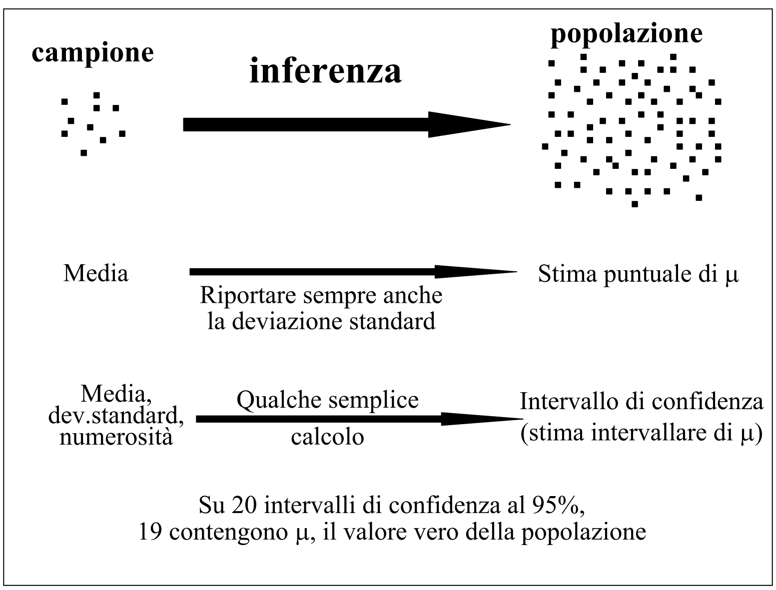
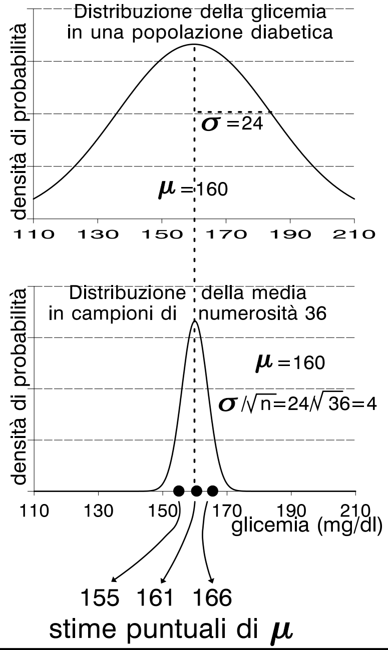
Intervallo di confidenza

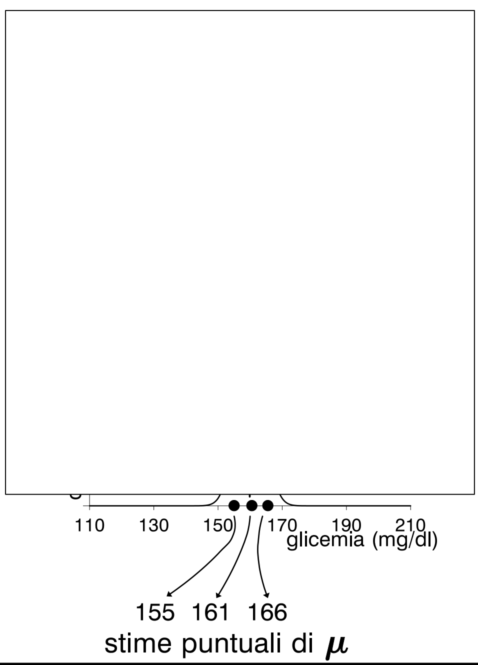
Prof. Giuseppe Verlato, Prof. Roberto de Marco
Sezione di Epidemiologia e Statistica Medica,
Università di Verona



Teoria del campionamento



Inferenza statistica



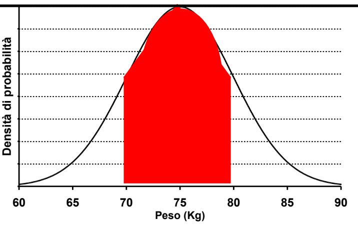
Dal momento che il campione viene estratto casualmente dalla popolazione, le conclusioni tratte da un campione possono essere errate.

L'inferenza statistica viene fatta "con umiltà":

1) si cerca di stimare la probabilità di commettere errori

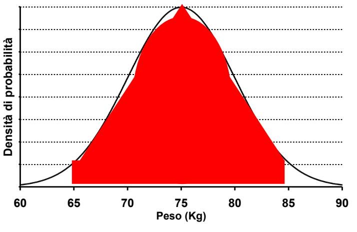
2) si cerca di limitare la probabilità di commettere errori

media ± 1 dev.standard =
75 ± 5 = 70-80 Kg



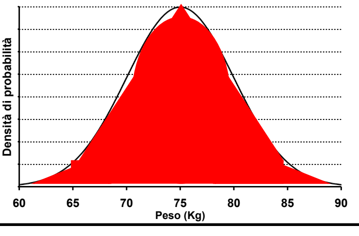
68,26%

media ± 2 dev.standard =
75 ± 10 = 65-85 Kg

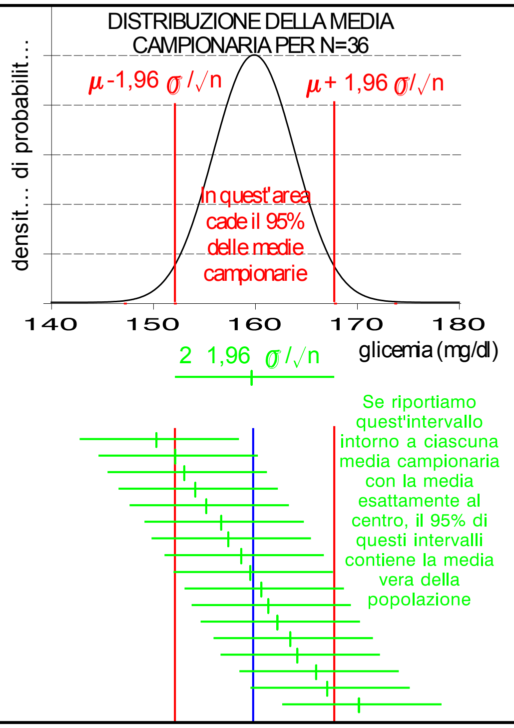


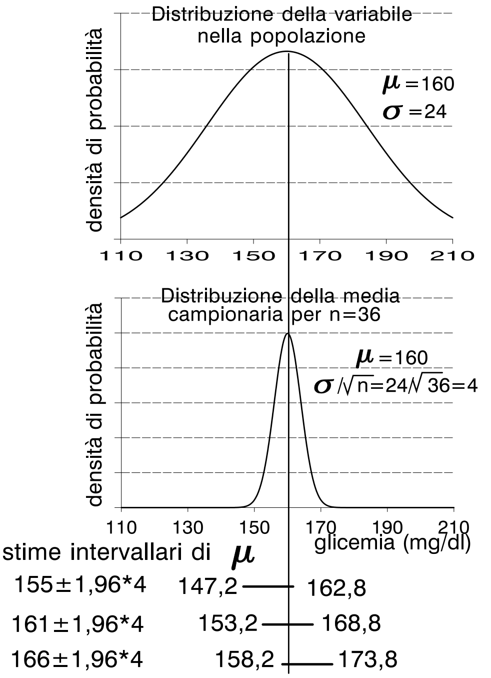
95,44%

media ± 3 dev.standard =
75 ± 15 = 60-90 Kg



99,74%





$$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

La **stima puntuale** fornisce un singolo valore. Tuttavia:

1) questo valore non coincide quasi mai con il valore vero (parametro) della popolazione;

2) campioni diversi forniscono stime puntuali diverse.

La **stima intervallare** fornisce un intervallo, che ha una predeterminata probabilità di contenere il valore vero della popolazione. Pertanto:

1) quest'intervallo ha una determinata probabilità (in genere, il 95%) di contenere il valore vero (parametro) della popolazione;

2) gli intervalli ottenuti da campioni diversi in genere si sovrappongono.

INTERVALLO di CONFIDENZA: DEFINIZIONE

Per intervallo di confidenza di un parametro Θ della popolazione, intendiamo un intervallo delimitato da due limiti L_{inf} (limite inferiore) ed L_{sup} (limite superiore) che abbia una definita probabilità $(1-\alpha)$ di contenere il vero parametro della popolazione:

$$p(L_{inf} < \Theta < L_{sup}) = 1-\alpha$$

dove:

$1-\alpha$ = grado di confidenza

α = probabilità di errore

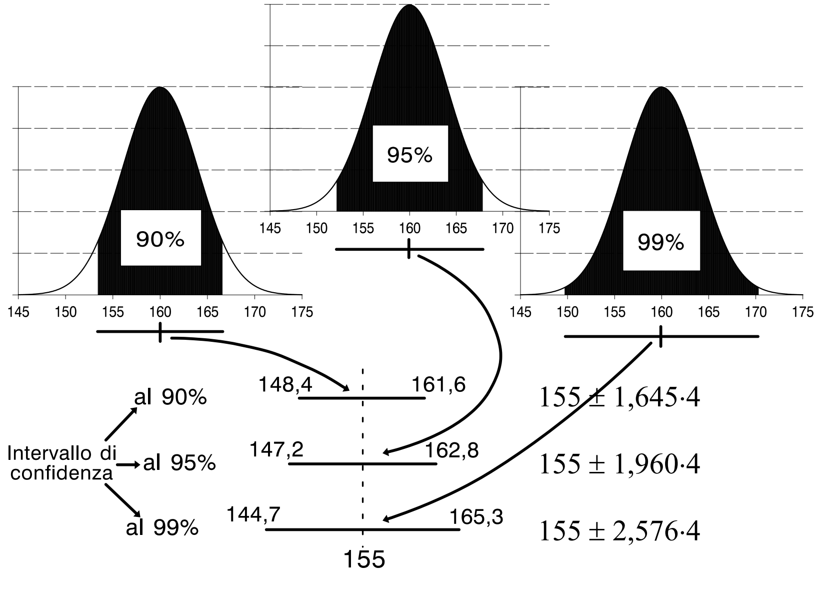
L'intervallo di confidenza diminuisce se:

1) diminuisce il **livello di confidenza** ($1-\alpha$)
(dal 99% al 95% al 90%)
(z da 2,576 a 1,960 a 1,645)

2) diminuisce la **variabilità nella popolazione**
(da σ=48 a σ=24 a σ=12)

3) aumenta la **numerosità** del campione
(da n=4 a n=36 a n=100)

$$\bar{x} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n}$$



Esempio: Calcolo dell'intervallo di confidenza della media di una popolazione

Problema: Qual è l'intervallo di confidenza al 95% della media del peso di una popolazione, se la media di un campione di 16 soggetti è pari a 75 Kg? Nella popolazione il peso è distribuito normalmente con deviazione standard pari a 12 Kg.

Dati: $x = 75$ Kg $\sigma = 12$ Kg $n = 16$ $1-\alpha = 95\%$ $z_{\alpha/2} = 1{,}96$

Formula da utilizzare: $I.C._{95\%} = x \pm z_{\alpha/2} \cdot \sigma/\sqrt{n} = x \pm z_{\alpha/2} \cdot E.S.$

I passo: calcolo l'errore standard
$E.S. = \sigma/\sqrt{n} = 12/\sqrt{16} = 12/4 = 3$ Kg

II passo: calcolo l'intervallo di confidenza

$$I.C._{95\%} = x \pm z_{\alpha/2} \cdot E.S. = 75 \pm 1{,}96 \cdot 3 = \begin{cases} 80{,}88 \text{ Kg} \\ 69{,}12 \text{ Kg} \end{cases}$$

L'intervallo che va da 69,12 Kg (limite inferiore) a 80,88 Kg (limite superiore) ha 95 probabilità su 100 di contenere la media vera della popolazione.

E se non conosco σ, la deviazione standard della popolazione?

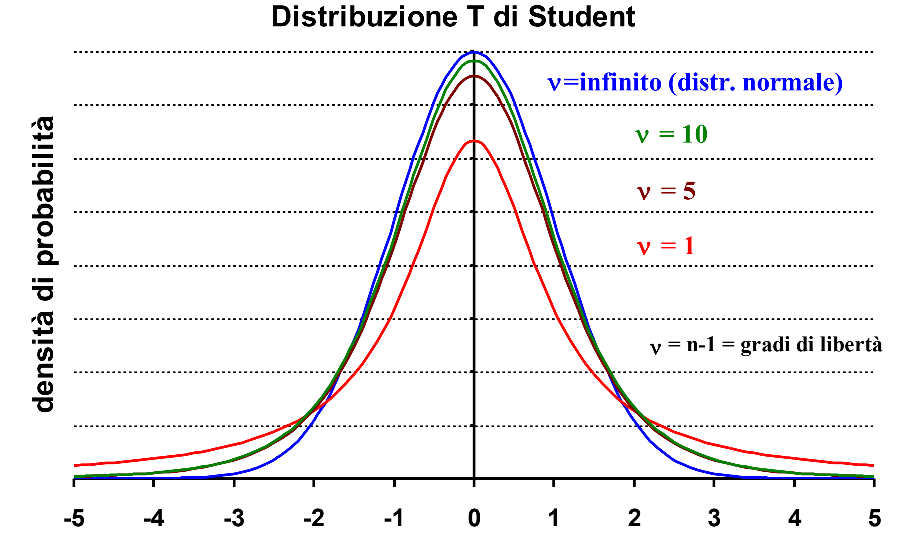
Posso usare s (dev. standard del campione) come stima di σ

Se la numerosità campionaria è sufficientemente grande (n≥60), s è una stima precisa di σ.

$$I.C. = \bar{x} \pm Z_{\alpha/2} * s / \sqrt{n}$$

Se la numerosità campionaria è piccola (n<60), stimare σ tramite s introduce un'ulteriore fonte di variabilità campionaria

Al posto della distribuzione z, devo utilizzare un'altra distribuzione di probabilità, la distribuzione t, caratterizzata da una maggiore dispersione.



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

Riassumendo:

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

σ nota $\Rightarrow$ $\bar{x} \pm Z_{\alpha/2} * \sigma / \sqrt{n}$

σ ignota $\Rightarrow$ $\bar{x} \pm t_{\alpha/2,\, \nu} * s / \sqrt{n}$

Prima della diffusione dei computer si cercava di utilizzare l'approssimazione normale ogni qualvolta possibile. Adesso non è più necessario, per cui la formula seguente è caduta in disuso:

σ ignota $n \geq 60$ $\Rightarrow$ $\bar{x} \pm Z_{\alpha/2} * s / \sqrt{n}$

Esempio: Calcolo dell'intervallo di confidenza della media di una popolazione

Problema: Qual è l'intervallo di confidenza al 95% della media del peso di una popolazione, se la media di un campione di 16 soggetti è pari a 75 Kg e la deviazione standard è pari a 12 Kg?

Dati: $x = 75$ Kg $s = 12$ Kg $n = 16$ $1-\alpha = 95\%$ $t_{15, \alpha/2} = 2{,}131$

Formula da utilizzare: $I.C._{95\%} = x \pm t_{\alpha/2} \cdot s/\sqrt{n} = x \pm t_{\alpha/2} \cdot E.S.$

I passo: calcolo l'errore standard
$E.S. = s/\sqrt{n} = 12/\sqrt{16} = 12/4 = 3$ Kg

II passo: calcolo l'intervallo di confidenza

$I.C._{95\%} = x \pm t_{15, \alpha/2} \cdot E.S. = 75 \pm 2{,}131 \cdot 3 =$ 81,39 Kg / 68,61 Kg

L'intervallo che va da 68,61 Kg (limite inferiore) a 81,39 Kg (limite superiore) ha 95 probabilità su 100 di contenere la media vera della popolazione.

Problema 3: Calcolo dell'intervallo di confidenza di una proporzione di una popolazione

Problema: Qual è l'intervallo di confidenza al 95% della probabilità (prevalenza) di asma in una popolazione, se la frequenza relativa di asma in un campione di 225 soggetti è pari a 0,05 (5%)?

Dati: $p = 0{,}05$ $n = 225$ $1-\alpha = 95\%$ $z_{\alpha/2} = 1{,}96$ I.C. = ?

Formula da utilizzare: $I.C._{95\%} = p \pm z_{\alpha/2} \cdot \sqrt{p(1-p)/n} = p \pm z_{\alpha/2} \cdot E.S.$

I passo: calcolo l'errore standard
$E.S. = \sqrt{p(1-p)/n} = \sqrt{0{,}05(1-0{,}05)/225} = \sqrt{0{,}05 * 0{,}95/225} = 0{,}01453 = 1{,}45\ \%$

II passo: calcolo l'intervallo di confidenza

$I.C._{95\%} = p \pm z_{\alpha/2} \cdot E.S. =$
Limite superiore = 5 + 1,96*1,45 = 7,85%
Limite inferiore = 5 - 1,96*1,45 = 2,15%

L'intervallo che va dal 2,15% (limite inferiore) al 7,85% (limite superiore) ha 95 probabilità su 100 di contenere la prevalenza vera di asma in quella determinata popolazione.

INTERVALLO DI CONFIDENZA DI LIVELLO ($1-\alpha$)

PER UNA PROPORZIONE

Se $np \geq 10$ e $n(1-p) \geq 10 \Rightarrow \hat{\pi} = p \sim N(\pi, \pi(1-\pi)/n)$

utilizzo $p(1-p)/n$ per stimare $\pi(1-\pi)/n$

$$p - Z_{\alpha/2} * \sqrt{p(1-p)/n} < \pi < p + Z_{\alpha/2} * \sqrt{p(1-p)/n}$$

per $1-\alpha = 95\%$

$$p - 1{,}96 * \sqrt{p(1-p)/n} < \pi < p + 1{,}96 * \sqrt{p(1-p)/n}$$