

LEZIONI DI STATISTICA MEDICA

Dott. SIMONE ACCORDINI

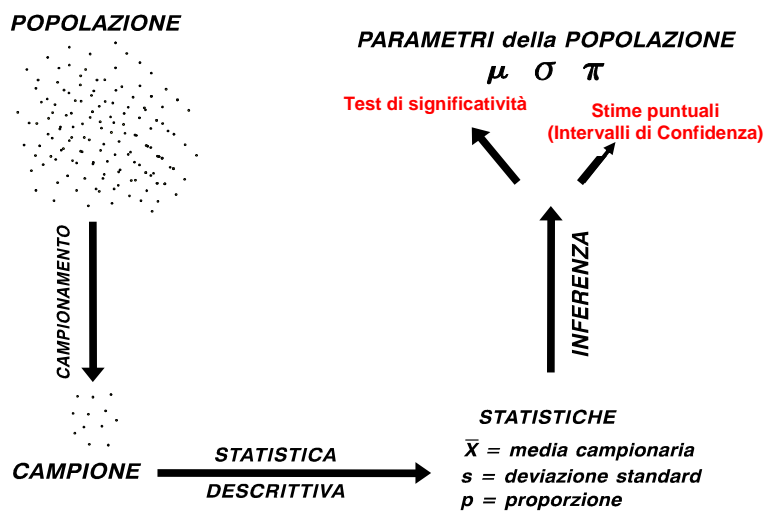
Lezione n.11

- Principi dell'inferenza statistica
- Test di significatività
- Intervallo di confidenza



Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

INFERENZA



METODI STATISTICI DELL'INFERENZA

1. Verificare se il parametro di occorrenza varia tra i gruppi ovvero se vi è associazione statistica → **le differenze osservate non sono dovute al caso**

⇒ **TEST DI SIGNIFICATIVITA'**

2. Stimare il parametro di occorrenza separatamente nei gruppi e la misura di effetto ...

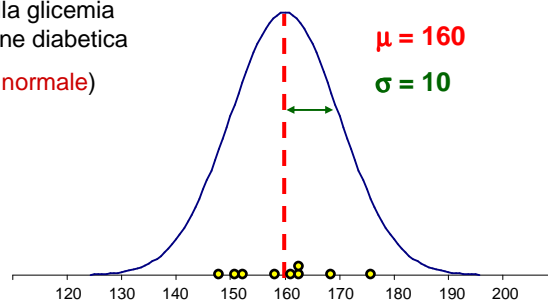
⇒ **STIMA PUNTUALE**

3. ... e associare alla stima puntuale una **misura di incertezza**

⇒ **INTERVALLO DI CONFIDENZA**



Distribuzione della glicemia
in una popolazione diabetica
(modello teorico normale)

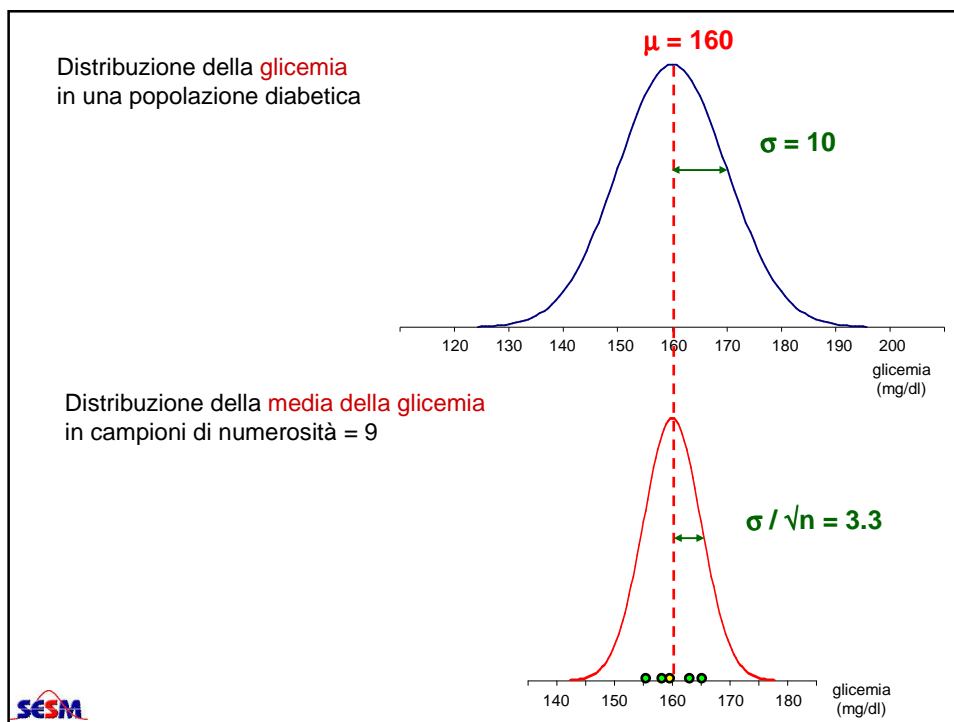


$\mu = 160$

$\sigma = 10$

I° campione:	162 152 158 162 168 161 148 176 150	→	$\bar{x} = 159.7$	$s = 8.9$
II° campione:	152 164 157 180 156 163 165 166 178	→	$\bar{x} = 164.6$	$s = 9.4$
III° campione:	157 142 163 162 152 149 152 180 151	→	$\bar{x} = 156.4$	$s = 11.0$
IV° campione:	162 154 168 160 155 172 162 152 140	→	$\bar{x} = 158.3$	$s = 9.5$
V° campione:	163 169 152 147 158 163 173 160 181	→	$\bar{x} = 162.8$	$s = 10.2$





DISTRIBUZIONE CAMPIONARIA DI UNA MEDIA

Sia \bar{x} la **media campionaria** di un campione casuale di dimensione n selezionato da una popolazione con media μ e deviazione standard σ :

1-2) la distribuzione campionaria di \bar{X} ha:

$$E[\bar{X}] = \mu$$

$$DS[\bar{X}] = ES[\bar{X}] = \sigma/\sqrt{n}$$

ERRORE STANDARD della media \rightarrow misura della precisione della stima

3) **TEOREMA DEL LIMITE CENTRALE**: se la dimensione campionaria è sufficientemente grande ($n \geq 30$), allora la distribuzione campionaria di \bar{X} è approssimativamente normale, indipendentemente dalla distribuzione della variabile nella popolazione

Esempio:

Cockburn et al (1980) riportano una ricerca clinica per la prevenzione dell'ipocalcemia infantile, nella quale donne in gravidanza che ricevevano un supplemento di vitamina D venivano messe a confronto con donne non trattate.

Calcemia del bambino misurata 6 giorni dopo la nascita:

	Numero di pazienti	Media (mg /100 ml)	DS (mg /100 ml)
<i>Vitamina D (D₁)</i>	233	9.36	1.15
<i>Controllo (D₀)</i>	394	9.01	1.33



Confronto della media di una variabile quantitativa tra 2 popolazioni - campioni indipendenti



Test T di Student

GRUPPO NON ESPOSTO (D₀):

$x_{01}, x_{02}, \dots, x_{0n_0} \rightarrow n_0$ determinazioni **indipendenti** della v.c. $X_0 \sim \text{Norm}(\mu_0, \sigma_0^2)$

GRUPPO ESPOSTO (D₁):

$x_{11}, x_{12}, \dots, x_{1n_1} \rightarrow n_1$ determinazioni **indipendenti** della v.c. $X_1 \sim \text{Norm}(\mu_1, \sigma_1^2)$

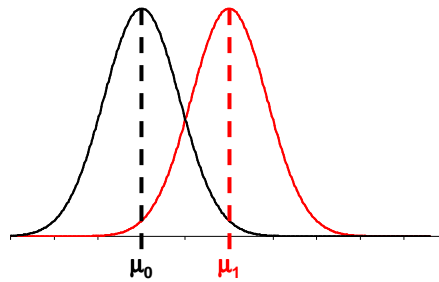


ASSUNZIONI SU CUI SI BASA IL TEST T DI STUDENT:

indipendenza

L'esito della misurazione in un soggetto non condiziona l'esito della misurazione negli altri soggetti.

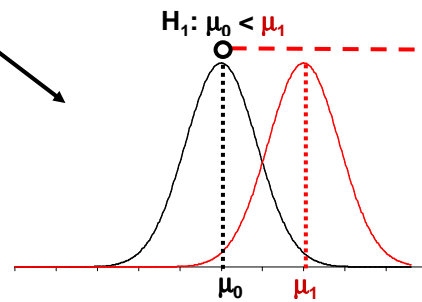
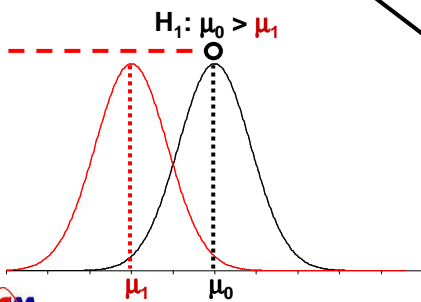
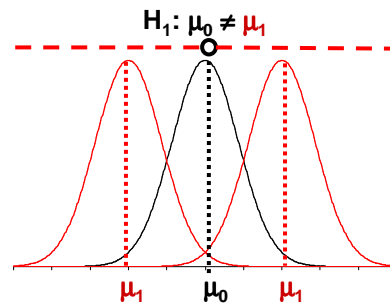
normalità



SESM

1°STEP: definire il sistema di ipotesi da verificare

$\left\{ \begin{array}{l} H_0: \mu_0 = \mu_1 \\ H_1: \dots \end{array} \right.$



SESM

II° STEP: definire la statistica test

$$\begin{cases} H_0: \mu_0 = \mu_1 & (\text{ipotesi nulla}) \\ H_1: \mu_0 \neq \mu_1 & (\text{ipotesi alternativa}) \end{cases}$$

differenza tra le medie = stima dell'effetto

$$T = \frac{\bar{X}_1 - \bar{X}_0}{ES[\bar{X}_1 - \bar{X}_0]}$$

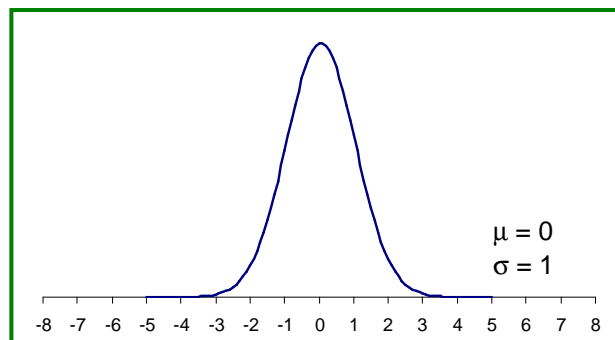
errore standard della differenza = misura della precisione della stima



$$\sigma_0^2 \neq \sigma_1^2$$

$$T = \frac{\bar{X}_1 - \bar{X}_0}{ES[\bar{X}_1 - \bar{X}_0]} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_1^2/n_1 + S_0^2/n_0}}$$

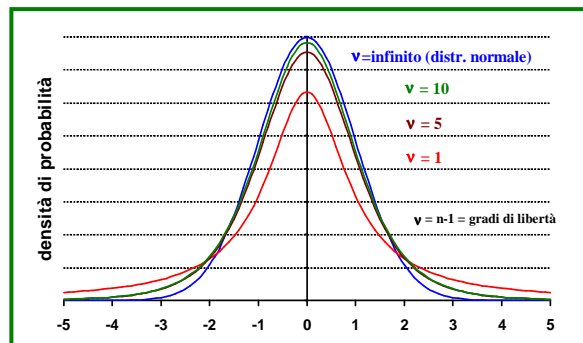
con campioni 'grandi' $(n_1 + n_0 - 2 \geq 30)$ \Rightarrow **distribuzione normale standard**
(sotto l'ipotesi nulla)



$$\sigma_0^2 \neq \sigma_1^2$$

$$T = \frac{\bar{X}_1 - \bar{X}_0}{\text{ES}[\bar{X}_1 - \bar{X}_0]} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_1^2/n_1 + S_0^2/n_0}}$$

con campioni 'piccoli' $(n_1 + n_0 - 2 < 30)$ \Rightarrow **distribuzione t di Student**
(sotto l'ipotesi nulla)



Esempio (continua):

	Numero di pazienti	Media (mg/100 ml)	DS (mg/100 ml)
<i>Vitamina D (D₁)</i>	233	9.36	1.15
<i>Controllo (D₀)</i>	394	9.01	1.33

$$\left\{ \begin{array}{l} H_0: \mu_0 = \mu_1 \\ H_1: \mu_0 \neq \mu_1 \end{array} \right. \leftarrow \text{I}^\circ \text{STEP}$$

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{S_1^2/n_1 + S_0^2/n_0}} = \frac{9.36 - 9.01}{0.1008} = 3.47 \leftarrow \text{II}^\circ \text{STEP}$$

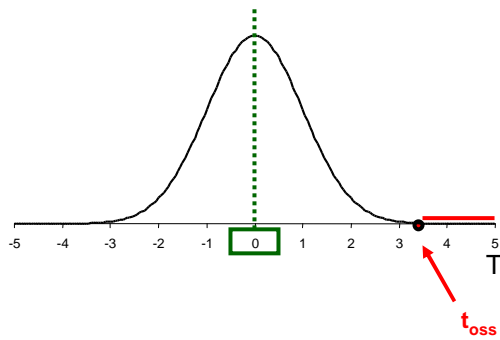


III° STEP: calcolare il p-value

misura continua del livello di consistenza dei dati con l'ipotesi nulla,
ovvero di quanto sono plausibili i dati sotto H_0

ipotesi
alternativa
unilaterale

$$\left\{ \begin{array}{l} H_0: \mu_0 = \mu_1 \\ H_1: \mu_0 < \mu_1 \end{array} \right.$$

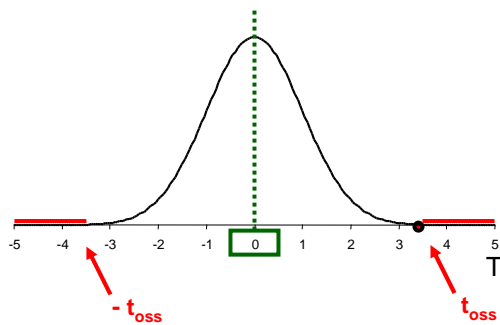


$$\text{upper one-sided } p\text{-value } (p_+) = \Pr(T \geq t_{oss} \mid H_0 \text{ è vera})$$



ipotesi
alternativa
bilaterale

$$\left\{ \begin{array}{l} H_0: \mu_0 = \mu_1 \\ H_1: \mu_0 \neq \mu_1 \end{array} \right.$$



$$\text{upper one-sided } p\text{-value } (p_+) = \Pr(T \geq t_{oss} \mid H_0 \text{ è vera})$$

$$\text{lower one-sided } p\text{-value } (p_-) = \Pr(T \leq -t_{oss} \mid H_0 \text{ è vera})$$

$$\text{two-sided } p\text{-value } (p) = 2 \min\{p_+, p_-\}$$



Esempio (continua):

	Numero di pazienti	Media (mg/100 ml)	DS (mg/100 ml)
Vitamina D (D ₁)	233	9.36	1.15
Controllo (D ₀)	394	9.01	1.33

$$\begin{cases} H_0: \mu_0 = \mu_1 \\ H_1: \mu_0 \neq \mu_1 \end{cases} \quad \leftarrow \text{I}^\circ \text{STEP}$$

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{S_1^2/n_1 + S_0^2/n_0}} = \frac{9.36 - 9.01}{0.1008} = 3.47 \quad \leftarrow \text{II}^\circ \text{STEP}$$

$$p\text{-value} = 2 \text{ prob}(T \geq 3.47 \mid H_0 \text{ è vera}) < 0.001 \quad \leftarrow \text{III}^\circ \text{STEP}$$

⇒ risultato 'significativo' e si rifiuta H₀

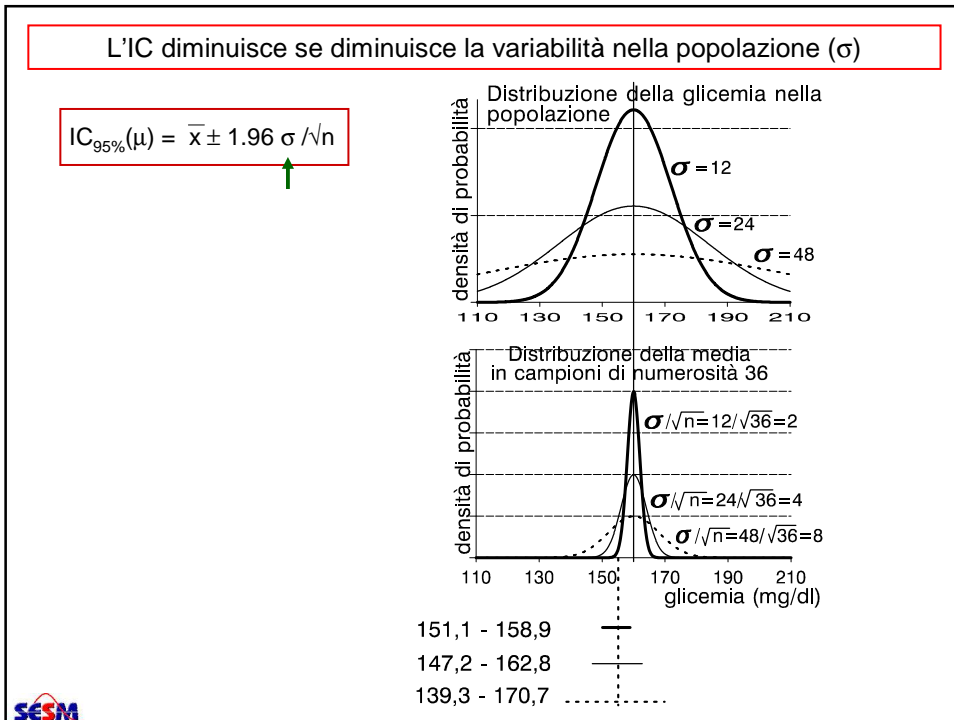
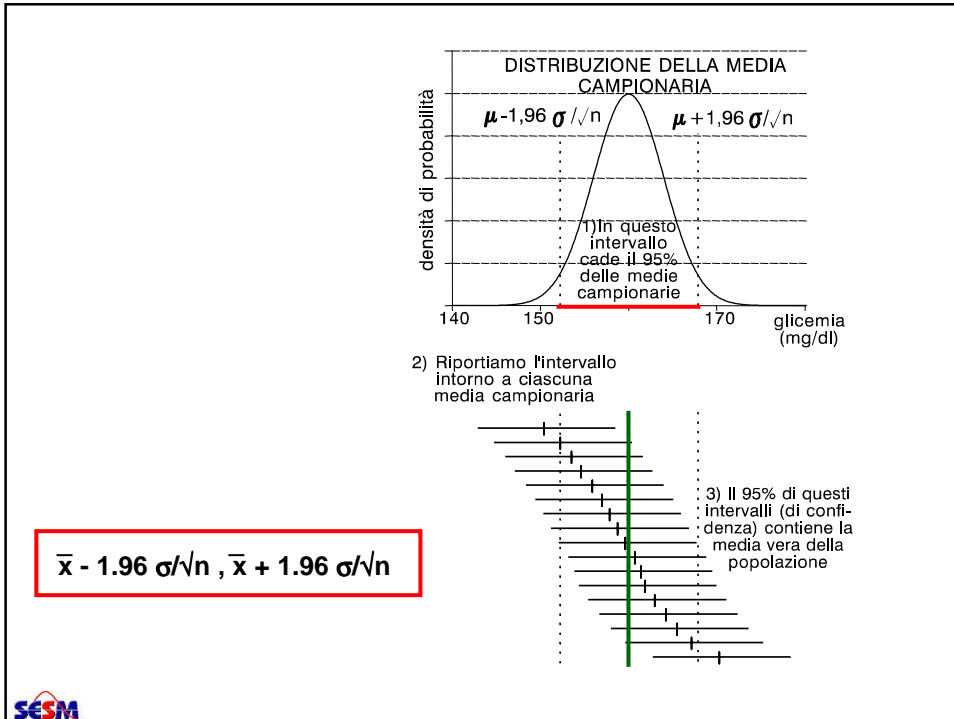


Intervallo di confidenza della media in 2 popolazioni - campioni indipendenti

Per intervallo di confidenza della media μ , si intende un intervallo delimitato da due limiti L_{inf} (limite inferiore) ed L_{sup} (limite superiore) che abbia una definita probabilità (livello di confidenza) di contenere il vero valore (ignoto) del parametro nella popolazione:

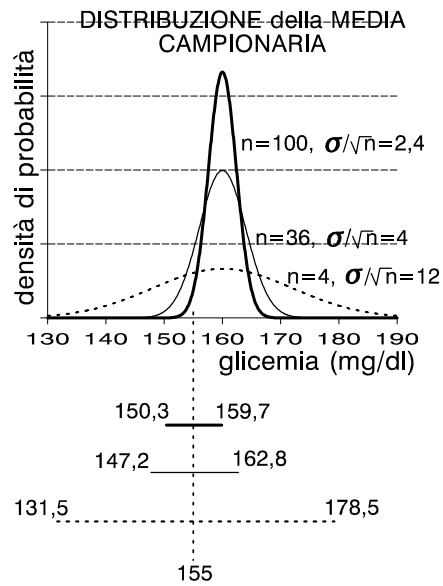
$$\text{prob}(L_{\text{inf}} < \mu < L_{\text{sup}}) = 0.95$$





L'IC diminuisce se aumenta la numerosità del campione (n)

$$IC_{95\%}(\mu) = \bar{x} \pm 1.96 \sigma / \sqrt{n}$$



Esempio (continua):

	Numero di pazienti	Media (mg/100 ml)	DS (mg/100 ml)
<i>Vitamina D (D₁)</i>	233	9.36	1.15
<i>Controllo (D₀)</i>	394	9.01	1.33

$$95\%CI(\mu_0) = \bar{x}_0 \pm 1.96 s_0 / \sqrt{n_0} = 9.01 \pm 1.96 * 1.33 / \sqrt{394}$$

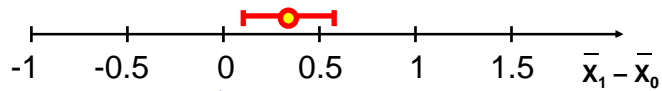
$$95\%CI(\mu_1) = \bar{x}_1 \pm 1.96 s_1 / \sqrt{n_1} = 9.36 \pm 1.96 * 1.15 / \sqrt{233}$$

	controllo (n = 394)	vitamina D (n = 233)	p-value
media	9.01	9.36	<0.001
[95%CI]	[8.88, 9.14]	[9.21, 9.51]	

Esempio (continua):

	Numero di pazienti	Media (mg/100 ml)	DS (mg/100 ml)
Vitamina D (D ₁)	233	9.36	1.15
Controllo (D ₀)	394	9.01	1.33

differenza delle medie [95%CI]	0.35 [0.15, 0.55]
--------------------------------	-------------------



assenza di associazione



Intervalli di confidenza che indicano cinque possibili interpretazioni in termini di significatività statistica e importanza pratica

