DISTRIBUZIONI CAMPIONARIE degli STIMATORI

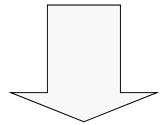
Una volta selezionato il campione, la variabile di interesse viene misurata sugli elementi che lo costituiscono.

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 ~ f(X)	$x_1, x_2, \ldots, x_n$
θ {μ, σ, π} (costanti)	$\hat{\theta}$ {x, s, p} (variabili casuali)
	f ($\hat{\theta}$) distribuzione campionaria degli stimatori

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

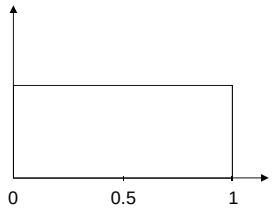
2) La distribuzione campionaria di $\bar{x}$ ha d.s. uguale alla d.s. della popolazione diviso la radice quadrata di n [errore standard - e.s]:

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

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.

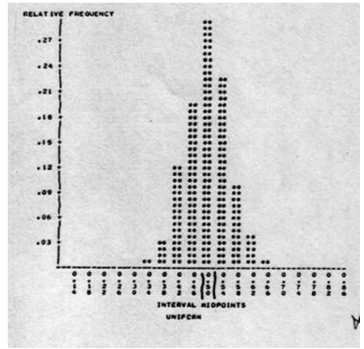
Distribuzione della variabile nella popolazione, f(X)



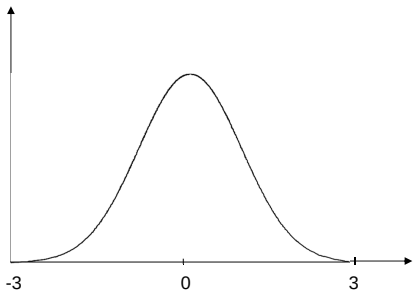
uniforme

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

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



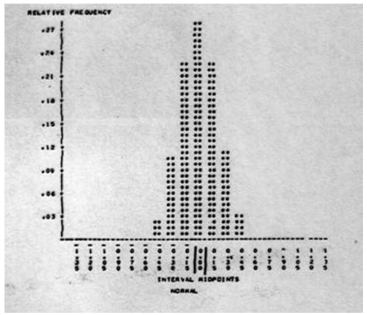
Distribuzione della variabile nella popolazione, f(X)



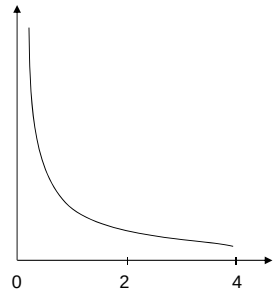
normale

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

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



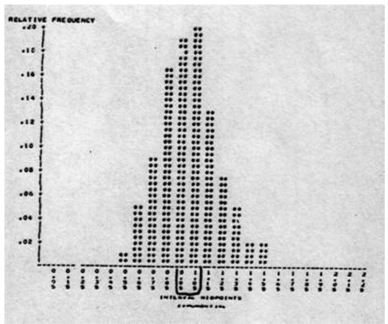
Distribuzione della variabile nella popolazione, f(X)



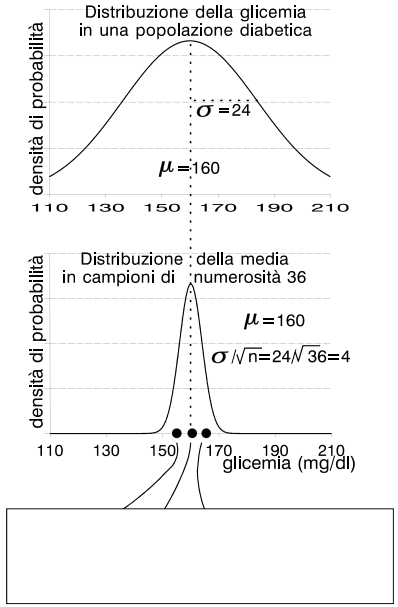
esponenziale

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

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:

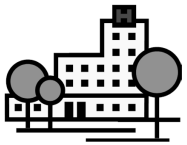
$$\bar{x} \sim N(38.3, 43.3/\sqrt{100})$$

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

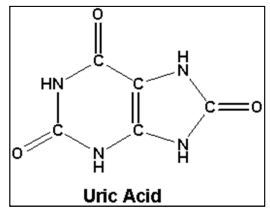
$$z = \frac{\bar{x} - E(x)}{d.s.(\bar{x})} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad z = \frac{46.9 - 38.3}{43.3/\sqrt{100}} = \frac{8.6}{4.3} = 2$$

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

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.

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!

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

L_i

(LIMITE INFERIORE DELL'INTERVALLO)

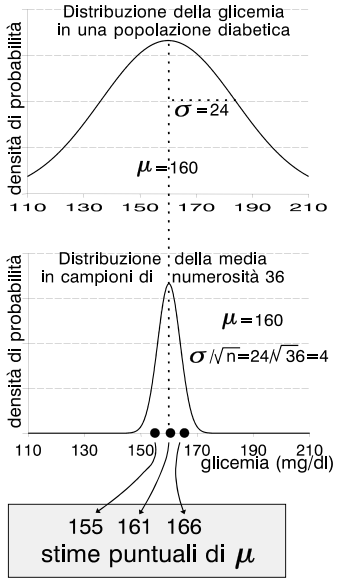
L_s

(LIMITE SUPERIORE DELL'INTERVALLO)

SESM

GLICEMIA nella POPOLAZIONE DIABETICA

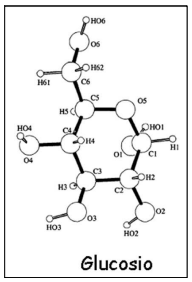
(stime puntuali)

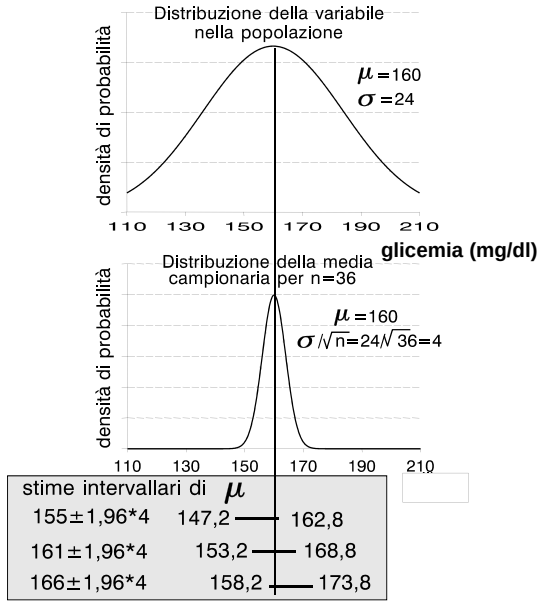


SESM

GLICEMIA nella POPOLAZIONE DIABETICA

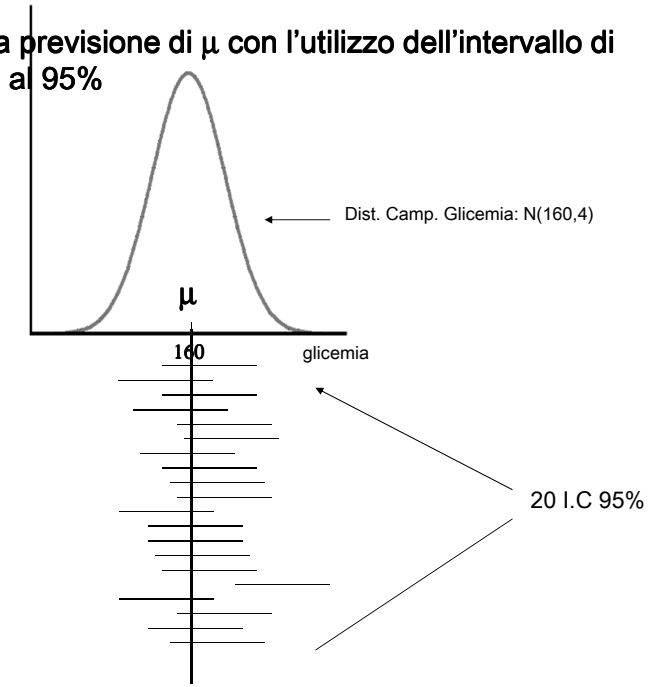
(stime intervallari)





SESM

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:

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

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)**

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:

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