bar{x}_1 - \bar{x}_2} = \dfrac{\sigma^2}{n_1 + n_2}$

la distribuzione della differenza tra medie

σ *noto*

il test associato alla differenza tra le medie è:

$$z_{\bar{x}_1-\bar{x}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\dfrac{\sigma}{\sqrt{n_1+n_2}}}$$

esempio 1

Un ricercatore vuole sapere se vi siano differenze nell'atteggiamento verso l'attività extradomestica tra le donne sposate con figli e quelle senza figli. Allo scopo somministra una scala di atteggiamento a due campioni casuali di donne coniugate, di cui $n_1 = 45$ con figli e $n_2 = 36$ senza figli, ottenendo i seguenti punteggi medi: $\bar{X}_1 = 65$, $\bar{X}_2 = 75$. Ipotizzando che la distribuzione dei punteggi sulla scala di atteggiamento sia normale in entrambi i gruppi, con $\sigma_1 = \sigma_2 = 10$, si vuole sapere se i due campioni siano estratti da popolazioni con media uguale oppure no.

esempio 1$_{(2)}$

1. formulazione delle ipotesi

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

2. calcolo del valore z

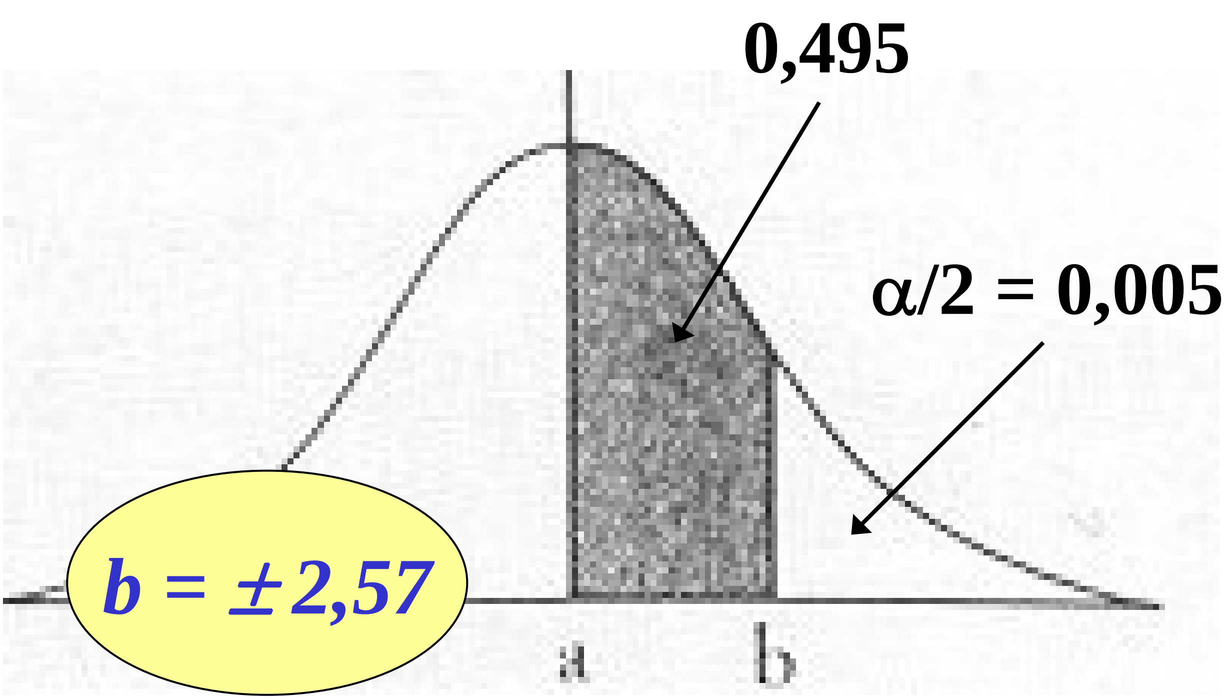
Usiamo la formula per grandi campioni e σ noto:

$$z_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma / \sqrt{n_1 + n_2}} = \frac{65 - 75}{10 / \sqrt{45 + 36}} = -9$$

esempio 1$_{(3)}$

3. determinazione dei valori critici

Fissato un livello di significatività $\alpha = 0{,}01$, e dato che il test è bidirezionale, bisogna trovare sulla tavola della normale il punto *b* che dà luogo ad un area di 0.495 ?



esempio 1$_{(4)}$

4. decisione

Poiché | z_{cal} | = 9 è maggiore del valore critico | z_c | = 2,58, dobbiamo rigettare l'ipotesi H_0.

CONCLUSIONE:

il punteggio medio delle donne con figli è significativamente diverso da quello delle donne senza figli;

l'atteggiamento dei due gruppi verso il lavoro extradomestico è differente.

confronto fra medie di due campioni indipendenti con varianze omogenee

σ ingnoto

La statistica che utilizziamo ha una distribuzione di probabilità che si approssima a quella del **t** di Student con $n_1 + n_2 - 2$ gradi di libertà.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

$\overline{X}_1 - \overline{X}_2$ è la differenza tra le medie calcolate nei due campioni

$(\mu_1 - \mu_2)$ è la differenza tra le medie delle due popolazioni

$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}$ è la stima della deviazione standard della distribuzione campionaria della differenza tra le medie

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}$$

$n_1 \ , n_2$ le numerosità dei due campioni

esempio 2

Un commerciante verifica la durata di due diverse marche di lampadine. Con 8 lampadine della marca A ottiene una media = *1237 ore* con s = *36*; con 7 lampadine della marca B ottiene una media di *1036 ore* con s = *40*. A fronte di tale risultato il commerciante vuole sapere se la differenza tra le due medie è tale da poter affermare con una probabilità del 95% che le lampadine di marca A hanno una durata superiore a quelle di marca B.

esempio 2$_{(2)}$

1. formulazione delle ipotesi

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A > \mu_B$$

2. calcolo del valore *t*

Usiamo la formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

esempio 2$_{(3)}$

2. calcolo del valore *t*

Per prima cosa dobbiamo stimare il valore della deviazione standard della differenza tra le medie:

$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2 n_1 + s_2^2 n_2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}} =$$

$$= \sqrt{\frac{36^2 \cdot 8 + 40^2 \cdot 7}{8+7-2} \cdot \frac{8+7}{8 \cdot 7}} = 19{,}63$$

esempio 2$_{(3)}$

2. calcolo del valore *t*

Quindi calcoliamo il valore di *t* con la formula:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\overline{x}_1 - \overline{x}_2}} =$$

$$= \frac{1237 - 1036}{19{,}63} = 10{,}24$$

esempio 2$_{(4)}$

3. determinazione dei valori critici

Il livello di significatività fissato è $\alpha = 0,05$; il test è unidirezionale, i gradi di libertà sono $(8 + 7 - 2) = 13$.

$$\boxed{t = 1,771}$$

4. decisione

Poiché $t_{cal} = 10,24$ è maggiore del valore critico $t_c = 1,771$, dobbiamo rigettare l'ipotesi H_0.

CONCLUSIONE:

Le lampadine della marca A sono migliori di quelle della marca B.

confronto fra medie di due campioni indipendenti con varianze non omogenee

σ ignoto

Se viene violato l'assunto di omogeneità delle varianze è necessario introdurre una correzione al test, rimane comunque necessario che la distribuzione delle popolazioni sia normale.

formula corretta nel caso di varianze non omogenee

La statistica viene calcolata con la formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1 - 1} + \dfrac{s_2^2}{n_2 - 1}}}$$

in cui i *gdl* sono dati dalla:

$$d = \frac{\left(\hat{\sigma}^2_{\bar{x}_1 - \bar{x}_2}\right)^2}{\dfrac{\left(\hat{\sigma}^2_{\bar{x}_1}\right)^2}{n_1 + 1} + \dfrac{\left(\hat{\sigma}^2_{\bar{x}_2}\right)^2}{n_2 + 1}} - 2$$

formula corretta nel caso di varianze non omogenee(2)

$$\hat{\sigma}^2_{\bar{x}_1-\bar{x}_2} = \frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}$$

è la stima della varianza della distribuzione campionaria della differenza tra le medie

$$\hat{\sigma}^2_{\bar{x}_1} = \frac{s_1^2}{n_1-1} \qquad \hat{\sigma}^2_{\bar{x}_2} = \frac{s_2^2}{n_2-1}$$

sono le stime delle varianze delle distribuzioni campionarie delle medie stimate a partire dalle varianze dei campioni

esempio 3

A due gruppi di $n_1 = 10$ e $n_2 = 26$ soggetti viene somministrato un test sull'ansia. Il primo gruppo ottiene un valore medio = *8* con $s_1 = 0{,}5$; il secondo gruppo un punteggio medio = *12* con $s_2 = 5$. Ci si chiede se i due gruppi differiscono relativamente al livello d'ansia.

Supponiamo che sia violato l'assunto di omogeneità delle varianze e che i due gruppi derivino da popolazioni con varianze non omogenee.

esempio 3$_{(2)}$

1. formulazione delle ipotesi

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

2. calcolo del valore *t*

Usiamo la formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\dfrac{s_1^2}{n_1 - 1} + \dfrac{s_2^2}{n_2 - 1}}}$$

esempio 3$_{(3)}$

2. calcolo del valore *t*

$$t = \frac{(8-12)}{\sqrt{\frac{(0,5)^2}{10-1} + \frac{(5)^2}{26-1}}} = -3,95$$

I cui gdl saranno dati da:

$$d = \frac{(1,03)^2}{\frac{(0,03)^2}{10+1} + \frac{(1)^2}{26+1}} - 2 = 26,58$$

esempio 3$_{(3)}$

3. determinazione dei valori critici

Il livello di significatività fissato è α = *0,05*; il test è bidirezionale, i gradi di libertà sono *26,58 = 27*.

$$\boxed{t = 2{,}052}$$

4. decisione

Poiché $| t_{cal} | = 3{,}95$ è maggiore del valore critico $| t_c | = 2{,}052$, dobbiamo rigettare l'ipotesi H_0.

CONCLUSIONE:
I due gruppi differiscono per livello d'ansia.

Approfondimento: metodo alternativo nel caso di varianze non omogenee

Nel caso le varianze non siano omogenee esiste un metodo alternativo, rispetto a quello appena visto, che prevede il calcolo di un *t* critico corretto attraverso la seguente formula:

$$t_c = \frac{\hat{\sigma}^2_{\bar{x}_1} t_{1c} + \hat{\sigma}^2_{\bar{x}_2} t_{2c}}{\hat{\sigma}^2_{\bar{x}_1 - \bar{x}_2}}$$

in cui:

t_{1c} è il valore critico di *t* con n_1-1 gdl (n_1 = numero di elementi del campione 1) e livello di sig. α

t_{2c} è il valore critico di *t* con n_2-1 gdl (n_2 = numero di elementi del campione 2) e livello di sig. α

esempio 3

metodo alternativo

Riprendiamo i dati dell'esercizio 3 e adottiamo il metodo alternativo.

determinazione dei valori critici

Il livello di significatività fissato è $\alpha = 0{,}05$; il test è bidirezionale,

t_{1c} = valore critico con 9 gdl: 2,26

t_{2c} = valore critico con 25 gdl: 2,06

$$t_c = \frac{(0{,}03)2{,}26 + (1)2{,}06}{1{,}01} = 2{,}11$$

decisione

Poiché $|\, t_{cal} \,| = 3{,}95$ è maggiore del valore critico $|\, t_c \,| = 2{,}11$, dobbiamo rigettare l'ipotesi H_0.

NOTA

In passato, la distribuzione t di Student veniva utilizzata solo per il confronto tra medie di due piccoli campioni ($n_1+n_2 < 30$), per evitare calcoli elaborati.

Attualmente, grazie alla diffusione dei calcolatori, la distribuzione t viene sempre utilizzata quando la varianza della popolazione è ignota, anche quando la numerosità campionaria è elevata.

Và infine ricordato, che per campioni molto numerosi ($n_1+n_2 > 100$), l'utilizzo della distribuzione t porta "praticamente" agli stessi risultati rispetto a quelli della distribuzione normale.

confronto fra medie di due campioni non indipendenti

Nel caso di coppie di osservazioni non indipendenti la ***media della distribuzione campionaria della differenza tra le medie*** risulta essere:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

la ***varianza*** della distribuzione campionaria della differenza tra le medie sarà:

$$\hat{\sigma}^2_{\bar{x}_1 - \bar{x}_2} = \hat{\sigma}^2_{\bar{x}_1} + \hat{\sigma}^2_{\bar{x}_2} - 2\hat{\sigma}_{\bar{x}_1}\hat{\sigma}_{\bar{x}_2} r_{\bar{x}_1\bar{x}_2}$$

in cui l'ultimo termine $r_{\bar{x}_1\bar{x}_2}$ è la correlazione tra le medie di tutti i possibili campioni non indipendenti tratti dalle popolazioni in esame.

Poiché non si conosce il valore di $r_{\bar{x}_1\bar{x}_2}$ ***è impossibile utilizzare la distribuzione campionaria della differenza tra le medie*;**

per superare questo inconveniente si considera ***un unico campione*** costituito da coppie di elementi appaiati;

il punteggio cui si fa riferimento è dato dalla ***differenza tra i punteggi*** di ciascuna coppia;

nell'ipotesi H_0, se non vi sono differenze tra le due serie di punteggi, la media delle differenze risulterà

ZERO

differenza dei punteggi

Di tale differenza possiamo calcolare la ***media*** con la formula:

$$\overline{X}_D = \frac{\sum D_i}{n} = \frac{\sum (X_{1i} - X_{2i})}{n} = \overline{X}_1 - \overline{X}_2$$

e calcolare la ***varianza*** con:

$$s_D^2 = \frac{\sum D_i^2}{n} - \left(\frac{\sum D_i}{n} \right)^2$$

distribuzione campionaria della differenza dei punteggi

1. la ***media*** è pari alla differenza tra le medie delle popolazioni da cui sono tratti i campioni

$$\mu_D = \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

2. la ***varianza*** è:

$$\hat{\sigma}_D^2 = \frac{s^2}{n-1} = \frac{1}{n-1}\left[\frac{\sum D_i^2}{n} - \left(\frac{\sum D_i}{n}\right)^2\right]$$

distribuzione campionaria della differenza dei punteggi

3. il test ***t*** avrà la seguente forma

$$t = \frac{\overline{X}_D}{\frac{s_D}{\sqrt{n-1}}}$$

la stessa statistica si può calcolare direttamente dai dati grezzi con la formula:

$$t = \frac{\sum D}{\sqrt{\frac{\left(n\sum D^2\right) - \left(\sum D\right)^2}{(n-1)}}}$$

esempio 4

Si vuole studiare l'effetto dell'affaticamento sul rendimento in una prova di precisione. A questo scopo si contano il numero di errori commessi da un gruppo di *10* soggetti in una prova di precisione. Dopo averli sottoposti ad un lavoro gravoso per un certo periodo di tempo, si contano nuovamente gli errori commessi dai *10* soggetti nella stessa prova di precisione.

I dati ottenuti sono riportati nella tabella seguente.

esempio 4$_{(2)}$

sogg.	*numero di errori*		*differenza*	
	prova 1	*prova 2*	$D = X_1 - X_2$	D^2
A	10	12	-2	4
B	8	9	-1	1
C	13	15	-2	4
D	12	13	-1	1
E	14	16	-2	4
F	12	11	1	1
G	11	13	-2	4
H	18	18	0	0
I	9	10	-1	1
L	16	16	0	0
	12.3	13.3	-10	20
	medie		*somme*	

esempio 4$_{(3)}$

1. formulazione delle ipotesi

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

2. calcolo del valore t

Usiamo la formula:

$$t = \frac{\sum D}{\sqrt{\dfrac{\left(n\sum D^2\right) - \left(\sum D\right)^2}{(n-1)}}}$$

esempio 4$_{(3)}$

1. formulazione delle ipotesi

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

2. calcolo del valore *t*

Usiamo la formula:

$$t = \frac{-10}{\sqrt{\dfrac{(10 \cdot 20) - (-10)^2}{(10-1)}}} = -3$$

esempio 4$_{(4)}$

3. determinazione dei valori critici

Il livello di significatività fissato è $\alpha = 0,05$; il test è bidirezionale, i gradi di libertà sono $(10\text{-}1) = 9$.

$$\boxed{t = 2,262}$$

4. decisione

Poiché $|t_{cal}| = 3$ è maggiore del valore critico $|t_c| = 2,262$, dobbiamo rigettare l'ipotesi H_0.

CONCLUSIONE:

L'affaticamento influisce sui risultati della prova.