STATISTICA DESCRITTIVA

Metodi per la descrizione e sintesi di un insieme di osservazioni su un campione

METODI E MODELLI PROBABILISTICI

Modelli che permettono di descrivere mediante pochi parametri la distribuzione di una variabile casuale nella popolazione

INFERENZA STATISTICA

INFERENZA STATISTICA

STUDIO DELLE RELAZIONI TRA CAMPIONE E POPOLAZIONE

possibilità, sulla base dei risultati ottenuti su un campione, di fare delle affermazioni sulla popolazione

Nella ricerca medica il CAMPIONE (l'esperienza particolare che viene considerata in uno studio) è un mezzo per apprendere e/o approfondire una relazione o un fenomeno che si vuole generalizzare a una POPOLAZIONE

La popolazione il più delle volte è puramente astratta, non limitata nè nello spazio nè nel tempo (universo)

POPOLAZIONE o UNIVERSO

μ, σ, π **parametri**

$y=\alpha+\beta x$

TEORIA DEL CAMPIONAMENTO
Quali soggetti selezionare?

STATISTICA INFERENZIALE
Cosa possiamo dire dei veri parametri della popolazione?
Qual è il margine d'incertezza?

CAMPIONE

$x_1, x_2, \ldots, x_n$

$y_1, y_2, \ldots, y_n$

STATISTICA DESCRITTIVA

$\bar{x}$, s, p
STATISTICHE
$Y=a+bx$

CENNI di TEORIA del CAMPIONAMENTO

Molte ricerche vengono programmate con lo scopo di pervenire a **conclusioni generali**, valide per tutte le unità statistiche della popolazione, sfruttando i risultati ottenuti da un numero ridotto di osservazioni

La teoria del campionamento concerne le modalità di selezione del CAMPIONE dalla popolazione, al fine di rendere possibile la generalizzazione dei risultati.

UTILIZZO del CAMPIONE

VANTAGGI:

1. risparmio di lavoro e di costi dell'indagine perché vengono ridotte le unità di osservazione
2. la raccolta dell'informazione può essere più attendibile e più accurata
3. unica possibilità quando la popolazione su cui si vogliono fare inferenze è infinita.

SVANTAGGI:

1. imprecisione delle stime; le misure calcolate sono solo una approssimazione delle vere misure della popolazione e variano da campione a campione.

L'utilizzo del campione introduce delle *fonti di errore* nella stima dei parametri incogniti della popolazione:

errori sistematici

vizi o bias legati alla **non rappresentatività** del campione prodotto dalla procedura di campionamento: le stime si allontanano in modo sistematico dal parametro della popolazione

errori campionari

intrinseci alla procedura di campionamento; **influenzano la precisione della stima**. La dimensione dell'errore può essere predetta in base alla teoria della probabilità

Il metodo migliore per la scelta di un campione è selezionare i soggetti con un metodo completamente casuale (**randomizzazione**) che assicuri a ciascun campione di una data dimensione la stessa probabilità di essere selezionato.

Un campione che soddisfa al precedente requisito prende il nome di **campione casuale semplice**.

DISTRIBUZIONI CAMPIONARIE degli STIMATORI

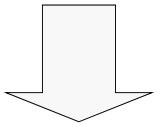
Una volta selezionato il campione, la variabile di interesse viene misurata sugli elementi che lo costituiscono.

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

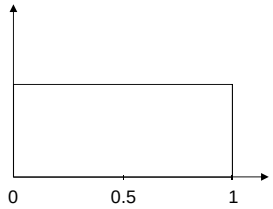
2) La distribuzione campionaria di $\bar{x}$ ha d.s. uguale alla d.s. della popolazione diviso la radice quadrata di n [errore standard - e.s]:

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

3) **TEOREMA CENTRALE DEL LIMITE**

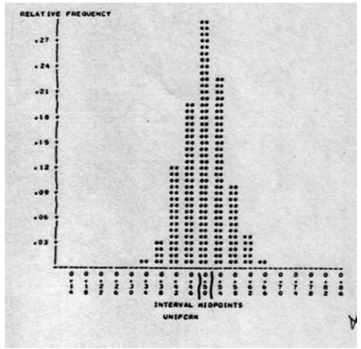
Se la dimensione campionaria è sufficientemente grande (n > 30) la distribuzione campionaria di $\bar{x}$ è approssimativamente **normale, i**ndipendentemente dalla forma della distribuzione della variabile nella popolazione.

Distribuzione della variabile nella popolazione, f(X)

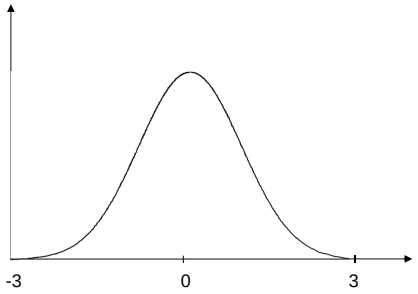


uniforme
($\mu = 0.5$, $\sigma = 0.29$)

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

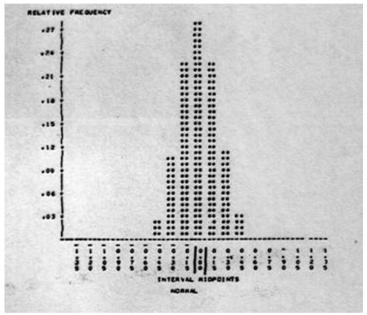


Distribuzione della variabile nella popolazione, f(X)

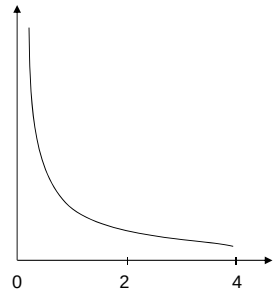


normale
($\mu = 0$, $\sigma = 1$)

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

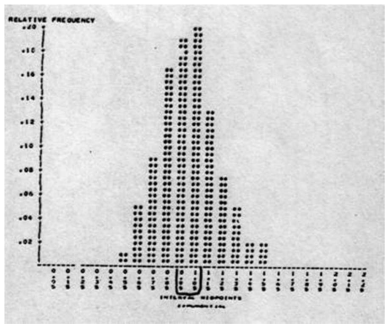


Distribuzione della variabile nella popolazione, f(X)

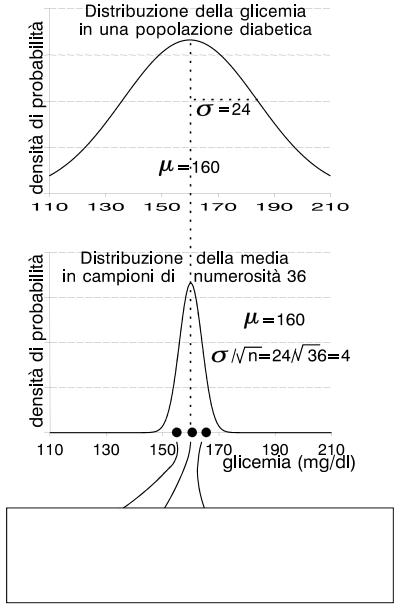


esponenziale
($\mu = 1$, $\sigma = 1$)

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 media ≥ 46.9 mesi?

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

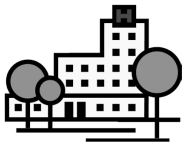
per il teorema del limite centrale:

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

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

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

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

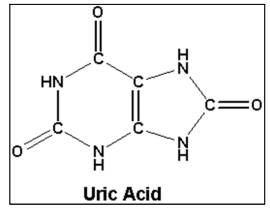


pr = 2.3%

Esercizio
I perimetri toracici della popolazione maschile italiana, di età compresa tra i 18 e i 74 anni, si distribuiscono normalmente con media = 75cm e scarto quadratico medio (deviazione standard) = 19 cm.

Determinare la probabilità che il parametro toracico medio calcolato in un campione casuale di numerosità n =100 superi i 79.75 cm.

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 le medie del 95% dei campioni di 30 soggetti.

DISTRIBUZIONE CAMPIONARIA di una PROPORZIONE

Sia X una **variabile bernoulliana** (X=1 ⇒ successo; X=0 ⇒ insuccesso) definita nella popolazione con media = π e varianza = π(1- π).

Sia p la <u>percentuale di successi</u> in un campione di dimensione n.

1. La distribuzione campionaria di p ha la media uguale alla media della popolazione da cui proviene il campione:

$$E(p) = \pi$$

2. La distribuzione campionaria di p ha d.s.:

$$d.s.(p) = \sqrt{\frac{\pi(1-\pi)}{n}} = E.S.$$

3. Se la dimensione campionaria è sufficientemente grande (n > 30) la distribuzione campionaria di P è approssimativa-mente **normale.**

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

Esempio:
E' noto che circa il 26% dei bambini nati da madri sieropositive per l'HIV risultano sieropositivi per l'HIV alla nascita o poco dopo la nascita.

Qual è la probabilità che in campioni casuali di 150 bambini più di 56 bambini siano sieropositivi?