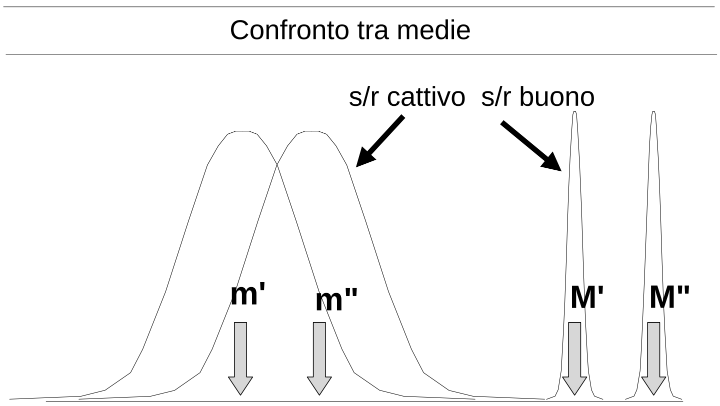
7. Test t di Student per campioni indipendenti

Si considerino due campioni con media m’ e media m” rispettivamente. La domanda che ci si pone è: la media m’ è significativamente diversa dalla media m”?



La risposta a questa domanda può essere assicurata sulla base di una tecnica statistica parametrica (il test t di Student) che si basa su un concetto che si rifà a quello di rapporto segnale/rumore. Molto semplicemente nel caso delle medie m’ e m”, sarà poco probabile che vi sia una differenza significativa, a causa del fatto che le medie sono caratterizzate da un grado di incertezza (che corrisponde all’embricarsi delle due distribuzioni a confronto) elevato.

Si considerino ora due campioni con media M’ e media M” rispettivamente. La domanda che ci si pone è nuovamente: la media M’ è significativamente diversa dalla media M”? Nel caso delle medie M’ e M”, esiste intuitivamente una differenza significativa, a causa del fatto che le medie sono caratterizzate da un grado di incertezza (che corrisponde all’embricarsi delle due distribuzioni a confronto) sostanzialmente nullo.

Quindi il test t di Student per il confronto fra le medie di campioni indipendenti è basato su un principio semplice e intuitivo: a parità di valore assoluto della differenza fra le medie, a tale differenza viene data tanto maggior peso quanto minore è la dispersione dei dati campionari (cioè in definitiva quanto meno le due distribuzioni campionarie si sovrappongono), e viceversa.

Dati allora due campioni, il primo comprendente n_1 osservazioni x_1, aventi media $\bar{x}_1$, e il secondo comprendente n_2 osservazioni x_2, aventi media $\bar{x}_2$, il valore t viene calcolato come

$$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s^2 / n_1 + s^2 / n_2)}$$

con gradi di libertà pari a

$$n_1 + n_2 - 2$$

Al numeratore compare la differenza fra le medie, mentre al denominatore compare l'errore standard di tale differenza, la grandezza che, come si è precedentemente accennato, consente di dare il peso maggiore o minore alla differenza fra le medie.
La varianza s^2 è la varianza combinata dei due campioni, e viene calcolata come rapporto fra la somma delle devianza dei due campioni e la somma dei loro gradi di libertà, cioè come

$$s^2 = (\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2) / (n_1 + n_2 - 2)$$

E' possibile semplificare il calcolo della varianza s_1^2 e della varianza s_2^2 ricordando che

$$\Sigma(x_1 - \bar{x}_1)^2 = \Sigma x_1^2 - (\Sigma x_1)^2 / n_1$$

$$\Sigma (x_2 - \bar{x}_2)^2 = \Sigma x_2^2 - (\Sigma x_2)^2 / n_2$$

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.

Un problema tuttavia complica lievemente la situazione: il test *t* di Student ordinario tende a fornire troppo pochi risultati significativi quando il campione più grande ha la varianza (cioè la dispersione dei dati) maggiore, e troppi risultati significativi quando il campione più grande ha la varianza minore. E' perciò necessario utilizzare, nei casi in cui la varianza dei due campioni differisca in maniera significativa, una forma del test che preveda una opportuna correzione. Per fare questo si calcola il rapporto tra varianze.

La varianza del primo campione viene calcolata come

$$s_1^2 = \Sigma (x_1 - \bar{x}_1)^2 / (n_1 - 1)$$

con gradi di libertà pari a

$$n_1 - 1$$

la varianza del secondo campione viene calcolata come

$$s_2^2 = \Sigma (x_2 - \bar{x}_2)^2 / (n_2 - 1)$$

con gradi di libertà pari a

$$n_2 - 1$$

Il valore del test *F* di omogeneità fra le varianze viene calcolato allora come rapporto fra la varianza maggiore e la varianza minore, cioè

$$F = s_1^2 / s_2^2 \qquad se \;\; s_1^2 > s_2^2$$

ovvero

$$F = s_2^2 / s_1^2 \qquad se \;\; s_2^2 > s_1^2$$

Il valore di *p* corrispondente alla statistica *F* rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze.

Nel caso di varianze significativamente diverse (varianze non omogenee) si calcola la statistica

$$t' = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s_1^2 / n_1 + s_2^2 / n_2)}$$

con gradi di libertà uguali a

$$(v_1 + v_2)^2 / (v_1^2 / (n_1 - 1) + v_2^2 / (n_2 - 1))$$

essendo rispettivamente

$$v_1 = s_1^2 / n_1$$
$$v_2 = s_2^2 / n_2$$

Il valore di *p* corrispondente alla statistica *t* rappresenta la probabilità di osservare per caso una differenza della grandezza di quella effettivamente osservata: se tale probabilità è sufficientemente piccola, si conclude per una significatività della differenza fra le medie.