Confronto fra una media campionaria ed un valore ipotetico

E' possibile stabilire se una media campionaria sia statisticamente uguale, maggiore, minore o, più genericamente, diversa da un valore ipotetico, μ_0, confrontando con i limiti dell'intervallo di fiducia il valore (detto realizzazione della statistica, t) che assume la variabile random impiegata nel calcolo dell'intervallo stesso quando si usa il valore μ_0 al posto del generico μ. Si possono configurare dunque 3 casi:

Caso	ipotesi a confronto	distribuzione dei dati tipo di confronto (test)	criteri di rigetto dell'ipotesi base
1: distribuzione normale σ^2 nota	$\mu = \mu_0$	$T = (\bar{X} - \mu_0) / (\sigma/\sqrt{n}) \sim N(0,1)$	
	$\mu > \mu_0$	una coda	$t \geq z_{(1-\alpha)}$
	$\mu < \mu_0$	una coda	$t \leq -z_{(1-\alpha)}$
	$\mu \neq \mu_0$	due code	$\lvert t \rvert \geq z_{(1-\alpha/2)}$
2: distribuzione normale σ^2 ignota, n > 30	$\mu = \mu_0$	$T = (\bar{X} - \mu_0) / (s/\sqrt{n}) \sim N(0,1)$	
	$\mu > \mu_0$	una coda	$t \geq z_{(1-\alpha)}$
	$\mu < \mu_0$	una coda	$t \leq -z_{(1-\alpha)}$
	$\mu \neq \mu_0$	due code	$\lvert t \rvert \geq z_{(1-\alpha/2)}$
3: distribuzione normale σ^2 ignota, n < 30	$\mu = \mu_0$	$T = (\bar{X} - \mu_0) / (s/\sqrt{n}) \sim t_{n-1}$	
	$\mu > \mu_0$	una coda	$t \geq t_{n-1, (1-\alpha)}$
	$\mu < \mu_0$	una coda	$t \leq -t_{n-1, (1-\alpha)}$
	$\mu \neq \mu_0$	due code	$\lvert t \rvert \geq t_{n-1, (1-\alpha/2)}$

Esempi numerici di confronto sulla media

Caso 1

Supponiamo che un programma abbia generato la seguente serie di 9 numeri random, che dovrebbero essere distribuiti secondo N(0,1):

+0.250 +1.620 -0.052 +0.014 -0.366 +0.756 +0.608 -2.150 +1.162

Si vuole stabilire su base statistica se

$\mu = 0$ o $\mu \neq 0$

In questo caso:

$n = 9$ $\overline{X} = 0.205$ $\sigma^2 = 1$ (il valore è noto in questo caso)

La realizzazione della statistica è:

$$\boxed{t = (\overline{X} - \mu_0) / (\sigma/\sqrt{n}) = 0.205/(1/3) = 0.615}$$

Se si adotta un livello di significatività del 5 %, $\alpha = 0.05$, il valore critico per la decisione è: $z_{(1-\alpha/2)} = z_{(0.975)} = 1.96$

Poiché $t = 0.615 < 1.96$, l'ipotesi $\mu = 0$ può essere accettata.

Caso 2

Riconsideriamo i dati del ΔH di neutralizzazione di HCl con NaOH per valutare se il loro valore medio sia uguale a quello di letteratura, 56.40 kJ/mol, o diverso.

In questo caso:

$n = 65$ (quindi $n > 30$) $\bar{X} = 57.34$ σ è ignota, mentre $s = 1.65$

La realizzazione della statistica T è:

$$\boxed{t = (\bar{X} - \mu_0) / (s/\sqrt{n}) = (57.34-56.40)/(1.65/\sqrt{65}) = 4.59}$$

Se si adotta un livello di significatività del 5%, $\alpha = 0.05$ e il valore critico per la decisione è:

$$z_{(1-\alpha/2)} = z_{(0.975)} = 1.96$$

Poiché $t = 4.59 > 1.96$ si può affermare che il valore sperimentale del ΔH NON è in accordo con quello di letteratura.

Una più attenta valutazione del dato di letteratura porterebbe a scoprire che esso si riferisce ad una temperatura di 25 °C ed è estrapolato a diluizione infinita.
Il ΔH sperimentale è stato invece ricavato a 20 °C e a concentrazione 0.1 M.

Il dato di letteratura a 20 °C e a concentrazione 0.1 M è 57.28 kJ/mol.

Sarebbe dunque più corretto valutare l'ipotesi:

μ = **57.28 kJ/mol**

Risulterebbe quindi:

$$t = (\overline{X} - \mu_0) / (s/\sqrt{n}) = (57.34-57.28)/(1.65/\sqrt{65}) = 0.29$$

Poiché $t = 0.29 < 1.96 = z_{(0.975)}$ l'eguaglianza può essere accettata.

Caso 3

I dati seguenti rappresentano il grado di associazione medio, m, delle molecole di butil-litio in soluzione benzenica, $(C_4H_9Li)_m$, ottenuto da misure di abbassamento della tensione di vapore:

5.980 6.089 6.198 6.151 6.479 6.105

Il confronto proposto è fra l'ipotesi

μ = 6 (valore teorico) e l'alternativa $\mu > 6$

n = 6 (quindi n < 30) $\overline{X}$ = 6.167 σ è ignota mentre s = 0.169

La realizzazione della statistica T è

$$\boxed{t = (\overline{X} - \mu_0) / (s/\sqrt{n}) = (6.167-6)/(0.169/\sqrt{6}) = 2.42}$$

Se si adotta un livello di significatività del 5 %, α = 0.05, il valore critico per la decisione (il test è ad una coda in questo caso) è $t_{5,0.95} = 2.02$

Poiché t = 2.42 > 2.02, il valore sperimentale di m è significativamente superiore a quello teorico.

Confronto fra due medie campionarie

Supponiamo di avere a disposizione due set di dati:

$x_1, x_2, \ldots, x_{n1} \sim N(\mu_1, \sigma_1^2)$ e $y_1, y_2, \ldots, y_{n2} \sim N(\mu_2, \sigma_2^2)$

e di volerne confrontare le medie. Essi possono ad esempio derivare da due metodi analitici diversi applicati allo stesso campione.

Il confronto può essere fatto valutando se la differenza fra le due medie sia statisticamente uguale a 0 o meno.

Come per l'intervallo di fiducia relativo alla differenza fra due medie, sono possibili quattro casi:

- ✓ caso 1: le varianze di popolazione (σ_1^2 e σ_2^2) sono note;
- ✓ caso 2: le varianze di popolazione (σ_1^2 e σ_2^2) sono ignote; le dimensioni dei campioni sono sufficientemente grandi (n_1 ed $n_2 > 30$);
- ✓ caso 3: le varianze di popolazione (σ_1^2 e σ_2^2) sono ignote ma uguali (σ^2); le dimensioni dei campioni sono esigue (n_1 e/o $n_2 < 30$);
- ✓ caso 4: le varianze di popolazione (σ_1^2 e σ_2^2) sono ignote e diverse; le dimensioni dei campioni sono esigue (n_1 ed $n_2 < 30$) $\Rightarrow$ Problema di Fisher-Behrens

Caso	ipotesi a confronto	distribuzione sotto l'ipotesi base	criteri di rigetto dell'ipotesi base
1	$\mu_1 - \mu_2 = 0$ $\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	$T = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$t \geq z_{(1-\alpha)}$ $t \leq -z_{(1-\alpha)}$ $\lvert t \rvert \geq z_{(1-\alpha/2)}$
2	$\mu_1 - \mu_2 = 0$ $\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	$T = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim N(0,1)$	$t \geq z_{(1-\alpha)}$ $t \leq -z_{(1-\alpha)}$ $\lvert t \rvert \geq z_{(1-\alpha/2)}$
3	$\mu_1 - \mu_2 = 0$ $\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	$T = \dfrac{(\bar{X} - \bar{Y})}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n1+n2-2}$	$t \geq t_{n1+n2-2,\,(1-\alpha)}$ $t \leq -t_{n1+n2-2,\,(1-\alpha)}$ $\lvert t \rvert \geq t_{n1+n2-2,\,(1-\alpha/2)}$
4	$\mu_1 - \mu_2 = 0$ $\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	$T = \dfrac{(\bar{X} - \bar{Y})}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t_\nu$	$t \geq t_{\nu,\,(1-\alpha)}$ $t \leq -t_{\nu,\,(1-\alpha)}$ $\lvert t \rvert \geq t_{\nu,\,(1-\alpha/2)}$

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

$$\frac{1}{\nu} = \frac{1}{n_1 - 1}\left(\frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2 / n_2}{s_1^2 / n_1 + s_2^2 / n_2}\right)^2$$

Esempi numerici di confronto fra due medie

Caso 2

Consideriamo i due set di dati di ΔH di neutralizzazione di HCl con NaOH già esaminati per stabilire l'intervallo di fiducia per la differenza fra le loro medie:

$\overline{X} = 57.34$, $s_1^2 = 2.726$, $n_1 = 65$

$\overline{Y} = 56.99$, $s_2^2 = 1.130$, $n_2 = 32$

Mettiamo a confronto le seguenti ipotesi:

$\mu_1 - \mu_2 = 0$ $\Rightarrow$ non c'e' differenza significativa fra le due medie

$\mu_1 - \mu_2 \neq 0$ $\Rightarrow$ c'e' differenza significativa fra le due medie

La realizzazione della statistica T è:

$$t = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{57.34 - 56.99}{\sqrt{\frac{2.726}{65} + \frac{1.130}{32}}} = 1.26$$

Scegliendo un livello di significatività del 5% il valore critico per la decisione è (l'ipotesi alternativa è che le due medie siano diverse):

$z_{0.975} = 1.96$.

Poiché $t = 1.26 < 1.96$ i due set di dati forniscono medie che non differiscono significativamente fra loro.

Caso 3

Consideriamo i due set di dati relativi alla densità di un polimero, già esaminati per stabilire l'intervallo di fiducia per la differenza fra le loro medie:

1) 1.2151 1.2153 1.2155 1.2145 1.2151

$\overline{X} = 1.21510$, $s_1^2 = 14.0 \times 10^{-8}$, $n_1 = 5$

2) 1.2167 1.2176 1.2157 1.2158 1.2167

$\overline{Y} = 1.21650$, $s_2^2 = 60.5 \times 10^{-8}$, $n_2 = 5$

Confrontiamo le seguenti ipotesi:

$\mu_1 - \mu_2 = 0$ $\Rightarrow$ non c'e' differenza significativa fra le due medie

$\mu_1 - \mu_2 \neq 0$ $\Rightarrow$ c'e' differenza significativa fra le due medie

Calcoliamo: $\boxed{s = \sqrt{\dfrac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} = 6.10 \times 10^{-4}}$

La realizzazione della statistica T è:

$$t = \frac{(\overline{X} - \overline{Y})}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1.21510 - 1.21650}{6.10 \times 10^{-4}\sqrt{\frac{2}{5}}} = -3.63$$

Scegliendo un livello di significatività del 5% il valore critico per la decisione è:

$t_{n1+n2-2,\ (1-a/2)} = t_{8,(0.975)} = 2.31.$

Poiché $t = -3.63 < -2.31$ i due set di dati forniscono medie che differiscono significativamente fra loro.

Caso 4

Supponiamo di avere a disposizione i seguenti set di dati, derivanti dall'applicazione di due diversi metodi analitici allo stesso campione:

Dato	Metodo A	Metodo B
1	5.11	5.18
2	5.14	5.13
3	5.13	5.27
4	5.17	5.12
5	5.12	5.27
6	5.08	5.29
7	5.15	5.17
8	5.20	5.28
9	5.16	5.14
10	5.14	5.18

Metodo A

$\overline{X} = 5.140$

$s_1^2 = 11.11 \times 10^{-4}$

$n_1 = 10$

Metodo B

$\overline{Y} = 5.203$

$s_2^2 = 45.34 \times 10^{-4}$

$n_2 = 10$

Questa situazione corrisponde al caso 4, per cui va prima calcolato il numero di gradi di libertà ν:

$$\boxed{\frac{1}{\nu} = \frac{1}{9}\left(\frac{1.111}{5.645}\right)^2 + \frac{1}{9}\left(\frac{4.534}{5.645}\right)^2} \quad \Rightarrow \quad \nu = 13.2 \approx 13$$

Le ipotesi da confrontare sono:

$\mu_A - \mu_B = 0$ e $\mu_A - \mu_B \neq 0$

La realizzazione della variabile random T è:

$$t = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.140 - 5.203}{\sqrt{\left(\frac{11.11 \times 10^{-4}}{10} + \frac{45.34 \times 10^{-4}}{10}\right)}} = -2.65$$

Se si decide di adottare un livello di significatività del 5% (ossia un livello di fiducia del 95 %), $1-\alpha/2 = 0.975$ per cui occorre calcolare

$t_{13,(0.975)} = 2.16$.

Poiché $t = -2.65 < -2.16$ si può affermare che:

i due metodi forniscono risultati diversi, al 95% di fiducia.

Decisioni successive ad un confronto

Dopo aver ottenuto i risultati di un confronto occorre decidere quali valori usare per i parametri coinvolti, a seconda del suo esito.

✓ Confronto fra una media ed un valore noto: $\mu = \mu_0$

1) se l'ipotesi indicata viene accettata si userà per la media il valore μ_0;

2) se l'ipotesi viene rigettata la migliore stima per μ rimarrà $\overline{X}$

✓ Confronto fra due medie: $\mu_1 - \mu_2 = \mu_0$ (caso generale)

1) se l'ipotesi viene accettata si useranno:

➢ il valore μ_0 per la differenza fra le due medie,

➢ il valore $\left(\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i\right) / (n_1 + n_2) + \mu_0/2$ per μ_1,

➢ il valore $\left(\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i\right) / (n_1 + n_2) - \mu_0/2$ per μ_2

Come si può notare, se $\mu_0 = 0$ e l'ipotesi viene accettata, per ciascuna delle due medie si userà il valore:

$$\boxed{\left(\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i\right) \Big/ \left(n_1 + n_2\right)}$$

che corrisponde alla grande media campionaria.

2) se l'ipotesi viene rigettata si useranno in tutti i casi le rispettive medie campionarie come migliori stime di μ_1 e μ_2.