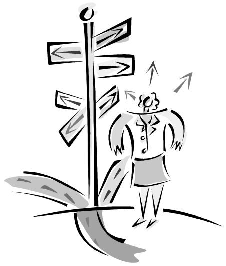
LEZIONI DI STATISTICA MEDICA

Gli Intervalli di confidenza



Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

INTERVALLO di CONFIDENZA

Lo scopo dell'inferenza statistica è la conoscenza dei **parametri** che caratterizzano una popolazione.

Per conoscere il parametro, però, dovremmo prendere in esame **tutte** le unità statistiche che costituiscono la popolazione; questo spesso è impossibile perché:

1. numerosità molto elevata
2. spesso la popolazione obiettivo è infinita

impossibile conoscere il **parametro**

Non potendo calcolare con esattezza il parametro, **ricorriamo ad una sua stima**.

La **statistica** (es $\overline{x}$, s) calcolata su un campione estratto dalla popolazione obiettivo è una **stima puntuale** del parametro della popolazione.

Questa stima puntuale del parametro <u>non sarà mai identica</u> al vero parametro della popolazione, ma sarà affetta da un **errore** per eccesso o per difetto.

In molte situazioni è preferibile **una stima intervallare** (cioè è preferibile indicare come stima del parametro un intervallo al posto di un *singolo punto* sull'asse dei valori) che esprima anche l'**errore associato alla stima** (precisione).

Tale stima prende il nome di:

INTERVALLO DI CONFIDENZA:

per IC di un parametro della popolazione θ, intendiamo un intervallo delimitato da $\mathbf{L_i}$ (limite inferiore) e $\mathbf{L_s}$ (limite superiore) che abbia una definita **probabilità (1 - α) di contenere il vero parametro della popolazione**:

$$\boxed{pr(L_i \le \theta \le L_s) = 1-\alpha}$$

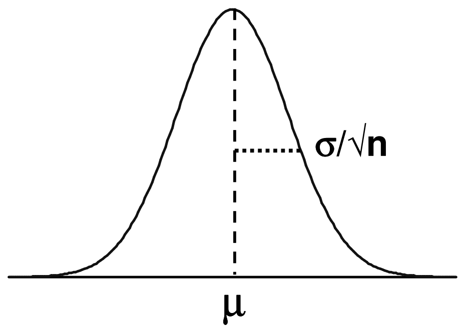
dove: 1 - α = **grado di confidenza**

α = **probabilità di errore**

quanto più grande è l'IC tanto più imprecisa è la nostra stima!

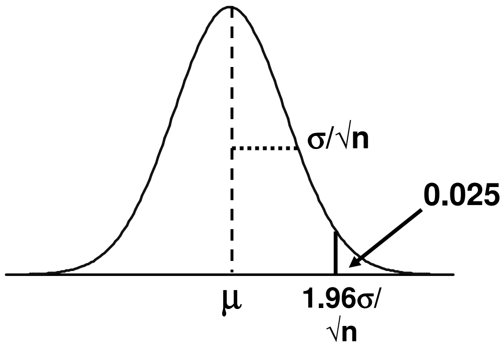
INTERVALLO di CONFIDENZA al 95% di una media

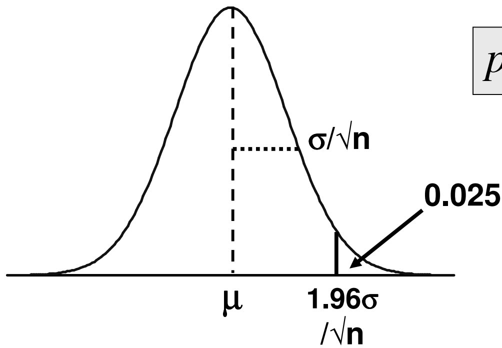
Si assuma che: $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$



Data la distribuzione, l'ntervallo simmetrico che comprende il 95% delle medie campionarie (p=0.95) sarà per definizione:

μ +/- 1.96 e.s.





$$\boxed{pr(L_i \le \theta \le L_s) = 1-\alpha}$$

$$pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

$$pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

e, riarrangiando le due disuguaglianze interne alla parentesi:

$$pr\left\{\boxed{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}}\right\} = 0.95$$

INTERVALLO DI CONFIDENZA

SESM

$$pr\left\{\boxed{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}} \le \mu \le \boxed{\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}}\right\} = 0.95$$

L_i

(LIMITE INFERIORE DELL'INTERVALLO)

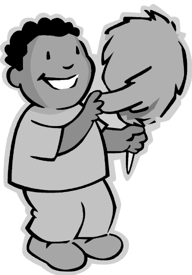
L_s

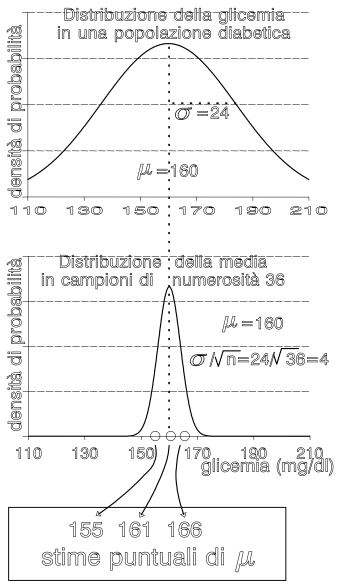
(LIMITE SUPERIORE DELL'INTERVALLO)

SESM

GLICEMIA nella POPOLAZIONE DIABETICA

(stime puntuali)

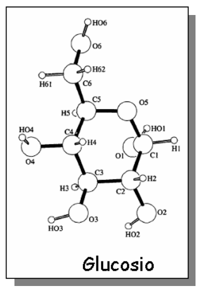


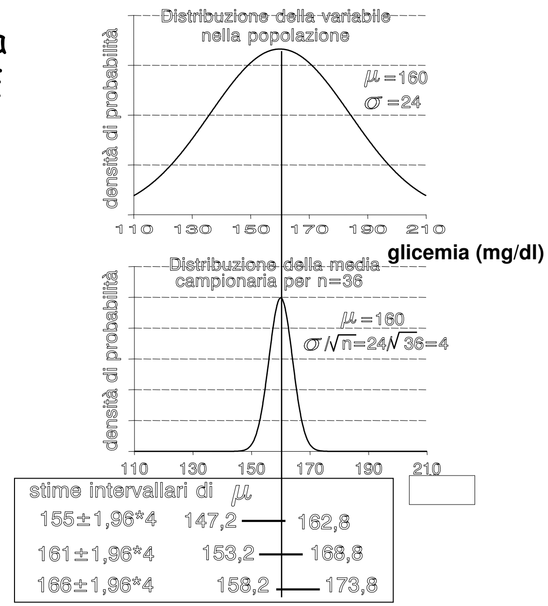


SESM

GLICEMIA nella POPOLAZIONE DIABETICA

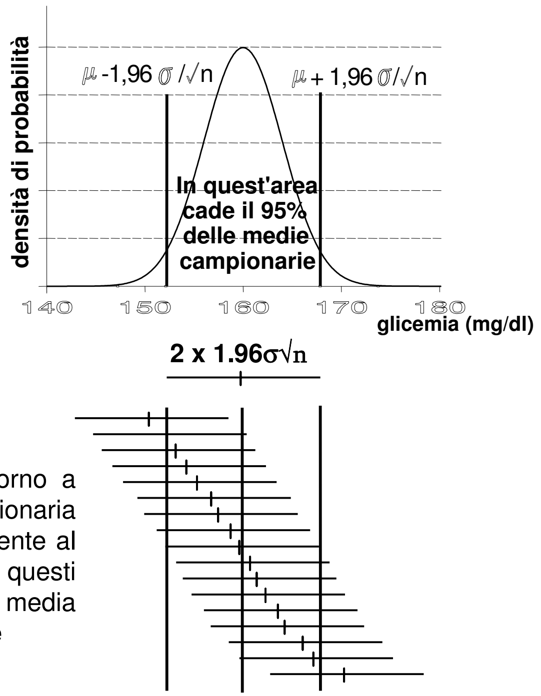
(stime intervallari)





SESM

DISTRIBUZIONE DELLA MEDIA CAMPIONARIA PER N=36



se riportiamo l'IC attorno a ciascuna media campionaria con la media esattamente al centro, il 95% di questi intervalli contiene la media vera della popolazione

RIASSUMENDO...

La **stima puntuale** fornisce un <u>singolo valore</u>. 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 <u>un intervallo</u>:

1. quest'intervallo ha una determinata probabilità (in genere, il 95%) di contenere il valore vero (parametro) della popolazione;
2. Il metodo generale per la costruzione dell'intervallo di confidenza al (1-α)% è:

$$\bar{x} \pm z_{\alpha/2} \cdot ES(\bar{x})$$

da cosa dipende l'ampiezza dell'IC?

$$\bar{x} \pm z_{\alpha/2} \cdot ES(\bar{x})$$

1. la **probabilità d'errore** α che determina il valore del coefficiente del limite fiduciale (z):

1-α	**α/2**	**$z_{\alpha/2}$**
0.90	**0.05**	**1.64**
0.95	**0.025**	**1.96**
0.98	**0.01**	**2.33**
0.99	**0.005**	**2.58**

2. la **dimensione del campione (n)**
3. la **variabilità della variabile nella popolazione (σ)**

NOTA BENE!

Nel calcolare l'intervallo di confidenza di una media si è supposto che la deviazione standard della popolazione fosse nota.

Infatti è stata usata la deviata standardizzata:

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

Molto spesso, però, σ è ignoto. In questo caso, se il campione è sufficientemente grande, **s può essere utilizzata come stima di σ** per il calcolo dell'intervallo di confidenza

intervallo di confidenza della media della popolazione <u>per piccoli campioni</u>

In medicina molto spesso si presenta la necessità di fare inferenze sulla base di **piccoli campioni (n < 30)** per:

- limitare i costi o il tempo dell'indagine
- studio di malattie molto rare

Quando il campione è piccolo sorgono i seguenti problemi:

1. poiché il **teorema del limite centrale non può essere applicato**, la distribuzione di $\overline{x}$ dipende dalla distribuzione della variabile nella popolazione.
2. la d.s. campionaria **s non è una buona approssimazione di σ** ed è tanto più insoddisfacente quanto più il campione è piccolo.

Per poter fare inferenze sulla media nel caso di piccoli campioni è necessario **assumere** che la variabile in studio abbia una **distribuzione approssimativamente normale**.

Sotto tale assunzione si può dimostrare che la variabile:

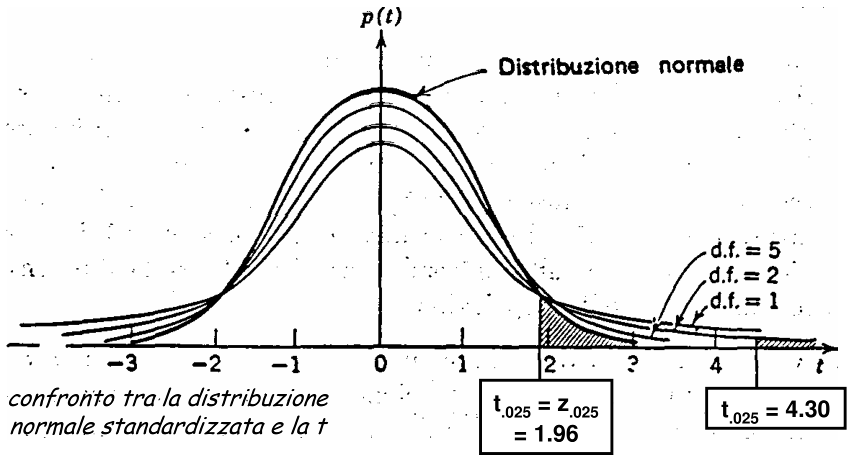
$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

ha una **distribuzione t di Student** per n – 1 gradi di libertà.

Nel caso di piccoli campioni l'intervallo di confidenza della media diventa quindi:

$$\overline{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

La **distribuzione t di Student** rappresenta una **famiglia di distribuzioni** simmetriche e differenti a seconda dei gradi di libertà



confronto tra la distribuzione normale standardizzata e la t

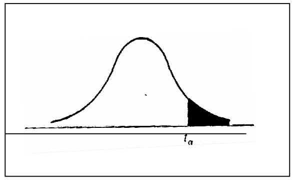


Tavola dei valori della funzione t di Student
in funzione dei gradi di libertà e della
probabilità in una coda della distribuzione
(.100, .050, .025, .010, .005)

DEGREES OF FREEDOM	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

INTERVALLO di CONFIDENZA di una PROPORZIONE

Per N > 30: $$p \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

In analogia con quanto visto per la media, segue che:

π sarà stimato da p

E che:

$$\boxed{p \pm z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}}$$