

# Statistics

*Francesco Pugliese, PhD*

*neural1977@gmail.com*

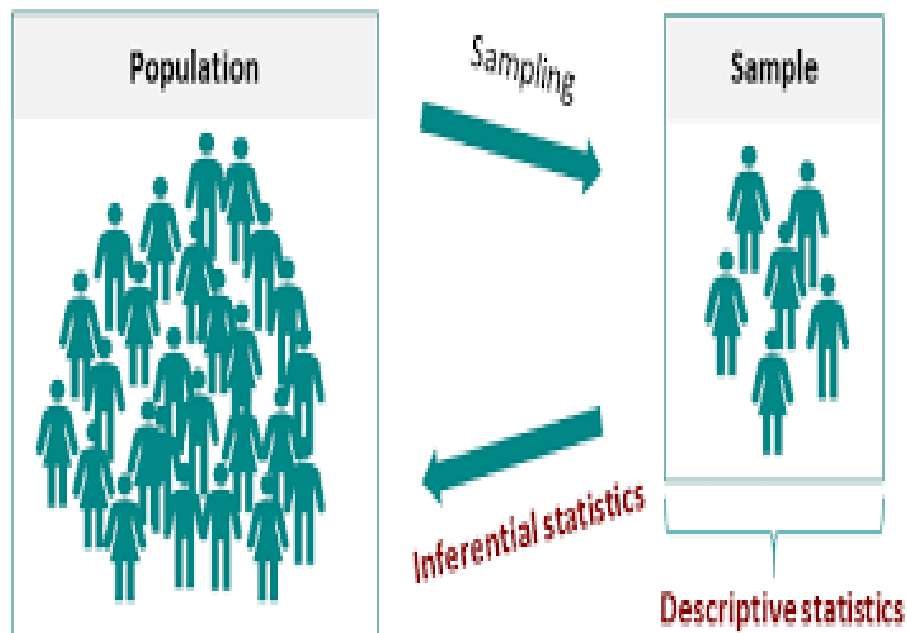
# Campo di analisi della statistica

---

- ✓ **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



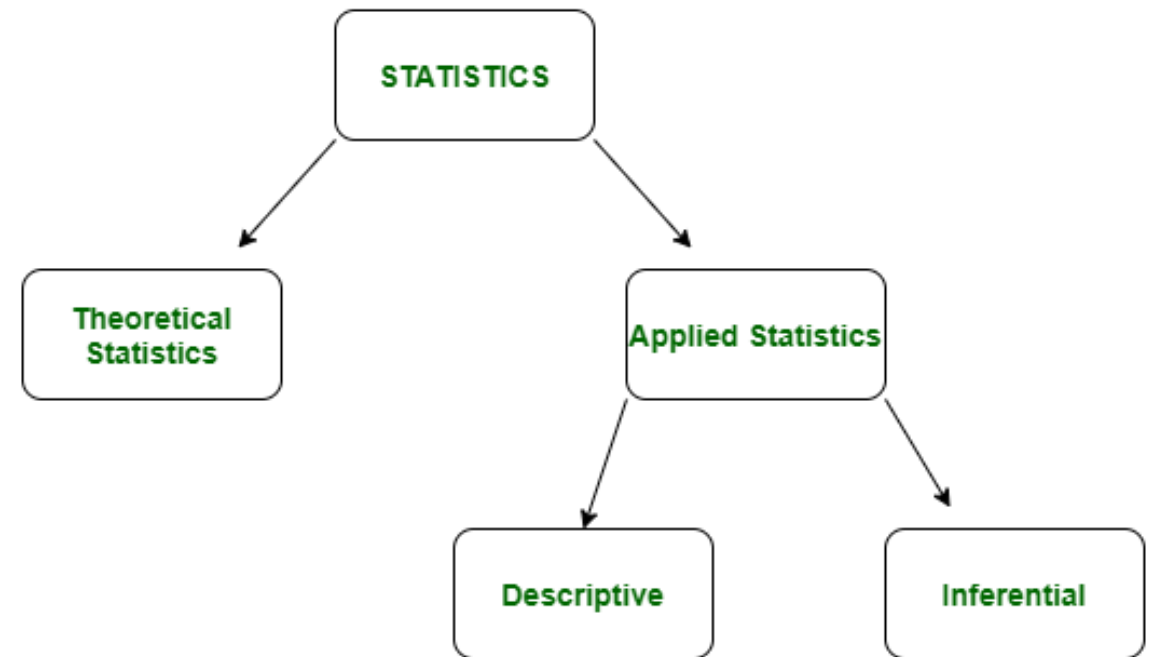
✓ La **Statistica Descrittiva** rappresenta le caratteristiche di un fenomeno collettivo attraverso strumenti statistici quali strumenti grafici o numerici che effettuano una sintesi (sintetizzano) di masse di dati grezzi chiamati microdati (come quelli derivanti dallo studio di un'intera popolazione)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

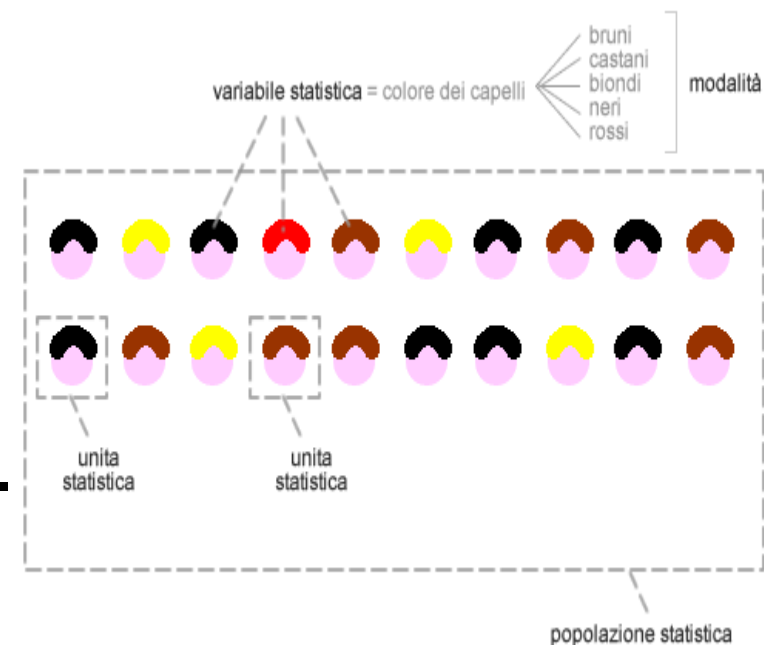
---

- ✓ **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	1331	60,5	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
Dialisi	19	1,7	48	2,1	67	1,9
Insufficienza respiratoria	61	5,4	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

# Mutabile Statistica

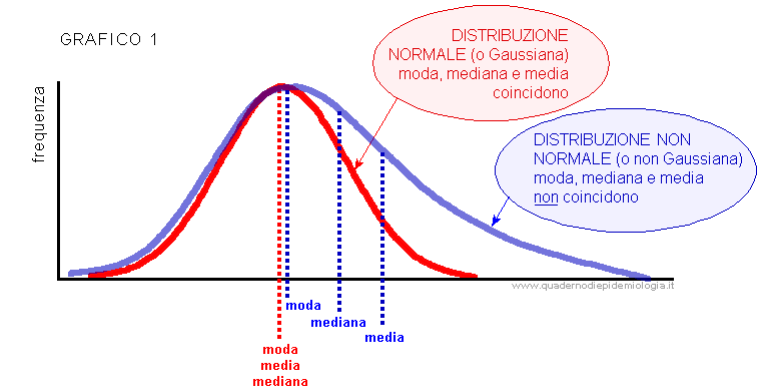
---

✓ Una **mutabile statistica** può essere:

1. **Rettilinea:** se esiste un ordine naturale o logico delle modalità (ad esempio il numero di operai metalmeccanici per livello)
2. **Serie storica o temporale:** se il principio regolatore è il tempo, il quale è inteso come progressione cronologica. Ovviamente una serie storica è una particolare **mutabile** rettilinea (ad esempio, il numero di autovetture di una data marca vendute in diversi anni)
3. **Ciclica:** se il tempo è inteso in termini di periodicità, per cui non esistono né una modalità iniziale, né una modalità finale (è il caso delle precipitazioni nevose nei diversi mesi in un anno)
4. **Sconnesse:** se non esistono né un ordine logico né un ordine naturale secondo cui sono disposte le modalità (il numero dei voti ottenuti dai partiti durante le elezioni)

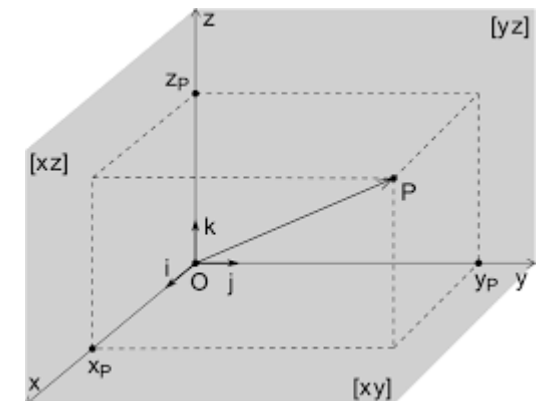
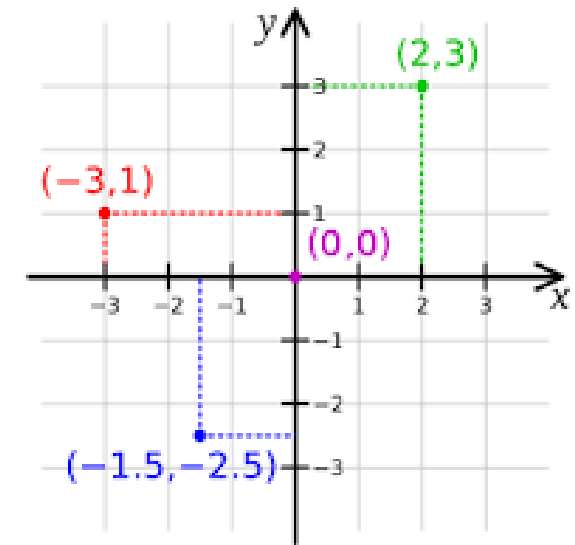
# Distribuzione Statistica

- ✓ **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.



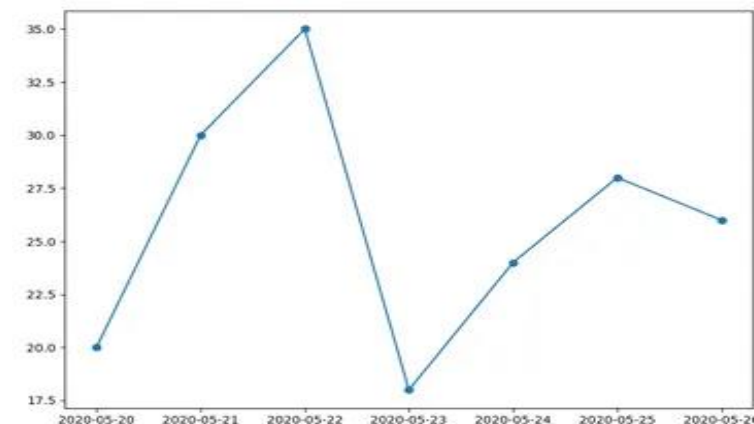
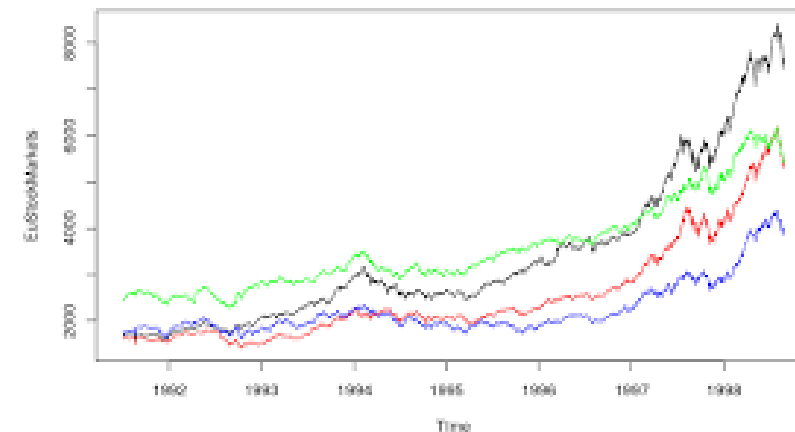
# Rappresentazioni Grafiche

- ✓ **Esistono molteplici rappresentazioni grafiche** dei dati statistiche, vediamo solo quelle che si prestano ad una interpretazione dei dati in maniera soddisfacente.
- ✓ **Riferimento Cartesiano Ortogonale (Scatter Plot, Line Plot, ecc):** tale sistema è costituito da due **rette ortogonali**, il cui punto di intersezione (0) è denominato **origine** e la linea orizzontale viene detta **asse delle ascisse** mentre quella verticale viene detta **asse delle ordinate**.
- ✓ Su entrambi gli assi si fissano un'unità di misura dei segmenti ed un orientamento.



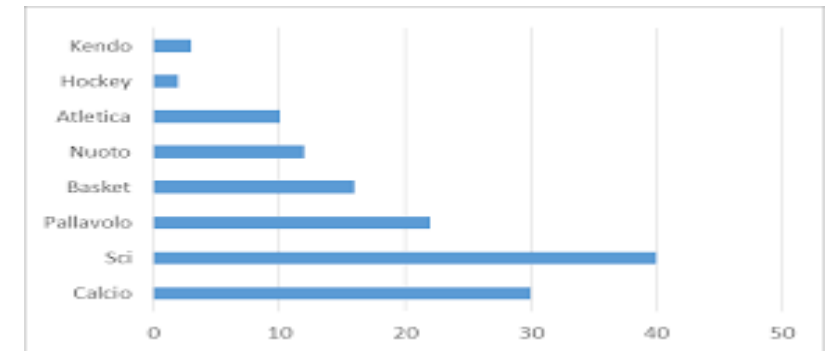
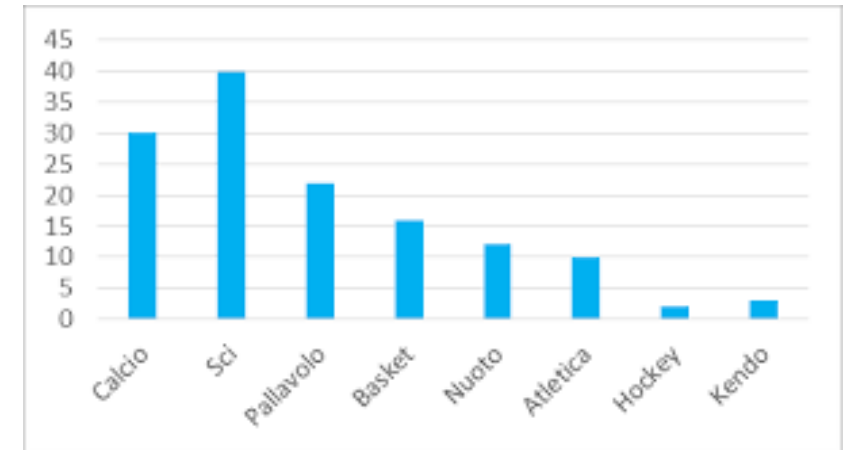
# Riferimento Cartesiano Ortogonale

- ✓ Una volta stabilite le opportune unità di misura per entrambi gli assi, la rappresentazione grafica in question è particolarmente utile nel caso delle **serie temporali**
- ✓ Sull'asse delle ascisse si fissa la relativa unità temporale (giorno, mese, anno)
- ✓ Sull'asse delle ordinate si fissano le modalità del carattere esaminato riferite ai diversi tempi.
- ✓ Unendo i punti-immagine tracciati, si ottiene una curva di evidente utilità ai fini dell'interpretazione dei dati statistici.



# Ortogrammi

- ✓ **Gli Ortogrammi** si basano tra intensità o frequenze e superfici rettangolari, e si attua attraverso:
- ✓ **Ortogrammi a Colonne:** rettangoli equidistanti, di uguale base e avanti altezze uguali o proporzionali alle intensità o frequenze da rappresentare (ortogramma a colonne)
- ✓ **Ortogrammi a Nastri:** rettangoli equidistanti, di uguale altezza e aventi basi uguali o proprzionali alle intensità o frequenze da rappresentare





# Indici di Posizione

---


- ✓ **Gli indici di posizione o medie** sono quantità idonee a dare un'idea di insieme (sintesi) di un dato collettivo statistico sostituendosi pertanto a tutti gli altri elementi che lo costituiscono.
- ✓ **Essi si distinguono in:**
  - 1. medie analitiche:** che si determinano considerando tutti i valori di una data variabile statistica e tra queste si annoverano la **media aritmetica**, la **media armonica**, la **media geometrica**, ecc.
  - 2. medie lasche:** che si determinano considerando solo dati elementi della distribuzione e tra queste vi sono la **mediana**, la **moda**, ecc.

# Media Artimetica


✓ La **media aritmetica** di una variabile **X** è un indice di posizione che può essere definita come quella **intensità che può essere sostituita ai singoli valori della variabile, in modo che resti invariata l'intensità globale.**

✓ Si distingue in:

**1. Media Artimetica Semplice:** si ottiene rapportando l'intensità globale di un carattere al numero totale dei casi osservati.


$$\sum_{i=1}^n \frac{x_i}{n}$$

**2. Media Aritmetica Ponderata:** si utilizza nel caso in cui le single modalità della variabile statistica esibiscano **frequenze diverse** differenti da 1


$$\frac{\sum_{i=1}^n x_i \cdot w_i}{\sum_{i=1}^n w_i}$$

# Proprietà della Media Artimetica

---

- ✓ Si noti che la media aritmetica di una variabile statistica viene indicato con il simbolo  $M_S$  mentre la media di una variabile casuale è invece indicate con la lettera greca  $\mu$
- ✓ **Differenza tra variabile statistica e variabile casuale:** una **variabile statistica** deriva dalla classificazione di dati rilevati, cioè viene definita empiricamente una volta conosciuti i dati ed averli classificati. Una **variabile casuale** è strettamente legata al concetto di di esperimento ossia di una prova il cui risultato è incerto.
- ✓ Una media aritmetica di una variabile statistica  **$X$**  è:
- ✓ **interna:** vale a dire il suo valore è sempre maggiore dell'intensità minima e sempre minore dell'intensità massima di una variabile statistica

# Proprietà della Media Artimetica

---

- ✓ **traslativa:** vale a dire che se ai valori della variabile  $X$  si addiziona o si sottrae uno stesso numero, si ottiene una nuova variabile avente media uguale alla media della variabile  $X$  rispettivamente aumentata o diminuita di quel numero.
- ✓ **omogenea:** ossia, se i valori della variabile  $X$  sono moltiplicati o divisi per uno stesso numero  $b$ , si ottiene una nuova variabile  $X'=bX$ , avente media aritmetica uguale alla media aritmetica della variabile  $X$ , rispettivamente, moltiplicata o divisa per il numero  $b$ ;
- ✓ **associativa:** vale a dire che, se i valori della variabile  $X$  sono suddivisi in due o più insiemi, la media aritmetica della variabile statistica è uguale alla media aritmetica delle medie parziali dei singoli insiemi, ponderate con la numerosità degli insiemi stessi.

# Media Quadratica

- ✓ Siano dati un collettivo statistico di  $N$  unità e una variabile statistica:
- ✓  $X = X_1, X_2, \dots, X_n$
- ✓ La **media quadratica** della variabile  $X$  è un indice di posizione e può essere definita **come quella intensità che può essere sostituita ai singoli valori della variabile, in maniera che resti invariata la somma dei quadrati delle intensità.**
- ✓ **Media Quadratica Semplice:** è la radice quadrata del rapporto tra la somma dei quadrati delle intensità e il numero totale dei casi osservati.
- ✓ **Media Quadratica Ponderata:** è identica alla media quadratica semplice solo che le modalità usano frequenze diverse
- ✓ La media quadratica è anche essa interna, associativa, omogenea ma non traslativa

$$M_q = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$$

$$M_q = \sqrt{\frac{\sum_{i=1}^n x_i^2 n_i}{\sum_{i=1}^n n_i}}$$

# Media Armonica

- ✓ Siano dati un collettivo statistico di  $N$  unità e una variabile statistica:
- ✓  $X = X_1, X_2, \dots, X_n$
- ✓ La **media armonica** della variabile  $X$  è un indice di posizione e può essere definita **come quella intensità che può essere sostituita ai singoli valori della variabile, in maniera che resti invariata la somma dei reciproci delle intensità.**
- ✓ **Media Armonica Semplice:** si ottiene rapportando il numero totale dei casi osservati alla somma dei reciproci delle intensità.
- ✓ **Media Armonica Ponderata:** è identica alla media armonica semplice solo che le modalità usano frequenze diverse
- ✓ La media armonica è anche essa interna, associativa, omogenea ma non traslativa

$$\mu_a = \frac{N}{\sum_{i=1}^n \frac{1}{x_i}}$$

$$\mu_a = \frac{\sum_{i=1}^k n_i}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

# Media Geometrica

- ✓ Siano dati un collettivo statistico di  $N$  unità e una variabile statistica:
- ✓  $X = X_1, X_2, \dots, X_n$
- ✓ La **media geometrica** della variabile  $X$  è un indice di posizione e può essere definita **come quella intensità che può essere sostituita ai singoli valori della variabile, in maniera che resti invariato il prodotto delle intensità.**
- ✓ **Media Geometrica Semplice:** si ottiene dalla radice  $N$ -esima del prodotto delle intensità.
- ✓ **Media Geometrica Ponderata:** è identica alla media geometrica solo che le modalità usano frequenze diverse
- ✓ La media geometrica è anche essa interna, associativa, omogenea ma non traslativa

$$\mu_g = \sqrt[N]{\prod_{i=1}^N x_i}$$

$$\mu_g = \sqrt[N]{\prod_{i=1}^N x_i^{N_i}}$$

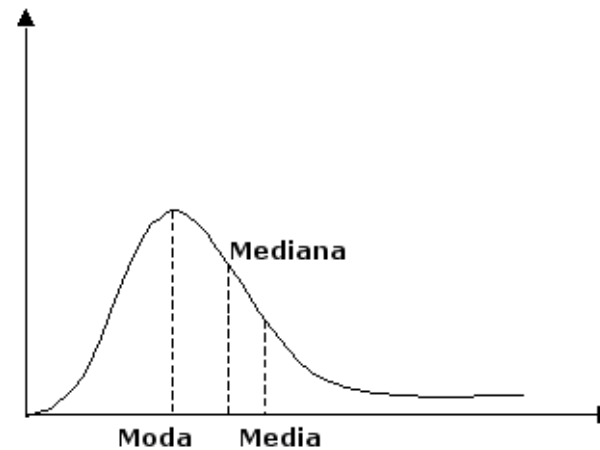
# Relazioni tra medie

- ✓ Per una data variabile statistica, i valori delle quattro medie considerate (aritmetica, quadratica, armonica e geometrica) stanno nella seguente relazione:

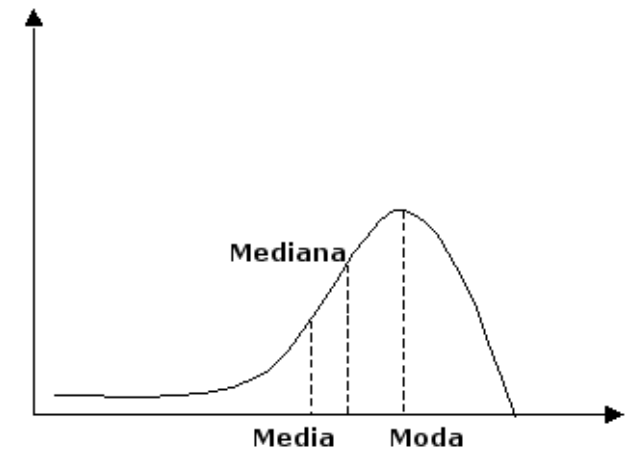
$$M_h < M_g < M < M_q$$

- ✓ Se le intensità di una variabile statistica sono tutte uguali, la relazione che lega le quattro medie è la seguente:

$$M_h = M_g = M = M_q$$



Curva asimmetrica positiva



Curva asimmetrica negativa



# Media di somme di potenze

---

- ✓ La **media delle somme** di potenze è una formula generale comprensiva di numerose medie.
- ✓ Essa è data dalla somma delle potenze di ordine  $s$  dei termini della distribuzione statistica, divisa per l'analoga somma delle potenze di ordine  $s-1$ .

$$M_{s,s-1} = \frac{x_1^s + x_2^s + \dots + x_N^s}{x_1^{s-1} + x_2^{s-1} + \dots + x_N^{s-1}} = \frac{\sum_{i=1}^N x_i^s}{\sum_{i=1}^N x_i^{s-1}}$$

# Moda

---

- ✓ La **moda o valore normale o norma** è un indice di posizione e può essere definito come quella modalità del carattere che presenta la **frequenza** più alta.
- ✓ La **moda** in alcuni casi, può non esistere e, anche se esiste, può non essere unica.
- ✓ Una distribuzione che presenta una sola moda si dice **unimodale**.
- ✓ Una distribuzione che presenta più di una moda si dice **plurimodale**.
- ✓ Una distribuzione che non presenta moda si dice **zeromodale**.
- ✓ La **moda** è una media (lasca) interna, traslativa, omogenea, ma non associativa.
- ✓ Per il calcolo della moda, nel caso di un collettivo che si distribuisce le modalità di un carattere discreto, si possono distinguere diversi casi:
  - ✓ se il **collettivo statistico** si presenta con modalità non raggruppate in classi, la moda si calcola considerando ad una ad una le modalità e verificando quale di queste ultime ha una frequenza maggiore

# Moda

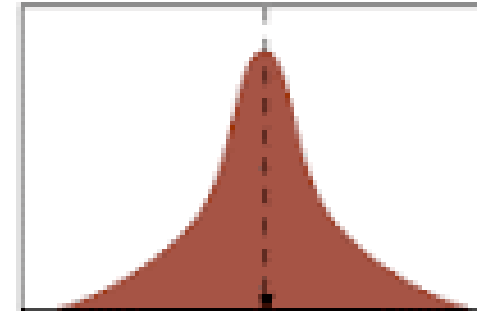
---

- ✓ se il **collettivo statistico** si presenta con modalità raggruppate in classi, con classi di uguale ampiezza, la moda cade nella classe avente frequenza più alta.
- ✓ se il **collettivo statistico** si presenta con modalità raggruppate con classi, di diversa ampiezza, si dividono le frequenze delle classi per il numero di valori del carattere contenuto in esse; la moda cade nella classe che rappresenta il quoziente più alto.
- ✓ Per il calcolo della moda, nel caso di un collettivo che si distribuisce secondo le modalità di un collettivo continuo, si procede in maniera leggermente diverse perchè non si tiene conto più delle frequenze ma delle densità di frequenza, ossia del rapporto tra frequenza di ciascuna classe e ampiezza della classe stessa; pertanto, la **moda è quel valore intorno a cui i casi sono addensati in maggior misura.**

# Moda

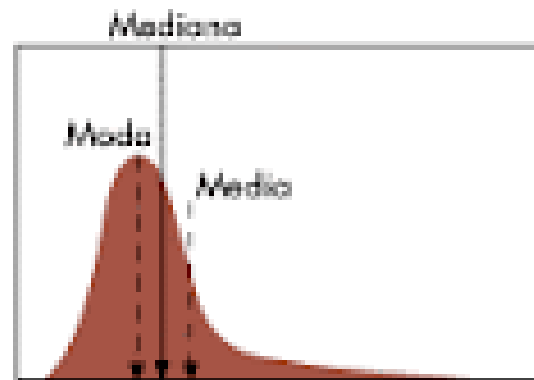
- ✓ Si possono verificare due casi:
- ✓ se il collettivo si presenta con modalità raggruppate in classi della stessa ampiezza, la classe modale è quella che presenta, indifferentemente, maggiore frequenza o densità di frequenza.
- ✓ se il collettivo si presenta con modalità raggruppate in classi di diversa ampiezza, la classe modale è quella che presenta la densità di frequenza più alta

Distribuzione Gaussiana (simmetrica)

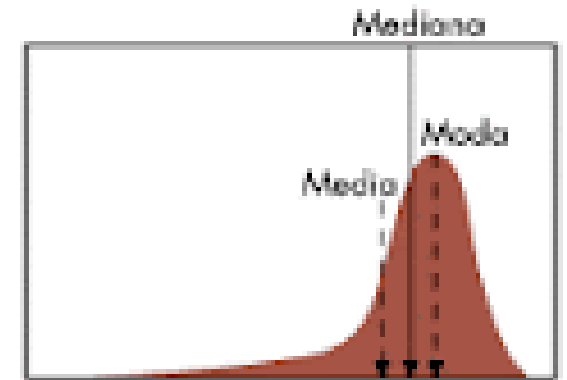


Media  
Mediana  
Moda

Distribuzione asimmetrica



Distribuzione asimmetrica



# Mediana

---

- ✓ Data una variabile statistica, le cui intensità siano in ordine crescente o decrescente, si definisce **mediana** quell'indice di posizione corrispondente al valore della variabile statistica che **bipartisce** la distribuzione, lasciando la metà dei casi a sinistra e la metà dei casi a destra.
- ✓ Per il calcolo della mediana per dati non raggruppati in classi, si identifica il posto centrale e a tal proposito si possono distinguere due casi:
- ✓ I valori osservati sono in numero  $N$  dispari. In tal caso, il posto centrale è unico ed è:

$$c = \frac{N + 1}{2}$$

- ✓ La **mediana** è quella intensità che occupa il posto centrale

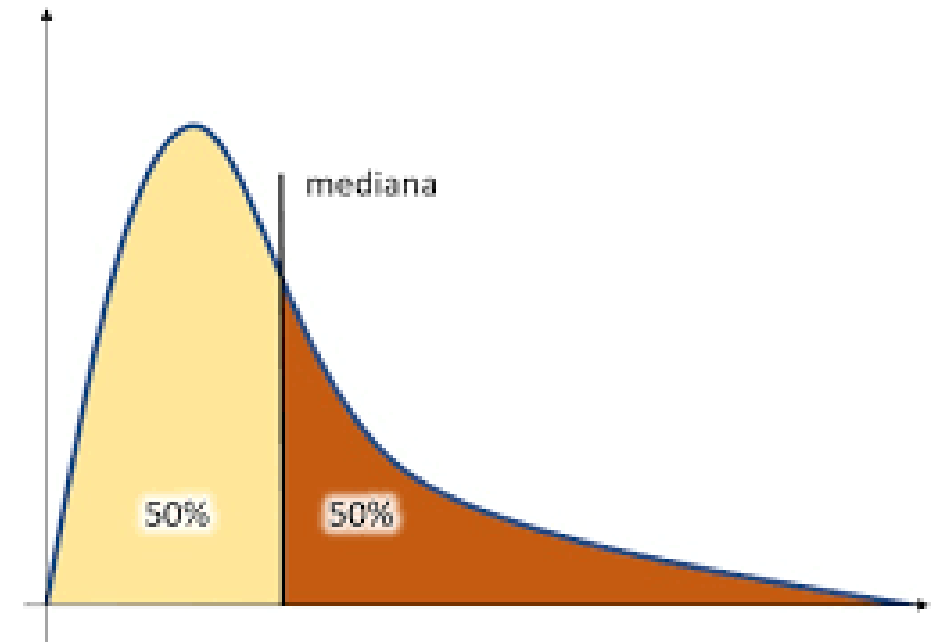
# Mediana

- ✓ Se i valori osservati sono in numero pari, e dunque i posti centrali sono due:

$$c_1 = \frac{N}{2}$$

$$c_2 = \frac{N}{2} + 1$$

- ✓ Si assume come mediana la media aritmetica delle intensità che occupano i due posti centrali considerati.
- ✓ Per il calcolo della mediana, nel caso di un collettivo che si distribuisce secondo le modalità di un carattere discreto con dati raggruppati in classi:



# Mediana

---

- ✓ Per il calcolo della mediana, nel caso di un collettivo che si distribuisce secondo le modalità di un carattere discreto si opera in questo modo:
  1. Si calcola il posto centrale, come nel caso di dati non raggruppati in classi,
  2. Sulla base delle frequenze cumulate si determina la classe in cui cade la mediana
  3. Si determina la differenza tra il posto centrale e la frequenza cumulata fino classe che precedere quella che contiene la mediana
  4. Si divide la frequenza della classe mediana per l'ampiezza della stessa classe
  5. Per ottenere la mediana si divide la differenza di cui al punto 3 per il quoziente di cui al punto 4
- ✓ Nel caso di un collettivo che si distribuisce secondo modalità di un carattere continuo, si considera la ***spezzata cumulativa*** che rappresenta graficamente le frequenze cumulate.

# Percentili

---

- ✓ I **percentili** sono indici di posizione che si definiscono come quelle **intensità che suddividono la distribuzione in due parti, lasciando da un lato una data percentuale dei casi e dall'altro la restante percentuale.**
- ✓ Si distinguono:
- ✓ i **terzili**: che sono 2. Il primo terzile (T1) lascia alla sua sinistra il **33.3%** dei casi e alla sua destra il rimanente **66.7%**; il secondo terzile (T2) lascia alla sua sinistra il **66.7%** dei casi e lascia alla sua destra il **33.7%** dei casi.
- ✓ i **quartile**: i quartili sono 3. Il primo quartile (Q1) lascia alla sua sinistra il **25%** dei casi e alla sua destra il restante **75%**; il secondo quartile (Q2) lascia alla sua sinistra il **50%** dei casi e alla sua destra il restante **50%**; infine il terzo quartile (Q3) lascia alla sua sinistra il **75%** dei casi e alla sua destra il restante **25%**
- ✓ Similmente si definiscono i decili che sono 9, i centili che sono 99 ecc.



# Probabilità e Teoria dell'Informazione

---

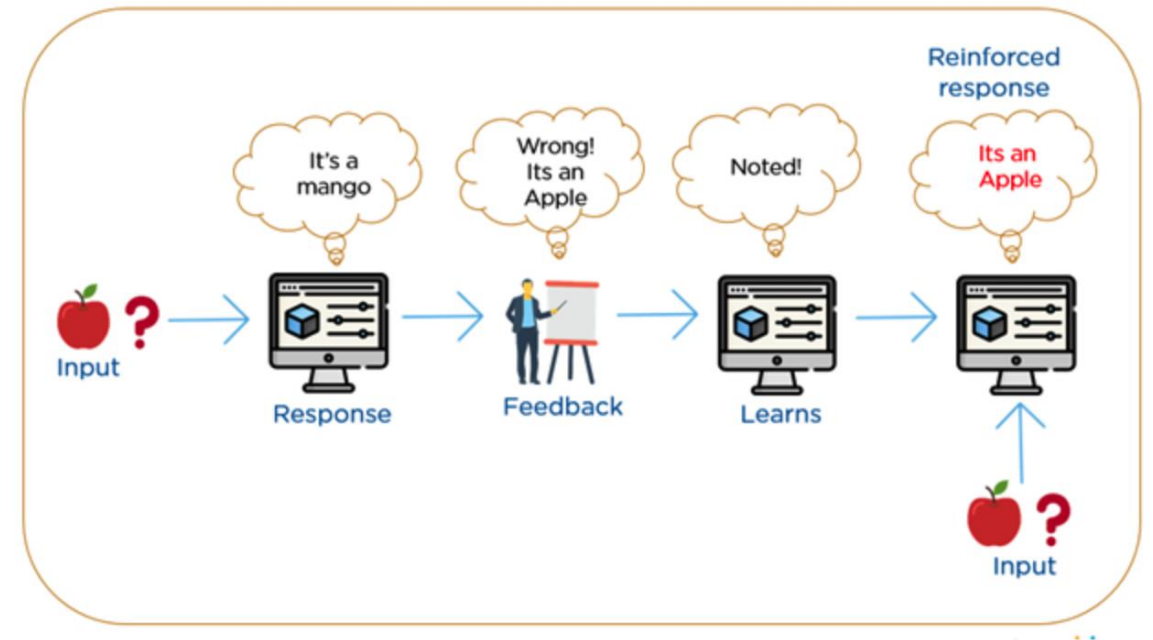
- ✓ La teoria della Probabilità è un framework matematico per rappresentare gli eventi **incerti**.
- ✓ Nelle applicazioni di **Intelligenza Artificiale**, usiamo la teoria della probabilità in due modi principali:
  - 1. La legge delle probabilità ci dice come l'AI dovrebbe ragionare**
  - 2. Possiamo usare la teoria delle probabilità per analizzare il comportamento e le decisioni dei sistemi di AI**
- ✓ **Example:** Robotica
- ✓ **La teoria delle Probabilità** ci permette di fare determinate assunzioni e di **ragionare in condizioni di incertezza**
- ✓ **La teoria delle Informazioni** ci permette di **quantificare la quantità di incertezza** in una distribuzione di probabilità

# Sorgenti di Incertezza

---

- ✓ Il Machine Learning deve sempre avere a che fare con quantità incerte e stocastiche (non deterministiche). Incertezza e stocasticità possono emergere da molte sorgenti.
- ✓ Esistono 3 tipi possibili di sorgenti di incertezza:
  - 1. Stocasticità inerente:** è una stocasticità intrinseca del Sistema che si vuole modellare.  
Esempio: Meccanica Quantistica
  - 2. Osservabilità incompleta:** anche i sistemi deterministici possono apparire stocastici quando non possiamo osservare tutte le variabili che influenzano il comportamento di un Sistema. Esempio: **Monty Hall problem, Reinforcement learning: Alpha Go**
  - 3. Modellazione Incompleta:** quando usiamo un modello che deve scartare alcune informazioni che abbiamo osservato, l'informazione scartata può produrre incertezza nelle predizioni del modello. Esempio: Un modello semplificato del cervello dal momento che non abbiamo risorse computazionali per simulare l'intero cervello

## The Monty Hall Problem



# Il Monty Hall Problem

---

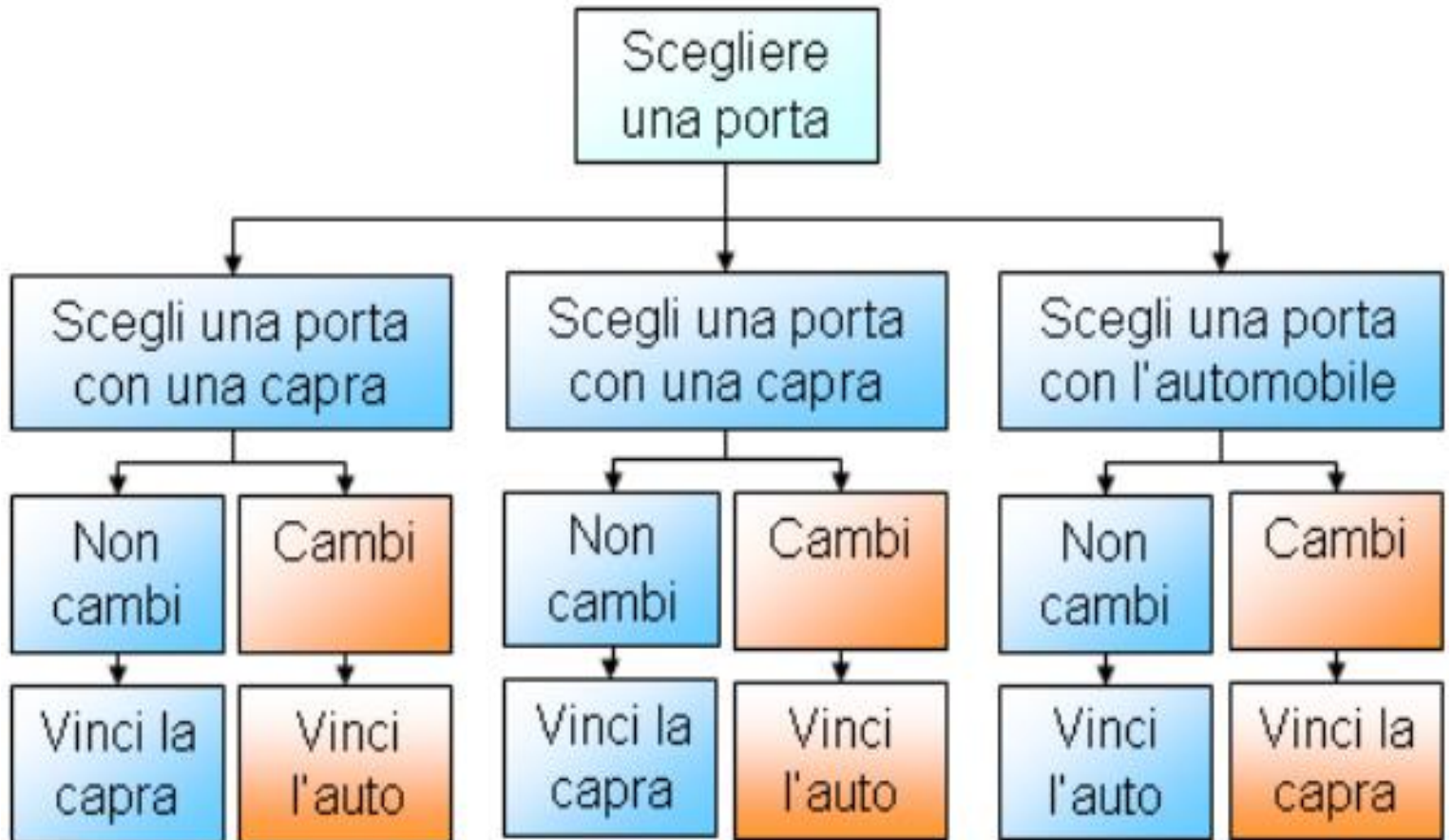
- ✓ Il **problema di Monty Hall (o paradosso di Monty Hall)** è un famoso problema della teoria delle probabilità, legato al gioco a premi statunitense "Let's make a deal". Prende il nome da quello del conduttore dello show, Maurice Halprin, noto con lo pseudonimo di Monty Hall.
- ✓ Il problema è anche noto come **paradosso di Monty Hall** in quanto la soluzione può apparire controintuitiva, ma non si tratta di una vera antinomia, ossia la compresenza di due affermazioni contraddittorie che possono essere entrambe giustificate o dimostrate, in quanto **non genera contraddizioni logiche**.
- ✓ Nel gioco vengono mostrate al concorrente tre porte chiuse; dietro ad una si trova un'automobile, mentre ciascuna delle altre due nasconde una **capra**. Il giocatore può scegliere una delle tre porte, vincendo il premio corrispondente. Dopo che il giocatore ha selezionato una porta, ma non l'ha ancora aperta, il conduttore dello show, che conosce ciò che si trova dietro ogni porta, apre una delle altre due rivelando una delle due capre, e offre al giocatore la possibilità di cambiare la propria scelta iniziale, passando all'unica porta restante.

# Il Monty Hall Problem

- ✓ Cambiare la porta migliora le **chance** del giocatore di vincere l'automobile, portandole da  $1/3$  a  $2/3$ .
- ✓ Le possibilità di vittoria aumentano per il giocatore se cambia la propria scelta? La risposta è **si**, le probabilità di trovare l'automobile raddoppiano.
- ✓ La **soluzione** può essere illustrate come segue: ci sono tre scenari possibili aventi probabilità di  $1/3$ :
  1. Il giocatore sceglie la capra numero 1. Il conduttore sceglie l'altra capra, la numero 2. Cambiando il giocatore vince l'auto.
  2. Il giocatore sceglie la capra numero 2. Il conduttore sceglie l'altra capra, la numero 1. Cambiando il giocatore vince l'auto.
  3. Il giocatore sceglie l'auto. Il conduttore sceglie una capra, non importa quale. Cambiando il giocatore trova l'altra capra.
- ✓ Nei primi due scenari, cambiando il giocatore vince l'auto, nel terzo scenario il giocatore che cambia non vince. Dal momento che la strategia di "cambiare" porta alla vittoria in due casi su tre, le chance di vittoria adottando la strategia sono  $2/3$ .

# Il Monty Hall Problem

- ✓ Dopo la scelta del giocatore, il presentatore apre una porta (egli sa dove si trova l'auto) mostrando una capra. Qualsiasi cosa ci sia dietro la scelta iniziale del giocatore, egli cambiando scelta ha il 66.7% di probabilità di vincere l'auto, non cambiandola avrebbe il 33.3%



# Perchè la probabilità ?

In molti casi, many cases, it is more practical to use a **simple** but uncertain rule rather than a complex but certain one, even if the true rule is deterministic and our modeling system has the fidelity to accommodate a complex rule.

For example, the simple rule “**Most birds fly**” is **cheap** to develop and is broadly useful, while a rule of the form, “Birds fly, except for very young birds that have not yet learned to fly, sick or injured birds that have lost the ability to fly, flightless species of birds including the cassowary, ostrich and kiwi. . .” is **expensive** to develop, maintain and communicate and, after all this effort, is still **fragile** and prone to **failure**.

Probability can be seen as the **extension** of logic to deal with uncertainty.

Probability theory provides a set of **formal** rules for determining the **likelihood** of a proposition being true given the likelihood of other propositions

# Types of probability

Frequentist probability:

**Frequency** of events

Example: The chance of drawing a certain hand in poker

Fixed model, different data ( We run the same experiments each time with different data)

Bayesian probability:

A degree of **belief**

Example: A doctor saying a patient has a 40 percent chance of having a flu

Fixed data and different models (We use the same belief to check the uncertainty of different models and update our beliefs)

Based on **Bayes rule** which we talk about later in the presentation



# Frequentist

[repeat repeat repeat]



# Bayesian

[observe, guess, experiment]



# Random variables

A random variable is a variable that can take on **different** values randomly.

On its own, a random variable is just a **description** of the states that are possible; it must be coupled with a **probability distribution** that specifies how **likely** each of these states are.

Random variables may be discrete or continuous.

A **discrete** random variable is one that has a finite or countably infinite number of states. Note that these states are not necessarily the integers; they can also just be named states that are not considered to have any numerical value.

A **continuous** random variable is associated with a real value

**Probability distribution:** description of how **likely** a random variable or set of random variables is to take on each of its **possible states**. The way we describe probability distributions depends on whether the variables are discrete or continuous.

# Types of random variables

## Discrete random variable:

Finite number of states, not necessarily integers they can also be a **named states** (that are not considered to have any numerical value)

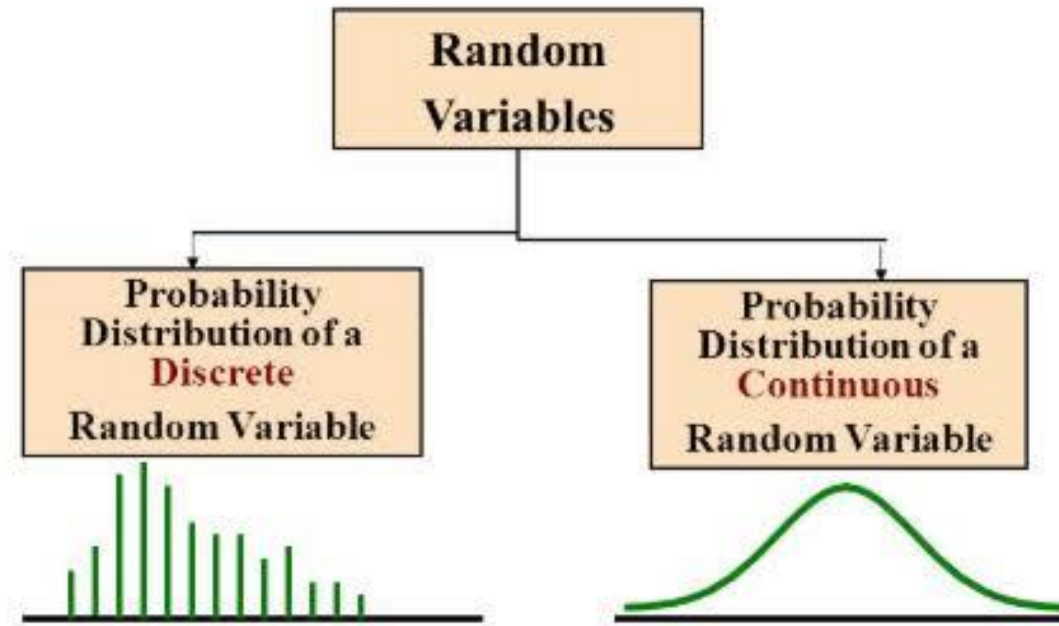
**Example:** Coin toss (2 states), Throwing a dice (6 states), Drawing a card from a deck of cards (52 states) etc.,

## Continuous random variable:

Must be associated with a **real** value

**Example:** Rainfall on a given day (in centimeters), Stock price of a company, Temperature of a given day

# Random variables



# Probability mass function (PMF)

Probability distribution over discrete random variables is referred to as a **probability mass function**(PMF)

Maps from a state of a random variable to the probability of that random variable taking on that state.

The probability that  $x=x$  is denoted as  $P(x)$ , with a probability of 1 indicating that  $x=x$  is **certain** and a probability of 0 indicating that  $x=x$  is **impossible**.

Criterion for being a PMF:

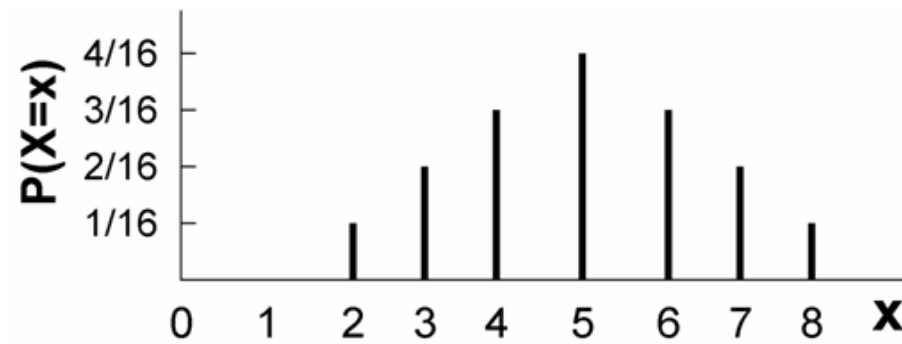
The domain of  $P$  must be the set of all possible states of  $x$

$$0 \leq P(x) \leq 1$$

Summation of all possible states of  $P(x) = 1$

# Probability mass function (PMF)

x	<u>P(x)</u>
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16



# Types of Distributions

## Joint probability distribution:

Probability mass function that can act on many variables at the same time.

Such a probability distribution over many variables is known as a joint probability distribution.

**Example:**  $P(x=x, y=y)$  denotes the probability that  $x=x$  and  $y=y$  simultaneously.

We may also write  $P(x, y)$  for brevity

## Uniform distribution:

Each state of the distribution is equally likely

$P(x = x_i) = 1/k$  where,  $k$  is the total number of possible states

Completely normalized equal distribution with equally likely states

# Probability density function (PDF)

When working with continuous random variables, we describe probability distributions using a probability density function (PDF) rather than a probability mass function (PMF).

Statistical expression used in probability theory as a way of representing the range of possible values of a continuous random variable.

To be a probability density function, a function  $p$  must satisfy the following properties:

The domain of  $p$  must be the set of all possible states of  $x$ .

$\forall x \in x, p(x) \geq 0$ .

Note that we do not require  $p(x) \leq 1$ .

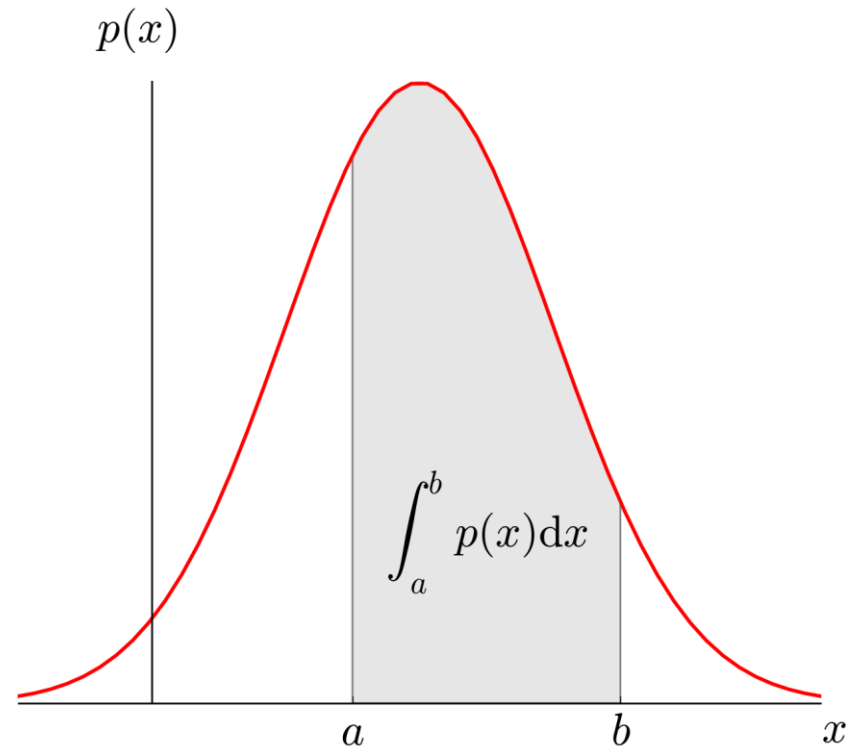
$$\int p(x)dx = 1$$

In other words a PDF,  $p(x)$  does not give the probability of a specific state directly; instead the probability of landing inside an **infinitesimal region** with volume  $\delta x$  is given by  $p(x)\delta x$ .

We can integrate the density function to find the actual probability mass of a set of points. Specifically, the probability that  $x$  lies in some set  $S$  is given by the integral of  $p(x)$  over that set.



# Probability density function (PDF)



# PDF and Machine Learning

The Probability Density Function works by **conceptualizing** the probabilities of a continuous random event occurring by defining a **range**, or **interval**.

For example, if one wanted to calculate the probability that a **specific** temperature, say 70 degrees, will be reached, they may turn to a PMF, as the variable is defined in discrete terms. However, if one wanted to calculate the probability that a temperature **between** 70-75 degrees will be reached, they may use a PDF, as the variable is defined as a **range** with infinite discrete values.

Since the PDF defines probabilities with intervals, the **probability of a single discrete value is defined as zero**, since it does not have a range.

A Probability Density Function is a tool used by machine learning algorithms and neural networks that are trained to calculate probabilities from continuous random variables.

For example, a neural network that is looking at **financial markets** and attempting to guide **investors** may calculate the probability of the **stock market rising 5-10%**. To do so, it could use a PDF in order to calculate the total probability that the continuous random variable range will occur.

# Marginal probability

The probability distribution over a subset of all variables

Oftentimes we will be working with Marginal probability distributions in AI applications since we don't have all the variables or data points that is one of the sources of uncertainty that we talked about earlier.

With discrete random variables:

If we know  $P(x,y)$ , we can find  $P(x)$  with the **sum rule**, more on this in the next slide

$P(x=x) = \sum_y P(x=x, y=y)$  (so  $P(x)$  is the summation over all possible  $y$  values)

With continuous random variables:

$$p(x) = \int p(x,y) dy$$

# Marginal probability

$f(x, y)$  is defined by the following table

	$x = \underline{0}$	$x = \underline{1}$	$x = \underline{2}$
$y = 1 \checkmark$	<u>0.3</u>	<u>0.2</u>	<u>0.1</u>
$y = 2 \checkmark$	0.1	0.1	0.2
$f_x(x)$	0.4	0.3	0.3

$$f_x(x) = \begin{cases} 0.4 & x=0 \\ 0.3 & x=1 \\ 0.3 & x=2 \end{cases}$$

$$f_y(y) = \begin{cases} 0 & y=1 \\ 0.6 & y=2 \end{cases}$$

$$f_y(1) = P(Y=1) = 0.6$$

# Conditional probability

The probability of some event, given that some other event has happened

$$P(y = y \mid x = x) = \frac{P(y = y, x = x)}{P(x = x)}.$$

Conditional probability that  $y = y$  given  $x = x$

Widely used in Bayesian inference to form the **beliefs**.

Its very important to understand marginal and conditional probability because in general we are not going to have **pristine** probability distributions. We are going to be working either with the **subset** of variables or we are going to be adding criterion which will essentially be saying what is the probability of  $y$  given  $x$  or given a certain criterion

# Expectation (of a random variable)

The expectation of some function  $f(x)$  ( $f(x)$  is the random variable for  $x$ ) with respect to some probability distribution  $P(x)$  is the mean value  $f$  takes on when it is drawn from  $P$ .

Mean or average value

For discrete random variables:

For continuous

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x) f(x)$$

It can be written as

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x) f(x) dx.$$

and consequently **standard deviation**

# Variance (of a random variable)

The expectation of **squared deviance** of a random variable from its mean is referred to as variance ( $\sigma^2$ ).

Measures how **far** random numbers drawn from a probability distribution  $P(x)$  are spread out from their average value (Expectation value).

$$\text{Var}(f(x)) = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right]$$

**Standard deviation( $\sigma$ ):** Square root of the variance

Variance and Standard deviation are two measures of **spread** which are widely used in machine learning. They come all the time because in machine learning we want to know what kind of distributions that our input variables have. so we will find these 2 and understand **patterns** in our data at a single glance.

These 2 are also a great measure to generating new probability distributions of our data.

For example, We can add Gaussian noise to the images to make the AI system generalize the new images in a better way.

# Covariance (of a random variable)

Measure of how much two variables are linearly related to each other.

$$\text{Cov}(f(x), g(y)) = \mathbb{E} [(f(x) - \mathbb{E} [f(x)]) (g(y) - \mathbb{E} [g(y)])] .$$

**High at**

**Positive** - both variables take on large values simultaneously.

**Negative** – variables take on large values at different times.

The notion of covariance and dependence are related but distinct concepts.

**Independence** is different from covariance because it also includes non linear relationships. It is possible that two variables are dependent but have 0 covariance since covariance only measures linear correlation between two variables it doesn't account for non linear correlation.

Covariance is effected by scale, so the larger your variables are the larger the covariance is going to be.

In machine learning, we can exploit the property of covariance to either **compress** your data or in getting better results



# Covariance matrix

Covariance matrix of a random vector  $\mathbf{x}$  is  $n \times n$  matrix, such that

$$\text{Cov}(\mathbf{x})_{i,j} = \text{Cov}(x_i, x_j).$$

The diagonal elements of the covariance matrix given the variance:

$$\text{Cov}(x_i, x_i) = \text{Var}(x_i).$$

While applying machine learning algorithms to our data, almost all of our input, weights, activations and outputs are going to be **vectors** and **matrices**. So often times we will be applying covariance. So we will be creating the covariance matrix so that it can be applied to our analysis. Because we will be working exclusively with **matrix forms and notations** exclusively for all the steps of a machine learning project.

# Common probability distributions

Probability distributions are useful in many contexts in machine learning.

Some of them are as follows –

- Bernoulli

- Normal

- Poisson

- Binomial

# Bernoulli

The Bernoulli distribution is a distribution over a single binary random variable.

It is controlled by a single parameter  $\phi \in [0, 1]$ , which gives the probability of the random variable being equal to 1.

It has the following properties

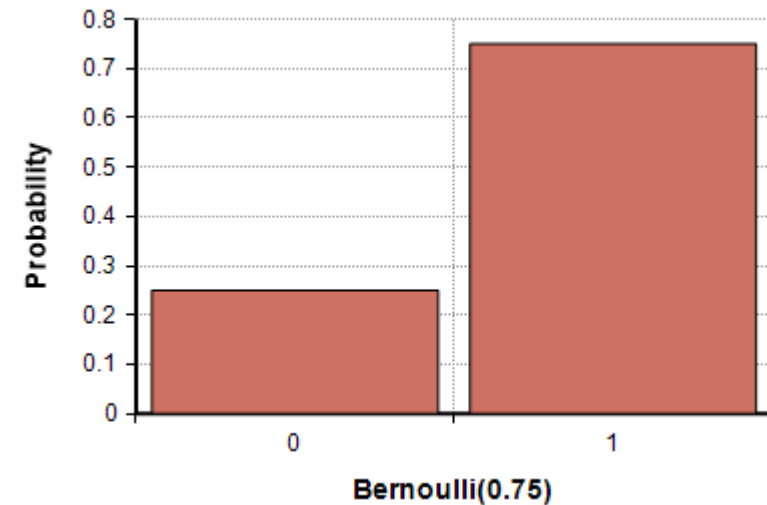
$$P(x = 1) = \phi$$

$$P(x = 0) = 1 - \phi$$

$$P(x = x) = \phi^x (1 - \phi)^{1-x}$$

$$\mathbb{E}_x[x] = \phi$$

$$\text{Var}_x(x) = \phi(1 - \phi)$$



# Normal

The most commonly used distribution over real numbers, also known as **Gaussian** distribution

The two parameters  $\mu \in \mathbf{R}$  and  $\sigma \in (0, \infty)$  control the normal distribution.

The parameter  $\mu$  gives the coordinate of the central peak.

This is also the mean of the distribution:  $\mathbf{E}[\mathbf{x}] = \mu$ . The standard deviation of the distribution is given by  $\sigma$ , and the variance by  $\sigma^2$

Normal distributions are a **sensible** choice for many applications.

The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many **complicated** systems can be modeled successfully as normally.

Normal distribution **encodes** the maximum amount of **uncertainty** over the real numbers. We can thus think of the normal distribution as being the one that inserts the **least** amount of prior knowledge into a model

# Normal

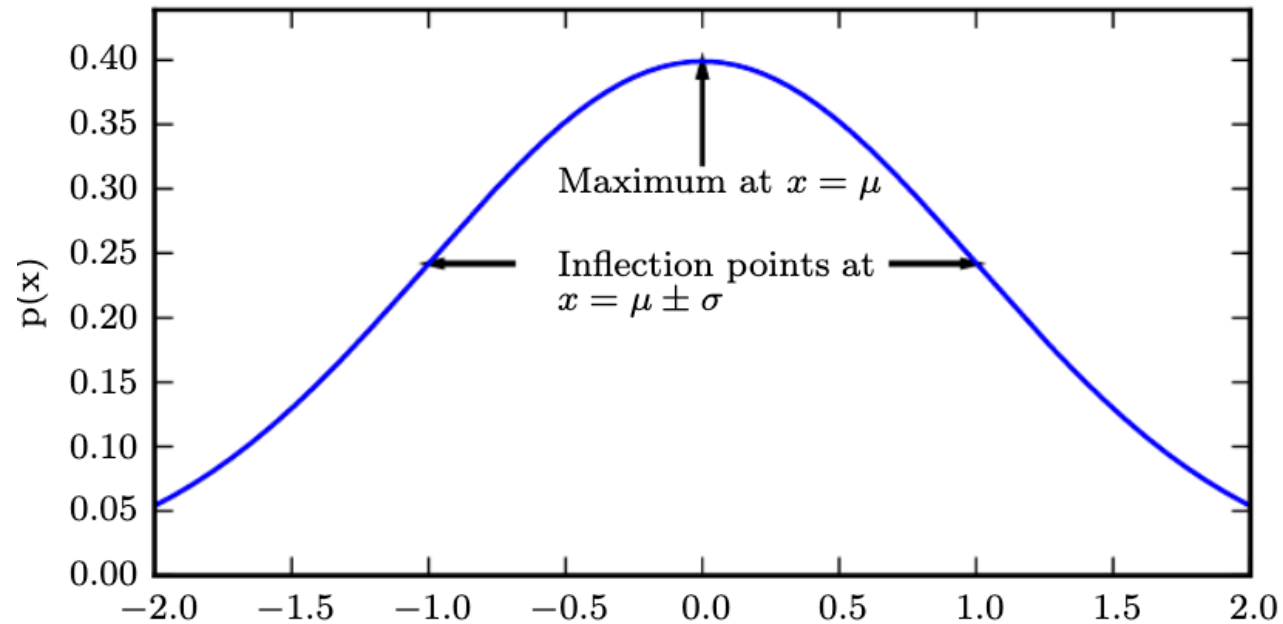


Figure: The normal distribution.

**Note:** The normal distribution  $N(x; \mu, \sigma^2)$  exhibits a classic “bell curve” shape, with the  $x$  coordinate of its central peak given by  $\mu$ , and the width of its peak controlled by  $\sigma$ . In this example, we depict the **standard normal distribution**, with  $\mu = 0$  and  $\sigma = 1$

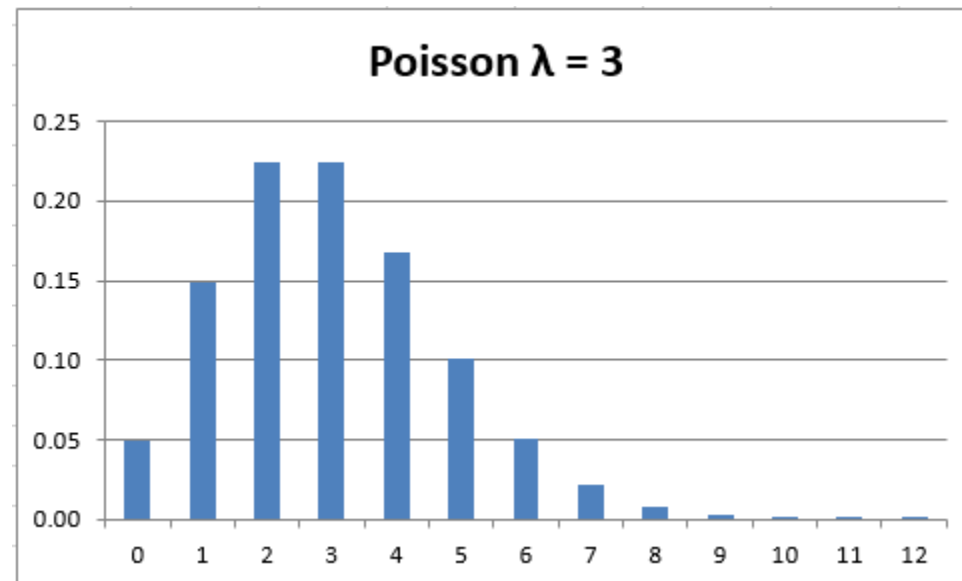
# Poisson

Is a discrete probability distribution that expresses the probability of a given number of events occurring in a **fixed interval of time or space**.

Some examples that may follow a Poisson distribution include,

Number of phone calls received by a call center per hour

Number of patients arriving in an emergency room between 10 and 11 pm



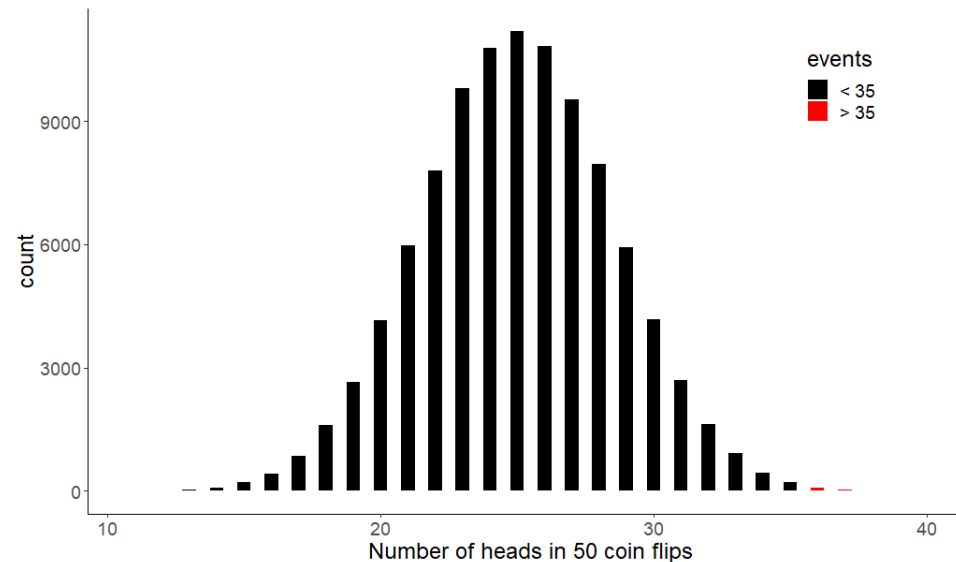
# Binomial

The binomial distribution with parameters  $n$  and  $p$  is the **discrete** probability distribution of the number of **successes** in a sequence of  $n$  independent experiments,

The binomial distribution is frequently used to model the number of successes in a sample of size  $n$  drawn with replacement from a population of size  $N$ .

Binomial probability distributions help us to understand the likelihood of rare events and to set probable expected ranges.

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$



# Useful properties of common distributions

Certain functions arise often while working with probability distributions, especially the probability distributions used in deep learning models.

Some of them are:

Sigmoid

Softmax

## **Sigmoid:**

The logistic sigmoid is commonly used to produce the  $\Theta$  parameter of a Bernoulli distribution because its range is  $(0,1)$ , which lies within the valid range of values for the  $\Theta$  parameter.

The sigmoid function saturates when its argument is very positive or very negative, meaning that the function becomes very flat and insensitive to small changes in its input.

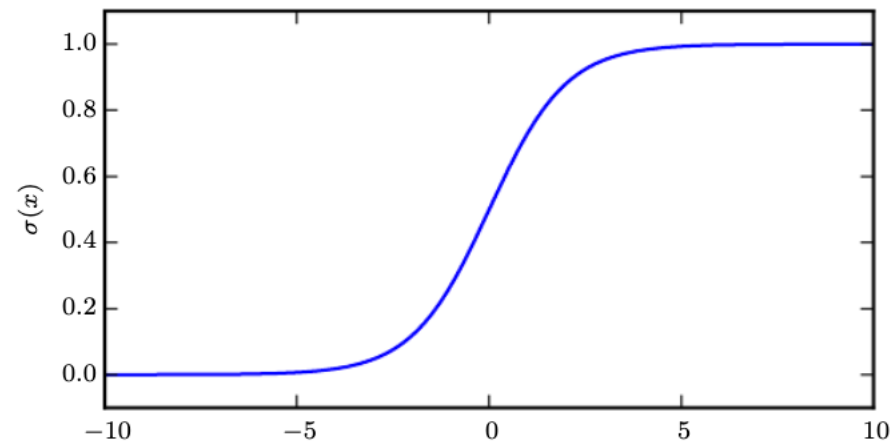
## **Softmax:**

The softmax function can be useful for producing the  $\beta$  or  $\sigma$  parameter of a normal distribution because its range is  $(0, \infty)$ .

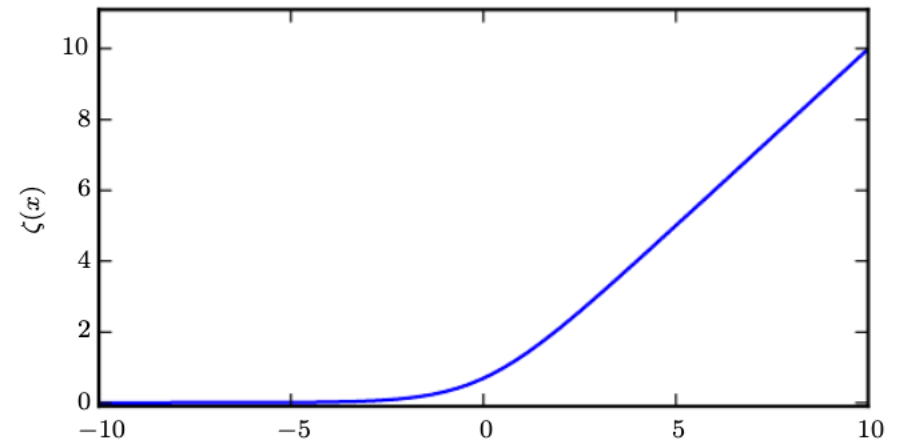
It also arises commonly when manipulating expressions involving sigmoids.



# Useful properties of common distributions



Sigmoid function



Softmax function

# Law of large numbers

The law of large numbers is a theorem from probability and statistics that suggests that the average result from repeating an experiment multiple times will better approximate the true or expected underlying result.

**The law of large numbers explains why casinos always make money in the long run.**

— Page 79, Naked Statistics: Stripping the Dread from the Data, 2014.

We have an intuition that **more observations** are better. This is the same intuition behind the idea that if we collect **more data**, our sample of data will be more representative of the problem domain.

Has important implications in applied machine learning.

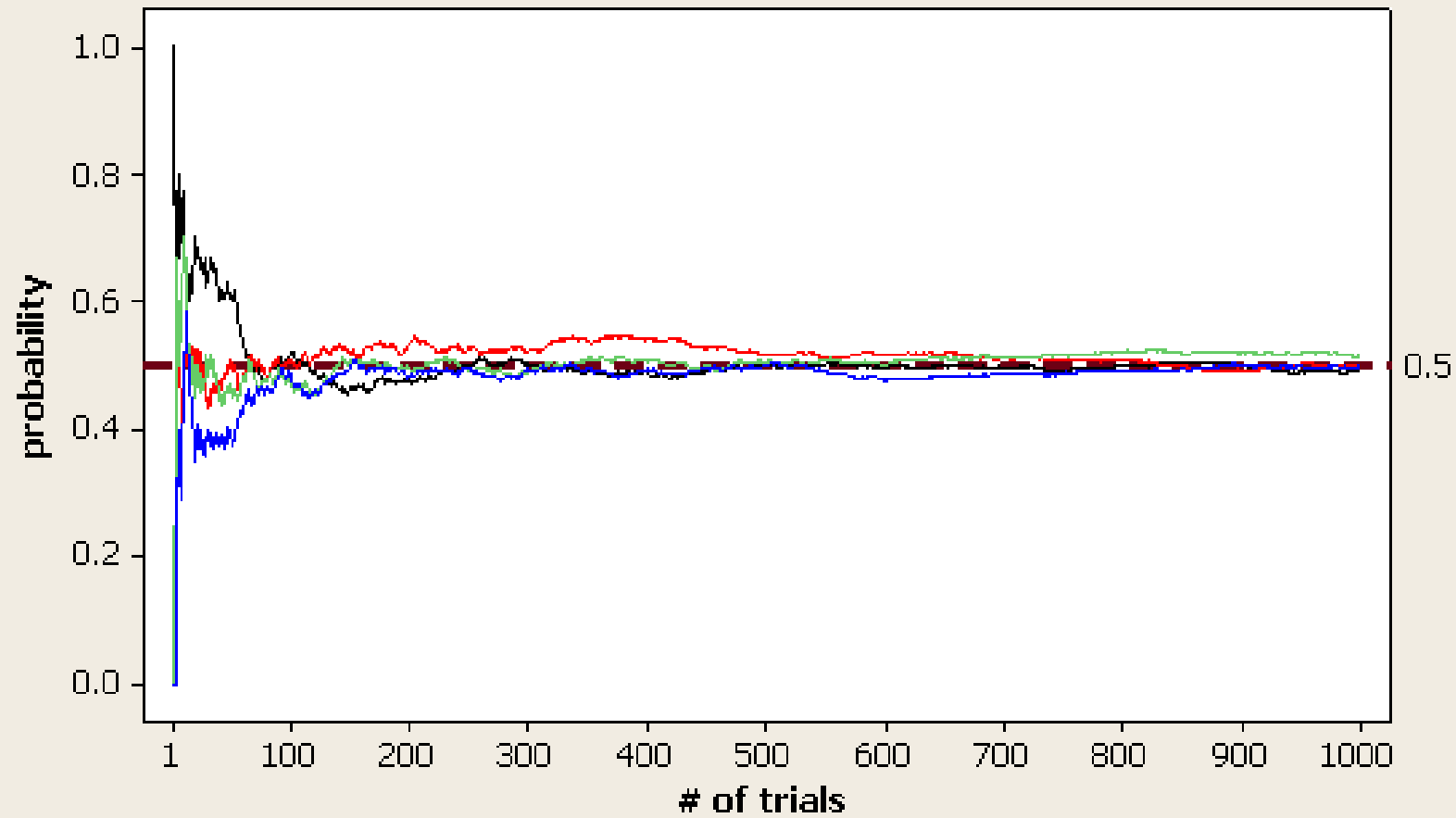
The law of large numbers is critical for understanding the selection of training datasets, test datasets, and in the evaluation of model skill in machine learning.

States that the mean of a large sample is close to the mean of the distribution.

The law reminds us to repeat the experiment in order to develop a large and representative sample of observations before we start making inferences about what the result means.

# Law of Large Numbers

$p = 0.5$



# Law of large numbers (Implications in Machine Learning)

The law of large numbers has important implications in applied machine learning.

Let's take a moment to highlight a few of these implications –

## **Training data:**

The data used to train the model must be representative of the observations from the domain. This really means that it must contain enough information to generalize to the true unknown and underlying distribution of the population.

This is easy to conceptualize with a single input variable for a model, but is also just as important when you have multiple input variables.

There will be unknown relationships or dependencies between the input variables and together, the input data will represent a multivariate distribution from which observations will be drawn to comprise your training sample. Keep this in mind during data collection, data cleaning, and data preparation.

You may choose to exclude sections of the underlying population by setting hard limits on observed values (e.g. for outliers) where you expect data to be too sparse to model effectively.

# Law of large numbers (Implications in Machine Learning)

## **Test data:**

The thoughts given to the training dataset must also be given to the test dataset.

This is often neglected with the blind use of 80/20 splits for train/test data or the blind use of 10-fold cross-validation, even on datasets where the size of 1/10th of the available data may not be a suitable representative of observations from the problem domain.

## **Model skill evaluation:**

Consider the law of large numbers when presenting the estimated skill of a model on unseen data.

It provides a defense for not simply reporting or proceeding with a model based on a skill score from a single train/test evaluation.

It highlights the need to develop a sample of multiple independent (or close to independent) evaluations of a given model such that the mean reported skill from the sample is an accurate enough estimate of population mean.

# The central limit theorem

The central limit theorem describes the shape of the distribution of sample means as a Gaussian or Normal distribution

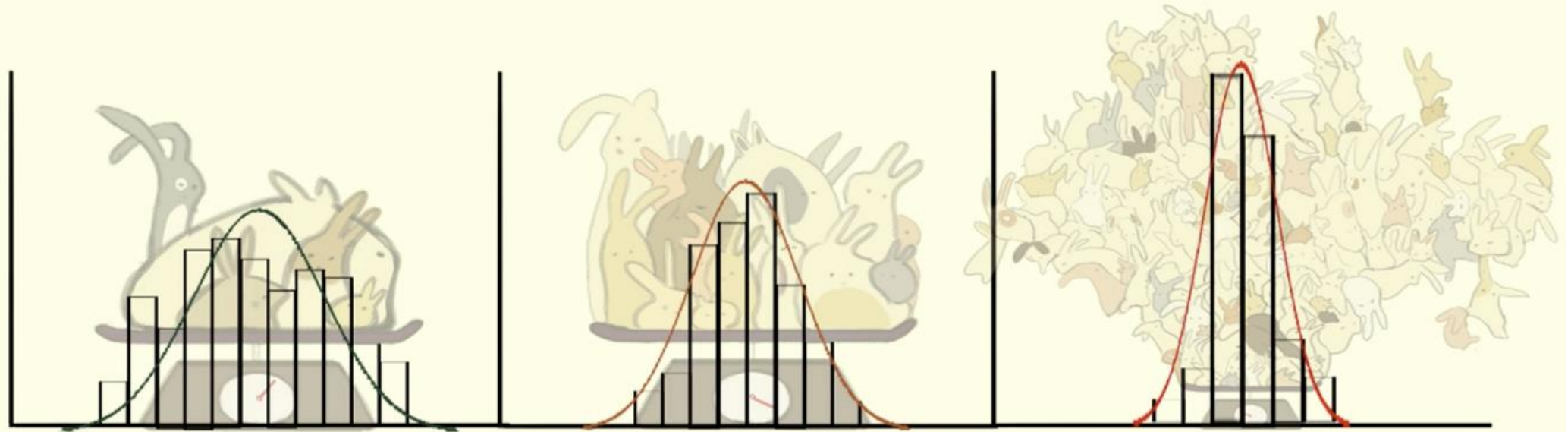
The theorem states that as the size of the sample increases, the distribution of the mean across multiple samples will approximate a Gaussian distribution.

It demonstrates that the distribution of errors from estimating the population mean fit a normal distribution.

This estimate of the Gaussian distribution will be more accurate as the size of the samples drawn from the population is increased. This means that if we use our knowledge of the Gaussian distribution in general to start making inferences about the means of samples drawn from a population, that these inferences will become more useful as we increase our sample size.

The central limit theorem does not state anything about a single sample mean (like law of large numbers); instead, it is broader and states something about the shape or the distribution of sample means.

# Central Limit Theorem



The averages of samples have **approximately normal distributions**

Sample size  $\longrightarrow$  **Bigger**  
Distribution of Averages  $\longrightarrow$  **More normal and narrower**

Image Credits: Casey Dunn & Creature Cast on [Vimeo](#)

# The central limit theorem (Implications in Machine Learning)

The central limit theorem has important implications in applied machine learning. The theorem does inform the solution to linear algorithms such as linear regression, but not exotic methods like artificial neural networks that are solved using numerical optimization methods.

## **Significance Tests:**

In order to make inferences about the skill of a model compared to the skill of another model, we must use tools such as statistical significance tests.

These tools estimate the likelihood that the two samples of model skill scores were drawn from the same or a different unknown underlying distribution of model skill scores.

If it looks like the samples were drawn from the same population, then no difference between the models skill is assumed, and any actual differences are due to statistical noise.

The ability to make inference claims like this is due to the central limit theorem and our knowledge of the Gaussian distribution and how likely the two sample means are to be a part of the same Gaussian distribution of sample means.



# The central limit theorem (Implications in Machine Learning)

## Confidence Intervals:

Once we have trained a final model, we may wish to make an inference about how skillful the model is expected to be in practice.

The presentation of this uncertainty is called a confidence interval.

We can develop multiple independent (or close to independent) evaluations of a model accuracy to result in a population of candidate skill estimates.

The mean of these skill estimates will be an estimate (with error) of the true underlying estimate of the model skill on the problem.

With knowledge that the sample mean will be a part of a Gaussian distribution from the central limit theorem, we can use knowledge of the Gaussian distribution to estimate the likelihood of the sample mean based on the sample size and calculate an interval of desired confidence around the skill of the model.

# Maximum Likelihood

The goal of maximum likelihood is to **fit** an **optimal** statistical distribution to some data.

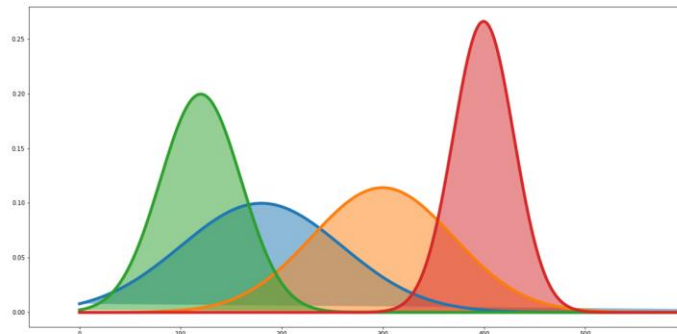
This makes the data **easier** to work with, makes it more general, allows us to see if **new** data follows the same distribution as the previous data, and lastly, it allows us to classify **unlabeled** data points.

The reason you want to fit a distribution to your data is it can be easier to work with and it is also more general – it applies to every experiment of the same type

To maximize the likelihood of the event of interest in our analysis.

In everyday conversation, the probability and likelihood mean the same thing. However in statistical analysis, "likelihood" refers to finding the optimal value for the mean or standard deviation for a distribution given a bunch of observed measurements. This is how we fit a distribution to data

MLE (Maximum likelihood estimate) tells us which curve has the highest likelihood of fitting our data.



# Bayes rule

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Here,

**P(A/B)** is **posterior** as the **conditional probability** of event A given event B

**P(A)** is our **prior**, or the initial **belief** of the probability of event A

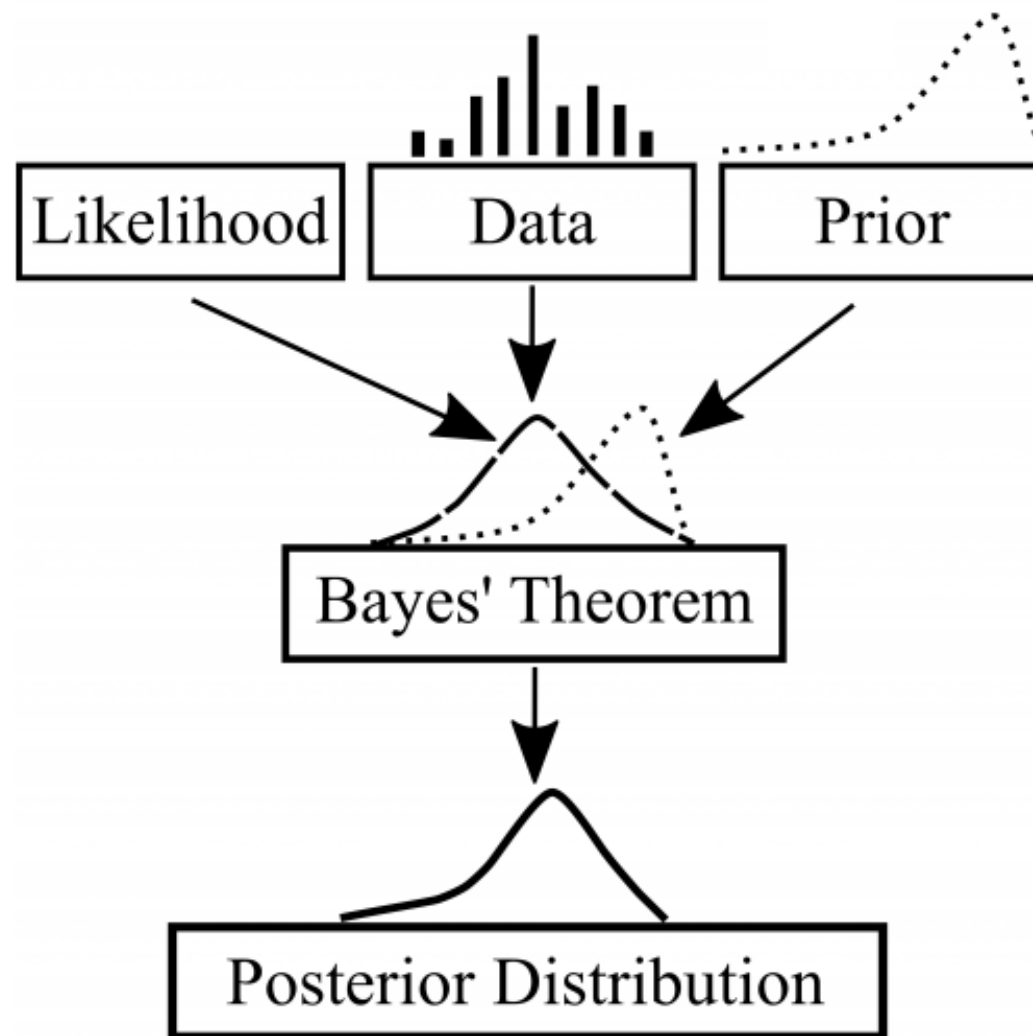
**P(B|A)** is the **likelihood** (also a conditional probability), which we **derive** from our **data**, and

**P(B)** is a **normalization** constant to make the probability distribution sum to 1

The general form of Bayes' Rule in statistical language is the posterior probability equals the likelihood times the prior divided by the normalization constant.

This short equation leads to the entire field of Bayesian Inference, an effective method for reasoning about the world.

# Bayesian inference



# Bayesian inference

Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their **beliefs** in the evidence of new data.

A Bayesian is one who, vaguely expecting a **horse**, and catching a glimpse of a **donkey**, strongly believes he has seen a **mule**

The fundamental idea of Bayesian inference is to become “**less wrong**” with **more data**.

Bayesian data analysis is based on the following two principles:

Probability is interpreted as a measure of uncertainty, whatever the source. Thus, in a Bayesian analysis, it is standard practice to **assign** probability distributions not only to unseen data, but also to parameters, models, and hypotheses.

**Uncertainty** is quantified both before and after the collection of data and Bayes’ formula is used to update our beliefs in light of the new data.

Bayesian analysis, a method of statistical inference that allows one to combine **prior** information about a **population** parameter with **evidence** from information contained in a sample to **guide** the statistical inference process

# Bayes theorem in Machine learning

Bayes' theorem can be used in both regression, and classification.

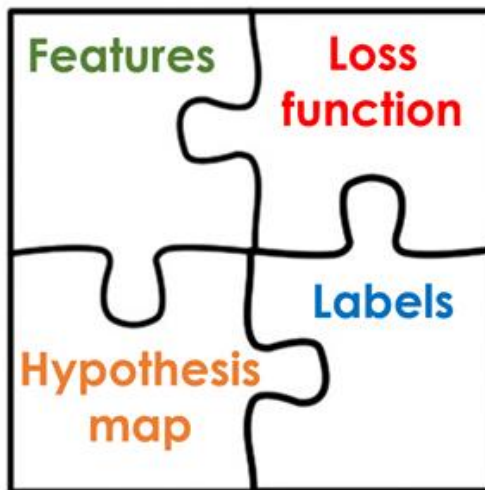
Generally, in Supervised Machine Learning, when we want to train a model the main building blocks are

a set of data points that contain **features** (the attributes that define such data points),

the **labels** of such data point (the numeric or categorical tag which we later want to predict on new data points),  
and

a **hypothesis** function or model that

We also have a **loss** function, which we want to reduce to achieve



corresponding labels.

predictions of the model and the real labels

# Bayes theorem in Machine learning

## Regression:

Usir

$$y = \theta_0 + \theta_1 x$$

Equation describing a linear model

After having trained the model with the available data we would get a value for both of the  $\theta$ s. This training can be performed by using an iterative process like gradient descent or another probabilistic method like Maximum Likelihood. In any way, we would just have ONE single value for each one of the parameters.

## Using Bayes approach –

When we use Bayes' theorem for regression, instead of thinking of the parameters (the  $\theta$ s) of the model as having a single, unique value, we represent them as parameters having a certain **distribution**: the **prior** distribution of the **parameters**.

What this means is that our parameter set (the  $\theta$ s of our model) is **not constant**, but instead has its own **distribution**. Based on previous knowledge (from experts for example, or from other works) we make a first hypothesis about the distribution of the parameters of our model. Then, as we train our models with more data, this distribution gets updated and grows more exact (in

# References and further reading

<http://www.deeplearningbook.org/>

<https://deepai.org/machine-learning-glossary-and-terms/probability-density-function>

<https://towardsdatascience.com/an-intuitive-real-life-example-of-a-binomial-distribution-and-how-to-simulate-it-in-r-d72367fbc0fa>

<https://towardsdatascience.com/bayes-rule-applied-75965e4482ff>

<https://towardsdatascience.com/central-limit-theorem-in-action-1d4832599b7f>

[https://en.wikipedia.org/wiki/List\\_of\\_probability\\_distributions](https://en.wikipedia.org/wiki/List_of_probability_distributions)

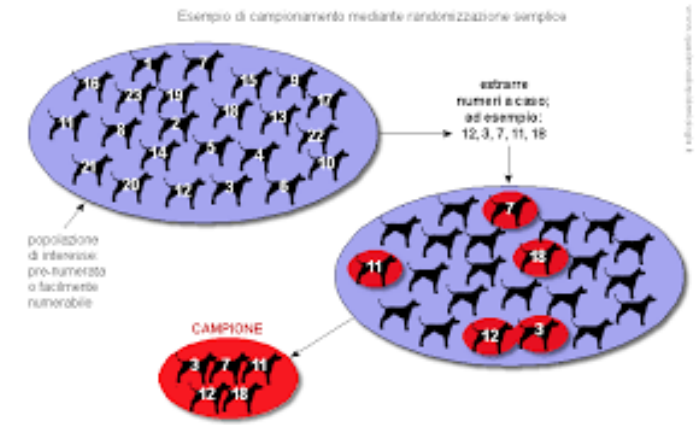
<https://towardsdatascience.com/probability-learning-ii-how-bayes-theorem-is-applied-in-machine-learning-bd747a960962>

<https://towardsdatascience.com/probability-learning-iii-maximum-likelihood-e78d5ebea80c>



# 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**.



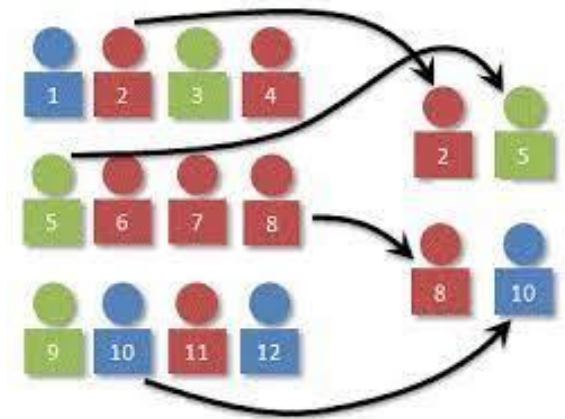
# La Teoria della Stima

---

- ✓ 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  $\sigma$  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

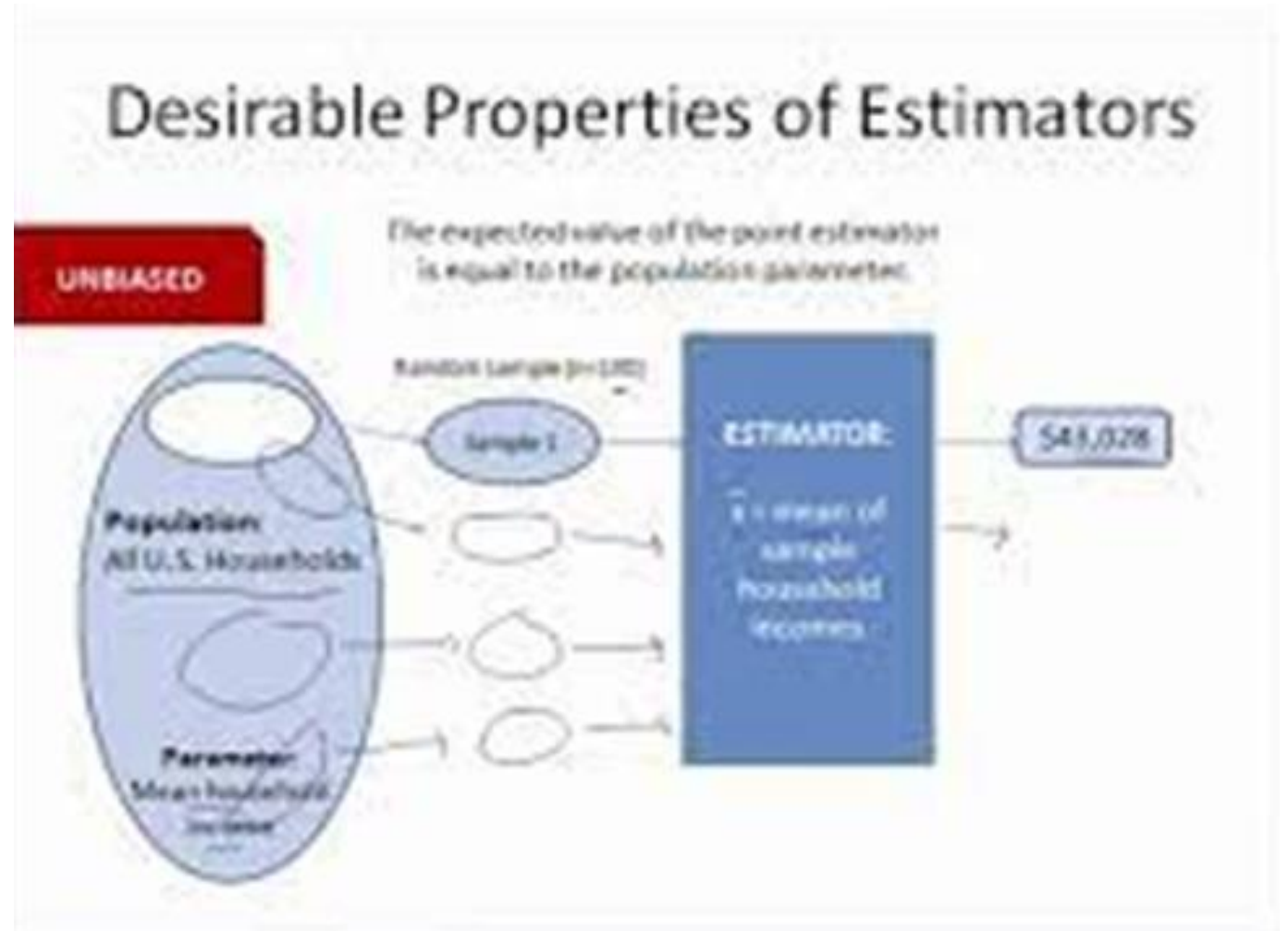
---

- ✓ 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:

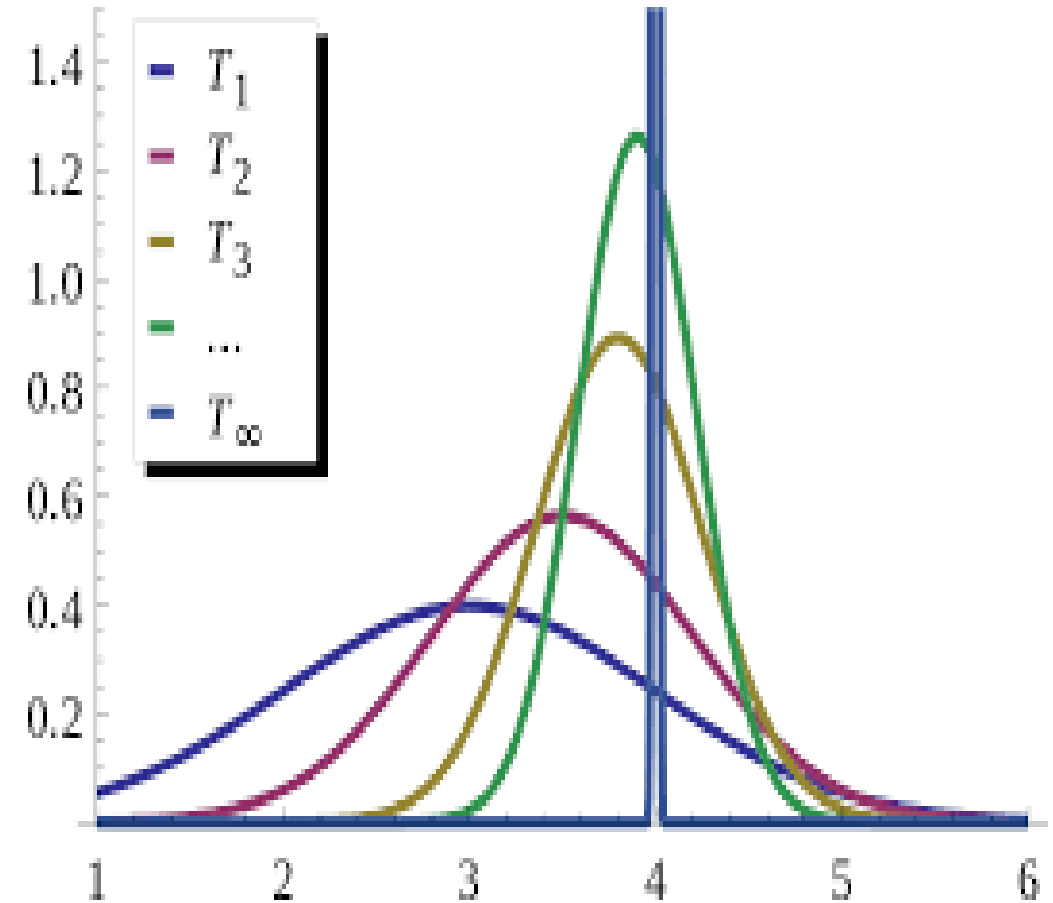
- **Correttezza**
- **Consistenza**
- **Efficienza**
- **Sufficienza**
- **Normalità asintotica**





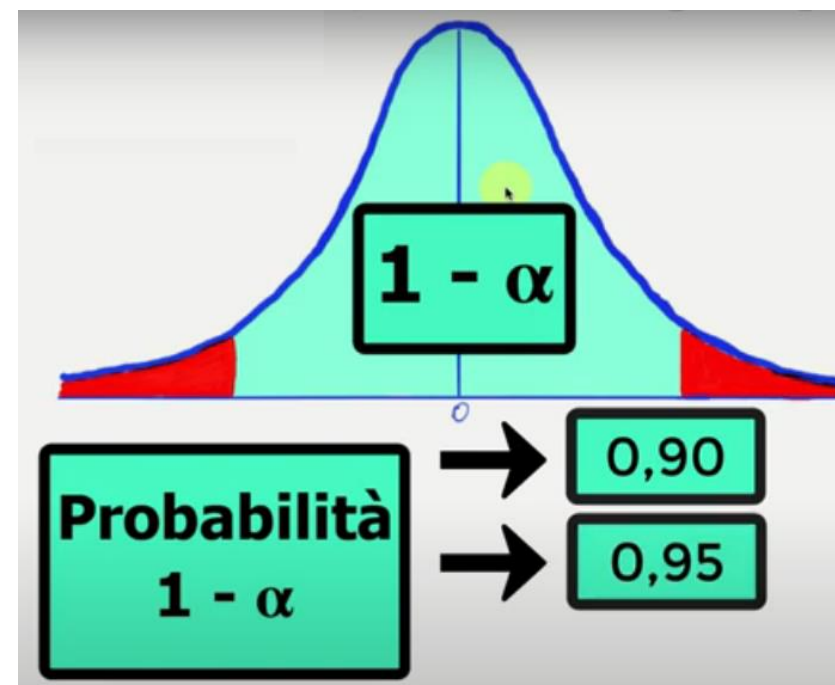
# Proprietà desiderabili degli Stimatori: Correttezza

- ✓ Uno stimatore  $T(X)$  si dice **corretto** o **non distorto** quando il suo **valore medio**  $E[T(X)]$  coincide con il valore del parametro  $\theta$  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  $\mu$  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  $\sigma^2$ . Lo **stimatore corretto** della varianza  $\sigma^2$  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  $E[S^2] = \frac{n-1}{n} \hat{S}^2$

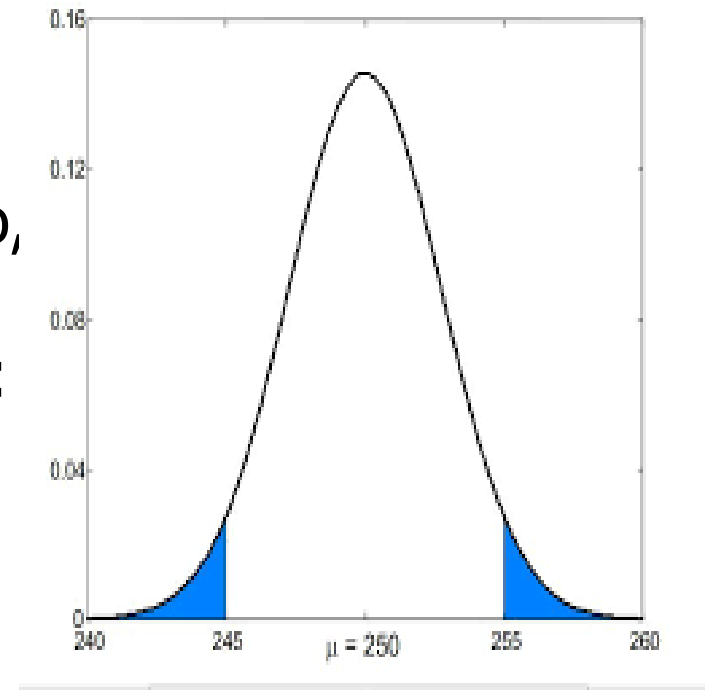


# Errore di prima specie ed errore di seconda specie

- ✓ 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.

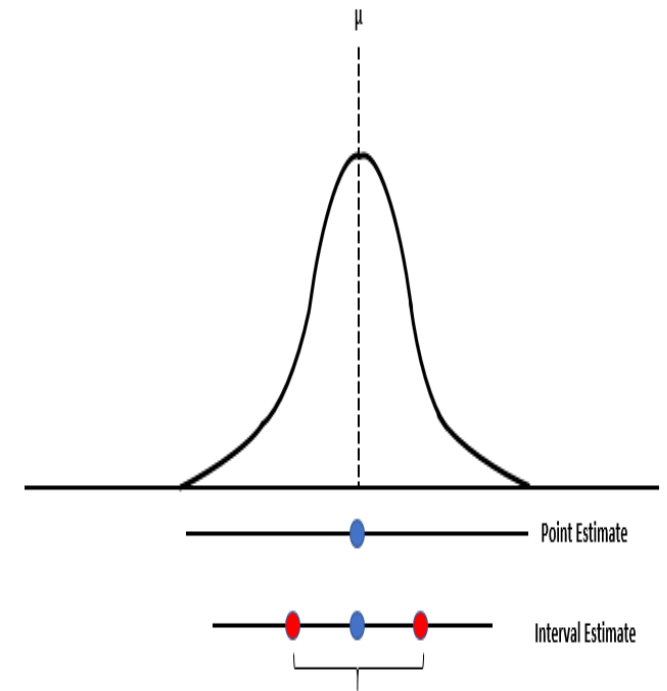
# Errore di prima specie ed errore di seconda specie

---

- ✓ 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  $\theta$  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  $\epsilon$  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  $\theta$ ; 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

---