Probabilità

24 ottobre 2023

Indice

1 Sommatorie e produttorie **9**
1.1 Sommatorie . . . 9
1.1.1 Sommatoria singola . . . 9
1.1.1.1 Definizione . . . 9
1.1.1.2 Tecniche utili . . . 10
1.1.1.3 Proprietà . . . 11
1.1.1.4 Applicazioni . . . 13
1.1.2 Sommatorie doppie . . . 15
1.1.2.1 Definizioni . . . 15
1.1.2.2 Proprietà . . . 16
1.2 Produttorie . . . 20
1.2.1 Produttoria singola . . . 20
1.2.1.1 Proprietà . . . 20
1.3 Esercizi . . . 22

2 Calcolo combinatorio **23**
2.1 Introduzione . . . 23
2.2 Casistica principale . . . 24
2.2.1 Permutazioni . . . 24
2.2.2 Disposizioni . . . 25
2.2.3 Combinazioni . . . 26
2.2.3.1 Combinazioni semplici . . . 26
2.2.3.2 Combinazioni con ripetizione . . . 26
2.3 Coefficiente binomiale e multinomiale . . . 27
2.3.1 Coefficiente binomiale . . . 27
2.3.1.1 Definizione . . . 27
2.3.1.2 Proprietà . . . 28
2.3.1.3 Origine del nome . . . 29
2.3.2 Il coefficiente multinomiale . . . 29
2.3.2.1 Definizione . . . 29
2.3.2.2 Origine del nome . . . 30
2.4 Calcolo combinatorio e funzioni . . . 30
2.4.1 Principio dell'overcounting . . . 31
2.4.2 Funzioni (disposizioni con ripetizione) . . . 31
2.4.3 Funzioni iniettive (disposizioni semplici) . . . 31
2.4.4 Permutazioni di un insieme (permutazioni semplici) . . . 31
2.4.5 Funzioni caratteristiche (coefficiente binomiale) . . . 31
2.5 Esercizi . . . 32

3 Introduction **35**
3.1 Probability space . . . 35
3.1.1 Sample space, events . . . 35
3.1.1.1 Events algebra . . . 36
3.1.1.2 Relationship between events . . . 37
3.1.2 σ-field $\mathscr{F}$ (or σ-algebra $\mathscr{A}$) . . . 38
3.1.3 Probability measure $\mathbb{P}$ . . . 40
3.2 Probability . . . 40
3.2.1 Immediate or useful general results . . . 40
3.2.2 Finite equiprobable Ω and probability evaluation . . . 43
3.2.3 Conditional probability . . . 45
3.2.4 Probability of intersection . . . 45
3.2.5 Law of total probability . . . 46
3.2.6 Bayes formula . . . 48
3.3 Indipendent events . . . 50
3.4 Further topics . . . 53
3.4.1 Odds ratio . . . 53
3.4.2 Conditional probability 2 . . . 54
3.4.2.1 È una probabilità . . . 54
3.4.2.2 Risultati . . . 55
3.4.2.3 Condizionare su più eventi . . . 56
3.4.2.4 Indipendenza condizionata, aggiornamento delle stime . . . 57
3.5 Esercizi vari . . . 59

4 Random variables **61**
4.1 Intro . . . 61
4.1.1 Discrete and continuous rvs . . . 62
4.2 Functions of random variables . . . 63
4.2.1 Discrete rvs: PMF, CDF . . . 63
4.2.2 Continuous rvs: PDF, CDF . . . 64
4.2.3 Distribution functions (Rigo's style) . . . 66
4.2.3.1 Discrete rvs . . . 68
4.2.3.2 Singular continuous rvs . . . 68
4.2.3.3 Absolutely continuous rvs . . . 69
4.2.3.4 n-variate random variables . . . 70
4.3 Other useful rv functions . . . 73
4.3.1 Support indicator . . . 73
4.3.2 Survival and hazard function . . . 73
4.4 Transformation of rvs . . . 74
4.4.1 Discrete rv transform . . . 74
4.4.2 Continuous rvs transform (linear case) . . . 75
4.5 Rvs independence . . . 75
4.5.1 Independence, iid rvs . . . 75
4.5.2 Conditional independence . . . 76
4.6 Moments . . . 76
4.6.1 Expected value . . . 77
4.6.2 Variance . . . 81
4.6.3 Asymmetry/skewness and kurtosis . . . 82
4.6.3.1 Asymmetry/Skewness . . . 83

4.6.3.2 Kurtosis 84
4.7 Exercises 84
4.8 Probability models and R 89

5 Discrete random variables **91**
5.1 Dirac 91
5.2 Bernoulli 91
5.2.1 Definition 91
5.2.2 Functions 92
5.2.3 Moments 92
5.3 Indicator rv for an event 92
5.3.1 Definition, properties 92
5.3.2 Probability/expected value link 93
5.3.3 Some application: probability 94
5.3.4 Applications: expected value evaluation 95
5.4 Binomial 96
5.4.1 Definition 96
5.4.2 Functions 97
5.4.3 Moments 97
5.4.4 Shape 98
5.4.5 Variabili derivate 100
5.5 Hypergeometric 101
5.5.1 Definition 101
5.5.2 Functions 101
5.5.3 Moments 102
5.5.4 Struttura essenziale ed esperimenti assimilabili 103
5.5.5 Connessioni con la binomiale 103
5.5.5.1 Dall'ipergeometrica alla binomiale 103
5.5.5.2 Dalla binomiale all'ipergeometrica 105
5.6 Geometric 105
5.6.1 Definition 105
5.6.2 Functions 106
5.6.3 Moments 107
5.6.4 Shape 109
5.6.5 Assenza di memoria 109
5.6.6 Alternative definition (first success distribution) 110
5.7 Negative binomial 111
5.7.1 Definition 111
5.7.2 Functions 111
5.7.3 Moments 112
5.7.4 Shape 112
5.7.5 Alternative definition 113
5.7.5.1 Definition 113
5.7.5.2 Functions 113
5.7.5.3 Moments 114
5.8 Poisson 114
5.8.1 Definition 114
5.8.2 Functions 114
5.8.3 Moments 115
5.8.4 Shape 116

5.8.5 Origine e approssimazione . . . 116
5.8.6 Legami con la binomiale . . . 118
5.8.6.1 Dalla Poisson alla binomiale . . . 118
5.8.6.2 Dalla binomiale alla Poisson . . . 119
5.8.7 Processo di Poisson . . . 120
5.9 Discrete uniform . . . 121
5.9.1 Definition . . . 121
5.9.2 Functions . . . 121
5.9.3 Moments . . . 122

6 Variabili casuali continue **123**
6.1 Logistica . . . 123
6.1.1 Origine/definizione . . . 123
6.1.2 Funzioni . . . 123
6.1.3 Versione generale . . . 123
6.2 Uniforme continua . . . 124
6.3 Esponenziale . . . 127
6.4 Normale/Gaussiana . . . 128
6.5 Gamma . . . 130
6.6 Chi-quadrato . . . 132
6.7 Beta . . . 133
6.8 T di Student . . . 134
6.9 F di Fisher . . . 135
6.10 Lognormale . . . 137
6.11 Weibull . . . 137
6.12 Pareto . . . 138

7 Random vectors **141**
7.1 Intro . . . 141
7.1.1 Relationship between rvs . . . 141
7.1.1.1 Covariance . . . 141
7.1.1.2 Correlation coefficient . . . 143
7.1.2 Relationship between rvs . . . 143
7.1.2.1 Covariance . . . 143
7.1.2.2 Correlation coefficient . . . 144

8 Misc topics **145**
8.1 Characteristic function . . . 145
8.1.1 Characteristic function . . . 145
8.1.2 Moment generating function . . . 148
8.2 Order statistics . . . 153
8.2.1 Minimum . . . 153
8.2.2 Maximum . . . 154
8.2.3 Generalized $X_{(i)}$ . . . 155
8.3 Inequalities . . . 156
8.3.1 Markov (Viroli) . . . 156
8.3.2 Tchebychev (Viroli) . . . 156
8.3.3 Tchebychev (Rigo) . . . 157
8.3.4 Jensen . . . 158
8.4 Transformation of random variables . . . 159

8.5 Conditional distribution 161
8.6 Multivariate normal 166
8.7 Exercises vari 168

9 Convergence **175**
9.1 Convergence in probability 175
9.1.1 Theorem: weak law of large numbers 177
9.2 Convergence in law/distribution 178
9.2.1 Theorem: central limit theorem 179
9.3 Convergence in mean of order k 180
9.3.1 Strong law of large numbers 182
9.4 Almost sure convergence 183
9.5 Relationship between convergences 183
9.6 Convergence exercises 184
9.7 Delta method 186
9.8 Rigo stuff 191

10 Simulation **195**
10.1 Sampling values from rvs 195
10.1.1 Inversion method 195
10.1.2 Accept-reject method 196
10.2 R exercises 200
10.2.1 CLT 200
10.2.2 Inversion method 200
10.2.3 Accept-reject 202

Capitolo 1

Sommatorie e produttorie

1.1 Sommatorie

1.1.1 Sommatoria singola

1.1.1.1 Definizione

Definizione 1.1.1 (Sommatoria singola)**.** Se $(a_j)_{j\in J}$, $a : J \to \mathbb{C}$ è una famiglia *finita* di numeri complessi (ossia l'insieme degli indici J è finito), è definita così la somma di tutti i numeri a_j per $j \in J$ e si indica con

$$\sum_{j\in J} a_j \tag{1.1}$$

Osservazione importante 1. Se $J = \emptyset$ si pone per definizione $\sum_{j\in J} a_j = 0$.

Osservazione 1. È importante osservare che il simbolo $\sum_{j\in J} a_j$ non dipende da j ma solo dall'intero insieme J e dalla funzione $a : J \to \mathbb{C}$; la variabile j si dice *muta*, si ha cioè

$$\sum_{j\in J} a_j = \sum_{k\in J} a_k = \sum_{\lambda\in J} a_\lambda$$

Osservazione 2. Laddove si possa riesce ad esprimere il generico a_j come una $f(j)$ dipendente dall'indice la sommatoria degli elementi può essere usata anche per la somma di valori assunti di funzione che utilizza l'indice come input,

$$\sum_{j\in J} f(j)$$

Esempio 1.1.1. Se $a_j = 1/j$, allora $\sum_{j\in J} \frac{1}{j}$

Proposizione 1.1.1 (Biezione e cambio di indici)**.** *In generale se si ha una funzione $\varphi : K \to J$ biettiva allora:*

$$\sum_{j\in J} a_j = \sum_{k\in K} a_{\varphi(k)}$$

Osservazione 3. Ossia possiamo anche utilizzare un altro set di indici K posto che, per garantire l'uguaglianza, vi sia una biezione che ci garantisce che questi vadano a puntare agli stessi elementi.

Osservazione 4 (Indici comuni). Spesso l'insieme J degli indici è $I_n = \{1, \ldots, n\}$ e si scrive allora anche

$$\sum_{j=1}^{n} a_j \qquad \text{oppure} \sum_{1 \leq j \leq n} a_j \qquad \text{intendendo} \sum_{j \in I_n} a_j$$

e per esteso si intende:

$$\sum_{j \in I_n} a_j = a_1 + \ldots + a_n$$

Definizione 1.1.2 (Sommatoria di sottofamiglia)**.** Si intende la sommatoria di un pezzo, ossia di una sottofamiglia di successione $a : \mathbb{N} \to \mathbb{C}$ compresa tra due indici m, n, con $m \leq n$:

$$\sum_{j=m}^{n} a_j = \sum_{m \leq j \leq n} a_j = a_m + \ldots + a_n$$

1.1.1.2 Tecniche utili

Osservazione 5. La traslazione di indici consiste nel cambiare gli indici senza cambiare gli oggetti puntati.

Proposizione 1.1.2 (Traslazione di indici)**.** *Per effettuarla occorre sostituire $j + offset$ al posto di j negli indici della sommatoria e sostituendo $j - offset$ nei termini indiciati (sia offset un termine positivo o negativo)*

$$\sum_{j=m}^{n} a_j = \sum_{j=m-p}^{n-p} a_{j+p} = \sum_{j=m+p}^{n+p} a_{j-p} \tag{1.2}$$

Dimostrazione. È una applicazione di 1.1.1. □

Osservazione 6. In sostanza per garantire l'uguaglianza delle sommatorie, basta che alla fine l'indice punti allo stesso elemento poi la formula può essere cambiata a piacere.

Proposizione 1.1.3 (Riflessione di indici)**.** *Mediante questa tecnica si sommano gli stessi elementi, posti però in ordine inverso (si somma dall'indice originario più alto al più basso):*

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} a_{n-i+1} = \sum_{i=0}^{n-1} a_{n-i} \tag{1.3}$$

Dimostrazione. È una permutazione su indici quindi funzione biettiva e si applica 1.1.1. L'ultima eguaglianza si giustifica mediante una traslazione di indici (sostituendo con $i - 1$ negli indici della sommatoria e $i + 1$ nei termini della stessa). □

Osservazione 7. Il cambiamento di indice può tornare utile nel caso di sommatoria di funzione laddove si vogliano normalizzare un po' gli indici

Proposizione 1.1.4 (Cambiamento di indice)**.** *Sia $\sum_{i\in I} f(i)$ la sommatoria di nostro interesse. Supponendo che vi sia una funzione biettiva $\varphi : I \to J$ che esprima gli indici in un nuovo insieme e che sia $J = \varphi(I)$; esisterà anche $\varphi^{-1} : J \to I$. Dato che un singolo $j = \varphi(i)$ si potranno applicare φ e φ^{-1} rispettivamente a indici ed elementi della sommatoria, ottenendo lo stesso risultato poiché, per definizione*

$$\sum_{i\in I} f(i) = \sum_{j=\varphi(i)\in J} f(\varphi^{-1}(j)) \tag{1.4}$$

Osservazione 8. L'equazione di sopra ci dice che dobbiamo applicare la biezione trovata agli indici e la sua inversa all'argomento della sommatoria

Esempio 1.1.2. Ipotizziamo di avere

$$\sum_{i=-10}^{-8} \frac{1}{i+1} = -\frac{1}{9} - \frac{1}{8} - \frac{1}{7}$$

Al fine di semplificare gli indici della sommatoria applichiamo a questi $\varphi : I \to J$ definita come $j = i + 10$ (si vede che φ è biettiva: è una retta), si applica poi $\varphi^{-1} : J \to I$ definita come $i = j - 10$ agli elementi della sommatoria

$$\sum_{i=-10}^{-8} \frac{1}{i+1} = \sum_{j=0}^{2} \frac{1}{j-9} = -\frac{1}{9} - \frac{1}{8} - \frac{1}{7}$$

Se si desidera, possiamo tornare all'indice iniziale con la sostituzione $i = j$:

$$\sum_{j=0}^{2} \frac{1}{j-9} = \sum_{i=0}^{2} \frac{1}{i-9}$$

Quindi:

$$\sum_{i=-10}^{-8} \frac{1}{i+1} = \sum_{i=0}^{2} \frac{1}{i-9}$$

1.1.1.3 Proprietà

Osservazione 9. Valgono le seguenti *proprietà* (che possono essere utili lette sia da sinistra a destra che viceversa).

Proposizione 1.1.5 (Sommatoria di costante)**.** *Se k è una costante che non dipende dall'indice i, allora:*

$$\sum_{i=1}^{n} k = nk \tag{1.5}$$

Dimostrazione. Si pone convenzionalmente $a_i = k$, per cui:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n = \underbrace{k + k + \ldots + k}_{\text{n volte}} = kn$$

□

Proposizione 1.1.6 (Sommatoria di prodotto per costante)**.** *Se k è una costante che non dipende dall'indice i, allora:*

$$\sum_{i=1}^{n} ka_i = k\sum_{i=1}^{n} a_i \tag{1.6}$$

Dimostrazione. Infatti

$$\sum_{i=1}^{n} ka_i = ka_1 + ka_2 + \ldots + ka_n = k(a_1 + a_2 + \ldots + a_n) = k\sum_{i=1}^{n} a_i$$

□

Proposizione 1.1.7 (Scomposizione/somme su sottoinsiemi)**.** *Se $m > n$, allora:*

$$\sum_{i=1}^{n} a_i + \sum_{i=n+1}^{m} a_i = \sum_{i=1}^{m} a_i \tag{1.7}$$

Dimostrazione. Infatti

$$\sum_{i=1}^{n} a_i + \sum_{i=n+1}^{m} a_i = (a_1 + \ldots + a_n) + (a_{n+1} + \ldots + a_m) = \sum_{i=1}^{m} a_i$$

□

Osservazione importante 2. Generalizzando, se Λ è un insieme di indici, $a : \Lambda \to \mathbb{C}$ una famiglia di complessi, e J, K sottoinsiemi finiti *disgiunti* di Λ si ha:

$$\sum_{\lambda \in J \cup K} a_\lambda = \sum_{\lambda \in J} a_\lambda + \sum_{\lambda \in K} a_\lambda \tag{1.8}$$

Proposizione 1.1.8 (Sommatoria di somme/additività rispetto alle famiglie)**.** *Si ha che:*

$$\sum_{i=1}^{n} (a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i \tag{1.9}$$

Dimostrazione. Infatti:

$$\begin{aligned} \sum_{i=1}^{n} (a_i + b_i) &= (a_1 + b_1) + (a_2 + b_2) + \ldots + (a_n + b_n) \\ &= (a_1 + a_2 + \ldots + a_n) + (b_1 + b_2 + \ldots + b_n) \\ &= \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i \end{aligned}$$

□

Osservazione importante 3. Generalizzando, se Λ è un insieme di indici, $a : \Lambda \to \mathbb{C}$ una famiglia di complessi e $b : \Lambda \to \mathbb{C}$ è un'altra famiglia di complessi si può definire la somma puntuale $a + b : \Lambda \to \mathbb{C}$ delle due famiglie ponendo $(a+b)(\lambda) = a_\lambda + b_\lambda$ per ogni $\lambda \in \Lambda$. Si ha anche che per ogni sottoinsieme finito J di Λ:

$$\sum_{j \in J} (a_j + b_j) = \sum_{j \in J} a_j + \sum_{j \in J} b_j \tag{1.10}$$

Proposizione 1.1.9 (Sommatoria di termini lineari)**.** *Se k e c sono costanti che non dipendono dall'indice i,*

$$\sum_{i=1}^{n}(ka_i + c) = nc + k\sum_{i=1}^{n} a_i \tag{1.11}$$

Dimostrazione. Alla luce delle proprietà precedentemente viste:

$$\sum_{i=1}^{n}(ka_i + c) = \sum_{i=1}^{n} ka_i + \sum_{i=1}^{n} c = nc + k\sum_{i=1}^{n} a_i$$

□

Osservazione 10. Si noti che prima abbiamo preposto nc alla sommatoria per evitare confusione; un altro modo sarebbe $k\left(\sum_{i=1}^{n} a_i\right) + nc$

1.1.1.4 Applicazioni

Proposizione 1.1.10 (Prodotti di sommatorie)**.** *Se $(a_j)_{j\in J}$ è una famiglia finita di numeri complessi e $(b_k)_{k\in K}$ è un'altra famiglia finita di numeri complessi si ha:*

$$\left(\sum_{j\in J} a_j\right)\cdot\left(\sum_{k\in K} b_k\right) = \sum_{j\in J,\, k\in K} a_j b_k = \sum_{(j,k)\in J\times K} a_j b_k \tag{1.12}$$

Dimostrazione. Accettiamo il fatto (che si può dimostrare per induzione sul numero di elementi di K) e verificare nei casi più semplici, es $(a+b)(c+d) = ac + ad + bc + bd$. □

Prodotti di sommatorie aventi medesimo insieme di indici Nel caso gli elementi siano indiciati dal medesimo set, es $J = I_n$, possiamo iniziare a pensare il relativo prodotto cartesiano $J \times J$ della precedente come una matrice quadrata:

$$\begin{aligned}
\left(\sum_{i=1}^{n} a_i\right)\left(\sum_{i=1}^{n} b_i\right) &= (a_1 + a_2 + \ldots + a_n)(b_1 + b_2 + \ldots + b_n) \\
&= a_1b_1 + a_1b_2 + \ldots + a_1b_n + \\
&\quad a_2b_1 + a_2b_2 + \ldots + a_2b_n + \\
&\quad + \ldots + \\
&\quad a_nb_1 + a_nb_2 + \ldots + a_nb_n \\
&= \sum_{i=1}^{n} a_ib_i + \sum_{i\neq j} a_ib_j
\end{aligned}$$

In altre parole abbiamo scomposto la sommatoria in due pezzi; quella degli elementi residenti sulla diagonale principale (primo termine) e i rimanenti (secondo termine).

Quadrato di sommatoria Nel caso particolare di quadrato di sommatoria degli elementi $(a_j)_{j\in J}$, si ha:

$$\left(\sum_{j\in J} a_j\right)^2 = \left(\sum_{j\in J} a_j\right)\cdot\left(\sum_{j\in J} a_k\right) = \sum_{(j,k)\in J\times J} a_j a_k \tag{1.13}$$

Per ritrovare l'usuale espressione del quadrato di una somma spezziamo indici $J\times J$ (e relative sommatorie) nella diagonale $\Delta = \{(j,j) : j\in J\}$ e nel suo complementare $J\times J\setminus\Delta$. Si ha

$$\sum_{(j,k)\in J\times J} a_j a_k = \sum_{(j,k)\in\Delta} a_j a_k + \sum_{(j,k)\in J\times J\setminus\Delta} a_j a_k$$

e dato che essendo $j=k$ per $(j,k)\in\Delta$ possiamo riscrivere il primo termine come

$$\sum_{(j,k)\in\Delta} a_j a_k = \sum_{j\in J} a_j a_j = \sum_{j\in J} a_j^2$$

mentre il secondo termine, che dipende dal set di indici $J\times J\setminus\Delta$, può essere diviso in due parti disgiunte, $S = \{(j,k) : j<k\}$ e $T = \{(j,k) : j>k\}$ (si pensi al triangolo superiore e inferiore della matrice che rappresenta il prodotto cartesiano $J\times J$):

$$\sum_{(j,k)\in J\times J\setminus\Delta} a_j a_k = \sum_{(j,k)\in S} a_j a_k + \sum_{(j,k)\in T} a_j a_k$$

e poiché $(j,k)\to(k,j)$ è una biezione dell'insieme S su T si ha

$$\sum_{(j,k)\in T} a_j a_k = \sum_{(k,j)\in S} a_j a_k = \sum_{(r,s)\in S} a_s a_r$$

dove nell'ultimo passaggio abbiamo effettuato un mero cambio di indici muti. Se ne effettuiamo uno lievemente simile nell'altro termine

$$\sum_{(j,k)\in S} a_j a_k = \sum_{(r,s)\in S} a_r a_s$$

possiamo tornare a

$$\begin{aligned}\sum_{(j,k)\in S} a_j a_k + \sum_{(j,k)\in T} a_j a_k &= \sum_{(r,s)\in S} a_r a_s + \sum_{(r,s)\in S} a_s a_r \\ &= \sum_{(r,s)\in S} (a_r a_s + a_s a_r) \\ &= \sum_{(r,s)\in S} 2a_r a_s\end{aligned}$$

e si conclude con infine a

$$\left(\sum_{j\in J} a_j\right)^2 = \sum_{j\in J} a_j^2 + \sum_{(j,k)\in J\times J : j<k} 2a_j a_k$$

cioè il quadrato di na somma è la somma dei quadrati di tutti i termini, più la somma di tutti i doppi prodotti dei termini stessi.

1.1.2 Sommatorie doppie

1.1.2.1 Definizioni

Osservazione 11. Date più quantità dipendenti da due indici, es:

$$\begin{array}{cccc} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array}$$

la loro somma si può scrivere utilizzando la notazione di sommatoria:

$$\begin{aligned} = \quad & (a_{11} + a_{12} + \dots a_{1n}) + \\ & +(a_{21} + a_{22} + \dots a_{2n}) + \\ & + \dots + \\ & +(a_{m1} + a_{m2} + \dots a_{mn}) \\ = \quad & \sum_{i=1}^{n} a_{1i} + \sum_{i=1}^{n} a_{2i} + \dots + \sum_{i=1}^{n} a_{mi} \end{aligned}$$

Ponendo $\sum_{i=1}^{n} a_{1i} = S_1, \sum_{i=1}^{n} a_{2i} = S_2, \dots, \sum_{i=1}^{n} a_{mi} = S_m$, la somma degli $m \times n$ elementi a diviene

$$S_1 + S_2 + \dots + S_m = \sum_{j=1}^{m} S_j = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{ji}$$

e si legge "sommatoria doppia delle a_{ji} con j che varia da 1 a m ed i che varia da 1 a n", essendo a_{ji} il termine generico che compare nella somma.

Osservazione 12. Si noti il caso particolare $\sum_{j=c}^{c} \sum_{i=k}^{k} a_{ji} = a_{ck}$.

Osservazione 13. Le due sommatorie si possono invertirsi (con l'effetto che prima di sommare una riga e poi passare alla successiva, prima si somma una colonna per passare poi alla susseguente; il quale ovviamente non ha riverbero sui risultati)

Proposizione 1.1.11 (Inversione delle sommatorie)**.**

$$\sum_{j=1}^{m} \sum_{i=1}^{n} a_{ji} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ji}$$

Dimostrazione. Si ha

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji} &= (a_{11}+a_{12}+\ldots a_{1n})+\\ &\quad +(a_{21}+a_{22}+\ldots a_{2n})+\\ &\quad +\ldots+\\ &\quad +(a_{m1}+a_{m2}+\ldots a_{mn})\\ &= (a_{11}+a_{21}+\ldots+a_{m1})+\\ &\quad +(a_{12}+a_{22}+\ldots+a_{m2})+\\ &\quad +\ldots+\\ &\quad +(a_{1n}+a_{2n}+\ldots+a_{mn})=\\ &= \sum_{j=1}^{m} a_{j1}+\sum_{j=1}^{m} a_{j2}+\ldots+\sum_{j=1}^{m} a_{jn}\end{aligned}$$

Ponendo $\sum_{j=1}^{m} a_{j1} = Z_1, \sum_{j=1}^{m} a_{j2} = Z_2, \ldots, \sum_{j=1}^{m} a_{jn} = Z_n$ si ha

$$\sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji} = Z_1 + Z_2 + \ldots + Z_n = \sum_{i=1}^{n} Z_i = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ji}$$

□

Osservazione 14. Anche in questo caso le lettere j e i, indici del termine generico, possono essere sostituite da qualsiasi altre lettere. Talvolta si può trovare $\sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji}$ espresso omettendo gli estremi del campo di variazione della i e della j (se ciò non crea confusione o equivoci), mediante $\sum_j \sum_i a_{ji}$ o anche $\sum_{j,i} a_{ji}$. Talvolta si può trovare la scrittura $\sum \sum a_{ji}$ che è bene evitare perché è sempre meglio indicare gli indici variabili (nel nostro caso j e i) rispetto ai quali si esegue la somma.

1.1.2.2 Proprietà

Proposizione 1.1.12 (Sommatoria di costante)**.** *Se k è una costante che non dipende dagli indici j e i:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n} k = kmn \tag{1.14}$$

Dimostrazione. Infatti è una sommatoria doppia in cui il termine generico $a_{ji} = k$:

$$\sum_{j=1}^{m}\sum_{i=1}^{n} k = \sum_{j=1}^{m} kn = n\sum_{j=1}^{m} k = kmn$$

□

Proposizione 1.1.13 (Sommatoria di prodotto per costante)**.** *Se k è una costante che non dipende dagli indici j e i:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n} k a_{ji} = k\sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji} \tag{1.15}$$

Dimostrazione. Infatti

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n} ka_{ji} &= ka_{11} + ka_{12} + \ldots + ka_{1n} + \\ &\quad + ka_{21} + ka_{22} + \ldots + ka_{2n} \\ &\quad + \ldots + \\ &\quad + ka_{m1} + ka_{m2} + \ldots + ka_{mn} \\ &= \sum_{i=1}^{n} ka_{1i} + \sum_{i=1}^{n} ka_{2i} + \ldots + \sum_{i=1}^{n} ka_{mi} = \\ &= k\sum_{i=1}^{n} a_{1i} + k\sum_{i=1}^{n} a_{2i} + \ldots + k\sum_{i=1}^{n} a_{mi} = \\ &= k(\sum_{i=1}^{n} a_{1i} + \sum_{i=1}^{n} a_{2i} + \ldots + \sum_{i=1}^{n} a_{mi}) = \\ &= k\sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji}\end{aligned}$$

□

Proposizione 1.1.14 (Scomposizione/somme su sottoinsiemi)**.** *Si ha che:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n_1} a_{ji} + \sum_{j=1}^{m}\sum_{i=n_1+1}^{n} a_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji} \tag{1.16}$$

Dimostrazione. Infatti

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n_1} a_{ji} + \sum_{j=1}^{m}\sum_{i=n_1+1}^{n} a_{ji} &= \sum_{j=1}^{m}\left(\sum_{i=1}^{n_1} a_{ji} + \sum_{i=n_1+1}^{n} a_{ji}\right) = \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji} \\ \sum_{j=1}^{m_1}\sum_{i=1}^{n} a_{ji} + \sum_{j=m_1+1}^{m}\sum_{i=1}^{n} a_{ji} &= \sum_{i=1}^{n}\sum_{j=1}^{m_1} a_{ji} + \sum_{i=1}^{n}\sum_{j=m_1+1}^{m} a_{ji} = \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji}\end{aligned}$$

□

Osservazione 15. Per visualizzare le operazioni di cui sopra si pensi ad una somma degli elementi di una matrice che procede attraverso le colonne (sommatoria interna) e poi passa alla prossima riga (ciclo sulla sommatoria esterna); nel primo caso qui sopra abbiamo aggiunto delle colonne ad una matrice, mentre nel secondo abbiamo aggiunto delle righe ad un'altra matrice.

Proposizione 1.1.15 (Sommatoria di somme)**.** *Vale la:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(a_{ji} + b_{ji}) = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ji} + \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji} \tag{1.17}$$

Dimostrazione. Infatti

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n}(a_{ji}+b_{ji}) &= \sum_{j=1}^{m}\left[\sum_{i=1}^{n}(a_{ji}+b_{ji})\right] \\ &= \sum_{j=1}^{m}\left[\sum_{i=1}^{n}a_{ji}+\sum_{i=1}^{n}b_{ji}\right] \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n}a_{ji}+\sum_{j=1}^{m}\sum_{i=1}^{n}b_{ji}\end{aligned}$$

□

Proposizione 1.1.16 (Sommatoria di termini lineari)**.** *Se k e c sono costanti che non dipendono dagli indici j e i, vale:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n}(ka_{ji}+c) = mnc + k\sum_{j=1}^{m}\sum_{i=1}^{n}a_{ji} \tag{1.18}$$

Dimostrazione. Infatti:

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n}(ka_{ji}+c) &= \sum_{j=1}^{m}\sum_{i=1}^{n}ka_{ji}+\sum_{j=1}^{m}\sum_{i=1}^{n}c \\ &= k\sum_{j=1}^{m}\sum_{i=1}^{n}a_{ji}+c\sum_{j=1}^{m}\sum_{i=1}^{n}1 \\ &= cmn + k\sum_{j=1}^{m}\sum_{i=1}^{n}a_{ji}\end{aligned}$$

□

Proposizione 1.1.17 (Portar fuori sommatoria)**.** *È lecito estrarre da ogni sommatoria i termini che* non *dipendono dall'indice della sommatoria:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n}a_j b_i = \sum_{j=1}^{m}a_j\sum_{i=1}^{n}b_i \tag{1.19}$$

Cioè dalla seconda sommatoria, fatta secondo l'indice i, si può estrarre il termine a_j che da i non dipende.

Dimostrazione. Infatti

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n} a_j b_i =& a_1b_1 + a_1b_2 + \ldots + a_1b_n + \\ & a_2b_1 + a_2b_2 + \ldots + a_2b_n + \\ & \vdots \\ & a_nb_1 + a_nb_2 + \ldots + a_nb_n = \\ =& a_1\sum_{i=1}^{n} b_i + a_2\sum_{i=1}^{n} b_i + \ldots + a_m\sum_{i=1}^{n} b_i \\ =& (a_1 + a_2 + \ldots + a_m)\sum_{i=1}^{n} b_i \\ =& \sum_{j=1}^{m} a_j \sum_{i=1}^{n} b_i\end{aligned}$$

□

Lemma 1.1.18. *Da ciò deriva ad esempio che si può scrivere*

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{j=1}^{n}\sum_{i=1}^{n} a_i a_j$$

Dimostrazione. Infatti

$$\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n} a_i \cdot \sum_{i=1}^{n} a_i = \sum_{j=1}^{n} a_j \cdot \sum_{i=1}^{n} a_i = \sum_{j=1}^{n}\sum_{i=1}^{n} a_j a_i$$

dove abbiamo posto j al posto di i in una delle due sommatorie per evitare confusioni. □

Lemma 1.1.19. *È lecito anche scrivere:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n} a_j = \sum_{j=1}^{m} a_j \sum_{i=1}^{n} 1 = n\sum_{j=1}^{m} a_j$$
$$\sum_{j=1}^{m}\sum_{i=1}^{n} b_i = \sum_{i=1}^{n} b_i \sum_{j=1}^{m} 1 = m\sum_{i=1}^{n} b_i$$

Lemma 1.1.20. *È corretto effettuare la seguente posizione:*

$$\sum_{j=1}^{m}\sum_{i=1}^{n} a_j b_{ji} = \sum_{j=1}^{m} a_j \sum_{i=1}^{n} b_{ji}$$

cioè estrarre a_j dalla seconda sommatoria da cui non dipende, perché quest'ultima è fatta rispetto all'indice i.

Osservazione importante 4. Si osservi che è scorretto scrivere

$$\sum_{i=1}^{n} b_{ji} \sum_{j=1}^{m} a_j$$

cioè non è possibile estrarre b_{ji} da alcuna sommatoria perché dipende da entrambi gli indici e quindi da entrambe le sommatorie.

1.2 Produttorie

1.2.1 Produttoria singola

Definizione 1.2.1 (Produttoria)**.** Se $(a_j)_{j\in J}$, $a : J \to \mathbb{C}$ è una famiglia *finita*, il prodotto di tutti i numeri a_j per $j \in J$ si indica con:

$$\prod_{j\in J} a_j \tag{1.20}$$

Osservazione 16. Si pone per convenzione

$$\prod_{j\in\emptyset} a_j = 1 \tag{1.21}$$

1.2.1.1 Proprietà

Osservazione 17. Analogamente al caso delle sommatorie valgono le seguenti *proprietà* (che possono essere utili lette sia da sinistra a destra che viceversa).

Proposizione 1.2.1 (Produttoria di costante)**.** *Se k è una costante che non dipende dall'indice i:*

$$\prod_{i=1}^{n} k = k^n \tag{1.22}$$

Dimostrazione. Infatti è una produttoria in cui il termine generico $a_i = k$

$$\prod_{i=1}^{n} a_i = a_1 a_2 \dots a_n = k \cdot k \cdot \dots k \cdot = k^n$$

□

Proposizione 1.2.2 (Produttoria di prodotto per costante)**.** *Se k è una costante che non dipende dall'indice i:*

$$\prod_{i=1}^{n} k a_i = k^n \prod_{i=1}^{n} a_i \tag{1.23}$$

Dimostrazione. Infatti

$$\prod_{i=1}^{n} k a_i = k a_1 \cdot k a_2 \cdot \dots \cdot k a_n = k^n (a_1 a_2 \cdots a_n) = k^n \prod_{i=1}^{n} a_i$$

□

Scomposizione in sottoinsiemi

Proposizione 1.2.3. *Vale la seguente:*

$$\prod_{i=1}^{m} a_i \prod_{i=m+1}^{n} a_i = \prod_{i=1}^{n} a_i \tag{1.24}$$

Dimostrazione. Infatti

$$\prod_{i=1}^{m} a_i \prod_{i=m+1}^{n} a_i = (a_1 a_2 \dots a_m)(a_{m+1} a_{m+2} \dots a_n) = \prod_{i=1}^{n} a_i$$

□

Osservazione 18. Generalizzando, se Λ è un insieme di indici, $a : \Lambda \to \mathbb{C}$ una famiglia di complessi, e J, K sottoinsiemi finiti *disgiunti* di Λ si ha:

$$\prod_{\lambda \in J \cup K} a_\lambda = \prod_{\lambda \in J} a_\lambda \cdot \prod_{\lambda \in K} a_\lambda \tag{1.25}$$

Proposizione 1.2.4 (Scomposizione: produttoria di prodotti)**.** *Vale la seguente:*

$$\prod_{i=1}^{n} a_i b_i = \prod_{i=1}^{n} a_i \prod_{i=1}^{n} b_i \tag{1.26}$$

Dimostrazione.

$$\begin{aligned}\prod_{i=1}^{n} a_i b_i =& a_1 b_1 \cdot a_2 b_2 \cdot \dots a_n b_n = \\ =& (a_1 a_2 \dots a_n)(b_1 b_2 \dots b_n) \\ =& \prod_{i=1}^{n} a_i \prod_{i=1}^{n} b_i\end{aligned}$$

□

Osservazione 19. Generalizzando, se Λ è un insieme di indici, $a : \Lambda \to \mathbb{C}$ una famiglia di complessi e $b : \Lambda \to \mathbb{C}$ è un'altra famiglia di complessi si può definire il prodotto $a \cdot b : \Lambda \to \mathbb{C}$ delle due famiglie ponendo $(a \cdot b)(\lambda) = a_\lambda \cdot b_\lambda$ per ogni $\lambda \in \Lambda$. Si ha anche che per ogni sottoinsieme finito J di Λ:

$$\prod_{j \in J} (a_j \cdot b_j) = \prod_{j \in J} a_j \cdot \prod_{j \in J} b_j \tag{1.27}$$

Proposizione 1.2.5 (Logaritmi e sommatorie)**.** *Vale la*

$$\log \prod_{i=1}^{n} a_i = \sum_{i=1}^{n} \log a_i \tag{1.28}$$

Dimostrazione. Infatti:

$$\log \prod_{i=1}^{n} a_i = \log(a_i a_2 \dots a_n) = \log a_i + \log a_2 + \dots + \log a_n = \sum_{i=1}^{n} \log a_i$$

□

1.3 Esercizi

Esercizio 1.3.1 (es 10 pag 34 bps1)**.** Ricavare la formula per la somma dei primi n numeri pari

$$\sum_{k=1}^{n}(2k)$$

e dimostrarla per induzione
Soluzione. Elaboriamola intanto

$$\sum_{k=1}^{n} 2k = 2\sum_{k=1}^{n} k = 2\frac{n(n+1)}{2} = n(n+1)$$

Dimostriamo per induzione (anche se non ci sarebbe bisogno, essendo che è moltiplicare per 2 entrambi i membri dell'equazione $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$:

- per il passo base

  $$\sum_{k=1}^{1} 2k = 2$$
  $$1(1+1) = 2$$

  sono uguali quindi il passo base è ok

- per il passo induttivo

  $$\sum_{k=1}^{n+1} 2k = \left(\sum_{k=1}^{n} 2k\right) + n(n+1) = (n+1)(n+2)$$

  Quest'ultima è proprio $n(n+1)$ con sostituzione $n \to n+1$, quindi anche il passo induttivo è ok

Capitolo 2

Calcolo combinatorio

2.1 Introduzione

Definizione 2.1.1 (Calcolo combinatorio)**.** Studio di come quantificare raggruppamenti aventi determinate caratteristiche degli elementi di un insieme finito di oggetti.

Osservazione 20. È fondamentale per il calcolo delle probabilità in quanto spesso la probabilità di un evento è calcolabile come il numero di modi in cui detto evento può verificarsi in rapporto al numero di casi possibili.

Definizione 2.1.2 (Principio fondamentale del calcolo combinatorio)**.** Se si realizzano due esperimenti:

- in cui il primo ha m esiti possibili;
- e per ognuno di questi il secondo ha n esiti possibili;
- e l'ordinamento conta per qualificare un esito (ossia sequenze diverse dei singoli esiti dei due esperimenti producono esiti finali distinti):

allora i due esperimenti (considerati congiuntamente) hanno $m \cdot n$ esiti possibili.

Osservazione 21. Generalizzato, con r esperimenti nel quale il primo abbia n_1 esiti possibili, per ciascuno di questi il secondo ne abbia n_2 ...per ogni esito dei primi due $r-1$ l'r-esimo n_r esiti possibili e l'ordinamento conta, allora gli esperimenti hanno in tutto $\prod_{i=1}^{r} n_i$ esiti possibili.

Definizione 2.1.3 (Funzione fattoriale)**.** Il fattoriale di n, indicato con $n!$ è una funzione $f : \mathbb{N} \to \mathbb{N}$ è definito come il prodotto dei primi n numeri interi:

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 \tag{2.1}$$

Si conviene che $0! = 1$.

Osservazione 22 (Definizione ricorsiva). Dato che $(n-1)\cdot(n-2)\cdot\ldots\cdot 1 = (n-1)!$ il fattoriale può esser definito anche come:

$$n! = \begin{cases} 1 & n \in \mathbb{N}, n = 0 \\ n \cdot (n-1)! & n \in \mathbb{N}, n \neq 0 \end{cases} \tag{2.2}$$

Osservazione 23 (Una semplificazione utile). Se $0 < k < n$, si ha:

$$\frac{n!}{(n-k)!} = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) \tag{2.3}$$

2.2 Casistica principale

Supponendo di voler costruire sottoinsiemi contenenti k elementi scelti tra gli n elementi di un insieme U:

- nel caso in cui l'*ordine* abbia importanza (configurazioni con gli stessi elementi posti in ordine diverso danno origine ad esiti diversi) abbiamo a che fare con:
  - **permutazioni**: disponiamo di $k = n$ slot ed n elementi ($\in U$) da utilizzare per riempirli. Ci interessa sapere in quanti modi si possono ordinare gli n oggetti: ognuno di questi ordinamenti si chiama *permutazione*. Possiamo avere due casi:
    1. permutazioni *semplici*: gli n elementi da ordinare sono unici (ad esempio gli anagrammi della parola "AMORE");
    2. permutazioni *con ripetizione*: ammettono che un elemento si presenti più volte tra gli n dai quali si può pescare (ad esempio gli anagrammi della parola "PEPPER").
  - **disposizioni** (che costituiscono una versione generalizzata della permutazioni): gli slot sono in numero $k \leq n$ inferiore (o uguale) rispetto agli elementi n con il quale possiamo riempirli. Di fatto qua si considera che gli n elementi siano tutti distinti/diversi. Abbiamo:
    1. disposizioni *semplici*: i k elementi sono pescati da un insieme di n elementi distinti e una volta che l'elemento è stato scelto esce dal pool degli utilizzabili;
    2. disposizioni *con ripetizione*: ciascun elemento dei n può essere estratto più volte
- se viceversa l'*ordine non ha rilevanza*, ossia sottoinsiemi composti da medesimi elementi posti in ordine differente sono considerati uguali (ad esempio quando si vogliono contare insiemi nell'accezione matematica del termine) si ha a che fare con le **combinazioni**. Le combinazioni semplici sono le più utilizzate e si hanno quando il pool dal quale si pesca è composto da oggetti diversi/distinti tra loro.

2.2.1 Permutazioni

Proposizione 2.2.1 (Permutazioni semplici)**.** *Il numero di permutazioni di n elementi distinti in n slot è:*

$$\boxed{P_n = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1 = n!} \tag{2.4}$$

Dimostrazione. Nella prima posizione possiamo porre n alternative, nella seconda $n-1$ (visto che una è già andata nella prima), e così via; arrivando così all'ultima posizione rimane un solo oggetto possibile degli n iniziali. Pertanto per il principio fondamentale del calcolo combinatorio si conclude. □

Osservazione 24. Nel caso in cui vi siano elementi ripetuti/uguali dai quali pescare (ad esempio se vogliamo permutare le lettere di "PEPPER") vogliamo che il numero di esiti complessivi diminuisca (evitando di contare come differenti due configurazioni con elementi uguali permutati tra loro)

Proposizione 2.2.2 (Permutazioni con ripetizione)**.** *Tra gli n dai quali pescare vi siano $i = 1, 2 \dots r$ elementi univoci che si possono ripetere, aventi numerosità rispettivamente $k_1, k_2 \dots k_r$ (ossia si ha $\sum_{i=1}^{r} i \cdot k_i = n$). Le permutazioni uniche (non ripetute) sono:*

$$\boxed{P_n^{k_1,k_2,\dots,k_r} = \frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_r!}} \tag{2.5}$$

Dimostrazione. Si parte dal numero di permutazioni degli n oggetti al numeratore Applicando il principio fondamentale del calcolo combinatorio al contrario, si tratta di dividere queste per il numero delle $k_1!$ permutazioni uguali fra loro (dovute al "girare" di uno stesso elemento), poi per le $k_2!$ permutazioni del secondo elemento multiplo, e così via. □

Esempio 2.2.1. Considerando le permutazioni PEPPER ad ogni sequenza univoca (ad esempio REPPEP) corrisponderanno 3!2! sequenze che sono di fatto uguali. Pertanto il numero di permutazioni univoche (con ripetizione) di PEPPER saranno $6!/(3! \cdot 2!)$.

Osservazione 25. La formula delle permutazioni è una generalizzazione e vale in realtà per qualsiasi permutazione, anche senza ripetizioni di elementi. Infatti, se abbiamo elementi univoci, ossia $k_1 = k_2 = \dots = k_r = 1$, otteniamo esattamente la formula delle permutazioni semplici in quanto:

$$P_n^{k_1,k_2,\dots,k_r} = \frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_r!} = \frac{n!}{1! \cdot 1! \cdot \dots \cdot 1!} = n! \tag{2.6}$$

2.2.2 Disposizioni

Definizione 2.2.1 (Disposizioni semplici)**.** Se il numero degli slot disponibili è inferiore (o uguale) al numero di elementi dai quali si pesca, gli elementi dai quali si pesca sono distinti tra loro e non vengono reinseriti nel pool dove pescare si hanno le disposizioni semplici.
Sono quello che in statistica si chiama *campionamento senza ripetizione.*

Proposizione 2.2.3 (Numero di disposizioni semplici)**.** *Il numero $D_{n,k}$ di disposizioni semplici di $k \leq n$ oggetti estratti da un insieme di n oggetti differenti è:*

$$\boxed{D_{n,k} = n \cdot (n-1) \cdot \dots \cdot (n-k+1) = \frac{n!}{(n-k)!}} \tag{2.7}$$

Dimostrazione. Il primo componente di una tale sequenza può essere scelto in n modi diversi, il secondo in $(n-1)$ e così via, sino al k-esimo che può essere scelto in $(n-k+1)$ modi diversi. □

Osservazione 26. Le permutazioni semplici (quando $k = n$) sono casi particolari delle disposizioni semplici (quando $k \leq n$):

$$P_n = D_{n,n} = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n! \tag{2.8}$$

Definizione 2.2.2 (Disposizioni con ripetizione)**.** Le disposizioni con ripetizione sono caratterizzate dal fatto che ciascuno degli n elementi possa essere estratto più volte per riempire i k slot.
Sono quello che in statistica si chiama *campionamento con ripetizione.*

Proposizione 2.2.4 (Numero di disposizioni con ripetizione)**.** *Il numero di disposizioni con ripetizione di n elementi in k slot:*

$$\boxed{D'_{n,k} = \underbrace{n \cdot n \cdot \ldots \cdot n}_{\text{k } volte} = n^k} \tag{2.9}$$

Dimostrazione. Si hanno n possibilità per scegliere il primo componente, n per il secondo, altrettante per il terzo e così via, sino al k-esimo; si conclude per il principio fondamentale del calcolo combinatorio. □

2.2.3 Combinazioni

2.2.3.1 Combinazioni semplici

Osservazione 27. Gli n elementi dai quali si pesca sono univoci: si pescano k elementi, l'ordine/disposizione di questi non è rilevante a qualificare un esito differente. Si hanno le combinazioni semplici che conteggiano il numero di sottoinsiemi di ampiezza definita di un determinato insieme base.

Proposizione 2.2.5 (Combinazioni semplici)**.** *Il numero delle combinazioni semplici di n elementi di lunghezza k, indicato con* $C_{n,k}$ *è:*

$$\boxed{C_{n,k} = \frac{D_{n,k}}{P_k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k! \cdot (n-k)!} = \binom{n}{k}} \tag{2.10}$$

Dimostrazione. Analogamente alle disposizioni semplici sceglieremo k elementi da n: si inizierà avendo n possibilità per il primo, sino a $n-k+1$ per il k-esimo. Tuttavia all'interno dei gruppi così determinati ci saranno combinazioni che sono formate dagli stessi elementi di altre, anche se in ordine inverso. Per non contare tali gruppi più volte (dato che l'ordine non interessa), sempre applicando il principio fondamentale del calcolo combinatorio, occorrerà dividere le disposizioni per il numero di permutazioni dei k elementi estratti ($k!$). □

2.2.3.2 Combinazioni con ripetizione

Osservazione 28. Nelle combinazioni semplici non è ammesso pescare lo stesso elemento più volte. Una volta estratto non rimane negli oggetti estraibili.
Nelle combinazioni con ripetizione invece vogliamo determinare quanti modi vi sono di scegliere k volte da un insieme di n oggetti diversi tra loro, ammettendo che però uno stesso oggetto possa essere pescato più volte.
L'ordine continua a non essere importante (ci interessa sono quante volte ogni oggetto è stato scelto, non l'ordine con cui esso appare).
Le combinazioni con ripetizione contano i *multiset* (insiemi che ammettono ripetizioni) sottoinsieme di un insieme dato.

Proposizione 2.2.6. *Il numero di combinazioni con ripetizione di k oggetti scelti tra n è*

$$\boxed{C^*_{n,k} = \binom{n+k-1}{k}} \tag{2.11}$$

Dimostrazione. Se l'ordine contasse il numero di combinazioni sarebbe n^k, ma questo non è il caso. Per dimostrare la formula risolviamo narrativamente un problema isomorfico (stesso problema con setup differente).
Il problema può essere posto come: porre k palline identiche in n scatole differenti: quello che conta è solamente il numero di palline in ciascuna scatola. Una qualsiasi configurazione può essere rappresentata come una sequenza di `|` per rappresentare i lati di una scatola e `o` per rappresentare le palline in essa. Ad esempio ipotizzando di avere $k = 7$ palline e $n = 4$ scatole, per rappresentare una pallina nella prima scatola, due nella seconda, tre nella terza e una nella quarta:

`|o|oo|ooo|o|`

Per essere valida ciascuna sequenza deve iniziare e finire con `|`: pertanto si tratta solo di contare il modo in cui si possono riarrangiare i termini rimanenti al suo interno (varie configurazioni di scatole). I termini all'interno dei bordi numero $n + k - 1$: di questi k (le palline) ed $((n - k - 1) - k) = n - 1$ anche (i bordi rimanenti utili per formare le n scatole, una volta che due sono stati impiegati per i lati). La soluzione è pertanto

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

□

2.3 Coefficiente binomiale e multinomiale

2.3.1 Coefficiente binomiale

2.3.1.1 Definizione

Osservazione 29. Approfondiamo il coefficiente che risulta dal calcolo del numero di combinazioni semplici di k elementi presi da n.

Definizione 2.3.1 (Coefficiente binomiale)**.** Indicato con $\binom{n}{k}$ e pronunciato "n su k" si definisce come

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

se $k \leq n$. Se $n < k$ si pone $\binom{n}{k} = 0$.

Osservazione 30. Per quanto riguarda il calcolo a mano, spesso è più utile/veloce la prima definizione, mentre la seconda è più compatta ed utilizzabile nelle parti teoriche.

2.3.1.2 Proprietà

Proposizione 2.3.1. *Si ha che:*

$$\boxed{\binom{n}{k} = \binom{n}{n-k}} \tag{2.12}$$

Dimostrazione.

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

□

Osservazione 31. Una intuizione sul significato di 2.12: per scegliere un comitato di k persone tra n sappiamo che ci sono $\binom{n}{k}$ modi. Un'altro modo di scegliere il comitato è specificare quali $n-k$ non ne faranno parte; specificare chi è nel comitato determina chi non vi è e viceversa. Pertanto i due lati sono uguali dato che sono due modi di contare la stessa cosa.

Osservazione 32. Esempi notevoli/utili della 2.12 sono:

$$\binom{n}{0} = \binom{n}{n} = 1, \qquad \binom{n}{1} = \binom{n}{n-1} = n \tag{2.13}$$

Proposizione 2.3.2.

$$\boxed{\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}} \tag{2.14}$$

Dimostrazione.

$$\begin{aligned}
\binom{n-1}{k-1} + \binom{n-1}{k} =& \frac{(n-1)!}{(k-1)! \cdot (n-k)!} + \frac{(n-1)!}{k! \cdot (n-k-1)!} \\
=& \frac{(n-1)! \cdot k}{k! \cdot (n-k)!} + \frac{(n-1)! \cdot (n-k)}{k! \cdot (n-k)!} \\
=& \frac{(n-1)! \cdot n}{k! \cdot (n-k)!} \\
=& \binom{n}{k}
\end{aligned}$$

□

Osservazione 33. Per il significato di 2.14: se ho un insieme di n oggetti $I_n = \{1, \ldots, n\}$ isolando un oggetto (diciamo l'n-esimo) posso dividere i sottoinsiemi di I_n che hanno k oggetti in quelli che non contengono l'n-esimo (che sono $\binom{n-1}{k}$, essendo esattamente i sottoinsiemi di I_{n-1} a p oggetti) ed in quelli che lo contengono, i quali si ottengono aggiungendo n ad un insieme di $p-1$ oggetti di I_{n-1} e quindi sono in numero di $\binom{n-1}{p-1}$[1]; questi due gruppi di sottoinsimi di I_n sono evidentemente disgiunte, quindi l'unione ha la somma come cardinale, e quindi si ha la formula.

[1] Sarebbero $\binom{n-1}{p-1} \cdot 1$ poiché vi è un solo modo di aggiungere l'n-esimo ad un insimeme di $p-1$ elementi già formati (scelti tra $n-1$ elementi disponibili)

Proposizione 2.3.3 (Identità di Vandermonde)**.**

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-k} \tag{2.15}$$

Dimostrazione. La prova mediante espansione dei termini e forza bruta ce la si può evitare. Una dimostrazione narrativa sul perché l'uguaglianza valga è comunque efficace.
Considerando un gruppo di m uomini ed n donne dal quale un comitato di k persone verrà scelto: ci sono $\binom{m+n}{k}$ per farlo. Se vi sono j uomini nel comitato, allora vi debbono essere $k-j$ donne. Il lato destro dell'uguaglianza smma per il numero j di uomini. □

Proposizione 2.3.4 (Squadra con capitano)**.** *Per $k, n \in N$ con $k \leq n$ si ha*

$$n\binom{n-1}{k-1} = k\binom{n}{k} \tag{2.16}$$

Dimostrazione. Una dimostrazione narativa: consideriamo un gruppo di n persone dal quale una squadra di k verrà scelta; uno di queste sarà capitano. Il numero possibile di team così formati può derivare da (lato sinistro) prima scegliere il capitano tra gli n e poi scegliere i $k-1$ rimanenti tra gli $n-1$ disponibili. Oppure ed equivalentemente scegliendo gli $\binom{n}{k}$ componenti e tra questi sceglierne uno dei k come capitano. □

2.3.1.3 Origine del nome

Osservazione 34. Il coefficiente binomiale prende nome dal fatto che determina i coefficienti dello sviluppo della potenza del binomio $(x+y)^n$

Proposizione 2.3.5 (Teorema binomiale)**.**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{2.17}$$

Dimostrazione. Per provare il teorema espandiamo il prodotto:

$$(x+y)^n = \underbrace{(x+y)\cdot(x+y)\cdot\ldots\cdot(x+y)}_{n \text{ fattori}}$$

I termini del prodotto $(x+y)^n$ sono ottenuti scegliendo la x o la y da ognuno dei fattori. Vi sono $\binom{n}{k}$ modi per scegliere esattamente k volte x (scegliendo y nei $n-k$ rimanenti): in questi casi si ottiene il termine $x^k y^{n-k}$. Il teorema si ottiene facendo variare il numero k di x scelti e sommando i termini risultati. □

2.3.2 Il coefficiente multinomiale

2.3.2.1 Definizione

Proposizione 2.3.6. *Il numero di modi in cui è possibile distribuire n oggetti distinti in r scatole distinte in modo che queste contengano, nell'ordine, $n_1, n_2, \ldots, n_r$ oggetti ($\sum_{i=1}^{r} n_i = n$) è:*

$$\boxed{\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_r!}} \tag{2.18}$$

Dimostrazione. Vi sono $\binom{n}{n_1}$ possibili scelte per gli oggetti della prima scatola; per ogni tale scelta vi sono $\binom{n-n_1}{n_2}$ scelte per la seconda; per ogni scelta effettuata nelle prime due vi sono $\binom{n-n_1-n_2}{n_3}$ nella terza e così via. Dal principio fondamentale del calcolo combinatorio discende che il risultato cercato è:

$$\binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdot \ldots \cdot \binom{n-n_1-\ldots-n_{r-1}}{n_r} \tag{2.19}$$

Sviluppando si ha

$$\frac{n!}{(n-n_1)!n_1!} \cdot \frac{(n-n_1)!}{(n-n_1-n_2)!n_2!} \cdot \ldots \cdot \frac{(n-n_1-n_2-\ldots-n_{r-1})!}{0!n_r!}$$

dalla quale, in seguito alle semplificazioni, si ottiene il coefficiente. □

Osservazione 35. Costituisce una generalizzazione del coefficiente binomiale (che si ottiene considerando due scatole).

Osservazione 36. Il coefficiente multinomiale è la formula che viene utilizzato nelle permutazioni con ripetizione (utile ad esempio per il numero di permutazioni di una parola con lettere ripetute).

2.3.2.2 Origine del nome

Osservazione 37. La formula del coefficiente multinomiale determina i coefficienti dello sviluppo di un polinomio di r termini

Proposizione 2.3.7 (Teorema multinomiale)**.**

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{\substack{(n_1,n_2,\ldots,n_r):\\ n_1+n_2+\cdots+n_r=n}} \binom{n}{n_1, n_2, \ldots, n_r} x_1^{n1} \cdot x_2^{n_2} \cdot \ldots \cdot x_r^{n_r}$$

Dimostrazione. Analoga al caso binomiale. □

Esempio 2.3.1. Nello sviluppo del cubo di un trinomio potremmo procedere manualmente:

$$(a+b+c)^3 = a^3 + b^3 + c^3 + 3a^2b + 3a^2c + 3b^2a + 3b^2c + 3c^2a + 3c^2b + 6abc$$

o calcolare più velocemente, ad esempio che:

- il termine $a^2b^0c^1$ presenta come coefficiente:

$$\binom{3}{2,0,1} = \frac{3!}{2! \cdot 0! \cdot 1!} = \frac{6}{2 \cdot 1 \cdot 1} = 3$$

- il termine $a^1b^1c^1$ ha invece coefficiente pari a:

$$\binom{3}{1,1,1} = \frac{3!}{1! \cdot 1! \cdot 1!} = \frac{6}{1 \cdot 1 \cdot 1} = 6$$

2.4 Calcolo combinatorio e funzioni

Il calcolo combinatorio può essere applicato per contare le funzioni aventi determinate caratteristiche tra due insiemi finiti. Vediamo innanzitutto un criterio utile per contare e poi alcune applicazioni al conteggio delle funzioni.

2.4.1 Principio dell'overcounting

Sia $f : X \to Y$ suriettiva; si ha allora

$$\mathrm{Card}\,(X) = \sum_{y \in Y} \mathrm{Card}\,(f^{\leftarrow}(\{y\})) \tag{2.20}$$

In particolare se tutte le fibre $f^{\leftarrow}(\{y\})$ hanno una stessa cardinalità, ossia $\mathrm{Card}\,(f^{\leftarrow}(\{y\})) = \alpha$, si ha:

$$\mathrm{Card}\,(X) = \alpha\,\mathrm{Card}\,(Y) \tag{2.21}$$

essendo X una unione disgiunta delle fibre $f^{\leftarrow}(\{y\})$ al variare di $y \in Y$.
Anche detto principio del pastore, questo torna utile quando conosciamo la cardinalità di uno dei due insiemi (ad esempio pecore) e desideriamo ricavare quella dell'altro (numero di zampe).

2.4.2 Funzioni (disposizioni con ripetizione)

Si indica con X^{I_p} l'insieme di tutte le funzioni $f : I_p \to X$ con $\mathrm{Card}\,(I_p) = p$ e $\mathrm{Card}\,(X) = m$. Il numero di tutte le funzioni possibili tra i due insiemi è

$$\mathrm{Card}\left(X^{I_p}\right) = \underbrace{m \cdot m \cdot \ldots \cdot m}_{p \text{ volte}} = m^p \tag{2.22}$$

e corrisponde alle disposizioni con ripetizione, a p a p degli m oggetti di X.

2.4.3 Funzioni iniettive (disposizioni semplici)

Siamo interessati a quantificare la cardinalità del sottoinsieme delle funzioni inettive $\Lambda(n,p) \subset I_n^{I_p}$ del tipo $f : I_p \to I_n$. Si ha che

$$\mathrm{Card}\,(\Lambda(n,p)) = n \cdot (n-1) \cdot \ldots \cdot (n-(p-1)) \tag{2.23}$$

vedendo che all'ultimo elemento di I_p ho già fatto $p-1$ collegamenti, quindi me ne rimangono possibili $n-(p-1)$.

2.4.4 Permutazioni di un insieme (permutazioni semplici)

In particolare se $p = n$si hanno le biiezioni di un insieme I_n in se stesso, ossia le permutazioni dell'insieme, che sono in numero $n!$

2.4.5 Funzioni caratteristiche (coefficiente binomiale)

Il calcolo del numero di sottoinsiemi a p elementi di un insieme di n oggetti equivale a quantificare la cardinalità delle funzioni caratteristiche che scelgono p elementi tra un insieme di n (ossia tali che $\sum \chi(I_n) = p$).
Indicando con $C(n,p)$ l'insieme dei sottoinsiemi di I_n che hanno cardinale p si ha una funzione suriettiva:

$$s : \Lambda(n,p) \to C(n,p) \tag{2.24}$$

Il dominio $\Lambda(n,p)$ è un insieme di funzioni mentre il codominio $C(n,p)$ è un insieme di insiemi: la funzione suriettiva è quella che ad ogni funzione iniettiva

$f : I_p \to I_n$ (con $f \in \Lambda(n,p)$) associa l'immagine $f(I_p) \in C(n,p)$, sottoinsieme a p oggetti di I_n.
Essendo che due funzioni iniettive facenti parte del dominio $f, g \in \Lambda(n,p)$ hanno la stessa immagine se e solo se differiscono per una permutazione sul proprio dominio, le fibre di s hanno tutte cardinale $p!$ (ossia ciascun insieme di p elementi si presenta in $p!$ ordini possibili), segue dal principio dell'overcounting che $\text{Card}\,(\Lambda(n,p)) = p!\,\text{Card}\,(C(n,p))$, quindi:

$$\text{Card}\,(C(n,p)) = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-(p-1))}{p!} = \binom{n}{p} = \frac{n!}{p!(n-p)!} \tag{2.25}$$

2.5 Esercizi

Esercizio 2.5.1 (Es 3.4 pg 49 de marco)**.** Sia $p \geq 0$ naturale fissato. Mostrare che per ogni naturale $n \geq p$ si ha

$$\sum_{p \leq k \leq n} \binom{k}{p} = \binom{n+1}{p+1}$$

Soluzione. Si ha che

- se $n = p$ l'eguaglianza è verificata in quanto

$$\sum_{p=k=n=a} \binom{a}{a} = \binom{a+1}{a+1} = 1$$

- supponendo sia vera per $n \geq p$ si ha che

$$\begin{aligned}
\sum_{p \leq k \leq n+1} \binom{k}{p} = \sum_{p \leq k \leq n} \binom{k}{p} + \binom{n+1}{p} &= \frac{(n+1)!}{(p+1)!(n-p)!} + \frac{(n+1)!}{p!(n-p+1)!} \\
&= \frac{(n+1)!}{(p+1)p!(n-p)!} + \frac{(n+1)!}{p!(n-p+1)(n-p)!} \\
&= \frac{(n-p+1)(n+1)! + (p+1)(n+1)!}{(p+1)p!(n-p)!(n-p+1)} = \ldots \\
&= \frac{n+2}{(p+1)!(n-p+1)} \\
&= \binom{n+2}{p+1} = \binom{(n+1)+1}{p+1}
\end{aligned}$$

  Dove avviene la sostituzione $n \to n+1$

Esercizio 2.5.2 (Es 1.26.1 pag 33 giusti1)**.** Dimostrare per induzione che $n^n \geq n!$.
Soluzione. Per l'induzione:

- se $n = 1$ si ha che $1 \geq 1$
- ipotizzando che valga per il generico $n \geq 1$, moltiplico per $n+1 > 0$ entrambi i membri ottenendo

$$n^n(n+1) \geq n!(n+1) = (n+1)!$$

ora notiamo che $n^n \cdot (n+1)$ sono $n+1$ termnini e si ha che

$$(n+1)^{n+1} \geq n^n(n+1)$$

perché per i primi n termini di entrambe si ha che $n+1 > n$. Pertanto considerando le due precedenti equazioni

$$(n+1)^{n+1} \geq n^n(n+1) \geq (n+1)!$$

si conclude guardando al primo e terzo membro

Esercizio 2.5.3 (Es 1.26.3 pag 33 giusti1)**.** Dimostrare per induzione che $2 \cdot 4 \cdot \ldots \cdot 2n = 2^n n!$
Soluzione. Si vuole dimostrare che

$$\prod_{i=1}^{n} 2i = 2^n n!$$

Per induzione:

- se $i = 1$ si ha che $2 = 2^1 \cdot 1! = 2$ quiondi ok
- per il passo induttivo moltiplichiamo entrambi i termini per $2(n+1)$

$$\left(\prod_{i=1}^{n} 2i\right) \cdot (2(n+1)) = \prod_{i=1}^{n+1} 2i = 2^n n! \cdot (2(n+1))$$
$$= 2^{n+1} \cdot (n+1)!$$

ed è ok.

Esercizio 2.5.4 (Es 1.26.3 pag 33 giusti1)**.** Dimostrare per induzione che $\forall n \geq 4, n! > 2^n$
Soluzione. Si ha:

- per il passo base, per $n = 4$ si ha

$$4! > 2^4 \iff 24 > 16$$

che è verificato

- per il passo induttivo moltiplichiamo entrambi i termini della disequazione generica per $(n+1)$

$$(n+1)! > 2^n(n+1)$$

ora si ha che $2^2(n+1) > 2^{n+1}$ dato che $(n+1) > 2$. Pertanto

$$(n+1)! > 2^n(n+1) > 2^{n+1}$$

e si conclude guardando primo e ultimo termine

Capitolo 3

Introduction

3.1 Probability space

Definizione 3.1.1 (Probability space)**.** Considering an experiment, it's a triplet $(\Omega, \mathscr{F}, \mathbb{P})$ composed by a σ-field $\mathscr{F}$ (or, same σ-algebra $\mathscr{A}$) and a probability function $\mathbb{P}$, used to describe the experiment it in mathematical way.

3.1.1 Sample space, events

Definizione 3.1.2 (Sample space, Ω)**.** The (non-null) set of possible outcomes of an experiment, $\Omega = \{\omega_1, \omega_2, \ldots\}$, of which *only one will occur*.

Osservazione 38. The assumption is that a-priori, before executing the experiment, we can know all the possible outcomes.

Definizione 3.1.3 (Outcome, ω)**.** One possible result of the experiment: $\omega \in \Omega$.

Definizione 3.1.4 (Event (E or A))**.** An event E is any subset of Ω.

Definizione 3.1.5 (Occurred event)**.** E occurred if it contains the result of the experiment.

Osservazione 39. Since an event is any subset of Ω the following are valid.

Definizione 3.1.6 (True event (Ω))**.** Always occurs, since at least an element of the Ω occurs during the event.

Definizione 3.1.7 (Impossible event ($\emptyset$))**.** Never occurs.

Definizione 3.1.8 (Singleton event (eventi elementari), $\{\omega\}$)**.** Events composed by a single experiment outcome.

Osservazione 40 (Plotting). With Venn diagram Ω is given by a rectangle, while events are represented by circles.

Esempio 3.1.1 (Coin toss)**.** Here $\Omega = \{h, t\}$ while h is one possible outcome. We could be interested in the events outcome is head $\{h\}$ (singleton), outcome is either head or tail, outcome is both head and tail (unilikely to occur), outcome is not a head.

Esempio 3.1.2 (Two dice throwing)**.** $\Omega = \{(1,1),(2,1),\ldots,(6,6)\}$. The event $E =$ first is one $= \{(1,1),\ldots,(1,6))\}$

Esempio 3.1.3 (Arrival order)**.** In arrival order of a race with 7 numbered horses $\Omega = \{7!$ permutations of (1,2,3,4,5,6,7)$\}$.

Esempio 3.1.4 (Number of cars counted at a crossroad during a minute)**.** $\Omega = \{0,1,2,\ldots\}$

Esempio 3.1.5 (Bulb lifetime)**.** Will be a positive real number so $\Omega = \{x \in \mathbb{R}^+ | x \geq 0\}$.

Definizione 3.1.9 (Sample space cardinality)**.** Sample spaces of experiments can be *finite* (eg 3.1.1, 3.1.2) 3.1.3) *countable* (in bijection with $\mathbb{N}$, eg 3.1.4) or *non countable* (bijection with $\mathbb{R}$, eg 3.1.5)

3.1.1.1 Events algebra

Osservazione 41. Rules that applies to create new events; inherits from set theory being the events a set.

Definizione 3.1.10 (Union $A \cup B$)**.** Event that occurs if occurs one of A or B.

Osservazione 42. The outcomes composing the event are given by union of the outcomes of starting events.

Osservazione 43. Union can be extended to a numerable infinite number of events

$$E_1 \cup E_2 \cup \ldots \cup E_n \cup \ldots = \bigcup_{i=1}^{\infty} E_i \tag{3.1}$$

and verifies if at least one of E_i happens.

Definizione 3.1.11 (Intersection $A \cap B$ (A, B or AB))**.** Event that occurs if occur both A and B.

Osservazione 44. The outcome composing the event are given by intersection of the outcomes of starting events.

Osservazione 45. Similarly intersection event can be extended to a numerable infinite set of events

$$E_1 \cap E_2 \cap \ldots \cap E_n \cap \ldots = \bigcap_{i=1}^{\infty} E_i \tag{3.2}$$

Definizione 3.1.12 (Complement/negation event)**.** The negation of the event A, typed $\overline{A}$ o A^c, is the event that happens if A does not: $A^c = \Omega \setminus A$.

Definizione 3.1.13 (Difference $A \setminus B$)**.** Events that occurs when A occurs but not B: $A \setminus B = A \cap \overline{B}$.

Osservazione 46. The outcome composing the event are given by the set difference $A \setminus B$ outcomes of starting events.

Definizione 3.1.14 (Symmetric difference $A \triangle B$ (xor))**.** Events that occur if A or B occurs, but not both

Osservazione 47. The outcome composing the event are given by $(A \cup B) \setminus (A \cap B)$.

Property	Union	Intersection
Idempotenza	$A \cup A = A$	$A \cap A = A$
Elemento neutro	$A \cup \emptyset = A$	$A \cap \Omega = A$
Commutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Associativa	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributiva	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Tabella 3.1: Proprietà di unione ed intersezione

Operation properties

Osservazione importante 5. Operation properties are the same as set properties and summarized in tab 3.1; same for DeMorgan Laws.

Proposizione 3.1.1 (DeMorgan laws)**.** *With two events*

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \tag{3.3}$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \tag{3.4}$$

while in the general form

$$\overline{\bigcap_i E_i} = \bigcup_i \overline{E_i} \tag{3.5}$$

$$\overline{\bigcup_i E_i} = \bigcap_i \overline{E_i} \tag{3.6}$$

3.1.1.2 Relationship between events

Definizione 3.1.15 (Inclusion, $A \subseteq B$)**.** Event A is included in B, $A \subseteq B$ if each time A happens, B happens as well.

Esempio 3.1.6. $E_1 = \{1, 2\}$ ("dice below 3") is included in $E_2 = \{1, 2, 3\}$ ("dice below 44")

Definizione 3.1.16 (Monotone increasing sequence of events)**.** A sequence of events $E_1, E_2, \ldots$ where $E_1 \subseteq E_2 \subseteq \ldots$.

Definizione 3.1.17 (Monotone decreasing sequence of events)**.** A sequence of events $E_1, E_2, \ldots$ where $E_1 \supseteq E_2 \supseteq \ldots$

Definizione 3.1.18 (Incompatibility/disjointness, $A \cap B = \emptyset$)**.** A and B are incompatible (or disjoint) if they can't verify together, that is, $A \cap B = \emptyset$.

Esempio 3.1.7. If $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$ (two dice sum to 7) and $B = \{(1,5), (2,4), (3,3), (4,2), (5,1)\}$ (sum to 6) are incompatible because $A \cap B = \emptyset$.

Osservazione 48. In Venn diagrams, two disjoint events are represented by non overlapping areas.

Definizione 3.1.19 (Pairwise disjointness/incompatibility/exclusiveness)**.** Given a collection of events E_i, $1 \leq i \leq \infty$, they are pairwise disjoint if

$$E_i \cap E_j = \emptyset \quad \forall i \neq j$$

Osservazione importante 6. The same can be defined for 3-folded incompatibility or n-folded. Clearly pairwise disjointness implies higher level disjointness (eg 3-folded, etc); viceversa does not happens.

Definizione 3.1.20 (Jointly exhaustive events (eventi necessari), $A \cup B = \Omega$)**.** A and B are jointly exhaustive if at least one event occurs, that is $A \cup B = \Omega$.

Osservazione 49. Same applies for a collection: $E_i, 1 \leq i \leq \infty$ is jointly exhaustive if at least one event occurs $\bigcup_{i=1}^{\infty} E_i = \Omega$

Definizione 3.1.21 (Ω partition)**.** It's a set of events $\{E_i\}_{i \in I}, E_i \subseteq \Omega$ which are both disjoint and jointly exhaustive:

$$E_i \cap E_j = \emptyset \quad i \neq j, \qquad \bigcup_{i=1}^{\infty} E_i = \Omega$$

Osservazione 50. On Venn diagrams it's a set of non overlapping shapes that sum up to Ω.

Osservazione 51. If the set of events E_i is finite, countable or uncountable (eg idem the set of index I), the partition of Ω will respectively be called finite, countable or uncountable.

Esempio 3.1.8. Suppose $\Omega = \mathbb{R}$, collection of all $\{x\}$ with $x \in \mathbb{R}$ is a partition (not finite nor countable, it's uncountable).

3.1.2 σ-field $\mathscr{F}$ (or σ-algebra $\mathscr{A}$)

Osservazione importante 7. Events are subset of Ω but it's not needed all the subsets of Ω, elements of $\mathcal{P}(\Omega)$, to be events (for technical complex reasons). It suffices for us to think of the collection of events as a subcollection $\mathscr{F} \subseteq \mathcal{P}(\Omega)$ of the power set of the sample space, having certain reasonable/minimal properties.

Definizione 3.1.22 (σ-field $\mathscr{F}$ (or σ-algebra $\mathscr{A}$))**.** Set of all the possible events of interest, $\mathscr{F} \subseteq \mathcal{P}(\Omega)$ having the following properties

1. $\emptyset \in \mathscr{F}$
2. $\mathscr{F}$ is closed under complements: $A \in \mathscr{F} \implies A^c \in \mathscr{F}$
3. $\mathscr{F}$ is closed under *finite* or *countable* unions (and intersection as well): if $E_1, E_2, \ldots \in \mathscr{F}$ is a finite or countable set of events then $\bigcup_{i=1}^{\infty} E_i \in \mathscr{F}$

Corollario 3.1.2. *$\Omega \in \mathscr{F}$ and $\mathscr{F}$ is closed under finite or countable intersection as well:*

$$\Omega = \emptyset^c \in \mathscr{F} \tag{3.7}$$

$$E_1, E_2, \ldots \in \mathscr{F} \implies \bigcap_{i=1}^{+\infty} E_i = \left(\bigcup_{i=1}^{+\infty} E_i^c \right)^c \in \mathscr{F} \tag{3.8}$$

the last by applying proprieties 2, 3 of the definition and DeMorgan's laws.

Osservazione 52. The idea is that

- if I make some operations of interest (unions, intersections, complement) can be confident of being inside the σ-algebra.
- $\mathscr{F}$ can be thought as the set of all possible events that are relevant regarding the considered experiment (probabilistic meaning of $\mathscr{F}$)
- if the set of possible events $\mathscr{E}$ of our interest is not a σ-algebra, then we set $\mathscr{F} = \sigma(\mathscr{E})$ as the minimum σ-algebra containing $\mathscr{E}$, and "work" with this one.

Esempio 3.1.9. $\mathscr{F} = \{\emptyset, \Omega\}$ is the least possible σ-field

Esempio 3.1.10. $\mathscr{F} = \{\emptyset, \Omega, A, A^c\}$ is the least possible σ-field including A.

Esempio 3.1.11 (Power set (insieme delle parti) of Ω as $\mathscr{F}$)**.** If $\mathscr{F} = \mathcal{P}(\Omega)$, then it's the most possible, that is no other $\mathscr{F}$ can be bigger (in terms of cardinality). If:

- Ω is finite, it can be $\mathscr{F} = \mathcal{P}(\Omega)$.
- Ω is countable (eg $\mathbb{N}$), its power set can be a *sigma*-field (see here).
- Ω is *non countable*, its power set is a too large collection for probabilities to be assigned reasonably (eg all being non negative and singleton events probabilities summing up to 1) to all its members

Osservazione importante 8. In case of $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^n$ we consider a particular case of σ-field called Borel σ-field

Definizione 3.1.23 (Intervals of $\mathbb{R}$)**.** The intervals of $\mathbb{R}$ are (a, b), $[a, b]$, $(a, b]$, $[a, b)$ $(-\infty, b]$, $(-\infty, b)$, (a, ∞), $[a, \infty)$, and $\mathbb{R}$ as well.

Definizione 3.1.24 (Borel σ-field on $\mathbb{R}$)**.** The borel sigma-field on $\mathbb{R}$, here denoted by $\beta(\mathbb{R})$, is the least (più piccolo) sigma-field including all the $\mathbb{R}$ intervals.

Osservazione importante 9. These are reasonable/desiderable properties; note that if $\Omega = \mathbb{R}$ and $\mathscr{E}$ is a set of intervals of R but *not* a σ-field by definition it could happen that $(-1, 5) \cup [7, 8] \notin \mathscr{E}$; same for $(-1, 5]^c = (-\infty, -1) \cup (5, +\infty) \notin \mathscr{E}$

Esempio 3.1.12. Are singleton events in $\beta\mathbb{R}$? Yes because $x = (x - 1, x] \cap [x, x + 1) \in \beta\mathbb{R}$ $\forall x \in \mathbb{R}$.
In addition to singletons, $\beta\mathbb{R}$ includes all sets which can be obtained, starting from intervals, by a countable numbers of unions, intersections and complements.

Definizione 3.1.25 (Borel σ-field on $\mathbb{R}^n$)**.** In the same way, if $\Omega = \mathbb{R}^n$, $\beta(\mathbb{R}^n)$ equals to the least σ-field on $\mathbb{R}^n$ including all sets of the form $I_1 \times I_2 \times \ldots \times I_n$, where I_i is an interval of $\mathbb{R}$

Osservazione importante 10. Note that $\exists A \subset \mathbb{R}$ such as that $A \notin \beta\mathbb{R}$; in other terms $\beta\mathbb{R}$ is *not* the power set of $\mathbb{R}$.

3.1.3 Probability measure $\mathbb{P}$

Osservazione 53. In our construction the third element is the probability function $\mathbb{P}$, defined according to three Kolmogorov axioms that specifies basic features of any probability function.

Definizione 3.1.26 (Probability function, $\mathbb{P}$)**.** It's a measure that is characterized by $\mathbb{P}(\Omega) = 1$; in other words it's a function $\mathbb{P} : \mathscr{F} \to [0,1]$ such as that

$$\mathbb{P}(A) \geq 0, \quad \forall A \in \mathscr{F} \tag{3.9}$$

$$\mathbb{P}(\Omega) = 1 \tag{3.10}$$

$$A_i \cap A_j = \emptyset, \forall i \neq j \implies \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) \tag{3.11}$$

The latter being a numerable set of incompatible events.

Osservazione importante 11. A measure, generally speaking, is a function assigning a positive number to each set and for which measure of disjoint set is sum of measure.

Esempio 3.1.13. A coin, possibly biased is tossed once. We have $\Omega = \{h, t\}$, $\mathscr{F} = \{\emptyset, \{h\}, \{t\}, \Omega\}$ and a *possible* probability measure (it fullfill the requirements) $\mathbb{P} : \mathscr{F} \to [0,1]$ is given by

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{h\}) = p, \quad \mathbb{P}(\{t\}) = 1 - p, \quad \mathbb{P}(\Omega = \{h, t\}) = 1$$

where p os a foxed real number in the interval $[0,1]$. If $p = \frac{1}{2}$ then we say the coin is *fair* or unbiased.

Definizione 3.1.27 (Null event)**.** Events A such as $\mathbb{P}(A) = 0$.

Definizione 3.1.28 (Event which occurs almost surely)**.** Event A such as $\mathbb{P}(A) = 1$.

Osservazione importante 12 (Null vs impossible events, true vs almost surely events). Null events should not be confused with the impossible event $\emptyset$: null events are happening all around us, even though they have zero probability (eg what's the chance that a dart strikes any given point of the target at which it's thrown).
That is: the impossible event is null, but null events need not to be impossible. Specular considerations for Ω with events A such as $\mathbb{P}(A) = 1$.

3.2 Probability

3.2.1 Immediate or useful general results

Osservazione 54. Let's see some properties following directly from the definition; in what follows we consider generic events $A, B \subseteq \Omega$.

Proposizione 3.2.1.

$$\boxed{\mathbb{P}\left(\overline{A}\right) = 1 - \mathbb{P}(A)} \tag{3.12}$$

Dimostrazione.

$$\begin{aligned} \Omega &= A \cup \overline{A} \\ \mathbb{P}(\Omega) &= \mathbb{P}\left(A \cup \overline{A}\right) \\ 1 &= \mathbb{P}(A) + \mathbb{P}\left(\overline{A}\right) \end{aligned}$$

□

Esempio 3.2.1. If the probability of having head with coin is $\frac{3}{8}$ then probability of tail have to be $\frac{5}{8}$.

Proposizione 3.2.2.

$$\boxed{\mathbb{P}(\emptyset) = 0} \tag{3.13}$$

Dimostrazione. Setting $A = \Omega$ in 3.12,

$$\begin{aligned} \mathbb{P}\left(\overline{\Omega}\right) &= 1 - \mathbb{P}(\Omega) \\ \mathbb{P}(\emptyset) &= 1 - 1 \end{aligned}$$

□

Proposizione 3.2.3.

$$\boxed{A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)} \tag{3.14}$$

Dimostrazione. If $A \subseteq B$, B can be written as union of two incompatible events A and $(B \setminus A)$; applying third axiom

$$\begin{aligned} B &= A \cup (B \setminus A) \\ \mathbb{P}(B) &= \mathbb{P}(A) + \mathbb{P}(B \setminus A) \end{aligned}$$

since $\mathbb{P}(B \setminus A) \geq 0$ by axioms, then $\mathbb{P}(B) \geq \mathbb{P}(A)$, □

Proposizione 3.2.4 (Probability that A occurs but not B)**.**

$$\boxed{\mathbb{P}(A \setminus B) = \mathbb{P}\left(A \cap \overline{B}\right) = \mathbb{P}(A) - \mathbb{P}(A \cap B)} \tag{3.15}$$

Dimostrazione. Looking at A as union of incompatible events (think using Venn diagram):

$$\begin{aligned} A &= (A \cap B) \cup (A \cap \overline{B}) \\ \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}\left(A \cap \overline{B}\right) \end{aligned}$$

then we conclude as in proposition. □

Proposizione 3.2.5 (Probability of union)**.**

$$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)} \tag{3.16}$$

Dimostrazione. Writing $A \cup B$ as union of two incompatible events,we apply axioms and 3.15:

$$\begin{aligned} A \cup B &= A \cup (B \cap \overline{A}) \\ \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}\left(B \cap \overline{A}\right) \\ \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \end{aligned}$$

□

Proposizione 3.2.6 (Inclusion/exclusion formula)**.** *Considering a finite union of events, probability of their union is calculated according to the following:*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n}(-1)^{r+1} \sum_{i_1<\ldots<i_r} \mathbb{P}(E_{i1} \cap E_{i2} \cap \ldots \cap E_{ir}) \tag{3.17}$$

$$= \sum_i \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k) - \ldots$$

$$\ldots + (-1)^{n+1}\mathbb{P}(E_1 \cap \ldots \cap E_n) \tag{3.18}$$

Dimostrazione. Can be proved by induction, as we'll see in 5.3.3. □

Esempio 3.2.2. In case of three events, E, F, G:

$$\begin{aligned} \mathbb{P}(E \cup F \cup G) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(E \cap F) \ldots \\ - \mathbb{P}(E \cap G) - \mathbb{P}(F \cap G) + \mathbb{P}(E \cap G \cap F) \end{aligned}$$

Proposizione 3.2.7 (Boole inequality (on union))**.**

$$\mathbb{P}(E_1 \cup E_2 \cup \ldots \cup E_n) \leq \sum_{i=1}^{n} \mathbb{P}(E_i) \tag{3.19}$$

Dimostrazione. Done in the following section 5.3.3. □

Proposizione 3.2.8 (Bonferroni inequality (on intersection))**.**

$$\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\overline{E_i}\right) \tag{3.20}$$

Dimostrazione. In section 5.3.3. □

Proposizione 3.2.9. *If $A_1, A_2, \ldots$ is an increasing sequence of events, so that $A_1 \subseteq A_2 \subseteq \ldots$ and we set A as the limit of the* union*:*

$$A = \bigcup_{i=1}^{+\infty} A_i = \lim_{i \to +\infty} A_i$$

then it follows that

$$\mathbb{P}(A) = \lim_{i \to +\infty} \mathbb{P}(A_i) \tag{3.21}$$

Proposizione 3.2.10. *Similarly if $B_1, B_2, \ldots$ is decreasing sequence of events $B_1 \supseteq B_2 \supseteq \ldots$ and we set as B the limit of the* intersection*:*

$$B = \bigcap_{i=1}^{+\infty} B_i = \lim_{i \to +\infty} B_i$$

then

$$\mathbb{P}(B) = \lim_{i \to +\infty} \mathbb{P}(B_i) \tag{3.22}$$

Dimostrazione. We prove only the first; we have that A can be seen as an union of a disjoint family of events

$$A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \ldots$$

Thus by definition of the probability function its probability is a sum of the disjoint events (again think with Venn, these are enclosing circles)

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(A_1) + \sum_{i=1}^{+\infty} \mathbb{P}(A_{i+1} \setminus A_i) \\
&= \mathbb{P}(A_1) + \lim_{n \to +\infty} \sum_{i=1}^{n-1} [\mathbb{P}(A_{i+1}) - \mathbb{P}(A_i)] \\
&= \lim_{n \to +\infty} \mathbb{P}(A_n)
\end{aligned}$$

The last passage involve semplification/elision. For the second results on B, take complements and use the first part. □

3.2.2 Finite equiprobable Ω and probability evaluation

Osservazione 55. In previous section we never evaluated a probability. In this one we show how it's done for the particular case where Ω is finite with every $\omega in \Omega$ having the same probability of occurring.
It's a reasonable assumption in several cases (eg balanced dice, coins etc)

Proposizione 3.2.11 (Probability of singleton a event)**.** *If Ω is* finite*, $\Omega = \{1, 2, \ldots, n\}$, and $\mathbb{P}(1) = \mathbb{P}(2) = \ldots = \mathbb{P}(n) = p$, being the singleton events disjoint and the probability of their union summing to 1 ($p \cdot n = 1$), we'll have*

$$p = \frac{1}{n}$$

Proposizione 3.2.12 (Probability of general event)**.** *Given a generic event E, its probability will be*

$$\mathbb{P}(E) = \frac{\#\ of\ outcomes\ composing\ E}{\#\ possible\ outcomes} = \frac{|E|}{|\Omega|}$$

Osservazione 56. In words, number of favorable outcome of event E out of possible outcomes of Ω. Often, count of numerator/denominator uses combinatorics.

Esempio 3.2.3. We have an urn with n numbered balls from 1 to n, we draw without replacement. Let's define $C_i =$ "concordance at trial i" as the selected ball at draw i is numbered i. We are interested in evaluating $\mathbb{P}(E)$ where $E =$ no concordance in n draws.
By applying the previous properties:

$$\mathbb{P}(E) = 1 - \mathbb{P}(\text{at least one concordance}) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} C_i\right)$$

$$= 1 - \left\{\sum_i \mathbb{P}(C_i) - \sum_{i<j} \mathbb{P}(C_i \cap C_j) + \sum_{i<j<k} \mathbb{P}(C_i \cap C_j \cap C_k) \ldots + (-1)^{n+1} \mathbb{P}(C_1 \cap \ldots \cap C_n)\right\}$$

Now

$$\mathbb{P}(C_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

we have n slots, the sequences of balls can be $n!$, while the sequence where i ball is at the i-th place are $(n-1)!$ (fix i in its place and then permute the remaining balls). Furthermore for similar reasons

$$\mathbb{P}(C_i \cap C_j) = \frac{(n-2)!}{n!}$$
$$\mathbb{P}(C_i \cap C_j \cap C_k) = \frac{(n-3)!}{n!}$$
$$\ldots$$
$$\mathbb{P}(C_1 \cap \ldots \cap C_n) = \frac{1}{n!}$$

Therefore

$$\mathbb{P}(E) = 1 - \left\{n \cdot \frac{1}{n} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} \ldots + (-1)^{n+1}\frac{1}{n!}\right\}$$

Esempio 3.2.4 (Birthday problem)**.** Ci sono k persone in una stanza. Assumendo che siano nate in uno dei 365 giorni dell'anno con probabilità uguale per ciascun giorno (escludiamo anni bisestili) e che i compleanni siano indipendenti (es non vi sono gemelli nella stanza), quale è la probabilità che due o più persone nel gruppo compiano gli anni lo stesso giorno?
La calcoliamo come complemento della probabilità che nessuno faccia compleanno lo stesso giorno: questa è data da casi favorevoli (numero di modi possibili per avere compleanni in date differente) fratto casi possibili (numero di possibili configurazioni di compleanni. Si ha:

$$\mathbb{P}(k \text{ compleanni diversi}) = \frac{365 \cdot \ldots \cdot (365-k+1)}{365^k}$$

da cui

$$\mathbb{P}(\text{Almeno due uguali tra } k) = 1 - \frac{365 \cdot \ldots \cdot (365-k+1)}{365^k}$$

Eseguendo i conti si nota come si supera la probabilità del 50% già con $k = 23$ persone (ossia in un gruppo di 23 persone c'è poco più del 50% di probabilità di averne due o più che fanno gli anni lo stesso giorno) mentre a $k = 57$ la probabilità è già oltre il 99%.

```r
prob_birthday <- function(k){
    # vectorized for several k
    num <- unlist(lapply(k, function(k2) prod(seq(365, 365 - k2 + 1))))
    den <- 365^k
    1 - num/den
}
k <- 1:60
round(prob_birthday(k = k), 4)
```

```
##  [1] 0.0000 0.0027 0.0082 0.0164 0.0271 0.0405 0.0562 0.0743 0.0946 0.1169
## [11] 0.1411 0.1670 0.1944 0.2231 0.2529 0.2836 0.3150 0.3469 0.3791 0.4114
## [21] 0.4437 0.4757 0.5073 0.5383 0.5687 0.5982 0.6269 0.6545 0.6810 0.7063
## [31] 0.7305 0.7533 0.7750 0.7953 0.8144 0.8322 0.8487 0.8641 0.8782 0.8912
## [41] 0.9032 0.9140 0.9239 0.9329 0.9410 0.9483 0.9548 0.9606 0.9658 0.9704
## [51] 0.9744 0.9780 0.9811 0.9839 0.9863 0.9883 0.9901 0.9917 0.9930 0.9941
```

3.2.3 Conditional probability

Osservazione 57. Often is needed to compute probability of an event in case another happens; or it's easier to compute a probability of event A conditioning on information of another event B.

Definizione 3.2.1 (Conditioned probability of A given B)**.** If $\mathbb{P}(B) > 0$ it's defined as

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}} \tag{3.23}$$

Osservazione 58. Can be interpreted as the probability of having A if actually B occurred/will occur.

Osservazione importante 13. $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$; denominators are different.

Osservazione 59. Limit/extreme cases:

$$\begin{aligned} A \cap B = \emptyset &\implies \mathbb{P}(A|B) = 0 \\ A \subseteq B &\implies \mathbb{P}(A|B) = 1 \end{aligned}$$

3.2.4 Probability of intersection

Proposizione 3.2.13 (For two events, $\mathbb{P}(A \cap B)$)**.** *If* $\mathbb{P}(B) \neq 0$*:*

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(B)\,\mathbb{P}(A|B)} \tag{3.24}$$

Symmetrically, if $\mathbb{P}(A) \neq 0$*:*

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B|A)} \tag{3.25}$$

Dimostrazione. Algebric manipulation of 3.23. □

Proposizione 3.2.14 (n events (*product rule*))**.** *Given* $E_1, \ldots, E_n \in \mathscr{F}$ *if* $\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1}) > 0$*, then:*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \mathbb{P}(E_1) \cdot \mathbb{P}(E_2|E_1) \cdot \mathbb{P}(E_3|E_1 \cap E_2) \cdot \ldots \cdot \mathbb{P}(E_n|E_1 \cap E_2 \cap \ldots \cap E_{n-1})$$

Dimostrazione. To verify it we apply recursively the definition 3.25 to the second member:

$$\mathbb{P}(E_1) \cdot \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} \cdot \frac{\mathbb{P}(E_1 \cap E_2 \cap E_3)}{\mathbb{P}(E_1 \cap E_2)} \cdot \ldots \cdot \frac{\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n)}{\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1})} \quad (3.26)$$

and after simplifying it remains $\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right)$.
Note that denominators in 3.26 are strictly positive thanks to the hypothesis $\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1}) > 0$: since intersection on $n-1$ events is not null, even the intersection of less events will be. □

Osservazione 60. In practice we can handle/manipulate events as we prefer, eg:

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2 \cap E_3) &= \mathbb{P}(E_1) \cdot \mathbb{P}(E_2|E_1) \cdot \mathbb{P}(E_3|E_1 \cap E_2) \\ &= \mathbb{P}(E_3) \cdot \mathbb{P}(E_2|E_3) \cdot \mathbb{P}(E_1|E_3 \cap E_2)\end{aligned}$$

3.2.5 Law of total probability

Osservazione 61 (Basic version). If E and C are two events we can split E in disjoint union as follows:

$$E = (E \cap C) \cup (E \cap \overline{C})$$

Being disjoint:

$$\boxed{\mathbb{P}(E)} = \mathbb{P}\left((E \cap C) \cup (E \cap \overline{C})\right) \quad (3.27)$$

$$= \mathbb{P}(E \cap C) + \mathbb{P}\left(E \cap \overline{C}\right)$$

$$= \boxed{\mathbb{P}(C)\,\mathbb{P}(E|C) + \mathbb{P}\left(\overline{C}\right)\mathbb{P}\left(E|\overline{C}\right)} \quad (3.28)$$

Osservazione 62 (Conditioning for problem solving). Sometimes is difficult to calculate $\mathbb{P}(E)$; this can become easier if we can condition on C (and $\overline{C}$), and summing up applying the previous formula. It's common practice to condition on hypothesis/hypothetical situation or, in sequential experiment, conditioning on previous steps.

Teorema 3.2.15 (Law of total probability (general version))**.** *If $C_1, C_2, \ldots$ is a* finite *or* countable *partition of Ω, the probability of a generic event E can be written as:*

$$\boxed{\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)} \quad (3.29)$$

Dimostrazione. If $C_1, C_2, \ldots, C_n$ is a partition of Ω, we can split E in disjoint pieces by intersection with C_i

$$E = \Omega \cap E = \left(\bigcup_{i=1}^{n} C_i\right) \cap E = (C_1 \cap E) \cup (C_2 \cap E) \cup \ldots \cup (C_n \cap E)$$

Being $(C_i \cap A)$ disjoint probability is the sum:

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(C_i \cap E) = \sum_{i=1}^{n} \mathbb{P}(C_i)\,\mathbb{P}(E|C_i) \quad (3.30)$$

and in the last passage we substitued 3.25. □

Osservazione importante 14. Looking at the formula, here it's not a problem if $\mathbb{P}(C_i) = 0$ (which is at the denominator of $\mathbb{P}(E|C_i)$, which would be undefined); undefined multiplied by zero is not considered in the sum.

Esempio 3.2.5. Domani potrebbe o piovere o nevicare, ma i due eventi non si possono verificare contemporaneamente. La probabilità che piova è $2/5$, mentre la probabilità che nevichi è $3/5$. Se pioverà, la probabilità che io faccia tardi a lezione è di $1/5$, mentre la probabilità corrispondente nel caso in cui nevichi è di $3/5$. Calcolare la probabilità che io sia in ritardo.

Si ha P = piove, $N = P^c$ = nevica, R = ritardo; avendo a che fare con una partizione

$$\mathbb{P}(R) = \mathbb{P}(P)\,\mathbb{P}(R|P) + \mathbb{P}(N)\,\mathbb{P}(R|N) = \frac{2}{5}\frac{1}{5} + \frac{3}{5}\frac{3}{5} = \frac{11}{25}$$

Esempio 3.2.6 (Esempio Rigo)**.** Having an urn with n_w white and n_b black balls, we draw without replacement. We are interested in $\mathbb{P}(W_2)$ where W_2 = white ball at second draw: it is not trivial without formula, since we don't know the result of the first trial. We however can calculate it conditioning on first draw results.
Let's set W_1 = white at first draw and B_1 = black at first draw; since $\{W_1, B_1\}$ is a finite partition of the sample space of the first trial, we can apply the law of total probabilities:

$$\mathbb{P}(W_2) = \mathbb{P}(W_1)\,\mathbb{P}(W_2|W_1) + \mathbb{P}(B_1)\,\mathbb{P}(W_2|B_1)$$

Given that we have $n = n_w + n_b$ balls and we draw without replacement

$$\mathbb{P}(W_1) = \frac{n_w}{n},\ \mathbb{P}(B_1) = \frac{n_b}{n},\ \mathbb{P}(W_2|W_1) = \frac{n_w - 1}{n-1},\ \mathbb{P}(W_2|B_1) = \frac{n_w}{n-1},$$

Therefore, overall

$$\mathbb{P}(W_2) = \frac{n_w}{n}\cdot\frac{n_w - 1}{n-1} + \frac{n_b}{n}\cdot\frac{n_w}{n-1} = \ldots = \frac{n_w}{n}$$

This is a counterintuitive result, since it's the same as drawing *with* replacement.
Furthermore, in general if W_j = white at draw j, $\mathbb{P}(W_j)$ is still $\frac{n_w}{n}$. In this case we have to condition on the partition of the first $j-1$ trials.
Eg regarding W_3 = white at draw 3 the first two draws will have $\Omega = \{ww, wb, bw, bb\}$, so

$$\begin{aligned}\mathbb{P}(W_3) &= \mathbb{P}(ww)\,\mathbb{P}(W_3|ww) + \mathbb{P}(wb)\,\mathbb{P}(W_3|wb) + \mathbb{P}(bw)\,\mathbb{P}(W_3|bw) + \mathbb{P}(bb)\,\mathbb{P}(W_3|bb)\\ &= \ldots = \frac{n_w}{n}\end{aligned}$$

Eg in this case $\mathbb{P}(W_3|ww) = \frac{n_w-2}{n-2}$

Osservazione 63. Suppose a partition $E_1, E_2, \ldots$ of Ω is *finite* or *countable* and we want to assign the same probability to all E_i. Is it possible?

Proposizione 3.2.16. *It's possible to assign to element/events of a* finite *partition of Ω the same probability; if the partition is* countable *this is no more possible.*

Dimostrazione. If the partition is *finite* in n events E_i, it suffices to assign $\mathbb{P}(E_i) = \frac{1}{n}$, so that $\mathbb{P}(\Omega) = \mathbb{P}(\cup_{i=1}^{n} E_i) = 1$.
If the partition is countable this is impossible: let's prove it by absurd/contradiction. Suppose be $\mathbb{P}(E_i) = c \geq 0, \forall i$. Then

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) = \sum_{i=1}^{\infty} c = \begin{cases} 0 & \text{if } c = 0 \\ +\infty & \text{if } c > 0 \end{cases}$$

Therefore we have a contraddiction: 1 can't be equal to 0 or $+\infty$ □

3.2.6 Bayes formula

Teorema 3.2.17 (Bayes formula)**.**

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B|A)}{\mathbb{P}(B)}} \tag{3.31}$$

Dimostrazione. Substitute 3.25 in 3.23. □

Osservazione 64 (Decision making and knowledge update). When performing a test to verify an hypothesis, bayes formula is used like this: let H be "my hypothesis is true", and T "positive test"; then:

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(H) \cdot \mathbb{P}(T|H)}{\mathbb{P}(T)}$$

in this case $\mathbb{P}(H)$ is called *a priori probability* $\mathbb{P}(T|H)$ *likelihood* and $\mathbb{P}(H|T)$ *posterior probability* (the denominator is merely a normalizing constant).

Osservazione 65 (Bayes in diagnostic: PPV and NPV). If D is "being diseased" and T è "being positive to diagnostic test", $\mathbb{P}(D|T)$ (applying bayes formula) is Positive predictive value while $\mathbb{P}\left(\overline{D}|\overline{T}\right)$ is negative predictive value..

Corollario 3.2.18. *Let E be a generic event and $C_1, C_2, \ldots, C_n$ a finite partition of Ω; the conditional probability of C_i given E is:*

$$\boxed{\mathbb{P}(C_i|E) = \frac{\mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}{\sum_{i=1}^{n} \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}}$$

Dimostrazione. We started from $\mathbb{P}(C_i|E)$ defined using Bayes law and then substituted the denominator using the law of total probability:

$$\mathbb{P}(C_i|E) = \frac{\mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}{\mathbb{P}(E)} = \frac{\mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}{\sum_{i=1}^{n} \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)}$$

□

Osservazione 66 (Interpretation). E can be thought as an occurred event/effect that is dued to only one of n causes C_i (disjoint, exaustive: that is one and only one of them surely happened) each one of the cause has probability $\mathbb{P}(C_i)$ to happen.
The theorem allows us to evaluate $\mathbb{P}(C_i|E)$, that is probability that having observed E, this has been caused by C_i. In the process we use prior probability $\mathbb{P}(C_i)$ and likelihood $\mathbb{P}(E|C_i)$ at numerator (denominator is a normalizing constant):

- when prior probability is not known, if the partition is *finite* (see 3.2.16), one can assign a common probability $\mathbb{P}(C_i) = 1/n, \forall i$;
- likelihood is generally easier to know/evaluate;
- we conclude C_i as the most reasonable cause if its $\mathbb{P}(C_i|E)$ is higher than the others;
- the final result depends only on the numerator, being the denominator a normalizing constant common for all C_i (and making posteriors $\mathbb{P}(C_i|E)$ to sum up to 1). For this reason we can write

$$\mathbb{P}(C_i|E) \propto \mathbb{P}(C_i)\,\mathbb{P}(E|C_i)$$

that is posterior probability is proportional to the prior time likelihood

Osservazione importante 15. It's often useful the simpler version of (where the partition of Ω composed by two events, only one of which is of interest, the other is the complement) reported here:

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(H)\cdot\mathbb{P}(T|H)}{\mathbb{P}(H)\cdot\mathbb{P}(T|H) + \mathbb{P}\left(\overline{H}\right)\cdot\mathbb{P}\left(T|\overline{H}\right)} \tag{3.32}$$

Esempio 3.2.7 (Moneta bilanciata)**.** Abbiamo una moneta bilanciata e una sbilanciata che cade su testa con probabilità $3/4$. Si sceglie una moneta a caso e la si lancia tre volte; restituisce testa tutte e tre le volte. Quale è la probabilità che la moneta scelta sia quella bilanciata?
Se H è l'evento "testa tre volte" e B è l'evento "scelta la moneta bilanciata"; siamo interessati alla probabilità $\mathbb{P}(B|H)$. Ci risulta tuttavia più semplice trovare $\mathbb{P}(H|B)$ e $\mathbb{P}\left(H|\overline{B}\right)$ dato che aiuta sapere quale moneta consideriamo per calcolare la probabilità di tre teste. Questo suggerisce l'utilizzo del teorema di Bayes e della legge delle probabilità totali. Si ha

$$\begin{aligned}\mathbb{P}(B|H) &= \frac{\mathbb{P}(B)\cdot\mathbb{P}(H|B)}{\mathbb{P}(B)\cdot\mathbb{P}(H|B) + \mathbb{P}\left(\overline{B}\right)\cdot\mathbb{P}\left(H|\overline{B}\right)} \\ &= \frac{(1/2)\cdot(1/2)^3}{(1/2)\cdot(1/2)^3 + (1/2)\cdot(3/4)^3} \\ &\approx 0.23\end{aligned}$$

Esempio 3.2.8 (Test di una malattia rara)**.** Un paziente è testato per una malattia che colpisce l'1% della popolazione. Sia D l'evento che "il paziente ha la malattia" e T il test è positivo (ossia suggerisce che il paziente abbia la malattia). Il paziente sottoposto al test risulta effettivamente positivo. Supponendo che il test sia accurato al 95%, ossia che $\mathbb{P}(T|D) = 0.95$ (la sensitività) ma anche che $\mathbb{P}\left(\overline{T}|\overline{D}\right) = 0.95$ (la specificità), qual è la probabilità che il paziente abbia effettivamente la malattia data la positività del test?
Applicando la formula di Bayes:

$$\begin{aligned}\mathbb{P}(D|T) &= \frac{\mathbb{P}(D)\,\mathbb{P}(T|D)}{\mathbb{P}(T)} \\ &= \frac{0.01\cdot 0.95}{\mathbb{P}(T)}\end{aligned}$$

$\mathbb{P}(T)$ non è così facile da ottenere (necessiterebbe di provare il test su tutta la popolazione), ma il teorema delle probabilità totali ci viene in soccorso:

$$\begin{aligned}\mathbb{P}(D|T) &= \frac{0.01 \cdot 0.95}{\mathbb{P}(D)\,\mathbb{P}(T|D) + \mathbb{P}\left(\overline{D}\right)\mathbb{P}\left(T|\overline{D}\right)} \\ &= \frac{0.01 \cdot 0.95}{0.01 \cdot 0.95 + 0.99 \cdot 0.05} \\ &\approx 0.16\end{aligned}$$

Pertanto vi è il 16% di probabilità che il paziente sia malato, anche se il test è positivo e lo strumento è affidabile: il fatto è che la malattia è estremamente rara e potrebbe essere un falso positivo, ossia un errore del test applicato (nella maggioranza dei casi) ad individui negativi.

3.3 Indipendent events

Definizione 3.3.1 (2 independent events, $A \perp\!\!\!\perp B$)**.** Two events A, B for which:

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)} \tag{3.33}$$

NB: Per rigo potrebbe essere unesercizio verificare indipendenza

Esempio 3.3.1. Tossing a fair coin two times we have $\Omega = \{ht, hh, th, tt\}$ each outcome with probability $1/4$. Defining $H_i =$ "i-th toss is a hed", we have $H_1 = \{ht, hh\}$, $H_2 = \{th, hh\}$; each has probability $\frac{1}{2}$. We have that $H_1 \cap H_2 = \{hh\}$ and since that

$$\mathbb{P}(H_1 \cap H_2) = \frac{1}{4} = \mathbb{P}(H_1) \cdot \mathbb{P}(H_2) = \frac{1}{2} \cdot \frac{1}{2}$$

the two events are independent: $H_1 \perp\!\!\!\perp H_2$. It makes sense since the result of the first outcome does not affect the next.

Proposizione 3.3.1 (Conditional probability of independent events)**.** *If A and B are independent and $\mathbb{P}(B) > 0$:*

$$\mathbb{P}(A|B) = \mathbb{P}(A) \tag{3.34}$$

Dimostrazione.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\,\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

□

Proposizione 3.3.2. *If $\mathbb{P}(B) = 0 \vee \mathbb{P}(B) = 1$, then A is independent of B, $\forall A$.*

Dimostrazione.

$$\begin{aligned}\mathbb{P}(B) = 0 &\implies \mathbb{P}(A \cap B) = 0 = 0 \cdot \mathbb{P}(A) = \mathbb{P}(B) \cdot \mathbb{P}(A) \\ \mathbb{P}(B) = 1 &\implies \mathbb{P}(A \cap B) = \mathbb{P}(A) = 1 \cdot \mathbb{P}(A) = \mathbb{P}(B) \cdot \mathbb{P}(A)\end{aligned}$$

□

Osservazione importante 16. The previous results applies even if the two events seems to be somewhat connected. Eg suppose $\mathbb{P}(B) = 0$ and $A \subseteq B$. According to intuition these seems not to be independent because if B happens A happens as well. However logic and math definition/point of view can be different in practice.

Osservazione importante 17 (Independence and disjointness). These are two different concepts, often confused:

- disjointness is a *relation between events*, depicted on Venn diagrams as non overlapping areas;
- independence is a *relation between probability of events*; since on Venn diagrams probability are not depicted, it's not graphically representable

In general, disjointness and independence have no relation, except in the following case

Proposizione 3.3.3. *Let be A, B events with positive probability; if they are disjoint/incompatible then they cannot be independent.*

Dimostrazione. If A, B are disjoint/incompatible it must be:

$$\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$$

If they were also independent it should be:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

but since we hypothized $\mathbb{P}(A), \mathbb{P}(B) > 0$, then $\mathbb{P}(A)\,\mathbb{P}(B) > 0$, which contraddict the previous statement on disjointness. □

Proposizione 3.3.4 (Independence and complements)**.** *If E e F are independent then the following couples are as well: E and $\overline{F}$, $\overline{E}$ and F, $\overline{E}$ e $\overline{F}$.*

Dimostrazione. Showing the first; suppose E,F are independent so $\mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F)$. We want to prove

$$\mathbb{P}\left(E \cap \overline{F}\right) = \mathbb{P}(E)\,\mathbb{P}\left(\overline{F}\right)$$

We split $E = (E \cap F) \cup (E \cap \overline{F})$ in a disjoint union and sum its component probability:

$$\mathbb{P}(E) = \mathbb{P}(E \cap F) + \mathbb{P}\left(E \cap \overline{F}\right)$$

therefore

$$\begin{aligned}
\mathbb{P}\left(E \cap \overline{F}\right) &= \mathbb{P}(E) - \mathbb{P}(E \cap F) \\
&= \mathbb{P}(E) - \mathbb{P}(E)\,\mathbb{P}(F) \\
&= \mathbb{P}(E)\,[1 - \mathbb{P}(F)] \\
&= \mathbb{P}(E)\,\mathbb{P}\left(\overline{F}\right)
\end{aligned}$$

Regarding $\overline{E}$ e F independence (and $\overline{E}$ e $\overline{F}$) it suffices to swap roles by negation/complement. □

Definizione 3.3.2 (3 independent events)**.** E, F, G are independent if:

$$\begin{aligned}\mathbb{P}(E \cap F) &= \mathbb{P}(E)\,\mathbb{P}(F)\\ \mathbb{P}(E \cap G) &= \mathbb{P}(E)\,\mathbb{P}(G)\\ \mathbb{P}(F \cap G) &= \mathbb{P}(F)\,\mathbb{P}(G)\\ \mathbb{P}(E \cap F \cap G) &= \mathbb{P}(E)\,\mathbb{P}(F)\,\mathbb{P}(G)\end{aligned}$$

Definizione 3.3.3 (Pairwise independence of 3 events)**.** E, F, G are pairwise independent if the first three equation above holds.

Osservazione 67. Pairwise independence is not enough to have independence.

NB: Altro esempio, volendo, rigo lez 2023-09-21.

Esempio 3.3.2. Throwing two coins ha $\Omega = \{tt, tc, ct, cc\}$. I seguenti eventi sono pairwise independent ma non independent:

- $A =$ "prima testa" $= \{tc, tt\}$
- $B =$ "seconda testa" $= \{ct, tt\}$
- $C =$ "le due monete danno lo stesso" $= \{cc, tt\}$

Infatti

$$\begin{aligned}\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) &= \frac{2}{4} = \frac{1}{2}\\ \mathbb{P}(A \cap B) = \mathbb{P}(\{tt\}) &= \frac{1}{4} = \mathbb{P}(A)\,\mathbb{P}(B)\\ \mathbb{P}(A \cap C) = \mathbb{P}(\{tt\}) &= \frac{1}{4} = \mathbb{P}(B)\,\mathbb{P}(C)\\ \mathbb{P}(B \cap C) = \mathbb{P}(\{tt\}) &= \frac{1}{4} = \mathbb{P}(A)\,\mathbb{P}(C)\\ \mathbb{P}(A \cap B \cap C) = \mathbb{P}(\{tt\}) &= \frac{1}{4} \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) = \frac{1}{8}\end{aligned}$$

Il punto è che sapere cosa è sucesso sia con A che con B determina/ci da informazione completa su C.

Esempio 3.3.3.

Osservazione importante 18. If E, F, G are indipendent, then E is indipendent from any event formed by union/intersection/complement of F e G.

Esempio 3.3.4. E is independent from $F \cup G$ being:

$$\begin{aligned}\mathbb{P}(E \cap (F \cup G)) &= \mathbb{P}((E \cap F) + (E \cap G))\\ &= \mathbb{P}(E \cap F) + \mathbb{P}(E \cap G) - \mathbb{P}(E \cap F \cap G)\\ &= \mathbb{P}(E)\,\mathbb{P}(F) + \mathbb{P}(E)\,\mathbb{P}(G) - \mathbb{P}(E)\,\mathbb{P}(F \cap G)\\ &= \mathbb{P}(E)\,[\mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G)]\\ &= \mathbb{P}(E)\,\mathbb{P}(F \cup G)\end{aligned}$$

Definizione 3.3.4 (Independence of n events)**.** n events $A_1, \ldots, A_n \subset \Omega$ are said to be independent if for any subgroup of m events, $1 < m \leq n$ we have:

$$\mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} \mathbb{P}(A_i) \tag{3.35}$$

Osservazione 68. Generally speaking, n-wise independence implies $n-1$-wise of its components but viceversa does not hold (eg pairwise does not imply 3-wise).

Definizione 3.3.5 (Independence of ∞ events)**.** Independent if any finite subset is.

3.4 Further topics

3.4.1 Odds ratio

Definizione 3.4.1 (Odds ratio (rapporto a favore))**.** L'odds ratio di un evento A è definito come

$$\mathrm{OR}(A) = \frac{\mathbb{P}(A)}{\mathbb{P}\left(\overline{A}\right)} = \frac{\mathbb{P}(A)}{1-\mathbb{P}(A)} \tag{3.36}$$

ed esprime quanto è più probabile che l'evento si realizzi rispetto al fatto che non si realizzi.

Osservazione 69. Per convertire da odds ratio a probabilità, come si può verificare sostituendo, si ha:

$$\mathbb{P}(A) = \frac{\mathrm{OR}(A)}{1+\mathrm{OR}(A)} \tag{3.37}$$

Osservazione 70. Può essere di interesse la modifica della probabilità che una ipotesi H sia vera $\mathbb{P}(H)$ quando si dispone di informazioni su una prova E; le probabilità condizionate dato E che H sia vera o meno

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(H\cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(H)\,\mathbb{P}(E|H)}{\mathbb{P}(E)}$$
$$\mathbb{P}\left(\overline{H}|E\right) = \frac{\mathbb{P}\left(\overline{H}\cap E\right)}{\mathbb{P}(E)} = \frac{\mathbb{P}\left(\overline{H}\right)\mathbb{P}\left(E|\overline{H}\right)}{\mathbb{P}(E)}$$

Definizione 3.4.2 (Odds ratio condizionato)**.** L'odds ratio dell'ipotesi H non è più $\frac{\mathbb{P}(H)}{\mathbb{P}\left(\overline{H}\right)}$, ma a seguito delle conoscenze su (o nell'ipotesi di) E è dato da:

$$\frac{\mathbb{P}(H|E)}{\mathbb{P}\left(\overline{H}|E\right)} = \frac{\frac{\mathbb{P}(H)\,\mathbb{P}(E|H)}{\mathbb{P}(E)}}{\frac{\mathbb{P}\left(\overline{H}\right)\mathbb{P}\left(E|\overline{H}\right)}{\mathbb{P}(E)}} = \frac{\mathbb{P}(H)}{\mathbb{P}\left(\overline{H}\right)}\cdot\frac{\mathbb{P}(E|H)}{\mathbb{P}\left(E|\overline{H}\right)} \tag{3.38}$$

Osservazione 71. A seguito dell'introduzione di una prova l'originale rapporto a favore $\frac{\mathbb{P}(H)}{\mathbb{P}\left(\overline{H}\right)}$ viene moltiplicato per un secondo termine che ne determina l'eventuale variazione: il rapporto a favore finale (e quindi la probabilità di H) aumenta se E è più probabile quando H è vera che quando H è falsa (secondo termine del prodotto) e diminuisce in caso contrario.

Esempio 3.4.1. Con riferimento all'esempio un altro modo conveniente era utilizzare 3.4.2 per il calcolo dell'odds ratio (e poi la 3.37 per passare a probabilità), evitando di dover utilizzare il teorema delle probabilità totali:

$$\frac{\mathbb{P}(D|T)}{\mathbb{P}\left(\overline{D}|T\right)} = \frac{\mathbb{P}(D)\,\mathbb{P}(T|D)}{\mathbb{P}\left(\overline{D}\right)\mathbb{P}\left(T|\overline{D}\right)} = \frac{0.01}{0.99}\cdot\frac{0.95}{0.05}\approx 0.19$$

da cui applicando la 3.37 si ha:

$$\mathbb{P}(D|T) \approx 0.19/(1+0.19) \approx 0.16$$

3.4.2 Conditional probability 2

3.4.2.1 È una probabilità

Osservazione 72. Quando condizioniamo su un evento F, aggiorniamo la nostra idea per essere coerente con questta conoscenza, ponendoci in un universo dove sappiamo che F è accaduto.
Entro questo nuovo universo, tuttavia, le leggi della probabilità funzionano come in precedenza dato che le probabilità condizionate sono probabilità a tutti gli effetti.

Proposizione 3.4.1. *La probabilità condizionata è una valida funzione di probabilità a tutti gli effetti in quanto rispetta gli assiomi di Kolmogorov. Si ha:*

$$0 \leq \mathbb{P}(E|F) \leq 1$$

$$\mathbb{P}(\Omega|F) = 1$$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i|F\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i|F) \quad \textit{se } E_i \cap E_j = \emptyset,\ \forall i \neq j$$

Dimostrazione. Per la prima dobbiamo mostrare che:

$$0 \leq \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} \leq 1$$

La prima disuguaglianza è ovvia, mentre la seconda discende dal fatto che $(E \cap F) \subseteq F$, da cui $\mathbb{P}(E \cap F) \leq \mathbb{P}(F)$.
La seconda segue dalla:

$$\mathbb{P}(\Omega|F) = \frac{\mathbb{P}(\Omega \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(F)}{\mathbb{P}(F)} = 1$$

Per la terza

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i|F\right) &= \frac{\mathbb{P}\left(\left(\bigcup_{i=1}^{\infty} E_i\right) \cap F\right)}{\mathbb{P}(F)} && \text{applicata la def. di } \mathbb{P}(A|B)\text{; per la prop. distributiva, poi } \ldots \\
&= \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty}(E_i \cap F)\right)}{\mathbb{P}(F)} && \ldots \text{ma dato che si tratta di unione eventi disgiunti} \\
&= \frac{\sum_{i=1}^{\infty} \mathbb{P}(E_i \cap F)}{\mathbb{P}(F)} && \text{e portando il denominatore sotto sommatoria} \\
&= \sum_{i=1}^{\infty} \mathbb{P}(E_i|F)
\end{aligned}$$

□

Osservazione 73 (Notazione). A volte si vuole esprimere compattamente la probabilità condizionata di un evento E condizionata al verificarsi di un altro evento F. Per farlo definiamo

$$\widetilde{\mathrm{P}}(E) = \mathbb{P}(E|F)$$

Osservazione 74. Pertanto si ha che ogni probabilità condizionata è una probabilità. Allo stesso modo *tutte le probabilità possono essere pensate come probabilità condizionate*. Vi è sempre qualche informazione di fondo sulla quale condizioniamo anche se non esplicitata. Quando scriviamo pertanto $\mathbb{P}(A)$ stiamo pensando a $\mathbb{P}(A|K)$ con K background knowledge.

3.4.2.2 Risultati

Osservazione 75. Il fatto che, in seguito a 3.4.1, la probabilità condizionata sia una funzione di probabilità a tutti gli effetti, fa sì che tutti i risultati sviluppati in precedenza (per la probabilità non condizionata) valgano anche per la probabilità condizionata.
Possiamo aggiornare tutti i risultati visti in precedenza aggiungendo F a destra della barra di condizionamento. Ne mostriamo alcuni.

Lemma 3.4.2.

$$\widetilde{\mathrm{P}}\left(\overline{A}\right) = 1 - \widetilde{\mathrm{P}}(A) \tag{3.39}$$

Dimostrazione. Infatti

$$\begin{aligned} 1 - \widetilde{\mathrm{P}}(A) &= 1 - \mathbb{P}(A|F) = 1 - \frac{\mathbb{P}(A \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(F) - \mathbb{P}(A \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}\left(\overline{A} \cap F\right)}{\mathbb{P}(F)} \\ &= \mathbb{P}\left(\overline{A}|F\right) = \widetilde{\mathrm{P}}\left(\overline{A}\right) \end{aligned}$$

□

Lemma 3.4.3 (Probabilità dell'unione e principio di inclusione/esclusione)**.** *Si ha*

$$\widetilde{\mathrm{P}}(A \cup B) = \widetilde{\mathrm{P}}(A) + \widetilde{\mathrm{P}}(B) - \widetilde{\mathrm{P}}(A \cap B)$$

o equivalentemente

$$\mathbb{P}(A \cup B|F) = \mathbb{P}(A|F) + \mathbb{P}(B|F) - \mathbb{P}(A \cap B|F)$$

Lemma 3.4.4 (Condizionamento ulteriore)**.** *La probabilità condizionata $A|B$ dove B è un nuovo condizionamento e F è già presente/sottointeso si sviluppa come*

$$\widetilde{\mathrm{P}}(A|B) = \frac{\widetilde{\mathrm{P}}(A \cap B)}{\widetilde{\mathrm{P}}(B)} = \frac{\mathbb{P}(A \cap B|F)}{\mathbb{P}(B|F)} = \frac{\frac{\mathbb{P}(A \cap B \cap F)}{\mathbb{P}(F)}}{\frac{\mathbb{P}(B \cap F)}{\mathbb{P}(F)}} = \mathbb{P}(A|B \cap F)$$

Lemma 3.4.5 (Regola di Bayes con condizionamento ulteriore)**.** *A patto che $\mathbb{P}(A \cap F) > 0$ e $\mathbb{P}(B \cap F) > 0$ si ha*

$$\widetilde{\mathrm{P}}(A|B) = \frac{\widetilde{\mathrm{P}}(A) \cdot \widetilde{\mathrm{P}}(B|A)}{\widetilde{\mathrm{P}}(B)} = \frac{\mathbb{P}(A|F) \cdot \mathbb{P}(B|A \cap F)}{\mathbb{P}(B|F)}$$

Lemma 3.4.6 (Odds ratio con condizionamento ulteriore)**.** *Si ha:*

$$\frac{\widetilde{\mathrm{P}}(A|B)}{\widetilde{\mathrm{P}}\left(\overline{A}|B\right)} = \frac{\mathbb{P}(A|B \cap F)}{\mathbb{P}\left(\overline{A}|B \cap F\right)} = \frac{\mathbb{P}(A|F) \cdot \mathbb{P}(B|A \cap F)}{\mathbb{P}\left(\overline{A}|F\right) \cdot \mathbb{P}\left(B|\overline{A} \cap F\right)} \tag{3.40}$$

Lemma 3.4.7 (Teorema delle probabilità totali 1)**.** *La probabilità condizionata dell'evento E può essere spezzata come somma delle probabilità di eventi incompatibili, analogamente a quanto fatto in 3.28*

$$\widetilde{\mathrm{P}}(E) = \widetilde{\mathrm{P}}(C)\,\widetilde{\mathrm{P}}(E|C) + \widetilde{\mathrm{P}}\left(\overline{C}\right)\widetilde{\mathrm{P}}\left(E|\overline{C}\right)$$

ossia, equivalentemente

$$\mathbb{P}(E|F) = \mathbb{P}(C|F)\,\mathbb{P}(E|C\cap F) + \mathbb{P}\left(\overline{C}|F\right)\mathbb{P}\left(E|\overline{C}\cap F\right)$$

Lemma 3.4.8 (Teorema delle probabilità totali (versione generica))**.** *Se $C_1,\ldots,C_n$ è una partizione di Ω e nell'ipotesi che $\mathbb{P}(C_i\cap F)>0$ per ogni i, allora analogamente a 3.29 si ha*

$$\widetilde{\mathrm{P}}(E) = \mathbb{P}(E|F) = \sum_{i=1}^{n}\mathbb{P}(C_i|F)\cdot\mathbb{P}(E|C_i\cap F)$$

Esempio 3.4.2 (Moneta bilanciata 2)**.** Riprendendo l'esempio 3.2.7, supponiamo di aver visto la moneta uscire testa tre volte. Se la rilanciamo quale è la probabilità che esca testa una volta ancora?
Sia H l'evento testa tre volte, e T esce testa anche la quarta volta. Siamo interessati a $\mathbb{P}(T|H)$; la legge delle probabilità totali ci permette di scriverla come media ponderata dei condizionamenti su B (scelta la moneta bilanciata)

$$\begin{aligned}\mathbb{P}(T|H) &= \mathbb{P}(B|H)\,\mathbb{P}(T|B\cap H) + \mathbb{P}\left(\overline{B}|H\right)\mathbb{P}\left(T|\overline{B}\cap H\right)\\ &= 0.23\cdot\frac{1}{2} + (1-0.23)\cdot\frac{3}{4}\\ &\approx 0.69\end{aligned}$$

con $\mathbb{P}(B|H)=0.23$ come derivato in esempio 3.2.7.

3.4.2.3 Condizionare su più eventi

Spesso si vuole condizionare su più eventi/informazioni, ora abbiamo vari modi per farlo. Ipotizzando di essere interessati a $\mathbb{P}(A|B\cap C)$, ossia di voler condizionare a sia B che C:

- possiamo utilizzare la definizione di probabilità condizionata

$$\mathbb{P}(A|B\cap C) = \frac{\mathbb{P}(A\cap B\cap C)}{\mathbb{P}(B\cap C)}$$

- possiamo utilizzare la regola di Bayes condizionando ulteriormente su C (questo è l'approccio naturale se pensiamo che ogni evento nel nostro problema sia condizionato su C)

$$\mathbb{P}(A|B\cap C) = \frac{\mathbb{P}(A|C)\cdot\mathbb{P}(B|A\cap C)}{\mathbb{P}(B|C)}$$

- viceversa utilizzare la regola di Bayes condizionando ulteriormente su B (questo è l'approccio naturale se pensiamo che ogni evento nel nostro problema sia condizionato su B)

$$\mathbb{P}(A|B\cap C) = \frac{\mathbb{P}(A|B)\cdot\mathbb{P}(C|A\cap B)}{\mathbb{P}(C|B)}$$

3.4.2.4 Indipendenza condizionata, aggiornamento delle stime

Definizione 3.4.3 (Indipendenza condizionata)**.** Gli eventi A e B sono indipendenti condizionatamente dato l'evento F se

$$\mathbb{P}(A \cap B|F) = \mathbb{P}(A|F) \cdot \mathbb{P}(B|F) \tag{3.41}$$

Osservazione 76. Attenzione, due eventi:

- possono essere indipendenti condizionatamente (dato F), ma non indipendenti;
- possono essere indipendenti, ma non indipendenti condizionatamente (dato F);
- possono essere indipendenti condizionatamente dato F ma non dato $\overline{F}$.

Lo vediamo nei seguenti esempi.

Esempio 3.4.3 (Eventi indipendenti condizionatamente ma non indipendenti)**.** Tornando al setup di esempio 3.2.7, sia F "ho scelto la moneta bilanciata", A_1 "primo lancio da testa" e A_2 "secondo lancio da testa". Condizionatamente a F, A_1 e A_2 sono indipendenti; ma A_1 e A_2 non sono indipendenti da soli perché A_1 fornisce informazioni su A_2.

Esempio 3.4.4 (Eventi indipendenti ma non condizionatamente)**.** Siano Alice e Bob sono le uniche due persone che mi telefonano; ogni giorno decidono indipendentemente se farlo e sia A "mi chiama Alice", B "mi chiama Bob". Questi sono eventi indipendenti. Ma supponendo che R "il telefono squilla", condizionatamente a questo A e B non sono più indipendenti, perché se non è Alice deve essere Bob, ossia

$$\mathbb{P}(B|R) < 1 = \mathbb{P}\left(B|\overline{A} \cap R\right)$$

per cui B e $\overline{A}$ non sono condizionalmente indipendenti dato R (e allo stesso modo A e B)

Esempio 3.4.5. Supponendo che vi siano solo due tipi di classi: classi buone dove se si lavora tanto si prendono buoni voti e classi cattive dove il professore assegna voti a caso. Sia G "classe è buona", W "si lavora tanto" e A "si prende un bel voto". Allora W, A sono indipendenti condizionatamente a $\overline{G}$, ma non lo sono dato G.

Esempio 3.4.6 (Aggiornamento delle stime (e indipendenza condizionale))**.** Riprendendo l'esempio 3.4.1 sul test della malattia rara, ipotizziamo che il paziente decida di intraprendere un secondo test; questo è indipendente dal primo test effettuato (condizionatamente allo stato di malattia) e ha la stessa sensibilità e specificità. Il paziente risulta positivo per la seconda volta. Come si aggiorna la sua probabilità di essere effettivamente malato?
Siamo interessati a $\widetilde{\mathrm{P}}(D|T_2)$, condizionata a T_1, dove D è essere malato, T_1 è essere risultati positivi al primo test e T_2 al secondo. Utilizziamo la forma per

l'odds ratio per ricondurci in secondo luogo alla probabilità; si ha

$$\begin{aligned}\frac{\widetilde{\mathrm{P}}\left(D|T_2\right)}{\widetilde{\mathrm{P}}\left(\overline{D}|T_2\right)} &= \frac{\mathbb{P}\left(D|T_1\cap T_2\right)}{\mathbb{P}\left(\overline{D}|T_1\cap T_2\right)} = \frac{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1\cap T_2|D\right)}{\mathbb{P}\left(\overline{D}\right)\cdot\mathbb{P}\left(T_1\cap T_2|\overline{D}\right)} \\ &= \frac{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1|D\right)\cdot\mathbb{P}\left(T_2|D\right)}{\mathbb{P}\left(\overline{D}\right)\cdot\mathbb{P}\left(T_1|\overline{D}\right)\cdot\mathbb{P}\left(T_2|\overline{D}\right)} = \boxed{\frac{\mathbb{P}\left(D|T_1\right)}{\mathbb{P}\left(\overline{D}|T_1\right)}\cdot\frac{\mathbb{P}\left(T_2|D\right)}{\mathbb{P}\left(T_2|\overline{D}\right)}} \\ &= 0.19\cdot\frac{0.95}{0.05}\approx 3.646\end{aligned}$$

Di particolare interesse è la seconda riga dove, in contesto di indipendenza condizionata, si vede che aggiorniamo i risultati cui eravamo giunti in precedenza mediante le informazioni sul nuovo test. Passiamo alla probabilità seguendo la consueta formula

$$\mathbb{P}\left(D|T_1\cap T_2\right) = \frac{3.646}{1+3.646} = 0.78$$

La probabilità di essere malati in seguito ad un secondo test positivo (indipendente condizionalmente) aumenta molto, da 0.16 a 0.78.

Esempio 3.4.7 (Calcolo diretto della probabilità)**.** Volendo invece calcolare direttamente la probabilità in un colpo solo si applica Bayes e torna comodo il teorema delle probabilità totali condizionando su D:

$$\begin{aligned}\mathbb{P}\left(D|T_1\cap T_2\right) &= \frac{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1\cap T_2|D\right)}{\mathbb{P}\left(T_1\cap T_2\right)} \\ &= \frac{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1\cap T_2|D\right)}{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1\cap T_2|D\right)+\mathbb{P}\left(\overline{D}\right)\cdot\mathbb{P}\left(T_1\cap T_2|\overline{D}\right)} \\ &= \frac{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1|D\right)\cdot\mathbb{P}\left(T_2|D\right)}{\mathbb{P}\left(D\right)\cdot\mathbb{P}\left(T_1|D\right)\cdot\mathbb{P}\left(T_2|D\right)+\mathbb{P}\left(\overline{D}\right)\cdot\mathbb{P}\left(T_1|\overline{D}\right)\cdot\mathbb{P}\left(T_2|\overline{D}\right)} \\ &= \frac{0.01\cdot 0.95\cdot 0.95}{0.01\cdot 0.95\cdot 0.95+0.99\cdot 0.05\cdot 0.05} = 0.78\end{aligned}$$

Soffermandoci un attimo sulla equazione prima del calcolo dell'ultima riga, se dividiamo algebricamente per $\mathbb{P}\left(T_1\right)$ sia numeratore che denominatore si ottiene:

$$\begin{aligned}\mathbb{P}\left(D|T_1\cap T_2\right) &= \frac{\mathbb{P}\left(D|T_1\right)\cdot\mathbb{P}\left(T_2|D\right)}{\mathbb{P}\left(D|T_1\right)\cdot\mathbb{P}\left(T_2|D\right)+\mathbb{P}\left(\overline{D}|T_1\right)\cdot\mathbb{P}\left(T_2|\overline{D}\right)} \\ &= \frac{0.16\cdot 0.95}{0.16\cdot 0.95+0.84\cdot 0.05}\approx 0.78\end{aligned}$$

che equivale ad un normale teorema di Bayes dove al posto delle probabilità a priori secca $\mathbb{P}\left(D\right)$ che avevamo utilizzato in esempio 3.4.1, abbiamo sostituito i risultati disponibili alla fine del primo test, ossia $\mathbb{P}\left(D|T_1\right) = 0.16$ e $\mathbb{P}\left(\overline{D}|T_1\right) = 1-0.16 = 0.84$; come si nota l'unica cosa che cambia nella formula (anche perché T_1 e T_2 performano allo stesso modo), sono tali parti, evidenziate in rosso. Aggiorniamo dunque i risultati al termine del primo test con le informazioni del secondo test, per arrivare alla probabilità a posteriori $\mathbb{P}\left(D|T_1\cap T_2\right)$. Seguendo questa impostazione, è facile generalizzare ad n test applicando ripetutamente il teorema.

3.5 Esercizi vari

Esempio 3.5.1 (Es rigo)**.** Stai viaggiando su un treno con un amico. Nessuno di voi ha il biglietto e il controllore vi ha beccato. Il controllore è autorizzato a infliggervi una punizione molto particolare. Vi porge una scatola contenente 9 cioccolatini identici, 3 dei quali avvelenati. Vi costringe a sceglierne uno a testa, a turno, e mangiarlo immediatamente.

1. Se scegli prima del tuo amico, qual è la probabilità che tu sopravviva?
2. Se scegli per primo e sopravvivi, qual è la probabilità che anche il tuo amico sopravviva?
3. Se scegli per primo e muori, qual è la probabilità che il tuo amico sopravviva?
4. E' nel tuo interesse far scegliere prima al tuo amico?
5. Se scegli per primo, qual è la probabilità che tu sopravviva, tenendo conto del fatto che il tuo amico resti in vita?

Se A="primo cioccolatino scelto è non avvelenato", e B="secondo scelto non avvelenato"

1. $\mathbb{P}(A) = 6/9$
2. $\mathbb{P}(B|A) = 5/8$
3. $\mathbb{P}(B|A^c) = 6/8$
4. $\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B|A) + \mathbb{P}(A^c)\mathbb{P}(B|A^c) = \frac{6}{9}\frac{5}{8} + \frac{6}{8}\frac{3}{9} = \frac{6}{9}$ quindi non vi è vantaggio nello scegliere dopo il tuo amico
5. $\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)} = \ldots = \frac{5}{8}$; notiamo che $\mathbb{P}(A|B) = \mathbb{P}(B|A)$ in accordo con l'osservazione precedente, ossia che l'ordine della scelta non influenzi le probabilità di sopravvivenza

Esempio 3.5.2 (Rs rigo)**.** Un dado a sei facce non truccato viene lanciato due volte.

1. Scrivere lo spazio di probabilità dell'esperimento.
2. Supponiamo che B sia l'evento corrispondente al fatto che il risultato del primo lancio sia un numero non maggiore di 3, e supponiamo anche che C sia l'evento corrispondente al fatto che la somma dei due numeri ottenuti nei due lanci sia uguale a 6. Determinare le probabilità di B e C, e le probabilità condizionali di C dato B, e di B dato C.

Lo spazio di probabilità in questo esperimento è la tripla $(\Omega, \mathscr{A}, \mathbb{P})$, dove:

- $\Omega = \{(1,1), \ldots (6,6)\}$
- $\mathscr{A} = \mathcal{P}(\Omega)$
- ciascun punto in Ω ha uguale probabilità di successo, ossia $\mathbb{P}((i,j)) = 1/36$

Per il secondo punto:

- B= primo lancio $\leq 3 = \{(1,1), \ldots, (1,6), (2,1) \ldots, (2,6), (3,1) \ldots, (3,6)\}$ pertanto $\mathbb{P}(B) = \frac{18}{36}$
- C = somma $= 6 = \{(1,5), (5,1), (2,4), (4,2), (3,3)\}$, $\mathbb{P}(C) = \frac{5}{36}$
- si ha che $C \cap B = \{(1,5), (2,4), (3,3)\}$ quindi $\mathbb{P}(C|B) = \frac{\mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \frac{3/36}{18/36} = \frac{1}{6}$
- $\mathbb{P}(B|C) = \frac{3/36}{5/36} = \frac{3}{5}$

Capitolo 4

Random variables

4.1 Intro

Osservazione 77. A probability space is a particular measurable space.

Definizione 4.1.1 (Measurable space)**.** A pair $(S, \mathscr{B})$, where S is a set and $\mathscr{B}$ is a σ-field defined over the set.

Osservazione 78. What random variables do is to create a mapping between a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a measurable space $(S, \mathscr{B})$.

Definizione 4.1.2 (Random variable (rv))**.** A random variable is a *measurable* function $X : \Omega \to S$, that is, a function such that:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F}, \quad \forall B \in \mathscr{B} \tag{4.1}$$

Osservazione 79. That means that if I take any event of $\mathscr{B}$, there's a corresponding event in $\mathscr{F}$ that does produce it through X.

Osservazione 80. In practice in this course we will interested in random variable that perform mapping toward $(\mathbb{R}, \beta(\mathbb{R}))$, where X is called real or univariate random variable, and $(\mathbb{R}^n, \beta(\mathbb{R}^n))$ (X sometimes called n-variate random variable or n-dimensional random vector)

Osservazione 81 (Interpretation). The interpretation of rv is the following: one makes the experiment and see the resulting outcome $\omega \in \Omega$. Then after observing ω, $X(\omega)$ make a measurement on the outcome.

Esempio 4.1.1. If the experiment is to draw one person from a class, $\Omega = \{everyone\}$, $= \mathcal{P}(\Omega)$, while the random variable X could be height, so if Luca is extracted ($\omega =$ Luca), then $X(\text{Luca}) = 1.78$.
Distribution function ν of X is:

$$\nu(B) = \mathbb{P}(X \in B) = \mathbb{P}(\text{quelli di noi la cui altezza cade in } B)$$

Eg, if $B = (190, 195]$ and only Paolo and Francesca have an height such as that, then

$$\nu(B) = \mathbb{P}(\text{Paolo}) + \mathbb{P}(\text{Francesca})$$

Esempio 4.1.2 (Two coin throws)**.** Two coin throws can generate the following $\Omega = \{tt, th, ht, hh\}$. On this one we can define $X =$ "sum of heads as follows"

$$X(tt) = 2;\ X(th) = 1;\ X(ht) = 1;\ X(hh) = 0;$$

Osservazione 82. While the random variable is a deterministic mapping, the random part comes from the experiment.

Osservazione importante 19 (A new probability space). Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, a measurable space $(S, \mathscr{B})$, and a random variable $X : \Omega \to$ connecting the twos, we can define a further probability space $(S, \mathscr{B}, \nu)$, where the added probability function $\nu : \mathscr{B} \to [0, 1]$ is defined, using $\mathbb{P}$, in the following way:

$$\nu(B) = \mathbb{P}\left(X^{-1}(B)\right) = \mathbb{P}\left(X \in B\right), \quad \forall B \in \mathscr{B} \tag{4.2}$$

ν is called *probability distribution* of X.

Osservazione importante 20 (Motivation for measurability request). Suppose we don't require X to be measurable. If it's not then it can be that $\exists B \in \mathscr{B} : X^{-1}(B) \notin \mathscr{F}$ (there's an event of $\mathscr{B}$ with no corresponding event in $\mathscr{F}$). Well in that case $X^{-1}(B)$ does not belong to the domain of $\mathbb{P}$ and thus we cannot define/write $\nu(B) = \mathbb{P}\left(X^{-1}(B)\right) = \mathbb{P}\left(X \in B\right)$.
Therefore the need to define ν forces us to require X to be measurable.

Osservazione importante 21 (Notation). If we say:

- $X \sim \nu$ means that ν is the probability distribution of the rv X; for istance considering a real random variable $X : \Omega \to \mathbb{R}$, if we say $X \sim \mathrm{N}(0, 1)$ we are stating that probability distribution of X is standard normal;
- $X \sim Y$ means that X and Y have the same distribution (whatever it is).

Definizione 4.1.3 (Rv support)**.** It's the image $X(\Omega)$, the set of possible mappings, denoted by $R_X = \{x_1, x_2, \ldots\}$

Esempio 4.1.3. Regarding example 4.1.2, $R_X = \{0, 1, 2\}$.

4.1.1 Discrete and continuous rvs

Definizione 4.1.4 (Discrete rv)**.** Rv which cardinality of support is finite or numerable (1-to-1 with $\mathbb{N}$.)

Esempio 4.1.4. Head count in two coin throwing is discrete since $\mathrm{Card}\left(R_X\right) = |\{0, 1, 2\}| = 3$.

Definizione 4.1.5 (Continuous rv)**.** Rv which cardinality of support is not numerable (1-to-1 with $\mathbb{R}$).

Esempio 4.1.5. Numbers of minutes T of bulb lifetime is continue because $R_T = \{t \in \mathbb{R} : t > 0\}$

4.2 Functions of random variables

4.2.1 Discrete rvs: PMF, CDF

Definizione 4.2.1 (Probability mass function)**.** Given a rv $X : \Omega \to \mathbb{R}$, PMF is a function $p : \mathbb{R} \to \mathbb{R}$ taking the outcome of the rv and giving its probability

$$p_X(x) = \mathbb{P}(X = x) = \begin{cases} \mathbb{P}(X(s) = x) & \text{se } x \in X(\Omega) \\ 0 & \text{se } x \in \mathbb{R} \setminus X(\Omega) \end{cases} \tag{4.3}$$

Proposizione 4.2.1 (Valid PMF)**.** *If X is a discrete rv with support $X(\Omega) = \{x_1, x_2, \ldots\}$, a valid PMF p_X satisfies:*

$$p_X(x) \geq 0, \quad \forall x \in \mathbb{R} \tag{4.4}$$

$$\sum_{x \in \mathbb{R}} p_X(x) = 1 \tag{4.5}$$

Dimostrazione. Il primo criterio deve esser valido dato che la probabilità è non negativa. Il secondo deve essere valido dato che gli eventi $X = x_1, X = x_2, \ldots$ sono disgiunti e X dovrà assumere pur qualche valore:

$$\begin{aligned} \sum_{x \in \mathbb{R}} p_X(x) = \sum_{x \in X(\Omega)} p_X(x) = \sum_j \mathbb{P}(X = x_j) = \mathbb{P}\left(\bigcup_j \{X = x_j\}\right) \\ = \mathbb{P}(X = x_1 \text{ or } X = x_2 \ldots) = 1 \end{aligned}$$

□

Esempio 4.2.1. In two coins throwing 4.1.2

$$\begin{aligned} p_X(X = 0) &= 1/4 \\ p_X(X = 1) &= 1/2 \\ p_X(X = 2) &= 1/4 \end{aligned}$$

and $p_X(x) = 0$ for $x \notin \{0, 1, 2\}$.

Definizione 4.2.2 ((Cumulative) distribution function (CDF))**.** Given a discrete rv X its defined as:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_j \in X(\Omega) : x_j \leq x} p_X(x_j) \tag{4.6}$$

Osservazione 83 (Function shape). If X is discrete, $F_X(x)$ has starway shape with finite or numerable steps on values of the support $x_1, x_2, \ldots$: the step height is $p_X(x_1), p_X(x_2), \ldots$.

Proposizione 4.2.2 (Valid CDF)**.** *If X is a discrete rv with support $X(\Omega) = \{x_1, x_2, \ldots\}$, a valid CDF F_X must satisfy*

$$x_1 \leq x_2 \implies F_X(x_1) \leq F_X(x_2) \tag{4.7}$$

$$\lim_{x \to x_j^+} F_X(x) = F_X(x_j) \quad \textit{(right continuous)} \tag{4.8}$$

$$\lim_{x \to -\infty} F_X(x) = 0, \quad \lim_{x \to +\infty} F_X(x) = 1 \tag{4.9}$$

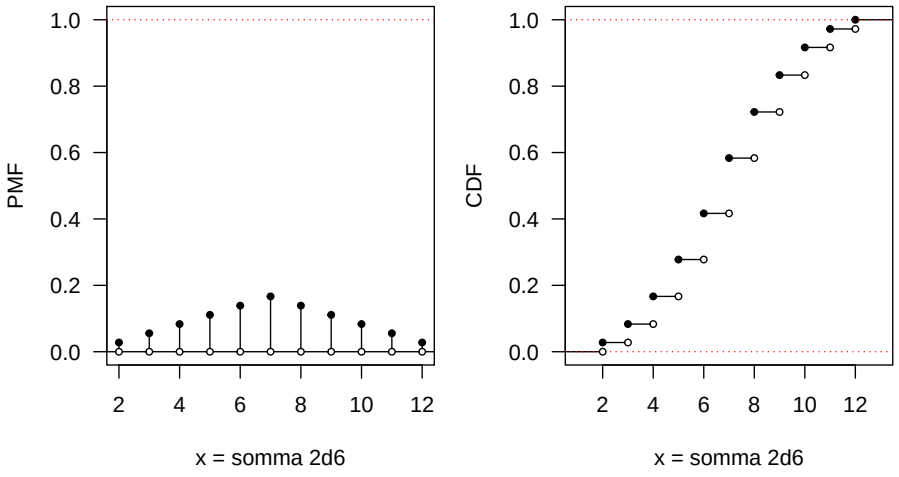


Figura 4.1: Somma del lancio di due d6

Dimostrazione. La prima è giustificata dal fatto che dato che, dato che l'evento $\{X \leq x_1\}$ si verifica sempre quando si verifica $\{X \leq x_2\}$ allora $\mathbb{P}(X \leq x_1) \leq \mathbb{P}(X \leq x_2)$.
La continuità da destra deriva dall'aver definito $F_X(x_0)$ come $\mathbb{P}(X \leq x_0)$ (coerentemente con la letteratura internazionale prevalente); altri autori definiscono $F_X(x_0) = \mathbb{P}(X < x_0)$, il che implica la continuità da sinistra.
Per la terza, dato che $F_X(x_{min}) = 0$ con $x_{min} = min(x_1, x_2, \ldots)$ e $-\infty < x_{min}$ allora per la prima proprietà si ha che $F(-\infty) \leq 0$, ma non potendo una probabilità esser negativa, sarà nulla, dunque si conclude che $\lim_{x \to -\infty} F_X(x) = 0$.
Altresì sfruttando sempre il fatto che $\{X = x_j\}$ sono eventi indipendenti

$$\lim_{x \to +\infty} F_X(x) = \sum_{x_j \in X(\Omega)} p_X(x_j) = 1$$

□

Esempio 4.2.2. Dato l'esperimento lancio di due dati, l'evento X somma degli esiti ha PMF e CMF riportate in figura 4.1. Ad esempio $\mathbb{P}(X = 2) = \mathbb{P}(\{1, 1\}) = (\frac{1}{6})^2 = 1/36 \approx 0.02778$. I "salti" nella CDF sono di entità pari alla PMF

4.2.2 Continuous rvs: PDF, CDF

Osservazione 84. PDF is the equivalent of PMF, CDF the same.

Definizione 4.2.3 ((Probability) density function (PDF))**.** If X is a continuous rv density is a $f : \mathbb{R} \to \mathbb{R}$, $f_X(x)$ such as, considered $X \in A \subseteq \mathbb{R}$:

$$\mathbb{P}(X \in A) = \int_{x \in A} f_X(x)\, \mathrm{d}x \tag{4.10}$$

Eg, if $a, b \in \mathbb{R}$, $a < b$:

$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)\,\mathrm{d}x \tag{4.11}$$

Proposizione 4.2.3 (Valid PDF)**.** *Must satisfy*

$$f_X(x) \geq 0 \tag{4.12}$$

$$\int_{-\infty}^{\infty} f_X(t)\,\mathrm{d}t = 1 \tag{4.13}$$

Dimostrazione. Il primo criterio è necessario perché la probabilità è non negativa: se $f_X(x_0)$ fosse negativa, allora potremmo integrare su un piccolo intorno di x_0 e ottenere una probabilità negativa.
Il secondo criterio è necessario dato che la X, variabile quantitativa, deve avere un esito che sta in $\mathbb{R}$. □

Osservazione 85. Differently from the discrete case (where PMF can't be more than 1) pdf can be more than 1, as long as integral sums on $\mathbb{R}$ sums up to 1.

Definizione 4.2.4 ((Cumulative) distribution function (CDF))**.** If X is a continuous rv, it's the function $F : \mathbb{R} \to \mathbb{R}$ defined as:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \tag{4.14}$$

Proposizione 4.2.4 (Valid CDF)**.** *It must satisfy*

$$x_1 \leq x_2 \implies F_X(x_1) \leq F_X(x_2) \tag{4.15}$$

$$\lim_{x \to x_0^+} F_X(x) = F_X(x_0) \quad \textit{(continuità da destra)} \tag{4.16}$$

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \lim_{x \to +\infty} F_X(x) = 1 \tag{4.17}$$

Osservazione 86 (Probability calculation with CDF). If we know CDF we can evaluate probability of an interval $a \leq X \leq b$, $a, b \in \mathbb{R}$ as follows:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a)$$

Osservazione 87 (Probability of a single value). A differenza delle variabili discrete, nel caso continuo si ha che:

$$\mathbb{P}(X = a) = \int_a^a f_X(x)\,\mathrm{d}x = F_X(a) - F_X(a) = 0$$

Intuitively, if there are infinite outcomes probability of each of them is null.

Osservazione 88 (Irrilevance of extremes of integration). For the same reason $a, b \in \mathbb{R}$, $a < b$:

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in (a, b]) = \mathbb{P}(X \in [a, b)) = \mathbb{P}(X \in (a, b)) = \int_a^b f_X(x)\,\mathrm{d}x$$

Esempio 4.2.3 (Logistic rv)**.** Logistic random variable, plotted in figure 4.2, is defined by:

$$F(x) = \frac{e^x}{1 + e^x}; \quad f(x) = \frac{e^x}{(1 + e^x)^2}$$

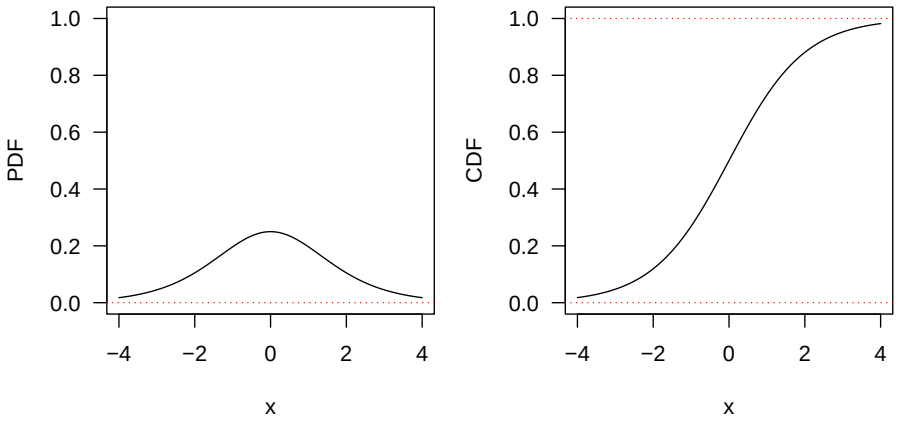


Figura 4.2: Logistic distribution

4.2.3 Distribution functions (Rigo's style)

Osservazione 89. In order to study random variables, an important concept is distribution function (which is the unifying one for continuous and discrete random variables); here we summarize/prove some results.

Osservazione importante 22 (Jargon). When it's sain distribution function we mean the cumulative distribution function.

Definizione 4.2.5 (Distribution function)**.** If X is a real valued rv, its distribution function is defined as

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x]) = \nu((-\infty, x]), \forall x \in \mathbb{R} \tag{4.18}$$

Proposizione 4.2.5 (Fundamental/characterizing properties)**.** *The properties* characterizing *distribution functions are*

1. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$,
2. *F is not decreasing: if $y > x$ then $F(y) \geq F(x)$;*
3. *F is right continuous* $F(x) = \lim_{y\to x^+} F(y)$, $\forall x \in \mathbb{R}$

Osservazione importante 23. This means that any function F which satisfies the three properties is a distribution function, that is, there exists a random variable X such that $F(x) = \mathbb{P}(X \leq x)$, $\forall x \in \mathbb{R}$.

Esempio 4.2.4 (Esempio crash course)**.** Let's check if

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x \geq 0 \end{cases}$$

is a distribution function. We have

1. $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = \lim_{x\to+\infty} 1 - e^{-x} = 1$, so check for the first

2. for $y > x$ we must show that $F(y) \geq F(x)$ to ensure non decreasing nature. Let's check the sign of $F(y) - F(x)$ (since if $F(y) - F(x) \geq 0$ then $F(y) \geq F(x)$): we have

$$1 - e^{-y} - 1 + e^{-x} = e^{-x} - e^{-y} \overset{(1)}{\geq} 0$$

with (1) since $e^{-y} < e^{-x}$ given that $y < x$

3. because $F(x)$ is continuous, it is also right continuous

So yes, $F(x)$ is a CDF ($X \sim \text{Exp}(1)$).

Proposizione 4.2.6. *Supposing we want to evaluate $\mathbb{P}(X = x)$, then the formula is*

$$\mathbb{P}(X = x) = F(x) - F(x^-) = F(x) - \lim_{y \to x^-} F(y), \qquad \textit{(jump of } F \textit{ at } x\textit{)} \tag{4.19}$$

with $y \to x$ from the left.

Dimostrazione. To prove this, recall (props 3.2.9 and 3.2.9) that for any probability measure $\mathbb{P}$

- if $A_1 \subseteq A_2 \subseteq \ldots$ is a increasing sequence of events, $\mathbb{P}(\cup_n A_n) = \lim_n \mathbb{P}(A_n)$
- if $A_1 \supseteq A_2 \supseteq \ldots$ is a decreasing sequence of events, $\mathbb{P}(\cap_n A_n) = \lim_n \mathbb{P}(A_n)$

Now suppose we want to evaluate

$$\mathbb{P}(X < x) = \mathbb{P}\left(\bigcup_{n=1}^{+\infty} \left\{X \leq x - \frac{1}{n}\right\}\right)$$

where we go nearer and nearer to x as n increases. These events are an increasing sequence of events, so

$$\begin{aligned}\mathbb{P}(X < x) = \mathbb{P}\left(\bigcup_{n=1}^{+\infty} \left\{X \leq x - \frac{1}{n}\right\}\right) &= \lim_{n \to +\infty} \mathbb{P}\left(X \leq x - \frac{1}{n}\right) = \lim_{n \to +\infty} F\left(X \leq x - \frac{1}{n}\right) \\ &= F(x^-)\end{aligned}$$

Finally in order to evaluate $\mathbb{P}(X = x)$ we have:

$$\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = F(x) - F(x^-)$$

□

Osservazione 90. As a consequence of this fact, the distribution function is *continuous* if and only if the jump is 0 at each point, or in other words $\mathbb{P}(X = x) = 0$, $\forall x \in \mathbb{R}$.

Osservazione importante 24. Considering the set $\{x \in R : \mathbb{P}(X = x) > 0\}$, this set is:

- empty, if the function is continuous
- its cardinality can be bounded from above: can at most be countable (eg Poisson, negative binomial); can be finite as well. Can't be uncountable.

Osservazione importante 25 (RV types). Real random variables can be *discrete*, *singular continuous* (we can ignore it) or *absolutely continuous*. Furthermore the following result is theorically important.

Proposizione 4.2.7. *If ν is any probability measure on $\beta\mathbb{R}$, the exists a unique triplets (a, b, c) such that:*

- $a, b, c \geq 0$
- $a + b + c = 1$
- $\nu = a\nu_1 + b\nu_2 + c\nu_3$ *where ν_1 is discrete, ν_2 is singular continuous, ν_3 is absolutely continuous*

Dimostrazione. We skip it. □

Osservazione 91. That is any ν can be written as this mix of this three kind of rv. Clearly, eg

$$a = 1, b = c = 0 \implies \nu = \nu_1 \text{is discrete} c = 1, a = b = 0 \implies \nu = \nu_3 \text{is absolutely continuous}$$

This is the reason to focuse on the three types, of which only discrete and absolutely continuous are of interest for practical applications.

Osservazione importante 26. In this course we speak indifferently like:

$$X \text{ is discrete} \iff \nu \text{ is discrete} \iff F \text{ is discrete}$$

Similarly for singular and absolutely continuous rv

4.2.3.1 Discrete rvs

Definizione 4.2.6 (Discrete rv)**.** X is discrete if and only if $\exists B \subset \mathbb{R}$, with B finite or countable such that $\mathbb{P}(X \in B) = 1$.

Esempio 4.2.5. If X is

- $\delta_(a)$, $B = \{a\}$
- binomial, then $B = \{0, 1, \ldots, n\}$;
- Poisson, $B = \{0, 1, \ldots\}$.

4.2.3.2 Singular continuous rvs

Osservazione 92. As we have said probability is a measure. In general

Definizione 4.2.7. A measure m is a function that, considered a set X:

$$m(X) \geq 0, \quad \forall X \tag{4.20}$$

$$X_i \cap X_j = \emptyset, \forall i \neq j \implies m\bigcup_{i=1}^{n} X_i = \sum_{i=1}^{n} mX_i \tag{4.21}$$

The latter being a numerable set of incompatible events.

Osservazione importante 27. The *Lebesgue measure* in $\mathbb{R}$ is the only measure on $\beta(\mathbb{R})$ that has this property, applied to an interval:

$$m(a,b] = b - a, \qquad \forall a < b \tag{4.22}$$

where m is the Lebesgue measure of the interval. Regarding the measure a point, countable and uncountable sets (the real line) Lebesgue measure

$$\begin{aligned} &m(\{x\}) = 0, \qquad \forall x \in \mathbb{R} \\ &m(X) = \sum_{x \in X} m(\{x\}) = \sum_{x \in X} 0 = 0 \qquad \forall X \subset \mathbb{R} : X \text{ is countable} \\ &m(\mathbb{R}) = +\infty \end{aligned}$$

Definizione 4.2.8 (Singular continuous rvs)**.** X is a singular continuous random variable if both

1. the distribution function F is continuous
2. his first derivative $F(x)' = 0$ *almost everywhere* with respect to the Lebesgue measure m, written "m - a.e.", that means

$$m(\{x \in \mathbb{R} : F'(x) \neq 0\}) = 0$$

Osservazione 93. I guess it can be taught as a sort of cardinality of the set: it has the same measure of a single point, so we mean that can have first derivative different from zero in very few points

Osservazione importante 28. First derivative fails to be 0 when:

1. it doesn't exists
2. exists but is not 0

For this kind of distribution it's not possible that distribution function is differentiable at everypoint and it's derivative is 0 at every point

Osservazione importante 29. For discrete rv effectively, $F' = 0$ m-a.e is true (think step F functions).

Osservazione 94. These seems to be a somewhat hybrid between discrete and absolutely continuous rv (since have characteristic from both the distribution), that is $F' = 0$ mae from the discrete, continuous F.

Osservazione importante 30. These variables are not usually used for describilng real phenomena, and we will not consider them here.

4.2.3.3 Absolutely continuous rvs

Esempio 4.2.6. eg exponential, beta, uniform, normal ...

Definizione 4.2.9 (Absolutely continuous rv)**.** X is absolutely continuous if and only if exists a function $f : \mathbb{R} \to \mathbb{R}$ such that:

1. $f \geq 0$
2. f is integrable

3. the distribution function valued at the point x is equal to the integral of function f
$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, \qquad \forall x \in \mathbb{R}$$

Osservazione importante 31. Some properties:

- Fs are written as integral only if X is absolutely continuous
- for this rvs, we have that $F' = f$ m.a.e, that is supposing we collect all the points where density doesn't equal the derivative of the distribution function, then
$$m(\{x \in \mathbb{R} : f(x) \neq F(x)\}) = 0$$
Therefore if f_1 and f_2 are both densities of X, can we say $f_1 = f_2$? Currently we have that $f_1 = F' = f_2$ m.a.e, hence
$$m(\{x \in \mathbb{R} : f_1(x) \neq f_2(x)\}) = 0$$
Namely, the density f is not necessary but almost everywhere unique.

Esempio 4.2.7. Consider $X \sim \mathrm{N}\,(()0, 1)$, a standard normal which is absolutely continuous with
$$f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$
if we define
$$g(x) = \begin{cases} f(x) & \text{if } x \in \mathbb{Q} \\ 1 + \sin(\log|x| + 3), & \text{if } x \notin \mathbb{Q} \end{cases}$$
$\mathbb{Q}$ has two properties: it's a countable set and it's dense ($\forall a, b \in \mathbb{Q}, \exists q \in \mathbb{Q}$ such that $a < q < b$). We have that
$$m(\{f \neq g\}) \leq \underbrace{m(\mathbb{Q})}_{=0}$$
that is the measure of the set where $f \neq g$ is ≤ 0 (so it's 0 since the measure is null or positive). Therefore the function f agrees with g m.a.e.
Hence f and g are both densities for X standard normal.

Teorema 4.2.8. *X is absolutely continuous if and only if, for every set $A \subset \beta(\mathbb{R})$ with lebesgue measure 0 ($m(A) = 0$) it's $\mathbb{P}(X \in A) = 0$*
$$\forall A \subset \beta(\mathbb{R}), m(A) = 0 \implies \mathbb{P}(X \in A) = 0$$

4.2.3.4 n-variate random variables

Let X be a n-variate random variable, X can be written as $X = (X_1, \ldots, X_n)^T$ or $X = \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix}$ indifferently (by default, vector are column vector), where $X_1, \ldots X_n$ are real random variables. We have that
$$\nu(B) = \mathbb{P}(X \in B), \qquad \forall B \in \beta(\mathbb{R}^n)$$

The distribution function $F : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) \tag{4.23}$$

Once again there's a 1-to-1 correspondance between F and ν expressed by

$$F(x_1, \ldots, x_n) = \nu((-\infty, x_1] \times \ldots \times (-\infty, x_n]), \qquad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

Again X is

- multivariate discrete if and only if $\exists B \subset \mathbb{R}^n$, B finite or countable such as that $\mathbb{P}(X \in B) = 1$
- multivariate absolutely continuous if and only if exists $f : \mathbb{R}^n \to \mathbb{R}$ such that
  1. $f \geq 0$
  2. f is integrable and

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\, \mathrm{d}t_1 \ldots\, \mathrm{d}t_n, \qquad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(t_1, \ldots, t_n)\, \mathrm{d}t_1 \ldots\, \mathrm{d}t_n = 1$$

Definizione 4.2.10 (Marginal)**.** A marginal of X is any subvector $(X_{j_1}, ..., X_{j_k})$ where $\{j_1, ..., j_k\}$ is a subset of $1, \ldots, n$

Osservazione 95. A marginal is just a vector with least random variables. There are

- n marginals of only 1 variable;
- $\binom{n}{2}$ marginals of 2 random variables:
- $\binom{n}{3}$ marginals of 3 random variables;

Teorema 4.2.9 (Density of marginal)**.** *If X is absolutely continuous then, then every marginal of X is still absolutely continuous; moreover if f is the multivariate density of X the density g of the marginal $(X_1, ldotsx_k)^t$ is*

$$g(x_1, \ldots, x_k) = \underbrace{\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty}}_{n-k \text{ integrals}} f(t_1, \ldots, t_n)\, \mathrm{d}t_{k+1}\, \mathrm{d}t_n \tag{4.24}$$

Osservazione 96. That is g is obtained making $n - k$ integral, by integrating out all the variable that you want to eliminate (in this case it were $n - k$).

Esempio 4.2.8. If $n = 3$, $X = (X_1, X_2, X_3)^T$, densitity of X_2 is

$$g(y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y, z)\, \mathrm{d}x\, \mathrm{d}z \tag{4.25}$$

Similarly the density of (X_2, X_3) is

$$g(y, z) = \int_{-\infty}^{+\infty} f(x, y, z)\, \mathrm{d}x$$

Definizione 4.2.11 (Lebesgue measure on $\mathbb{R}^n$)**.** It's the only measure on $\beta\mathbb{R}^n$ such that the measure of the cartesian product of interval is equal to the product of the length of the intervals

$$m(I_1 \times \ldots \times I_n) = l(I_1) \cdot \ldots \cdot l(I_n), \quad \forall I_i$$

where $l(I_i)$ is the legth of the interval I_i.

Esempio 4.2.9. Intuitively, if $A \in \beta(\mathbb{R}^2)$ is a borel set, then $m(A)$ is the area of A; in $\beta(\mathbb{R}^3)$ is a volume and on

Osservazione 97. Now we extend to random vector a theorem (already discussed in the $n = 1$ case) useful for proving that X is absolutely continuous.

Teorema 4.2.10. *A random vector X is absolutely continuous if and only if*

$$\mathbb{P}(X \in A) = 0, \forall A \in \beta(\mathbb{R}^n), \quad stm(A) \tag{4.26}$$

Teorema 4.2.11. *If $X_1, ...X_n$ are absolutely continuous, this does not imply that the vector X is absolutely continuous.*

Esempio 4.2.10. It may be that X is not absolutely continuous even if $X_1, ..., X_n$ are. An example of this follows.
Let $n = 2$, $X_1 \sim \mathrm{N}(()0, 1)$, $X_2 = X_1$; now X_1 is absolutely continuos because its a standard normal, X_2 is equal. Is X is absolutely continuous? We apply the theorem.
To check that X is not absolutely continuous we let

$$A = \left\{(x, y) \in \mathbb{R}^2 : x = y\right\} \quad \text{(it's the diagonal } y = x)$$

We have that $\mathbb{P}(X \in A) = 1$: infact once we extracted $X_1 = x_1$ we have $X_2 = x_1$ as well so thevector will be on the diagonal $y = x$; however we have that $m(A) = 0$ (that is the area of the line $y = x$ compared to 2 space dimension $\mathbb{R}^2$ is 0).

Teorema 4.2.12 (Independence)**.** *Let $X = (X_1, ..., X_n)$ be* any *random vector. Then:*

1. *$X_1, .., X_n$ are indipendent if and only if the joint distribution function is the product of the marginal distribution functions*

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n), \quad \forall \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix} \in \mathbb{R}^n \tag{4.27}$$

   where F_i is the marginal for X_i

2. *if X is* absolutely continuous *we can replace distribution functions with densities; therefore X is composed of independents random variables if and only if*

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdot \ldots \cdot f_n(x_n), \quad \forall \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix} \in \mathbb{R}^n \tag{4.28}$$

3. *finally if $(X_1, ..., X_n)$ are independent, then X is absolutely continuous if and only if the 1-dimensional marginal are absolutely continuous*

4.3 Other useful rv functions

4.3.1 Support indicator

Osservazione 98. Nel seguito servirà essere compatti/sicuri sul fatto che, al di fuori del supporto R_X della vc X, la probabilità/densità sia nulla. Per farlo si moltiplicherà la PMF/PDF per la funzione indicatrice applicata al supporto della variabile casuale.

Definizione 4.3.1 (Funzione indicatrice del supporto di una vc)**.** Definita come:

$$\mathbb{1}_{R_X}(x) = \begin{cases} 1 & \text{se } x \in R_X \\ 0 & \text{se } x \notin R_X \end{cases}$$

4.3.2 Survival and hazard function

Osservazione 99. If rv T has non negative support (eg lifetime), then two function are useful (survival for both discrete and continuous rvs, hazard for continuous)

Definizione 4.3.2 (Survival function)**.** Given a rv T such as $\mathbb{P}(T \geq 0) = 1$, it's defined as complement to 1 of cumulative distribution function

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F_T(t) \tag{4.29}$$

Definizione 4.3.3 (Funzione di azzardo (o rischio))**.** Given a continuous rv T such as $\mathbb{P}(T \geq 0) = 1$, hazard function is defined as

$$H(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{\mathrm{d}t} \log(1 - F_T(t)) = -\frac{d}{\mathrm{d}t} \log(S(t)) \tag{4.30}$$

Osservazione 100. Hazard function can be interpreted as the probability that T stops at t given that it arrived to t

Osservazione 101. Relationship between Hazard, survival, density and distribution function can be retrieved by the equation. Eg integrating both members between tra $-\infty$ and x we have

$$\begin{aligned} H(t) &= -\frac{d}{\mathrm{d}t} \log(S(t)) \\ \int_{-\infty}^{x} H(t)\,\mathrm{d}t &= \int_{-\infty}^{x} -\frac{d}{\mathrm{d}t} \log(S(t)) \\ \int_{-\infty}^{x} H(t)\,\mathrm{d}t &= -\log(S(t)) \end{aligned}$$

Therefore:

$$\begin{aligned} \log(S(t)) &= -\int_{-\infty}^{x} H(t)\,\mathrm{d}t \\ S(t) &= \exp\left(-\int_{-\infty}^{x} H(t)\,\mathrm{d}t\right) \end{aligned} \tag{4.31}$$

X	$\mathbb{P}(X=x)$	$Y=2X$	$\mathbb{P}(Y=y)$	$Z=X^2$	$\mathbb{P}(Z=z)$
-1	0.33	-2	0.33	1	0.66
0	0.33	0	0.33	0	0.33
1	0.33	2	0.33		

Tabella 4.1: PMF of discrete rv transform, an example

While for what concerns $F_T(t)$ e $f_T(t)$ we have:

$$F_T(t) = 1 - \exp\left(-\int_{-\infty}^{x} H(t)\,\mathrm{d}t\right) \tag{4.32}$$

$$f_T(t) = H(t) \cdot \exp\left(-\int_{-\infty}^{x} H(t)\,\mathrm{d}t\right) \tag{4.33}$$

Btw, in the lower limit of integration we could have write 0 instead of $-\infty$.

4.4 Transformation of rvs

Definizione 4.4.1 (Trasform of rv $g(X)$)**.** Considered an experiment with sample space Ω, a random variable X on it and a function $g : \mathbb{R} \to \mathbb{R}$, then $g(X)$ is the random variable mapping $\omega \to g(X(\omega))$, $\forall \omega \in \Omega$ and having support $R_{g(X)} = \{g(X(\omega 1)), g(X(\omega 2)), \ldots\}$.

4.4.1 Discrete rv transform

Osservazione 102. Given a discrete rv X with known PMF, how to get PMF of $Y = g(X)$? If:

- g è *injective*, $X(s_1) \neq X(s_2) \implies g(X(s_1)) \neq g(X(s_2))$, then PMF Y will be the same of X:

$$\mathbb{P}(Y = g(x)) = \mathbb{P}(g(X) = g(x)) = \mathbb{P}(X = x)$$

- otherwise there could be cases where $X(s_1) \neq X(s_2)$ but $\implies g(X(s_1)) = g(X(s_2))$: here we have to sum probability of different x that with g ends in the same y.

The following result is general and is ok for both cases

Proposizione 4.4.1 (PMF of $g(X)$)**.** *Let X be a discrete rv and $g : \mathbb{R} \to \mathbb{R}$. Then support of $g(X)$ is the set of y such as that $g(x) = y$ for at least one $x \in R_X$ and PMF of $g(X)$ is*

$$\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x), \qquad \forall y \in R_{g(X)} \tag{4.34}$$

Esempio 4.4.1. In table 4.1 an example with X, $Y = 2X$ ($g(x) = 2 \cdot x$, inctive) e $Z = X^2$ ($g(x) = x^2$ not injective).

Osservazione 103. It's a common error to apply g to the PMF (it could take probability over 1): g have to be applied to domain/support of PMF.

4.4.2 Continuous rvs transform (linear case)

Definizione 4.4.2 (Scale-location transform for continuous rv)**.** Let X be a continuous rv; $Y = \sigma X + \mu$ with $\sigma, \mu \in \mathbb{R}$ is a random variable obtained using a (linear) transform of both position and scale.

Osservazione 104. Here σ set the scale (if positive spread Y compared to X) while μ the location (if positive moves Y distribution toward right compared to X).

Osservazione 105. In order to go back to X we standardize Y, aka apply the transformation $X = \dfrac{Y - \mu}{\sigma}$.

Proposizione 4.4.2. *Y has the same family of distribution as X.*

Dimostrazione. It has been obtained by a linear, injective transformation. □

Osservazione 106. If this kind of transformation is applied to a discrete rv we have a distribution no more of the same family, considered that support changes (eg linear transform of a binomial does not give a binomial, defined on support $0, 1, \ldots$).

4.5 Rvs independence

Osservazione 107. It's similar to events independence.

4.5.1 Independence, iid rvs

Definizione 4.5.1 (Indipendenza di 2 vc, $X \perp\!\!\!\perp Y$)**.** Two rvs X, Y are independent, and we write $X \perp\!\!\!\perp Y$, if

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y), \qquad \forall x, y \in \mathbb{R} \tag{4.35}$$

Osservazione 108 (Notation). $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x \cap Y \leq y)$

Osservazione 109. In the discrete case 4.35 is equivalent to

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y), \qquad \forall x, y \in \mathbb{R}$$

Esempio 4.5.1. Let be X the result of first dice thrown and Y the second; sum and difference of results random variables $X + Y$, $X - Y$ are not independent considered that:

$$\mathbb{P}(X + Y = 12, X - Y = 1) = 0$$
$$\mathbb{P}(X + Y = 12) \cdot \mathbb{P}(X - Y = 1) = \frac{1}{6} \cdot \frac{5}{6}$$

This does make sense: knowing that the sum is 12, tells that their difference must be 0 so the two rv gives information of each other

Proposizione 4.5.1 (Transform of independent rv)**.** *If X and Y are independent, then any transformation of X and Y are independent as well.*

Dimostrazione. Not shown. □

Definizione 4.5.2 (rvs independence (general case))**.** Given *any* collection (finite, countable non countable) of random variables $\mathcal{V} = \{X_1, X_2, \ldots\}$, the elements of $\mathcal{V}$ are said to be independent if, for any *finite* subset of events $\mathcal{X} \subset \mathcal{V}$, with $\text{Card}(\mathcal{X}) = n < \text{Card}(\mathcal{V})$

$$\begin{aligned} &\mathbb{P}(X_j \leq x_j, \ldots, X_k \leq x_k) = \mathbb{P}(X_j \leq x_j) \cdot \ldots \cdot \mathbb{P}(X_k \leq x_k) \\ &X_j, \ldots, X_k \in \mathcal{X}, \quad \forall x_j, \ldots, x_k \in \mathbb{R} \end{aligned} \tag{4.36}$$

or equivalently with rigo's notation

$$\begin{aligned} &\mathbb{P}(X_j \in B_j, \ldots, X_k \in B_k) = \mathbb{P}(X_j \in B_j) \cdot \ldots \cdot \mathbb{P}(X_k \in B_k) \\ &X_j, \ldots, X_k \in \mathcal{X} \quad \forall B_j, \ldots, B_k \in \mathscr{B} \end{aligned}$$

Proposizione 4.5.2. *If $X_1, \ldots, X_n$ are independent, then they are pairwise, 3-3, $\ldots (n-1)$-$(n-1)$ independent. Viceversa does not apply.*

Dimostrazione. If $X_1, \ldots, X_n$ are independent si ha (considerando a titolo di esempio la coppia X_1, X_2) che

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbb{P}(X_1 \leq x_1) \cdot \mathbb{P}(X_2 \leq x_2)$$

Per vedere perché sia così basta far tendere a $+\infty$ gli $x_3, \ldots, x_n$ in maniera tale che a sinistra dell'uguale, nella definizione 4.36, entro parentesi si abbiano eventi certi e a destra dell'uguale si moltiplichi per 1. □

Definizione 4.5.3 (i.i.d. rvs)**.** Random variables that are *independent* and *identically* distributed (same CDF).

Osservazione importante 32. If the elements of $\mathcal{X} = \{X_1, X_2, \ldots\}$ are iid, to communicate the common distribution of the X_i it suffices to write $X_i \sim \nu$

4.5.2 Conditional independence

Definizione 4.5.4 (Conditional independence)**.** X and Y are conditional independent given Z if $\forall x, y \in \mathbb{R}$ and $\forall z \in R_Z$ it is:

$$\mathbb{P}(X \leq x, Y \leq y | Z = z) = \mathbb{P}(X \leq x | Z = z) \cdot \mathbb{P}(Y \leq y | Z = z) \tag{4.37}$$

Osservazione 110. For discrete rvs, an equivalent definition based on the mass function is

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z) \cdot \mathbb{P}(Y = y | Z = z) \tag{4.38}$$

Proposizione 4.5.3. *Rvs indipendence does not imply conditional independence and viceversa.*

Dimostrazione. By counterexamples, see Blitzstein pag 121. □

4.6 Moments

Osservazione 111. Distribution functions are the unifying concepts for continuous and discrete rvs; furthermore knowing F_X is to know the entire probabilistic structure of the rv.
In order to compare different rv, however, often synthetic indicator are needed and these are the moments.

Definizione 4.6.1 (Moment of a rv)**.** A statistic of this kind, if it exists

$$\begin{cases} \sum_{i=1}^{\infty} g(x_i) \cdot p_X(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} g(x) \cdot f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

Different g functions defines different moments

Osservazione importante 33 (Moment existence). We here suppose that integrals/series converges and therefore the moment exist; not all random variable have moments

Osservazione importante 34 (Important moments). These are expected value, variance, asymmetry and kurtosis.

Proposizione 4.6.1. *Se X Y then they have the same moments.*

Dimostrazione. This comes from the fact that in moments definition we use only the distribution that, if equal, will conduct to same results. □

4.6.1 Expected value

Definizione 4.6.2 (Moment of order r (r-th moment) of X)**.** Where g is the r-power of X:

$$\mu_r = \mathbb{E}[X^r] = \begin{cases} \sum_i {x_i}^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} x^r \cdot f_X(x)\, \mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{4.39}$$

Definizione 4.6.3. In general we say that X has the moment of order r if $\mathbb{E}[|X|^r] < +\infty$.

Teorema 4.6.2. *If $\mathbb{E}[|X|^r] < +\infty$ for some $r > 0$, then all the moments of order $q < r$ exists/are finite:*

$$\mathbb{E}[|X|^q] < +\infty, \forall q \in (0, r]$$

Definizione 4.6.4 (Expected value)**.** First moment of X, denoted by $\mathbb{E}[X]$ or μ: gives a probability weighted mean of X:

$$\mathbb{E}[X] = \begin{cases} \sum_i x_i \cdot p_X(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x \cdot f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases} \tag{4.40}$$

Esempio 4.6.1 (Single dice)**.** Let X be the result of a single fair dice with $p_X(1) = \ldots = p_X(6) = 1/6$:

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$$

Osservazione importante 35. We are not sure that series or integrals of the definition above exists. Wheter the random variable is discrete or continuous in order to check it we need previously to evaluate the expectation of the absolute value of the random variable, that is

$$\mathbb{E}\left[|X|\right] = \int_0^{+\infty} \mathbb{P}\left(|X| \geq t\right) \mathrm{d}t$$

There are two possible situation; if this integral:

1. is infinite: then $\mathbb{E}\left[X\right]$ does not exist and we stop;
2. is finite ($< \infty$): the mean exists and can be evaluated through the following formula, distinguishing by type of variable

Esempio 4.6.2. For the Cauchy random variable, the expected value does not exists. If $X \sim Cauchy$, X is absolutely continuous with density

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

In order to check it, we start evaluating the test for expected value existence

$$\mathbb{E}\left[|X|\right] = \int_0^{+\infty} \mathbb{P}\left(|X| > t\right) \mathrm{d}t \overset{(1)}{=} \int_{-\infty}^{+\infty} |x| \cdot \frac{1}{\pi}\frac{1}{1+x^2} \overset{(2)}{=} \frac{2}{\pi}\int_0^{+\infty} \frac{x}{1+x^2}\, \mathrm{d}x$$

where (1) take it as given (we don't prove it), (2) because it's an even function (symmetryc with respect to y axis) so we can double the integral on the positive part (taking x out of absolute value). Integrating by parts we have:

$$\int \frac{x}{1+x^2}\, \mathrm{d}x = \frac{1}{2}\log(x^2+1) + c$$

Therefore

$$\mathbb{E}\left[|X|\right] = \frac{2}{\pi}\left(\left[\frac{1}{2}\log(x^2+1)\right]_0^{+\infty}\right) = \frac{2}{\pi}(+\infty - 0) = +\infty$$

Therefore the expected value does not exists.

Proposizione 4.6.3 (Expected value properties)**.**

$$\mathbb{E}\left[aX+b\right] = a\,\mathbb{E}\left[X\right] + b \tag{4.41}$$
$$\mathbb{E}\left[X+Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right] \tag{4.42}$$
$$X \geq 0 \implies \mathbb{E}\left[X\right] \geq 0 \tag{4.43}$$
$$X \geq 0, \mathbb{P}\left(X > 0\right) > 0 \implies \mathbb{E}\left[X\right] > 0 \tag{4.44}$$
$$\mathbb{E}\left[g(X)\right] = \sum_i g(x_i) \cdot p_X(x_i) \tag{4.45}$$
$$X \perp\!\!\!\perp Y \implies \mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{4.46}$$
$$\min\left(X\right) \leq \mathbb{E}\left[X\right] \leq \max\left(X\right) \tag{4.47}$$
$$\mathbb{E}\left[X - \mathbb{E}\left[X\right]\right] = 0 \tag{4.48}$$
$$\textit{minimizes } \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \tag{4.49}$$

Osservazione 112. Congiuntamente alle 4.41 e 4.42 ci si riferisce come linearità del valore atteso, che torna spesso comodo per il calcolo soprattutto se si riesce a scrivere una vc come somma di due o più vc. La linearità è un mero fatto algebrico e di bello c'è che, ad esempio per 4.42, non è necessaria l'indipendenza tra X e Y affinché valga.

Osservazione importante 36. If $f : \mathbb{R} \to \mathbb{R}$ is a measurable function, to evaluate the expectation of $f(X)$, that is $E(f(X))$, we can repeat the previous properties with $f(X)$ instead of X.

TODO: da chiarire sta nota di colore

Dimostrazione. Mostriamo con riferimento alle variabili discrete. Per la 4.41

$$\begin{aligned}\mathbb{E}[aX+b] &= \sum_i (ax_i+b)\cdot\mathbb{P}(aX+b=ax_i+b) = \sum_i (ax_i+b)\cdot\mathbb{P}(X=x_i)\\ &= \sum_i ax_i\cdot\mathbb{P}(X=x) + \sum_i b\cdot\mathbb{P}(X=x)\\ &= a\sum_i x_i\cdot\mathbb{P}(X=x) + b\underbrace{\sum_i \mathbb{P}(X=x)}_{1}\\ &= a\,\mathbb{E}[X]+b\end{aligned}$$

Viceversa nel caso continuo

$$\mathbb{E}[aX+b] = \int_{D_x}(ax+b)f(x)\,\mathrm{d}x = a\int_{D_x} xf(x)\,\mathrm{d}x + b\underbrace{\int_{D_x} f(x)\,\mathrm{d}x}_{=1} = a\,\mathbb{E}[X]+b$$

Per 4.42 facendo un passo indietro, possiamo scrivere un generico valore atteso facendo riferimento all'evento $s \in \Omega$ e applicando la funzione X ad esso, al fine di ottenere x_i:

$$\mathbb{E}[X] = \sum_i x_i\cdot\mathbb{P}(X=x_i) = \sum_s X(s)\cdot\mathbb{P}(\{s\})$$

Da questa possiamo generalizzare alla somma di due funzioni:

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_s (X+Y)(s)\cdot\mathbb{P}(\{s\}) = \sum_s (X(s)+Y(s))\cdot\mathbb{P}(\{s\})\\ &= \sum_s X(s)\cdot\mathbb{P}(\{s\}) + \sum_s Y(s)\cdot\mathbb{P}(\{s\})\\ &= \mathbb{E}[X]+\mathbb{E}[Y]\end{aligned}$$

Per il valore atteso della trasformazione g, 4.45, sfruttiamo la stessa tecnica facendo un passo indietro (rispetto all'applicazione della funzione X agli eventi dello spazio campionario): sia $s \in \Omega$ un evento dello spazio campionario e X la vc considerata. Come detto possiamo scrivere il valore atteso $\mathbb{E}[X]$ come prodotto del risultato di X per la probabilità che si verifichi quell'evento:

$$\mathbb{E}[X] = \sum_s X(s)\,\mathbb{P}(\{s\})$$

L'applicazione della trasformazione g porta il valore atteso $\mathbb{E}[g(X)]$:

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_s g(X(s)) \cdot \mathbb{P}(\{s\}) \\ &\overset{(1)}{=} \sum_i \sum_{s:X(s)=x_i} g(X(s))\,\mathbb{P}(\{s\}) \\ &= \sum_i g(x_i) \sum_{s:X(s)=x_i} \mathbb{P}(\{s\}) \\ &= \sum_i g(x_i) \cdot \mathbb{P}(X = x_i) \\ &= \sum_i g(x_i) \cdot p_X(x_i)\end{aligned}$$

dove in (1) semplicemente raggruppiamo per i diversi s che attraverso X forniscono lo stesso x_i.
Per 4.46 (facendo mostrando il caso delle discrete) se $X \perp\!\!\!\perp Y$, allora $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$, da questo

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x\in D_x} \sum_{y\in D_y} x \cdot y \cdot \mathbb{P}(X = x, Y = y) = \sum_{x\in D_x} \sum_{y\in D_y} x \cdot y \cdot \mathbb{P}(X = x)\,\mathbb{P}(Y = y) \\ &= \sum_{x\in D_x} x \cdot \mathbb{P}(X = x) \sum_{y\in D_y} y \cdot \mathbb{P}(Y = y) = \mathbb{E}[X] \cdot \mathbb{E}[Y]\end{aligned}$$

La 4.47 è ovvia essendo $\mathbb{E}[X]$ una media pesata da probabilità dei valori assunti da X; l'uguaglianza vale in caso di variabili degeneri.
La 4.48 è una applicazione della linearità

$$\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

□

Esempio 4.6.3 (Valore atteso di trasformazione)**.** Supponiamo che X sia una vc che assuma i valori $-1, 0, 1$ con probabilità pari a $\mathbb{P}(x = -1) = 0.2$, $\mathbb{P}(x = 0) = 0.5$, $\mathbb{P}(x = 1) = 0.3$. Calcoliamo $\mathbb{E}\left[X^2\right]$ applicando prima la trasformazione e poi moltiplicando per la probabilità:

$$\mathbb{E}\left[X^2\right] = (-1)^2(0.2) + 0^2 \cdot (0.5) + 1^2(0.3) = 0.5$$

Proposizione 4.6.4 (Valore atteso di funzioni non lineari di vc)**.** *In generale non vale* $\mathbb{E}[g(X)] = g\,\mathbb{E}[X]$ *per una qualsiasi funzione* g*.*

Esempio 4.6.4. Sia X il lancio di un dado: calcoliamo $\exp(\mathbb{E}[X])$ e $\mathbb{E}[\exp X]$; ricordando che $\mathbb{E}[X] = 7/2$ si ha

$$\begin{aligned}g(\mathbb{E}[X]) &= \exp(7/2) \approx 33.12 \\ \mathbb{E}[g(X)] &= e^1 \cdot \frac{1}{6} + \ldots + e^6 \cdot \frac{1}{6} \approx 106.1\end{aligned}$$

Considerando invece una trasformazione lineare $g(x) = 2x + 1$ i due risultati coincidono, come in mostrato 4.41. Si ha:

$$\begin{aligned}g(\mathbb{E}[X]) &= 2 \cdot \frac{7}{2} + 1 = 8 \\ \mathbb{E}[g(X)] &= 3\frac{1}{6} + 5\frac{1}{6} + 7\frac{1}{6} + 9\frac{1}{6} + 11\frac{1}{6} + 13\frac{1}{6} = 8\end{aligned}$$

4.6.2 Variance

Definizione 4.6.5 (r-th moments of X with respect to mean)**.** We have them if $g = (x - \mathbb{E}[X])^r$:

$$\bar{\mu_r} = \mathbb{E}[(X - \mathbb{E}[X])^r] = \begin{cases} \sum_i (x_i - \mathbb{E}[X])^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^r \cdot f_X(x)\,\mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{4.50}$$

Osservazione 113. Since $\bar{\mu_0} = 1, \bar{\mu_1} = 0$, these moments become interesting starting from $r = 2$.

Definizione 4.6.6 (Variance)**.** If $\mathbb{E}\left[X^2\right] < +\infty$ (here absolute value is superfluous), we can define the variance of X as

$$\bar{\mu_2} = \mathrm{Var}[X] = \sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{4.51}$$

measure dispersion of the rv around its mean value.

Proposizione 4.6.5 (Formula to use for evaluation)**.**

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \tag{4.52}$$

Dimostrazione. We have:

$$\begin{aligned} \mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \sum_i (x_i - \mathbb{E}[X])^2 \cdot p_X(x_i) = \sum_i (x_i^2 - 2\,\mathbb{E}[X]\,x_i + \mathbb{E}[X]^2) \cdot p_X(x_i) \\ &= \sum_i x_i^2 \cdot p_X(x_i) - 2\,\mathbb{E}[X] \sum_i x_i \cdot p_X(x_i) + \mathbb{E}[X]^2 \sum_i p_X(x_i) \\ &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \end{aligned}$$

Otherwise we could have expandad $(X - \mathbb{E}[X])^2$ and used expected value linearity:

$$\begin{aligned} \mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \mathbb{E}\left[X^2 - 2\,\mathbb{E}[X]\,X + \mathbb{E}[X]^2\right] = \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - [\mathbb{E}[X]]^2 \end{aligned}$$

□

Esempio 4.6.5 (Dice variance)**.** If X is result of a dice throw, previously we computed $\mathbb{E}[X] = 7/2$; furthermore we have

$$\mathbb{E}\left[X^2\right] = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \left(\frac{1}{6}\right)(91)$$

Therefore

$$\mathrm{Var}[X] = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Proposizione 4.6.6 (Properties of variance)**.** *Given $a, b, c \in \mathbb{R}$:*

$$\operatorname{Var}[X] \geq 0 \tag{4.53}$$
$$\operatorname{Var}[X] = 0 \iff \mathbb{P}(X = c) = 1 \tag{4.54}$$
$$\operatorname{Var}[aX + b] = a^2 \operatorname{Var}[X] \tag{4.55}$$
$$X \perp\!\!\!\perp Y \implies \operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] \tag{4.56}$$

Dimostrazione. Per la 4.53, la varianza è il valore atteso della vc nonnegativa $(X - \mathbb{E}[X])^2$, motivo per cui è non negativa date le proprietà del valore atteso. Per 4.54 se $\mathbb{P}(X = c) = 1$ per qualche costante c allora $\mathbb{E}[X] = c$ e $\mathbb{E}\left[X^2\right] = c^2$, pertanto $\operatorname{Var}[X] = 0$; viceversa se $\operatorname{Var}[X] = 0$ allora $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = 0$ che mostra che $(X - \mathbb{E}[X])^2 = 0$ ha probabilità 1, che a sua volta mostra che X è uguale alla sua media con probabilità 1.
Per la 4.55 e per la linearità del valore atteso si ha:

$$\begin{aligned}
\operatorname{Var}[aX + b] &= \mathbb{E}\left[(aX + b - (a\,\mathbb{E}[X] + b))^2\right] \\
&= \mathbb{E}\left[(aX + b - a\,\mathbb{E}[X] - b)^2\right] \\
&= \mathbb{E}\left[(aX - a\,\mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[a^2(X - \mathbb{E}[X])^2\right] \\
&= a^2\,\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= a^2 \operatorname{Var}[X]
\end{aligned}$$

La 4.56 verrà dimostrata/generalizzata in seguito, per ora verifichiamola:

$$\begin{aligned}
\operatorname{Var}[X + Y] &= \mathbb{E}\left[(X + Y)^2\right] - (\mathbb{E}[X + Y])^2 = \mathbb{E}\left[X^2 + 2XY + Y^2\right] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&\overset{(1)}{=} \mathbb{E}\left[X^2\right] + 2\,\mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}\left[Y^2\right] - \mathbb{E}[X]^2 - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\
&= \operatorname{Var}[X] + \operatorname{Var}[Y]
\end{aligned}$$

where in (1) we used that if $X \perp\!\!\!\perp Y$ we have $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. □

Osservazione 114 (Variance is nonlinear). Differently from expected value a is squared and b omitted, therefore variance of sum of different random variable could be different from sum of their variance.

Definizione 4.6.7 (Standard deviation)**.**

$$\sigma = \sigma_X = \sqrt{\operatorname{Var}[X]} \tag{4.57}$$

4.6.3 Asymmetry/skewness and kurtosis

Definizione 4.6.8 (Standardized rvs)**.** If X has $\mathbb{E}[X] = \mathbb{E}[X]$ and variance $\operatorname{Var}[X] = \sigma^2 \in (0, +\infty)$, standardized rv Z is defined as:

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}} = \frac{X - \mathbb{E}[X]}{\sigma} \tag{4.58}$$

Osservazione 115. This transform make rv independent from measure unit.

Definizione 4.6.9 (r-th standardized moments of X)**.** We have them if $g = \left(\frac{x-\mathbb{E}[X]}{\sigma}\right)^r$:

$$\bar{\bar{\mu}}_r = \mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sigma}\right)^r\right] = \begin{cases} \sum_i \left(\frac{x_i-\mathbb{E}[X]}{\sigma}\right)^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} \left(\frac{x-\mathbb{E}[X]}{\sigma}\right)^r \cdot f_X(x)\,\mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{4.59}$$

Osservazione 116. Since for any rv $\bar{\bar{\mu}}_0 = 1$, $\bar{\bar{\mu}}_1 = 0$, $\bar{\bar{\mu}}_2 = 1$ moments of interest are where $r = 3$ and $r = 4$.

4.6.3.1 Asymmetry/Skewness

Definizione 4.6.10 (Symmetric rv)**.** X is symmetric (respect to $\mathbb{E}[X]$) if $X - \mathbb{E}[X]$ has the same distribution of $\mathbb{E}[X] - X$.

Osservazione 117 (Intuizione significato)*.* $X-\mathbb{E}[X]$ sposta la densità/probabilità, così com'è, centrandola sullo 0. Intuitivamente $-X$ ha l'effetto di ottenere la densità probabilità simmetrica/specchiata rispetto a $x = 0$; infine $-X + \mathbb{E}[X]$ specchia la densità/probabilità rispetto a 0 e poi la ricentra su 0. Pertanto se $X - \mathbb{E}[X]$ e $-X + \mathbb{E}[X]$ coincidono, è perché la distribuzione di partenza X è simmetrica rispetto al centro.

Proposizione 4.6.7 (Simmetria di una vc continua (PDF))**.** *Sia X una vc continua con PDF f. Allora è simmetrica su $\mathbb{E}[X]$ se e solo se $f(x) = f(2\,\mathbb{E}[X] - x)$.*

Osservazione 118. La definizione è meramente quella di una funzione simmetrica rispetto a $x = \mu$ (vedi calcolo).

Dimostrazione. Sia F la CDF di X; dimostriamo la doppia implicazione.
Se la simmetria vale ($X - \mathbb{E}[X] = \mathbb{E}[X] - X$) abbiamo:

$$\begin{aligned} F(x) = \mathbb{P}(X - \mathbb{E}[X] \le x - \mathbb{E}[X]) &\overset{(1)}{=} \mathbb{P}(\mathbb{E}[X] - X \le x - \mathbb{E}[X]) \overset{(2)}{=} \mathbb{P}(X \ge 2\,\mathbb{E}[X] - x) \\ &= 1 - F(2\,\mathbb{E}[X] - x) \end{aligned}$$

dove in (1) abbiamo sfruttato la simmetria ($X - \mathbb{E}[X] = \mathbb{E}[X] - X$) e in (2) abbiamo elaborato algebricamente. Facendo la derivata dei membri estremi dell'equazione si ottiene $f(x) = f(2\,\mathbb{E}[X] - x)$.
Viceversa supponendo che $f(x) = f(2\,\mathbb{E}[X] - x)$ valga $forall x$, vogliamo dimostrare che $\mathbb{P}(X - \mathbb{E}[X] \le t) = \mathbb{P}(\mathbb{E}[X] - X \le t)$, ossia vi è simmetria e le cumulate CDF coincidono. Si ha

$$\begin{aligned} \mathbb{P}(X - \mathbb{E}[X] \le t) = \mathbb{P}(X \le \mathbb{E}[X] + t) = \int_{-\infty}^{\mathbb{E}[X]+t} f(x)\,\mathrm{d}x &\overset{(1)}{=} \int_{-\infty}^{\mathbb{E}[X]+t} f(2\,\mathbb{E}[X] - x)\,\mathrm{d}x \\ &\overset{(2)}{=} \int_{\mathbb{E}[X]-t}^{\infty} f(w)\,\mathrm{d}w = \mathbb{P}(\mathbb{E}[X] - X \le t) \end{aligned}$$

dove in abbiamo sfruttato che $f(x) = f(2\,\mathbb{E}[X] - x)$, mentre in (2) deve avvenire qualche trick di integrazione (integra $f(-x)$ ad indici invertiti e moltiplicati direi). □

Definizione 4.6.11 (Skewness)**.** It's the 3-rd standardized moment:

$$\text{Asym}(X) = \bar{\mu_3} = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^3\right] \tag{4.60}$$

Osservazione 119. A negative skewness means a left longer tail, while positive a right longer one.

4.6.3.2 Kurtosis

Definizione 4.6.12 (Kurtosis)**.** It's the 4-th standardized moment

$$\text{Kurt}(X) = \bar{\mu_4} = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^4\right] \tag{4.61}$$

Osservazione 120. Some defines kurtosis by centering on 3 (value assumed by the normal) as in:

$$\text{Kurt}(X) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^4\right] - 3 \tag{4.62}$$

In this way the normal will have 0 kurtosis and the remaining a value a negative or positive value, related to givin less or more weight to the tail of the distribution.

Osservazione 121. Una distribuzione con eccesso di curtosi (4.62) negativo (detta *platicurtica*) tende ad avere un profilo più piatto della normale e una minore importanza delle code. Produce outlier in misura minore o meno estremi rispetto alla normale. Un esempio è l'uniforme.
Viceversa una distribuzione con eccesso di curtosi positivo è detta *leptocurtica* (ad esempio distribuzione T di Student, logistica, Laplace): ha code che si avvicinano allo zero più lentamente rispetto una gaussiana, per cui produce più outlier della stessa.
In fig 4.3 alcune distribuzioni (con media 0 e varianza 1) e relativa curtosi.

4.7 Exercises

Esempio 4.7.1 (Es crash course, giorno 1)**.** Let X be a rv that has the density

$$f(x) = \begin{cases} ce^{-\lambda x} & if x \geq 0 \\ 0 & x < 0 \end{cases}$$

Find:

1. c
2. $\mathbb{E}[X]$
3. $\text{Var}[X]$
4. $F(X)$

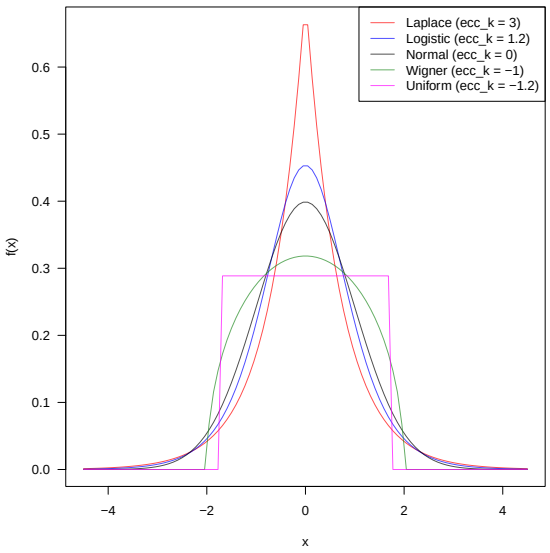


Figura 4.3: PDF for some rv (mean 0, variance 1) and their kurtosis

We have

1. it must be that

$$1 = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_{0}^{+\infty} ce^{-\lambda x}\,\mathrm{d}x = c\int_{0}^{+\infty} e^{-\lambda x}\,\mathrm{d}x = c\left[\left(\frac{1}{-\lambda}e^{-\lambda x}\right)\right]_{0}^{+\infty}$$
$$= 0 - \frac{c}{-\lambda}\cdot 1$$

   therefore $c = \lambda$ (this is the exponential distribution)

2. we have,

$$\mathbb{E}[X] = \int_{0}^{+\infty} x\cdot\lambda e^{-\lambda x}\,\mathrm{d}x = \lambda\int_{0}^{+\infty} x\cdot e^{-\lambda x}\,\mathrm{d}x$$

   using integration by parts we have

$$\begin{aligned}\int xe^{-\lambda x}\,\mathrm{d}x &= x\left(-\frac{1}{\lambda}e^{-\lambda x}\right) - \int -\frac{1}{\lambda}e^{-\lambda x}\\ &= \left(-\frac{x}{\lambda}e^{-\lambda x}\right) + \frac{1}{\lambda}\int e^{-\lambda x}\\ &= \left(-\frac{x}{\lambda}e^{-\lambda x}\right) + \frac{1}{\lambda}\left(-\frac{1}{\lambda}e^{-\lambda x}\right)\end{aligned}$$

   che opportunamente valutato

$$\left[\left(-\frac{x}{\lambda}e^{-\lambda x}\right) + \frac{1}{\lambda}\left(-\frac{1}{\lambda}e^{-\lambda x}\right)\right]_{0}^{+\infty} = 0 + 0 - \left(0 - \frac{1}{\lambda^2}\right)$$

Per cui tornando al valore atteso

$$\mathbb{E}[X] = \lambda\left(\frac{1}{\lambda^2}\right) = \frac{1}{\lambda}$$

3. first we find $\mathbb{E}\left[X^2\right]$

$$\mathbb{E}\left[X^2\right] = \lambda\int_{-\infty}^{+\infty} x^2 e^{-\lambda x}\,\mathrm{d}x \overset{(1)}{=} \lambda\left[(x^2\frac{-1}{\lambda}e^{-\lambda x})\Big|_0^\infty + \frac{2}{\lambda}\int_0^{+\infty} xe^{-\lambda x}\,\mathrm{d}x\right] = 2\int_0^{+\infty} xe^{lambdax}\,\mathrm{d}x =$$

where in (1) again by integration by parts. So

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

4. we have

$$\begin{aligned} F(x) &= \int_0^x f(s)\,\mathrm{d}s = \lambda\int_0^x e^{-\lambda s}\,\mathrm{d}s = \lambda\left(\frac{-1}{\lambda}e^{-\lambda s}\right)\Big|_0^x \\ &= 1 - e^{\lambda x}, \quad \text{for } x \geq 0 \end{aligned}$$

for $x < 0, F(x) = \int_{-\infty}^x f(s)\,\mathrm{d}s = 0$ so

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{\lambda x} & x \geq 0 \end{cases}$$

Esempio 4.7.2 (crash course, day 1 es 3 pag 6)**.** Let $f(k) = \frac{c^k e^{-\lambda}}{k!}$ for $k \in \{0, 1, \ldots\}$ be the pmf that X satisfies:

1. find c
2. find $\mathbb{E}[X]$
3. find $\mathrm{Var}[X]$

we have

1.

$$\sum_{k=0}^{\infty} f(k) = 1 = e^{-\lambda}\underbrace{\sum_{k=0}^{\infty}\frac{c^k}{k!}}_{e^c} = e^{-\lambda}e^c = e^{c-\lambda} = 1 = e^0 \implies c = \lambda$$

2.

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} kf(k) = \sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty}\frac{\lambda^k e^{-\lambda}}{(k-1)!} = \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!} \overset{(1)}{=} \lambda\underbrace{\sum_{u=0}^{\infty}\frac{\lambda^u e^{-}}{u!}}_{F(\infty)=1}$$

with (1) substituting $u = k - 1$. This is the poisson distribution, we say $X \sim \mathrm{Pois}(\lambda)$

3. first we find $\mathbb{E}\left[X^2\right]$, but first consider the following

$$\mathbb{E}\left[X(X-1)\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right] = \sum_{k=0}^{\infty} k(k-1)f(k) = \sum_{k=2}^{\infty} k(k-1)f(k) = \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=2}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-2)!} =$$

where in (1) doin subst $u = k-2$. Therefore

$$\mathbb{E}\left[X^2\right] = \lambda^2 + \lambda \implies \operatorname{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Esempio 4.7.3 (crashcourse, day 1 es 3 pag 7)**.** Let $X \sim \operatorname{Bin}(n,p)$, that is $\mathbb{P}(X = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$.

TODO: da finire ma valuta se ne vale la pena, la binomiale è già sviluppata nella prossima sezione

Esempio 4.7.4 (crashcourse, day 1 es 4 pag 7)**.** Let $F(x) = \frac{c}{2}\left(1 - \frac{1}{x^2}\right)$ for $x \in [1, \infty)$:

1. obtain $f(x)$
2. obtain c
3. $\mathbb{E}\left[X\right]$
4. $\operatorname{Var}\left[X\right]$

we have

1. $$f(x) = \frac{\partial}{\partial x} F(x) = \frac{\partial}{\partial x} \frac{c}{2}\left(1 - \frac{1}{x^2}\right) = \frac{c}{x^3}$$
2. $$c\int_1^{\infty} \frac{1}{x^3}\,\mathrm{d}x = c\left[-\frac{1}{2}\frac{1}{x^2}\right]_1^{\infty} = \frac{c}{2} = 1 \implies c = 2$$
3. $$2\int_1^{\infty} x\frac{1}{x^3}\,\mathrm{d}x = 2\int_1^{\infty} \frac{1}{x^2}\,\mathrm{d}x = 2\left[\frac{-1}{x}\right]_1^{\infty} = 2$$
4. first we find
$$\mathbb{E}\left[X^2\right] = 2\int_1^{\infty} x^2\frac{1}{x^3} = 2\int_1^{\infty} \frac{1}{x}\,\mathrm{d}x = 2\left[\log x\right]_1^{\infty} = +\infty$$

Esempio 4.7.5 (crashcourse, day 1 es 5 pag 8)**.** Let $f(x) = ce^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$:

1. find c
2. $\mathbb{E}\left[X\right]$
3. $\operatorname{Var}\left[X\right]$

Respectively

1. we know $\int_{-\infty}^{+\infty} cf(x)\,\mathrm{d}x = 1$ so we can do this trick

$$1 = \underbrace{\int_{-\infty}^{+\infty} cf(x)\,\mathrm{d}x}_{1}\underbrace{\int_{-\infty}^{+\infty} cf(y)\,\mathrm{d}y}_{1}$$
$$= c^2\int_{-\infty}^{+\infty} f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y = c^2\int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y$$

   Now trasforming variable to polar coordinates that is applying

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases}, \quad r \in [0,\infty), \theta \in [0, 2\pi)$$

   so that $x^2 + y^2 = r^2$ and $\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}r\,\mathrm{d}\theta$ we have

$$1 = c^2\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r\,\mathrm{d}r\,\mathrm{d}\theta$$
$$\overset{(1)}{=} c^2\int_0^{2\pi}\underbrace{\int_0^{+\infty} e^{-u}\,\mathrm{d}u}_{=1}\,\mathrm{d}\theta = c^2\int_0^{2\pi}\mathrm{d}\theta = c^2 2\pi = 1 \implies c = \frac{1}{\sqrt{2\pi}}$$

   where in (1) we substitute $u = \frac{r^2}{2}$ so $\mathrm{d}u = r\,\mathrm{d}r$.

2. $\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x$. We have that $f(x)$ is an even function:

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{2\pi}}e^{-\frac{(-x)^2}{2}} = f(-x)$$

   it's symmetric. However we are interested in $\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x$ that is trying to find the area under an odd function. Now in general if we're trying to find

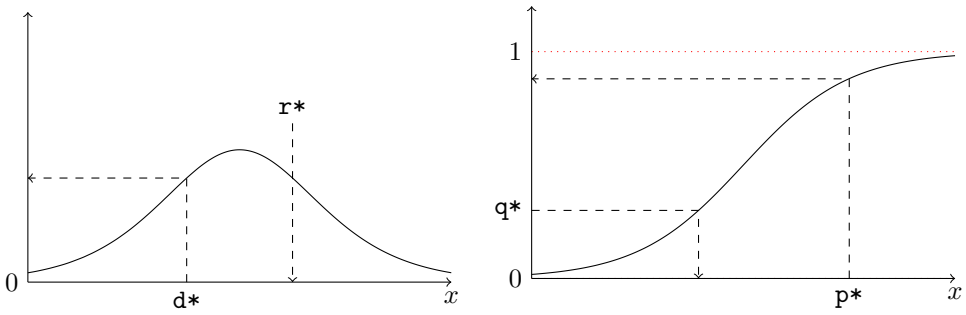
   - odd functions: given that it's symmetric around origin, positive areas compensates with negative areas so it's integral (over $\mathbb{R}$) is 0 (this holds for any odd function).

   - even functions: synce symmetric around y axis to calculate integral on region $(-\infty, \infty)$ we can double the integral on region $(0, \infty)$

   Therefore our $\mathbb{E}[X] = 0$.

3. for the variance first get

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{+\infty}\overbrace{\underbrace{x^2}_{even}\underbrace{f(x)}_{even}}^{even}$$
$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x \overset{(1)}{=} \frac{2}{\sqrt{2\pi}}\int_0^{+\infty} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x \overset{(2)}{=} \frac{2}{\sqrt{2\pi}}\int_0^{+\infty}\sqrt{2}u^{1/2}e^{-u}\,\mathrm{d}u$$
$$= \frac{2}{\sqrt{\pi}}\int_{-\infty}^{+\infty} u^{1/2}e^{-u}\,\mathrm{d}u = \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = 1$$



(a) PMF/PDF (b) CDF

Figura 4.4: Funzioni in R

where in (1) since it's even and (2) using the variable change $u = \frac{x^2}{2}$ therefore $\mathrm{d}u = x\,\mathrm{d}x$ and $x = \sqrt{2u}$.
$\Gamma(x)$ is called gamma function, we will be familiar with it in the next years, for now trust me that $\Gamma x = (x-1)\Gamma(x-1)$ and $\Gamma(1) = 1$ $\Gamma(1/2) = \sqrt{\pi}$ so for integers n $\Gamma n = (n-1)!$ but for our case $\Gamma(3/2) = \frac{1}{2}\Gamma\frac{1}{2} = \frac{\sqrt{\pi}}{2}$

Therefore in the end for $X \sim \mathrm{N}(0,1)$, $\mathbb{E}[X] = 0$ and $\mathrm{Var}[X] = 1$. This is called a standard rv. But in general normal rvs can have different mean and variance: the general case is denoted as $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ and this correspond to translation of a standard normal rv and then scaling it.
Let $Z \sim \mathrm{N}(0,1)$ and $X = \sigma Z + \mu$ then

$$\mathbb{E}[X] = \mathbb{E}[\sigma Z + \mu] = \sigma \underbrace{\mathbb{E}[Z]}_{=0} + \mu = \mu$$
$$\mathrm{Var}[X] = \mathrm{Var}[\sigma Z + \mu] = \sigma^2 \underbrace{\mathrm{Var}[Z]}_{} = 1 = \sigma^2$$

4.8 Probability models and R

Osservazione 122. In the following chapters we study main probabilistic models, which are the most used family of distribution

Definizione 4.8.1 (Family of random variables)**.** Set of distribution function $F(x; \boldsymbol{\Theta})$ having the same functional form+ but different for one or more parameters.

Definizione 4.8.2 (Parameters space)**.** $\boldsymbol{\Theta}$, it's the set of possible value for the parameters of a distribution function.

Osservazione 123. In:

- table 4.2 we report prefix of main functions and suffixes of main families;
- figure 4.4 function input needed (where arrows starts) and output returned (where arrow ends) for the 4 main functions.

Function	Prefix	Family	Suffix	Family	Suffix
Density/Probability	`d`	Bernoulli	`binom`	Uniforme cont.	`unif`
PDF	`p`	Binomiale	`binom`	Esponenziale	`exp`
Quantile	`q`	Geometrica	`geom`	Normale	`norm`
RNG	`r`	Binomiale neg.	`nbinom`	Gamma	`gamma`
		Ipergeometrica	`hyper`	Chi-quadrato	`chisq`
		Poisson	`pois`	Beta	`beta`
		Uniforme disc.	$*$	T di Student	`t`
				F	`f`
				Logistica	`logis`
				Lognormale	`lnorm`
				Weibull	`weibull`
				Pareto (pac. `VGAM`)	`pareto`

Tabella 4.2: Utilities for family of rvs in R

Osservazione 124 (Variabili discrete con supporto finito in R). Per quanto riguarda la simulazione di queste (tra le quali l'uniforme discreta) si fa utilizzo della funzione `sample` alla quale, oltre a specificare l'urna `x`, il numero `size` di estrazioni desiderate, l'estrazione con reinserimento (`replace`) o meno, si possono specificare le probabilità `prob` di ciascun elemento nell'urna.

```
## DUnif(100)
sample(x = 1:100, size = 10, replace = TRUE)

##  [1] 23 52 90 13  5  6 33 80 12  6

## Urna discreta custom
sample(x = 1:3, prob = c(0.4, 0.4, 0.2), size = 10, replace = TRUE)

##  [1] 1 3 2 2 1 1 1 2 2 2
```

Capitolo 5

Discrete random variables

5.1 Dirac

Definizione 5.1.1 (Dirac rv (degenere))**.** $X \sim \delta_c$ if $\mathbb{P}(X = c) = 1$.

Proposizione 5.1.1.

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 & \text{if } x \geq c \\ 0 & \text{if } x < c \end{cases} \tag{5.1}$$

Proposizione 5.1.2 (Moments)**.**

$$\mathbb{E}[X] = c$$
$$\mathrm{Var}[X] = 0$$

Dimostrazione.

$$\mathbb{E}[X] = c \cdot 1 = c$$
$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = c^2 \cdot 1 - c^2 = 0$$

□

Osservazione 125. Dirac is the only random variable having null variance.

5.2 Bernoulli

5.2.1 Definition

Osservazione 126. Viene utilizzata quando si ha a che fare con un esperimento il cui esito possibile è dicotomico (es $X = 1$ successo, $X = 0$ insuccesso).

Definizione 5.2.1 (vc di Bernoulli)**.** X is distributed as Bernoulli with parameter $0 \leq p \leq 1$, written $X \sim \mathrm{Bern}(p)$, if $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

Osservazione 127. If $p = 0 \vee p = 1$ we obtain a Dirac.

5.2.2 Functions

Osservazione 128 (Support and parametric space).

$$R_X = \{0, 1\}$$
$$\boldsymbol{\Theta} = \{p \in \mathbb{R} : 0 \leq p \leq 1\}$$

Definizione 5.2.2 (PMF)**.**

$$p_X(x) = \mathbb{P}(X = x) = p^x \cdot (1-p)^{1-x} \cdot \mathbb{1}_{R_X}(x) \tag{5.2}$$

Definizione 5.2.3 (PDF)**.**

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{se } x < 0 \\ 1-p & \text{se } 0 \leq x < 1 \\ 1 & \text{se } x \geq 1 \end{cases} \tag{5.3}$$

5.2.3 Moments

Proposizione 5.2.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = p \tag{5.4}$$
$$\text{Var}[X] = p(1-p) \tag{5.5}$$
$$\text{Asym}(X) = \frac{1-2p}{\sqrt{p(1-p))}} \tag{5.6}$$
$$\text{Kurt}(X) = \frac{3p^2 - 3p + 1}{p(1-p)} \tag{5.7}$$

Dimostrazione. Per il valore atteso

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

Per la varianza, dato che $X^2 = X$ e dunque $\mathbb{E}\left[X^2\right] = \mathbb{E}[X]$ si ha:

$$\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

□

Osservazione 129. In particolare il valore atteso coincide con la probabilità di successo e la varianza è sempre compresa nell'intervallo $[0; 0.25]$, raggiungendo il massimo per $p = 1/2$.

5.3 Indicator rv for an event

5.3.1 Definition, properties

Osservazione importante 37. Any event A is associated to a Bernoulli indicator random variable.

Definizione 5.3.1 (Indicator rv of event A)**.** Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be the sample space of the experiment considered and $A \subseteq \Omega$ a possible event; suppose that ω is the outcome that currently happens as a result of the experiment. Then:

$$I_A = I(A) = \begin{cases} 1 & \text{if } A \text{ verifies: } \omega \in A \\ 0 & \text{if } A \text{ does not: } \omega \notin A \end{cases}$$

therefore if $\mathbb{P}(A) = p$, then $I_A \sim \text{Bern}(p)$

Proposizione 5.3.1 (Indicator rv properties)**.**

$$(I_A)^n = I_A, \quad \forall n \in \mathbb{N} : n > 0 \tag{5.8}$$
$$I_{\overline{A}} = 1 - I_A \tag{5.9}$$
$$I_{A \cap B} = I_A \cdot I_B \tag{5.10}$$
$$I_{A \cup B} = I_A + I_B - I_A \cdot I_B \tag{5.11}$$

Dimostrazione. La 5.8 vale dato che $0^n = 0$ e $1^n = 1$ per qualsiasi intero positivo n. La 5.9 vale dato che $1 - I_A$ è 1 se A non accade e 0 se accade. Per la 5.10, $I_A \cdot I_B$ è 1 solo se sia I_A che I_B sono 1 e 0 altrimenti. Per la 5.11,

$$\begin{aligned} I_{A \cup B} &\overset{(1)}{=} 1 - I_{\overline{A} \cap \overline{B}} = 1 - I_{\overline{A}} \cdot I_{\overline{B}} = 1 - (1 - I_A)(1 - I_B) \\ &= I_A + I_B - I_A I_B \end{aligned}$$

dove in (1) abbiamo sfruttato De Morgan. □

5.3.2 Probability/expected value link

Osservazione 130. Indicator function/rv provide a link between probability of an event and expected value

Proposizione 5.3.2 (Fundamental bridge)**.** *There's a 1-1 link between events and indicator rv: probability of an event A and the expected value of its indicator rv I_A:*

$$\mathbb{P}(A) = \mathbb{E}[I_A] \tag{5.12}$$

Dimostrazione. For any event A we have a rv I_A, and viceversa for each I_A there's one event A (that is $A = \{\omega \in \Omega : I_A(\omega) = 1\}$).
Considered $I_A \sim \text{Bern}(p)$ with $p = \mathbb{P}(A)$, we have

$$\mathbb{E}[I_A] = \mathbb{E}[\text{Bern}(p)] = p = \mathbb{P}(A)$$

□

Osservazione 131 (Usefulness). Previous result enable to express any probability as expected value; some examples come in the following section.
Furthermore indicator rvs are useful in exercises on expected value: often we can define a complex rv of unknown/complex distribution function as sum of indicator function (simpler). The so-called fundamental bridge enable then, applying expected value properties, to find expected value of unknown complex distribution function

5.3.3 Some application: probability

Proposizione 5.3.3 (Boole inequality)**.** *If $E_1, \ldots, E_n$ are events we have:*

$$\mathbb{P}(E_1 \cup \ldots \cup E_n) \leq \mathbb{P}(E_1) + \ldots + \mathbb{P}(E_n) \tag{5.13}$$

Dimostrazione. Let $E_1, \ldots, E_n$ be the events considered; we note that

$$I_{E_1 \cup \ldots \cup E_n} \leq I_{E_1} + \ldots + I_{E_n}$$

since left branch is 1 is all the events occur while right one is 1 even if only one does. Taking expected value:

$$\begin{aligned}
&\mathbb{E}[I_{E_1 \cup \ldots \cup E_n}] \leq \mathbb{E}[I_{E_1} + \ldots + I_{E_n}] && \text{by linearity of expectation } \ldots \\
&\mathbb{E}[I_{E_1 \cup \ldots \cup E_n}] \leq \mathbb{E}[I_{E_1}] + \ldots + \mathbb{E}[I_{E_n}] && \text{applying 5.12 } \ldots \\
&\mathbb{P}(E_1 \cup \ldots \cup E_n) \leq \mathbb{P}(E_1) + \ldots + \mathbb{P}(E_n)
\end{aligned}$$

□

Proposizione 5.3.4 (Bonferroni inequality)**.** *If $E_1, \ldots, E_n$ are events:*

$$\mathbb{P}(E_1 \cap \ldots \cap E_n) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\overline{E_i}\right) \tag{5.14}$$

Dimostrazione. Similarty to the Boole inequality, applying DeMorgan

$$I_{E_1 \cap \ldots \cap E_n} = 1 - I_{\overline{E_1} \cup \ldots \cup \overline{E_n}}$$

Taking expected value:

$$\begin{aligned}
&\mathbb{E}[I_{E_1 \cap \ldots \cap E_n}] = \mathbb{E}\left[1 - I_{\overline{E_1} \cup \ldots \cup \overline{E_n}}\right] && \text{per linearità } \ldots \\
&\mathbb{E}[I_{E_1 \cap \ldots \cap E_n}] = 1 - \mathbb{E}\left[I_{\overline{E_1} \cup \ldots \cup \overline{E_n}}\right] && \text{passando alle probabilità } \ldots \\
&\mathbb{P}(E_1 \cap \ldots \cap E_n) = 1 - \mathbb{P}\left(\overline{E_1} \cup \ldots \cup \overline{E_n}\right)
\end{aligned}$$

Finally applying 5.13

$$\mathbb{P}(E_1 \cap \ldots \cap E_n) = 1 - \mathbb{P}\left(\overline{E_1} \cup \ldots \cup \overline{E_n}\right) \geq 1 - \mathbb{P}\left(\overline{E_1}\right) - \ldots - \mathbb{P}\left(\overline{E_n}\right)$$

□

Proposizione 5.3.5 (Inclusion/exclusion principle)**.** *In case of two events*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{5.15}$$

In general:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \ldots < i_r} \mathbb{P}(E_{i1} \cap E_{i2} \cap \ldots \cap E_{ir}) \tag{5.16}$$

$$= \sum_{i} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k) - \ldots + (-1)^{n+1} \mathbb{P}(E_1 \cap \ldots \cap E_n) \tag{5.17}$$

Dimostrazione. Given 5.15 we take expected value of both branch of 5.11. Considering 5.16, we can apply indicator rv properties

$$\begin{aligned} 1 - I_{E_1 \cup \ldots \cup E_n} &= I_{\overline{E_1} \cap \ldots \cap \overline{E_n}} \\ &= I_{\overline{E_1}} \cdot \ldots \cdot I_{\overline{E_n}} \\ &= (1 - I_{E_1}) \cdot \ldots \cdot (1 - I_{E_n}) \\ &\overset{(1)}{=} 1 - \sum_i I_{E_i} + \sum_{i<j} I_{E_i} I_{E_j} - \ldots + (-1)^n I_{E_1} \cdot \ldots \cdot I_{E_n} \end{aligned}$$

where in (1):

- il 1 significa selezionare tutti gli 1 negli n fattori;
- il $\sum_i I_{E_i}$ si ottiene selezionando tutti gli 1 a meno di un fattore a turno che ha sempre il segno $-$ davanti;
- $\sum_{i<j} I_{E_i} I_{E_j}$ si ottiene selezionando tutti gli 1 ad eccezione di due fattori.

Prendendo i valori attesi di ambo i membri si ha

$$\mathbb{E}\left[1 - I_{E_1 \cup \ldots \cup E_n}\right] = \mathbb{E}\left[1 - \sum_i I_{E_i} + \sum_{i<j} I_{E_i} I_{E_j} - \ldots + (-1)^n I_{E_1} \cdot \ldots \cdot I_{E_n}\right]$$

$$1 - \mathbb{E}\left[I_{E_1 \cup \ldots \cup E_n}\right] \overset{(1)}{=} 1 - \mathbb{E}\left[\sum_i I_{E_i} - \sum_{i<j} I_{E_i} I_{E_j} + \ldots + (-1)^{n+1} I_{E_1} \cdot \ldots \cdot I_{E_n}\right]$$

$$\mathbb{E}\left[I_{E_1 \cup \ldots \cup E_n}\right] = \mathbb{E}\left[\sum_i I_{E_i}\right] - \mathbb{E}\left[\sum_{i<j} I_{E_i} I_{E_j}\right] + \ldots + \mathbb{E}\left[(-1)^{n+1} I_{E_1} \cdot \ldots \cdot I_{E_n}\right]$$

$$\mathbb{P}\left(\bigcup_{i=1}^n E_i\right) = \sum_i \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \ldots + (-1)^{n+1} \mathbb{P}(E_1 \cap \ldots \cap E_n)$$

dove in (1) abbiamo raccolto un meno al secondo membro entro parentesi. □

5.3.4 Applications: expected value evaluation

Esempio 5.3.1 (Matching carte)**.** Abbiamo un mazzo di n carte numerate da 1 a n ben mischiato. Una carta è un match se la sua posizione nell'ordine del mazzo matcha con il suo numero. Sia X il numero totale di match nel mazzo: qual è il valore atteso di X?
Se scriviamo $X = I_1 + \ldots + I_n$ con

$$I_i = \begin{cases} 1 & \text{se l'}i\text{-esima carta matcha col proprio numero} \\ 0 & \text{altrimenti} \end{cases}$$

Si ha che, non condizionando a nulla e pensando ad un singolo shuffle/match

$$\mathbb{E}[I_i] = \frac{1}{n}$$

Fisso ...	con reinserimento	senza reinserimento
n trial	binomiale	ipergeometrica
n successi	binomiale negativa	ipergeometrica negativa

Tabella 5.1

pertanto per linearità

$$\mathbb{E}[X] = \mathbb{E}[I_1] + \ldots + \mathbb{E}[I_n] = n \cdot \frac{1}{n} = 1$$

Quindi il numero di match medi è 1, indipendentemente da n. Anche se I_i sono dipendenti in maniera complicata, la linearità del valore atteso vale sempre.

Esempio 5.3.2 (Valore atteso di Ipergeometrica Negativa)**.** Un'urna contiene w palline bianche e b palline nere che sono estratte senza reinserimento. Il numero di palline nere estratte prima di pescare la prima bianca ha una distribuzione Ipergeometrica negativa (in tab 5.1 una sintesi dei casi). Trovare il valore atteso. Trovarlo dalla definizione della variabile è complicato, ma possiamo esprimere la variabile come somma di indicatrici. Etichettiamo le palline nere con $1, 2, \ldots b$ e sia I_i l'indicatrice che la pallina nera i è stata estratta prima di qualsiasi bianca. Si ha che

$$\mathbb{P}(I_i = 1) = \frac{1}{w+1}$$

dato nel listare l'ordine in cui la pallina nera i e le altre bianche son pescate (ignorando le altre) tutti gli ordine sono equiprobabili. Pertanto per linearità

$$\mathbb{E}\left[\sum_{i=1}^{b} I_i\right] = \sum_{i=1}^{b} \mathbb{E}[I_i] = \frac{b}{w+1}$$

La risposta ha n senso dato che aumenta con b, diminuisce con w ed è corretta nei casi estremi $b = 0$ (nessuna pallina nera sarà estratta) e $w = 0$ (tutte le palline nere saranno esaurite prima di pescare una non esistente bianca).

5.4 Binomial

5.4.1 Definition

Osservazione 132. Used to know the probability of having x success among $n \geq x$ independent Bernoulli trial with common probabily success p.

Definizione 5.4.1 (vc binomiale)**.** Eseguiamo n prove bernoulliane indipendenti, aventi comune probabilità di successo p. Sia X la somma dei successi ottenuti: allora X si distribuisce come una vc binomiale di parametri n e p, e si scrive $X \sim \text{Bin}(n, p)$.

Osservazione 133. Se $n = 1$ la distribuzione Binomiale coincide con quella di Bernoulli, ossia $\text{Bin}(1, p) = \text{Bern}(p)$

Proposizione 5.4.1. *La binomiale può essere generata sommando bernoulliane iid; se X_i, $i = 1, \ldots, n$ sono vc bernoulliane iid $X_i \sim \text{Bern}(p)$ allora la loro somma $X = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$*

Dimostrazione. Sia $X_i = 1$ se l'i-esimo trial ha successo o 0 in caso contrario. Se pensiamo di avere una persona per ciascun trial, chiediamo di alzare la mano se si ha successo e contiamo le mani alzate (che equivale a sommare X_i) otteniamo il numero totale di successi in n trial che è X. □

5.4.2 Functions

Osservazione 134 (Supporto e spazio parametrico).

$$R_X = \{0, 1, \dots, n\}$$
$$\boldsymbol{\Theta} = \{n \in \mathbb{N} \setminus \{0\}, \ p \in \mathbb{R} : 0 \leq p \leq 1\}$$

Definizione 5.4.2 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \binom{n}{x} \cdot p^x (1-p)^{n-x} \cdot \mathbb{1}_{R_X}(x) \tag{5.18}$$

con: x è il numero di successi, n è il numero di esperimenti, p probabilità di successo in ogni esperimento.

Osservazione 135. Nella 5.18 la prima parte (il coefficiente binomiale) serve per quantificare il numero di casi in cui si verificano il numero di successi desiderati; questa viene moltiplicata per la seconda che costituisce la probabilità di un tale esito (determinato come probabilità di eventi indipendenti di successo/insuccesso).

Definizione 5.4.3 (Funzione di ripartizione)**.**

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} \cdot p^k (1-p)^{n-k}$$

Validità PMF. Si ha che

$$\sum_{x=0}^{n} p_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \overset{(1)}{=} (p + (1-p))^n = 1$$

dove in (1) si è sfruttata la proprietà del coefficiente binomiale:

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

□

5.4.3 Moments

Proposizione 5.4.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = np \tag{5.19}$$
$$\mathrm{Var}[X] = np(1-p) \tag{5.20}$$
$$\mathrm{Asym}(X) = \frac{1-2p}{\sqrt{np(1-p))}} \tag{5.21}$$
$$\mathrm{Kurt}(X) = 3 + \frac{1 - 6p + 6p^2}{np(1-p)} \tag{5.22}$$

Dimostrazione. Per il valore atteso, sfruttando il fatto che $X \sim \text{Bin}(n,p)$ sia descrivibile come la somma di n vc $X_i \sim \text{Bern}(p)$, sfruttando la linearità del valore atteso, il risultato è la somma di n valori attesi uguali:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\,\mathbb{E}[X_i] = np$$

Alternativamente potevamo sviluppare l'algebra:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{(n-x)} = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)} \\ &= \sum_{x=0}^{n} x \cdot \frac{n(n-1)!}{x(x-1)!\,[(n-1)-(x-1)]!} p p^{x-1} (1-p)^{[(n-1)-(x-1)]}\end{aligned}$$

Ora dato che per $x = 0$ il termine entro sommatoria è nullo possiamo portare avanti di uno l'indice inferiore della stessa:

$$\mathbb{E}[X] = \sum_{x=1}^{n} x \cdot \frac{n(n-1)!}{x(x-1)!\,[(n-1)-(x-1)]!} p p^{x-1} (1-p)^{[(n-1)-(x-1)]}$$

ponendo $y = x - 1$ si giunge

$$\begin{aligned}\mathbb{E}[X] &= np \sum_{y=0}^{n-1} \underbrace{\frac{(n-1)!}{y!\,[(n-1)-y]!} p^y (1-p)^{[(n-1)-y]}}_{\text{Bin}(n-1,p)} \\ &\overset{(1)}{=} np\end{aligned}$$

con (1) dato che la sommatoria è $= 1$. □

Dimostrazione. Sfruttando sempre il fatto che $X \sim \text{Bin}(n,p)$ sia descrivibile come la somma di n vc iid $X_i \sim \text{Bern}(p)$, con varianza comune $p(1-p)$, e applicando le proprietà della varianza:

$$\text{Var}[X] = \text{Var}\left[\sum_{i=1}^{n} X_i\right] \overset{(1)}{=} \sum_{i=1}^{n} \text{Var}[X_i] = n\,\text{Var}[X_i] = n \cdot p(1-p) \tag{5.23}$$

where in (1) there's no covariance since they are independent. □

5.4.4 Shape

Proposizione 5.4.3 (Shape)**.** *La distribuzione è simmetrica se $p = 0.5$, è asimmetrica positiva (coda a destra) se $p < 0.5$, asimmetrica negativa (a sinistra) se $p > 0.5$. (Figura 5.1)*

Dimostrazione. Per $p = 0.5$ è simmetrica in quanto $p = 1 - p = \frac{1}{2}$ e

$$p_X(x) = \binom{n}{x}\left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x} = p_X(n-x) = \binom{n}{n-x}\left(\frac{1}{2}\right)^{n-x}\left(\frac{1}{2}\right)^x \tag{5.24}$$

per le proprietà del coefficiente binomiale. E dato che $p_X(x) = p_X(n-x)$, $\forall x \in R_X$, allora la distribuzione è simmetrica attorno al centro del supporto. □

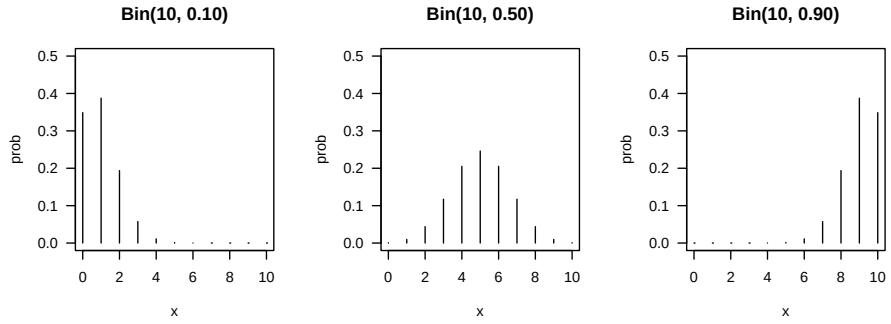


Figura 5.1: Forma distribuzione Bin (n, p)

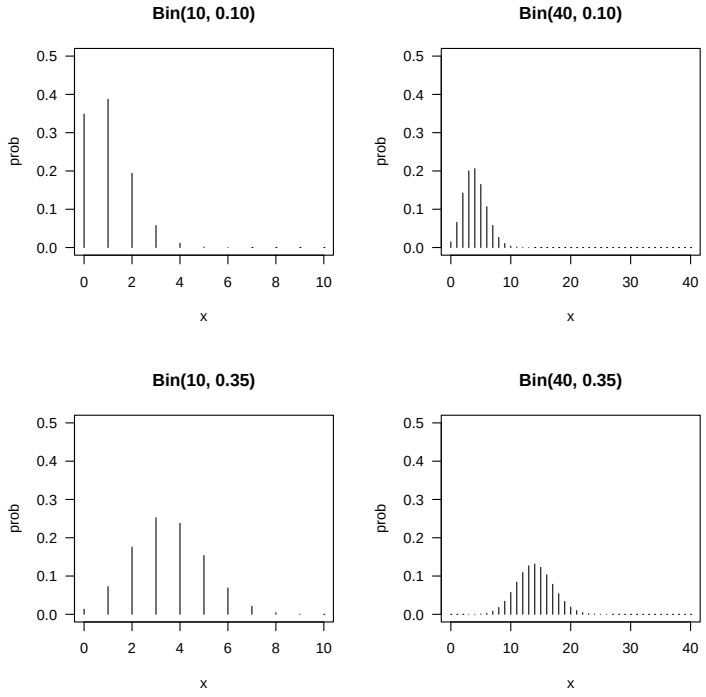


Figura 5.2: Convergenza alla normale della binomiale

Proposizione 5.4.4. *In una binomiale di parametri n, p, la funzione di densità (per x che varia da 0 a n) è inizialmente strettamente crescente e successivamente strettamente decrescente. Si raggiunge il massimo in corrispondenza del più grande intero $x \leq (n+1)p$*

Dimostrazione. Consideriamo il rapporto $\mathbb{P}(X = x)/\mathbb{P}(X = x-1)$ e determiniamo per quali valori di x esso risulti maggiore (funzione crescente) o minore (decrescente) di 1:

$$\frac{\mathbb{P}(X=x)}{\mathbb{P}(X=x-1)} = \frac{\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x}}{\frac{n!}{(n-x+1)!(x-1)!}p^{x-1}(1-p)^{n-x+1}} = \frac{(n-x+1)p}{x(1-p)}$$

Quindi tale rapporto ≥ 1 se e solo se:

$$\begin{aligned}(n-x+1)p &\geq x(1-p)\\ np - xp + p &\geq x - xp\end{aligned}$$

ossia $x \leq (n+1)p$ □

Osservazione 136 (Convergenza alla normale). La distribuzione converge verso la Normale (diviene simmetrica e la curtosi tende a 3) al crescere di $n \to \infty$; la convergenza è tanto più veloce per quanto più p è prossimo a 0.5. (figura 5.2)

5.4.5 Variabili derivate

Proposizione 5.4.5 (Vc numero di insuccessi)**.** *Sia $X \sim \text{Bin}(n,p)$. Allora $n - X \sim \text{Bin}(n, 1-p)$.*

Dimostrazione. Ad intuito basta invertire i ruoli di successo e insuccesso (si inverte anche la probabilità). Volendo tuttavia verificare, sia $Y = n - X$, la PMF è:

$$\begin{aligned}\mathbb{P}(Y=x) &\overset{(1)}{=} \mathbb{P}(X = n-x) = \binom{n}{n-x}p^{n-x}(1-p)^x\\ &\overset{(1)}{=} \binom{n}{x}(1-p)^x p^{n-x} = \text{Bin}(n, 1-p)\end{aligned}$$

dove in (1) diciamo che in n estrazioni la probabilità di avere x fallimenti è uguale alla probabilità di avere $n-x$ successi , mentre in (2) abbiamo sfruttato la proprietà del coefficiente binomiale. □

Osservazione 137. Un fatto importante della binomiale è che la somma di binomiali indipendenti aventi la stessa probabilità di successo è un'altra binomiale

Proposizione 5.4.6 (Somma di binomiali)**.** *Se $X \sim \text{Bin}(n,p)$, $Y \sim \text{Bin}(m,p)$ e X è indipendente da Y, allora $X + Y \sim \text{Bin}(n+m, p)$*

Dimostrazione. Un modo semplice è rappresentare X e Y come le somma di $X = X_1 + \ldots + X_n$ e $Y = Y_1 + \ldots + Y_n$ con $X_i, Y_i \sim \text{Bern}(p)$ iid. Allora $X + Y$ è la somma di $n+m$ $\text{Bern}(p)$ iid, pertanto la distribuzione è $\text{Bin}(n+m, p)$ per teorema 5.4.1.

Alternativamente, mediante la legge delle probabilità totali, possiamo trovare la PMF di $X+Y$ condizionando su X (oppure ugualmente su Y) e sommando:

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{j=0}^{k} \mathbb{P}(X+Y=k|X=j)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \mathbb{P}(Y=k-j|X=j)\cdot\mathbb{P}(X=j) \\
&\overset{(1)}{=} \sum_{j=0}^{k} \mathbb{P}(Y=k-j)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j}(1-p)^{m-k+j} \cdot \binom{n}{j} p^{j}(1-p)^{n-j} \\
&= p^k(1-p)^{n+m-k} \sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j} \\
&\overset{(2)}{=} \binom{n+m}{k} p^k(1-p)^{n+m-k} = \mathrm{Bin}(n+m,p)
\end{aligned}$$

dove in (1) abbiamo sfruttato l'indipendenza tra X e Y e in (2) l'identità di Vandermonde (eq 2.15). □

5.5 Hypergeometric

5.5.1 Definition

Osservazione 138. La variabile ipergeometrica descrive l'estrazione *senza reinserimento* di palline dicotomiche da un urna. A differenza della binomiale dove la probabilità di successo p non cambiava da una sottoprova Bernoulliana all'altra, qui il non reinserimento fa si che la probabilità di successo vari ad ogni prova.

Definizione 5.5.1 (Distribuzione ipergeometrica)**.** Supponiamo di dover estrarre un campione di n palline senza reinserimento da un'urna che contiene w palline bianche (successo) e b nere. Il numero X di palline bianche (successi) tra le estratte si distribuisce come una ipergeometrica con parametri w, b ed n e si scrive $X \sim \mathrm{HGeom}(w,b,n)$.

5.5.2 Functions

Osservazione 139 (Supporto e spazio parametrico)*.*

$$\begin{aligned}
R_X &= \{0,1,\dots,n\} \\
\boldsymbol{\Theta} &= \{w,b\in\mathbb{N} : w+b\geq 1;\ n\in\{0,\dots,w+b\}\}
\end{aligned}$$

Definizione 5.5.2 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}} \cdot \mathbb{1}_{R_X}(x) \tag{5.25}$$

Osservazione 140 (Interpretazione)*.* Al denominatore sono quantificati il numero di modi con cui posso estrarre n palline qualsiasi dall'urna. Di queste estrazioni, al numeratore sono quantificati il numero di modi in cui nelle n palline estratte ci sono x bianche (successi); ossia devo averne x bianche scelte tra b, e $n - x$ nere scelte tra b.

Validità PMF. Facendo la somma del numeratore si ha:

$$\sum_{x=0}^{n} \binom{w}{x}\binom{b}{n-x} \overset{(1)}{=} \binom{w+b}{n}$$

con (1) per l'identità di Vandermonde (eq 2.15), per cui la PMF somma a 1. □

Osservazione 141. In R per la PMF si usa `dhyper(x, m, n, k)` dove `x` è il supporto (ossia il numero di palline bianche estratte), `m` il numero di palline bianche nell'urna, `n` il numero di palline nere e k il numero di estrazioni.

5.5.3 Moments

Proposizione 5.5.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = n\frac{w}{w+b} \tag{5.26}$$

$$\mathrm{Var}[X] = np(1-p)\left(\frac{w+b-n}{w+b-1}\right), \qquad \textit{con } p = \frac{w}{w+b} \tag{5.27}$$

Dimostrazione. Per il valore atteso, come nel caso binomiale possiamo scrivere X come somma di Bernoulliane $I_i \sim \mathrm{Bern}(p)$ con $p = w/(w+b)$.

$$X = I_1 + \ldots + I_n$$

A differenza della binomiale le I_i non sono indipendenti, tuttavia la linearità del valore atteso non lo richiede, quindi

$$\mathbb{E}[X] = \mathbb{E}[I_1 + \ldots + I_n] = \mathbb{E}[I_1] + \ldots + \mathbb{E}[I_n] = np = n\frac{w}{w+b}$$

□

Dimostrazione. Per la varianza invece essendo variabili non indipendenti non possiamo sommare le varianze direttamente. Vedremo in seguito la dimostrazione della formula riportata. □

5.5.4 Struttura essenziale ed esperimenti assimilabili

Osservazione 142. L'idea dell'Ipergeometrica è classificare una popolazione utilizzando due set di tag consecutivi (entrambi dicotomici successo/insuccesso) e ottenere il numero degli elementi caratterizzati dal successo in entrambi i tag. Nell'esempio delle palline il primo tag è il colore della pallina (bianco = successo), mentre il secondo è estrazione (estratta = sucesso).
Problemi aventi la stessa struttura presenteranno medesima distribuzione.

Esempio 5.5.1. Il numero A di assi estratti (sono 4 in un mazzo di 52 carte) in una mano di poker (5 carte estratte) si distribuirà come $A \sim \text{HGeom}(4, 48, 5)$.

Osservazione 143. La struttura essenziale ci permette di dimostrare facilmente l'uguaglianza di due ipergeometriche dove l'ordine dei set di tag viene invertito

Proposizione 5.5.2. $\text{HGeom}(w, b, n)$ *e* $\text{HGeom}(n, w + b - n, w)$ *sono identiche.*

Dimostrazione. Sia $X \sim \text{HGeom}(w, b, n)$ è il numero di palline bianche tra le estratte campione; sia $Y \sim \text{HGeom}(n, w + b - n, w)$ il numero di palline estratte tra le bianche (pensando ad estratto/non estratto come il primo tag e al colore come secondo. Entrambe X, Y contano il numero di bianche estratte pertanto avranno la stessa distribuzione.
Alternativamente possiamo controllare algebricamente che

$$\mathbb{P}(X = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}} = \frac{\frac{w!}{k!(w-k)!}\frac{b!}{(n-k)!(b-n+k)!}}{\frac{(w+b)!}{n!(w+b-n)!}} = \frac{w!b!n!(w+b-n)!}{k!(w-k)!(n-k)!(b-n+k)!}$$

$$\mathbb{P}(Y = y) = \frac{\binom{n}{y}\binom{w+b-n}{w-y}}{\binom{w+b}{w}} = \frac{\frac{n!}{k!(n-k)!}\frac{(w+b-n)!}{(w-k)!(b-n+k)!}}{\frac{(w+b)!}{w!b!}} = \frac{w!b!n!(w+b-n)!}{k!(w-k)!(n-k)!(b-n+k)!}$$

e dunque $\mathbb{P}(X = x) = \mathbb{P}(Y = y)$. □

5.5.5 Connessioni con la binomiale

Osservazione 144. Binomiale ed ipergeometrica sono connesse: possiamo ottenere la binomiale calcolando un limite sull'ipergeometrica, oppure ottenere una ipergeometrica condizionando una binomiale.

5.5.5.1 Dall'ipergeometrica alla binomiale

Proposizione 5.5.3. *Se* $X \sim \text{HGeom}(w, b, n)$ *e* $w + b \to \infty$ *ma* $p = w/(w + b)$ *rimane fisso, allora la PMF di* X *converge a* $\text{Bin}(n, p)$.

Dimostrazione. Sviluppiamo algebricamente per essere comodi prima di applicare il limite:

$$\mathbb{P}(X = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}} \overset{(1)}{=} \binom{n}{x}\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}}$$

dove in (1) abbiamo sfruttato che $\text{HGeom}(w, b, n) = \text{HGeom}(n, w + b - n, w)$ come nella dimostrazione di 5.5.2. Ora sviluppiamo il rapporto al secondo

fattore ricordando che $\binom{n}{d} = \frac{n!}{d!(n-d)!}$; si ha:

$$\begin{aligned}
\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}} &= \frac{(w+b-n)!}{(w-x)!(w+b-n-w+x)!} : \frac{(w+b)!}{w!(w+b-w)!} \\
&= \frac{(w+b-n)!}{(w-x)!(b-n+x)!} \cdot \frac{w!b!}{(w+b)!} \\
&= \frac{w!}{(w-x)!} \frac{b!}{(b-n+x)!} \frac{(w+b-n)!}{(w+b)!} \\
&= \frac{w \cdot \ldots \cdot (w-x+1)(w-x)!}{(w-x)!} \frac{b \cdot \ldots \cdot (b-n+x+1)(b-n+x)!}{(b-n+x)!} \frac{(w+b-n)!}{(w+b) \cdot \ldots \cdot (w+b-n+1)} \\
&= \frac{w \cdot \ldots \cdot (w-x+1)}{1} \frac{b \cdot \ldots \cdot (b-n+x+1)}{1} \frac{1}{(w+b) \cdot \ldots \cdot (w+b-n+1)}
\end{aligned}$$

ora al numeratore del primo rapporto abbiamo $w - (w-x+1) + 1 = x$ fattori, al numeratore del secondo ne abbiamo $b - (b-n+x+1) + 1 = n - x$ elementi. Pertanto complessivamente al numeratore abbiamo n fattori. Al denominatore invece abbiamo $(w+b) - (w+b-n+1) + 1 = n$ fattori anche qui. Pertanto possiamo dividere per $(w+b)$, applicandolo n volte sia al numeratore che al denominatore, ottenendo

$$\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}} = \frac{\frac{w}{w+b} \cdot \ldots \cdot \left(\frac{w}{w+b} - \frac{x-1}{w+b}\right) \cdot \left(\frac{b}{w+b}\right) \cdot \ldots \cdot \left(\frac{b}{w+b} - \frac{n-x-1}{w+b}\right)}{1 \cdot \ldots \cdot \left(1 - \frac{n-1}{w+b}\right)}$$

ora sostituendo $p = \frac{w}{w+b}$, $1-p = \frac{b}{w+b}$ e al denominatore $w+b = N$ dove utile si ha:

$$\frac{\binom{w+b-n}{w-x}}{\binom{w+b}{w}} = \frac{p \cdot \ldots \cdot \left(p - \frac{x-1}{N}\right) \cdot (1-p) \cdot \ldots \cdot \left(1 - p - \frac{n-x-1}{N}\right)}{\left(1 - \frac{1}{N}\right) \ldots \left(1 - \frac{n-1}{N}\right)}$$

Ora tornando da dove siamo partiti abbiamo:

$$\mathbb{P}(X = x) = \binom{n}{x} \frac{p \cdot \ldots \cdot \left(p - \frac{x-1}{N}\right) \cdot (1-p) \cdot \ldots \cdot \left(1 - p - \frac{n-x-1}{N}\right)}{\left(1 - \frac{1}{N}\right) \ldots \left(1 - \frac{n-1}{N}\right)}$$

Infine per $N \to +\infty$ il denominatore va a 1 mentre il numeratore va a $p^x(1-p)^{n-x}$ pertanto

$$\mathbb{P}(X = x) \to \binom{n}{x} p^x (1-p)^{n-x}$$

che è la $\mathrm{Bin}(n, p)$.
Intuitivamente data un'urna con w palline bianche e b nere, la binomiale sorge dall'estrarre n palline con replacement, mentre l'ipergeometrica senza. Se il numero di palline nell'urna sale notevolmente rispetto al numero di palline estratte, il campionamento con ripetizione e senza diventano essenzialmente equivalenti. (l'estrazione di una pallina non cambia la probabilità delle prossime estrazioni perché data la grande numerosità nell'urna non modifica praticamente la probabilità di successo) □

Osservazione 145. In termini pratici il teorema ci dice che se $N = w + b$ è grande rispetto a n possiamo approssimare la PMF di $\text{HGeom}(w, b, n)$ con $\text{Bin}(n, w/(w + b))$.

5.5.5.2 Dalla binomiale all'ipergeometrica

Proposizione 5.5.4. *Se $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$ e X è indipendente da Y, allora la distribuzione condizionata di X dato che $X + Y = r$ è* $\text{HGeom}(n, m, r)$

Osservazione 146. Dimostriamo attraverso un esempio (distribuzione del test esatto di Fisher).

Dimostrazione. Un ricercatore vuole studiare se la prevalenza di una data malattia sia uguale o meno tra maschi e femmine. Raccoglie un campione di n donne ed m uomini e testa la malattia. Sia $X \sim \text{Bin}(n, p_1)$ il numero di donne con la malattia nel campione e $Y \sim \text{Bin}(n, p_2)$ il numero di uomini. Qui p_1 e p_2 sono sconosciuti.
Supponiamo che siano osservate $X + Y = r$ persone malate. Siamo interessati a testare se $p_1 = p_2 = p$ (la cd ipotesi nulla); il test di Fisher si fonda sul condizionare sui totali di riga e colonna (quindi n, m, r sono considerati fissi) e verificare se il valore osservato X (numero di donne malate) sia estremo (dato che il tot malati è r) sotto ipotesi nulla. Assumendo l'ipotesi nulla vera troviamo la PMF condizionale di X dato che $X + Y = r$.
La tabella 2×2 di riferimento è la 5.2. Costruiamo PMF condizionata attraverso la regola di Bayes:

$$\mathbb{P}(X = x|X + Y = r) = \frac{\mathbb{P}(X + Y = r|X = x)\,\mathbb{P}(X = x)}{\mathbb{P}(X + Y = r)} = \frac{\mathbb{P}(Y = r - x|X = x)\,\mathbb{P}(X = x)}{\mathbb{P}(X + Y = r)}$$
$$\overset{(1)}{=} \frac{\mathbb{P}(Y = r - x)\,\mathbb{P}(X = x)}{\mathbb{P}(X + Y = r)}$$

dove in (1) abbiamo sfruttato l'indipendenza di X e Y. Assumendo per buona l'ipotesi nulla e impostando $p_1 = p_2 = p$ si hanno le vc indipendenti $X \sim \text{Bin}(n, p)$ e $Y \sim \text{Bin}(m, p)$, per cui $X + Y \sim \text{Bin}(n + m, p)$ (per il risultato 5.4.6). Pertanto sostituendo le formule per esteso si ha

$$\mathbb{P}(X = x|X + Y = r) = \frac{\binom{m}{r-x}p^{r-x}(1-p)^{m-r+x} \cdot \binom{n}{x}p^{x}(1-p)^{n-x}}{\binom{n+m}{r}p^{r}(1-p)^{n+m-r}}$$
$$= \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}} = \text{HGeom}(n, m, r)$$

Intuitivamente questo avviene perché condizionatamente ad avere $X + Y = r$ malati (primo tag), X è il numero di donne (secondo tag) tra quelli. □

5.6 Geometric

5.6.1 Definition

Osservazione 147. Supponiamo di ripetere in maniera indipendente diverse prove bernoulliane, ciascuna avente p probabilità di successo, sino a che che si ve-

	Donne	Uomini	Tot
Malato	x	$r-x$	r
Sano	$n-x$	$m-r+x$	$n+m-r$
Tot	n	m	$n+m$

Tabella 5.2

rifica il primo successo. Sia X il numero di *fallimenti* necessari per ottenere il primo successo; X si distribuisce come una variabile geometrica con parametro p e si scrive $X \sim \text{Geom}(p)$.

Esempio 5.6.1. Il numero di croci sino alla prima testa si distribuisce come $\text{Geom}(1/2)$.

5.6.2 Functions

Osservazione 148 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{N}\}$$
$$\boldsymbol{\Theta} = \{p \in (0,1)\}$$

Definizione 5.6.1 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = (1-p)^x p \cdot \mathbb{1}_{R_X}(x) \tag{5.28}$$

Validità PMF. Si ha che

$$\sum_{x=0}^{\infty}(1-p)^x p = p\sum_{x=0}^{\infty}(1-p)^x \overset{(1)}{=} p \cdot \frac{1}{p} = 1$$

con l'uguaglianza (1) dovuta alla serie geometrica. □

Osservazione 149. Come il teorema binomiale mostra che la PMF binomiale sia valida, la serie geometrica mostra che la PMF Geometrica sia valida.

Osservazione 150 (Interpretazione). La probabilità di avere x fallimenti consecutivi seguiti da un successo è data dalla probabilità di x fallimenti per la probabilità di un successo.

Definizione 5.6.2 (Funzione di ripartizione)**.** Si ha

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - (1-p)^{x+1} \tag{5.29}$$

Derivazione della CDF. Si ha

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - \mathbb{P}(X > x) = 1 - \sum_{k=x+1}^{\infty}(1-p)^k p$$

Espandendo la sommatoria:

$$\begin{aligned}\sum_{k=x+1}^{\infty}(1-p)^k p &= (1-p)^{x+1}\cdot p + (1-p)^{x+2}\cdot p + \ldots + (1-p)^{\infty}\cdot p \\ &= p(1-p)^x\left[(1-p)+(1-p)^2+\ldots+(1-p)^{\infty}\right] \\ &= p(1-p)^x\left[\sum_{i=1}^{\infty}(1-p)^i\right] \\ &= p(1-p)^x\left[\sum_{i=0}^{\infty}(1-p)^i - 1\right] \\ &= p(1-p)^x\left(\frac{1}{p}-1\right) = p(1-p)^x\frac{1-p}{p} \\ &= (1-p)^{x+1}\end{aligned}$$

Pertanto:

$$F_X(x) = 1-(1-p)^{x+1}$$

□

5.6.3 Moments

Proposizione 5.6.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{1-p}{p}$$
$$\mathrm{Var}[X] = \frac{1-p}{p^2}$$

Dimostrazione. Per il valore atteso abbiamo

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x\cdot(1-p)^x p$$

Non può essere ricondotta a serie geometrica direttamente per la presenza entro sommatoria di x come primo fattore. Ma notiamo che il termine entro sommatoria assomiglia a $x(1-p)^{x-1}$ ossia la derivata di $(1-p)^x$ rispetto a $1-p$, quindi partiamo da li:

$$\sum_{x=0}^{\infty}(1-p)^x = \frac{1}{p}$$

Questa serie converge dato che $0 < p < 1$. Derivando entrambi i membri rispetto a p.

$$\sum_{x=0}^{\infty} x(1-p)^{x-1}\cdot(-1) = -\frac{1}{p^2}$$
$$\sum_{x=0}^{\infty} x(1-p)^{x-1} = \frac{1}{p^2}$$

e se moltiplichiamo entrambi i lati per $p(1-p)$ otteniamo la somma dalla quale siamo partiti

$$p(1-p)\sum_{x=0}^{\infty} x(1-p)^{x-1} = \frac{1}{p^2}p(1-p)$$
$$\sum_{x=0}^{\infty} xp(1-p)^x = \frac{1-p}{p}$$

□

Dimostrazione. Per la varianza dobbiamo calcolare $\mathbb{E}\left[X^2\right]$:

$$\mathbb{E}\left[X^2\right] = \sum_{x=0}^{\infty} x^2 \cdot \mathbb{P}\left(X = x\right) = \sum_{x=0}^{\infty} x^2 \cdot (1-p)^x \cdot p \overset{(1)}{=} \sum_{x=1}^{\infty} x^2 \cdot (1-p)^x \cdot p$$

con (1) dato dal fatto che se $x = 0$ il termine entro sommatoria è nullo e si può portare avanti l'indice della stessa. Anche qui cerchiamo di sfruttare la serie geometrica per arrivare ad una espressione compatta equivalente all'ultimo termine di sopra. La serie è

$$\sum_{x=0}^{\infty}(1-p)^x = \frac{1}{p}$$

Derivando rispetto a p entrambi i membri, come visto in precedenza si ha:

$$\sum_{x=0}^{\infty} x \cdot (1-p)^{x-1} = \frac{1}{p^2}$$

Possiamo portare avanti di 1 l'indice di sommatoria dato che se $x = 0$ è nullo il termine dentro

$$\sum_{x=1}^{\infty} x \cdot (1-p)^{x-1} = \frac{1}{p^2}$$

Ora, derivando ancora si andrebbe a $x(x-1)$ entro sommatoria, invece di x^2 desiderato, pertanto moltiplichiamo per $(1-p)$ entrambi i membri giungendo a:

$$\sum_{x=1}^{\infty} x \cdot (1-p)^{x} = \frac{1-p}{p^2}$$

Derivando ambo i membri nuovamente rispetto a p si va a

$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} \cdot (-1) = \frac{(-1)\cdot p^2 - 2p\cdot(1-p)}{p^4}$$
$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} = (-1)\frac{p^2-2p}{p^4}$$
$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} = \frac{2-p}{p^3}$$

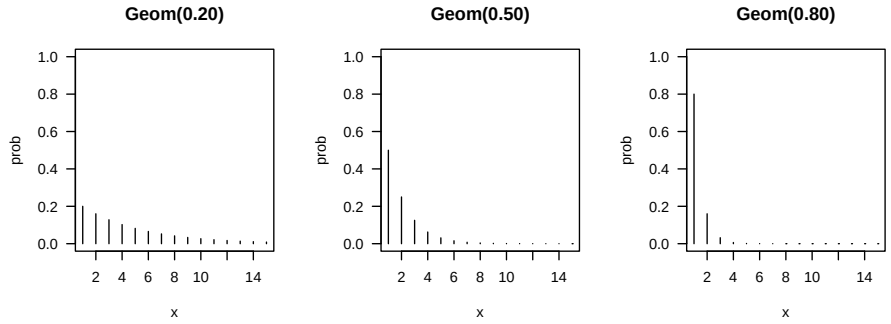


Figura 5.3: Forma distribuzione Geom (p)

Moltiplicando entrambi i membri per $(1-p)\cdot p$ si arriva al punto dove eravamo rimasti con $\mathbb{E}\left[X^2\right]$

$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^x \cdot p = \frac{2-p}{p^3} \cdot (1-p) \cdot p = \frac{(2-p)(1-p)}{p^2}$$

Per cui

$$\mathbb{E}\left[X\right] = \sum_{x=1}^{\infty} x^2 \cdot (1-p)^x \cdot p = \frac{(2-p)(1-p)}{p^2}$$

e dunque:

$$\begin{aligned}\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 &= \frac{(2-p)(1-p)}{p^2} - \frac{(1-p)^2}{p^2} \\ &= \frac{(1-p)(2-p-1+p))}{p^2} = \frac{1-p}{p^2}\end{aligned}$$

□

5.6.4 Shape

Osservazione 151 (Shape). Tutte le geometriche hanno forma simile: la funzione è decrescente, con probabilità più alte associate ai valori più piccoli di x. Ha asimmetria positiva che aumenta al crescere di p (più p è alto più velocemente la PMF discende verso 0). Ha una notevole curtosi (figura 5.3)

5.6.5 Assenza di memoria

Osservazione 152. Una proprietà peculiare della geometrica è di esser l'unica vc discreta senza memoria (a parte la sua riformulazione).

Proposizione 5.6.2 (Assenza di memoria)**.**

$$\mathbb{P}\left(X > t+s | X > t\right) = \mathbb{P}\left(X > s\right) \tag{5.30}$$

Dimostrazione. Si ha:

$$\begin{aligned}\mathbb{P}(X > t+s | X > t) &= \frac{\mathbb{P}(X > t+s)}{\mathbb{P}(X > t)} = \frac{1 - F_X(t+s)}{1 - F_X(t)} = \frac{1 - 1 + (1-p)^{t+s+1}}{1 - 1 + (1-p)^{t+1}} \\ &= (1-p)^s = 1 - F_X(s) = \mathbb{P}(X > s)\end{aligned}$$

□

5.6.6 Alternative definition (first success distribution)

Osservazione 153. Altri definiscono X come il numero di *prove* necessarie per ottenere il primo successo (incluso quest'ultimo). Qui la chiamiamo FS distribution e la indichiamo con $X \sim \mathrm{FS}(p)$

Osservazione 154. Se $Y \sim \mathrm{FS}(p)$ allora $Y - 1 \sim \mathrm{Geom}(p)$ e possiamo convertire tra le PMF di Y e $Y - 1$ scrivendo

$$\mathbb{P}(Y = k) = \mathbb{P}(Y - 1 = k - 1)$$

Viceversa se $X \sim \mathrm{Geom}(p)$ allora $X + 1 \sim \mathrm{FS}(p)$

Osservazione 155 (Supporto e spazio parametrico).

$$\begin{aligned}R_X &= \{x \in \mathbb{N} \setminus \{0\}\} \\ \boldsymbol{\Theta} &= \{p \in (0,1)\}\end{aligned}$$

Definizione 5.6.3 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = (1-p)^{x-1} p \cdot \mathbb{1}_{R_X}(x) \tag{5.31}$$

Osservazione 156 (Interpretazione). La probabilità di avere il primo successo all'n-esima estrazione e data dalla probabilità di $n - 1$ fallimenti per la probabilità di un successo.

Definizione 5.6.4 (Funzione di ripartizione)**.**

$$\begin{aligned}F_X(x) = \mathbb{P}(X \leq x) &= \sum_{k=1}^{x} \mathbb{P}(X = k) = \sum_{k=1}^{x} (1-p)^{k-1} p \\ &= 1 - (1-p)^x \end{aligned} \tag{5.32}$$

Proposizione 5.6.3 (Momenti caratteristici)**.**

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{p} \\ \mathrm{Var}[X] &= \frac{1-p}{p^2} \\ \mathrm{Asym}(X) &= \frac{2-p}{\sqrt{1-p}} \\ \mathrm{Kurt}(X) &= 9 + \frac{p^2}{1-p}\end{aligned}$$

Dimostrazione. Sia $Y = X + 1 \sim \mathrm{FS}(p)$ con $X \sim \mathrm{Geom}(p)$. Allora sfruttando le conoscenze sulla geometrica e le proprietà di valore atteso e varianza

$$\mathbb{E}[Y] = \mathbb{E}[X+1] = \mathbb{E}[X] + 1 = \frac{1-p}{p} + 1 = \frac{1}{p}$$
$$\mathrm{Var}[Y] = \mathrm{Var}[X+1] = \mathrm{Var}[X] = \frac{1-p}{p^2}$$

□

Proposizione 5.6.4 (Assenza di memoria)**.** *Analogamente a quanto avviene per la geometrica* $\mathbb{P}(X > t+s | X > t) = \mathbb{P}(X > s)$.

Dimostrazione. Si ha:

$$\begin{aligned}\mathbb{P}(X > t+s | X > t) &= \frac{\mathbb{P}(X > t+s)}{\mathbb{P}(X > t)} = \frac{1 - F_X(t+s)}{1 - F_X(t)} \\ &= \frac{(1-p)^{t+s}}{(1-p)^t} = (1-p)^s \\ &= \mathbb{P}(X > s)\end{aligned}$$

ovvero il ritardo accertato di un evento in t sottoprove indipendenti non modifica la probabilità che esso si verifichi entro ulteriori s sottoprove. □

5.7 Negative binomial

Osservazione 157. Generalizza la distribuzione Geometrica: invece di aspettare il primo successo conta i fallimenti prima di ottenere il k-esimo successo.

5.7.1 Definition

Definizione 5.7.1. In una sequenza di prove Bernoulliane indipendenti con probabilità di successo p, se X è il numero di fallimenti prima del k-esimo successo, allora X ha una distribuzione binomiale negativa con parametri k e p e si scrive $X \sim \mathrm{Nb}(k, p)$

Osservazione 158. Anche a livello di notazione, nei parametri, si nota subito la differenza con la binomiale: questa fissa il numero di trial mentre la binomiale negativa fissa il numero di successi.

5.7.2 Functions

Osservazione 159 (Supporto e spazio parametrico)*.*

$$\begin{aligned}R_X &= \mathbb{N} \\ \boldsymbol{\Theta} &= \{k \in \mathbb{N} : k \geq 1, p \in \mathbb{R} : 0 \leq p \leq 1\}\end{aligned}$$

Definizione 5.7.2 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \binom{x+k-1}{k-1} p^k (1-p)^x \cdot \mathbb{1}_{R_X}(x) \tag{5.33}$$

Osservazione 160 (Interpretazione). Ci sono $\binom{x+k-1}{k-1}$ sequenze possibili di x fallimenti e $k-1$ successi. Ciascuna di esse ha probabilità $p^{k-1}(1-p)^x$. Si termina con un success, quindi moltiplicando per p.

Osservazione 161. Come una binomiale può essere rappresentata da una somma di Bernoulli iid, una binomiale negativa può essere rappresentata come somma di Geometriche iid, come mostrato dal seguente teorema.

Proposizione 5.7.1. *Sia $X \sim \mathrm{Nb}(k,p)$ il numero di fallimenti prima del k-esimo successo in una sequenza di probe bernoulliane indipendenti con probabilità di successo p. Allora possiamo scrivere $X = X_1 + \ldots + X_k$ dove gli X_i sono iid e $X_i \sim \mathrm{Geom}(p)$.*

Dimostrazione. Sia X_1 il numero di fallimenti prima del primo successo, X_2 il numero di fallimenti tra il primo successo e il secondo e, in generale, X_i il numero di fallimenti tra $(i-1)$-esimo succeso e l'i-esimo.
Allora $X_1 \sim \mathrm{Geom}(p)$ per la definizione della geometrica, $X_2 \sim \mathrm{Geom}(p)$ e così via. Inoltre le X_i sono indipendenti dato che le prove bernoulliane sono indipendenti l'un l'altra. Sommando gli X_i si ottiene il totale di fallimenti prima del k-esimo successo, che è X. □

5.7.3 Moments

Proposizione 5.7.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = k\frac{1-p}{p} \tag{5.34}$$

$$\mathrm{Var}[X] = k\frac{1-p}{p^2} \tag{5.35}$$

Dimostrazione. Per il valore atteso sfruttiamo che X è scrivibile come somma di k vc Geometriche X_i. Il valore atteso è la somma dei valori attesi delle geometriche:

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \ldots + X_k] = \mathbb{E}[X_1] + \ldots \mathbb{E}[X_k] = k\frac{1-p}{p}$$

Per la varianza avviene lo stesso, dato che le variabili sono indipendenti:

$$\mathrm{Var}[X] = \mathrm{Var}[X_1 + \ldots + X_k] = \mathrm{Var}[X_1] + \ldots \mathrm{Var}[X_k] = k\frac{1-p}{p^2}$$

□

5.7.4 Shape

Osservazione 162 (Shape). Si nota che così al crescere di k, la distribuzione diviene più simmetrica e la curtosi tende a 3 indicando convergenza alla normalità. All'aumentare di p assume asimmetria positiva. (figura 5.4)

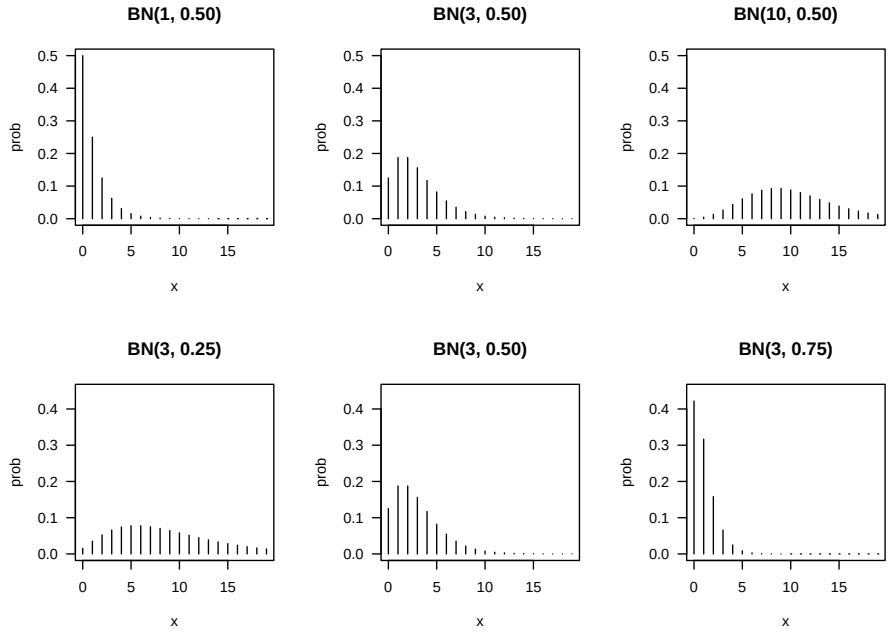


Figura 5.4: Distribuzione binomiale negativa

5.7.5 Alternative definition

5.7.5.1 Definition

Definizione 5.7.3 (Distribuzione binomiale negativa)**.** Il numero di prove indipendenti X (ciascuna con probabilità p di essere successo) necessarie per avere $k \geq 1$ successi si distribuisce come una binomiale negativa di parametri k e p, ossia $X \sim \mathrm{Nb}(k, p)$.

5.7.5.2 Functions

Osservazione 163 (Supporto e spazio parametrico)*.*

$$R_X = \{k, k+1, \ldots\}$$
$$\boldsymbol{\Theta} = \{k \in \mathbb{N} \setminus \{0\}, p \in \mathbb{R} : 0 \leq p \leq 1\}$$

Definizione 5.7.4 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \cdot \mathbb{1}_{R_X}(x) \tag{5.36}$$

Osservazione 164 (Interpretazione)*.* La formula deriva dalla considerazione che per ottenere il k-esimo successo nella n-esima prova, ci dovranno essere $k-1$ successi nelle prime $n-1$ prove, la cui probabilità

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

è moltiplicata per la probabilità di un successo nella n-esima, ossia p.

5.7.5.3 Moments

Proposizione 5.7.3 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{k}{p}$$
$$\text{Var}[X] = \frac{k(1-p)}{p^2}$$
$$\text{Asym}(X) = \frac{2-p}{\sqrt{k(1-p)}}$$
$$\text{Kurt}(X) = 3 + \frac{6}{k} + \frac{p^2}{k(1-p)}$$

5.8 Poisson

5.8.1 Definition

Osservazione 165. È una vc utilizzabile per modellare conteggi (motivo per cui il supporto è $\mathbb{N}$); sull'origine definizione ragioniamo in seguito. Per ora ci accontentiamo di definire la Poisson come la distribuzione caratterizzata dalle funzioni presentate in seguito: se la vc X è distribuita come una Poisson con parametro λ scriveremo $X \sim \text{Pois}(\lambda)$.

Osservazione 166. Un risultato che ci servirà per questa distribuzione è il seguente

Proposizione 5.8.1 (Sviluppo di Maclaurin della funzione esponenziale)**.**

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{5.37}$$

Dimostrazione. Si ha:

$$e^x = e^0 + \frac{e^0}{1!}(x-0) + \frac{e^0}{2!}(x-0)^2 + \ldots \frac{e^0}{m!}(x-0)^m + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

□

5.8.2 Functions

Osservazione 167 (Supporto e spazio parametrico)*.*

$$R_X = \mathbb{N}$$
$$\boldsymbol{\Theta} = \{\lambda \in \mathbb{R} : \lambda > 0\}$$

Definizione 5.8.1 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \frac{e^{(-\lambda)} \cdot \lambda^x}{x!} \cdot \mathbb{1}_{R_X}(x) \tag{5.38}$$

Validità PMF. Si ha:

$$\sum_{x=0}^{\infty} p_X(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \overset{(1)}{=} e^{-\lambda}e^{\lambda} = 1$$

dove in (1) abbiamo sfruttato la 5.37 con le dovute sostituzioni di lettere. □

5.8.3 Moments

Proposizione 5.8.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \lambda \tag{5.39}$$
$$\mathrm{Var}[X] = \lambda \tag{5.40}$$
$$\mathrm{Asym}(X) = \frac{1}{\sqrt{\lambda}} \tag{5.41}$$
$$\mathrm{Kurt}(X) = 3 + \frac{1}{\lambda} \tag{5.42}$$

Dimostrazione. Per il valore atteso

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} \overset{(1)}{=} e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$
$$\overset{(2)}{=} \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

dove in (1) abbiamo anche portato avanti di 1 la sommatoria dato che il primo termine è nullo e in (2) abbiamo sostituito $y = x - 1$ e sfruttato 5.37. □

Dimostrazione. Per la varianza troviamo innanzitutto $\mathbb{E}\left[X^2\right]$:

$$\mathbb{E}\left[X^2\right] = \sum_{x=0}^{\infty} x^2 \cdot \mathbb{P}(X = x) = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} x^2 \frac{\lambda^x}{x!}$$

Ora prendiamo la serie dell'esponenziale e la deriviamo rispetto a λ ad entrambi i membri (x costante)

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \overset{(1)}{=} \sum_{x=0}^{\infty} x \frac{\lambda^{x-1}}{x!} \overset{(2)}{=} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{x!}$$

dove in (1) abbiamo effettuato la derivazione (il primo membro rimane invariato), in (2) abbiamo portato avanti l'indice di sommatoria perché il primo termine è nullo. Ora moltiplicando per λ entrambi i lati si ottiene

$$\lambda e^{\lambda} = \sum_{x=1}^{\infty} x \frac{\lambda^x}{x!}$$

Effettuando gli stessi passaggi, nell'ordine derivare entrambi i membri rispetto a λ e moltiplicandoli per λ si prosegue come

$$\sum_{x=1}^{\infty} x^2 \frac{\lambda^{x-1}}{x!} = e^{\lambda} + \lambda e^{\lambda} = e^{\lambda}(1 + \lambda)$$
$$\sum_{x=1}^{\infty} x^2 \frac{\lambda^x}{x!} = e^{\lambda}\lambda(1 + \lambda)$$

E infine riprendendo da dove eravamo arrivati con la main quest

$$\mathbb{E}\left[X^2\right] = e^{-\lambda} \sum_{x=0}^{\infty} x^2 \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} \lambda(1 + \lambda) = \lambda(1 + \lambda)$$

per cui

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \lambda(1+\lambda) - \lambda^2 = \lambda$$

□

Dimostrazione. Dimostrazione alternativa per la varianza:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[X^2\right] - [\mathbb{E}[X]]^2 \\
&= \left(\sum_{x=0}^{\infty} x^2 \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x^2 + x - x) \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x(x-1) + x) \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x(x-1)) \frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda\lambda^{x-1}}{x(x-1)!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} e^{-\lambda}\lambda^2 + \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda}\lambda\right) - \lambda^2 \\
&\overset{(1)}{=} \left(\sum_{z=0}^{\infty} \frac{\lambda^z}{z!} e^{-\lambda}\lambda^2 + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda}\lambda\right) - \lambda^2 \\
&= (e^{\lambda}e^{-\lambda}\lambda^2 + e^{\lambda}e^{-\lambda}\lambda) - \lambda^2 \\
&= (\lambda^2 + \lambda) - \lambda^2 \\
&= \lambda
\end{aligned}$$

dove in (1) abbiamo posto $y = x-1$, $z = x-2$ per sfruttare 5.37 nel seguito. □

5.8.4 Shape

Osservazione 168 (Shape). Quindi valore medio e varianza della vc di Poisson coincidono con il parametro λ; la distribuzione ha picco intorno a λ. Al crescere di questo, la distribuzione diventa più simmetrica e la curtosi tende a 3 (convergendo ad una Normale). Se $\lambda < 1$ la distribuzione ha un andamento decrescente, mentre se > 1 è prima crescente e poi decrescente. (figura 5.5)

5.8.5 Origine e approssimazione

Osservazione 169. È utilizzata per modellare il numero di eventi registrati in un ambito circoscritto (temporale o spaziale), in cui vi è un largo numero di prove indipendenti (o quasi) caratterizzate ciascuna da una bassa probabilità di successo (per questa è chiamata legge degli eventi rari)

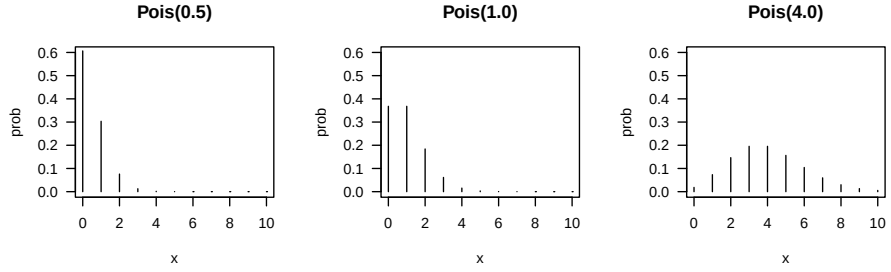


Figura 5.5: Distribuzione Poisson

Proposizione 5.8.3 (Paradigma di Poisson)**.** *Siano $E_1, \ldots, E_n$ eventi con $p_i = \mathbb{P}(E_i)$, dove n è largo, p_i sono piccoli e gli E_i sono vc indipendenti o debolmente dipendenti. Sia*

$$X = \sum_{i=1}^{n} I_{E_i}$$

la somma di quanti eventi E_i siano accaduti. Allora X è abbastanza bene distribuita come una Pois (λ) *con $\lambda = \sum_i p_i$.*

Dimostrazione. La prova dell'approssimazione di sopra è complessa, richiede definire la dipendenza debole e buona approssimazione; è omessa qui. □

Osservazione 170 (Ruolo di λ). Il parametro λ è interpretato come *rate di occorrenza*: ad esempio $\lambda = 2$ mail di spam per giorno.

Osservazione 171. Nell'esempio sopra il numero di eventi X non è esattamente distribuito come Poisson perché una variabile di Poisson non ha limite superire, mentre $I_{E_1} + \ldots + I_{E_n}$ somma al più a n. Ma la distribuzione di Poisson da spesso una buona approssimazione e le condizioni per il verificarsi della situazione di sopra sono abbastanza flessibili: infatti i p_i non devono essere uguali e le prove non devono essere strettamente indipendenti. Questo fa sì che il modello di Poisson sia spesso un buon punto di partenza per dati che assumono valore intero non negativo (chiamati conteggi)
È comunque possibile quantificare l'errore commesso.

Proposizione 5.8.4 (Errore di approssimazione)**.** *Se E_i sono indipendenti e sia $N \sim$ Pois (λ), allora l'errore di approssimazione che si fa nell'utilizzare la poisson per stimare la probabilità di un dato set di interi non negativi $I \subset \mathbb{N}$, è dato dalla seguente:*

$$\mathbb{P}(X \in I) - \mathbb{P}(N \in I) \leq \min\left(1, \frac{1}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \tag{5.43}$$

Dimostrazione. Anche questa è per ora complessa (necessita di una tecnica chiamata metodo di Stein). □

Osservazione 172. La 5.43 fornisce un limite superiore dell'errore commesso nell'utilizzare unaapprossimazione di Poisson: non solo per l'intera distribuzione

(se $I = \mathbb{N}$) ma per qualsiasi suo sottoinsieme. Altresì precisa quanto i p_i dovrebbero essere piccoli: vogliamo che $\sum_{i=1}^{n} p_i^2$ sia molto piccolo, o quanto meno lo sia rispetto a λ.

5.8.6 Legami con la binomiale

Osservazione 173. La relazione tra Poisson e Binomiale è simile a quella intercorrente tra Binomiale e Ipergeometrica: possiamo andare dalla Poisson alla binomiale condizionando, e viceversa dalla Binomiale alla Poisson prendendo un limite. Prima un risultato strumentale.

Proposizione 5.8.5 (Somma di Poisson indipendenti)**.** *Siano* $X \sim \text{Pois}(\lambda_1)$ *e* $Y \sim \text{Pois}(\lambda_2)$ *vc indipendenti. Allora* $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$

Dimostrazione. Per ottenere la PMF di $X+Y$ condizioniamo su X e utilizziamo il teorema delle probabilità totali

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{j=0}^{k} \mathbb{P}(X+Y=k|X=j)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \mathbb{P}(Y=k-x|X=j)\cdot\mathbb{P}(X=j) \\
&\overset{(1)}{=} \sum_{j=0}^{k} \mathbb{P}(Y=k-x)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!}\frac{e^{-\lambda_1}\lambda_1^{j}}{(j)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda_1^j\lambda_2^{k-j} \\
&\overset{(2)}{=} \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k)}{k!} = \text{Pois}(\lambda_1+\lambda_2)
\end{aligned}$$

con (1) data l'indipendenza e in (2) si è utilizzato il teorema binomiale $(a+b)^n = \sum_{i=0}^{n}\binom{n}{i}a^i b^{n-i}$ □

Osservazione 174. A intuito se vi sono due tipi di eventi che accadono ai rate λ_1 e λ_2 indipendentemente, allora il rate complessivo di eventi è $\lambda_1 + \lambda_2$.

5.8.6.1 Dalla Poisson alla binomiale

Proposizione 5.8.6. *Se* $X \sim \text{Pois}(\lambda_1)$ *e* $Y \sim \text{Pois}(\lambda_2)$ *sono indipendenti, allora la ditribuzione condizionata di* X *dato che* $XY = n$ *è* $\text{Bin}(n, \lambda_1/(\lambda_1\lambda_2))$.

Dimostrazione. Utilizziamo la regola di Bayes per calcolare la PMF condizionata $\mathbb{P}(X = x|X + Y = n)$:

$$\begin{aligned}\mathbb{P}(X = x|X + Y = n) &= \frac{\mathbb{P}(X + Y = n|X = x) \cdot \mathbb{P}(X = x)}{\mathbb{P}(X + Y = n)} \\ &= \frac{\mathbb{P}(Y = n - x|X = x) \cdot \mathbb{P}(X = x)}{\mathbb{P}(X + Y = n)} \\ &\overset{(1)}{=} \frac{\mathbb{P}(Y = n - x) \cdot \mathbb{P}(X = x)}{\mathbb{P}(X + Y = n)}\end{aligned}$$

con (1) per indipendenza delle due. Ora sostituendo le PMF di X, Y e $X + Y$; questa al denominatore è distribuita come $\text{Pois}(\lambda_1 + \lambda_2)$ per proposizione 5.8.5. Si ha:

$$\begin{aligned}\mathbb{P}(X = k|X + Y = n) &= \frac{\left(\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\right)\left(\frac{e^{\lambda_1}\lambda_2^{k}}{k!}\right)}{\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}} = \frac{\frac{e^{-(\lambda_1+\lambda_2)}\cdot\lambda_1^k\cdot\lambda_2^{n-k}}{k!(n-k)!}}{\frac{e^{-(\lambda_1+\lambda_2)}\cdot(\lambda_1+\lambda_2)^n}{n!}} \\ &= \frac{e^{-(\lambda_1+\lambda_2)} \cdot \lambda_1^k \cdot \lambda_2^{n-k}}{k!(n-k)!} \cdot \frac{n!}{e^{-(\lambda_1+\lambda_2)} \cdot (\lambda_1 + \lambda_2)^n} \\ &= \frac{n!}{k!(n-k)!} \cdot \frac{\lambda_1^k \cdot \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1^k}{\lambda_1 + \lambda_2}\right)^k\left(\frac{\lambda_2^k}{\lambda_1 + \lambda_2}\right)^{n-k} \\ &= \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)\end{aligned}$$

□

5.8.6.2 Dalla binomiale alla Poisson

Osservazione 175. Viceversa se prendiamo il limite della $\text{Bin}(n, p)$ per $n \to \infty$ e $p \to 0$ con np fisso arriviamo alla Poisson.

Proposizione 5.8.7 (Approssimazione Poissoniana della binomiale)**.** *Se* $X \sim \text{Bin}(n, p)$ *e facciamo tendere* $n \to \infty$, $p \to 0$ *ma* $\lambda = np$ *rimane fisso, allora la PMF di* X *converge a* $\text{Pois}(\lambda)$.
La stessa conclusione si ha se $n \to \infty$, $p \to 0$ *ed* np converge *ad una costante* λ.

Osservazione 176. Questo è un *caso speciale* del paradigma di Poisson dove E_i sono indipendenti e hanno la stessa probabilità, quindi $\sum_{i=1}^n I_{E_i}$ ha distribuzione binomiale. In questo caso speciale possiamo dimostrare che l'approssimazione di Poisson ha senso limitandoci a prendere il limite della Binomiale.

Dimostrazione. Effettueremo la dimostrazione per $\lambda = np$ fisso (considerando $p = \lambda/n$), mostrando che la PMF $\text{Bin}(n, p)$ converge alla $\text{Pois}(\lambda)$. Per $0 \leq x \leq$

n:

$$\begin{aligned}\mathbb{P}(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\cdot\ldots\cdot(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{\lambda^x}{x!}\frac{n(n-1)\cdot\ldots\cdot(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\end{aligned}$$

Per $n\to\infty$ con k fisso

$$\begin{aligned}&\frac{\overbrace{n(n-1)\cdot\ldots\cdot(n-x+1)}^{x \text{ termini}}}{n^x} \overset{(1)}{=} \frac{n\cdot n\left(1-\frac{1}{n}\right)\cdot\ldots\cdot n\left(1-\frac{k-1}{n}\right)}{n^x} \to 1 \\ &\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda} \\ &\left(1-\frac{\lambda}{n}\right)^{-k} = \left[\left(1-\frac{\lambda}{n}\right)^n\right]^{-\frac{k}{n}} \to e^{-\frac{k}{n}} = 1\end{aligned}$$

dove in (1) abbiamo raccolto un n a partire dal secondo fattore, lasciando fuori parentesi k n che si moltiplicano. Pertanto

$$\mathbb{P}(X=x) \to \frac{e^{-\lambda}\lambda^x}{x!} = \text{Pois}(\lambda)$$

□

Osservazione 177. Il precedente risultato implica che se n è grande, p piccolo e np moderato, possiamo approssimare $\text{Bin}(n,p)$ con $\text{Pois}(np)$; come visto in precedenza l'errore nell'approssimare $\mathbb{P}(X\in I)$ con $\mathbb{P}(N\in I)$ per $X\sim\text{Bin}(n,p)$ e $N\sim\text{Pois}(np)$ è al massimo $\min(p,np^2)$.

Esempio 5.8.1. Il proprietario di un sito vuole studiare la distribuzione del numero di visitatori. Ogni giorno un millione di persone in maniera indipendente decide se visitare il sito o meno, con probabilità $p=2\times10^{-1}$. Fornire una approssimazione della probabilità di avere almeno tre visitatori al giorno.
Se $X\sim\text{Bin}(n,p)$ è il numero di visitatori con $n=10^6$, fare i calcoli con la binomiale va incontro a difficoltà computazionali ed errori numerici del pc (dato che n è largo e p molto basso). Ma data la situazione con n largo p basso e $np=2$ moderato, $\text{Pois}(2)$ è una buona approssimazione. Questo porta a

$$\mathbb{P}(X\geq3) = 1-\mathbb{P}(X<3) \approx 1-e^{-2}-e^{-2}\cdot2-e^{-2}\cdot\frac{2^2}{2!} = 1-5e^{-2} \approx 0.3233$$

che è una approssimazione molto accurata.

5.8.7 Processo di Poisson

Definizione 5.8.2 (Processo di Poisson)**.** È una insieme di prove E_i che si possono verificare ciascuna in un dato arco temporale $[0,T]$. Le prove sono svolte nelle medesime condizioni e soddisfano di assiomi:

- il verificarsi di E nell'intervallo (t_1, t_2) è indipendente dal verificarsi di E nell'intervallo (t_3, t_4) (se gli intervalli non si sovrappongono);
- la probabilità del verificarsi di E in un intervallo infinitesimo $(t_0, t_0 + \mathrm{d}t)$ è proporzionale ad un parametro $\lambda > 0$ che caratterizza la prova;
- la probabilità che due eventi si verifichino nello stesso intervallo di tempo è un infinitesimo di ordine superiore rispetto alla probabilità che se ne verifichi soltanto uno.

5.9 Discrete uniform

5.9.1 Definition

Osservazione 178. La prova che genera la vc Uniforme discreta si può assimilare all'estrazione di una pallina da un'urna che contiene n palline identiche numerate da 1 a n. Viene in genere utilizzata quanto tutti i risultati dell'esperimento sono equiprobabili

Definizione 5.9.1 (Uniforme discreta)**.** Il numero X della pallina estratta dall'urna contenente n palline numerate (da 1 a n) si distribuisce come Uniforme discreta $X \sim \mathrm{DUnif}(n)$.

5.9.2 Functions

Osservazione 179 (Supporto e spazio parametrico)*.*

$$R_X = \{1, \dots, n\}$$
$$\boldsymbol{\Theta} = \{n \in \mathbb{N} \setminus \{0\}\}$$

Proposizione 5.9.1 (Funzione di massa di probabilità)**.**

$$p_X(x) = \mathbb{P}(X = x) = \frac{1}{n} \cdot \mathbb{1}_{R_X}(x) \tag{5.44}$$

Definizione 5.9.2 (Funzione di ripartizione)**.**

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{se } x < 1 \\ \frac{k}{n} & \text{se } k \leq x < k+1, (k = 1, 2, \dots, n-1) \\ 1 & \text{se } x \geq n \end{cases} \tag{5.45}$$

Osservazione 180. La funzione di ripartizione è nulla in $(\infty; 1)$ ed è una funzione a gradini di altezza costante pari a $1/n$, in corrispondenza di ogni valore intero $1 \leq x \leq n$ e vale 1 in $[n; +\infty)$.

5.9.3 Moments

Proposizione 5.9.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{n+1}{2} \tag{5.46}$$

$$\mathrm{Var}[X] = \frac{n^2-1}{12} \tag{5.47}$$

$$\mathrm{Asym}(X) = 0 \tag{5.48}$$

$$\mathrm{Kurt}(X) = 1.8 \tag{5.49}$$

Dimostrazione.

$$\mathbb{E}[X] = \sum_{x=1}^{n} x\frac{1}{n} = \frac{1}{n}(1+2+\ldots+n) = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

□

Dimostrazione.

$$\begin{aligned}
\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \left[\mathbb{E}[x]\right]^2 &= \left(\sum_{x=1}^{n} x^2\frac{1}{n}\right) - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{1}{n}(1^2+2^2+\ldots+n^2)\right) - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6}\right) - \left(\frac{n^2+1+2n}{4}\right) \\
&= \left(\frac{(n+1)(2n+1)}{6}\right) - \left(\frac{n^2+1+2n}{4}\right) \\
&= \frac{2(2n^2+2n+n+1) - 3(n^2+1+2n)}{12} \\
&= \frac{4n^2+4n+2n+2-3n^2-3-6n}{12} \\
&= \frac{n^2-1}{12}
\end{aligned}$$

□

Capitolo 6

Variabili casuali continue

6.1 Logistica

6.1.1 Origine/definizione

Osservazione 181. Viene utilizzata per modelli di crescita di grandezze nel tempo, dove la crescita segue le fasi di crescita esponenziale, saturazione e arresto. Un buon modello per rappresentare fenomeni di questo tipo è rappresentato dalla funzione di ripartizione logistica.

Osservazione 182. Deriva il nome dall'avere la funzione di ripartizione che soddisfa l'equazione logistica: $F'(x) = \frac{1}{s}F(x)(1 - F(x))$.

Osservazione 183. E' matematicamente semplice e ci permette di focalizzarci su aspetti non numerici; è altresì importante nella regressione logistica.

6.1.2 Funzioni

Definizione 6.1.1 (Funzione di ripartizione)**.** Ha CDF

$$F_X(x) = \mathbb{P}(X \leq x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}, \qquad x \in \mathbb{R} \tag{6.1}$$

Osservazione 184. Si trovano entrambe le definizioni (si passa dall'una all'altra moltiplicando/dividendo a numeratore e denominatore per e^x)

Definizione 6.1.2 (Funzione di densità)**.** Derivando entrambe le espressioni si hanno, equivalentemente:

$$f_x(x) = \frac{e^x}{(1 + e^x)^2} = \frac{e^{-x}}{(1 + e^{-x})^2} \tag{6.2}$$

6.1.3 Versione generale

Osservazione 185 (Supporto e spazio parametrico).

$$R_X = \mathbb{R}$$
$$\boldsymbol{\Theta} = \{\mu \in \mathbb{R}, s \in \mathbb{R} : s > 0\}$$

Definizione 6.1.3 (Funzione di ripartizione)**.** La funzione di densità di una vc $X \sim \text{Logistic}(\mu, \sigma)$ è

$$F_X(x) = \frac{e^{\frac{x-\mu}{\sigma}}}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)} \cdot \mathbb{1}_{R_X}(x) \tag{6.3}$$

Definizione 6.1.4 (Funzione di densità)**.** La funzione di densità di una vc $X \sim \text{Logistic}(\mu, \sigma)$ è

$$f_X(x) = \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} \cdot \mathbb{1}_{R_X}(x) \tag{6.4}$$

Proposizione 6.1.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \mu$$
$$\text{Var}[X] = \frac{\pi^2}{3}\sigma^2$$

TODO: perché la varianza non è σ^2 applicando le regole su trasf lineari?

Mia dimostrazione, controllare. Sia $Z \sim \text{Logistic}(0, 1)$ e sia $X = \sigma Z + \mu$, con σ parametro di scala e μ di posizione. Allora si ha che

$$Z = \frac{X - \mu}{\sigma} \sim \text{Logistic}(0, 1)$$

Per cui possiamo scrivere che

$$F_X(x) = \frac{e^{\frac{x-\mu}{\sigma}}}{1 + e^{\frac{x-\mu}{\sigma}}}$$

Derivando per ottenere $f_X(x)$ si ha

$$f_X(x) = \frac{\left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(1 + e^{\frac{x-\mu}{\sigma}}\right) - \left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(e^{\frac{x-\mu}{\sigma}}\right)}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} = \frac{\left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(1 + e^{\frac{x-\mu}{\sigma}} - e^{\frac{x-\mu}{\sigma}}\right)}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2}$$
$$= \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2}$$

□

6.2 Uniforme continua

Osservazione 186. È una vc continua X definita sul supporto (a, b), con $a < b$ ed ed esiti aventi la medesima densità, indicata con $X \sim \text{Unif}(a, b)$

Osservazione 187. Una formulazione usuale per tale modello probabilistico è la uniforma continua sull'intervallo con $a = 0, b = 1$.

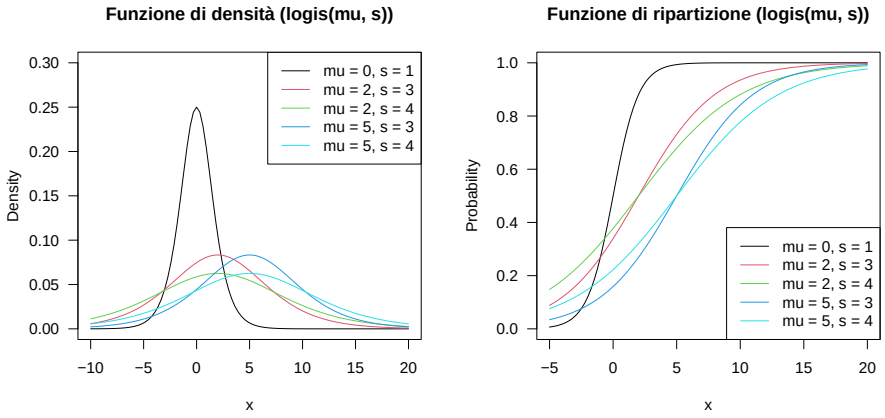


Figura 6.1: Distribuzione logistica

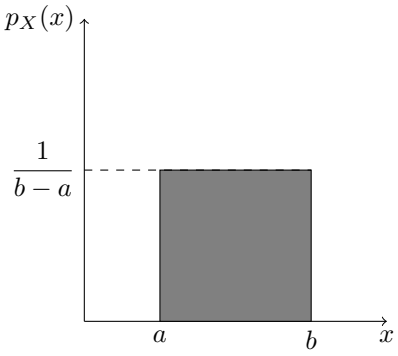


Figura 6.2: Uniforme continua

Osservazione 188 (Supporto e spazio parametrico).

$$R_X = [a, b]$$
$$\boldsymbol{\Theta} = \{a, b \in \mathbb{R}, a < b\}$$

Definizione 6.2.1 (Funzione di densità)**.** In figura 6.2

$$f_X(x) = \frac{1}{b-a} \cdot \mathbb{1}_{R_X}(x) \tag{6.5}$$

Proposizione 6.2.1. *L'area è 1.*

Dimostrazione.

$$(b-a) \cdot \frac{1}{(b-a)} = 1$$

□

Definizione 6.2.2 (Funzione di ripartizione)**.**

$$F_X(x) = \begin{cases} 0 & \text{per } x \leq a \\ \dfrac{x-a}{b-a} & \text{se } a < x < b \\ 1 & \text{per } x \geq b \end{cases} \tag{6.6}$$

Proposizione 6.2.2 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{a+b}{2} \tag{6.7}$$

$$\mathrm{Var}[X] = \frac{(b-a)^2}{12} \tag{6.8}$$

$$\mathrm{Asym}(X) = 0 \tag{6.9}$$

$$\mathrm{Kurt}(X) = 1.8 \tag{6.10}$$

Dimostrazione.

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b x \frac{1}{b-a}\,\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b \\ &= \left(\frac{b^2}{2(b-a)} + c\right) - \left(\frac{a^2}{2(b-a)} + c\right) \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2} \end{aligned}$$

□

Dimostrazione.

$$\begin{aligned} \mathrm{Var}[X] &= \left(\int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \left[\frac{x^3}{3(b-a)}\right]_a^b - \left(\frac{a+b}{2}\right)^2 \\ &= \left(\frac{b^3}{3(b-a)} + c\right) - \left(\frac{a^3}{3(b-a)} + c\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\ &= \frac{(b-a)(a^2+b^2+ab)}{3(b-a)} - \frac{(a+b)^2}{4} \\ &= \frac{a^2+b^2+ab}{3} - \frac{a^2+b^2+2ab}{4} \\ &= \frac{4a^2+4b^2+4ab-3a^2-3b^2-6ab}{12} \\ &= \frac{a^2+b^2-2ab}{12} = \frac{(a-b)^2}{12} = \frac{(b-a)^2}{12} \end{aligned}$$

□

Osservazione 189. Si tratta di una variabile simmetrica e platicurtica (ovvero con una distribuzione molto piatta).

6.3 Esponenziale

Osservazione 190. L'esponenziale è generalmente usata per fenomeni di cui interessa un tempo/durata t (di vita, resistenza, funzionamento).
La derivazione può avvenire se si ipotizza una funzione di rischio/azzardo costante $H(t) = \lambda > 0$, con λ tasso di occorrenza dell'evento (reciproco del numero di eventi per unità di tempo).

Osservazione 191 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{\lambda \in \mathbb{R} : \lambda > 0\}$$

Definizione 6.3.1 (Distribuzione esponenziale)**.** Se $H(t) = \lambda > 0$ la funzione di ripartizione si ricava dalla 4.32 come

$$\begin{aligned} F_X(t) &= 1 - \exp\left(-\int_0^t H(w)\,\mathrm{d}w\right) = 1 - \exp\left(-\int_0^t \lambda\,\mathrm{d}w\right) \\ &= 1 - \exp\left(-\lambda t\right) \end{aligned}$$

Definizione 6.3.2 (Funzione di ripartizione)**.**

$$F_X(x) = \begin{cases} 1 - \exp(-\lambda x) & \text{per } x \geq 0 \\ 0 & \text{per } x < 0 \end{cases} \tag{6.11}$$

Osservazione 192. La funzione di densità si ottiene derivando dalla 6.11; pertanto una vc continua X si dice vc Esponenziale con parametro $\lambda > 0$, e si scrive $X \sim \mathrm{Exp}\,(\lambda)$ se caratterizzata dalla seguente funzione di densità.

Definizione 6.3.3 (Funzione di densità)**.**

$$f_X(x) = \lambda \exp(-\lambda x) \cdot \mathbb{1}_{R_X}(x) \tag{6.12}$$

Proposizione 6.3.1 (Momenti caratteristici)**.**

$$\mathbb{E}\,[X] = \frac{1}{\lambda} \tag{6.13}$$
$$\mathrm{Var}\,[X] = \frac{1}{\lambda^2} \tag{6.14}$$
$$\mathrm{Asym}\,(X) = 2 \tag{6.15}$$
$$\mathrm{Kurt}\,(X) = 9 \tag{6.16}$$

Osservazione 193 (Forma distribuzione). Tale funzione è decrescente a partire da $x = 0$, in corrispondenza del quale si registra la moda; è asimmetrica positiva e fortemente leptocurtica (a punta), con asimmetria e curtosi costanti al variare di λ. (figura 6.3)

Osservazione 194. La vc Esponenziale presenta una struttura molto semplice ma rigida, per cui non si adatta facilmente a tutte le situazioni reali; infatti, talvolta non è realistico assumere che la funzione di rischio si costante rispetto al tempo. Pertanto si hanno almeno due generalizzazioni: la Weibull e la Gamma.

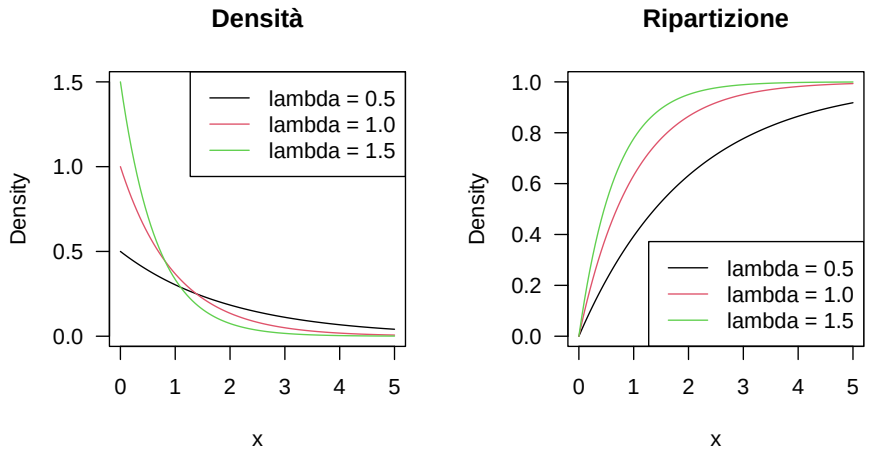


Figura 6.3: Distribuzione esponenziale

6.4 Normale/Gaussiana

Osservazione 195. Viene utilizzata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio.

Osservazione 196. Una vc continua si dice vc Normale con parametri μ e σ^2, e la si indica con $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ se è definita su tutto l'asse reale e presenta la seguente funzione di densità.

Osservazione 197 (Supporto e spazio parametrico).

$$R_X = \{\mathbb{R}\}$$
$$\boldsymbol{\Theta} = \left\{\mu \in \mathbb{R}; \sigma^2 \in \mathbb{R} : \sigma^2 > 0\right\}$$

Definizione 6.4.1 (Funzione di densità)**.**

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \cdot \mathbb{1}_{R_X}(x) \tag{6.17}$$

Osservazione 198 (Forma della distribuzione). Ha una forma campanulare e simmetrica rispetto al punto di ascissa $x = \mu$, è crescente in $(-\infty, \mu)$ e decrescente in (μ, ∞). In corrispondenza di μ $f_X(x)$ ha il massimo (perché l'esponente negativo è minimo). Pertanto μ è il valore centrale la moda, mediana e valore medio della vc.
Si dimostra che $f_X(x)$ presenta due flessi in corrispondenza di $x = \mu \pm \sigma$. Ha come asintoto l'asse x
μ è un parametro di posizione mentre σ^2 misura la dispersione attorno a μ. La modifica di μ a parità di σ^2 implica una traslazione della funzione di densità lungo l'asse x; invece, al crescere di σ a parità di μ, i flessi si allontanano da μ e la funzione di den attribuisce maggiore probabilità ai valori lontani dal valore centrale (e viceversa al diminuire di σ^2). (figura 6.4)

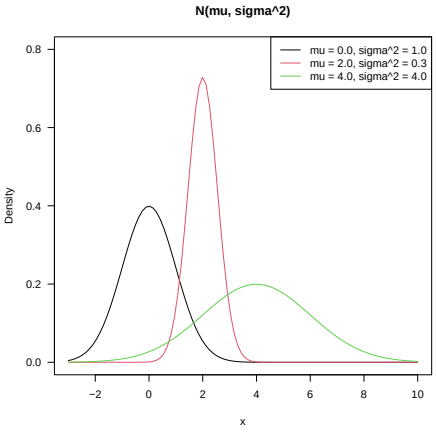


Figura 6.4: Distribuzione normale

Definizione 6.4.2 (Normale standardizzata)**.** Se $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, la trasformazione lineare $Z = (X - \mu)/\sigma$ definisce la vc Normale standardizzata $Z \sim \mathrm{N}\left(0, 1\right)$

Definizione 6.4.3 (Funzione di densità (Normale standardizzata))**.**

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \cdot \mathbb{1}_{R_X}(x) \tag{6.18}$$

Definizione 6.4.4 (Funzione di ripartizione (Normale standardizzata))**.**

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) \mathrm{d}w \tag{6.19}$$

Osservazione 199. La funzione di ripartizione della vc Z non ammette una formulazione esplicita ed è necessario predisporre delle tavole che per opportuni valori di z forniscano l'integrale con sufficiente accuratezza.

Osservazione 200. Sfruttando la simmetria della funzione di densità, è sufficiente conoscere $\Phi(z)$ per i soli valori di $z > 0$. Infatti $\Phi(0) = 0.5$ ed inoltre:

$$\Phi(-z) = 1 - \Phi(z) \quad \forall z \geq 0 \tag{6.20}$$

Osservazione 201. La conoscenza della funzione di ripartizione della vc $Z \sim \mathrm{N}\left(0, 1\right)$ è sufficiente per calcolare la probabilità di qualsiasi vc $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ mediante una semplice trasformazione:

$$\begin{aligned}
\mathbb{P}\left(x_0 < X \leq x_1\right) &= \mathbb{P}\left(\frac{x_0 - \mu}{\sigma} < \underbrace{\frac{X - \mu}{\sigma}}_{Z} \leq \frac{x_1 - \mu}{\sigma}\right) \\
&= \Phi\left(\frac{x_1 - \mu}{\sigma}\right) - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)
\end{aligned}$$

In pratica per calcolare la probabilità ce una vc normale assuma valori in un intervallo basta standardizzare gli estremi dell'intervallo ed utilizzare le tavole di $\Phi(z)$.

Proposizione 6.4.1 (Momenti caratteristici (Normale standardizzata))**.**

$$\mathbb{E}[Z] = 0 \tag{6.21}$$
$$\mathrm{Var}[Z] = 1 \tag{6.22}$$
$$\mathrm{Asym}(Z) = 0 \tag{6.23}$$
$$\mathrm{Kurt}(Z) = 3 \tag{6.24}$$

Proposizione 6.4.2 (Momenti caratteristici (Normale))**.** *Da $X = \mu + \sigma Z$ si ha*

$$\mathbb{E}[X] = \mu \tag{6.25}$$
$$\mathrm{Var}[X] = \sigma^2 \tag{6.26}$$
$$\mathrm{Asym}(X) = 0 \tag{6.27}$$
$$\mathrm{Kurt}(X) = 3 \tag{6.28}$$

Osservazione 202. Nel prosieguo tratteremo della vc Normale standardizzata, per semplicità.

Proposizione 6.4.3. *Se $X_i \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, allora:*

$$\sum_{i=1}^{n} a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

Osservazione 203. La famiglia delle vc normali è chiusa rispetto ad ogni combinazione lineare: in particolare la combinazione lineare di vc normali e indipendenti è ancora una vc normale che ha per valore medio la combinazione lineare dei valori medi e per varianza la combinazione lineare delle varianze con i quadrati dei coefficienti (proprietà riproduttiva della vc normale).

6.5 Gamma

Osservazione 204. Viene utilizzata quando si deve verificare la lunghezza dell'intervallo di tempo fino all'istante in cui si verifica la n-esima manifestazione di un evento aleatorio di interesse.
Similmente alla Beta è chiamata cosi perché coinvolge l'omonima funzione matematica.

Osservazione 205 (Supporto e spazio parametrico)*.*

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{n, \lambda \in \mathbb{R} : n, \lambda > 0\}$$

Definizione 6.5.1 (Funzione di densità)**.** Una vc continua X si distribuisce come una Gamma con parametri $n > 0, \lambda > 0$, indicata con $X \sim \mathrm{Gamma}(n, \lambda)$, se presenta una funzione di densità come la:

$$f_X(x) = \frac{\lambda^n}{\Gamma(n)} \cdot x^{n-1} \exp(-\lambda x) \cdot \mathbb{1}_{R_X}(x) \tag{6.29}$$

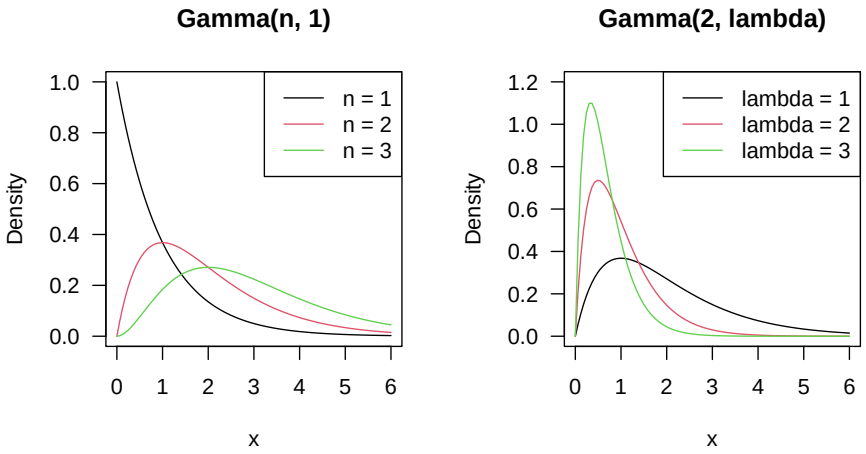


Figura 6.5: Distribuzione gamma

Definizione 6.5.2 (Funzione Gamma)**.** È definita come

$$\Gamma(n) = \int_0^{+\infty} x^{n-1} e^{-x} \, \mathrm{d}x \tag{6.30}$$

e presenta le seguenti proprietà: se $n \in \mathbb{R}, n > 1$, $\Gamma(n) = (n-1)\Gamma(n-1)$ (ossia è ricorsiva); se $n \in \mathbb{N} \setminus \{0\}$, $\Gamma(n) = (n-1)!$; ha valore notevole $\Gamma(1/2) = \sqrt{\pi}$.

Osservazione 206 (Funzione di ripartizione). Non si può definire una funzione di ripartizione perché questa dipende dalla funzione Γ (a meno che n sia intero).

Proposizione 6.5.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{n}{\lambda} \tag{6.31}$$

$$\mathrm{Var}[X] = \frac{n}{\lambda^2} \tag{6.32}$$

$$\mathrm{Asym}(X) = \frac{2}{\sqrt{n}} \tag{6.33}$$

$$\mathrm{Kurt}(X) = 3 + \frac{6}{n} \tag{6.34}$$

Osservazione 207 (Forma della distribuzione). λ è un parametro di scala mentre n determina la forma della distribuzione. All'aumentare del parametro λ la distribuzione si concentra sui valori più piccoli. Quando $n \to \infty$ la distribuzione diviene simmetrica e di forma campanulare (curtosi pari a 3). (figura 6.5)

Osservazione 208. Si nota che se $n = 1$, la distribuzione gamma diviene una esponenziale, ovvero $\mathrm{Gamma}(1, \lambda) \sim \mathrm{Exp}(\lambda)$; pertanto la gamma è una generalizzazione della esponenziale.

Osservazione 209. Altro caso particolare, se $n = \frac{\nu}{2}$ (con $\nu \in \mathbb{N} \setminus \{0\}$, numero dei gradi di libertà) e $\lambda = \frac{1}{2}$ la distribuzione Gamma coincide con la Chi-quadrato.

Proposizione 6.5.2. *La gamma gode della proprietà riproduttiva nel senso che la somma di gamma indipendenti ancora una gamma:*

$$\sum \text{Gamma}\left(n_i, \lambda\right) \sim \text{Gamma}\left(\sum_i n_i, \lambda\right) \tag{6.35}$$

6.6 Chi-quadrato

Osservazione 210. La somma di ν vc normali standardizzate indipendenti ed elevate al quadrato è una vc continua sul supporto $(0, +\infty)$ che si distribuisce come una vc Chi-quadrato con ν gradi di libertà

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi_\nu^2 \tag{6.36}$$

Osservazione 211 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{\nu \in \mathbb{N} \setminus \{0\}\}$$

Definizione 6.6.1 (Funzione di densità)**.**

$$f_X(x) = \frac{1}{2^{\left(\frac{\nu}{2}\right)}\Gamma\left(\frac{\nu}{2}\right)} x^{\left(\frac{\nu}{2}-1\right)} e^{\left(-\frac{x}{2}\right)} \cdot \mathbb{1}_{R_X}(x) \tag{6.37}$$

con $x > 0$

Osservazione 212. Anche se ν può esser qualsiasi numero reale positivo, in pratica le applicazioni hanno tipicamente ν intero positivo.

Osservazione 213 (Forma della distribuzione). La vc Chi-quadrato è asimmetrica positiva e, al crescere di $\nu \to \infty$, tende ad assumere una forma sempre più vicina alla Normale. La forma della funzione di densità è monotona decrescente a zero se $\nu \leq 2$; se $\nu > 2$, presenta un picco intermedio in corrispondenza della moda (pari a $\nu - 2$). (figura 6.6)

Proposizione 6.6.1 (Momenti caratteristici)**.**

$$\mathbb{E}\left[X\right] = \nu \tag{6.38}$$
$$\text{Var}\left[X\right] = 2\nu \tag{6.39}$$
$$\text{Asym}\left(X\right) = \sqrt{\frac{8}{\nu}} \tag{6.40}$$
$$\text{Kurt}\left(X\right) = 3 + \frac{12}{\nu} \tag{6.41}$$

Proposizione 6.6.2. *Anche la distribuzione Chi-quadrato gode della proprietà riproduttiva:*

$$\sum_{i=1}^{n} \chi_{\nu_i}^2 \sim \chi_{\sum_i \nu_i}^2$$

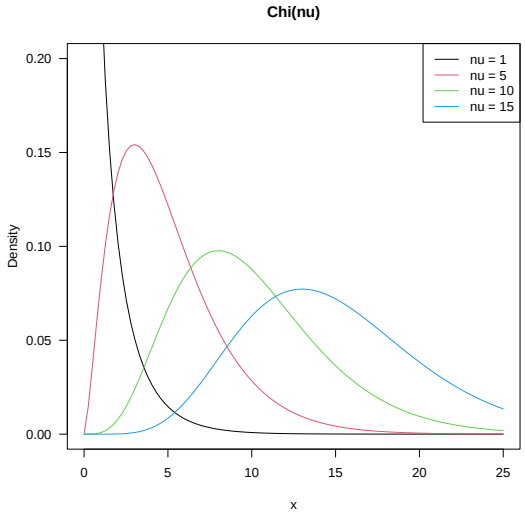


Figura 6.6: Distribuzione χ^2

6.7 Beta

Osservazione 214. Viene utilizzata quando si vogliono definire a priori i valori possibili delle probabilità di successo per variabili Bernoulliane.

Osservazione 215 (Supporto e spazio parametrico).

$$R_X = [0,1]$$
$$\boldsymbol{\Theta} = \{\alpha, \beta \in \mathbb{R} : \alpha, \beta > 0\}$$

Definizione 6.7.1 (Funzione di densità)**.** Una vc continua X si definisce Beta con due parametri $\alpha > 0, \beta > 0$, e la indichiamo con $X \sim \text{Beta}(\alpha, \beta)$ se la sua funzione di densità è:

$$f_X(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \cdot \mathbb{1}_{R_X}(x) \tag{6.42}$$

Definizione 6.7.2 (Funzione Beta)**.** Definita come

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \tag{6.43}$$

Presenta le seguenti proprietà

$$B(\alpha, \beta) = B(\beta, \alpha)$$
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

Osservazione 216. Una vc Beta è definita nell'intervallo $[0,1]$, ma effettuando la trasformazione $Y = X(b-a) + a$, la si può ricondurre all'intervallo $[a,b]$.

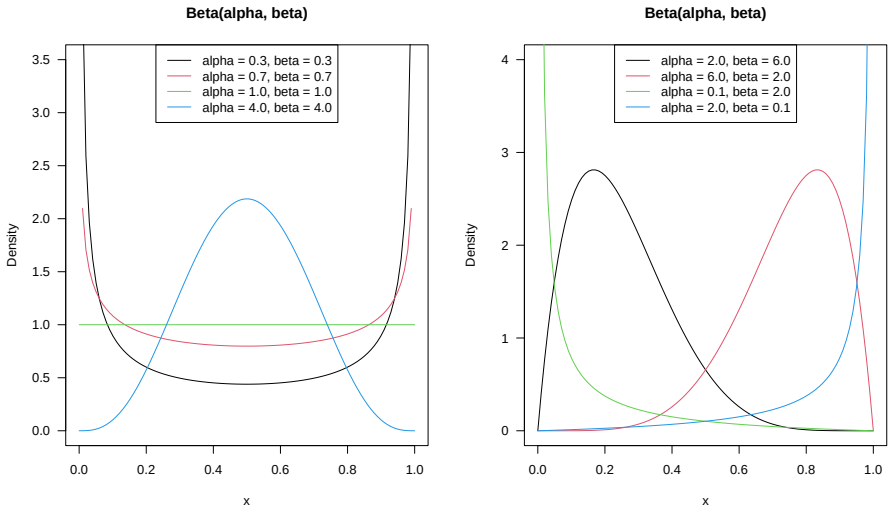


Figura 6.7: Distribuzione beta

Proposizione 6.7.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \tag{6.44}$$

$$\text{Var}[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{6.45}$$

Osservazione 217 (Forma della distribuzione). La forma (figura 6.7) dipende dai parametri α, β:

- se $\alpha = \beta$ la distribuzione è simmetrica rispetto al valore centrale $x = 1/2$; nel caso particolare $\alpha = \beta = 1$, la distribuzione coincide con l'uniforme: $\text{Beta}(1,1) \sim \text{Unif}(0,1)$;
- altrimenti il segno di $\beta - \alpha$ denota l'asimmetria (es se negativo, perché $\alpha > \beta$, la coda è a sinistra, se positivo la coda a destra); scambiando α con β si inverte l'asse di simmetria.

6.8 T di Student

Osservazione 218. Il suo uso è prettamente teorico, in quanto è la risultante di una trasformazione su due variabili, una normale e una chi quadrato.

Osservazione 219 (Supporto e spazio parametrico).

$$R_X = \mathbb{R}$$
$$\boldsymbol{\Theta} = \{g \in \mathbb{N} \setminus \{0\}\}$$

Definizione 6.8.1 (Distribuzione T)**.** Se $Z \sim \text{N}(0,1)$ ed C è una distribuzione indipendente tale che $C \sim \chi^2_g$ allora si definisce vc di Student la seguente X:

$$X = \frac{Z}{\sqrt{C/g}} \sim \text{T}(g) \tag{6.46}$$

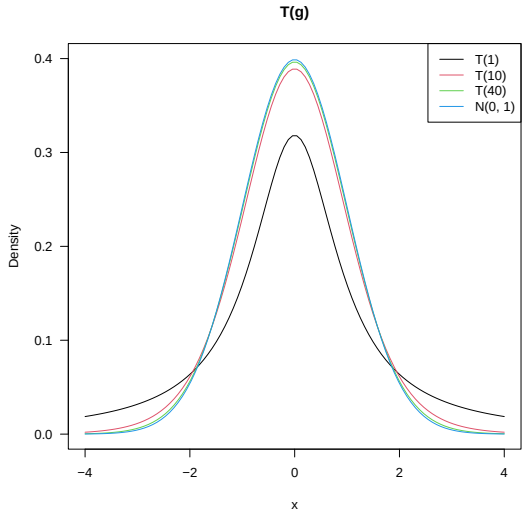


Figura 6.8: Distribuzione t

Definizione 6.8.2 (Funzione di densità)**.**

$$f_X(x) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)\sqrt{\pi g}}\left(1+\frac{x^2}{g}\right)^{-\frac{g+1}{2}} \cdot \mathbb{1}_{R_X}(x) \tag{6.47}$$

Proposizione 6.8.1 (Momenti caratteristici)**.**

$$\begin{aligned} \mathbb{E}[X] &= 0 \quad se\ g > 1 \\ \mathrm{Var}[X] &= \frac{g}{g-2} \quad seg > 2 \\ \mathrm{Kurt}(X) &= 3 + \frac{6}{g-4} \quad se\ g > 4 \end{aligned}$$

Osservazione 220 (Forma della distribuzione). Per $g \to \infty$ si nota la convergenza alla normale standardizzata. Verso $g = 30$, l'approssimazione è già buona; per g via via inferiore permane qualche differenza (code più alte rispetto alla normale, moda e media più basse). (figura 6.8)

6.9 F di Fisher

Osservazione 221. Il suo uso è prettamente teorico, in quanto è risultate di una trasformazione. È la distribuzione che deriva dal rapporto tra due vc Chi quadrato indipendenti tra loro e divise per i rispettivi gradi di libertà.

Osservazione 222. Se $X_1 \sim \chi^2_{g_1}$ e $X_2 \sim \chi^2_{g_2}$, allora

$$X = \frac{X_1/g_1}{X_2/g_2} \sim \mathrm{F}(g_1, g_2) \tag{6.48}$$

ovvero X si distribuisce come una F con g_1 e g_2 gradi di libertà.

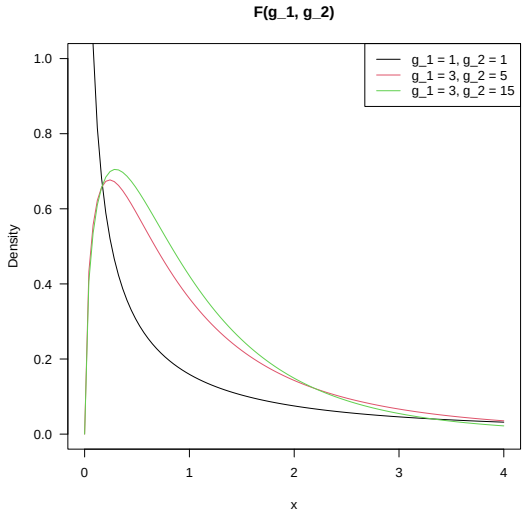


Figura 6.9: Distribuzione F

Osservazione 223 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{g_1, g_2 \in \mathbb{N} \setminus \{0\}\}$$

Definizione 6.9.1 (Funzione di densità)**.**

$$f_X(x) = \frac{\Gamma\left(\frac{g_1+g_2}{2}\right)}{\Gamma\left(\frac{g_1}{2}\right)\Gamma\left(\frac{g_2}{2}\right)} \cdot \left(\frac{g_1}{g_2}\right)^{\frac{g_1}{2}} \cdot \frac{x^{(g_1-2)/2}}{\left(1+\frac{g_1}{g_2}x\right)^{\frac{g_1+g_2}{2}}} \cdot \mathbb{1}_{R_X}(x) \tag{6.49}$$

Osservazione 224 (Funzione di ripartizione). Anche per la F non vi è una forma chiusa della ripartizione e ci si affida alle tavole.

Proposizione 6.9.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{g_2}{g_2-2} \quad se\ g_2 > 2$$
$$\mathrm{Var}[X] = \frac{2g_2^2(g_1+g_2-2)}{g_1(g_2-2)^2(g_2-4)} \quad se\ g_2 > 4$$

Osservazione 225 (Forma della distribuzione). Si nota che se $g_1 = g_2 = 1$ la funzione è monotona decrescente, se $g_1, g_2 \neq 1$ la funzione è asimmetrica positiva. (figura 6.9)
La distribuzione converge a quella di una normale solo se contemporaneamente $g_1 \to \infty$ e $g_2 \to \infty$.

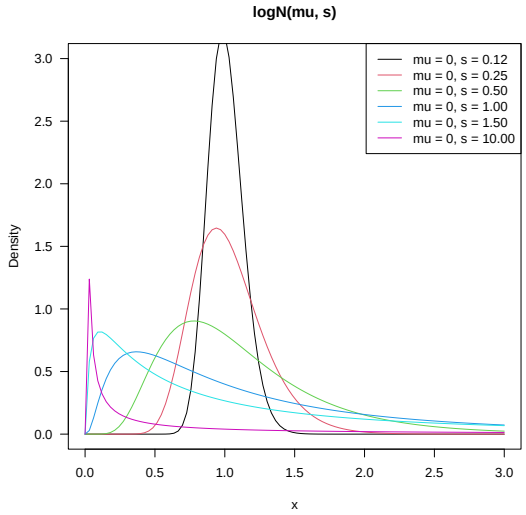


Figura 6.10: Distribuzione lognormale

6.10 Lognormale

Osservazione 226. Viene utilizzata quando la grandezza oggetto di studio è il risultato del prodotto di n fattori indipendenti

Osservazione 227 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \left\{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R} : \sigma^2 > 0\right\}$$

Definizione 6.10.1 (Funzione di densità)**.**

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2} \cdot \mathbb{1}_{R_X}(x) \tag{6.50}$$

Proposizione 6.10.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = e^{\mu + \frac{\sigma^2}{2}}$$
$$\mathrm{Var}[X] = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}$$

Osservazione 228. Si ha che se $X \sim \mathrm{LogN}(mu, \sigma)$ allora $\log X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, mentre se $Y \sim \mathrm{N}\left(\mu, \sigma^2\right)$, $e^Y \sim \mathrm{LogN}\left(\mu, \sigma^2\right)$

Osservazione 229 (Forma della distribuzione). Con μ fisso all'aumentare di σ l'asimmetria si incrementa (figura 6.10)

6.11 Weibull

Osservazione 230. Viene utilizzata per studiare l'affidabilità dei sistemi di produzione nei processi industriali, in particolare per valutare i tassi di rottura

Osservazione 231. La Weibull presenta la caratteristica di avere una funzione di rischio variabile in funzione di un ulteriore parametro a: se la vc $(X/b)^a \sim \text{Exp}(1)$, allora diremo che la vc continua X, definita sula semiretta positiva è una vc di Weibull con parametri $a > 0, b > 0$.

Osservazione 232 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{a, b \in \mathbb{R} : a, b > 0\}$$

Osservazione 233 (Forma della funzione). Il parametro a determina la forma (figura 6.11):

- se $a < 1$ il tasso di rottura è decrescente nel tempo, ci sono componenti difettose che si rompono subito e, una volta sostituito, il tasso diminuisce
- se $a = 1$ il tasso di rottura è costante nel tempo: le cause dei difetti sono casuali (e la distribuzione coincide con una esponenziale di parametro $1/b$, ossia $\text{Weibull}(1, b) \sim \text{Exp}\left(\frac{1}{b}\right)$)
- se $a > 1$ il tasso di rottura è crescente nel tempo, le cause della rottura dei componenti derivano dall'usura

Definizione 6.11.1 (Funzione di densità)**.**

$$f_X(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{\left[-\left(\frac{x}{b}\right)^a\right]} \cdot \mathbb{1}_{R_X}(x) \tag{6.51}$$

Proposizione 6.11.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{\Gamma\left(1+\frac{1}{b}\right)}{a^{1/b}}$$
$$\text{Var}[X] = \frac{\Gamma\left(1+\frac{2}{b}\right) - \Gamma^2\left(1+\frac{1}{b}\right)}{a^{2/b}}$$

6.12 Pareto

Osservazione 234. Viene utilizzata quando si studiano distribuzioni di variabili che hanno un minimo (ad esempio come, con x_m = reddito minimo)

Osservazione 235 (Supporto e spazio parametrico).

$$R_X = (x_m, +\infty)$$
$$\boldsymbol{\Theta} = \{x_m, k \in \mathbb{R} : x_m, k > 0\}$$

Definizione 6.12.1 (Funzione di densità)**.**

$$f_X(x) = k\frac{x_m^k}{x^{k+1}} \cdot \mathbb{1}_{R_X}(x) \tag{6.52}$$

Proposizione 6.12.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \frac{kx_m}{k-1} \qquad \textit{per } k > 1$$
$$\text{Var}[X] = \left(\frac{x_m}{k-1}\right)^2 \frac{k}{k-2} \qquad \textit{per } k > 2$$

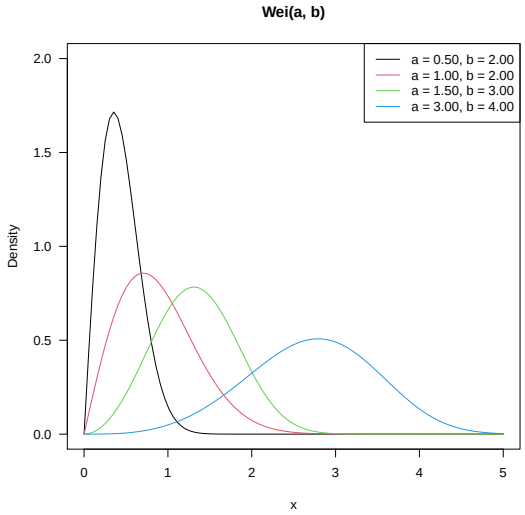


Figura 6.11: Distribuzione Weibull

Osservazione 236 (Forma della distribuzione). Al crescere di k la distribuzione è disuguale, ed è molto probabile trovare valori vicini al limite inferiore x_m, poco proabile trovare valori molto grandi.

```
## Error in loadNamespace(x):  non c'è alcun pacchetto chiamato 'VGAM'
## Error in (function (s, units = "user", cex = NULL, font = NULL,
vfont = NULL, :  plot.new has not been called yet
```

Capitolo 7

Random vectors

7.1 Intro

Definizione 7.1.1. A random vector is a function that maps $\Omega \to D \subset \mathbb{R}^n$.

Esempio 7.1.1. When we roll two dice $\Omega = \{\{1,1\}, \ldots, \{6,6\}\}$, the following are *bivariate* random vectors (or random vector with 2 dimensions):

- if X = outcome for the first die, Y = outcome of the second one, then (X, Y) is a bivariate rv;
- if X = sums of both outcomes, Y = absolute difference ofthe two outcomes, again (X, Y) is a bivariate rv;

Esempio 7.1.2. Regarding (X, Y) with X = sums of both outcomes, Y = absolute difference ofthe two outcomes we have that

- $\mathbb{P}(X = 5, Y = 3) = \mathbb{P}(\{X = 5\} \cap \{Y = 3\}) = \mathbb{P}((4,1),(1,4)) = \frac{1}{18}$
-

7.1.1 Relationship between rvs

7.1.1.1 Covariance

Definizione 7.1.2 (Covariance)**.** If we have two random variables X, Y and

$$\mathbb{E}[|X|] \leq +\infty,\ \mathbb{E}[|Y|] \leq +\infty,\ \mathbb{E}[|XY|] \leq +\infty$$

we can define the covariance as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - E(Y))] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \tag{7.1}$$

Proposizione 7.1.1. *Assuming the covariance exists, some remarks regarding it*

1. *if X is indipendent Y,* $\mathrm{Cov}(X,Y) = 0$. *The converse is false.*

2. *the generalized version of 4.56 for the variance of the sum of random variables involves their covariance, that is*

$$\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \operatorname{Var}[X_i] + \sum_{i \neq j} a_i a_j \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}[X_i] + 2\sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j) \tag{7.2}$$

where (1) *because* $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$

Dimostrazione. To prove the first one, if $X \perp\!\!\!\perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ so the covariance is 0.
A counterexample of null covariance but not independent random variable follows. □

Esempio 7.1.3 ($\operatorname{Cov}(X,Y) = 0$ but X, Y are not independent)**.** Let $X \sim \mathrm{N}(0,1)$ and $Y = X^2$. Let's prove:

- $\operatorname{Cov}(X,Y) = 0$. We have that

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \underbrace{\mathbb{E}[X]}_{=0}\mathbb{E}[Y] = \mathbb{E}[XY] = \mathbb{E}\left[X^3\right]$$

  Since $X$is absolutely continuous (normal) the expectation of X to the power 3 can be wrritten as

$$\mathbb{E}\left[X^3\right] = \int_{-\infty}^{+\infty} x^3 \cdot \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

  and since the integrand it's an odd function evaluated on a simmetric interval, the integral is 0

- $X \not\perp\!\!\!\perp Y$. It's intuitive these are not independent, however let's prove it formally. To prove that we consider this probability

$$\mathbb{P}(|X| \leq 1, Y > 1) \overset{(1)}{=} \mathbb{P}(|X| \leq 1, |X| > 1) = \mathbb{P}(\emptyset) = 0$$

  where in (1) since $Y = X^2$.
  If X and Y *were* independent we would have that this result is equal to the product

$$\mathbb{P}(|X| \leq 1, Y > 1) = \underbrace{\mathbb{P}(|X| \leq 1)}_{>0} \cdot \underbrace{\mathbb{P}(|X| > 1)}_{>0} > 0$$

  But since that probability is 0, we conclude they are not independent.

Esempio 7.1.4. An example of 7.5 is $\operatorname{Var}[X - Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] - 2\operatorname{Cov}(X,Y)$

Osservazione 237. In the lucky case these are indipendent covariance is null and variance of sum is sum of variance.

7.1.1.2 Correlation coefficient

Definizione 7.1.3 (Correlation coefficient)**.** if $\mathbb{E}\left[X^2\right] < +\infty$, $\mathbb{E}\left[Y^2\right] < +\infty$, $\operatorname{Var}[X] > 0$, $\operatorname{Var}[Y] > 0$, we can define the correlation coefficent as

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}[X]}\sqrt{\operatorname{Var}[Y]}} \tag{7.3}$$

Proposizione 7.1.2. *Some properties:*

- *It can be written as the covariance between the two standardized variables*

$$\operatorname{Corr}(X,Y) = \operatorname{Cov}\left(\frac{X-\mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}}, \frac{Y-\mathbb{E}[Y]}{\sqrt{\operatorname{Var}[Y]}},\right)$$

  in essence correlation is nothing other than a covariance on standardized variables.

- *it ranges in* $-1 \leq \operatorname{Corr}(X,Y) \leq 1$ *with the following limit cases:*

$$\operatorname{Corr}(X,Y) = 1 \iff Y = a + bX, b > 0$$
$$\operatorname{Corr}(X,Y) = -1 \iff Y = a + bX, b < 0$$

7.1.2 Relationship between rvs

7.1.2.1 Covariance

Definizione 7.1.4 (Covariance)**.** If we have two random variables X, Y and

$$\mathbb{E}[|X|] \leq +\infty,\ \mathbb{E}[|Y|] \leq +\infty,\ \mathbb{E}[|XY|] \leq +\infty$$

we can define the covariance as

$$\operatorname{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-E(Y))] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \tag{7.4}$$

Proposizione 7.1.3. *Assuming the covariance exists, some remarks regarding it*

1. *if* X *is indipendent* Y, $\operatorname{Cov}(X,Y) = 0$. *The converse is false.*
2. *the generalized version of 4.56 for the variance of the sum of random variables involves their covariance, that is*

$$\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 \operatorname{Var}[X_i] + \sum_{i \neq j} a_i a_j \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \operatorname{Var}[X_i] + 2\sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j) \tag{7.5}$$

   where (1) *because* $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$

Dimostrazione. To prove the first one, if $X \perp\!\!\!\perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ so the covariance is 0.
A counterexample of null covariance but not independent random variable follows. □

Esempio 7.1.5 ($\operatorname{Cov}(X,Y) = 0$ but X, Y are not independent)**.** Let $X \sim \mathrm{N}(0,1)$ and $Y = X^2$. Let's prove:

- $\mathrm{Cov}(X,Y)=0$. We have that

$$\mathrm{Cov}(X,Y)=\mathbb{E}[XY]-\underbrace{\mathbb{E}[X]}_{=0}\mathbb{E}[Y]=\mathbb{E}[XY]=\mathbb{E}\left[X^3\right]$$

Since $X$is absolutely continuous (normal) the expectation of X to the power 3 can be wrritten as

$$\mathbb{E}\left[X^3\right]=\int_{-\infty}^{+\infty}x^3\cdot\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

and since the integrand it's an odd function evaluated on a simmetric interval, the integral is 0

- $X \not\perp\!\!\!\perp Y$. It's intuitive these are not independent, however let's prove it formally. To prove that we consider this probability

$$\mathbb{P}(|X|\leq 1, Y>1)\overset{(1)}{=}\mathbb{P}(|X|\leq 1,|X|>1)=\mathbb{P}(\emptyset)=0$$

where in (1) since $Y=X^2$.
If X and Y *were* independent we would have that this result is equal to the product

$$\mathbb{P}(|X|\leq 1, Y>1)=\underbrace{\mathbb{P}(|X|\leq 1)}_{>0}\cdot\underbrace{\mathbb{P}(|X|>1)}_{>0}>0$$

But since that probability is 0, we conclude they are not independent.

Esempio 7.1.6. An example of 7.5 is $\mathrm{Var}[X-Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]-2\,\mathrm{Cov}(X,Y)$

Osservazione 238. In the lucky case these are indipendent covariance is null and variance of sum is sum of variance.

7.1.2.2 Correlation coefficient

Definizione 7.1.5 (Correlation coefficient)**.** if $\mathbb{E}\left[X^2\right]<+\infty$, $\mathbb{E}\left[Y^2\right]<+\infty$, $\mathrm{Var}[X]>0$, $\mathrm{Var}[Y]>0$, we can define the correlation coefficent as

$$\mathrm{Corr}(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}} \tag{7.6}$$

Proposizione 7.1.4. *Some properties:*

- *It can be written as the covariance between the two standardized variables*

$$\mathrm{Corr}(X,Y)=\mathrm{Cov}\left(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}},\frac{Y-\mathbb{E}[Y]}{\sqrt{\mathrm{Var}[Y]}},\right)$$

in essence correlation is nothing other than a covariance on standardized variables.

- *it ranges in* $-1\leq\mathrm{Corr}(X,Y)\leq 1$ *with the following limit cases:*

$$\mathrm{Corr}(X,Y)=1\iff Y=a+bX, b>0$$
$$\mathrm{Corr}(X,Y)=-1\iff Y=a+bX, b<0$$

Capitolo 8

Misc topics

8.1 Characteristic function

8.1.1 Characteristic function

Definizione 8.1.1 (Characteristic function)**.** Let X be a random variable, the characteristic function $\phi_X(t) : \mathbb{R} \to \mathbb{C}$, existing $\forall t \in \mathbb{R}$ is defined as

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itx} f(x)\, \mathrm{d}x = \int_{-\infty}^{+\infty} \cos tx f(x)\, \mathrm{d}x + i \int_{-\infty}^{+\infty} \sin(tx) f(x)\, \mathrm{d}x$$

with $i^2 = -1$

Osservazione importante 38 (characteristic function for n-variante random). If $\mathbf{X} = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix}$ is a n-variate random vector, the characteristic function of $\mathbf{X}$ is

$$\phi_{\mathbf{X}}(t) = \mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{X}}\right] = \mathbb{E}\left[e^{i\sum_i t_i X_i}\right] = \mathbb{E}\left[\cos \mathbf{t}^T\mathbf{X} + i \sin \mathbf{t}^T\mathbf{X}\right], \quad \forall t \in R^n$$

where $\mathbf{t} = \begin{bmatrix} t_1 \\ \dots \\ t_n \end{bmatrix}$ so $\mathbf{t}^T$ is a column vector so $\mathbf{t}^T\mathbf{X} = \sum_{i=1}^n t_i X_i$.
However, from now on we assume single variable (because it's more convenient) not n-variate random vector.

Esempio 8.1.1 (Characteristic function of a binomial)**.** Let $X \sim \mathrm{Bin}(n, p)$, $D_x = \{0, 1, \dots, n\}$, the characteristic function is

$$\begin{aligned} \phi_X(t) = \mathbb{E}\left[e^{itX}\right] &= \sum_{x=0}^{n} e^{itx} \cdot \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} (pe^{it})^x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\ &\overset{(1)}{=} (1 - p + pe^{it})^n \end{aligned}$$

where in (1) we applied binomial formula $(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$

Osservazione importante 39 (Usefulness). Despite being complicated, they are useful for several reasons (both theoric and practical:

1. they determine the distribution of the random variable: this is the reason this stuff is so important to statistic (important for Rigo);
2. they provide a *link with the moment* of order k of the variable via *differentiation* (with respect to t evaluated at $t = 0$);
3. they provide a *link with the density function* via the *inversion formula.*

Teorema 8.1.1 (Link with distribution)**.** *Supposing we have two random vector X, Y, then*

$$X \sim Y \iff X \text{ and } Y \text{ have the same characteristic function} \tag{8.1}$$

Dimostrazione. Rigo non l'ha fatta. □

Proposizione 8.1.2 (Link with the moments)**.**

$$\left[\frac{\partial^k}{\partial t^k}\phi_X(t)\right]_{t=0} = i^k \, \mathbb{E}\left[X^k\right]$$

Dimostrazione. we have

$$\frac{\partial^k}{\partial t^k}\phi_X(t) = \frac{\partial^k}{\partial t^k}\,\mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\frac{\partial^k}{\partial t^k}e^{itX}\right] = \mathbb{E}\left[i^k X^k e^{itx}\right]$$

and evaluated in $t = 0$

$$\mathbb{E}\left[i^k X^k e^{itx}\right] = i^k \, \mathbb{E}\left[X^k\right]$$

□

Proposizione 8.1.3 (Link with density (inversion formula))**.**

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\phi_X(t)\,\mathrm{d}t$$

Dimostrazione. Virols skips it. □

Proposizione 8.1.4 (Important properties (Rigo))**.** *We have the following:*

1. *if $X \perp\!\!\!\perp Y$, the characteristic function of the sum is equal to the product of the single characteristic functions*

   $$\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$$

   This because

   $$\begin{aligned}\phi_{X+Y}(t) = \mathbb{E}\left[e^{it(X+Y)}\right] = \mathbb{E}\left[e^{itX}e^{itY}\right] &\overset{((1))}{=} \mathbb{E}\left[e^{itX}\right]\mathbb{E}\left[e^{itY}\right] \\ &= \phi_X(t)\cdot\phi_Y(t), \quad \forall t \in \mathbb{R}\end{aligned}$$

   where in (1)*, since $X \perp\!\!\!\perp Y$, any combination is independent as well, and so we apply the expected value property of product of independent variables. Because of this property characteristic function becomes* very handy *when woriking with sums of independent rvs.*

2. *connection between characteristic function and moments: if the random variable has the moment of order j, then the characteristic function is $\in C^j$ (that is has derivatives of order up to j which are continuous) and the derivative of order r (for $r = 1, \dots, j$), is known:*

$$\mathbb{E}\left[|X|^j\right] < +\infty \implies \begin{cases} \phi_X(t) \in C^j \\ \phi_X(t)^{(r)} = \mathbb{E}\left[(iX)^r e^{itX}\right] \end{cases}$$

the latter means that in each derivative up to order j we can interchange the operator of derivative and the operator of expectation. The derivative of characteristic function is a derivative of expectation; in order to make the derivative one can change the operator of derivative with the operator of differentiation. For instance for $r = 1$ (suppose we want to calculate the first derivative)

$$\phi_X(t)' = \frac{\partial}{\partial t}\mathbb{E}\left[e^{itX}\right] \overset{(1)}{=} \mathbb{E}\left[\frac{\partial}{\partial t}e^{itX}\right] = \mathbb{E}\left[iXe^{itX}\right]$$

where in (1) the swap occurs.

The converse implication holds not always: if the characteristic function has derivative j in zero and j is even, then we can conclude that the random variable has moments of order j:

$$\begin{cases} \exists \phi_X(0)^{(j)} \\ j \text{ is even} \end{cases} \implies \mathbb{E}\left[|X|^j\right] < +\infty$$

Note that, since $j = 1$is odd, it may be that $\exists \phi_X(0)'$ but $\mathbb{E}[|X|] = +\infty$.

3. ***inversion theorem*** *gives a closed formula for determining the distribution function given characteristic function. The important fact to recall for the exam is that characteristic function can be inverted: if you know the characteristic function, there exists a formula that allows to write down the distribution function (no need to memorize it for the exam).*

If $a < b$ and $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$ then:

$$F(b) - F(a) = \mathbb{P}(a < X \leq b) = \frac{1}{2\pi i}\lim_{c \to +\infty}\int_{-c}^{c} \frac{e^{-itb} - e^{-ita}}{t}\phi_X(t)\,\mathrm{d}t$$

4. ***continuity theorem****: we have the equivalence*

$$X_n \xrightarrow{d} X \iff \lim_{n \to +\infty}\phi_{X_n}(t) = \phi_X(t), \quad \forall t \in \mathbb{R}$$

this theorem is important because any time we want to prove convergence in distribution (an important type of convergence) we can (if convenient) prove the limit of characteristic function.

Esempio 8.1.2. In this example we show that if X_n is iid and the characteristic function has the first derivative at 0, $\exists \phi_X(0)'$, then the sample mean converges (in distribution and probability) to a constant/degenerate rv.
Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of iid rvs; we define the sample

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

The characteristic function of the sample mean is

$$\phi_{\bar{X}_n}(t) = \mathbb{E}\left[e^{it\sum_i \frac{X_i}{n}}\right] = \phi_{\sum_i X_i}\left(\frac{t}{n}\right) \overset{(\perp\!\!\!\perp)}{=} \prod_{i=1}^{n} \phi_{X_i}\left(\frac{t}{n}\right) \overset{(id)}{=} \left[\phi_{X_i}\left(\frac{t}{n}\right)\right]^n$$

Suppose now that the first derivative of the characteristic function of X_i exists in 0, that is $\exists \phi_{X_i}(0)'$; then by Taylor expansion formula

$$\phi_{\bar{X}_n}(t) = \left[\phi_{X_i}\left(\frac{t}{n}\right)\right]^n = \left[\phi_{X_i}(0) + \frac{t}{n}\phi_{X_i}(0)' + o\left(\frac{t}{n}\right)\right]^n = \left[1 + \frac{t\phi_{X_i}(0)' + no\left(\frac{t}{n}\right)}{n}\right]^n$$

where $o\left(\frac{t}{n}\right)$ is the Peano rest. In general $g = o(f)$ if $\lim_{x\to x_0} \frac{g(x)}{f(x)} = 0$.
Now, what is the limit of the formula above for $n \to +\infty$? Using the fact that

$$\text{if } a_n \to a \implies \left(1 + \frac{a_n}{n}\right)^n \to e^a$$

we have (with $a_n = t\phi_{X_i}(0)' + no\left(\frac{t}{n}\right)$ and noted that $a_n \to t\phi_{X_i}(0)' + 0$)

$$\phi_{\bar{X}_n}(t) \to e^{t\phi_{X_i}(0)'}$$

Now it can be shown (we won't) that the first derivative in 0 is

$$\phi_{X_i}(0)' = i\alpha, \quad \alpha \in \mathbb{R}$$

and thus we our characteristic function converges to

$$\phi_{\bar{X}_n}(t) \to e^{it\alpha}, \forall t \in \mathbb{R}$$

Is $e^{it\alpha}$ a characteristic function? Yes the δ_α has this characteristic function since if $X \sim \delta_\alpha$

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[e^{it\alpha}\right] = e^{it\alpha}$$

Hence $\bar{X}_n \xrightarrow{d} \alpha$, by continuity theorem, and since the limit is a degenerate rv, we have not only convergence in distribution byt also convergence in probability $\bar{X}_n \xrightarrow{p} \alpha$.

Osservazione importante 40. The above should be *weak low* of large number (convergence not a.s. but only in probability, check with Virols).
Furthermore, if the sequence is not only iid, but also the mean exists, $\mathbb{E}[|X_i|] < +\infty$, then $\bar{X}_n \xrightarrow{a.s.} \mathbb{E}[X_i]$ then the sample mean converges almost surely to the mean (this is the *strong law of large number*).
But as noted above, it may be that $\exists \phi_{X_i}(0)'$ even if $\mathbb{E}[|X_i|] = +\infty$.

8.1.2 Moment generating function

Definizione 8.1.2 (Moment generating function (mgf))**.** It's obtained from the characteristic function by evaluating it at $-it$, $\phi_X(-it)$, so that there are no complex number:

$$\phi_X(-it) = \mathbb{E}\left[e^{-iitX}\right] = \mathbb{E}\left[e^{tX}\right] = M_X(t), \quad \forall t \in \mathbb{R} \tag{8.2}$$

Osservazione importante 41. It's simpler than characteristic function (no i here) but has its drawbacks:

- we don't have an inversion theorem, so it's useful only for the moments
- it always exists for $t = 0$ but it may fail to exist for $t \neq 0$ (eg it could be $M_X(t) = +\infty$, while characteristic function always exist).
  If for some reason we know that the moment generating function is finite in a neighborhood of zero (not true/necessaire in general), it's convenient to use it instead of the characteristic function. In fact, in this lucky case, that is where it's finite in a neighborhood of 0:

  $$M_X(t) < +\infty, \forall t \in (-\varepsilon, \varepsilon)$$

  the following hold:

  - the random variable has moments of every order: $\mathbb{E}\left[|X|^n\right] < +\infty, \forall n$
  - the sequence of moments $\mathbb{E}\left[X^n\right]$, with $n = 1, 2, \ldots$, determines the distribution, in the sense that if X and Y does not have the same distribution then *either* one of them have some moments not finite or moments both are finite but different for some n:

    $$X \nsim Y \implies \left(\mathbb{E}\left[|X|^n\right] = +\infty, \text{for some n}\right) \vee \left(\mathbb{E}\left[X^n\right] \neq \mathbb{E}\left[Y^n\right], \text{for some n}\right)$$

Osservazione importante 42. If we have two random variables X, Y, and we know that have both the moments of every order and the same order (mean, variance, third moment etc).

$$\mathbb{E}\left[|X|^n\right] < +\infty,\ \mathbb{E}\left[|Y|^n\right] < +\infty,\ \mathbb{E}\left[|X|^n\right] = \mathbb{E}\left[|Y|^n\right], \quad \forall n$$

Can we conclude that the two random variables has the same distribution? No we cannot conclude that.
This is contrary to intuition; however this annoying fact doesn't occur if one between X and Y has finite moment generating function. In that case we can say they have the same distribution.

Proposizione 8.1.5 (Properties)**.**

$$\left[\frac{\partial^k}{\partial t^k} M_X(t)\right]_{t=0} = \mathbb{E}\left[X^k\right] \tag{8.3}$$

$$M_X(0) = \mathbb{E}\left[e^{0X}\right] = \mathbb{E}\left[1\right] = 1 \tag{8.4}$$

$$M_X(t) = M_Y(t), \forall t \iff F_X(x) = F_Y(y) \qquad \textit{(uniqueness)} \tag{8.5}$$

$$M_{aX+b}(t) = e^{tb} M_X(at), \quad a, b \in \mathbb{R} \tag{8.6}$$

$$X \perp\!\!\!\perp Y \implies M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \tag{8.7}$$

TODO: l'implicazione per l'indipendenza è anche coimplicazione?

Dimostrazione. For 8.6

$$M_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = \mathbb{E}\left[e^{taX} \cdot \underbrace{e^{tb}}_{\text{constant}}\right] = e^{tb} \cdot \mathbb{E}\left[e^{taX}\right] = e^{tb} M_X(at)$$

For 8.7

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right]$$

Now note that:

TODO: questo andrebbe portato più in vista nella la sezione indipendenza o prop v. atteso

- first

$$\begin{aligned}\mathbb{E}\left[g(X)h(Y)\right] &= \int_{D_x}\int_{D_y} g(x)h(y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \overset{(1)}{=} \int_{D_x}\int_{D_y} g(x)h(y)f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{D_x} g(x)f_X(x)\,\mathrm{d}x \int_{D_y} h(y)f_Y(y)\,\mathrm{d}y = \mathbb{E}\left[g(X)\right]\mathbb{E}\left[h(Y)\right]\end{aligned}$$

  where (1) due to be $X \perp\!\!\!\perp Y$.

- furthermore

$$\begin{aligned}\mathbb{E}\left[g(X)+h(Y)\right] &= \int_{D_x}\int_{D_y} (g(x)+h(y))f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{D_x}\int_{D_y} g(x)f(x,y)\,\mathrm{d}x\,\mathrm{d}y + \int_{D_x}\int_{D_y} h(y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{D_x} g(x)\underbrace{\int_{D_y} f(x,y)\,\mathrm{d}y}_{f(x)}\,\mathrm{d}x + \int_{D_x}\int_{D_y} h(y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{D_x} g(x)f(x)\,\mathrm{d}x + \int_{D_y} h(y)f(y)\,\mathrm{d}y = \mathbb{E}\left[g(X)\right] + \mathbb{E}\left[h(Y)\right]\end{aligned}$$

Therefore coming back to our focus, under independence and using the first one

$$M_{X+Y}(=)\,\mathbb{E}\left[e^{tX}e^{tY}\right] \overset{(1)}{=} \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_y(t)$$

in (1) because of $\perp\!\!\!\perp$ □

Osservazione 239. The following is a result that become useful sometimes (eg clt)

Proposizione 8.1.6 (Mc Laurin expansion of mgf)**.**

$$M_X(t) = 1 + t\,\mathbb{E}\left[X\right] + \frac{t^2}{2!}\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}\left[X^3\right] + \dots \tag{8.8}$$

Dimostrazione. In general decomposition of $M_X(t)$ is like the following. Considered that mclaurin expansion of e^{tx}

$$e^{tx} = 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \dots$$

then

$$\begin{aligned}M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_{D_X} e^{tx}f(x)\,\mathrm{d}x = \underbrace{\int_{D_X} 1f(x)\,\mathrm{d}x}_{=1} + \int_{D_X} tx \\ &= 1 + t\int_{D_X} xf(x)\,\mathrm{d}x + \frac{t^2}{2!}\int_{D_X} x^2f(x)\,\mathrm{d}x + \frac{t^2}{3!}\int_{D_X} x^3f(x)\,\mathrm{d}x + \dots \\ &= 1 + t\,\mathbb{E}\left[X\right] + \frac{t^2}{2!}\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}\left[X^3\right] + \dots\end{aligned}$$

□

Osservazione 240. Now we see an example where mgf does not always exists

Esempio 8.1.3 (Mgf of Gamma)**.** Let $X \sim \text{Gamma}(\alpha, \beta)$, $\alpha, \beta > 0$

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

with $D_x = [0, +\infty)$ and

$$\Gamma(x) = \int_0^{+\infty} x^{\alpha-1} e^{-x} \, \mathrm{d}x, \quad \forall \alpha > 0$$
$$\alpha \in \mathbb{N} \implies \Gamma(\alpha) = (\alpha - 1)!$$

Let's evaluate $M_X(t)$

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_0^{+\infty} e^{tx} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} e^{-(\beta-t)x} \cdot x^{\alpha-1} \, \mathrm{d}x \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} e^{-(\beta-t)x} \cdot x^{\alpha-1} \frac{(\beta-t)^\alpha}{(\beta-t)^\alpha} \, \mathrm{d}x \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha} \underbrace{\int_0^{+\infty} \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} \cdot e^{-(\beta-t)x} x^{\alpha-1} \, \mathrm{d}x}_{=1, \text{ since } f(x) \text{ of a Gamma}(\alpha, \beta-t)} \end{aligned}$$

Therefore

$$M_X(t) = \frac{\beta^\alpha}{(\beta-t)^\alpha} = \left(\frac{\beta}{(\beta-t)}\right)^\alpha = \left(\frac{\beta-t}{(\beta)}\right)^{-\alpha} = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$$

where, since $\alpha > 0$ (and it's an exponent), $M_X(t)$ is well defined only if the base is positive

$$1 - \frac{t}{\beta} > 0 \iff t < \beta$$

Esempio 8.1.4. Exercise:

1. compute the second moment of the binomial distribution using $\mathbb{E}\left[X^2\right]$= second derivative of mgf evaluated in 0
2. verify property 2 of mgf, mgf(0)=1
3. eval $\mathbb{E}\left[(\right] X)$ where X is Gamma by using first derivative of mgf

Esempio 8.1.5 (Normal distributions)**.** Let $X \sim \mathrm{N}(0, 1)$, then lets derive the mgf

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_{-\infty}^{+\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \, \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{tx - \frac{1}{2}x^2} \, \mathrm{d}x \end{aligned}$$

Now we apply this substitution trick

$$tx - \frac{1}{2}x^2 = \frac{t^2 - (x-t)^2}{2}$$

because of the expansion

$$\frac{t^2 - (x-t)^2}{2} = \frac{t^2 - x^2 - t^2 + 2xt}{2} = tx - \frac{x^2}{2}$$

So

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2-(x-t)^2}{2}} \, \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} e^{\frac{-(x-t)^2}{2}} \, \mathrm{d}x \\ &= e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-t)^2}{2}} \, \mathrm{d}x}_{=1 \text{ since integral of } \mathrm{N}(t,1)} \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

Therefore

$$X \sim \mathrm{N}(0,1) \iff M_X(t) = e^{t^2/2}$$

while applying properties of mgf it turns out that, if $X \sim \mathrm{N}(0,1)$

$$\sigma X + \mu \sim \mathrm{N}\left(\mu, \sigma^2\right) \iff M_{\sigma X+\mu}(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$$

Esempio 8.1.6. Regarding the normal prove:

- $\frac{\partial M_{\sigma X+\mu}(t)}{=} \mu$
- derive $\mathbb{E}\left[X^2\right]$ by mgf
- check that $\mathrm{Var}\left[\sigma X + \mu\right] = \sigma^2$ (applying $\mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$

Osservazione 241. In the following we use the property of mgf for sum of random variables

Esempio 8.1.7 (Mgf of bernoulli and binomial)**.** if $X \sim \mathrm{Bern}(p)$, $p(x) = p^x(1-p)^{1-x}$, $D_x = \{0,1\}$, while the mgf is

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = e^{t\cdot 0} \cdot (1-p)p^0 0 + e^{t\cdot 1} p^1 (1-p)^0 = 1 - p + e^t p$$

Being the binomial $Y = X_1 + \ldots + X_n$ with X_i, by properties of mgf

$$M_Y(t) = \prod_{i=1}^{n} (pe^t + 1 - p) = (pe^t + 1 - p)^n$$

Esempio 8.1.8. Homework: by using 8.7 find $M_Y(t)$, with $Y = \sum_{i=1}^n X_i$, $X_i \sim \mathrm{Pois}(\lambda_i)$, and $X_i \perp\!\!\!\perp X_j$.
Hint:

$$\sum_{x=0}^{+\infty} \frac{c^x}{x!} = e^c$$

8.2 Order statistics

Osservazione 242. The kth order statistic of statistical sample is equal to its kth-smallest value. Together with rank statistics, order statistics are fundamental tools in non-parametric statistics and inference.

Osservazione 243. Important special cases of the order statistics are the minimum $X_{(1)}$ maximum $X_{(n)}$ sample median and other sample quantiles.

Osservazione importante 43 (Setup). In this setion we consider a sequence of n iid rvs $X_1, \dots, X_n$ and define the following random variable

$$\begin{aligned} X_{(1)} &= \min\{X_1, \dots, X_n\} \\ X_{(2)} &= \min\left\{\{X_1, \dots, X_n\} \setminus \{X_{(1)}\}\right\} \\ \dots X_{(n)} &= \max\{X_1, \dots, X_n\} \end{aligned}$$

we have that $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. We are interested in studying properties of these newly defined rvs.

Esempio 8.2.1. Throwing a dice 6 times, having the sequence $X_1, \dots, X_6$. To study the distribution of the minimum $X_{(1)}$ we can say that

$$\begin{aligned} \mathbb{P}\left(X_{(1)} = 6\right) &= \frac{1}{6} \cdot \dots \cdot \frac{1}{6} = \left(\frac{1}{6}\right)^6 \\ \mathbb{P}\left(X_{(1)} = 6\right) &= 1 - \left(\frac{5}{6}\right)^6 \end{aligned}$$

8.2.1 Minimum

Proposizione 8.2.1 (Distribution function)**.** *We have that*

$$F_{(1)}(x) = 1 - [1 - F_X(x)]^n \tag{8.9}$$

Dimostrazione.

$$\begin{aligned} F_{(1)}(x) &= \mathbb{P}\left(X_{(1)} \leq x\right) = 1 - \mathbb{P}\left(X_{(1)} > x\right) \\ &= 1 - \mathbb{P}\left(X_1 > x, X_2 > x, \dots, X_n > x\right) \overset{(1)}{=} 1 - \prod_{i=1} \mathbb{P}\left(X_i > x\right) \\ &\overset{(2)}{=} 1 - \prod_{i=1} \mathbb{P}\left(X > x\right) = 1 - \left[\mathbb{P}\left(X > x\right)\right]^n = 1 - \left[1 - \mathbb{P}\left(X \leq x\right)\right]^n \\ &= 1 - [1 - F_X(x)]^n \end{aligned}$$

with (1) we considered independent rvs and (2) identically distributed. □

Osservazione 244. Interpretazione affinché il minimo sia al più x si fa il complemento in cui si guarda la probabilità che siano tutte contemporaneamente $> x$

Proposizione 8.2.2 (Density function)**.**

$$f_{(1)}(x) = n f_X(x) \cdot [1 - F_X(x)]^{n-1}$$

Dimostrazione.

$$f_1(x) = \frac{\partial F_{(1)}(x)}{\partial x} = -n\left[1 - F_X(x)\right]^{n-1}(-1)f_X(x) = nf_X(x)\cdot\left[1 - F_X(x)\right]^{n-1}$$

□

Esempio 8.2.2. A room is lit by 5 light bulbs, each bulb lifetime has ha distribution $X \sim \text{Exp}\left(\lambda = \frac{1}{100}\right)$. What is the probability that after 200 days all the bulbs are still working?
We can setup this as $\mathbb{P}\left(X_{(1)} > 200\right)$, therefore:

$$\mathbb{P}\left(X_{(1)} > 200\right) = 1 - \mathbb{P}\left(X_{(1)} \leq 200\right) = 1 - F_{(1)}(200)$$

we have that, being X an exponential

$$F_{(1)}(200) = 1 - (1 - F_X(200))^5 = 1 - \left(1 - 1 + e^{-200/100}\right)^5 = 1 - \frac{1}{e^{10}}$$

Therefore

$$\mathbb{P}\left(X_{(1)} > 200\right) = 1 - 1 + \frac{1}{e^{10}} = \frac{1}{e^{10}}$$

8.2.2 Maximum

Proposizione 8.2.3 (Distribution function)**.**

$$F_{(n)}(x) = \left[F_X(x)\right]^n \tag{8.10}$$

Dimostrazione.

$$\begin{aligned} F_{(n)}(x) &= \mathbb{P}\left(X_{(n)} \leq x\right) = \mathbb{P}\left(X_1 \leq x, \ldots, X_N \leq x\right) \\ &\overset{(iid)}{=} \left[\mathbb{P}\left(X \leq x\right)\right]^n = \left[F_X(x)\right]^n \end{aligned}$$

□

Osservazione 245. Il massimo sia $\leq x$ se tutte le vc sono $\leq x$

Proposizione 8.2.4 (Density function)**.**

$$f_{(n)}(x) = n\left[F_X(x)\right]^{n-1} f_X(x) \tag{8.11}$$

Dimostrazione.

$$f_{(n)}(x) = \frac{\partial}{\partial x}F_{(n)}(x) = n\left[F_X(x)\right]^{n-1} f_X(x)$$

□

Esempio 8.2.3. Considering again a room lit by 5 light bulbs, each bulb lifetime has ha distribution $X \sim \text{Exp}\left(\lambda = \frac{1}{100}\right)$. What is the probability that after 200 days at least a bulb will be working?
This can be setup with

$$\begin{aligned} \mathbb{P}\left(X_{(n)} > 200\right) &= 1 - \mathbb{P}\left(X_{(n)} \leq 200\right) = 1 - F_{(n)}(200) \\ &= 1 - \left[F_X(200)\right]^5 = 1 - \left(1 - e^{-2}\right)^5 \simeq 0.52 \end{aligned}$$

Esempio 8.2.4. Draw randomly 12 numbers between from $X \sim \text{Unif}(0,1)$. What is the probability that at least a number > 0.9?
If $X \sim \text{Unif}(0,1)$, $F_X(x) = x$. We have

$$\mathbb{P}\left(X_{(n)} > 0.9\right) = 1 - \mathbb{P}\left(X_{(n)} \leq 0.9\right) = 1 - [F_X(0.9)]^{12} = 1 - 0.9^{12} = 0.718$$

8.2.3 Generalized $X_{(i)}$

Osservazione importante 44. if we write $X_{(i)} \sim F_{(i)}(x)$, with $i = 1, \ldots, n$ we mean that $X_{(i)}$ is distributed following the i-th ordered statistic

Proposizione 8.2.5 (Distribution function)**.**

$$F_{(i)}(x) = \mathbb{P}\left(X_{(i)} \leq x\right) = \sum_{j=i}^{n} \binom{n}{j} F_X(x)^j \cdot (1 - F_X(x))^{n-j} \tag{8.12}$$

Osservazione 246. Affinché l'i-esima ordinata sia $\leq x$ devo avere i variabili che abbiano un valore sotto x ($F_X(x)^i = \mathbb{P}(X \leq x)^i$) e $n-1$ sopra ($\mathbb{P}(X > x)^{n-1}$) = $(1 - F_X(x))^{n-i}$.
Penso che la proof sotto sia una spiegazione della sommatoria a partire da i invece che da 1

Dimostrazione. Not proved formally but to give intuitio, imagine $n = 3$ with $x_{(1)} = 3$, $x_{(2)} = 5$, $x_{(3)} = 7$. We have that $\mathbb{P}\left(X_{(2)} \leq x\right)$ is the probability that 2 rvs are $\leq x$ *OR* the probability that 3 random variables are $\leq x$. □

Proposizione 8.2.6 (Density function)**.**

$$f_{(i)}(x) = \binom{n}{i} i F_X(x)^{i-1} \cdot f_X(x)(1 - F_X(x))^{n-i} \tag{8.13}$$

Osservazione importante 45. Eg when $i = 1$ we obtain the formula for minimum

$$\begin{aligned} f_{(i)}(x) &= \binom{n}{1} 1 F_X(x)^0 \cdot f_X(x)(1 - F_X(x))^{n-1} \\ &= n f_X(x) \cdot [1 - F_X(x)]^{n-1} \end{aligned}$$

while for $i = n$ the maximum

$$f_{(n)}(x) = \binom{n}{n} n F_X(x)^{n-1} \cdot f_X(x)(1 - F_X(x))^0 = n\left[F_X(x)\right]^{n-1} f_X(x)$$

Esempio 8.2.5. Let $X_1, \ldots, X_n \sim \text{Unif}(0,1)$ be n iid uniforms, therefore having

$$, f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{elsewhere} \end{cases}, \quad F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 < x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

The k-th ordered statistic is distributed as a beta. Let's see it:

$$f_{(k)}(x) = k \binom{n}{k} x^{k-1}(1 - x)^{n-k}$$

Now we have that

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

Therefore

$$f_{(k)}(x) = \frac{1}{B(k, n-k+1)} x^{k-1}(1-x)^{n-k}$$

or $X_{(k)} \sim \text{Beta}\,(k, n-k+1)$. As special cases

$$X_{(1)} \sim \text{Beta}\,(()1, n)$$
$$X_{(n)} \sim \text{Beta}\,(()n, 1)$$

8.3 Inequalities

8.3.1 Markov (Viroli)

Teorema 8.3.1. *Given $X \in \mathbb{R}^+$, $D_X = \mathbb{R}^+$, $\lambda > 0$*

$$\mathbb{P}\left(X \geq \lambda \cdot \mathbb{E}\left[X\right]\right) \leq \frac{1}{\lambda} \tag{8.14}$$

Dimostrazione. Let

$$\mathbb{E}\left[X\right] = m = \int_{D_X} x \cdot f(x)\, \mathrm{d}x = \int_0^{+\infty} x \cdot f(x)\, \mathrm{d}x$$

Now

$$m \geq \int_{\lambda m}^{+\infty} x \cdot f(x)\, \mathrm{d}x \geq \int_{\lambda m}^{+\infty} x \cdot m \cdot f(x)\, \mathrm{d}x = \lambda m \underbrace{\int_{\lambda m}^{+\infty} \cdot f(x)\, \mathrm{d}x}_{\mathbb{P}\left(X \geq \lambda \cdot m\right)}$$

therefore

$$m \geq \lambda m\, \mathbb{P}\left(X \geq \lambda \cdot m\right) \iff \frac{1}{\lambda} \geq \mathbb{P}\left(X \geq \lambda \cdot m\right)$$

□

8.3.2 Tchebychev (Viroli)

Osservazione importante 46. We have two equivalent formulations

Teorema 8.3.2 (Tchebychev inequality)**.** *Respectively*

$$\mathbb{P}\left(|X - \mathbb{E}\left[X\right]| \geq \lambda \cdot \sigma_X\right) \leq \frac{1}{\lambda^2} \tag{8.15}$$

$$\mathbb{P}\left(|X - \mathbb{E}\left[X\right]| < \lambda \cdot \sigma_X\right) \geq 1 - \frac{1}{\lambda^2} \tag{8.16}$$

where σ_X is the standard deviation of X

Dimostrazione. We do by applying Markov inequality to $Y = (X - \mathbb{E}[X])^2$. We have that $\mathbb{E}[Y] = \sigma_X^2$ (by definition of variance), so by Markov

$$\mathbb{P}(Y \geq \lambda \mathbb{E}[Y]) \leq \frac{1}{\lambda}$$
$$\mathbb{P}\left((X - \mathbb{E}[X])^2 \geq \lambda \sigma_X^2\right) \leq \frac{1}{\lambda}$$
$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \sqrt{\lambda}\sigma_X\right) \leq \frac{1}{\lambda}$$

Then by setting $\lambda^* = \sqrt{\lambda}$ we conclude

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda^* \sigma_x) \leq \frac{1}{(\lambda^*)^2}$$

□

8.3.3 Tchebychev (Rigo)

Teorema 8.3.3. *For any real random variable X*

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}[|X|^\alpha]}{c} \tag{8.17}$$

Rigo's proof. In general, given an event A in $\mathscr{F}$ we have the indicator random variable

$$I_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}, \quad \mathbb{E}[I_A] = \mathbb{P}(A)$$

To prove Tchebychev lets define

$$A = \{w : |X(w)| \geq c\}$$

then

$$\mathbb{E}[|X|^\alpha] \overset{(1)}{\geq} \mathbb{E}[I_A \cdot |X|^\alpha] \overset{(2)}{\geq} \mathbb{E}[c^\alpha I_A] = c^\alpha \mathbb{E}[I_A] = c^\alpha \mathbb{P}(A)$$

where (1) because $|X|^\alpha \geq I_A \cdot |X|^\alpha$, (2) since $|X(w)|^\alpha \geq c^\alpha$. Therefore we conclude that

$$\mathbb{P}(A) \leq \frac{\mathbb{E}[X]^\alpha}{c^\alpha}$$

□

Osservazione 247. Useful because it applies to any random variable without any assumption and gives an upper bound of the prob.

Osservazione 248. An important special case is when $X = Y - E(Y)$ and $\alpha = 2$, in this case the inequaliy goes to

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq c) \leq \frac{\mathrm{Var}[Y]}{c^2}$$

But to apply Tchebychev in this form we need to know that the variance exists.

8.3.4 Jensen

Proposizione 8.3.4. *Let X be a real random variable and $f : ItoR$ a function defined on interval I. Now suppose that*

1. *f is a convex function, $\mathbb{P}(X \in I) = 1$*
2. *$\mathbb{E}[|X|] < +\infty$, $\mathbb{E}[|f(X)|] < +\infty$*

Then:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Osservazione importante 47 (Convex function (conca tipo $y = x^2$)). Btw f is convex function if

$$f[(1-\alpha)x + \alpha y] \leq (1-\alpha)f(x) + \alpha f(y), \quad \forall \alpha \in [0,1], x, y \in I$$

If f is twice differentiable, f is convex if and only if the second derivatibe is ≥ 0.

Osservazione importante 48 (Strictly convex). The same as above but instead of $\geq$ we have $>$ for both criteria

Esempio 8.3.1. Let's see some application of Jensen inequality.

- $f(x) = x^2$ is convex (second derivative $= 2 \geq 0$). If we apply Jensen we find out that
$$\mathbb{E}\left[X^2\right] \geq [\mathbb{E}[X]]^2 \tag{8.18}$$
This was already known sice variance is ≥ 0 (by computational formula of variance).
- absolute value $f(x) = |x|$ (second derivative $= 0$); applying Jensen we discover something new
$$\mathbb{E}[|X|] \geq |\mathbb{E}[X]| \tag{8.19}$$
- $f(x) = x^{b/a}$ for any $x >= 0$ with $(0 < a < b)$. Applying Jensen
$$\mathbb{E}\left[|X|^b\right] = \mathbb{E}\left[(|X|^a)^{\frac{b}{a}}\right] \geq [\mathbb{E}[|X|^a]]^{\frac{b}{a}} \tag{8.20}$$
thus Jensen implies that
$$\mathbb{E}\left[(|X|^a)^{\frac{1}{a}}\right] \leq \mathbb{E}\left[\left(|X|^b\right)^{\frac{1}{b}}\right]$$

Proposizione 8.3.5. *Under the condition of Jensen inequality suppose also that X is non degenerate/Dirac and f is strictly convex. (eg not the absolute value), then*

$$\mathbb{E}[f(X)] > f(\mathbb{E}[X]) \tag{8.21}$$

Osservazione 249. Now we prove that the rv is degenerate iff its variance is 0.

Proposizione 8.3.6.

$$X \sim \delta. \iff \mathrm{Var}[X] = 0$$

Dimostrazione. Respectively:

- if $X = a$ almost surely $\mathbb{P}(X = a) = 1$ then $\mathbb{E}[X] = a$ and also $\mathbb{E}\left[X^2\right] = a^2$ so that $\mathrm{Var}[X] = 0$
- otherwise suppose $\mathrm{Var}[X] = 0$: we prove that by contradition. We use the fact that $f(x) = x^2$ is strictly convex and apply Jensen inequality, if by absurd X is non degenerate we get

  $$\mathbb{E}\left[X^2\right] = \mathbb{E}[f(X)] > f(\mathbb{E}[X]) = [\mathbb{E}[X]]^2$$

  this happens if and only if $\mathrm{Var}[X] > 0$, but we assumed $\mathrm{Var}[X] = 0$ so we found a contradition

□

8.4 Transformation of random variables

Osservazione 250. The logic behind is that, if X is a rv and g is a "well behaved" function (mainly *strictly increasing* or *strictly decreasing*), then $g(X)$ is also a rv. Our main aim is determine density function of $g(X)$. In the discrete case finding PMF of $g(X)$ is usually easy, the following is an example.

Esempio 8.4.1 (Transformation of a bernoulli)**.** Let $X \sim \mathrm{Bern}(p)$ and we're interested in $g(X) = e^X$. What is the dist of $g(X)$. We have that

$$X = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases}, \qquad g(X) = \begin{cases} e^1 = e & \text{with prob } p \\ e^0 = 1 & \text{with prob } 1-p \end{cases}$$

Therefore

$$\mathbb{P}(g(X) = e) = \mathbb{P}\left(X = g^{-1}(e)\right) = \mathbb{P}(X = 1) = p$$

Osservazione 251. For the continuous case we have that, in order to obtain $f_{g(X)}(x)$ we need to differentiate $F_{g(X)}(x)$

$$F_{g(X)}(x) = \mathbb{P}(g(X) \leq x)$$

Da qui in poi sviluppo mio ma non mi tornano i conti con la formula finale Now

- if the function g is *decreasing* we have

  $$F_{g(X)}(x) = \mathbb{P}(g(X) \leq x) = \mathbb{P}\left(X \geq g^{-1}(x)\right) = 1 - \mathbb{P}\left(X < g^{-1}(x)\right) = 1 - F_X(g^{-1}(x))$$

- viceversa if g is *increasing*

  $$F_{g(X)}(x) = \mathbb{P}(g(X) \leq x) = \mathbb{P}\left(X \leq g^{-1}(x)\right) = F_X(g^{-1}(x))$$

In any case after that we have that

$$f_{g(X)}(x) = \frac{\partial}{\partial x} F_{g(X)}(x) = \begin{cases} \frac{\partial\left(1 - F_X(g^{-1}(x))\right)}{\partial x} & \text{if increasing} \\ \frac{\partial\left(F_X(g^{-1}(x))\right)}{\partial x} & \text{if decreasing} \end{cases}$$
$$= \begin{cases} -f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x) \\ f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x) \end{cases}$$

Finally the supra dupra formula for rv transformation is:

$$f_{g(X)}(x) = f_X(g^{-1}(x)) \cdot \left| \frac{\partial g^{-1}(x)}{\partial x} \right| \tag{8.22}$$

Esempio 8.4.2 (Esercizio virol)**.** If $X \sim \text{Unif}\,(0,1)$ and $Y = -2\log X$, show that $Y \sim \chi^2_2$. We apply 8.22 and compare with χ^2_n one.
We have the transformation $y = -2\log x$ so to obtain the inverse

$$-\frac{1}{2}y = \log x \iff x = e^{-\frac{1}{2}y}$$

therefore $g^{-1}(Y) = \exp\left(-\frac{Y}{2}\right)$. We have, being X a uniform on 0,1, that $f_X(x) = 1 \cdot \mathbb{1}_{[0,1]}(x)$. Now

$$\frac{\partial}{\partial y} g^{-1}(y) = -\frac{1}{2} e^{-y/2}$$

So applying the formula we arrive at

$$f_Y(y) = \mathbb{1}_{[0,1]}\left(e^{-y/2}\right) \cdot \frac{1}{2} e^{-y/2}$$

Now we need to express $\mathbb{1}_{[0,1]}\left(e^{-y/2}\right)$ in terms of y. The domain of y so

$$\begin{aligned} &0 \leq e^{-y/2} \leq 1 \\ &-\infty < -y/2 \leq 0 \\ &0 < y \leq +\infty \end{aligned}$$

Finally

$$f_Y(y) = \mathbb{1}_{[0,+\infty)}(y) \cdot \frac{1}{2} e^{-y/2} = \begin{cases} \frac{1}{2} e^{-y/2} & \text{if } y \in [0, +\infty) \\ 0 & \text{elsewhere} \end{cases}$$

which is a χ^2 with 2 degrees of freedom.

Esempio 8.4.3 (Esercizio Berk Tan)**.** Let $X \sim \text{Unif}\,(0,1)$ and be $g(x) = e^x$; then what is the pdf of $Y = g(X)$? We have that $g^{-1}(Y) = \log Y$, so

$$\frac{\partial}{\partial y}(g^{-1}(y)) = \frac{1}{y} =$$

Applying the formula

$$f_Y(y) = \mathbb{1}_{[0,1]}(\log y)\frac{1}{y}$$

and expressing $\mathbb{1}_{[0,1]}(\log y)$ in terms of y we have

$$\begin{aligned} &0 \leq \log y \leq 1 \\ &1 \leq y \leq e \end{aligned}$$

so finally

$$f_Y(y) = \mathbb{1}_{[1,e]}(y)\frac{1}{y} = \begin{cases} \frac{1}{y} & \text{if } y \in [1, e] \\ 0 & \text{elsewhere} \end{cases}$$

Esempio 8.4.4 (Esercizio 2 assignment 1, Braglia)**.** Let $X \sim \text{Unif}(0,1)$ and $Y = X^{\frac{1}{\alpha}}$, with $\alpha > 0$. Let's obtain $f_Y(y)$ by applying:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{\partial g^{-1}(y)}{\partial y} \right| \tag{8.23}$$

Being $X \sim \text{Unif}(0,1)$ we have that $f_X(x) = \mathbb{1}_{[0,1]}(x)$. Given the transformation $y = x^{1/\alpha}$, its inverse is

$$y = x^{1/\alpha} \iff y^\alpha = x$$

so $g^{-1}(Y) = Y^\alpha$; doing the partial derivative with respect to y we obtain:

$$\frac{\partial}{\partial y} g^{-1}(y) = \alpha y^{\alpha-1}$$

so putting things together:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{\partial g^{-1}(y)}{\partial y} \right| = \mathbb{1}_{[0,1]}(y^\alpha) \cdot \alpha y^{\alpha-1}$$

Now we need to express the indicator $\mathbb{1}_{[0,1]}(y^\alpha)$ in terms of y, therefore:

$$0 \leq y^\alpha \leq 1$$
$$0 \leq y \leq 1$$

Finally:

$$f_Y(y) = \mathbb{1}_{[0,1]}(y) \cdot \alpha y^{\alpha-1} = \begin{cases} \alpha y^{\alpha-1} & \text{if } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases}$$

If $\alpha = 1$, as expected

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases} = \mathbb{1}_{[0,1]}(y) \implies Y \sim \text{Unif}(0,1)$$

8.5 Conditional distribution

Osservazione 252. Roughly speaking the problemi is: given 2 real random variable X, Y we aim to evaluate the distribution of Y given that $X = x$.

The conditional distribution of Y given X is any function of two variables

$$\mathbb{P}((X,Y) \in C | X = x), \qquad x \in \mathbb{R}, C \in \beta\mathbb{R}^n$$

satisfying the following properties:

1. $\forall x in \mathbb{R}$, the map $C \to \mathbb{P}((X,Y) \in C | X = x)$ is a probability measure on $\beta\mathbb{R}^2$
2. $\mathbb{P}((X, Y \in C)) = E_X\{\mathbb{P}((X,Y) \in C | X = x)\}$, $\forall C \in \beta(\mathbb{R}^2)$, where E_X means expectation with respect to X.
   Thanks to this property, any time we aim to evaluate the probability $\mathbb{P}((X, Y \in C))$ we can use this equation.

3. $\mathbb{P}((X,Y) \in C|X = x) = \mathbb{P}((x,Y) \in C|X = x)$: since we're conditioning on $X = x$ we know that $X = x$ and can substitute it within parenthesis.

Osservazione importante 49 (Important remarks). Some important remarks:

1. if we know that $\mathbb{P}(X \in A) = 1$ for some $A \in \beta(R^n)$, then is suffices to assign $\mathbb{P}([(X,Y)inC|X = x])$, $\forall x \in A$ (and not necessarily $\forall xin\mathbb{R}$).
   For instance if $X \sim$ Unif $(()0,1)$, its enough to assign $P[(X,Y)inC|X = x]$, $\forall x \in (0,1)$

2. if $X \perp\!\!\!\perp Y$, then $\mathbb{P}((X,Y) \in C|X = x) = \mathbb{P}((x,Y)inC|X = x)$ is true by property 3 of the definition. Then i can drop the conditioning because X and Y are independent

$$P[(x;Y)inC|X = x] = P[(x;Y)inC]$$

3. it can be shown that the conditional distribution of Y given X, namely a function satisfying definition *always* exists and is *almost surely unique*.
   This remark is important because looking at the defn it's not sure that any function such as that defined exists. But this time fortunately the object exists: there are problem that can be solved only using the existence of conditional distribution and this is guaranteed;

4. the notation $\mathbb{P}((X,Y)inC|X = x)$ is very useful but also quite dangereous. Infact, if $P(X = x) = 0$ (which is possible eg in continuous distribution), then $P[(X,Y)inC|X = x]$ is *not* probability of intersection over probability $P(X = x)$; it's not

$$\text{not } \frac{\mathbb{P}(X = x, (X,Y) \in C)}{\mathbb{P}(X = x)}$$

This notation have not to be misleaded; for instance suppose $P(X = x) = 0, \forall xin\mathbb{R}$ (or equivalently the distribution function is continuous) then by the previous remark $\mathbb{P}((X,Y)inC|X = x)$ exists, but it certainly does not coincide with the ratio above. This because the ratio is not defined (you would have 0 at denominator and 0 at the numerator).

Osservazione 253. Unfortunately, in generale there is not an intuitive formula to evaluate conditional distribution (there is in some cases as we'll see later).

Esempio 8.5.1 (A usual question at the Rigo exam)**.** Suppose $X \perp\!\!\!\perp Y$ and Y has a continuous distribution function. What is the $\mathbb{P}(X = Y)$? This should be 0. Let's show it.
To answer let's define $C = \{(x,y)inR^2 : x = y\}$ which is the set of points constituting the diagonal

$$\mathbb{P}(X = Y) = \mathbb{P}((X,Y) \in C) \overset{(1)}{=} E_X\{prob(X,Y)inC|X = x\} \overset{(2)}{=} E_X\{prob(x,Y)inC|X = x\} \overset{(3)}{=} E_X\{pr$$

with:

- (1) by property 2 of defn,
- (2) by property 3 (since we're conditioning I can write x instead of X)

- (3) since thery are independent i can drop the conditioning
- (4) since being Y continuous, the probability that $Y = X$ (aka a single value) is zero

Osservazione importante 50. Note that:

- in statistical inference the elements of the sample are often assumed to be iid. Under this assumption, if the distribution of the character in the population is *continuous* what is the prob of having the sample with all different observation?
  It's 1 (almost sure event). This because $\mathbb{P}(X_i = X_j) = 0$, $\forall i \neq j$, so that the probability that $\mathbb{P}(X_1, ldots, X_n$ are all distinct$) = 1$
- if X and Y are independent but they are both discrete, then

$$\mathbb{P}(X = Y) = \sum_{xinB} \mathbb{P}(X = Y, X = x)$$

  where B is any set satisfying B finite or countable and $\mathbb{P}(X \in B) = 1$. Hence the $P(X = Y)$ can be written as above

$$\mathbb{P}(X = Y) = \sum_{x \in B} \mathbb{P}(X = Y, X = x) = \sum_{x \in B} \mathbb{P}(x = Y, X = x) \overset{(\perp\!\!\!\perp)}{=} \sum_{x \in B} \mathbb{P}(Y = x)\, \mathbb{P}(X = x)$$

  and this may be > 0.

Esempio 8.5.2. Suppose $X \perp\!\!\!\perp Y$, Y has a continuous distribution function. X, Y as above but we want to evaluate $\mathbb{P}(X = \sin(Y))$. It's 0 again. How to prove it?
A quick way to do it is the following: since X is independent of Y then X is still indipendent of any trasformation (and thus $\sin(Y)$).
Thus, to conclude that the $\mathbb{P}(X = sin(Y))$ it suffices to prove that, equivalently

- $\sin(Y)$ has a continuous distribution function (because if it's continuous we can repeat the argument of the previous exercise)
- $\mathbb{P}(\sin(Y) = a)) = 0$, $\forall ainR$ (this is trivially true if $a \notin [-1, 1]$)

We follow the second way, supposing $a \in [-1, 1]$ and define a set of random variable outcomes which the sinus is equal to a:

$$I_a = \{y \in \mathbb{R} : \sin y = a\}$$

we have that I_a is countable (pensa y sull'asse delle x, ci sono infiniti punti di seno che hanno altezza a). Thus the probability:

$$\mathbb{P}(\sin(Y) = a) = \mathbb{P}(Y \in I_a) \overset{(1)}{=} \sum_{y \in I_a} \mathbb{P}(Y = y) \overset{(2)}{=} \sum_{y \in I_a} 0 = 0$$

with (1) since I_a is countable and (2) because Y is a continuous distribution function

Esempio 8.5.3. Suppose $X \perp\!\!\!\perp Y$, $X \sim \text{Unif}\,(()0,1)$ and $Y \sim N(0,1)$. We want to evaluate the distribution function of the product XY.
Here conditional distribution become handy. For all $ain\mathbb{R}$, by definition the distribution function is

$$\mathbb{P}\,(XY \leq a) \overset{(1)}{=} E_X(\mathbb{P}\,(XY \leq a|X=x)) = E_X(\mathbb{P}\,(xY \leq a|X=x)) \overset{(\perp\!\!\!\perp)}{=} E_X(\mathbb{P}\,(xY \leq a)) \overset{(2)}{=} \int_{-\infty}^{+\infty} \mathbb{P}\,(xY$$

with (1) by definition, (2) since X is uniform (absolutely continuous), (3) because Y is normal.
After this we go to our friend mathematician asking for help.

Esempio 8.5.4. Let A, B, C iid with all $\sim \text{Exp}\,(()1)$. Lets define the random parabola

$$f(x) = Ax^2 + Bx + C, \qquad \forall x \in \mathbb{R}$$

random parabola because coefficiente a,b,c are rvs. What about the probability that f has real roots? it is the probability $\mathbb{P}\left(B^2 - 4AC \geq 0\right)$. To evaluate, we have to choose on one of the three variable and condition on it; eg let' condition on C

$$\mathbb{P}\left(B^2 - 4AC \geq 0\right) = E_C\left\{\mathbb{P}\left(B^2 \geq 4AC|C=c\right)\right\} = E_C\left\{\mathbb{P}\left(B^2 \geq 4Ac|C=c\right)\right\} \overset{(\perp\!\!\!\perp)}{=} E_C\left\{\mathbb{P}\left(B^2 \geq 4Ac\right)\right\}$$

where in (1) since C is exponential (continuous) arrived at (2) we have to evaluate $\mathbb{P}\left(B^2 \geq 4Ac\right)$ and its convenient to do it conditioning further on A, and finally using the fact that B is exponential (if $Z \sim \text{Exp}\,(\lambda)$ then $\mathbb{P}\,(Z > z) = e^{-\lambda z}$.

How to calculate $\mathbb{P}\,((X,Y)inC|X=x)$? We know this object exists and in many problem its enough to know it.

Unfortunately there is not a general formula which allows to calculate the probability above in every situation. Such a formula exists in *two special cases*:

1. X discrete
2. (X,Y) absolutely continuous

Definizione 8.5.1 (Discrete case)**.** If X is discrete, there is a set $B \subset \mathbb{R}$, B finite or countable, $\mathbb{P}\,(X \in B) = 1$, and $\mathbb{P}\,(X = x) > 0$, $\forall x \in B$ (true by definition of discreteness). Hence it suffices to let

$$\mathbb{P}\,((X,Y) \in C|X=x) = \frac{\mathbb{P}\,(X=x|(X,Y) \in C)}{\mathbb{P}\,(X=x)}, \forall x \in B, \forall C \in \beta(\mathbb{R}^2)$$

(this is the base definition of conditional probability with positive denominator, being the distribution discrete and focusing on $x \in B$).

Definizione 8.5.2 (Continuous case)**.** If (X,Y) is absolutely continuous with joint density $f(x,y)$, then the conditional distribution of Y given $X = x$ is still absolutely continuous with *conditional density*:

$$h(y|x) = \frac{f(x,y)}{f_1(x)}$$

where f_1 is the marginal density of X, namely the integral of the joint density in $\mathrm{d}y$:

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y$$

Hence the distribution function of Y given $X = x$ is

$$\mathbb{P}\left(Y \leq y | X = x\right) = \int_{-\infty}^{y} \frac{f(x,t)}{f_1(x)}\,\mathrm{d}t, \qquad \forall x, y \in \mathbb{R} : f_1(x) > 0$$

in this special case, we have an explicit formula for the conditional distribution

Definizione 8.5.3. In general, given any random vector $\mathbf{X} = (X_1, ldots, X_n)^T$, the corresponding order statistics are $X_{(1)}leqX_{(2)}leq \ldots leqX_{(n)}$ where $X_{(1)}, \ldots, X_{(n)}$ are obtained by arranging $X_1, ldots, X_n$ in increasing order.

Esempio 8.5.5. If $n = 2$, $X_{(1)} = min(X_1, X_2)$ and $X_{(2)} = max(X_1, X_2)$

Teorema 8.5.1 (Teorema di Rigo)**.** *If $X_1, \ldots, X_n$ are iid and absolutely continuous with density g, then the vector of order statistics is $(X_1, ...X_{(n)}^T)$ is still absolutely continuous with joint density:*

$$f(X_1, \ldots, X_n) = \begin{cases} n! \prod_{i=1}^n g(x_i) & \text{if } x_1 < \ldots < x_n \\ 0 & \text{otherwise} \end{cases}$$

somewhat intuitively the result is not too strange: intuitively $\prod_{i=1}^n g(x_i)$ is the density of the original vector, composed of iid vars; we have $n!$ permutation to produce the same arrangement ... meh per adesso

Esempio 8.5.6 (Example with order statistics)**.** Let S and T be iid with $S \sim$ Unif $(()0, 1)$. Define $X = \min(S,T)$ and $Y = \max(S,T)$. We want the conditional distribution of Y given $X = x$ we aim to write it explicitly, in this example X is absolutely continuous by the previous theorem.
Since (X,Y) are exactly the order statistic corresponding to (S,T), the above thm implies that (X,Y) are absolutely continuous.
Hence since its absolutely continuous, we have the formula and the conditional distribution of Y given X is still absolutely continuous with density

$$h(y|x) = \frac{f(x,y)}{f_1(x)}$$

in this case $f(x,y)$ (look at ordered statistics, rigo theorem)

$$f(x,y) = \begin{cases} 2!g(x)g(y) & if x < y \\ 0 & \text{otherwise} \end{cases}$$

where g is density of Unif $(0,1)$

$$g(x) = \begin{cases} 1 & if x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

hence the joint density and marginal density of X are respectively

$$f(x,y) = \begin{cases} 2 & if 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$
$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\, \mathrm{d}y = \int_x^1 2\, \mathrm{d}y = 2(1-x)$$

and finally

$$h(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x}, \quad 0 < x < y < 1$$

Now a bits of *interpretation* on the results: Since S and T are iid Unif $(()0,1)$ observing the pair (S,T) is like to select "at random" a point form the unit square.
Suppose now that $X = \min(S,T)$; what can be said about $Y = \max(S,T)$? Certainly $Y > X$ so if we fix a point $x \in [0,1]$, y is above the diagonal $y = x$, that is it is in the $[x,1]$. In fact the distribution of $Y|X \sim$ Unif $(()x,1)$: and this is why we obtained $1/(1-x)$ as density (coming from that distribution)

8.6 Multivariate normal

Osservazione 254. Let's start from univariate and see that multivariate formula are univariate generalization

Proposizione 8.6.1 (Characteristic functions of univariate normal)**.** *If* $Z \sim \mathrm{N}(0,1)$ *and* $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ *then* $\forall t \in \mathbb{R}$*:*

$$\phi_Z(t) = e^{-t^2/2} \tag{8.24}$$
$$\phi_X(t) = e^{it\mu - \frac{1}{2}(t\sigma)^2} \tag{8.25}$$

Dimostrazione. If $Z \sim \mathrm{N}(0,1)$: then its characteristic function is

$$\phi_Z(t) = \mathbb{E}\left[e^{itZ}\right] = \int_{-\infty}^{+\infty} e^{itx} \frac{\exp\left(-\frac{1}{2}x^2\right)}{\sqrt{2\pi}}\, \mathrm{d}x \overset{(1)}{=} \ldots = e^{-t^2/2}, \quad \forall t \in \mathbb{R}$$

(in (1) after doing calculation). If $X \sim \mathrm{N}\left(()\mu, \sigma^2\right)$ then X can be written as $X = \mu + \sigma Z$ with $Z \sim \mathrm{N}\left(()0,1\right)$ and thus we can derive the formula given the definition (in the univariate case as)

$$\phi_X(t) = \mathbb{E}\left[e^{it(\mu+\sigma Z)}\right] = \mathbb{E}\left[\underbrace{e^{it\mu}}_{\text{constant}}\, e^{it\sigma Z}\right] = e^{it\mu}\, \mathbb{E}\left[e^{i(t\sigma)Z}\right]$$
$$= e^{it\mu} \cdot \phi_Z(t\sigma) = e^{it\mu - \frac{1}{2}(t\sigma)^2} \qquad \forall t \in \mathbb{R}$$

□

Osservazione 255. MVN is not so important for this course: it's very important for statistician, but from point of view of probability it's just a special distribution among the others.

Definizione 8.6.1. Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be n dimensional random vector, then X is said to be normally distribuited with parameters μ and Σ, where $\mu \in \mathbb{R}^n$ and Σ is a $n \times n$ symmetric non-negative definite (geq 0) matrix (or also said semidefinite positive), if the characteristic function of X is given by:

$$\phi_X(t) = \mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{X}}\right] = \mathbb{E}\left[e^{i\mathbf{t}^T\mu - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}}\right], \quad \forall \mathbf{t} \in \mathbb{R}^n$$

Osservazione 256. Our definition includes not only absolutely continuous normal vector, but also other (eg degenerate in some cases)

Osservazione importante 51. Some remarks:

- the meaning of the two parametrs: mu is the vector of the mean, Sigma is the so called covariance matrix which have variances on the diagonal, covariance out of main diagonal

$$\mu = \begin{bmatrix} \mathbb{E}[X_1] \\ \ldots \\ \mathbb{E}[X_n] \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \operatorname{Var}[X_1] & \operatorname{Cov}(X_1, X_2) & \ldots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}[X_2] & \ldots & \operatorname{Cov}(X_2, X_n) \\ \ldots & \ldots & \ldots & \ldots \\ \operatorname{Cov}(X_n, X_1) & \ldots & \ldots & \operatorname{Var}[X_n n] \end{bmatrix}$$

- if Σ is positive-definite > 0 (not only ≥ 0) then $\mathbf{X}$ is absolutely continuous with density

$$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right\}, \qquad \forall \mathbf{x} \in \mathbb{R}^n$$

  The univariate density we know is a special case where the matrix Σ is positive definite (otherwise matrix can be inverted). For $n = 1$ the density Σ reduce to a scalar (σ^2, variance of the variable) and μ to a single number

$$f(x) = \frac{\exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)}{\sigma\sqrt{2\pi}}$$

- if Σ is non negative $>= 0$ definite but $\det \Sigma = 0$ then X is still normal, but the distribution of X is no longer absolutely continuous. For instance if $n = 1$ and $\sigma^2 = \Sigma = 0$ then

$$\phi_X(t) = e^{-it\mu}$$

  and $X = mu$ is degenerate. In other terms if $n = 1$, the above definition implies that the degenerate random variable are normal in a way.

- a **linear trasformation** of a normal random variable is still normal: if $\mathbf{X} \sim \mathrm{N}(\mu, \Sigma)$ and $\mathbf{Y} = \alpha + \mathbf{A}\mathbf{X}$ where the matrix $\mathbf{A}$ is $m \times n$ and $\alpha \in \mathbb{R}^n$, then $\mathbf{Y} \sim \mathrm{N}\left(\alpha + \mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T\right)$

Linear transformation proof. In order to prove that if $\mathbf{X} \sim \mathrm{MVN}(\mu, \mathbf{\Sigma})$ then $\mathbf{Y} = \alpha + \mathbf{A}\mathbf{X} \sim \mathrm{MVN}\left(\alpha + \mathbf{A}\mu, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\right)$ we write the characteristic function of

$\mathbf{Y}$ according to the definition above. Let's evaluate it:

$$\begin{aligned}\mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{Y}}\right] = \mathbb{E}\left[e^{i\mathbf{t}^T\alpha+\mathbf{AX}}\right] = \mathbb{E}\left[\underbrace{e^{it^T\alpha}}_{\text{constant}}\ e^{it^TAX}\right] = e^{it^T\alpha}\underbrace{\mathbb{E}\left[e^{it^TAX}\right]}_{\phi_X(A^Tt)} \\ = e^{it^T\alpha}e^{it^TA\mu-\frac{1}{2}t^TA\sigma A^Tt} = \exp\left(it^T(\alpha+A\mu)-\frac{1}{2}t^T(A\sigma A^T)t\right) \\ \iff Y\sim \mathrm{N}\left(\alpha+\mathbf{A}\mu, \mathbf{A\Sigma A}^T\right)\end{aligned}$$

□

Osservazione importante 52. As a consequence of the linear transformation, if $\mathbf{X}$ is normal, all marginals are still normal being the marginal obtained via a linear transformation (therefore we get a normal) that merely extract the marginal/subset. Eg

$$\mathbf{Y} = \begin{bmatrix}X_1\\X_2\\X_4\end{bmatrix} = \begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix}\begin{bmatrix}X_1\\X_2\\X_3\\X_4\end{bmatrix} = \mathbf{AX}$$

8.7 Exercises vari

Esempio 8.7.1 (Esercizio viroli, primo set)**.** Let $\mathbf{X} = \begin{bmatrix}X_1\\X_2\end{bmatrix}$ be a bivariate vector with joint density $f_{\mathbf{X}}(x_1,x_2) = 2e^{-(x_1+x_2)}$ where $X_1 > X_2 > 0$

1. find $M_{\mathbf{X}}(t)$
2. compute $\mathbb{E}[X_1]$ by $M_{\mathbf{X}}(t)$
3. compute $\mathbb{E}[X_1]$ by definition
4. are $X_1 \perp\!\!\!\perp X_2$, both by density and by moment generating function

1. we have

$$\begin{aligned}M_{\mathbf{X}}(t) &= 2\int_0^{+\infty}\int_{x_2}^{\infty} e^{tx_1}e^{tx_2}e^{-(x_1+x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= 2\int_0^{+\infty} e^{-x_2(1-t_2)}\cdot\int_{x_2}^{+\infty}\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= 2\int_0^{\infty} e^{-x_2(1-t_2)}\cdot\left[-\frac{e^{x_1(1-t_1)}}{1-t_1}\right]_{x_2}^{\infty}\mathrm{d}x_2 \\ &= 2\frac{1}{1-t_1}\int_0^{+\infty} e^{-x_2(2-t_1-t_2)}\,\mathrm{d}x_2 \\ &= \frac{2}{(1-t_1)(2-t_1-t_2)}\end{aligned}$$

2. we have

$$\begin{aligned}\frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial t_1}\Big|_{\mathbf{t=0}} &= 2(1-t_1)^{-2}(2-t_1-t_2)^{-1} + 2(1-t_1)^{-1}(2-t_1-t_2)^{-2}\Big|_{\mathbf{t=0}} \\ &= \frac{2}{2} + \frac{2}{4} = \frac{3}{2}\end{aligned}$$

3. it's longer, we have:

$$\mathbb{E}[X_1] = \int_{D_{X_1}} x_1 f_{X_1}(x_1)\,\mathrm{d}x_1$$

where

$$\begin{aligned}f_{X_1}(x_1) &= \int_{D_{X_2}} f_{\mathbf{X}}(x_1,x_2)\,\mathrm{d}x_2 \\ &= \int_0^{x_1} 2e^{-(x_1+x_2)}\,\mathrm{d}x_2 = \int_0^{x_1} 2e^{-x_1}e^{-x_2}\,\mathrm{d}x_2 \\ &= 2e^{-x_1}\cdot\int_0^{x_1} e^{-x_2}\,\mathrm{d}x_2 = 2e^{-x_1}\left[-e^{-x_2}\right]_0^{x_1} \\ &= 2e^{-x_1}(1-e^{-x_1}) = 2e^{-x_1} - 2e^{-2x_1}\end{aligned}$$

therefore

$$\begin{aligned}\mathbb{E}[X_1] &= \int_0^{+\infty} x_1\left(2e^{-x_1} - 2e^{-2x_1}\right)\mathrm{d}x_1 \\ &= 2\underbrace{\int_0^{+\infty} x_1 e^{-x_1}\,\mathrm{d}x_1}_{expectedvalueof\ \mathrm{Exp}\,(1)} - \underbrace{\int_0^{+\infty} x_1 2e^{-2x_1}\,\mathrm{d}x_1}_{expectedvalueof\ \mathrm{Exp}\,(2)} \\ &= 2\cdot 1 - \frac{1}{2} = \frac{3}{2}\end{aligned}$$

4. by the density

$$f_{X_2}(x_2) = \int_{x_2}^{+\infty} 2e^{-(x_1+x_2)}\,\mathrm{d}x_1 = 2e^{-x_1}\cdot\left[-e^{-x_1}\right]_{x_2}^{+\infty} = e^{-x_2}e^{-x_2} = e^{-2x_2}$$

Now we check if $f_{X_1}(x_1)\cdot f_{X_2}(x_2) = f_{\mathbf{X}}(x_1,x_2)$:

$$2e^{-x_1}(1-e^{-x_1})e^{-2x_2}2 \neq 2e^{-(x_1+x_2}$$

therefore they are not independent.
Now let's check according to the moment generating function; we observe that:

$$M_{X_1}(()t_1) = M_{\mathbf{X}}(t_1,0) = \frac{2}{(1-t_2)\frac{1}{2-t_1}} \quad M_{X_2}(()t_2) = M_{\mathbf{X}}(0,t_2) = \frac{2}{2-t_2}$$

Since $M_{\mathbf{X}}(\mathbf{t}) \neq M_{X_1}(()t_1)M_{X_2}(()t_2)$ are not independent.
Note: in case of mutually independent rvs:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} f_{X_i}(x_i)$$
$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} F_{X_i}(x_i)$$
$$M_{\mathbf{X}}(\mathbf{t}) = \prod_{i=1}^{p} M_{X_i}(t_i)$$

Esempio 8.7.2 (Virols, eserciziario 1, es 2)**.** Given a sequence of independent rvs $X_i \sim \text{Pois}(\lambda_i)$ find $M_Y(t)$ where $Y = \sim_{i=1}^{n} X_i$.
We first determine $M_{X_i}(t)$

$$\begin{aligned} M_{X_i}(t) = \mathbb{E}\left[e^{tX_i}\right] &= \sum_{x_i=0}^{\infty} e^{tx_i} \frac{1}{x_i!} e^{-\lambda_i} \lambda_i^{x_i} \\ &= \sum_{x_i=0}^{\infty} (e^t \lambda_i)^{x_i} \frac{1}{x_i!} e^{-\lambda_i} \\ &\overset{(1)}{=} e^{-\lambda_i} \cdot e^{\lambda_i e^t} = e^{-\lambda_i(1-e^t)} \end{aligned}$$

where in (1) we used $\sum_{x=0}^{\infty} \frac{c^x}{x!} = e^c$. Therefore $Y \sim \text{Pois}\left(\sum_{i=1}^{n} \lambda_i\right)$

Esempio 8.7.3. Consider the function

$$f(x|y) = \begin{cases} \frac{y^x e^{-y}}{x!} & \text{for } x = 0, 1, 2, \ldots \text{ and } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

1. if the marginal pdf of Y is $\text{Exp}(1)$, what is the joint probability function of (X, Y)
2. derive the marginal probability function of X

We have:

1. for the joint probability

$$f_{X,Y}(x, y) = f(y) \cdot f(x|y) = e^{-y} \frac{y^x e^{-y}}{x!} = \frac{y^x e^{-2y}}{x!}$$

2. for the marginal probability of X

$$\begin{aligned} f_X(x) = \int_0^{+\infty} \frac{y^x e^{-2y}}{x!} \, \mathrm{d}y &= \frac{1}{x!} \underbrace{\int_0^{+\infty} y^x e^{-2y} \, \mathrm{d}y}_{(1)} \\ &= \frac{1}{x!} \frac{\Gamma(x+1)}{2^{x+1}} = \frac{1}{2^{x+1}} \end{aligned}$$

where (1) is the kernel of a $\Gamma\alpha = x + 1, \beta = 2$

Esempio 8.7.4 (Virols, eserciziario 1, es 4)**.** Let $X_1, \dots, X_n \sim \text{Geom}(p)$ iid rvs. Find $M_Y(t)$ where $Y = \sum_{i=1}^n X_i$. What can you say about the distribution of Y?
For a geometric rv we have

$$\mathbb{P}(X = x) = p(1-p)^{x-1}, \qquad D_X = \{1, 2, \dots\}$$

so

$$\begin{aligned} M_X(t) &= \sum_{x=1}^{\infty} e^{tx} p(1-p)^{x-1} = \sum_{x=1}^{\infty} e^{tx} p \frac{1-p}{1-p}(1-p)^{x-1} \\ &= \frac{p}{1-p} \cdot \sum_{x=1}^{\infty} \left[e^t(1-p)\right]^x = \frac{p}{1-p} \cdot \left(\sum_{x=0}^{\infty} \left[e^t(1-p)\right]^x - 1 \right) \end{aligned}$$

Now we define $q = 1-p$; if $|e^t(1-p)| < 1$ the previous series converges to $\frac{1}{1-qe^t}$. Therefore the $M_X(t)$ exists only for $e^t < \frac{1}{1-p}$, that is $t < -\log(1-p)$). For such values we have

$$M_X(t) = \frac{p}{q}\left(\frac{1}{1-qe^t} - 1\right) = \frac{p}{q}\left(\frac{qe^t}{1-qe^t}\right) = \frac{pe^t}{1-qe^t}$$

Now

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left[\frac{pe^t}{1-q^{e^t}}\right]^n$$

with the last being the moment generating function of a negative binomial distribution with parameters n and p

Esempio 8.7.5 (Virols eserciziario 1, es 6)**.** Let $X_1, \dots, X_n$ be a random sample from a Weibull (α, β) distribution, that is

$$f(x) = \alpha\beta x^{\beta 1} e^{-\alpha x^\beta}, \quad x > 0, \alpha, \beta > 0$$

Derive the probability density function of $X_{(1)}$ and recognize it.
The distribution function of a Weibull rv is

$$F_X(x) = 1 - e^{-\alpha x^\beta}$$

therefore

$$F_{(1)}(x) = 1 - [1 - F_X(x)]^n = 1 - \left[e^{-\alpha x^\beta}\right]^n = 1 - e^{-\alpha n x^\beta}$$

which is a weibull with parameters $n\alpha$ and β

Esempio 8.7.6 (Virols S01E07)**.** A rv X is said to have the two-parameter Pareto distribution with parameters α and β if its pdf is given by

$$f_X(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}, \qquad x > \beta, \alpha, \beta > 0$$

1. show that the function just given is indeed a pdf

2. set $Y = \frac{X}{\beta}$ and show that its pdf is given by $f_Y(y) = \frac{\alpha}{y^{\alpha+1}}$, $y > 1$ and $\alpha > 0$, which is referred to as the one-parameter Pareto distribution
3. show that $\mathbb{E}[X] = \frac{\alpha\beta}{\alpha-1}$
4. show that $\text{Var}[X] = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ with $\alpha > 2, \beta > 0$
5. let $\{X_n\}_{n\in\mathbb{N}}$ with $X_n \text{ Bin}\left(n, \frac{lambda}{n}\right)$. Prove that $X_n \xrightarrow{d} \text{Pois}(\lambda)$.
6.
7.

we have:

1. in order to be a proper pdf

$$\int_\beta^{+\infty} \frac{x\beta}{x^{\alpha+1}}\,\mathrm{d}x = 1$$

So

$$\int_\beta^{+\infty} \frac{x\beta}{x^{\alpha+1}}\,\mathrm{d}x = \alpha\beta^\alpha \cdot \int_\beta^{\infty} \frac{1}{x^{\alpha+1}} = \alpha\beta \cdot \left[-\frac{1}{\alpha}\cdot\frac{1}{x^\alpha}\right]_\beta^\infty$$
$$= \alpha\beta^\alpha \frac{1}{\alpha}\frac{1}{\beta^\alpha} = 1$$

2. we apply

$$f_Y(y) = \left|\frac{\partial g^{-1}(y)}{\partial y}\right| f_X(g^{-1}(y))$$

having

$$g^{-1}(y) = \beta y$$

so

$$f_Y(y) = \beta \cdot \alpha\beta^\alpha \frac{1}{\beta^{\alpha+1}y^{\alpha+1}} \underbrace{\mathbb{1}_{\beta,+\infty}(by)}_{=\mathbb{1}_{1,+\infty}(y)}$$

3. we have

$$\mathbb{E}[X] = \int_\beta^{+\infty} \frac{x\cdot\alpha\beta^\alpha}{x^{\alpha+1}}\,\mathrm{d}x = \alpha\beta^\alpha \cdot \underbrace{\int_\beta^\infty \frac{1}{x^\alpha}\,\mathrm{d}x}_{(1)}$$
$$= \alpha\beta^\alpha \frac{1}{\alpha-1}\frac{1}{\beta^{\alpha-1}} = \frac{\alpha}{\alpha-1}\beta$$

where (1) is the kernel of a Pareto with parameters $\alpha = 1$ and $\beta = 0$, therefore $a - 1 > 0$, $\alpha > 1$

4. we have that

$$\mathbb{E}\left[X^2\right] = \alpha\beta^\alpha \int_\beta^\infty x^2 \frac{1}{x^{\alpha+1}}\,\mathrm{d}x = \alpha\beta^\alpha \int_\beta^\infty \frac{1}{x^{\alpha-1}}\,\mathrm{d}x = \alpha\beta^\alpha \frac{1}{\alpha-2}\frac{1}{\beta^{\alpha_2}} = \frac{\beta^2\alpha}{\alpha-2}$$

Therefore:

$$\begin{aligned}\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 &= \frac{\beta^2\alpha}{\alpha-2} - \frac{\alpha^2\beta^2}{(\alpha-1)^2} \\ &= \frac{\beta^2(\alpha-1)^2 - \alpha^2\beta^2(\alpha-2)}{(\alpha-2)(\alpha-1)^2} \\ &= \frac{\beta^2\alpha^3 + \beta^2\alpha - 2\beta^2\alpha^2 - \alpha^3\beta^2 + 2\alpha^2\beta^2}{(\alpha-2)(\alpha-1)^2}\end{aligned}$$

5. take $M_{X_n}(t)$ of the binomial:

$$M_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right)^n = \left(1 - \frac{\lambda}{n}\big(1-e^t\big)\right)^n \overset{(1)}{=} \left(1 - \frac{a}{n}\right)^n$$

with (1) taking $a = \lambda(1-e^t)$. Therefore

$$\lim_{n\to\infty}\left(1 - \frac{a}{n}\right)^n = e^{-a}$$

therefore $X_n \xrightarrow{d} X$ with $M_X(t) = e^{-\lambda(1-e^t)}$. But this happens $\iff$ $X \sim \mathrm{Pois}\left(\lambda\right)$.

Esempio 8.7.7 (Virols S01E09)**.** Given $X_i \sim \mathrm{Pois}\left(\lambda\right)$, be iid rvs and $Y = \sum_{i=1}^n X_i$, find $M_Y(t)$.
First we derive $M_X(t)$ where $X \sim \mathrm{Pois}\left(\lambda\right)$

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{x=0}^\infty e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^\infty \frac{e^{-\lambda}(e^t\lambda)^x}{x!} = e^{-\lambda}\sum_{x=0}^\infty \frac{c^x}{x!} = e^{-\lambda}e^c$$

where $c = e^t\lambda$. Therefore

$$M_X(t) = e^{-\lambda}e^{e^t\lambda} = e^{\lambda(e^t-1)}$$

Now

$$M_Y(t) = M_{\sum_i X_i}(t) = \prod_{i=1}^n e^{\lambda_i(e^t-1)} = e^{(e^t-1)\sum_{i=1}^n \lambda_i}$$

therefore $\implies$ $Y \sim \mathrm{Pois}\left(\sum_{i=1}^n \lambda_i\right)$

Capitolo 9

Convergence

Osservazione importante 53 (Setup). Given a sequence of rvs, $X_1, X_2, \ldots$, the aim is to study

$$\{X_n\}_{n\in\mathbb{N}} \xrightarrow[n\to\infty]{} X$$

We have four types of convergence:

1. convergence in probability (weak)
2. convergence in law/distribution (weak)
3. almost sure convergence (strong)
4. convergence in mean of order k (strong)

9.1 Convergence in probability

Osservazione 257. it's the first type of convergence: this is a weak type (it implies convergency in distribution byt not implies strong kinds of convergency)

Definizione 9.1.1 (Convergence in probability)**.** We say that a sequence $\{X_n\}_{n\in\mathbb{N}}$ converges in probability to the *limit distribution* X and we write:

$$\{X_n\}_{n\in\mathbb{N}} \xrightarrow[n\to\infty]{p} X$$

if alternatively (equivalent definitions), $\forall \varepsilon > 0$:

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \tag{9.1}$$

$$\mathbb{P}\left(|X_n - X| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1 \tag{9.2}$$

Osservazione 258. The limit distribution X can be any rv (gaussian etc) but as a special case it's when X_n converges to a δ_θ (the constant θ); it's peculiar since in inference the sequence can be an estimator collasing to a point (eg population mean) and can be a good property for an estimator.

Definizione 9.1.2 (Weak consistence)**.** If $\{X_n\}_{n\in\mathbb{N}} \xrightarrow[n\to\infty]{p} \delta_\theta$ we say that X_n is (weakly) consistent for θ

Osservazione importante 54. Weak consistency means converging probability (link between probability and inference)

Esempio 9.1.1. Considering a sequence of iid rvs $\{X_n\}_{n\in\mathbb{N}} \sim \text{Unif}\,(0,\theta)$, with $\theta > 0$, the transformation (max of the first n)

$$\max_{1\leq i\leq n} X_i = X_{(n)}$$

Let's prove that $X_{(n)}$ is a consistent estimator for θ

$$X_{(n)} \xrightarrow{p} \delta_\theta$$

Remembering that $F_{(n)}(x) = [F_X(x)]^n$ we want to prove that

$$\mathbb{P}\left(\left|X_{(n)} - \theta\right| < \varepsilon\right) \to 1$$

Since $X_{(n)} - \theta$ is negative or null (being θ the max of the uniform rvs) we can avoid the absolute value multiplying by -1

$$\begin{aligned}\mathbb{P}\left(\left|X_{(n)} - \theta\right| < \varepsilon\right) &= \mathbb{P}\left(-X_{(n)} + \theta < \varepsilon\right) = \mathbb{P}\left(-X_{(n)} < \varepsilon - \theta\right) = \mathbb{P}\left(X_{(n)} > \theta - \varepsilon\right) \\ &= 1 - \mathbb{P}\left(X_{(n)} \leq \theta - \varepsilon\right) = 1 - F_{(n)}(\theta - \varepsilon) = 1 - [F_X(\theta - \varepsilon)]^n\end{aligned}$$

If $X \sim \text{Unif}\,(0,\theta)$, then $F_X(x) = \frac{x}{\theta}$, $0 \leq x \leq \theta$ so

$$\mathbb{P}\left(\left|X_{(n)} - \theta\right| < \varepsilon\right) = 1 - [F_X(\theta - \varepsilon)]^n = 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

and since $\frac{\theta-\varepsilon}{\theta} < 1$ with $0 < \varepsilon \leq \theta$

$$\lim_{n\to\infty} 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n = 1$$

Proposizione 9.1.1 (Sufficient conditions for conv. in prob. to Dirac)**.** *If both:*

$$\lim_{n\to+\infty} \mathbb{E}\left[X_n\right] = \theta \tag{9.3}$$

$$\lim_{n\to+\infty} \text{Var}\left[X_n\right] = 0 \tag{9.4}$$

then $\{X_n\}_{n\in\mathbb{N}} \xrightarrow[n\to\infty]{p} \delta_\theta$.

Osservazione 259. The viceversa does not hold: the conditions are sufficient, not needed (eg X_n can converge in prob even if these conditions are not met)

Dimostrazione. Applying Tchebychev inequality

$$\mathbb{P}\left(|X_n - \mathbb{E}\left[X_n\right]| < \lambda\sigma(X_m)\right) \geq 1 - \frac{1}{\lambda^2}$$

Now we define/substitute $\varepsilon = \lambda\sigma(X_m)$ so that $\lambda^2 = \frac{\varepsilon^2}{\sigma^2(X_m)}$; therefore

$$\mathbb{P}\left(|X_n - \mathbb{E}\left[X_n\right]| < \varepsilon\right) \geq 1 - \frac{\sigma^2(X_n)}{\varepsilon^2}$$

if $n \to +\infty$ the last term go to zero so

$$\mathbb{P}\left(|X_n - \mathbb{E}\left[X_n\right]| < \varepsilon\right) \geq 1$$

and since this probability can't be larger than 1, it must be 1 so

$$\mathbb{P}\left(|X_n - \mathbb{E}\left[X_n\right]| < \varepsilon\right) = 1 \implies X_n \xrightarrow{p} \theta$$

□

Esempio 9.1.2. Let $X_n \sim \text{Geom}\left(p_n\right)$ with $p_n = 1 - \frac{1}{n}$ (that is theta changes ...). We have that

$$\mathbb{P}\left(X_n = x\right) = p_n(1-p_n)^{x-1}$$

with $\mathbb{E}\left[X_n\right] = \frac{1}{p_n}$, $\text{Var}\left[X_n\right] = \frac{1-p_n}{p_n^2}$. Let's prove that $X_n \xrightarrow{p} \delta_1$.

$$\lim_{n\to\infty} \mathbb{E}\left[X_n\right] = \frac{1}{p_n} = \frac{1}{1-\frac{1}{n}} \to 1$$
$$\lim_{n\to\infty} \text{Var}\left[X_n\right] = \frac{\frac{1}{n}}{(1-\frac{1}{n})^2} = \frac{\frac{1}{n}}{(\frac{(n-1)^2}{n^2})} = \frac{n}{(n-1)^2} \to 0$$

9.1.1 Theorem: weak law of large numbers

Teorema 9.1.2 (Weak law of large numbers)**.** *Let X_n be a sequence of iid rvs with $\mathbb{E}\left[X_n\right] = \theta$ and $\text{Var}\left[X_n\right] = \sigma^2 < +\infty$; if we define the partial mean as the mean of the first n rvs*

$$M_n = \frac{\sum_{i=1}^n X_i}{n} \tag{9.5}$$

then we have that

$$M_n \xrightarrow{p} \delta_\theta \tag{9.6}$$

Dimostrazione. We have that

$$\mathbb{E}\left[M_n\right] = \frac{\sum_{i=1}^n \mathbb{E}\left[X_i\right]}{n} = \frac{n\theta}{n} = \theta$$
$$\text{Var}\left[M_n\right] = \frac{1}{n^n}\sum_{i=1}^n \text{Var}\left[X_i\right] = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

therefore since both

$$\lim_{n\to+\infty} \mathbb{E}\left[M_n\right] = \theta$$
$$\lim_{n\to+\infty} \text{Var}\left[M_n\right] = 0$$

the sufficient conditions are met and $M_n \xrightarrow{p} \delta_\theta$ □

9.2 Convergence in law/distribution

Osservazione 260. We have two equivalent definition, by limit of distribution function or convergence in law/distribution of the moment generating function.

Definizione 9.2.1 (Convergence in law (or distribution))**.** The sequence X_n converge in law (or distribution) to X, and we write $X_n \xrightarrow{d} X$, if and only if ($\iff$),

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$$

$\forall x \in D_X$ in which $F_X(x)$ is continuous.

Definizione 9.2.2 (Alternate definition)**.**

$$X_n \xrightarrow{d} X \iff M_{X_n}(t) \to M_X(t), \forall t : |t| < \varepsilon \tag{9.7}$$

in a intorno di $t = 0$

Osservazione 261. Two theorem without proof before going on

Teorema 9.2.1. *Convergence in probability is stronger than convergence in distribution since* $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$

Teorema 9.2.2. *... but in the case of dirac we have both implication* $X_n \xrightarrow{p} \delta_\theta \iff X_n \xrightarrow{d} \delta_\theta$

Esempio 9.2.1. Let $\{X_n\}_{n\in\mathbb{N}}$ be iid standard normal, $X_n \sim \mathrm{N}(0,1)$. Defining the following variable

$$Y_n = \frac{X_1^2 + \ldots + X_n^2}{n} = \frac{\chi_n^2}{n}$$

(at numerator we have a χ_n^2), prove that $Y_n \xrightarrow{d} \delta_1$.
We do it by moment generating function. Looking at the mgf of a chi square we have that:

$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

We have that $Y_n = \frac{\chi_n^2}{n}$ so its moment generating function (applying properties)

$$M_{Y_n}(()t) = M_{\frac{\chi_n^2}{n}}(()t) = M_{\chi_n^2}(()\frac{t}{n}) = \left(1 - 2\frac{t}{n}\right)^{-\frac{n}{2}}$$

We then have

$$\lim_{n\to+\infty} M_{Y_n}(t) = \lim_{n\to+\infty} \left(1 - 2\frac{t}{n}\right)^{-\frac{n}{2}} = e^t$$

this remembering that

$$\lim_{n\to+\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}$$
$$\lim_{n\to+\infty} \left(1 + \frac{a}{n}\right)^n = e^{a}$$

So we have found that

$$\lim_{n\to+\infty} M_{Y_n}(t) = e^t$$

Now looking at δ_θ it has a simple moment generating function; if $X \sim \delta_\theta$

$$M_X(t) = \mathbb{E}\left[e^t X\right] = e^{t\theta}$$

therefore if $X \sim \delta_1$, its $M_X(t) = e^t$ and is the limit developed above.

Esempio 9.2.2. Let $X_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$. Prove that $X_n \xrightarrow{d} \text{Pois}(\lambda)$.
Here again there's a moving probability $X_1 \sim \text{Bin}(n, \lambda)$, $X_2 \sim \text{Bin}(n, \lambda/2)$, $\dots X_n \sim \text{Bin}(n, \lambda/n)$. The mgf of generic binomial rv

$$M_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right)^n = \left(1 + \frac{\lambda}{n}\left(e^t - 1\right)\right)^n$$

Using $\lim(1 + a/n)^n = e^a$ we have that

$$\lim_{n\to+\infty} \left(1 + \frac{\lambda}{n}\left(e^t - 1\right)\right)^n = e^{\lambda(e^t-1)}$$

But this is the mgf for $\text{Pois}(\lambda)$.

9.2.1 Theorem: central limit theorem

Osservazione importante 55. fundamental theorem basis for inference; this is why low number of patients does not permit to have a good approximation (it would be for $n \to +\infty$ but are needed at least 20/30 patients for the approximation start working)

Osservazione 262. This can be defined equivalently in terms of partial sum $\sum_{i=1}^n X_i$ or partial mean $\frac{\sum_{i=1}^n X_i}{n}$ of iid random variables with finite expected value and variance.

Proposizione 9.2.3. *Let X_i be iid random variables, with mean $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$; let $S_n = \sum_{i=1}^n X_i$ be the partial sum and $M_n = \frac{\sum_{i=1}^n X_i}{n}$ the partial mean. If we define the standardized sum as*

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}[S_n]}} = \frac{S_n - n\mu}{\underbrace{\sqrt{n\sigma^2}}_{\text{no cov, } \perp\!\!\!\perp}} \overset{(1)}{=} \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where in (1) *we divided everything by n.*
Then $Z_n \xrightarrow{d} \text{N}(0, 1)$

Dimostrazione.

$$Z_n = \frac{S_n - n\mu}{\sigma \cdot \sqrt{n}} = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu -}{\sigma \cdot \sqrt{n}} = \sum_{i=1}^n \underbrace{\left(\frac{X_i - \mu}{\sigma}\right)}_{U_i} \cdot \frac{1}{\sqrt{n}} = \frac{\sum_{i=1}^n U_i}{\sqrt{n}}$$

with $\mathbb{E}[U_i] = 0$ and $\text{Var}[U_i] = 1$ (being standardized) and $\mathbb{E}\left[U_i^2\right] = 1$ as consequence of the first two using the variance formula $\text{Var}[U_i] = \mathbb{E}\left[U_i^2\right] -$

$\mathbb{E}[U_i]^2$.
Now for the moment generating function of Z_n we have

$$M_{Z_n}(t) = M_{\frac{\sum U_i}{\sqrt{n}}}(t) \overset{(1)}{=} M_{\sum U_i}\left(t/\sqrt{n}\right) \overset{(2)}{=} \prod_{i=1}^{n} M_{U_i}\left(t/\sqrt{n}\right) \overset{(3)}{=} \left[M_U\left(t/\sqrt{n}\right)\right]^n$$

with (1) by prop of mgf, (2) by independence and (3) since they are identically distributed. Since the mgf of standard normal is $e^{t^2/n}$, we want to prove that

$$\lim_{n\to+\infty} M_{Z_n}((t) = \left[M_U\left(t/\sqrt{n}\right)\right]^n = e^{t^2/2}$$

We decomposing $M_U(t/\sqrt{n})$ by Taylor (in point $t = 0$ so maclaurin) expansion. In general we have that

$$M_X(t) = 1 + t\,\mathbb{E}[X] + \frac{t^2}{2!}\,\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\,\mathbb{E}\left[X^3\right] + \dots$$

Applying this to $M_U(t/\sqrt{n})$ (two terms here are enough for what follows):

$$M_U\left(t/\sqrt{n}\right) = 1 + \frac{t}{\sqrt{n}}\underbrace{\mathbb{E}[U]}_{=0} + \frac{t^2}{n\cdot 2}\underbrace{\mathbb{E}\left[U^2\right]}_{=1} + \dots \simeq 1 + \frac{t^2}{2n}$$

therefore

$$M_{Z_n}(t) \simeq \left(1 + \frac{t^2}{2n}\right)^n$$

Finally

$$\lim_{n\to+\infty} M_{Z_n}(t) = \lim_{n\to+\infty}\left(1 + \frac{t^2}{2n}\right)^n = e^{t^2/2}$$

which is the mgf of $\mathrm{N}(0,1)$. □

9.3 Convergence in mean of order k

Definizione 9.3.1. Let $k \in \mathbb{N}^+$. It' said that $X_n \xrightarrow{L_k}$ if and only if

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^k\right] = 0$$

Osservazione importante 56 (Convergence in quadratic mean). One of the most famous is the L_2 where we say $n = 2$, $X_n \xrightarrow{L_2} X$ iff $\lim_{n\to\infty} \mathbb{E}[(X_n - X)] = 0$

Definizione 9.3.2 (Strongly consistence for constant)**.** If $X_n \xrightarrow{L_2} \delta_\theta$ that is

$$\lim_{n\to+\infty} \mathbb{E}\left[(X_n - \theta)^2\right] = 0$$

we say that X_n is strongly consistent for θ.

Proposizione 9.3.1. *In this type of convergence we have this result*

$$X_n \xrightarrow{L_2} \delta_\theta \iff \begin{cases} \lim_{n\to+\infty} \mathbb{E}[X_n] = \theta \\ \lim_{n\to+\infty} \operatorname{Var}[X_n] = 0 \end{cases}$$

Osservazione importante 57. In inference there are two types of consistency, *weak* consistency and *strong* consistency:

- weak type is convergence in probability
- strong is convergence in L_2 (quadratic mean)

In inference X_n is an estimator and θ iss the parameter you want to estimate. Consistency is a good property for an estimator to have; it's better to have strong because it implies weak.

Esempio 9.3.1. Let $X_n \sim \text{Pois}(2/n)$; let's check that

1. $X_n \xrightarrow{d} \delta_0$
2. $X_n \xrightarrow{L_2} \delta_0$

We have that

1. for the Poisson distribution we have that $M_{X_n}(t) = e^{\frac{2}{n}(e^t-1)}$. Taking the limit
$$\lim_{n\to+\infty} e^{\frac{2}{n}(e^t-1)} = e^0 = 1$$
For δ_0, the mgf is
$$M(t) = \mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^0\right] = 1$$
so same mgf we have proved the convergence

2. we have
$$\lim_{n\to+\infty} \mathbb{E}\left[(X_n - 0)^2\right] = \lim_{n\to+\infty} \mathbb{E}\left[X_n^2\right]$$
To obtain this we can use exploit formula; since X_n is a Poisson
$$\begin{aligned} \mathbb{E}[X_n] &= \frac{2}{n} \\ \operatorname{Var}[X_n] &= \frac{2}{n} \\ \mathbb{E}\left[X_n^2\right] &= \operatorname{Var}[X_n] + \mathbb{E}[X_n]^2 = \frac{2}{n} + \frac{4}{n^2} = \frac{2n+4}{n^2} \end{aligned}$$
And finally
$$\lim_{n\to+\infty} \frac{2n+4}{n^2} = 0$$
so it goes to δ_0

Esempio 9.3.2. Let $X_n \sim \text{Bern}\left(\frac{1}{n}\right) \cdot n$ so, its pmf be

$$X_n = \begin{cases} n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n \end{cases}$$

We have that

$$\mathbb{E}[X_n] = n \cdot \frac{1}{n} = 1$$
$$\mathrm{Var}[X_n] = n^2 \frac{1}{n}\left(1 - \frac{1}{n}\right) = n\left(\frac{n-1}{n}\right) = n - 1$$
$$\mathbb{E}\left[X_n^2\right] = n - 1 + 1 = n$$

We can't conclude X_n converges in L_2 because of the (limit of the) variance

$$\begin{cases} \lim_{n\to\infty} \mathbb{E}[X_n] = 1 \\ \lim_{n\to\infty} \mathrm{Var}[X_n] = +\infty \end{cases} \implies X_n \not\xrightarrow{L_2} \delta_1$$

Let's study if it converge in probability, where? to two possible distribution

- what about δ_1? we have that

  $$\mathbb{P}(|X_n - 1| < \varepsilon) \not\xrightarrow[n\to\infty]{} 1$$

  convergence is not true because look at $X_n \sim \mathrm{Bern}(1/n)$: 0 with larger and larger prob, n with lowering prob. Therefore $X_n - 1$ will be 1 with increasing prob and so $1 \not\leq \varepsilon$, $\forall varepsilon \in \mathbb{R}$.

- what about δ_0? we have that

  $$\mathbb{P}(|X_n| < \varepsilon) = \mathbb{P}(X_n < \varepsilon) \xrightarrow[n\to\infty]{} 1$$

  this is true so $X_n \xrightarrow{p} \delta_0$.
  So here we probed the convergence without the two sufficient condition (they're just sufficient, not needed; we can have convergence in prob even if we don't have the two sufficient conditions).

9.3.1 Strong law of large numbers

Osservazione 263. It's the most important theorem related to convergence in quadratic mean.

Teorema 9.3.2 (Strong law of large numbers)**.** *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of* indepent *random variables and assume* $\mathbb{E}[X_n] = \mu$, $\mathrm{Var}[X_n] = \sigma^2 < +\infty$. *Then we say that the partial mean:*

$$M_n \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{L_2} \mu$$

Dimostrazione.

$$\mathbb{E}\left[(M_n - \mu)^2\right] = \mathbb{E}\left[\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{n\mu}{n}\right)\right] \overset{(1)}{=} \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (X_i - \mu)^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[(X_i - \mu)^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}$$

where in (1) due to independence, the expectations of the cross products are all zeros, so the square of sums is the sum of squares. Finally

$$\lim_{n\to+\infty} \mathbb{E}\left[(M_n - \mu)^2\right] = \lim_{n\to+\infty} \frac{\sigma^2}{n} = 0$$

so $M_n \xrightarrow{L_2} \mu$ □

9.4 Almost sure convergence

Osservazione 264. It's a strong convergence

Definizione 9.4.1. A sequence converges almost surely to a limit distribution X, and we write $X_n \xrightarrow{a.s.} X$ iff $\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1$

Osservazione 265. Difficult to prove because it's not the limit of a probability but the probability of a limit.
The most important associated theorem with a.s. convergence is the following; somewhat similar to the strong/weak law large number.

Teorema 9.4.1 (Kolmogorov theorem)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be iid rvs such as $\mathbb{E}[X_n] = \mu$ is constant/fixed (no assumption on variance here); then it's possible to prove that the partial mean $M_n \xrightarrow{a.s.} \mu$*

Dimostrazione. No proof here, quite complicate. □

9.5 Relationship between convergences

Proposizione 9.5.1 (Properties)**.** *Convergence implications are summarized in the following schema: to be read as "if $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p}$ to the same X":*

$$\begin{array}{ccccc} \xrightarrow{L_k} & \underset{k>s}{\Longrightarrow} & \xrightarrow{L_s} & & \\ & & \Downarrow & & \\ \xrightarrow{a.s.} & \Longrightarrow & \xrightarrow{p} & \Longrightarrow & \xrightarrow{d} \end{array}$$

Finally, there's only a special case of double implication between $\xrightarrow{p}$ and $\xrightarrow{d}$:

$$\xrightarrow{p} \delta_\theta \iff \xrightarrow{d} \delta_\theta$$

Teorema 9.5.2 (Continuous mapping theorem)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be rvs with some domain D_{X_n}. If g is a* continuous *function on the same domain D_{X_n}, the follow applies:*

$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X) \tag{9.8}$$

$$X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X) \tag{9.9}$$

$$X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X) \tag{9.10}$$

$$\begin{cases} X_n \xrightarrow{L_k} X \\ g \text{ is linear} \end{cases} \implies g(X_n) \xrightarrow{L_k} g(X) \tag{9.11}$$

Osservazione 266. For the last case: if g is quadratic, logarithm, exponential and so on, being not a linear function, then the implication convergence doesn't hold.

Proposizione 9.5.3 (Further properties)**.** *We have that*

1. *for convergence in probability*

$$(X_n \xrightarrow{p} X \wedge Y_n \xrightarrow{p} Y) \implies aX_n + bY_n \xrightarrow{p} aX + bY \tag{9.12}$$

$$(X_n \xrightarrow{p} X \wedge Y_n \xrightarrow{p} Y) \implies X_n \cdot Y_n \xrightarrow{p} X \cdot Y \tag{9.13}$$

2. *same as above applies for* $\xrightarrow{a.s.}$

3. *for* $\xrightarrow{L_k}$ *we only have*

$$(X_n \xrightarrow{L_k} X \wedge Y_n \xrightarrow{L_k} Y) \implies aX_n + bY_b \xrightarrow{L_k} aX + bY \tag{9.14}$$

$$\tag{9.15}$$

but the product does not hold

4. *for* $\xrightarrow{d}$ *we have Slutsky theorem:*

$$(X_n \xrightarrow{d} X \wedge Y_n \xrightarrow{d} \delta_c) \implies \begin{cases} X_n + Y_n \xrightarrow{d} X + c \\ X_n \cdot Y_n \xrightarrow{d} cX \end{cases} \tag{9.16}$$

9.6 Convergence exercises

Esempio 9.6.1. Let be $X_n \sim \text{Pois}(\lambda)$ a sequence of iid rvs; study the convergence of $Z_n = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+X_i}$.
Let's define a continuous transformation of X_i that is $Y_i = \frac{1}{1+X_i}$ and so $Z_n = \frac{\sum_i Y_i}{n}$ is like a partial mean (we have many theorem associated to partial mean: weak/strong laws of large numbers and Kolmogorov theorem). Note that if $X_1, \dots X_n$ are iid then also $Y_1, \dots, Y_n$ are iid as well (the trasformation applied is the same and when we transform independent rv the independence is preserved, unless we combine different rvs).
If we can prove almost sure convergence then we have also the other one so it's convenient to start from the strongest, in case.
So according to Kolmogorov $M_n \xrightarrow{a.s.} \mu$ where in our case $\mu = \mathbb{E}[Y_i]$. Now let's see what is μ:

$$\begin{aligned}
\mu = \mathbb{E}\left[\frac{1}{1+X_i}\right] &= \sum_{D_X} \frac{1}{1+x_i}\mathbb{P}(X_i = x_i) = \sum_{x=0}^{+\infty} \frac{1}{1+x}\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{+\infty} \frac{1}{(x+1)!} e^{-\lambda}\lambda^x \cdot \frac{\lambda}{\lambda} = \frac{1}{\lambda}\sum_{x=0}^{+\infty} \frac{1}{(x+1)!} e^{-\lambda}\lambda^{x+1} \overset{(1)}{=} \frac{1}{\lambda}\underbrace{\sum_{t=1}^{+\infty} \underbrace{\frac{1}{t!}e^{-\lambda}\lambda^t}_{\text{Pois}(\lambda)}}_{1-\left(\frac{1}{0!}e^{-\lambda}\lambda^0\right)} \\
&= \frac{1}{\lambda}\left(1 - e^{-\lambda}\right)
\end{aligned}$$

where in (1) we made substitution $t = x + 1$ and considered that the sum is a Poisson without the probability for $t = 0$, starting the sum from 1). Therefore

$$Z_n \xrightarrow{a.s.} \mu = \frac{1}{\lambda}(1 - e^{-\lambda})$$

and then

$$Z_n \xrightarrow{a.s.} \delta_\mu \implies Z_n \xrightarrow{p} \delta_\mu \implies Z_n \xrightarrow{d} \delta_\mu$$

We can stop here since we proved all the convergences; if one can a strong type it's perfect.

Osservazione importante 58. We don't need here to study L_k convergence since we already have a strong kind of convergence; it's enough to prove one of them. (We could try but it's not easy in the previous case).

Esempio 9.6.2. Study the convergence of $Y_n = (X_1 \cdot \ldots \cdot X_n)^{1/n}$ where $X_i \sim \text{Unif}(0,1)$ are iid rvs.
We need to think about a possible trick and it's given by the continuous mapping theorem which states that we can mantain convergence if we apply some continuous transformation (except for convergence in mean of order k, where g have to be both continuous and linear).
The transformation we should apply here is the logarithm because we have products and logarithm of a product is a sum.
Therefore consider the transformation $\log Y_n = \frac{1}{n}\sum_{i=1}^n \log X_i$; again we notice this is a partial mean and therefore could think of the strongest theorem we have, which is Kolmogorov; then we can say $M_n \xrightarrow{a.s.} \mu$, and as before we have to find $\mu = \mathbb{E}[\log X]$ wher $X \sim \text{Unif}(0,1)$. Therefore:

$$\mu = \mathbb{E}[\log X] = \int_0^1 \log x \cdot 1 \, \mathrm{d}x \overset{(1)}{=} [x \log x - x]_0^1 = -1$$

where in (1) we did it by parts i guess. Therefore

$$\frac{1}{n}\sum_{i=1}^n \log X_i \xrightarrow{a.s.} -1$$

So by applying the continuous mapping theorem (we apply the inverse of the logarithm which is the exponential to both the sides of the convergence)

$$Y_n \xrightarrow{a.s.} e^{-1} = \frac{1}{e} \implies Y_n \xrightarrow{a.s,p,d} \delta_{\frac{1}{e}}$$

Esempio 9.6.3. Let $X_1, \ldots, X_n \sim \text{Exp}(1)$. Find the distribution of $X_{(1)} = \min(X_1, \ldots, X_n)$ and study its convergence.
Remembering that $F_{(1)} = 1 - [1 - F_X(x)]^n$ and being X exponential we have $F_X(x) = 1 - e^{-x}$, therefore:

$$F_{(1)}(x) = 1 - \left[1 - 1 + e^{-x}\right]^n = 1 - e^{-xn}$$

which is the pdf of $\text{Exp}(n)$. So even the minimum is distributed according to an exponential but of parameter n, which are the number of rvs we consider; that is $X_{(1)} \sim \text{Exp}(n)$.
What is its asymptotic distribution of $X_{(1)}$? We study the limit of the distribution of the minimum; that is *in this case we study convergence in distribution* (using the cumulative distribution function, not the mgf or the characteristic function). In this exercise, since we have the pdf of the minimum, it's convenient

for us to try to study the limit of it, and if we find that the limit is a certain pdf we have the solution (finding which random variable gives that pdf).
So let's study the limit:

$$\lim_{n\to\infty} F_{(1)}(x) = \lim_{n\to\infty} 1 - e^{-xn} = 1$$

At the same time 1 is equal to e^0 which is the cumulative distribution function of a δ_0 in 0: $e^0 = F_{\delta_0}(x)$.
. Therefore the minumum converges in distribution to a Dirac in 0 but this also implies that it converge in probability:

TODO: non chiarissimo, la cumulata dovrebbe essere una step function non una costante, poi ok che da 0 in poi sia a 1

$$X_{(1)} \xrightarrow{d} \delta_0 \implies X_{(1)} \xrightarrow{p} \delta_0$$

Now we could study a strong kind of convergence; in this case it's convenient to try studying the L_2 convergence, since we know the limiting distribution (the constant 0), so the expectation should be simpler. Furthermore the limit should be the same: if I know that it converges in distribution to a point, if it converges also in quadratic mean, then it should be at the same point (given the implication schema), it can't be another point.

$$\mathbb{E}\left[(X_{(1)} - 0)^2\right] = \underbrace{\mathbb{E}\left[X_{(1)}^2\right]}_{\text{second moment of Exp}(n)} = \underbrace{\frac{1}{n^2}}_{\text{variance}} + \underbrace{\frac{1}{n^2}}_{\text{second moment squared}} = \frac{2}{n^2}$$

Finally for the convergence in quadratic mean we should study the limit and check that it goes to 0. So:

$$\lim_{n\to+\infty} \frac{2}{n^2} = 0$$

Therefore we can conclude that

$$X_{(1)} \xrightarrow{L_2} \delta_0 \implies X_{(1)} \xrightarrow{L_1} \delta_0$$

9.7 Delta method

Osservazione 267. This is a very useful tool for inference.

Osservazione importante 59 (Motivation). From now on we think of this sequence X_n of random variable as an estimator for a parameter θ of interest; most of time n is the sample size. Imagine that you know that your estimator converges in distribution, as sample goes larger, to the constant θ

$$\{X_n\}_{n\in\mathbb{N}} \xrightarrow{d} \delta_\theta$$

So we can use our estimator to estimate θ. However, point is that we are interested not on θ but on a transformation on the parameter $g(\theta)$, with g continuous.
Using the continuous mapping theorem is not alway simple. Let's see a more detailed example of this.

Esempio 9.7.1 (odds)**.** Let $X_1, \dots, X_n \sim \text{Bern}(p)$ be independent, with $\mathbb{E}[X_i] = p$. Consider $Y_n = \frac{\sum_i X_i}{n} = \bar{X}$. We know that, respectively by weak law of large number and by central limit theorem (it's a sum, not standardized) that:

$$Y_n \xrightarrow{p} p$$
$$Y_n \xrightarrow[\text{by CLT}]{d} \text{N}\left(p, \frac{p(1-p)}{n}\right)$$

First, they should have the same limit, but above *seems* different: no they aren't. By the clt we have a distribution but if $n \to \infty$ the variance of the gaussian goes to 0 and the distribution converges to a Dirac like the first one. In other terms these two results above are asymptotically equivalent (they are the same limit) since $\lim_{n\to\infty} \frac{p(1-p)}{n} = 0$.
The second result however is more useful to know, it's better for us to have a distribution rather than a point: according to gaussian distribution, we can construct intervals, we can test hypotheses, so we can use the idea that we have a distribution for this kind of things, very important from the inferential pov.

Now, *why we need the delta method*? What can we say about the odds? These are

$$g(p) = \frac{p}{1-p}$$

We know that thanks to the continuous mapping theorem, the transformation of the sequence converges to the transformation of the limit distribution:

$$Y_n \xrightarrow{p} p \implies \begin{cases} g(Y_n) \xrightarrow{p} g(p) \\ odds \xrightarrow{p} \frac{p}{1-p} \\ \frac{\bar{x}}{1-\bar{x}} \xrightarrow{p} \frac{p}{1-p} \end{cases}$$

However these are point results, while we may be interested in constructing intervals/hypothesis testing and all that shit that need a proper distribution, not a point.
Therefore here comes the delta method, which will give a second result.

Osservazione 268. To define the delta method first we need the generalized version of CLT.

Teorema 9.7.1 (Generalized version of the central limit theorem)**.** *If we have that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} Y$ converges to a limit distribution Y, then we also have the following equivalent facts (si riporta anche il primo) with $Z \sim \text{N}(0,1)$*

$$\begin{cases} \sqrt{n}(Y_n - \theta) \xrightarrow{d} Y \\ \sqrt{n}(Y_n - \theta) \xrightarrow{d} \sigma Z \\ \frac{Y_n - \theta}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \text{N}(0,1) \\ Y_n \xrightarrow{d} Y \sim \text{N}\left(\theta, \sigma^2/n\right) \end{cases}$$

we can say that a standardized random variable converges Z where $Z \sim \text{N}(0,1)$ by writing it according to the first or the second expression. If one write according to first or second expression, one is using the so called generalized version of the central limit theorem. This is for notation.

Osservazione 269. Coming back to our example we have that $Y_n \xrightarrow{d} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$; then we can rewrite sequentially this according to what we've seen above:

$$
\begin{aligned}
&Y_n \xrightarrow{d} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) && \text{centering } \dots \\
&Y_n - p \xrightarrow{d} \mathrm{N}\left(0, \frac{p(1-p)}{n}\right) && \text{multiply both by } \sqrt{n} \dots \\
&\sqrt{n}(Y_n - p) \xrightarrow{d} \mathrm{N}\left(0, p(1-p)\right) && (1) \\
&\sqrt{n}(Y_n - p) \xrightarrow{d} Z \cdot \sqrt{p(1-p)} && (2)
\end{aligned}
$$

where in (1) and (2) remember that $c\,\mathrm{N}(0, b) = \mathrm{N}\left(0, bc^2\right)$ by the property of the standard gaussian and, again, $Z \sim \mathrm{N}(0, 1)$.
This is another example where starting from a gaussian i can rewrite it in a generalized form. Last one is the generalized-version-of-CLT style.
Now we arrive at the delta method?

Osservazione 270. What is the converging/limit distribution of $\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d}$?
We needed generalized clt because we need the result for a transformation of Y.

Proposizione 9.7.2 (Delta method)**.** *We have that*

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot Y \tag{9.17}$$

this because

$$\sqrt{n}(Y - \theta) \xrightarrow{d} Y$$

Osservazione importante 60. Delta method is a method to derive the limit distribution of a transformation starting from the limit distribution of the orginal variable; the convergency is a convergency in distribution/law and it basically says that if the generalized clt, you have as result the same limit Y multiplied by the derivative of the transformation.

Osservazione importante 61 (Motivation recap 2 (not dirac, general X))*.* Immagine we have a sequence which converges to a random variable X

$$\{X_n\}_{n \in \mathbb{N}} \xrightarrow{d} X$$

But are interested on $g(X_n)$ with g continuous (eg the odds). The question is what is the limit distribution of $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d}$?
This is the eneralized version of CLT

Esempio 9.7.2 (Example refresh)**.** We have a sequence $X_1, \dots, X_n \sim \mathrm{Bern}(p)$ of iid rvs with $\mathbb{E}[X_i] = p$. Consider the so called partial mean $Y_n = \frac{\sum_i X_i}{n} = \bar{X}$. We know that

$$
\begin{aligned}
&Y_n \xrightarrow{p} p \\
&Y_n \xrightarrow[byCLT]{d} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)
\end{aligned}
$$

these two above are asymptotically equivalent since $\lim_{n\to\infty} \frac{p(1-p)}{n} = 0$.
Howevere having a distribution here is better because hypothesis testing and confidence intervals.
Now, what can we say about the odds? We're interested

$$g(p) = \frac{p}{1-p}$$

If we know that Y_n is a good estimator for p, what is a good estimator for the odds instead of p? What is the distribution for the estimator of the odds? We know these results:

- thanks to the continuous mapping theorem

  $$Y_n \xrightarrow{p} p \implies \frac{\bar{x}}{1-\bar{x}} \xrightarrow{p} \frac{p}{1-p}$$

  problem here is that we don't have a distribution here but a point.

- the delta method is a tool that gives us a distribution for the odds. Given the generalized version of clt, starting from

  $$Y_n \xrightarrow{d} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) \implies \sqrt{n}(Y_n - p) \xrightarrow{d} \sqrt{p(1-p)}\,\mathrm{N}(0,1)$$

  we obtain the last after a few passage. BTW in general terms the generalized CLT can be written

  $$\sqrt{n}(Y_n - \theta) \xrightarrow{d} Y$$

  Imagine that we know that for a rv the above holds. Now the question: is what is the distribution of the transformation of the result above

  $$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} ?$$

  Last time we concluded with the formula of delta method

Delta method proof. To answer consider Taylor expansion of the first order of $g(Y_n)$ at the point θ. It's sufficient to stop at first derivative:

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \dots$$

therefore

$$g(Y_n) - g(\theta) \simeq g'(\theta)(Y_n - \theta)$$

so multiplying by $\sqrt{n}$

$$\sqrt{n}(g(Y_n) - g(\theta)) \simeq g'(\theta) \underbrace{\sqrt{n}(Y_n - \theta)}_{\xrightarrow{d} Y}$$

Given the generalized version of the CLT the last part converges to Y so we have the final formula of the delta method which is

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot Y$$

□

Esempio 9.7.3. Coming back to the odds example, apply the formula we have to find the first derivative of the transformation

$$\begin{aligned} g'(p) &= \frac{\partial}{\partial p}\frac{p}{1-p} = \frac{\partial}{\partial p} p(1-p)^{-1} = p(1-p)^{-2} + (1-p)^{-1} \\ &= \frac{p+(1-p)}{(1-p)^2} = \frac{1}{(1-p)^2} \end{aligned}$$

Now we can find the estimator for the odds and also its asymptotic distribution. Now with $\bar{x}$ as our estimator for the percentage, we can say that

$$\sqrt{n}(\bar{x} - p) \xrightarrow{d} \mathrm{N}\left(0, p(1-p)\right)$$

and according to the delta method we can say that

$$\begin{aligned} \sqrt{n}(g(Y_n) - g(\theta)) &\xrightarrow{d} g'(\theta) \cdot Y \\ \sqrt{n}\left(\frac{\bar{x}}{1-\bar{x}} - \frac{p}{1-p}\right) &\xrightarrow{d} \frac{1}{(1-p)^2} \cdot \mathrm{N}\left(0, p(1-p)\right) \\ &\xrightarrow{d} \mathrm{N}\left(0, \frac{p(1-p)}{(1-p)^4}\right) \\ &\xrightarrow{d} \mathrm{N}\left(0, \frac{p}{(1-p)^3}\right) \end{aligned}$$

Esempio 9.7.4. Having $X_1, \ldots, X_n$ are iid with dist $f(x)$ (whatever dist), $\mathbb{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$ if we take the average $Y_n = \bar{X} = \sum_i X_i$ as our estimator, with the clt we have the

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} \sigma\,\mathrm{N}\left(0, 1\right)$$

Now taking the transformation $g(\mu) = \log(\mu)$, we ask: what is an estimator/distribution for the logarithm of μ?
Applying the delta method:

$$\sqrt{n}(g(\bar{x}) - g(\mu)) \xrightarrow{d} g'(\mu) \cdot \sigma \cdot \mathrm{N}\left(0, 1\right)$$

Calculating the derivative of the transformation:

$$g'(\mu) = \frac{\partial}{\partial \mu} \log \mu = \frac{1}{\mu}$$

So:

$$\begin{aligned} \sqrt{n}(\log(\bar{x}) - \log \mu) &\xrightarrow{d} \frac{1}{\mu} \cdot \sigma \cdot \mathrm{N}\left(0, 1\right) \\ &\xrightarrow{d} \mathrm{N}\left(0, \frac{\sigma^2}{\mu^2}\right) \end{aligned}$$

OR better in explicit way

$$\log \bar{x} \xrightarrow{d} \mathrm{N}\left(\log \mu, \frac{\sigma^2}{n\mu^2}\right)$$

Esempio 9.7.5. Let $X_1, \dots, X_n$ iid, with $X_i \sim f_X(x)$, $\mathbb{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$. Find the asymptotic distribution of the second moment $\bar{X}^2$.
We know $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} \sigma \,\mathrm{N}(0,1)$ by CLT. According to Delta method

$$\sqrt{n}(\bar{x}^2 - \mu^2) \xrightarrow{d} g'(\mu)\sigma \,\mathrm{N}(0,1)$$

with

$$g'(\mu) = \frac{\partial}{\partial \mu}\mu^2 = 2\mu$$

then we conclude that

$$\begin{aligned}\sqrt{n}(\bar{x}^2 - \mu^2) &\xrightarrow{d} 2\mu\sigma \,\mathrm{N}(0,1) \\ &\xrightarrow{d} \mathrm{N}\left(0, 4\mu^2\sigma^2\right)\end{aligned}$$

Osservazione 271. Esercizi delta method sul casella berger e controlla le soluzioni.

9.8 Rigo stuff

Osservazione importante 62. We are given a sequence $X_1, ldots, X_n$ of real random variables and a further real random variable X, and we are interested in checking wether or not X_n converges to X as n goes to $+\infty$, written $X_n \to X$. All the standard calculus limits involved in the sequel are meant for $n \to +\infty$. We have 4 types/modes of convergence. In each case as n become larger, X_n get "closer" to X; but the way this happen is different so one convergence does not necessarely imply others (we will see relationship between them in the following. Let's start from definitions

Definizione 9.8.1 (Type of convergences)**.** We have:

1. **almost sure convergence**: X_n converge almost surely to X and we write $X_n \xrightarrow{a.s.} X$ if and only if

   $$\mathbb{P}(\omega \in \Omega : X_n(\omega) \to X(\omega)) = 1$$

   Interpretation: if we choose/fix omega, then $X_n(\omega)$ is a sequence of real number (not random variables) that can converge to the real number $X(\omega)$ as in standard calculus. If this is going to happen for all the elements of Ω then we met the condition.

2. L_p **convergence**: X_n converges to X in L_p, written $X_n \xrightarrow{L_p} X$ and with $p > 0$, if and only if:

   (a) all the involved X_n has $\mathbb{E}[|X_n|^p]$ has moment of order p;
   (b) X has $\mathbb{E}[|X|^p]$ as well;
   (c) and most importantly

   $$\mathbb{E}[|X_n - X|^p] \to 0$$

   Here, again, above is a simple/standard limit for calculus with $n \to +\infty$.

3. **convergence in probability** X_n converges to X in probability, written $X_n \xrightarrow{p} X$, if and only if

$$\lim_{n\to+\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0, \qquad \forall \varepsilon > 0$$

4. **convergence in distribution** X_n converges to X in distribution, written $X_n \xrightarrow{d} X$, if and only if

$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \qquad \forall x \in \mathbb{R} : F_X \text{ is continuous in } x$$

Intuitively it would be more natural to require the convergence to hold on all the domain (not only where F_X is continuous) but this would be a too much severe requirement, as we will see in the following.

TODO: audio arrivato a 26:11 di 9/10/23

Osservazione importante 63. Some important Rigo's remarks on converges implications graph:

1. if $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$ then, $\mathbb{P}(X = Y) = 1$ (they are almost surely equal). So the limit in probability is unique (provided it exists).
   A nice consequence is the following: suppose that we have proved $X_n \xrightarrow{p} X$ and we aim to prove X_n converges to some limit in L_p or a.s. (a stronger type). In order to prove that, the only possible limit is X: suppose in fact that $X_n \xrightarrow{a.s.} Y$ then, we have $X_n \xrightarrow{p} Y$, and by the previous result, we have that $X = Y$ almost surely.

TODO: TODOHERE 30:00

2. as the above picture illustrates $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$ but the converse is not true. However there is an important special case where

$$X_n \xrightarrow{d} X \implies X_n \xrightarrow{p} X$$

and this is X is degenerate. Hence if $X = a$ almost surely (is degenerate) we obtain $X_n \xrightarrow{p} X \iff X_n \xrightarrow{d} X$

TODO: ...

3. the definition we gave $X_n \xrightarrow{d} X \iff \lim F_n(x) = F(x)$ is true for may appear strange. It may seem more natural require convergence for all x that is:

$$\lim_{n\to+\infty} F_{X_n}(x) = F(x), \qquad \forall x \in \mathbb{R}$$

But this second alternative definition is too strong. To understand why, suppose we have both degenerate $X_n = \frac{1}{n}$ and $X = 0$. Here

$$F_{X_n}(x) = \begin{cases} 1 & \text{if } x \geq \frac{1}{n} \\ 0 & \text{if } x < \frac{1}{n} \end{cases}, \quad F_X(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

So the va value of F at 0 is 1, while for F_{X_n} is 0. Therefore $\lim_{n\to+\infty} F_{X_n}(0) \neq F_X(0)$. Thus if we would require $\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \forall x \in \mathbb{R}$ we would get the disturbing consequence that $X_n = \frac{1}{n}$ does not converge in distribution to $X = 0$ ($X_n = \frac{1}{n} \not\to X = 0$) and this is a consequence we don't like.

Osservazione 272. Now some counterexamples to show that som double implications doesnt' work (as stated in the graph of convergence implications).

Esempio 9.8.1. Let $\mathbb{P}(X_n = 0) = fracn - 1n$ and $P(X_n = n) = 1/n$ and $X = 0$. Then $X_n convp X$ but $X_n \overset{L_1}{\not\longrightarrow} X$ here convergence in probability but not in L_p. Its an contraexample of why the implication does not hold:

- given $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}(|X_n - X| > \varepsilon) &\overset{(1)}{=} \mathbb{P}(|X_n|) = \mathbb{P}(|X_n| > \varepsilon, X_n = 0) + \mathbb{P}(|X_m| > \varepsilon, X_n = n) \\ &\leq 0 + \mathbb{P}(X_n = n) = \frac{1}{n}\end{aligned}$$

  where in (1) X_n by assumption takes 2 values, 0 and n. Hence since $\frac{1}{n} \to 0$ we can state $X_n \xrightarrow{p} X$

- however

$$\begin{aligned}\mathbb{E}[|X_n - X|] = \mathbb{E}[X_n] &= |0|\,\mathbb{P}(X_n = 0) + |n|\,\mathbb{P}(X_n = n) \\ &= 0 + |n|\,\mathbb{P}(X_n = n) = n\frac{1}{n} = 1, \quad \forall n\end{aligned}$$

  Hence $X_n \overset{L_p}{\not\longrightarrow} X$

Capitolo 10

Simulation

10.1 Sampling values from rvs

Osservazione importante 64. This is important for pratical/inferential reasons: some methods in statistics need sampling from rvs to get estimates, and not always distribution are available/easy to use (eg complicate, not popular, not well known, or because we don't know completely the analytical stuff eg we know the kernel not the normalization constant).

Osservazione importante 65. So it's important in difficult situation to have a method for sampling from the distribution (to have some values), because if we draw infinite time we can obtain the distribution.

Osservazione importante 66. The methods available are summarized in table 10.1. Viroli will do univariate methods. The MCMC stuff (Gibbs sampling, Metropolis-Hasting) will be done in Bayesian statistics.

10.1.1 Inversion method

Osservazione 273. This is the simpler method

Osservazione importante 67. If $X \sim f_X(x)$ whatever f, then its $F_X(x) = U$ can be thought as a new random variable $U \sim \text{Unif}(0,1)$ (this result is called *probability integral transform*).

Definizione 10.1.1 (Inversion method)**.** If our aim is to draw values from X, a solution with a two step procedure is as follows:

1. draw different values $u_1, \dots, u_n$ from $\text{Unif}(0,1)$;
2. compute $F_X^{-1}(u_1), \dots, F_X^{-1}(u_n)$ obtaining $x_1, \dots, x_n \sim f_X(x)$

Univariate	Multivariate
Inversion	Gibbs sampling
Accept-reject	Metropolis-Hasting
Sampling and resampling	...

Tabella 10.1: Sampling methods

Therefore this method requires we know F (and obtain its inverse).

Esempio 10.1.1. Let $X \sim \text{Exp}(\lambda)$, with known $F_X(x) = 1 - e^{-\lambda x} = u$; but imagine we are not able to draw from the exponential distribution. Knowing F we can obtain its inverse and apply inversion method. For the inverse:

$$\begin{aligned} 1 - u &= e^{-\lambda x} \\ \log(1-u) &= -\lambda x \\ -\frac{1}{\lambda}\log(1-u) &= x \end{aligned}$$

Following the algorithm:

1. we generate $u_1, \ldots, u_n$ from $\text{Unif}(0,1)$
2. we calculate $x_1 = -\frac{1}{\lambda}\log(1-u_1), \ldots, x_n = -\frac{1}{\lambda}\log(1-u_n)$

Osservazione 274. From a pratical point of view it's not very useful:

1. it's already implemented for common distribution in statistical software: eg `rexp` uses this method;
2. there are very few rvs for which the pdf is known and is invertible.

10.1.2 Accept-reject method

Osservazione importante 68 (Setup). We

- are interested in generating values from $\pi(x)$ which is the target distribution (not a known one eg exp, normal etc). It's known in part, analytically (eg we know at least the kernel concerning x, not necessarily integral-normalizing-to-1 constants), but we are not able to draw values from it.
- we choose $p(x)$, a perfectly-known distribution from which we can draw values.

Osservazione 275. One could invent a distribution by specifying the kernel (a function of x), setup the domain, and deriving the normalization constant by integration like this

$$1 = \int_{D_X} c \cdot (\text{kernel in } x)\, \mathrm{d}x = c \int_{D_X} (\text{kernel in } x)\, \mathrm{d}x = \ldots$$

Immagininig we're not able to solve the integral and don't know or we don't know to generate values from this distribution. Then we can use the accept-reject method.

Definizione 10.1.2 (Accept-reject method)**.** The algorithm is the following:

1. draw a value x from the proposal $p(x)$
2. draw a value u from $\text{Unif}(0,1)$

3. check if
$$u < \frac{\pi(x)}{M \cdot p(x)} \tag{10.1}$$
where for $\pi(x)$ and $p(x)$ we mean densities and M is a positive constant (so overall the right hand ratio is positive). M has to be fixed in advance, such that:
$$\pi(x) < M \cdot p(x), \quad \forall x \in D_X$$
One should check this condition
4. if 10.1 is true then *accept* x, if false then *reject* x
5. you repeat from 1, until you have enough elements for the application's need

Osservazione importante 69. Some remarks:

- accept and reject because the rule specify when keeping our simulation as suitable value or not
- first of all we shoud choose the proposal p. How we should choose p? First we should know something about the target:
  1. know the domain space D_X of the target (eg if one wants positive values or values between $-\infty$ and $+\infty$). *The proposal should respect the domain space.*
  2. if we know that the target is symmetric (or asymmetric), the proposal should be symmetric (respectively asymmetric) as well.
- regarding M: its said that M should guarantee that the ratio
$$\frac{\pi(x)}{M \cdot p(x)}$$
is a value between 0 and 1, since it's compared with a draw from Unif $(0, 1)$; so we shoud choose M high enough. So from here the condition to be checked
$$\pi(x) < M \cdot p(x), \quad \forall x \in D_X$$
If this condition is not satisfied, the method doesn't work.
Since we don't know what is the target so it's difficult to practically choose M:
  - an *option in practice* is to choose M very large (eg 1000). in this case i'm quite sure the inequality will be respected.
  - but if it's too large we will have a method where acceptance is very rare. So the method could be slow (method has a tradeoff).

Proof of accept-reject. We aim is to prove that what we generate is a realization of the distribution of interest, and in math terms that the density f of the value we accept x (conditioned to being accepted) is equal to the target
$$f\left(x \middle| u < \frac{\pi(x)}{Mp(x)}\right) = \pi(x)$$

In order to prove that we start by writing/expanding the conditional density, which is the ratio between joint density and at denominator the probability of conditioning. Thus by definition we have:

$$f\left(x\middle|u<\frac{\pi(x)}{Mp(x)}\right)=\frac{\mathbb{P}\left(x\cap u<\frac{\pi(x)}{Mp(x)}\right)}{\mathbb{P}\left(u<\frac{\pi(x)}{Mp(x)}\right)}$$

Now, given that the intersection can be written twofold (conditioning on the first or the second event)

$$\mathbb{P}\left(A\cap B\right)=\mathbb{P}\left(A|B\right)\mathbb{P}\left(B\right)=\mathbb{P}\left(B|A\right)\mathbb{P}\left(A\right)$$

we can rewrite the numerator, which is a joint density, this way:

$$\begin{aligned}\frac{\mathbb{P}\left(x\cap u<\frac{\pi(x)}{Mp(x)}\right)}{\mathbb{P}\left(u<\frac{\pi(x)}{Mp(x)}\right)}&\overset{(1)}{=}\frac{p(x)\cdot\mathbb{P}\left(u<\frac{\pi(x)}{Mp(x)}|x\right)}{\int_{D_X}p(x)\cdot\mathbb{P}\left(u<\frac{\pi(x)}{Mp(x)}|x\right)\,\mathrm{d}x}\overset{(2)}{=}\frac{p(x)\cdot\frac{\pi(x)}{Mp(x)}}{\int_{D_x}p(x)\frac{\pi(x)}{Mp(x)}\,\mathrm{d}x}\\&\overset{(3)}{=}\frac{\pi(x)}{\underbrace{\int_{D_x}\pi(x)\,\mathrm{d}x}_{=1}}\overset{(4)}{=}\pi(x)\end{aligned}$$

where:

- (1) because we write the denominator as well as (integral of) joint density, given the fact that it's a marginal density so the way to do it is $\int_{D_Y}f(x,y)\,\mathrm{d}y=f(x)$.
  Furthermore informally/put another way (Luca's view) it seems basically the theorem of total probability for $\mathbb{P}\left(u<\frac{\pi(x)}{Mp(x)}\right)$;
- (2) remembering that u is coming from a Unif $(0,1)$ distribution; therefore the probability that a Unif $(0,1)$ is lower than a constant $c\in[0-1]$, is the constant c itself (here our constant is $\frac{\pi(x)}{Mp(x)}$;
- (3) we moved constant M and simplified;
- (4) the denominator is the integral of the target distribution over the domain so it must be 1.

□

Osservazione importante 70. As said it works iff M is carefully chosen to make the ratio $\frac{\pi(x)}{Mp(x)}$ between 0 and 1: if it doesnt, the property of the uniform used at (2) in proof above doesn't work any more.

Acceptance probability A things important for the algorithm to be computed: we have said that M the probability of acceptance of drawn values from the proposal distribution.
Idea is that if you take M too large you will accept few values (we will see in lab), so there is an inverse correlation between these two quantities. But again

it's important that M is large enough to make the ratio is lower than 1.
This is the tradeoff: now we view this tradeoff in math terms. We want to compute the acceptance probability.
Let's call acceptance probability of the algorithm *alpha*; it's defined as

$$\begin{aligned}\alpha = \mathbb{P}\left(\text{accepted}\right) &\overset{(1)}{=} \mathbb{P}\left(U < \frac{\pi(x)}{Mp(x)}\right) \\ &\overset{(2)}{=} \int_{D_x} p(x)\,\mathbb{P}\left(U < \frac{\pi(x)}{Mp(x)} \middle| X = x\right) \mathrm{d}x \\ &= \int_{D_x} p(x)\frac{\pi(x)}{Mp(x)}\,\mathrm{d}x = \frac{1}{M}\underbrace{\int_{D_x} \pi(x)\,\mathrm{d}x}_{=1} \\ &= \frac{1}{M}\end{aligned}$$

where in:

- (1) we wrote capital U (meaning Unif $(0,1)$) since it's not a single extraction but a random variable that originate a probability
- (2) we rewrite as integral of joint probability (as made for the accept reject method proof), or more explicitly here

$$\int_{D_y} f(x,y)\,\mathrm{d}y = \int_{D_y} f(y)f(x|y)\,\mathrm{d}y$$

TODO: to be reported above maybe

So given that $\alpha = \frac{1}{M}$ we have a very precise relation regarding the trade off we talked above.

Osservazione importante 71. Observe: it works even if we don't know fully the target, but we know the target unless a normalization constant.
What about a situation in which the target can be decomposed in a kernel part $k(x)$ times a constant, that is in situations such as $\pi(x) = k(x) \cdot c$ (where we know $k(x)$ but not c)? Eg we want to generate a rv from a distribution with kernel: $\exp(-\log(x)) \cdot c$ (we don't know c).
This doesn't matter since the method works in any case: we repeat the proof with a different perspective.

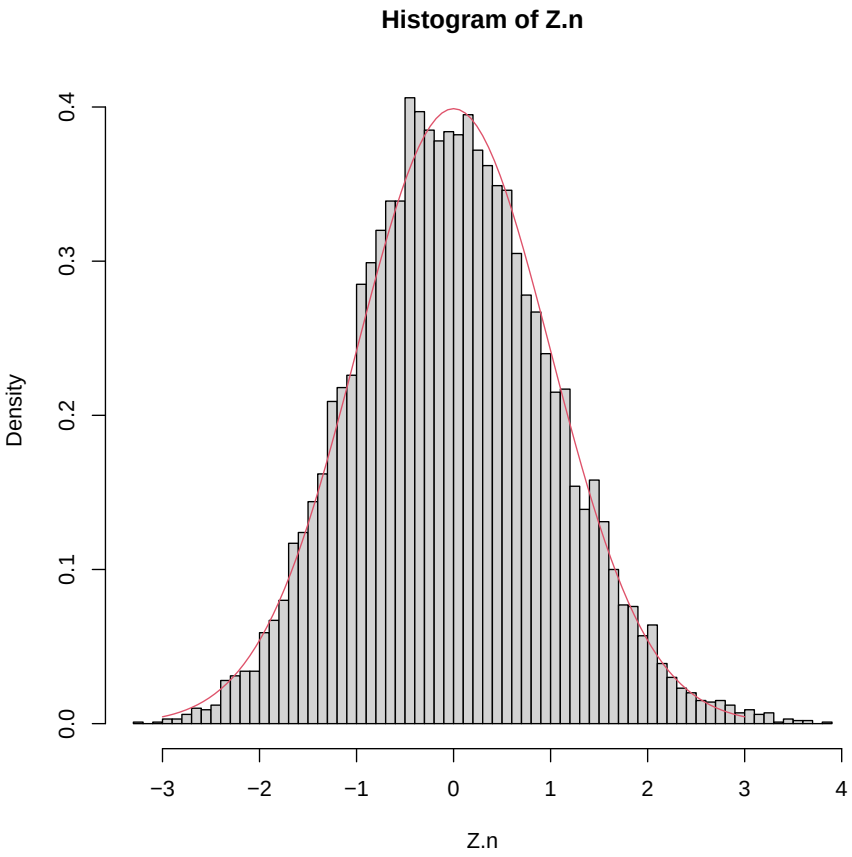
Dimostrazione. In proof the difference is at numerator of where we substituted $k(x) \cdot c$ instead of $\pi(x)$. So we aim is to prove that $f\left(x|u < \frac{k(x)\cdot c}{Mp(x)}\right) = \pi(x)$:

$$\begin{aligned}f\left(x|u < \frac{k(x)\cdot c}{Mp(x)}\right) &= \frac{\mathbb{P}\left(x \cap u < \frac{k(x)\cdot c}{Mp(x)}\right)}{\mathbb{P}\left(u < \frac{k(x)\cdot c}{Mp(x)}\right)} = \frac{p(x)\,\mathbb{P}\left(u < \frac{k(x)\cdot c}{Mp(x)}|x\right)}{\int_{D_x} p(x)\,\mathbb{P}\left(u < \frac{k(x)\cdot c}{Mp(x)}|x\right)\mathrm{d}x} \\ &= \frac{p(x)\frac{k(x)\cdot c}{Mp(x)}}{\int_{D_x} p(x)\frac{k(x)\cdot c}{Mp(x)}\,\mathrm{d}x} = \frac{k(x)\cdot c}{\underbrace{\int_D k(x)\cdot c\,\mathrm{d}x}_{=1}} = k(x)\cdot c \\ &= \pi(x)\end{aligned}$$

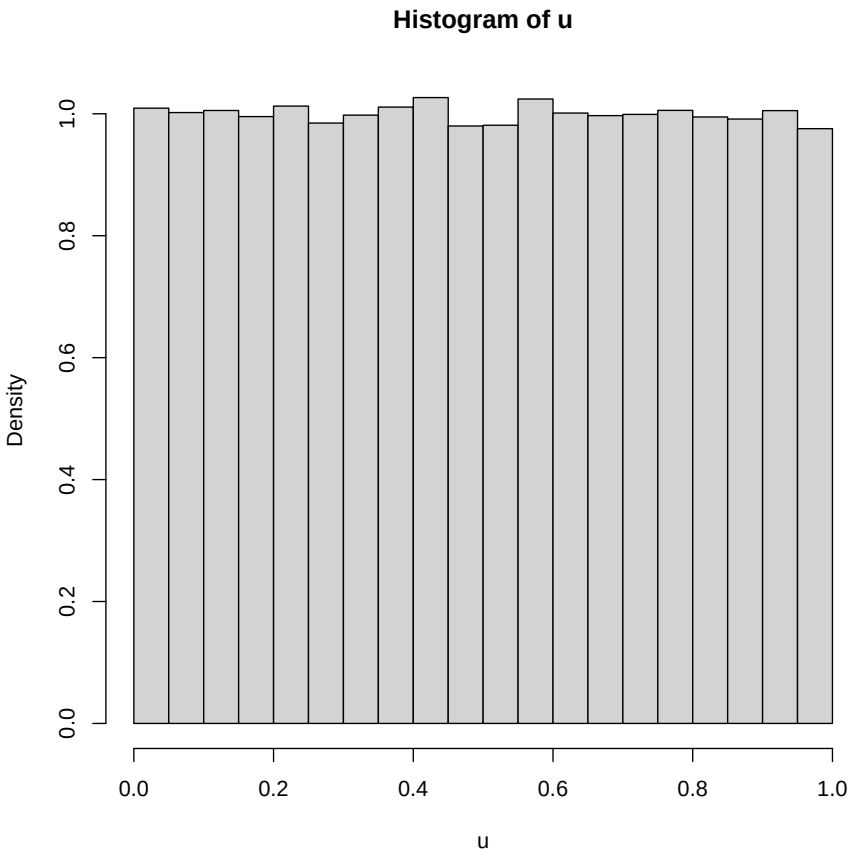
So we proved that we can use the algorithm even without knowing the normalization constant. □

10.2 R exercises

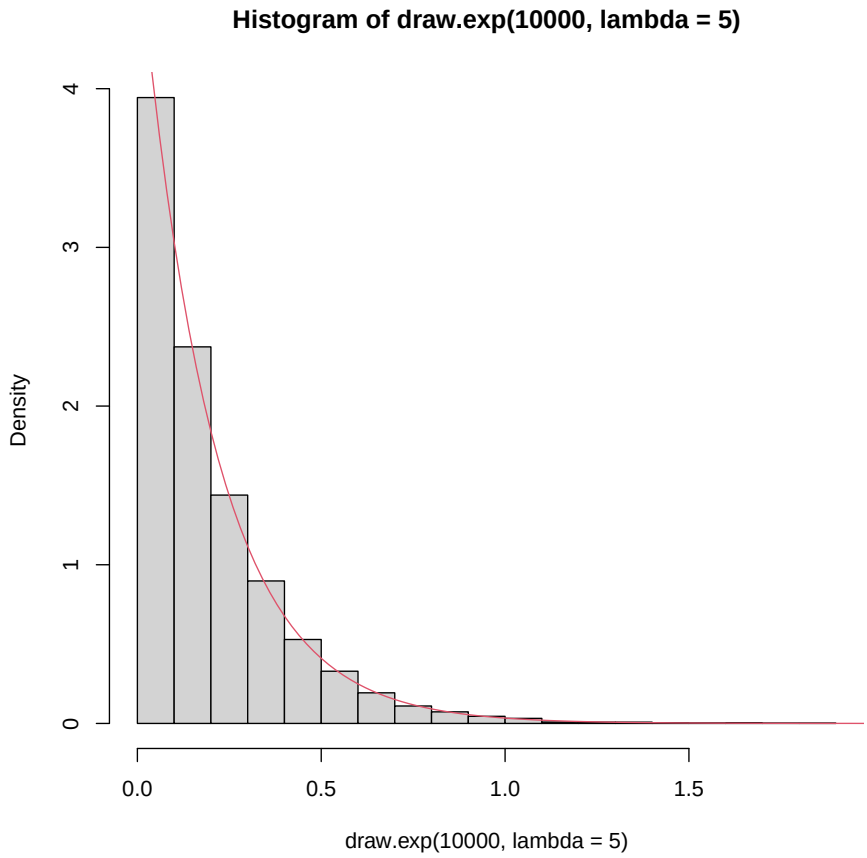
10.2.1 CLT



10.2.2 Inversion method

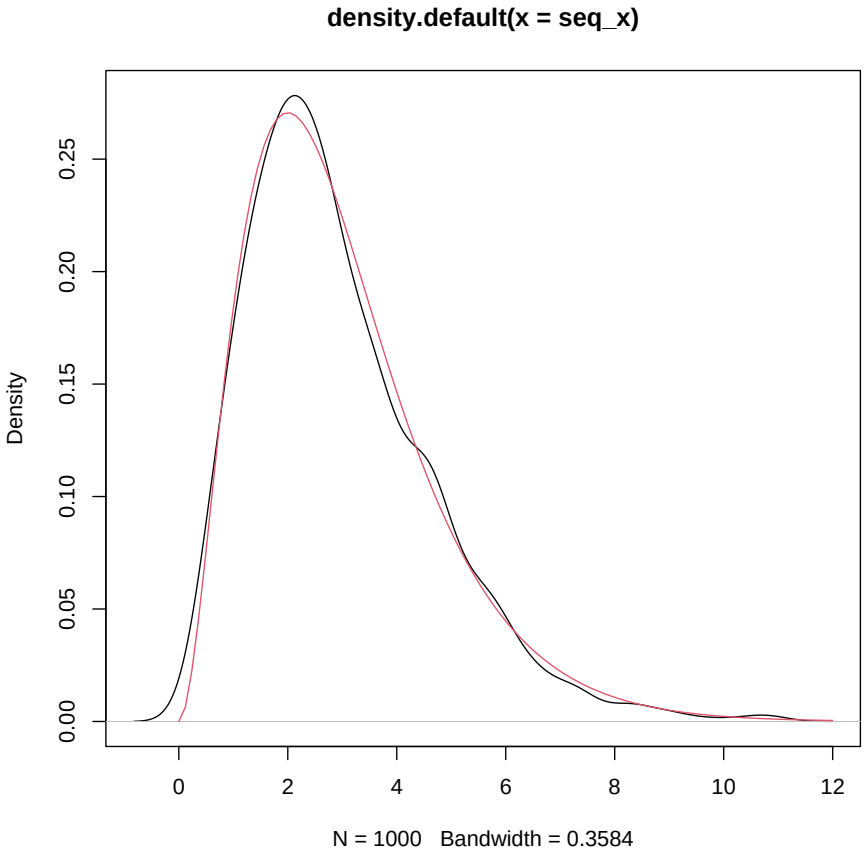


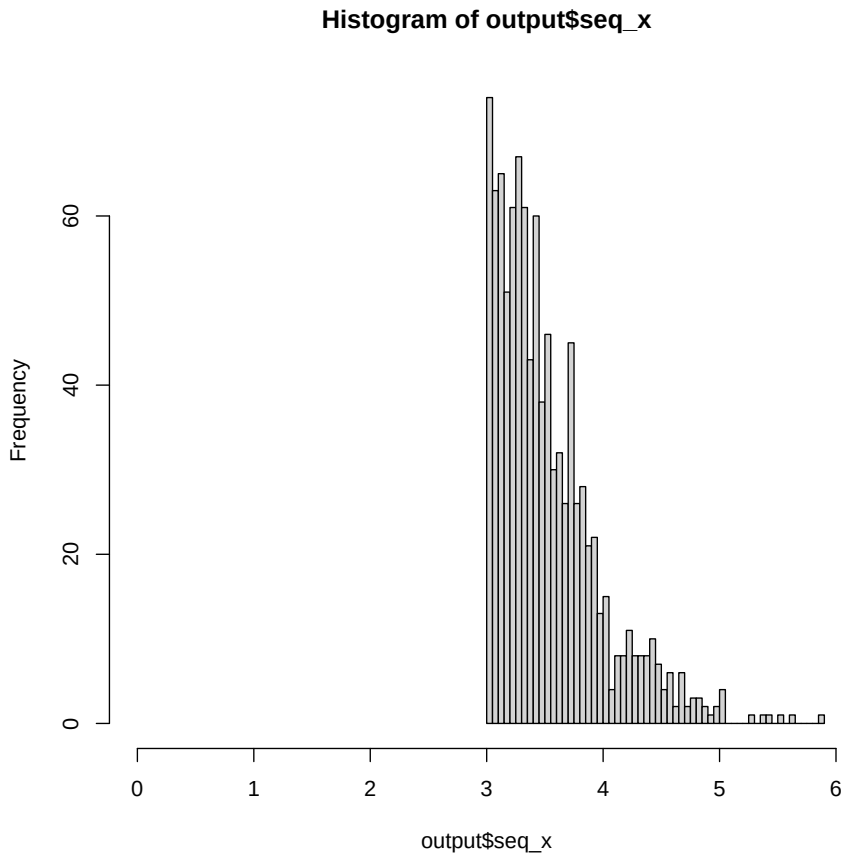

```
##   [1] 0.00769923 0.05389461 0.37726227 0.64083592 0.48585141 0.40095990
##   [7] 0.80671933 0.64703530 0.52924708 0.70472953 0.93310669 0.53174683
##  [13] 0.72222778 0.05559444 0.38916108 0.72412759 0.06889311 0.48225177
##  [19] 0.37576242 0.63033697 0.41235876 0.88651135 0.20557944 0.43905609
##  [25] 0.07339266 0.51374863 0.59624038 0.17368263 0.21577842 0.51044896
##  [31] 0.57314269 0.01199880 0.08399160 0.58794121 0.11558844 0.80911909
##  [37] 0.66383362 0.64683532 0.52784722 0.69493051 0.86451355 0.05159484
##  [43] 0.36116388 0.52814719 0.69703030 0.87921208 0.15448455 0.08139186
##  [49] 0.56974303 0.98820118 0.91740826 0.42185781 0.95300470 0.67103290
##  [55] 0.69723028 0.88061194 0.16428357 0.14998500 0.04989501 0.34926507
##  [61] 0.44485551 0.11398860 0.79792021 0.58544146 0.09809019 0.68663134
##  [67] 0.80641936 0.64493551 0.51454855 0.60183982 0.21287871 0.49015098
##  [73] 0.43105689 0.01739826 0.12178782 0.85251475 0.96760324 0.77322268
##  [79] 0.41255874 0.88791121 0.21537846 0.50764924 0.55354465 0.87481252
##  [85] 0.12368763 0.86581342 0.06069393 0.42485751 0.97400260 0.81801820
##  [91] 0.72612739 0.08289171 0.58024198 0.06169383 0.43185681 0.02299770
##  [97] 0.16098390 0.12688731 0.88821118 0.21747825
```



10.2.3 Accept-reject

```
## [1] 101852
```


density.default(x = seq_x)
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
N = 1000 Bandwidth = 0.3584


Histogram of output$seq_x
Frequency
0
20
40
60
0
1
2
3
4
5
6
output$seq_x