Intervalli di confidenza

Francesco Lagona

1 Introduzione

Questa dispensa riassume schematicamente i principali risultati discussi a lezione sulla costruzione di intervalli di confidenza.

2 Intervalli di confidenza per la media di una popolazione

Supponiamo di aver a che fare con una variabile statistica quantitativa X che si distribuisce nella popolazione di riferimento con media μ e varianza σ^2. Si desidera costruire un intervallo di confidenza per μ al livello $1-\alpha$ sulla base di un campione casuale semplice

$$(x_1 \dots x_n)$$

di dimensione n. È necessario distinguere il caso in cui la varianza della popolazione σ^2 è nota da quello in cui tale varianza è incognita.

2.1 Varianza nota

Si tratta di un caso abbastanza raro nelle applicazioni, ma in certe circostanze è possibile che indagini precedenti a quella effettuata rendano possibile una conoscenza esatta della varianza σ^2. La costruzione di un intervallo di confidenza per μ sotto l'assunzione di varianza nota, si basa sul seguente risultato: la **media campionaria**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

è una variabile aleatoria che si distribuisce approssimativamente come una normale

$$N(\mu, \frac{\sigma^2}{n})$$

e tale approssimazione migliora all'aumentare della dimensione campionaria n.

Se dunque usiamo la media campionaria come stimatore della media della popolazione, il fatto che la sua distribuzione sia centrata sul valore vero del parametro μ indica che $\bar{x}$ è uno stimatore non distorto. Inoltre, il rapporto $\frac{\sigma^2}{n}$ misura la precisione dello stimatore: come ci si potrebbe aspettare, tale precisione è tanto minore quanto più elevata è la varianza σ^2 e tanto maggiore quanto più elevata è la dimensione campionaria n.

In taluni casi, la variabile X si distribuisce esattamente come una normale: solo in queste circostanze $\bar{x}$ si distribuisce esattamente secondo la normale $N(\mu, \sigma^2/n)$. In tutti gli altri casi, la distribuzione della media coampionaria è solo approssimata e dunque i risultati che seguono valgono in modo approssimato, sebbene è importante ricordare che la qualità dell'approssimazione migliora al crescere di n.

Dal fatto che $\bar{x} \sim N(\mu, \sigma^2/n)$, si deduce che

$$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1).$$

Per ogni valore di probabilità $1 - \alpha$, possiamo allora scrivere che

$$P(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\alpha/2}) = 1 - \alpha$$

dove $z_{\alpha/2}$ è il quantile della normale di ordine $1 - \alpha/2$, ovvero il punto che si lascia a sinistra un'area sotto la normale pari a $1 - \alpha/2$. Ad esempio, se $1 - \alpha = 0.95$, allora $z_{\alpha/2} = 1.96$ (il calcolo del quantile $z_{\alpha/2}$ corrispondente al livello di probabilità $1 - \alpha$ va compiuto usando le opportune tavole o un PC).

Un intervallo di confidenza può allora essere costruito sulla base della seguente catena di uguaglianze:

$$\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{\alpha/2}) = 1 - \alpha \\
&= P(-z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \bar{x} - \mu \leq z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}) \\
&= P(-\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq -\mu \leq -\bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}) \\
&= P(\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}})
\end{aligned}$$

In altre parole, è approssimativamente uguale a $1-\alpha$ la probabilità che i due estremi dell'intervallo

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

contengano il valore "vero" della media μ della popolazione.

Quello appena costruito è un **intervallo di confidenza** per la media μ al livello $1-\alpha$. Il valore $1-\alpha$ indica il livello di copertura fornito dall'intervallo: esiste sempre una probabilità pari ad α che i dati campionari provengano da una popolazione con una media che si trova al di fuori dell'intervallo.

Si osservi che l'intervallo che abbiamo costruito è centrato sulla stima puntuale della media $\bar{x}$ e ha un "raggio" pari a

$$z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

la cui lunghezza dipende sia dal livello di copertura desiderato (da cui dipende il quantile $z_{\alpha/2}$), sia dal grado di precisione dello stimatore misurato dalla quantità

$$\sqrt{\frac{\sigma^2}{n}}$$

meglio nota come **errore standard** della stima.

Come applicazione numerica, consideriamo il seguente esempio.

Esempio Da informazioni derivanti da una precedente analisi, si sa che la durata delle telefonate che arrivano ad un *call center* si distribuisce con una varianza pari a $\sigma^2 = 16$ minuti quadrati. Si vuole calcolare un intervallo di confidenza al livello $1 - \alpha = 0.95$ per la durata media delle telefonate. A tale scopo, si estrae un campione di $n = 10$ telefonate che fornisce le seguenti durate:

$$7.36, 11.91, 12.91, 9.77, 5.99, 10.91, 9.57, 11.01, 6.11, 12.12$$

Il calcolo dell'intervallo desiderato è a questo punto piuttosto semplice: si calcola dapprima la media campionaria ed il suo errore standard

$$\bar{x} = 9.766$$

$$\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{16}{10}} = 1.265$$

Se inoltre $1 - \alpha = 0.95$, il quantile desiderato è dato da

$$z_{0.025} = 1.96$$

per cui il raggio dell'intervallo è dato da

$$z_{0.025}\sqrt{\frac{16}{10}} = 2.479$$

e l'intervallo è dunque dato da

$$(9.766 - 2.479, 9.766 + 2.479) = (7.287, 12.245).$$

2.2 Varianza incognita

Nella maggior parte delle applicazioni, è difficile avere una stima attendibile della varianza σ^2 della popolazione e si preferisce in genere stimarla sulla base del campione estratto. Una stima non distorta della varianza della popolazione è data da

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2\right)$$

che non è altro che la varianza campionaria corretta dal fattore $\frac{n}{n-1}$. Tale correzione dipende dal fatto che, per piccoli campioni, la varianza campionaria è uno stimatore distorto della varianza della popolazione, cioè la sua distribuzione campionaria non ha come valore atteso il valore vero del parametro σ^2. Per grandi campioni, il fattore di correzione $\frac{n}{n-1} \approx 1$ e dunque l'uso della varianza campionaria fornisce stime attendibili della varianza della popolazione.

In questo caso, per costruire un intervallo di confidenza della media μ della popolazione, occorre utilizzare il fatto che la distribuzione della variabile aleatoria

$$\frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$$

segue approssimativamente quella di una t di Student con $n - 1$ gradi di libertà, dove n è la dimensione del campione estratto e che tale approssimazione migliora all'aumentare di n. La distribuzione t di Student è molto simile a quella di una normale standardizzata. Essa è infatti centrata sullo 0 e simmetrica rispetto ad esso. Si differenzia dalla distribuzione normale in quanto ha delle code "più" pesanti, ovvero valori lontani dallo 0 hanno una probabilità di essere estratti più elevata di quella che avrebbero avuto se fossero stati estratti da una normale standardizzata. Tali differenze si attenuano sempre più all'aumentare della numerosità campionaria, per cui quando n è molto elevato, si può utilizzare la distribuzione normale standardizzata in luogo della t.

La costruzione dell'intervallo di confidenza segue linee analoghe a quelle mostrate nella sezione precedente. Si indichi pertanto con $t_{n-1,\alpha/2}$ il quantile di ordine $1-\alpha/2$ di una t di Student di $n-1$ gradi di libertà, ovvero il punto che si lascia a sinistra un'area sotto la t pari a $1-\alpha/2$. Ad esempio, se $1-\alpha = 0.95$ e il campione ha numerosità $n = 10$, allora $t_{n-1,\alpha/2} = 2.262$ (il calcolo del quantile $t_{n-1,\alpha/2}$ corrispondente al livello di probabilità $1-\alpha$ va compiuto usando le opportune tavole o un PC).

Un intervallo di confidenza può allora essere costruito sulla base della seguente catena di uguaglianze:

$$\begin{aligned}
1-\alpha &= P(-t_{n-1,\alpha/2} \leq \frac{\bar{x}-\mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \leq t_{n-1,\alpha/2}) = 1-\alpha \\
&= P(-t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} \leq \bar{x}-\mu \leq t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}) \\
&= P(-\bar{x}-t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} \leq -\mu \leq -\bar{x}+t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}) \\
&= P(\bar{x}-t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}} \leq \mu \leq \bar{x}+t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}})
\end{aligned}$$

In altre parole, è approssimativamente uguale a $1-\alpha$ la probabilità che i due estremi dell'intervallo

$$\left(\bar{x}-t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}, \bar{x}+t_{n-1,\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}\right)$$

contengano il valore "vero" della media μ della popolazione.

Considerando l'esempio precedente sulle durate delle telefonate, un intervallo di confidenza costruito stimando la varianza della popolazione al livello $1-\alpha = 0.95$ può essere costruito stimando dapprima la varianza della popolazione

$$\hat{\sigma}^2 = \frac{n}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \frac{10}{9}5.633 = 6.259$$

calcolando poi l'errore standard della stima

$$\sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{\frac{6.259}{10}} = 0.791$$

e infine il raggio dell'intervallo dato da:

$$t_{9,0.025}\sqrt{\frac{\hat{\sigma}^2}{n}} = 2.262 \cdot 0.791 = 1.789.$$

Si osservi come il raggio di questo intervallo di confidenza è minore di quello trovato nella sezione precedente: la ragione risiede nel fatto che il campione ha fornito una stima della varianza inferiore alla varianza vera della popolazione (la dimensione campionaria deve essere sufficientemente elevata per dare stime affidabili della varianza della popolazione). Ne segue un intervallo di confidenza più stretto di quello trovato in precedenza:

$$(9.766 - 1.789, 9.766 + 1.789) = (7.977, 11.555).$$

3 Calcolare la numerosita' campionaria

L'ampiezza dell'intervallo di confidenza per la media di una popolazione è data da

$$d = 2z_{\alpha/2}\sqrt{\sigma^2/n}$$

nel caso di varianza nota. E' facile osservare che, a parita' del livello $1-\alpha$ scelto per l'intervallo di confidenza e della varianza nella popolazione, l'ampiezza dell'intervalo dipende dalla dimensione campionaria n, al crescere della quale l'ampiezza si riduce.

In molti casi applicativi, la dimensione campionaria n e' fissata in partenza e dipende dal *budget* a disposizione per l'estrazione del campione. In altri casi (ad esempio in test clinici o in controllo della qualità) è più importante fissare l'ampiezza d^* che l'intervallo non può superare e determinare la dimensione campionaria minima n^* che garantisce tale requisito, cioè tale per cui quando $n < n^*$ si ottiene un intervallo con ampiezza $d > d^*$ (ovviamente, per tutti gli $n > n^*$ si ottiene un intervallo con ampiezza $d < d^*$).

Per effettuare il calcolo di n^* e' sufficiente osservare che se deve essere

$$2z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq d^*$$

allora

$$\sqrt{\frac{\sigma^2}{n}} \leq \frac{d^*}{2z_{\alpha/2}}$$

ovvero

$$\frac{\sigma^2}{n} \leq \left(\frac{d^*}{2z_{\alpha/2}}\right)^2$$

o infine

$$\left(\frac{2\sigma z_{\alpha/2}}{d^*}\right)^2 \leq n \tag{1}$$

In altre parole, per ottenere un intervallo di confidenza di un'ampiezza non superiore a d^*, è necessario considerare il minimo intero n che verifica la (1), ovvero

$$n^* = \left\lceil \left(\frac{2\sigma z_{\alpha/2}}{d^*} \right)^2 \right\rceil$$

dove con $\lceil x \rceil$ indichiamo il piu' piccolo intero superiore ad x (ad esempio: $\lceil 4.1 \rceil = 5$; la funzione $\lceil x \rceil$ si chiama 'cielo' di x). Come applicazione numerica, consideriamo il seguente esempio.

Esempio Da informazioni derivanti da una precedente analisi, si sa che la durata delle telefonate che arrivano ad un *call center* si distribuisce in modo approssimativamente normale con media μ incognita e varianza $\sigma^2 = 16$ minuti quadrati. Si desidera calcolare la dimensione campionaria minima necessaria per costruire un intervallo della durata media delle chiamate al livello 95% che abbia un'ampiezza massima di 5 minuti. La dimensione richiesta e' data da

$$n^* = \left\lceil \left(\frac{2\sigma z_{\alpha/2}}{d^*} \right)^2 \right\rceil = \left\lceil \left(\frac{2 \cdot 4 \cdot 1.96}{5} \right)^2 \right\rceil = \lceil 9.83 \rceil = 10$$

Si osservi che la conoscenza di σ^2 è cruciale per la determinazione della dimensione campionaria ottimale. Quando la varianza della popolazione è incognita, si usa considerare un valore cautelativo per σ^2, ponendo σ pari a 4 o 6 volte il campo di variazione atteso per la variabile di interesse. Ad esempio, se pensiamo che le telefonate al call center possano durare da un minimo di 0 minuti ad un massimo di 30 minuti, utilizzeremo $\sigma^2 = (4 * 30)^2$ o $\sigma^2 = (6 * 30)^2$. Naturalmente ci si aspetta che la varianza abbia valori più bassi, ma è meglio utilizzare una dimensione campionaria troppo elevata che una troppo bassa.

4 Intervalli di confidenza per proporzioni

Supponiamo di aver a che fare con una variabile statistica dicotomica X che si distribuisce nella popolazione di riferimento secondo la tabella di frequenze relative

x	
0	$1-\theta$
1	θ
	1

dove θ indica la proporzione (incognita) degli individui che posseggono la modalità 1.

Si desidera costruire un intervallo di confidenza per θ al livello $1-\alpha$ sulla base di un campione casuale semplice

$$(x_1 \dots x_n)$$

di dimensione n. Come vedremo, non è qui necessario distinguere casi diversi, poichè la precisione dello stimatore che utilizzeremo per θ dipende comunque dal valore incognito assunto da θ.

La costruzione dell'intervallo si basa sul seguente risultato: la **frequenza relativa campionaria**

$$\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

è una variabile aleatoria che si distribuisce approssimativamente come una normale

$$N(\theta, \frac{\hat{\theta}(1-\hat{\theta})}{n})$$

e tale approssimazione migliora all'aumentare della dimensione campionaria n. La frequenza relativa campionaria $\hat{\theta}$ non è altro che una media campionaria, essendo le osservazioni dicotomiche. Continueremo tuttavia a far riferimento a $\hat{\theta}$ invece che a $\bar{x}$ per tenere ben distinto il caso di stima di medie da quello di stima di proporzioni (per la verità non si tratta di casi distinti, ma queste sono questioni da risolvere in eventuali futuri corsi di statistica successivi a questo).

Se dunque usiamo $\hat{\theta}$ come stimatore di θ, il fatto che la sua distribuzione sia centrata sul valore vero del parametro θ indica che $\hat{\theta}$ è uno stimatore non distorto. Inoltre, il rapporto $\frac{\hat{\theta}(1-\hat{\theta})}{n}$ è una stima della precisione dello stimatore: come sempre, tale precisione è tanto maggiore quanto più elevata è la dimensione campionaria n. C'è tuttavia un'importante differenza da osservare qui rispetto a quanto discusso nel caso della stima di medie. Mentre infatti la precisione dello stimatore di una media non dipende dal valore vero assunto dal parametro di interesse, qui la precisione varia al variare del valore assunto da θ. In particolare, ci si accorge che la funzione $\theta(1-\theta)$ è una funzione concava che vale 0 quando $\theta = 0, 1$ e raggiunge il suo massimo quando $\theta = 0.5$. Se ne deduce che a parità di dimensione campionaria e di livello di copertura otteremo intervalli di confidenza generalmente più stretti quando θ si trova vicino agli estremi 0 e 1, e più larghi quando θ si trova in un intorno di 0.5.

Dal fatto che $\hat{\theta} \sim N(\theta, \hat{\theta}(1-\hat{\theta})/n)$, si deduce che

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} \sim N(0,1).$$

Per ogni valore di probabilità $1-\alpha$, possiamo allora scrivere che

$$P(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} \leq z_{\alpha/2}) = 1-\alpha$$

dove $z_{\alpha/2}$ è al solito il quantile della normale di ordine $1-\alpha/2$.

Un intervallo di confidenza può allora essere costruito sulla base della seguente catena di uguaglianze:

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} \leq z_{\alpha/2}) = 1-\alpha \\
&= P(-z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \hat{\theta} - \theta \leq z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}) \\
&= P(-\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq -\theta \leq -\hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}) \\
&= P(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}})
\end{aligned}$$

In altre parole, è approssimativamente uguale a $1-\alpha$ la probabilità che i due estremi dell'intervallo

$$\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right)$$

contengano il valore "vero" della proporzione θ della popolazione.

5 Ancora sulla determinazione della dimensione campionaria

Il calcolo della dimensione campionaria ottimale può essere compiuto anche quando l'intervallo di confidenza è calcolato per una proporzione incognita θ.

Naturalmente, in questo caso la precisione dello stimatore (e quindi l'ampiezza dell'intervallo) dipende dal valore assunto da θ, che è incognito. È dunque necessario usare come misura cautelativa la quantità

$$\theta(1-\theta) = 0.5^2 = 0.25$$

e procedere sulle linee della sezione dedicata alla dimensione campionaria nel calcolo di intervalli di confidenza per medie.

Più precisamente, per ogni dimensione n l'ampiezza dell'intervallo (ad un prefissato livello $1-\alpha$) raggiungerà al più il valore

$$d = 2z_{\alpha/2}\sqrt{\frac{0.25}{n}}.$$

Se dunque desideriamo calcolare la dimensione minima richiesta per avere un intervallo per θ che non superi l'ampiezza massima d^*, dobbiamo cercare il minimo valore di n tale che

$$2z_{\alpha/2}\sqrt{\frac{0.25}{n}} \leq d^*$$

ovvero tale che

$$4z^2_{\alpha/2}\frac{0.25}{n} \leq (d^*)^2$$

o ancora tale che

$$n \geq 4z^2_{\alpha/2}\frac{0.25}{(d^*)^2} = \left(\frac{z_{\alpha/2}}{d^*}\right)^2$$

La dimensione ottimale n^* è dunque data da

$$n^* = \lceil \left(\frac{z_{\alpha/2}}{(d^*}\right)^2 \rceil$$

Secondo tale formula, se ad esempio programmiamo un'indagine d'opinione per stimare la proporzione degli elettori di un collegio elettorale che voteranno per un certo partito politico e desideriamo un intervallo di confidenza che al livello $1-\alpha = 0.95$ non superi l'ampiezza di 2 punti percentuali ($d^* = 0.02$), avremo bisogno di un minimo di

$$n^* = \lceil \left(\frac{1.96}{0.02}\right)^2 \rceil = 9604$$

elettori da intervistare.

6 Inferenza sulla differenza tra medie

Supponiamo di aver a che fare con due campioni di osservazioni, diciamo $(x_1...x_{n_1})$ e $(y_1...y_{n_2})$, estratti indipendentemente da due popolazioni dove la stessa variabile quantitativa si distribuisce rispettivamente con medie μ_1 e μ_2 e con varianze σ_1^2 e σ_2^2. Indichiamo inoltre, rispettivamente, con $\bar{x}$ e $\bar{y}$ le due medie aritmetiche campionarie. Si desidera costruire un intervallo di confidenza al livello $1-\alpha$ per la differenza tra le medie $\mu_1 - \mu_2$.

Si pensi all'interpretazione di un intervallo di confidenza di questo tipo: se esso contiene lo 0, diremo che **le due medie non sono significativamente diverse tra loro** al livello $1-\alpha$, poichè non possiamo escludere che il valore vero del parametro d'interesse sia pari a $\mu_1 - \mu_2 = 0$.

Per la costruzione dell'intervallo in questione (e sotto l'ipotesi che i due campioni siano stati estratti indipendentemente l'uno dall'altro) possiamo distinguere i seguenti casi:

varianze uguali e note: $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$ in questo caso, la variabile aleatoria

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

si distribuisce come una normale standardizzata e l'intervallo di confidenza desiderato e' dato da:

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

varianze diverse e note: $(\sigma_1^2 \neq \sigma_2^2)$ in questo caso la variabile aleatoria

$$\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

si distribuisce come una normale standardizzata e l'intervallo di confidenza desiderato e' dato da:

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

varianze uguali ma incognite: $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$ in questo caso, una stima della varianza comune σ^2 e' data dalla cosiddetta varianza campionaria *pooled*

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_1 + n_2 - 2}$$

e si ha che la variabile aleatoria

$$\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

si distribuisce come una t di Student con $n_1 + n_2 - 2$ gradi di liberta' e l'intervallo di confidenza desiderato e' dato da:

$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2,\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Si osservi che non e' stato considerato il caso di varianze diverse e incognite: la soluzione di questo problema esula dal programma del corso. Per comprendere l'uso delle formule introdotte, consideriamo il seguente esempio numerico.

Esempio Supponiamo che siano stati estratti due campioni di studenti universitari, iscritti al secondo anno in due università italiane, e di ogni studente è stata registrata la media dei voti conseguiti agli esami. Il primo campione è costituito da $n_1 = 50$ studenti e ha fornito una media campionaria pari a $\bar{x} = 23.5$, mentre il secondo è costituito da $n_2 = 100$ studenti ed ha fornito una media campionaria pari a $\bar{y} = 25.2$. Si desidera costruire un intervallo di confidenza al livello $1 - \alpha = 0.95$ per la differenza $\mu_1 - \mu_2$ tra i voti medi riportati dagli studenti nelle due università. Le tre procedure più semplici che possiamo seguire fanno riferimento alle formule viste in precedenza.

Varianze note e uguali L'ipotesi più semplice (ma anche la più rischiosa) consiste nell'assumere che il voto medio si distribuisca nelle due universita' con la stessa varianza che assumiamo nota: tale varianza potrebbe essere ad esempio quella pubblicata dall'ufficio statistico del MIUR con riferimento al voto medio degli studenti iscritti al secondo anno in tutti gli atenei italiani. Supponiamo che tale varianza σ^2 sia uguale a 16. Formalmente, stiamo assumendo che il voto medio degli studenti della prima università sia una variabile aleatoria che si distribuisce seguendo la normale $N(\mu_1, \sigma^2)$, mentre il voto relativo agli iscritti nella seconda università segua la normale $Y \sim N(\mu_2, \sigma_2)$. L'intervallo di confidenza cercato e' dato allora da

$$\bar{x} - \bar{y} \pm z_{\alpha/2} \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 23.5 - 25.2 \pm 1.96 \cdot 4 \cdot \sqrt{\frac{1}{50} + \frac{1}{100}} = -1.7 \pm 1.36$$

ovvero $(-3.06, -0.34)$. Sulla base di questo risultato possiamo affermare (con un livello di fiducia del 95%) che gli studenti della prima

università hanno conseguito in media un voto medio al secondo anno inferiore a quello conseguito dagli iscritti alla seconda università. Si osservi che, sulla base di tale intervallo che non comprende lo zero, si può affermare che i voti medi nelle due università sono significativamente differenti, al livello $1 - \alpha$.

Varianze note e diverse Se invece gli uffici statistici delle due università hanno pubblicato recentemente (rispetto alla nostra analisi) delle tabelle da cui si evince che le due popolazioni hanno varianze diverse, possiamo decidere di considerare queste come note. Supponendo di avere $\sigma_1^2 = 16$ e $\sigma_2^2 = 4$, l'intervallo di confidenza desiderato sara' dato da

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 23.5 - 25.2 \pm 1.96 \cdot \sqrt{\frac{16}{n_1} + \frac{4}{n_2}} = -1.7 \pm 1.18$$

ovvero $(-2.88, -0.52)$.

varianze uguali ma incognite Se non reputiamo attendibili le statistiche del MIUR nè quelle dei due atenei, non ci rimane altra scelta che assumere incognite le due varianze. Se ci sono informazioni sufficienti per assumere che tuttavia i voti hanno la stessa dispersione nelle due università, possiamo usare la formula contenente la varianza *pooled* per l'intervallo desiderato, se conosciamo le deviazioni standard dei due campioni. Supponendo che le seguenti siano le devazioni standard dei due campioni:

$$\sqrt{\frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2} = \sqrt{50/50} = 1$$

$$\sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \bar{y})^2}\sqrt{400/100} = 2$$

allora la varianza *pooled* e' data da

$$\hat{\sigma}^2 = \frac{50 + 400}{50 + 100 - 2} = 3.04$$

e possiamo calcolare gli estremi dell'intervallo desiderato come segue:

$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = -1.7 \pm 1.96 \cdot 1.74 \cdot \sqrt{\frac{1}{50} + \frac{1}{100}} = -1.7 \pm 0.59$$

dato che, essendo $n_1 + n_2 - 2 > 100$, si ha $t_{n_1+n_2-2,\alpha/2} \approx z_{\alpha/2}$

7 Differenza tra due proporzioni

Supponiamo di aver a che fare con due campioni indipendenti, diciamo $(x_1...x_{n_1})$ e $(y_1...y_{n_2})$, estratti rispettivamente da due popolazioni in cui una stessa variabile dicotomica si distribuisce secondo le due tabelle:

pop.ne 1

x	
0	$1-\theta_1$
1	θ_1
	1

pop.ne 2

y	
0	$1-\theta_2$
1	θ_2
	1

Indichiamo inoltre, rispettivamente, con $\hat{\theta}_1 = \bar{x}$ e $\hat{\theta}_2 = \bar{y}$ le due frequenze relative campionarie. Si desidera costruire un intervallo di confidenza al livello $1-\alpha$ per la differenza tra le proporzioni $\theta_1 - \theta_2$. L'importanza di un intervallo del genere è chiara: se l'intervallo contiene lo 0, diremo che le due proporzioni non sono significativamente diverse, al livello $1-\alpha$.

Il risultato che usiamo per costruire il nostro intervallo è il seguente. Sia

$$\hat{\theta} = \frac{n_1\hat{\theta}_1 + n_2\hat{\theta}_2}{n_1 + n_2}$$

allora

$$\frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1).$$

Si tratta al solito di un risultato approssimato, ma la qualità di tale risultato è sempre migliore man mano che crescono le dimensioni campionarie n_1 e n_2.

Da tale risultato, si deduce che un intervallo di confidenza per la differenza tra due proporzioni al livello $1-\alpha$ è dato dagli estremi

$$\hat{\theta}_1 - \hat{\theta}_2 - z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\hat{\theta}_1 - \hat{\theta}_2 + z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Supponiamo ad esempio di aver effettuato due sondaggi di opinione in date successive chiedendo agli intervistati la preferenza per un determinato partito politico. In particolare, supponiamo di aver intervistato 100 elettori durante il primo sondaggio e 200 elettori durante il secondo sondaggio, ottenendo una percentuale di elettori favorevoli del 40% nel primo e del 42% nel

secondo sondaggio. Ci chiediamo se tale incremento di preferenze sia stato significativo al livello $1 - \alpha = 0.95$. Si ottiene:

$$\hat{\theta} = \frac{n_1\hat{\theta}_1 + n_2\hat{\theta}_2}{n_1 + n_2} = \frac{100 \cdot 0.4 + 200 \cdot 0.42}{100 + 200} = 0.413$$

$$\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.413 \cdot 0.587\left(\frac{1}{100} + \frac{1}{200}\right)} = 0.060$$

Se ne deduce che l'intervallo di confidenza desiderato è dato da

$$(0.40 - 0.42 - 1.96 \cdot 0.060, 0.40 - 0.42 + 1.96 \cdot 0.060) = (-0.138, 0.098)$$

ovvero l'aumento osservato nei campioni non può essere considerato significativo, poichè l'intervallo contiene lo 0.