Statistics

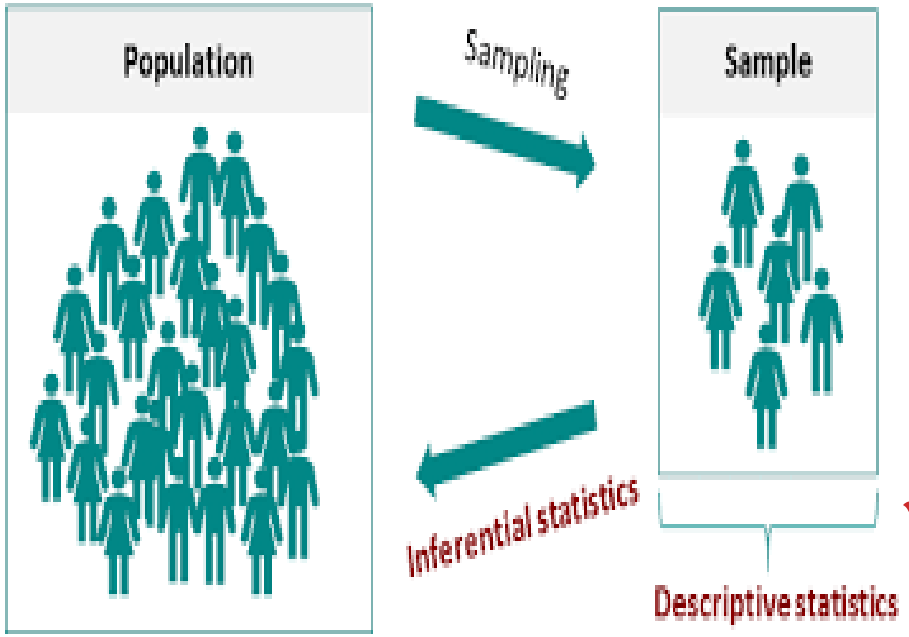
Francesco Pugliese, PhD

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.

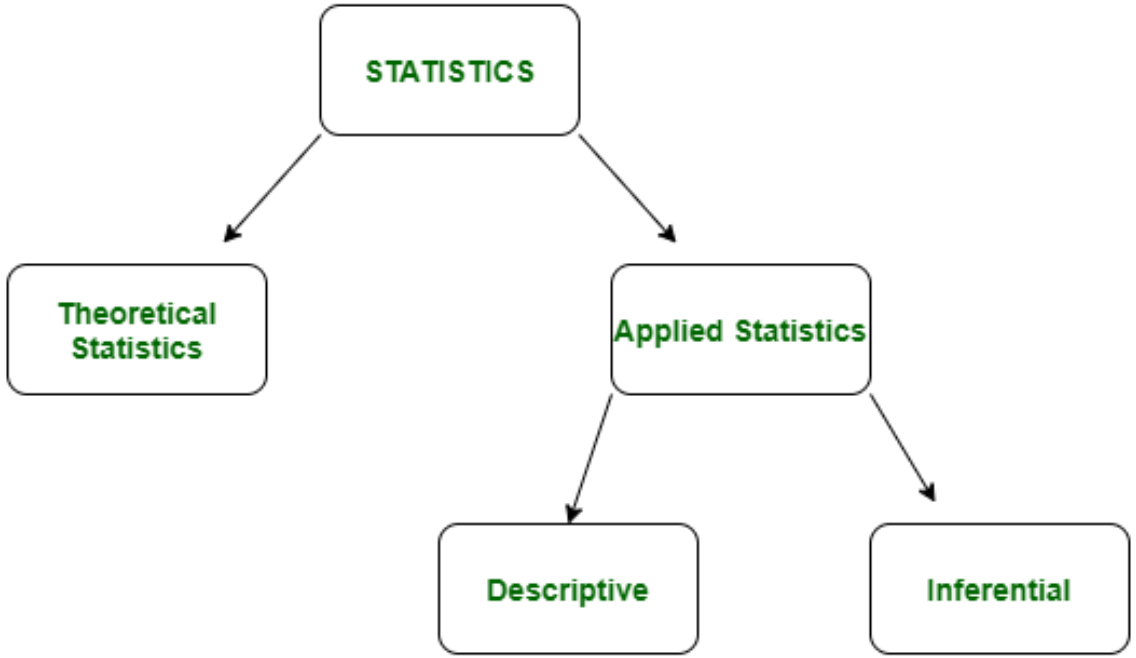
Statistica Descrittiva e Statistica Inferenziale



- 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.

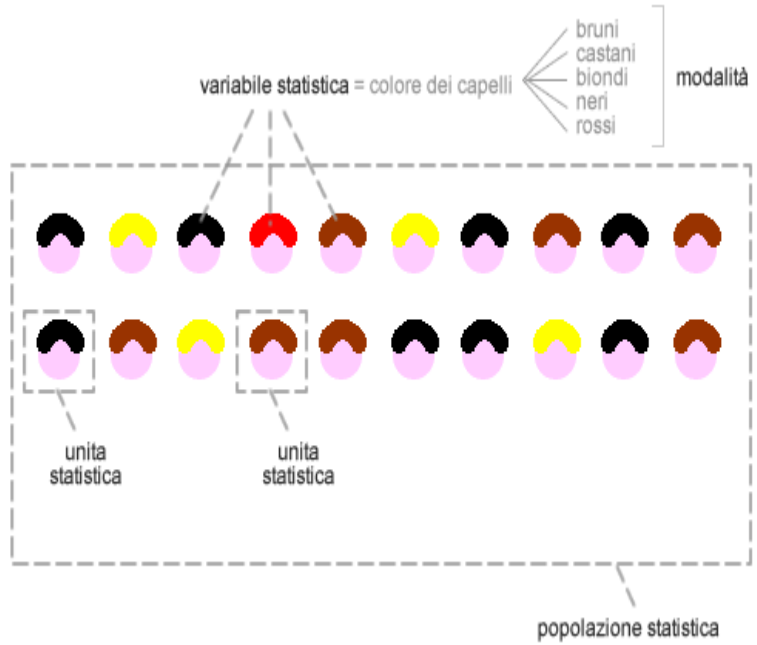
Statistica Pura vs Statistica Applicata

- **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.

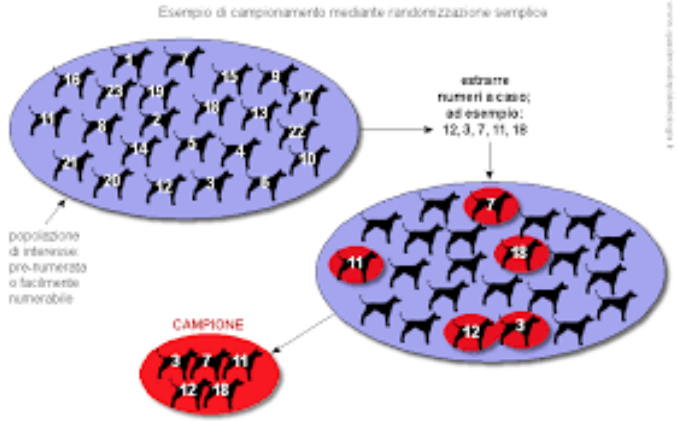
Le 5 fasi dell'Analisi Statistica

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.

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.**



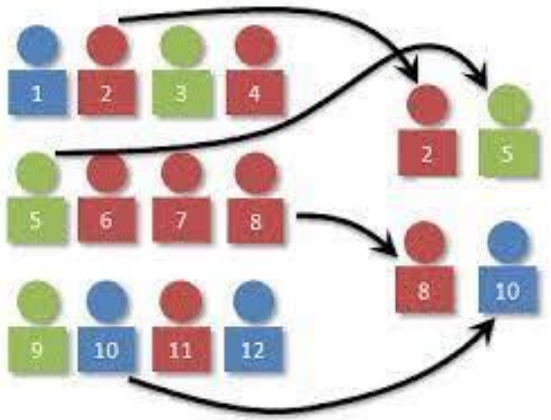
La Teoria della Stima

- 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 campuione.
- **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 si che le variabili casuali rappresentate dalla prima estrazione, dalla seconda e così via siando 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 riassuma 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 rapresenta 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 uno dei possibili valori che può assumere la variabile casuale.

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

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

Riassunto campionario

- ✓ La **media campionaria** è dunque un **riassunto campionario.**
- ✓ E' necessario stabilire la distribuzione della media campionaria pertanto, dato che tutte le $\boldsymbol{X_1}$, .., $\boldsymbol{X_n}$ 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:

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

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.

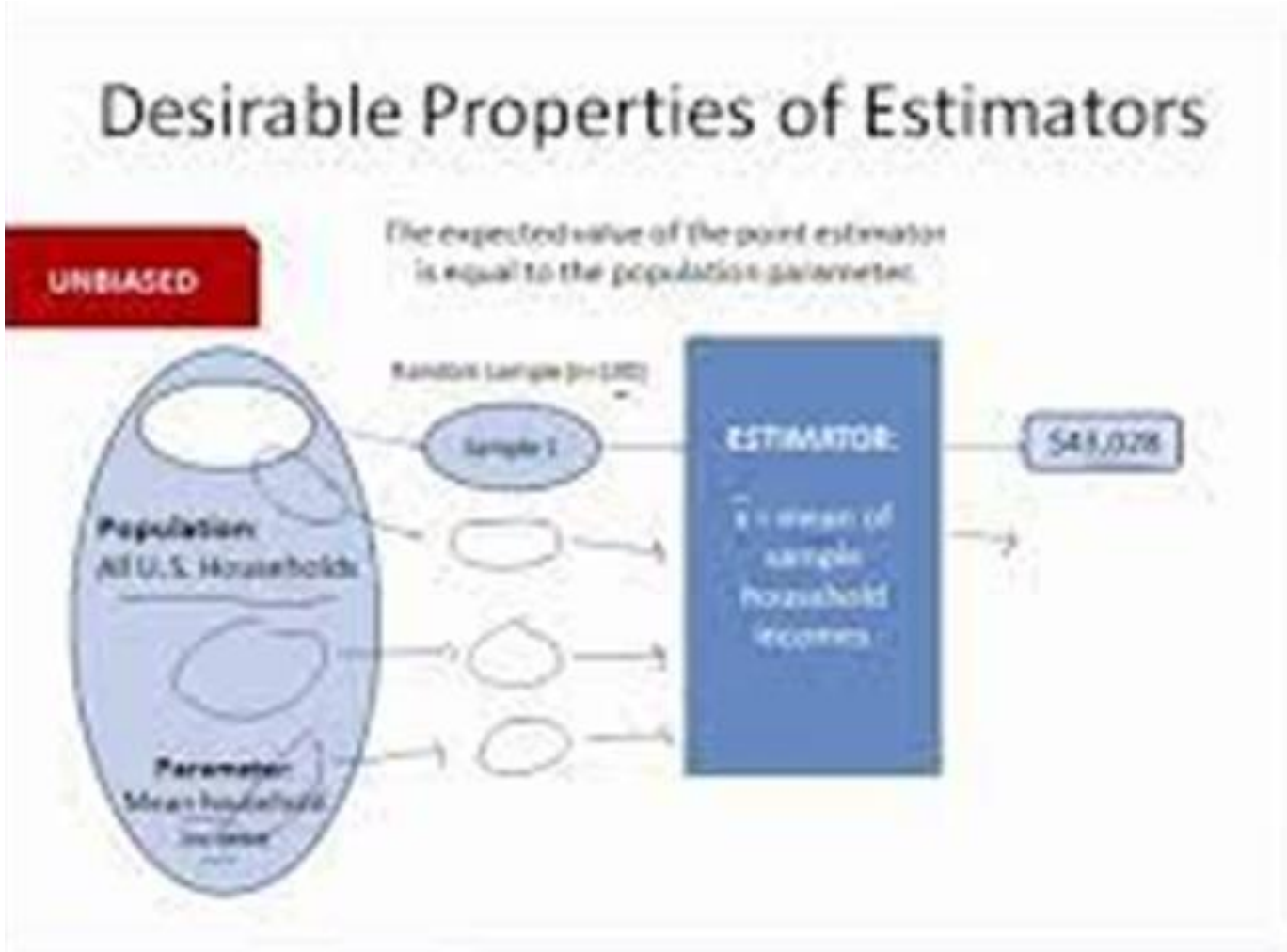
Proprietà di uno stimatore

- ✓ 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:

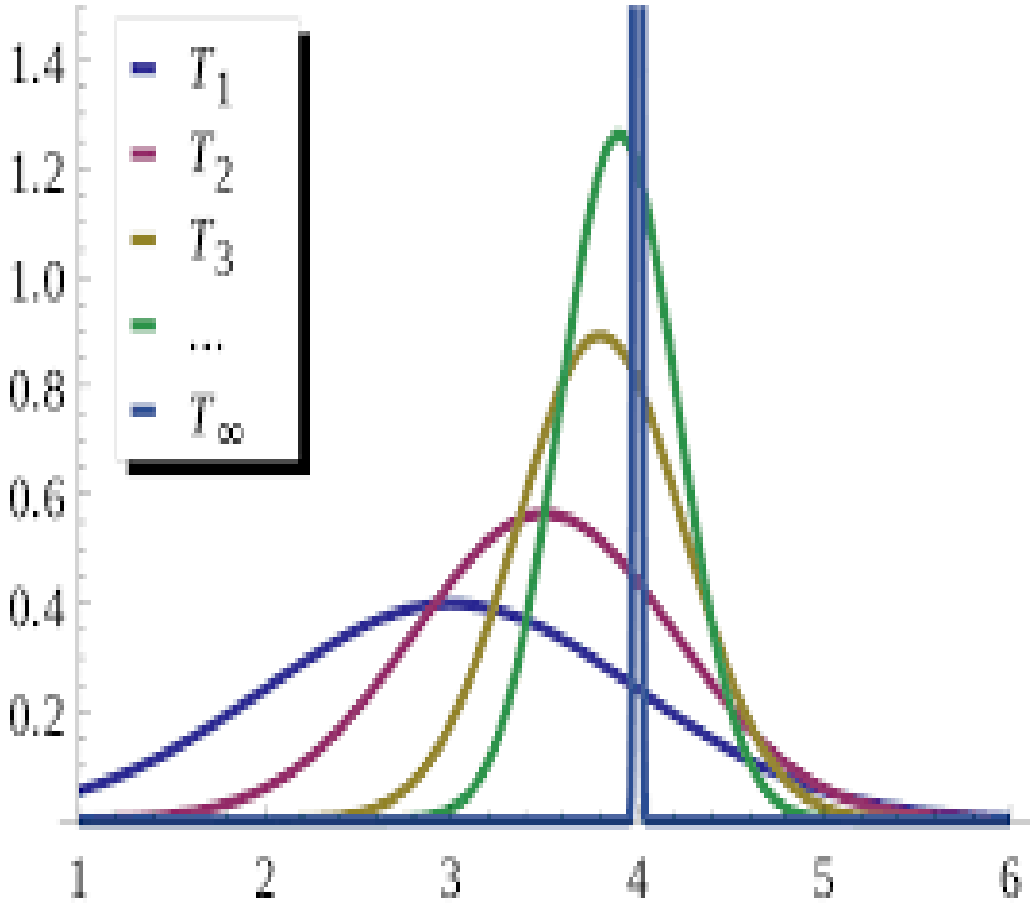
- **Correttezza**
- **Consistenza**
- **Efficienza**
- **Sufficienza**
- **Normalità asintotica**



Proprietà desiderabili degli Stimatori: Correttezza

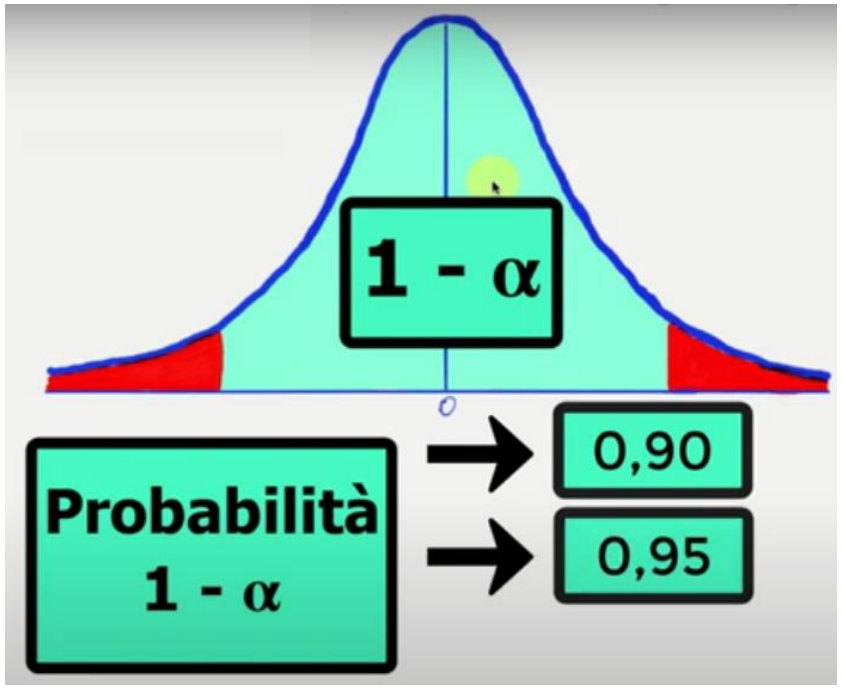
- ✓ Uno stimatore *T(X)* si dice **corretto** o **non distorto** quando il suo **valore medio** *E[T(X)]* coincide con il valore del parametro Θ da stimare per qualsiasi suo valore: $E[T(X) = \Theta$.
- ✓ Se invece tale uguaglianza non si verifica, allora l'espressione:

  $\mathrm{d}(\Theta) = \Theta - E[T(X)]$ e prende il nome di "**tendeziosità**" 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 paramtero **media della popolazione.**
- ✓ Invece, lo stimatore $\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ ha come valore atteso $\mathrm{E}[\hat{S}^2] = \frac{n-1}{n}\sigma^2$ 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 $\mathrm{E}[S^2] = \frac{n-1}{n}\hat{S}^2$

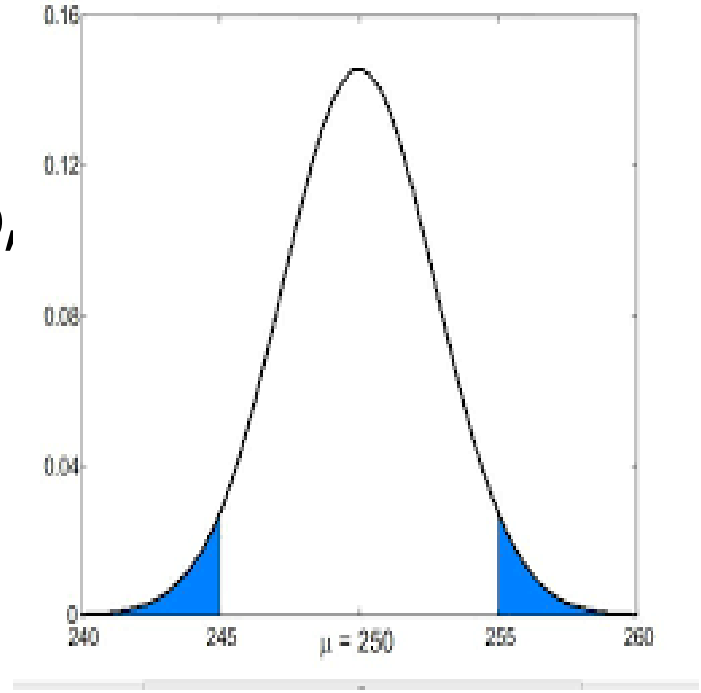


Errore di prima specie ed errore di seconda specie

- 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 **X** che assume valori in **S**.
- In generale, **X** può avere struttura complessa, ad esempio, se l'esperimento consiste nell'estrarre *n* unità da una popolazione e registrare le varie misure di interese, allora:

$$X = (X_1, X_2, \dots X_n)$$

- 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.

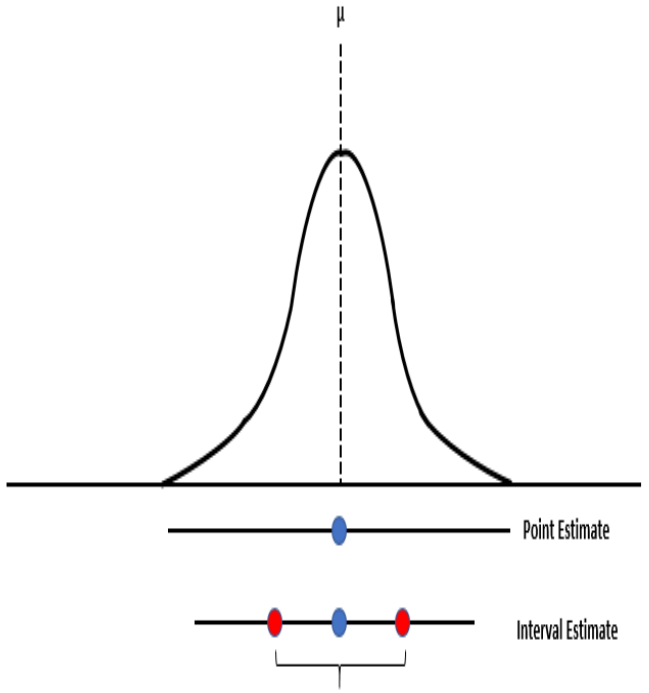
Errore di prima specie ed errore di seconda specie

- ✓ 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 statstica 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 (x1, x2,…,xn) 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 distringere 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 $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ 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:

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

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) + **ϵ**, 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 paramtri ***W*** oppure come stima di una funzione $\hat{f}$ che fa un mapping dalla ***x*** alla ***y.***

Bibliografia