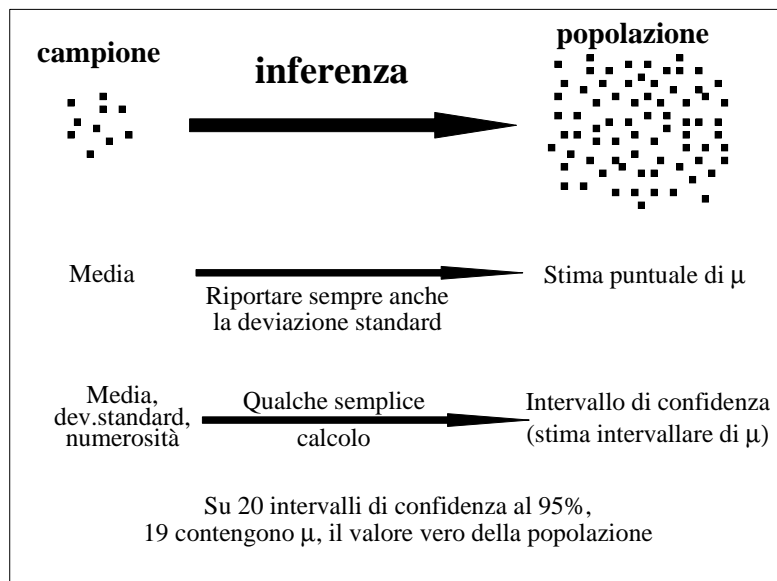
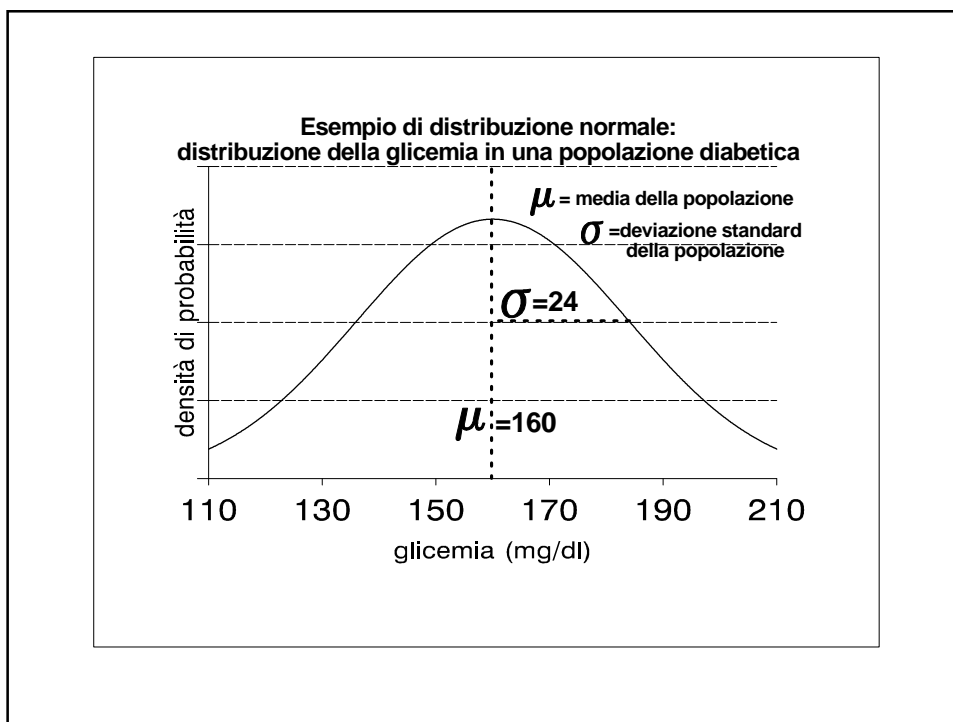


Intervallo di confidenza

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

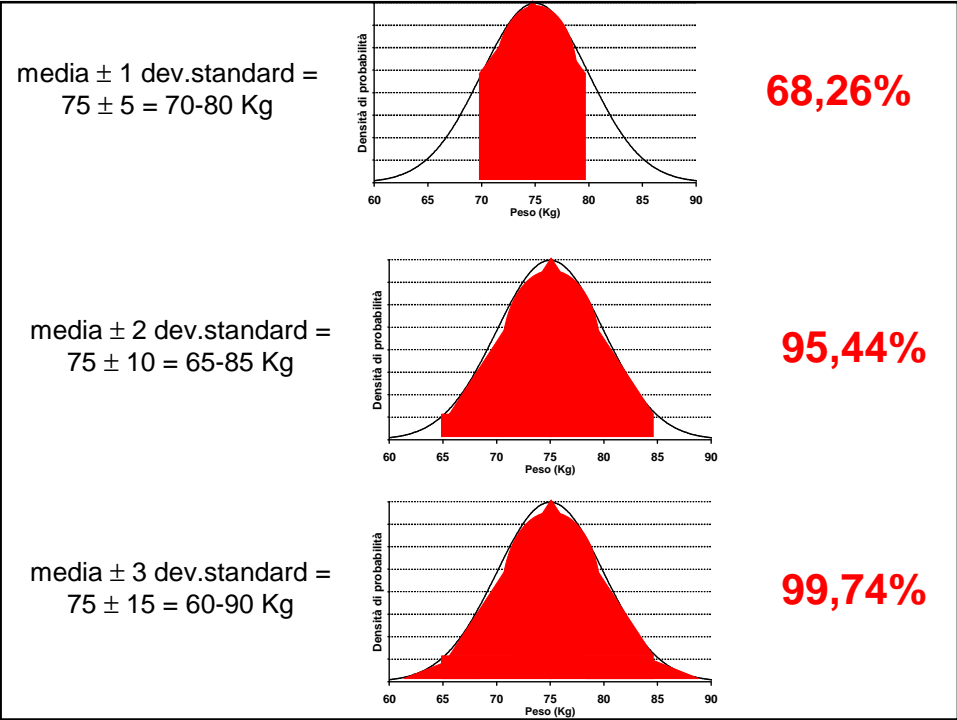
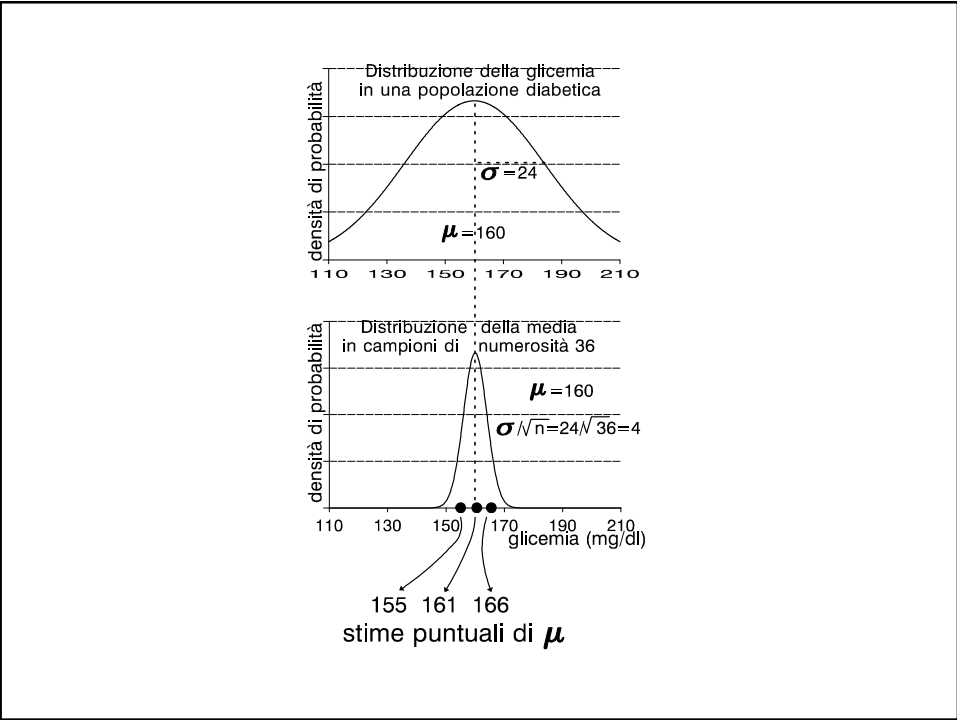


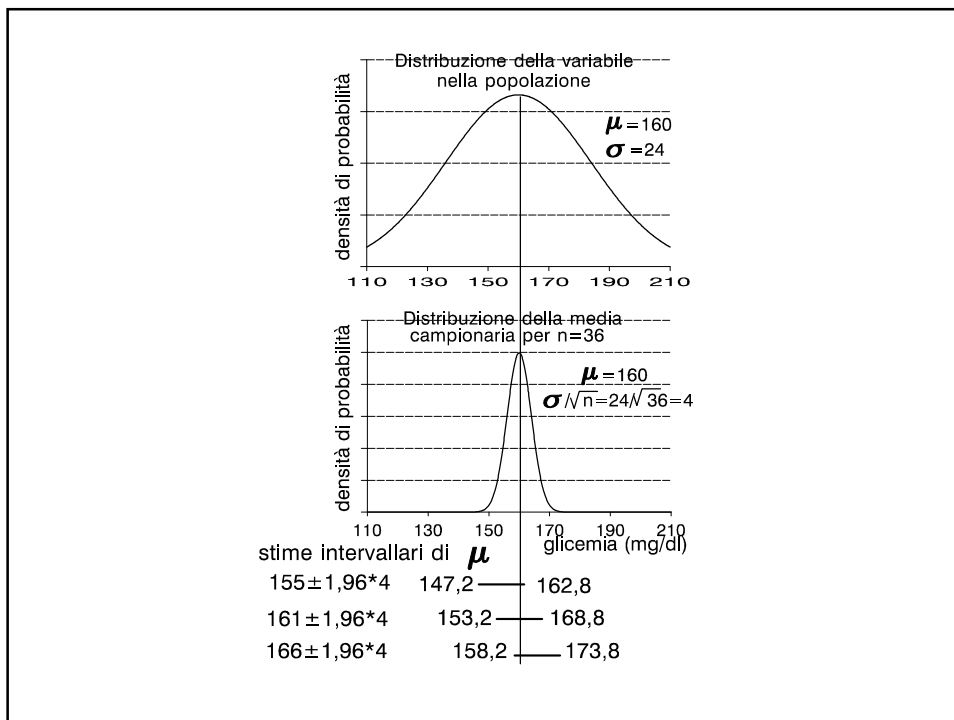
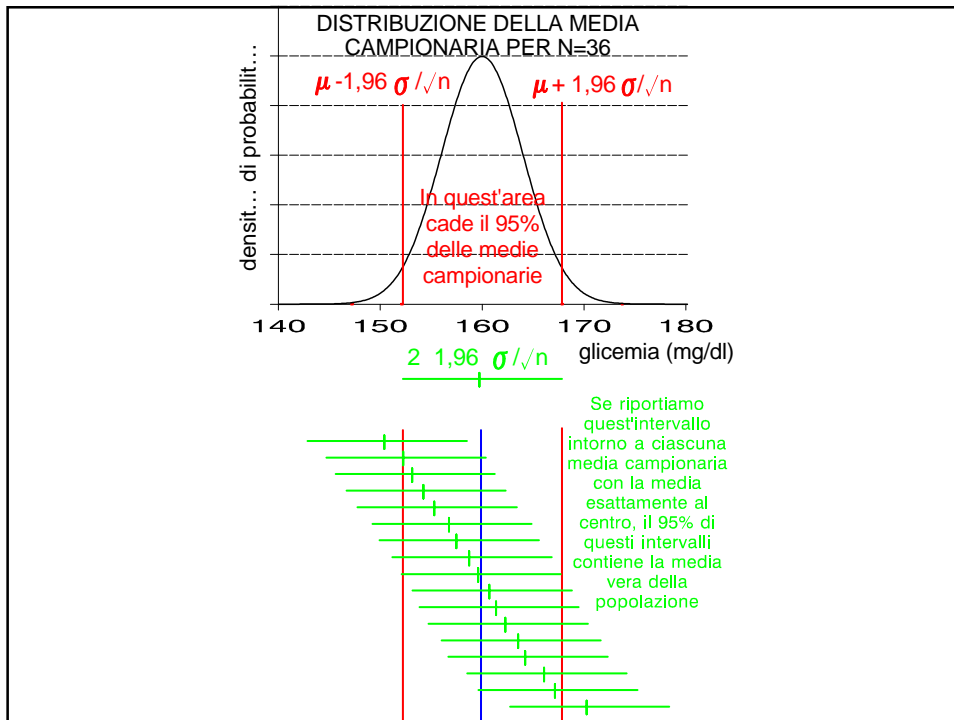


Dal momento che il campione viene estratto casualmente dalla popolazione, le conclusioni tratte da un campione possono essere errate.

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**





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:

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

dove:

$1 - \alpha$ = grado di confidenza

α = probabilità di errore

DERIVAZIONE DELL'INTERVALLO DI CONFIDENZA AL 95%
PER LA MEDIA DI UNA POPOLAZIONE (Dev.St. NOTA)

$$\Pr (\mu - 1.96 * \sigma / \sqrt{n} < \bar{x} < \mu + 1.96 * \sigma / \sqrt{n}) = 0,95$$

$$\mu - 1.96 * \sigma / \sqrt{n} < \bar{x} < \mu + 1.96 * \sigma / \sqrt{n}$$

↓ -μ

$$- 1.96 * \sigma / \sqrt{n} < \bar{x} - \mu < 1.96 * \sigma / \sqrt{n}$$

↓ -x̄

$$-\bar{x} - 1.96 * \sigma / \sqrt{n} < -\mu < -\bar{x} + 1.96 * \sigma / \sqrt{n}$$

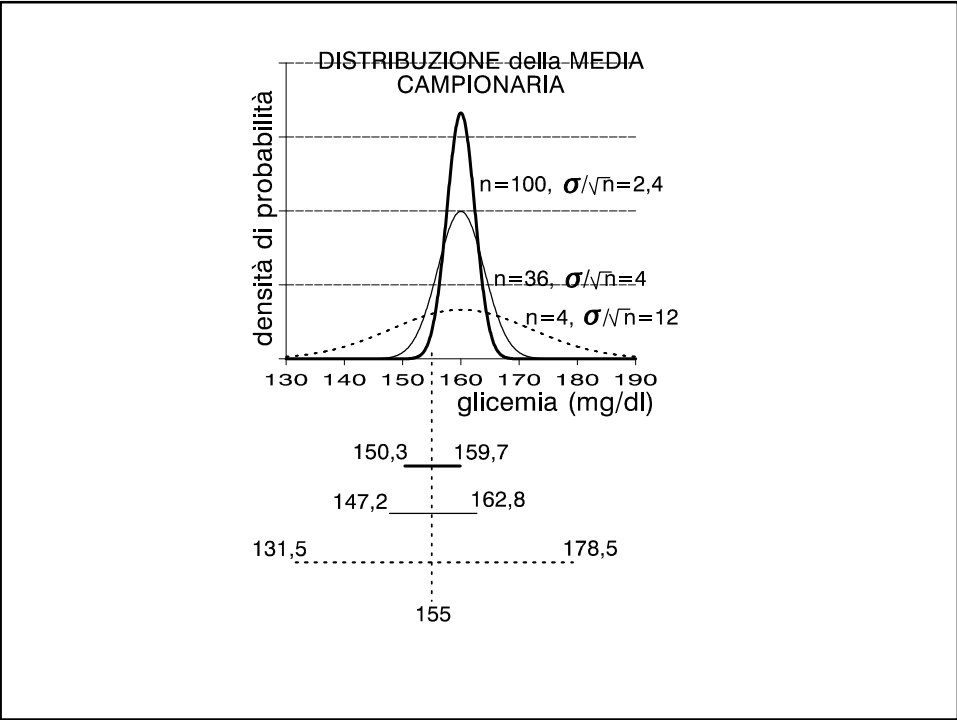
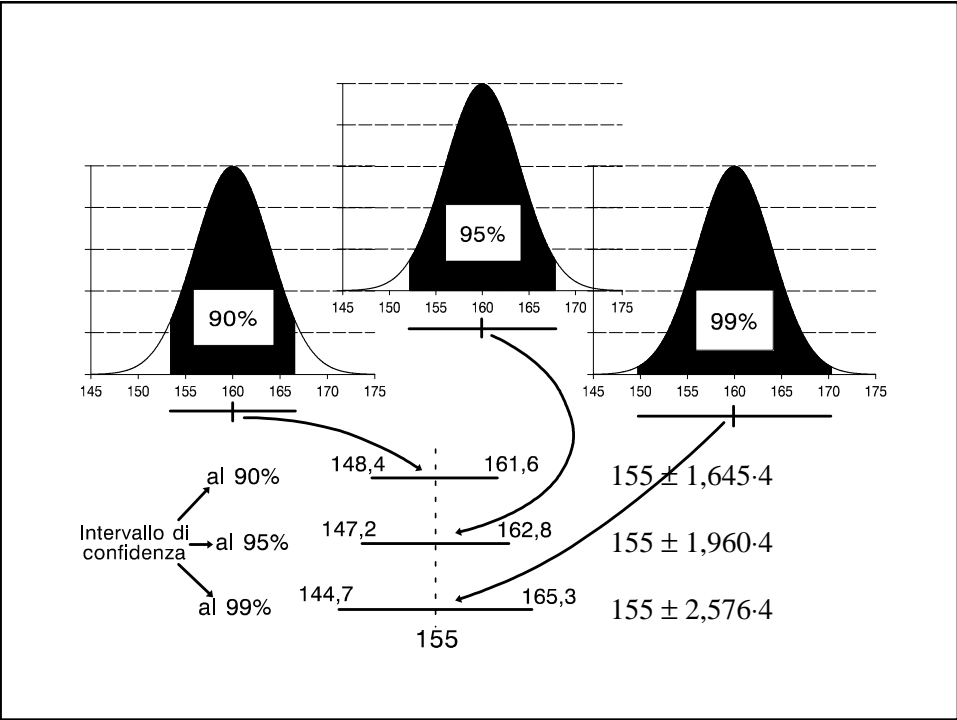
↓ **Moltiplico per -1**

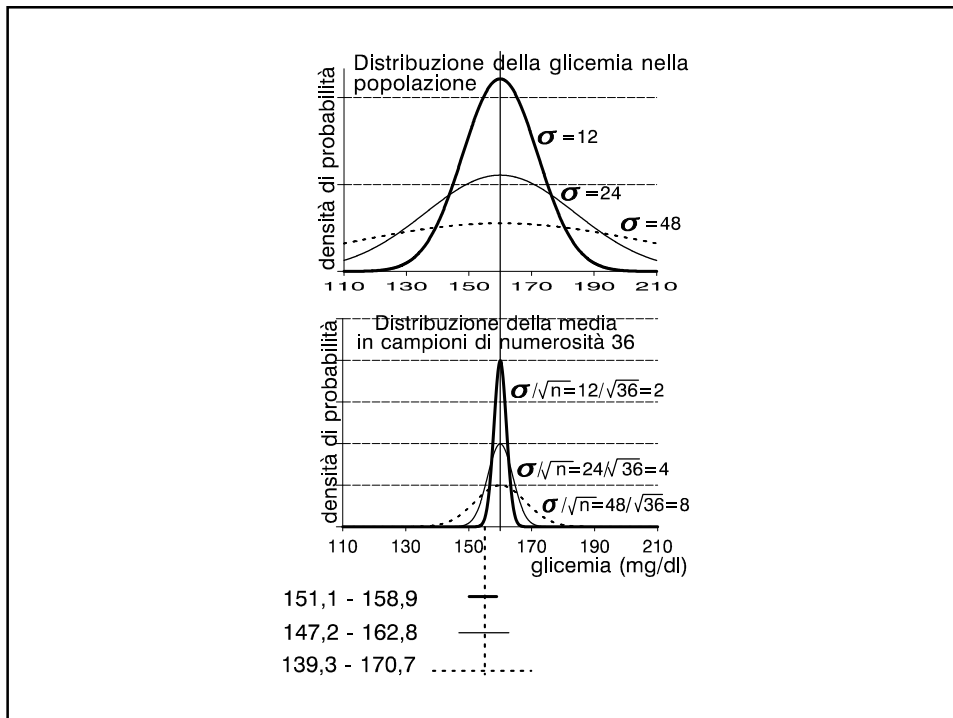
$$\bar{x} + 1.96 * \sigma / \sqrt{n} > \mu > \bar{x} - 1.96 * \sigma / \sqrt{n}$$

$$\bar{x} - 1.96 * \sigma / \sqrt{n} < \mu < \bar{x} + 1.96 * \sigma / \sqrt{n}$$

L'intervallo di confidenza **diminuisce** se

- 1) **diminuisce** il **livello di confidenza** (1-α)
(dal 99% al 95% al 90%)
- 2) **aumenta** la **numerosità** del campione
(da n=4 a n=36 a n=100)
- 3) **diminuisce** la **variabilità nella popolazione**
(da σ=4 a σ=36 a σ=100)





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

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

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

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

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**?

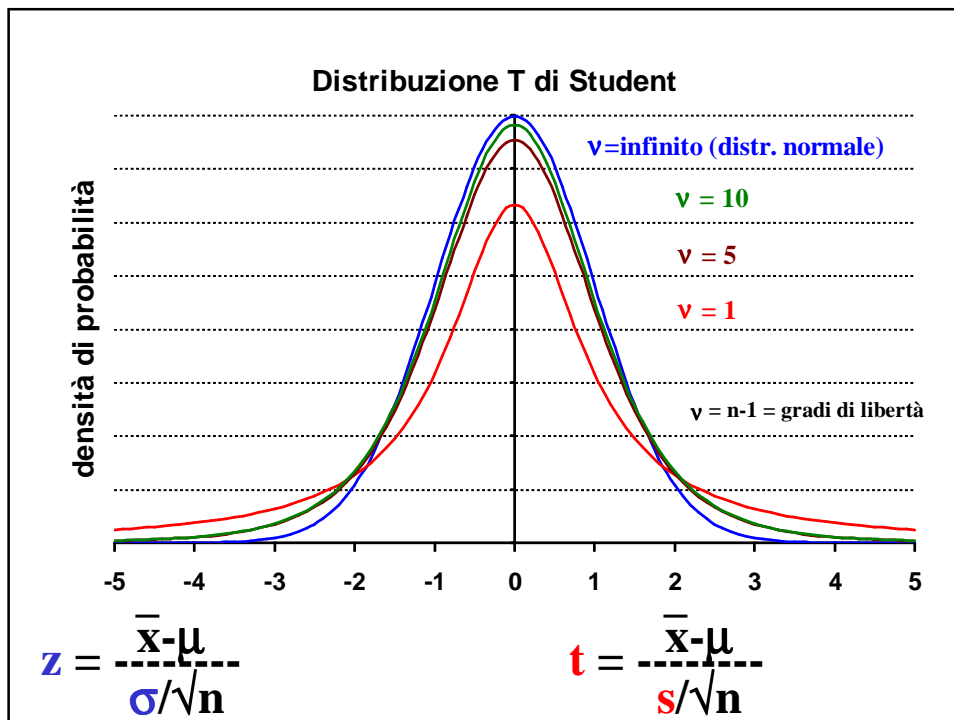
Posso usare S (**dev. standard del campione**) come stima di σ

Se la numerosità campionaria è sufficientemente grande ($n \geq 60$), S è una stima precisa di σ .

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

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.



Riassumendo:

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

$$\sigma \text{ nota} \qquad \Rightarrow \qquad \bar{x} \pm z_{\alpha/2} * \sigma / \sqrt{n}$$

$$\sigma \text{ ignota} \qquad \Rightarrow \qquad \bar{x} \pm t_{\alpha/2, v} * s / \sqrt{n}$$

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:

$$\sigma \text{ ignota} \quad n \geq 60 \quad \Rightarrow \quad \bar{x} \pm z_{\alpha/2} * s / \sqrt{n}$$

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

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

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

II passo: calcolo l'intervallo di confidenza

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

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.

Intervallo di confidenza

$$\theta = \mu$$

livello di confidenza = 0,95

$$\bar{x} - 1,96 * \sigma / \sqrt{n} < \mu < \bar{x} + 1,96 * \sigma / \sqrt{n}$$

per un generico livello di confidenza = $1-\alpha$

$$\bar{x} - Z_{\alpha/2} * \sigma / \sqrt{n} < \mu < \bar{x} + Z_{\alpha/2} * \sigma / \sqrt{n}$$

per un generico parametro θ

$$\hat{\theta} - Z_{\alpha/2} * E.S.(\hat{\theta}) < \theta < \hat{\theta} + Z_{\alpha/2} * E.S.(\hat{\theta})$$

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

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

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- α)

PER UNA PROPORZIONE

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

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 $1-\alpha = 95\%$

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

Intervallo di confidenza per proporzioni								
APPROSSIMAZIONE NORMALE: casi\geq10 e non-casi\geq10								
	tutti i			limite	limite			
casi	soggetti	p %	ESp %	infer. %	sup. %			
40	211	18,957	2,698	13,669	24,246			
48	300	16,000	2,117	11,851	20,149			
METODO ESATTO, basato sulla distribuzione binomiale								
	tutti i		limite	limite	calcoli statistici			
casi	soggetti	p %	infer. %	sup. %	pLOW	2,50%	pHIGH	2,50%
3	55	5,455	1,139	15,123	0,011393	0,02500	0,15123	0,02501
3	75	4,000	0,833	11,248	0,008326	0,02500	0,11248	0,02500

In una distribuzione binomiale con $\pi=0,0083$ ed $n=75$ la probabilità di osservare 3 o più casi è di 0,025

In una distribuzione binomiale con $\pi=0,1125$ ed $n=75$, $P(X \leq 3)=0,025$

Intervallo di confidenza per tassi di incidenza						
APPROSSIMAZIONE NORMALE: casi >= 30				ES = (√casi) / persone-anno IC 95% = inc ± 1,96*ES		
per 100000 persone-anno						
casi	persone	incidenza	ES	limite		
				infer. %	sup. %	
9	30000	30,000	10,000	10,400	49,600	
50	30000	166,667	23,570	120,469	212,864	
METODO ESATTO, basato sulla distribuzione di Poisson						
per 100000 persone-anno						
casi	persone	incidenza	mi		limite	
			mi0	mi1	infer. %	sup. %
9	30000	30,000	4,120	17,080	13,733	56,933
50	30000	166,667	37,110	65,920	123,700	219,733

In una distribuzione di Poisson con $\mu=4,12$ la probabilità di osservare 9 o più casi è di 0,025

In una distribuzione di Poisson con $\mu=17,08$, $P(X \leq 9) = 0,025$

Problema 4: Utilizzo dell'Intervallo di Confidenza per decidere la numerosità di un campione.

Problema: Si vuole stimare la prevalenza (probabilità) di asma in una popolazione. Dati preliminari provenienti dalla letteratura suggeriscono che la prevalenza di asma si aggiri intorno al 5%. Qual è la numerosità campionaria necessaria per ottenere un intervallo di confidenza al 95% di ampiezza inferiore o uguale al 2%?

Dati: $p = 0,05$ $1-\alpha = 95\%$ $z_{\alpha/2} = 1,96$ ampiezza IC $\leq 2\%$ $n = ?$

$$(p + z_{\alpha/2} \cdot \sqrt{p(1-p)/n}) - (p - z_{\alpha/2} \cdot \sqrt{p(1-p)/n}) \leq \delta$$

$$2 z_{\alpha/2} \cdot \sqrt{p(1-p)/n} \leq \delta$$

$$\sqrt{p(1-p)/n} \leq \delta / (2 z_{\alpha/2})$$

$$p(1-p)/n \leq \delta^2 / (2 z_{\alpha/2})^2$$

$$p(1-p) (2 z_{\alpha/2})^2 / \delta^2 \leq n$$

$$n \geq 0,05 * 0,95 * (2 * 1,96)^2 / 0,02^2$$

$$n \geq 0,0475 * 15,36 / 0,0004$$

divido il I e il II membro per $2 z_{\alpha/2}$

elevo il I e il II membro al quadrato

moltiplico per n e divido per il II membro

$$n \geq 0,0475 * (3,92)^2 / 0,0004$$

$$n \geq 1824,76$$

$$n \geq 1825$$