

# Statistica e Analisi dei dati

Università degli studi di Milano - Informatica

Luca Favini, Matteo Zaghenò

Ultima modifica: 14/06/2024 - [Codice sorgente](#)

# Statistica e Analisi dei dati

Insegnamento del corso di laurea triennale in Informatica, Università degli studi di Milano. Tenuto dal Professore Dario Malchiodi, anno accademico 2023-2024.

La statistica si occupa di raccogliere, analizzare e trarre conclusioni su dati, attraverso vari strumenti:

- Statistica descrittiva: esposizione e **condensazione** dei dati, cercando di limitarne l'incertezza;
- Calcolo delle probabilità: creazione e analisi di modelli in situazioni di **incertezza**;
- Statistica inferenziale: **approssimazione** degli esiti mancanti, attraverso modelli probabilistici;
- Appendice: Cheatsheet Python: raccolta funzioni/classi Python utili ai fini dell'esame (*e non*).
- Appendice: Cheatsheet integrali: come svolgere gli integrali.

## Indice

1. Statistica descrittiva .....	4
1.1. Classificazione dei dati: qualitativi e quantitativi .....	4
1.2. Frequenze .....	4
1.2.1. Frequenze assolute e relative .....	4
1.2.2. Frequenze cumulate .....	4
1.2.2.1. Funzione cumulativa empirica .....	4
1.2.3. Frequenze congiunte e marginali .....	4
1.2.4. Stratificazione .....	4
1.3. Grafici .....	4
1.4. Indici di centralità .....	4
1.4.1. Media campionaria .....	4
1.4.2. Mediana campionaria .....	4
1.4.3. Moda campionaria .....	5
1.5. Indici di dispersione .....	5
1.5.1. Scarto assoluto medio .....	5
1.5.2. Varianza campionaria .....	5
1.5.2.1. Varianza campionaria standard .....	5
1.5.3. Coefficiente di variazione .....	6
1.5.4. Quantile .....	6
1.6. Indici di correlazione .....	6
1.6.1. Covarianza campionaria .....	6
1.6.2. Indice di correlazione di Pearson (indice di correlazione lineare) .....	7
1.7. Indici di eterogeneità .....	8
1.7.1. Indice di Gini (per l'eterogeneità) .....	8
1.7.2. Entropia .....	8
1.8. Indici di concentrazione .....	9
1.8.1. Curva di Lorentz .....	9
1.8.2. Indice di Gini (per la concentrazione) .....	10
1.8.3. Analisi della varianza (ANOVA) .....	10
1.9. Alberi di decisione .....	12
1.10. Classificatori .....	12
1.10.1. Casi particolari .....	13
1.10.2. Classificatori a soglia (Curva ROC) .....	13
1.11. Trasformazione dei dati .....	14
1.12. Grafici .....	14

2. Calcolo delle probabilità .....	14
2.1. Calcolo combinatorio .....	14
2.1.1. Disposizioni .....	14
2.1.2. Combinazioni .....	14
2.1.3. Permutazioni .....	15
2.2. Elementi di probabilità .....	15
2.2.1. Algebra di eventi .....	16
2.2.2. Assiomi di Kolmogorov .....	17
2.2.3. Teoremi derivati dagli assiomi .....	17
2.2.4. Spazi di probabilità ed Esiti equiprobabili .....	18
2.3. Probabilità condizionata .....	18
2.3.1. Regola di fattorizzazione .....	18
2.3.2. Teorema delle probabilità totali .....	19
2.3.3. Teorema di Bayes .....	19
2.3.4. Classificatore naive-Bayes .....	20
2.3.5. Eventi indipendenti .....	21
2.3.6. Indipendenza a tre o più eventi .....	21
2.4. Variabili aleatorie .....	22
2.4.1. Variabili aleatorie discrete .....	22
2.4.1.1. Funzione indicatrice .....	22
2.4.1.2. Funzione di massa di probabilità .....	23
2.4.1.3. Funzione di ripartizione .....	23
2.4.1.4. Valore atteso .....	24
2.4.1.5. Varianza .....	24
2.4.2. Variabili aleatorie multivariate .....	25
2.4.2.1. Funzione di ripartizione congiunta .....	25
2.4.2.2. Funzione di massa di probabilità congiunta .....	26
2.4.2.3. Indipendenza .....	26
2.4.2.4. Valore atteso .....	27
2.4.2.5. Covarianza .....	27
2.4.2.6. Varianza .....	27
2.4.3. Variabili aleatorie continue .....	28
2.4.3.1. Funzione densità di probabilità .....	28
2.4.3.2. Valore atteso .....	28
2.4.3.3. Varianza .....	29
2.4.3.4. Disuguaglianza di Markov .....	29
2.4.3.5. Disuguaglianza di Chebyshev .....	30
2.4.4. Modelli di distribuzione .....	31
2.4.4.1. Modello di Bernoulli $X \sim B(p)$ .....	31
2.4.4.2. Modello binomiale $X \sim B(n, p)$ .....	32
3. Statistica inferenziale .....	33
4. Cheatsheet Python .....	33
5. Cheatsheet integrali .....	33

# 1. Statistica descrittiva

**Popolazione** insieme di elementi da analizzare, spesso troppo numerosa per essere analizzata tutta

**Campione** parte della popolazione estratta per essere analizzata, deve essere rappresentativo

**Campione casuale (semplice)** tutti i membri della popolazione hanno la stessa possibilità di essere selezionati

## 1.1. Classificazione dei dati: qualitativi e quantitativi

**Dati quantitativi / Scalari / Numerici** l'esito della misurazione è una quantità numerica

**Discreti** si lavora su valori singoli (spesso interi), ad esempio: *numeri di figli*

**Continui** si lavora su range di intervalli, ad esempio: *peso* o *altezza*

**Dati qualitativi / Categorici / Nominali** l'esito della misurazione è un'etichetta

**Booleani / Binari** due valori possibili, ad esempio: *sex*

**Nominali / Sconnessi** valori **non** ordinabili, ad esempio: *nome*

**Ordinali** valori ordinabili, ad esempio: *livello di soddisfazione*

### *i* Nota

Spesso alcuni dati *numerici* vengono considerati *qualitativi*, dato che non ha senso effettuare su di essi considerazioni algebriche o numeriche. Un esempio potrebbe essere la data di nascita.

## 1.2. Frequenze

### 1.2.1. Frequenze assolute e relative

### 1.2.2. Frequenze cumulate

#### 1.2.2.1. Funzione cumulativa empirica

### 1.2.3. Frequenze congiunte e marginali

### 1.2.4. Stratificazione

## 1.3. Grafici

## 1.4. Indici di centralità

Sono indici che danno un'idea approssimata dell'ordine di grandezza (quindi dove ricadono) dei valori esistenti.

### 1.4.1. Media campionaria

Viene indicata da  $\bar{x}$ , ed è la **media aritmetica** di tutte le osservazioni del campione.

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

La media opera linearmente, quindi può essere scalata ( $\cdot a$ ) e/o traslata ( $+b$ ):

$$\forall i \ y_i = ax_i + b \Rightarrow \bar{y} = a\bar{x} + b$$

Non è un stimatore robusto rispetto agli outlier. Può essere calcolata solo con dati quantitativi.

### 1.4.2. Mediana campionaria

È il valore a **metà** di un dataset ordinato in ordine crescente, ovvero un valore  $\geq$  e  $\leq$  di almeno la metà dei dati.

Dato un dataset di dimensione  $n$  la mediana è:

- l'elemento in posizione  $\frac{n+1}{2}$  se  $n$  è dispari
- la media aritmetica tra gli elementi in posizione  $\frac{n}{2}$  e  $\frac{n}{2} + 1$  se  $n$  è pari

È robusta rispetto agli outlier ma può essere calcolata solo su *campioni ordinabili*.

### 1.4.3. Moda campionaria

È l'osservazione che compare con la maggior frequenza. Se più di un valore compare con la stessa frequenza allora tutti quei valori sono detti modali.

## 1.5. Indici di dispersione

Sono indici che misurano quanto i valori del campione si discostano da un valore centrale.

### 1.5.1. Scarto assoluto medio

Per ogni osservazione, lo scarto è la distanza dalla media:  $x_i - \bar{x}$ . La somma di tutti gli scarti farà sempre 0.

$$\sum_{i=1}^n x_i - \bar{x} = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0$$

### 1.5.2. Varianza campionaria

Misura di quanto i valori si discostano dalla media campionaria

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Metodo alternativo per calcolare la varianza:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - n\bar{x}^2)$$

#### Nota

Verrebbe intuitivo applicare il *valore assoluto* ad ogni scarto medio, ma questo causa dei problemi. Per questo motivo la differenza viene elevata al *quadrato*, in modo da renderla sempre positiva.

La varianza *non* è un operatore lineare: la traslazione non ha effetto mentre la scalatura si comporta:

$$s_y^2 = a^2 s_x^2$$

#### 1.5.2.1. Varianza campionaria standard

È possibile applicare alla varianza campionaria la radice quadrata, ottenendo la varianza campionaria standard.

$$s = \sqrt{s^2}$$

#### Attenzione

Applicando la radice quadrata solo dopo l'elevamento a potenza, non abbiamo reintrodotta il problema dei valori negativi:  $\sqrt{a^2} \neq (\sqrt{a})^2 = a$

### 1.5.3. Coefficiente di variazione

Valore **adimensionale**, utile per confrontare misure di fenomeni con unità di misura differenti.

$$s^* = \frac{s}{|\bar{x}|}$$

#### **i** Nota

Sia la varianza campionaria standard che la media campionaria sono dimensionali, ovvero hanno unità di misura. Dividendoli tra loro otteniamo un valore adimensionale.

### 1.5.4. Quantile

Il quantile di ordine  $\alpha$  (con  $\alpha$  un numero reale nell'intervallo  $[0, 1]$ ) è un valore  $q_\alpha$  che divide la popolazione in due parti, proporzionali in numero di elementi ad  $\alpha$  e  $(1-\alpha)$  e caratterizzate da valori rispettivamente minori e maggiori di  $q_\alpha$ .

**Percentile** quantile descritto in percentuale

**Decile** popolazione divisa in 10 parti con ugual numero di elementi

**Quartile** popolazione divisa in 4 parti con ugual numero di elementi

#### **i** Nota

È possibile visualizzare un campione attraverso un **box plot**, partendo dal basso composto da:

- eventuali *outliers*, rappresentati con le x prima del baffo
- il *baffo* “inferiore”, che parte dal valore minimo e raggiunge il primo quartile
- il *box* (scatola), che rappresenta le osservazioni comprese tra il primo e il terzo quartile
- la linea che divide in due il box, che rappresenta la *mediana*
- il *baffo* “superiore”, che parte terzo quartile e raggiunge il massimo
- eventuali *outliers* “superiori”, rappresentati con le x dopo il baffo

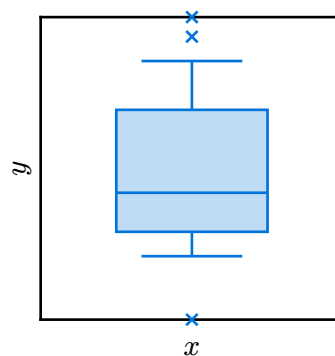


Figure 1: Grafico boxplot

## 1.6. Indici di correlazione

**Campione bivariato** campione formato da coppie  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ .

**Correlazione** relazione tra due variabili tale che a ciascun valore della prima corrisponda un valore della seconda seguendo una certa regolarità.

### 1.6.1. Covarianza campionaria

È un valore numerico che fornisce una misura di quanto le due variabili varino assieme. Dato un campione bivariato definiamo la **covarianza campionaria** come:

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Metodo alternativo di calcolo:

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i y_i - n\bar{x}\bar{y})$$

### 💡 Informalmente

Intuitivamente c'è una **correlazione diretta** se al crescere di  $x$  cresce anche  $y$  o al decrescere di  $x$  decresce anche  $y$ , dato che il contributo del loro prodotto alla sommatoria sarà positivo. Quindi se  $x$  e  $y$  hanno segno concorde allora la correlazione sarà *diretta*, altrimenti *indiretta*.

- $\text{Cov}(x, y) > 0$  probabile correlazione diretta
- $\text{Cov}(x, y) \simeq 0$  correlazione improbabile
- $\text{Cov}(x, y) < 0$  probabile correlazione indiretta

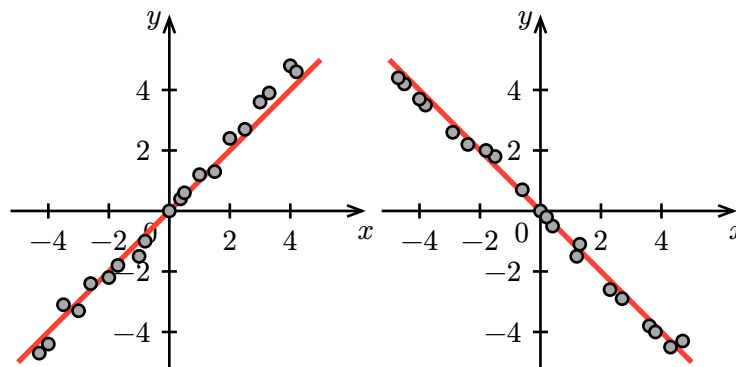


Figure 2: Correlazione lineare *diretta* (sinistra) e *indiretta* (destra)

### i Nota

Una relazione diretta/indiretta non è necessariamente *lineare*, può essere anche *logaritmica* o seguire altre forme.

#### 1.6.2. Indice di correlazione di Pearson (indice di correlazione lineare)

Utilizziamo l'indice di correlazione di Pearson per avere un valore *adimensionale* che esprime una correlazione. Possiamo definirlo anche come una misura normalizzata della covarianza nell'intervallo  $[-1, +1]$ .  $\rho$  è **insensibile** alle trasformazioni lineari.

$$\rho(x, y) = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{s_{XY}}{s_X s_Y}$$

Dove  $s$  è la varianza campionaria standard.

- $\rho \simeq +1$  probabile correlazione linearmente diretta
- $\rho \simeq 0$  correlazione improbabile
- $\rho \simeq -1$  probabile correlazione linearmente indiretta

**! Attenzione**

L'indice di correlazione lineare ( $\rho$ ) cattura **solo** relazioni dirette/indirette *lineari* ed è insensibile alle trasformazioni lineari.

**! Attenzione**

La covarianza campionaria o l'indice di correlazione lineare  $\simeq 0$  non implicano l'indipendenza del campione, ma è vero il contrario:

$$\text{Cov}(x, y) \simeq 0 \not\Rightarrow \text{Indipendenza}$$

$$\rho(x, y) \simeq 0 \not\Rightarrow \text{Indipendenza}$$

$$\text{Indipendenza} \Rightarrow \rho(x, y) \simeq \text{Cov}(x, y) \simeq 0$$

**1.7. Indici di eterogeneità**

**Massima eterogeneità** il campione è composto da tutti elementi diversi

**Minima eterogeneità** il campione non contiene due elementi uguali (*campione omogeneo*)

L'eterogeneità può essere calcolata anche su un insieme di dati qualitativi.

**1.7.1. Indice di Gini (per l'eterogeneità)**

$$I = 1 - \sum_{j=1}^n f_j^2$$

Dove  $f_j$  è la frequenza relativa di  $j$  ed  $n$  è il numero di elementi distinti. Quindi  $\forall j, 0 \leq f_j \leq 1$ . Prendiamo in considerazione i due estremi:

- eterogeneità *minima* (solo un valore con frequenza relativa 1):

$$I = 1 - 1 = 0$$

- eterogeneità *massima* (tutti i valori hanno la stessa frequenza relativa  $\frac{1}{n}$  dove  $n$  è la dimensione del campione):

$$I = 1 - \sum_{j=1}^n \left(\frac{1}{n}\right)^2 = 1 - \frac{n}{n^2} = \frac{n-1}{n}$$

Generalizzando,  $I$  non raggiungerà mai 1:

$$0 \leq I \leq \frac{n-1}{n} < 1$$

Dal momento che l'indice di Gini tende a 1 senza mai arrivarci introduciamo l'**indice di Gini normalizzato**, in modo da arrivare a 1 nel caso di eterogeneità massima:

$$I' = \frac{n}{n-1} I$$

**1.7.2. Entropia**

$$H = \sum_{j=1}^n f_j \log\left(\frac{1}{f_j}\right) = \sum_{j=1}^n -f_j \log(f_j)$$



Dove  $f_j$  è la frequenza relativa e  $n$  è il numero di elementi distinti. L'entropia assume valori nel range  $[0, \log(n)]$  quindi utilizziamo l'**entropia normalizzata** per confrontare due misurazioni con diverso numero di elementi distinti  $n$ .

$$H' = \frac{1}{\log(n)} H$$

### Nota

In base alla base del logaritmo utilizzata, l'entropia avrà unità di misura differente:

- $\log_2$ : bit
- $\log_e$ : nat
- $\log_{10}$ : hartley

### Informalmente

Intuitivamente sia l'indice di Gini che l'entropia sono una “*media pesata*” tra la frequenza relativa di ogni elemento ed un peso: la *frequenza stessa* nel caso di Gini e il *logaritmo del reciproco* nell'entropia. La frequenza relativa è già nel range  $[0, 1]$ , quindi non c'è bisogno di dividere per il numero di elementi.

## 1.8. Indici di concentrazione

Un indice di concentrazione è un indice statistico che misura in che modo un *bene* è distribuito nella *popolazione*.

**Distribuzione del bene**  $a_1, a_2, \dots, a_n$  indica la quantità ordinata in modo **non decrescente**, del bene posseduta dall'individuo  $i$

**Media**  $\bar{a}$  indica la quantità media posseduta da un individuo

**Totale**  $TOT = n\bar{a}$  indica il totale del bene posseduto

- Concentrazione **massima (sperequato)**: un individuo possiede tutta la quantità  $a_{1..n-1} = 0, a_n = n\bar{a}$
- Concentrazione **minima (equo)**: tutti gli individui possiedono la stessa quantità  $a_{1..n} = \bar{a}$

### 1.8.1. Curva di Lorentz

La curva di Lorenz è una rappresentazione **grafica** della *distribuzione* di un bene nella popolazione.

Dati:

- $F_i = \frac{i}{n}$ : posizione percentuale dell'osservazione  $i$  nell'insieme
- $Q_i = \frac{1}{TOT} \sum_{k=1}^i a_k$

La tupla  $(F_i, Q_i)$  indica che il  $100 \cdot F_i\%$  degli individui detiene il  $100 \cdot Q_i\%$  della quantità totale.

Inoltre:  $\forall i, 0 \leq Q_i \leq F_i \leq 1$ .

### Informalmente

Possiamo vedere  $F_i$  come “*quanta*” popolazione è stata analizzata fino all'osservazione  $i$ , espressa nel range  $[0, 1]$ .  $Q_i$  è invece una “frequenza cumulata” della ricchezza, fino all'osservazione  $i$ .

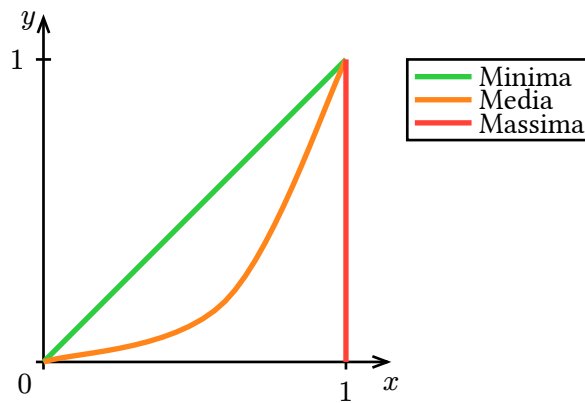


Figure 3: Curva di Lorenz

### 1.8.2. Indice di Gini (per la concentrazione)

Dato che la curva di Lorenz non assume mai alcun valore nella parte di piano superiore alla retta che collega  $(0, 0)$  a  $(1, 1)$ , allora introduciamo l'**indice di Gini**, che invece assume valori nel range  $[0, 1]$ .

Anche esso indica la *concentrazione* di un bene nella popolazione.

$$G = \frac{\sum_{i=1}^{n-1} F_i - Q_i}{\sum_{i=1}^{n-1} F_i}$$

È possibile riscrivere il denominatore come:

$$\sum_{i=1}^{n-1} F_i = \frac{1}{n} \sum_{i=1}^{n-1} i = \frac{1}{n} \frac{n(n-1)}{2} = \frac{n-1}{2}$$

Ottendendo come formula alternativa:

$$G = \frac{2}{n-1} \sum_{i=1}^{n-1} F_i - Q_i$$



#### Informalmente

Facendo un parallelo con la curva di Lorenz, possiamo vedere  $F_i - Q_i$  come la distanza tra la bisettrice ( $F_i$ ) e la ricchezza dell'osservazione  $i$  ( $Q_i$ ). La somma di queste distanze viene poi "normalizzata", dividendo per  $\frac{n-1}{2}$ .

### 1.8.3. Analisi della varianza (ANOVA)

Dato un campione, è possibile suddividerlo in più *gruppi* ed effettuare delle analisi sulle *diversità* tra i vari gruppi. Ad esempio, dato un campione di dati sulla natalità, si potrebbe analizzare formando gruppi per regione o per reddito.

L'analisi della varianza (**ANOVA** - ANalysis Of VAriance) è un insieme di tecniche statistiche che permettono, appunto, di confrontare due o più *gruppi* di dati. Definiamo a questo scopo:

**Numerosità dei gruppi** dato un campione diviso in  $G$  gruppi, ognuno ha numerosità  $n_1, \dots, n_G$

**Osservazione** viene definita  $x_i^g$  come l' $i$ -esima osservazione del  $g$ -esimo gruppo

**Media campionaria di tutte le osservazioni** la media del campione

$$\bar{x} = \frac{1}{n} \sum_{g=1}^G \sum_{i=1}^{n_g} x_i^g$$

**Media campionaria di un gruppo** la media dei valori del gruppo

$$\bar{x}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} x_i^g$$

### Somme degli scarti

- Somma **totale** degli scarti (tra ogni elemento e la media di tutto il campione):

$$SS_T = \sum_{g=1}^G \sum_{i=1}^{n_g} (x_i^g - \bar{x})^2$$

- Somma degli scarti **entro/within** i gruppi (tra ogni elemento e la media del proprio gruppo):

$$SS_W = \sum_{g=1}^G \sum_{i=1}^{n_g} (x_i^g - \bar{x}_g)^2$$

- Somma degli scarti **tra/between** i gruppi (tra la media di ogni gruppo e la media del campione, “pesato” per la numerosità del gruppo):

$$SS_B = \sum_{g=1}^G n_g (\bar{x}_g - \bar{x})^2$$

Vale la seguente regola:  $SS_T = SS_W + SS_B$ .

### Indici di variazione

- **Total** (la varianza totale del campione):

$$\frac{SS_T}{n-1}$$

- **Within** (la varianza di ogni elemento del gruppo):

$$\frac{SS_W}{n-G}$$

- **Between** (la varianza tra ogni gruppo e il campione completo):

$$\frac{SS_B}{G-1}$$

L'ipotesi alla base è che dati  $G$  gruppi, sia possibile scomporre la varianza in due componenti: *Varianza interna ai gruppi* (varianza **Within**) e *Varianza tra i gruppi* (varianza **Between**).



#### Informalmente

Analizzando diversi gruppi attraverso l'ANOVA, si possono raggiungere due conclusioni:

- i gruppi risultano significativamente **diversi** tra loro: la *varianza between* contribuisce più significativamente alla varianza totale (il fenomeno è legato a caratteristiche proprie di ciascun gruppo)
- i gruppi risultano **omogenei**: la *varianza within* contribuisce più significativamente alla varianza totale (il fenomeno è legato a caratteristiche proprie di tutti i gruppi)

```
import numpy as np

def anova(groups):
    all_elements = pd.concat(groups)
    sum_total = sum((all_elements - all_elements.mean())**2)
    sum_within = sum([sum((g - g.mean())**2) for g in groups])
    sum_between = sum([len(g) * (g.mean() - all_elements.mean())**2 for g in groups])
    assert(np.abs(sum_total - sum_within - sum_between) < 10**-5)
    n = len(all_elements)
    total_var = sum_total / (n-1)
    within_var = sum_within / (n-len(groups))
    return (total_var, within_var*(n-len(groups))/(n-1))
```

Python

## 1.9. Alberi di decisione

### 1.10. Classificatori

Dato un *classificatore binario* che divide in due classi (positiva e negativa) e un *insieme di oggetti* di cui è **nota** la classificazione, possiamo valutare la sua bontà tramite il numero di casi classificati in modo errato. La classificazione errata può essere:

- **Falso negativo:** oggetto *positivo* classificato come *negativo*
- **Falso positivo:** oggetto *negativo* classificato come *positivo*

#### *i* Nota

Il peso di un falso positivo può **non** essere lo stesso di un falso negativo, si pensi al caso di una malattia contagiosa: un *falso negativo* sarà molto più pericoloso di un *falso positivo* (che verrà scoperto con ulteriori analisi).

Introduciamo la **matrice di confusione**, che riassume la bontà del classificatore:

		Valore effettivo		
		Positivo	Negativi	
Predizione del classificatore	Positivo	Veri positivi (VP)	Falsi positivi (FP)	<i>Totali classificati positivi (TOT CP)</i>
	Negativi	Falsi negativi (FN)	Veri negativi (VN)	<i>Totali classificati negativi (TOT CN)</i>
		<i>Totale positivi (TP)</i>	<i>Totale negativi (TN)</i>	<i>Totale casi (TOT casi)</i>

Table 1: Matrice di confusione

```
pd.DataFrame(metrics.confusion_matrix(Y_test, preds))
```

Python

**Sensibilità** capacità del classificatore di predire bene i positivi  $\frac{VP}{TP}$   
**Specificità** capacità del classificatore di predire bene i negativi  $\frac{VN}{TN}$

È possibile valutare la bontà di un classificatore attraverso il punto:

$$(1 - \text{Specificità}, \text{Sensibilità}) = \left(1 - \frac{VN}{TN}, \frac{VP}{TP}\right) = \left(\frac{FP}{TN}, \frac{VP}{TP}\right)$$

### 1.10.1. Casi particolari

**Classificatore costante** associa indiscriminatamente gli oggetti ad una classe (positiva o negativa)

**Classificatori positivi (CP)** tutti i casi sono classificati come positivi

- Sensibilità: 1, Specificità: 0, Punto (1, 1) ●

**Classificatori negativi (CN)** tutti i casi sono classificati come negativi

- Sensibilità: 0, Specificità: 1, Punto (0, 0) ●

**Classificatore ideale (CI)** tutti i casi sono classificati correttamente

- Sensibilità: 1, Specificità: 1, Punto (0, 1) ●

**Classificatore peggiore (CE)** tutti i casi sono classificati erroneamente

- Sensibilità: 0, Specificità: 0, Punto (1, 0) ●

**Classificatore casuale** ogni caso viene assegnato in modo casuale

- Sensibilità: 0.5, Specificità: 0.5, Punto  $(\frac{1}{2}, \frac{1}{2})$  ●

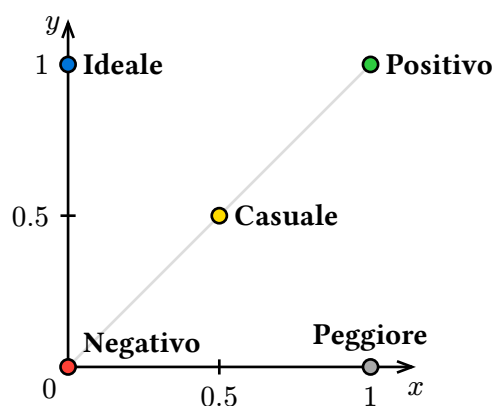


Figure 4: Rappresentazione classificatori

### 1.10.2. Classificatori a soglia (Curva ROC)

Un classificatore a soglia discrimina un caso in base ad una **soglia** stabilita a priori, in caso la misurazione sia *superiore* alla soglia allora verrà classificato *positivamente*, altrimenti *negativamente*.

Per trovare il valore con cui *fissare* la soglia, possiamo sfruttare questo metodo:

- definiamo  $\theta$  come una generica soglia
- è necessario stabilire un intervallo  $[\theta_{\min}, \theta_{\max}]$ 
  - utilizzando  $\theta_{\min}$  tutti i casi saranno positivi, ottenendo un classificatore positivo ●
  - utilizzando  $\theta_{\max}$  tutti i casi saranno negativi, ottenendo un classificatore negativo ●
- definiamo  $D$  come una discretizzazione di questo intervallo continuo

Per ogni soglia  $\theta \in D$  è possibile calcolare la *sensibilità* e *specificità*. Questo classificatore viene quindi *rappresentato* sul piano cartesiano attraverso il punto  $(1 - \text{Specificità}, \text{Sensibilità})$ .

Il risultato è una **curva**, detta **ROC** (Receiver Operator Carapteristic) —, che ha sempre come estremi in  $(0, 0)$  (caso in cui viene usato  $\theta_{\max}$ ) e  $(1, 1)$  (caso in cui viene usato  $\theta_{\min}$ ).

Per misurare la *bontà* del classificatore viene misurata l'area di piano sotto la curva (**AUC** - Area Under the ROC Curve ■■■), più si avvicina a 1, *migliore* è il classificatore.

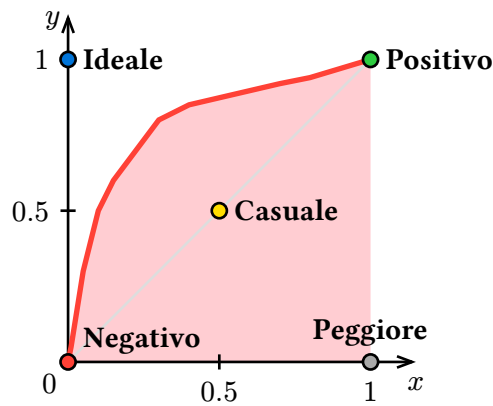


Figure 5: Curva ROC

## 1.11. Trasformazione dei dati

## 1.12. Grafici

# 2. Calcolo delle probabilità

## 2.1. Calcolo combinatorio

Analizzare *come* e in *quanti* modi si possono effettuare raggruppamenti di elementi.

**Principio di enumerazione (principio fondamentale del calcolo combinatorio)** se dobbiamo compiere  $t$  esperimenti e per ognuno di essi ci possono essere  $s_i$  possibili risultati, il numero di risultati totali è  $s_1 \cdot s_2 \cdot \dots \cdot s_t$



### Informalmente

Vogliamo selezionare  $k$  elementi da un insieme  $A$  di  $n$  elementi:

**Disposizioni** l'ordine è importante  $(a, b) \neq (b, a)$

**Combinazioni** l'ordine *non* è importante  $(a, b) = (b, a)$

**Permutazioni** tutti gli elementi vengono disposti  $k = n$

È possibile sia *avere* che *non avere* delle **ripetizioni** in tutti i casi.

### 2.1.1. Disposizioni

Dato un insieme di  $n$  oggetti distinti  $A = \{a_1, \dots, a_n\}$ , vogliamo selezionare  $k$  oggetti (con  $k \leq n$ ), tenendo in considerazione l'**ordine**.

**Disposizione senza ripetizioni (semplici)** gli oggetti di  $A$  possono essere usati una volta sola

$$d_{n,k} = \frac{n!}{(n-k)!}$$

**Disposizione con ripetizione** gli oggetti di  $A$  possono essere usati più di una volta

$$D_{n,k} = n^k$$

### 2.1.2. Combinazioni

Dato un insieme di  $n$  oggetti distinti  $A = \{a_1, \dots, a_n\}$ , vogliamo selezionare  $k$  oggetti (con  $k \leq n$ ), **senza** considerare l'ordine.

**i Nota**

Il numero di combinazioni  $c_{n,k}$  è sempre minore del numero di disposizioni  $d_{n,k}$ , dato che l'ordine non conta.

**Combinazione senza ripetizioni (semplici)** gli oggetti di  $A$  possono essere usati una volta sola

$$c_{n,k} = \frac{d_{n,k}}{k!} = \frac{n!}{k! \cdot (n-k)!} = \binom{n}{k}$$

**i Nota**

$\binom{n}{k}$  viene detto **coefficiente binomiale**

**Combinazione con ripetizioni** gli oggetti di  $A$  possono venir usati più di una volta

$$C_{n,k} = \frac{(n+k-1)!}{k! \cdot (n-1)!} = \binom{n+k-1}{k}$$

**2.1.3. Permutazioni**

Dato un insieme di  $n$  oggetti  $A = \{a_1, \dots, a_n\}$ , una **permutazione** è una sequenza *ordinata* in cui compaiono *tutti* gli oggetti (quindi vogliamo selezionare  $k$  elementi).

**Permutazioni semplici (senza ripetizioni)** l'insieme  $A$  non contiene elementi duplicati

$$P_n = n!$$

**Permutazioni di oggetti distinguibili a gruppi (con ripetizioni)** l'insieme  $A$  contiene  $k$  gruppi di oggetti indistinguibili, ognuno con numerosità  $n_1, \dots, n_k$  (con  $\sum_{i=1}^k n_i = n$ ), allora dobbiamo disporre tutti questi elementi

$$P_{n:n_1, \dots, n_k} = \frac{n!}{n_1! \cdot \dots \cdot n_k!} = \binom{n}{(n_1, \dots, n_k)}$$

**i Nota**

$\binom{n}{(n_1, \dots, n_k)}$  viene detto **coefficiente multinomiale**

**2.2. Elementi di probabilità**

**Esito**  $\omega \in \Omega$  risultato effettivo di un esperimento

**Evento**  $E \subseteq \Omega$  è un qualsiasi insieme formato da tutti, alcuni o nessuno dei possibili esiti di un esperimento

**Probabilità** quantificazione dell'incertezza di un evento

**Spazio campionario**  $\Omega$  (**insieme degli esiti o insieme universo**) è l'insieme di tutti gli esiti possibili. Può essere *finito* o *infinito*, *continuo* o *discreto*

**💡 Informalmente**

*Esempio:* lanciando un dado, l'*esito* è il numero risultante, un *evento* può essere “esce 3 o 6” e la *probabilità* di questo evento è  $\frac{2}{6}$ .

**Evento certo**  $E = \Omega$  si verifica sempre

**Evento impossibile**  $E = \emptyset$  non si verifica mai

### **i** Nota

Indichiamo sempre con una *minuscola* un *esito*, mentre con una *maiuscola* un *evento*.

Dati degli eventi, è possibile applicare le operazioni e proprietà degli insiemi su di essi:

**Unione**  $E \cup F$  quando si verifica l'evento  $E$  o l'evento  $F$

**Intersezione**  $E \cap F$  quando si verificano entrambi gli eventi  $E$  ed  $F$

**Mutualmente esclusivi**  $E \cap F = \emptyset$  i due eventi sono *mutualmente esclusivi*

**Differenza**  $E - F$  si verifica l'evento  $E$ , ma l'evento  $F$  non si verifica (l'operazione di sottrazione non è commutativa,  $E - F \neq F - E$ )

**Complemento**  $\Omega - E = E^c = \overline{E}$  quando l'evento  $E$  non si verifica

**Sottoinsieme**  $E \subseteq F = E \rightarrow F$  quando si verifica  $E$ , allora si verifica anche  $F$

**Proprietà per unione e intersezione**

**Commutatività**  $E \cup F = F \cup E$

**Associatività**  $(D \cup E) \cup F = D \cup (F \cup E)$

**Distributività**  $D \cup (E \cap F) = (D \cup E) \cap (D \cup F)$

**De Morgan**  $\overline{E \cup F} = \overline{E} \cap \overline{F}$ : l'evento che si verifica quando non si verifica  $E$  o  $F$  è lo stesso evento che si verifica quando non si verifica  $E$  e non si verifica  $F$

È possibile dare diverse *interpretazioni* alla probabilità:

**Approccio soggettivista** la probabilità di un esito non è oggettiva: è il livello di *fiducia* che un soggetto (*lo studioso*) ripone nel verificarsi di un evento

**Approccio frequentista** la probabilità di un esito è una *proprietà* dell'esito stesso: viene calcolata come il rapporto tra il numero di casi *favorevoli* e il numero di casi *possibili* ripetendo l'esperimento un numero di volte tendente all'infinito

#### 2.2.1. Algebra di eventi

Un algebra di eventi  $A$  è un insieme di eventi  $\{E_1, E_2, \dots\}$  a cui sono associate delle operazioni che soddisfa le proprietà:

- $\forall E \in A, E \subseteq \Omega$ : ogni evento appartenente all'*algebra*  $A$  appartiene all'insieme di tutti gli *eventi possibili*  $\Omega$
- $\Omega \in A$ : l'insieme di tutti gli *eventi possibili*  $\Omega$  appartiene all'*algebra*  $A$
- $\forall E \in A, \overline{E} \in A$ : chiusura rispetto al *complemento*
- $\forall E, F \in A, E \cup F \in A$ : chiusa rispetto all'*unione*
- $\forall E, F \in A, E \cap F \in A$ : chiusura rispetto all'*intersezione*

### **i** Nota

La chiusura rispetto all'*intersezione* non è una vera proprietà, ma deriva dalla chiusura rispetto all'*unione* a cui viene applicata la *legge di De Morgan*



**i Nota**

Se la chiusura sull'*unione* vale anche per  $|\Omega| = \infty$ , allora  $\mathcal{A}$  viene chiamata  $\sigma$ -algebra

**💡 Informalmente**

L'algebra degli eventi non è un *vero* insieme di eventi, ma è un “*dizionario*” che sfruttiamo per definire quali *operazioni* e *variabili* sono ammesse su un  $\Omega$

**2.2.2. Assiomi di Kolmogorov**

Definiamo la funzione **probabilità**  $P : \mathcal{A} \rightarrow [0, 1]$ , che stabilisce la probabilità che un evento avvenga.

$P : \mathcal{A} \rightarrow [0, 1]$  è una funzione di probabilità se e solo se:

1.  $\forall E \in \mathcal{A}, 0 \leq P(E) \leq 1$ : la frequenza è sempre *positiva* e compresa tra 0 e 1
2.  $P(\Omega) = 1$ : un evento che si verifica tutte le  $n$  volte:  $\frac{n}{n} = 1$
3.  $\forall E, F \in \mathcal{A}, (E \cap F) = \emptyset \Rightarrow P(E \cup F) = P(E) + P(F)$

**i Nota**

La probabilità che accadano diversi eventi *distinti*  $E_i, E_j$  e *disgiunti*  $E_i \cap E_j = \emptyset$  è la *somma* delle loro probabilità:

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$$

**i Nota**

Formalmente la funzione probabilità è definita  $P : \mathcal{A} \rightarrow \mathbb{R}^+$  (numeri *reali positivi*), applicando gli assiomi il *codominio* viene ristretto a  $[0, 1]$ .

In modo analogo, il *primo assioma* stabilisce che il risultato dell'applicazione della funzione debba essere *positiva*, senza imporre un *limite superiore*, che poi viene aggiunto dal *secondo assioma*

**2.2.3. Teoremi derivati dagli assiomi****Probabilità del complemento**

$$\forall E \in \mathcal{A}, P(\overline{E}) = 1 - P(E)$$

**Probabilità dell'evento impossibile**

$$P(\emptyset) = 0$$

**Proprietà di monotonicità**

$$\forall E, F \in \mathcal{A} \mid E \subseteq F \Rightarrow P(E) \leq P(F)$$

**Probabilità dell'unione di eventi**

$$\forall E, F \in A, P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

#### 2.2.4. Spazi di probabilità ed Esiti equiprobabili

Definiamo lo **spazio di probabilità** come la tripla  $(\Omega, A, P)$  composta dallo spazio di *esiti possibili*  $\Omega$ , l'*algebra*  $A$  e la *funzione probabilità*  $P$ .

**Spazio equiprobabile** uno spazio è *equiprobabile* se gli eventi elementari (gli elementi  $\Omega$ ) hanno tutti la stessa probabilità:

$$P(E) = \frac{1}{N} \quad P(\{E_1, \dots, E_k\}) = \frac{k}{N}$$

Si dimostra con il secondo assioma di *Kolmogorov*:

$$P(\Omega) = 1 = P(\{e_1\}) + \dots + P(\{e_N\}) = \sum_{i=1}^N P(\{e_i\})$$

#### Nota

Uno spazio può essere *equiprobabile* solo se  $\Omega$  è un *insieme finito*

### 2.3. Probabilità condizionata

Dati due eventi  $E, F$ , la probabilità che si verifichi l'evento  $E$  sapendo che *si è verificato* l'evento  $F$  è detta **probabilità condizionata**:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

#### Nota

- $P(E|F)$  si legge “*probabilità di E dato F*”
- $E$  si dice evento *condizionato*
- $F$  si dice evento *condizionante*

#### Attenzione

In caso  $P(F) = 0$ , ovvero  $F = \emptyset$ , allora  $P(E|F) = \text{indefinita}$

#### Informalmente

Intuitivamente  $P(E|F)$  è la probabilità che preso un punto qualsiasi all'interno di  $F$ , il punto appartenga a  $E \cap F$ , quindi  $\frac{E \cap F}{F}$

#### 2.3.1. Regola di fattorizzazione

Dati due eventi  $E, F \in \Omega$ , la probabilità che accadano *entrambi* (la loro intersezione) è data dalla regola di *fattorizzazione*:

$$P(E \cap F) = P(F) \cdot P(E|F)$$

**Informalmente**

A differenza di una possibilità condizionata “semplice”, *non sappiamo* se  $F$  si sia già verificato o meno, quindi dobbiamo considerare anche la *sua possibilità* oltre a quella condizionata di  $E$

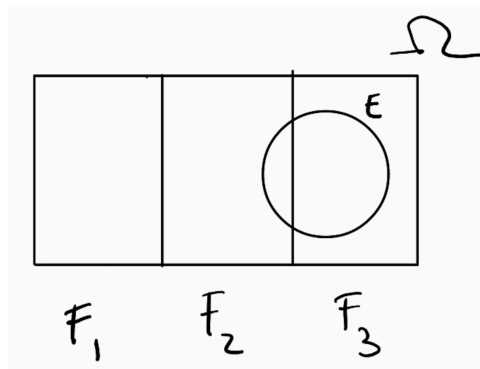
**2.3.2. Teorema delle probabilità totali**

Dato  $\Omega$  partizionato in  $F_1, \dots, F_n$  partizioni disgiunte, la probabilità che accada un evento  $E \in \Omega$  è:

$$P(E) = \sum_{i=1}^n P(F_i) \cdot P(E|F_i)$$

**Nota**

Insieme  $A$  partizionato:  $\bigcup_{i=1}^n F_i = A$  con  $\forall i, j, i \neq j, F_i \cap F_j = \emptyset$ . L'unione di tutte le partizioni è uguale all'insieme iniziale e tutte le partizioni sono *disgiunte*

Figure 6: Probabilità di  $E$ :

$$\begin{aligned} P(E) &= (P(F_1) \cdot P(E|F_1)) + (P(F_2) \cdot P(E|F_2)) + (P(F_3) \cdot P(E|F_3)) \\ &= \left(\frac{1}{3} \cdot 0\right) + \left(\frac{1}{3} \cdot \frac{1}{6}\right) + \left(\frac{1}{3} \cdot \frac{1}{2}\right) = \frac{2}{9} \end{aligned}$$

È possibile esprimere  $E$  come:

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap \bar{F}) \\ &= P(E | F)P(F) + P(E | \bar{F})P(\bar{F}) \\ &= P(E | F)P(F) + P(E | \bar{F})P(1 - P(F)) \end{aligned}$$

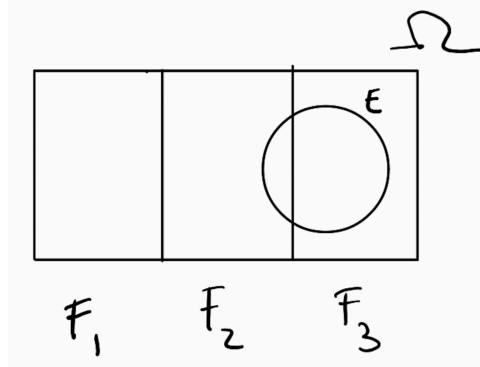
Altre trasformazioni utili:

$$\begin{aligned} (E \cap F) \cup (E \cap \bar{F}) &= E \cap (F \cup \bar{F}) = E \cap \Omega = E \\ (E \cap F) \cap (E \cap \bar{F}) &= E \cap (F \cap \bar{F}) = E \cap \emptyset = \emptyset \end{aligned}$$

**2.3.3. Teorema di Bayes**

Dato  $\Omega$  partizionato in  $F_1, \dots, F_n$  partizioni disgiunte, e un evento  $E$ , la probabilità che accada una certa  $F_k \subseteq \Omega$  è:

$$\begin{aligned}
 P(F_k | E) &= \frac{P(E | F_k)P(F_k)}{P(E)} \\
 &= \frac{P(E | F_k)P(F_k)}{\sum_{i=1}^n P(E | F_i)P(F_i)}
 \end{aligned}$$

Figure 7: Probabilità di  $F_2$ :

$$\begin{aligned}
 P(F_2) &= P(E|F_2) \cdot \frac{P(F_2)}{P(E)} \\
 &= \frac{\frac{P(E \cap F_2)}{P(F_2)} P(F_2)}{(P(F_1) \cdot P(E|F_1)) + (P(F_2) \cdot P(E|F_2)) + (P(F_3) \cdot P(E|F_3))} \\
 &= \frac{\frac{\frac{1}{6} \cdot \frac{1}{3}}{\frac{1}{3}} \cdot \frac{1}{3}}{\frac{2}{9}} = \frac{1}{4}
 \end{aligned}$$

#### 2.3.4. Classificatore naive-Bayes

Possiamo generalizzare il *teorema di Bayer* per ricavarne un **classificatore**: date delle *caratteristiche*  $X_1, \dots, X_n$  che assumono valore  $x_1, \dots, x_n$ , vogliamo assegnare l'oggetto  $y_k$  alla *classe* che massimizza la probabilità:

$$P(Y = y_k | X_1 = x_1, \dots, X_n = x_n)$$

Applicando il teorema di Bayes:

$$= \frac{P(X_1 = x_1, \dots, X_n = x_n | Y = y_k) \cdot P(Y = y_k)}{P(X_1 = x_1, \dots, X_n = x_n)}$$

La formula viene semplificata in modo “ingenuo” (*naive*), assumendo che le caratteristiche siano eventi indipendenti  $P(X_1 = x_1 \wedge X_2 = x_2 | Y) = P(X_1 = x_1) \cdot P(X_2 = x_2)$ :

$$\begin{aligned}
 &P(Y = y_k) \cdot \prod_{i=1}^n P(X_i = x_i | Y = y_k) \\
 &= \frac{\quad}{P(X_1 = x_1, \dots, X_n = x_n)}
 \end{aligned}$$

**Informalmente**

Questa *assunzione* è, appunto, *ingenua*: ad esempio, una persona con i capelli chiari è *più probabile* che abbia anche gli occhi chiari rispetto ad una persona con i capelli scuri. Le due caratteristiche *non* sono *indipendenti*.

Come capire *formalmente* se due eventi (o più) eventi sono *indipendenti* è descritto nel [paragrafo successivo](#).

Per trovare la classe alla quale *assegnare* l'oggetto, bisogna calcolare la probabilità per ogni possibile  $y_k$  e trovare il massimo:

$$= \arg \max_k P(Y = y_k) \cdot \prod_i^n P(X_i = x_i \mid Y = y_k)$$

**Nota**

Dato che ci interessa solo  $y_k$  massimo e il *denominatore* non dipende da  $k$ , allora possiamo *ignorarlo* dato che non influenzerà la scelta del massimo

**2.3.5. Eventi indipendenti**

Quando il verificarsi di un evento  $F$  *non influenza* la probabilità del verificarsi di un altro evento  $E$ , allora gli eventi si dicono **indipendenti**:

$$P(E \mid F) = P(E)$$

$$P(E \cap F) = P(E) \cdot P(F)$$

**Nota**

Sfruttando le formule viste in precedenza, è possibile verificare che i conti tornino:

$$P(E) = \frac{P(E \cap F)}{P(F)} = \frac{P(E) \cdot \cancel{P(F)}}{\cancel{P(F)}} = P(E)$$

**Informalmente**

È molto difficile rappresentare *graficamente* attraverso *diagrammi di Venn* eventi indipendenti, meglio non farlo :)

**Proprietà**

- Se  $E$  è indipendente da  $F$ ,  $F$  è indipendente da  $E$
- Se  $E$  e  $F$  sono indipendenti, allora anche  $E$  e  $\bar{F}$  sono indipendenti

**2.3.6. Indipendenza a tre o più eventi**

Tre eventi  $E, F, G$  sono *indipendenti* se valgono le proprietà:

- $P(E \cap F \cap G) = P(E) \cdot P(F) \cdot P(G)$
- $P(E \cap F) = P(E) \cdot P(F)$
- $P(F \cap G) = P(F) \cdot P(G)$

$$\bullet P(E \cap G) = P(E) \cdot P(G)$$

È possibile *estendere* la definizione ad un numero *arbitrario* di eventi:

Gli eventi  $E_1, \dots, E_n$  si dicono indipendenti se per ogni loro sottogruppo  $E_{a_1}, \dots, E_{a_r}$  con  $1 \leq a_1 \leq \dots \leq a_r \leq n$  vale l'equazione:

$$P\left(\bigcap_{i=1}^r E_{a_i}\right) = \prod_{i=1}^r P(E_{a_i})$$

## 2.4. Variabili aleatorie

Una *variabile aleatoria* o *casuale* (random variable) è una variabile che assume un valore *diverso* ogni osservazione. Permettono di codificare gli *eventi* in termini di numeri reali.

Dato uno spazio di probabilità  $(\Omega, A, P)$ , una variabile aleatoria è  $X : \Omega \rightarrow \mathbb{R}$  che associa ad ogni *esito* un *numero reale*.

### ! Attenzione

Non tutte le funzioni definite come  $X : \Omega \rightarrow \mathbb{R}$  rappresentano una *variabile aleatoria*

**Specificazioni** valori che possono essere assunti da una variabile aleatoria

**Supporto** l'*insieme* delle *specificazioni*, dove la probabilità non sia nulla  $P(X = x_i) \neq 0$

### i Nota

È possibile calcolare il *supporto* di una variabile aleatoria **discreta** calcolando l'*insieme* di punti in cui la funzione di massa non assuma valore nullo

### 2.4.1. Variabili aleatorie discrete

Una variabile aleatoria si dice discreta se il suo supporto è finito e numerabile (ovvero ha un numero finito di valori possibili). Dato  $[m, n]$  il range di valori che possono *essere assunti* dalla variabile  $X$ , vale:

$$1 = P(\Omega) = P\left(\bigcup_{i=m}^n \{X = i\}\right) = \sum_{i=m}^n P(X = i)$$

#### 2.4.1.1. Funzione indicatrice

Dati  $A, B$  due insiemi tali che  $A \subseteq B$ , la funzione indicatrice di  $A$  rispetto a  $B$  è la funzione  $I : B \rightarrow \{0, 1\}$  che vale:

$$I_{A(x)} = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases}$$

### 💡 Informalmente

La *funzione indicatrice* agisce da *filtro*, “limitando il dominio”. Quando viene *moltiplicata* con una probabilità la *annulla* o la *lascia inalterata*.

Ad esempio, la probabilità di un *dado* è:  $P(X = x) = \frac{1}{6} \cdot I_{\{1, \dots, 6\}}(x)$ , la funzione indicatrice *annulla* la probabilità di  $\frac{1}{6}$  in caso  $x$  non sia nel “dominio” del dado ( $1 \leq x \leq 6$ )

### 2.4.1.2. Funzione di massa di probabilità

Data una variabile aleatoria discreta  $X$  con supporto  $D$ , la sua *funzione di massa di probabilità*  $P_X : \mathbb{R} \rightarrow [0, 1]$  è la funzione che associa ad ogni valore  $x \in \mathbb{R}$  la probabilità che l'esito di  $X$  ne sia uguale:

$$\forall x \in \mathbb{R}, p_X(x) = P(X = x) \cdot I_D(x)$$

#### ⚠ Attenzione

La funzione di massa di probabilità **NON** vale per le variabili aleatorie continue

**Proprietà** che la funzione di massa di probabilità deve *rispettare*:

- $\forall x \in \mathbb{R}, f_X(x) \geq 0$ : non può essere negativa
- $\sum_{x \in D} f_X(x) = 1$ : la somma della funzione di massa per tutti i valori che  $x$  può assumere deve fare 1

### 2.4.1.3. Funzione di ripartizione

Data una variabile aleatoria  $X$ , la sua *funzione di ripartizione* o *distribuzione cumulativa*  $F_X : \mathbb{R} \rightarrow [0, 1]$  è la funzione che associa ad ogni valore  $x \in \mathbb{R}$  la probabilità che l'esito di  $X$  ne sia minore o uguale:

$$\forall x \in \mathbb{R}, F_X(x) = P(X \leq x)$$

#### ⚠ Attenzione

La funzione di ripartizione è *valida* sia per le variabili aleatorie discrete che per le variabili aleatorie continue

#### i Nota

È possibile calcolare la probabilità per un *valore maggiore* di una certa soglia sfruttando il complementare:

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x)$$

#### i Nota

È possibile calcolare la probabilità per un valore *compreso* tra due estremi:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

La funzione di ripartizione può essere vista come la *somma* applicando alla *funzione di massa* tutti i valori minori uguali di un certo valore  $a$ :

$$F(a) = \sum_{x \leq a} p(x)$$

**Informalmente**

Per una variabile aleatoria discreta,  $F$  è una *funzione a gradini*, costante tra gli intervalli dei valori assunti da  $X$ , che salta di  $p(x)$  ad ogni nuovo valore

**2.4.1.4. Valore atteso**

Il *valore atteso* di una variabile aleatoria  $X$  è un *indice dimensionale* di centralità delle *specificazioni* della variabile aleatoria.

$$E[X] = \sum_i x_i \cdot P(X = x_i) = \sum_i x_i \cdot p(x_i)$$

**Informalmente**

Il *valore atteso* è semplicemente la “*media pesata*” per ogni possibile valore nel *dominio* (specificazione) di una variabile aleatoria

**Nota**

Il valore atteso può essere indicato con  $E$  o con  $\mu$

**Proprietà** del valore atteso:

- il *valore atteso* di una funzione indicatrice è uguale alla *probabilità dell'evento*:

$$E[I_A] = P(A)$$

- il *valore atteso* di una variabile aleatoria discreta  $X$  opera in modo *lineare*:

$$Y = a \cdot X + b \quad E[Y] = a \cdot E[X] + b$$

- data una qualsiasi *funzione* reale  $g$  e una variabile aleatoria  $X$  con funzione di massa  $p$ , allora vale:

$$E[g(X)] = \sum_i g(x_i) \cdot p(x_i)$$

- data una qualsiasi *funzione* reale  $g$  di due variabili e due variabili aleatorie discrete  $X, Y$ , allora vale:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) \cdot p(x, y)$$

**2.4.1.5. Varianza**

Sia  $X$  una variabile aleatoria di media  $\mu$ , la varianza di  $X$  è:

$$\text{Var}(X) = G_X^2 = E[(X - \mu)^2]$$



**i Nota**

Formula alternativa per la varianza:

$$\begin{aligned}
 \text{Var}(X) &= E[(X - \mu)^2] \\
 &= E[X^2 - 2\mu X + \mu^2] \\
 &= E[X^2] - 2\mu E[X] + \mu^2 \\
 &= E[X^2] - 2\mu^2 + \mu^2 \\
 &= E[X^2] - E[X]^2 \\
 &= \sum_i x_i^2 \cdot P(X = x_i) - \left( \sum_i x_i \cdot P(X = x_i) \right)^2
 \end{aligned}$$

**Proprietà** varianza:

- la varianza della funzione indicatrice è la probabilità dell'*evento* moltiplicata per la probabilità dell'*evento complementare*

$$\text{Var}(I) = P(A) \cdot P(\overline{A})$$

- la varianza non opera in modo lineare:

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

**Deviazione standard**  $\sigma_X = \sqrt{\text{Var}(X)}$

#### 2.4.2. Variabili aleatorie multivariate

Oltre alle variabili aleatorie **univariate**, è possibile utilizzare un *vettore* di lunghezza arbitraria, ottenendo una variabile aleatoria **multivariata**.

**i Nota**

Quando il vettore è lungo 2 elementi, la variabile aleatoria si dice **bivariata**

##### 2.4.2.1. Funzione di ripartizione congiunta

Sia  $A$  una variabile aleatoria bivariata formata da  $X, Y$  variabili aleatorie *univariate discrete*, allora la loro *funzione di ripartizione congiunta* è:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

**i Nota**

La virgola dentro la probabilità denota l'intersezione:

$$P(X \leq x, Y \leq y) = P(X \leq x \cap Y \leq y)$$

È possibile *estendere* a variabili aleatorie *multivariate* di dimensione arbitraria:

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

**i Nota**

Si dice *funzione di ripartizione (o massa) marginale* quando da una *funzione di ripartizione (o massa) congiunta* estraggo una *funzione di ripartizione (o massa)* di una variabile **univariata**

Possiamo ottenere una  $F_X(x)$  **funzione di ripartizione marginale** di  $X$  nel seguente caso:

$$\begin{aligned}\lim_{y \rightarrow +\infty} F_{X,Y}(x, y) &= \lim_{y \rightarrow +\infty} P(X \leq x, Y \leq y) \\ &= P(X \leq x) \cdot \lim_{y \rightarrow +\infty} P(Y \leq y) \\ &= P(X \leq x) \cdot \Omega \\ &= P(X \leq x) \\ &= F_X(x)\end{aligned}$$

**2.4.2.2. Funzione di massa di probabilità congiunta**

Siano  $X, Y$  due variabili aleatorie *univariate discrete*, allora la loro *funzione di massa di probabilità congiunta*  $p : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$  è:

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

È possibile *estendere* a variabili aleatorie *multivariate* di dimensione arbitraria:

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$$

Possiamo ottenere una  $p_X(x)$  **funzione di massa di probabilità marginale** di  $X$  nel seguente caso:

$$\begin{aligned}\sum_{y_i \in D} p_{X,Y}(x, y) &= \sum_{y_i \in D} P(\{X = x\} \cap \{Y = y\}) \\ &= P\left(\{X = x\} \cap \bigcup_{y_i \in D} \{Y = y\}\right) \\ &= P(\{X = x\} \cdot \Omega) \\ &= P(\{X = x\}) \\ &= p_X(x)\end{aligned}$$

**2.4.2.3. Indipendenza**

Due variabili aleatorie  $X, Y$  si dicono **indipendenti** se  $\forall A, B \subseteq \mathbb{R}$ ,  $X \in A$  e  $Y \in B$  sono indipendenti:

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

$$P(X \leq A, Y \leq B) = P(X \leq A) \cdot P(Y \leq B)$$

$$p(x, y) = p_X(x) \cdot p_Y(y)$$

$$F(a, b) = F_X(a) \cdot F_Y(b)$$

**i Nota**

Questi risultati sono *dimostrabili* usando gli *assiomi delle probabilità*

È possibile *estendere* a variabili aleatorie *multivariate* di dimensione arbitraria:

$$P(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n P(X_i \in A_i)$$

**2.4.2.4. Valore atteso**

Il valore atteso della *somma* di variabili aleatorie discrete è:

$$E\left[\sum_i X_i\right] = \sum_i E[X_i]$$

Il valore atteso del *prodotto* di variabili aleatorie discrete è:

$$E\left[\prod_i X_i\right] = \prod_i E[X_i]$$

**2.4.2.5. Covarianza**

Siano  $X$  e  $Y$  due variabili aleatorie di media  $\mu_X$  e  $\mu_Y$ , la loro *covarianza* è:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

**i Nota**

Formula alternativa:

$$\begin{aligned}\text{Cov}(X, Y) &= [XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

**Proprietà della covarianza:**

- simmetria:  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- generalizzazione concetto di varianza:  $\text{Cov}(X, X) = \text{Var}(X)$
- linearità:
  - $\text{Cov}(aX, Y) = \text{Cov}(X, aY) = a \text{Cov}(X, Y)$
  - $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$

**2.4.2.6. Varianza**

Siano  $X$  e  $Y$  due variabili aleatorie la loro *varianza* della loro *somma* è:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$$

È possibile *estendere* a variabili aleatorie *multivariate* di dimensione arbitraria:

$$\text{Var}\left(\sum_i^n X_i\right) = \sum_i^n \text{Var}(X_i) + \sum_i^n \sum_{j, j \neq i}^n \text{Cov}(X_i, X_j)$$

### 2.4.3. Variabili aleatorie continue

Una variabile aleatoria si dice **continua** quando ha un *supporto non numerabile*.

#### 2.4.3.1. Funzione densità di probabilità



##### Informalmente

La funzione di massa (come spiegato sotto) non ha più senso per le *variabili aleatorie continue*, quindi lo stesso concetto prende il nome di *funzione densità di probabilità*

Siccome  $X$  deve per forza assumere un valore in  $\mathbb{R}$ , allora la *funzione di densità* ( $f_X(x)$ ) deve rispettare:

$$P(X \in \mathbb{R}) = \int_{-\infty}^{+\infty} f_X(x) dx = 1$$

Per variabili aleatorie *continue* non ha senso cercare la probabilità assunta da un *singolo valore*, infatti:

$$P(X = a) = \int_a^a f_X(x) dx = 0$$

Per questo motivo si ragiona in termini di *intervalli* di probabilità:

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

Esiste una relazione tra la funzione di ripartizione  $F$  (che **vale anche** per le variabili aleatorie continue) con la funzione di densità  $f$ :

$$\begin{aligned} F(a) &= P(X \leq a) \\ &= P(X \in (-\infty, a]) \\ &= \int_{-\infty}^a f_X(x) dx \end{aligned}$$

Quindi la **funzione di densità** è uguale alla derivata della funzione di ripartizione  $F$ :

$$f_X(a) = F'(a)$$

#### 2.4.3.2. Valore atteso

Il *valore atteso* di una variabile aleatoria continua vale:

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx$$



##### Nota

Formula alternativa per il valore atteso:

$$E(X) = \int_0^{+\infty} 1 - F_X(x) dx$$

### 2.4.3.3. Varianza

La *varianza* di una variabile aleatoria continua vale:

$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x) dx$$

### 2.4.3.4. Disuguaglianza di Markov



#### Informalmente

Permette di ottenere un limite superiore alla probabilità dalla sola conoscenza del valore atteso

Sia  $X$  una *variabile aleatoria*  $X \geq 0$ , allora  $\forall a > 0 \in \mathbb{R}$ , vale:

$$P(X \geq a) \leq \frac{E[X]}{a}$$



#### Nota

Possiamo trarre che:

$$P(X < a) = 1 - P(X \geq a) \geq 1 - \frac{E[x]}{a}$$



#### Dimostrazione

Variabili aleatorie **discrete**:

$$\begin{aligned} E[X] &= \sum_{x \geq 0} x \cdot p(x) \\ &= \sum_{x \leq a} x \cdot p(x) + \sum_{x \geq a} x \cdot p(x) \geq \sum_{x \geq a} x \cdot p(x) = \\ &\geq \sum_{x \geq a} a \cdot p(x) = \\ &\geq a \cdot \sum_{x \geq a} p(x) = \\ &\geq a \cdot P(X \geq a) \end{aligned}$$

Quindi

$$E[X] \geq a \cdot P(X \geq a) \Rightarrow P(X \geq a) \leq \frac{E[X]}{a}$$

**! Dimostrazione**

Variabili aleatorie **continue**:

$$\begin{aligned}
 E[X] &= \int_{-\infty}^{+\infty} x f_X(x) \, dx \\
 &= \int_0^a x f_X(x) \, dx + \int_a^{+\infty} x f_X(x) \, dx \geq \int_a^{+\infty} x f_X(x) \, dx = \\
 &\geq \int_a^{+\infty} a f_X(x) \, dx = \\
 &\geq a \int_a^{+\infty} f_X(x) \, dx = \\
 &\geq a \cdot P(X \geq a)
 \end{aligned}$$

Quindi

$$E[X] \geq a \cdot P(X \geq a) \Rightarrow P(X \geq a) \leq \frac{E[X]}{a}$$

**2.4.3.5. Disuguaglianza di Chebyshev****Informalmente**

Permette di ottenere un limite superiore alla probabilità che il valore di una variabile aleatoria si discosti dal suo valore atteso di una quantità maggiore o uguale a una soglia scelta

Sia  $X$  una *variabile aleatoria* di valore atteso  $E[X] = \mu$  e varianza  $\text{Var}(X) = \sigma^2$ , allora:

$$\forall r > 0, \quad P(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2}$$

**Informalmente**

$|X - \mu|$  è la distanza tra la variabile aleatoria e il suo valore atteso

**Nota**

Un'applicazione della *disuguaglianza di Chebyshev* riguarda la *deviazione standard*: esprime l'andamento della probabilità allontanandosi dal valore atteso di quantità ripetute della deviazione standard:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

**! Dimostrazione**

$$|X - \mu| \geq r \iff (X - \mu)^2 \geq r^2$$

dunque:

$$P(|X - \mu| \geq r) = P((X - \mu)^2 \geq r^2)$$

$$P((X - \mu)^2 \geq r^2) \leq E\left[\frac{(X - \mu)^2}{r^2}\right] \text{ per Markov}$$

$$P(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2}$$

**2.4.4. Modelli di distribuzione**

Alcune *distribuzioni/modelli* di variabili aleatorie sono molto *frequenti*, di conseguenza esistono dei risultati notevoli.

**2.4.4.1. Modello di Bernoulli  $X \sim B(p)$** 

La variabile aleatoria può assumere solo due specificazioni: **fallimento** o **successo**, ovvero il loro supporto è  $D_X = \{0, 1\}$

Il parametro  $p$  indica la probabilità che  $X = 1$  con  $p \in [0, 1]$

**Funzione di massa**

$$\begin{aligned} p_X(x) = P(X = x) &= p^x(1-p)^{(1-x)} I_{\{0,1\}}(x) \\ &= \begin{cases} 1-p & \text{per } x = 0 \\ p & \text{per } x = 1 \\ 0 & \text{altrimenti} \end{cases} \end{aligned}$$

**Funzione di ripartizione**

$$\begin{aligned} F_X(x) = P(X \leq x) &= (1-p) I_{[0,1]}(x) + I_{(1,+\infty)}(x) \\ &= \begin{cases} 0 & \text{se } x < 0 \\ 1-p & \text{se } 0 \leq x < 1 \\ 1 & \text{se } x \geq 1 \end{cases} \end{aligned}$$

**Valore atteso**

$$E[X] = p$$

**! Dimostrazione**

$$\begin{aligned} E[X] &= \sum_x x \cdot P(X = x) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 1 \cdot P(X = 1) = p \end{aligned}$$

**Varianza**

$$\text{Var}(X) = p(1-p)$$

### ! Dimostrazione

$$\begin{aligned}\text{Var}(X) &= E[(X - \mu)^2] \\ &= E[(X - p)^2] \\ &= \sum_x (x - p)^2 \cdot P(X = x) \\ &= (0 - p)^2 \cdot P(X = 0) + (1 - p)^2 \cdot P(X = 1) \\ &= p^2(1 - p) + p(1 - p)^2 \\ &= p(1 - p)(p + 1 - p) = p(1 - p)\end{aligned}$$

#### 2.4.4.2. Modello binomiale $X \sim B(n, p)$

Il modello ripete  $n$  volte un esperimento bernoulliano di probabilità  $p$ , dove  $n$  e  $p$  sono i due parametri del modello. Il supporto del modello è  $D_X = \{0, 1, \dots, n\}$ .

#### Funzione di massa

$$p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)} I_{\{0, \dots, n\}}(x)$$

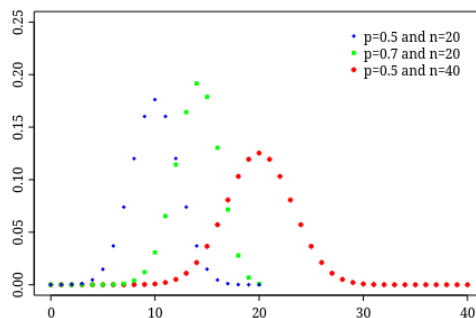


Figure 8: Funzione di massa modello binomiale

#### Funzione di ripartizione

$$\begin{aligned}F_X(x) &= P(X \leq x) = \\ &= \sum_{i=0}^x p_X(i) \cdot I_{[0, n]}(x) + I_{(n, +\infty)}(x) \\ &= \begin{cases} \sum_{i=0}^x p_X(i) & \text{per } x \leq n \\ 1 & \text{per } x > n \end{cases}\end{aligned}$$

### ! Attenzione

Nella formula della funzione di ripartizione,  $p_X$  indica la *funzione di massa*, solo  $p$  indica la *probabilità di successo*

#### Valore atteso



$$E[X] = n \cdot p$$

**! Dimostrazione**

$$\begin{aligned} E[X] &= E\left[\sum_i^n X_i\right] \\ &= \sum_i^n E[X_i] = n \cdot p \end{aligned}$$

**Varianza**

$$\text{Var}(X) = n \cdot p(1 - p)$$

**Relazioni**

Siano  $X_1 \sim B(n, p)$  e  $X_2 \sim B(m, p)$  indipendenti, allora:

$$X_1 + X_2 = \sum_{i=1}^n X_{1,i} + \sum_{j=1}^m X_{2,j} = \sum_{i=1}^{n+m} Y_i = Y$$

dove  $Y \sim B(n + m, p)$

**3. Statistica inferenziale****4. Cheatsheet Python****5. Cheatsheet integrali**