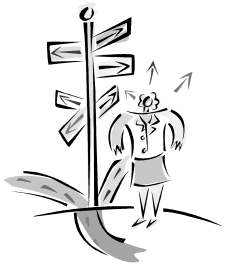
LEZIONI DI STATISTICA MEDICA

L'inferenza statistica



Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

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

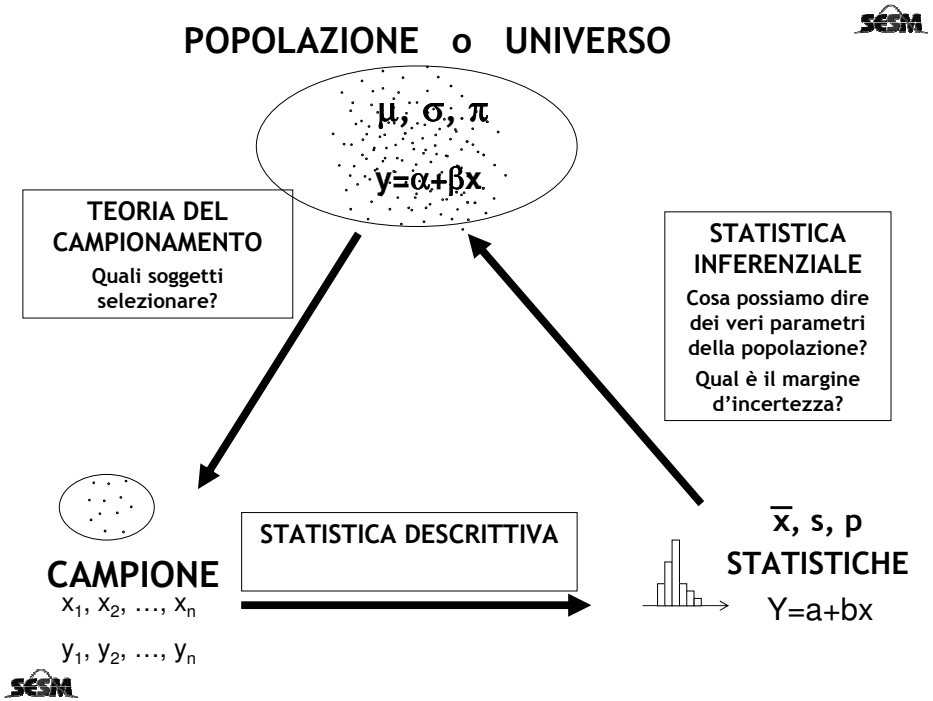
INFERENZA STATISTICA

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)



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:

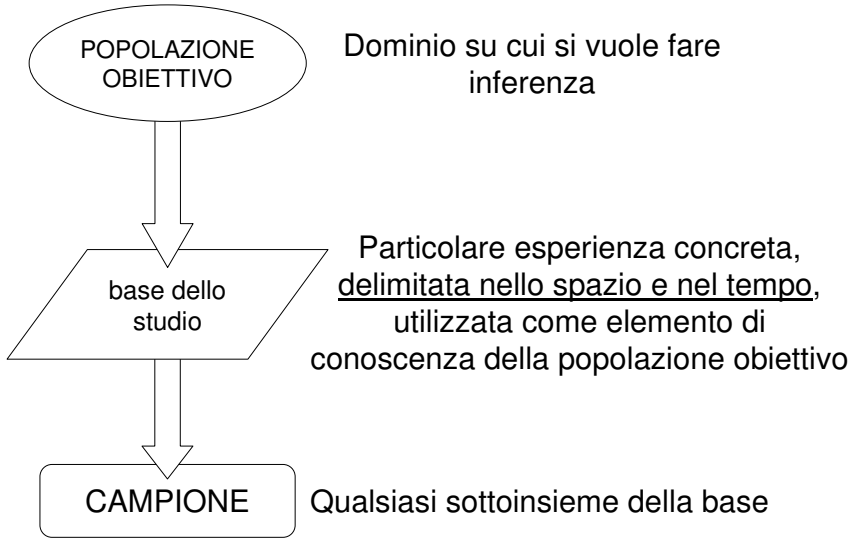
errori sistematici

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

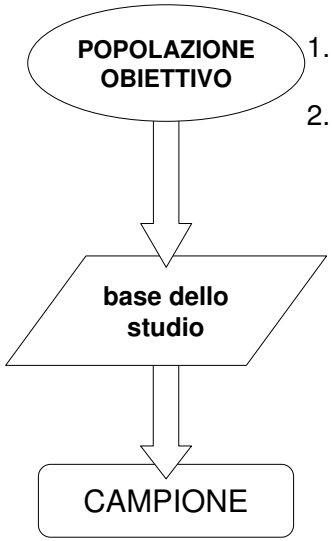
errori campionari

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

SCHEMA della PROCEDURA di CAMPIONAMENTO



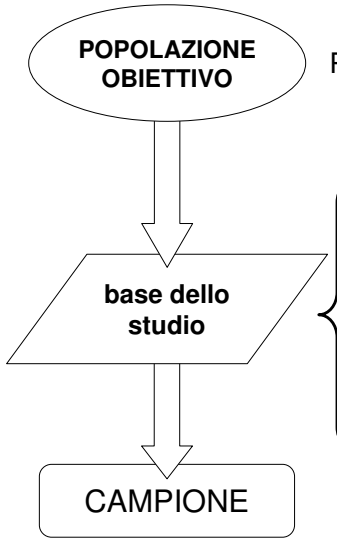
esempi:



1. Pop. degli studenti universitari di VR nel 2004 (finita)
2. Pop. adulta di una città (finita)

1. Elenco degli iscritti.
2. Registro aggiornato dei certificati elettorali; Elenco telefonico.

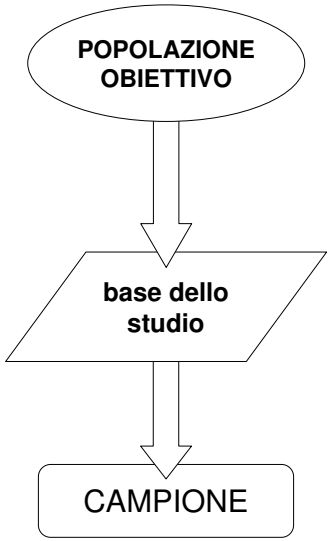
esempi:



Pop. degli ipertesi (infinita)

- Tutti i soggetti ipertesi che si rivolgono al loro medico a seguito di disturbi legati alla malattia in una data area e tempo;
- Tutti i soggetti di una data area che hanno valori di pressione arteriosa superiori a un certo cut-off;
- Tutti i soggetti ricoverati con diagnosi di ipertensione.

esempi:



Dominio della relaz. tra fumo e carcinoma polmonare

Opportuno insieme in cui sono rappresentate alcune categorie di fumo (alto, medio, basso, nullo) simili tra di loro eccetto che per l'abitudine al fumo

SCELTA della BASE dello STUDIO

La **base di uno studio** è scelta con criteri logici in funzione della sua idoneità rispetto alla popolazione obiettivo.

In genere:

1. nelle **indagini campionarie o di prevalenza**, mirate alla stima dei parametri della popolazione, deve essere garantita la rappresentatività della popolazione obiettivo.

2. nelle **indagini etiologiche,** mirate allo studio dei fattori responsabili dell'insorgenza di specifiche patologie, deve essere garantita la confrontabilità dei gruppi che hanno esposizioni (fattori di rischio) differenti.

indagini campionarie: esempio 1

In un'indagine mirata a valutare i consumi alimentari di una provincia del meridione le possibili basi potrebbero essere:

- elenco telefonico provinciale
- elenco dei certificati elettorali
- elenco dei certificati di residenza

SESM

indagini campionarie: esempio 2

- In un indagine mirata a valutare le complicazioni in bambini tra 3 e 5 anni affetti da morbillo :

- tutti i bambini ricoverati in ospedale con diagnosi di morbillo
- tutti i bambini iscritti alle scuole materne della zona
- elenco dei pediatri della zona

SESM

N.B.: Se la base non è rappresentativa della popolazione obiettivo le stime ottenute sono sistematicamente errate!

Anno 1936, U.S.A.: ELEZIONI PRESIDENZIALI

Candidati: **Roosevelt** e **Landon**

Literary Digest condusse un'indagine campionaria per predire i risultati delle elezioni.

Popolazione obiettivo: tutti i votanti degli Stati Uniti.

Base: liste riportate negli elenchi telefonici.

SESM

L'indagine predisse una vittoria globale di **Landon**

Roosevelt vinse con il più largo margine mai raggiunto in un'elezione presidenziale fino a quel tempo.

SESM

Nel 1936 vi era un gran numero di persone, per la maggior parte sostenitrici di Roosevelt, che non possedeva un telefono.

La base da cui è stato scelto il campione era "biased" (viziata). Il ceto sociale più elevato era sovrarappresentato nella base scelta.

N.B.: la grandezza del campione è relativamente priva di rilevanza nel compensare gli effetti dei vizi di selezione avvenuti nel campionamento **(il campione era costituito da oltre un milione di risposte!)**

SCELTA DEL CAMPIONE

Il campione è un qualsiasi sottoinsieme della base.

Perché il campione possa essere utilizzato per fare "inferenze" sulla popolazione da cui è stato estratto è necessario che esso sia **rappresentativo della base** e non sia frutto di una **selezione** cosciente o incosciente effettuata dal ricercatore (equazione personale di errore).

Es.: scelta di un campione di studenti per la misura del Q.I.

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

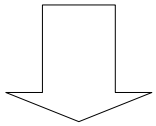
Una volta selezionato il campione, la variabile di interesse viene misurata sugli elementi che lo costituiscono.

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**
	$x_1, x_2, \ldots, x_n$
$X \sim f(X)$	$\hat{\theta}$ {x, s, p} (variabili casuali)
θ {μ, σ, π} (costanti)	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

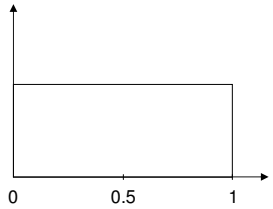
2) La distribuzione campionaria di $\bar{x}$ ha d.s. uguale alla d.s. della popolazione diviso la radice quadrata di n [errore standard - e.s]:

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

3) **TEOREMA CENTRALE DEL LIMITE**

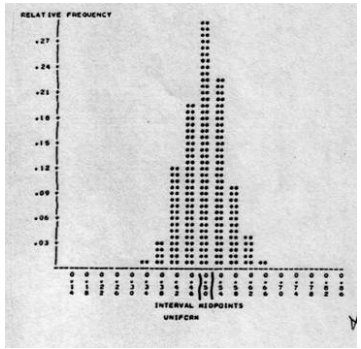
Se la dimensione campionaria è sufficientemente grande (n > 30) la distribuzione campionaria di $\bar{x}$ è approssimativamente **normale, i**ndipendentemente dalla forma della distribuzione della variabile nella popolazione.

Distribuzione della variabile nella popolazione, f(X)

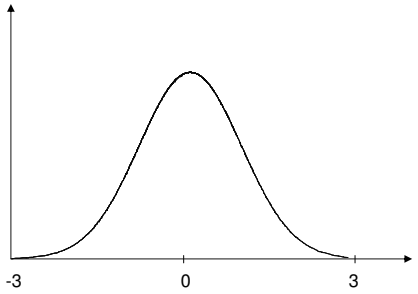


uniforme
($\mu = 0.5$, $\sigma = 0.29$)

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

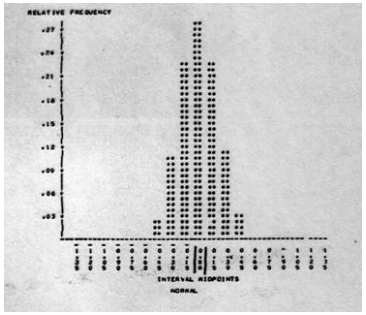


Distribuzione della variabile nella popolazione, f(X)

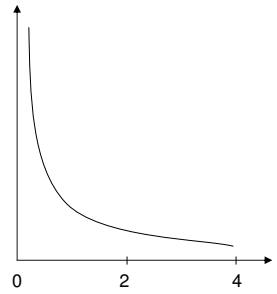


normale
($\mu = 0$, $\sigma = 1$)

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

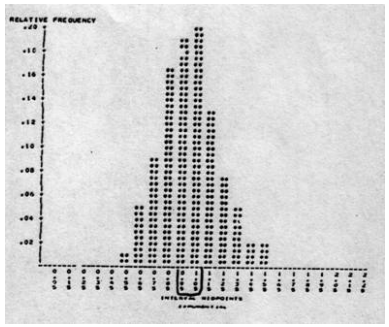


Distribuzione della variabile nella popolazione, f(X)

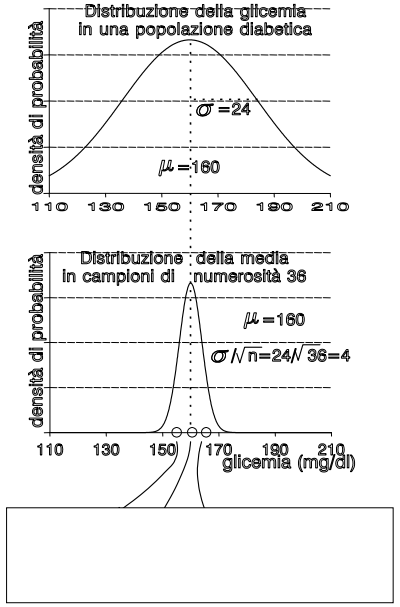


esponenziale
($\mu = 1$, $\sigma = 1$)

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?

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

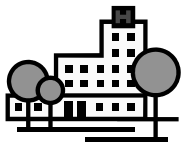
per il teorema del limite centrale:

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

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

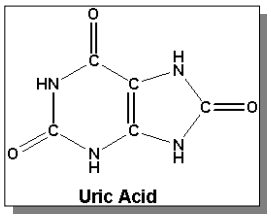
$$z = \frac{\bar{x} - E(x)}{d.s.(\bar{x})} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad z = \frac{46.9 - 38.3}{43.3/\sqrt{100}} = \frac{8.6}{4.3} = 2$$

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



pr = 2.3%

ESERCIZIO:
Sapendo che nella popolazione maschile l'acido urico serico è distribuito **normalmente** 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 percentuale di successi 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.**

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