LEZIONI DI STATISTICA MEDICA

Prof. SIMONE ACCORDINI

Lezione n.11

- Principi dell'inferenza statistica

- Campionamento

- Distribuzione campionaria di una media e di una proporzione

- Intervallo di confidenza di una media e di una proporzione

Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

STATISTICA DESCRITTIVA

Metodi per la descrizione e la sintesi dei valori di una variabile misurati in un campione

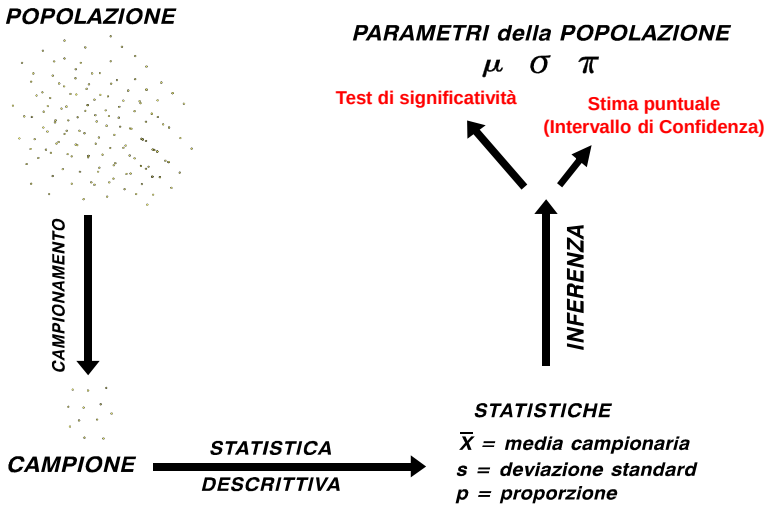
MODELLI PROBABILISTICI

Modelli che permettono di descrivere la distribuzione di una V.C. nella popolazione mediante pochi parametri

INFERENZA STATISTICA

Studio delle caratteristiche della popolazione (parametri) sulla base delle informazioni raccolte in un campione

INFERENZA



METODI STATISTICI DELL'INFERENZA

1. Stimare il parametro di interesse (μ, σ, π) in una o più popolazioni

⇒ **STIMA PUNTUALE**

2. Associare alla stima puntuale ($\bar{x}$, s, p) una **misura di precisione**
→ **misura dell'errore di stima**

⇒ **INTERVALLO DI CONFIDENZA**

METODI STATISTICI DELL'INFERENZA

3. Verificare se il parametro di interesse (μ, σ, π) in una popolazione ha un valore diverso da quello ipotizzato

[ad esempio: la probabilità di avere l'asma per un adulto in Italia (π) è $\neq 0.02$?] …

… verificare se il parametro di interesse varia tra due o più popolazioni:

$\mu_1 \neq \mu_0$, $\sigma_1 \neq \sigma_0$, $\pi_1 \neq \pi_0$

→ **la differenza osservata è dovuta al caso oppure è dovuta ad altri fattori (trattamento, fattori di rischio, …)?**

⇒ **TEST STATISTICO**

UTILIZZO DEL CAMPIONE

VANTAGGI

- riduzione dei costi dell'indagine
- maggiore accuratezza nella raccolta delle informazioni
- unica possibilità quando la popolazione in studio è infinita

SVANTAGGI

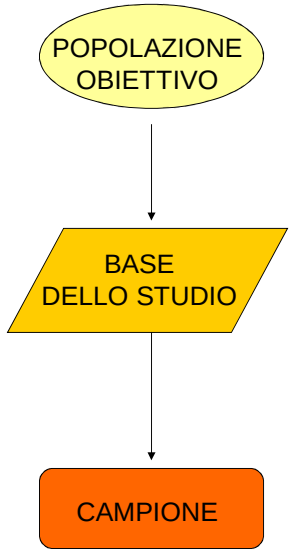
- imprecisione delle stime

  → ERRORI CAMPIONARI: le stime campionarie variano da campione a campione

- possibile distorsione delle stime

  → ERRORI SISTEMATICI (BIAS): errori legati alla non rappresentatività del campione ottenuto dalla procedura di campionamento

Schema di una procedura di campionamento



Dominio su cui si vuole fare inferenza:

1) pop. adulta di una città in un determinato periodo (finita)
2) pop. degli ipertesi (infinita)

Particolare esperienza concreta, delimitata nello spazio e nel tempo, utilizzata come elemento di conoscenza della popolazione obiettivo:

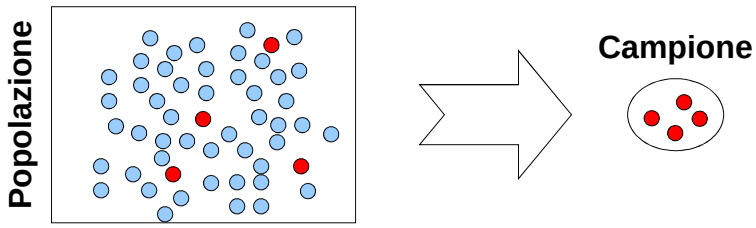
1) lista anagrafica
2) tutti i soggetti ipertesi che si rivolgono al loro medico per disturbi ipertensivi in una determinata area e in un determinato periodo

Qualsiasi sottoinsieme della base dello studio, ma ...

… per l'inferenza, è necessario che sia rappresentativo della base dello studio

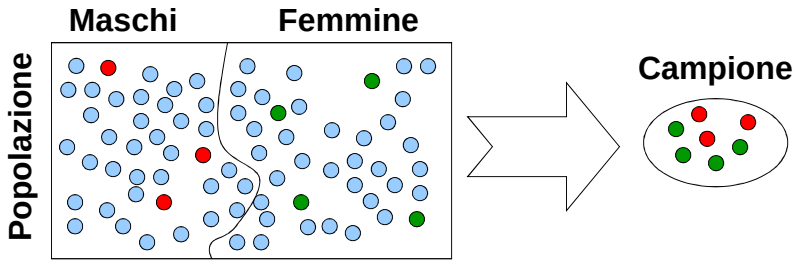
⇒ SELEZIONE CASUALE

Campionamento casuale semplice



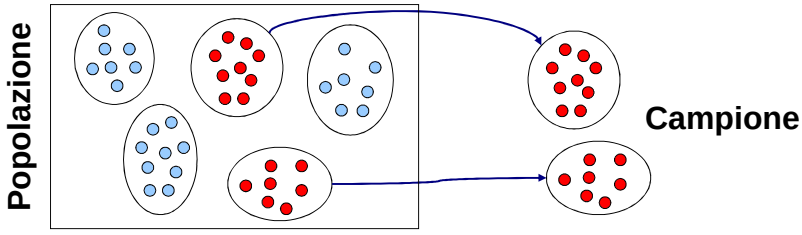
→ tutte le unità statistiche della popolazione hanno uguale probabilità di selezione

Campionamento stratificato



→ gruppi sufficientemente omogenei al loro interno ma diversi tra loro

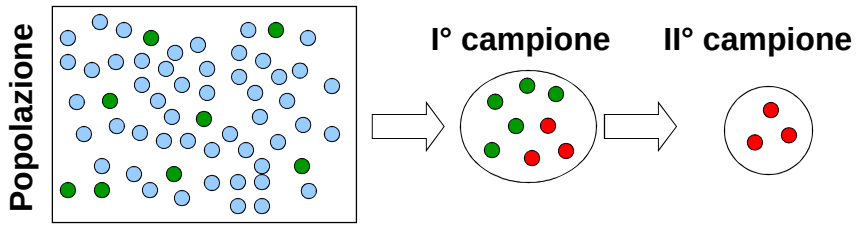
Campionamento a grappolo



→ gruppi eterogenei al loro interno ma omogenei tra loro

Esempio: gruppi = ospedali di una determinata area
unità elementari = tutti i pazienti che hanno subito un certo intervento

Campionamento a due o più fasi

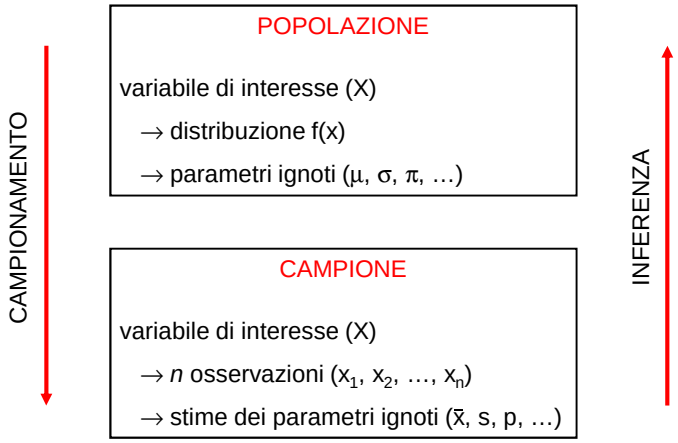


PROCEDURA:

- estrazione del **I° campione** → **indagini meno approfondite**
- estrazione successiva del **II° campione** dal I° campione → **indagini più approfondite**

Esempio: l'indagine ECRHS è stata condotta in due fasi:

- somministrazione di un questionario postale di screening
- fase clinica effettuata su un campione dei rispondenti allo screening



Le statistiche (media, dev. std, proporzione, …) ottenute in un campione sono STIME DEI PARAMETRI IGNOTI (μ, σ, π, ...) della popolazione di interesse

⇒ il valore delle statistiche dipende dal particolare campione selezionato

DISTRIBUZIONE CAMPIONARIA DEGLI STIMATORI

Ripetendo idealmente più volte la procedura di campionamento nelle medesime condizioni si potrebbe definire una distribuzione di frequenza delle statistiche

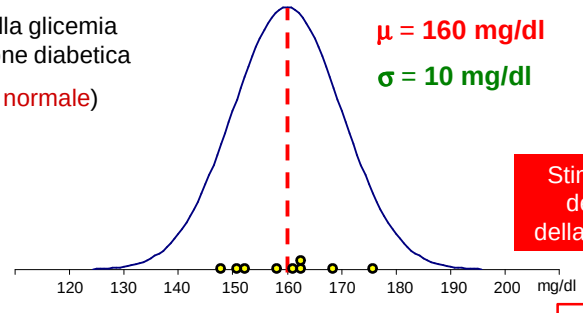
⇒ le statistiche campionarie sono V.C. (STIMATORI) caratterizzate da una specifica distribuzione di probabilità

DEF: La distribuzione campionaria di una statistica, basata su *n* osservazioni, è la distribuzione di frequenza dei valori che assume la statistica, generata teoricamente prendendo infiniti campioni di dimensione *n* nelle stesse identiche condizioni e calcolando il valore della statistica per ogni campione

INFERENZA SULLA MEDIA DI UNA POPOLAZIONE

Distribuzione della glicemia in una popolazione diabetica

(modello teorico normale)



Stima puntuale della media della popolazione

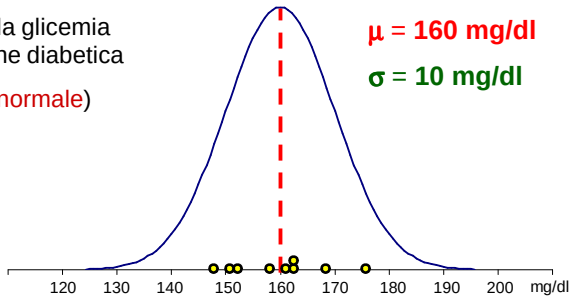
campione: **162 152 158 162 168 161 148 176 150** → $\bar{x}$ = **159.7**

Distribuzione della glicemia
in una popolazione diabetica

(modello teorico normale)

μ = 160 mg/dl

σ = 10 mg/dl



Campione studiato

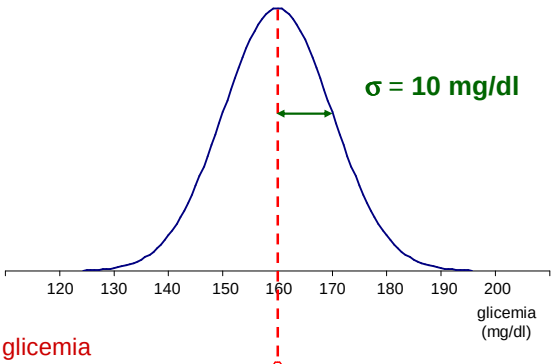
I° campione:	162 152 158 162 168 161 148 176 150	$\bar{x}$ = 159.7
II° campione:	152 164 157 180 156 163 165 166 178	$\bar{x}$ = 164.6
III° campione:	157 142 163 162 152 149 152 180 151	$\bar{x}$ = 156.4
IV° campione:	162 154 168 160 155 172 162 152 140	$\bar{x}$ = 158.3
V° campione:	163 169 152 147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

PRINCIPIO DEL CAMPIONAMENTO RIPETUTO

Distribuzione della glicemia
in una popolazione diabetica

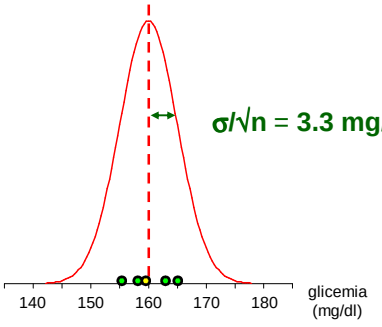
μ = 160 mg/dl

σ = 10 mg/dl



Distribuzione della media della glicemia
in campioni di numerosità = 9

$\sigma/\sqrt{n}$ = 3.3 mg/dl



DISTRIBUZIONE CAMPIONARIA DI UNA MEDIA

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

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

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

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

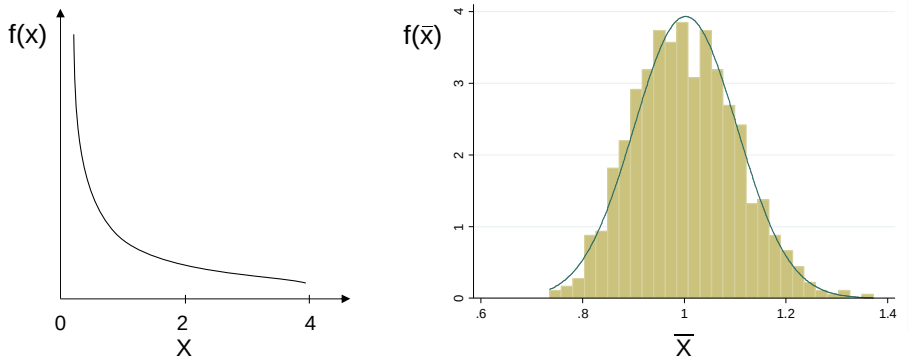
ERRORE STANDARD della media $\rightarrow$ misura della precisione della stima

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

Esempio (teorema del limite centrale - media):

800 campioni di dimensione $n = 100$ generati casualmente da una distribuzione esponenziale

DISTRIBUZIONE CAMPIONARIA



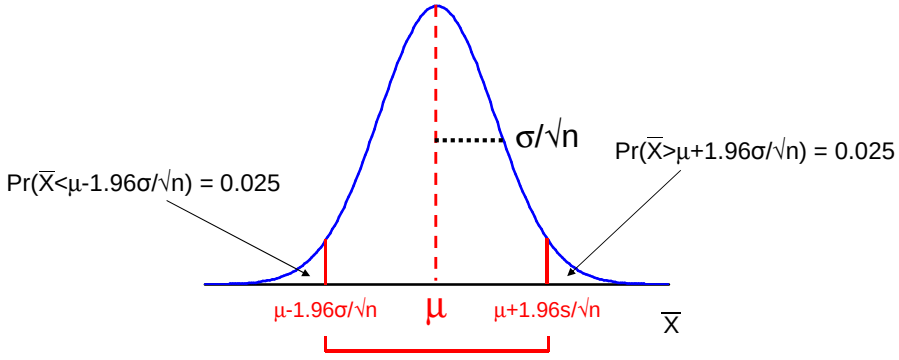
Intervallo di confidenza della media in una popolazione: IC95%(μ)

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:

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

L'intervallo simmetrico centrato sulla vera media (μ) che comprende il 95% delle medie campionarie è:

$$\Pr\left\{-1.96 \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right\} = \Pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

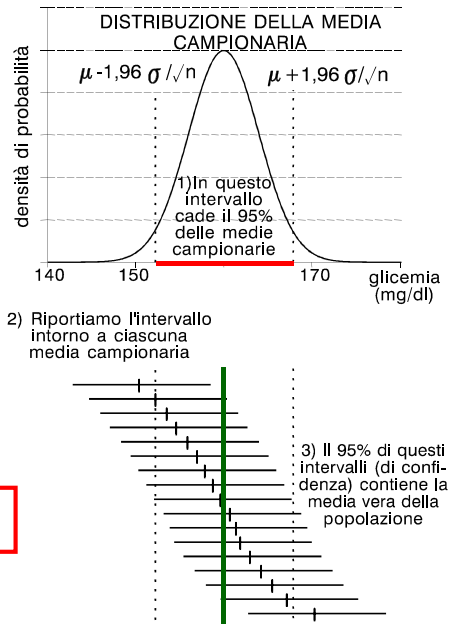


$$\Pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

riarrangiando le due disuguaglianze interne alla parentesi:

$$\Pr\left\{\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

IC95%(μ)



IC95%(μ)

$\overline{x} - 1.96\,\sigma/\sqrt{n}$, $\overline{x} + 1.96\,\sigma/\sqrt{n}$

L_{inf} L_{sup}

Esempio: Inferenza sulla media della glicemia in una popolazione diabetica

1. Stimare il parametro di occorrenza (μ)

⇒ **STIMA PUNTUALE** (*n* = 9) **$\bar{x}$ = 159.7** mg/dl

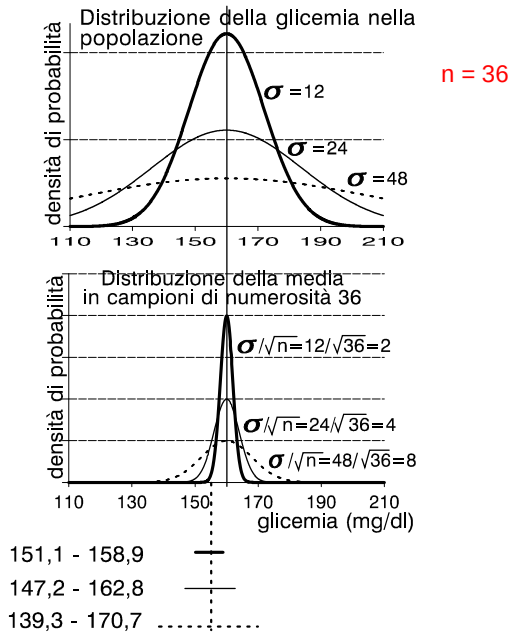
2. Associare alla stima puntuale una **misura di precisione**

⇒ **INTERVALLO DI CONFIDENZA**

IC95%(μ) = 159.7 ± 1.96*10/√9 = [153.2 mg/dl**, 166.2** mg/dl]

L'IC diminuisce se diminuisce la variabilità nella popolazione (σ)

IC95%(μ) = $\bar{x} \pm 1.96\ \sigma/\sqrt{n}$

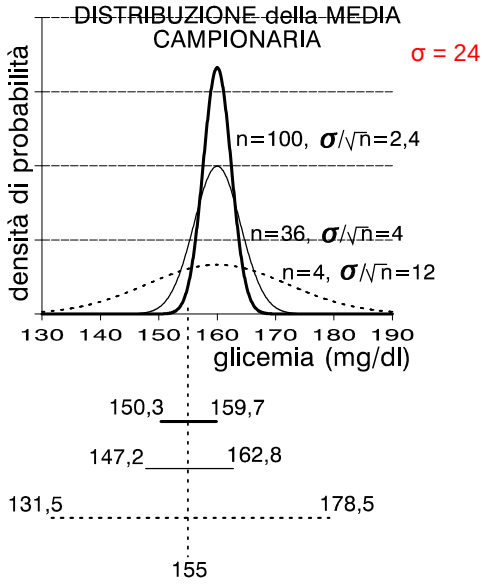


n = 36

151,1 - 158,9
147,2 - 162,8
139,3 - 170,7

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

IC95%(μ) = $\bar{x} \pm 1.96\ \sigma/\sqrt{n}$



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

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \Rightarrow \quad Z_{0.025} = 1.96$$

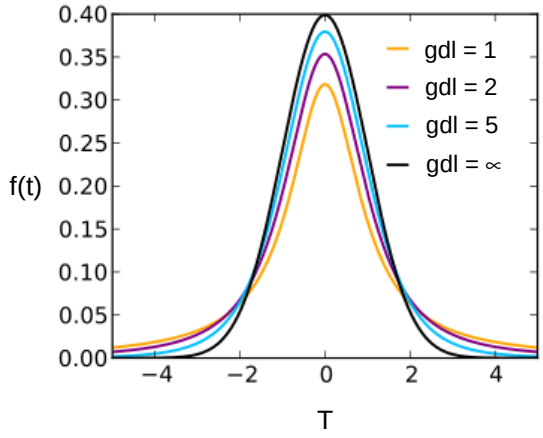
Molto spesso **σ è ignota**

→ si utilizza la deviazione standard campionaria S (stima di σ)

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

distribuzione t di Student con (n-1) gradi di libertà

Famiglia di distribuzioni simmetriche
che dipendono dai gradi di libertà



gdl	$t_{gdl,0.025}$
1	12.706
2	4.303
3	3.182
4	2.776
5	2.571
6	2.447
7	2.365
8	2.306
9	2.262
10	2.228
11	2.201
12	2.179
13	2.160
14	2.145
15	2.131
16	2.120
17	2.110
18	2.101
19	2.093
20	2.086
21	2.080
22	2.074
23	2.069
24	2.064
25	2.060
26	2.056
27	2.052
28	2.048
29	2.045
30	2.042
40	2.021
50	2.009
60	2.000
100	1.984
∞	1.960

La **stima puntuale** di un parametro fornisce un singolo valore:

- è una determinazione di una V.C.
- il valore campionario non coincide quasi mai con il vero valore (ignoto) del parametro nella popolazione
- campioni diversi forniscono stime puntuali diverse

La **stima intervallare** di un parametro fornisce un intervallo di valori:

- i limiti di confidenza sono una determinazione di una V.C.
- l'intervallo di confidenza ha una prefissata probabilità (95%) di contenere il vero valore (ignoto) del parametro nella popolazione
- il metodo per il calcolo dell'intervallo di confidenza di una media è:

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \qquad \bar{x} \pm t_{n-1,0.025} \cdot \frac{s}{\sqrt{n}}$$

σ nota σ ignota

Esercizio (intervallo di confidenza di una media):

Qual è l'intervallo di confidenza al 95% della media del peso in una certa popolazione, se la media in un campione di 16 soggetti è pari a 75 Kg? Nella popolazione il peso è distribuito normalmente con deviazione standard pari a 12 Kg (parametro).

$$IC95\%(\mu) = \bar{x} \pm z_{0.025}\, ES[\bar{X}] = 75 \pm 1.96 * (12/\sqrt{16}) = [69.12\ Kg, 80.88\ Kg]$$

Supponiamo che la deviazione standard nella popolazione non sia nota e che la deviazione standard nel campione di 16 soggetti sia pari a 12 Kg (stima).

$$IC95\%(\mu) = \bar{x} \pm t_{15,0.025}\, ES[\bar{X}] = 75 \pm 2.131 * (12/\sqrt{16}) = [68.61\ Kg, 81.39\ Kg]$$

DISTRIBUZIONE CAMPIONARIA DI UNA PROPORZIONE

Sia X una variabile bernoulliana:

$$\begin{cases} X = 1 \text{ (successo)} \rightarrow Pr(X = 1) = \pi \\ X = 0 \text{ (insuccesso)} \rightarrow Pr(X = 0) = 1 - \pi \end{cases}$$

$E[X] = \pi$

$VAR[X] = \pi(1 - \pi)$

Sia $P = \Sigma X/n$ la proporzione di successi in un campione di osservazioni indipendenti di dimensione *n*:

1) $E[P] = \pi$

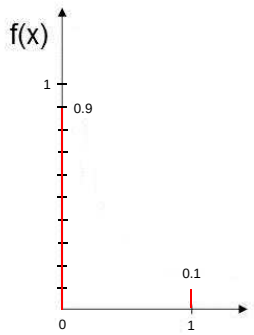
2) $DS[P] = ES[P] = \sqrt{\pi(1-\pi)/n}$ (ERRORE STANDARD della proporzione)

3) se $np>5$ e $n(1-p)>5$, allora la distribuzione campionaria di P è approssimativamente normale:

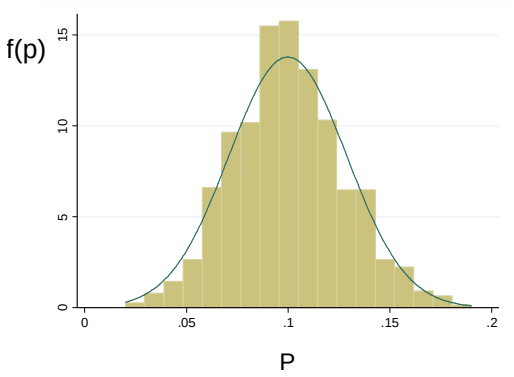
$$P \sim N(\pi, \sqrt{\pi(1-\pi)/n})$$

Esempio (teorema del limite centrale - proporzione):

800 campioni di dimensione *n* = 100 generati casualmente
da una distribuzione bernoulliana con π = 0.1



DISTRIBUZIONE CAMPIONARIA



Intervallo di confidenza della proporzione in una popolazione: IC95%(π)

Se *np>5* e *n(1-p)>5*, allora la distribuzione campionaria di P ha distribuzione approssimativamente normale: $P \sim N(\pi, \sqrt{\pi(1-\pi)/n})$

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

- la proporzione campionaria p è una stima di π
- l'intervallo di confidenza di una proporzione è

$$IC95\%(\pi) = p \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

Esercizio (intervallo di confidenza di una proporzione):

Un agenzia sanitaria sostiene che la proporzione di soggetti che presentano una determinata malattia nella città di Verona sia pari a 0.05. Su un campione casuale di 3000 residenti, 120 riportano la malattia in questione.

Ritenete che le affermazioni dell'agenzia sanitaria siano confermate dall'indagine?

p = 120/3000 = 0.04

$IC95\%(\pi) = p \pm z_{0.025}\ ES[P] = 0.04 \pm 1.96 * \sqrt{0.04*(1-0.04)/3000} = [0.033, 0.047]$

L'indagine non conferma le affermazioni dell'agenzia sanitaria.