STIME E LORO AFFIDABILITA’

- L’idea chiave su cui si basa l’analisi statistica è che si possono eseguire osservazioni su un campione di soggetti e che da questo si possono compiere inferenze sulla popolazione rappresentata da tutti i soggetti con caratteristiche analoghe a quelle del campione

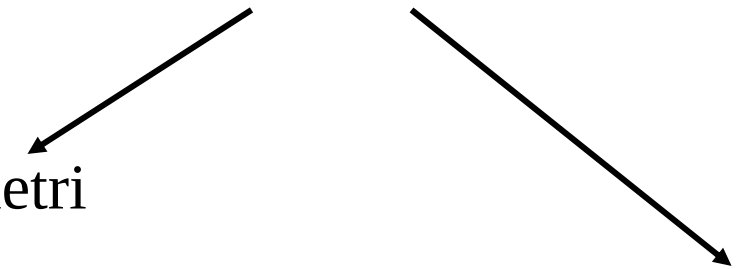
- Anche se ben pianificato uno studio può dare solo una idea della risposta cercata, a causa essenzialmente della *variabilità casuale* del campione stesso strettamente collegata, tra l’altro, al *numero di soggetti* inclusi in uno studio

- Le quantità statistiche ottenute (medie, proporzioni, odds, coefficienti di regressione, etc.) sono ***stime*** imprecise dei ***veri*** valori nella popolazione generale

STIMA

Una misura descrittiva calcolata dai dati di una **popolazione** è detta ***parametro*.**
Una misura descrittiva calcolata dai dati di un **campione** è detta ***stima del parametro***.

L'insieme dei metodi che ci consentono di estendere i risultati ottenuti dal campione a tutta la popolazione oggetto dello studio costituiscono la inferenza statistica.



Stima dei parametri

Verifica delle ipotesi

La stima è il calcolo, dai dati di un campione, di una qualche statistica, ed è una approssimazione del corrispondente parametro della popolazione da cui il campione è stato estratto.

- ☞ Stima puntuale: si calcola un singolo valore numerico per stimare il corrispondente parametro. Es. una media, una proporzione, una deviazione standard.

- ☞ Stima di intervallo: si calcola un intervallo di valori che, con un certo grado di probabilità, conterrà il parametro da stimare.

INTERVALLI DI CONFIDENZA

Le stime di intervallo forniscono informazioni sia sul valore numerico del parametro incognito che sul grado di attendibilità della stima.

La procedura di calcolo degli intervalli, detti di confidenza, si basa sulla determinazione di due limiti entro i quali, con una probabilità $1-\alpha$, è contenuto il parametro, a partire dalle informazioni campionarie.

$$\mathbf{1-\alpha = P(L_1 \leq \theta \leq L_2) \ \ con \ \ 0 \leq \alpha \leq 1}$$

L_1 e L_2 dipendenti dalla dimensione del campione

$1-\alpha$ grado di attendibilità della stima ed è detto livello di confidenza

Supponiamo di costruire un intervallo di confidenza per la media di una variabile casuale che segue una distribuzione di Gauss $N(m,s^2)$, con varianza della popolazione nota.

$1 - a = P(L1 \leq m \leq L2)$

Devo individuare i valori di L1 e L2 che mi garantiscano che, estraendo dalla popolazione altri campioni di uguale dimensione n, con probabilità pari al 95% la media campionaria sarà contenuta nell'intervallo.

Dopo aver eseguito la stima puntuale $\bar{x}$ della media della variabile X posso considerare la "nuova" variabile casuale Z, e applicando il teorema del limite centrale posso scrivere:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$Z \longrightarrow N(0,1)$$

$$1-\alpha = P\left(L_1 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le L_2 \right)$$

$$1-\alpha = P\left(-L \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le +L \right)$$

$$1-\alpha = P\left(-L\,\sigma/\sqrt{n} \le \bar{x}-\mu \le L\,\sigma/\sqrt{n} \right)$$

$$1-\alpha = P\left(-L\,\sigma/\sqrt{n} - \bar{x} \le -\mu \le -\bar{x} + L\,\sigma/\sqrt{n} \right)$$

$$1-\alpha = P\left(\bar{x} - L\,\sigma/\sqrt{n} \le \mu \le \bar{x} + L\,\sigma/\sqrt{n} \right)$$

$$1-\alpha = P\left(\bar{x} - z\,\sigma/\sqrt{n} \le \mu \le \bar{x} + z\,\sigma/\sqrt{n} \right)$$

Dato che la variabile casuale Z per il Teorema Centrale del limite segue una distribuzione di Gauss standard, che è simmetrica,

abbiamo potuto trasformare L_1 e L_2 in $-L$ e $+L$,

infine abbiamo potuto sostituirli con z_{tab}, che si ricavano dalle tavole della distribuzione di Gauss standard.

- per $1-\alpha=0,95$ (cioè $\alpha = 0,05$) $z_{tab} = \pm \underline{1,96.}$
- per $1-\alpha=0,99$ (cioè $\alpha = 0,01$) $z_{tab} = \pm \underline{2,58.}$

Intervallo di confidenza per la media di una variabile casuale con distribuzione di Gauss N(m,s^2), e varianza incognita.

Nella realtà quotidiana anche la varianza della popolazione è incognita e si stima (stima puntuale) con la varianza dei dati campionari.

La varianza ha una sua distribuzione campionaria che prende il nome di distribuzione χ^2 e dipende dai gradi di libertà (denominatore della varianza).

Per costruire l'intervallo di confidenza seguiremo lo stesso ragionamento ma il rapporto z diventa t :

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{S^2}{n}}}$$

che ha distribuzione t-student che dipende dai gradi di libertà (della varianza).

La distribuzione t-student:

- ha la stessa forma della distribuzione di Gauss con picco meno alto e code più pesanti;
- è simmetrica intorno alla media;
- ha media zero;
- ha varianza >1, che si avvicina all'unità per $n \to \infty$
- esiste una famiglia di distribuzioni t distinte dai gradi di libertà;
- $t_n \to N(0,1)$ per campioni grandi (gradi di libertà$\to \infty$).

Anche per la distribuzione t-student esistono delle tavole per determinare i valori di area sotto la curva (probabilità) per i corrispondenti punti sull'asse delle ascisse e in relazione ai diversi di gradi di libertà

$$1-\alpha = P\left(c_1 \le \frac{\bar{x}-\mu}{S/\sqrt{n}} \le c_2 \right)$$

$$1-\alpha = P\left(-c \le \frac{\bar{x}-\mu}{S/\sqrt{n}} \le +c \right)$$

$$1-\alpha = P\left(-c\, S/\sqrt{n} \le \bar{x}-\mu \le c\, S/\sqrt{n} \right)$$

$$1-\alpha = P\left(\bar{x}-c\, S/\sqrt{n} \le \mu \le \bar{x}+c\, S/\sqrt{n} \right)$$

$$1-\alpha = P\left(\bar{x}-t\, S/\sqrt{n} \le \mu \le \bar{x}+t\, S/\sqrt{n} \right)$$

I limiti c saranno ricavati dalle tavole t-student in corrispondenza dei gradi di libertà e del livello di confidenza.

L’intervallo di confidenza per la media di una variabile con distribuzione di Gauss con media incognita e varianza nota

$$\bar{x} - z\,\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z\,\sigma/\sqrt{n}$$

per 1-α=0,95 (cioè α = 0,05) z è pari a 1,96.
per 1-α=0,99 (cioè α = 0,01) z è pari a 2,58.

L’intervallo di confidenza per la media di una variabile con distribuzione di Gauss con media e varianza incognita

$$\bar{x} - t\,S/\sqrt{n} \leq \mu \leq \bar{x} + t\,S/\sqrt{n}$$

t si ottiene dalla distribuzione t-student con n-1 gradi di libertà

Interpretazione probabilistica degli intervalli di confidenza:

estraendo tutti i possibili campioni da una popolazione distribuita normalmente la media μ della popolazione cadrà $(1-\alpha)\%$ volte nell'intervallo calcolato.

Interpretazione pratica degli intervalli di confidenza:

se effettuiamo il campionamento da una popolazione con distribuzione normale abbiamo una probabilità del $(1-\alpha)\%$ che l'intervallo calcolato contenga la media μ della popolazione.

Lo scopo principale degli intervalli di confidenza è quello di indicare la **Im**precisione delle stime campionarie come rappresentazione dei valori della popolazione

L'imprecisione della stima campionaria è indicata dall'ampiezza degli intervalli:

Più ampi sono gli intervalli ⇨ Minore è la precisione

L'ampiezza dipende essenzialmente da tre fattori:

- dal numero di soggetti studiati
(campioni poco numerosi, conclusioni inattendibili)
- dalla variabilità dei soggetti in studio
(minore variabilità, stima più precisa)
- dal livello di confidenza
(maggiore è il livello di confidenza, tanto più ampi sono gli intervalli)

STIME DI INTERVALLO

In conclusione è possibile indicare una formula generica per la daterminazione di un intervallo di confidenza :

Stima ± (fattore di correzione * errore della stima)

Media
Percentuale
…

Valore che serve a determinare i due limiti (inferiore e superiore) tenendo in considerazione il grado di attendibilità della stima

ESERCIZIO 1

Sono state misurate le pulsazioni cardiache in 10 soggetti ansiosi

90 86 88 86 88 87 87 90 88 89

Determinare l'intervallo di confidenza per la media della popolazione, ipotizzando che:

- la varianza della popolazione sia pari a 3
- la varianza della popolazione non sia nota

$$1-\alpha = P\left(\bar{x} - z\,\sigma/\sqrt{n} \le \mu \le \bar{x} + z\,\sigma/\sqrt{n}\right)$$

$\bar{x} = 87.9$

$\sigma^2 = 3 \rightarrow \sigma = 1.73$

$1-\alpha = 0.95 \rightarrow z_\alpha = 1.96$

$$0.95 = P\left(87.9 - 1.96\,1.73/\sqrt{10} \le \mu \le 87.9 + 1.96\,1.73/\sqrt{10}\right)$$

$$0.95 = P\left(87.9 - 1.1 \le \mu \le 87.9 + 1.1\right)$$

$$0.95 = P\left(86.8 \le \mu \le 89\right)$$

$$1-\alpha = P\left(\bar{x} - t\, s/\sqrt{n} \le \mu \le \bar{x} + t\, s/\sqrt{n}\right)$$

$S^2 = 2.1 \rightarrow S = 1.45$

$g.l. = 9 \rightarrow t = 2.262$

$$0.95 = P\left(87.9 - 2.262\, 1.45/\sqrt{10} \le \mu \le 87.9 + 2.262\, 1.45/\sqrt{10}\right)$$

$$0.95 = P\left(87.9 - 1 \le \mu \le 87.9 + 1\right)$$

$$0.95 = P\left(86.9 \le \mu \le 88.9\right)$$

L’intervallo di confidenza per la differenza di due medie

Dai due campioni si determinano le medie della variabile in studio e avremo:

μ_1-μ_2 =differenza delle medie delle popolazioni

$\bar{x}_1 - \bar{x}_2$ =stima della differenza delle medie delle popolazioni

$\sigma_1^2/n_1+\sigma_2^2/n_2$=varianza della differenza delle medie campionarie.

L’intervallo sarà:

$$1-\alpha=$$

$$=P\left[\left(\bar{x}_1-\bar{x}_2\right)-z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\le\mu_1-\mu_2\le\left(\bar{x}_1-\bar{x}_2\right)+z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

Abbiamo usato z perché eravamo a conoscenza della varianza della popolazione.
Nel caso non si conosca la varianza della popolazione si deve stimare dai campioni la varianza "comune" (pooled):

$$S_p^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Gradi di libertà

L'intervallo di confidenza diventa:

$$1 - \alpha =$$

$$= P\left[\left(\bar{x}_1 - \bar{x}_2\right) - t_{1-\alpha/2, g.l.} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{x}_1 - \bar{x}_2\right) + t_{1-\alpha/2, g.l.} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right]$$

ESERCIZIO 3

Si dispone dei valori relativi ad i battiti cardiaci in un campione di soggetti ansiosi ed in un campione di atleti:

ansiosi	90	86	88	86	88	87	87	90	88	89
atleti	70	65	66	66	68	66	66	64	62	65

determinare l'intervallo di confidenza per la differenza delle medie delle popolazioni ipotizzando che:

1. le varianze siano note e rispettivamente pari a 2.5 e 4
2. le varianze non siano note

$$1-\alpha = P\left(\left(\bar{x}_1 - \bar{x}_2\right) - z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{x}_1 - \bar{x}_2\right) + z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \right)$$

$\bar{x}_1 = 87.9$ $\bar{x}_2 = 65.8$

$\sigma_1^2 = 2.5$ $\sigma_2^2 = 4$

$$0.95 = P\left(\left(87.9-65.8\right) - 1.96\sqrt{\frac{2.5}{10}+\frac{4}{10}} \le \mu_1 - \mu_2 \le \left(87.9-65.8\right) + 1.96\sqrt{\frac{2.5}{10}+\frac{4}{10}} \right)$$

$$0.95 = P\left(22.1-1.58 \le \mu_1 - \mu_2 \le 22.1+1.58\right)$$

$$0.95 = P\left(20.52 \le \mu_1 - \mu_2 \le 23.68\right)$$

QUESITO 2

$$1-\alpha = P\left(\left(\bar{x}_1 - \bar{x}_2\right) - t\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{x}_1 - \bar{x}_2\right) + t\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right)$$

$\bar{x}_1 = 87.9$ $\quad \bar{x}_2 = 65.8$ $\quad t_{18,0.05} = 2.101$

$S_1^2 = 2.1$ $\quad S_2^2 = 4.62$ $\quad S_p^2 = \dfrac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2} = 3.36$

$$0.95 = P\left(\left(87.9 - 65.8\right) - 2.101\sqrt{3.36 \cdot 2/10} \le \mu_1 - \mu_2 \le \left(87.9 - 65.8\right) + 2.101\sqrt{3.36 \cdot 2/10} \right)$$

$$0.95 = P\left(22.1 - 1.72 \le \mu_1 - \mu_2 \le 22.1 + 1.72\right)$$

$$0.95 = P\left(20.38 \le \mu_1 - \mu_2 \le 23.82\right)$$

INTERVALLI DI CONFIDENZA PER UNA PROPORZIONE

In medicina spesso è importante quantificare un fenomeno per mezzo delle percentuali.

Es.: percentuale di soggetti affetti da una certa malattia,
percentuale di soggetti sottoposti ad un trattamento…

Il parametro da stimare diventa p la proporzione nella popolazione e per costruire lintervallo di confidenza procederemo nel seguente modo:

$$1-\alpha = P\left(L_1 \leq p \leq L_2\right)$$

La variabile X che conta il numero dei successi sul numero di prove ha una distribuzione Binomiale con $\mu = np$ e $\sigma^2 = npq$

- **può essere approssimata ad una Gauss:**

se n, la numerosità campionaria, è grande ($n \to \infty$) e $p \cong 0.5$

- **quindi standardizzata**

$$z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}};$$

Segue N(0,1)

$$1 - \alpha = P\left(L_1 \le \frac{X - np}{\sqrt{np(1-p)}} \le L_2 \right);$$

$$1 - \alpha = P\left(-c \le \frac{X - np}{\sqrt{np(1-p)}} \le +c \right);$$

Limiti uguali ed opposti

Nel nostro caso il parametro p si trova anche a denominatore e inoltre sotto radice.

Bisogna quindi "razionalizzare", elevare al quadrato nel nostro caso, per togliere il segno di radice.

$$1-\alpha = P\left[\left(\frac{X-np}{\sqrt{np(1-p)}}\right)^2 \leq c^2\right];$$

$$1-\alpha = P\left[\frac{X^2+n^2p^2-2npX}{np(1-p)} \leq c^2\right];$$

Per n sufficientemente grande l'intervallo di confidenza può essere determinato secondo la seguente formula:

$$1-\alpha = P\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Analogamente a quanto fatto per gli intervalli di confidenza per le medie, è possibile costruire l'intervallo di confidenza per la differenza tra due proporzioni.

La stima di tale intervallo viene dalla seguente formula:

$$1-\alpha = P\left[\begin{array}{l} (\hat{p}_1-\hat{p}_2)-z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n2}} \le p_1-p_2 \\ p_1-p_2 \le (\hat{p}_1-\hat{p}_2)+z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n2}} \end{array}\right]$$

QUESITO 1.

Dieta A $n_1 = 300$ $X_1 = 105$ $p_1 = 105/300 = 0.35$

Dieta B $n_2 = 200$ $X_2 = 60$ $p_2 = 60/200 = 0.30$

$$1-\alpha = \boldsymbol{P}\left[\hat{\boldsymbol{p}} - \boldsymbol{z}_{1-\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{\boldsymbol{n}}} \le \boldsymbol{p} \le \hat{\boldsymbol{p}} + \boldsymbol{z}_{1-\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}\right]$$

$$1-\alpha = \boldsymbol{P}\left[0.35 - 1.96\sqrt{\frac{0.35\times 0.65}{300}} \le \hat{\boldsymbol{p}}_1 \le 0.35 + 1.96\sqrt{\frac{0.35\times 0.65}{300}}\right] =$$

$$0.296 \le \hat{\boldsymbol{p}}_1 \le 0.404$$

$$1-\alpha = \boldsymbol{P}\left[0.30 - 1.96\sqrt{\frac{0.30\times 0.70}{200}} \le \hat{\boldsymbol{p}}_2 \le 0.30 + 1.96\sqrt{\frac{0.30\times 0.70}{200}}\right] =$$

$$0.236 \le \hat{\boldsymbol{p}}_2 \le 0.363$$

QUESITO 2.

Dieta A $n_1 = 300$ $X_1 = 105$ $p_1 = 105/300 = 0.35$

Dieta B $n_2 = 200$ $X_2 = 60$ $p_2 = 60/200 = 0.30$

$$1-\alpha = $$

$$= P\left[\begin{array}{l} (\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n2}} \le p_1 - p_2 \\ p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n2}} \end{array}\right]$$

$$(0.35 - 0.30) - 1.96\sqrt{\frac{0.35 \times 0.65}{300} + \frac{0.30 \times 0.70}{200}} \le \boldsymbol{p}_1 - \boldsymbol{p}_2 \le$$

$$\le (0.35 - 0.30) + 1.96\sqrt{\frac{0.35 \times 0.65}{300} + \frac{0.30 \times 0.70}{200}}$$

$$-0.03 \le \boldsymbol{p}_1 - \boldsymbol{p}_2 \le 0.13$$