DISTRIBUZIONI CAMPIONARIE degli STIMATORI

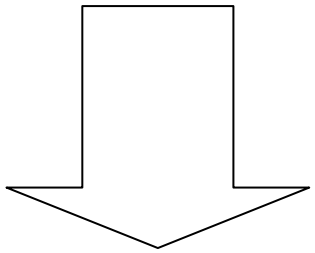
Una volta selezionato il campione, la variabile di interesse viene misurata sugli elementi che lo costituiscono.

I valori che la variabile assume vengono poi sintetizzati utilizzando le statistiche opportune (media, d.s, etc.).

Le statistiche campionarie sono stime dei parametri ignoti della popolazione al cui valore siamo interessati.

Le statistiche campionarie, tuttavia, dipendono dal particolare campione selezionato e variano da campione a campione!

Ripetendo per molte volte la procedura di campionamento si potrebbe costruire una distribuzione di frequenza con i valori della statistica calcolata sui differenti campioni.



le statistiche campionarie sono **variabili casuali** caratterizzate da una specifica distribuzione di probabilità (**distribuzione campionaria dello stimatore**).

La **distribuzione campionaria di una statistica** basata su n osservazioni è la distribuzione di frequenza dei valori che la statistica assume.

Tale distribuzione è generata teoricamente prendendo infiniti campioni di dimensione n e calcolando i valori della statistica per ogni campione.

POPOLAZIONE	**CAMPIONE**
	$x_1, x_2, \ldots, x_n$
$X \sim f(X)$	
$\theta \{\mu, \sigma, \pi\}$ (costanti)	$\hat{\theta}$ {x, s, p} (variabili casuali)
	f ($\hat{\theta}$) distribuzione campionaria degli stimatori

SESM

PROPRIETÀ della DISTRIBUZIONE CAMPIONARIA di una MEDIA

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

1) La distribuzione campionaria di $\bar{x}$ ha la media uguale alla media della popolazione da cui proviene il campione:

$$E(\bar{x}) = \mu$$

PROPRIETÀ della DISTRIBUZIONE CAMPIONARIA di una MEDIA

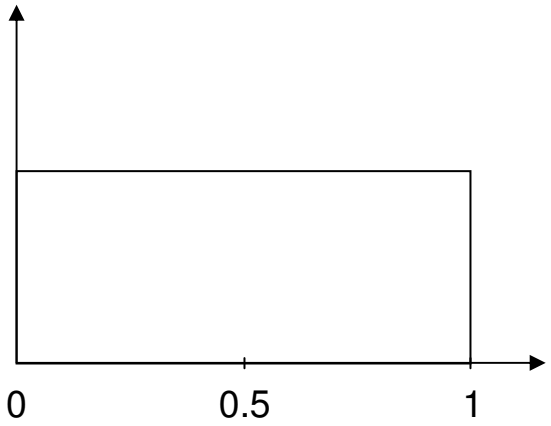
2) La distribuzione campionaria di $\bar{x}$ ha d.s. uguale alla d.s. della popolazione diviso la radice quadrata di n [errore standard - e.s]:

$$d.s.(\bar{x}) = \sigma/\sqrt{n} = e.s.$$

3) **TEOREMA CENTRALE DEL LIMITE**

Se la dimensione campionaria è sufficientemente grande (n > 30) la distribuzione campionaria di $\bar{x}$ è approssimativamente **normale, i**ndipendentemente dalla forma della distribuzione della variabile nella popolazione.

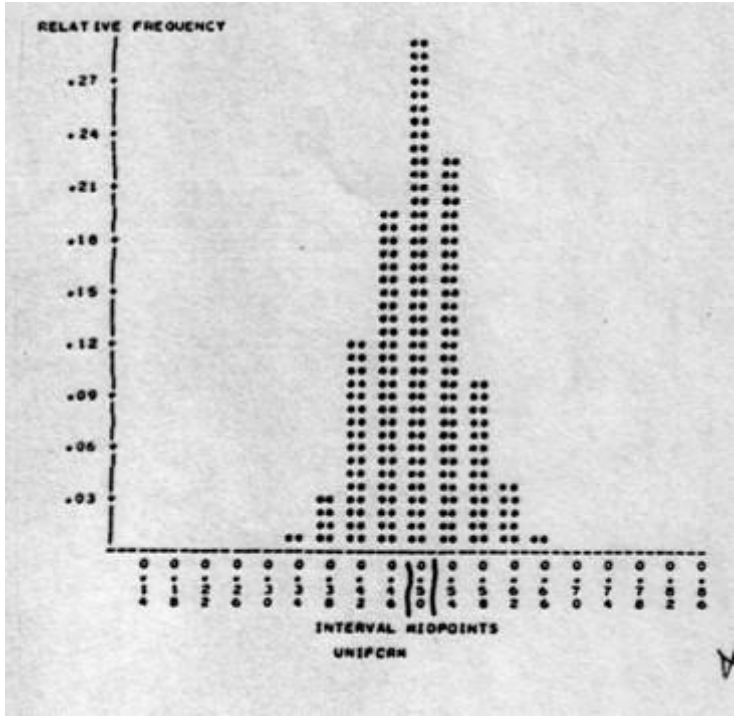
Distribuzione della variabile nella popolazione, f(X)



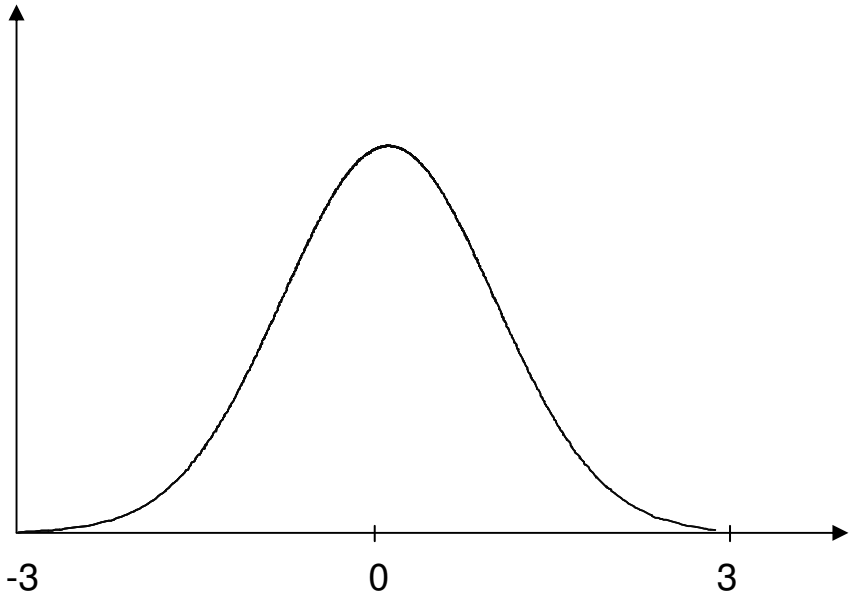
uniforme

($\mu = 0.5$, $\sigma = 0.29$)

Distribuzione empirica di $\overline{x}$ in 1000 campioni di n = 25



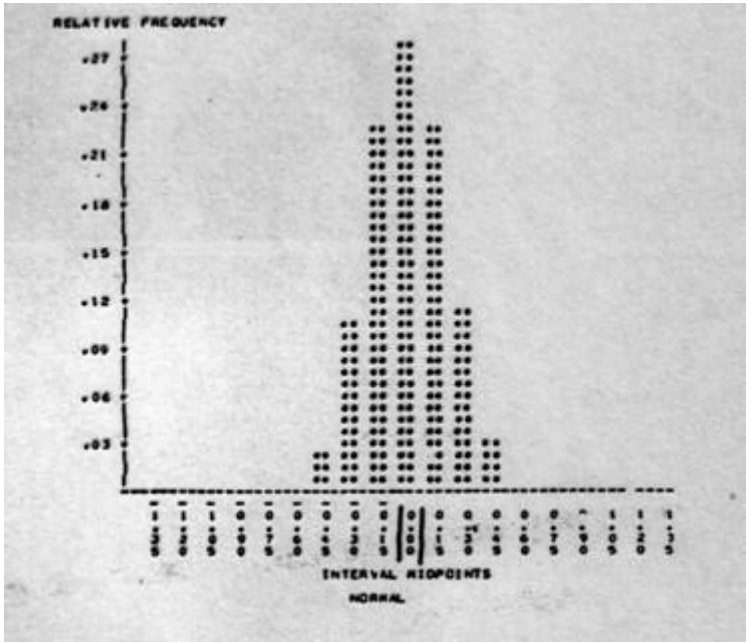
Distribuzione della variabile nella popolazione, f(X)



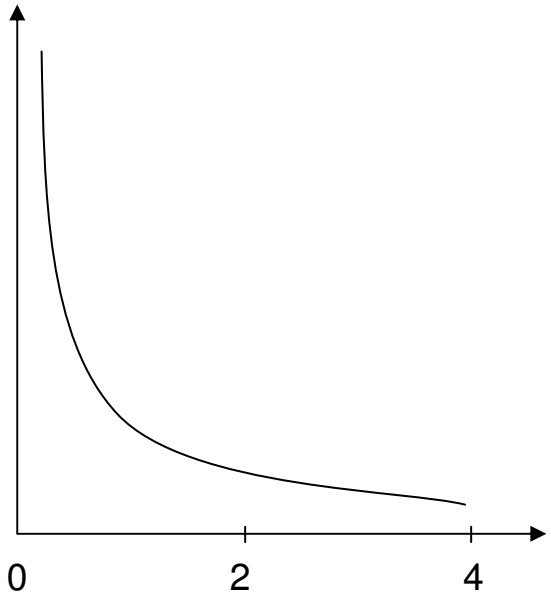
normale

($\mu = 0$, $\sigma = 1$)

Distribuzione empirica di $\bar{x}$ in 1000 campioni di n = 25



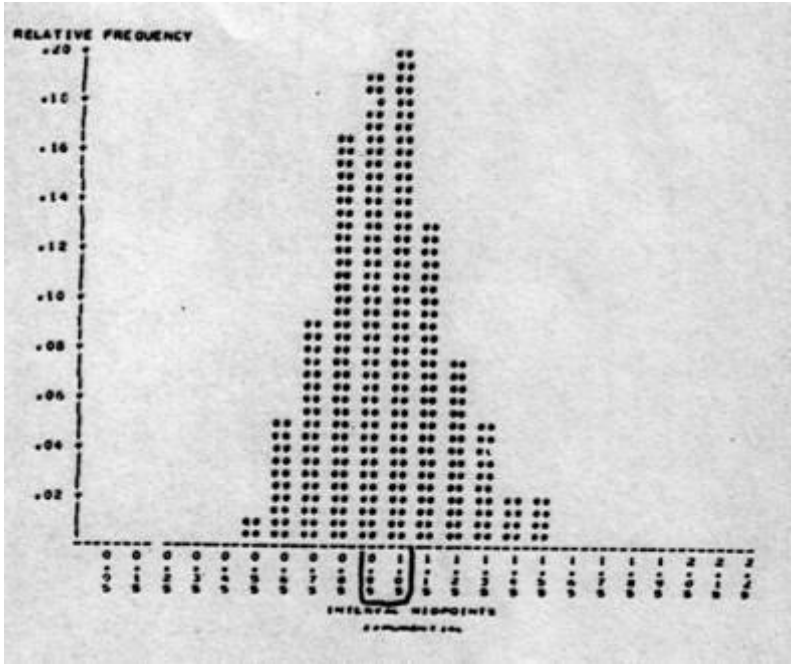
Distribuzione della variabile nella popolazione, f(X)



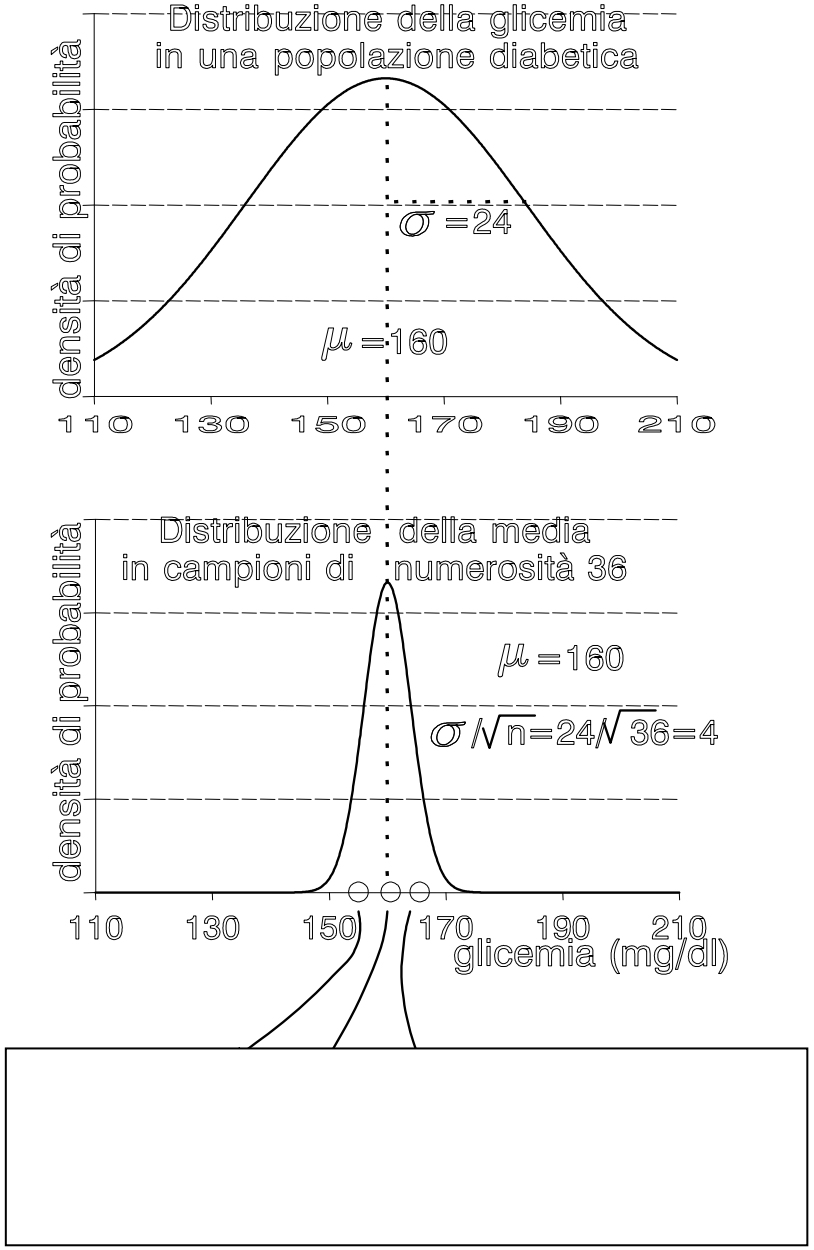
esponenziale

($\mu = 1$, $\sigma = 1$)

Distribuzione empirica di $\bar{x}$ in 1000 campioni di n = 25



Relazione tra
distribuzione di X
e distribuzione campionaria
di $\bar{x}$



esempio:
Si è stabilito sperimentalmente su un gran numero di pazienti affetti da un determinato tipo di tumore ad un certo stadio che il tempo medio di sopravvivenza dalla diagnosi è di 38.3 mesi con d.s. pari a 43.3 mesi.

Qual è la probabilità che un campione casuale di 100 soggetti abbia una sopravvivenza $\geq$ 46.9 mesi?

$$\bar{x} = 46.9$$
$$d.s. = 43.3$$
$$n = 100$$

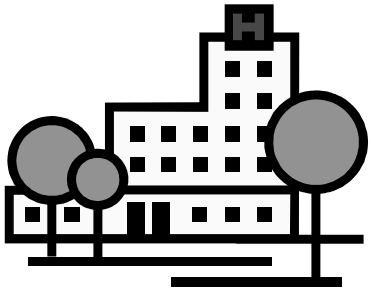
per il teorema del limite centrale:

$$\bar{x} \sim N(38.3, 43.3/\sqrt{100})$$

La variabile casuale in studio è $\bar{x}$, e la corrispondente deviata standardizzata sarà:

$$z = \frac{\bar{x} - E(x)}{d.s.(\bar{x})} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad z = \frac{46.9 - 38.3}{43.3/\sqrt{100}} = \frac{8.6}{4.3} = 2$$

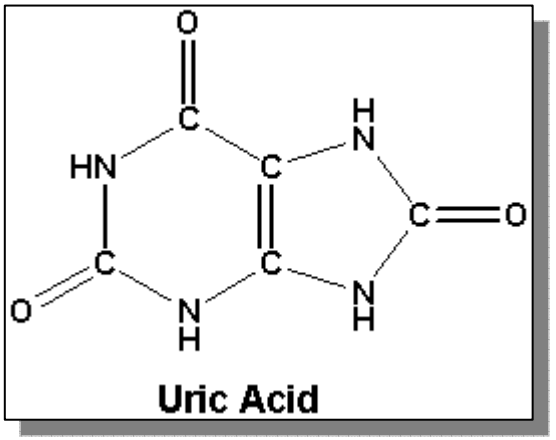
$$pr(\bar{x} \geq 46.9) = pr(z \geq 2) = 0.0227$$



pr = 2.3%

ESERCIZIO:

Sapendo che nella popolazione maschile l'acido urico serico è distribuito **normal-mente** con media = 5.4 mg/100 ml e d.s. = 1 mg/100 ml:



a) calcolare la probabilità di estrarre un campione di **30** soggetti che abbia una media > di 5.9 mg/100 ml.

b) Si calcoli l'intervallo simmetrico in cui ricadono il 95% dei campioni di 30 soggetti.

DISTRIBUZIONE CAMPIONARIA DI UN CONTEGGIO BINOMIALE

Sia X il numero di successi su n prove. Si supponga che X provenga da una distribuzione B (n,π). Allora, per n. sufficientemente grande (n>30) vale (approssimazione alla normale):

$$E(X) = n\pi$$

$$Var(X) = n\pi(1-\pi)$$

$$X \sim N\left(n\pi; \sqrt{n\pi(1-\pi)}\right)$$

DISTRIBUZIONE CAMPIONARIA DI UNA PROPORZIONE

Sia X una variabile B (n,π). Sia Y=X/n (proporzione di successi).
Per np >=10 vale:

$$E(Y) = \pi$$

$$Var(Y) = \frac{\pi(1-\pi)}{n}$$

$$Y \sim N\left(\pi; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

SESM

Esercizio

Si sa che il 20% della popolazione Italiana soffre di rinite allergica. Qual è la probabilità di estrarre a caso dalla popolazione un campione di 100 soggetti con una prevalenza di rinite inferiore al 15%?

Risposta P =0.1056

DISTRIBUZIONE CAMPIONARIA DI UN CONTEGGIO POISSONIANO

Sia X il numero di eventi. Si supponga che X provenga da una distribuzione di Poisson con parametro λ. Allora, per λ sufficientemente grande (> 5:20) vale (approssimazione alla normale):

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

$$X \sim N\left(\lambda; \sqrt{\lambda}\right)$$

esercizio

- Si supponga che numero di atteso di incidenti giornalieri in un dato terrritorio sia 8. Qual è la probabilità che in un giorno si verifichino meno di 2 incidenti stradali?

p=0.017

DISTRIBUZIONE CAMPIONARIA DI UN TASSO (Incidenza , Mortalità)

Un tasso è il *rapporto tra il numero di eventi e l'esperienza tempo che li ha generati: R=X/T* . In tale contesto è conveniente assumere che il numeratore (X=conteggio) è distribuito come una ***Poisson con parametro λT***. Se gli eventi sono calcolati in un periodo di 2.5 anni, il parametro è: 2.5λ; se il periodo è di 5 anni, il parametro è 5λ..

$$E(R) = \lambda$$

$$Var(R) = \frac{R}{T}$$

$$X \sim N\left(\lambda; \sqrt{\frac{R}{T}}\right)$$

Vale per λT > 5:20

Esempio

- **Un evento si verifica con tasso costante di 8.7 al giorno. Qual' è la probabilità di osservare meno di 50 eventi in una settimana?**

  come chiedersi: ***Qual è la probabilità che in una settimana il tasso medio sia inferiore a R=50/7=7.14 ?***

$$\boldsymbol{R} \rightarrow \boldsymbol{N}(8.7; \sqrt{\frac{8.7}{7}})$$

$$\boldsymbol{Z} = \frac{7.14 - 8.7}{\sqrt{\frac{8.7}{7}}} = -1.25 \quad \longrightarrow \quad \text{p=0.1056}$$

INTERVALLO di CONFIDENZA

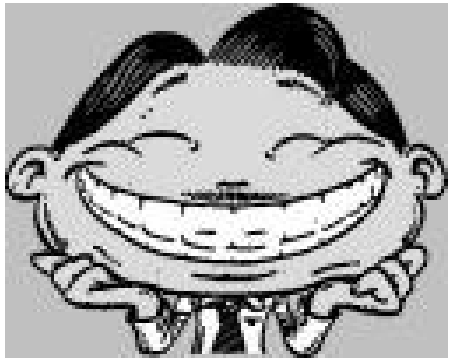
Lo scopo dell'inferenza statistica è la conoscenza dei **parametri** che caratterizzano una popolazione.

Per conoscere il parametro, 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 $\bar{x}$, s) calcolata su un campione estratto dalla popolazione obiettivo è una **stima puntuale** del parametro della popolazione.

Questa stima puntuale del parametro non sarà mai identica 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**:

$$pr(L_i \leq \theta \leq L_s) = 1 - \alpha$$

dove: 1 - α = **grado di confidenza**

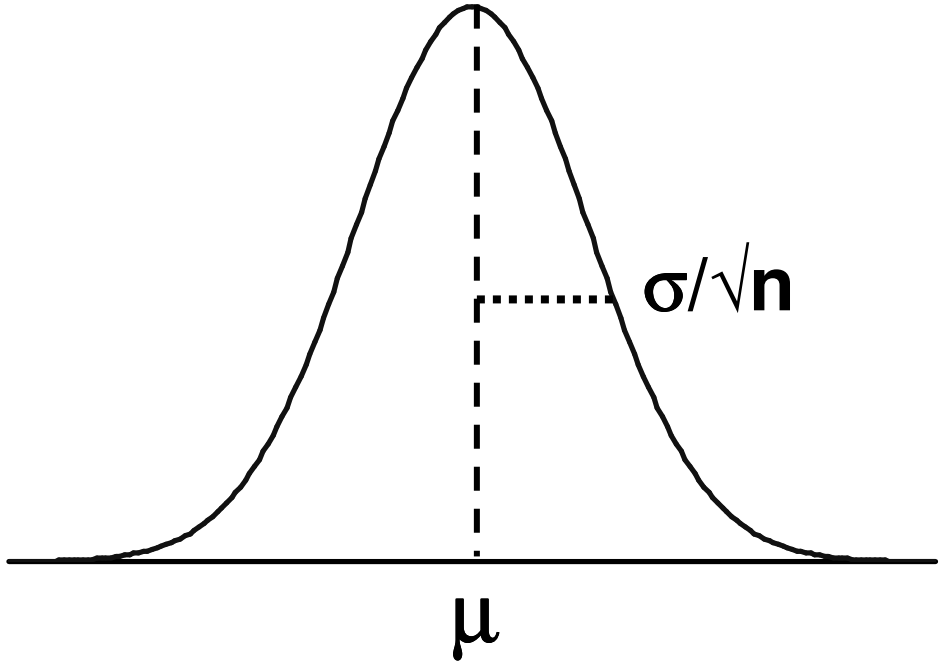
α = **probabilità di errore**

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

SESM

INTERVALLO di CONFIDENZA al 95% di una media

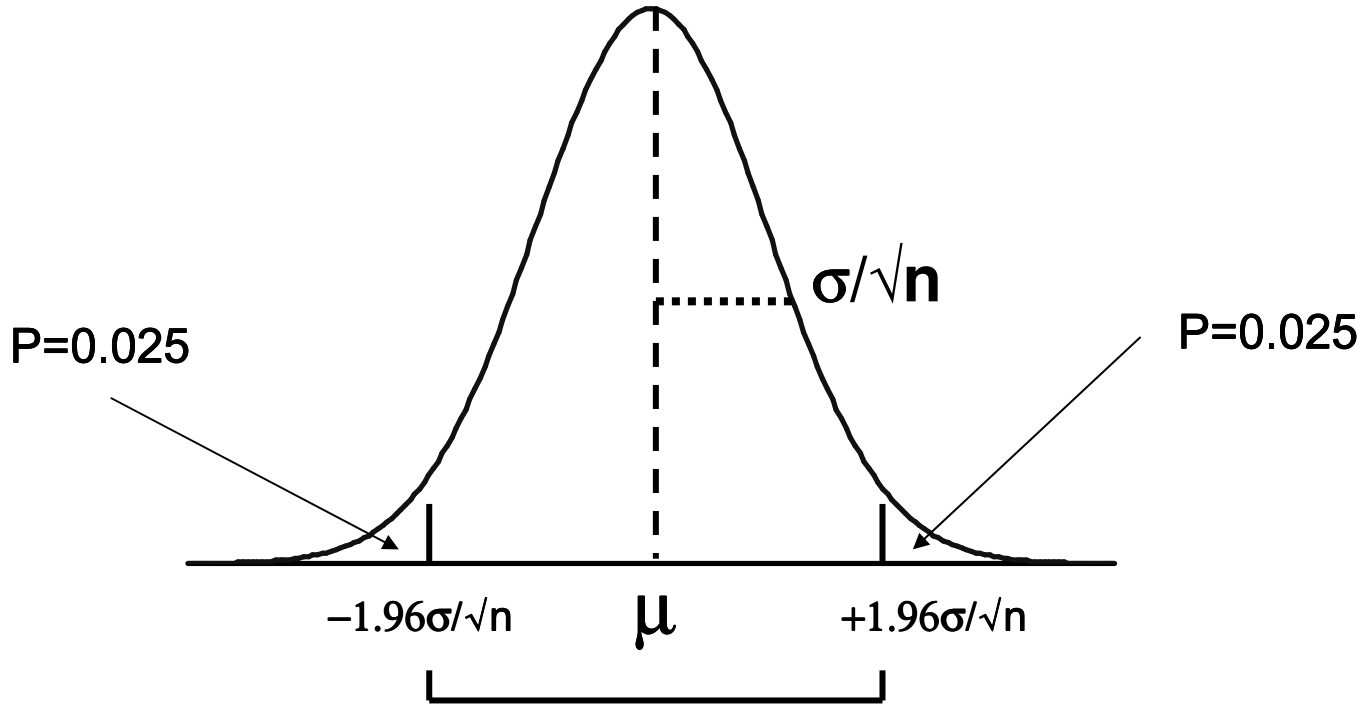
Si assuma che: $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

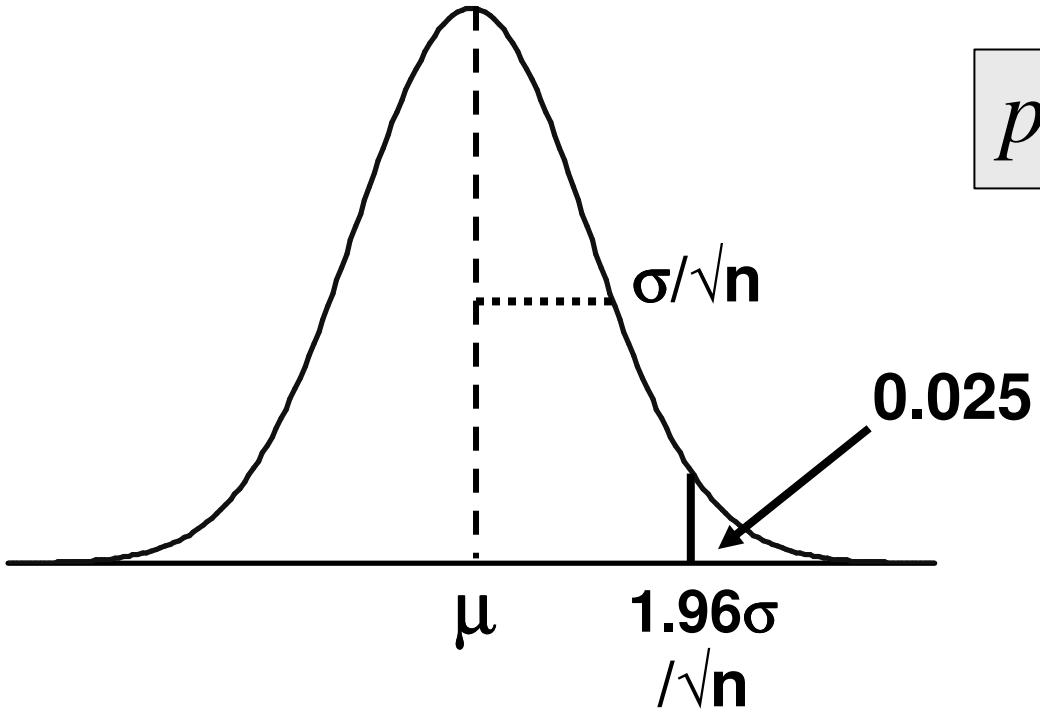


SESM

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

μ +/- 1.96 e.s.





$$pr(L_i \leq \theta \leq L_s) = 1 - \alpha$$

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

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

e, riarrangiando le due disuguaglianze interne alla parentesi:

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

INTERVALLO DI CONFIDENZA

SESM

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

L_i

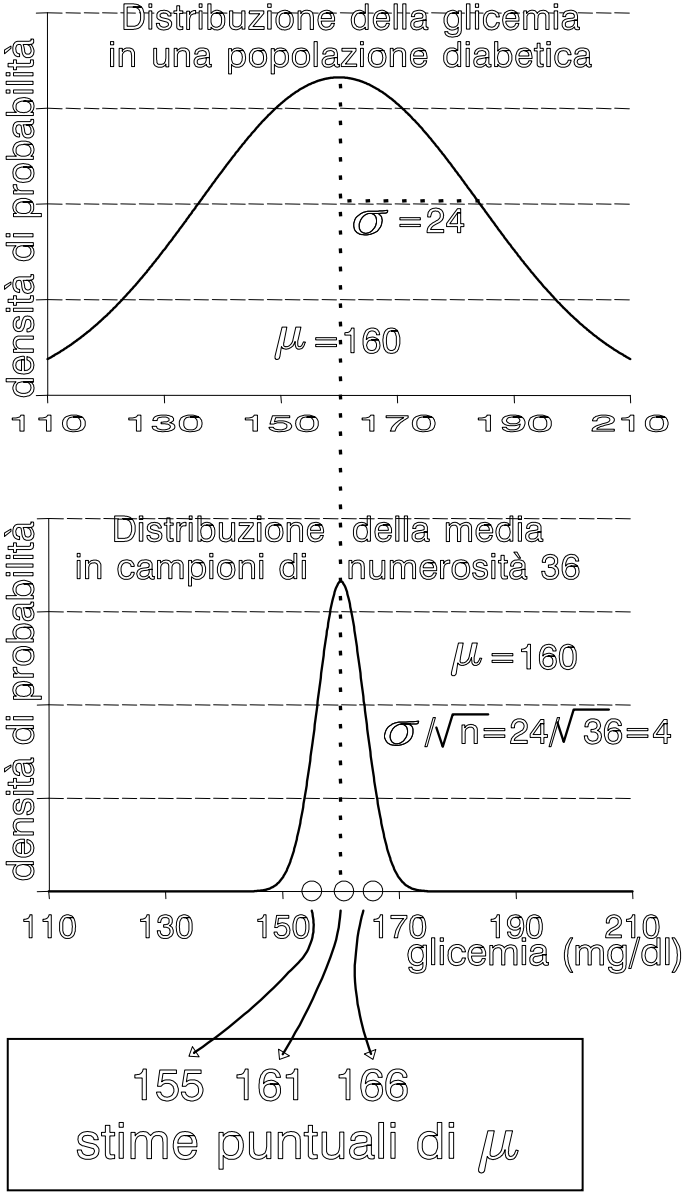
(LIMITE INFERIORE DELL'INTERVALLO)

L_s

(LIMITE SUPERIORE DELL'INTERVALLO)

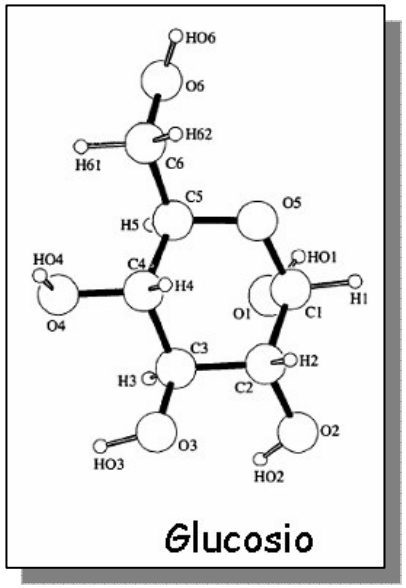
GLICEMIA nella POPOLAZIONE DIABETICA

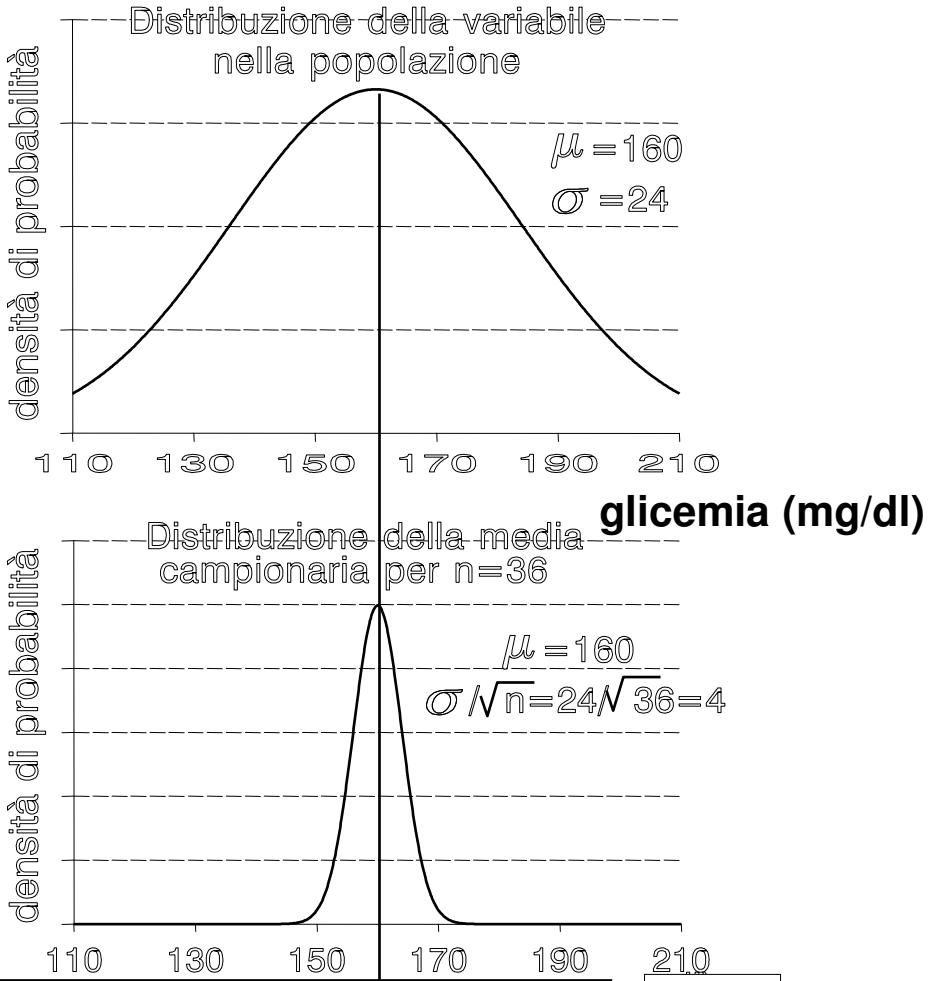
(stime puntuali)



GLICEMIA nella POPOLAZIONE DIABETICA

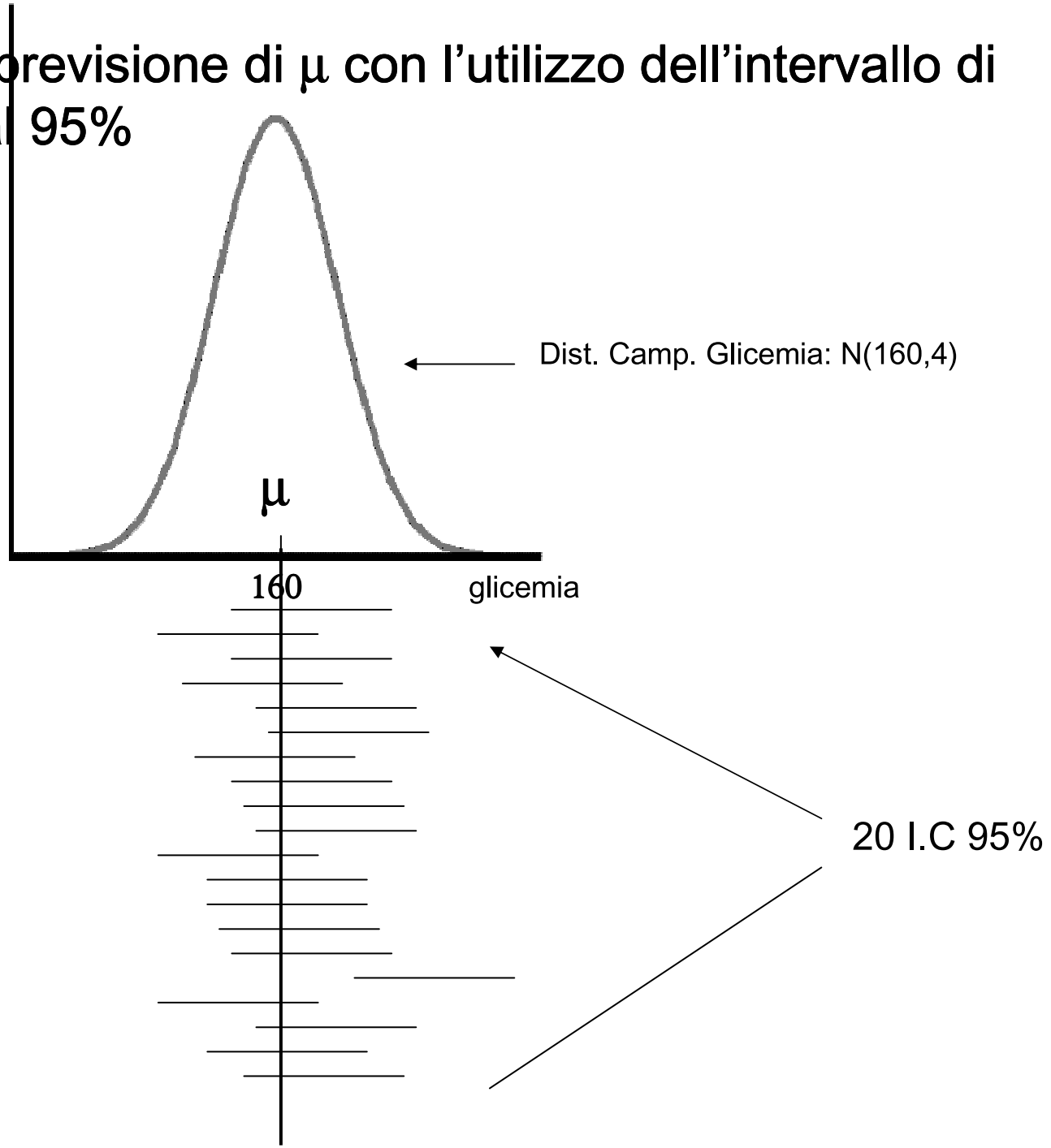
(stime intervallari)





stime intervallari di μ		
155±1,96*4	147,2	162,8
161±1,96*4	153,2	168,8
166±1,96*4	158,2	173,8

Errore nella previsione di μ con l'utilizzo dell'intervallo di confidenza al 95%



RIASSUMENDO...

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:

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 di una media al (1-a) è:

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

SESM

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):

$1-\alpha$	$\alpha/2$	$z_{\alpha/2}$
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)**

SESM

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 per piccoli campioni

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 $\bar{x}$ 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:

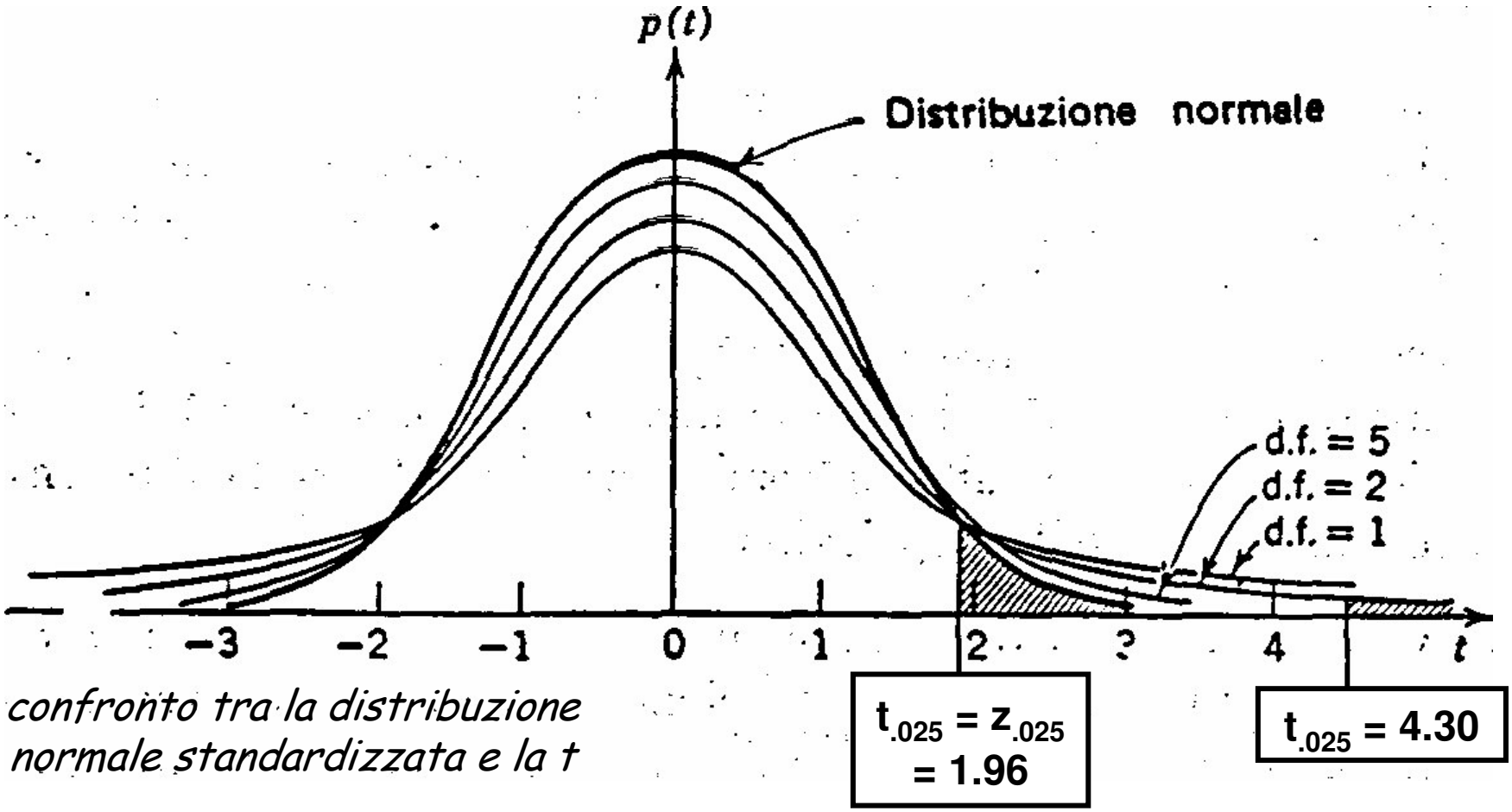
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

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:

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

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

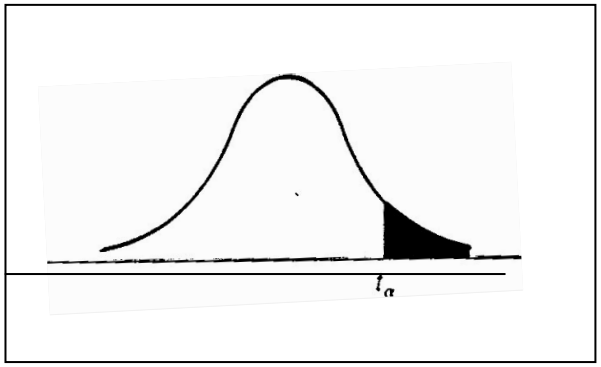


Tavola dei valori della funzione t di Student in funzione dei gradi di libertà e della probabilità in una coda della distribuzione

(.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

SESM

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:

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

INTERVALLO di CONFIDENZA UN CONTEGGIO POISSONIANO E DI UN TASSO

Sia C un conteggio che si suppone derivato da una Poisson, e sia R=C/T il tasso corrispondente. L'intervallo di confidenza approssimato di C e R sarà:

$$C \pm Z_{\alpha/2}\sqrt{C}$$

$$R \pm Z_{\alpha/2}\sqrt{\frac{R}{T}}$$

I precedenti valgono sotto l'assunzione che la normale sia una buona approssimazione della Poisson. Se l'assunzione non è Valida è necessario utilizzare metodi esatti.

SESM

esercizi

- Una macchina confezionatrice di bevande è programmata in modo tale che il contenuto di ogni bottiglia sia μ. Un campione di 100 bottiglie ha un contenuto medio di 48cl. Calcolate l'intervallo di confidenza al 95% e 99% (σ=5cl).
- Un agenzia sanitaria sostiene che la prevalenza di una determinata malattia nella città di Verona è del 5%. Su un campione casuale di 3000 residenti, 120 riportano la malattia in questione. Ritenete che le affermazioni dell'agenzia sanitaria siano confermate dall'indagine?
- La "scrapie" è un malattia degli ovini simile al morbo della mucca pazza. In una sperimentazione è stata utilizzata, su cavie, una sostanza per il trattamento della scrapie. In un gruppo di 10 cavie infettate e trattate con il farmaco sperimentale, il tempo necessario per la comparsa dei sintomi (tempo di induzione) era di 81.9 giorni (errore standard=2.3 giorni). 1- Calcolate l'intervallo di confidenza del tempo di induzione nelle cavie trattate. 2- Qual è la d.s. del tempo di induzione? 3- In un gruppo di 10 cavie infettate di controllo, il tempo medio di induzione era di 102.8 giorni con errore standard=3.8, Calcolate l'IC95% e commentate (Tagliavini 1997).