

Statistics

Francesco Pugliese, PhD

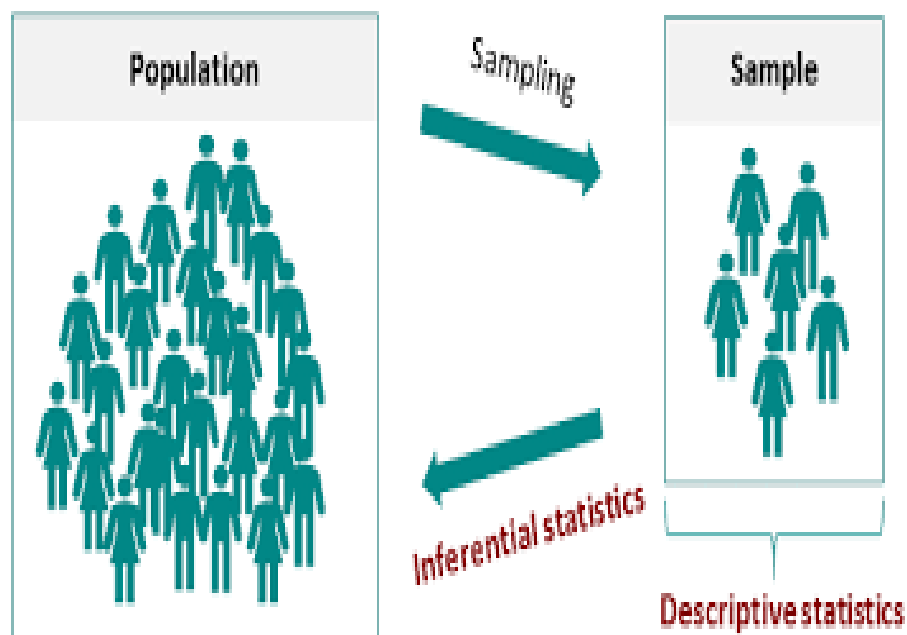
neural1977@gmail.com

Campo di analisi della statistica

- ✓ **Definizione di Statistica:** La **statistica** è una scienza che per oggetto l'acquisizione, l'elaborazione e la valutazione **qualitativa** e **quantitativa** dei dati riguardanti fenomeni di massa suscettibili alla misurazione. Nell'ambito della **statistica** si distinguono due settori: la **statistica descrittiva** e la **statistica inferenziale** (o induttiva).
- ✓ Il **collettivo statistico** o **collettività** o **popolazione statistica** rappresenta l'insieme di unità statistiche omogenee rispetto ad alcuni caratteri di cui si acquisiscono informazioni per studiarne le modalità; non è necessariamente riferito a esseri umani.



Statistica Descrittiva e Statistica Inferenziale

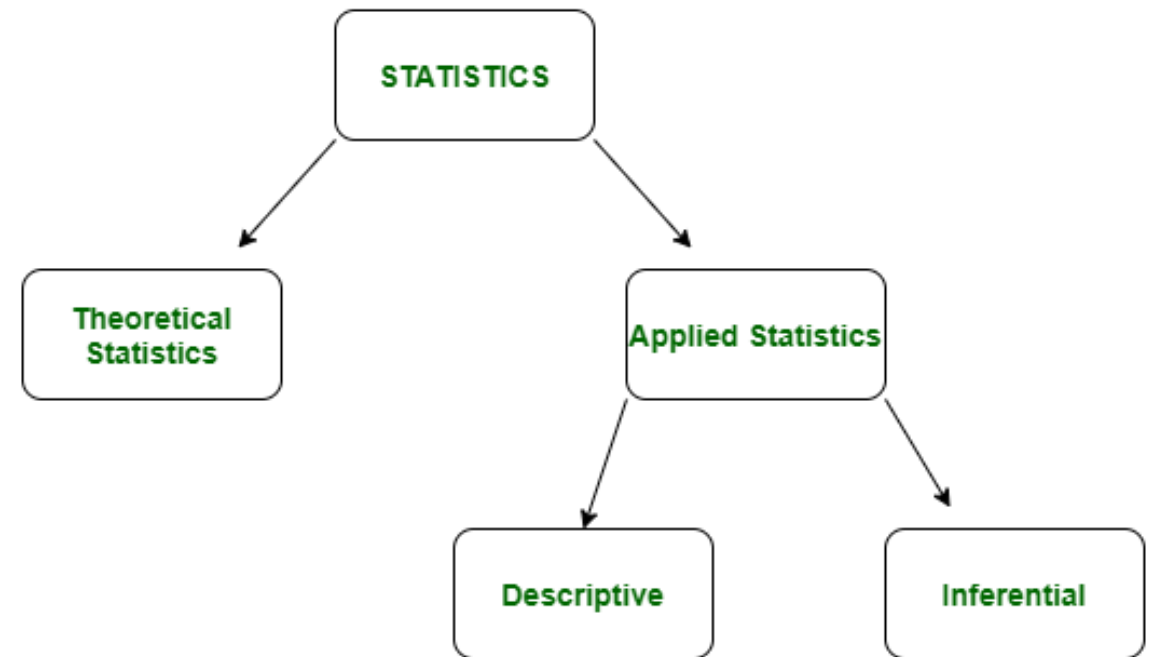


✓ La **Statistica Descrittiva** rappresenta le caratteristiche di un fenomeno collettivo attraverso strumenti statistici quali strumenti grafici o numerici che effettuano una sintesi (sintetizzano) di masse di dati grezzi chiamati microdati (come quelli derivanti dallo studio di un'intera popolazione) senza alterarne il significato complessivo.

✓ La **Statistica Inferenziale** partendo dall'osservazione di un **campione** di individui rappresentativo di un gruppo o di una popolazione, permette, tramite **induzione probabilistica**, di trarre indicazioni valide per l'intero gruppo o popolazione.

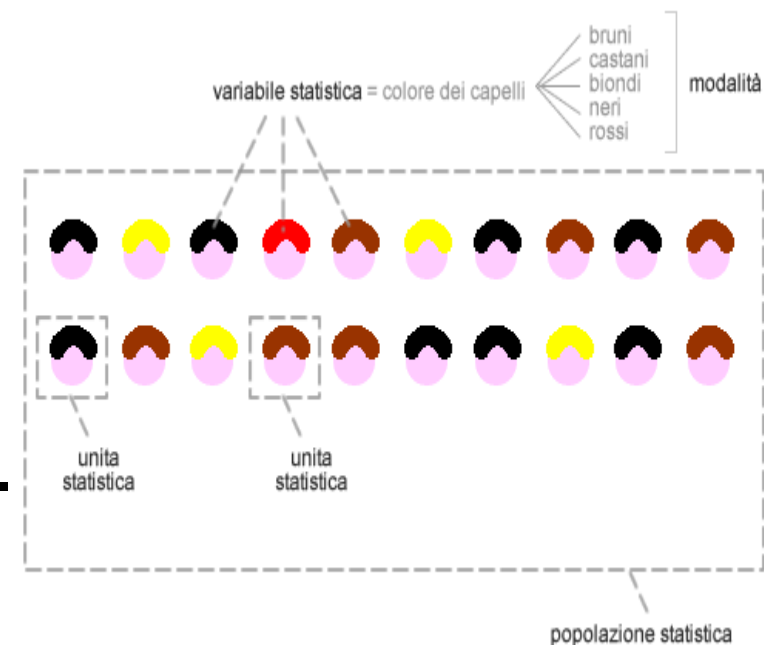
Statistica Pura vs Statistica Applicata

- ✓ **Statistica pura o teorica:**
racchiude regole e principi generali propri della scienza statistica astratta, indipendentemente dal fenomeno di riferimento.
- ✓ **Statistica applicata:** a seconda della materia a cui si applica la statistica possono distinguersi varie specializzazioni: statistica economica, statistica medica, statistica demografica, ecc. Il campo di applicazione della statistica si è notevolmente esteso negli ultimi anni.



Le 5 fasi dell'Analisi Statistica

- 1. Definizione degli Obiettivi:** Si tratta di una fase delicata in cui lo statistico deve individuare gli obiettivi delimitando lo spazio di ricerca in termini spaziali e temporali.
- 2. Rilevazione:** E' l'osservazione dei caratteri relativi alle unità statistiche mediante opportuni strumenti di rilevazione statistica. Questa fase può essere **completa (censimento)** se eseguita su tutte le unità statistiche che costituiscono la popolazione del fenomeno in esame. Oppure questa fase può essere **parziale** se viene condotta su un campione estratto dalla popolazione e il suo impiego si basa sull'approccio induttivo (dalla parte al tutto, dal principio specifico al principio generale) tipico dell' **Inferenza Statistica**.



Le 5 fasi dell'Analisi Statistica

NOTA: I dati sono raccolti su modelli che sono dei veri e propri formulari completi di domande e risposte, predisposti in modo da ottenere quei dati che interessano ai fini dell'analisi.

La rilevazione dei dati può essere svolta da enti privati (aziende, società commerciali, studi professionali, ecc.) o pubblici. In Italia, l'organo statistico ufficiale dello Stato è **l'ISTAT** (Istituto Nazionale di Statistica), persona giuridica di diritto pubblico con ordinamento autonomo, sottoposta alla vigilanza della Presidenza del Consiglio dei Ministri e al controllo della Corte dei Conti.

- 3. Elaborazione dei dati:** in questa fase i dati rilevati sono sintetizzati allo scopo di ottenere dati più significative.
- 4. Presentazione e interpretazione dei dati:** Consiste nella rappresentazione dei dati attraverso tabelle, grafici e indici, e nella spiegazione dei risultati ottenuti dall'intera analisi statistica.

Le 5 fasi dell'Analisi Statistica

5. Applicazione degli esiti dell'analisi:

La statistica non è una scienza fine a se stessa, ma richiede di essere applicate a diversi campi. In questa fase è compito dello statistico definire i limiti e i criteri di applicazione dei risultati dell'analisi.

La statistica è utilizzata sia nello studio dei **fenomeni naturali**, dei **fenomeni scientifici** (chimica, biologia, fisica, medicina, ecc.) e dei **fenomeni sociali** (economia, sociologia, ecc.), in ambito **tecnico** e **ingegneristico**, ecc.



Indagine statistica e tabelle

- ✓ **Un'indagine statistica** è un'operazione condotta, mediante l'osservazione, su elementi indefiniti di un determinato collettivo, con l'obiettivo di distinguerli e classificarli secondo le modalità di uno o più caratteri.
- ✓ **Un'unità statistica** è la componente elementare del collettivo, è su di essa che si acquisiscono le informazioni. Le unità statistiche possono essere:
 - 1. Unità semplice:** una singola persona o un'abitazione per esempio
 - 2. Unità composte:** sono insiemi di unità semplici, per esempio una famiglia o un edificio.
 - 3. Unità complesse:** che sono insiemi di unità semplice diverse, atte però a caratterizzarle nella loro totalità. Un esempio può essere il processo di produzione di prodotti assemblati, effettuato da un'impresa che si occupa delle singole componenti ma anche del loro montaggio.

Le 4 fasi di un'indagine statistica

- 1. Identificazione del collettivo statistico:** ossia l'operazione di **Un'indagine statistica** è un'operazione condotta, mediante l'osservazione, su elementi indefiniti di un determinato collettivo, con l'obiettivo di distinguerli e classificarli secondo le modalità di uno o più caratteri.
- 2. Rilevazione:** che è l'operazione mediante la quale si acquisiscono le modalità di uno o più caratteri del collettivo statistico. Essa può riguardare l'intera popolazione oggetto di osservazione (censimento) o può essere per campione, ossia riguardante un sottoinsieme della popolazione.
- 3. Elaborazione:** che è l'operazione di classificazione (in tabelle e grafici) e sintesi dei dati risultanti dallo spoglio.
- 4. Interpretazione:** la quale, sulla base delle conoscenze in merito al fenomeno oggetto di studio, chiarisce i risultati acquisiti.

Tabelle e caratteri statistici

- ✓ **I caratteri statistici** sono gli aspetti del fenomeno oggetto di rilevazione. A loro volta i caratteri statistici si dividono in **qualitativi** (tipo di attività, genere, direzione del vento) e **quantitativi** (come il reddito, la produzione, l'età, ecc.).
- ✓ **Le tabelle statistiche** emergono dalle operazioni di spoglio dei risultati di indagine statistiche, attraverso la classificazione dei dati rilevati in base alle modalità (o manifestazione dei caratteri). Le tabelle possono essere:

1. **semplice:** riportano le informazioni statistiche su un fenomeno collettivo in relazione ad un solo carattere
2. **multiple o a più entrate:** riportano le informazioni statistiche su un fenomeno collettivo in relazione a più di un carattere, combinando ciascuna modalità di un carattere con le modalità dell'uno o degli altri caratteri.

Tabella 2. Patologie preesistenti osservate più frequentemente

Patologie	Donne		Uomini		Totale	
	N.	%	N.	%	N.	%
Cardiopatia ischemica	236	20,8	711	31,8	957	27,8
Fibrillazione atriale	340	22,9	498	21,6	758	22,0
Scorpeno cardiaco	309	17,8	333	14,3	599	15,7
Ictus	318	20,6	231	10,0	349	10,2
Ipertensione arteriosa	774	88,1	1131	66,5	2305	67,0
Diabete mellito-tipo 2	322	28,3	718	31,3	1040	30,3
Demenza	266	23,6	296	12,9	562	16,3
PCO	343	22,6	433	18,8	576	16,8
Cancro attivo negli ultimi 5 anni	325	26,3	864	35,9	1189	35,0
Patologia cronica	87	3,3	111	4,8	146	4,3
Insufficienza renale cronica	200	17,9	488	21,2	688	20,0
Dalton	19	1,7	48	2,1	67	1,9
Insufficienza respiratoria	61	5,6	119	5,3	180	5,2
HIV	0	0,0	7	0,3	7	0,2
Malattie autoimmuni	67	5,9	79	3,6	137	4,0
Osteoporosi	327	31,2	259	10,5	577	15,9

Mutable statistica
"Sesso"

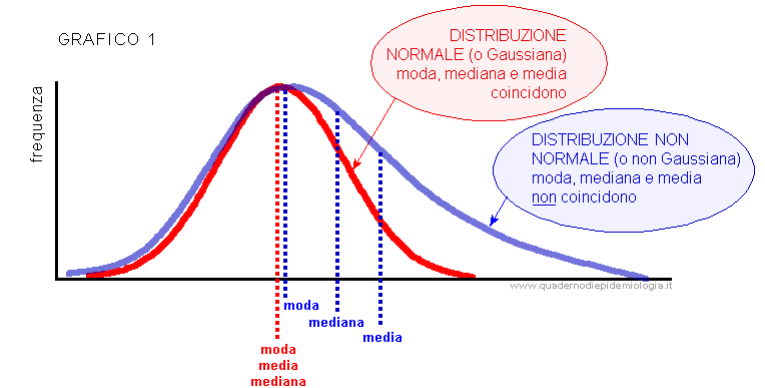
Mutable statistica
"Patologie"

Tabelle e caratteri statistici

- ✓ **Una tabella statistica** si definisce mediante qualificazioni non determinabili numericamente e/o mediante numeri che sono:
 - 1. intensità:** se mostrano la misura o la grandezza di un carattere (come il peso di una persona, l'ammontare degli investimenti di un'azienda, ecc..)
 - 2. frequenze:** se mostrano il numero di volte in cui una modalità del carattere si presenta nelle unità statistiche (come il numero degli iscritti alle liste di leva di uno specifico anno, il numero di iscritti ai licei scientifici in un dato anno scolastico, ecc..)
- ✓ A sua volta le frequenze si distinguono in: **frequenze assolute**, che indicano il numero di unità di un collettivo che presenta una data modalità (valore) di un carattere, **frequenze relative**, che derivano dalla frequenza assoluta diviso il totale delle stesse, **frequenze percentuali** che sono le frequenze relative per 100, e infine le **frequenze cumulate** che indicano le frequenze delle osservazioni che hanno un valore del carattere minore a una prestabilita modalità

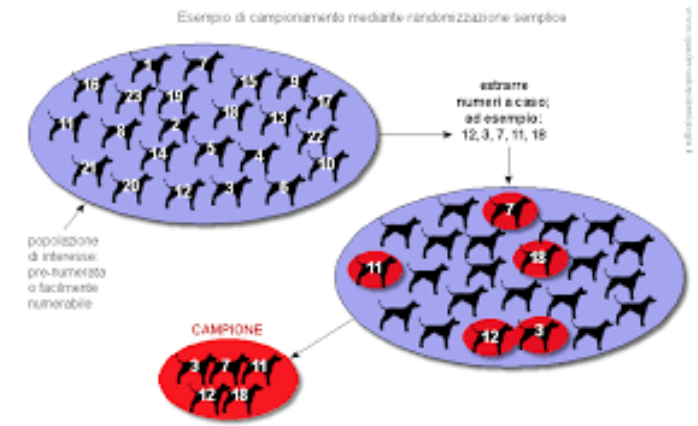
Distribuzione Statistica

- ✓ **Una distribuzione statistica** è l'insieme delle determinazioni del carattere e delle rispettive frequenze.
- ✓ Se il carattere è **quantitativo** allora la distribuzione statistica prende il nome di **variabile statistica**.
- ✓ Se il carattere è **qualitativo** allora distribuzione prende il nome di **mutabile statistica**.
- ✓ Inoltre una variabile statistica può essere **continua** o **discreta** a seconda dell'insieme di dati di riferimento.



La Teoria della Stima

- ✓ Spesso non si hanno le risorse disponibili per effettuare una rilevazione di dati che riguardi l'intera **popolazione** interessata da un fenomeno. Per esempio potrebbe succedere che tale popolazione è **infinita**, ed una rilevazione completa (esaustiva) risulta impossibile.
- ✓ In questi casi si procede ad una **rilevazione di dati per campione**.
- ✓ **Il campione** è quella parte del collettivo statistico che viene sottoposto ad osservazione.
- ✓ L'insieme dei **campioni** di una certa ampiezza che si possono estrarre da un dato collettivo mediante una determinata procedura prende il nome di **Universo dei Campioni**.

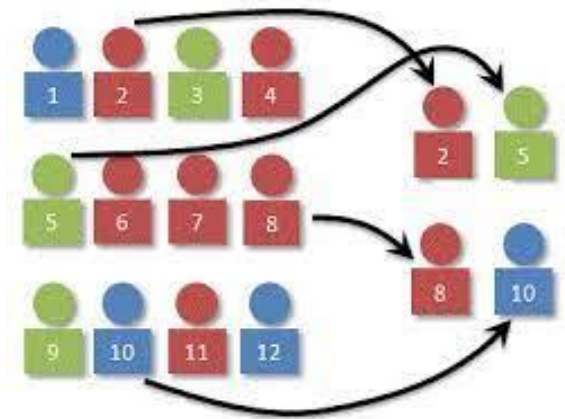


La Teoria della Stima

- ✓ La **numerosità** (o consistenza) del **campione n** dipende dalla **numerosità della popolazione N** .
- ✓ **L'Inferenza Statistica** (o statistica inferenziale) è quella parte dell'analisi statistica che tenta di derivare dalle informazioni raccolte sul **campione** altre informazioni riguardanti la **popolazione**, in modo da "inferire" quali sono le caratteristiche salienti della popolazione a partire da quelle del campione.
- ✓ **Campionamento:** è il procedimento in base al quale si perviene alla costituzione del campione e alla rilevazione dei dati relativi ad esso.
- ✓ L'estrazione di un **campione** può avvenire in due modalità:
 - 1. con reimmissione**
 - 2. senza reimmissione**

La rilevazione dei dati per campioni

- ✓ Nel campionamento con **reimmissione**, detto anche "**campionamento bernoulliano**", non si esclude che un elemento del campione venga ripescato una o più volte. Questo è il caso che interessa maggiormente, in quanto la reimmissione fa sì che le variabili casuali rappresentate dalla prima estrazione, dalla seconda e così via siano una indipendente dall'altra, cosa che non avverrebbe in caso di estrazione senza reimmissione, detto anche "**campionamento in blocco**".
- ✓ **Non esiste un unico modo per campionare da una popolazione.** Il **campionamento casuale semplice** è quello più utilizzato, quando si vuole che le unità statistiche della popolazione abbiano la stessa probabilità di entrare nel **campione**.

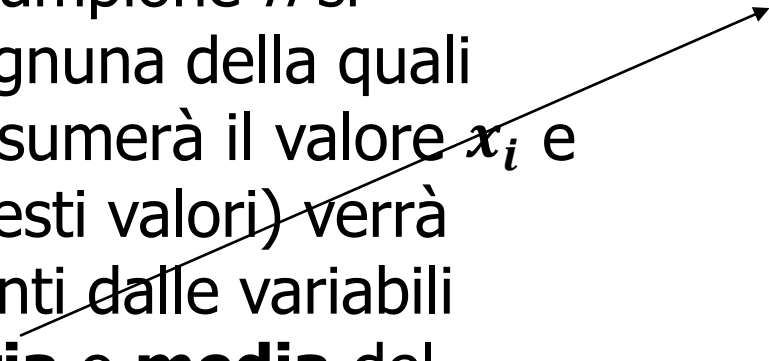



Campionamento statistico

- ✓ Il primo individuo estratto è una **variabile casuale** X_1
- ✓ Il secondo individuo estratto è una **variabile casuale** X_2
- ✓ L' n -esimo estratto rappresenta la **variabile casuale** X_n
- ✓ Estratto il campione la **variabile casuale** X_1 assumerà il valore x_1 , X_2 assumerà il valore x_2 , e così via fino ad n .
- ✓ Nel caso di un campionamento con reimmissione o ripetizione le n variabili casuali sono **indipendenti** ed hanno identica funzione di probabilità $f(X)$.
- ✓ Dalle n funzioni di probabilità è possibile ottenere con metodi matematici un'espressione che riassume le **caratteristiche** del campione.
- ✓ Per esempio è importante fornire informazioni sui **parametri** della popolazione che riteniamo sconosciuti come **media** o **varianza**.

I parametri campionari

- ✓ Il **riassunto campionario**, ossia desumere i parametri della popolazione mediante parametri campionari prende il nome di "**stima**".
- ✓ Dunque, determinata l'ampiezza del campione n si definiscono n **variabili casuali** X_i , ognuna della quali rappresenta l' i -esima estrazione che assumerà il valore x_i e la media del campione (dunque di questi valori) verrà detta media aritmetica dei valori assunti dalle variabili casuali, ovvero la **media campionaria** o **media del campione**.
- ✓ Questa media non è altro che uno dei possibili valori che può assumere la variabile casuale.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$


$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$


Riassunto campionario

- ✓ La **media campionaria** è dunque un **riassunto campionario**.
- ✓ E' necessario stabilire la distribuzione della media campionaria pertanto, dato che tutte le X_1, \dots, X_n hanno la stessa distribuzione come si è supposto e il valore atteso della media campionarie è $E(\bar{X}) = \mu_{\bar{X}} = \mu$ allora tutte le variabili hanno lo stesso valore atteso e la stessa varianza: $E(X_i) = \mu$ e $var(X_i) = \sigma^2$
- ✓ La **varianza** della distribuzione campionaria delle medie è invece data, nel caso di popolazione finita e campionamento senza ripetizione, da:

$$S_{\bar{X}}^2 = var(\bar{X}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

Dove N indica la **numerosità della popolazione**, n è la **numerosità del campione** e σ è lo **scarto quadratico medio della popolazione** (o deviazione standard) che è un indice di dispersione statistico, vale a dire una stima della variabilità di una popolazione di dati o di una variabile casuale.

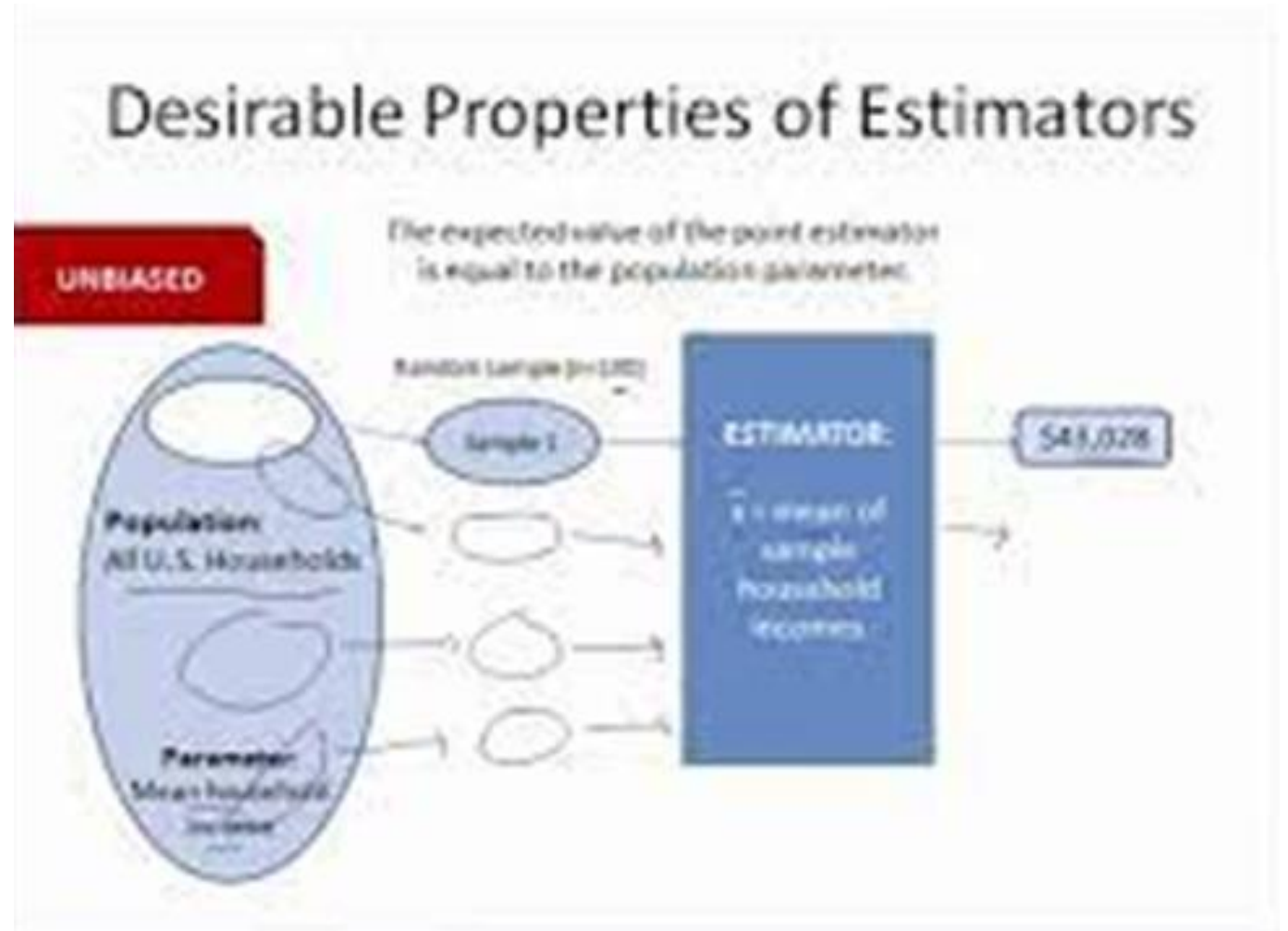
Proprietà di uno stimatore

- ✓ Esistono i parametri che riguardano la **popolazione** e che sono **sconosciuti**
- ✓ Ed esistono i parametri che riguardano il **campione** che sono calcolabili a partire dai dati rilevati.
- ✓ L' **inferenza statistica**, esegue delle stime sui parametri della **popolazione** a partire dai parametri del **campione**.
- ✓ Dunque l' **inferenza statistica** ha il compito di determinare uno **stimatore**, cioè una funzione che associa ad ogni possibile campione un valore del parametro da stimare.
- ✓ La **stima** è appunto il valore che uno **stimatore** assume in corrispondenza di un particolare campione. Dunque uno stimatore è una **variabile casuale funzione del campione** a valori nello spazio parametrico, ossia nell'insieme dei possibili valori del parametro (codominio dello stimatore).

Proprietà di uno stimatore

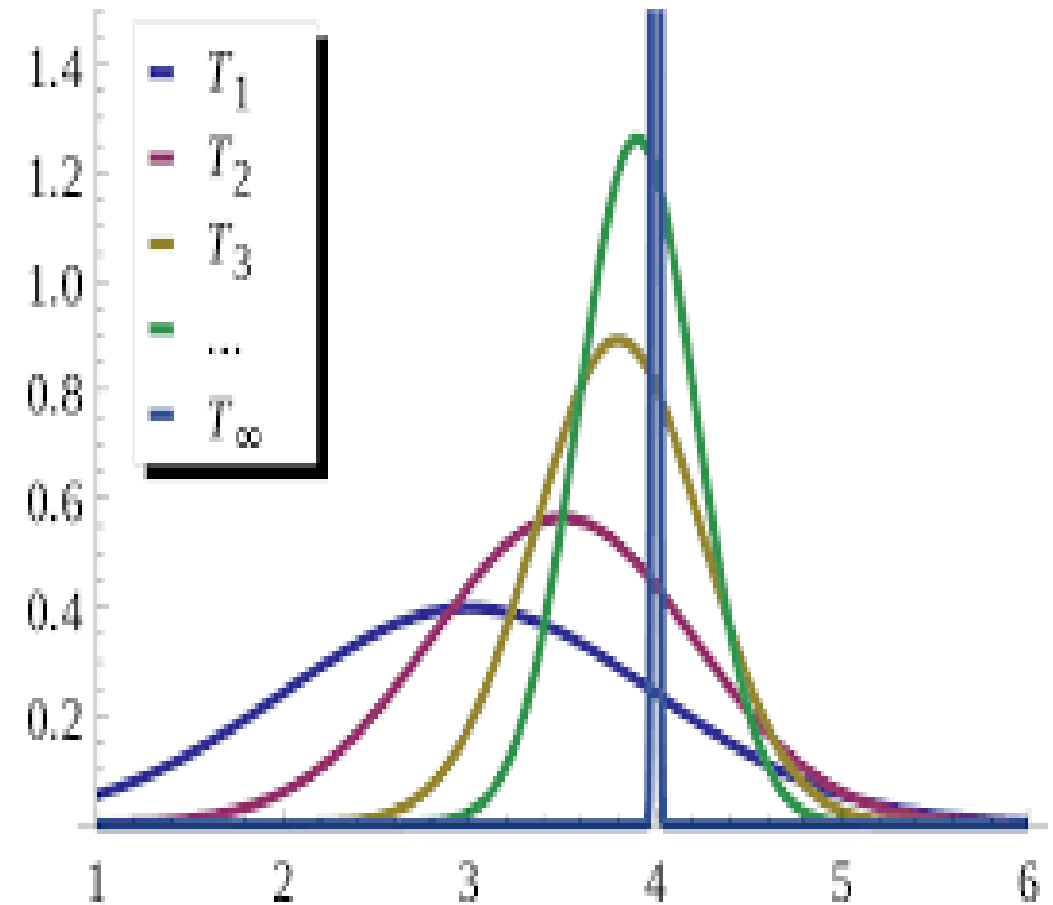
✓ Le proprietà desiderabili di uno stimatore possono essere:

- **Correttezza**
- **Consistenza**
- **Efficienza**
- **Sufficienza**
- **Normalità asintotica**



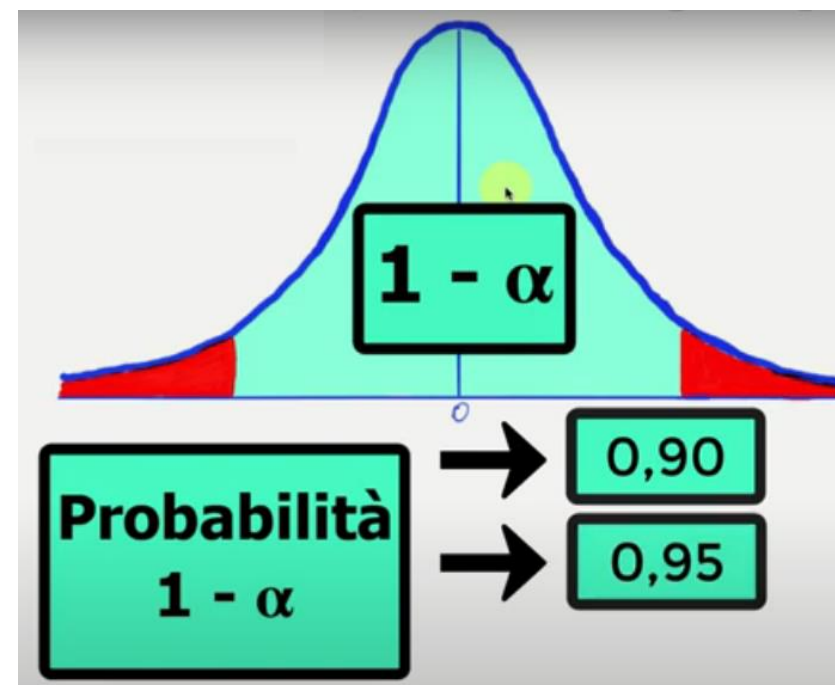
Proprietà desiderabili degli Stimatori: Correttezza

- ✓ Uno stimatore $T(X)$ si dice **corretto** o **non distorto** quando il suo **valore medio** $E[T(X)]$ coincide con il valore del parametro θ da stimare per qualsiasi suo valore: $E[T(X)] = \theta$.
- ✓ Se invece tale uguaglianza non si verifica, allora l'espressione:
 $d(\theta) = \theta - E[T(X)]$ e prende il nome di **"tendenziosità"** o **"distorsione"** dello stimatore.



Proprietà desiderabili degli Stimatori: Correttezza

- ✓ Lo stimatore **media campionaria** \bar{X} della media μ è **corretto** in quanto il valore atteso della media campionaria coincide con il parametro **media della popolazione**.
- ✓ Invece, lo stimatore $\hat{S}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ ha come valore atteso $E[\hat{S}^2] = \frac{n-1}{n} \sigma^2$ che è diverso da σ^2 . Lo **stimatore corretto** della varianza σ^2 è invece: $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{n-1}{n} \hat{S}^2$ che ha come valore atteso $E[S^2] = \frac{n-1}{n} \hat{S}^2$

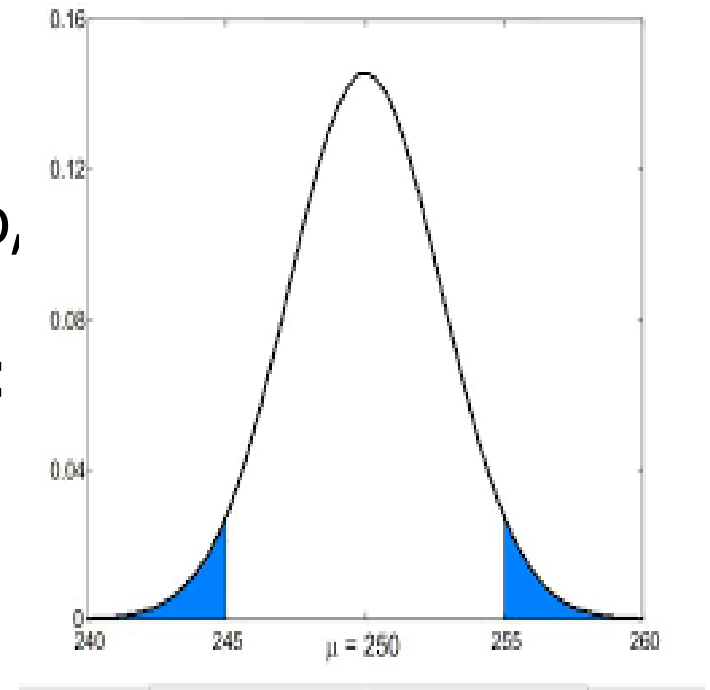


Errore di prima specie ed errore di seconda specie

- ✓ Dato un **esperimento casuale** definito su un certo **spazio campionario** e con misura di probabilità P , nel **modello statistico di base**, abbiamo una **variabile casuale** osservabile \mathbf{X} che assume valori in \mathbf{S} .
- ✓ In generale, \mathbf{X} può avere struttura complessa, ad esempio, se l'esperimento consiste nell'estrarre n unità da una popolazione e registrare le varie misure di interesse, allora:

$$\mathbf{X} = (X_1, X_2, \dots, X_n)$$

- ✓ dove X_i è il vettore di misurazioni per l' i -esima unità.
- ✓ Il caso più importante si ha quando X_1, X_2, \dots, X_n sono indipendenti e identicamente distribuite. Si ha allora un **campione casuale** di dimensione n dalla distribuzione comune



Errore di prima specie ed errore di seconda specie

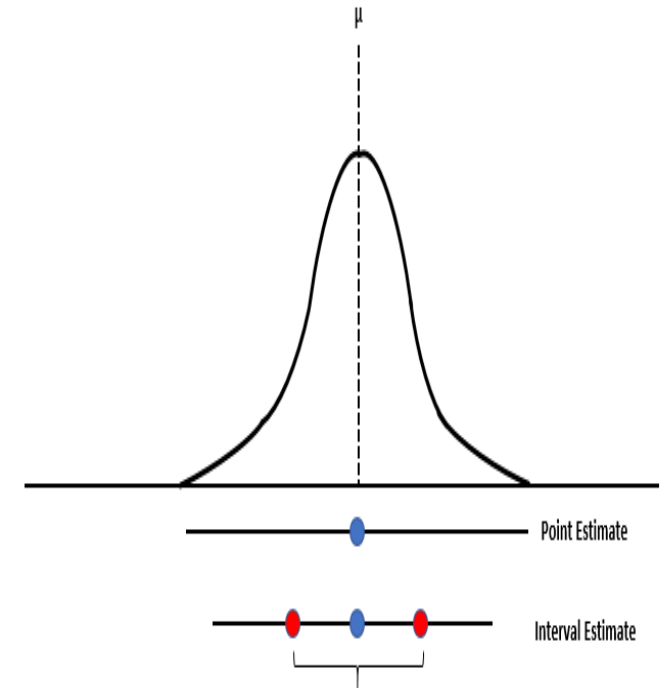
- ✓ **Un'ipotesi statistica** è un'asserzione sulla distribuzione della variabile **X** (ipotesi appunto).
- ✓ Equivalentemente **un'ipotesi statistica** individua un **insieme** di possibili distribuzioni per **X**.
- ✓ L'obiettivo dei **test delle ipotesi** è valutare se vi è sufficiente **evidenza statistica** per rifiutare l'**ipotesi nulla** in favore dell'**ipotesi alternativa**.
- ✓ **L'ipotesi nulla** si indica generalmente con H_0 , mentre **l'ipotesi alternativa** H_1 .
- ✓ Un'ipotesi che specifica una singola distribuzione per **X** si dice semplice; mentre un'ipotesi che ne specifica più di una **X** si dice invece **composta**.
- ✓ Un test di ipotesi conduce ad una **decisione statistica**, la cui conclusione potrà essere di rifiutare l'ipotesi nulla in favore di quella alternativa, o **di non poter rifiutare** l'ipotesi nulla.

Errore di prima specie ed errore di seconda specie

- ✓ La decisione che prendiamo è basata sui dati di cui disponiamo **X** .
- ✓ Pertanto dobbiamo trovare un **sottoinsieme R** dello spazio campionario **S** e rifiutare H_0 se e solo se **X appartiene a R** . R prende il nome di **regione di rifiuto** o **regione critica**.
- ✓ Usualmente, la regione critica è definita in funzione di una statistica detta **statistica di test: $W(X)$** .
- ✓ La decisione che prendiamo può essere corretta o errata. Esistono due tipi di errore, a seconda di quale delle due ipotesi è vera:
 - 1. Errore di prima specie:** consiste nel rifiutare l'ipotesi nulla quando è vera
 - 2. Errore di seconda specie:** consiste nel non rifiutare l'ipotesi nulla quando è falsa

Stimatori, Bias e Varianza per il Machine Learning

- ✓ Il campo della statistica fornisce molti strumenti che possono essere usati anche gli obiettivi del machine learning di risolvere un compito non solo sul training set ma anche di generalizzare. Concetti fondamentali come stima dei parametri, bias e varianza sono utili per caratterizzare formalmente le nozioni di generalizzazione, underfitting e overfitting
- ✓ La **Stima puntuale** dei parametri rappresenta l'insieme dei metodi di statistica inferenziale che permettono di attribuire un valore ad un parametro della popolazione, utilizzando i dati di un campione casuale osservato (x_1, x_2, \dots, x_n) ed elaborandoli.



Stima Puntuale

- ✓ La **Stima Puntuale** è dunque il tentativo di fornire la migliore predizione singola ad alcune quantità di interesse. In generale le quantità di interesse possono essere un **singolo parametro** o un vettore di parametri in alcuni modelli parametrici, come i pesi di una **rete neurale** o i coefficienti di una **regressione lineare**.
- ✓ Al fine di distinguere le stime dei parametri dai loro valori veri, la nostra convenzione sarà di denotare una stima puntuale di un parametro θ con $\hat{\theta}$.
- ✓ Siano $\{x^{(1)}, \dots, x^{(m)}\}$ un insieme di m data point (punti dati) che sono indipendenti e identicamente distribuiti. Uno **stimatore puntuale** o **statistica** è una qualsiasi funzione sui dati di tipo:

$$\hat{\theta}_m = g(x^{(1)}, \dots, x^{(m)}).$$

Proprietà desiderabili degli Stimatori: Correttezza

- ✓ La **stima puntuale** può anche riferirsi alla stima delle relazioni tra input e variabili di target. Ci riferiamo a questi tipi di stime puntuali come stimatori di funzione (o approssimatori di funzione).
- ✓ Stiamo cercando di predire una variabile y dato un vettore di input x . Assumiamo che ci sia una funzione $f(x)$ che descrive la relazione approssimata tra y e x . Per esempio assumiamo che $y = f(x) + \epsilon$, dove ϵ sta per la parte di y che non è predicibile a partire dalla x .
- ✓ Nella stima di funzioni siamo interessati ad approssimare f attraverso un modello o stima \hat{f} . Stimare una funzione è lo stesso di stimare il parametro θ ; in altre parole lo stimatore di funzione \hat{f} è semplicemente uno stimatore puntuale nello spazio puntuale delle funzioni. La regressione lineare e la regressione polinomiale sono entrambi possono essere interpretati come stima di parametri W oppure come stima di una funzione \hat{f} che fa un mapping dalla x alla y .

Bibliografia
