Fundamental concepts

January 19, 2024

Contents

Part I

Probability

Chapter 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)*

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

Proof. 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):*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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,*

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

Proof. 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:*

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

Proof. 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)**.**

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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.

Proof. 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*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof.

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

□

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*

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

Proof. 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

Chapter 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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:*

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

Proof. 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*

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

Proof. 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:*

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

Proof.

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

□

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.

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

Proof.

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

□

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)**.**

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

Proof. 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*

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

Proof. 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)**.**

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

Proof. 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:*

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

Proof. 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)**.**

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

Proof. 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

Chapter 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	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Table 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. 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.

Osservazione 51. On Venn diagrams it's a set of non overlapping shapes that sum up to Ω.

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 $\bigcup_{i=1}^{\infty} E_i \in \mathscr{F}$

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

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

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

$$\tag{3.9}$$

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 σ-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 $\mathbb{R}$ 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

$$\mathbb{P}(A) \geq 0, \quad \forall A \in \mathscr{F} \tag{3.10}$$

$$\mathbb{P}(\Omega) = 1 \tag{3.11}$$

$$A_i \cap A_j = \emptyset, \forall i \neq j \implies \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) \tag{3.12}$$

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.

$$\boxed{\mathbb{P}\left(\overline{A}\right) = 1 - \mathbb{P}(A)} \tag{3.13}$$

Proof.

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

□

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.

$$\boxed{\mathbb{P}(\emptyset) = 0} \tag{3.14}$$

Proof. Setting $A = \Omega$ in 3.13,

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

□

Proposizione 3.2.3.

$$\boxed{A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)} \tag{3.15}$$

Proof. 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)**.**

$$\boxed{\mathbb{P}(A \setminus B) = \mathbb{P}\left(A \cap \overline{B}\right) = \mathbb{P}(A) - \mathbb{P}(A \cap B)} \tag{3.16}$$

Proof. 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)**.**

$$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)} \tag{3.17}$$

Proof. Writing $A \cup B$ as union of two incompatible events,we apply axioms and 3.16:

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

□

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

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n}(-1)^{r+1} \sum_{i_1<\ldots<i_r} \mathbb{P}(E_{i1} \cap E_{i2} \cap \ldots \cap E_{ir}) \tag{3.18}$$

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

$$\ldots + (-1)^{n+1} \mathbb{P}(E_1 \cap \ldots \cap E_n) \tag{3.19}$$

Proof. 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))**.**

$$\mathbb{P}(E_1 \cup E_2 \cup \ldots \cup E_n) \leq \sum_{i=1}^{n} \mathbb{P}(E_i) \tag{3.20}$$

Proof. Done in the following section 5.3.3. □

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

$$\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\overline{E_i}\right) \tag{3.21}$$

Proof. 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

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

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

$$\mathbb{P}(B) = \lim_{i \to +\infty} \mathbb{P}(B_i) \tag{3.23}$$

Proof. 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.

Osservazione 57. 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 3.2.13. *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.*

Proof. 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 $+\infty$ □

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%.

```
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 58. 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

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

Osservazione 59. 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 60. 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 3.2.14 (For two events, $\mathbb{P}(A \cap B)$)**.** *If* $\mathbb{P}(B) \neq 0$*:*

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

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

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B|A)} \tag{3.26}$$

Proof. Algebric manipulation of 3.24. □

Proposizione 3.2.15 (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:*

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

Proof. To verify it we apply recursively the definition 3.26 to the second member:

$$\mathbb{P}(E_1) \cdot \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} \cdot \frac{\mathbb{P}(E_1 \cap E_2 \cap E_3)}{\mathbb{P}(E_1 \cap E_2)} \cdot \ldots \cdot \frac{\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_n)}{\mathbb{P}(E_1 \cap E_2 \cap \ldots \cap E_{n-1})} \tag{3.27}$$

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 3.27 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 61. 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 62 (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:

$$\begin{aligned}\boxed{\mathbb{P}(E)} &= \mathbb{P}\left((E \cap C) \cup (E \cap \overline{C})\right) && (3.28)\\ &= \mathbb{P}(E \cap C) + \mathbb{P}\left(E \cap \overline{C}\right) \\ &= \boxed{\mathbb{P}(C)\,\mathbb{P}(E|C) + \mathbb{P}\left(\overline{C}\right)\mathbb{P}\left(E|\overline{C}\right)} && (3.29)\end{aligned}$$

Osservazione 63 (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 3.2.16 (Law of total probability (general version))**.** *If $C_1, C_2, \dots$ is a* finite *or* countable *partition of Ω, the probability of a generic event E can be written as:*

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

Proof. If $C_1, C_2, \dots, C_n$ 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:

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(C_i \cap E) = \sum_{i=1}^{n} \mathbb{P}(C_i)\,\mathbb{P}(E|C_i) \tag{3.31}$$

and in the last passage we substitued 3.26. □

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

3.2.6 Bayes formula

Teorema 3.2.17 (Bayes formula)**.**

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\cdot\mathbb{P}(B|A)}{\mathbb{P}(B)}} \tag{3.32}$$

Proof. Substitute 3.26 in 3.24. □

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:*

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

Proof. 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 3.2.13), 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:

$$\mathbb{P}(H|T) = \frac{\mathbb{P}(H)\cdot\mathbb{P}(T|H)}{\mathbb{P}(H)\cdot\mathbb{P}(T|H) + \mathbb{P}\left(\overline{H}\right)\cdot\mathbb{P}\left(T|\overline{H}\right)} \tag{3.33}$$

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:

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)} \tag{3.34}$$

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 $\mathbb{P}(B) > 0$:*

$$\mathbb{P}(A|B) = \mathbb{P}(A) \tag{3.35}$$

Proof.

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

□

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

Proof.

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

□

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.*

Proof. 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 $\overline{F}$.*

Proof. 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 $E = (E \cap F) \cup \left(E \cap \overline{F}\right)$ 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:

$$\mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} \mathbb{P}(A_i) \tag{3.36}$$

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

$$\mathrm{OR}(A) = \frac{\mathbb{P}(A)}{\mathbb{P}\left(\overline{A}\right)} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)} \tag{3.37}$$

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:

$$\mathbb{P}(A) = \frac{\mathrm{OR}(A)}{1 + \mathrm{OR}(A)} \tag{3.38}$$

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:

$$\frac{\mathbb{P}(H|E)}{\mathbb{P}\left(\overline{H}|E\right)} = \frac{\dfrac{\mathbb{P}(H)\,\mathbb{P}(E|H)}{\mathbb{P}(E)}}{\dfrac{\mathbb{P}\left(\overline{H}\right)\mathbb{P}\left(E|\overline{H}\right)}{\mathbb{P}(E)}} = \frac{\mathbb{P}(H)}{\mathbb{P}\left(\overline{H}\right)} \cdot \frac{\mathbb{P}(E|H)}{\mathbb{P}\left(E|\overline{H}\right)} \tag{3.39}$$

Osservazione 71. A seguito dell'introduzione di una prova l'originale rapporto a favore $\frac{\mathbb{P}(H)}{\mathbb{P}(\overline{H})}$ 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 3.38 per passare a probabilità), evitando di dover utilizzare il teorema delle probabilità totali:

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

da cui applicando la 3.38 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:*

$$\begin{aligned}
&0 \leq \mathbb{P}(E|F) \leq 1 \\
&\mathbb{P}(\Omega|F) = 1 \\
&\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i|F\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i|F) \quad \textit{se } E_i \cap E_j = \emptyset,\ \forall i \neq j
\end{aligned}$$

Proof. 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

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

□

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.

$$\widetilde{\mathrm{P}}\left(\overline{A}\right) = 1 - \widetilde{\mathrm{P}}(A) \tag{3.40}$$

Proof. 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:*

$$\frac{\widetilde{\mathrm{P}}(A|B)}{\widetilde{\mathrm{P}}\left(\overline{A}|B\right)} = \frac{\mathbb{P}(A|B \cap F)}{\mathbb{P}\left(\overline{A}|B \cap F\right)} = \frac{\mathbb{P}(A|F) \cdot \mathbb{P}(B|A \cap F)}{\mathbb{P}\left(\overline{A}|F\right) \cdot \mathbb{P}\left(B|\overline{A} \cap F\right)} \tag{3.41}$$

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 3.29*

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

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 3.30 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

$$\mathbb{P}(A \cap B|F) = \mathbb{P}(A|F) \cdot \mathbb{P}(B|F) \tag{3.42}$$

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

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

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

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

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:

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

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

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

che equivale ad un normale teorema di Bayes dove al posto delle probabilità a priori secca $\mathbb{P}(D)$ che avevamo utilizzato in esempio 3.4.1, abbiamo sostituito i risultati disponibili alla fine del primo test, ossia $\mathbb{P}(D|T_1) = 0.16$ 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 $\mathbb{P}(D|T_1 \cap T_2)$. 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:

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

Per il secondo punto:

- B= primo lancio $\leq 3 = \{(1,1), \dots, (1,6), (2,1) \dots, (2,6), (3,1) \dots, (3,6)\}$ 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 $\mathbb{P}(C|B) = \frac{\mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \frac{3/36}{18/36} = \frac{1}{6}$
- $\mathbb{P}(B|C) = \frac{3/36}{5/36} = \frac{3}{5}$

Chapter 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 (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, $\Omega = \{everyone\}$,while the random variable X could be height, so if Luca is extracted ($\omega = \text{Luca}$), 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 81. 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 $X \sim \mathrm{N}\left(0, 1\right)$ 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

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

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:*

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

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

Proof. 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 82 (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*

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

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

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

Proof. 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 $-\infty < x_{min}$

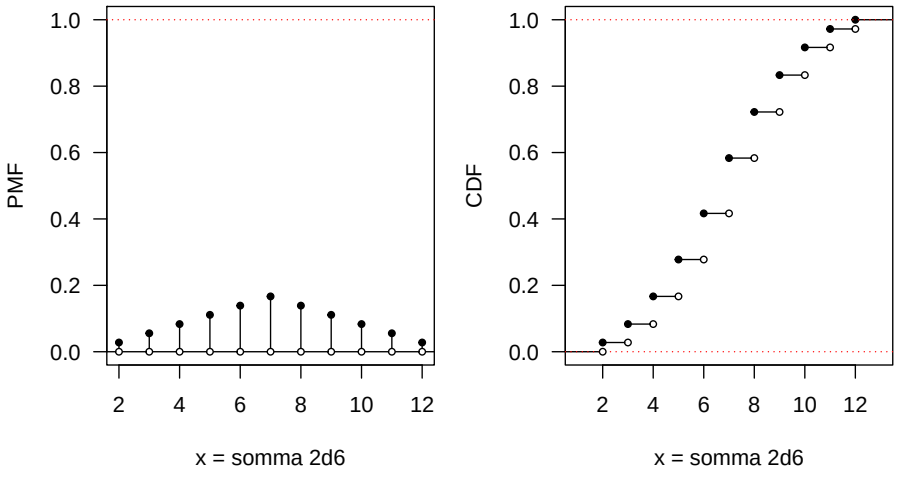


Figure 4.1: Somma del lancio di due d6

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 83. 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*

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

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

Proof. 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 84. 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*

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

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

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

Esempio 4.2.3 (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 ($X \sim \mathrm{Exp}(1)$).

Osservazione 85 (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 86 (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.

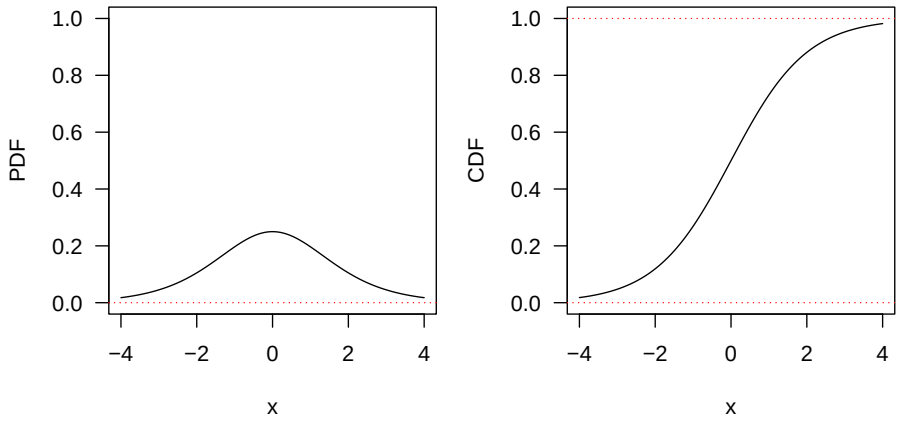


Figure 4.2: Logistic distribution

Osservazione 87 (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 4.2.4 (Logistic rv)**.** Logistic random variable, plotted in figure 4.2, is defined by:

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

```
flogis <- function(x) exp(x)/(1 + exp(x))^2
Flogis <- function(x) exp(x)/(1 + exp(x))
par(mfrow = c(1, 2), mar = c(5,4,1,1))
plot_fun(flogis, from = -4, to = +4, ylim = c(0, 1),
         cartesian_plane = FALSE,
         ylab = 'PDF', las = 1)
abline(h = c(0), col = 'red', lty = 'dotted')
plot_fun(Flogis, from = -4, to = +4, ylim = c(0, 1),
         cartesian_plane = FALSE,
         ylab = 'CDF', las = 1)
abline(h = c(0,1), col = 'red', lty = 'dotted')
```

4.2.3 Distribution functions (Rigo's style)

Osservazione 88. 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 said 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 $F(x) = \mathbb{P}(X \leq x)$, $\forall x \in \mathbb{R}$.

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.

Proof. 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 89. 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*

Proof. We skip it. □

Osservazione 90. 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

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

4.2.3.2 Singular continuous rvs

Osservazione 91. 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 92. 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 93. 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 $\mathbb{P}(X \in A) = 0$*
$$\forall A \subset \beta(\mathbb{R}), m(A) = 0 \implies \mathbb{P}(X \in A) = 0$$

4.3 Other useful rv functions

4.3.1 Support indicator

Osservazione 94. 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 95. 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

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = 1 - F_T(t) \tag{4.23}$$

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

$$H(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{\mathrm{d}t} \log(1 - F_T(t)) = -\frac{d}{\mathrm{d}t} \log(S(t)) \tag{4.24}$$

Osservazione 96. Hazard function can be interpreted as the probability that T stops at t given that it arrived to t

Osservazione 97. 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:

$$\begin{aligned} \log(S(t)) &= -\int_{-\infty}^{x} H(t)\,\mathrm{d}t \\ S(t) &= \exp\left(-\int_{-\infty}^{x} H(t)\,\mathrm{d}t\right) \end{aligned} \tag{4.25}$$

While for what concerns $F_T(t)$ e $f_T(t)$ we have:

$$F_T(t) = 1 - \exp\left(-\int_{-\infty}^{x} H(t)\,\mathrm{d}t\right) \tag{4.26}$$

$$f_T(t) = H(t) \cdot \exp\left(-\int_{-\infty}^{x} H(t)\,\mathrm{d}t\right) \tag{4.27}$$

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 $R_{g(X)} = \{g(X(\omega_1)), g(X(\omega_2)), \ldots\}$.

Osservazione 98. 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 $g(X)$.

Osservazione 99. In the discrete case finding PMF of $g(X)$ is usually easy, the following are some example.

4.4.1 Discrete rv transform

Osservazione 100. 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*

$$\mathbb{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbb{P}(X = x), \qquad \forall y \in R_{g(X)} \tag{4.28}$$

Esempio 4.4.1. In table 4.1 an example with X, $Y = 2X$ ($g(x) = 2 \cdot x$, injective) e $Z = X^2$ ($g(x) = x^2$ not injective).

Osservazione 101. 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.

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		

Table 4.1: PMF of discrete rv transform, an example

Esempio 4.4.2 (Transformation of a bernoulli)**.** Let $X \sim \text{Bern}(p)$ 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

$$\mathbb{P}(g(X)=e) = \mathbb{P}\left(X = g^{-1}(e)\right) = \mathbb{P}(X=1) = p$$

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 102. 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 103. 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.*

Proof. It has been obtained by a linear, injective transformation. □

Osservazione 104. 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 $0, 1, \ldots$).

4.4.3 Continuous rvs (monotonic) transform

Proposizione 4.4.3. *if X is a continuous random variable, g a monotonic function (strictly increasing or decreasing), the density function of the random variable $g(X)$,$f_{g(X)}$, is obtained as:*

$$f_{g(X)}(x) = f_X(g^{-1}(x)) \cdot \left| \frac{\partial g^{-1}(x)}{\partial x} \right| \tag{4.29}$$

Proof. For the continuous case we have that, in order to obtain $f_{g(X)}(x)$ we need to differentiate $F_{g(X)}(x)$

$$F_{g(X)}(x) = \mathbb{P}(g(X) \leq x)$$

Now

- if the function g is *decreasing* we have

$$F_{g(X)}(x) = \mathbb{P}\left(g(X) \leq x\right) = \mathbb{P}\left(X \geq g^{-1}(x)\right) = 1 - \mathbb{P}\left(X < g^{-1}(x)\right)$$
$$= 1 - F_X(g^{-1}(x))$$

- viceversa if g is *increasing*

$$F_{g(X)}(x) = \mathbb{P}\left(g(X) \leq x\right) = \mathbb{P}\left(X \leq g^{-1}(x)\right) = F_X(g^{-1}(x))$$

In any case after that we have that

$$f_{g(X)}(x) = \frac{\partial}{\partial x} F_{g(X)}(x) = \begin{cases} \frac{\partial\left(1 - F_X(g^{-1}(x))\right)}{\partial x} & \text{if increasing} \\ \frac{\partial\left(F_X(g^{-1}(x))\right)}{\partial x} & \text{if decreasing} \end{cases}$$
$$= \begin{cases} -f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x) \\ f_X(g^{-1}(x)) \cdot \frac{\partial}{\partial x} g^{-1}(x) \end{cases}$$

The two cases can be combined in the single formula (not clear how to me for the moment) which is the thorem □

Esempio 4.4.3 (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

$$0 \leq \log y \leq 1$$
$$1 \leq y \leq e$$

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 4.4.4 (Esame vecchio viroli)**.** Let X have the probability density function given by

$$f_X(x) = \frac{x}{2}$$

with $X \in [0,2]$. Find the density function of $Y = 6X - 3$.
Qua il dominio diventa palesemente $Y \in [-3,9]$, per quanto riguarda la funzione si ha che

$$f_Y(y) = \left|\frac{\partial}{\partial y} g^{-1}(y)\right| f_X(g^{-1}(y))$$
$$g(X) = 6X - 3 \quad g^{-1}(Y) = \frac{Y+3}{6}$$
$$f_Y(y) = \frac{1}{6}\left(\frac{Y+3}{6 \cdot 2}\right) = \frac{1}{6}\left(\frac{Y+3}{12}\right)$$

the answer is $f_Y(y) = \frac{3+y}{12}\frac{1}{6}$.
Si può verificare che $\int_{-3}^{9} f_Y(y) = 1$ mediante sympy. Qui non c'è il problema di resprimere le variabili indicatrici (perché non è una uniforme 0,1 e la densità non ne fa uso).

Esempio 4.4.5 (Assignment 1 Viroli, Exercise 2)**.** Let $X \sim \text{Unif}(0,1)$. Find the PDF of $X^{1/\alpha}$ with $\alpha > 0$.
Let $X \sim \text{Unif}(0,1)$ and $Y = X^{\frac{1}{\alpha}}$, with $\alpha > 0$. Let's obtain $f_Y(y)$ by applying:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{\partial g^{-1}(y)}{\partial y} \right| \tag{4.30}$$

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 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:

$$\begin{aligned} 0 \leq y^\alpha \leq 1 \\ 0 \leq y \leq 1 \end{aligned}$$

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

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [0,1] \\ 0 & \text{elsewhere} \end{cases} = \mathbb{1}_{[0,1]}(y) \implies Y \sim \text{Unif}(0,1)$$

Esempio 4.4.6 (Esercizio virol)**.** If $X \sim \text{Unif}(0,1)$ and $Y = -2\log X$, show that $Y \sim \chi^2_2$. We apply 4.29 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.

4.5 Rvs independence

Osservazione 105. 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

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y), \qquad \forall x, y \in \mathbb{R} \tag{4.31}$$

Osservazione 106 (Notation). $\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x \cap Y \leq y)$

Osservazione 107. In the discrete case 4.31 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:

$$\begin{aligned} &\mathbb{P}(X + Y = 12, X - Y = 1) = 0 \\ &\mathbb{P}(X + Y = 12) \cdot \mathbb{P}(X - Y = 1) = \frac{1}{6} \cdot \frac{5}{6} \end{aligned}$$

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.*

Proof. 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 $\text{Card}(\mathcal{X}) = n < \text{Card}(\mathcal{V})$

$$\begin{aligned} &\mathbb{P}(X_j \leq x_j, \ldots, X_k \leq x_k) = \mathbb{P}(X_j \leq x_j) \cdot \ldots \cdot \mathbb{P}(X_k \leq x_k) \\ &X_j, \ldots, X_k \in \mathcal{X}, \quad \forall x_j, \ldots, x_k \in \mathbb{R} \end{aligned} \tag{4.32}$$

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.*

Proof. 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 4.32, 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:

$$\mathbb{P}(X \leq x, Y \leq y | Z = z) = \mathbb{P}(X \leq x | Z = z) \cdot \mathbb{P}(Y \leq y | Z = z) \tag{4.33}$$

Osservazione 108. For discrete rvs, an equivalent definition based on the mass function is

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z) \cdot \mathbb{P}(Y = y | Z = z) \tag{4.34}$$

Proposizione 4.5.3. *Rvs indipendence does not imply conditional independence and viceversa.*

Proof. By counterexamples, see Blitzstein pag 121. □

4.6 Moments

Osservazione 109. 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.*

Proof. 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:

$$\mu_r = \mathbb{E}[X^r] = \begin{cases} \sum_i {x_i}^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} x^r \cdot f_X(x)\, \mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{4.35}$$

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:

$$\mathbb{E}[X] = \begin{cases} \sum_i x_i \cdot p_X(x_i) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} x \cdot f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases} \tag{4.36}$$

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)**.**

$$\mathbb{E}\left[aX+b\right] = a\,\mathbb{E}\left[X\right] + b \tag{4.37}$$
$$\mathbb{E}\left[X+Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right] \tag{4.38}$$
$$X \geq 0 \implies \mathbb{E}\left[X\right] \geq 0 \tag{4.39}$$
$$X \geq 0, \mathbb{P}\left(X > 0\right) > 0 \implies \mathbb{E}\left[X\right] > 0 \tag{4.40}$$
$$\mathbb{E}\left[g(X)\right] = \sum_i g(x_i) \cdot p_X(x_i) \tag{4.41}$$
$$X \perp\!\!\!\perp Y \implies \mathbb{E}\left[XY\right] = \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \tag{4.42}$$
$$\min\left(X\right) \leq \mathbb{E}\left[X\right] \leq \max\left(X\right) \tag{4.43}$$
$$\mathbb{E}\left[X - \mathbb{E}\left[X\right]\right] = 0 \tag{4.44}$$
$$\textit{minimizes } \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] \tag{4.45}$$

Osservazione 110. Congiuntamente alle 4.37 e 4.38 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 4.38, non è necessaria l'indipendenza tra X e Y affinché valga.

TODO: da chiarire sta nota di colore

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.

Proof. Mostriamo con riferimento alle variabili discrete. Per la 4.37

$$\begin{aligned}\mathbb{E}[aX+b] &= \sum_i (ax_i+b)\cdot\mathbb{P}(aX+b=ax_i+b) = \sum_i (ax_i+b)\cdot\mathbb{P}(X=x_i)\\ &= \sum_i ax_i\cdot\mathbb{P}(X=x) + \sum_i b\cdot\mathbb{P}(X=x)\\ &= a\sum_i x_i\cdot\mathbb{P}(X=x) + b\underbrace{\sum_i \mathbb{P}(X=x)}_{1}\\ &= a\,\mathbb{E}[X]+b\end{aligned}$$

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 4.38 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, 4.41, 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 4.42 (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 4.43 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 4.44 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 4.37. 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 $g = (x - \mathbb{E}[X])^r$:

$$\bar{\mu_r} = \mathbb{E}[(X - \mathbb{E}[X])^r] = \begin{cases} \sum_i (x_i - \mathbb{E}[X])^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^r \cdot f_X(x)\,\mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{4.46}$$

Osservazione 111. 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

$$\bar{\mu_2} = \mathrm{Var}[X] = \sigma^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{4.47}$$

measure dispersion of the rv around its mean value.

Proposizione 4.6.5 (Formula to use for evaluation)**.**

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \tag{4.48}$$

Proof. 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 expanded $(X - \mathbb{E}[X])^2$ 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 $a, b, c \in \mathbb{R}$:*

$$\operatorname{Var}[X] \geq 0 \tag{4.49}$$
$$\operatorname{Var}[X] = 0 \iff \mathbb{P}(X = c) = 1 \tag{4.50}$$
$$\operatorname{Var}[aX + b] = a^2 \operatorname{Var}[X] \tag{4.51}$$
$$X \perp\!\!\!\perp Y \implies \operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] \tag{4.52}$$

Proof. Per la 4.49, 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 4.50 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 4.51 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 4.52 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 112 (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)**.**

$$\sigma = \sigma_X = \sqrt{\operatorname{Var}[X]} \tag{4.53}$$

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:

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}[X]}} = \frac{X - \mathbb{E}[X]}{\sigma} \tag{4.54}$$

Osservazione 113. This transform make rv independent from measure unit.

Definizione 4.6.9 (r-th standardized moments of X)**.** We have them if $g = \left(\frac{x-\mathbb{E}[X]}{\sigma}\right)^r$:

$$\bar{\bar{\mu}}_r = \mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sigma}\right)^r\right] = \begin{cases} \sum_i \left(\frac{x_i-\mathbb{E}[X]}{\sigma}\right)^r \cdot p_X(x_i) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} \left(\frac{x-\mathbb{E}[X]}{\sigma}\right)^r \cdot f_X(x)\,\mathrm{d}x & \text{se } X \text{ è continua} \end{cases} \tag{4.55}$$

Osservazione 114. 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 115 (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 116. La definizione è meramente quella di una funzione simmetrica rispetto a $x = \mu$ (vedi calcolo).

Proof. 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:

$$\text{Asym}(X) = \bar{\mu}_3 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^3\right] \tag{4.56}$$

Osservazione 117. 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

$$\text{Kurt}(X) = \bar{\mu}_4 = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^4\right] \tag{4.57}$$

Osservazione 118. Some defines kurtosis by centering on 3 (value assumed by the normal) as in:

$$\text{Kurt}(X) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma}\right)^4\right] - 3 \tag{4.58}$$

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 119. Una distribuzione con eccesso di curtosi (4.58) 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 Random vectors and relationship between rvs

4.7.1 Random vectors

Definizione 4.7.1. A random vector is a function that maps $\Omega \to D \subset \mathbb{R}^n$.

Esempio 4.7.1. 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 4.7.2. Regarding (X, Y) with X = sums of both outcomes, Y = absolute difference ofthe two outcomes we have that

- $\mathbb{P}(X = 5, Y = 3) = \mathbb{P}(\{X = 5\} \cap \{Y = 3\}) = \mathbb{P}((4,1),(1,4)) = \frac{1}{18}$
-

TODO: questo dove l'ho preso? blitstein?

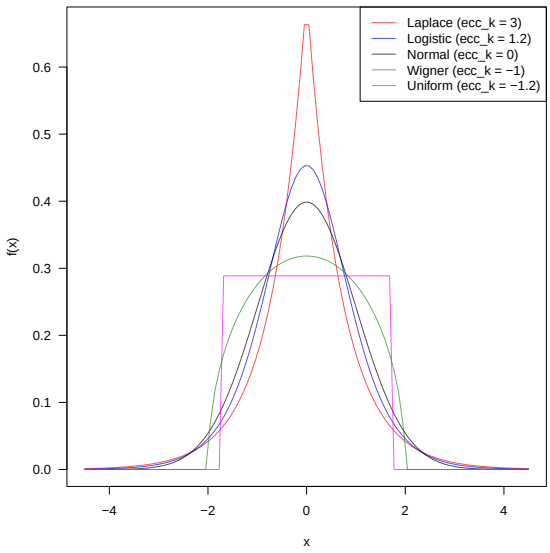


Figure 4.3: PDF for some rv (mean 0, variance 1) and their kurtosis

4.7.2 n-variate random variables (Rigo)

Let X be a n-variate random variable (or random vector), X can be written as $X = (X_1, \dots, X_n)^T$ or $X = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix}$ indifferently (by default, vector are column vector), where $X_1, \dots X_n$ 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

$$F(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) \tag{4.59}$$

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

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

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

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

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

Definizione 4.7.2 (Marginal)**.** A marginal of X is any subvector $(X_{j_1}, ..., X_{j_k})$ where $\{j_1, ..., j_k\}$ is a subset of $1, \dots, n$

Osservazione 120. 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 4.7.1 (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*

$$g(x_1, \dots, x_k) = \underbrace{\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty}}_{n-k \text{ integrals}} f(t_1, \dots, t_n) \, \mathrm{d}t_{k+1} \, \mathrm{d}t_n \tag{4.60}$$

Osservazione 121. 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 4.7.3. If $n = 3$, $X = (X_1, X_2, X_3)^T$, densitity of X_2 is

$$g(y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y, z) \, \mathrm{d}x \, \mathrm{d}z \tag{4.61}$$

Similarly the density of (X_2, X_3) is

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

Definizione 4.7.3 (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 4.7.4. 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 122. Now we extend to random vector a theorem (already discussed in the $n = 1$ case) useful for proving that X is absolutely continuous.

Teorema 4.7.2. *A random vector X is absolutely continuous if and only if*

$$\mathbb{P}(X \in A) = 0, \forall A \in \beta(\mathbb{R}^n), \quad stm(A) \tag{4.62}$$

Teorema 4.7.3. *If $X_1, ...X_n$ are absolutely continuous, this does not imply that the vector X is absolutely continuous.*

Esempio 4.7.5. 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

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

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 4.7.4 (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*

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n), \quad \forall \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix} \in \mathbb{R}^n \tag{4.63}$$

   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*

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdot \ldots \cdot f_n(x_n), \quad \forall \begin{bmatrix} X_1 \\ \ldots \\ X_n \end{bmatrix} \in \mathbb{R}^n \tag{4.64}$$

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*

Esempio 4.7.6 (Esame vecchio viroli)**.** Let $\mathbf{X} = (X, Y)^T$ be a random vector with joint density

$$f(x, y) = ky$$

where $0 < x < y < 1$. Compute k.
In order to compute k it must be:

$$\begin{aligned} 1 &= \int_0^1 \int_0^y ky \, \mathrm{d}x \, \mathrm{d}y = k \int_0^1 y \int_0^y 1 \, \mathrm{d}x \, \mathrm{d}y \\ &= k \int_0^1 y[x]_0^y \, \mathrm{d}y = k \int_0^1 y^2 \, \mathrm{d}y = k \left[\frac{y^3}{3}\right]_0^1 = \frac{k}{3} \end{aligned}$$

da cui $k = 3$

Esempio 4.7.7 (Esame vecchio viroli)**.** Boh

$$1 - x_2$$

with $x_2 < x_3 < 1$ compute the $f(x_1, x_3|x_2)$

1. $\frac{x_3^2}{x_2(1-x_2)}$
2. $\frac{3x_3}{x_2(1-x_2)}$ suggerita da taluni
3. $3x_3/x_2$
4. $3x_1/(x_1(1-x_2))$

Esempio 4.7.8. Consider the function

$$f(x|y) = \begin{cases} \frac{y^x e^{-y}}{x!} & \text{for } x = 0, 1, 2, \dots \text{ and } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

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

$$f_X(x) = \int_0^{+\infty} \frac{y^x e^{-2y}}{x!} \, \mathrm{d}y = \frac{1}{x!} \underbrace{\int_0^{+\infty} y^x e^{-2y} \, \mathrm{d}y}_{(1)}$$

$$= \frac{1}{x!} \frac{\Gamma(x+1)}{2^{x+1}} = \frac{1}{2^{x+1}}$$

where (1) is the kernel of a $\text{Gamma}(\alpha = x + 1, \beta = 2)$

4.7.3 Covariance (Rigo)

Definizione 4.7.4 (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

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - E(Y))] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{4.65}$$

Proposizione 4.7.5 (Proprietà covarianza (wikipedia, non fatte da rigo))**.** *If X, Y, W, V are real-valued random variables and $a, b, c, d \in \mathbb{R}$, then the following facts are a consequence of the definition of covariance:*

$$\operatorname{Cov}(X, a) = 0 \tag{4.66}$$
$$\operatorname{Cov}(X, X) = \operatorname{Var}[X] \tag{4.67}$$
$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X) \tag{4.68}$$
$$\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y) \tag{4.69}$$
$$\operatorname{Cov}(X + a, Y + b) = \operatorname{Cov}(X, Y) \tag{4.70}$$
$$\operatorname{Cov}(aX + bY, cW + dV) = ac \operatorname{Cov}(X, W) + ad \operatorname{Cov}(X, V) + bc \operatorname{Cov}(Y, W) + bd \operatorname{Cov}(Y, V) \tag{4.71}$$
$$X \perp\!\!\!\perp Y \implies \operatorname{Cov}(X, Y) = 0 \tag{4.72}$$

Esempio 4.7.9 (Esame vecchio viroli)**.** Let X_1 and X_2 be two random variables with distribution $X_1 \sim \mathrm{N}(0, 2)$ and $X_n \sim \mathrm{N}(-2, 1)$ (parameters are mean and variance) and covariance -1. Compute $\operatorname{Cov}(X_1 + X_2, X_1 - X_2)$.
We have that

$$\begin{aligned}\operatorname{Cov}(X_1 + X_2, X_1 - X_2) &= \operatorname{Cov}(X_1, X_1) - \operatorname{Cov}(X_1, X_2) + \operatorname{Cov}(X_2, X_1) - \operatorname{Cov}(X_2, X_2) \\ &= \operatorname{Var}[X_1] - \operatorname{Var}[X_2] = 2 - 1 = 1\end{aligned}$$

Esempio 4.7.10 (Esame vecchio viroli)**.** Let X_1, X_2 be two standard gaussian variables with covariance -1. Compute $\operatorname{Cov}(X_1 + X_2, X_1 - X_2)$.
With the same developmet as above we have:

$$\operatorname{Cov}(X_1 + X_2, X_1 - X_2) = \operatorname{Var}[X_1] - \operatorname{Var}[X_2] = 1 - 1 = 0$$

Esempio 4.7.11 (Esame vecchio viroli)**.** Let X and Y be two independent bernoulli random variables with same parameter p. Compute $\operatorname{Cov}(Y - X, 2X + 2Y)$.

$$\begin{aligned}\operatorname{Cov}(Y - X, 2X + 2Y) &= 2 \operatorname{Cov}(X, Y) + 2 \operatorname{Cov}(Y, Y) - 2 \operatorname{Cov}(X, X) - 2 \operatorname{Cov}(X, Y) \\ &= 2 \operatorname{Var}[X] - 2 \operatorname{Var}[Y] = 0\end{aligned}$$

taluni suggeriscono -1 ma mi pare na gran cacata

```
p = 0.5
x = rbinom(100000, 1, 0.5)
y = rbinom(100000, 1, 0.5)
cov(y-x, 2*x+2*y)

## [1] 8.904089e-06
```

Proposizione 4.7.6. *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 4.52 for the variance of the sum of random variables involves their covariance, that is*

$$\begin{aligned}\operatorname{Var}\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n} a_i^2 \operatorname{Var}[X_i] + \sum_{i\neq j} a_i a_j \operatorname{Cov}(X_i, X_j) \\ &\overset{(1)}{=} \sum_{i=1}^{n} a_i^2 \operatorname{Var}[X_i] + 2\sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j) \qquad (4.73)\end{aligned}$$

where (1) *because* $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$

Proof. 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 4.7.12 (Esame vecchio viroli)**.** Let $X = (X_1, X_2)$ be a bivariate gaussian vector with $\mu = [0, 0]$ and

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

What is the distribution of $Y = 3X_1 - 2X_2$?
Si ha che

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[3X_1 - 2X_2] = 3\,\mathbb{E}[X_1] - 2\,\mathbb{E}[X_2] = 0 \\ \operatorname{Var}[Y] &= \operatorname{Var}[3X_1 - 2X_2] \\ &= 3^2 \operatorname{Var}[X_1] + (-2)^2 \operatorname{Var}[X_2] + 2(3\cdot(-2))\operatorname{Cov}(X_1, X_2) \\ &= 9\cdot 1 + 4\cdot 1 + 2\cdot(-6)\cdot 0.5 = 7\end{aligned}$$

quindi è $Y \sim \mathrm{N}(0, 7)$ come confermato da taluni

Esempio 4.7.13 ($\operatorname{Cov}(X, Y) = 0$ 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 4.7.14. An example of 4.73 is $\operatorname{Var}[X - Y] = \operatorname{Var}[X] + \operatorname{Var}[Y] - 2\operatorname{Cov}(X, Y)$

Osservazione 123. In the lucky case these are indipendent covariance is null and variance of sum is sum of variance.

4.7.4 Correlation coefficient

Definizione 4.7.5 (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

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}[X]}\sqrt{\operatorname{Var}[Y]}} \tag{4.74}$$

Proposizione 4.7.7. *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:*

$$\operatorname{Corr}(X, Y) = 1 \iff Y = a + bX, b > 0$$
$$\operatorname{Corr}(X, Y) = -1 \iff Y = a + bX, b < 0$$

Esempio 4.7.15 (Esame vecchio viroli)**.** Let X and Y be two gaussian variables with zero mean $\operatorname{Var}[X] = 1$, $\operatorname{Var}[Y] = 9$, covariance -1, compute $\rho(X + Y, X)$.

$$\begin{aligned}\operatorname{Corr}(X + Y, X) &= \frac{\operatorname{Cov}(X + Y, X)}{\sqrt{\operatorname{Var}[X + Y]}\sqrt{\operatorname{Var}[X]}} = \frac{\operatorname{Cov}(X, X) + \operatorname{Cov}(Y, X)}{\sqrt{\operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}(X, Y)}\sqrt{\operatorname{Var}[X]}} \\ &= \frac{\operatorname{Var}[X] + \operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}[X] + \operatorname{Var}[Y] + 2\operatorname{Cov}(X, Y)}\sqrt{\operatorname{Var}[X]}} = \frac{1 + (-1)}{\sqrt{1 + 9 + 2(-1)}\sqrt{1}} \\ &= 0\end{aligned}$$

Il risultato è confermato dal Bigo.

Esempio 4.7.16 (Esame vecchio viroli)**.** Let X and Y be two gaussian variables with zero mean $\operatorname{Var}[X] = 1$, $\operatorname{Var}[Y] = 9$, covariance -1, compute $\rho(1 - 2X + 2, 3 + Y)$.
We have:

$$\begin{aligned}\operatorname{Corr}(1 - 2X + 2, 3 + Y) = \operatorname{Corr}(3 - 2X, 3 + Y) &= \frac{\operatorname{Cov}(3 - 2X, 3 + Y)}{\sqrt{\operatorname{Var}[3 - 2X]}\sqrt{\operatorname{Var}[3 + Y]}} \\ &= \frac{\operatorname{Cov}(3,3) + \operatorname{Cov}(3,Y) + \operatorname{Cov}(-2X,3) + \operatorname{Cov}(-2X,Y)}{\sqrt{4\operatorname{Var}[X]}\sqrt{\operatorname{Var}[Y]}} \\ &= \frac{0 + 0 + 0 + 2}{2 \cdot 1 \cdot 3} = \frac{1}{3}\end{aligned}$$

come confermato da taluni

4.8 Exercises

Esempio 4.8.1 (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. $\operatorname{Var}[X]$
4. $F(X)$

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

$$\mathbb{E}\left[X\right]=\lambda\left(\frac{1}{\lambda^2}\right)=\frac{1}{\lambda}$$

3. first we find $\mathbb{E}\left[X^2\right]$

$$\begin{aligned}\mathbb{E}\left[X^2\right]&=\lambda\int_{-\infty}^{+\infty}x^2e^{-\lambda x}\,\mathrm{d}x\overset{(1)}{=}\lambda\left[(x^2\frac{-1}{\lambda}e^{-\lambda x})\big|_0^{\infty}+\frac{2}{\lambda}\int_0^{+\infty}xe^{-\lambda x}\,\mathrm{d}x\right]\\&=2\int_0^{+\infty}xe^{-\lambda x}\,\mathrm{d}x=\frac{2}{\lambda^2}\end{aligned}$$

where in (1) again by integration by parts. So

$$\mathrm{Var}\left[X\right]=\mathbb{E}\left[X^2\right]-\mathbb{E}\left[X\right]^2=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}$$

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 4.8.2 (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 $\mathbb{E}\left[X\right]$
3. find $\mathrm{Var}\left[X\right]$

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.

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k f(k) = \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}$$
$$\overset{(1)}{=} \lambda \underbrace{\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!}}_{F(\infty)=1} = \lambda$$

with (1) substituting $u = k - 1$. This is the poisson distribution, we say $X \sim \text{Pois}(\lambda)$

3. first we find $\mathbb{E}\left[X^2\right]$, but first consider the following

$$\mathbb{E}[X(X-1)] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X] = \sum_{k=0}^{\infty} k(k-1) f(k) = \sum_{k=2}^{\infty} k(k-1) f(k)$$
$$= \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=2}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-2)!} = \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2} e^{-\lambda}}{(k-2)!}$$
$$\overset{(1)}{=} \lambda^2 \underbrace{\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!}}_{F(\infty)=1} = \lambda^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]$$

where in (1) doin subst $u = k - 2$. Therefore

$$\mathbb{E}\left[X^2\right] = \lambda^2 + \lambda \implies \text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Esempio 4.8.3 (crashcourse, day 1 es 3 pag 7)**.** Let $X \sim \text{Bin}(n, p)$, 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 4.8.4 (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. $\mathbb{E}[X]$
4. $\text{Var}[X]$

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 4.8.5 (crashcourse, day 1 es 5 pag 8)**.** Let $f(x) = ce^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$:

1. find c
2. $\mathbb{E}[X]$
3. $\mathrm{Var}[X]$

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

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{+\infty} \overbrace{\underbrace{x^2}_{even}\underbrace{f(x)}_{even}}^{even}$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x \overset{(1)}{=} \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x \overset{(2)}{=} \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} \sqrt{2}u^{1/2}e^{-u}\,\mathrm{d}u$$

$$= \frac{2}{\sqrt{\pi}}\int_{-\infty}^{+\infty} u^{1/2}e^{-u}\,\mathrm{d}u = \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = 1$$

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 $\Gamma(3/2) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}$

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

$$\mathbb{E}[X] = \mathbb{E}[\sigma Z + \mu] = \sigma \underbrace{\mathbb{E}[Z]}_{=0} + \mu = \mu$$

$$\mathrm{Var}[X] = \mathrm{Var}[\sigma Z + \mu] = \sigma^2 \underbrace{\mathrm{Var}[Z]}_{} = 1 = \sigma^2$$

4.9 Probability models and R

Osservazione 124. In the following chapters we study main probabilistic models, which are the most used family of distribution

Definizione 4.9.1 (Family of random variables)**.** Set of distribution function $F(x; \boldsymbol{\Theta})$ having the same functional form+ but different for one or more parameters.

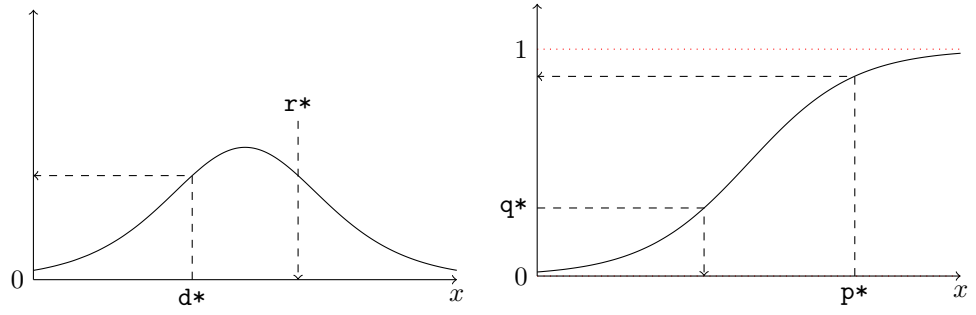
Definizione 4.9.2 (Parameters space)**.** $\boldsymbol{\Theta}$, it's the set of possible value for the parameters of a distribution function.

Osservazione 125. In:

- table 4.2 we report prefix of main functions and suffixes of main families;

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`

Table 4.2: Utilities for family of rvs in R



(a) PMF/PDF (b) CDF

Figure 4.4: Funzioni in R

- figure 4.4 function input needed (where arrows starts) and output returned (where arrow ends) for the 4 main functions.

Osservazione 126 (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]  30 100  48  41  46  46  96  71  89  94

## Urna discreta custom
sample(x = 1:3, prob = c(0.4, 0.4, 0.2), size = 10, replace = TRUE)

##  [1] 1 1 3 2 1 3 3 2 1 1
```

Chapter 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.

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geq c \end{cases} \tag{5.1}$$

Proposizione 5.1.2 (Moments)**.**

$$\mathbb{E}[X] = c$$
$$\mathrm{Var}[X] = 0$$

Proof.

$$\mathbb{E}[X] = c \cdot 1 = c$$
$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = c^2 \cdot 1 - c^2 = 0$$

□

Osservazione 127. Dirac is the only random variable having null variance.

5.2 Bernoulli

5.2.1 Definition

Osservazione 128. 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 129. If $p = 0 \vee p = 1$ we obtain a Dirac.

5.2.2 Functions

Osservazione 130 (Support and parametric space).

$$R_X = \{0, 1\}$$
$$\mathbf{\Theta} = \{p \in \mathbb{R} : 0 \leq p \leq 1\}$$

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)**.**

$$\mathbb{E}[X] = p \tag{5.4}$$
$$\text{Var}[X] = p(1-p) \tag{5.5}$$
$$\text{Asym}(X) = \frac{1-2p}{\sqrt{p(1-p))}} \tag{5.6}$$
$$\text{Kurt}(X) = \frac{3p^2 - 3p + 1}{p(1-p)} \tag{5.7}$$

Proof. 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 131. 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)**.**

$$(I_A)^n = I_A, \quad \forall n \in \mathbb{N} : n > 0 \tag{5.8}$$
$$I_{\overline{A}} = 1 - I_A \tag{5.9}$$
$$I_{A \cap B} = I_A \cdot I_B \tag{5.10}$$
$$I_{A \cup B} = I_A + I_B - I_A \cdot I_B \tag{5.11}$$

Proof. 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 132. 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:*

$$\mathbb{P}(A) = \mathbb{E}[I_A] \tag{5.12}$$

Proof. 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 133 (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:*

$$\mathbb{P}(E_1 \cup \ldots \cup E_n) \leq \mathbb{P}(E_1) + \ldots + \mathbb{P}(E_n) \tag{5.13}$$

Proof. 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:*

$$\mathbb{P}(E_1 \cap \ldots \cap E_n) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left(\overline{E_i}\right) \tag{5.14}$$

Proof. 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:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \ldots < i_r} \mathbb{P}(E_{i1} \cap E_{i2} \cap \ldots \cap E_{ir}) \tag{5.16}$$

$$= \sum_{i} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k) - \ldots + (-1)^{n+1} \mathbb{P}(E_1 \cap \ldots \cap E_n) \tag{5.17}$$

Proof. 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

Table 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 134. 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 135. 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 $X = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$*

Proof. 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 136 (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 137. 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)**.**

$$\mathbb{E}[X] = np \tag{5.19}$$
$$\mathrm{Var}[X] = np(1-p) \tag{5.20}$$
$$\mathrm{Asym}(X) = \frac{1-2p}{\sqrt{np(1-p))}} \tag{5.21}$$
$$\mathrm{Kurt}(X) = 3 + \frac{1 - 6p + 6p^2}{np(1-p)} \tag{5.22}$$

Proof. 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 $= 1$. □

Proof. 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

```
plot_binom <- function(n, p, plot_main = TRUE, ...){
    the_seq <- seq(from = 0, to = n)
    probs <- dbinom(x = the_seq, size = n, p = p)
    plot(x = the_seq, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob',
         main = if (plot_main) sprintf('Bin(%d, %.2f)', n, p) else '',
         ...)
}
```

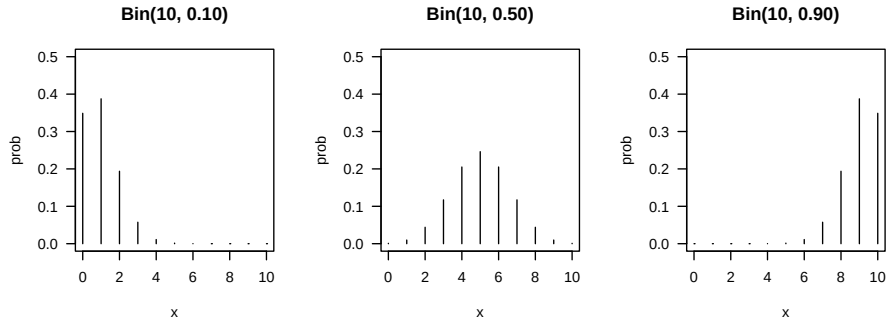


Figure 5.1: Forma distribuzione Bin (n, p)

```
par(mfrow = c(1,3))
ylim <- c(0, 0.5)
plot_binom(n = 10, p = 0.1, ylim = ylim)
plot_binom(n = 10, p = 0.5, ylim = ylim)
plot_binom(n = 10, p = 0.9, ylim = ylim)
```

```
par(mfrow = c(2,2))
first_n <- 10
second_n <- 40
first_p <- 0.1
second_p <- 0.35
ylim <- c(0, 0.5)
plot_binom(n = first_n,  p = first_p,  ylim = ylim)
plot_binom(n = second_n, p = first_p,  ylim = ylim)
plot_binom(n = first_n,  p = second_p, ylim = ylim)
plot_binom(n = second_n, p = second_p, ylim = ylim)
```

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)*

Proof. Per $p = 0.5$ è simmetrica in quanto $p = 1 - p = \frac{1}{2}$ e

$$p_X(x) = \binom{n}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{n-x} = p_X(n-x) = \binom{n}{n-x}\left(\frac{1}{2}\right)^{n-x}\left(\frac{1}{2}\right)^x \quad (5.24)$$

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. □

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 $x \leq (n+1)p$*

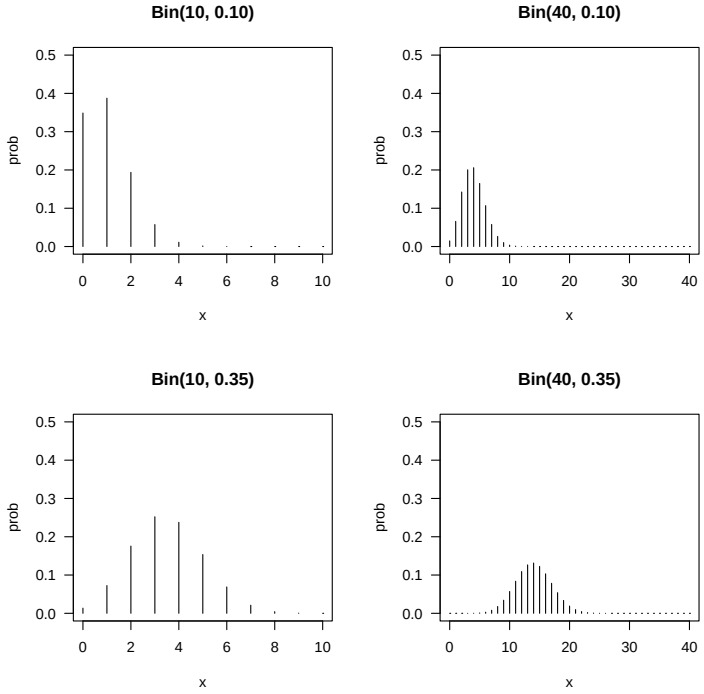


Figure 5.2: Convergenza alla normale della binomiale

Proof. 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:

$$(n-x+1)p \geq x(1-p)$$
$$np - xp + p \geq x - xp$$

ossia $x \leq (n+1)p$ □

Osservazione 138 (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* $n - X \sim \text{Bin}(n, 1 - p)$.

Proof. 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:

$$\mathbb{P}(Y = x) \overset{(1)}{=} \mathbb{P}(X = n - x) = \binom{n}{n-x} p^{n-x}(1-p)^x$$
$$\overset{(1)}{=} \binom{n}{x}(1-p)^x p^{n-x} = \text{Bin}(n, 1-p)$$

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 139. 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 $X + Y \sim \text{Bin}(n+m, p)$*

Proof. 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:

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{j=0}^{k} \mathbb{P}(X+Y=k|X=j)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \mathbb{P}(Y=k-j|X=j)\cdot\mathbb{P}(X=j) \\
&\overset{(1)}{=} \sum_{j=0}^{k} \mathbb{P}(Y=k-j)\cdot\mathbb{P}(X=j) \\
&= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j}(1-p)^{m-k+j} \cdot \binom{n}{j} p^j (1-p)^{n-j} \\
&= p^k(1-p)^{n+m-k} \sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j} \\
&\overset{(2)}{=} \binom{n+m}{k} p^k (1-p)^{n+m-k} = \text{Bin}(n+m, p)
\end{aligned}$$

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 140. 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 $X \sim \text{HGeom}(w, b, n)$.

5.5.2 Functions

Osservazione 141 (Supporto e spazio parametrico)*.*

$$R_X = \{0, 1, \dots, n\}$$
$$\boldsymbol{\Theta} = \{w, b \in \mathbb{N} : w + b \geq 1;\ n \in \{0, \dots, w + b\}\}$$

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 142 (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 143*.* 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)**.**

$$\mathbb{E}[X] = n\frac{w}{w+b} \tag{5.26}$$

$$\text{Var}[X] = np(1-p)\left(\frac{w+b-n}{w+b-1}\right), \qquad \textit{con } p = \frac{w}{w+b} \tag{5.27}$$

Proof. Per il valore atteso, come nel caso binomiale possiamo scrivere X come somma di Bernoulliane $I_i \sim \text{Bern}(p)$ 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

$$\mathbb{E}[X] = \mathbb{E}[I_1 + \ldots + I_n] = \mathbb{E}[I_1] + \ldots + \mathbb{E}[I_n] = np = n\frac{w}{w+b}$$

□

Proof. 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 144. 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 145. 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.*

Proof. 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 146. 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* $\text{Bin}(n, p)$.

Proof. 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 $\text{Bin}(n, p)$.
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 147. 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 148. Dimostriamo attraverso un esempio (distribuzione del test esatto di Fisher).

Proof. 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

	Donne	Uomini	Tot
Malato	x	$r-x$	r
Sano	$n-x$	$m-r+x$	$n+m-r$
Tot	n	m	$n+m$

Table 5.2

5.4.6). Pertanto sostituendo le formule per esteso si ha

$$\begin{aligned}\mathbb{P}(X=x|X+Y=r) &= \frac{\binom{m}{r-x}p^{r-x}(1-p)^{m-r+x}\cdot\binom{n}{x}p^{x}(1-p)^{n-x}}{\binom{n+m}{r}p^{r}(1-p)^{n+m-r}} \\ &= \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}} = \mathrm{HGeom}(n,m,r)\end{aligned}$$

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 149. Supponiamo di ripetere in maniera indipendente diverse prove bernoulliane, ciascuna avente p probabilità di successo, sino a che che si verifica 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 $X \sim \mathrm{Geom}(p)$.

Esempio 5.6.1. Il numero di croci sino alla prima testa si distribuisce come $\mathrm{Geom}(1/2)$.

5.6.2 Functions

Osservazione 150 (Supporto e spazio parametrico).

$$\begin{aligned}R_X &= \{x \in \mathbb{N}\} \\ \boldsymbol{\Theta} &= \{p \in (0,1)\}\end{aligned}$$

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 151. Come il teorema binomiale mostra che la PMF binomiale sia valida, la serie geometrica mostra che la PMF Geometrica sia valida.

Osservazione 152 (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)**.**

$$\mathbb{E}[X] = \frac{1-p}{p}$$
$$\mathrm{Var}[X] = \frac{1-p}{p^2}$$

Proof. 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

$$p(1-p)\sum_{x=0}^{\infty} x(1-p)^{x-1} = \frac{1}{p^2}p(1-p)$$
$$\sum_{x=0}^{\infty} xp(1-p)^{x} = \frac{1-p}{p}$$

□

Proof. 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

$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} \cdot (-1) = \frac{(-1) \cdot p^2 - 2p \cdot (1-p)}{p^4}$$
$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} = (-1)\frac{p^2 - 2p}{p^4}$$
$$\sum_{x=1}^{\infty} x^2 \cdot (1-p)^{x-1} = \frac{2-p}{p^3}$$

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 153 (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)

```
## occhio alla parametrizzazione di R per cui p(x) = p (1-p)^x
plot_geom <- function(p, plot_main = TRUE, ...){
    the_seq <- seq(from = 0, length = 15)
    probs <- dgeom(x = the_seq, prob = p)
    plot(x = the_seq + 1, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob',
         main = if (plot_main) sprintf('Geom(%.2f)', p) else '',
         ...)
}

all_p <- c(0.2, 0.5, 0.8)
par(mfrow = c(1, 3))
rm <- lapply(all_p, function(p) plot_geom(p = p, ylim = c(0, 1)))
```

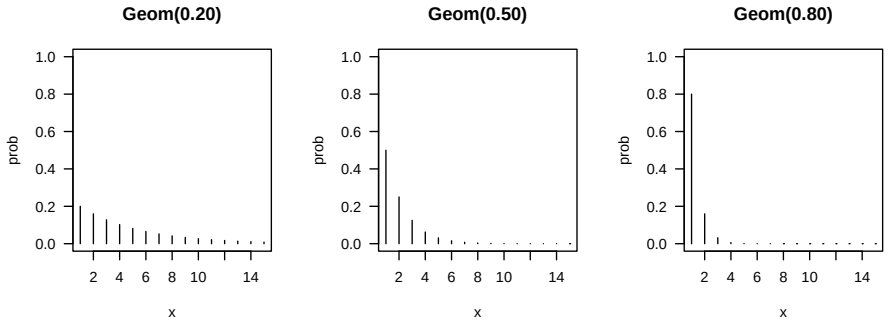


Figure 5.3: Forma distribuzione Geom (p)

5.6.5 Assenza di memoria

Osservazione 154. 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)**.**

$$\mathbb{P}(X > t+s | X > t) = \mathbb{P}(X > s) \tag{5.30}$$

Proof. 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 155. 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 156. 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 157 (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 158 (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)**.**

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{p} \\ \mathrm{Var}[X] &= \frac{1-p}{p^2} \\ \mathrm{Asym}(X) &= \frac{2-p}{\sqrt{1-p}} \\ \mathrm{Kurt}(X) &= 9 + \frac{p^2}{1-p} \end{aligned}$$

Proof. 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

$$\begin{aligned} \mathbb{E}[Y] = \mathbb{E}[X+1] = \mathbb{E}[X] + 1 = \frac{1-p}{p} + 1 = \frac{1}{p} \\ \mathrm{Var}[Y] = \mathrm{Var}[X+1] = \mathrm{Var}[X] = \frac{1-p}{p^2} \end{aligned}$$

□

Proposizione 5.6.4 (Assenza di memoria)**.** *Analogamente a quanto avviene per la geometrica* $\mathbb{P}(X > t+s | X > t) = \mathbb{P}(X > s)$.

Proof. 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 159. 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 $X \sim \text{Nb}(k, p)$

Osservazione 160. 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 161 (Supporto e spazio parametrico).

$$R_X = \mathbb{N}$$
$$\boldsymbol{\Theta} = \{k \in \mathbb{N} : k \geq 1, p \in \mathbb{R} : 0 \leq p \leq 1\}$$

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 162 (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 163. 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 $X \sim \text{Nb}(k, p)$ 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 $X_i \sim \text{Geom}(p)$.*

Proof. 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 $X_1 \sim \text{Geom}(p)$ per la definizione della geometrica, $X_2 \sim \text{Geom}(p)$ 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)**.**

$$\mathbb{E}[X] = k\frac{1-p}{p} \tag{5.34}$$

$$\text{Var}[X] = k\frac{1-p}{p^2} \tag{5.35}$$

Proof. 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 164 (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)

```
## The probability of obtaining the fourth cross before the
## third head (and then after two head) is equal to 11.72%.

plot_binom_neg <- function(k, p, plot_main = TRUE, ...){
    fails <- seq(from = 0, length = 20)
    probs <- dnbinom(x = fails, size = k, p = p)
    plot(x = fails, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob', xlim = range(fails),
         main = if (plot_main) sprintf('BN(%d, %.2f)', k, p) else '',
         ...)
}

par(mfrow = c(2,3))
plot_binom_neg(k =  1, p = 0.5, ylim = c(0, 0.5))
plot_binom_neg(k =  3, p = 0.5, ylim = c(0, 0.5))
plot_binom_neg(k = 10, p = 0.5, ylim = c(0, 0.5))
## incremento di p
plot_binom_neg(k =  3, p = 0.25, ylim = c(0, 0.45))
plot_binom_neg(k =  3, p = 0.5, ylim = c(0, 0.45))
plot_binom_neg(k =  3, p = 0.75, ylim = c(0, 0.45))
```

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 $X \sim \mathrm{Nb}(k, p)$.

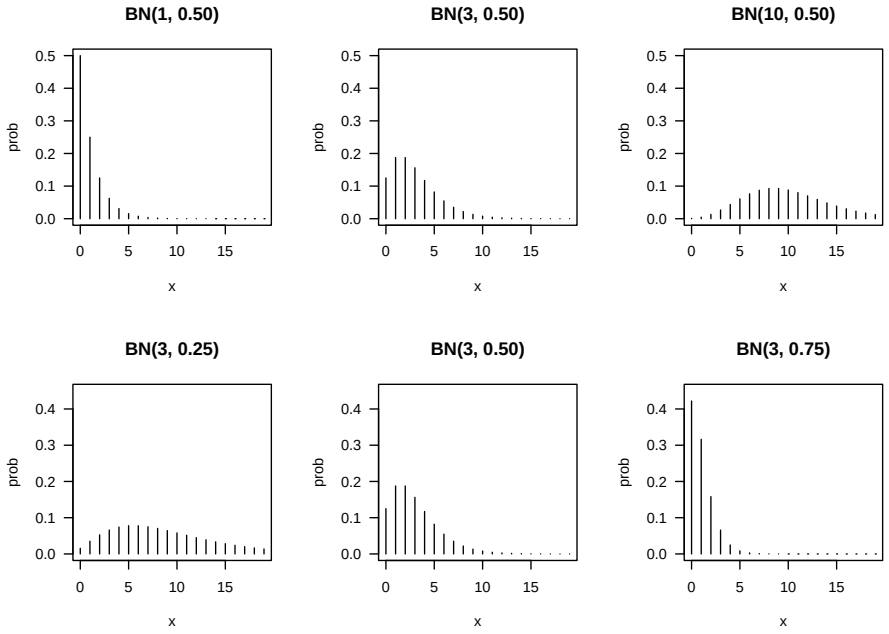


Figure 5.4: Distribuzione binomiale negativa

5.7.5.2 Functions

Osservazione 165 (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 166 (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)**.**

$$\mathbb{E}[X] = \frac{k}{p}$$
$$\mathrm{Var}[X] = \frac{k(1-p)}{p^2}$$
$$\mathrm{Asym}(X) = \frac{2-p}{\sqrt{k(1-p)}}$$
$$\mathrm{Kurt}(X) = 3 + \frac{6}{k} + \frac{p^2}{k(1-p)}$$

5.8 Poisson

5.8.1 Definition

Osservazione 167. 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 $X \sim \mathrm{Pois}(\lambda)$.

Osservazione 168. Un risultato che ci servirà per questa distribuzione è il seguente

Proposizione 5.8.1 (Sviluppo di Maclaurin della funzione esponenziale)**.**

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{5.37}$$

Proof. 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 169 (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)**.**

$$\mathbb{E}[X] = \lambda \tag{5.39}$$
$$\mathrm{Var}[X] = \lambda \tag{5.40}$$
$$\mathrm{Asym}(X) = \frac{1}{\sqrt{\lambda}} \tag{5.41}$$
$$\mathrm{Kurt}(X) = 3 + \frac{1}{\lambda} \tag{5.42}$$

Proof. 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. □

Proof. 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

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2 = \lambda(1+\lambda) - \lambda^2 = \lambda$$

□

Proof. Dimostrazione alternativa per la varianza:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[X^2\right] - \left[\mathbb{E}[X]\right]^2 \\
&= \left(\sum_{x=0}^{\infty} x^2 \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x^2 + x - x) \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x(x-1) + x) \cdot \frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} (x(x-1))\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^2\lambda^{x-2}}{x(x-1)(x-2)!} + \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda\lambda^{x-1}}{x(x-1)!}\right) - \lambda^2 \\
&= \left(\sum_{x=0}^{\infty} \frac{\lambda^{x-2}}{(x-2)!}e^{-\lambda}\lambda^2 + \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}e^{-\lambda}\lambda\right) - \lambda^2 \\
&\overset{(1)}{=} \left(\sum_{z=0}^{\infty} \frac{\lambda^z}{z!}e^{-\lambda}\lambda^2 + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}e^{-\lambda}\lambda\right) - \lambda^2 \\
&= (e^{\lambda}e^{-\lambda}\lambda^2 + e^{\lambda}e^{-\lambda}\lambda) - \lambda^2 \\
&= (\lambda^2 + \lambda) - \lambda^2 \\
&= \lambda
\end{aligned}$$

dove in (1) abbiamo posto $y = x-1$, $z = x-2$ per sfruttare 5.37 nel seguito. □

5.8.4 Shape

Osservazione 170 (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)

```
plot_pois <- function(lambda, plot_main = TRUE, ...){
    x <- 0:10
    probs <- dpois(x = x, lambda = lambda)
    plot(x = x, y = probs, type = 'h', las = 1,
         xlab = 'x', ylab = 'prob', xlim = range(x),
         main = if (plot_main) sprintf('Pois(%.1f)', lambda) else '',
         ...)
}
```

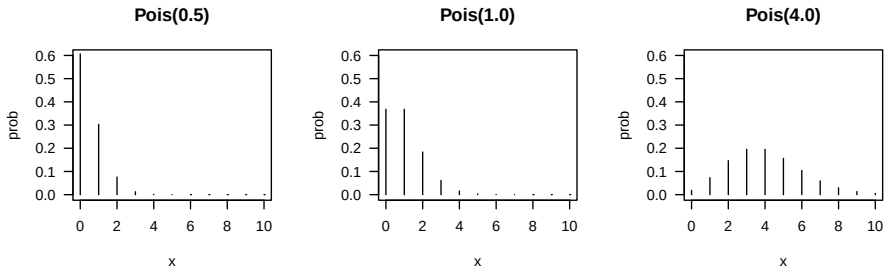


Figure 5.5: Distribuzione Poisson

```
par(mfrow = c(1,3))
tmp <- lapply(c(0.5, 1, 4), plot_pois, ylim = c(0, 0.6))
```

5.8.5 Origine e approssimazione

Osservazione 171. 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)

Proposizione 5.8.3 (Paradigma di Poisson)**.** *Siano $E_1, \dots, E_n$ 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 $\text{Pois}(\lambda)$ *con $\lambda = \sum_i p_i$.*

Proof. La prova dell'approssimazione di sopra è complessa, richiede definire la dipendenza debole e buona approssimazione; è omessa qui. □

Osservazione 172 (Ruolo di λ). Il parametro λ è interpretato come *rate di occorrenza*: ad esempio $\lambda = 2$ mail di spam per giorno.

Osservazione 173. 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 $N \sim \text{Pois}(\lambda)$, 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:

$$\mathbb{P}(X \in I) - \mathbb{P}(N \in I) \leq \min\left(1, \frac{1}{\lambda}\right) \sum_{i=1}^{n} p_i^2 \tag{5.43}$$

Proof. Anche questa è per ora complessa (necessita di una tecnica chiamata metodo di Stein). □

Osservazione 174. 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 175. 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 $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$*

Proof. 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 176. 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* $\text{Bin}(n, \lambda_1/(\lambda_1\lambda_2))$.

Proof. 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 177. 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 178. 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.

Proof. 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 179. 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 $p = 2 \times 10^{-1}$. 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

$$\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X < 3) \approx 1 - e^{-2} - e^{-2}\cdot 2 - e^{-2}\cdot\frac{2^2}{2!} = 1 - 5e^{-2} \approx 0.3233$$

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 180. 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 181 (Supporto e spazio parametrico)*.*

$$\begin{aligned} R_X &= \{1, \ldots, n\} \\ \boldsymbol{\Theta} &= \{n \in \mathbb{N} \setminus \{0\}\} \end{aligned}$$

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)**.**

$$F_X(x) = \mathbb{P}\,(X \leq x) = \begin{cases} 0 & \text{se } x < 1 \\ \frac{k}{n} & \text{se } k \leq x < k+1, (k = 1, 2, \ldots, n-1) \\ 1 & \text{se } x \geq n \end{cases} \tag{5.45}$$

Osservazione 182. 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)**.**

$$\mathbb{E}[X] = \frac{n+1}{2} \tag{5.46}$$

$$\mathrm{Var}[X] = \frac{n^2-1}{12} \tag{5.47}$$

$$\mathrm{Asym}(X) = 0 \tag{5.48}$$

$$\mathrm{Kurt}(X) = 1.8 \tag{5.49}$$

Proof.

$$\mathbb{E}[X] = \sum_{x=1}^{n} x\frac{1}{n} = \frac{1}{n}(1+2+\ldots+n) = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

□

Proof.

$$\begin{aligned}
\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - [\mathbb{E}[x]]^2 &= \left(\sum_{x=1}^{n} x^2\frac{1}{n}\right) - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{1}{n}(1^2+2^2+\ldots+n^2)\right) - \left(\frac{n+1}{2}\right)^2 \\
&= \left(\frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6}\right) - \left(\frac{n^2+1+2n}{4}\right) \\
&= \left(\frac{(n+1)(2n+1)}{6}\right) - \left(\frac{n^2+1+2n}{4}\right) \\
&= \frac{2(2n^2+2n+n+1) - 3(n^2+1+2n)}{12} \\
&= \frac{4n^2+4n+2n+2-3n^2-3-6n}{12} \\
&= \frac{n^2-1}{12}
\end{aligned}$$

□

Chapter 6

Absolute continuous random variables

6.1 Logistica

6.1.1 Origine/definizione

Osservazione 183. 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 184. 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 185. 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 186. 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 187 (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

$$f_X(x) = \frac{\left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(1 + e^{\frac{x-\mu}{\sigma}}\right) - \left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(e^{\frac{x-\mu}{\sigma}}\right)}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2} = \frac{\left(e^{\frac{x-\mu}{\sigma}} \cdot \frac{1}{\sigma}\right)\left(1 + e^{\frac{x-\mu}{\sigma}} - e^{\frac{x-\mu}{\sigma}}\right)}{\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2}$$
$$= \frac{e^{\frac{x-\mu}{\sigma}}}{\sigma\left(1 + e^{\frac{x-\mu}{\sigma}}\right)^2}$$

□

```
par(mfrow = c(1,2))
## mu <- c(5, 9, 9, 6, 2)
## s  <- c(2, 3, 4, 2, 1)
mu <- c(0, 2, 2 ,5, 5)
s  <- c(1, 3, 4, 3, 4)
```

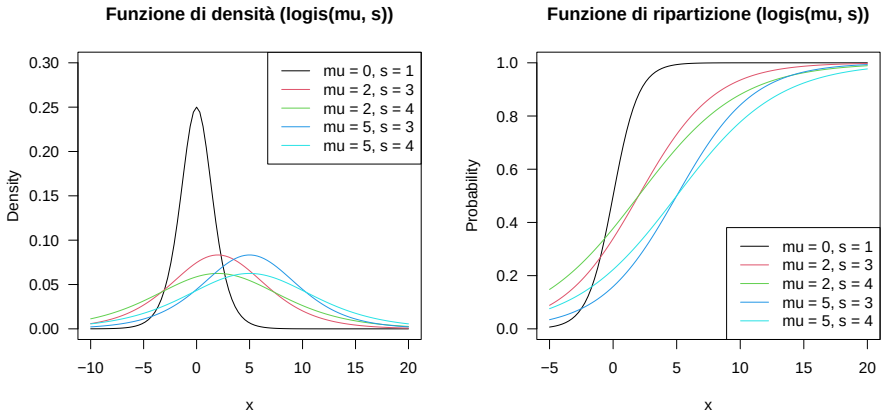


Figure 6.1: Distribuzione logistica

```
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) dlogis(x, location = mu, scale = s),
             from = -10, to = 20,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.3),
             ylab = 'Density', las = 1,
             main = 'Funzione di densità (logis(mu, s))')
}, as.list(mu), as.list(s), as.list(c(F, T, T, T, T)), as.list(1:5))
leg <- unlist(Map(function(mu, s) sprintf('mu = %d, s = %d', mu, s), mu, s))
legend('topright', legend = leg,  col = 1:5, lty = 'solid')

tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) plogis(x, location = mu, scale = s),
             from = -5, to = 20,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1),
             ylab = 'Probability', las = 1,
             main = 'Funzione di ripartizione (logis(mu, s))')
}, as.list(mu), as.list(s), as.list(c(F, T, T, T, T)), as.list(1:5))
leg <- unlist(Map(function(mu, s) sprintf('mu = %d, s = %d', mu, s), mu, s))
legend('bottomright', legend = leg,  col = 1:5, lty = 'solid')
```

6.2 Uniforme continua

Osservazione 188. 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 189. Una formulazione usuale per tale modello probabilistico è la uniforma continua sull'intervallo con $a = 0, b = 1$.

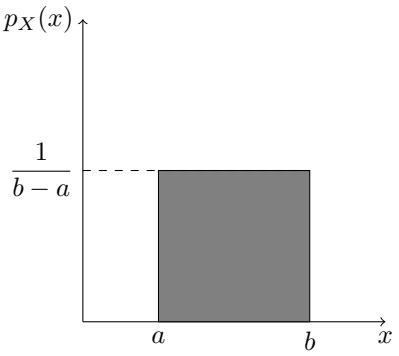


Figure 6.2: Uniforme continua

Osservazione 190 (Supporto e spazio parametrico).

$$R_X = [a, b]$$
$$\mathbf{\Theta} = \{a, b \in \mathbb{R}, a < b\}$$

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.*

Proof.

$$(b-a) \cdot \frac{1}{(b-a)} = 1$$

□

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)**.**

$$\mathbb{E}[X] = \frac{a+b}{2} \tag{6.7}$$
$$\text{Var}[X] = \frac{(b-a)^2}{12} \tag{6.8}$$
$$\text{Asym}(X) = 0 \tag{6.9}$$
$$\text{Kurt}(X) = 1.8 \tag{6.10}$$

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b x\frac{1}{b-a}\,\mathrm{d}x = \left[\frac{x^2}{2(b-a)}\right]_a^b \\ &= \left(\frac{b^2}{2(b-a)}+c\right) - \left(\frac{a^2}{2(b-a)}+c\right) \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}\end{aligned}$$

□

Proof.

$$\begin{aligned}\mathrm{Var}[X] &= \left(\int_a^b x^2\frac{1}{b-a}\,\mathrm{d}x\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \left[\frac{x^3}{3(b-a)}\right]_a^b - \left(\frac{a+b}{2}\right)^2 \\ &= \left(\frac{b^3}{3(b-a)}+c\right) - \left(\frac{a^3}{3(b-a)}+c\right) - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} \\ &= \frac{(b-a)(a^2+b^2+ab)}{3(b-a)} - \frac{(a+b)^2}{4} \\ &= \frac{a^2+b^2+ab}{3} - \frac{a^2+b^2+2ab}{4} \\ &= \frac{4a^2+4b^2+4ab-3a^2-3b^2-6ab}{12} \\ &= \frac{a^2+b^2-2ab}{12} = \frac{(a-b)^2}{12} = \frac{(b-a)^2}{12}\end{aligned}$$

□

Osservazione 191. Si tratta di una variabile simmetrica e platicurtica (ovvero con una distribuzione molto piatta).

6.3 Esponenziale

Osservazione 192. 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 193 (Supporto e spazio parametrico).

$$\begin{aligned}R_X &= \{x \in \mathbb{R} : x > 0\} \\ \boldsymbol{\Theta} &= \{\lambda \in \mathbb{R} : \lambda > 0\}\end{aligned}$$

Definizione 6.3.1 (Distribuzione esponenziale)**.** Se $H(t) = \lambda > 0$ la funzione di ripartizione si ricava dalla 4.26 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 194. 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 195 (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)

```
par(mfrow = c(1,2))
lambda <- c(0.5, 1, 1.5)
tmp <- Map(function(l, cp, add, col) {
    plot_fun(function(x) dexp(x, rate = l), from = 0, to = 5,
             cartesian_plane = cp, add = add, col = col, ylim = c(0, 1.5),
             ylab = 'Density', las = 1, main = 'Densità')
}, as.list(lambda), as.list(c(F, F, F)), as.list(c(F, T, T)), as.list(1:3))
legend('topright', legend = sprintf("lambda = %.1f", lambda),
       col = 1:3, lty = 'solid' )

tmp <- Map(function(l, cp, add, col) {
    plot_fun(function(x) pexp(x, rate = l), from = 0, to = 5,
             cartesian_plane = cp, add = add, col = col, ylim = c(0, 1),
             ylab = 'Density', las = 1, main = 'Ripartizione')
}, as.list(lambda), as.list(c(F, F, F)), as.list(c(F, T, T)), as.list(1:3))
legend('bottomright', legend = sprintf("lambda = %.1f", lambda),
       col = 1:3, lty = 'solid' )
```

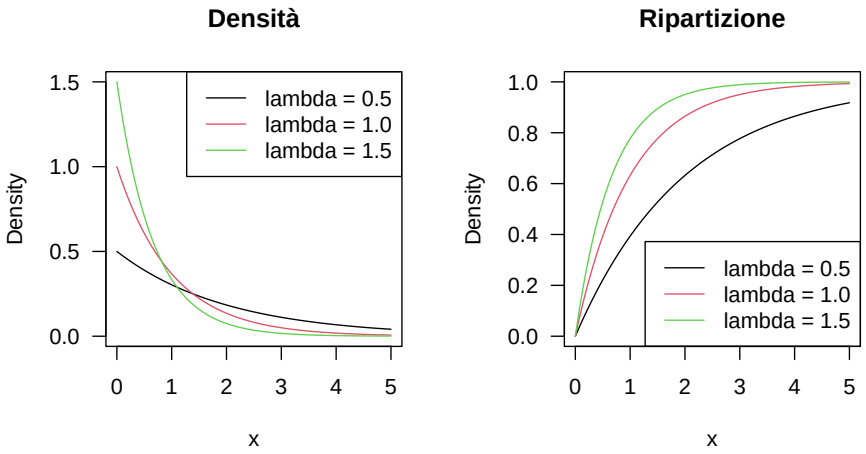


Figure 6.3: Distribuzione esponenziale

Osservazione 196. 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.

6.4 Normale/Gaussiana

Osservazione 197. Viene utilizzata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio.

Osservazione 198. 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 199 (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 200 (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)

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 $Z \sim \mathrm{N}(0, 1)$

Definizione 6.4.3 (Funzione di densità (Normale standardizzata))**.**

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \cdot \mathbb{1}_{R_X}(x) \tag{6.18}$$

```
params <- list(c('mu' = 0, 's2' = 1),
               c('mu' = 2, 's2' = 0.3),
               c('mu' = 4, 's2' = 4))

tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dnorm(x, mean = p["mu"], sd = sqrt(p["s2"])),
             from = -3, to = 10,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.8),
             ylab = 'Density', las = 1, main = 'N(mu, sigma^2)'
             )
}, params, as.list(c(F, T, T)), as.list(1:3))

leg <- unlist(lapply(params, function(x)
    sprintf('mu = %.1f, sigma^2 = %.1f', x['mu'], x['s2'])))
legend('topright', legend = leg,  col = 1:3, lty = 'solid' )
```

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 201. 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 202. 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 203. La conoscenza della funzione di ripartizione della vc $Z \sim \mathrm{N}(0, 1)$ è sufficiente per calcolare la probabilità di qualsiasi vc $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$

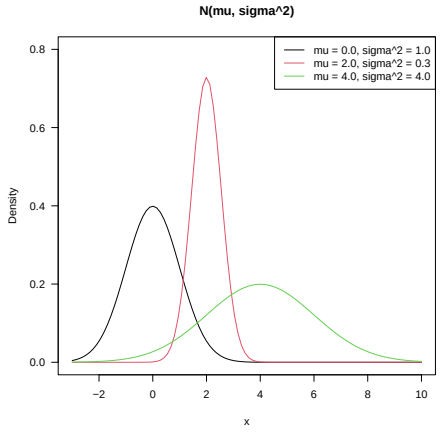


Figure 6.4: Distribuzione normale

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))**.**

$$\mathbb{E}\left[Z\right] = 0 \tag{6.21}$$
$$\mathrm{Var}\left[Z\right] = 1 \tag{6.22}$$
$$\mathrm{Asym}\left(Z\right) = 0 \tag{6.23}$$
$$\mathrm{Kurt}\left(Z\right) = 3 \tag{6.24}$$

Proposizione 6.4.2 (Momenti caratteristici (Normale))**.** *Da* $X = \mu + \sigma Z$ *si ha*

$$\mathbb{E}\left[X\right] = \mu \tag{6.25}$$
$$\mathrm{Var}\left[X\right] = \sigma^2 \tag{6.26}$$
$$\mathrm{Asym}\left(X\right) = 0 \tag{6.27}$$
$$\mathrm{Kurt}\left(X\right) = 3 \tag{6.28}$$

Osservazione 204. 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 205. 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).

Esempio 6.4.1 (Esame vecchio viroli)**.** A random variable X is distributed according to $\mathrm{N}(0,2)$ where 2 is the variance. What is the distribution of $Y = 2X$? Il risultato è $Y \sim \mathrm{N}(0,8)$ (come confermato dal Bigo).

Esempio 6.4.2 (Esame vecchio viroli)**.** A random variable X is distributed according to $\mathrm{N}(-1,1)$. What is the distribution of $Y = -2X + 1$.
Correct answer is $Y \sim \mathrm{N}(3,4)$

Esempio 6.4.3 (Esame vecchio viroli)**.** let $X \sim \mathrm{N}(0,2)$ and $Y \sim \mathrm{N}(1,1)$ be independent random variables where the parameters in the bracket are the expectation and the variace. What is the distribution of $Z = 2X + Y$

1. $Z \sim \mathrm{N}(1,9)$
2. $Z \sim \mathrm{N}(1,5)$
3. not possible to determine
4. $Z \sim \mathrm{N}(1,2)$

should be the first

6.5 Gamma

Osservazione 206. 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 207 (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:

$$f_X(x) = \frac{\lambda^n}{\Gamma(n)} \cdot x^{n-1} \exp(-\lambda x) \cdot \mathbb{1}_{R_X}(x) \tag{6.29}$$

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 208 (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)**.**

$$\mathbb{E}[X] = \frac{n}{\lambda} \tag{6.31}$$

$$\text{Var}[X] = \frac{n}{\lambda^2} \tag{6.32}$$

$$\text{Asym}(X) = \frac{2}{\sqrt{n}} \tag{6.33}$$

$$\text{Kurt}(X) = 3 + \frac{6}{n} \tag{6.34}$$

Osservazione 209 (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)

```
params1 <- list(c('n' = 1, 'lambda' = 1),
                c('n' = 2, 'lambda' = 1),
                c('n' = 3, 'lambda' = 1))

params2 <- list(c('n' = 2, 'lambda' = 1),
                c('n' = 2, 'lambda' = 2),
                c('n' = 2, 'lambda' = 3))

par(mfrow = c(1,2))
tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dgamma(x, shape = p["n"], rate = p['lambda']),
             from = 0, to = 6,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1),
             ylab = 'Density', las = 1, main = 'Gamma(n, 1)')
}, params1, as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(lapply(params1, function(x)
    sprintf('n = %d', x['n'])))
legend('topright', legend = leg,  col = 1:3, lty = 'solid' )

tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dgamma(x, shape = p["n"], rate = p['lambda']),
             from = 0, to = 6,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1.2),
             ylab = 'Density', las = 1, main = 'Gamma(2, lambda)')
}, params2, as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(lapply(params2, function(x)
    sprintf('lambda = %d', x['lambda'])))
legend('topright', legend = leg,  col = 1:3, lty = 'solid' )
```

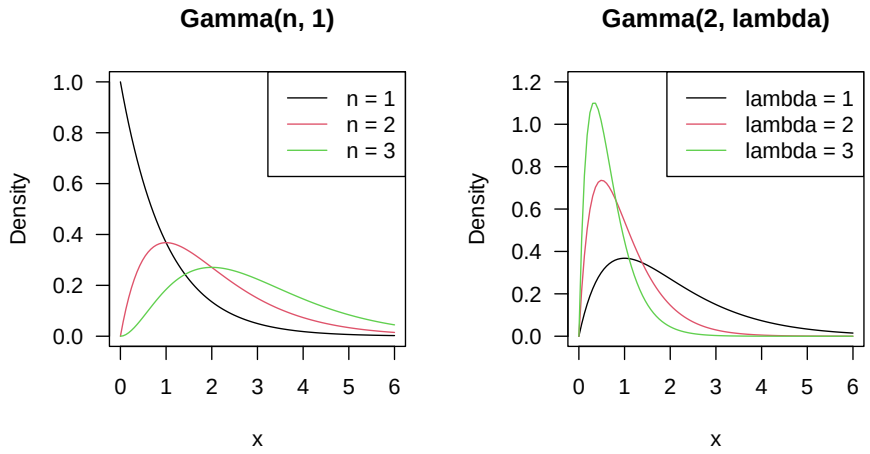


Figure 6.5: Distribuzione gamma

Osservazione 210. Si nota che se $n = 1$, la distribuzione gamma diviene una esponenziale, ovvero $\text{Gamma}(1, \lambda) \sim \text{Exp}(\lambda)$; pertanto la gamma è una generalizzazione della esponenziale.

Osservazione 211. 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:*

$$\sum \text{Gamma}(n_i, \lambda) \sim \text{Gamma}\left(\sum_i n_i, \lambda\right) \tag{6.35}$$

6.6 Chi-quadrato

Osservazione 212. 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

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu \tag{6.36}$$

Osservazione 213 (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 $x > 0$

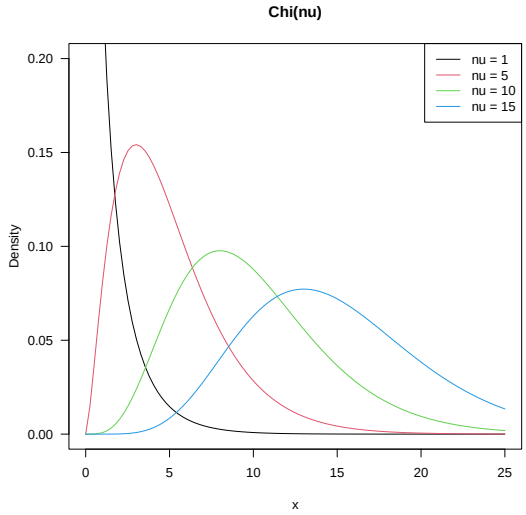


Figure 6.6: Distribuzione χ^2

Osservazione 214. Anche se ν può esser qualsiasi numero reale positivo, in pratica le applicazioni hanno tipicamente ν intero positivo.

Osservazione 215 (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)

```
nu <- c(1, 5, 10, 15)
tmp <- Map(function(p, add, col) {
    ## browser()
    plot_fun(function(x) dchisq(x, df = p),
             from = 0, to = 25,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.2),
             ylab = 'Density', las = 1, main = 'Chi(nu)')
}, nu, as.list(c(F, T, T, T)), as.list(1:4))
leg <- unlist(lapply(nu, function(x) sprintf('nu = %d', x)))
legend('topright', legend = leg,  col = 1:4, lty = 'solid')
```

Proposizione 6.6.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = \nu \tag{6.38}$$

$$\mathrm{Var}[X] = 2\nu \tag{6.39}$$

$$\mathrm{Asym}(X) = \sqrt{\frac{8}{\nu}} \tag{6.40}$$

$$\mathrm{Kurt}(X) = 3 + \frac{12}{\nu} \tag{6.41}$$

Proposizione 6.6.2. *Anche la distribuzione Chi-quadrato gode della proprietà riproduttiva:*

$$\sum_{i=1}^{n} \chi^2_{\nu_i} \sim \chi^2_{\sum_i \nu_i}$$

6.7 Beta

Osservazione 216. Viene utilizzata quando si vogliono definire a priori i valori possibili delle probabilità di successo per variabili Bernoulliane.

Osservazione 217 (Supporto e spazio parametrico).

$$R_X = [0,1]$$
$$\mathbf{\Theta} = \{\alpha, \beta \in \mathbb{R} : \alpha, \beta > 0\}$$

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 218. Una vc Beta è definita nell'intervallo $[0,1]$, ma effettuando la trasformazione $Y = X(b-a) + a$, la si può ricondurre all'intervallo $[a,b]$.

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 219 (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.

```
p <- c(0.3, 0.7, 1, 4)
alphas <- p
betas <- p

par(mfrow = c(1,2))
tmp <- Map(function(a, b, add, col) {
    ## browser()
    plot_fun(function(x) dbeta(x, shape1 = a, shape2 = b),
             from = 0, to = 1,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 3.5),
             ylab = 'Density', las = 1, main = 'Beta(alpha, beta)')
}, as.list(alphas), as.list(betas), as.list(c(F, T, T, T)), as.list(1:4))
leg <- unlist(lapply(p, function(x) sprintf('alpha = %.1f, beta = %.1f', x, x)))
legend('top', legend = leg,  col = 1:4, lty = 'solid')

alphas <- c(2, 6, 0.1, 2)
betas  <- c(6, 2, 2  , 0.1)
tmp <- Map(function(a, b, add, col) {
    ## browser()
    plot_fun(function(x) dbeta(x, shape1 = a, shape2 = b),
             from = 0, to = 1,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 4),
             ylab = 'Density', las = 1, main = 'Beta(alpha, beta)')
}, as.list(alphas), as.list(betas), as.list(c(F, T, T,T)), as.list(1:4))
leg <- unlist(Map(function(a, b) sprintf('alpha = %.1f, beta = %.1f', a, b),
                  as.list(alphas), as.list(betas)))
legend('top', legend = leg,  col = 1:4, lty = 'solid')
```

6.8 T di Student

Osservazione 220. Il suo uso è prettamente teorico, in quanto è la risultante di una trasformazione su due variabili, una normale e una chi quadrato.

Osservazione 221 (Supporto e spazio parametrico).

$$\begin{aligned} R_X &= \mathbb{R} \\ \boldsymbol{\Theta} &= \{g \in \mathbb{N} \setminus \{0\}\} \end{aligned}$$

Definizione 6.8.1 (Distribuzione T)**.** Se $Z \sim \mathrm{N}(0,1)$ ed C è una distribuzione indipendente tale che $C \sim \chi^2_g$ allora si definisce vc di Student la seguente X:

$$X = \frac{Z}{\sqrt{C/g}} \sim \mathrm{T}(g) \tag{6.46}$$

Definizione 6.8.2 (Funzione di densità)**.**

$$f_X(x) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)\sqrt{\pi g}}\left(1 + \frac{x^2}{g}\right)^{-\frac{g+1}{2}} \cdot \mathbb{1}_{R_X}(x) \tag{6.47}$$

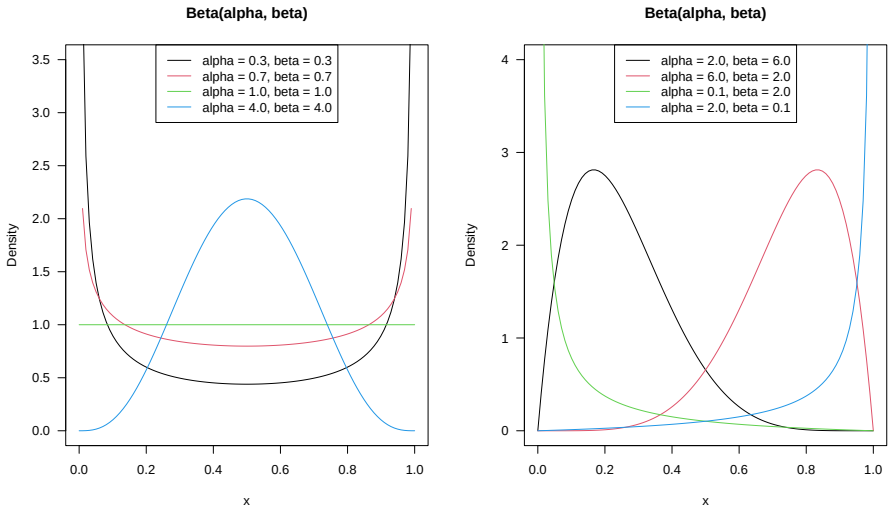


Figure 6.7: Distribuzione beta

Proposizione 6.8.1 (Momenti caratteristici)**.**

$$\mathbb{E}[X] = 0 \quad se\ g > 1$$
$$\text{Var}[X] = \frac{g}{g-2} \quad seg > 2$$
$$\text{Kurt}(X) = 3 + \frac{6}{g-4} \quad se\ g > 4$$

Osservazione 222 (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)

```
g <- c(1, 10, 40, NA)
tmp <- Map(function(g, add, col) {
    plot_fun(function(x) if (!is.na(g))  dt(x, df = g)
                         else dnorm(x),
             from = -4, to = 4,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 0.4),
             ylab = 'Density', las = 1, main = 'T(g)')
}, as.list(g), as.list(c(FALSE, TRUE, TRUE, TRUE)), as.list(1:4))
leg <- unlist(lapply(g, function(x)
    if (!is.na(x)) sprintf('T(%d)', x)
    else 'N(0, 1)'))
legend('topright', legend = leg,  col = 1:4, lty = 'solid')
```

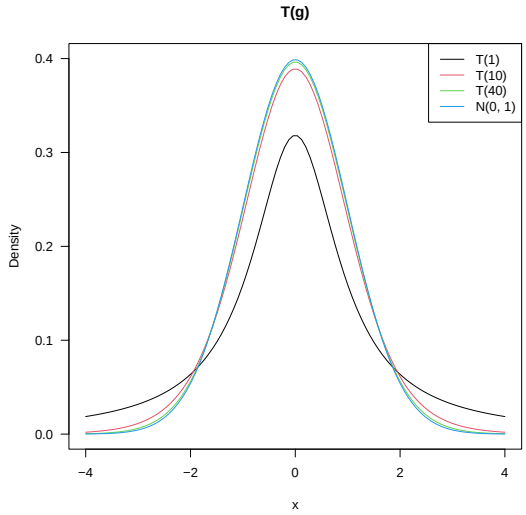


Figure 6.8: Distribuzione t

6.9 F di Fisher

Osservazione 223. 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 224. Se $X_1 \sim \chi^2_{g_1}$ e $X_2 \sim \chi^2_{g_2}$, allora

$$X = \frac{X_1/g_1}{X_2/g_2} \sim \mathrm{F}\left(g_1, g_2\right) \tag{6.48}$$

ovvero X si distribuisce come una F con g_1 e g_2 gradi di libertà.

Osservazione 225 (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 226 (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)**.**

$$\mathbb{E}\left[X\right] = \frac{g_2}{g_2-2} \quad \textit{se } g_2 > 2$$
$$\mathrm{Var}\left[X\right] = \frac{2g_2^2(g_1+g_2-2)}{g_1(g_2-2)^2(g_2-4)} \quad \textit{se } g_2 > 4$$

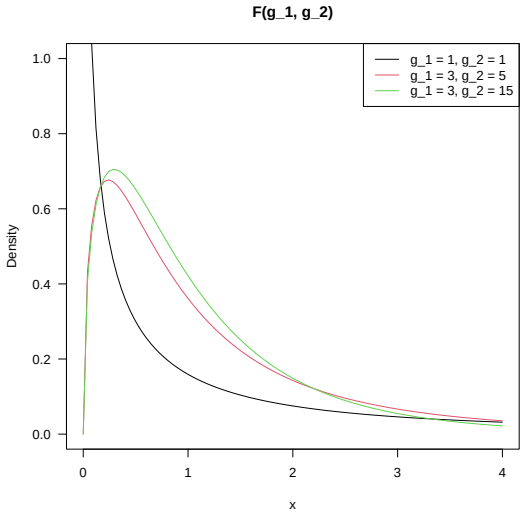


Figure 6.9: Distribuzione F

```
g1 <- c(1, 3,  3)
g2 <- c(1, 5, 15)
tmp <- Map(function(g1, g2, add, col) {
    plot_fun(function(x) df(x, df1 = g1, df2 = g2),
             from = 0, to = 4,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 1),
             ylab = 'Density', las = 1, main = 'F(g_1, g_2)')
}, as.list(g1), as.list(g2), as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(Map(function(g1, g2) sprintf('g_1 = %d, g_2 = %d', g1, g2),
                  g1, g2))
legend('topright', legend = leg,  col = 1:3, lty = 'solid')
```

Osservazione 227 (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 $g_2 \to \infty$.

6.10 Lognormale

Osservazione 228. Viene utilizzata quando la grandezza oggetto di studio è il risultato del prodotto di n fattori indipendenti

Osservazione 229 (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 230. Si ha che se $X \sim \mathrm{LogN}(mu, \sigma)$ allora $\log X \sim \mathrm{N}(\mu, \sigma^2)$, mentre se $Y \sim \mathrm{N}(\mu, \sigma^2)$, $e^Y \sim \mathrm{LogN}(\mu, \sigma^2)$

Osservazione 231 (Forma della distribuzione). Con μ fisso all'aumentare di σ l'asimmetria si incrementa (figura 6.10)

```
mu <- rep(0, 6)
s <- c(0.125, 0.25, 0.5, 1, 1.5, 10)
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) dlnorm(x, meanlog = mu, sdlog = s),
             from = 0, to = 3,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 3),
             ylab = 'Density', las = 1,
             main = 'logN(mu, s)')
}, as.list(mu), as.list(s), as.list(c(F, T, T, T, T, T)), as.list(1:6))
leg <- unlist(Map(function(mu, s)
    sprintf('mu = %d, s = %.2f', mu, s), mu, s))
legend('topright', legend = leg,  col = 1:6, lty = 'solid')
```

6.11 Weibull

Osservazione 232. Viene utilizzata per studiare l'affidabilità dei sistemi di produzione nei processi industriali, in particolare per valutare i tassi di rottura

Osservazione 233. La Weibull presenta la caratteristica di avere una funzione di rischio variabile in funzione di un ulteriore parametro a: se la vc $(X/b)^a \sim \mathrm{Exp}(1)$, allora diremo che la vc continua X, definita sula semiretta positiva è una vc di Weibull con parametri $a > 0, b > 0$.

Osservazione 234 (Supporto e spazio parametrico).

$$R_X = \{x \in \mathbb{R} : x > 0\}$$
$$\boldsymbol{\Theta} = \{a, b \in \mathbb{R} : a, b > 0\}$$

Osservazione 235 (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

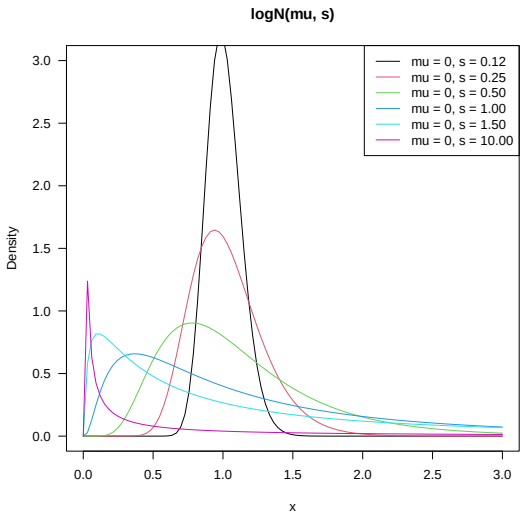


Figure 6.10: Distribuzione lognormale

- 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 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)**.**

$$\mathbb{E}[X] = \frac{\Gamma\left(1+\frac{1}{b}\right)}{a^{1/b}}$$
$$\text{Var}[X] = \frac{\Gamma\left(1+\frac{2}{b}\right) - \Gamma^2\left(1+\frac{1}{b}\right)}{a^{2/b}}$$

```
b <- c(2, 2, 3, 4)
a <- c(0.5, 1, 1.5, 3)
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) dweibull(x, scale = mu, shape = s),
             from = 0, to = 5,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 2),
             ylab = 'Density', las = 1,
             main = 'Wei(a, b)')
}, as.list(a), as.list(b), as.list(c(F,T, T, T)), as.list(1:4))
leg <- unlist(Map(function(mu, s)
```

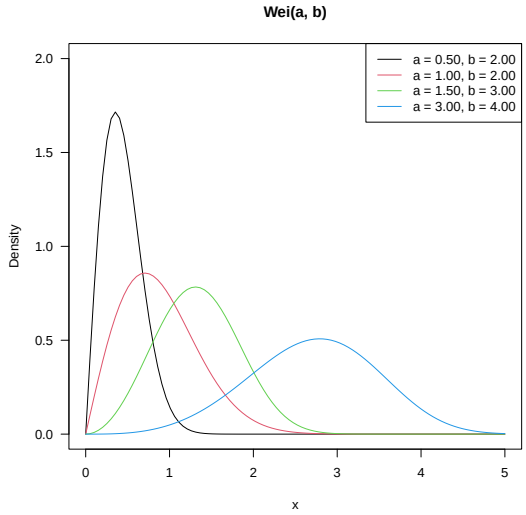


Figure 6.11: Distribuzione Weibull

```
    sprintf('a = %.2f, b = %.2f', mu, s), a, b))
legend('topright', legend = leg,  col = 1:4, lty = 'solid')
```

6.12 Pareto

Osservazione 236. Viene utilizzata quando si studiano distribuzioni di variabili che hanno un minimo (ad esempio come, con x_m = reddito minimo)

Osservazione 237 (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)**.**

$$\mathbb{E}[X] = \frac{kx_m}{k-1} \qquad per\ k > 1$$
$$\mathrm{Var}[X] = \left(\frac{x_m}{k-1}\right)^2 \frac{k}{k-2} \qquad per\ k > 2$$

Osservazione 238 (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.

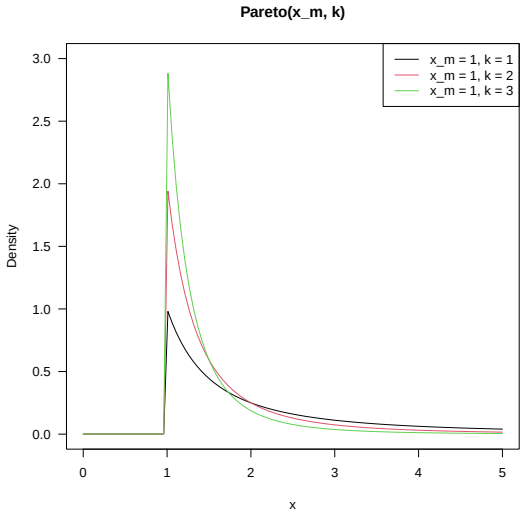


Figure 6.12: Distribuzione di Pareto

```
mu <- 1
k <- 1:3
tmp <- Map(function(mu, s, add, col) {
    plot_fun(function(x) VGAM::dpareto(x, scale = mu, shape = s),
             from = 0, to = 5,
             cartesian_plane = FALSE,
             add = add, col = col, ylim = c(0, 3),
             ylab = 'Density', las = 1,
             main = 'Pareto(x_m, k)')
}, as.list(mu), as.list(k), as.list(c(F, T, T)), as.list(1:3))
leg <- unlist(Map(function(mu, s)
    sprintf('x_m = %d, k = %d', mu, s), mu, k))
legend('topright', legend = leg,  col = 1:3, lty = 'solid')
```

Chapter 7

Misc topics

7.1 Characteristic and moment generating function

7.1.1 Characteristic function

Definizione 7.1.1 (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 7.1.1 (Characteristic function of a binomial)**.** Let $X \sim \mathrm{Bin}(n, p)$, $D_x = \{0, 1, \dots, n\}$, the characteristic function is

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right] = \sum_{x=0}^{n} e^{itx} \cdot \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (\underbrace{pe^{it}}_{a})^x (\underbrace{1-p}_{b})^{n-x}$$

$$\overset{(1)}{=} (1 - p + pe^{it})^n$$

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 theorical 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 7.1.1 (Link with distribution)**.** *Supposing we have two random variables/vectors X, Y; then*

$$X \sim Y \iff X \text{ and } Y \text{ have the same characteristic function} \tag{7.1}$$

Proof. Rigo non l'ha fatta. □

Proposizione 7.1.2 (Link with the moments)**.** *We have:*

$$\left[\frac{\partial^k}{\partial t^k}\phi_X(t)\right]_{t=0} = i^k \,\mathbb{E}\left[X^k\right]$$

and therefore

$$\mathbb{E}\left[X^k\right] = \frac{\left[\frac{\partial^k}{\partial t^k}\phi_X(t)\right]_{t=0}}{i^k}$$

Proof. We have

$$\frac{\partial^k}{\partial t^k}\phi_X(t) = \frac{\partial^k}{\partial t^k}\,\mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\frac{\partial^k}{\partial t^k}e^{itX}\right] = \mathbb{E}\left[i^k X^k e^{itx}\right]_{t=0} \overset{(1)}{=} i^k\,\mathbb{E}\left[X^k\right]$$

where in (1) we evaluated for $t = 0$. □

Proposizione 7.1.3 (Link with density (inversion formula))**.**

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\phi_X(t)\,\mathrm{d}t$$

Proof. Virols skips it. □

Proposizione 7.1.4 (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 7.1.2. 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 law* 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 $\mathbb{E}[|X_i|] = +\infty$.

7.1.2 Moment generating function

Definizione 7.1.2 (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:

$$\phi_X(-it) = \mathbb{E}\left[e^{-iitX}\right] = \mathbb{E}\left[e^{tX}\right] = M_X(t), \quad \forall t \in \mathbb{R}$$

so

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \quad \forall t \in \mathbb{R}$$

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 7.1.5 (Properties)**.**

$$\left[\frac{\partial^k}{\partial t^k} M_X(t)\right]_{t=0} = \mathbb{E}\left[X^k\right] \tag{7.2}$$

$$M_X(0) = \mathbb{E}\left[e^{0X}\right] = \mathbb{E}\left[1\right] = 1 \tag{7.3}$$

$$M_X(t) = M_Y(t), \forall t \iff F_X(x) = F_Y(y) \qquad \textit{(uniqueness)} \tag{7.4}$$

$$M_{aX+b}(t) = e^{tb} M_X(at), \quad a, b \in \mathbb{R} \tag{7.5}$$

$$X \perp\!\!\!\perp Y \implies M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \tag{7.6}$$

TODO: l'implicazione per l'indipendenza è anche coimplicazione?

Proof. For 7.5

$$M_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = \mathbb{E}\left[e^{taX} \cdot \underbrace{e^{tb}}_{\text{constant}}\right] = e^{tb} \cdot \mathbb{E}\left[e^{taX}\right] = e^{tb} M_X(at)$$

For 7.6

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right]$$

Now note that:

TODO: questo andrebbe portato più in vista nella 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

$$M_{X+Y}(t) = \mathbb{E}\left[e^{tX}e^{tY}\right] \overset{(1)}{=} \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_y(t)$$

in (1) because of $\perp\!\!\!\perp$ □

Esempio 7.1.3 (Mgf of bernoulli and binomial)**.** If $X \sim \mathrm{Bern}\,(p)$, $p(x) = p^x(1-p)^{1-x}$, $D_x = \{0,1\}$. Its mgf is:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = e^{t\cdot 0}\cdot(1-p)p^0 0 + e^{t\cdot 1}p^1(1-p)^0 = 1-p+pe^t$$

Being the binomial $Y = X_1 + \ldots + X_n$ with X_i iid, by properties of mgfs, the mgf of a binomial is

$$M_Y(t) = \prod_{i=1}^{n}(1-p+pe^t) = (1-p+pe^t)^n$$

Esempio 7.1.4 (Mgf of poisson)**.** Let $X \sim \mathrm{Pois}\,(\lambda)$, let's determine $M_X(t)$

$$\begin{aligned}M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \sum_{x=0}^{\infty} e^{tx}\frac{1}{x!}e^{-\lambda}\lambda^x = \sum_{x=0}^{\infty}(e^t\lambda)^x\frac{1}{x!}e^{-\lambda} \\ &\overset{(1)}{=} e^{-\lambda}\cdot e^{\lambda e^t} = e^{-\lambda(1-e^t)} = e^{\lambda(e^t-1)}\end{aligned}$$

where in (1) we used $\sum_{x=0}^{\infty}\frac{c^x}{x!} = e^c$.

Esempio 7.1.5 (Esercizio richiesto Viroli)**.** By using 7.6 find $M_Y(t)$, with $Y = \sum_{i=1}^{n} X_i$, $X_i \sim \text{Pois}(\lambda_i)$, and $X_i \perp\!\!\!\perp X_j$.
La mgf di una poisson con parametro λ è $M_X(t) = e^{\lambda(e^t-1)}$, da cui per l'indipendenza possiamo applica la produttoria

$$M_Y(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{\sum_{i=1}^{n}\lambda_i(e^t-1)} = e^{(e^t-1)\cdot\sum_{i=1}^{n}\lambda_i}$$

che è la mgf di una poisson con parametro lambda la somma delle lambda componenti (come atteso).
Therefore $\implies Y \sim \text{Pois}(\sum_{i=1}^{n}\lambda_i)$.

Esempio 7.1.6 (Esame vecchio viroli)**.** Let X be a bernoulli rv with parameter $\frac{1}{2}$. Find the moment generating functions of $Y = \frac{1}{2} + \frac{X}{2}$.
We have that for the bernoulli

$$M_X(t) = 1 - p + pe^t$$

and consider

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

Now here we have $Y = \frac{1}{2} + \frac{X}{2}$ so $a = b = \frac{1}{2}$, therefore:

$$M_Y(t) = e^{t/2}M_X\left(\frac{t}{2}\right) = e^{\frac{t}{2}}(1 - p + pe^{t/2})$$

Finally, if $p = \frac{1}{2}$

$$M_Y(t) = e^{t/2}\left(\frac{1}{2} + \frac{e^{t/2}}{2}\right) = \frac{1}{2}\left(e^{t/2} + e^t\right)$$

Therefore we have that $M_Y(t) = \frac{1}{2}(e^t + e^{\frac{t}{2}})$

Esempio 7.1.7 (Esame vecchio viroli)**.** Let X_1 and X_2 be two independent Bernoulli rv with parameters 1/2. find the moment generating function of $Z = X_1 - X_2$.
If $X \sim \text{Bern}(p)$, its $M_X(t) = (1 - p + pe^t)$. Here for the difference of two bernoulli we apply the mgf properties

$$M_{X_1-X_2}(t) = M_{X_1+(-X_2)}(t) \overset{(1)}{=} M_{X_1}(t) \cdot M_{-X_2}(t) \overset{(2)}{=} M_{X_1}(t) + M_{X_2}(-t)$$

with 1 for independence and 2 for linear transformation properties. So considering both as bernoulli with $p = 1/2$

$$\begin{aligned} M_{X_1-X_2}(t) &= (1 - p + pe^t)(1 - p + pe^{-t}) = \left(\frac{1}{2} + \frac{1}{2}e^t\right)\left(\frac{1}{2} + \frac{1}{2}e^{-t}\right) \\ &= \frac{1}{4} + \frac{1}{4}e^{-t} + \frac{1}{4}e^t + \frac{1}{4} = \frac{1}{2} + \frac{1}{4}\left(e^t + e^{-t}\right) \end{aligned}$$

so $M_{X_1-X_2}(t) = 1/2 + 1/4(e^t + e^{-t})$. And Bigo confirms.

Osservazione 239. The following is a result that become useful sometimes (eg clt)

Proposizione 7.1.6 (Mc Laurin expansion of mgf)**.**

$$M_X(t) = 1 + t\,\mathbb{E}[X] + \frac{t^2}{2!}\,\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\,\mathbb{E}\left[X^3\right] + \dots \tag{7.7}$$

Proof. 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

$$\begin{aligned}
M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_{D_X} e^{tx} f(x)\,\mathrm{d}x \\
&= \underbrace{\int_{D_X} 1 f(x)\,\mathrm{d}x}_{=1} + \int_{D_X} tx f(x)\,\mathrm{d}x + \int_{D_X} \frac{(tx)^2}{2!} f(x)\,\mathrm{d}x + \int_{D_X} \frac{(tx)^3}{3!} f(x)\,\mathrm{d}x + \dots \\
&= 1 + t\int_{D_X} x f(x)\,\mathrm{d}x + \frac{t^2}{2!}\int_{D_X} x^2 f(x)\,\mathrm{d}x + \frac{t^2}{3!}\int_{D_X} x^3 f(x)\,\mathrm{d}x + \dots \\
&= 1 + t\,\mathbb{E}[X] + \frac{t^2}{2!}\,\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\,\mathbb{E}\left[X^3\right] + \dots
\end{aligned}$$

□

Osservazione 240. Now we see an example where mgf does not always exists

Esempio 7.1.8 (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

$$\begin{aligned}
&\Gamma(x) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\,\mathrm{d}x, \quad \forall \alpha > 0 \\
&\alpha \in \mathbb{N} \implies \Gamma(\alpha) = (\alpha-1)!
\end{aligned}$$

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 7.1.9 (Esercizio richiesto Viroli)**.** For this exercise:

1. compute the second moment $\mathbb{E}\left[X^2\right]$ of the binomial distribution using the second derivative of mgf evaluated in 0;
2. for the binomial, verify property 2 of mgf, that is $M_X(0) = 1$;
3. eval $\mathbb{E}[X]$ where X is Gamma by using first derivative of mgf

We have

1. per la prima deriviamo due volte e valutiamo in 0 la mgf della binomiale che è $(1 - p + pe^t)^n$. Si ha

   $$\left[(1 - p + pe^t)^n\right]' = n(1 - p + pe^t)^{n-1}(pe^t)$$
   $$\left[(1 - p + pe^t)^n\right]'' = n\left[(n-1)(1 - p + pe^t)^{n-2}(pe^t)^2 + (pe^t)(1 - p + pe^t)^{n-1}\right]$$

   che valutata per $t = 0$ da

   $$n(n-1)p^2 + np = n^2p^2 - np^2 + np$$

   Possiamo verificare il risultato applicando la formula di calcolo della varianza (dato che della binomiale si conoscono varianza e valore atteso)

   $$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \\ np(1-p) &= \mathbb{E}\left[X^2\right] - n^2p^2 \\ \mathbb{E}\left[X^2\right] &= np(1-p) + n^2p^2 = np - np^2 + n^2p^2\end{aligned}$$

2. per $t = 0$, si ha $(1 - p + pe^t)^n = (1 - p + p)^n = 1$
3. la mgf della gamma è $\left(\frac{\lambda}{\lambda - t}\right)^\alpha$ la sua derivata prima

   $$\alpha\left(\frac{\lambda}{\lambda - t}\right)^{\alpha-1}\left(-\frac{\lambda(-1)}{(\lambda - t)^2}\right) = \alpha\left(\frac{\lambda}{\lambda - t}\right)^{\alpha-1}\left(\frac{\lambda}{(\lambda - t)^2}\right)$$

   che valutata in $t = 0$ da α/λ, il valore atteso della gamma

Esempio 7.1.10 (Normal distributions)**.** Let $X \sim \mathrm{N}(0,1)$, then lets derive the mgf

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_{-\infty}^{+\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, \mathrm{d}x$$
$$= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{tx - \frac{1}{2}x^2}\, \mathrm{d}x$$

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 7.1.11 (Esercizio richiesto Viroli)**.** Regarding the normal (consider $X \sim \mathrm{N}(0,1)$):

- prove $\frac{\partial M_{\sigma X+\mu}(t)}{\partial t} = \mu$
- derive $\mathbb{E}\left[X^2\right]$ by mgf
- check that $\mathrm{Var}\left[\sigma X + \mu\right] = \sigma^2$ (applying $\mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$)

If the mgf of the general normal is $e^{\mu t} e^{\frac{\sigma^2 t^2}{2}}$

1. we derive it one time and evaluate for $t = 0$ to find μ

$$e^{\mu t} \cdot \mu \cdot e^{\frac{1}{2}\sigma^2 t^2} + e^{\frac{1}{2}\sigma^2 t^2} \cdot \left(\frac{1}{2}\sigma^2 2t\right) \cdot e^{\mu t} = e^{\mu t + \frac{1}{2}\sigma^2 t^2}\left(\mu + \frac{1}{2}\sigma^2 2t\right)$$

   che valutata per $t = 0$ restituisce $e^0(\mu + 0) = 1 \cdot \mu = \mu$

2. la derivata seconda è

$$e^{\mu t+\frac{1}{2}\sigma^2 t^2}\left(\mu+\frac{1}{2}\sigma^2 2t\right)^2+\left(\frac{1}{2}\sigma^2 2\right)e^{\mu t+\frac{1}{2}\sigma^2 t^2}$$
$$e^{\mu t+\frac{1}{2}\sigma^2 t^2}\left[\left(\mu+\frac{1}{2}\sigma^2 2t\right)^2+\sigma^2\right]$$

se $t=0$

$$e^0\left[(\mu+0)^2+\sigma^2\right]=\mu^2+\sigma^2$$

3. abbiamo

$$\mathrm{Var}\left[Y\right]=\mathbb{E}\left[Y^2\right]-\mathbb{E}\left[Y\right]^2=\mu^2+\sigma^2-\mu^2=\sigma^2$$

Esempio 7.1.12 (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 $\mathbb{E}\left[X_1\right]$ by $M_{\mathbf{X}}(t)$
3. compute $\mathbb{E}\left[X_1\right]$ by definition
4. are $X_1 \perp\!\!\!\perp X_2$, both by density and by moment generating function

We have:

1.

$$\begin{aligned}M_{\mathbf{X}}(t)&=2\int_0^{+\infty}\int_{x_2}^{\infty}e^{tx_1}e^{tx_2}e^{-(x_1+x_2)}\,\mathrm{d}x_1\,\mathrm{d}x_2\\&=2\int_0^{+\infty}e^{-x_2(1-t_2)}\cdot\int_{x_2}^{+\infty}\mathrm{d}x_1\,\mathrm{d}x_2\\&=2\int_0^{\infty}e^{-x_2(1-t_2)}\cdot\left[-\frac{e^{x_1(1-t_1)}}{1-t_1}\right]_{x_2}^{\infty}\mathrm{d}x_2\\&=2\frac{1}{1-t_1}\int_0^{+\infty}e^{-x_2(2-t_1-t_2)}\,\mathrm{d}x_2\\&=\frac{2}{(1-t_1)(2-t_1-t_2)}\end{aligned}$$

2.

$$\begin{aligned}\frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial t_1}\Big|_{\mathbf{t=0}}&=2(1-t_1)^{-2}(2-t_1-t_2)^{-1}+2(1-t_1)^{-1}(2-t_1-t_2)^{-2}\Big|_{\mathbf{t=0}}\\&=\frac{2}{2}+\frac{2}{4}=\frac{3}{2}\end{aligned}$$

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

$$\begin{aligned} \mathbb{E}[X_1] &= \int_0^{+\infty} x_1\left(2e^{-x_1} - 2e^{-2x_1}\right)\mathrm{d}x_1 \\ &= 2\underbrace{\int_0^{+\infty} x_1 e^{-x_1}\,\mathrm{d}x_1}_{\text{expected value of Exp}(1)} - \underbrace{\int_0^{+\infty} x_1 2e^{-2x_1}\,\mathrm{d}x_1}_{\text{expected value of Exp}(2)} \\ &= 2\cdot 1 - \frac{1}{2} = \frac{3}{2} \end{aligned}$$

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:

$$M_{X_1}(t_1) = M_{\mathbf{X}}(t_1, 0) = \frac{2}{(1-t_2)\frac{1}{2-t_1}} \qquad M_{X_2}(t_2) = M_{\mathbf{X}}(0, t_2) = \frac{2}{2-t_2}$$

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 7.1.13 (Mgf of Geometric and Negative binomial)**.** Let $X_1, \dots, X_n \sim$ 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

7.2 Order statistics

Osservazione 241. 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 242. Important special cases of the order statistics are the minimum $X_{(1)}$, the maximum $X_{(n)}$, the 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 $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$. We are interested in studying properties of these newly defined rvs.

Esempio 7.2.1. 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

$$\mathbb{P}\left(X_{(1)} = 6\right) = \frac{1}{6} \cdot \ldots \cdot \frac{1}{6} = \left(\frac{1}{6}\right)^6$$
$$\mathbb{P}\left(X_{(1)} = 1\right) = 1 - \left(\frac{5}{6}\right)^6$$

7.2.1 Minimum

Proposizione 7.2.1 (Distribution function)**.** *We have that*

$$F_{(1)}(x) = 1 - [1 - F_X(x)]^n \tag{7.8}$$

Proof.

$$\begin{aligned} F_{(1)}(x) &= \mathbb{P}\left(X_{(1)} \leq x\right) = 1 - \mathbb{P}\left(X_{(1)} > x\right) \\ &= 1 - \mathbb{P}\left(X_1 > x, X_2 > x, \dots, X_n > x\right) \overset{(1)}{=} 1 - \prod_{i=1} \mathbb{P}\left(X_i > x\right) \\ &\overset{(2)}{=} 1 - \prod_{i=1} \mathbb{P}\left(X > x\right) = 1 - \left[\mathbb{P}\left(X > x\right)\right]^n = 1 - \left[1 - \mathbb{P}\left(X \leq x\right)\right]^n \\ &= 1 - \left[1 - F_X(x)\right]^n \end{aligned}$$

with (1) we considered independent rvs and (2) identically distributed. □

Osservazione 243. Interpretazione affinché il minimo sia al più x si fa il complemento in cui si guarda la probabilità che siano tutte contemporaneamente $> x$

Proposizione 7.2.2 (Density function)**.**

$$f_{(1)}(x) = n f_X(x) \cdot [1 - F_X(x)]^{n-1}$$

Proof.

$$\begin{aligned} f_1(x) = \frac{\partial F_{(1)}(x)}{\partial x} &= -n\left[1 - F_X(x)\right]^{n-1}\left(-f_X(x)\right) \\ &= n f_X(x) \cdot \left[1 - F_X(x)\right]^{n-1} \end{aligned}$$

□

Esempio 7.2.2. A room is lit by 5 light bulbs, each bulb lifetime has a 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

$$F_{(1)}(200) = 1 - \left(1 - F_X(200)\right)^5 = 1 - \left(1 - 1 + e^{-200/100}\right)^5 = 1 - \frac{1}{e^{10}}$$

Therefore

$$\mathbb{P}\left(X_{(1)} > 200\right) = 1 - 1 + \frac{1}{e^{10}} = \frac{1}{e^{10}}$$

Esempio 7.2.3 (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

7.2.2 Maximum

Proposizione 7.2.3 (Distribution function)**.**

$$F_{(n)}(x) = [F_X(x)]^n \tag{7.9}$$

Proof.

$$\begin{aligned} F_{(n)}(x) &= \mathbb{P}\left(X_{(n)} \le x\right) = \mathbb{P}\left(X_1 \le x, \dots, X_N \le x\right) \\ &\overset{(iid)}{=} \left[\mathbb{P}\left(X \le x\right)\right]^n = \left[F_X(x)\right]^n \end{aligned}$$

□

Osservazione 244. Il massimo sia $\le x$ se tutte le vc sono $\le x$

Proposizione 7.2.4 (Density function)**.**

$$f_{(n)}(x) = n\left[F_X(x)\right]^{n-1} f_X(x) \tag{7.10}$$

Proof.

$$f_{(n)}(x) = \frac{\partial}{\partial x} F_{(n)}(x) = n\left[F_X(x)\right]^{n-1} f_X(x)$$

□

Esempio 7.2.4. Considering again a room lit by 5 light bulbs, each bulb lifetime has a 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

$$\begin{aligned} \mathbb{P}\left(X_{(n)} > 200\right) &= 1 - \mathbb{P}\left(X_{(n)} \le 200\right) = 1 - F_{(n)}(200) \\ &= 1 - \left[F_X(200)\right]^5 = 1 - \left(1 - e^{-2}\right)^5 \simeq 0.52 \end{aligned}$$

Esempio 7.2.5. 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

$$\mathbb{P}\left(X_{(n)} > 0.9\right) = 1 - \mathbb{P}\left(X_{(n)} \le 0.9\right) = 1 - \left[F_X(0.9)\right]^{12} = 1 - 0.9^{12} = 0.718$$

Esempio 7.2.6 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x,\theta) = \frac{3x^2}{\theta^3}$$

with $X \in [0,\theta]$. Compute the cumulative distribution function of the maximum $X_{(n)}$.
Per ottenerla occorre sviluppare la cumulata della funzione di partenza

$$F_X(x) = \int \frac{3x^2}{\theta^3} = \frac{3}{\theta^3}\int x^2 = \frac{3}{\theta}\frac{x^3}{3} = \frac{x^3}{\theta^3}$$

Da cui

$$F_{X_{(i)}}(x) = [F_X(x)]^n = \left(\frac{x}{\theta}\right)^{3n}$$

come confermato da taluni

Esempio 7.2.7 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x,\theta) = \frac{2x}{\theta^2}$$

with $X \in [0,\theta]$. Compute the probability distribution function of the maximum $X_{(n)}$

1. $F_n(x) = \frac{x^{2n}}{\theta^n}$
2. $F_n(x) = \frac{x^{n-1}}{\theta^n}$
3. $F_n(x) = \frac{x^{3n-1}}{\theta^{3n}}$
4. $F_n(x) = \frac{x^{2n}}{\theta^{2n}}$; taluni suggeriscono questa

Analogamente

$$F_X(x) = \int \frac{2x^2}{\theta^2} = \frac{2}{\theta^2}\int x^2 = \frac{2}{\theta}\frac{x^2}{2} = \frac{x^2}{\theta^2}$$

da cui

$$F_{X_{(i)}}(x) = [F_X(x)]^n = \left(\frac{x}{\theta}\right)^{2n}$$

7.2.3 Generalized $X_{(i)}$

Osservazione importante 44. If we write $X_{(i)} \sim F_{(i)}(x)$, with $i = 1,\dots,n$ we mean that $X_{(i)}$ is distributed following the i-th ordered statistic.

Proposizione 7.2.5 (Distribution function)**.**

$$F_{(i)}(x) = \mathbb{P}\left(X_{(i)} \le x\right) = \sum_{j=i}^{n} \binom{n}{j} F_X(x)^j \cdot (1 - F_X(x))^{n-j} \tag{7.11}$$

Osservazione 245. 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

Proof. Not proved formally but to give intuition, 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 7.2.6 (Density function)**.**

$$f_{(i)}(x) = \binom{n}{i} i F_X(x)^{i-1} \cdot f_X(x)(1 - F_X(x))^{n-i} \tag{7.12}$$

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 7.2.8. 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

$$\begin{aligned} X_{(1)} &\sim \text{Beta}(1, n) \\ X_{(n)} &\sim \text{Beta}(n, 1) \end{aligned}$$

7.3 Inequalities

7.3.1 Markov (Viroli)

Teorema 7.3.1. *Given $X \in \mathbb{R}^+$, $D_X = \mathbb{R}^+$, $\lambda > 0$*

$$\mathbb{P}(X \geq \lambda \cdot \mathbb{E}[X]) \leq \frac{1}{\lambda} \tag{7.13}$$

Proof. Let

$$\mathbb{E}[X] = m = \int_{D_X} x \cdot f(x)\, \mathrm{d}x = \int_0^{+\infty} x \cdot f(x)\, \mathrm{d}x$$

Now

$$m \geq \int_{\lambda m}^{+\infty} x \cdot f(x)\, \mathrm{d}x \geq \int_{\lambda m}^{+\infty} x \cdot m \cdot f(x)\, \mathrm{d}x = \lambda m \underbrace{\int_{\lambda m}^{+\infty} f(x)\, \mathrm{d}x}_{\mathbb{P}(X \geq \lambda \cdot m)}$$

therefore

$$m \geq \lambda m \, \mathbb{P}(X \geq \lambda \cdot m) \iff \frac{1}{\lambda} \geq \mathbb{P}(X \geq \lambda \cdot m)$$

□

Esempio 7.3.1 (Esame vecchio viroli)**.** Let $\{X_n\}$ be a sequence of independent exponential random variables with parameter $\lambda_n = \frac{n}{2}$. Find the value of n such that $\mathbb{P}(X_n > 0.25) \leq 0.8$.
According to markov inequality

$$\mathbb{P}(X \geq c\,\mathbb{E}[X]) \leq \frac{1}{c}$$

We have that

$$\mathbb{E}[X_n] = \frac{1}{\lambda_n} = \frac{2}{n}$$

So

$$\mathbb{P}\left(X_n \geq c \cdot \frac{2}{n}\right) \leq \frac{1}{c}$$

TODO: boh qui non mi è chiarissimo

Now if $\frac{1}{c} = 0.8$ then $c = 1.25$ and we have:

$$\mathbb{P}\left(X_n \geq 1.25\frac{2}{n}\right) \leq 0.8$$
$$\mathbb{P}\left(X_n \geq \frac{2.5}{n}\right) \leq 0.8$$
$$\frac{2.5}{n} = 0.25$$
$$n = 10$$

Risposta $n = 10$

7.3.2 Tchebychev (Viroli)

Osservazione importante 46. We have two equivalent formulations

Teorema 7.3.2 (Tchebychev inequality)**.** *Respectively*

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \lambda \cdot \sigma_X\right) \leq \frac{1}{\lambda^2} \tag{7.14}$$

$$\mathbb{P}\left(|X - \mathbb{E}[X]| < \lambda \cdot \sigma_X\right) \geq 1 - \frac{1}{\lambda^2} \tag{7.15}$$

where σ_X is the standard deviation of X

Proof. 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

$$\mathbb{P}\left(Y \geq \lambda \mathbb{E}[Y]\right) \leq \frac{1}{\lambda}$$
$$\mathbb{P}\left((X - \mathbb{E}[X])^2 \geq \lambda \sigma_X^2\right) \leq \frac{1}{\lambda}$$
$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \sqrt{\lambda}\sigma_X\right) \leq \frac{1}{\lambda}$$

Then by setting $\lambda^* = \sqrt{\lambda}$ we conclude

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \lambda^* \sigma_x\right) \leq \frac{1}{(\lambda^*)^2}$$

□

Esempio 7.3.2 (esame viroli)**.** Let X_n be a sequence of independent poisson random variable with parameter 9 and $\overline{x}_n = \sum_{i=1}^n X_i/n$ is the partial mean. By the chebychev inequality find the value of n such that

$$\mathbb{P}\left(|\overline{x} - 9| < 15\right) \geq 0.99$$

- n = 36
- n = 10
- n = 4 taluni suggeriscono questa, confermata sotto
- n = 40

Qui effettivamente si ha che 9 è il valore atteso della somma di queste poissoniane poiché

$$\mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \frac{n\,\mathbb{E}[X_i]}{n} = \mathbb{E}[X_i] = 9$$

Il setup è dunque giusto e data la richiesta dobbiamo applicare la disuguaglianza nella seconda versione; abbiamo che

$$1 - \frac{1}{\lambda^2} = 0.99 \iff \lambda = 10$$

Dunque si ha che

$$10\sqrt{\text{Var}\left[\sum_{i=1}^{n} X_i/n\right]} = 15$$

Ora per ricavare n (ricordando che $\text{Var}[X_i] = \mathbb{E}[X_i] = 9$)

$$\text{Var}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\text{Var}\left[\sum_{i=1}^{n} X_i\right]}{n^2} = \frac{n\,\text{Var}[X_i]}{n^2} = \frac{varX_i}{n} = \frac{9}{n}$$

$$\sqrt{\text{Var}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right]} = \frac{3}{\sqrt{n}}$$

Dunque

$$10\frac{3}{\sqrt{n}} = 15 \iff \sqrt{n} = 2 \iff n = 4$$

7.3.3 Tchebychev (Rigo)

Teorema 7.3.3. *For any real random variable X*

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}[|X|^{\alpha}]}{c} \tag{7.16}$$

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 246. Useful because it applies to any random variable without any assumption and gives an upper bound of the prob.

Osservazione 247. An important special case is when $X = Y - \mathbb{E}[Y]$ and $\alpha = 2$, in this case the inequaliy goes to

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq c) \leq \frac{\text{Var}[Y]}{c^2}$$

But to apply Tchebychev in this form we need to know that the variance exists.

7.3.4 Jensen (Rigo)

Proposizione 7.3.4. *Let X be a real random variable and $f : I \to R$ a function defined on interval I. Now suppose that*

1. *f is a convex function*
2. *$\mathbb{P}(X \in I) = 1$*
3. *$\mathbb{E}[|X|] < +\infty$, $\mathbb{E}[|f(X)|] < +\infty$*

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 7.3.3. 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
$$\mathbb{E}\left[X^2\right] \geq [\mathbb{E}[X]]^2 \tag{7.17}$$
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
$$\mathbb{E}[|X|] \geq |\mathbb{E}[X]| \tag{7.18}$$
- $f(x) = x^{b/a}$ for any $x >= 0$ with $(0 < a < b)$. Applying Jensen
$$\mathbb{E}\left[|X|^b\right] = \mathbb{E}\left[(|X|^a)^{\frac{b}{a}}\right] \geq [\mathbb{E}[|X|^a]]^{\frac{b}{a}} \tag{7.19}$$
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 7.3.5. *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*

$$\mathbb{E}[f(X)] > f(\mathbb{E}[X]) \tag{7.20}$$

Osservazione 248. Now we prove that the rv is degenerate iff its variance is 0.

Proposizione 7.3.6.

$$X \sim \delta. \iff \mathrm{Var}[X] = 0$$

Proof. 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

□

7.4 Rigo: Conditional distribution

Osservazione 249. Roughly speaking the problemi is: given 2 real random variable X, Y we aim to evaluate the distribution of Y given that $X = x$.

Definizione 7.4.1 (Conditional distribution)**.** 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 $\mathbb{P}([(X, Y) \in C | X = x])$, $\forall x \in A$ (and not necessarily $\forall x \in \mathbb{R}$).
   For instance if $X \sim \mathrm{Unif}(0, 1)$, its enough to assign $P[(X, Y) \in C | X = x]$, $\forall x \in (0, 1)$
2. if $X \perp\!\!\!\perp Y$, then $\mathbb{P}((X, Y) \in C | X = x) = \mathbb{P}((x, Y) \in C | X = x)$ is true by property 3 of the definition. Then i can drop the conditioning because X and Y are independent

   $$P[(x; Y) \in C | X = x] = P[(x; Y) \in C]$$

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 $\mathbb{P}((X,Y) \in C|X = x)$ is very useful but also quite dangereous. Infact, if $P(X = x) = 0$ (which is possible eg in continuous distribution), then $P[(X,Y) \in C|X = x]$ 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 $P(X = x) = 0, \forall x \in \mathbb{R}$ (or equivalently the distribution function is continuous) then by the previous remark $\mathbb{P}((X,Y) \in C|X = x)$ 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 250. Unfortunately, in generale there is not an intuitive formula to evaluate conditional distribution (there is in some cases as we'll see later).

Esempio 7.4.1 (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 $C = \{(x,y) \in R^2 : x = y\}$ which is the set of points constituting the diagonal

$$\begin{aligned}\mathbb{P}(X = Y) = \mathbb{P}((X,Y) \in C) &\overset{(1)}{=} E_X\{\mathbb{P}((X,Y) \in C|X = x)\} \\ &\overset{(2)}{=} E_X\{\mathbb{P}((x,Y) \in C|X = x)\} \overset{(3)}{=} E_X\{\mathbb{P}((x,Y) \in C)\} \\ &= E_X(\underbrace{\mathbb{P}(Y = X)}_{0}) \overset{(4)}{=} E_X(0) = 0\end{aligned}$$

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

$$\mathbb{P}(X = Y) = \sum_{x \in B} \mathbb{P}(X = Y, X = x)$$

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 7.4.2. 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)
- $\mathbb{P}(\sin(Y) = a)) = 0$, $\forall a \in \mathbb{R}$ (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 7.4.3. Suppose $X \perp\!\!\!\perp Y$, $X \sim \mathrm{Unif}(0, 1)$ and $Y \sim \mathrm{N}(0, 1)$. We want to evaluate the distribution function of the product XY.
Here conditional distribution become handy. For all $a \in \mathbb{R}$, by definition the distribution function is

$$\mathbb{P}(XY \le a) \overset{(1)}{=} E_X(\mathbb{P}(XY \le a | X = x)) = E_X(\mathbb{P}(xY \le a | X = x))$$
$$\overset{(\perp\!\!\!\perp)}{=} E_X(\mathbb{P}(xY \le a)) \overset{(2)}{=} \int_{-\infty}^{+\infty} \mathbb{P}(xY \le a) f(x)\,\mathrm{d}x$$
$$= \int_0^1 \mathbb{P}(xY \le a) 1\,\mathrm{d}x \overset{(3)}{=} \int_0^1 \mathbb{P}\left(\mathrm{N}(0, 1) \le \frac{a}{x}\right)\mathrm{d}x$$

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 7.4.4. 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

$$\begin{aligned}
\mathbb{P}\left(B^2 - 4AC \geq 0\right) &= E_C\left\{\mathbb{P}\left(B^2 \geq 4AC | C = c\right)\right\} = E_C\left\{\mathbb{P}\left(B^2 \geq 4Ac | C = c\right)\right\} \\
&\overset{(\perp\!\!\!\perp)}{=} E_C\left\{\mathbb{P}\left(B^2 \geq 4Ac\right)\right\} \overset{(1)}{=} \int_{-\infty}^{+\infty} \mathbb{P}\left(B^2 \geq 4Ac\right) f(c)\, \mathrm{d}c \\
&= \int_0^{+\infty} \mathbb{P}\left(B^2 \geq 4Ac\right) e^{-c}\, \mathrm{d}c \overset{(2)}{=} \int_0^{+\infty} E_A\left\{\mathbb{P}\left(B^2 \geq 4Ac\right) | A = a\right\} e^{-c}\, \mathrm{d}c \\
&= \int_0^{+\infty} E_A\left\{\mathbb{P}\left(B^2 \geq 4ac\right)\right\} e^{-c}\, \mathrm{d}c = \int_0^{+\infty}\int_0^{+\infty} \mathbb{P}\left(B^2 \geq 4ac\right) e^{-a} e^{-c}\, \mathrm{d}a\, \mathrm{d}c \\
&= \int_0^{+\infty}\int_0^{+\infty} \mathbb{P}\left(B \geq 2\sqrt{ac}\right) e^{-a} e^{-c}\, \mathrm{d}a\, \mathrm{d}c \overset{(3)}{=} \int_0^{+\infty}\int_0^{+\infty} e^{-2\sqrt{ac}} e^{-a} e^{-c}\, \mathrm{d}a\, \mathrm{d}c
\end{aligned}$$

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 $\mathbb{P}(Z > z) = e^{-\lambda z}$.

Osservazione importante 51. How to calculate $\mathbb{P}((X, Y) \in C | X = x)$? 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 7.4.2 (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 7.4.3 (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

$$\mathbb{P}(Y \leq y | X = x) = \int_{-\infty}^{y} \frac{f(x,t)}{f_1(x)} \, \mathrm{d}t, \qquad \forall x, y \in \mathbb{R} : f_1(x) > 0$$

in this special case, we have an explicit formula for the conditional distribution

Definizione 7.4.4. In general, given any random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$, the corresponding order statistics are $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ where $X_{(1)}, \ldots, X_{(n)}$ are obtained by arranging $X_1, \ldots, X_n$ in increasing order.

Esempio 7.4.5. If $n = 2$, $X_{(1)} = min(X_1, X_2)$ and $X_{(2)} = max(X_1, X_2)$

Teorema 7.4.1 (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 7.4.6 (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)

7.5 Rigo: Multivariate normal

Osservazione 251. Let's start from univariate and see that multivariate formula are univariate generalization

Proposizione 7.5.1 (Characteristic functions of univariate normal)**.** *If* $Z \sim \mathrm{N}(0,1)$ *and* $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ *then* $\forall t \in \mathbb{R}$*:*

$$\phi_Z(t) = e^{-t^2/2} \tag{7.21}$$
$$\phi_X(t) = e^{it\mu - \frac{1}{2}(t\sigma)^2} \tag{7.22}$$

Proof. 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 $Z \sim \mathrm{N}(0,1)$ and thus we can derive the formula given the definition (in the univariate case) as:

$$\begin{aligned}\phi_X(t) = \mathbb{E}\left[e^{it(\mu+\sigma Z)}\right] &= \mathbb{E}\left[\underbrace{e^{it\mu}}_{\text{constant}}\, e^{it\sigma Z}\right] = e^{it\mu}\,\mathbb{E}\left[e^{i(t\sigma)Z}\right] \\ &= e^{it\mu} \cdot \phi_Z(t\sigma) = e^{it\mu - \frac{1}{2}(t\sigma)^2} \qquad \forall t \in \mathbb{R}\end{aligned}$$

□

Osservazione 252. 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 7.5.1. 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 253. Our definition includes not only absolutely continuous normal vector, but also other (eg degenerate in some cases)

Osservazione importante 52. 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

$$\mu = \begin{bmatrix} \mathbb{E}[X_1] \\ \dots \\ \mathbb{E}[X_n] \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \operatorname{Var}[X_1] & \operatorname{Cov}(X_1, X_2) & \dots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}[X_2] & \dots & \operatorname{Cov}(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \operatorname{Cov}(X_n, X_1) & \dots & \dots & \operatorname{Var}[X_n n] \end{bmatrix}$$

- 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:

$$\begin{aligned} \mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{Y}}\right] = \mathbb{E}\left[e^{i\mathbf{t}^T\alpha + \mathbf{A}\mathbf{X}}\right] = \mathbb{E}\left[\underbrace{e^{it^T\alpha}}_{\text{constant}}\, e^{it^T AX}\right] = e^{it^T\alpha} \underbrace{\mathbb{E}\left[e^{it^T AX}\right]}_{\phi_X(A^T t)} \\ = e^{it^T\alpha} e^{it^T A\mu - \frac{1}{2}t^T A\sigma A^T t} = \exp\left(it^T(\alpha + A\mu) - \frac{1}{2} t^T (A\sigma A^T) t \right) \\ \iff Y \sim \mathrm{N}\left(\alpha + \mathbf{A}\mu, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\right) \end{aligned}$$

□

Osservazione importante 53. 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

$$\mathbf{Y} = \begin{bmatrix} X_1 \\ X_2 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \mathbf{AX}$$

Esempio 7.5.1 (Assignment 1 Viroli, Exercise 3)**.** Suppose that $\mathbf{X}$ is a bivariate Gaussian vector with components (X_1, X_2) which are marginally standard normally distributed and with correlations $1/2$:

1. What is the distribution of $Y_1 = 2X_1 - X_2$ and $Y_2 = X_1 - X_2/2$
2. find the linear transformation from $\mathbf{X}$ to $\mathbf{Y}$ and ask what is the distribution of $\mathbf{Y}$

Since $X_1, X_2 \sim \mathrm{N}(0,1)$ and considered that

$$\begin{aligned} \frac{1}{2} = \mathrm{Corr}(X_1, X_2) &= \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}[X_1]}\sqrt{\mathrm{Var}[X_1]}} = \frac{\mathrm{Cov}(X_1, X_2)}{1 \cdot 1} \\ &= \mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1) \end{aligned}$$

1. if $Y_1 = 2X_1 - X_2$ and $Y_2 = X_1 - X_2/2$, then Y_1, Y_2 will be linear combinations of correlated normals; the distributions of Y_1, Y_2 will be normals with mean the linear combinations of means:

$$\begin{aligned} \mathbb{E}[Y_1] &= \mathbb{E}[2X_1 - X_2] = 2\,\mathbb{E}[X_1] - \mathbb{E}[X_2] = 0 \\ \mathbb{E}[Y_2] &= \mathbb{E}\left[X_1 - \frac{1}{2}X_2\right] = \mathbb{E}[X_1] - \frac{1}{2}\,\mathbb{E}[X_2] = 0 \end{aligned}$$

Applying $\mathrm{Var}[aX + bY] = a^2\,\mathrm{Var}[X] + b^2\,\mathrm{Var}[Y] + 2ab\,\mathrm{Cov}(X, Y)$ we have:

$$\begin{aligned} \mathrm{Var}[Y_1] &= \mathrm{Var}[2X_1 - X_2] = 4\,\mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2 \cdot 2(-1)\,\mathrm{Cov}(X_1, X_2) \\ &= 4 + 1 - 4\frac{1}{2} = 5 - 2 = 3 \\ \mathrm{Var}[Y_2] &= \mathrm{Var}\left[X_1 - \frac{1}{2}X_2\right] = \mathrm{Var}[X_1] + \frac{1}{4}\,\mathrm{Var}[X_2] + 2\left(-\frac{1}{2}\right)\mathrm{Cov}(X_1, X_2) \\ &= 1 + \frac{1}{4} - \frac{1}{2} = \frac{3}{4} \end{aligned}$$

Therefore: $Y_1 \sim \mathrm{N}(0, 3)$, $Y_2 \sim \mathrm{N}\left(0, \frac{3}{4}\right)$

2. in general, a linear trasformation of a multivariate normal is still normal; if $\mathbf{X} \sim \mathrm{MVN}(\mu, \mathbf{\Sigma})$ is a n-dimensional random vector and $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$, with $\mathbf{A}$ an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, then $\mathbf{Y}$ is a m-dimensional random vector and specifically $\mathbf{Y} \sim \mathrm{MVN}\left(\mathbf{A}\mu + \mathbf{b}, \mathbf{A\Sigma A}^T\right)$.
In our case $m = n = 2$ and we have:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}\right), \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 2X_1 - X_2 \\ X_1 - \frac{1}{2}X_2 \end{bmatrix}$$

so $\mathbf{Y} = \mathbf{AX} + \mathbf{0}$, represent the linear transformation needed to obtain $\mathbf{Y}$, where

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & -1/2 \end{bmatrix}$$

Therefore to evaluate the parameters of $\mathbf{Y}$:

$$\mathbf{A}\mu + \mathbf{b} = \begin{bmatrix} 2 & -1 \\ 1 & -1/2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathbf{A}\Sigma\mathbf{A}^T = \begin{bmatrix} 2 & -1 \\ 1 & -1/2 \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & -1/2 \end{bmatrix} = \begin{bmatrix} 3 & 3/2 \\ 3/2 & 3/4 \end{bmatrix}$$

Finally:

$$\mathbf{Y} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 3/2 \\ 3/2 & 3/4 \end{bmatrix} \right)$$

Chapter 8

Convergence

Osservazione importante 54 (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. convergence in mean of order k (strong)
4. almost sure convergence (strong)

8.1 Convergence in probability

8.1.1 Definition

Osservazione 254. It's the first type of convergence: this is a weak type (it implies convergency in distribution but not stronger kinds of convergency)

Definizione 8.1.1 (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), $\forall \varepsilon > 0$:

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \tag{8.1}$$

$$\mathbb{P}\left(|X_n - X| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1 \tag{8.2}$$

Osservazione 255. 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 collassing to a point (eg population mean) and can be a good property for an estimator.

8.1.2 Weak consistence

Definizione 8.1.2 (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 55. Weak consistency means converging probability (link between probability and inference)

Esempio 8.1.1. 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)

$$\max_{1\le i\le n} X_i = X_{(n)}$$

Let's prove that $X_{(n)}$ is a consistent estimator for θ, that is:

$$X_{(n)} \xrightarrow{p} \delta_\theta$$

Remembering that $F_{(n)}(x) = [F_X(x)]^n$ we want to prove that

$$\mathbb{P}\left(\left|X_{(n)} - \theta\right| < \varepsilon\right) \to 1$$

Now we have:

$$\begin{aligned}\mathbb{P}\left(\left|X_{(n)} - \theta\right| < \varepsilon\right) &\overset{(1)}{=} \mathbb{P}\left(-X_{(n)} + \theta < \varepsilon\right) = \mathbb{P}\left(-X_{(n)} < \varepsilon - \theta\right) = \mathbb{P}\left(X_{(n)} > \theta - \varepsilon\right)\\ &= 1 - \mathbb{P}\left(X_{(n)} \le \theta - \varepsilon\right) = 1 - F_{(n)}(\theta - \varepsilon)\\ &= 1 - [F_X(\theta - \varepsilon)]^n\end{aligned}$$

where in (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.
If $X \sim \text{Unif}(0,\theta)$, then $F_X(x) = \frac{x}{\theta}$, $0 \le x \le \theta$ 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 $0 < \varepsilon \le \theta$

$$\lim_{n\to\infty} 1 - \left(\frac{\theta - \varepsilon}{\theta}\right)^n = 1$$

Proposizione 8.1.1 (Sufficient conditions for weak consistence)**.** *If*

$$\begin{cases}\lim\limits_{n\to+\infty} \mathbb{E}[X_n] = \theta\\ \lim\limits_{n\to+\infty} \text{Var}[X_n] = 0\end{cases} \implies X_n \xrightarrow{p} \delta_\theta \tag{8.3}$$

Osservazione 256. The viceversa does not hold: eg X_n can converge in probility even if these conditions are not met.

Proof. Applying Tchebychev inequality

$$\mathbb{P}\left(|X_n - \mathbb{E}[X_n]| < \lambda\sigma(X_m)\right) \ge 1 - \frac{1}{\lambda^2}$$

Now we define/substitute $\varepsilon = \lambda\sigma(X_m)$ so that $\lambda^2 = \frac{\varepsilon^2}{\sigma^2(X_m)}$; therefore

$$\mathbb{P}\left(|X_n - \mathbb{E}[X_n]| < \varepsilon\right) \geq 1 - \frac{\sigma^2(X_n)}{\varepsilon^2}$$

if $n \to +\infty$ the last term go to zero so

$$\mathbb{P}\left(|X_n - \mathbb{E}[X_n]| < \varepsilon\right) \geq 1$$

and since this probability can't be larger than 1, it must be 1 so

$$\mathbb{P}\left(|X_n - \mathbb{E}[X_n]| < \varepsilon\right) = 1 \implies X_n \xrightarrow{p} \theta$$

□

Esempio 8.1.2. Let $X_n \sim \text{Geom}(p_n)$ with $p_n = 1 - \frac{1}{n}$, having pmf

$$\mathbb{P}(X_n = x) = p_n(1 - p_n)^{x-1}$$

with $\mathbb{E}[X_n] = \frac{1}{p_n}$, $\text{Var}[X_n] = \frac{1-p_n}{p_n^2}$. Let's prove that $X_n \xrightarrow{p} \delta_1$.

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}[X_n] &= \frac{1}{p_n} = \frac{1}{1 - \frac{1}{n}} \to 1 \\
\lim_{n\to\infty} \text{Var}[X_n] &= \frac{1 - p_n}{p_n^2} = \frac{1 - \left(1 - \frac{1}{n}\right)}{\left(1 - \frac{1}{n}\right)^2} = \frac{\frac{1}{n}}{(1 - \frac{1}{n})^2} \\
&= \frac{\frac{1}{n}}{(\frac{(n-1)^2}{n^2})} = \frac{n}{(n-1)^2} \to 0
\end{aligned}$$

Esempio 8.1.3 (Esame vecchio viroli)**.** Let θ be the parameter of a population random variable X that follows a continuous uniform distribution on the interval $[\theta - 2, \theta + 1]$ and let $X = (X_1, \ldots, X_n)$ be a simple random sample. Given the estimator $T_n(X) = \overline{X} + \frac{1}{2}$ decide if it is weakly consistent.
We need

$$\begin{aligned}
\mathbb{E}[T_n(X)] &= \mathbb{E}\left[\overline{X} + \frac{1}{2}\right] = \mathbb{E}\left[\overline{X}\right] + \frac{1}{2} = \mathbb{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] + \frac{1}{2} = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right] + \frac{1}{2} \\
&= \frac{1}{n} \cdot n \cdot \mathbb{E}[X_i] + \frac{1}{2} = \frac{\theta - 2 + \theta + 1}{2} + \frac{1}{2} = \frac{2\theta}{2} = \theta \\
\text{Var}[T_n(X)] &= \text{Var}\left[\overline{X} + \frac{1}{2}\right] = \text{Var}\left[\overline{X}\right] = \text{Var}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^n X_i\right] \\
&= \frac{1}{n^2} \cdot n \cdot \text{Var}[X_i] = \frac{\text{Var}[X_i]}{n} = \frac{1}{12n}(\theta + 1 - \theta + 2)^2 = \frac{9}{12}\frac{1}{n} = \frac{3}{4n} \to 0
\end{aligned}$$

Therefore $T(X)$ is weakly consistent

Esempio 8.1.4 (Esame vecchio viroli)**.** Let X_n be a sequence of iid exponential random variables with parameter 1. Study the convergence in probability of the minimum $X_{(1)}$.
The minimum of an exponential should converge to the minimum of the domain

so for the exponential is 0. We check the two sufficient condition but first let write the density function of the minimum

$$f_{X_1}(x) = n \cdot f_X(x) \cdot [1 - F_X(x)]^{n-1}$$

where for the Exp (1) we have

$$f(x) = e^{-x}$$
$$F_X(x) = 1 - e^{-x}$$

and therefore

$$f_{X_{(1)}}(x) = n \cdot e^{-x} \cdot (1 - 1 + e^{-x})^{n-1} = n \cdot e^{-x(n-1)-x} = n \cdot e^{-nx}$$

We have

$$\mathbb{E}\left[X_{(1)}\right] = \int_0^{+\infty} x \cdot n \cdot e^{-nx} = -\int_0^{+\infty} x \cdot (-n) \cdot e^{-nx}$$

Sviluppiamo l'integrale indefinito e poi valutiamolo

$$\begin{aligned} -\int x(-n)e^{-nx} &= -\left[e^{-nx}x - \int e^{-nx}\right] = -\left[e^{-nx}x + \frac{1}{n}\int (-n)e^{-nx}\right] \\ &= -\left[e^{-nx}x + \frac{1}{n}e^{-nx}\right] = -e^{-nx}\left(x + \frac{1}{n}\right) \end{aligned}$$

Che valutato

$$\left[-e^{-nx}\left(x + \frac{1}{n}\right)\right]_0^{+\infty} = \frac{1}{n}$$

Per cui $\mathbb{E}\left[X_{(1)}\right] \to 0$.
Per la varianza calcoliamo il secondo momento

$$\mathbb{E}\left[X_{(1)}^2\right] = \int_0^{+\infty} x^2 \cdot n \cdot e^{-nx} = \ldots = \frac{2}{n^2}$$

Per cui

$$\text{Var}\left[X_{(1)}\right] = \frac{2}{n^2} - \frac{1}{n} \to 0$$

Answer: $X_{(1)} \xrightarrow{p} 0$

Esempio 8.1.5 (Esame vecchio viroli)**.** Let $(X_1, \ldots, X_n)$ a simple random sample from an exponential random variable

$$f_X(x) = \theta e^{-\theta x}$$

Study the convergence in probability of

$$T_n = 2\frac{\sum_{i=1}^n X_i}{n} + 3$$

1. Tn cnverges in probability to a dirac at $(3+\theta)/2$
2. Tn converges to a dirac at 2θ
3. Tn does not converge in probability
4. Tn converge to a dirac at $\frac{2}{\theta}+3$ (dovrebbe essere questa)

For the exponential we have that $\mathbb{E}[X_i] = \frac{1}{\theta}$ and $\mathrm{Var}[X_i] = \frac{1}{\theta^2}$. We check the two sufficient condition

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}\left[2\frac{\sum_{i=1}^n X_i}{n} + 3\right] = 2\,\mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] + 3 = \frac{2}{n} n\,\mathbb{E}[X_i] + 3 \\ &= \frac{2}{\theta} + 3 \\ \mathrm{Var}[T_n] &= \mathrm{Var}\left[2\frac{\sum_{i=1}^n X_i}{n} + 3\right] = \frac{4}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \frac{4}{n^2} n\,\mathrm{Var}[X_i] \\ &= \frac{4}{n}\frac{1}{\lambda^2} \to 0\end{aligned}$$

So $T_n \xrightarrow{p} \delta_{\frac{2}{\theta}+3}$

8.1.3 Theorem: weak law of large numbers

Teorema 8.1.2 (Weak law of large numbers)**.** *Let X_n be a sequence of iid rvs with $\mathbb{E}[X_n] = \theta$ and $\mathrm{Var}[X_n] = \sigma^2 < +\infty$; if we define the partial mean as the mean of the first n rvs*

$$M_n = \frac{\sum_{i=1}^n X_i}{n} \tag{8.4}$$

then we have that

$$M_n \xrightarrow{p} \delta_\theta \tag{8.5}$$

Proof. We have that

$$\begin{aligned}\mathbb{E}[M_n] &= \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n} = \frac{n\theta}{n} = \theta \\ \mathrm{Var}[M_n] &= \frac{1}{n^n}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

therefore since both

$$\begin{aligned}\lim_{n\to+\infty} \mathbb{E}[M_n] &= \theta \\ \lim_{n\to+\infty} \mathrm{Var}[M_n] &= 0\end{aligned}$$

the sufficient conditions are met and $M_n \xrightarrow{p} \delta_\theta$ □

Esempio 8.1.6. Let $X_1, \dots X_n$ be independent rvs each distributed as Bernoulli with parameter p. Prove that $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{p} p$ as $n \to \infty$.
According to the WLLN we have that

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{p} \mathbb{E}\left[X^2\right]$$

Now if $X \sim \text{Bern}(p)$, then $\mathbb{E}[X] = p$ and $\text{Var}[X] = p(1-p)$, so $\mathbb{E}\left[X^2\right] = \text{Var}[X] + \mathbb{E}[X]^2 = p(1-p) + p^2 = p$. Therefore

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} p$$

8.2 Convergence in law/distribution

Osservazione 257. We have two equivalent definition, by limit of distribution function or convergence in law/distribution of the moment generating function.

Definizione 8.2.1 (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 8.2.2 (Alternate definition)**.**

$$X_n \xrightarrow{d} X \iff M_{X_n}(t) \to M_X(t), \forall t : |t| < \varepsilon \tag{8.6}$$

in a intorno di $t = 0$

Osservazione 258. Two theorem without proof before going on

Teorema 8.2.1. *Convergence in probability is stronger than convergence in distribution since* $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$

Teorema 8.2.2. *... but in the case of dirac we have both implication* $X_n \xrightarrow{p} \delta_\theta \iff X_n \xrightarrow{d} \delta_\theta$

Esempio 8.2.1. Let $\{X_n\}_{n\in\mathbb{N}}$ be iid standard normal, $X_n \sim \text{N}(0,1)$. 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)

$$M_{Y_n}(t) = M_{\frac{\chi_n^2}{n}}(t) = M_{\chi_n^2}\left(\frac{t}{n}\right) = \left(1 - 2\frac{t}{n}\right)^{-\frac{n}{2}}$$

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 8.2.2. 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 $\text{Pois}(\lambda)$.

Esempio 8.2.3 (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 $X_n \xrightarrow{d} \text{Pois}(\lambda)$.

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:

$$\begin{aligned}\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 &= \frac{\beta^2\alpha}{\alpha-2} - \frac{\alpha^2\beta^2}{(\alpha-1)^2} \\ &= \frac{\beta^2(\alpha-1)^2 - \alpha^2\beta^2(\alpha-2)}{(\alpha-2)(\alpha-1)^2} \\ &= \frac{\beta^2\alpha^3 + \beta^2\alpha - 2\beta^2\alpha^2 - \alpha^3\beta^2 + 2\alpha^2\beta^2}{(\alpha-2)(\alpha-1)^2}\end{aligned}$$

5. take $M_{X_n}(t)$ of the binomial:

$$M_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right)^n = \left(1 - \frac{\lambda}{n}(1-e^t)\right)^n \overset{(1)}{=} \left(1 - \frac{a}{n}\right)^n$$

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 $\iff$ $X \sim \mathrm{Pois}(\lambda)$.

DO: rivedere

Esempio 8.2.4. Let X be a continuous uniform random variable in $[0,1]$. Let $Y = \frac{X}{1-X}$ and $Y_n = Y^{1/n}$:

1. Determine $f_Y(y)$ and $F_Y(y)$
2. Determine $F_{Y_n}(y)$
3. Study the convergence in law of Y_n

We have

1. that
$$Y = \frac{X}{1-X} = g(X) \implies g^{-1}(Y) = \frac{1}{1+Y} = X$$
Then
$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}\left(\frac{X}{1-X} \le y\right) = \mathbb{P}(X \le y - yX) \\ &= \mathbb{P}(X + yX \le y) = \mathbb{P}\left(X \le \frac{y}{1+y}\right) = F_X\left(\frac{y}{1+y}\right) \\ &= \frac{y}{1+y} \end{aligned}$$
and so
$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cot\left|\frac{\partial g^{-1}(y)}{\partial y}\right| = \mathbb{1}_{[y/(1+y)]}(x) \cdot \left|-y(1+y)^{-2} + (1+y)^{-1}\right| \\ &= \left|\frac{1+y-y}{(1+y)^2}\right| \mathbb{1}_{x/1-x}(y) = \frac{1}{(1+y)^2}\mathbb{1}_{[0,+\infty]}(y) \end{aligned}$$
2. for the second point
$$F_{Y_n}(y) = \mathbb{P}(Y_n \le y) = \mathbb{P}\left(Y^{1/n} \le y\right) = \mathbb{P}(Y \le y^n) = F_Y(y^n) = \frac{y^n}{} + y^n$$
3. for the third
$$\lim_{n\to\infty} F_Y(y^n) = \begin{cases} 0 & \text{if } y < 1 \\ 1/2 & \text{if } y = 1 \\ 1 & \text{if } y > 1 \end{cases}$$
so the F of a δ_1 and $F_{Y_n}(y)$ coincides except for $y = 1$, but it is a discontinuity point and we can ignore it.
Therefore $Y_n \xrightarrow{d} \delta_1$

8.2.1 Theorem: central limit theorem

Osservazione importante 56. 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 259. 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 8.2.3. *Let X_i be iid random variables, with mean $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$; 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*

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} = \frac{S_n - n\mu}{\underbrace{\sqrt{n\sigma^2}}_{no\ cov,\ \perp\!\!\!\perp}} \overset{(1)}{=} \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where in (1) we divided everything by n.
Then $Z_n \xrightarrow{d} \mathrm{N}(0,1)$

Proof.

$$Z_n = \frac{S_n - n\mu}{\sigma \cdot \sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu}{\sigma \cdot \sqrt{n}} = \sum_{i=1}^{n} \underbrace{\left(\frac{X_i - \mu}{\sigma}\right)}_{U_i} \cdot \frac{1}{\sqrt{n}} = \frac{\sum_{i=1}^{n} U_i}{\sqrt{n}}$$

with $\mathbb{E}[U_i] = 0$ and $\mathrm{Var}[U_i] = 1$ (being standardized) and $\mathbb{E}\left[U_i^2\right] = 1$ as consequence of the first two using the variance formula $\mathrm{Var}[U_i] = \mathbb{E}\left[U_i^2\right] - \mathbb{E}[U_i]^2$.
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 decompose $M_U(t/\sqrt{n})$ 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 $\mathrm{N}(0,1)$. □

8.3 Convergence in mean of order k

8.3.1 Definition

Definizione 8.3.1. Let $k \in \mathbb{N}^+$. It' said that $X_n \xrightarrow{L_k} X$ if and only if

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^k\right] = 0$$

Osservazione importante 57 (Convergence in quadratic mean). One of the most famous is for $n = 2$, $X_n \xrightarrow{L_2} X \iff \lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0$

8.3.2 Strong consistence

Osservazione importante 58. 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 is 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.*

Definizione 8.3.2 (Strong consistence)**.** 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 8.3.1. *In this type of convergence we have this result*

$$X_n \xrightarrow{L_2} \delta_\theta \iff \begin{cases} \lim_{n\to+\infty} \mathbb{E}[X_n] = \theta \\ \lim_{n\to+\infty} \mathrm{Var}[X_n] = 0 \end{cases}$$

Esempio 8.3.1 (Esame vecchio viroli)**.** Let Y_n be a sequence of independent poisson random variables with parameter $\lambda_n = 1/\sqrt{n}$. Study the convergence in quadratic mean of Y_n.
First we need to decide where it converges. Let's try $\mathbb{E}[Y_n]$

$$\mathbb{E}[Y_n] = \lambda_n = \frac{1}{\sqrt{n}}$$

$$\lim_{n\to+\infty} \mathbb{E}[Y_n] = \lim_{n\to+\infty} \frac{1}{\sqrt{n}} = 0$$

Does it converge to a δ_0? let's apply the definition

$$\lim_{n\to+\infty} \mathbb{E}\left[(Y_n - 0)^2\right] = \lim_{n\to+\infty} \mathbb{E}\left[Y_n^2\right]$$

To obtain the second moment of a poisson (considered that $\mathrm{Var}[Y] = \lambda$):

$$\mathrm{Var}[Y] = \mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2$$
$$\lambda = \mathbb{E}\left[Y^2\right] - \lambda^2 \implies \mathbb{E}\left[Y^2\right] = \lambda + \lambda^2$$

and so in our case $\mathbb{E}\left[Y_n^2\right] = \frac{1}{\sqrt{n}} + \frac{1}{n}$. Then

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} + \frac{1}{n} = 0$$

Therefore: $Y_n \xrightarrow{L_2} 0$

Esempio 8.3.2. 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
$$\begin{aligned}
&\mathbb{E}\left[X_n\right] = \frac{2}{n} \\
&\text{Var}\left[X_n\right] = \frac{2}{n} \\
&\mathbb{E}\left[X_n^2\right] = \text{Var}\left[X_n\right] + \mathbb{E}\left[X_n\right]^2 = \frac{2}{n} + \frac{4}{n^2} = \frac{2n+4}{n^2}
\end{aligned}$$
And finally
$$\lim_{n\to+\infty} \frac{2n+4}{n^2} = 0$$
so it goes to δ_0

Esempio 8.3.3. Let $X_n \sim \text{Bern}\left(\frac{1}{n}\right) \cdot n$ so, its pmf be

$$X_n = \begin{cases} n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n \end{cases}$$

Study convergence in L_2 and probability.
We have that

$$\begin{aligned}
&\mathbb{E}\left[X_n\right] = n \cdot \frac{1}{n} = 1 \\
&\mathbb{E}\left[X_n^2\right] = n^2 \cdot \frac{1}{n} = n \\
&\text{Var}\left[X_n\right] = n - 1^2 = n - 1
\end{aligned}$$

Now

- we can't conclude X_n converges in L_2 because of the (limit of the) variance

$$\begin{cases} \lim_{n\to\infty} \mathbb{E}[X_n] = 1 \\ \lim_{n\to\infty} \operatorname{Var}[X_n] = +\infty \end{cases} \implies X_n \not\xrightarrow{L_2} \delta_1$$

- if it converges 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 $X_n \sim \text{Bern}(1/n)$: 0 with larger and larger prob, n with lowering prob. Therefore $X_n - 1$ will be 1 with increasing prob and so $1 \not\lesssim \varepsilon$, $\forall \varepsilon \in \mathbb{R}$.
  - 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).

Esempio 8.3.4. Let $X_1, X_2, \dots$ be a sequence of random variables such that

$$\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2} \quad \mathbb{P}(X_n = n) = \frac{1}{n^2}$$

- Does X_n converge in quadratic mean?
- Does it converge in probability

Respectively

1. For L_2 we should prove $\lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0$ but who is X? By reasoning we see that $X_n \to 0$ with probability $\to 1$ therefore we try with a δ_0

$$\lim_{n\to\infty} \mathbb{E}\left[(X_n - 0)^2\right] = \lim_{n\to\infty} \mathbb{E}\left[X_n^2\right]$$

   The second moment is

$$\mathbb{E}\left[X_n^2\right] = \frac{1}{n^2} \cdot \left(1 - \frac{1}{n^2}\right) + n^2 \frac{1}{n^2} = \frac{n^2 - 1}{n^4} + 1$$

   so

$$\lim_{n\to\infty} \mathbb{E}\left[X^2\right] = 1$$

   so we conclude that $X_n \not\xrightarrow{L_2} \delta_0$.

2. let's check the two sufficient assumptions for the convergence in probability:

(a) for the first we have

$$\mathbb{E}[X_n] = \frac{1}{n}\left(1 - \frac{1}{n^2}\right) = \frac{n^2-1}{n^3}$$

and $\lim_{n\to\infty} \mathbb{E}[X_n] = 0$

(b) for the second

$$\mathrm{Var}[X_n] = \mathbb{E}\left[X_n^2\right] - \mathbb{E}[X_n]^2 = \frac{n^2-1}{n^2} + 1 - \frac{(n^2-1)^2}{n^6}$$

from which $\lim_{n\to\infty} \mathrm{Var}[X_n] \to 1$

However by applying the definition

$$\lim_{n\to\infty} \mathbb{P}(|X_n| < \varepsilon) = 1$$

$= \lim_{n\to\infty} \mathbb{P}(X_n < \varepsilon) = 1$ and this is true since $X_n \to 0$ with probability $\to 1$ as $n \to \infty$

8.3.3 Theorem: strong law of large numbers

Osservazione 260. It's the most important theorem related to convergence in quadratic mean.

Teorema 8.3.2 (Strong law of large numbers)**.** *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of* indepedent *random variables and assume* $\mathbb{E}[X_n] = \mu$, $\mathrm{Var}[X_n] = \sigma^2 < +\infty$. *Then we say that the partial mean:*

$$M_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{L_2} \mu$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[(M_n - \mu)^2\right] &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n X_i}{n} - \frac{n\mu}{n}\right)\right] \overset{(1)}{=} \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^n (X_i - \mu)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[(X_i-\mu)^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{1}{n^2}\cdot n\cdot\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

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 $M_n \xrightarrow{L_2} \mu$ □

Esempio 8.3.5. Let $X_1, \ldots, X_n \sim \mathrm{Exp}(1)$. 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 $X_{(1)} \sim \text{Exp}(n)$.
Regarding the convergence to study,

- in this exercise, since we have the pdf of the minimum, it's convenient for us to try to study the limit of it, that is *in this case we study convergence in distribution* (using the cumulative distribution function, not the mgf or the characteristic function). 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 of F:

  $$\lim_{n\to\infty} F_{(1)}(x) = \lim_{n\to\infty} 1 - e^{-xn} = 1$$

  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:

  $$X_{(1)} \xrightarrow{d} \delta_0 \implies X_{(1)} \xrightarrow{p} \delta_0$$

  TODO: non chiarissimo, la cumulata dovrebbe essere una step function non una costante, poi ok che da 0 in poi sia a 1.

- 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

  $$X_{(1)} \xrightarrow{L_2} \delta_0 \implies X_{(1)} \xrightarrow{L_1} \delta_0$$

8.4 Almost sure convergence

Osservazione 261. It's a strong convergence

Definizione 8.4.1. 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 262. Difficult to prove because it's not the limit of a probability but the probability of a limit.

Osservazione 263. The most important associated theorem with a.s. convergence is the following; somewhat similar to the strong/weak law large number.

Teorema 8.4.1 (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 $M_n \xrightarrow{a.s.} \mu$*

Proof. No proof here, quite complicate. □

Esempio 8.4.1. 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:

$$\mu = \mathbb{E}\left[\frac{1}{1+X_i}\right] = \sum_{D_X} \frac{1}{1+x_i}\mathbb{P}(X_i = x_i) = \sum_{x=0}^{+\infty} \frac{1}{1+x}\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=0}^{+\infty} \frac{1}{(x+1)!} e^{-\lambda}\lambda^x \cdot \frac{\lambda}{\lambda} = \frac{1}{\lambda}\sum_{x=0}^{+\infty} \frac{1}{(x+1)!} e^{-\lambda}\lambda^{x+1} \overset{(1)}{=} \frac{1}{\lambda}\underbrace{\sum_{t=1}^{+\infty} \underbrace{\frac{1}{t!}e^{-\lambda}\lambda^t}_{\text{Pois}(\lambda)}}_{1-\left(\frac{1}{0!}e^{-\lambda}\lambda^0\right)}$$

$$= \frac{1}{\lambda}\left(1 - e^{-\lambda}\right)$$

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 59. 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 8.4.2. 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 (section below) 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 8.4.3 (Assignment 1 Viroli, Exercise 4)**.** Let $X_1, \ldots, X_n$ be a sequence of independent random variables with $X \sim \text{Exp}(\theta)$. Let $T_n = \frac{\sum_{i=1}^n e^{-x_i}}{n}$. Study the convergence of T_n as n goes to infinity.
By setting $Y_i = e^{-X_i}$ we have that $T_n = \frac{\sum_i Y_i}{n}$ so, being a partial mean of iid rvs with $\mathbb{E}[Y_i]$ constant (to be evaluated), we have that $T_n \xrightarrow{a.s.} \mathbb{E}[Y_i]$ by Kolmogorov theorem. Let's evaluate $\mathbb{E}[Y_i]$:

$$\begin{aligned}
\mathbb{E}[Y_i] = \mathbb{E}\left[e^{-X_i}\right] &= \int_{D_X} e^{-x} \cdot \underbrace{f(x)}_{\text{Exp}(\theta)} \, \mathrm{d}x = \int_0^{+\infty} e^{-x} \cdot \theta \cdot e^{-\theta x} \, \mathrm{d}x \\
&= \theta \int_0^{+\infty} e^{-x-\theta x} \, \mathrm{d}x = \theta \int_0^{+\infty} e^{-x-\theta x} \cdot \frac{(-1-\theta)}{(-1-\theta)} \, \mathrm{d}x \\
&= \frac{\theta}{-1-\theta} \int_0^{+\infty} e^{-x-\theta x} \cdot (-1-\theta) \, \mathrm{d}x = -\frac{\theta}{1+\theta} \left[e^{-x-\theta x}\right]_0^{+\infty} \\
&= -\frac{\theta}{1+\theta}[0-1] = \frac{\theta}{1+\theta}
\end{aligned}$$

So we can conclude that

$$T_n \xrightarrow{a.s.,p,d} \delta_{\frac{\theta}{1+\theta}}$$

Clearly as $\theta \to +\infty \implies T_n \to 1$; in figure **??** some heuristic checks for $\theta = 0.1, 1, 10$, (where if calculation above is ok, T_n should converge to $\frac{0.1}{1.1}, \frac{1}{2}, \frac{10}{11}$, horizontal dotted black lines).

```
set.seed(15346)
## for each theta we do "nreps" simulated sequences in order to have
## variability and check that at the end is a Dirac (eg no
## variability around the estimate theta/(1+theta)
## each simulated sequence is composed of "n" random extraction
## from Exp(theta)

n <- 30000
nreps <- 1000
thetas <- c(0.1, 1, 10)
cols <- c("green", "yellow", "red")

sim <- function(theta){
    res <- list()
    for (rep in 1:nreps) {
        x_i <- rexp(n = n, rate = theta)
        y_i <- exp(-x_i)
        sample_size  <- seq_along(x_i)
        partial_mean <- cumsum(y_i)/sample_size
        res[[sprintf("r%d", rep)]] <- partial_mean
    }
    data.frame(res)
}

res <- lapply(thetas, sim)

plotter <- function(data, theta, col, first){
    ## setup the plot if the first theta
    if (first){
        plot(c(0, n), 0:1, pch = NA,
             xlab = 'sample_size',
             ylab = 'T_n -> theta/(1+theta)')
    }
    sample_size <- seq_len(nrow(data))
    lapply(data, function(y){
        points(x = sample_size, y = y, col = col, pch = ".", )
    })
    abline(h = theta/(1 + theta), lty = 'dotted')
}

tmp <- Map(plotter,
           res, thetas, as.list(cols), list(TRUE, FALSE, FALSE))
legend(n * 0.7, 0.8, legend = sprintf("theta=%.1f", thetas),
       col = cols, lty=1)
# seems ok, theta = 10 somewhat quicker
```

8.5 Convergences properties

Proposizione 8.5.1 (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* $\xrightarrow{d}$*:*

$$\xrightarrow{p} \delta_\theta \iff \xrightarrow{d} \delta_\theta$$

Esempio 8.5.1 (Esame vecchio viroli)**.** Indicate which of the following definitions is false: the convergence in mean of order 4 implies:

1. convergence in quadratic mean
2. the convergence in mean of order 3
3. the almost sure convergence
4. the convergence in distribution

We have that $\xrightarrow{L_4} \not\Longrightarrow \xrightarrow{a.s.}$.

Teorema 8.5.2 (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:*

$$X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X) \tag{8.7}$$

$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X) \tag{8.8}$$

$$\begin{cases} X_n \xrightarrow{L_k} X \\ g \text{ is linear} \end{cases} \implies g(X_n) \xrightarrow{L_k} g(X) \tag{8.9}$$

$$X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X) \tag{8.10}$$

Osservazione 264. For the L_k case: if g is quadratic, log, exponential etc, being not a linear function, then the implication convergence doesn't hold.

Proposizione 8.5.3 (Further properties)**.** *We have that*

1. *for convergence in probability*

$$(X_n \xrightarrow{p} X \wedge Y_n \xrightarrow{p} Y) \implies aX_n + bY_n \xrightarrow{p} aX + bY \tag{8.11}$$

$$(X_n \xrightarrow{p} X \wedge Y_n \xrightarrow{p} Y) \implies X_n \cdot Y_n \xrightarrow{p} X \cdot Y \tag{8.12}$$

2. *same as above applies for* $\xrightarrow{a.s.}$

3. *for* $\xrightarrow{L_k}$ *we only have*

$$(X_n \xrightarrow{L_k} X \wedge Y_n \xrightarrow{L_k} Y) \implies aX_n + bY_b \xrightarrow{L_k} aX + bY \tag{8.13}$$

but the product does not hold

4. *for* $\xrightarrow{d}$ *we have Slutsky theorem:*

$$(X_n \xrightarrow{d} X \wedge Y_n \xrightarrow{d} \delta_c) \implies \begin{cases} X_n + Y_n \xrightarrow{d} X + c \\ X_n \cdot Y_n \xrightarrow{d} cX \end{cases} \tag{8.14}$$

8.6 Delta method

Osservazione 265. This is a very useful tool for inference.

Osservazione importante 60 (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.
Delta method is needed if we are interested not on θ but on a transformation on the parameter $g(\theta)$, with g continuous; this because using the continuous mapping theorem is not alway optimal.

Esempio 8.6.1 (Motivating example: odd)**.** Let $X_1, \dots, X_n \sim \text{Bern}(p)$ be independent, with $\mathbb{E}[X_i] = p$ and consider the partial mean $Y_n = \overline{X} = \frac{\sum_i X_i}{n}$. We know that, respectively by weak law of large number and by central limit theorem (it's a sum, not standardized) that:

$$\begin{aligned} Y_n &\xrightarrow{p} \delta_p \\ Y_n &\xrightarrow[\text{by CLT}]{d} \text{N}\left(p, \frac{p(1-p)}{n}\right) \end{aligned}$$

Some remarks:

1. the two limits above are not conflicting: 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 $\lim_{n\to\infty} \frac{p(1-p)}{n} = 0$.

2. 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 suppose we're interested not in p of event, but in its odd, that is:

$$g(p) = \frac{p}{1-p}$$

We know that (continuous mapping theorem), the transformation of the sequence converges to the transformation of the limit distribution:

$$Y_n \xrightarrow{p} p \implies g(Y_n) \xrightarrow{p} g(p) \iff odd \xrightarrow{p} \frac{p}{1-p} \iff \frac{\overline{x}}{1-\overline{x}} \xrightarrow{p} \frac{p}{1-p}$$

However this is a point results; we may be interested in constructing confidence intervals and hypothesis testing and for all that shit we need a proper distribution, not a point.
Therefore here comes the delta method.

Osservazione 266. To define the delta method first we need the generalized version of CLT.

Teorema 8.6.1 (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 $Z \sim \mathrm{N}(0,1)$*

$$\begin{cases} \sqrt{n}(Y_n - \theta) \xrightarrow{d} Y \\ \sqrt{n}(Y_n - \theta) \xrightarrow{d} \sigma Z \\ \frac{Y_n - \theta}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathrm{N}(0,1) \\ Y_n \xrightarrow{d} Y \sim \mathrm{N}\left(\theta, \sigma^2/n\right) \end{cases}$$

Osservazione importante 61 (jargon/style). So we can say that a standardized random variable converges to Z, where $Z \sim \mathrm{N}(0,1)$, 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.

Esempio 8.6.2 (Odd example continued)**.** 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 using the generalized CLT

$$\begin{aligned} & Y_n \xrightarrow{d} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) && \text{centering } \dots \\ & Y_n - p \xrightarrow{d} \mathrm{N}\left(0, \frac{p(1-p)}{n}\right) && \text{multiply both by } \sqrt{n} \dots \\ & \sqrt{n}(Y_n - p) \xrightarrow{d} \mathrm{N}(0, p(1-p)) && (1) \\ & \sqrt{n}(Y_n - p) \xrightarrow{d} Z \cdot \sqrt{p(1-p)} && (2) \end{aligned}$$

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-CLT version-style; we need it for the delta method.

Proposizione 8.6.2 (Delta method)**.** *If the generalized CLT holds, that is:*

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} Y$$

we have that

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot Y \tag{8.15}$$

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

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot Y$$

□

Osservazione importante 62 (Motivation recap (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 odd). The question is what is the limit distribution of $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d}$?
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 says that if the generalized clt holds, you have as result the same limit Y multiplied by the derivative of the transformation.

Esempio 8.6.3 (Odd example conclusion)**.** The delta method is a tool that gives us a distribution for the odds. We can apply it since the generalized clt holds, as shown above:

$$Y_n \xrightarrow{d} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) \implies \sqrt{n}(Y_n - p) \xrightarrow{d} \sqrt{p(1-p)}\,\mathrm{N}(0,1)$$

To apply the delta method formula we have to find the first derivative of the transformation

$$g(p) = \frac{p}{1-p}$$
$$g'(p) = \frac{1(1-p) - (-1)p}{(1-p)^2} = \frac{1-p+p}{(1-p)^2} = \frac{1}{1-p}$$

Now we can find the estimator for the odds and also its asymptotic distribution. Now with $\overline{x}$ as our estimator for p we can say that

$$\sqrt{n}(\overline{x} - p) \xrightarrow{d} \mathrm{N}\left(0, p(1-p)\right)$$

and according to the delta method we can say that

$$\begin{aligned}
\sqrt{n}(g(Y_n) - g(\theta)) &\xrightarrow{d} g'(\theta) \cdot Y \\
\sqrt{n}\left(\frac{\overline{x}}{1-\overline{x}} - \frac{p}{1-p}\right) &\xrightarrow{d} \frac{1}{(1-p)^2} \cdot \mathrm{N}\left(0, p(1-p)\right) \\
&\xrightarrow{d} \mathrm{N}\left(0, \frac{p(1-p)}{(1-p)^4}\right) \\
&\xrightarrow{d} \mathrm{N}\left(0, \frac{p}{(1-p)^3}\right)
\end{aligned}$$

Esempio 8.6.4 (Logarithm of the mean)**.** Having $X_1, \ldots, X_n$ are iid with dist $f(x)$ (whatever distribution), $\mathbb{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$ if we take the average $Y_n = \overline{X} = \sum_{i=1}^{n} X_i/n$ as our estimator, with the clt we have the

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{d} \sigma\, \mathrm{N}\left(0, 1\right)$$

Now what is the distribution of the estimator for the logarithm of μ $g(\mu) = \log(\mu)$? Applying the delta method we have:

$$\begin{aligned}
g(\mu) &= \log(\mu) \\
g'(\mu) &= \frac{1}{\mu}
\end{aligned}$$

So:

$$\begin{aligned}
\sqrt{n}(g(\overline{x}) - g(\mu)) &\xrightarrow{d} g'(\mu) \cdot \sigma \cdot \mathrm{N}\left(0, 1\right) \\
\sqrt{n}(\log(\overline{x}) - \log\mu) &\xrightarrow{d} \frac{1}{\mu} \cdot \sigma \cdot \mathrm{N}\left(0, 1\right) \\
&\xrightarrow{d} \mathrm{N}\left(0, \frac{\sigma^2}{\mu^2}\right)
\end{aligned}$$

OR better, in explicit way:

$$\log\overline{x} \xrightarrow{d} \mathrm{N}\left(\log\mu, \frac{\sigma^2}{n\mu^2}\right)$$

Esempio 8.6.5. Let $X_1, \ldots, X_n$ 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 $\overline{X}^2$.
We know that by CLT

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{d} \sigma\, \mathrm{N}\left(0, 1\right)$$

According to Delta method

$$\sqrt{n}(\overline{x}^2 - \mu^2) \xrightarrow{d} g'(\mu)\sigma\, \mathrm{N}\left(0, 1\right)$$

with

$$g(\mu) = \mu^2$$
$$g'(\mu) = 2\mu$$

then we conclude that

$$\begin{aligned}\sqrt{n}(\overline{x}^2 - \mu^2) &\xrightarrow{d} 2\mu\sigma \,\mathrm{N}(0,1) \\ &\xrightarrow{d} \mathrm{N}\left(0, 4\mu^2\sigma^2\right)\end{aligned}$$

Esempio 8.6.6 (Esame vecchio viroli)**.** Let $\hat{\theta}_n$ be an estimator for θ with the following asymptotic distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \sqrt{\theta}\,\mathrm{N}(0,1)$$

Use the delta method to derive the asymptotic distribution of $g(\hat{\theta}_n) = \log \hat{\theta}_n$:

1. $\sqrt{n}(\log \hat{\theta}_n - \log \theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{4}\right)$
2. $\sqrt{n}(\log \hat{\theta}_n - \log \theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\theta}\right)$
3. $\sqrt{n}(\log \hat{\theta}_n - \log \theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\theta^2}\right)$
4. $\sqrt{n}(\log \hat{\theta}_n - \log \theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{2}{\theta}^2\right)$

We have that $g(\theta) = \log(\theta)$ and $g'(\theta) = \frac{1}{\theta}$ so, by the delta method

$$\begin{aligned}\sqrt{n}(\log(\hat{\theta}_n) - \log(\theta)) &\xrightarrow{d} g'(\theta) \cdot \sqrt{\theta}\,\mathrm{N}(0,1) \\ &\xrightarrow{d} \frac{1}{\theta} \cdot \sqrt{\theta}\,\mathrm{N}(0,1) \\ &\xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\theta}\right)\end{aligned}$$

as reported by Bigo as well

Esempio 8.6.7 (Esame vecchio viroli)**.** Let $\hat{\theta}_n$ be an estimator for θ with the following asymptotic distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \sqrt{\theta}\,\mathrm{N}(0,1)$$

Use the delta method to derive the asymptotic distribution of $g(\hat{\theta}_n) = \frac{\hat{\theta}_n^2}{2} + 2$:

1. $\sqrt{n}(\frac{\hat{\theta}_n^2}{2} - \frac{\theta^2}{2}) \xrightarrow{d} \mathrm{N}\left(0, \theta^3\right) + 2$
2. $\sqrt{n}(\frac{\hat{\theta}_n^2}{2} - \frac{\theta^2}{2}) \xrightarrow{d} \mathrm{N}\left(0, \theta^3\right)$
3. $\sqrt{n}(\frac{\hat{\theta}_n^2}{2} - \frac{\theta^2}{2}) \xrightarrow{d} \mathrm{N}\left(0, \frac{\theta^2}{2}\right)$
4. $\sqrt{n}(\frac{\hat{\theta}_n^2}{2} - \frac{\theta^2}{2}) \xrightarrow{d} \mathrm{N}\left(0, \frac{\theta^4}{4}\right)$

qui si ha che $g(x) = \frac{x^2}{2} + 2$ da cui $g'(x) = x$ e $g'(\theta) = \theta$. Per cui

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} g'(\theta)\sqrt{\theta}\,\mathrm{N}(0,1)$$
$$\sqrt{n}\left(\frac{\hat{\theta}^2}{2} + 2 - \frac{\theta^2}{2} - 2\right) \xrightarrow{d} \theta^{\frac{3}{2}}\,\mathrm{N}(0,1)$$
$$\sqrt{n}\left(\frac{\hat{\theta}^2}{2} - \frac{\theta^2}{2}\right) \xrightarrow{d} \mathrm{N}\left(0, \theta^3\right)$$

Esempio 8.6.8 (Esame vecchio viroli)**.** Let $\hat{\theta}_n$ be an estimator for θ with the following asymptotic distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \frac{2}{\theta}\,\mathrm{N}(0,1)$$

Use the delta method to derive the asymptotic distribution of $g(\hat{\theta}_n) = \sqrt{\hat{\theta}_n}$:

- $\sqrt{n}(\sqrt{\hat{\theta}_n} - \sqrt{\theta}) \xrightarrow{d} \mathrm{N}\left(0, \frac{4}{boh}\right)$
- $\sqrt{n}(\sqrt{\hat{\theta}_n} - \sqrt{\theta}) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\theta_3}\right)$
- $\sqrt{n}(\sqrt{\hat{\theta}_n} - \sqrt{\theta}) \xrightarrow{d} \mathrm{N}\left(0, \frac{2}{boh}\right)$
- $\sqrt{n}(\sqrt{\hat{\theta}_n} - \sqrt{\theta}) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\theta^2}\right)$

qui si ha $g(x) = \sqrt{x}$, $g'(x) = \frac{1}{2\sqrt{x}}$ e $g'(\theta) = \frac{1}{2\sqrt{\theta}}$. Da cui

$$\sqrt{n}(\sqrt{\hat{\theta}_n} - \sqrt{\theta}) \xrightarrow{d} \frac{1}{2\sqrt{\theta}}\frac{2}{\theta}\,\mathrm{N}(0,1)$$
$$\xrightarrow{d} \mathrm{N}\left(0, \theta^{-3}\right)$$

Esempio 8.6.9. Let $X_1, \ldots, X_n$ be independent Geom (p)

1. Does $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converge in probability?
2. what is its limiting distribution?
3. and what is the distribution of $\frac{1}{\overline{X}}$

We have that

1. According to the WLLN $\overline{X} \xrightarrow{p} \mathbb{E}[X] = \frac{1}{p}$.
2. The limiting distribution can be derived by the CLT

$$\sqrt{n}\left(\overline{X} - \frac{1}{p}\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{1-p}{p^2}\right)$$

with $\frac{1-p}{p^2}$ as variance.

3. The limiting distribution of $\frac{1}{\overline{X}}$ can be found by the Delta method. We have that

$$g(x) = \frac{1}{x}$$
$$g'(x) = -\frac{1}{x^2}$$

So considering $\theta = \frac{1}{p}$ we have that

$$\sqrt{n}\left(\overline{X} - \frac{1}{p}\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{1-p}{p^2}\right)$$
$$\Longrightarrow$$
$$\sqrt{n}\left(\frac{1}{\overline{X}} - p\right) \xrightarrow{d} g'(\theta)\,\mathrm{N}\left(0, \frac{1-p}{p^2}\right)$$
$$\sqrt{n}\left(\frac{1}{\overline{X}} - p\right) \xrightarrow{d} -\frac{1}{(1/p)^2}\,\mathrm{N}\left(0, \frac{1-p}{p^2}\right)$$
$$\sqrt{n}\left(\frac{1}{\overline{X}} - p\right) \xrightarrow{d} -p^2\,\mathrm{N}\left(0, \frac{1-p}{p^2}\right)$$
$$\sqrt{n}\left(\frac{1}{\overline{X} - p}\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{p^4(1-p)}{p^2}\right)$$

Esempio 8.6.10. Let $X_1, \ldots, X_n$ a sequence of independent rvs with $X \sim \mathrm{Exp}(\theta)$. Let $T_n = \sum_{i=1}^n \frac{X_i}{2n}$

1. Does T_n converge in probability?
2. Find the limiting distribution of T_n by CLT
3. find the limiting distribution of $\log(T_n)$

For

1. the convergence in probability we have that

$$\mathbb{E}[T_n] = \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{2n} = \frac{\sum_{i=1}^n \frac{1}{\theta}}{2n} = \frac{1}{2\theta}$$
$$\mathrm{Var}[T_n] = \frac{1}{4n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{n}{4n^2\theta^2} = \frac{1}{4n\theta^2}$$

   therefore $T_n \xrightarrow{p} \delta_{1/2\theta}$

2. for the convergence in distribution by CLT let's first study $T_n^* = 2T_n = \frac{\sum_{i=1}^n X_i}{n} = \overline{X}$. By CLT

$$\sqrt{n}(\overline{X} - 1/\theta) \xrightarrow{d} \mathrm{N}\left(0, \frac{1}{\theta^2}\right)$$

since

$$\frac{\overline{X} - 1/\theta}{\frac{1}{\sqrt{n}\theta}} \xrightarrow{d} \mathrm{N}(0,1)$$

with $\mathbb{E}\left[\overline{X}\right] = \frac{1}{\theta}$, $\mathrm{Var}\left[\overline{X}\right] = \frac{1}{n\theta^2}$. So by the continuous mapping theorem

$$\frac{T_n - 1/2\theta}{\frac{1}{2\sqrt{n}\theta}} \xrightarrow{d} \mathrm{N}(0,1)$$

and the generalized form is

$$\sqrt{n}\left(T_n \frac{1}{2\theta}\right) \xrightarrow{d} \frac{1}{2\theta} \cdot \mathrm{N}(0,1)$$

3. for the convergence of $\log T_n$, by the delta method

$$\begin{aligned}\sqrt{n}\left(\log T_n - \log \frac{1}{2\theta}\right) &\xrightarrow{d} g'(\theta) \cdot \frac{1}{2\theta}\,\mathrm{N}(0,1) \\ &\xrightarrow{d} 2\theta \frac{1}{2\theta}\,\mathrm{N}(0,1)\end{aligned}$$

Chapter 9

Rigo stuff

9.1 Convergence

Osservazione importante 63. 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 below are meant for $n \to +\infty$.

Osservazione importante 64. 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.

Definizione 9.1.1 (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

$$\mathbb{P}\left(\omega \in \Omega : X_n(\omega) \to X(\omega)\right) = 1$$

   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 X_n have moment of order p: $\mathbb{E}\left[|X_n|^p\right] < +\infty$;

   (b) X has moment of order p as well: $\mathbb{E}\left[|X|^p\right] < +\infty$;

   (c) and most importantly

$$\mathbb{E}\left[|X_n - X|^p\right] \to 0$$

   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.

Osservazione 267. Qui l'immagine delle implicazioni sulle convergenze

Osservazione importante 65. 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 know/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 that we can prove is still 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.

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 occurs if X is degenerate. Hence if $X = a$ almost surely (is degenerate) we obtain $X_n \xrightarrow{p} X \iff X_n \xrightarrow{d} X$

3. the definition we gave regarding convergence in distribution may appear strange. It may seem more natural require convergence for all x (not only where F is continuous) that is:
$$\lim_{n\to+\infty} F_{X_n}(x) = F_X(x), \qquad \forall x \in \mathbb{R}$$
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, for these degenerate, the distribution functions are:
$$F_{X_n}(x) = \begin{cases} 1 & \text{if } x \geq \frac{1}{n} \\ 0 & \text{if } x < \frac{1}{n} \end{cases}, \quad F_X(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$
So the value of F at 0 is $F_X(0) = 1$, while for $F_{X_n}(0) = 0$. Therefore
$$\lim_{n\to+\infty} F_{X_n}(0) \neq F_X(0), \quad F_n(0) \not\to F(0)$$
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.

TODO: non chiarissimo qui

Osservazione 268. Now some counterexamples to show that some double implications don't work (as stated in the graph of convergence implications).

Esempio 9.1.1. Let $\mathbb{P}(X_n = 0) = \frac{n-1}{n}$, $\mathbb{P}(X_n = n) = \frac{1}{n}$ and $X = 0$. Then $X_n \xrightarrow{p} X$ but $X_n \not\xrightarrow{L_1} X$: here there's convergence in probability but not in L_p. Its an contraexample of why the implication does not hold:

- given $\varepsilon > 0$

$$\begin{aligned}\mathbb{P}(|X_n - X| > \varepsilon) &= \mathbb{P}(|X_n| > \varepsilon) \\ &\overset{(1)}{=} \mathbb{P}(|X_n| > \varepsilon \cap X_n = 0) + \mathbb{P}(|X_n| > \varepsilon \cap X_n = n) \\ &\leq 0 + \mathbb{P}(X_n = n) = \frac{1}{n}\end{aligned}$$

  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 $X_n \not\xrightarrow{L_p} X$

TODO: non capisco perché $\leq$ all'ultimo, suggerito al prof dalla matematica gnocca

Esempio 9.1.2. To prove that convergence in Lp implies convergence in probability it suffices to use Tchebychev inequality. Suppose infact that $X_n \xrightarrow{L_p} X$, then given $\varepsilon > 0$ to have convergence in probability $\mathbb{P}(|X_n - X| > \varepsilon)$ must go to 0. Now we have that an upper bound for $\mathbb{P}(|X_n - X| > \varepsilon)$ is

$$\mathbb{P}(|X_n - X| > \varepsilon) \overset{(1)}{\leq} \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \overset{(2)}{\to} 0$$

where (1) due to Tchebychev and (2) since by definition/assumption on $X_n \xrightarrow{L_p} X$. So given that the right part goes to 0, even the left part goes to 0 and saying that means that there is convergence in probability.

Esempio 9.1.3. An (counter-)example where A.s. convergence does not imply L_1 convergence. Considering the space:

$$(\Omega, \mathscr{A}, \mathbb{P}) = ([0,1], \mathscr{B}([0,1]), m)$$

with m the Lebesgue measure. In general the Lebesuge meassure is *not* a probability measure, because on the real line it gives $+\infty$; but if defined on 0-1 its max is 1 so can be a probability measure.
We define also

$$\begin{aligned}X_n &= n \cdot \mathbb{1}_{[0,\frac{1}{n}]}(\omega) \\ X &= 0\end{aligned}$$

Here by construction we have that $\omega \in [0,1]$; if

- $\omega \in (0,1]$ then $\omega > \frac{1}{n}$ for large n. Therefore for large n we have that $X_n(\omega) = 0$.

- $\omega = 0$, we have that $X(0) = n\mathbb{1}_{[0,\frac{1}{n})}(0) = n$ that goes to $+\infty$ as $n \to +\infty$.

Hence

$$\begin{aligned}\mathbb{P}\left(\omega \in \Omega : X_n(\omega) \to X(\omega)\right) &= \mathbb{P}\left(\omega \in \Omega : X(\omega) = 0\right) = \mathbb{P}(0,1] \\ &= m(0,1] = 1 - 0 = 1\end{aligned}$$

That is $X_n \xrightarrow{a.s.} X$.
However:

$$\begin{aligned}\mathbb{E}\left[|X_n - X|\right] = \mathbb{E}\left[|X_n|\right] &= \mathbb{E}\left[n \cdot \mathbb{1}_{[0,1/n)}(\omega)\right] = n \cdot \mathbb{E}\left[\mathbb{1}_{[0,1/n)}(\omega)\right] \\ &= n \cdot \mathbb{P}\left([0,1/n)\right) = n \cdot m[0,1/n) = n \cdot \frac{1}{n} \\ &= 1\end{aligned}$$

Hence here $X_n \xrightarrow{a.s.} X$ but $X_n \not\xrightarrow{L_1} X$.

Esempio 9.1.4. An example where convergence in distribution does not imply convergence in probability. Considering the same space:

$$(\Omega, \mathscr{A}, \mathbb{P}) = \left([0,1], \mathscr{B}([0,1]), m\right)$$

now define $X_n = \mathbb{1}_{[0,1/2]}(\omega)$ and $X = \mathbb{1}_{(1/2,1]}(\omega)$. In this case we have that

$$|X_n - X| = 1, \quad \forall n$$

so X_n fails to converge to X in probability: $X_n \not\xrightarrow{p} X$. However the distribution functions are:

$$F(x) = \mathbb{P}\left(X \leq x\right) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

and the other is the same:

$$F_n(x) = \mathbb{P}\left(X \leq x\right) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Hence since $F_n = F$, $\forall n$, we have that $X_n \xrightarrow{d} X$.

9.2 Laws of large numbers

Osservazione 269. There are several, some are more famous/attractive, but in general there are many.

Definizione 9.2.1. Let $\{X_n\}$, be a sequence of real rvs. We say it satisfies the law of large number if the sample mean $\overline{X} = \frac{1}{n}\sum_i X_i$ converges to V for some random variable V:

- if it converges in probability, $\overline{X} \xrightarrow{p} V$, we speak of *weak law of large number*;
- if instead $\overline{X} \xrightarrow{a.s.} V$ we speak of *strong low of large numbers.*

Osservazione 270. Roughly speaking, any time we prove sample mean converges to a limit we have a law of large number. There are research papers that discover new large of large numbers frequently: they simply prove that a sample mean of certain sequences $X_1, \ldots, X_n$ converges to something.

Osservazione importante 66. In general the limit V is an arbitrary real rv; however the most *important special case* is when the rvs have all the same mean $\mathbb{E}[X_i], \forall i$ and $V = \delta_{\mathbb{E}[X_i]}$.
The most important strong law of large number is the strong law dued to Kolmogorov: it's what most people think about law of large number.

Teorema 9.2.1 (Kolmogorov strong law of large numbers)**.** *If $\{X_n\}$ is iid and $\mathbb{E}[|X_1|] < +\infty$, then $\overline{X_n} \xrightarrow{a.s.} \mathbb{E}[X_1]$.*

Osservazione 271. Another strong law of large number is the following. These are examples of laws of large number which are different for the assumptions (but again the sample mean converges to a certain limit). Compared to Kolmogorov, here we drop the iid hypothesis and replace it with some other condition.

Teorema 9.2.2 (A second example of strong lln)**.** *Given a sequence $\{X_n\}$, if*

- $\mathbb{E}\left[X_n^2\right] \leq c$, $\forall n$, *where c is a fixed constant*
- *the random variable have the same mean* $\mathbb{E}[X_1] = \mathbb{E}[X_n]$
- $\mathrm{Cov}(X_i, X_j) \leq 0$, $\forall i \neq j$

then $\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X_1]$

Esempio 9.2.1. Let's prove the above laws (only convergence in probability, the almost sure is easier).
It suffices to apply Tchebychev inequality: given $\varepsilon > 0$ we have that

$$\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| > \varepsilon\right) \leq \mathrm{Var}\left[\overline{X}_n\right]$$

now we evaluate the variance

$$\begin{aligned}\mathrm{Var}\left[\overline{X}_n\right] &= \frac{1}{\varepsilon^2}\frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \frac{1}{n^2\varepsilon^2}\left\{\sum_{i=1}^n \mathrm{Var}[X_i] + 2\sum_{1\leq i<j\leq n}\mathrm{Cov}(X_i, X_j)\right\} \\ &= \frac{1}{n^2\varepsilon^2}\sum_{i=1}^n \mathrm{Var}[X_i] \leq \frac{1}{n^2\varepsilon^2}\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \\ &\leq \frac{nc}{n^2\varepsilon^2} = \frac{c}{\varepsilon^2}\frac{1}{n} \to 0\end{aligned}$$

This proves that $\overline{X_n} \xrightarrow{p} \mathbb{E}[X_1]$. Indeed, as claimed in the thorem, one also obtains $X_n \xrightarrow{a.s.} \mathbb{E}[X_1]$ but we will not prove almost sure convergence (the latter fact).

Osservazione 272. In the next example we have a strong law but the limit is not the mean.

Esempio 9.2.2. A sequence $\{X_n\}_{n\in N}$ is said to be *stationary* if the probability distribution of the sequence starting from two, is the same of the distribution of the unshifted sequence:

$$(X_2, X_3, X_4, \ldots) \sim (X_1, X_2, X_3, \ldots)$$

hence the probability distribution of the sequence is invariant (doesn't change under shifts); in some framework this is the classical assumptions. Here the following result holds.
If X_n is stationary $\mathbb{E}[|X_1|] < +\infty$ (mean of X_1 exists), then we have almost sure convergence to V: $\overline{X}_n \xrightarrow{a.s.} V$ where V is a not necessarily degenerate rv.

Osservazione 273. Two reasons why we mention the result above:

1. stationarity is an important assumption (like iid)
2. this is an example where we have a strong lln (being the convergence as) but the limit is not the mean (this does not need to be the case).

Osservazione 274. Finally we state a weak law of large numbers.

Proposizione 9.2.3. *If $\{X_n\}$ is iid and the characteristic function of X_1 is differentiable at point 0, that is exists $\phi_{X_1}(0)'$, then $X_n \xrightarrow{p} \alpha$ for some constant α.*

Osservazione importante 67. Some remarks regarding this latter:

1. if $\mathbb{E}[|X_1|] < +\infty$ (X_1 has the mean) then $\alpha = \mathbb{E}[X_1]$ and convergence is almost sure and not only in probability (by the strong law of Kolmogorov). However, it may be that the characteristic function has first derivative at point zero even if $\mathbb{E}[|X_1|] < +\infty$; in this case we have the weak law of large number but not the strong one. Let's make an example for this: let X be an absolutely continuous random variable with density

$$f(x) = \begin{cases} \frac{c}{x^2 \log|x|} & \text{if } x \notin [-2, 2] \\ 0 & \text{if } x \in [-2, 2] \end{cases}$$

where c is the normalizing constant. Then:

$$\begin{aligned} \mathbb{E}[|X|] &= \int_{-\infty}^{+\infty} |x|\, f(x)\, \mathrm{d}x \overset{(1)}{=} 2c \int_2^{\infty} \frac{x}{x^2 \log x}\, \mathrm{d}x = 2c \int_2^{+\infty} \frac{1}{x \log x}\, \mathrm{d}x \\ &= +\infty \end{aligned}$$

where (1) because it's an even function. So this random variable does not have mean.
However it can be show that the characteristic function of X has the first derivative at 0, so $\exists \phi_X(0)'$. Hence if X_n is iid and $X_1 \sim X$ (common distribution is X), we have that $\overline{X}_n \xrightarrow{p} \alpha$ for some α but we don't have any strong law of large number. It can be also shown that, in this example, $\alpha = 0$.

2. the previous weak law of large number is very easy to prove. Suppose infact $\{X_n\}$ is iid and exists the first derivative in point 0 of the characteristic function. Then the characteristic function of the sample mean is:

$$\phi_{\overline{X}_n}(t) = \mathbb{E}\left[e^{i\frac{t}{n}\sum_{i=1}^n X_i}\right] = \phi_{\sum_{i=1}^n X_i}\left(\frac{t}{n}\right) 27 \overset{(\perp\!\!\!\perp)}{=} \prod_{i=1}^n \phi_{X_i}\left(\frac{t}{n}\right)$$
$$\overset{(1)}{=} \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n$$

in (1) equally distributed. Now we apply Taylor up to the first order

$$\phi_{\overline{X}_n}(t) = \left[\phi_{X_1}(0) + \frac{t}{n}\phi_{X_1}(0)' + \sigma\left(\frac{t}{n}\right)\right]^n = \left[1 + \frac{t\phi_{x_1}(0)' + n\sigma\left(\frac{t}{n}\right)}{n}\right]^n$$

now, using the fact that if $a_n \to a$ then $\left(1 + \frac{a_n}{n}\right)^n \to e^a$, we have that

$$\lim_{n\to\infty}\left[1 + \frac{t\phi_{x_1}(0)' + n\sigma\left(\frac{t}{n}\right)}{n}\right]^n = e^{t\phi_{X_1}(0)'}$$

Finally it can be shown that the derivative of the characteristic function at point 0 (provided it exists) is equal to

$$\phi_{X_1}(0)' = i\alpha, \quad \alpha \in \mathbb{R}$$

hence the characteristic function for the sample limit converge as follows

$$\phi_{\overline{X}_n}(t) \to e^{it\alpha}, \quad \forall t \in \mathbb{R}$$

which is the characteristic function of a degenerate/dirac random variable $X = \alpha$. So using the properties of the characteristic function, we get that $\overline{X}_n \xrightarrow{d} \alpha$, but since α is degenerate, we also get that sample mean converges to α not only in distribution but also in probability $X_n \xrightarrow{p} \alpha$.

Esempio 9.2.3 (A very classical example)**.** We have an urn of black an white balls. The proportion p of white balls is not known. To make inference on p, we make a sequence of drawings *with* replacement. Let:

$$X_i = \begin{cases} 1 & \text{if } \textit{white} \text{ ball drawn at trial } i \\ 0 & \text{if } \textit{black} \text{ ball drawn at trial } i \end{cases}$$

Since the drawing are with replacement sequence (X_i) are iid, and $\mathbb{E}[X_1] = p$. Hence by Kolmogorov's strong law we obtain that the sample mean converges to p, that is $\overline{X_n} \xrightarrow{a.s.} p$. Kolmogorov only confirm a fact that our intuition considered obvious.
This example can be generalized as follows: let $\{X_n\}$ be iid but the distribution function F of X_1 is unknown. To make inference on F we fix a real number $x \in \mathbb{R}$ and we define the following random indicator variables

$$Y_i = \mathbb{1}_{\cdot}(X_i \leq x)$$

Then $\{Y_i\}$ are still iid and

$$\mathbb{E}[Y_1] = \mathbb{E}[\mathbb{1}_{\cdot}(X_1 \leq x)] = \mathbb{P}(X_1 \leq x) = F(x)$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\cdot}(X_1 \leq x) \xrightarrow{a.s.} F(x)$$

In general:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\cdot}(X_1 \leq x)$$

is called the *empirical distribution function*, and can be regarded as an estimate of F. Infact, in statistical terms, the empirical distribution function is a *consistent* estimator of the true distribution function (that is, as the sample size goes $n \to \infty$, the procedure converge to the true value).

9.3 Central limit theorem

9.3.1 CLT

Osservazione 275. Big topic of probability, one of the main findings with law of large numbers.

Osservazione 276. In reality there are several CLTs, all fullfill the following general definition.

Definizione 9.3.1. Given a sequence $X_1, X_2 \dots$ of real random variable, we say that the sequence $\{X_n\}$ satisfies the CLT if there are two constants $a_n \in \mathbb{R}$ and $b_n > 0$ such that

$$\frac{\sum_{i=1}^{n} X_i - a_n}{b_n} \xrightarrow{d} \mathrm{N}(0,1)$$

Osservazione importante 68 (CLT of standardized sum). The sequence $\{X_n\}$ is arbitrary we need to find a_n and b_n for the ratio above to go in distribution to the standard normal.
The constants a_n and b_n are generally arbitrary but the main/most important special case is when:

$$a_n = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \quad \text{mean of the sum}$$

$$b_n = \sigma\left(\sum_{i=1}^{n} X_i\right) \quad \text{sd of the sum}$$

Under these choices we have that the standardization of the sum

$$\frac{\sum_{i=1}^{n} X_i - a_n}{b_n}$$

fullfill the CLT definition

Osservazione 277 (Natural/tipical application of CLT). One can think of sequence $X_1, X_2, \dots$ as the sequence of observation. We are interested in the probability distribution of the sum $\sum_{i=1}^n X_i$ (but we don't know it/aren't able to evaluate it). However if CLT holds, a possibility is to replace such an unknown distribution with a normal distrubution with mean a_n and variance b_n^2, that is $\mathrm{N}\left(a_n, b_n^2\right)$.
Why this is true? If n is large the CLT implies that the distribution of standardized sample mean is close to standard normal

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \sim \mathrm{N}(0,1)$$

so that the distribution of the sum is close to

$$\sum_{i=1}^n X_i \sim a_n + b_n \mathrm{N}(0,1) = \mathrm{N}\left(a_n, b_n^2\right)$$

If I adopt the normal for a fixed n we surely make an error, the distribution is not normal: but the distribution becomes normal as n gets larger, and the error smaller.

Osservazione 278. Now we start with some examples of CLT: in case LLN the most important is Kolmogorov one, similarly in CLT the main/most popular statement of this kind is the so-called CLT1.

Proposizione 9.3.1 (CLT1)**.** *If $\{X_n\}$ is sequence of iid rvs with $\mathbb{E}\left[X_i^2\right] < +\infty$ (finite second moments) and X_i is not degenerate, then the*

$$\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sigma(\sum_{i=1}^n X_i)} \xrightarrow{d} \mathrm{N}(0,1)$$

Proof. Let ϕ denote the characteristic function of $\frac{(X_1 - \mathbb{E}[X_1])}{\sigma(X_1)}$. Here I can divide for standard deviation cause looking at the assumption, the rv is not degenerate so the variance is positive. Then let's evaluate:

$$\begin{aligned} Z_n &= \frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sigma(\sum_{i=1}^n X_i)} \overset{(iid)}{=} \frac{\sum_{i=1}^n \left(X_i - \mathbb{E}\left[X_i\right]\right)}{\sqrt{n \operatorname{Var}\left[X_1\right]}} \\ &= \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n X_i - \mathbb{E}\left[X_i\right]}{\sigma(X_i)} \end{aligned}$$

Hence the characteristic function of Z_n is

$$\begin{aligned} \phi_{Z_n}(t) &= \phi_{\frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n X_i - \mathbb{E}[X_i]}{\sigma(X_i)}}(t) = \phi_{\frac{\sum_{i=1}^n X_i - \mathbb{E}[X_i]}{\sigma(X_i)}}\left(\frac{t}{\sqrt{n}}\right) \\ &\overset{(1)}{=} \left[\phi_{\frac{X_i - \mathbb{E}[X_i]}{\sigma(X_i)}}\left(\frac{t}{\sqrt{n}}\right)\right]^n \end{aligned}$$

where in (1) since the rv are independent the characteristic function of the sum is the product of the char function, and being identically distributed we have the power.
Now as in the weak LLN proof, we use that rv by assumption have second

moment finite; so we can say that the its characteristic function is C^2 and we can apply Taylor expansion (up to the the second order). So by Taylor (with Peano remainder):

$$\phi_{Z_n}(t) = \left[\phi(0) + \frac{t}{\sqrt{n}}\phi'(0) + \frac{t^2}{n}\frac{1}{2}\phi''(0) + o\left(\frac{t^2}{n}\right)\right]^n$$

Now since second moment exists, it exist the first as well and in the previous step we did the substitution following,

$$\phi'(0) = i\,\mathbb{E}\left[\frac{X_1 - \mathbb{E}[X_1]}{\sigma(X_1)}\right] \overset{(1)}{=} 0$$
$$\phi''(0) = i^2\,\mathbb{E}\left[\left(\frac{X_i - \mathbb{E}[X_i]}{\sigma(X_i)}\right)^2\right] \overset{(1)}{=} -1 \cdot 1 = -1$$

where in (1) the substitution are done considering that the expectation of a standardized variable is zero while its second moment 1. So we have

$$\phi_{Z_n}(t) = \left[1 + 0 - \frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right]^n = \left[1 + \frac{-t^2/2 + n \cdot o\left(\frac{t^2}{n}\right)}{n}\right]^n$$

now, as $n \to +\infty$, considering that if $a_n \to a$ then $\left(1 + \frac{a_n}{n}\right)^n \to e^a$, then overall it suffices to let $a_n = -\frac{t^2}{2} + n \cdot o\left(\frac{t^2}{n}\right) \to -\frac{t^2}{2}$ the characteristic function converges to

$$\phi_{Z_n}(t) \to e^{-\frac{t^2}{2}}$$

which is the characteristic function of the standard normal, and this concludes the proof. □

Osservazione 279. Let's see another CLT. There are several other version of the CLT btw.

Proposizione 9.3.2 (CLT2)**.** *If (X_n) are independent, with $\mathbb{E}[X_n] = 0, \forall n$ and it holds the following strange stuff*

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|X_i|^3\right]}{\left(\sum_{i=1} n\,\mathbb{E}[X_i^2]\right)^{\frac{3}{2}}} \to 0$$

then (qui sotto non sottraiamo la media perché 0 per ipotesi)

$$\frac{\sum_{i=1}^n X_i}{\sigma(\sum_{i=1}^n X_i)} \xrightarrow{d} \mathrm{N}(0,1)$$

TODO: check here che al denominatore il quadrato sia della variabile o del valore atteso

Osservazione 280. Here the conclusions are the same as the CLT1, the differences are in the preconditions. What is the very big assumption different from the first case?
It's that here the rvs are not forced to be identically distributed. So we need to replace that assumption with the new strange condition (dont' try to attach a meaning to this condition: it's just a technical condition for the theorem to

hold).
So this second example is **useful because** it can be used when X_i are not identically distributed.
In the following some examples.

Esempio 9.3.1. Suppose (X_n) be independent, all the rvs with null mean $\mathbb{E}[X_n] = 0, \forall n$, $|X_n| \leq c, \forall n$ and $\sum_{i=1}^n X_i^2 \to +\infty$. We are interested in convergence in distribution of

$$Z_n = \frac{\sum_i X_i}{\sigma(\sum_{i=1}^n X_i)}$$

The tool to study convergence in distribution for sum/mean is clt; in these cases we use the second version because we didn't say they are identically distributed. The random variable are independent, their mean is zero, thus in order to conclude that Z_n converge to standard normal is enough to verify the "strange condition"; to answer we note that

$$|X_i|^3 = |X_i| X_i^2 \overset{(1)}{\leq} cX_i^2$$

with (1) by assumptions. Hence:

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|X_i|^3\right]}{(\sum_{i=1}^n \mathbb{E}[X_i^2])^{\frac{3}{2}}} \leq \frac{\sum_{i=1}^n \mathbb{E}\left[c \cdot X_i^2\right]}{(\sum_{i=1}^n \mathbb{E}[X_i^2])^{\frac{3}{2}}} = \frac{c\sum_{i=1}^n X_i^2}{(\sum_{i=1}^n \mathbb{E}[X_i^2])^{\frac{3}{2}}} = \frac{c}{(\sum_{i=1}^n \mathbb{E}[X_i^2])^{\frac{1}{2}}}$$

and this latter $\to 0$ since the denominator goes to $+\infty$ by assumption. So since $\frac{\sum_{i=1}^n \mathbb{E}[|X_i|^3]}{(\sum_{i=1}^n \mathbb{E}[X_i^2])^{\frac{3}{2}}}$ is upper bunded by 0, the strange condition goes to 0 as well.
Hence $Z_n \xrightarrow{d} \mathrm{N}(0,1)$

Esempio 9.3.2. Suppose $\{X_n\}$ is iid with $\mathbb{E}[X_i] = 0$ and second moment $\mathbb{E}\left[X_i^2\right] = 2$, so variance $\mathrm{Var}[X_i] = 2$. We're interested in convergence in distribution of this ratio:

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n X_i^2}}$$

We use clt1 because of iid rvs. In fact Z_n can be written as (by dividing by $\sqrt{n}$ both numerator and denominator):

$$Z_n = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}}$$

and dividing by $\sqrt{2}$ as well both numerator and denominator

$$\ldots = \frac{\frac{1}{\sqrt{2}\sqrt{n}}\sum_{i=1}^n X_i}{\frac{1}{\sqrt{2}}\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}} \xrightarrow{d} \frac{\mathrm{N}(0,1)}{\frac{1}{\sqrt{2}}\sqrt{\mathbb{E}[X_i^2]}} = \frac{\mathrm{N}(0,1)}{\frac{1}{\sqrt{2}}\sqrt{2}} = \mathrm{N}(0,1)$$

TODO: non chiaro sto esempio di merda

in fact since (X_n) is iid (X_n^2) is iid as well. Moreover $\mathbb{E}\left[X_1^2\right] = 2 < \infty$. Hence the strong law of large number yields:

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{a.s.} \mathbb{E}\left[X_i^2\right] = 2$$

Osservazione 281. In the above example as in the proof of CLT1, among other things, we used that if X_n is iid

NB: boh sta cons azione . . .

$$\frac{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]}{\sigma(\sum_{i=1}^n X_i)} = \frac{\sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right])}{\sqrt{n}\sigma(X_i)} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

where $\sigma = \sigma(X_i)$ and $\mu = \mathbb{E}\left[X_i\right]$. In many theorem we write the quantity in that way.
Now $\sqrt{n} \to +\infty$ while $\overline{X_n} - \mu \xrightarrow{a.s.} 0$ if X_n is iid (and the moment exists).

$$\underbrace{\frac{\sqrt{n}}{\sigma}}_{\to +\infty} \cdot \underbrace{(\overline{X}_n - \mu)}_{\xrightarrow{a.s.} 0} \xrightarrow{d} \mathrm{N}(0,1)$$

Esempio 9.3.3. Suppose $\{X_n\}$ iid and

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{\alpha_i}{2}$$
$$\mathbb{P}(X_i = 0) = 1 - \alpha_i$$

$\forall i$. Let's find conditions on the constant α_i under which

$$Z_n = \frac{\sum_{i=1}^n X_i}{\sigma(\sum_{i=1}^n X_i)} \xrightarrow{d} \mathrm{N}(0,1)$$

These rvs can take only three values. We have that:

$$\mathbb{E}\left[X_i\right] = 0 \cdot \mathbb{P}(X_i = 0) + 1 \cdot \mathbb{P}(X_i = 1) + (-1)\,\mathbb{P}(X_i = -1) \overset{(1)}{=} 0$$

with (1) due to the fact that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1)$.
Moreover $|X_i| \leq c, \forall i$ if $c = 1$. Hence by example 1 (localizzalo) $Z_n \xrightarrow{d} \mathrm{N}(0,1)$ if the sum $\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \to +\infty$. What is $\mathbb{E}\left[X_i^2\right]$? remembering that X_i^2 values are 0 and 1 we have that

$$\mathbb{E}\left[X_i^2\right] = 1 \cdot P(X_i^2 = 1) = \ldots = \alpha_i$$

Since $\mathbb{E}\left[X_i^2\right] = \alpha_i$, we finally obtain

$$\sum_{i=1}^n \alpha_i \to +\infty$$

So this condition imply that $Z_n \xrightarrow{d} \mathrm{N}(0,1)$.

We now prove that the converse holds, that is

$$Z_n \xrightarrow{d} \mathrm{N}(0,1) \implies \sum_{i=1}^n \alpha_i \to +\infty$$

Toward the contraddiction suppose that $\sum_{i=1}^n \alpha_i \not\to +\infty$; the sum is then

$$\alpha = \lim_{n\to\infty} \sum_{i=1}^n \alpha_i = \sum_1^\infty \alpha_i < +\infty$$

why this limit exists finite? this sequence is increasing: we are summing non negative constants α_i, thus this sequance $\sum_{i=1}^n \alpha_i$ is an increasing sequence and of course an increasing sequence has limit equal to the sup. Now consider

$$\sum_{i=1}^n X_i = \sum_{i=1}^n X_i \cdot \frac{\sum_{i=1}^n \alpha_i}{\sum_{i=1}^n \alpha_i} = \underbrace{\sum_{i=1}^n \alpha_i}_{\to \alpha} \cdot \underbrace{\frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n \alpha_i}}_{\xrightarrow{d} \mathrm{N}(0,1)} \xrightarrow{d} \alpha\, \mathrm{N}(0,1) = \mathrm{N}\left(0, \alpha^2\right)$$

So under the assumption above the sequence go in distribution to a normal. But this is a contraddition: X_i can assume only integer values (0, 1, -1) and so will be the sum $\sum_{i=1}^n X_i \in \mathbb{Z}$.
However probability of the integer under the standard normal is 0, and for a infinite countable set of points it will be the same (integers under normal have probability 0): that is, if $U \sim \mathrm{N}(0,1)$ then $\mathbb{P}(U \in \mathbb{Z}) = 0$. So the sum of these variables cannot coverge in distribution to the standard normal which has domain on $\mathbb{R}$.
In questo esempio la condizione per la convergenza alla normale non solo e sufficiente ma anche necessaria.

9.3.2 Berry-Esseen theorem

Osservazione importante 69. One of reason of importance of CLT is a result which allows to evaluate the error we make in adopting the normal distribution for the sum of random variables.

Teorema 9.3.3 (Berry Theorem)**.** *If*

- $\{X_n\}$ *is iid*
- X_1 *is non degenerate*
- $\mathbb{E}\left[|X_1|^3\right] < +\infty$

(condition of clt1 + existence third moment). Then consider the difference/error at point x*:*

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \le x\right) - \Phi(x)$$

where Φ *is distribution function of* $\mathrm{N}(0,1)$.
By CLT1 the first term $\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \le x\right)$ *goes to standard normal* $\Phi(x)$ *so the difference above goes to 0. But the error we make using the normal is:*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \le x\right) - \Phi(x) \right| \le c \cdot \mathbb{E}\left[\left|\frac{X_i - \mu}{\sigma}\right|^3\right] \frac{1}{\sqrt{n}}$$

where $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \mathrm{Var}[X_i]$, *and* c *is a constant such that* $c < \frac{1}{2}$.

Osservazione importante 70. This theorem is useful because, when CLT1 does hold we know that

$$\mathbb{P}\left(\frac{\sum X_i - \mu n}{\sigma\sqrt{n}} \le x\right) \to \Phi(x), \quad \forall x \in \mathbb{R}$$

But thanks to this result we can say more. For every x the error we make by applying standard normal instead of its upper bounded

Osservazione 282. Tipical situation: we don't know $\mathbb{P}\left(\frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \leq x\right)$ (the distribution function of the standardized sum) so we replace it with $\mathrm{N}(0,1)$. The error we make is supped by the upper bound

$$\leq \frac{1}{2}\,\mathbb{E}\left[\left(\left|\frac{X_i-\mu}{\sigma}\right|^3\right)\right]\frac{1}{\sqrt{n}}$$

which does not depend on x any more.

Esempio 9.3.4. For instance if the assumption by Berry holds and $n = 100$, we can say that the error made at any point x is

$$\leq \frac{1}{2}\frac{1}{10}\,\mathbb{E}\left[\left|\frac{X_1-\mu}{\sigma}\right|^3\right], \quad \forall x \in \mathbb{R}$$

Thus in practice to have a good estimate it's enough to know σ (or making assumption/educated guess).

Osservazione 283. One last remark on CLT: CLT1 allows to obtain some infos about speed of converge (also said convergence rate) in the Kolmogorov strong law of large numbers (the most important one). We see it below

Proposizione 9.3.4. *Let's assume the condition of CLT1 and fix a sequence a_n of constants such that*

$$\frac{a_n}{\sqrt{n}} \to 0$$

Now by kolmogorov's strong law we can say that

$$\overline{X}_n - \mu \xrightarrow{a.s.} 0$$

where as before $\mu = \mathbb{E}[X_1]$. Moreover by CLT1 we have that

$$a_n(\overline{X}-\mu) = \frac{a_n}{\sqrt{n}}\sqrt{n}(\overline{X}-\mu)$$

and by assumption $\frac{a_n}{\sqrt{n}} \to 0$, while for CLT1 $\sqrt{n}(\overline{X}-\mu) \to \mathrm{N}\left(0,\sigma^2\right)$ where $\sigma^2 = \mathrm{Var}[X_i]$. Thus the product goes to 0

$$a_n(\overline{X}-\mu) \xrightarrow{p} 0$$

further it can be shown that one also obtains

$$a_n(\overline{X}-\mu) \xrightarrow{a.s.} 0$$

Osservazione 284. If we have only LLN we can say only $X_i - \mu \xrightarrow{a.s.} 0$; using clt we can say much more $a_n(\overline{X}-\mu) \xrightarrow{a.s.} 0$.

Esempio 9.3.5. If I take $a_n = \sqrt{n}/\log n$ we have that

$$\frac{a_n}{\sqrt{n}} = \frac{1}{\log n} \to 0$$

and i get that

$$\frac{\sqrt{n}}{\log n}(\overline{X} - \mu) \xrightarrow{a.s.} 0$$

but $\sqrt{n}/\log n \to +\infty$ and

$$\frac{\sqrt{n}}{\log n}(\overline{X} - \mu) \to 0$$

even if $(\overline{X} - \mu)$ is multiplied by something that goes to $+\infty$.

9.4 Additional topics

9.4.1 Borel-Cantelli lemma

Let $\{A_n\}$ be a sequence of events, be them any subset of the sample space Ω. Then we define two new events:

1. first is limsup of the sequence (remembering that intersection means $\forall$ and union means $\exists$):

$$\overline{\lim_n} A_n = \bigcap_{n=1}^{\infty} \bigcup_{j=n}^{+\infty} A_j = \{\omega \in \Omega : \forall n \geq 1, \exists j \geq n \text{ such that } \omega \in A_j\}$$
$$= \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\}$$

   For instance if Bologna plays every sunday, A_n is Bologna wins at time n: limsup is event that Bologna wins infinite number of games.

2. the second event is liminf, defined as

$$\underline{\lim_n} A_n = \bigcup_{n=1}^{\infty} \bigcap_{j=n}^{+\infty} A_j \quad = \{\omega \in \Omega : \exists n \geq 1 \text{ such that } \omega \in A_j, \forall j \geq n\}$$

   Eg liminf is event there is an n such that from n on, Bologna wins every time.

Osservazione importante 71. By the Demorgan Law the complement of the limsup is the liminf of the complement, and the two events are connected by this equation

$$\left(\overline{\lim_n} A_n\right)^c = \left(\bigcap_{n=1}^{\infty} \bigcup_{j=n}^{+\infty} A_j\right)^c = \bigcup_{n=1}^{+\infty} \left(\bigcup_{j=n}^{+\infty} A_j\right)^c = \bigcup_{n=1}^{+\infty} \bigcap_{j=n}^{+\infty} A_j^c = \underline{\lim_n} A_n^c$$

Osservazione 285. Borel-Cantelli lemma is a tool to evaluate the probability of the limsup $\mathbb{P}\left(\overline{\lim_n} A_n\right)$ under some assumptions.

Teorema 9.4.1 (Borel-Cantelli)**.** *If*

- $\sum_{i=1}^{n} \mathbb{P}(A_n) < +\infty$ *(that is converges) then the probability of the limsup is null:* $\mathbb{P}\left(\overline{\lim_n} A_n\right) = 0$*;*
- $\sum_{i=1}^{n} \mathbb{P}(A_n) = +\infty$ *(that is diverges) and the* A_n *are independent, then* $\mathbb{P}\left(\overline{\lim} A_n\right) = 1$*.*

Osservazione importante 72 (Two remarks). Regarding Borel-Cantelli:

1. Why the series of probability *necessarily* converges or diverges (can't be oscillating)? This is because it's the limit of a partial sum of positive or null numbers (probabilities).
   In other words let $\alpha_n \geq 0$ be a sequence of non-negatives $\forall n$ then the sequence $\sum_{i=1}^{n} \alpha_i$ is increasing and every increasing sequence has a limit equal to the sup (whether it is finite or not). Hence $\exists \lim_n \sum_{i=1}^{n} \alpha_i = \sup_n \sum_{i=1}^{n} \alpha_i$.
   Hence letting $\alpha_n = \mathbb{P}(A_n)$ there are only two situations. Either $\sum_{i=1}^{n} \mathbb{P}(A_i) < +\infty$ or $\sum_{i=1}^{n} \mathbb{P}(A_i) = \infty$;
2. if $\sum_{i=1}^{n} \mathbb{P}(A_n) = +\infty$ but the A_n are not independent, the Borel-Cantelli lemma does not apply (it does not cover any possible situation).

Osservazione 286. Proof is relatively easy but instead of it we make some examples to appreciate the use of the lemma.

Esempio 9.4.1. Suppose we have a coin and we throw it infinitely many times; we assume that the probability of tail is constant, say $\mathbb{P}(T) = \alpha \in (0, 1)$ indipendently from the past.
Under these assumptions, we observe any finite string of heads and tails infinitely many time with probability 1. We want to apply the Borel-Cantelli obtaining probability 1
To see this, fix a finite string, say `TTHHT`; Define the random variable X_n equal to indicator of the event

$$X_n = \mathbb{1}_{.(\text{tail at time } n)}$$

X_n are independent evs (iid). We also define also (all sequences `TTHHT` below, constructed to be independent)

$$\begin{aligned} A_1 &= \{X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, X_5 = 1\} \\ A_2 &= \{X_6 = 1, X_7 = 1, X_8 = 0, X_9 = 0, X_{10} = 1\} \\ A_3 &= \{X_{11} = 1, X_{12} = 1, X_{13} = 0, X_{14} = 0, X_{15} = 1\} \\ &\ldots \text{and so on} \end{aligned}$$

A_1 is the event where the string occurs at the first five trials; we want A_i to be independent to apply the second version/point of Borel-Cantelli so we defined them using different X_i (which are independent).
Now, we have that, for any A_i:

$$\mathbb{P}(A_i) = \alpha \cdot \alpha \cdot (1-\alpha) \cdot (1-\alpha) \cdot \alpha = a^3(1-\alpha)^2 > 0$$

The last is strictly positive because $\alpha \in (0,1)$.
Hence $\sum_{i=1}^{n} \mathbb{P}(A_n) = \sum_{i=1}^{n} a^3(1-\alpha)^2 = +\infty$ since is an infinite sum of positive constant. By Borel-Cantelli one finally obtains

$$\mathbb{P}(\text{observe } \texttt{TTHHT} \text{ infinitely many times}) \geq \mathbb{P}\left(\overline{\lim} A_n\right) \overset{(1)}{=} 1$$

in (1) by Borel-Cantelli.

Esempio 9.4.2. Thanks to Borel-Cantelli it easily built an example where X_n converges in L_1 but does not almost surely.
Take any sequence A_n of *independent* events such that $\mathbb{P}(A_n) = \frac{1}{n}$ and define $X_n = \mathbb{1}_{\cdot}(A_n)$. We have that $X_n \xrightarrow{L_1} 0$ since:

$$\mathbb{E}[|X_n - 0|] = \mathbb{E}[|X_n|] = \mathbb{E}[\mathbb{1}_{\cdot}(A_n)] = \mathbb{P}(A_n) = \frac{1}{n} \to 0$$

It remains to see that it does not converge almost surely: the A_n are independent by assumption and

$$\sum_{i=1}^{n} \mathbb{P}(A_n) = \sum_{i=1}^{n} \frac{1}{n} \overset{(1)}{=} +\infty$$

being (1) the armonic series.
Hence by Borel-Cantelli we can say that $\mathbb{P}\left(\overline{\lim_n} A_n\right) = 1$; similarly the A_n^c are independent (if A_n are independent the complements are still independent) and

$$\sum_{i=1}^{n} \mathbb{P}(A_n^c) = \sum_{i=1}^{n} \frac{n-1}{n} = +\infty$$

so that

$$\mathbb{P}\left(\overline{\lim_n} A_n^c\right) = 1$$

It follows that the intersection of two almost sure events is still almost sure, that is:

$$\mathbb{P}\left(\overline{\lim_n} A_n \cap \overline{\lim_n} A_n^c\right) = 1$$

Now fix an omega in this intersection

$$\omega \in (\overline{\lim_n} A_n \cap \overline{\lim_n} A_n^c)$$

Then the numerical sequence $X_n(\omega)$ does not converge because $X_n(\omega) = 1$ for infinitely many n and $X_n(\omega) = 0$ for infinitely many n so X_n does not converge almost surely.

Esempio 9.4.3. Let $\{X_n\}$ be iid rvs and suppose X_1 is non degenerate. Under these assumption:

$$\mathbb{P}(X_n \text{ converges to a finite limit}) = 0$$

It's intuitive (if everyone of us choose a random number from the same distribution, then this will not converge to something).

To prove it formally, since X_1 is non degenerate it can be show that there are two numbers a, b with $a < b$ such that

$$\mathbb{P}(X_1 \leq a) > 0 \vee \mathbb{P}(X_1 \geq b) > 0$$

Now define two events

$$A_n = \{X_n \leq a\},$$
$$B_n = \{X_n \geq b\}$$

What is the probability of limsup of A_n? Being A_n independent (because rvs are independent) and identically distributed we have that:

$$\sum_{i=1}^{n} \mathbb{P}(A_n) = \sum_{i=1}^{n} \mathbb{P}(X_n \leq a) = \sum_{i=1}^{n} \mathbb{P}(X_1 \leq a) \overset{(1)}{=} +\infty$$

with (1) because summing the same positive number infinite times. Hence $\mathbb{P}\left(\overline{\lim} A_n\right) = 1$.
By exactly the same arguments $\mathbb{P}\left(\overline{\lim_n} B_n\right) = 1$.
Hence $\mathbb{P}\left(\overline{\lim_n} A_n \cap \overline{\lim_n} B_n\right) = 1$.
Now as before, we fix ω in that intersection

$$\omega \in (\overline{\lim_n} A_n \cap \overline{\lim_n} B_n)$$

then $X_n(\omega)$ become a numerical sequence. Again this sequence does not converge:

- since $\omega \in \overline{\lim_n} A_n$, then $X_n(\omega) \leq a$ for infinitely many n
- otoh since $\omega \in \overline{\lim_n} A_n$, then $X_n(\omega) \geq b$ for infinitely many n

So having that $a < b$ this can't converge.

Osservazione 287. Incidentally (related to Borel-Cantelli) recall that:

- if $\mathbb{P}(A_i) = 0, \forall i$ then then the probability of the union

$$\mathbb{P}(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) = 0$$

- if $\mathbb{P}(A_i) = 1, \forall i$ then the probability of intersection

$$\mathbb{P}(\cap_{i=1}^{n} A_i) = 1 - \mathbb{P}((\cap_{i=1}^{n} A_i)^c) = 1 - \mathbb{P}(\cup_{i=1}^{n} A_i^c) = 1 - 0 = 1$$

  infact $\mathbb{P}((\cap A_{i=1}^{n})^c) = \mathbb{P}(\cup_{i=1}^{n} A_i^c) = 0$ since $\mathbb{P}(A_i^c) = 0, \forall n$

9.4.2 Infinite divisible rvs

Osservazione 288. In probability theory, a distribution is infinitely divisible if it can be expressed as the sum of an arbitrary number of independent and identically distributed (i.i.d.) random variables.

Definizione 9.4.1 (Infinite divisible rv)**.** Let X be a real rv, then X is infinite divisible if and only if $\forall n \geq 1$, $\exists X_{n_1}, \ldots, X_{n_n}$ iid rvs such that $X \sim \sum_{i=1}^{n} X_{n_i}$.

Esempio 9.4.4. $X \sim \text{Pois}(\lambda)$ is infinite divisible. Infact, if $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \text{Pois}(\lambda_i)$ then $\sum_{i=1}^{n} Y_i \sim \text{Pois}(\sum_{i=1}^{n} \lambda_i)$.
Hence if $X \sim \text{Pois}(\lambda)$, if is sufficient to take $X_{n_1}, \ldots, X_{n_n}$ iid rvs with $X_{n_i} \sim \text{Pois}\left(\frac{\lambda}{n}\right)$

Esempio 9.4.5. $\text{N}\left(\mu, \sigma^2\right)$ is infinite divisible. Infact if $X_1, \ldots, X_n$ idependent, $\text{N}\left(\mu_i, \sigma_i^2\right)$, then $\sum_{i=1}^{n} X_i \sim \text{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$

Esempio 9.4.6. Another example is the gamma. In fact $X \sim \text{Gamma}(\alpha, \beta)$ iff X is absolutely continuous with density

$$f(x) = \begin{cases} \frac{\alpha^\beta}{\Gamma(\beta)} e^{-\alpha x} x^{\beta 1} & x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Note for $\beta = 1$ we get the $\text{Gamma}(\alpha, 1) = \text{Exp}(\alpha)$ so exponential is a special case of gamma.
Now if $Y_1, \ldots Y_n$ indep and $Y_i \sim \text{Gamma}(\alpha, \beta_i)$ (with common α) then the sum of Y_i is still a gamma, that is $\sum_{i=1}^{n} Y_i \sim \text{Gamma}(\alpha, \sum_{i=1}^{n} \beta_i)$.
By the way, if $Y_1, \ldots Y_n$ are iid $Y_i \sim \text{Exp}(\alpha)$, then the distribution of the sum $\sum_{i=1}^{n} Y_i = \text{Gamma}(\alpha, n)$ (n because $\beta = 1$ and the sum is n).
Using the above results it follows that Gamma is infinite divisible.

Osservazione 289. Another nice fact on infinite divisible is the following.

Teorema 9.4.2. *If:*

1. *X is infinite divisible and*
2. *$\mathbb{E}\left[X^2\right] < +\infty$ (has finite second moment)*

This is possible if and only if $X \sim X_1 + X_2 + X_3$ with X_1, X_2, X_3 independent, X_1 degenerate, X_2 normal with mean 0 and variance σ^2 and X_3 generalized Poisson.

Osservazione 290. So this describes the structure of infinite divisible random variables (with finite second moment). Let's see what is a generalized Poisson.

Definizione 9.4.2. X is generalized poisson if $X \sim \mathbb{1}.(N > 0) \cdot \sum_{j=1}^{N} Z_j$ where:

- $N \sim \text{Pois}(\lambda)$
- (Z_j) are iid
- the sequence of (Z_j) are independent of N.

Osservazione importante 73. We expect to find the poisson rv as a special case of this. To do it: if $Z_j = 1, \forall j$, then $X \sim \mathbb{1}.(N > 0) \cdot N = N$, but N is poisson. So the poisson is just special case of the generalized poisson.

Teorema 9.4.3. *If X is infinite divisible and $\mathbb{P}(a \leq X \leq b) = 1$ for some a and b (X is bounded) then X is degenerate.*

Proof. Since X is infinite divisible, by definition $\forall n \geq 1$ (questo n è il pediced di n_i) we have $X \sim \sum_{i=1}^{n} X_{n_i}$ where $X_{n_1}, \ldots, X_{n_n}$ are iid. Therefore:

$$\operatorname{Var}[X] = \operatorname{Var}\left[\sum_{i=1}^{n} X_{n_i}\right] = \sum_{i=1}^{n} \operatorname{Var}[X_{n_i}] = n \operatorname{Var}[X_{n_1}]$$

Now we have that $n \operatorname{Var}[X_{n_1}] \leq n \mathbb{E}\left[X_{n_1}^2\right]$ (consider the variance calculation formula i guess).
Since however $\mathbb{P}(a \leq X \leq b) = 1$, we have that $\mathbb{P}\left(X_{n_1} > \frac{b}{n}\right) = 0$. Infact

$$0 = \mathbb{P}(X > b) = \mathbb{P}\left(\sum_{i=1}^{n} X_{n_i} > b\right) \geq \mathbb{P}\left(X_{n_i} > \frac{b}{n}, \forall i\right) \overset{(iid)}{=} \left[\mathbb{P}\left(X_{n_1} > \frac{b}{n}\right)\right]^n$$

and therefore

$$\mathbb{P}\left(X_{n_1} > \frac{b}{n}\right) = 0$$

Similarly $\mathbb{P}\left(X_{n_1} < \frac{a}{n}\right) = 0$ by the same argument. Hence X_{n_1} stays between $\frac{a}{n}$ and $\frac{b}{n}$, therefore therefore

$$\mathbb{P}\left(\frac{a}{n} \leq X_{n_1} \leq \frac{b}{n}\right) = 1 \quad \text{and therefore}$$
$$\mathbb{P}\left(|X_{n_1}| \leq \frac{\max(|a|, |b|)}{n}\right) = 1$$

And finally:

$$\mathbb{E}\left[X_{n_1}^2\right] \overset{(1)}{\leq} \frac{\max(|a|, |b|)^2}{n^2} \overset{(2)}{\leq} \frac{n \max(|a|, |b|)^2}{n^2} = \frac{\max(|a|, |b|)^2}{n}$$

where

- in (1) if a rv is maggiorata by a constant, so it is its expected value (we used the last equation above with some algebra trick regarding the square)
- in (2) we added a n, respecting unequality

Hence:

$$\operatorname{Var}[X] \leq \lim_{n} \frac{\text{constant}}{n} = 0$$

where at numerator the constant is given by the max above. Therefore X is degenerate. □

9.4.3 Stable rvs

Osservazione 291. It's another important type of random variables.

Definizione 9.4.3 (Stable rv)**.** X is said to be stable iff exists numbers sequences $\exists a_n \in \mathbb{R}$, $b_n > 0$, and a rvs $\{Y_n\}$ iid sequence such that

$$\frac{\sum_{i=1}^{n} Y_i - a_n}{b_n} \xrightarrow{d} X$$

Esempio 9.4.7. An example of stable rv is the normal (look clt): $\mathrm{N}\left(\mu, \sigma^2\right)$ is stable essentially by definition.

Osservazione 292. Is the normal the only stable? no, other example are the Cauchy and degenerate.

Esempio 9.4.8 (Cauchy)**.** The Cauchy is the rv which does not have the mean. It's easy to prove its stable: it suffices to note, if X is cauchy then the characteristic function of X is (take it as given)

$$\phi_X(t) = e^{-|t|}$$

Now we have to verify the definition finding a_n, b_n and $\{Y_i\}$ etc. Take $a_n = 0$, $b_n = n$ and $\{Y_n\}$ iid with Cauchy distribution. Then the sum

$$\frac{\sum_{i=1}^n Y_i - a_n}{b_n} = \frac{\sum_{i=1}^n Y_i}{n} = \overline{Y_n}$$

is the sample mean, and we get

$$\phi_{\frac{\sum_{i=1}^n Y_i - a_n}{b_n}}(t) = \phi_{\overline{Y_n}}(t) \overset{(1)}{=} \left[\phi_{Y_1}\left(\frac{t}{n}\right)\right]^n = \left[e^{-\left|\frac{t}{n}\right|}\right]^n = e^{-|t|}$$

where in (1) since Y_i are iid. Hence the sum:

$$\frac{\sum_{i=1}^n Y_i - a_n}{n} = \overline{Y_n} \sim Y_1$$

so trivially

$$\frac{\sum_{i=1}^n Y_i - a_n}{b_n} \xrightarrow{d} Y_1$$

so the Cauchy is example of stable rv.

Osservazione 293. Two final important remarks.

Proposizione 9.4.4. *If X is stable then X is infinite divisible, but the viceversa does not holds. (so stable are a proper subset of infinite divisible)*

Proof. To prove that viceversa does not hold by counterexample we need a infinite divisible but not stable.
It is sufficient to note that the only stable random variable X with finite second moment ($\mathbb{E}\left[X^2\right] < \infty$) are the normal $\mathrm{N}\left(\mu, \sigma^2\right)$ and the degenerate.

Based on this fact an example of infinite divisible but not stable is the exponential (or the poisson): the exponential has the second moment but it is neither normal nor degenerate thus it's not stable; however as we noted before is infinite divisible. □

Teorema 9.4.5. *X is stable $\iff$ $\forall n \geq 1$, $\exists \alpha_n \in \mathbb{R}$ and β_n such that the sum:*

$$\frac{\sum_{i=1}^n Y_i - \alpha_n}{\beta_n} \sim X$$

if $\{Y_i\}$ is iid and $Y_1 \sim X$.

Osservazione 294. The idea of this theorem: given any rv X, take $Y_1, \ldots, Y_n$ iid with the same distribution as X, $Y_i \sim X$. Then, in general, $\sum_{i=1}^n Y_i \nsim X$. However, if X is stable we can find constants α_n, β_n such that:

$$\frac{\sum_{i=1}^n Y_i - \alpha_n}{\beta_n} \sim X$$

Chapter 10

Simulation

10.1 Sampling values from rvs

Osservazione importante 74. 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 75. 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 76. 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 295. This is the simpler method

Osservazione importante 77. 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	...

Table 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 296. 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 78 (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 297. 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 79. 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 80. 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 81. 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.

Proof. 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 $f\left(x|u < \frac{k(x)\cdot c}{Mp(x)}\right) = \pi(x)$:

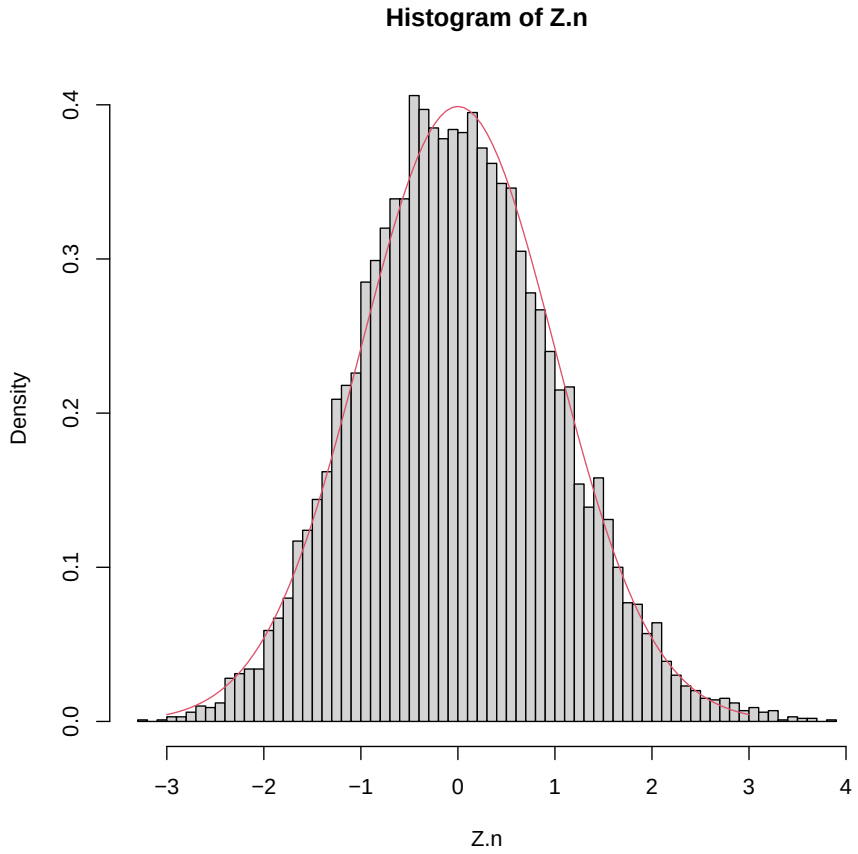
$$\begin{aligned}f\left(x|u < \frac{k(x)\cdot c}{Mp(x)}\right) &= \frac{\mathbb{P}\left(x \cap u < \frac{k(x)\cdot c}{Mp(x)}\right)}{\mathbb{P}\left(u < \frac{k(x)\cdot c}{Mp(x)}\right)} = \frac{p(x)\,\mathbb{P}\left(u < \frac{k(x)\cdot c}{Mp(x)}|x\right)}{\int_{D_x} p(x)\,\mathbb{P}\left(u < \frac{k(x)\cdot c}{Mp(x)}|x\right)\mathrm{d}x} \\ &= \frac{p(x)\frac{k(x)\cdot c}{Mp(x)}}{\int_{D_x} p(x)\frac{k(x)\cdot c}{Mp(x)}\,\mathrm{d}x} = \frac{k(x)\cdot c}{\underbrace{\int_D k(x)\cdot c\,\mathrm{d}x}_{=1}} = k(x)\cdot c \\ &= \pi(x)\end{aligned}$$

So we proved that we can use the algorithm even without knowing the normalization constant. □

10.2 R exercises

10.2.1 CLT

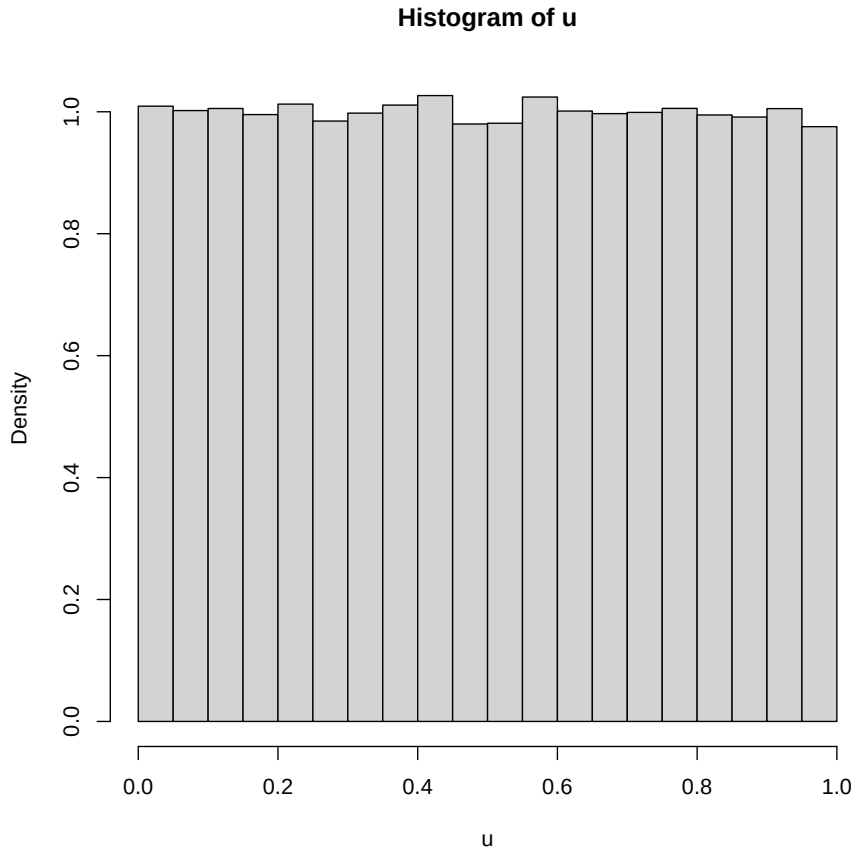
```
set.seed(1) # start from the same point every time
M.n=0       # inizialization (will contain all the values)

### try with n=10 and n=100: clt works for n to infty
n=100
for (i in 1:10000) {
    ## we generate 10000 samples, each of dimension 100;
    ## we use Exp because it's very skewed and we want to visualize
    ## clt works as well
    x = rexp(n,5)
    M.n[i] = mean(x) # here we have the partial mean up to the 100th
                     # extraction
}

## we standardize the partial mean
Z.n = (M.n - mean(M.n))/sd(M.n)
hist(Z.n, 100, freq=FALSE) # histogram of densities
curve(dnorm, -3, 3, col=2, add=TRUE) # add a standard gaussian above
```




```
## we can see clt works in this case

## try to repeat with n = 5: the problem is that assumption of clt are
## not satisfied (n should go to +infty) so if one has 5-10
## observation (and the originating distribution is skewed, one cannot
## invoke the clt for normal-based inference (such area/p-value etc)
```

10.2.2 Inversion method

```
## -------------------------
## 1) quick proof: rv distribution of any pdf is uniform(0,1)
## -------------------------
x = rexp(100000, 5) ## again the exponential because it's very skewed
u = pexp(x, 5)      ## we take the percentile of the extractions
# Theory says that distribution of percentiles are always unif(0,1)
hist(u, 20, freq=FALSE)  # it's uniform as hell (all the bar are height=1)
```

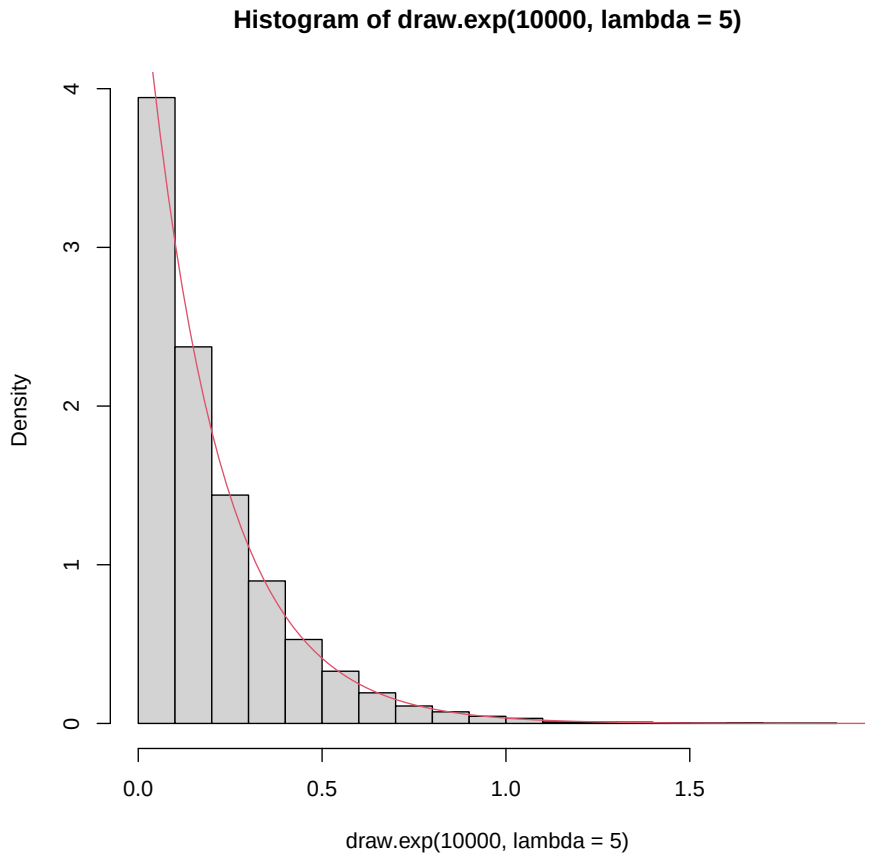



```
## ------------------------
## 2) Pseudo number generation algorithm
## ------------------------
## Before doing it we want to draw pseudo-random numbers in 0,1 using
## a pc which is a deterministic machine (without runif). Some methods
## from math are available
## To be a pseudo random number it should have 2 characteristics
##
## 1) the numbers generated should appear as independent;
## 2) the values should be in interval [0,1]
##
## Example on how to generate
## - fix two integer m and b, with b < m (but b should be not a
##   number such as m/b is integer).
## - fix a seed x0: it's a first number set for starting the algorithm
## - iterate:
##   x1 = b * x0 mod m   (mod is the operator which gives the rest of division, 10%%3
##   u1 = x1/m           (u1 is the first ru extraction)
##   repeat such as
##   x2 = (b * x1) mod m
```

```
##   u1 = x1/m
##   .. and so on
##
## This is an example of algorithm which works and gives pseudo random
## numbers. The seed/starting point make the difference; eg could be
## the seconds. Let's apply this.
draw.unif <- function(n,      ## number of values we want to generate
                      x0 = 11,     ## seed/starting point
                      m = 10001,   ## a large number
                      b = 7) ## lower than m but (m mod b) should be not equal to 0
{
    seq_u = 0     # just to initialize
    x = x0        # seed as starting point
    for (i in 1:n) {
        x = (b*x) %% m    # the resulting x is our x
        seq_u[i] = x/m    # this is our u/extraction
    }
    return(seq_u)
}

draw.unif(100)
```

```
##   [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
```

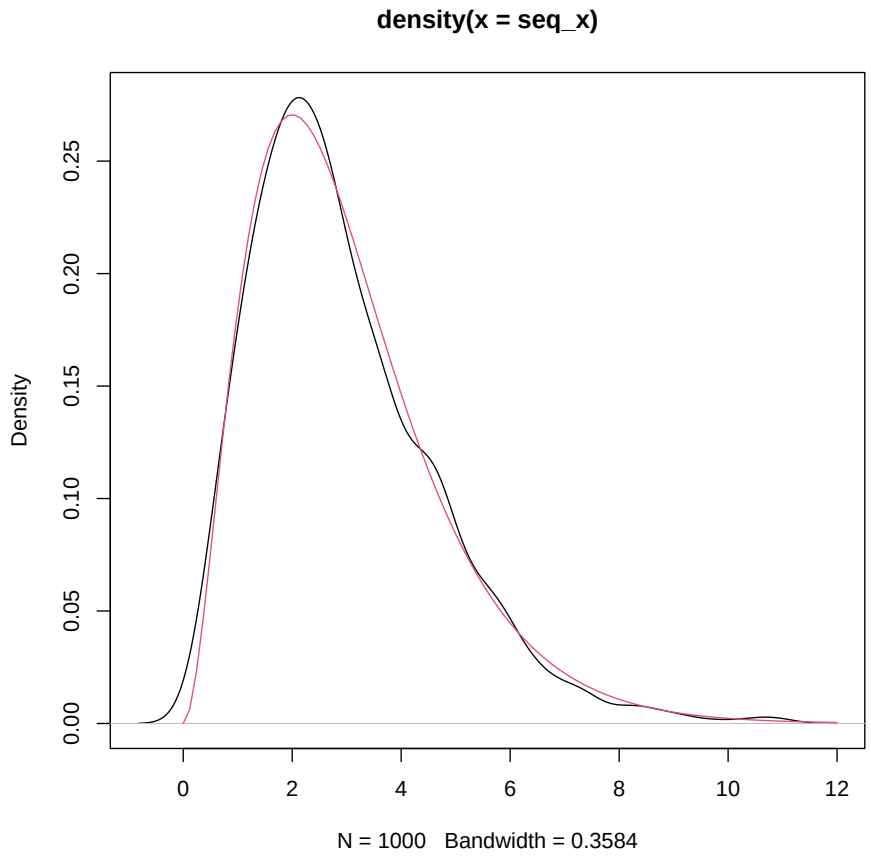
```
## 3) inversion method
draw.exp <- function(n, lambda){
    u = runif(n)
    x = -1/lambda*log(1-u)
    return(x)
}

hist(draw.exp(10000, lambda = 5), freq = FALSE)
curve(dexp(x, rate=5), 0, 2, col = 2, add = TRUE)
```



10.2.3 Accept-reject

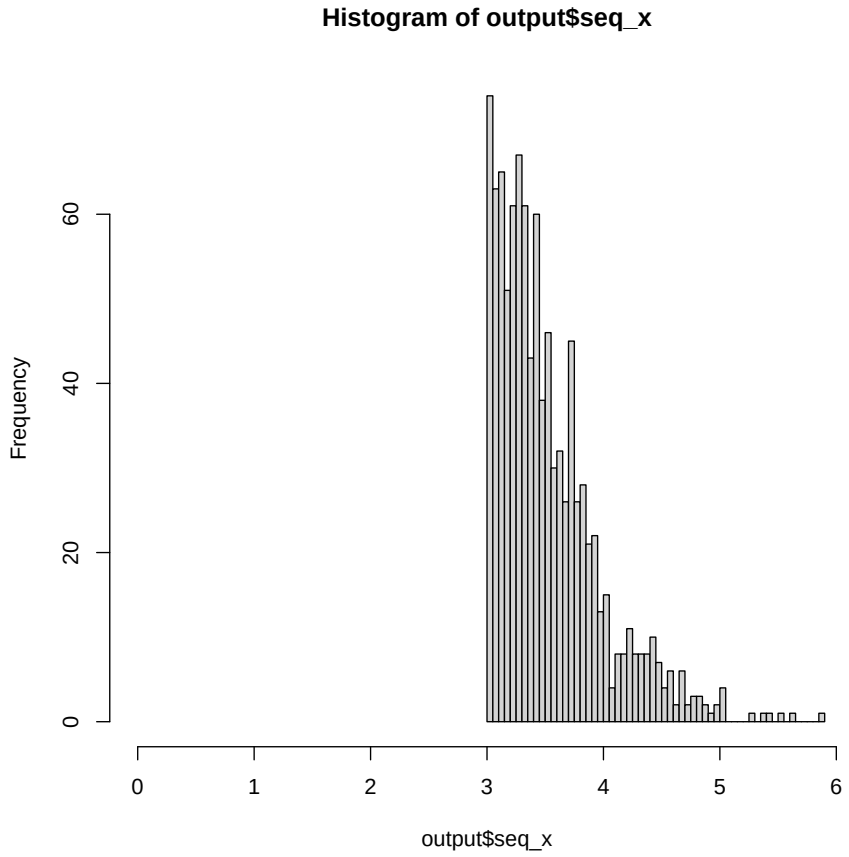
```
############
## Example 1
############

## We consider the problem of generating from a gamma distribution
## with parameters 3 and 1, this one, imagining we don't know the
## command rgamma.  As proposal we could take a normal with mean 3 and
## sd 2 however it's not good in some sense for both symmetri
## (gaussian symmetric, gamma not) and domain (gaussian on R, gamma on
## R+)

mu = 3          ## mean of gaussian
sigma = 2       ## sd of gaussian
n = 1000        ## number of values that we want to obtain
ng = 0          ## number of actually generated values
nit = 0         ## number of algorithm iterations performed
```

```
seq_x = NULL  ## contains all the accepted extraction

while (ng < n) {
  nit = nit + 1              ## increase performed iterations
  x = rnorm(1, mu, sigma)    ## extraction from proposal: this is x
  u = runif(1)               ## extraction from unif(0,1)
  f.val = dgamma(x, 3, 1)    ## target: we know density, don't have rng
                             ## only. Here we could write the density
                             ## kernel as well
  M = 100
  if (u <= f.val / (M * dnorm(x, mu, sigma))){
      seq_x = c(seq_x, x)    ## add to results if condition is met
      ng = ng + 1            ## increment # generated to stop cycle
  }
}

nit # a lot of iteration to have 1000 extraction

## [1] 101852

## acceptance probability (# values generated / # iteration to
## obtain them)
alpha = n / nit
## low, reasons could be: bad proposal, different from the target
## distribution; M choosen (if we lower M alpha increase and algo is
## more efficient, but there's a risk as pointed out before)

## Check results
plot(density(seq_x))
curve(dgamma(x, 3, 1), 0, 12, col=2,add=TRUE)
```




```
#############
## Example 2
#############

## generate values from a truncated gaussian (mean = 2, sd = 1)
## truncated at a=3 starting from a normal (mean = 5, sd = 1)

ar.tn <- function(n){ # n number of values
    mu=2    ## gaussian mean
    sigma=1 ## gaussian sd
    a=3     ## point of truncation
    ng=0    ## number of generated
    seq_x=NULL
    nit=0   ## number of iteration performed
    while (ng < n) {
        nit = nit + 1
        x = rnorm(1, 5, 1)  # proposal
        u = runif(1)
        ## below target (unless/without normalizing constant) it's the
        ## kernel! it's similar to gaussian but it's truncated.
```

```
        ## truncation, keep if > 3, otherwise density is 0 so we keep
        ## the right tail
        f.val = exp(-0.5*((x-mu)/sigma)^2) * (x > a)
        M=100
        if (u <= f.val/(M * dnorm(x, 5, 1))) {
            seq_x = c(seq_x, x)
            ng = ng + 1
        }
    }
    return(list(alpha = n / nit, seq_x = seq_x))
}

output = ar.tn(1000)
hist(output$seq_x, 100, xlim = c(0,6))
```




```
## plot(density(output$seq_x))
```

Part II

Inference

Chapter 11

Introduction to inference

Osservazione importante 82 (Statistics and machine learning). We have that:

- **statistics** was born in 1917, where the first contribution in classical inference (in modern sense) was due to Fisher. The general idea is that we have input data, we work on data (descriptive, inference). Most of time what we do is construct a model.
  The main interest of statistics is have good *interpretation* of data; understanding what data suggest us.
- **machine learning** (ML) on the other hand, speaks about algorithms; was born (from statistician and computer scientist) in the 80's because of computer availability.
  You have a kind of black box, you don't care about interpretation, don't know how it works: important is, as output, to have a good prediction (eg for image/audio recognition)

Osservazione 298. What is best? it's subjective, there are no closed boundaries, there is overlap between two; in stat we speak about models, in ml about algorithms.

11.1 Classical inference setup

Osservazione importante 83 (Setup). We observe a sample, subset of population, composed by n (sample size) observation $(x_1, \dots, x_n)$ (denoted in *lowercase* letters).
We can view the sample as realization of n random variables $(X_1, \dots X_n)$ (in *capital* letters): we assume that each rv is distributed according to a common an F_X we don't know (the distribution function in the population). We want to infer characteristics of F_X from the sample.

Definizione 11.1.1 (Parametric inference)**.** It's when one assume that F_X is a probabilistic model characterized by a parameter θ from a parameter space $\mathbf{\Theta}$:

$$F_X(\theta) = \{F(x;\theta) : \theta \in \mathbf{\Theta}\}$$

In this framework to make inference it's enough to estimate θ (eg for point estimation, interval estimation); if you know θ you know everything.

Definizione 11.1.2 (Nonparametric)**.** The set of all possible distribution F of interest is not restricted to belong to a probabilistic model, it's the complete set of all possible distribution function:

$$F = \{\text{all the CSF's}\}$$

So in this framework one doesn't have a probabilistic model and therefore a parameter θ to be estimated.

11.2 Parametric inference

Osservazione importante 84. Imagine we observed a sample of n observation $(x_1, \dots, x_n)$: we want to find the best guess for the parameter θ or a transformation of the parameter $T(\theta)$. In order to do our guess we're aware that when we make inference we can make mistakes; the idea however is to reduce occurrence by

- choosing best formula/estimator to do our guess
- reducing the variability and increase the precision of our guess (with sample size).

11.2.1 Point estimation

Osservazione importante 85 (Estimator, estimate). In the following we write:

- $T_n = T(X_1, \dots, X_n)$ to mean our *estimator* for θ, that is the procedure/statistics/transformation we apply on random variable to obtain an estimate; being a function of random variable, it's a random variable itself with an own distribution;
- $t_n = \hat{\theta} = T(x_1, \dots, x_n)$ to mean our *estimate*, that is the result of applying the estimator to our sample.

11.2.2 Property of estimators

Osservazione importante 86 (On quality of estimators). When we have an estimate t_n we don't know if it's good or not for θ; our trust on t_n is based on the behaviour of $T(X_1, \dots, X_n)$ in the family of possible (infinite) samples, that is in the sample space Ω.
Therefore it's crucial to search for estimators with good behaviour.

Osservazione importante 87 (Property of estimators). The desiderable properties for estimators are

1. unbiasedness
2. efficiency (comparative)
3. consistency

11.2.2.1 Unbiasedness

Definizione 11.2.1 (Unbiasedness)**.** $T_n = T(X_1, \dots, X_n)$ is unbiased for θ if and only if $\mathbb{E}[T_n] = \theta$, $\forall \theta \in \boldsymbol{\Theta}$.

Osservazione 299. We look at expectation because, as said, the estimator is a function of random variables, so it's a random variable itself.

Definizione 11.2.2 (Bias)**.** It's the difference between the expected value and the real parameter to be estimated:

$$\text{Bias}(T_n) = \mathbb{E}[T_n] - \theta \tag{11.1}$$

Osservazione importante 88. One cannot compute the bias, because we don't know θ.

Esempio 11.2.1 (Sample mean)**.** Let X_i be independent rvs with expected value $\mathbb{E}[X_i] = \mu$ and variance $\text{Var}[X_i] = \sigma^2$. The sample mean is defined as

$$T_n = \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \tag{11.2}$$

We have that

$$\mathbb{E}\left[\bar{X}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n}\sum_i \mathbb{E}[X_i] = \frac{1}{n} n\mu = \mu$$

$$\text{Var}\left[\bar{X}\right] = \text{Var}\left[\frac{\sum_i X_i}{n}\right] = \frac{1}{n^2}\text{Var}\left[\sum_i X_i\right] \overset{(\perp\!\!\!\perp)}{=} \frac{\sum_i \text{Var}[X_i]}{n^2} = \frac{n\,\text{Var}[X_i]}{n^2} = \frac{\text{Var}[X_i]}{n}$$

So the sample mean is a unbiased estimator of the mean μ; its variance $\frac{\sigma^2}{n}$ is directly associated to population variance but it collapses on μ as $n \to \infty$.

Esempio 11.2.2. Let $X \sim \text{Bern}(p)$, our parameter of interest is p and we have two estimators for it:

$$T_n^{(1)} = \frac{\sum_i X_i}{n}$$
$$T_n^{(2)} = \frac{\sum_i X_i^2}{n}$$

Checking for bias we have

$$\mathbb{E}\left[T_n^{(1)}\right] = \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] = \frac{\sum_i \mathbb{E}[X_i]}{n} \overset{(1)}{=} \frac{\sum_i p}{n} = \frac{np}{n} = p$$
$$\mathbb{E}\left[T_n^{(2)}\right] = \mathbb{E}\left[\frac{\sum_i X_i^2}{n}\right] = \frac{\sum_i \mathbb{E}\left[X_i^2\right]}{n} \overset{(1)}{=} \frac{\sum_i p}{n} = \frac{np}{n} = p$$

where in (1) since it's a bernoulli, and in (2) given that for the bernoulli distribution (and only for her) all the moments are equal (first moment = p, second = p etc), or otherwise using computational formula for variance and obtaining $\mathbb{E}\left[X_i^2\right]$.
Therefore both estimators are unbiased: even the second estimator is so the mean is the not the only estimator. We could take the mean of power two three as well.

Esempio 11.2.3 (Sample variance)**.** Given X_i iid rvs with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, if our interest is in estimating σ^2. One estimator could be the sample variance, defined as follows using the sample mean:

$$T_n = \frac{\sum_i^n (X_i - \bar{X})^2}{n} \tag{11.3}$$

However this is a biased of the variance of population σ^2. Let's see why. First some results we have: since

$$\mathbb{E}\left[\bar{X}\right] = \mu, \quad \mathrm{Var}\left[\bar{X}\right] = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[X_i^2\right] = \sigma^2 + \mu^2, \quad \mathbb{E}\left[\bar{X}^2\right] = \frac{\sigma^2}{n} + \mu^2$$

with the first two as properties of sample mean, the latters by the calculation formula of the variance (applied to X_i or $\bar{X}$).
Now we work algebrically the sample variance formula to calculate expectation easier:

$$\begin{aligned} T_n &= \frac{\sum_{i=1}^n (X_i^2 + \bar{X}^2 - 2\bar{X}X_i)}{n} = \frac{\sum_{i=1}^n X_i^2 + \sum_{i=1}^n \bar{X}^2 - 2\bar{X}\sum_{i=1}^n X_i}{n} \\ &= \frac{\left(\sum_{i=1}^n X_i^2\right) + n\bar{X}^2 - 2\bar{X}n\bar{X}}{n} = \frac{\left(\sum_{i=1}^n X_i^2\right) - n\bar{X}^2}{n} = \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \end{aligned}$$

Now take expectation

$$\begin{aligned} \mathbb{E}[T_n] &= \frac{\sum_i \mathbb{E}\left[X_i^2\right]}{n} - \mathbb{E}\left[\bar{X}^2\right] = \frac{n(\sigma^2 + \mu^2}{n} - \left[\frac{\sigma^2}{n} + \mu^2\right] = \sigma^2 + \mu^2 - \frac{\sigma^2}{\mu} - \mu^2 \\ &= \frac{n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2}}{n} = \frac{\sigma^2(n-1)}{n} = \frac{n-1}{n}\sigma^2 \end{aligned}$$

Being $\mathbb{E}[T_n] \neq \sigma^2$, sample variance is biased with bias

$$\mathrm{Bias}(T_n) = \mathbb{E}[T_n] - \sigma^2 = \frac{\cancel{n\sigma^2} - \sigma^2 - \cancel{n\sigma^2}}{n} = -\frac{\sigma^2}{n}$$

Two consideration:

- for a certain n, the estimator is biased; however it's asymptotically unbiased/correct, that is, when n increases the biasing factor $(n-1)/n$ goes to 1 or, otherwise stated, if one computes $\lim_{n\to+\infty} \mathrm{Bias}(T_n) = 0$
- for a certain sample size, by putting $n-1$ to denominator in 11.3, the estimator becomes unbiased.

Osservazione 300. An *estimator* for correctness/unbiasedness of our estimators is mean squared error: it measures how much T_n is concentrated around θ.
If it's low the estimator is precise: it's a kind of proximity measure of the estimator around the parameter of interest. It's an expected value on all the sample we could observe before performing the experiment

Definizione 11.2.3 (Mean squared error)**.** It's defined as:

$$\mathrm{MSE}(T_n) = \mathbb{E}\left[(T_n - \theta)^2\right] \tag{11.4}$$

Osservazione 301. MSE can be decomposed according to a famous decomposition

Proposizione 11.2.1 (Decomposition of mse)**.**

$$\mathrm{MSE}\left(T_n\right) = \mathbb{E}\left[(T_n - \theta)^2\right] = \mathrm{Var}\left[T_n\right] + \mathrm{Bias}\left(T_n\right)^2$$

Osservazione 302. The decomposition highlights the source of estimates imprecision of our estimator:

1. variability of the estimator (with respect to its expectation): as the most spread the distribution of the estimator is, the most error we'll make using it to estimate θ;
2. bias of the estimator.

Proof.

$$\begin{aligned}
\mathbb{E}\left[(T_n - \theta)^2\right] &\overset{(1)}{=} \mathbb{E}\left[(T_n - \theta + \mathbb{E}\left[T_n\right] - \mathbb{E}\left[T_n\right])^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{T_n - \mathbb{E}\left[T_n\right]}_{} + \underbrace{\mathbb{E}\left[T_n\right] - \theta}_{}\right)^2\right] \\
&\overset{(2)}{=} \mathbb{E}\left[\left(T_n - \mathbb{E}\left[T_n\right]\right)^2 + \left(\mathbb{E}\left[T_n\right] - \theta\right)^2 + 2(T_n - \mathbb{E}\left[T_n\right])(\mathbb{E}\left[T_n\right] - \theta)\right] \\
&\overset{(3)}{=} \mathbb{E}\left[\left(T_n - \bar{\theta_n}\right)^2 + \left(\bar{\theta_n} - \theta\right)^2 + 2\left(T_n - \bar{\theta_n}\right)\left(\bar{\theta_n} - \theta\right)\right] \\
&\overset{(4)}{=} \underbrace{\mathbb{E}\left[\left(T_n - \bar{\theta_n}\right)^2\right]}_{\mathrm{Var}[T_n]} + \underbrace{\left(\bar{\theta_n} - \theta\right)^2}_{\mathrm{Bias}\left(T_n\right)^2} + 2\left(\bar{\theta_n} - \theta\right)\underbrace{\mathbb{E}\left[T_n - \bar{\theta_n}\right]}_{=0} \\
&= \mathrm{Var}\left[T_n\right] + \mathrm{Bias}\left(T_n\right)^2
\end{aligned}$$

before expanding the square in (1) we use a trick; in (2) we expand the square of the grouped stuff where in (3) we called $\mathbb{E}\left[T_n\right] = \bar{\theta_n}$; finally in (4) the bias factor doesn't need expectation since it's a difference of constant (and its expectation is a constant) and the last factor is zero simply by applying expected value properties and remembering that $\bar{\theta_n} = \mathbb{E}\left[T_n\right]$. □

Osservazione 303 (Jargon). In inference (eg bootstrap) we speak standard error when speaking about estimator

Definizione 11.2.4 (Standard error of an estimator)**.** It's the standard deviation of the estimator T_n

$$\mathrm{SE}\left(T_n\right) = \sqrt{\mathrm{Var}\left[T_n\right]}$$

Esempio 11.2.4. Let $X_1, \ldots, X_n \sim \mathrm{Unif}\left(0, \theta\right)$. We want to estimate θ and consider this estimator

$$T_n(X_1, \ldots, X_n) = X_{(n)} = \max\left\{X_1, \ldots, X_n\right\}$$

Let's find $\mathrm{Bias}\left(T_n\right)$ and $\mathrm{MSE}\left(T_n\right)$ of our estimator. This is a typical exercise; given the distribution, the parameter of interest and the estimator find the properties of the latter.

First being uniforms and considering the maximum (order statistics) we remember that

$$F_{X_i}(x) = \frac{x-0}{\theta-0} = \frac{x}{\theta}$$
$$F_{(n)}(x) = [F_X(x)]^n = \left(\frac{x}{\theta}\right)^n$$
$$f_{(n)}(x) = n \cdot \left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} = n \cdot \frac{x^{n-1}}{\theta^n}$$

To compute the bias we should compute expectation of the estimator

$$\mathbb{E}\left[X_{(n)}\right] = \int_0^\theta x \cdot f_{(n)}(x)\,\mathrm{d}x = \int_0^\theta x \cdot n \cdot \frac{x^{n-1}}{\theta^n}\,\mathrm{d}x = \frac{n}{\theta^n}\int_0^\theta x^n\,\mathrm{d}x$$
$$= \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{\theta^n}\cdot\frac{\theta^{n+1}}{n+1} = \frac{\theta\cdot n}{n+1}$$

Now we can answer the first question. The bias of our estimator is:

$$\mathrm{Bias}\,(T_n) = \frac{\theta\cdot n}{n+1} - \theta = \frac{n\theta-(n+1)\theta}{n+1} = -\frac{\theta}{n+1}$$

For the MSE we need only to compute the variance of the estimator, since $\mathrm{MSE}\,(T_N) = \mathrm{Var}\,[Tn] + \mathrm{Bias}\,(T_n)^2$. To compute the variance we can use the fact that $\mathrm{Var}\,[T_n] = \mathbb{E}\left[T_n^2\right] - \mathbb{E}\,[T_n]^2$; we have already the first moment, we need only the second moment:

$$\mathbb{E}\left[T_n^2\right] = \mathbb{E}\left[X_{(n)}^2\right] = \int_0^\theta x^2 f_{(n)}(x)\,\mathrm{d}x = \int_0^\theta x^2\cdot n\cdot\frac{x^{n-1}}{\theta^n}\,\mathrm{d}x$$
$$= \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,\mathrm{d}x = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n}{\theta^n}\left[\frac{\theta^{n+2}}{n+2}\right]$$
$$= \frac{\theta^2 n}{n+2}$$

And so the variance of T_n is

$$\mathrm{Var}\,[T_n] = \mathbb{E}\left[T_n^2\right] - \mathbb{E}\,[T_n]^2 = \frac{\theta^2 n}{n+2} - \frac{\theta^2 n^2}{(n+1)^2}$$
$$= \frac{\theta^2 n(n+1)^2 - \theta^2 n^2(n+2)}{(n+2)(n+1)^2} = \ldots = \frac{\theta^2 n}{(n+2)(n+1)^2}$$

It cannot be decomposed more than above. Now for the MSE of the estimator

$$\mathrm{MSE}\,(T_n) = \frac{\theta^2 n}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} = \ldots = \frac{2\theta^2}{(n+1)(n+2)}$$

Esempio 11.2.5. Try to compute bias and mse of the minimum $X_{(1)}$ and check it will be not a good estimator: it will estimate the minimum of the interval of the uniform, that is 0, not θ.

Esempio 11.2.6 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x;\theta) = \frac{2x}{\theta}$$

with $X \in [0,\theta]$, $\mathbb{E}[X] = \frac{2\theta}{3}$ and $\mathrm{Var}[X] = \frac{\theta^2}{18}$. Calculate the mean square errors (MSE) of the estimator $\hat{\theta} = \frac{3\sum_{i=1}^{n} x_i}{2n}$

1. $\frac{\theta^2}{8}$
2. $\frac{\theta^2}{8n}$
3. $\frac{\theta^2}{18n}$
4. $\frac{\theta^2(1+2n)}{18n}$

We have that

$$\mathbb{E}[T_n] = \mathbb{E}\left[\frac{3\sum_{i=1}^{n} X_i}{2n}\right] = \frac{3}{2n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{3}{2n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{3}{2n}n\,\mathbb{E}[X_i] = \frac{3}{2n}\frac{2\theta}{3} = \theta$$

so being an estimator of θ, it's unbiased.
For the variance we have

$$\mathrm{Var}[T_n] = \mathrm{Var}\left[\frac{3\sum_{i=1}^{n} X_i}{2n}\right] = \frac{9}{4n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{9}{4n^2}n\,\mathrm{Var}[X_i] = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}$$

So the MSE is

$$\mathrm{MSE}(T_n) = \mathrm{Var}[T_n] + \mathrm{Bias}(T_n)^2 = \frac{\theta^2}{8n} + 0 = \frac{\theta^2}{8n}$$

11.2.2.2 Efficiency

Definizione 11.2.5 (First definition)**.** Let $T_n^{(1)}$ and $T_n^{(2)}$ be two estimators for θ. We say that $T_n^{(1)}$ is *more efficient* than $T_n^{(2)}$ if:

$$\mathrm{Var}\left[T_n^{(1)}\right] \leq \mathrm{Var}\left[T_n^{(2)}\right]$$

Definizione 11.2.6 (Second definition)**.** Let $T_n^{(1)}$ and $T_n^{(2)}$ be two estimators for θ. We say that $T_n^{(1)}$ is more efficent than $T_n^{(2)}$ if

$$\mathrm{MSE}\left(T_n^{(1)}\right) \leq \mathrm{MSE}\left(T_n^{(2)}\right)$$

Osservazione 304. Btw this second definition consider also the information about the bias, since the MSE can be decomposed in a part of variance and in one of bias of the estimator; if we know that the two estimators have the same bias then the two definitions are equivalent (since cancelling out the bias at the two members from the second gives the first).

Osservazione 305 (Relative efficiency). One could equivalently look at *relative efficiency* defined as ratio. Adopting the MSE definition we have:

$$e\left(T_n^{(1)}, T_n^{(1)}\right) = \frac{\text{MSE}\left(T_n^{(2)}\right)}{\text{MSE}\left(T_n^{(1)}\right)}$$

if $e\left(T_n^{(1)}, T_n^{(2)}\right) > 1 \implies T_n^{(1)}$ is preferable

11.2.2.3 Consistency

Osservazione importante 89. It considers the behaviour of an estimator as n (sample size) increases: if the estimator is consistent, as n increases we have better results.
Consistency and unbiasedness are *indepent* properties/definitions: we can have an estimator which is consistent but is biased and viceversa.

Definizione 11.2.7 (Simple (weak) consistency)**.** T_n is weakly consistent for θ if it converges in probability $T_n \xrightarrow{p} \delta_\theta$, that is:

$$\lim_{n\to\infty} \mathbb{P}\left(|T_n - \theta| < \varepsilon\right) = 1, \quad \forall \theta \in \boldsymbol{\Theta}, \varepsilon > 0$$

Osservazione importante 90. Remember that we have the two sufficient condition that can be used to prove weak convergency:

$$\begin{cases} \lim_{n\to\infty} \mathbb{E}\left[T_n\right] = \theta \text{ (or } \lim \text{Bias}\left(T_n\right) = 0) \\ \lim_{n\to\infty} \text{Var}\left[T_n\right] = 0 \end{cases} \implies T_n \xrightarrow{p} \delta_\theta$$

Given the decomposition of the MSE (in variance and bias) the above result is equivalent to the following one

$$\lim_{n\to\infty} \text{MSE}\left(T_n\right) = 0 \implies T_n \xrightarrow{p} \delta_\theta$$

Esempio 11.2.7. In this example the difference between unbiasdness and consistency, with an estimator which is biased but consistent.
We have that $X_n \sim \text{Bern}\left(\frac{1}{n}\right)n$ and we want to estimate the parameter of the distribution $\theta = \frac{1}{n}$ (which is a moving value); we use as estimator the last observed value $T_n = X_n$. The estimator is biased since

$$\mathbb{E}\left[T_n\right] = \mathbb{E}\left[X_n\right] = 1 \cdot \frac{1}{n} \cdot n + 0 \cdot \left(1 - \frac{1}{n}\right) \cdot n = 1 \neq \theta = \frac{1}{n}$$

so the expectation is 1 for every n (we have bias).
However it's consistent since $T_n = X_n \xrightarrow{p} \delta_0$ as proved previously (in example 8.3.3); when n increases the estimator goes to 0 (as the parameter $\theta = \frac{1}{n}$ does.)

Esempio 11.2.8. Suppose that X_n is an estimator for θ with probability function:

$$X_n \sim \begin{cases} \theta & \text{with probability } \frac{n-1}{n} \\ \theta + n & \text{with probability } \frac{1}{n} \end{cases}$$

Show that X_n is weakly consistent for θ but that $\text{Bias}\left(X_n\right) \not\to 0$ as $n \to \infty$.
Here:

- We can try the two sufficient conditions. We have

$$\mathbb{E}[X_n] = \theta\frac{n-1}{n} + \theta\frac{1}{n} + 1 = \theta + 1$$

  So the bias is

$$\text{Bias}(X_n) = \mathbb{E}[X_n] - \theta = 1 \not\to 0$$

  Therefore we cannot apply the 2 sufficient conditions

- Let's try this

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \theta| < \varepsilon) = 1$$

  This is true since $X_n = 0$ with probability $\frac{n-1}{n}$ so $\lim_{n\to\infty} X_n = \theta$

Definizione 11.2.8 (Strong consistency)**.** T_n is strongly consistent for θ iff $T_n \xrightarrow{L_2} \delta_\theta$:

$$\lim_{n\to\infty} \mathbb{E}\left[(T_n - \theta)^2\right] = 0, \quad \forall\theta \in \boldsymbol{\Theta}$$

Osservazione importante 91. In this case the definition using MSE is (both sufficient and necessary):

$$\lim_{n\to\infty} \text{MSE}(T_n) = 0 \iff T_n \xrightarrow{L_2} \delta_\theta$$

11.3 Methods for finding estimators

Osservazione 306. So far we discussed property of given estimators; but how to find them? Now we focus on this problem, with an historical approach.
The most important methods are least square, moments and likelihood, but not the only ones. Maximum likelihood is the most important/used one

11.3.1 Method of least squares

Osservazione 307. One of the first methods used/developed. We encounter this methods in linear models course, since there is very used.
Here we use Y_i instead of X_i since it's general notation (this's what she said).

Definizione 11.3.1. If $Y_1, \dots, Y_n$ are independent rvs with same variance and higher moments, with $\mathbb{E}[Y_i] = \mathcal{T}(\theta)$[1], where $\mathcal{T}$ is a *linear* function, then the least square estimate for θ is obtained by minimizing

$$\sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i])^2$$

Osservazione importante 92. Pros and cons:

- Pro: least square estimators are BLUE (best linear unbiased estimators; est since the most efficient one)
- Cons: limited applicability (we cannot use this square in many problems because we don't have a regression model)

[1] This is general, but the function can identity as well, eg $\mathbb{E}[Y_i] = \theta$. Eg the transformation is used for glm.

11.3.2 Method of minimum distance

Osservazione 308. Not very used/known but we cite it. First some prereqs.

Definizione 11.3.2 (Empirical distribution function)**.** It's defined as the distribution function we can construct with our sample without assuming any probabilistic distributional form:

$$F_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

TODO: mettere a posto l'indicatrice nelle sintesi latex

where $I(A)$ is the indicator function

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Esempio 11.3.1. Let $(x_1, \ldots, x_5) = (2, 3, 5, 5, 7)$; then we have $F_n(0) = 0$, $F_n(3) = \frac{2}{5}$, $F_n(5) = \frac{4}{5}$ and $F_n(10) = 1$

Osservazione importante 93. It can be shown that the empirical distribution function is a good estimator for the true distribution function, that is F_n is a good estimator for F_X.

Definizione 11.3.3 (Method of minimum distance)**.** Let $(x_1, \ldots, x_n)$ be a random sample from $F_X(x; \theta)$ with θ our object of interest. Now let $F_n(x)$ be the *empirical* distribution function (not the maximum, without tonde). An estimate for θ can be obtained by minimizing (by θ) the *distance* between the empirical distribution function and the theorical/true distribution function:

$$\min_{\theta} d\left[F_n(x), F_X(x; \theta)\right]$$

Osservazione importante 94. The distance can be any distance function: eg euclidean, manhattan or the maximum distance. The latter is defined as the maximum value of the the absolute difference

$$\sup_{x \in D_X} |F_n(x) - F_X(x; \theta)|$$

Different choice about distance imply different results

Osservazione importante 95. Pros and cons:

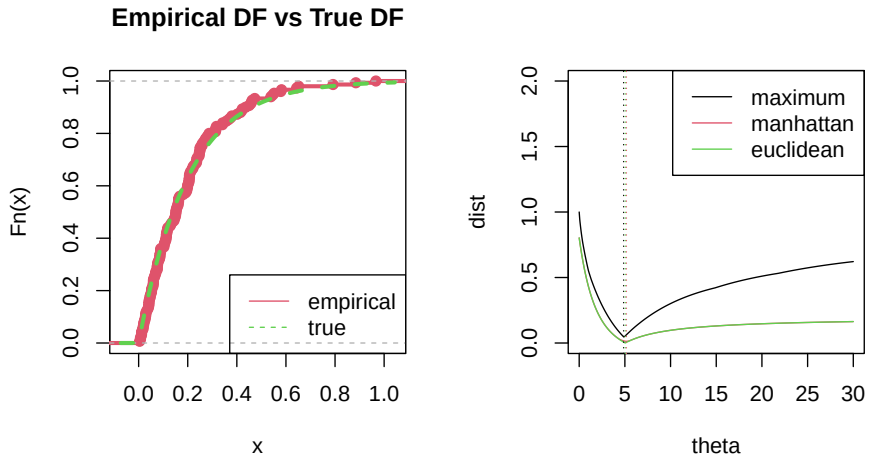
1. pros: very large applicability. Can be used to estimate one or more parameters with sophisticadted method of optimization. It needs assumption on theoretical distribution for my data only;

2. cons: we have just the result of the optimization problem, is difficult to get an analytical function for the estimator $\hat{\theta_n}$ so we cannot study the theoretical properties of the estimator (it's similar to neural networks where one specifies cost function and minimize it)

Esempio 11.3.2 (Example of method of minimum distance)**.** we have

```r
set.seed(1)

# our sample
theta = 5
n = 150
x = rexp(n, theta)

## its empirical cdf
(Fn = ecdf(x))

## Empirical CDF
## Call: ecdf(x)
##  x[1:150] = 0.0040501, 0.0074305, 0.0074537,  ..., 0.88479, 0.96656

## Note that after estimating the Fn we can use it as a normal
## function! eg Fn(x)
Fn(0.5)

## [1] 0.9333333

## ------------------------------------------------------------------
## 1) Ecdf is a good approximation of real df for n to infty
## ------------------------------------------------------------------
## this full comparison plot can't be done in reality because we don't
## know theta. However as n go to infty (here is 150) the empirical
## distribution function is a good estimator of population cdf (if we
## do the same with n = 1000 its almost perfectly overlapping)

par(mfrow = c(1,2))
plot(Fn, lwd = 3, col = 2, main = "Empirical DF vs True DF")
curve(pexp(x, theta), add = TRUE, lwd = 3, lty = 2, col = 3)
legend("bottomright", legend = c('empirical', "true"),
       col = c(2,3), lty = c(1,2))

## --------------------------
## 2) minimum distance method
## --------------------------
## imagine that we know the population distribution is exponential but
## we don't know the parameter theta (here = 5). How do we find theta =
## 5 by using this method?

### distance

## To optimize for theta we need to write a function depending only on
## it. Therefore we define some instrumental stuff

## points on F, F_n codomain where the distance are evaluated
xx = seq(0, 1, length.out = 1000)

## we use as distance the following
```

```
d1 = function(theta) max(abs(Fn(xx) - pexp(xx, theta))) # maximum
d2 = function(theta) mean(abs(Fn(xx) - pexp(xx, theta))) # manhattan
d3 = function(theta) sqrt(mean(Fn(xx) - pexp(xx, theta))^2) # euclidean

# minimizing for theta, the didactic way
# --------------------------------------
## theta searched for in maximization
theta.val = seq(0, 30, length.out = 1000)
## vectors of distance (for each theta) between Fn and F
out1 = out2 = out3 = 0
for (i in 1:1000) {
    out1[i] = d1(theta.val[i])
    out2[i] = d2(theta.val[i])
    out3[i] = d3(theta.val[i])
}
## plot of distances for several thetas
plot(theta.val,  out1, type='l', col=1,
     ylim=c(0,2), xlab = 'theta', ylab = 'dist')
lines(theta.val, out2, type='l', col=2)
lines(theta.val, out3, type='l', col=3)
legend('topright', legend = c("maximum", "manhattan", "euclidean"),
       col = 1:3, lty=1)
## find where the distance is minimal
theta.val[which.min(out1)]

## [1] 4.864865

theta.val[which.min(out2)]

## [1] 5.165165

theta.val[which.min(out3)]

## [1] 5.045045

abline(v = theta.val[which.min(out1)], col = 1, lty='dotted')
abline(v = theta.val[which.min(out2)], col = 2, lty='dotted')
abline(v = theta.val[which.min(out3)], col = 3, lty='dotted')
```




```
## all the estimates are not on 5; they come from an estimator and
## depends on the sample

## we don't have a superiority of a distance over another one, it
## depends on the sample (so look at every)

## minimizing for theta, the automatic way
## ----------------------------------------
## with optim you do the same in a quick manner.
## optim minimizes by default so if one has to optimize
## change sign
## - the first is the initial value for the parameters
##   to be optimized over (it doesn't matter)
## - the second is the function to optimize
## - the third is optimization method
## $par returns the estimated parameter
optim(3, d1, method = "BFGS")$par

## [1] 4.873426

optim(3, d2, method = "BFGS")$par

## [1] 5.172686

optim(3, d3, method = "BFGS")$par

## [1] 5.042332

## if we below increase n (eg 1000) we expect a better estimates,
## nearer to 5) at least on the poor man way
```

11.3.3 Method of moments (MM)

Osservazione 309. Originates with Karl Pearson in 1894, one of the first method invented

Osservazione importante 96. Let $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_k \end{bmatrix}$ be unknown quantities/parameters of interest. So we start esplicitly in a multivariate situation (good, extension of minimum distance on multivariate can be complicate).
Now the steps of the method of moments are as follows:

1. we define $\mu_j = \mathbb{E}\left[X^j\right]$ the j-th *moment* of our variable X (which will be function of our parameter of interest);

2. we define the *j-sample* moment (not the population moment) as:

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

3. method of moments (MM) define the estimator $\hat{\boldsymbol{\theta}}_n$ such that we can construct a system where we equate population moments (functions of the population parameters) and sample moments (which we can calculate), solving the system for population parameter of interest.

$$\begin{cases} \mu_1 = M_1 \\ \mu_2 = M_2 \\ \dots \\ \mu_k = M_k \end{cases}$$

The ending formula will be the MM estimators for our parameters

Esempio 11.3.3. Imagine we have a gaussian $\mathrm{N}\left(\mu, \sigma^2\right)$ and we are interested in estimating the couple of its parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \quad k = 2$$

then we define a system where we have the first two moments, and we make an equivalence between these two moments and the moments of the population (respectively μ and given the computation variance formula $\sigma^2 + \mu^2$)

$$\begin{cases} \mu_1 = \mathbb{E}\left[X^1\right] = \mu \\ \mu_2 = \mathbb{E}\left[X^2\right] = \sigma^2 + \mu^2 \end{cases}$$

In case where we have a gaussian $\mathrm{N}\left(\mu, \sigma^2\right)$ and we want to estimate μ, σ^2 we set up the following equation system

$$\begin{cases} M_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} = \mu \\ M_2 = \frac{\sum_{i=1}^{n} x_i^2}{n} = \sigma^2 + \mu^2 \end{cases}$$

According to this system, the estimators of μ and σ^2 becomes:

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \frac{\sum_i x_i^2}{n} - \hat{\mu}^2 = \frac{\sum_i x_i^2}{n} - \bar{x}^2 = \frac{\sum_i (x_i - \bar{x})^2}{n} \end{cases}$$

Here we find the sample variance which we know is a biased estimator (only asymptotically unbiased).

Esempio 11.3.4. Let $(x_1, \ldots, x_n)$ be a sample of X with pdf

$$f(x;\theta) = \begin{cases} \frac{2}{\theta^2}(\theta - x) & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

This not a known density; use MM to get an estimator of θ, $\hat{\theta}$.
This is a univariate problem; to solve it we equate the first sample moment equal to the first population moment. First of all we compute the first population moment

$$\begin{aligned} \mathbb{E}[X] = \mu_1 &= \int_0^\theta x \frac{2}{\theta^2}(\theta - x)\, \mathrm{d}x = \int_0^\theta \frac{2x}{\theta^2}\theta\, \mathrm{d}x - \int_0^\theta \frac{2x^2}{\theta^2}\, \mathrm{d}x \\ &= \frac{2}{\theta}\left[\frac{x^2}{2}\right]_0^\theta - \frac{2}{\theta}\left[\frac{x^3}{3}\right]_0^\theta = \frac{2}{\theta}\frac{\theta^2}{2} - \frac{2}{\theta^2}3 = \theta - \frac{2}{3}\theta \\ &= \frac{\theta}{3} \end{aligned}$$

Therefore to get the formula for the estimator we equate

$$\frac{\theta}{3} = \bar{x}$$

in order to obtain the estimator

$$\hat{\theta} = 3\bar{x}$$

Esempio 11.3.5. Let $X_1, \ldots, X_n$ be independent rvs each distributed as **TODO**: da rivedere ...

$$f(x|\alpha) = \begin{cases} \frac{1+\alpha x}{2} & -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

with $-1 \le \alpha \le 1$. Now:

1. find the method of moments estimator for the parameter α
2. find the mean squared error of such estimator

We have:

1. for the estimator

$$\begin{aligned} \mathbb{E}[X] &= \int_{-1}^1 \frac{1+\alpha x}{2} x\, \mathrm{d}x = \int_{-1}^1 \left(\frac{x}{2} + \frac{\alpha x^2}{2}\right) \mathrm{d}x = \left[\frac{x^2}{4}\right]_1^{+1} + \left[\frac{\alpha x^3}{6}\right]_{-1}^{+1} \\ &= 0 + \alpha\frac{1}{6} + \alpha\frac{1}{6} = \frac{\alpha}{3} \end{aligned}$$

so $\overline{x} = \frac{\alpha}{3}$ therefore $\hat{\alpha} = 3\overline{x}$

2. for the MSE:

$$\mathrm{MSE}\left(\hat{\alpha}\right) = \mathrm{MSE}\left(3\overline{X}\right)$$

we have that

$$\mathbb{E}\left[3\overline{X}\right] = 3\,\mathbb{E}\left[\overline{X}\right] = 3\,\mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{3}{n} n\,\mathbb{E}\left[X_i\right] = \frac{3\alpha}{3} = \alpha$$

So $\mathrm{Bias}\left(\hat{\alpha}\right) = 0$. For the variance

$$\mathrm{Var}\left[3\overline{X}\right] = 9\,\mathrm{Var}\left[\overline{X}\right] = 9\,\mathrm{Var}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{9}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{9}{n}\,\mathrm{Var}\left[X\right]$$

In order to compute the variance we need $\mathbb{E}\left[X^2\right]$

$$\mathbb{E}\left[X^2\right] = \int_{-1}^{1} x^2\left(\frac{1}{2} + \frac{\alpha x}{2}\right)\mathrm{d}x = \left[\frac{1}{2}\frac{x^3}{3}\right]_{-1}^{1} + \left[\frac{\alpha}{2}\frac{x^4}{4}\right]_{-1}^{1} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

so

$$\mathrm{Var}\left[X\right] = \frac{1}{3} - \frac{\alpha^2}{9} = \frac{3-\alpha^2}{9}$$

therefore

$$\mathrm{MSE}\left(\hat{\alpha}\right) = \frac{9}{n}\cdot\frac{3-\alpha^2}{9} = \frac{3-\alpha^2}{n}$$

Esempio 11.3.6 (Esame vecchio viroli)**.** A random variable X is supposed to follow a continuous distribution whose density function is

$$f(x;\theta) = \theta x^{\theta-1}$$

for $0 < X < 1$. A sample of 4 observation $x_1 = 0.2, x_2 = 0.5, x_3 = 0.7, x_4 = 0.8$ is collected from X. Apply the method of moments fo find an estimate of the parameter θ.
We need $\mathbb{E}\left[X\right]$

$$\begin{aligned}\mathbb{E}\left[X\right] &= \int_0^1 x\cdot\theta x^{\theta-1}\,\mathrm{d}x = \int_0^1 x\frac{1}{x}\theta x^{\theta}\,\mathrm{d}x = \theta\int_0^1 x^{\theta}\,\mathrm{d}x \\ &= \theta\left[\frac{x^{\theta+1}}{\theta+1}\right]_0^1 = \frac{\theta}{\theta+1}\end{aligned}$$

Regarding the sample we have that

$$\overline{x} = \frac{0.2+0.5+0.7+0.8}{4} = 0.55$$

So for the estimate we have

$$\begin{aligned}\frac{\theta}{\theta+1} &= 0.55 \\ \theta &= 0.55\theta + 0.55 \\ 0.45\theta &= 0.55 \\ \hat{\theta} &= 1.22\end{aligned}$$

Risposta: $\hat{\theta}_M = 1.22$

Esempio 11.3.7 (Esame vecchio viroli)**.** a random variable X has density function

$$f(x;\theta) = \frac{2}{\theta^2}(\theta - x)$$

with $X \in [0, \theta]$. Let $X = (X_1, \dots, X_n)$ be a simple random sample. Apply the method of the moments to find an estimator for the parameter θ

1. $\theta_M = 2\overline{X}$
2. $\theta_M = 2/\overline{X}$
3. $\theta_M = \overline{X}/3$
4. $\theta_M = 3\overline{X}$

We have that

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\theta x\frac{2}{\theta^2}(\theta - x) = \int_0^\theta x \cdot \frac{2}{\theta^2} \cdot \theta - x^2 \cdot \frac{2}{\theta^2} \\ &= \int_0^\theta \frac{2x}{\theta} - \frac{2x^2}{\theta^2} = \frac{2}{\theta}\left[\frac{x^2}{2}\right]_0^\theta - \frac{2}{\theta^2}\left[\frac{x^3}{3}\right]_0^\theta \\ &= \frac{2}{\theta}\frac{\theta^2}{2} - \frac{2}{\theta^2}\frac{\theta^3}{3} = \frac{1}{3}\theta\end{aligned}$$

So

$$\overline{x} = \frac{1}{3}\theta \implies \hat{\theta} = 3\overline{x}$$

Esempio 11.3.8 (Esame vecchio viroli)**.** Let θ be the parameter of a population random variable X that follows a continuous uniform distribution on the interval $[\theta - 2, \theta + 1]$ and let $X = (X_1, \dots, X_n)$ be a simple random sample. Find the method of moment estimator for θ

1. $\theta_M = \overline{X}$
2. $\theta_M = \overline{X} + 1/2$
3. $\theta_M = \overline{X} + 1/n$
4. $\theta_M = \sum_{i=1}^n x_i + \frac{1}{2}$

We have that

$$\begin{aligned}\mathbb{E}[X] &= \int_{\theta-2}^{\theta+1} x \cdot \frac{1}{\theta + 1 - \theta + 2} = \frac{1}{3}\int_{\theta-2}^{\theta+1} x = \frac{1}{3}\left[\frac{x^2}{2}\right]_{\theta-2}^{\theta+1} \\ &= \frac{1}{3}\left(\frac{(\theta+1)^2}{2} - \frac{(\theta-2)^2}{2}\right) = \ldots = \theta - \frac{1}{2}\end{aligned}$$

therefore

$$\overline{x} = \theta - \frac{1}{2} \implies \hat{\theta} = \overline{x} + \frac{1}{2}$$

Come confermato da taluni

Esempio 11.3.9 (Esame vecchio viroli)**.** A random variable X is supposed to follow a distribution whose probability function is for $X \in [0,3]$

$$f(x;\theta) = \frac{\theta x^{\theta-1}}{3^\theta}$$

Apply the method of the moments to find an estimator of the parameter θ. We have

$$\begin{aligned}\mathbb{E}[X] &= \int_0^3 x \cdot \frac{\theta x^{\theta-1}}{3^\theta} = \frac{\theta}{3^\theta}\int_0^3 x^\theta = \frac{\theta}{3^\theta}\left[\frac{x^{\theta+1}}{\theta+1}\right]_0^3 \\ &= \frac{\theta}{3^\theta}\left(\frac{3^{\theta+1}}{\theta+1}\right) = \frac{3\theta}{\theta+1}\end{aligned}$$

So

$$\begin{aligned}\overline{x} &= \frac{3\theta}{\theta+1} \\ \overline{x}(\theta+1) &= 3\theta \\ \overline{x}\theta + \overline{x} - 3\theta &= 0 \\ \theta(\overline{x}-3) &= -\overline{x} \\ \hat{\theta} &= \frac{\overline{x}}{3-\overline{x}}\end{aligned}$$

Esempio 11.3.10 (esame vecchio viroli)**.** Let $X_1, \ldots, X_n$ be a random sample from the density function

$$f(x) = \theta^2 x e^{-\theta x}$$

with $x > 0$ and $\theta > 0$. Find the method of moments estimator for θ (you may use the second moment of an exponential distribution)

- $\hat{\theta} = 1/\overline{x}$
- $\hat{\theta} = 2/\overline{x}$: taluni suggeriscono questa
- $+\infty$
- $\hat{\theta} = 2\overline{x}$

Cursed exercise, non ci salto fuori.

11.4 Inference: direct and inverse problem

Osservazione importante 97 (General framework based on probabilistic models). The probabilistic framework

1. X rv that describes a feature of interest on the population;

2. one set a **probabilistic model**: we make an assumption on the distribution of X: the distribution of X is indexed by a parameter $\theta \in \mathbf{\Theta}$ (or a vector of parameters), so $f(x;\theta)$

3. we observe a random sample $(x_1, \dots, x_n)$ according to a proper **sampling model**.
   If the sampling model is the *simple random sampling* with replacement (most common and what we assume here), then $X_1, \dots, X_n$ are iid rvs distributed according to $f(x;\theta)$.
   Simple sampling scheme is not true in practice: eg we don't have replacement (when you observe/sample people you observe a person just one time and don't repeat the observation), but if the population is large enough, with/without replacement becomes equivalent (the probability of observing the same person is very rare/impossible so, even if we replace, taking the second time the same person is very rare) , so we assume it's with replacement even it's not true in practice.
   In other courses other sampling scheme are studied as well (eg stratified, clustering and so on). According to different sampling you develop different theory in inference.

4. the assumption of both *probabilistic model* and *sampling model* are equal to our the **statistical model**, from which we derive our inference

How to make inference on θ? In the general probabilistic framework there are two other methods for finding estimators: maximum likelihood and bayesian estimator.
At the beginning of history people were divided in these two methods; todays situation is different especially in advanced statistics were methods are mixed but in the foundational aspect of statistics were in contraddiction

1. frequentist (classic) framework: we work in a direct problem with a likelihood function $f(\mathbf{x}|\theta)$. The likelihood function is not what we want to know; it's the probability given the parameter of observing our sample

2. bayesian framework: we work in a inverse problem with the posterior function $f(\theta|\mathbf{x})$. The posterior what we want to know/do: it is the probability of the possible values of theta given our sample

It's seems contrary what is direct and inverse: Fisher was inventor of likelihood function and justified the idea of likelihood function which is strange. The idea is: i want to find the parameter that better justify what i have observed. if the probability of my sample given theta = 5 is low, then i will discard it. I try different theta and keep what according to which what i've observed have maximum

Both functions are probability: likelihood function is a probability for the data (even if we already observed them), the posterior function is a probability for the values of the parameter.

Both methods are very used in statistics; we have advantages and disadvantages for both of them.

In most complicated models bayesian stuff allows us to have an inference, while solution based on likelihood could be very difficult. In most simple models likelihood is the most used method

EM algorithm is likelihood based; stochastic EM algorithm is a mixed frequentist-bayesian

11.4.1 Likelihood: frequentist (classic) framework

Osservazione 310. Some notation:

$$\begin{aligned} \mathrm{L}(\theta) &= f(\mathbf{x}|\theta) && \text{Likelihood function} \\ \ell(\theta) &= \log(\mathrm{L}(\theta)) && \text{Log-likelihood function} \end{aligned}$$

Osservazione 311. Likelihood means the level of agreement between what we have observed and the possible level of θ that generated it. $\hat{\theta}_n$ can be estimated by maximizing $\mathrm{L}(\theta)$ or $\ell(\theta)$; log-likelihood is used since maximization happen for the same value of θ (the maximum of a positive function is preserved if we transform it according to a logarithm) and most of time it simplifies the derivation/optimization just from a math standpoint.

Osservazione 312. This framework arise under simple sampling scheme *only* (otherwise it's not likelihood). Under simple sampling scheme *with replacement* $X_1, \dots, X_n$ are iid; then joint density is equal to product of marginal density that is:

$$\mathrm{L}(\theta) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

the joint density (left: which is the density of our observation) is the product of marginal density (according to independence of observation).
$f(\mathbf{x}|\theta)$ is the probability function of my data $\mathbf{x}$ given θ. Why do we use it? choose θ that makes most likely what we observed.

Esempio 11.4.1. Let $X_1, \dots, X_n \sim \mathrm{Bern}(\theta)$ be iid rvs then

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i} \cdot (1-\theta)^{n-\sum x_i}$$

Here for this distribution, we don't have normalization constant, only the kernel.
Now for maximization we set first derivative with respect to $\theta = 0$; maximizing directly the likelihood (find its derivative) could be complicate so in most problem work with log likelihood simplify from a math standpoint. Therefore taking logs of both members we get:

$$\ell(\theta) = \sum x_i \log(\theta) + \left(n - \sum x_i\right) \log(1-\theta)$$

And finally

$$0 = \frac{\partial \ell(\theta)}{\partial \theta} = \frac{\sum x_i}{\theta} + \frac{n - \sum x_i}{1-\theta}(-1)$$

so solve for θ

$$\begin{aligned}
\frac{\sum x_i}{\theta} &= \frac{n - \sum x_i}{1-\theta} \\
(1-\theta)\sum x_i &= \theta(n - \sum x_i) \\
\sum x_i - \theta \sum x_i &= n\theta - \theta \sum x_i \\
\sum x_i &= n\theta
\end{aligned}$$

therefore

$$\hat{\theta} = T_n(\theta) = \frac{\sum x_i}{n} = \bar{x}$$

Esempio 11.4.2 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x-2}{\theta}}$$

with $X \geq 2$ and $\theta > 0$. compute the maximum likelihood estimator for the parameter θ

1. $\hat{\theta} = \overline{X}$
2. none of these
3. $\hat{\theta} = \overline{X}/2$
4. $\hat{\theta} = \overline{X} - 2$

Si ha che

$$\begin{aligned}
\mathrm{L}(x,\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{\frac{2-x_i}{\theta}} = \frac{1}{\theta^n} \prod_{i=1}^{n} e^{\frac{2-x_i}{\theta}} \\
\ell(x,\theta) &= -n\log\theta \sum_{i=1}^{n} \log e^{\frac{2-x_i}{\theta}} = -n\log\theta \sum_{i=1}^{n} \frac{2-x_i}{\theta} \\
&= -n\log\theta + \frac{1}{\theta} n2 + \frac{1}{\theta} n\overline{x} \\
\frac{\partial \ell(x,\theta)}{\partial\theta} &= -\frac{n}{\theta} - 1(n2)\frac{1}{\theta^2} + n\overline{x}\frac{1}{\theta^2} = \frac{-n\theta - 2n + n\overline{x}}{\theta^2}
\end{aligned}$$

And

$$\begin{aligned}
0 &= \frac{-n\theta - 2n + n\overline{x}}{\theta^2} \\
0 &= -n\theta - 2n + n\overline{x} \\
\hat{\theta} &= \overline{x} - 2
\end{aligned}$$

Esempio 11.4.3 (Esame vecchio viroli)**.** A random variable X has density function

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x-3}{\theta}}$$

with $X \geq 2$ and $\theta > 0$. compute the maximum likelihood estimator for the parameter θ

1. $\hat{\theta} = \overline{X} - 3$ (corretta stando a quando suggerito da altri)
2. $\hat{\theta} = \frac{1}{\overline{X}-3}$
3. $\hat{\theta} = \overline{X}$
4. $\hat{\theta} = 1/\overline{X}$

Sviluppo uguale a quello di sopra

11.4.2 Bayesian framework

On the contrary $f(\theta|\mathbf{x})$ is a probability function of θ given the data we have observed. This is more logical.
We can rewrite the posterior function according to the bayesian rule/theorem:

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

as follows:

$$f(\theta|\mathbf{x}) = \frac{f(\theta, \mathbf{x})}{f(\mathbf{x})} = \frac{f(\theta) \cdot f(\mathbf{x}|\theta)}{f(\mathbf{x})} \propto f(\theta) \cdot f(\mathbf{x}|\theta)$$

The last used "proportional to" because we can ignore the denominator.

- $f(\mathbf{x})$ at denominator does not depend on the model, eg we don't depend on θ. So the denominator is the probability of observing the data without any specific probabilistic model (according to any possible probabilistic model); btw it can be rewritten as the integral of all possible models/values of theta (think law of total probabilities)

  $$f(\mathbf{x}) = \int_{\Theta} f(\theta) f(\mathbf{x}|\theta)\, \mathrm{d}\theta$$

  is the marginal distribution of X.

- $f(\theta)$ is called prior on θ: it's the probability without observing data (before the experiment);

- $f(\mathbf{x}|\theta)$ is the likelihood so the probability of data given θ

So we say the posterior function is proportional to the product between the *prior* and *likelihood*.
The two approaches (frequentist and bayesian) are linked (at the numerator): in one approach one have a prior, in the other not, this is the only difference.

Esempio 11.4.4. Imagine we have some observation from a Bernoulli of unknown parameter, $X_1, \dots X_n \sim \text{Bern}(\theta)$ iid; imagine we have a *prior distribution* for the parameter, that is $\theta \sim \text{Beta}(a, b)$, with a, b known. Find the Bayes

estimator for θ.
We want to find the posterior distribution that is

$$
\begin{aligned}
f(\theta|\mathbf{x}) = \frac{f(\theta)\cdot f(\mathbf{x}|\theta)}{f(\mathbf{x})} &\overset{(1)}{\propto} f(\theta) f(\mathbf{x}|\theta) \\
&\overset{(2)}{=} \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\theta^{a-1}\cdot(1-\theta)^{b-1}}_{\text{prior}}\cdot\underbrace{\prod_{i=1}^{n}\theta^{x_i}\cdot(1-\theta)^{1-x_i}}_{\text{likelihood}} \\
&= c\cdot\theta^{a-1}\cdot(1-\theta)^{b-1}\cdot\theta^{\sum x_i}\cdot(1-\theta)^{n-\sum x_i} \\
&\overset{(4)}{=} c\cdot\underbrace{\theta^{\sum x_i+a-1}\cdot(1-\theta)^{n-\sum x_i+b-1}}_{\text{kernel of a Beta}\,(\cdot)}
\end{aligned}
$$

where

- (1) since we want to maximize $f(\theta|\mathbf{x})$ we can ignore $f(\mathbf{x})$ (at its denominator) and all the terms that do not depend on θ.
- (2) we substituted the density for theta (prior) from a beta distribution with parameters a and b (having θ as our x) and likelihood assuming iid observation (therefore the product) of bernoulli distribution with a given θ fixed
- (3) we rewrite the normalization constant of the gamma $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = c$ (because it does not depend on θ) and sums of x_i due to the product
- in (4) after putting terms together, we recognize the kernel of a Beta, especially Beta $(a+\sum x_i, b+n-\sum x_i)$. The *posterior is proportional to a kernel of a Beta so the posterior is a Beta*: this because we know the posterior is a proper density (constructed according to the Bayes rule) and is equal to the kernel of a beta times normalization constant not considered (we don't care about normalization constant, it doesn't alter the distribution, only set his integral to 1). Therefore the posterior is a beta. So In bayesian statistics our aim is to identify the kernel of the posterior; that is enough. therefore we conclude that $\theta|\mathbf{x} \sim$ Beta $(a+\sum x_i, b+n-\sum x_i)$: here in bayesian approach we have a full distribution for the parameter estimator.

Three important facts :

1. when the prior $f(\theta)$ and the posterior $f(\theta|\mathbf{x})$ belong to the same probabilistic model (as in the case above, both have Beta), we say they are **conjugate**
2. in the posterior parameters (a, b, $\sum x_i$, $n-\sum x_i$) we have the contribution of both prior (a,b) and likelihood ($\sum x_i$, $n-\sum x_i$)
3. we have a full probability distribution function for θ, that is $f(\theta|\mathbf{x})$. Then what is the best representative value/ our *estimate* for θ? In the literature we have two solutions:
   (a) mode: this is a good choice because it maximizes $f(\theta|\mathbf{x})$

(b) mean: $\mathbb{E}[\theta|\mathbf{x}] = \hat{\theta}$ is used if the posterior distribution is asymmetric (as was in this example: having the two parameters of the beta become very different the beta becomes asymmetric); in our example

$$\hat{\theta} = \frac{a + \sum x_i}{a + \sum x_i + b + n - \sum x_i} = \frac{a + \sum x_i}{a + b + n}$$

Nowadays mean is more used than mode: computing mean is easier than computing a mode and it's handy for both symmetric and asymmetric distributions.

Esempio 11.4.5. What happens to the previous problem if we take as prior a lognormal density, for instance $\theta \sim \text{LogN}(\log(0.5), 0.1)$? The density of the lognormal theta (*prior*) is reported below (only the kernel *without* the normalization constant because in a bayesian framework we forget about it).

$$f(\theta) \propto \frac{1}{\theta} \exp\left(-\frac{1}{2}\frac{(\log\theta - \log 0.5)^2}{0.1}\right)$$

In this case the posterior will be

$$f(\theta|\mathbf{x}) \propto \frac{1}{\theta} \exp\left(-\frac{1}{2}\frac{(\log\theta - \log 0.5)^2}{0.1}\right) \cdot \theta^{\sum x_i} \cdot (1-\theta)^{n - \sum x_i}$$

In this case we cannot do further semplification, so $f(\theta|\mathbf{x})$ is known from an analytical point of view but we are not able to draw values from it: it's a kernel of a distribution we don't recognize, it's complicate.
To reconstruct the distribution (to obtain the estimate) we have to sample from it. We can do it by accept-reject algorithm (or one could do sampling-resampling, we didn't). Once we have generated many values from the target we have reconstructed the distribution of the target, we can take the mean, and it is the bayesian estimator.
How can we do it: remember that for accept-reject algorithm we should have a target $\pi(x)$ such as $\pi(x) \leq M \cdot$ proposal: in our case the target is the posterior $f(\theta|\mathbf{x})$ and we want to generate many values from it. It's simple, look at the step by step explanation below:

$$f(\theta|\mathbf{x}) \overset{(1)}{\propto} f(\theta) \cdot \text{L}(\mathbf{x}|\theta) \overset{(2)}{\leq} \underbrace{f(\theta)}_{\text{proposal}} \cdot \underbrace{\text{L}\left(\mathbf{x}|\hat{\theta}\right)}_{\text{maximum likelihood } M}$$

- in (1) the target $f(\theta|\mathbf{x})$ is proportional to prior $f(\theta)$ times the likelihood $\text{L}(\mathbf{x}|\theta)$ as usual;
- in (2) it is lower or equal to the same prior times the likelihood evaluated at it maximum point for θ, $\text{L}\left(\mathbf{x}|\hat{\theta}\right)$ (that is the likelihood of our maximum likelihood estimate). At this point, then, the idea is one can take the maximum likelihood as the constant M of the accept reject method, and the prior as the proposal.

So to generate value from the posterior:

TODO: da rivedere

1. we first solve the maximum likelihood problem to obtain $\mathrm{L}\left(\mathbf{x}|\hat{\theta}\right)$ for our ml estimate; so for a Bernoulli distribution we should find theta which maximizes the likelihood of a bernoulli distribution (and for the bernoulli distribution its the mean, look below) and calculate the likelihood of the sample with it. The found values constitutes M, a number

2. then we start generating values from a proposal (eg Lognormal)

3. compute the ratio between the target evaluated at the value we have drawn, divided by M times the probability of the proposal with that value, and we accept or reject according to this ratio. Accept reject algorithm can be used even if we don't know the normalization constant (it works) and this is the case

11.4.3 Final remarks

Osservazione 313. We have seen maximum likelihood estimation and bayesian estimation (with two different priors) of a bernoulli distribution

Teorema 11.4.1 (Relation between direct and inverse problems (frequentist and Bayes approach))**.** *When $n \to +\infty$, likelihood and Bayesian estimation become equivalent*

Proof. When n increases the importance of the prior becomes negligible so likelihood and posterior become similar □

Esempio 11.4.6. Es see the parameters of the posterior of the previous examples $\theta|\mathbf{x} \sim \mathrm{Beta}\left(a + \sum x_i, b + n - \sum x_i\right)$. When $n \to +\infty$ $\sum x_i$ increases a lot, as well as $n - \sum x_i$ increases: the effect of the prior (characterized by a and b become small compared to the contribution of the likelihood). This happen all the times.

Osservazione importante 98. So we can argue likelihood vs bayesian, but when n increases it doesn't matter, they go in the same direction.

11.5 Property of maximum likelihood estimators

- invariance
- efficiency

11.5.1 Invariance

Definizione 11.5.1 (Invariance)**.** If T_n is a maximum likelihood estimator for θ, then *any* function $\mathcal{T}$ of the estimator T_n, $\mathcal{T}(T_n)$, is the maximum likelihood estimator for $\mathcal{T}(\theta)$.

Esempio 11.5.1. $\log \bar{x}$ is the mle of $\log \theta$ in the previous example.

11.5.2 Efficiency

11.5.2.1 Fisher information

Osservazione importante 99. In order to talk about efficiency we need to talk about an important quantity we have in statistics which is the **Fisher information**. In different book there are different notations for this concept.
There are *three* different definitions.

Definizione 11.5.2 (Simple Fisher information)**.** It's the second derivative with respect to θ of the loglikelihood, with changed sign, that can be observed in my sample (so it's a function):

$$i(\theta) = i_n(\theta) = -l''(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta)$$

it's also denoted as $i_n(\theta)$ to emphasize we have a (it's derived on a) sample size of dimensionality n.

Definizione 11.5.3 (Expected Fisher information)**.** Its the expected value for the previous one:

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$$

It's a teoretical quantity that could be observed having all the sample.
If the observation are iid it's denoted also in some book as

$$I_n(\theta) = nI_1(\theta)$$

here again to stress the dimensionality of sample size we would compute the expected value; the equation above state that if obs are iid, the information coming from a sample of dimensionality n is n time the information coming from a sample of dimensionality 1

Osservazione importante 100. It's possible to prove that $I(\theta)$ is also equal to the expectation of the first derivative of the log likelihood squared:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2\right]$$

Osservazione importante 101. Therefore it's possible to prove that the square of the first derivative of the log likelihood is always equal to minus the second derivative of the log likelihood

$$\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2 = -\frac{\partial^2}{\partial^2\theta}\ell(\theta) \tag{11.5}$$

TODO: non mi risulta

Esempio 11.5.2 (Esame vecchio viroli)**.** a random variable X is supposed to follow the distribution

$$f(x) = \theta(1-\theta)^{x-1}$$

with $x \in \mathbb{N}^+$ and $\theta \in (0,1)$ and $\mathbb{E}\left[X\right] = 1/\theta$. Compute the expected fisher information $I_n(\theta)$

1. $I_n = \frac{\theta^2(\theta-1)}{n}$
2. $I_n = \frac{n}{\theta(1-\theta)}$
3. $I_n = \frac{n}{\theta^2(1-\theta)}$: suggerita da taluni
4. cannot be derived

we have that

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i-1} = \theta^n \prod_{i=1}^{n} (1-\theta)^{x_i-1}$$

$$\ell(\theta) = n\log\theta + \sum_{i=1}^{n}(x_i-1)\log(1-\theta) = n\log\theta - n\log(1-\theta) + \log(1-\theta)\sum_{i=1}^{n} x_i$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{n}{\theta} - n\frac{1}{1-\theta}(-1) + n\overline{x}\frac{1}{1-\theta}(-1) = \frac{n}{\theta} + \frac{n}{1-\theta} - \frac{n\overline{x}}{1-\theta}$$

$$\frac{\partial^2 \ell(\theta)}{\partial^2 \theta} = -\frac{n}{\theta^2} - \frac{n}{(1-\theta)^2} + \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2}$$

$$i(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial^2 \theta} = \frac{n}{\theta^2} + \frac{n}{(1-\theta)^2} - \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2}$$

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[\frac{n}{\theta^2} + \frac{n}{(1-\theta)^2} - \frac{\sum_{i=1}^{n} x_i}{(1-\theta)^2}\right] = \frac{n}{\theta^2} + \frac{n}{(1-\theta)^2} - \frac{n\,\mathbb{E}\left[X_i\right]}{(1-\theta)^2} = \frac{n}{\theta^2} + \frac{n}{(1-\theta)^2} - \frac{n}{(1-\theta)^2 \cdot \theta}$$

$$= \ldots = \frac{n(1+2\theta^2-3\theta)}{\theta^2(1-\theta)^2}$$

Definizione 11.5.4 (Observed Fisher information)**.** Differently from previous which are function of theta and we don't explicit the value of theta, here is a value/evaluated : it minus the second derivative of the log likelihood evaluated at the estimate of theta:

$$i(\hat{\theta}_n) = -l''(\theta)\big|_{\theta=\hat{\theta}_n}$$

We take the first information and we evaluate it for a value, which is the estimate of theta and so we have a value.

Esempio 11.5.3 (Esame vecchio viroli)**.** A random variable X is supposed to follow a distribution whose probability function is for $\theta > 1$ and $X > 1$

$$f(x;\theta) = \theta x^{-\theta-1}$$

consider the situation

- n = 4
- $x_1 = 1.5$, $x_2 = 2.1$ $x_3 = 1.9$ and $x_4 = 1.8$
- true $\theta = 2$

Compute the observed fisher information $i_n(\hat{\theta})$.
We have

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \theta x_i^{-(\theta+1)} = \theta^n \prod_{i=1}^{n} x_i^{-(\theta+1)}$$

$$\ell(\theta) = n\log\theta - \sum_{i=1}^{n}(\theta+1)\log x_i = n\log\theta - (1+\theta)\sum_{i=1}^{n}\log(x_i) = n\log\theta - \sum_{i=1}^{n}\log(x_i) + \theta\sum_{i=1}^{n}\log(x_i)$$

$$\frac{\partial \ell(\theta)}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n}\log(x_i)$$

$$\frac{\partial^2 \ell(\theta)}{\partial^2\theta} = -\frac{n}{\theta^2}$$

Now for the mle estimate we put $= 0$ the first loglikelihood derivative

$$0 = \frac{n}{\theta} - \sum_{i=1}^{n}\log(x_i) \iff \hat{\theta} = \frac{n}{\sum_{i=1}^{n}\log x_i} = \frac{4}{\log 1.5 + \log 2.1 + \log 1.9 + \log 1.8} = 1.683$$

and finally the observed fisher information is

$$i(\hat{\theta}) = \frac{n}{\hat{\theta}^2} = \frac{4}{1.683^2} = 1.412$$

Esempio 11.5.4. Consider $X_i \sim \mathrm{Exp}(\theta)$ with iid obs, $i = 1, \ldots, n$. We want to write the likelihood and then the loglikelihood (to compute the second derivative):

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \theta \cdot e^{-\theta x_i} = \theta^n \cdot e^{-\theta \sum_i x_i}$$

$$\ell(\theta) = n \cdot \log\theta - \theta \cdot \sum_{i=1}^{n} x_i$$

Going with the derivatives

$$\frac{\partial^1}{\partial\theta}(\ell(\theta)) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$

$$\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = -\frac{n}{\theta^2}$$

Starting from the first derivative one can obtain the maximum likelihood estimator of θ equating it to 0:

$$0 = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$$

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

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

Now we start deriving the simple Fisher information:

$$i(\theta) = -\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = \frac{n}{\theta^2}$$

It's a function of θ: if we change theta we have a different values for the information. The information can also be algebrically rewritten as product of $i_1(\theta)$ as follows:

$$\begin{aligned} i_1(\theta) &= \frac{1}{\theta^2} \\ i(\theta) &= n \cdot i_1(\theta) \end{aligned}$$

Now, for the second definition:

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[\frac{n}{\theta^2}\right] \overset{(1)}{=} \frac{n}{\theta^2}$$

in (1) expectation of a constant since we don't have x (otherwise we should compute the expectation). So in this case definition 1 is equivalent to definition 2, that is $i(\theta) = I(\theta)$ (this is not always the case).
Before the third definition we check that we get the same $I(\theta) = \frac{n}{\theta^2}$ by using the definition

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}(\ell(\theta))\right)^2\right]$$

Since we have already computed the first derivative:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}(\ell(\theta))\right)^2\right] = \mathbb{E}\left[\left(\frac{n}{\theta} - \sum x_i\right)^2\right]$$

Now we assume that $n = 1$ and then we will compute $I(\theta) = n \cdot I_1(\theta)$. To obtain $I_1(\theta)$ the information of theta from a single random variable X (of the above n) we adapt the information formula using first derivative:

$$\begin{aligned} I_1(\theta) &= \mathbb{E}\left[\left(\frac{1}{\theta} - X\right)^2\right] = \mathbb{E}\left[\frac{1}{\theta^2} + X^2 - 2\frac{X}{\theta}\right] = \frac{1}{\theta^2} + \mathbb{E}\left[X^2\right] - \frac{2}{\theta}\mathbb{E}\left[X\right] \\ &\overset{(1)}{=} \frac{1}{\theta^2} + \frac{2}{\theta^2} - \frac{2}{\theta^2} = \frac{1}{\theta^2} \end{aligned}$$

where in (1) we remembered the first moment of an $X \sim \text{Exp}\,(\theta)$ is $\mathbb{E}\left[X\right] = \frac{1}{\theta}$ while the second moment $\mathbb{E}\left[X^2\right] = \frac{2}{\theta^2}$.
So even using this first derivative based formula we get the same result as before:

$$nI_1(\theta) = \frac{n}{\theta^2} = I(\theta)$$

Finally the observed information (evaluated at the mle estimate $\hat{\theta} = \frac{1}{\bar{x}}$ which was derived previously) is:

$$i(\hat{\theta}) = \frac{n}{\hat{\theta}^2} = \frac{n}{\left(\frac{1}{\bar{x}}\right)^2} = n\bar{x}^2$$

Esempio 11.5.5 (Bernoulli distribution)**.** Let $X_i \sim \text{Bern}(\theta)$ iid rvs; the log-likelihood of the sample is

$$\ell(\theta) = \sum_i x_i \cdot (\log \theta) + (n - \sum x_i) \cdot \log(1-\theta)$$

The maximum likelihood estimator $\hat{\theta}_n = \bar{x}$. The first and second derivatives of loglikelihood are:

$$\frac{\partial^1}{\partial\theta}(\ell(\theta)) = \frac{\sum x_i}{\theta} + \frac{n - \sum x_i}{1-\theta}(-1)$$
$$\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = -\frac{\sum x_i}{\theta^2} - \frac{n - \sum x_i}{(1-\theta)^2}$$

The simple information is:

$$i(\theta) = -\frac{\partial^2}{\partial\theta\theta}(\ell(\theta)) = \frac{\sum x_i}{\theta^2} + \frac{n - \sum x_i}{(1-\theta)^2}$$

The expected information is:

$$\begin{aligned} I(\theta) = n \cdot I_1(\theta) &= n \cdot \mathbb{E}\left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}\right] \\ &= n \cdot \left[\frac{\theta}{\theta^2} + \frac{1}{(1-\theta)^2}(1-\theta)\right] = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)} \end{aligned}$$

The observed information is:

$$i(\hat{\theta}_n) = \frac{\sum x_i}{\hat{\theta}^2} + \frac{n - \sum x_i}{(1-\hat{\theta})^2} = \frac{n\bar{x}}{\bar{x}^2} + \frac{n - n\bar{x}}{(1-\bar{x})^2} = \ldots = \frac{n}{\bar{x}(1-\bar{x})}$$

11.5.2.2 Rao-Cramer theorem and efficiency

Osservazione importante 102. Back to efficiency, we needed fisher information for one of the most important theorem regarding efficiency of maximum likelihood estimators. This theorem gives a lower bound for the variance of an estimator $\text{Var}[T_n]$ under regularity conditions.

Teorema 11.5.1 (Rao-Cramer Theorem)**.** *Assume T_n is an unbiased estimator for θ, so (or for a transformation $\tau(\theta)$: it doesn't matter because of invariance property), so $\mathbb{E}[T_n] = \tau(\theta)$ (where τ can be identity). Supposing that (*regularity conditions*):*

1. *the domain or support D_X (set of values for X), does not depend on θ;*
2. *the likelihood $\text{L}(\theta)$ have first and second derivatives (not infinite etc);*
3. *there should be exchangability between integral and derivative*

$$\frac{\partial}{\partial\theta}\int f(\mathbf{x}|\theta)\,\mathrm{d}x = \int \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)\,\mathrm{d}x$$

This is regularity condition needed for math development of the theorem (this situation holds most of the time)

4. *we should have that the expected information is positive but finite that is, using the formula based on the first derivative:*

$$0 \leq \mathbb{E}\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right] < +\infty, \quad \forall \theta \in \mathbf{\Theta}$$

Under these condition, not only the maximum likelihood estimator is unbiased (that is $\mathbb{E}[T_n] = \tau(\theta)$*) but there is a lower bound for its variance equal to the square of first derivative of* τ *over the expected Fisher information:*

$$\operatorname{Var}[T_n] \geq \frac{[\tau'(\theta)]^2}{I(\theta)} = \frac{[\tau'(\theta)]^2}{\mathbb{E}[-\ell''(\theta)]} = \frac{[\tau'(\theta)]^2}{\mathbb{E}[\ell'(\theta)^2]} \tag{11.6}$$

Furthermore in 11.6 the equality holds (so the variance is equal to the ratio) if and only if we have that the first derivative of the loglikelihood can be written in the product of two parts, that is:

$$\frac{\partial \ell(\theta)}{\partial \theta} = k(\theta) \cdot (T_n - \tau(\theta))$$

with $k(\theta)$ *just a generic function of* θ *times the difference between the estimator and the quantity to be estimated.*
In this case we say the estimator T_n *is* UMVUE*, uniformly, miniumum variance, unbiased estimator and we say it's* fully efficient*.*

Esempio 11.5.6. If T_n is

- an unbiased estimator for θ (that is τ is the identity function) then $\tau(\theta) = \theta$ and $\tau'(\theta) = 1$, so at the numerator one gets 1, so in this case one have that

$$\operatorname{Var}[T_n] \geq \frac{1}{I(\theta)}$$

- an unbiased estimator for $\log\theta$ then $\tau(\theta) = \log\theta$ and $\tau'(\theta) = \frac{1}{\theta}$ and at the numerator one has $\frac{1}{\theta^2}$

Esempio 11.5.7. A situation where condition 1 does not hold is $\operatorname{Unif}(0, \theta)$ with θ the parameter of interest; θ is the domain upper bound.

Definizione 11.5.5 (Full efficiency)**.** When $\operatorname{Var}[T_n] = \frac{[\tau'(\theta)]^2}{I(\theta)}$, we say T_n is fully efficient.

11.5.2.3 Examples

Esempio 11.5.8. An example where regularity conditions are satisfied, here everything is regular.
We want to show that $\bar{x}$ (sample mean) is fully efficient for μ if the sample $(x_1, \ldots, x_n)$ comes from a gaussian distribution $\mathrm{N}(\mu, \sigma^2)$.
We start by computing the expected information:

$$I(\theta) = nI_1(\theta)$$

The loglikelihood function, for a sample of $n = 1$ considered I_1, is logarithm of the density of a normal (we don't have anymore the product). So given the density:

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{1}\left(\frac{x-\mu}{\sigma}\right)^2}$$

we have the loglikelihood for $n = 1$ being:

$$\log f(x|\mu, \sigma^2) = \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2} = -\log(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2}$$

And its first derivative

$$\frac{\partial}{\partial\mu} \log f(x|\mu, \sigma^2) = -2\frac{(x-\mu)}{2\sigma^2}(-1) = \frac{(x-\mu)}{\sigma^2}$$

The expected fisher information (for $n = 1$) is

$$I_1(\mu) = \mathbb{E}\left[\left(\frac{\partial}{\partial\mu} \log f(x|\mu, \sigma^2)\right)^2\right] = \mathbb{E}\left[\frac{(x-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^4}\mathbb{E}\left[(x-\mu)^2\right] \overset{(1)}{=} \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

where in (1), we have $\mathbb{E}\left[(x-\mu)^2\right] = \sigma^2$ by definition for a gaussian. Finally the expected information for n observation is:

$$I(\mu) = n \cdot I_1(\mu) = \frac{n}{\sigma^2}$$

Who is the lower bound of the variance for our estimator? Here we us $\bar{x}$ to estimate μ so $\tau(\mu) = \mu$ (identity function)

$$\mathrm{Var}\left[\bar{x}\right] \geq \frac{1}{I(\mu)} = \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n}$$

Now we are interested in the full efficiency for the estimator on μ not on σ^2.
We have computed the variance of sample mean for iid sample previously and it's

$$\mathrm{Var}\left[\bar{x}\right] = \mathrm{Var}\left[\frac{\sum X_i}{n}\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum X_i\right] = \frac{\sigma^2}{n}$$

Since in this case the variance is exactly equal to the lower possible bound, therefore $\bar{x}$ is fully efficient for μ.

Esempio 11.5.9. In this example one regularity condition is not satisfied and things doesn't work as expected.
Suppose $X \sim \mathrm{Unif}\,(0, \theta)$ with $\theta > 0$ and we are interested in estimating θ. The density of X for uniform

$$f(x, \theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$$

Let's find the expected information for the uniform distribution. Again it's convenient specify $n = 1$ and go for $I(\theta) = nI_1(\theta)$. If $n = 1$, the loglikelihood is $\log f(x, \theta) = -\log\theta$ therefore

$$I_1(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \log f(x; \theta)\right)^2\right] = \mathbb{E}\left[-\frac{1}{\theta}^2\right] = \frac{1}{\theta^2}$$

so the expected information is

$$I(\theta) = n \cdot \frac{1}{\theta^2} = \frac{n}{\theta^2}$$

Imagine now we have a sample of n observation we know have a likelihood with:

$$\mathrm{L}\left(\theta\right) = \prod_{i=1}^{n} \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n$$

and according to RC theorem, the variance of an unbiased estimator for θ shoud have

$$\mathrm{Var}\left[T_n\right] \geq \frac{1}{I(\theta)} = \frac{\theta^2}{n}$$

Now we take any unbiased estimator for θ; the estimator is

$$T_n = X_{(n)} \cdot \frac{n+1}{n}$$

this is unbiased: let's check. Remembering that the density of the maximum formula and applying it to the uniform:

$$f_{(n)}(x) = n \cdot F(x)^{n-1} \cdot f(x) = n\Big(\frac{x}{\theta}\Big)^{n-1}\frac{1}{\theta}$$

Now the expected value of the maximum is

$$\begin{aligned}\mathbb{E}\left[X_{(n)}\right] &= \int_0^\theta x \cdot f_{(n)}(x)\,\mathrm{d}x = \int_0^\theta x \cdot n \cdot \left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta}\,\mathrm{d}x = \int_0^\theta n\Big(\frac{x}{\theta}\Big)^{n}\,\mathrm{d}x \\ &= \frac{n}{\theta^n} \cdot \left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \theta \cdot \frac{n}{n+1}\end{aligned}$$

if this is expectation of the maximum, its clear that our estimator T_n is unbiased, because it fix the $n/(n+1)$ ratio:

$$\mathbb{E}\left[T_n\right] = \mathbb{E}\left[\frac{(n+1)}{n}X_{(n)}\right] = \frac{(n+1)}{n}\,\mathbb{E}\left[X_{(n)}\right] = \frac{n+1}{n}\frac{n}{n+1}\theta = \theta$$

So we have an unbiased estimator; according to Rao Cramer the variance lower bound of any unbiased estimator is $\geq \frac{\theta^2}{n}$. Now we want to compute the variance to check if it's equal to the lower bound and to say the estimator is fully efficient. We can compute the second moment of the maximum

$$\begin{aligned}\mathbb{E}\left[X_{(n)}^2\right] &= \int_0^\theta x^2 \cdot n \cdot \left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta}\,\mathrm{d}x = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\,\mathrm{d}x = \frac{n}{\theta^2} \cdot \left[\frac{x^{n+2}}{n+2}\right]_0^\theta \\ &= \frac{n}{n+2}\theta^2\end{aligned}$$

To the variance of the maximum is

$$\mathrm{Var}\left[X_{(n)}\right] = \mathbb{E}\left[X_{(n)}^2\right] - \mathbb{E}\left[X_{(n)}\right]^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\right)^2\theta^2 = \ldots = \frac{n\theta^2}{(n+1)^2(n+2)}$$

and the variance of our estimator is

$$\begin{aligned}\operatorname{Var}[T_n] &= \operatorname{Var}\left[\frac{n+1}{n}X_{(n)}\right] = \left(\frac{n+1}{n}\right)^2 \operatorname{Var}\left[X_{(n)}\right] \\ &= \left(\frac{n+1}{n}\right)^2 \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}\end{aligned}$$

So we have an estimator which is unbiased, which has a variance $\frac{\theta^2}{n(n+2)}$ that is lower than the lower bound $\frac{\theta^2}{n}$. This is called *super efficient* or *more than efficient*.
So if not all the regularity conditions are satisfied (here the domain $[0, \theta]$ depends on the parameter of interest) we could find an estimator with variance lower than the lower bound given by Rao-Cramer (if equal the estimator is *fully efficient*). Otherwise if all the regularity cond are satisfied the lower bounds *holds*.

Esempio 11.5.10 (Assignment 2 Viroli, Exercise 1 (cramer rao))**.** Let $X_1, \dots, X_n$ be independently poisson distributed with parameter λ, ie

$$f(x) = e^{-\lambda}\lambda^x/x!, \quad x = 0, 1, \dots$$

1. Determine the Maximum Likelihood (MLE) and the Method of Moments estimators of λ.
2. Determine the Mean Squared Error (MSE) of the MLE.
3. Find the Cramer-Rao lower-bound (CRLB) for an unbiased estimate of λ.
4. Determine whether the MLE is efficient, i.e., whether it attains the CRLB.

Considering $X_1, \dots, X_n \sim \operatorname{Pois}(\lambda)$ iid rvs:

1. for the **maximum likelihood estimator** we need to equate to 0 the first derivative of log likelihood and solving for λ:

$$\begin{aligned}\mathrm{L}(\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ \ell(\lambda) &= \sum_{i=1}^{n} \left[\log e^{-\lambda} + \log \lambda^{x_i} - \log x_i!\right] = \sum_{i=1}^{n} \left[-\lambda + x_i \log \lambda - \log x_i!\right] \\ &= \sum_{i=1}^{n} -\lambda + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log x_i! = -n\lambda + \log \lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log x_i! \\ \ell'(\lambda) &= -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i\end{aligned}$$

   Therefore

$$\ell'(\lambda) = 0 \iff -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \implies \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

   For the **method of moments** we equate in a system the j-sample moments $M_j = \frac{\sum_{i=1}^{n} x_i^j}{n}$ and the population moments $\mathbb{E}\left[X^j\right]$, solving for

the parameters of the latters. Here is a univariate problem so a single equation suffices; the equation is therefore $M_1 = \mathbb{E}[X]$; considered that if $X \sim \text{Pois}(\lambda)$ then $\mathbb{E}[X] = \lambda$ we have

$$M_1 = \frac{\sum_{i=1}^n x_i}{n} = \mathbb{E}[X] = \lambda \iff \hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}$$

So the two estimators have the same formula

2. let $T_n = \frac{\sum_i X_i}{n}$ be our estimator for λ. We want to compute

$$\text{MSE}(T_n) = \text{Var}[T_n] + \text{Bias}(T_n)^2$$

The variance of the estimator is

$$\text{Var}[T_n] = \text{Var}\left[\frac{\sum_i X_i}{n}\right] = \frac{1}{n^2}\text{Var}\left[\sum_i X_i\right] \overset{(1)}{=} \frac{\sum_i \text{Var}[X_i]}{n^2} \overset{(2)}{=} \frac{n\lambda}{n^2} = \frac{\lambda}{n}$$

in (1) because of independence of rvs, in (2) given that for $X_i \sim \text{Pois}(\lambda)$, we have that $\text{Var}[X_i] = \lambda$.
The estimator is unbiased since:

$$\begin{aligned}\text{Bias}(T_n) &= \mathbb{E}[T_n] - \lambda = \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] - \lambda = \frac{\sum_i \mathbb{E}[X_i]}{n} - \lambda = \frac{n\lambda}{n} - \lambda \\ &= 0\end{aligned}$$

Therefore:

$$\text{MSE}(T_n) = \text{Var}[T_n] + \text{Bias}(T_n)^2 = \frac{\lambda}{n} + 0^2 = \frac{\lambda}{n}$$

3. we have seen T_n is unbiased; to find the Cramer-Rao lower-bound (CRLB) for variance of the estimator, that is

$$\text{Var}[T_n] \geq \frac{[\tau'(\lambda)]^2}{I(\lambda)} \overset{(1)}{=} \frac{1}{I(\lambda)}$$

where (1) holds since T_n is unbiased estimator for λ (that is τ is the identity function with first derivative 1), we need to find the expected Fisher information $I(\lambda)$ at the denominator. To do so we need the second derivative of the loglikelihood, which minus is expected valued to find the information. We have

$$\ell''(\lambda) = \frac{\partial}{\partial\lambda}\left[-n + \lambda^{-1}\sum_{i=1}^n x_i\right] = -1 \cdot \lambda^{-2} \cdot \sum_{i=1}^n x_i = -\frac{\sum_{i=1}^n x_i}{\lambda^2}$$

and for the Fisher expected information:

$$I(\lambda) = \mathbb{E}[-\ell''(\lambda)] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{\lambda^2}\right] = \frac{1}{\lambda^2}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

Therefore for CR theorem:

$$\text{Var}[T_n] \geq \frac{1}{I(\lambda)} = \frac{\lambda}{n} = \text{CRLB}$$

4. since at point 2 we had that $\text{Var}[T_n] = \frac{\lambda}{n}$, and at point 3 $\text{CRLB} = \frac{\lambda}{n}$, we have that $\text{Var}[T_n] = \text{CRLB}$ and the estimator is currently said to be *fully efficient*.

11.5.3 Properties of ML estimators

Teorema 11.5.2. *1. invariance;*

2. *if an unbiased and fully efficient (variance equal to the lower bound) for θ exists, then it can be found by maximum likelihood*

3. *under non restrictive conditions maximum likelihood estimators are:*

   (a) *asympototically unbiased: we are not sure to find an unbiased estimator, but we are sure they are asympototically unbiased (eg estimator of the variance for the gaussian distribution is biased having n instead of $(n-1)$ but when n increases this small difference is negligible)*

   (b) *asymptotic efficient: they reach the lower RC bound for variance when $n \to +\infty$*

   (c) *weakly consistent $T_n \xrightarrow{p} \theta$*

   (d) *they are asymptotically gaussian: if one take T_n and rewrite it in the canonical form for the central limit theorem*

   $$\sqrt{n}(T_n - \theta) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2(\theta)\right)$$

   where $\sigma^2(\theta) = \frac{1}{I(\theta)}$ so it's the lower bound (this because of the first two)

   (e) *T_n is BAN estimators: best asymptotically normal estimators.*

Proof. Here we proof only that if an unbiased and fully efficient for θ exists, then it can be found by maximum likelihood.
In Rao Cramer the equality to the lower bound holds when

$$\frac{\partial}{\partial\theta} \log f(x;\theta) = k(\theta) \cdot [T_n - \theta]$$

At the same time in M.L. we solve

$$\frac{\partial}{\partial\theta} \log f(x;\theta) = 0$$

so

$$k(\theta) \cdot [T_n - \theta] = 0$$

implies T_n is M.L. estimator. □

Esempio 11.5.11 (Esame vecchio viroli)**.** Let $X_1, \ldots, X_n$ be a random sample from the density function

$$f(x) = \theta^2 x e^{-\theta x}$$

with $x > 0$ and $\theta > 0$. Find the asymptotic variance of the maximum likelihood estimator of θ (under the hypothesis that it is unbiased).

Asymptotically they go to the RC lower bound, that is being $\hat{\theta}$ unbiased $1/\mathbb{E}\left[-\ell''(\theta)\right]$, therefore we have

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \theta^2 x e^{-\theta x} = \theta^{2n} \prod_{i=1}^{n} x e^{-\theta x}$$

$$\ell(\theta) = 2n \log\theta + \sum_{i=1}^{n} \log\left(x e^{-\theta x_i}\right) = 2n\log\theta + \sum_{i=1}^{n} \left[\log x - \theta x_i\right]$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{2n}{\theta} - \sum_{i=1}^{n} x_i$$

$$\frac{\partial^2 \ell(\theta)}{\partial^2 \theta} = -\frac{2n}{\theta^2}$$

and finally the asymptotic variance

$$\mathrm{Var}\left[\hat{\theta}\right] = \frac{1}{\mathbb{E}\left[-\ell''(\theta)\right]} = \frac{1}{\mathbb{E}\left[\frac{2n}{\theta^2}\right]} = \frac{1}{\frac{2n}{\theta^2}} = \frac{\theta^2}{2n}$$

11.6 Assignment viroli

Esempio 11.6.1 (Assignment 2 Viroli, Exercise 2 (point estimation))**.** Let $X_1, \ldots, X_n$ be a sample of independent, identically distributed random variables, with density

$$f(x) = \frac{2}{3\theta}\left(1 - \frac{x}{3\theta}\right) \quad 0 < x < 3\theta$$

1. Determine the Method of Moments estimator $\hat{\theta}$ of θ.
2. Determine whether $\hat{\theta}$ is unbiased.
3. Determine whether $\hat{\theta}$ is consistent.
4. Why doesn't the Cramer-Rao lower bound apply to unbiased estimates of θ for this distribution?

Let $X_1, \ldots, X_n$ be iid rvs from an unknown model with density

$$f(x) = \frac{2}{3\theta}\left(1 - \frac{x}{3\theta}\right), \quad 0 < x < 3\theta$$

1. being a univariate problem, to solve it we equate the first sample moment to the first population moment. First of all we compute the first population moment

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{3\theta} x\left[\frac{2}{3\theta} - \frac{2x}{9\theta^2}\right] \mathrm{d}x = \int_0^{3\theta} \frac{2x}{3\theta}\,\mathrm{d}x - \int_0^{3\theta} \frac{2x^2}{9\theta^2}\,\mathrm{d}x \\
&= \frac{2}{3\theta}\int_0^{3\theta} x\,\mathrm{d}x - \frac{2}{9\theta^2}\int_0^{3\theta} x^2\,\mathrm{d}x = \frac{2}{3\theta}\left[\frac{x^2}{2}\right]_0^{3\theta} - \frac{2}{9\theta^2}\left[\frac{x^3}{3}\right]_0^{3\theta} \\
&= \frac{2}{3\theta}\left[\frac{9\theta^2}{2}\right] - \frac{2}{9\theta^2}\left[\frac{27\theta^3}{3}\right] = \frac{18\theta^2}{6\theta} - \frac{2\theta^3}{\theta^2} = 3\theta - 2\theta \\
&= \theta
\end{aligned}$$

So for the estimator we equate theorical moment above with sample moment $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, that is

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

2. $T_n = \frac{\sum_i X_i}{n}$, the estimator for θ is unbiased since:

$$\mathbb{E}[T_n] = \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] = \frac{\sum_i \mathbb{E}[X_i]}{n} = \frac{n\theta}{n} = \theta$$

3. the estimator T_n is consistent (both weakly and strongly) if $\lim_{n\to\infty} \mathrm{MSE}(T_n) = 0$; now given the decomposition of MSE and the fact that the estimator is unbiased we have that the estimator is consistent if $\lim_{n\to\infty} \mathrm{Var}[T_n] = 0$. But the estimator is the sample mean, with variance:

$$\mathrm{Var}[T_n] = \mathrm{Var}\left[\frac{\sum_i X_i}{n}\right] = \frac{\sum_i \mathrm{Var}[X_i]}{n^2} = \frac{n\,\mathrm{Var}[X_i]}{n^2} = \frac{\mathrm{Var}[X_i]}{n}$$

which clearly goes to 0 when $n \to \infty$ being the variance at the numerator a constant and specifically

$$\mathrm{Var}[X_i] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_0^{3\theta} x^2 \left[\frac{2}{3\theta} - \frac{2x}{9\theta^2}\right] dx - \theta^2 = \ldots = \frac{\theta^2}{2}$$

Therefore T_n is consistent.

4. RC lower bound doesn't apply because at least one of its regularity conditions (being $0 < x < 3\theta$, the support of the random variable depends on the parameter of interest θ) doesn't hold.

Esempio 11.6.2 (Assignment 2 Viroli, Exercise 3 (two parameter model, errato))**.** We observe a random sample, i.e., independent and identically distributed, $X = (X_1; ...; X_n)$ of size n from the following distribution,

$$f(x) = \theta_1 e^{-\theta_1(x-\theta_2)} \quad x \geq \theta_2$$

Determine the maximum likelihood estimator for $\boldsymbol{\theta} = (\theta_1; \theta_2)$. Given $X_1, \ldots X_n$ iid rvs coming from the distribution

$$f(x) = \theta_1 e^{-\theta_1}, \quad x \geq \theta_2$$

we determine maximum likelihood estimator for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ by maximizing log-likelihood.

$$\begin{aligned}
\mathrm{L}(\boldsymbol{\theta}, \boldsymbol{x}) &= \prod_{i=1}^{n} \theta_1 e^{-\theta_1(x_i-\theta_2)} \\
\ell(\boldsymbol{\theta}, \boldsymbol{x}) &= \sum_{i=1}^{n} \left[\log\theta_1 + \log e^{-\theta_1(x_i-\theta_2)}\right] = \sum_{i=1}^{n} \log\theta_1 + \sum_{i=1}^{n} (-\theta_1(x_i - \theta_2)) \\
&= n\log\theta_1 + \sum_{i=1}^{n} \theta_1(\theta_2 - x_i) = n\log\theta_1 + \sum_{i=1}^{n} \theta_1\theta_2 - \sum_{i=1}^{n} \theta_1 x_i \\
&= n\log\theta_1 + n\theta_1\theta_2 - \theta_1 n\bar{x} \\
&= n\left[\log\theta_1 + \theta_1\theta_2 - \theta_1\bar{x}\right]
\end{aligned}$$

Now the first derivative with respect to θ_1, θ_2 are

$$\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_1} = n\left[\frac{1}{\theta_1} + \theta_2 - \overline{x}\right]$$
$$\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \theta_2} = n\theta_1$$

I've tried following the suggestions given in class but I didn't find a way, either via system of equations or profile likelihood:

- by system of equations of nulling partial derivatives

$$\begin{cases} n\left[\frac{1}{\theta_1} + \theta_2 - \overline{x}\right] = 0 \\ n\theta_1 = 0 \end{cases} \quad \begin{cases} \frac{1}{\theta_1} + \theta_2 - \overline{x} = 0 \\ \theta_1 = 0 \end{cases} \quad \begin{cases} \theta_1 = \frac{1}{\overline{x} - \theta_2} \\ \theta_1 = 0 \end{cases}$$

- by profile likelihood we could try substituting $\hat{\theta}_1 = \frac{1}{\overline{x} - \theta_2}$ in place of θ_1 in the loglikelihood and maximize for θ_2. The profile likelihood $\ell^*(\theta_2, \boldsymbol{x})$ becomes

$$\ell^*(\theta_2, \boldsymbol{x}) = n\left[-\log(\overline{x} - \theta_2) + \frac{\theta_2}{\overline{x} - \theta_2} - \frac{\overline{x}}{\overline{x} - \theta_2}\right] = n\left[-\log(\overline{x} - \theta_2) - 1\right]$$

and

$$\frac{\partial \ell^*(\theta_2, \boldsymbol{x})}{\partial \theta_2} = n\left[-\frac{1}{\overline{x} - \theta_2}(-1)\right] = \frac{n}{\overline{x} - \theta_2}$$

but then we have

$$\frac{n}{\overline{x} - \theta_2} = 0$$

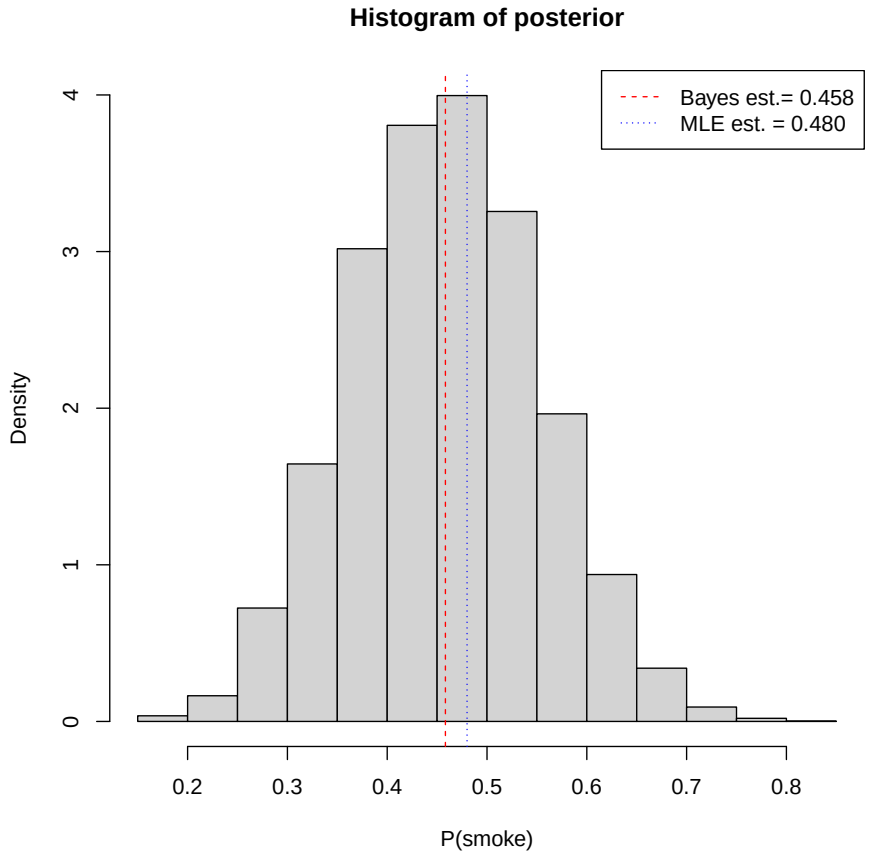
Esempio 11.6.3 (Assignment 2 Viroli, Exercise 4 (bayesian estimation))**.** A study on n=25 students revealed the smoker/non-smoker status, presenting the following results (where 1 indicates a smoker):

```
0  1  1 1 1 0 0 1 0 1 1 0 0 1 1 0 0 1 0 0 1 1 0 0 0
```

Knowing that the a priori probability of being a smoker follows a log-normal distribution with parameters -1 and 1, implement in R an algorithm for estimating the probability of being a smoker. The R program should be fully functional and able to generate 1000 observations from the posterior distribution to obtain the required estimate as a final output. Make sure that it is completely runnable starting from the data loading to the final estimate.
Reported here as well (with $n = 10000$), for completeness/security.

```
data <- c(rep(1, 12), rep(0,13))
mle <- mean(data) # mle hat(theta)

draw_posterior <- function(n = 1000, seed = 0001131115){
    # for reproducibility
    set.seed(seed)
```

```
    # main function and constants for the accept/reject algorithm
    lik <- function(theta) theta^{sum(data==1)}*(1-theta)^{sum(data==0)}
    rng_prior <- function(n) rlnorm(n = n, meanlog = -1, sdlog = 1)
    f_prior <- function(x) dlnorm(x = x,  meanlog = -1, sdlog = 1)
    propto_posterior <- function(x) f_prior(x = x) * lik(theta = x)
    M <- lik(mle)
    ## accept-reject
    ng = nit = 0                          # number of draws generated and number of it
    seq_draw = NULL                       # returned draws generated
    while (ng < n){
        nit = nit + 1
        u <- runif(n = 1)                 # extract u
        x <- rng_prior(n = 1)             # x: rng from the proposal/prior
        p_x <- f_prior(x = x)             # p(x): density proposal/prior
        pi_x <- propto_posterior(x = x)   # pi(x): target is posterior, here the numer
        if (u <= (pi_x /(M * p_x))){
            seq_draw <- c(seq_draw, x)
            ng = ng+1
        }
    }
    seq_draw
}

posterior <- draw_posterior(n = 10000)
(estimate <- mean(posterior))

## [1] 0.4583061

hist(posterior, freq = FALSE, xlab = "P(smoke)")
abline(v = estimate, col = 'red', lty = 'dashed')
abline(v = mle, col = 'blue', lty = 'dotted')
legend("topright",
       legend = c(sprintf("Bayes est.= %.3f", estimate),
                  sprintf("MLE est. = %.3f", mle)),
       col = c('red', 'blue'),
       lty = c('dashed', 'dotted'))
```


Histogram of posterior
Bayes est.= 0.458
MLE est. = 0.480
Density
0
1
2
3
4
0.2
0.3
0.4
0.5
0.6
0.7
0.8
P(smoke)

Chapter 12

Optimization methods

Osservazione 314. We start to study methods/computational tools to find maximum likelihood, from classical numerical techniquest to the EM algorithm (one of the most used algorithm in statistics).

12.1 Optimization techniques for maximum likelihood

Esempio 12.1.1 (Motivating example)**.** In some cases M.L. estimators cannot be written in closed form. Consider as example the Gamma distribution $X \sim \text{Gamma}\,(()\alpha, \beta)$ with $\alpha > 0$ (shape) and $\beta > 0$ (rate); the density is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

where

- $\frac{\beta^\alpha}{\Gamma(\alpha)}$ is a normalization constant
- $x^{\alpha-1}e^{-\beta x}$ is the kernel
- $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}\,\mathrm{d}x$
- our parameter of interest are $\boldsymbol{\theta} = \big[\alpha, \beta\big]$

In this case the likelihood and the logklikelihood functions (under iid obs) are:

$$\mathrm{L}\,(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \cdot \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \cdot e^{-\beta \sum_i x_i}$$

$$\ell(\mathbf{x}|\boldsymbol{\theta}) = n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\log x_i - \beta\sum_{i=1}^{n} x_i$$

Now

- the maximum likelihood estimators β can be found by setting the first derivative with respect to β (also called *score*) equal to zero:

$$\frac{\partial \ell(\mathbf{x}|\boldsymbol{\theta})}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0$$

and this leads, resolving for β, to a closed-form solution for β as function of α (and we don't know α)

$$\hat{\beta} = T_n = \frac{n\alpha}{\sum_{i=1}^{n} x_i} = \frac{\alpha}{\bar{x}}$$

- when it comes to α we have several possibilities:

  1. we do the same procedure and hope it don't come out an estimator of α as function of β; in that case one has the estimator for α and put the estimator for α in the solution for β above
  2. a second possibility is that in the process the estimator of α is a function of β (as above in switched roles). In this case we could either:
     - set up a *linear equation system*
     - a third possibility is to use/get the *profile loglikelihood*, that is: we substitute the estimator of β in the loglikelihood derived above, instead of β itself. Then one could compute the first derivative with respect to α (the only parameter remaining)

  For this example we do the same procedure as before computing the derivative of the original loglikelihood with respect to α: unfortunately here the ML estimator cannot be obtained in closed form since:

$$\frac{\ell(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha} = n \log \beta - n \frac{1}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha} + \sum_{i=1}^{n} \log x_i = 0$$

  in the loglikelihood we have three terms which depends on α and the middle one $-n \log \Gamma(\alpha)$ when deriving become the complex $-n \frac{1}{\Gamma(\alpha)} \frac{\partial \Gamma(\alpha)}{\partial \alpha}$. So we don't have an explicit solution for the parameter/we can't isolate alpha: it's inside an integral (the Γ) and inside a derivative of the Γ

12.1.1 Newton-Raphson algorithm

12.1.1.1 The algorithm

This is the simplest solution: the idea is optimize a function, in our case the loglikelihood. The idea is: if one cannot solve/maximize the $\ell(\theta)$ directly one can try to approximate locally by taking a quadratic function. We approximate the loglikelihood by a very good quadratic function of it

The assumptions are that:

1. $\ell(\theta)$ is differentiable (have first derivative)
2. the third derivative of the loglik should be not infinite: $\ell'''(\theta) < \infty$

3. the second derivative should not be null: $\ell''(\theta) \neq 0$

Considering a point $\theta_0 \in \mathbf{\Theta}$: we want to approximate the loglikelihood by a quadratic in this point. Developing the quadratic using Taylor expansion we have:

$$\begin{aligned}\ell(\theta) &\overset{(1)}{=} \ell(\theta_0) + (\theta - \theta_0)\ell'/\theta_0) + \frac{1}{2}(\theta - \theta_0)^2\ell''(\theta_0) + r_2(\theta, \theta_0) \\ &\overset{(2)}{=} \ell(\theta_0) + (\theta - \theta_0)S(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 H(\theta_0) + r_2(\theta, \theta_0)\end{aligned}$$

where in (1) we compacted the rest of the expansion

$$r_2(\theta, \theta_0) = \frac{1}{3!}\frac{\partial^3 \ell(\theta)}{\partial \theta^3}\Big|_{\theta=\theta_0}(\theta - \theta_0)^3 + \dots$$

and in (2) we merely replaced naming by:

- indicating $\ell'(\theta_0)$, the first derivative of the loglikelihood evaluated in θ_0 with $S(\theta_0)$ as *score function* (first derivative)
- indicating $\ell''(\theta_0)$, the second derivative, as $H(\theta_0)$ meaning *Hessian function*

Ignoring the remainder term $r_2(\theta, \theta_0)$ we have that the function at the point θ_0 is approximated by the following quadratics

$$\ell(\theta) \approx \ell(\theta_0) + (\theta - \theta_0)S(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 H(\theta_0) \tag{12.1}$$

which represents a quadratic approximation of $\ell(\theta)$ at θ_0. Taking the first derivative of with respect to θ and putting it equal to 0 for maximization (that is we're maximizing an approximation in θ_0) we get

$$\ell'(\theta) \approx S(\theta_0) + (\theta - \theta_0)H(\theta_0)$$

So equating to 0 and solving for θ leads to a solution which maximizes the log likelihood (its approximation)

$$S(\theta_0) + \theta H(\theta_0) - \theta_0 H(\theta_0) = 0 \implies \theta_1 = \theta_0 - \frac{S(\theta_0)}{H(\theta_0)}$$

from which we derive θ_1 which is a local approximation of $\hat{\theta}$.
If we apply the idea recursively we obtain the general iterative rule

$$\theta_{t+1} = \theta_t - \frac{S(\theta_t)}{H(\theta_t)} \tag{12.2}$$

from which we derive a sequence of values $\{\theta_t\}_{t\in\mathbb{N}}$.
The idea is approximate a function at a point, obtain a solution, use the solution as a new starting point and continue up to a stop.

12.1.1.2 Stopping criteria

We need a criteria to stop the sequence and obtain our estimate of θ. Ideally we want to stop where the point we get θ_t is near enough to the maximum likelihood point $\hat{\theta}$, that is $\left|\theta_t - \hat{\theta}\right| \leq \varepsilon$ with very small ε (eg $\varepsilon = 0.0001$). In practice we don't know $\hat{\theta}$ so we have several alternative stopping criteria (not involving it):

- we stop when the absolute difference between estimates at step is lower than a bound $|\theta_{t+1} - \theta_t| \leq \varepsilon$
- we stop with relative difference $\frac{|\theta_{t+1} - \theta_t|}{|\theta_t|} \leq \varepsilon$ which is better than the previous one since its robust for theta scale: if theta is small or high they are handled appropriately as well by going on relative variations
- we stop with low change of loglikelihood $|\ell(\theta_{t+1}) - \ell(\theta_t)| \leq \varepsilon$
- we stop with low score function $|S(\theta_{t+1})| \leq \varepsilon$ since at maximum point should be equal to 0

The last three are better (using one or the other depends on the coding), the first one is to be avoided.

12.1.1.3 Conditions for convergence

We have a problem: by Newton-Raphson one is not perfectly sure that we converge/end the algorithm, and therefore NR algorithm is not guaranteed to reach the global maximum.
Convergence might depend on the starting point θ_0 and if there are several local maximum: if we start near a local maximum that isn't a global maximum, the algorithm could converge to the local maximum not the global maximum. Other times one could diverge to $+\infty$ or $-\infty$ or go outside the support of the parameter etc.

We should have therefore some math conditions to be sure that NR converges. These are given by a theorem

Teorema 12.1.1. *If*

- $f(\theta)$ *has the first two derivatives*
- f *it's concave, that is* $H(\theta) < 0, \forall \theta \in \mathbf{\Theta}$

then we are sure that, by NR algorithm, we'll have that:

$$\lim_{t \to \infty} \theta_t = \hat{\theta}, \forall \theta_0 \in \mathbf{\Theta}$$

that is the algorithm converges from all the starting points.

12.1.2 Quasi-Newton algorithms

NR algorithm is the most popular but it is a special case of the *family* of methods (called quasi-newton algorithm) which implement an iterative solution searching following the step prescribed by the equation below

$$\theta_{t+1} = \theta_t - \alpha \frac{S(\theta_t)}{m(\theta_t)}$$

where

- α is a number you fix;
- m is another function/quantity one fixes;
- S is the score function

In this family it is possible to prove that the loglikelihood is increasing step after step, that is $\ell(\theta_{t+1}) > \ell(\theta_t)$, provided that $m(\theta_t) < 0$ and α is sufficiently small. Some cases of the QN algorithms are the following:

- if $\alpha = 1$ and $m(\theta_t) = H(\theta_t)$ we have the *Newton-Raphson algorithm*;
- if $\alpha > 0$ (but better small) and $m(\theta_t) = -1$ we have the *gradient descent algorithm*

  $$\theta_{t+1} = \theta_t + \alpha S(\theta_t)$$

  It's very used in neural networks because it's very simple, but its slower than NR: not using info provided by the hessian of the loglikelihood i don't need to compute the hessian all the time but use less information and it takes more iteration to converge.
  In the univariate case we have two directions: right step is the score is positive or left step is the score is negative.
- if $\alpha = 1$ and

  $$m(\theta_t) = \frac{S(\theta_t) - S(\theta_{t-1})}{\theta_t - \theta_{t-1}}$$

  we have the *secant method*. This formula for m resemble the definition of the derivative (pendenza della retta passante tra $(\theta_{t-1}, S(\theta_{t-1}))$ e $(\theta_t, S(\theta_t))$. So the secant method is basically the NR when i approximate the hessian (second derivative) with the slope of the first derivative (which is conceptually close).
  So the secant method is used when one cannot compute the hessian but want an approximation nonetheless, which is better than -1 implemented in gradient descent: so we expect that the secant method will be faster than the gradient method (it uses more information)
- if $\alpha > 0$ (but better small) and if $m(\theta_t) = -I(\theta_t)$ (minus expected Fisher information), we have the *Fisher scoring method*, very used by statistician:

  $$\theta_{t+1} = \theta_t + \alpha \frac{S(\theta_t)}{I(\theta_t)}$$

Convergence order/speed

Osservazione 315. It represents the 'average' speed of convergence of a method to the optimal point.

Definizione 12.1.1 (Convergence order)**.** An algorithm has order of convergence β if

$$\lim_{t\to\infty} \frac{\left|\theta_{t+1} - \hat{\theta}\right|}{\left|\theta_t - \hat{\theta}\right|^{\beta}} = c$$

where $c \neq 0$ is a constant and $\beta > 0$.

Osservazione 316. The higher β, the faster the convergence.

Osservazione importante 103. Regarding convergence order we have that for

- NR algorithm, $\beta = 2$ (quadratic order)
- Gradient descent algorithm, $\beta = 1$ (linear order)
- Secant method, $\beta = 1.62$ (superlinear order)
- Fisher scoring, $\beta = 2$ (quadratic order)

Extension to the p-dimensional case Most of the QN algorithms (not all) can be generalized to the case in which θ is a p-dimensional vector and we want to estimate several parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_p \end{bmatrix}$$

In this case the score is a vector containing the first p partial derivatives

$$S(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \dots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}$$

The Hessian is $p \times p$ matrix of the partial second derivatives:

$$H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1^2} & \dots & \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \dots & \dots & \dots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \dots & \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p^2} \end{bmatrix}$$

Esempio 12.1.2 (Multivariate problem with gaussian)**.** Let $x_1, \dots x_n$ be a iid sample from the gaussian density $\mathrm{N}(\theta_1, \theta_2)$. We are interested in estimating θ_1 and θ_2. We are interested in computing the score vector and the hessian matrix. The density function is

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2}\frac{(x_i-\theta_1)^2}{\theta_2}}$$

The loglikelihood $\ell(\theta_1, \theta_2)$ is:

$$\ell(\theta_1, \theta_2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

OO: check

The score (vector of partial first derivative of the loglikelihood):

$$S(\theta_1, \theta_2) = \begin{bmatrix} \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2} \\ \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{2\theta_2^2} - \frac{n}{2\theta_2} \end{bmatrix}$$

The Hessian:

TODO: check

$$H(\theta_1, \theta_2) = \begin{bmatrix} -\frac{n}{\theta_2} & -\frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} \\ -\frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} & -\frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^3} - \frac{n}{2\theta_2^2} \end{bmatrix}$$

The expected information (used in fisher-scoring) which is the expected value of minus the second-derivative/Hessian in the multivariate case:

$$I(\theta_1, \theta_2) = \mathbb{E}\left[-H(\theta_1, \theta_2)\right] = \mathbb{E}\left[\begin{bmatrix} \frac{n}{\theta_2} & \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} \\ \frac{\sum_{i=1}^{n}(x_i - \theta_1)}{\theta_2^2} & \frac{\sum_{i=1}^{n}(x_i - \theta_1)^2}{\theta_2^3} + \frac{n}{2\theta_2^2} \end{bmatrix}\right]$$
$$\overset{(1)}{=} \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{\theta_2^2} - \frac{n}{2\theta_2^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{bmatrix}$$

where in (1):

- in position (1,1) we have the expected value of a constant
- in (1,2) and (2,1) the expected value is 0 because the expected value of the sum (numerator) is the sum of expected values and it semplifies with $\mathbb{E}[x_i] - \mathbb{E}[\theta_1] = \theta_1 - \theta_1$ (since θ_1 is the mean).
- in $(2,2)$ the expected value $\mathbb{E}\left[(x_i - \theta_1)^2\right]$ at the numerator is by definition σ^2 that in this notation is θ_2, so considering iid its the expected value of the sum is n times θ_2; the second term again is a constant

In this case we have both the Hessian and the expected information; so we could use newton raphson (having the hessian) the fisher scoring (having the expected information) or the gradient descent as well (having the score).
The only method that can't be used here is the secant: in a multivariate problem, the approximation of the hessian cannot be done in this manner (because it's multivariate, it's a matrix not a function).
We have seen that the inverse of the expected information is the lower bound for rao cramer. Here in the multivariate setup rao cramer works as well and is the inverse of the information matrix gives us the lower bound of variances according to RC theorem

$$I^{-1}(\theta_1, \theta_2) = \begin{bmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{bmatrix}$$

Here having a diagonal matrix as information matrix, the inverse is easily obtained by inverting the diagonal terms.
Therefore:

- the lower bound for the variance of an estimator for θ_1 is θ_2/n (actually $\bar{x}$ has this variance and therefore it is fully efficient)
- the lower bound for the variance of an estimator for θ_2 is $2\theta_2^2/n$
- the ml estimator for θ_2 is the uncorrected variance $s(x) = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n}$ but it is biased (as shown previously). So we here cannot apply the lower bound given by rao cramer theorem (being the estimator biased)
- the sample variance $s(x) = \frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{n-1}$ is unbiased but it has variance $\frac{2\theta_2^2}{n-1}$ which is larger than the lower bound (so it's not fully efficient)
- so we can conclude that an unbiased and fully efficient estimator for θ_2 does not exist

Osservazione importante 104 (Extension to the p-dimensional case). Most algorithm can be generalized in this way. The extention of the secant algorithm is instead challenging and problematic.
For these situations other proposals exist. In particular, the BFCS algorithm works well in the p-dimensional context. BFGS stands for Broyden, Fletcher, Goldfarb and Shannon (the authors). In R it is implemented in the command `optim`.

12.1.3 Exercises oilspills

Osservazione 317. In this section two exercises on same dataset. The data regards crude oil spills of at least 1000 barrels from tankers in US waters during 1974-1999. Columns are:

- `year` considered (1974-1999)
- the count of `spills` (denoted by y)
- `importexport` (x) , the estimated amount of oil shipped through US waters as part of US import/export operations (adjusted for spillage in international or foreign waters)
- `domestic` (z), the amount of oil shipped through US waters during domestic shipments

Oil shipments are measured in billions of barrels of oil (Bbbl).

```
##   year spills importexport domestic
## 1 1974      2        0.720     0.22
## 2 1975      5        0.850     0.17
## 3 1976      3        1.120     0.15
## 4 1977      3        1.345     0.20
## 5 1978      1        1.290     0.59
## 6 1979      5        1.260     0.64
```

Esempio 12.1.3 (First part - Univariate case)**.** Here we assume that the variable spills, y_i, follows a Poisson process with parameter $\lambda_i = \theta_1 \cdot x_i$ (in this case lambda is not fixed, but the number of accidents can depend on the number of operations) where $i = 1, \ldots, 26$. We are interested in estimating θ_1:

1. write the likelihood function and the log-likelihood function:

$$\begin{aligned}
\mathrm{L}(\theta_1) &= \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} = \prod_{i=1}^{n} \frac{(\theta_1 x_i)^{y_i} e^{-\theta_1 x_i}}{y_i!} \\
\ell(\theta_1) &= \sum_{i=1}^{n} y_i \log(\theta_1 x_i) - \theta_1 x_i - \log(y_i!) \\
\mathbb{E}[Y_i] &= \lambda_i = \theta_1 x_1
\end{aligned}$$

2. compute the score and find the maximum likelihood estimator for θ_1.

$$\frac{\partial}{\partial \theta_1} \ell(\theta_1) = S(\theta_1) = \sum_{i=1}^{n} \left[y_i \frac{1}{\theta_1 x_i} x_i - x_i \right] = \sum_{i=1}^{n} \frac{y_i}{\theta_1} - \sum_{i=1}^{n} x_i$$

therefore our estimator can be obtaining equating above to 0 and solving for θ that is

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

so our estimate is computed as follows

```
y = oilspills$spills
x = oilspills$importexport
z = oilspills$domestic

### maximum likelihood estimation
(hattheta1 = sum(y) / sum(x))

## [1] 1.715778
```

in this case it would be not necessary for us to use numerical methods because a close formula for the estimator exists. However in the following we pretend that this is not the case and try to get the same estimate by the methods presented before

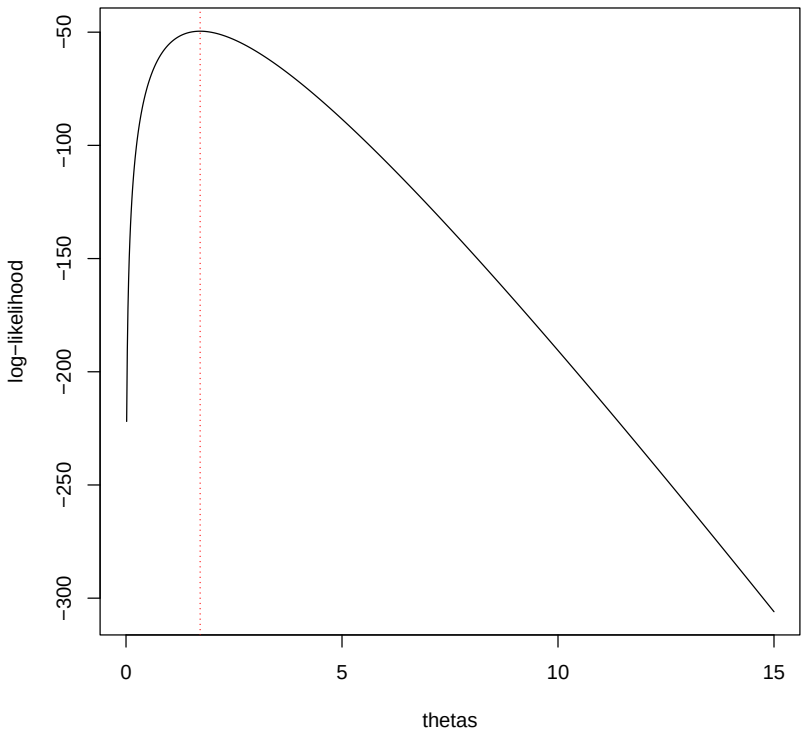
3. compute the Hessian. We have

$$H(\theta_1) = \frac{\partial}{\partial \theta_1} S(\theta_1) = \frac{\partial}{\partial \theta_1} \left(\sum_{i=1}^{n} \frac{y_i}{\theta_1} - x_i \right) = -\sum_{i=1}^{n} \frac{y_i}{\theta_1^2}$$

4. apply the Newton-Raphson algorithm and with starting value 0.6 and then with starting value 4. Compare the two solutions

```
## loglikelihood function, score and hessian as derived above
## loglik <- function(theta1, y, x){
loglik <- function(theta1){
    a = sum(y * log(theta1 * x))
    b = theta1 * sum(x)
```

```
    c = sum(log(factorial(y)))
    return(a-b-c)
}
S <- function(theta1) sum(y)/theta1 - sum(x)
H <- function(theta1) -sum(y)/(theta1^2)

## loglikelihood plotting for several possible values of theta
thetas <- seq(0, 15, length.out = 1000)
l = 0  # computed loglikelihood for the values of theta
for (i in 1:1000){
    l[i] = loglik(thetas[i])
}
plot(thetas, l,
     ylab = 'log-likelihood',
     type = "l", xlim = c(0,15))
## add the mle for check
abline(v = hattheta1, col = 'red', lty = 'dotted')
```




```
## the loglikelihood is regular/concave, with 1 mode, so newton
## raphson should work
nr <- function(theta_0){
```

```
    delta = 100       # value for stopping rules
    epsilon = 0.0001 # tolerance level
    it = 0            # iteration counter
    theta.all = theta_0 # vector with all thetas estimates
    theta = theta_0     # current value of theta considered
    while (delta > epsilon) {
        it = it + 1
        theta = theta - S(theta) / H(theta) # new theta estimate
        theta.all = c(theta.all, theta)  # save it
        delta = abs(S(theta)) # score criteria implemented
        print(paste("it=", it, " theta=", theta))
    }
    return(invisible(list(theta = theta.all, it = it)))
}

## starting from 0.6, in 5 iteration we have basically the same
## estimate
out06 = nr(theta_0 = 0.6)

## [1] "it= 1  theta= 0.990182608695652"
## [1] "it= 2  theta= 1.4089266204895"
## [1] "it= 3  theta= 1.6609001998968"
## [1] "it= 4  theta= 1.71402249113269"
## [1] "it= 5  theta= 1.71577589935419"

## instead here, it doesnt converge it diverges, we don't have
## solution
## out10 = nr(theta_0 = 4)
```

So not all the starting point for θ_0 are good, some starting point can be bad. Why we have this result: from the math point of view there's no reason, the likelihood plot is very nice. here the problem is related to the domain space of the poisson distribution (the parameter lambda of a poisson should be positive, but in some iteration the theta computed using the iterative rule is negative).
We are not sure that it will work

```
## while with theta =3 its ok
out06 = nr(theta_0 = 3)

## [1] "it= 1  theta= 0.754565217391304"
## [1] "it= 2  theta= 1.17728752238637"
## [1] "it= 3  theta= 1.54677464353514"
## [1] "it= 4  theta= 1.69913099791267"
## [1] "it= 5  theta= 1.71561618648414"
## [1] "it= 6  theta= 1.71577767968697"
```

So before givin up saying i made mistakes try with different starting values for θ_0.

Esempio 12.1.4 (Second part – Multivariate case)**.** In this case we assume that both variables (x and z) affect the outcome in this way

$$\lambda_i = \theta_1 x_i + \theta_2 z_i, \quad i = 1, \dots, 26$$

1. Write the likelihood and log-likelihood function. We have:

$$\begin{aligned} \mathrm{L}(\theta_1, \theta_2) &= \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \\ \ell(\theta_1, \theta_2) &= \sum_{i=1}^{n} (y_i \log \lambda_i - \lambda_i - \log y_i!) \\ &= \sum_{i=1}^{n} (y_i \log(\theta_1 x_i + \theta_2 z_i) - (\theta_1 x_i + \theta_2 z_i) - \log y_i!) \end{aligned}$$

   so $\mathbb{E}[Y_i] = \lambda_i = \theta_1 x_i + \theta_2 z_i$.

```
loglik <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    sum(y * log(lambda) - lambda - log(factorial(y)))
}

## check
loglik(c(1,1))

## [1] -48.06826
```

2. Write the score vector. The components are

$$\mathbf{S}(\theta_1, \theta_2) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ell(\theta_1, \theta_2) \\ \frac{\partial}{\partial \theta_2} \ell(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \frac{y_i x_i}{\theta_1 x_i + \theta_2 z_i} - \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} \frac{y_i z_i}{\theta_1 x_i + \theta_2 z_i} - \sum_{i=1}^{n} z_i \end{bmatrix}$$

   So

```
S <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    c(sum(y*x/lambda) - sum(x),
      sum(y*z/lambda) - sum(z))
}

# Score for starting point
S(c(1,1))

## [1] 1.1128229 0.3871771

# not so close to 0 for the starting thetas c(1,1), especially for the
# first: so probably the first theta will change more than the second
```

Observe that this time if we equate the two partial derivatives to zero we don't get a close solution for the two parameters. The only part of the score vector elemets where θ_1 and θ_2 are included is the denominator; in order to simiplify it we should multiply by the denominator, but the denominator depends on sum. So there is no way to construct a system or substitute or whatever. So there are no close form for the estimator and simple value for maximum likelihood estimates; therefore approximation and numerical methods comes very handy here.

$$\begin{cases} \frac{\partial}{\partial \theta_1}\ell(\theta_1,\theta_2) = 0 \\ \frac{\partial}{\partial \theta_2}\ell(\theta_1,\theta_2) = 0 \end{cases} \not\Longrightarrow \text{closed form solution for } \theta_1, \theta_2$$

3. Write the Hessian matrix:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

with $h_{12} = h_{21}$ we have

$$h_{11} = \frac{\partial^2}{\partial \theta_1^2}\ell(\theta_1,\theta_2) = -\sum_{i=1}^{n} \frac{y_i x_i^2}{(\theta_1 x_i + \theta_2 z_i)^2}$$

$$h_{22} = \frac{\partial^2}{\partial \theta_2^2}\ell(\theta_1,\theta_2) = -\sum_{i=1}^{n} \frac{y_i z_i^2}{(\theta_1 x_i + \theta_2 z_i)^2}$$

$$h_{12} = h_{21} = \frac{\partial^2}{\partial \theta_1 \partial \theta_2}\ell(\theta_1,\theta_2) = -\sum_{i=1}^{n} \frac{y_i x_i z_i}{(\theta_1 x_i + \theta_2 z_i)^2}$$

So

```
H <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    rval = matrix(0, nrow = 2, ncol = 2)
    rval[1,1] = -sum(y * x^2 / lambda^2)
    rval[2,2] = -sum(y * z^2 / lambda^2)
    rval[1,2] = rval[2,1] = -sum(y*x*z/lambda^2)
    rval
}

## Check with the (1,1) starting point
H(c(1,1))

##             [,1]      [,2]
## [1,] -18.315084 -9.607739
## [2,]  -9.607739 -8.469438
```

In this multivariate case Newton Raphson will be applied in this manner

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}^{-1}\mathbf{S}$$

4. Implement the gradient descent algorithm with $\alpha = 0.1$, $\alpha = 0.01$ and $\boldsymbol{\theta}_0 = (1, 1)$ and comment the results.

```
## gradient descent
gd <- function(thetas_0, alpha, tol = 10^-5){
    it = 1                                          # iteration id
    p = length(thetas_0)                            # n of params
    thetas = matrix(thetas_0, nrow = 1, ncol = p)  # matrix with saved res (it x p
    l.it = loglik(thetas_0)                          # loglik of the considered par
    print(c(it, l.it[it], thetas_0))                 #
    delta = 100
    while (delta > tol){
        S.it = S(thetas[it,])                       # score: is only needed for gra
        theta = thetas[it, ] + alpha * S.it         # new values of theta estimated
        thetas = rbind(thetas, theta)               # save them
        l.it = c(l.it, loglik(theta))               # compute the loglik and add to
        delta = abs((l.it[it+1]-l.it[it])/l.it[it])  # compute the check with re
        it = it+1
        print(c(it, l.it[it], theta))               # iteration results
    }
    out = list(likelihood = l.it, theta = thetas)
    return(invisible(out))
}

## with alpha = 0.1 it doesnt work

## gd(c(1,1), alpha = 0.1)

## loglik moves between  -49.7352 and -52.286
## and estimated parameters between
## 0.8088     0.7291
## 1.702     1.304
## the estimates of the parameters oscillates and dont converge. the
## jumps are due to the fact that alpha is too much high and the
## change in the theta after each step is too much over, so it bounces
## from one side to another of the maximum


## diminishing alpha it converges to 1.052 and 1.003 after 12 iteration
## (loglikelihood of the last couple is -48)
gd(c(1,1), alpha = 0.01)

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.000000 -48.055982   1.011128   1.003872
## [1]   3.000000 -48.048641   1.019867   1.006357
## [1]   4.000000 -48.044136   1.026811   1.007815
## [1]   5.000000 -48.041283   1.032393   1.008507
## [1]   6.000000 -48.039403   1.036935   1.008628
## [1]   7.000000 -48.038105   1.040677   1.008322
## [1]   8.000000 -48.037158   1.043800   1.007699
```

```
## [1]   9.000000 -48.036428   1.046443   1.006842
## [1]  10.000000 -48.035836   1.048711   1.005814
## [1]  11.000000 -48.035332   1.050684   1.004664
## [1]  12.000000 -48.034888   1.052425   1.003427

## using lower alpha again it takes much iteration but onverges to
## other values..
gd(c(1,1), alpha = 0.001)

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.000000 -48.066890   1.001113   1.000387
## [1]   3.000000 -48.065581   1.002202   1.000760
## [1]   4.000000 -48.064332   1.003267   1.001120
## [1]   5.000000 -48.063139   1.004309   1.001466
## [1]   6.000000 -48.062000   1.005329   1.001800
## [1]   7.000000 -48.060913   1.006328   1.002121
## [1]   8.000000 -48.059874   1.007305   1.002430
## [1]   9.000000 -48.058882   1.008262   1.002727
## [1]  10.000000 -48.057935   1.009198   1.003012
## [1]  11.000000 -48.057030   1.010115   1.003286
## [1]  12.000000 -48.056164   1.011013   1.003549
## [1]  13.000000 -48.055337   1.011892   1.003801
## [1]  14.000000 -48.054546   1.012753   1.004043
## [1]  15.000000 -48.053790   1.013596   1.004275
## [1]  16.000000 -48.053067   1.014421   1.004497
## [1]  17.000000 -48.052375   1.015230   1.004709
## [1]  18.000000 -48.051714   1.016023   1.004912
## [1]  19.000000 -48.051080   1.016799   1.005106
## [1]  20.000000 -48.050474   1.017560   1.005291
## [1]  21.000000 -48.049894   1.018305   1.005467
## [1]  22.000000 -48.049338   1.019036   1.005635
## [1]  23.000000 -48.048806   1.019752   1.005794
## [1]  24.000000 -48.048296   1.020453   1.005946
## [1]  25.000000 -48.047807   1.021141   1.006089
## [1]  26.000000 -48.047339   1.021815   1.006226
```

5. Implement the Newton-Raphson algorithm and comment the results. Try also with the starting points $\boldsymbol{\theta}_0 = (100, 1)$

```
# newton raphson
nr <- function(thetas_0, tol = 10^-5){
    it = 1
    p = length(thetas_0)
    thetas = matrix(thetas_0, nrow = 1, ncol =p)
    l.it = loglik(thetas_0)
    print(c(it, l.it[it], thetas_0))
    delta = 100
    while (delta>tol)  {
        S.it = S(thetas[it,])
```

```
        H.it = H(thetas[it,])
        theta = c(thetas[it,] - solve(H.it) %*% S.it)
        thetas = rbind(thetas, theta)
        l.it = c(l.it, loglik(theta))
        delta = abs((l.it[it+1] - l.it[it])/l.it[it])
        it = it+1
        print(c(it, l.it[it], theta))
    }
    out=list(likelihood = l.it, theta = thetas)
    return(invisible(out))
}

# same as befor the convergence is faster: in 3 iterations we have our solution
nr(c(1,1))

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.0000000 -48.0273087   1.0908311   0.9426757
## [1]   3.0000000 -48.0271623   1.0971283   0.9375748

## starting from a different point: manda in crash probabilmente per
## il valore della poisson con lambda ancora negativo
nr(c(100,1))

## [1]     1.000 -2514.898   100.000     1.000

## Warning in log(lambda):  Si è prodotto un NaN

## [1]     2.000       NaN -7920.792  2815.468

## Error in while (delta > tol) {:  valore mancante dove è richiesto
TRUE/FALSE
```

6. Derive the Fisher expected Information.

$$I(\theta_1, \theta_2) = -\mathbb{E}\left[H(\theta_1, \theta_2)\right] = \begin{bmatrix} \sum_{i=1}^{n} \frac{x_i^2}{\theta_1 x_i + \theta_2 z_i} & \sum_{i=1}^{n} \frac{x_i z_i}{\theta_1 x_i + \theta_2 z_i} \\ \sum_{i=1}^{n} \frac{x_i z_i}{\theta_1 x_i + \theta_2 z_i} & \sum_{i=1}^{n} \frac{z_i^2}{\theta_1 x_i + \theta_2 z_i} \end{bmatrix}$$

because $\mathbb{E}\left[Y_i\right] = \lambda_i = \theta_1 x_i + \theta_2 z_i$. So

```
I <- function(thetas){
    lambda = thetas[1] * x + thetas[2] * z
    rval = matrix(0,2,2)
    rval[1,1] = sum(x^2/lambda)
    rval[1,2] = rval[2,1] = sum(z*x/lambda)
    rval[2,2] = sum(z^2/lambda)
    rval
}
```

```
## test
I(c(1,1))

##           [,1]     [,2]
## [1,] 17.140991 9.669009
## [2,]  9.669009 8.020991
```

7. Implement Fisher-Scoring algorithm with $\alpha = 0.1$, $\theta = (1, 1)$ and $\theta = (100, 1)$.

```
## Fisher scoring
fs <- function(thetas_0, alpha, tol=10^-5){
    it = 1
    p = length(thetas_0)
    thetas = matrix(thetas_0, nrow = 1, ncol = p)
    l.it = loglik(thetas_0)
    print(c(it, l.it[it], thetas_0))
    delta = 100
    while (delta>tol)  {
        S.it = S(thetas[it,])  # score
        I.it = I(thetas[it,])  # expected info
        theta = c(thetas[it,] + alpha * solve(I.it) %*% S.it) # solve(I.it) is inverse of
        thetas = rbind(thetas, theta)
        l.it = c(l.it, loglik(theta))
        delta=abs((l.it[it+1]-l.it[it])/l.it[it])
        it = it+1
        print(c(it, l.it[it], theta))
    }
    out=list(likelihood=l.it,theta=thetas)
    return(invisible(out))
}

output <- fs(c(1,1), alpha = 0.1)

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.0000000 -48.0593625   1.0117785   0.9906284
## [1]   3.000000 -48.052445   1.022090   0.982632
## [1]   4.0000000 -48.0470535   1.0311241   0.9758091
## [1]   5.0000000 -48.0428410   1.0390436   0.9699881
## [1]   6.0000000 -48.0395421   1.0459904   0.9650233
## [1]   7.0000000 -48.0369530   1.0520873   0.9607901
## [1]   8.0000000 -48.0349170   1.0574411   0.9571825
## [1]   9.0000000 -48.0333131   1.0621448   0.9541095
## [1]  10.0000000 -48.0320473   1.0662791   0.9514938
## [1]  11.0000000 -48.0310468   1.0699148   0.9492689
## [1]  12.0000000 -48.0302547   1.0731132   0.9473781
## [1]  13.0000000 -48.0296268   1.0759282   0.9457727
## [1]  14.0000000 -48.0291284   1.0784068   0.9444112
## [1]  15.0000000 -48.0287323   1.0805900   0.9432578
```

```
fs(c(1,1), alpha = 0.2) # increasing alpha it is faster

## [1]   1.00000 -48.06826   1.00000   1.00000
## [1]   2.0000000 -48.0516085   1.0235571   0.9812569
## [1]   3.0000000 -48.0418533   1.0412553   0.9680014
## [1]   4.000000 -48.036068   1.054596   0.958636
## [1]   5.0000000 -48.0326017   1.0646818   0.9520339
## [1]   6.0000000 -48.0305060   1.0723257   0.9473956
## [1]   7.0000000 -48.0292292   1.0781324   0.9441523
## [1]   8.0000000 -48.0284461   1.0825532   0.9418981
## [1]   9.0000000 -48.0279630   1.0859257   0.9403434
## [1]  10.0000000 -48.0276636   1.0885035   0.9392818

## starting here it takes longer because very different from the solution
## fs(c(100,1), alpha = 0.1)
```

8. Compute the standard errors of the parameter estimates.

```
## standard errors
final_theta = output$theta[nrow(output$theta), ]
I(final_theta)

##           [,1]     [,2]
## [1,] 16.55713 9.455030
## [2,]  9.45503 7.922531

diag(solve(I(final_theta)))^0.5

## [1] 0.4354762 0.6295424
```

9. Solve the same problem with the function `optim` of R.

```
## To optimize the loglikelihood we use optim givin it the loglikelihood functio

## bfgs is an optimization method
## c(1,1) is the starting parameter values for loglik function
## at the end "control=list(fnscale=-1)" is to say to R to maximize instead of m
optim(c(1,1), fn = loglik, method = "BFGS", control = list(fnscale = -1))

## $par
## [1] 1.0971528 0.9375544
##
## $value
## [1] -48.02716
##
## $counts
## function gradient
```

```
##       18        5
##
## $convergence
## [1] 0
##
## $message
## NULL

## the maximum log likelihood is -48 and is obtained when the thetas are 1.0972 0.9376
```

12.2 EM algorithm

12.2.1 Introduction

Osservazione 318. In many situations for maximum likelihood where no close formula is available and the problem is complicated we cannot use the optimization techniques seen before so we need another tool.

- *Osservazione importante* 105 (EM (Expectation maximization) algorithm). The EM algorithm (Dempster et al., 1977) is one of the most used algorithm in statistics (and not only) for the maximum likelihood estimation in complex problems (for instance: latent variable models).

- It is an iterative method composed by two steps: E-step (E for expectation) and M-step (M for maximization):
  - E-step: we compute a conditional expectation given the data at the 'current' value of parameters.
  - M-step: we maximize the conditional expectation with respect to the parameters of interest.
  - goto the E-step again and repeat the procedure since convergence

Osservazione 319. We apply EM in mixture model estimation; so first we introduce mixture models and then EM and how to use it.

Osservazione 320. Nowadays exists other evolution of EM such as Stochastic EM which include a first bayesian step (so it's a mixed classical/bayesian method), but we focus on the simple EM algorithm

12.2.2 Mixture models

12.2.2.1 Introduction

Osservazione 321 (Mixture model idea). The distribution of weight for men and women is different; the distribution of men is normal with mean 73 and sd 9; the distribution for women is right skewed with mean 56 and sd 6.8. We have different distribution but we observe the total; we can use the information that behind the total distribution we have the composition of the two distributions. Mixture model is a typical example where classic maximum likelihood estimation procedures does not work.

Definizione 12.2.1 (Mixture model)**.** Let $(y_1, \ldots, y_n)$ be a random sample of size n for a rv Y. Imagine that Y as heterogeneous, that is the population from which it is observed is composed by different k sub-populations not known in advance (not observed heterogeneity). We assume that the density of one value can be seen as a blend of the composing distribution, weighted by probability of groups

$$f(y_j) = \sum_{i=1}^{k} \pi_i f_i(y_j) \tag{12.3}$$

where

- $j = 1, \ldots, n$ are patients
- $i = 1, \ldots, k$ are groups
- $f_i(y_j)$ are the *component densities*;
- π_i are the mixing proportions, called *a priori probabilities*, satisfying $\pi_i > 0$ and $\sum_{i=1}^{k} \pi_i = 1$.

So the mixture model is a mean of several distribution, in a way, weighted by π_i.

Osservazione 322. Off course we don't know π_i (in the weight example the proportion of men and women); these are parameter to be estimated, together with the parameter of the component densities. So in a mixture model we have to estimate:

1. parameters of the component densities
2. the mixing proportion π_i

Osservazione importante 106 (Mixture model: Interpretation). These are special kind of latent variable model, with latent variable Z.
Let Z_j be a *categorical variable* (pedice is ok, there's one for each patient) that assumes values $1, \ldots, k$ (which denotes the membership group) with probability $\pi_1, \ldots, \pi_k$, respectively.
The conditional density of y_j given the value assumed by Z, that is $z_j = i$, is the i-th component density $f_i(y_j)$ with $i = 1, \ldots, k$:

$$f(y_j | z_j = i) = f_i(y_j)$$

The random variable z_j can be thought of as 'label' of the component y_j belongs to.

Osservazione importante 107 (Allocation variable). For practical reasons we represent $\mathbf{Z}_j$ as a random vector with k components and one of its realization (for j-th patient) is:

$$\mathbf{z}_j = (0, 0, \ldots, 0, 1, 0, \ldots, 0)^T$$

$\mathbf{z}_j$ contains all 0's and value 1 in the i-th position (located in the position of the group), i.e. $z_{ij} = 1$.
The advantage of this reparameterization is that probabilistically speaking, the

random variable $\mathbf{Z}_j$ is now distributed as a multinomial random variable, with only $n = 1$ trial that leads to a success for one of the k categories with probability $\pi_1, \ldots, \pi_k$:

$$\mathbb{P}(\mathbf{Z}_j = \mathbf{z}_j) = \pi_1^{z_{1j}} \cdot \ldots \cdot \pi_k^{z_{kj}} \tag{12.4}$$

where the exponent are all 0 but one 1 so at the end we select just one π. The weights $\pi_1, \ldots, \pi_k$ can be viewed as *prior* probability that a statistical unit belongs to the i-th component and the equation above is equivalent to $\mathbb{P}(z_{ij} = 1) = \pi_i$.
Again $\boldsymbol{Z}_j$ is not an observable random variable (not observed heterogeneity!): it's a *latent variable*, here called *allocation variable*.

Osservazione 323. Other than prior probabilities, in a mixture model we can define also the *posterior* probability.

Definizione 12.2.2 (Posterior probabilities)**.** They are defined as probability that a unit belongs to the i-th component, given that we have observed its value y_j:

$$\begin{aligned}\tau_i(y_j) &= \mathbb{P}(\text{unit } j \in i\text{-th component} \mid y_j) = \mathbb{P}(z_{ij} = 1 | y_j) \\ &= \frac{\mathbb{P}(z_{ij} = 1 \cap y_j)}{\mathbb{P}(y_j)} = \frac{\mathbb{P}(z_{ij} = 1)\,\mathbb{P}(y_j | z_{ij} = 1)}{\mathbb{P}(y_j)} = \frac{\pi_i f_i(y_j)}{f(y_j)} \\ &= \frac{\pi_i f_i(y_j)}{\sum_{i=1}^{k} \pi_i f_i(y_j)}\end{aligned} \tag{12.5}$$

Osservazione 324. In the last equation we see the its a ratio between a part and the total so it's a probability

Esempio 12.2.1. In our example the posterior probability it's the probability of being a man given that our weight is 45 kg.

Osservazione importante 108 (Application: clustering). Mixture models are useful tools for clustering: we can assign each statistical unit to the most likely group based on posterior probabilities. This kind of clustering is called *model based clustering*.

Osservazione importante 109 (Application: density estimation). Aside from clustering, mixture models are also useful for density estimation. When k increases the mixtures are universal approximation tools: that is even the most strange empirical distribution can be approximated by a mixture of simpler distribution with parameters to be estimated.

12.2.2.2 Problems with classical estimation

Osservazione 325. Now we see that classical maximum likelihood is difficult and justify the need for EM algorithm.

Definizione 12.2.3 (Density revised)**.** If we consider $\boldsymbol{\theta}_i$ as parameter of the i-th group density function, then 12.3 becomes

$$f(y_j) = \sum_{i=1}^{k} \pi_i f_i(y_j; \boldsymbol{\theta}_i)$$

Now we can properly write the log likelihood

Definizione 12.2.4 (Log-likelihood of mixture)**.** The loglikelihood of our sample is given by:

$$\begin{aligned}\log f(y;\Psi) &= \log \prod_{j=1}^{n} f(y_j;\Psi) = \sum_{j=1}^{n} \log f(y_j;\Psi) \\ &= \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} \pi_i f_i(y_j;\boldsymbol{\theta_i}) \right) \end{aligned} \tag{12.6}$$

where all the parameters are described as $\Psi = \{\boldsymbol{\pi}, \boldsymbol{\theta}\}$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{k-1})$ and $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$ (the prior probabilities), and $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_k})$ the parameter vectors of the k composing distribution.

Osservazione importante 110 (Problems with ML estimation in mixture models). Suppose for a moment, for semplicity, that $\boldsymbol{\theta_i}$ is known. Then the estimate for the weights of the mixture can be found by maximizing the loglikelihood 12.6 only by pis:

$$\begin{aligned}\frac{\partial}{\partial \pi_i} \log f(y;\Psi) &\overset{(1)}{=} \frac{\partial}{\partial \pi_i} \left[\sum_{j=1}^{n} \log \left(\sum_{i=1}^{k-1} \pi_i f_i(y_j;\boldsymbol{\theta_i}) + \left(1 - \sum_{i=1}^{k-1} \pi_i\right) f_k(y_j;\boldsymbol{\theta_k}) \right) \right] \\ &\overset{(2)}{=} \sum_{j=1}^{n} \left[\frac{f_i(y_j;\boldsymbol{\theta_i}) - f_k(y_j;\theta_k)}{\sum_{i=1}^{k} \pi_i f_i(y_j;\boldsymbol{\theta_i})} \right] = 0 \end{aligned} \tag{12.7}$$

where

- in (1) we merely rewrited π_k (for convenience of what is going on afterwards with derivative) as difference between 1 and remaining and
- in (2) we performed the derivative where derivative of the sum is the sum of derivative and applying the $log(f(x))$ pattern of $\frac{1}{f(x)} f'(x)$.

Problem with ML and numerical estimation are the following

1. from 12.7 we cannot isolate π_i is at the denominator to obtain a close form estimator for it. So (even hypothesizing to know $\boldsymbol{\theta_i}$, which is not the case in practice) we can't get a closed formula. Therefore we should rely on optimization methods as done previously.
2. however if we have to estimate $\boldsymbol{\theta_i}$ $(i = 1, \ldots, k)$ as well besides the weights, then it's even more complicated (especially if they depend on the functional form for the component densities $f_i(y_j;\boldsymbol{\theta_i})$).
   So usually, not only π_i do not have closed form, but also $\boldsymbol{\theta_i}$ share the same issue.

In all this shit, the EM algorithm can simplify the estimation problem.

12.2.3 The EM algorithm

12.2.3.1 Intro

Definizione 12.2.5 (Misc naming)**.** We have:

- the observed sample $(y_1, \dots, y_n)$ is called *incomplete data*;
- the set $(z_1, \dots, z_n)$ represents the *latent variables*;
- the couple $\{y_j, z_j\}_{j=1,\dots,n}$ is called the *complete data*.
- $f(\mathbf{y}; \Psi)$ is the *incomplete density*
- $f(\mathbf{y}, \mathbf{z}; \Psi)$ is the *complete density*.

Osservazione 326. The idea of the EM algorithm is to (invent a strategy to) work with complete data even if we don't have it.
This is so because if we knew the complete data (knowing z), we could have close form solution on all the parameters; if one could work on the complete log-likelihood, instead on the incomplete one, the estimation problem would be pretty simple.

Osservazione importante 111 (Estimation of the complete log-likelihood). Imagine, that $\mathbf{z}_j$ is observed and not latent. The complete density can be decomposed into the product of two densities:

$$f(y, \mathbf{z}; \Psi) \overset{(1)}{=} f(\mathbf{z}; \Psi) \cdot f(y|\mathbf{z}; \Psi) \overset{(2)}{=} f(\mathbf{z}; \boldsymbol{\pi}) \cdot f(y|\mathbf{z}; \boldsymbol{\theta})$$

where the conjoint density is splitted as follows:

- in (1) the density of $y, \mathbf{z}$ is a conjoint density, so we split as product of the density of $\mathbf{z}$ times the conditional density of y given $\mathbf{z}$
- in (2) the density of $\mathbf{z}$, it is only parametrized using $\boldsymbol{\pi}$, while the conditional density of y given $\mathbf{z}$ is parametrized using only $\boldsymbol{\theta}$

So here we have a split of the parameters in two parts, which belongs to separate densities; this split semplifies a lot the estimation problem.
We now can write the log-likelihood which will be the sum of two terms:

$$\begin{aligned} \log f(y, \mathbf{z}; \Psi) &= \log \prod_{j=1}^{n} f(y_j, \mathbf{z}_j; \Psi) = \log \prod_{j=1}^{n} f(\mathbf{z}_j; \boldsymbol{\pi}) \cdot f(y_j|\mathbf{z}_j; \boldsymbol{\theta}) \\ &= \sum_{j=1}^{n} [\log f(\mathbf{z}_j; \boldsymbol{\pi}) + \log f(y_j|\mathbf{z}_j; \boldsymbol{\theta})] \end{aligned} \tag{12.8}$$

with

- $f(\mathbf{z}_j; \boldsymbol{\pi}) = \prod_{i=1}^{k} \pi_i^{z_{ij}}$ (the density of $\mathbf{z}_j$ take the multinomial form where the proper probability is selected)
- $f(y_j|\mathbf{z}_j; \boldsymbol{\theta}) = \prod_{i=1}^{k} f_i(y_j; \theta_i)^{z_{ij}}$ where again the exponent (uno 0 o 1) funge da activator/selector for the proper density

Notice that the first term of 12.8 involves only the weights $\boldsymbol{\pi}$, while the second term involves only the parameters θ of the component densities. Therefore if we compute the first derivative in 12.8 with respect to pis, the second part goes to zero and we have just the second part; with respect to thetas the first part is zero. This simplify a lot.

Precisely, the score (first derivative of loglik) with respect to $\boldsymbol{\pi}$ becomes:

$$\frac{\partial \log f(y,z;\Psi)}{\partial \pi_i} \overset{(1)}{=} \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{n} \log f(\mathbf{z}_j;\boldsymbol{\pi})\right) \overset{(2)}{=} \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{n} \log \prod_{i=1}^{k} \pi_i^{z_{ij}}\right)$$
$$= \frac{\partial}{\partial \pi_i}\left(\sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij} \log \pi_i\right) \tag{12.9}$$

where in (1) since the derivative with respect to the second term in 12.8 is zero, while in (2) we substituted according to the points below equation 12.8.
Here we have another problem: computing the derivative with respect to pis and equating to 0 we would get:

$$\sum_j \sum_i \frac{z_{ij}}{\pi_i} = 0$$

Then if we try to isolate π_i at the denominator there are problems. However here we can use the idea that the weights π_i are positive and sum to 1: $\pi_i > 0$ and $\sum_{i=1}^{k} \pi_i = 1$. This can be find by Lagrange multiplier or by doing a trick with the following transformation. If we parametry the π_i in this manner:

$$\pi_i = \frac{e^{w_i}}{\sum_{i'=1}^{k} e^{w_{i'}}}$$

and compute the score with respect to w_i (and not π_i) we get likelihood maximizing solution where π_i are positive and sums to 1. This second parametrization of the problem assure us that the pis satisfy the constraints; w can be anything and everything returns/is fine.
Therefore with this second trick, we plug the new definition of pis in 12.9 and maximize for wis. We get:

$$\begin{aligned}
\frac{\partial \log f(y,z;\Psi)}{\partial w_i} &= \frac{\partial}{\partial w_i}\left(\sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij} \log \frac{e^{w_i}}{\sum_{i'=1}^{k} e^{w_{i'}}}\right) \\
&= \frac{\partial}{\partial w_i}\left(\sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij} \log e^{w_i} - \sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij}\left(\log \sum_{i'=1}^{k} e^{w_{i'}}\right)\right) \\
&\overset{(1)}{=} \frac{\partial}{\partial w_i}\left(\sum_{j=1}^{n}\sum_{i=1}^{k} z_{ij} w_i - n\left(\log \sum_{i'=1}^{k} e^{w_{i'}}\right)\right) \\
&= \sum_{j=1}^{n} z_{ij} - n\frac{e^{w_i}}{\sum_{i'=1}^{k} e^{w_{i'}}} \\
&= \sum_{j=1}^{n} z_{ij} - n\pi_i
\end{aligned}$$

where in (1) we took $\log \sum_{i'=1}^{k} e^{w_{i'}}$ out of the double sum because has a different index i' and the double sum of z_{ji} is simply n (its a sum of vector with only one 1, one for each patient).

Then we obtain the maximum likelihood estimator in closed form by equating to 0:

$$\sum_{j=1}^{n} z_{ij} - n\pi_i = 0 \iff \hat{\pi}_i = \frac{\sum_{j=1}^{n} z_{ij}}{n}$$

so pi is a kind of mean of the zed (the probability of being in this group i is given by the proportion of 1 of the patients of this group)

Osservazione 327. Therefore, if we could consider the complete Likelihood instead of the incomplete one, we would have a closed form for the weights (and most of times also for the other parameters, but this depends on the functional form of the component densities).
But, of course, the problem is that **we do not know** z_{ij}! They are latent variables, not observed in practice.
The strength of the EM algorithm is that it allows to work with the complete likelihood in some way.

12.2.3.2 EM algorithm (Dempster, Laird, Rubin, 1977)

Imagine that y is the observed rv and z is the latent rv. The EM algorithm starts by considering the *maximization of the incomplete log-likelihood* $\max_{\Psi} \log f(y;\Psi)$; this can be reformulated as follows:

$$\begin{aligned}
\max_{\Psi} \log f(y;\Psi) &\overset{(1)}{=} \max_{\Psi} \log \sum_{z} f(y,z;\Psi) \\
&\overset{(2)}{=} \max_{\Psi} \log \sum_{z} f(z|y;\Psi') \frac{f(y,z;\Psi)}{f(z|y;\Psi')} \\
&\overset{(3)}{=} \max_{\Psi} \log E_{z|y;\Psi'} \left[\frac{f(y,z;\Psi)}{f(z|y;\Psi')} \right]
\end{aligned}$$

where in:

- (1) the marginal density of y is decomposed in the sum of joint density of y and z, by value of z (which being discrete lead us to a sum)
- (2) we multiply and divide by the same quantity $f(z|y;\Psi')$, that is the density of z given y with a set of parameters Ψ' different from Ψ (the set of parameters we don't know and want to maximize). Imagine that the set of parameter Ψ' is fixed/known (eg we set the parameters we want to estimate, we don't care if they are wrong. Eg we set $\pi_1 = 0.3, \pi_2 = 0.01$ and so on). If we fix these parameter we get a conditional density and multiply/divide by it.
- now the second line is an expectation: it's a ratio multiplied by its probability summed up, so we write as it is.
  So we discovered that maximizing the incomplete log likelihood is equal to maximizing the logarithm of a conditional expectation of a ratio. They are equivalent from a math pov.

To continue through the fire and the flames we apply the *Jensen inequality*. It says that if g is a concave function (its second derivative is negative) then we

have that $g(\mathbb{E}[x]) \geq \mathbb{E}[g(x)]$. Given that it holds for whatever x, if we want to maximize $g(\mathbb{E}[x])$, we can maximize $\mathbb{E}[g(x)]$ instead as well; if we maximize the second one, automatically we maximize the first one. In our case we want to maximize log of expectation so our $g(x) = \log x$ and therefore $g''(x) = -\frac{1}{x^2} < 0$ (so the logarithm is a concave function).
Then we can and do apply the inequality:

$$\log \sum_z f(z|y;\Psi') \frac{f(y,z;\Psi)}{f(z|y;\Psi')} \geq \sum_z f(z|y;\Psi') \log \frac{f(y,z;\Psi)}{f(z|y;\Psi')}$$

So we decide to maximize the second one (right) instead of the first one; the advantage is that now the second one can be semplified a lot. How?
First some notation: imagine we rewrite the two quantities above in this manner

$$\log L(\Psi) \geq h(\Psi|\Psi')$$

with the first the logarithm of the incomplete likelihood, so our loglikelihood; we call the second expression above as $h(\Psi|\Psi')$ (a function of the parameter we want estimate and/given the parameters we fixed), again it's a minorant function of the incomplete log-likelihood we want to maximize.
Therefore to maximize $\log L(\Psi)$ we can maximize $h(\Psi|\Psi')$ with respect to Ψ the parameters of interest.
Now we have that h can be rewritten splitting by the log of the ratio

$$h(\Psi|\Psi') = \sum_z f(z|y;\Psi') \log f(y,z;\Psi) - \sum_z f(z|y;\Psi') \log f(z|y;\Psi')$$

Now we want to maximize for Ψ; we see that the second term does not depend on Ψ but on Ψ' only, which is known. So in the maximization it's a constant and can be ignored (when we compute the first derivative with respect to Ψ the second quantity goes to zero). The latter term is called *entropy* of the posterior density $f(z|y;\Psi')$.
Therefore, maximizing $h(\Psi|\Psi')$ with respect to Ψ is equivalent to maximizing $\sum_z f(z|y;\Psi') \log f(y,z;\Psi)$ only, which is the conditional expectation of the complete likelihood. In other terms we want to maximize:

$$\begin{aligned}
\arg\max_{\Psi} \; h(\Psi|\Psi') &= \arg\max_{\Psi} \; \sum_z f(z|y;\Psi') \log f(y,z;\Psi) \\
&\overset{(1)}{=} \arg\max_{\Psi} \; \sum_z f(z|y;\Psi') \log L_c(\Psi) \\
&= \arg\max_{\Psi} \; E_{z|y;\Psi'}[\log L_c(\Psi)]
\end{aligned}$$

where in (1) abbiamo sostituito indicando la verosimiglianza completa come L_c.
Therefore the estimation problem consists in computing the conditional expectation (*E-STEP*) of the complete likelihood and then maximize it (*M-STEP*), or more precisely from a computer science pov:

- we start choosing/fixing a $\Psi^{(0)} = \Psi^{(h)}$ with $h = 0$ (initialization)
- Repeat until convergence the following steps:

- E-STEP: compute

$$Q(\Psi; \Psi^{(h)}) = E_{z|y;\Psi^{(h)}}[\log L_c(\Psi)]$$

- M-STEP: find $\Psi^{(h+1)}$ that maximizes the $Q(\Psi; \Psi^{(h)})$ above, such that

$$Q(\Psi^{(h+1)}; \Psi^{(h)}) \geq Q(\Psi; \Psi^{(h)}), \quad \forall \Psi \in \Omega$$

- $h = h + 1$

We will show that if we start from whatever values of $\Psi^{(0)}$ the likelihood will always increase (it's monotonic) so we at certain point will have a maximum (may be a local maximum).

12.2.4 Application to Gaussian Mixture models

Suppose that the model we want to estimate is a mixture of k Gaussian (univariate) components, each with parameters μ_i e σ_i^2:

$$f(y) = \sum_{i=1}^{k} \pi_i \mathcal{N}(y; \mu_i, \sigma_i^2)$$

we want to estimate the weights π_i and the parameters μ_i, σ_i^2.
Then, given a sample of n observations, we have:

- as we have seen the density of z is a multinomial, that is $f(\mathbf{z}_j; \boldsymbol{\pi}) = \prod_{i=1}^{k} \pi_i^{z_{ij}}$, and so we have this semplification for the probability $f(z_{ij} = 1; \boldsymbol{\pi}) = \pi_i$ (it's the weight)
- the density of the data is $f(y_j|\mathbf{z}_j; \boldsymbol{\theta}) = \prod_{i=1}^{k} \mathcal{N}(y_j; \theta_i)^{z_{ij}}$ and $f(y_j|z_{ij} = 1; \boldsymbol{\theta}) = \mathcal{N}(y_j; \theta_i)$, with $\theta_i = (\mu_i, \sigma_i^2)$.

Conditional expectation What is the conditional expectation of the complete loglikelihood in a mixture model? it becomes:

$$\begin{aligned}
E_{z|y;\Psi'}[\log L_c(\Psi)] &= E_{z|y;\Psi'}\left[\sum_{j=1}^{n} \log f(y_j, \mathbf{z}_j; \Psi)\right] \\
&\overset{(1)}{=} \sum_{j=1}^{n}\sum_{i=1}^{k} f(z_{ij}|y_j; \Psi') \log f(y_j, z_{ij}; \Psi) \\
&\overset{(2)}{=} \sum_{j=1}^{n}\sum_{i=1}^{k} f(z_{ij}|y_j; \Psi') \log \left[f(z_{ij}; \Psi) \cdot f(y_j|z_{ij}; \Psi)\right] \\
&= \sum_{j=1}^{n}\sum_{i=1}^{k} f(z_{ij}|y_j; \Psi') \left[\log f(z_{ij}; \Psi) + \log f(y_j|z_{ij}; \Psi)\right] \\
&= \sum_{j=1}^{n}\sum_{i=1}^{k} f(z_{ij}|y_j; \Psi') \log f(z_{ij}; \Psi) + \sum_{j=1}^{n}\sum_{i=1}^{k} f(z_{ij}|y_j; \Psi') \log f(y_j|z_{ij}; \Psi) \\
&\overset{(3)}{=} \mathcal{F}_L(\boldsymbol{\pi}) + \mathcal{F}_O(\boldsymbol{\theta})
\end{aligned}$$

where in

- (1) we start by computing the expected value multiplying the density for the value of interest, for each possible z, as showed before
- (2) we have decomposed the logarithm of the joint density of y and z in the product of the density of z, times the density of y given z;
- (3) we observe that the first term, $\mathcal{F}_O(\boldsymbol{\theta})$, depends only by the parameters μ_i and σ_i^2 of the gaussian ($f(y_j|z_{ij};\Psi)$ is the gaussian); we name it with $\mathcal{F}_O$ to mean *observable* because it includes the observed data.
  The term $\mathcal{F}_L(\boldsymbol{\pi})$, by the weights π_i ($f(y_j|z_{ij};\Psi)$ is the multinomial which depends on π_i). In the naming $\mathcal{F}_L$, L means latent which depends on the unknown parameters

So we have splitted the parameters in the two components and the conditional expectation is the sum of these two quantities.

E-step To complete the function above, in the E-step we need to compute the density $f(z_{ij}|y_j;\Psi')$ (present in both terms).
In the mixture models these are posterior probabilities of the group membership defined in equation 12.5, and evaluated at a current set of parameters, so:

$$f(z_{ij}|y_j;\Psi') = \frac{\pi_i \mathcal{N}(y_j;\mu_i,\sigma_i^2)}{\sum_{i=1}^k \pi_i \mathcal{N}(y_j;\mu_i,\sigma_i^2)}$$

with μ_i and σ_i^2 parameters estimated at the previous EM-step of the algorithm.

M-step In the M-step, we need to compute the score of $\mathcal{F}_O(\boldsymbol{\theta})$ and $\mathcal{F}_L(\boldsymbol{\pi})$ with respect to the parameters.
For the maximization of $\mathcal{F}_L$, we have that $\frac{\partial \mathcal{F}_L(\boldsymbol{\pi})}{\partial \pi_i}$ is equivalent to 12.9. Therefore we get:

$$\hat{\pi}_i = \frac{\sum_{j=1}^n f(z_{ij}|y_j;\Psi')}{n}$$

The maximization of $\mathcal{F}_L$ with respect to μ_i is obtained by:

$$\frac{\partial \mathcal{F}_O(\boldsymbol{\theta}_i)}{\partial \mu_i}$$

In order to solve the problem, we rewrite $\mathcal{F}_O(\boldsymbol{\theta}_i)$ in explicit form:

$$\begin{aligned}\mathcal{F}_O(\boldsymbol{\theta}_i) &= \sum_{j=1}^n \sum_{i=1}^k f(z_{ij}|y_j;\Psi') \log \mathcal{N}(y_j;\mu_i,\sigma_i^2) \\ &= \sum_{j=1}^n \sum_{i=1}^k f(z_{ij}|y_j;\Psi') \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_i^2 - \frac{1}{2}\frac{(y_j-\mu_i)^2}{\sigma_i^2}\right)\end{aligned}$$

Therefore:

$$\frac{\partial \mathcal{F}_O(\boldsymbol{\theta}_i)}{\partial \mu_i} = \sum_{j=1}^n f(z_{ij}|y_j;\Psi') \left(\frac{(y_j-\mu_i)}{\sigma_i^2}\right) = 0$$

from which:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} f(z_{ij}|y_j;\Psi')y_j}{\sum_{j=1}^{n} f(z_{ij}|y_j;\Psi')}$$

the estimate for μ_i is the mean of the observations weighted by the posterior probabilities.
Finally, for the maximization with respect to σ_i^2 the derivative is (try)

$$\frac{\partial \mathcal{F}_O(\boldsymbol{\theta}_i)}{\partial \sigma_i^2} = \sum_{j=1}^{n} f(z_{ij}|y_j;\Psi')\left(-\frac{1}{2}\frac{1}{\sigma_i^2} + \frac{1}{2}\frac{(y_j-\mu_i)^2}{\sigma_i^4}\right) = 0$$

from which

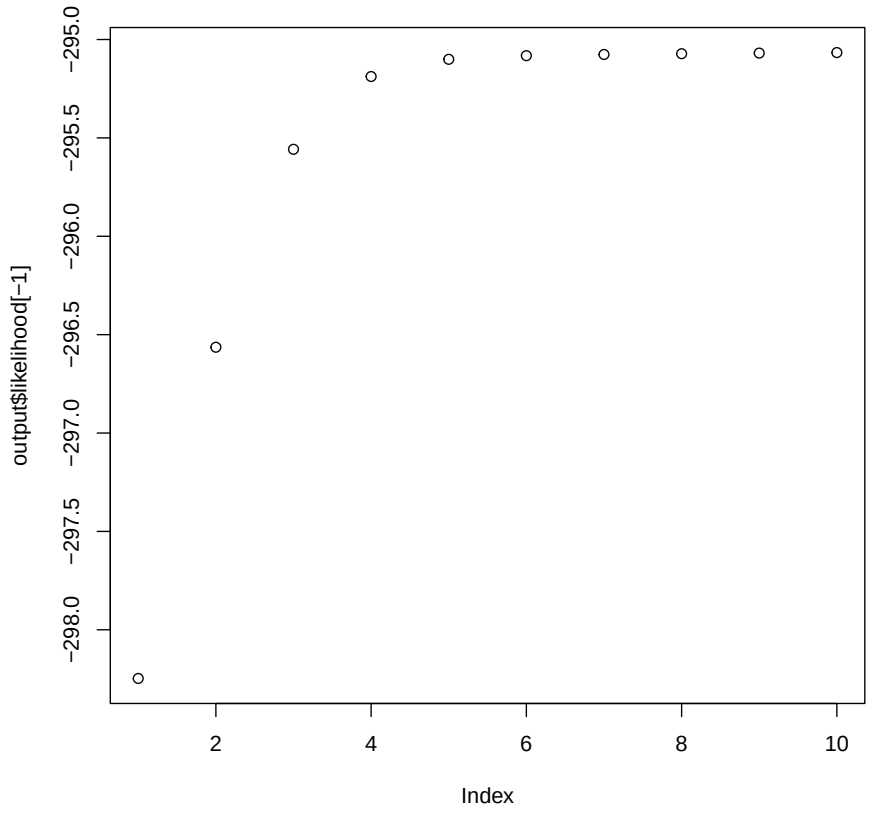
$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n} f(z_{ij}|y_j;\Psi')(y_j-\mu_i)^2}{\sum_{j=1}^{n} f(z_{ij}|y_j;\Psi')}$$

12.2.5 Example in R

```
nm.em <- function(y,          # data
                  k,          # number of components/groups
                  it = 50,    # number of iterations of em algorithm
                  eps = 0.00001) # tolerance level to stop
{
    # in this algorithm we stop if we reach the maximum number of
    # iterations or we reached convergence (checking tolerance level)

    # we have univariate data
    numobs = length(y)

    ## initialization
    ## --------------
    ## the parameter to be estimated in our model these are
    ## k sigmas and k mu.
    ## Here we initialize them with equal values for sigma (using the sample)
    sigma = rep(var(y), k)
    ## mus are set to a random starting point in the range of data
    mu = runif(k, min(y), max(y))

    ## initialization of w=pi (we dont' use pi in code because it's a
    ## R constant)
    w = runif(k)
    w = w/sum(w) # normalize so the pi sums is 1

    ## ################################################

    ## here we initialize some objects with very low number
    ## these are n x k matrix (a density for each observation and for each group)
    f.y.z <- matrix(eps, nrow = numobs, ncol = k) ## this will contain density of y|z (normal)
```

```
f.z.y <- matrix(eps, nrow = numobs, ncol = k) ## this will contain density of z|y

likelihood <- NULL # this will contain the likelihood
ratio = 1000       # to decide when to stop for convergence
lik = -10^10       # it's the previous iteration likelihood: initialize by a very
h = 0

## do this until condition is not respected we check number of
## iterations (first) and convergence (second)
while ((h < it) & (ratio > eps)) {

    h <- h + 1

    ## E-STEP: where we compute the posterior probability

    ## first density of several k gaussian for the data, given the current params
    for (i in 1:k) f.y.z[, i] <- dnorm(y, mu[i], sqrt(sigma[i]))
    ## then posterior probability (formula in the e-step)
    ## at the denominator
    for (i in 1:k) f.z.y[, i] <- w[i] * f.y.z[,i] / (w %*% t(f.y.z))

    f.z.y <- ifelse(is.na(f.z.y),mean(f.z.y, na.rm=T),f.z.y)   ## check per gli NA

    ## M-STEP: where we compute the new values of the parameter
    for (i in 1:k) {
        mu[i] <- (f.z.y[,i]%*%y) / sum(f.z.y[,i])
        sigma[i] <- f.z.y[,i]%*%((y-mu[i])^2)/sum(f.z.y[,i])
        w[i] <- sum(f.z.y[,i])/numobs
    }

    ## in order to decide when to stop we measure the relative
    ## difference of the current likelihood, divided by the
    ## previous likelihood (because of monotonicity we have seen
    ## that it increases). We decide to stop when it doesn't
    ## increase anymore a lot

    temp = sum(log(w%*%t(f.y.z))) # current loglikelihood: sum of the log of comp
    likelihood = c(likelihood, temp) # add the current to the saved likelihoods
    ratio <- (temp-lik) / abs(lik) # we check the increase (lik is the previous l
    lik <- temp   # save the current lik for the next iteration as previous

    if (is.na(lik)) ratio<-eps/2
}

## here EM algorithm is finished. Before exiting we compute AIC
## and BIC (information criteria to choose the number of
## parameters in a model). For a mixture model they are these

h = k-1+k+k
```

```
    BIC = -2*lik+h*log(numobs)
    AIC = -2*lik+2*h

    ## returning stuff
    invisible(list(
        likelihood = likelihood, # let's see if monotonicity is respected
        w = w,                   # pis
        sigma = sigma,           # other params
        mu = mu,
        f.z.y = f.z.y, # we put the posterior probability as well that
                       # can be used for classification
        BIC = BIC,
        AIC = AIC))
}

set.seed(1)
## To apply we simulate data from two groups: height. say 30 men and
## 50 woman
men = rnorm(30, 73.7, 9.8)
women = rnorm(50, 56.8, 6.8)
## but we don't know which groups they belong, these below is our
## final dataset
x = c(men, women)

## we want to estimate a mixture model with two components, estimate
## the parameter and also to classify people in the two groups
## according to weight
output <- nm.em(x, k = 2, it = 100)
plot(output$likelihood[-1]) # remove the first value very low
```




```
## we see the likelihood is increasing: ok it's working

## value of the parameters: we should have an estimate more or less
## corresponding to
30/80 #w1

## [1] 0.375

50/80 #w2

## [1] 0.625

output$w # the order can be different from before (here is not

## [1] 0.6210914 0.3789086

         # respected, no probs)
output$mu # le approssimazioni sono buone

## [1] 56.71351 75.81719

sqrt(output$sigma) # queste un po meno
```

```
## [1] 5.125657 6.719552

## le cose dovrebbero essere simili a quelle ottenute mediante la
## generazione di dati

## Here below posterior probability of the two groups for each person: these
## are used to classify
head(round(output$f.z.y, 3))

##       [,1]  [,2]
## [1,] 0.333 0.667
## [2,] 0.003 0.997
## [3,] 0.621 0.379
## [4,] 0.000 1.000
## [5,] 0.001 0.999
## [6,] 0.601 0.399

## above the first observation which are all mens. the first weight
x[1]

## [1] 67.56075

## is in the middle: with a probability 33% is a man with 66 is a
## woman.

## we can classify people looking at the greatest probability
est.cl = apply(output$f.z.y, 1, which.max) # estimate classification (1 men, 2 woman)
true.cl = rep(c(1,2), c(30,50))             # true classification
table(true.cl,est.cl)

##        est.cl
## true.cl  1  2
##       1  5 25
##       2 45  5

## qualcosa non va temo ... se si cambia il seme poi i risultati cambiano
## notevolmente (in termini di clustering)
set.seed(2)
men = rnorm(30, 73.7, 9.8)
women = rnorm(50, 56.8, 6.8)
x = c(men, women)
output <- nm.em(x, k = 2, it = 100)
output$w

## [1] 0.3686124 0.6313876

output$mu

## [1] 52.33889 70.16741

sqrt(output$sigma)
```

```
## [1]  4.773234 12.013574

est.cl = apply(output$f.z.y, 1, which.max) # estimate classification (1 men, 2 woman)
true.cl = rep(c(1,2), c(30,50))             # true classification
table(true.cl,est.cl)

##        est.cl
## true.cl  1  2
##       1  2 28
##       2 32 18
```

12.2.6 Monotonicity property

Osservazione 328. This is the reason why we can use EM algorithm: this theorem says wherever we start, it doesn't matter, the likelihood will increase (or will stay constant). This mean that we will end in a maximum despite it could be a local/relative one; for this reason people apply EM algorithm starting from different point.
If the algorithm converges to several set of final estimates we should check their likelihood to choose the set with the maximum one.

Proposizione 12.2.1. *The EM algorithm satisfies the condition of monotonicity of the log-likelihood, which states that the EM algorithm assures, that at each step, the log-likelihood is not decreasing:*

$$\log L\left(\Psi^{(h+1)}\right) \geq \log L\left(\Psi^{(h)}\right) \qquad \forall\ h$$

Proof. Observe that the joint density $f(y, z; \Psi)$ can be decomposed as $f(y, z; \Psi) = f(y|z; \Psi)f(z; \Psi)$ or as $f(y, z; \Psi) = f(z|y; \Psi)f(y; \Psi)$.
Consider the last form and rewrite the conditional expectation:

$$\begin{aligned} Q(\Psi; \Psi^{(h)}) &= E_{z|y;\Psi^{(h)}}[\log L_c(\Psi)] \\ &= \sum_z f(z|y; \Psi^{(h)}) \log f(z|y; \Psi) + \sum_z f(z|y; \Psi^{(h)}) \log f(y; \Psi) \\ &= \sum_z f(z|y; \Psi^{(h)}) \log f(z|y; \Psi) + \log f(y; \Psi) \sum_z f(z|y; \Psi^{(h)}) \\ &= H(\Psi; \Psi^{(h)}) + \log f(y; \Psi) = H(\Psi; \Psi^{(h)}) + \log L(\Psi) \end{aligned}$$

where the first term is denoted in compact form as $H(\Psi; \Psi^{(h)})$ because it represents a kind of entropy.
From the previous equation we have that:

$$\log L(\Psi) = Q(\Psi; \Psi^{(h)}) - H(\Psi; \Psi^{(h)}) \tag{12.10}$$

Now, the monotonicity property is proven if we can show that

$$\log L(\Psi^{(h+1)}) - \log L(\Psi^{(h)}) \geq 0$$

Starting from the last difference, rewritten using (12.10):

$$\begin{aligned}\log L(\Psi^{(h+1)}) - \log L(\Psi^{(h)}) = & \quad \left\{Q(\Psi^{(h+1)};\Psi^{(h)}) - Q(\Psi^{(h)};\Psi^{(h)})\right\} - \\ - & \quad \left\{H(\Psi^{(h+1)};\Psi^{(h)}) - H(\Psi^{(h)};\Psi^{(h)})\right\} \quad (12.11)\end{aligned}$$

In (12.11) the first difference is greater than or equal to zero, thanks to the M-step.
The monotonicity is guaranteed if we can prove that

$$H(\Psi^{(h+1)};\Psi^{(h)}) - H(\Psi^{(h)};\Psi^{(h)}) \leq 0$$

Now, for every Ψ we have

$$\begin{aligned}H(\Psi^{(h+1)};\Psi^{(h)}) - H(\Psi^{(h)};\Psi^{(h)}) &= E_{z|y;\Psi^{(h)}}\left[\log \frac{f(z|y;\Psi^{(h+1)})}{f(z|y;\Psi^{(h)})}\right] \\ &\leq \log\left(E_{z|y;\Psi^{(h)}}\left[\frac{f(z|y;\Psi^{(h+1)})}{f(z|y;\Psi^{(h)})}\right]\right) \\ &= \log\sum_z f(z|y;\Psi^{(h+1)}) = 0\end{aligned}$$

where we applied Jensen inequality. □

12.2.7 GEM Algorithm

G means generalized: it's a large family of generalization.
Sometimes, in the M-step is not possible to find a closed form for the parameter of the component densities. In this case could be necessary to incorporate a numerical optimization method (eg newton raphson). In such situations, we have not the condition:

$$Q(\Psi^{(h+1)};\Psi^{(h)}) \geq Q(\Psi;\Psi^{(h)}) \quad \forall\Psi \in \Omega \tag{12.12}$$

but the less restrictive one:

$$Q(\Psi^{(h+1)};\Psi^{(h)}) \geq Q(\Psi^{(h)};\Psi^{(h)}) \tag{12.13}$$

Equation (12.13) respects the monotonicity property, but at each iteration, we do not reach the optimal point but values that still allow to increment the log-likelihood.

Chapter 13

Hypothesis test

Osservazione 329. This is one of most important part of statistics for practical applications.

13.1 Intro

Osservazione 330. Starting from the parametric approach our aim is to find the parameter of the distribution. We have collected data and we have an idea on the possible parameter for a distribution.

Definizione 13.1.1 (Statistical hypothesis)**.** An hypothesis is a idea/conjecture about the unknown value of θ. Can be

- *simple* hypothesis, $H_0 : \theta = \theta_0$: simple if parameter (or a vector of parameters) has a single value (single defined vector)
- *composite* hypothesis, $H_0 : \theta \in \Theta_0$: composite if the parameter belongs to a set/interval of possible values

Esempio 13.1.1 (Esame vecchio viroli)**.** Which of the following hypotheses is a simple hypothesis

1. X follows a gaussian distribution with mean $\mu > 0$ and known variance
2. X follows a gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 > 10$
3. none of these
4. an unbiased coin against a biased coin such that heads in 3 times as likely as tails

Credo la quarta, altri dicono la seconda

Osservazione 331. From the historical point of view we had two approaches:

1. in the *significance theory of Fisher* we had one hypothesis, H_0, the null hypothesis, and we want to check against it

	retain the null	reject null
H_0 true	ok	type I error (α)
H_0 not true	type II error (β)	ok

Table 13.1: errori dei test

2. in the *Neymann-Pearson Theory* we have two contradicting hypotheses: H_0 the null hypothesis and H_1 the alternative hypothesis. The null is the one that we want to check, and if we reject the null, we accept the alternative hypothesis

How to decide regarding our hypothesis? We collect a sample of data $\mathbf{x}$ and then we can use a *test statistic* or statistical test on our data.
Test statistics is a transformation/function on the sample of the data, that maps the sample space Ω in two region: the R rejection region and $\overline{R}$ not rejection region

$$T(\mathbf{X}) : \Omega \to \left\{R, \overline{R}\right\}$$

Due to the sampling, any statistical test may lead to wrong conclusion as summarized in table 13.1.

Esempio 13.1.2 (Esame vecchio viroli)**.** decide which of these definitions is true

- the probability that the null hypothesis is correctly rejected is equal to the size of the test
- the probability that the null hypothesis is correctly rejected is equal to the p-value of the test
- the probability that the null hypothesis is falsely rejected is equal to the power of the test
- the probability that the null hypothesis is correctly rejecteed is equal to the power of the test

the last one of course

Esempio 13.1.3 (Esame vecchio viroli)**.** which of the following statements about the p-value is true

1. the p-value is always less than 0.05
2. is the probability that the alternative hyothesis is true
3. is the probability that the null hyothesis is true
4. is the probability of obtaining the observed results or results which are more estreme if the null hypothesis is true

l'ultima

Esempio 13.1.4 (Esame vecchio viroli)**.** Decide which of these definitions is true

1. the significance level of a statistical test is equal to the probability that the null hypothesis is accepted while the null hypothesis is true
2. the significance level of a statistical test is larger than the size of the test
3. the significance level of a statistical test is equal to the probability that the null hypothesis is true
4. the significance level of a statistical test is equal to the probability that the null hypothesis is rejected while it is true

L'ultima ovviamente

13.1.1 Significance theory by Fisher

In this approach the procedure is the following:

- we have a single hypothesis H_0;
- consider a test statistics T_n, which is a function of the sample, being a transformation of random variable is a random variable and has a distribution. It's distributed with a known pdf $f(T_n|H_0)$ under the null;
- we compute T_n on the data, that is $T(\mathbf{x}) = t_c$;
- define the *p-value* as the probability that under null hypothesis our computed statistics takes a more extreme value than that obtained here in other hypothetical infinite samples. We can be interested in:
  - $\mathbb{P}(T_n \geq t_c|H_0)$ *right-tail* event;
  - $\mathbb{P}(T_n \leq t_c|H_0)$ left-tail event;
  - $\mathbb{P}(|T_n| \geq |t_c||H_0)$ double-tail event.
- according to Fisher, p-value can be used to decide and is *evidential* against a theory: if it's very small, "either something very rare has happened (due to our extracted sample) or H_0 is false";
- fix the *type-I error* of the test, α, called also *significance level* or *size* of the test: it's the larger type I error we can admit/tolerate;
- compare the p-value with α; if $p < \alpha$ we reject the null (the event is considered too rare to be true).

Esempio 13.1.5 (Simple example)**.** Consider:

1. we assume that $X \sim \mathrm{N}\left(\mu, \sigma^2 = 5\right)$ with μ unknown. We are interested in checking wether $H_0 : \mu = 3$
2. we observe $\mathbf{x} = (x_1, \ldots, x_n)$, suppose $n = 10$ and in our sample $\bar{x} = 4$. Can we rule out that the population μ is 3 given our data?
3. we know that the statistical test obtained standardized according to the null hypothesis $\mu_0 = 3$ divided by its standard deviation is $T_n = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$ is distributed according to a standard normal

4. the test statistic in our sample is $t_c = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{4-3}{\sqrt{5}/\sqrt{10}} = 1.414$,
5. the p-value here is set as two tailed because we want to reject the $\mu = 3$ both if the real μ is above or below. So for a two tailed:

$$\begin{aligned}\mathbb{P}\left(|T_n| \geq |t_c|\,|H_0\right) &= \mathbb{P}\left(|\mathrm{N}\left(0,1\right)| \geq 1.414\right) = 2 \cdot \mathbb{P}\left(\mathrm{N}\left(0,1\right) \geq 1.414\right) \\ &= 2 \cdot (1 - \texttt{pnorm}(1.414)) = 0.1573\end{aligned}$$

   since it's not so rare under the null we are less moved to refuse that hypothesis
6. if we set $\alpha = 0.05$, $p > \alpha$: in other terms, under the null, the probability to observe what we have observed is 0.1573. It is not rare enough to believe the null hypothesis not true
7. we *do not reject* the null hypothesis: according to Fisher we reject or we do not reject, the word *acceptance is not used*.

Esempio 13.1.6 (Esame vecchio viroli)**.** A coin is thrown independently 10 times to test the hypothesis that the probability of heads is 3/4 versus the alternative that the probability is not 3/4. The test rejects if less of 5 heads are observed. Given that $X \sim \mathrm{Bin}\left(10, 0.75\right)$ under H_0 compute the significance level of the test.

```
## probabilita di rifiutare sotto nulla
pbinom(4, 10, 0.75)

## [1] 0.01972771
```

Risposta 0.0197

13.1.2 Neyman-Pearson Theory

Here we have with two different hypotheses, the null and the alternative, and we specify the value of the parameter for both of them. In general terms:

$$\begin{aligned}H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1\end{aligned}$$

We have a *bipartition* of the parameter space in Θ_0 and Θ_1; it's the null or the alternative, no other possibilities are possible. Here are some system of hypothesis we could construct:

$$\begin{aligned}&\textit{simple}: \begin{cases}H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1\end{cases} \quad \textit{one}-\textit{tailed}: \begin{cases}H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \quad (\text{or } \theta < \theta_0)\end{cases} \\ &\textit{two}-\textit{tailed}: \begin{cases}H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0\end{cases} \quad \textit{composite}: \begin{cases}H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1\end{cases}\end{aligned}$$

The procedure is as follows:

1. we consider a test statistics T_n: it has two different pdfs under the two different hypotheses $f(T_n|H_0)$ and $f(T_n|H_1)$.

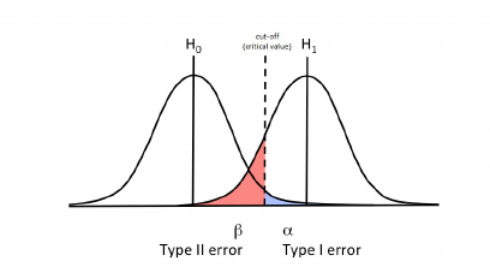


Figure 13.1: Alpha (blue) and beta (red).

2. the errors in decision making are formalized as $\alpha = \mathbb{P}(T_n \in R|H_0)$ and $\beta = \mathbb{P}\left(T_n \in \overline{R}|H_1\right)$. Imagine we have a cutoff/critical point before which we accept the null and after it we refuse in favour of the alternative, then α, β are visualized in 13.1 in blue and red respectively.

3. The critical point defines the values that call for rejecting the null hypothesis. By moving the line on the left one increase type I error and diminish type II, viceversa on the right. Therefore simultaneous minimization of α and β is not possible; there's a tradeoff between them.

Esempio 13.1.7 (Esame vecchio viroli)**.** Which of the following is a one tailed test

1. A die is unbiased
2. a six die is unbiased against it is biased in such a way that even numbers are three times as likely to be rolled as odd numbers
3. a government officail claims that the dropout rate forschool is 15%
4. X follows a gaussian distribution with known mean and variance $\sigma^2 > 10$

Corretta l'ultima

Esempio 13.1.8 (Esame vecchio viroli)**.** Which of the following is a two tailed test

1. a biased boin such that heads in 3 times, suggerita da taluni
2. ...

Osservazione 332. Important quantity in the Neyman-Pearson theory is the power. It can be confusing because in different book it's defined as power measure or function.

Definizione 13.1.2 (Power measure)**.** We define power as the complementary of the type II error:

$$1 - \beta = \mathbb{P}(T_n \in R|H_1)$$

By decreasing the second type error one increase the power of the test

Definizione 13.1.3 (Power function $\tau(\theta)$)**.** It's defined as the probability to reject the null as a function of θ, give values in the null or the alternative parameter space (here we don't have conditioning on an hypothesis):

$$\tau(\theta) = \mathbb{P}\left(T_n \in R; \theta\right)$$

Osservazione 333. Power function and power measure are the same when θ is the value specified by H_1. So the power function generalises the power measure

Osservazione 334. The power function has the same role of mean square error in point estimation, because it allows to *compare two statistical tests* (we will see later)

Osservazione importante 112 (α and the power function). α, the significance level, can be defined even through the power function $\tau(\theta)$: it's the maximum value the function have when considering a theta in the rejection zone

$$\alpha = \sup_{\theta \in \Theta_0} \tau(\theta)$$

If the null is a value, α is simply the power function evaluated at the value of the null.

Esempio 13.1.9. Imagine that a feature of interest is distributed $Y \sim \mathrm{N}\left(\theta, 1\right)$ and we have a system of hypotheses which is a *bipartition* of the parameter space

$$\begin{cases} H_0 : \theta \leq 0 \\ H_1 : \theta > 0 \end{cases}$$

We observe a single value y ($n = 1$) and we define a test/decision rule T_n as follows

- if $T_n = y \leq 0.5 \implies \overline{R}$, we don't reject the null since we find our result in line with the null $\theta \leq 0$
- if $y > 0.5 \implies R$: here we find our result enough to be against the null

Now the power function of this test is:

$$\begin{aligned} \tau(\theta) &= \mathbb{P}\left(T_n \in R; \theta\right) = \mathbb{P}\left(Y > 0.5; \theta\right) \\ &\overset{(1)}{=} \mathbb{P}\left(\underbrace{Y - \theta}_{Z} > 0.5 - \theta; \theta\right) \\ &\overset{(2)}{=} \mathbb{P}\left(Z > 0.5 - \theta; \theta\right), \quad Z \sim \mathrm{N}\left(0, 1\right) \\ &= 1 - \mathbb{P}\left(z \leq 0.5 - \theta; \theta\right) \\ &= 1 - \Phi(0.5 - \theta) \end{aligned}$$

where in

- (1) we subtracted θ from both the member of the inequality to center the random variable

- (2) we did this subtraction because we know that $Y \sim \mathrm{N}(\theta, 1)$ so $Z = Y - \theta \sim \mathrm{N}(0, 1)$, so we standardized the test
- Φ is the distribution function of the standard gaussian $\mathrm{N}(0, 1)$

So the power function derived is plotted in fig 13.2. It's an increasing function that starts from 0 end end up to 1

- for $\theta = 0$ (value given of the null), the power $1 - \Phi(0.5 - 0) = 0.308 = \alpha$. It's an high first type error if we choose this cutoff point of 0.5; maybe 0.5 is too low and we reject the null too much;
- if we want α to be lower than 0.308, for example we fix $\alpha = 0.05$, what is its cutoff point? We must have

  $$\alpha = 0.05 = 1 - \Phi(c)$$

  with c unknown to be determined.

  $$1 - \Phi(c) = 0.05 \implies \Phi(c) = 0.95 \implies c = \Phi^{-1}(0.95) \iff c = 1.645$$

```r
## power function
thetas <- seq(-2, 2, length.out = 100)
powf <- 1 - pnorm(0.5 - thetas)
plot(x = thetas, y = powf, ylim = c(0,1))
abline(h = c(0,1), v = 0, lty = 'dotted', col = 'red')
```

```r
## cutoff point for alpha = 0.05
qnorm(0.95)

## [1] 1.644854
```

Esempio 13.1.10. We have a variable of interest $X \sim \mathrm{N}\left(\theta, \sigma^2 = 25\right)$. Considering the hypothesis

$$\begin{cases} H_0 : \theta \leq 17 \\ H_1 : \theta > 17 \end{cases}$$

We take the sample mean as test, $T_n = \overline{x}$, and as decision rule we use:

$$\begin{cases} T_n \leq 17 + \frac{5}{\sqrt{n}} \implies \overline{R} \\ T_n > 17 + \frac{5}{\sqrt{n}} \implies R \end{cases}$$

We have two questions:

1. derive the power function
2. compute the size α of the test

Respectively:

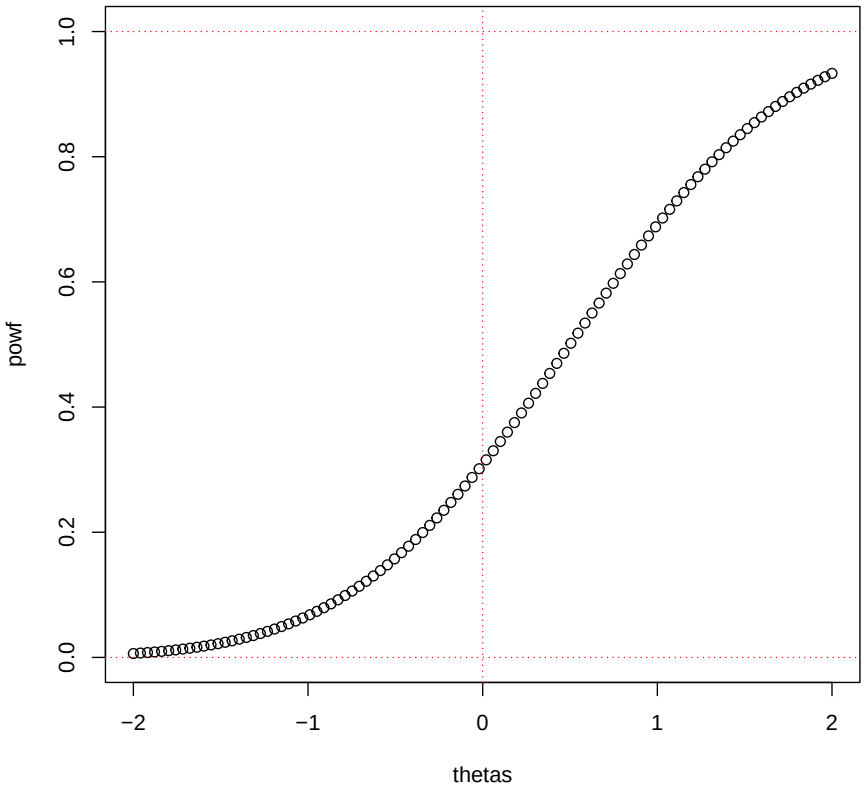


Figure 13.2: Power function

1. The power function is

$$\tau(\theta) = \mathbb{P}\left(T_n \in R; \theta\right) = \mathbb{P}\left(\overline{x} > 17 + \frac{5}{\sqrt{n}}; \theta\right)$$

since $X \sim \mathrm{N}\left(\theta, 25\right)$ then we have also the distribution of our test $\overline{X} \sim \mathrm{N}\left(\theta, \frac{25}{n}\right)$. Now back to power function we standardize our estimator by subtracting the mean θ and dividing by the standard deviation $5/\sqrt{n}$, and standardize the reject value as well in the same way

$$\begin{aligned}\tau(\theta) &= \mathbb{P}\left(\frac{\overline{x} - \theta}{5/\sqrt{n}} > \frac{17 + \frac{5}{\sqrt{n}} - \theta}{5/\sqrt{n}}; \theta\right) \\ &= \mathbb{P}\left(z > \frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}}; \theta\right) \\ &= 1 - \Phi\left(\frac{17 + 5/\sqrt{n} - \theta}{5/\sqrt{n}}\right)\end{aligned}$$

This is the power function; observe that $\frac{17+5\sqrt{n}-\theta}{5/\sqrt{n}}$ is decreasing in θ. On the other hand $\tau(\theta)$ is increasing in θ (as it should be).

2. for the size we have:

$$\begin{aligned}\alpha = \sup_{\theta \in \Theta_0}\left(\tau(\theta)\right) &\overset{(1)}{=} 1 - \Phi\left(\frac{17 + 5/\sqrt{n} - 17}{5/\sqrt{n}}\right) \\ &= 1 - \Phi(1) = \texttt{1 - pnorm(1)} \\ &= 0.159\end{aligned}$$

where in (1) the value of theta under null hypothesis where we get the maximum power is 17 because the H_0 says that $\theta \leq 17$ and the power is an increasing function of theta so the maximum of the null values is obtained in 17.

Esempio 13.1.11. A coin is thrown independently 10 times to test if the coin is fair:

$$\begin{aligned}H_0 &: p = 1/2 \\ H_1 &: p \neq 1/2\end{aligned}$$

As rule the null is rejected if we obtain either 0 or 10 heads:

1. what is α?

2. if $p = 0.1$ (probability of head) what is the power measure of the test?

We have

1. under H_0 we have that $X = \{\text{number of heads}\} \sim \text{Bin}\,(n = 10, p = 1/2)$

$$\begin{aligned}
\alpha &= \mathbb{P}\,(\text{reject } H_0 | H_0) \\
&= \mathbb{P}\,(X = 0 \vee X = 10 | p = 0.5) \\
&= \mathbb{P}\,(X = 0 | p = 0.5) + \mathbb{P}\,(X = 10 | p = 0.5) \\
&= \binom{10}{0} \cdot 0.5^0 \cdot (1 - 0.5)^{10-0} + \binom{10}{10} \cdot 0.5^0 \cdot (1 - 0.5)^{10-0} \\
&= 1 \cdot 1 \cdot 0.5^{10} + 1 \cdot 1 \cdot 0.5^{10} \\
&= 2 \cdot 0.5^{10} = 0.00195
\end{aligned}$$

   since it's so low we say the test is *very conservative*.

2. we have that the power

$$\begin{aligned}
1 - \beta &= \mathbb{P}\,(\text{reject } H_0 | H_1) \\
&= \mathbb{P}\,(\text{reject } H_0 | p = 0.1) \\
&= \mathbb{P}\,(X = 0 | p = 0.1) + \mathbb{P}\,(X = 10 | p = 0.1) \\
&= \binom{10}{0} \cdot 0.1^0 \cdot 0.9^{10} + \binom{10}{10} \cdot 0.1^{10} \cdot 0.9^0 \\
&= 0.1^{10} + 0.9^{10} \\
&= 0.349
\end{aligned}$$

   power is very low (we have seen alpha is low and there is tradeoff between alpha and beta). So this is not a good test in terms of power.

Esempio 13.1.12. Suppose that $X \sim \text{Bin}\,(100, p)$. We consider the test

$$\begin{aligned}
H_0 &: p = 1/2 \\
H_1 &: p \neq 1/2
\end{aligned}$$

we reject H_0 if $|X - 50| > 10$ (eg $X = 38, 37, \ldots$ or $X = 61, 62, \ldots$. Since $n = 100$ use the gaussian approximation of the binomial (due to CLT) in order to derive α.
We remember that the approximation of X is $X \simeq \text{N}\,(100 \cdot p, 100 \cdot p \cdot (1 - p))$. This is dued to the fact that the binomial distribution is a sum of bernoulli distribution, and since the binomial is a sum we can apply the central limit theorem to the Bernoulli distribution. Since it has mean p and variance $p(1 - p)$ the sum will have mean $100 \cdot p$ and variance $100 \cdot p \cdot (1 - p)$ because the bernoulli distribution are independent (sum of the means, sum of the variances, no covariances because of independence).

then

$$\begin{aligned}
\alpha &= \mathbb{P}\left(\text{reject } H_0 | H_0\right) \\
&= \mathbb{P}\left(|x-50| > 10 | p = 1/2\right) \\
&= \mathbb{P}\left(X - 50 > 10 \vee X - 50 < -10 | p = 1/2\right) \\
&= \mathbb{P}\left(X > 60 | H_0\right) + \mathbb{P}\left(X < 40 | p = 1/2\right) \\
&\overset{(1)}{=} \mathbb{P}\left(Z > \frac{60 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5 \cdot 0.5}}\right) + \mathbb{P}\left(Z < \frac{40 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5 \cdot 0.5}}\right) \\
&= \mathbb{P}\left(Z > \frac{10}{5}\right) + \mathbb{P}\left(z < -\frac{10}{5}\right) = \mathbb{P}\left(Z > 2\right) + \mathbb{P}\left(Z < -2\right) \\
&= 1 - \Phi(2) + \Phi(-2) \\
&= 1 - \texttt{pnorm}(2) + \texttt{pnorm}(-2) \\
&= 0.0455
\end{aligned}$$

were in (1) we standardized both terms by subtracting the mean under H_0 that is $100 \cdot 0.5$ and by dividing by sd that is $\sqrt{100 \cdot p \cdot (1-p)}$ again under H_0 $\sqrt{100 \cdot p \cdot (1-p)}$

Esempio 13.1.13 (Esame vecchio viroli)**.** Let $x_1, \ldots, X_n$ be a random sample from the density

$$f(x) = \theta^2 x e^{-\theta x}$$

with $X > 0$, $\theta > 0$, $\mathbb{E}\left[X\right] = 2/\theta$ and $\text{Var}\left[X\right] = 2/\theta^2$. Consider the system of hypothesis

$$\begin{cases} H_0 : \theta = 1 \\ H_1 : \theta = 2 \end{cases}$$

and a test with a rejection region $RR\left\{\overline{x} < k_\alpha\right\}$ where $\overline{x}$ is asymptotically $\text{N}\left(\frac{2}{\theta}, \frac{2}{n\theta^2}\right)$. For $n = 20$ evaluate the power function at $\alpha = 0.01$

1. 0.99 (a me ad ora viene questa)
2. 0.43
3. 0.23
4. 0.95 (taluni suggeriscono questa)

Questo sotto è un mio tentativo.

$$1 - \beta = \mathbb{P}\left(\text{reject } H_0 | H_1\right) = \mathbb{P}\left(\overline{x} < k_\alpha | \theta = 2\right)$$

Dobbiamo determinare k_α, sappiamo che $\overline{x}$ è distribuita normalmente $\text{N}\left(\frac{2}{\theta}, \frac{2}{n\theta^2}\right)$ quindi determiniamo il quantile corrispondente a 0.01 (determinato con `qnorm(0.01) = -2.32`).
Il valore soglia (valutato sotto ipotesi della nulla, ergo $\theta = 1$) e ricordando che

$\sigma = \sqrt{\frac{2}{n\theta^2}} = \frac{\sqrt{2}}{\theta\sqrt{n}}$:

$$\begin{aligned} k_\alpha &= \frac{2}{\theta} + q_\alpha \cdot \frac{\sigma}{\sqrt{n}} = \frac{2}{\theta} + q_\alpha \cdot \frac{\sqrt{2}}{\theta n} \\ &= \frac{2}{1} - 2.326 \cdot \frac{\sqrt{2}}{1 \cdot n} \\ &= \frac{2}{1} - 2.326 \cdot \frac{\sqrt{2}}{1 \cdot 20} \end{aligned}$$

```
## qui sotto  ocio a mettere +qnorm, perché il valore restituito è gia negativo
(k_a = 2 + qnorm(0.01)*sqrt(2)/20)

## [1] 1.835502
```

Quindi il valore soglia è 1.83. Riprendendo l'equazione della potenza, standardizziamo la media per utilizzare la normale standardizzata

$$\begin{aligned} 1 - \beta = \mathbb{P}\left(\overline{x} < k_\alpha | \theta = 2\right) &= \mathbb{P}\left(\underbrace{\frac{\overline{x} - \frac{2}{\theta}}{\sqrt{2}/(\theta n)}}_{Z} < \frac{k_\alpha - \frac{2}{\theta}}{\sqrt{2}/(\theta n)} | \theta = 2 \right) \\ &\overset{(1)}{=} \mathbb{P}\left(Z < \frac{1.83 - 1}{\sqrt{2}/(2 \cdot 20)} \right) \end{aligned}$$

dove in (1) abbiamo sostituito il valore soglia k_α precedentemente determinato, e sostituito/sfruttato l'ipotesi alternativa Pertanto la potenza è

```
(pow = pnorm((k_a - 1)/(sqrt(2)/40)))

## [1] 1
```

boh, qui taluni dicono 0.95.

13.2 UMP tests (Neyman-Pearson)

Osservazione importante 113. Ideally we want that the power function $\tau(\theta)$ is the largest as possible when $\theta \in \Theta_1$ (theta is in the set of value of alternative hypothesis) while $\tau(\theta)$ is smallest when $\theta \in \Theta_0$ (set of value of the null). This would be a good test.
Problem is that the possible test are infinite (if we change cutoff point/math formulation and so on) and we need something to make order.

13.2.1 Definition and existence

Definizione 13.2.1 (UMP (uniformely most powerful) test)**.** It's a test T^* with the largest power (in Θ_1) among all possible tests having the same size α. So it should have two properties:

1. $\sup_{\theta \in \Theta_0} \tau(\theta) = \alpha$;
2. given the UMP T^* and any other test T with size α, we have that $\tau_{T^*}(\theta) \geq \tau_T(\theta)$, $\forall \theta \in \Theta_1$,

Osservazione importante 114. However does $T*$ (the UMP test) exists? Not always! we need further conditions. Without them we could have one test with higher power on a subset region of Θ_1 while another outperforms in other regions of Θ_1.

Teorema 13.2.1 (Neyman theorem (1935))**.** *For a system of* simple hypotheses *(not composite, eg single value for theta under the null and single value under the alternative) is possible to derive a uniform most powerful test, it exists!*
Furthermore the theorem can be extended to a system with one-tail alternative hypothesis*, provided that we have/add a monotonicity condition (we'll see it later).*
But for a two-sided test *or* composite hypotheses *it may not necessarily exist (unless specific cases).*

Osservazione 335. After that paper we started study how to construct such best tests

13.2.2 How to construct the UMP test

We have a sample from IID rvs $\{X_1, \ldots, X_n\}$. We assume a probabilistic model for X, $f_X(x)$: in this condition we can use the likelihood $\mathrm{L}(\theta)$ to construct a test. Suppose the two hypotheses are

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

we could base our decision on a *simple likelihood ratio*

$$\lambda(\mathbf{x}) = \frac{\mathrm{L}(\theta_0; \mathbf{x})}{\mathrm{L}(\theta_1; \mathbf{x})}$$

so if $\lambda(\mathbf{x}) > 1$ then θ_0 is more likely, while if $\lambda(\mathbf{x}) < 1$ then θ_1 is preferable. Specifically we have to choose a cutoff point k ($k > 0$, is positive because a ratio of likelihood which are positive) to choose between the two hypotheses:

- we reject H_0 if $\lambda(\mathbf{x}) < k$
- we accept H_0 if $\lambda(\mathbf{x}) \geq k$

Notice: for different k we have a different test! We should find the best one, that is the cutoff point that give us the largest power measure.

Teorema 13.2.2 (Neyman-pearson lemma)**.** *Given* $\mathbf{x}$ *(a iid sample, the empirical information) and a system of hypotheses*

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

we can without loss of generality assume that $\theta_0 < \theta_1$ (otherwise we exchange the hypotheses system). Define

$$\lambda(\mathbf{x}) = \frac{\mathrm{L}(\theta_0; \mathbf{x})}{\mathrm{L}(\theta_1; \mathbf{x})}$$

and the reject region $R = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k_\alpha\}$ where the cutoff point k_α depends on α, the size o the test. We choose k_α such that:

$$\mathbb{P}(\lambda(\mathbf{x}) \leq k_\alpha | H_0) = \alpha$$

This test has the hightest $\tau(\theta)$ among all the infinite test with the same α, that is it's UMP.

Osservazione 336. So we reverse the problem: we fix α and given this kind of test (the ratio) we choose k such as α is equal to the value we desider.

Esempio 13.2.1. Imagine that $X \sim \mathrm{Exp}(\theta)$, so $f(x) = e^{-\theta x}$. We assume a system of simple hypotheses:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

with $\theta_0 < \theta_1$. Now the likelihoods under the different hypotheses are

$$\mathrm{L}(\theta_0; \mathbf{x}) = \prod_{i=1}^{n} \theta_0 e^{-\theta_0 x_i} = \theta_0^n e^{-\theta_0 \sum_{i=1}^n x_i}$$
$$\mathrm{L}(\theta_1; \mathbf{x}) = \ldots = \theta_1^n e^{-\theta_1 \sum_{i=1}^n x_i}$$

We have that the ratio test

$$\lambda(\mathbf{x}) = \frac{\mathrm{L}(\theta_0; \mathbf{x})}{\mathrm{L}(\theta_1; \mathbf{x})} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left(-(\theta_0 - \theta_1) \sum_{i=1}^{n} x_i\right)$$

Since $-(\theta_0 - \theta_1) > 0$ by hypotheses ($\theta_0 < \theta_1$), large values of $\sum_{i=1}^n x_i$ will favour H_0 because $\lambda(\mathbf{x})$ will be larger and we reject when $\lambda(\mathbf{x}) \leq k_\alpha$.
Therefore ported considering only $\sum_{i=1}^n x_i$ (the random part) we have that

- if $\sum_{i=1}^n x_i < c_\alpha \implies R$ (c_α is a sort of cutoff only for $\sum_{i=1}^n x_i$)
- if $\sum_{i=1}^n x_i \geq c_\alpha \implies \overline{R}$

Now we observe that if X_i are n iid Exponential then their sum $Y = \sum_{i=1}^n X_i \sim \mathrm{Gamma}(n, \theta)$;

$$\begin{aligned} \alpha = \mathbb{P}(\text{reject } H_0 | H_0) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i < c_\alpha | H_0\right) \\ &= \mathbb{P}(Y < c_\alpha | H_0) = \int_0^{c_\alpha} \mathrm{Gamma}(n, \theta_0) \\ &= \int_0^{c_\alpha} \frac{1}{\Gamma(n)} \cdot \theta_0^n \cdot y^{n-1} \cdot e^{y\theta_0} \, \mathrm{d}Y \end{aligned}$$

and α is the integral in the first part of this gamma distribution up to a c_α.
This is finished: this test were

1. $\sum_{i=1}^{n} x_i < c_\alpha \implies R$
2. $\sum_{i=1}^{n} x_i \geq c_\alpha \implies \overline{R}$

with c_α determined as above to have α =whatever is our UMP test.

Esempio 13.2.2. We have a sample of $n = 25$ units from $X \sim \mathrm{N}\left(\mu, \sigma^2 = 100\right)$ and the following system of hypotheses:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu = 1.5 \end{cases}$$

Find the rejection area of the UMP test and the power with sizes $\alpha = 0.1$ and $\alpha = 0.01$ (remembering that if we reduce alpha, we reduce power as well).
Starting from the ratio test:

$$\begin{aligned} \lambda(\mathbf{x}) = \frac{\mathrm{L}\left(\mu_0\right)}{\mathrm{L}\left(\mu_1\right)} = \frac{\mathrm{L}\left(0\right)}{\mathrm{L}\left(1.5\right)} &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi 100}} e^{-\frac{1}{2}\frac{(x_i-0)^2}{100}}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi 100}} e^{-\frac{1}{2}\frac{(x_i-1.5)^2}{100}}} \\ &\overset{(1)}{=} \frac{e^{-\frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2}{100}}}{e^{-\frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2 + n1.5^2 - 2\cdot 1.5\sum_{i=1}^{n} x_i}{100}}} \\ &= \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2}{100} + \frac{1}{2}\frac{\sum_{i=1}^{n} x_i^2}{100} + \frac{1}{2}\cdot\frac{n\cdot 1.5^2}{100} - \frac{1.5\sum_{i=1}^{n} x_i}{100}\right) \\ &= \exp\left(\frac{1}{2}\cdot\frac{25\cdot 1.5^2}{100} - \frac{1.5\sum_{i=1}^{n} x_i}{100}\right) \\ &= \exp\left(0.28 - 0.015\sum_{i=1}^{n} x_i\right) \end{aligned}$$

where in 1 the first term under products simplify and we expand the denominator.
So for the rejection area, remembering that $\sum_{i=1}^{n} x_i = 25 \cdot \overline{x}$

$$\begin{aligned} \exp\left(0.28 - 0.015 \cdot 25 \cdot \overline{x}\right) \leq k_\alpha &\implies R \\ \exp\left(0.28 - 0.015 \cdot 25 \cdot \overline{x}\right) > k_\alpha &\implies \overline{R} \end{aligned}$$

We note that if $\overline{x}$ increases the ratio decrease and we are less in favour of the null, and more in favour of the alternative. Alternatively of above we could compare $\overline{x}$ with a proper c_α (this because we have a distribution for $\overline{x}$!): if

$$\begin{aligned} \overline{x} \geq c_\alpha &\implies R \\ \overline{x} < c_\alpha &\implies \overline{R} \end{aligned}$$

Now we know that

$$\begin{aligned} \overline{x}|H_0 &\sim \mathrm{N}\left(0, 100/25 = 4\right) \\ \overline{x}|H_1 &\sim \mathrm{N}\left(1.5, 100/25 = 4\right) \end{aligned}$$

so we can find the cutoff point for $\overline{x}$ and the latter above will be our test.

$$\begin{aligned}\alpha &= \mathbb{P}\left(\text{reject } H_0 | H_0\right) = \mathbb{P}\left(\overline{x} \geq c_\alpha | \mu = 0\right) \\ &\overset{(1)}{=} \mathbb{P}\left(\frac{\overline{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \geq \frac{c_\alpha - \mu}{\sqrt{\frac{\sigma^2}{n}}} | \mu = 0\right) = \mathbb{P}\left(z \geq \frac{c_\alpha - 0}{\sqrt{4}}\right)\end{aligned}$$

where in (1) we standardized both terms and so $Z \sim \mathrm{N}(0,1)$. Therefore:

$$\alpha = \mathbb{P}\left(z \geq \frac{c_\alpha}{2}\right) = 1 - \Phi\Big(\frac{c_\alpha}{2}\Big) = \texttt{1 - pnorm(c\_alpha/2)}$$

therefore

- if $\alpha = 0.1$

  $$0.1 = 1 - \Phi\Big(\frac{c_\alpha}{2}\Big) \iff 0.9 = \Phi\Big(\frac{c_\alpha}{2}\Big) \iff \frac{c_\alpha}{2} = \Phi^{-1}(0.9)$$

  where Φ^{-1} is done with `qnorm` in R. So

  $$\frac{c_\alpha}{2} = \texttt{qnorm(0.9)} = 1.28 \implies c_\alpha = 2.56$$

- viceversa if $\alpha = 0.01$ following the same procedure we get that $c_\alpha = 4.66$

Finally, regarding the *power* of the test:

$$\begin{aligned}\tau(\mu) &= \mathbb{P}\left(\text{reject } H_0 | H_1\right) = \mathbb{P}\left(\overline{x} \geq c_\alpha | \mu = 1.5\right) \\ &\overset{(1)}{=} \mathbb{P}\left(z \geq \frac{c_\alpha - 1.5}{2}\right) = 1 - \Phi\left(\frac{c_\alpha - 1.5}{2}\right) \\ &= \begin{cases} 1 - \Phi\big(\frac{2.56-1.5}{2}\big) = 0.298, & \text{if } \alpha = 0.1 \\ 1 - \Phi\big(\frac{4.66-1.5}{2}\big) = 0.0571, & \text{if } \alpha = 0.01 \end{cases}\end{aligned}$$

where again in (1) we standardized. Notice that the powers are very very low (this because of n which is small). If we want to increase the power we should increase n.
Furthermore as we have seen if we increase α the power increases as well.

Esempio 13.2.3. Given a random sample from a Poisson distribution, the UMP test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is

$$\sum x_i \geq c_\alpha$$

remembering that for poisson we have that $\sum X_i \sim \mathrm{Pois}\left(n\theta\right)$. Given this idea consider:

$$\begin{cases} H_0 : \theta = 1 \\ H_1 : \theta = 2 \end{cases}$$

and

1. use the normal approximation (CLT) in order to determine both n and c_α such that $\alpha = 0.05$ and the power $1 - \beta = 0.9$ (here the news is that we fix both size and power, so this will end into a system);

2. calculate the p-value of the test when the sample is $(x_1, x_2, x_3) = (3, 5, 1)$ $(n = 3)$

We have:

1. we start by determining the α-related stuff:

$$\begin{aligned}\alpha = 0.05 &= \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq c_\alpha | H_0\right) = \mathbb{P}\left(\text{Pois}\left(n\theta\right) \geq c_\alpha | \theta = 1\right) \\ &\approxeq \mathbb{P}\left(\text{N}\left(n, n\right) \geq c_\alpha\right)\end{aligned}$$

this because $\theta = 1$ under H_0 and $\text{Pois}\left(n\theta\right) \approxeq \text{N}\left(n\theta, n\theta\right)$ as $n \to \infty$ by CLT (if $X \sim \text{Pois}\left(n\theta\right)$ then $\mathbb{E}\left[X\right] = n\theta$ and $\text{Var}\left[X\right] = n\theta$).
From the last equation we get, by standardization of inequality terms under probability, that:

$$\begin{aligned}0.05 &= \mathbb{P}\left(z \geq \frac{c_\alpha - n}{\sqrt{n}}\right) \\ 0.05 &= 1 - \mathbb{P}\left(z < \frac{c_\alpha - n}{\sqrt{n}}\right) \\ \Phi\left(\frac{c_\alpha - n}{\sqrt{n}}\right) &= 0.95 \\ \frac{c_\alpha - n}{\sqrt{n}} &= \Phi^{-1}(0.95) = \texttt{qnorm}(0.95) = 1.645 \\ \frac{c_\alpha - n}{\sqrt{n}} &= 1.645\end{aligned}$$

Regarding beta and power we have that:

$$\begin{aligned}1 - \beta = 0.9 = \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq c_\alpha | H_1\right) &= \mathbb{P}\left(\text{Pois}\left(\theta n\right) \geq c_\alpha | \theta = 2\right) \\ &= \mathbb{P}\left(\text{Pois}\left(2n\right) \geq c_\alpha\right) \\ &\approxeq \mathbb{P}\left(\text{N}\left(2n, 2n\right) \geq c_\alpha\right) \\ &= \mathbb{P}\left(z \geq \frac{c_\alpha - 2n}{\sqrt{2n}}\right)\end{aligned}$$

Therefore

$$\begin{aligned}0.9 &= 1 - \mathbb{P}\left(z < \frac{c_\alpha - 2n}{\sqrt{2n}}\right) \\ \Phi\left(\frac{c_\alpha - 2n}{\sqrt{2n}}\right) &= 0.1 \\ \frac{c_\alpha - 2n}{\sqrt{2n}} &= \Phi^{-1}(0.1) = \texttt{qnorm}(0.1) \\ \frac{c_\alpha - 2n}{\sqrt{2n}} &= -1.282\end{aligned}$$

Then we put all this α and β shit in a system of equation, obtaining both n and c_α with the desired properties:

$$\begin{cases} \frac{c_\alpha - n}{\sqrt{n}} = 1.645 \\ \frac{c_\alpha - 2n}{\sqrt{2n}} = -1.282 \end{cases} \quad \begin{cases} c_\alpha = 1.645\sqrt{n} + n \\ \frac{1.645\sqrt{n} + n - 2n}{\sqrt{2}\sqrt{n}} = -1.282 \end{cases} \quad \begin{cases} // \\ \frac{1.645}{\sqrt{2}} - \frac{\sqrt{n}}{\sqrt{2}} = -1.282 \end{cases}$$

$$\begin{cases} // \\ \sqrt{n} = 1.282 \cdot \sqrt{2} + 1.645 = 3.46 \end{cases} \quad \begin{cases} // \\ n = 11.96 \approxeq 12 \end{cases} \quad \begin{cases} n = 12 \\ c_\alpha = 1.645\sqrt{12} + 12 = 17.7 \end{cases}$$

2. for the second point the p-value is:

$$\begin{aligned} p &= \mathbb{P}\left(T_n \geq t_n | H_0\right) = \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq 3 + 5 + 1 | H_0\right) \\ &= \mathbb{P}\left(\text{Pois}\left(3\theta\right) \geq 9 | \theta = 1\right) = \mathbb{P}\left(\text{Pois}\left(3\right) \geq 9\right) \\ &= 1 - \mathbb{P}\left(\text{Pois}\left(3\right) \leq 8\right) \\ &= 1 - \texttt{ppois}(8, 3) \\ &= 0.0038 \end{aligned}$$

Esempio 13.2.4. Given $(X_1, \dots X_n) \sim \text{Pois}\left(\theta\right)$. Derive the UMP test for

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

with $\theta_0 < \theta_1$.
We have that

$$\begin{aligned} \lambda(\mathbf{x}) = \frac{\text{L}\left(\theta_0\right)}{\text{L}\left(\theta_1\right)} &= \frac{\prod_{i=1}^{n} e^{-\theta_0} \theta_0^{x_i} / \cancel{x_i!}}{\prod_{i=1}^{n} e^{-\theta_1} \theta_0^{x_i} / \cancel{x_i!}} = \frac{e^{-n\theta_0} \theta_0^{\sum_{i=1}^{n} x_i}}{e^{-n\theta_0} \theta_1^{\sum_{i=1}^{n} x_i}} \\ &= e^{n(\theta_1 - \theta_0)} \cdot \left(\frac{\theta_0}{\theta_1}\right)^{\sum_{i=1}^{n} x_i} \end{aligned}$$

with the ratio $\frac{\theta_0}{\theta_1} < 1$. Since that, large values of $\sum_{i=1}^{n} x_i$ corresponds to small values of $\lambda(\mathbf{x})$ which in furtns favours H_1. Therefore we reject when $\sum_{i=1}^{n} x_i > c_\alpha$. Since with iid rvs $\sum_{i=1}^{n} X_i \sim \text{Pois}\left(n\theta\right)$ we have

$$\alpha = \mathbb{P}\left(\sum_{i=1}^{n} x_i > c_\alpha | \theta_0\right) = 1 - F(c_\alpha) = 1 - \texttt{ppois}(c_\alpha)$$

In example, if $n = 100, \alpha = 0.05$ and the hypotheses are

$$\begin{aligned} H_0 &: \theta = 20 \\ H_1 &: \theta = 30 \end{aligned}$$

Then $Y = \sum_{i=1}^{n} X_i$. Under:

1. H_0 we have that $Y \sim \text{Pois}\left(100 \cdot 20\right)$
2. H_1 we have that $Y \sim \text{Pois}\left(100 \cdot 30\right)$

if $\alpha = 0.05 = 1 - F(c_\alpha)$ then

$$0.95 = F(c_\alpha) \implies F^{-1}(0.95) = c_\alpha \iff \texttt{qpoiss}(0.95, 2000) = c_\alpha$$

$2074 = c_\alpha$

If $\sum_{i=1}^n x_i > 2074$ then we reject the null with $\alpha = 0.05$.

Esempio 13.2.5 (Esame vecchio viroli)**.** Let $X_1, \dots, X_n$ be a random sample from the density function

$$\theta^2 x e^{-\theta x}$$

with $X > 0$, $\theta > 0$, $\mathbb{E}[X] = \frac{2}{\theta}$ and $\mathrm{Var}[X] = \frac{2}{\theta^2}$. Consider the system of hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1 > \theta_0$$

Prove that the rejection region of the neyman pearsons test is of the form $RR = \{\overline{x} < C_\alpha\}$and write analytically C_α:

1. it cannot be derived
2. $C_a = \frac{\log k/n - 2(\log\theta_0 - \log\theta_1)}{\theta_1 - \theta_0}$: taluni suggeriscono questa
3. $C_a = \frac{\log k/n - 2(\log\theta_1 - \log\theta_0)}{\theta_1 - \theta_0}$
4. $C_a = \frac{\log k/n + 2(\log\theta_0 - \log\theta_1)}{\theta_1 - \theta_0}$

Si ha che

$$\begin{aligned}
\lambda(x) = \frac{\mathrm{L}(\theta_0)}{\mathrm{L}(\theta_1)} &= \frac{\prod_{i=1}^n \theta_0^2 x_i e^{-\theta_0 x_i}}{\prod_{i=1}^n \theta_1^2 x_i e^{-\theta_1 x_i}} \\
&= \left(\frac{\theta_0}{\theta_1}\right)^{2n} \frac{(\prod_{i=1}^n x_i) \cdot e^{-\sum_{i=1}^n \theta_0 x_i}}{(\prod_{i=1}^n x_i) \cdot e^{-\sum_{i=1}^n \theta_i x_i}} \\
&= \left(\frac{\theta_0}{\theta_1}\right)^{2n} \exp\left(\theta_1 \sum_{i=1}^n x_i - \theta_0 \sum_{i=1}^n x_i\right) \\
&= \left(\frac{\theta_0}{\theta_1}\right)^{2n} \exp\left((\theta_1 - \theta_0) \sum_{i=1}^n x_i\right)
\end{aligned}$$

e se $\sum_{i=1}^n x_i$ aumenta aumenta la verosimiglianza di H_0 quindi siamo portati a non rifiutare, ergo:

- $\sum_{i=1}^n X_i > k_\alpha \implies \overline{R}$
- $\sum_{i=1}^n X_i \leq k_\alpha \implies R$

Si ha allora che

$$\begin{aligned}
\alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) &= \mathbb{P}\left(\sum_{i=1}^n X_i \leq k_\alpha | \theta = \theta_0\right) \\
&= F_{\sum_{i=1}^n X_i}(k_\alpha)
\end{aligned}$$

poi bo come determinare la cumulata (analiticamente), forse clt però non so quanto si eviti l'integrale e si possa determinare analiticamente

Esempio 13.2.6 (Assignment 4 Viroli, Exercise 1 (optimal testing))**.** The Laplace distribution appears naturally when the measurement of quantity is not only subject to measurement error of the measurement instrument, but also due to variability of heterogeneous instrument operators. In total $n = 81$ measurements $X = (X_1, ..., X_{81})$ were made according to a Laplace(θ),

$$f_\theta(x) = \frac{1}{2}\theta e^{-\theta|x|}$$

We want to test the following simple hypotheses,

$$\begin{cases} H_0 : \theta = 1 \\ H_1 : \theta = 2/3 \end{cases}$$

1. Use the Central Limit Theorem to determine the approximate distribution of $\sum_{i=1}^{81} |x_i|$ under some unknown value of θ
2. Use the result in (1) to determine the optimal test at the 5% level. In other words, determine the optimal rejection region $RR \in \mathbb{R}^{81}$
3. use again the result in 1 to determine the power of this test.

Respectively:

1. following clt we have that

$$\frac{\sum_i |X_i| - \mathbb{E}\left[\sum_i |X_i|\right]}{\sqrt{\text{Var}\left[\sum_i |X_i|\right]}} \xrightarrow{d} \text{N}(0, 1)$$

with

$$\begin{aligned} \mathbb{E}\left[|X_i|\right] &= \int_{-\infty}^{+\infty} |x| \cdot \frac{1}{2}\theta e^{-\theta|x|}\, \mathrm{d}x = 2\int_0^{+\infty} x \cdot \frac{1}{2}\theta e^{-\theta|x|}\, \mathrm{d}x \\ &= \int_0^{+\infty} x \cdot \theta e^{-\theta|x|}\, \mathrm{d}x \\ &= \ldots \text{by parts} \ldots \\ &= \frac{1}{\theta} \end{aligned}$$

$$\mathbb{E}\left[\sum_i |X_i|\right] = \sum_i \mathbb{E}\left[|X_i|\right] = \frac{n}{\theta}$$

For the variance

$$\begin{aligned} \mathbb{E}\left[|X_i|^2\right] &= \int_{-\infty}^{+\infty} |x|^2 \cdot \frac{1}{2}\theta e^{-\theta|x|} = \int_{-\infty}^{+\infty} x^2 \cdot \frac{1}{2}\theta e^{-\theta|x|} \\ &= 2 \cdot \frac{1}{2}\int_0^{+\infty} x^2 \theta e^{-\theta|x|} \\ &= \ldots \text{by parts two times} \ldots \\ &= \frac{2}{\theta^2} \end{aligned}$$

and then

$$\text{Var}\left[|X_i|\right] = \mathbb{E}\left[|X_i|^2\right] - \mathbb{E}\left[|X_i|\right]^2 = \frac{2}{\theta^2} - \frac{1}{\theta^2} = \frac{1}{\theta^2}$$
$$\text{Var}\left[\sum_{i=1}^{n} |X_i|\right] = \sum_{i=1}^{n} \text{Var}\left[|X_i|\right] = n \cdot \text{Var}\left[|X_i|\right] = \frac{n}{\theta^2}$$

Therefore, finally

$$\frac{\sum_i |X_i| - \frac{n}{\theta}}{\sqrt{n}/\theta} \xrightarrow{d} \text{N}(0,1)$$

and thus

$$\sum_i |X_i| \xrightarrow{d} \text{N}\left(\frac{n}{\theta}, \frac{n}{\theta^2}\right)$$

2. we have that

$$\lambda(\mathbf{x}) = \frac{\prod_{i=1}^{81} \frac{1}{2} e^{-|x_i|}}{\prod_{i=1}^{81} \frac{1}{2} \cdot \frac{2}{3} e^{-\frac{2}{3}|x_i|}} = \ldots = \left(\frac{3}{2}\right)^{81} \exp\left(-\frac{1}{3}\sum_{i=1}^{81} |x_i|\right)$$

Here if $\sum_{i=1}^{81} |x_i|$ increase, $\lambda(\mathbf{x})$ decrease and go against H_0 so focusing on the only random part of the equation above that is $\sum_{i=1}^{81} |x_i|$:

- if $\sum_{i=1}^{81} |x_i| > c_\alpha \implies R$, we reject H_0
- if $\sum_{i=1}^{81} |x_i| \leq c_\alpha \implies \overline{R}$

with c_α to be determined. We know that under the null hypothesis, when n is large

$$\sum_{i=1}^{n} X_i | H_0 \sim \text{N}\left(\frac{n}{\theta}, \frac{n}{\theta^2}\right) = \text{N}\left(\frac{n}{1}, \frac{n}{1}\right) = \text{N}(n, n)$$

in our case $\text{N}(81, 81)$ will be ok since 81 is large. We determine c_α as follows:

$$\begin{aligned}
\alpha &= \mathbb{P}\left(\text{reject } H_0 | H_0\right) = \mathbb{P}\left(\sum_{i=1}^{n} |X_i| > c_\alpha | \theta = 1\right) \\
&= 1 - \mathbb{P}\left(\sum_{i=1}^{n} |X_i| < c_\alpha | \theta = 1\right) \\
&= 1 - \int_{-\infty}^{c_\alpha} \text{N}(81, 81)
\end{aligned}$$

Thus when $\alpha = 0.05$ we have that

$$\int_{-\infty}^{c_\alpha} \text{N}(81, 81) = 0.95 \iff c_\alpha = \Phi^{-1}_{\text{N}(81,81)}(0.95) = \texttt{qnorm}(0.95, 81, \texttt{sqrt}(81)) = 95.8$$

So our test will be:

- if $\sum_{i=1}^{81} |x_i| > 95.8 \implies R$, we reject H_0;
- if $\sum_{i=1}^{81} |x_i| \leq 95.8 \implies \overline{R}$.

3. under $H_1 : \theta = 2/3$ we have that

$$\sum_{i=1}^{n} X_i | H_1 \sim \mathrm{N}\left(\frac{3}{2}n, \frac{9}{4}n\right)$$

in our case $\mathrm{N}(3 \cdot 81/2, 9 \cdot 81/4) \cong \mathrm{N}(121.5, 182.2)$. The power is

$$\begin{aligned} 1 - \beta &= \mathbb{P}(\text{reject } H_0 | H_1) \\ &= \mathbb{P}\left(\sum_{i=1}^{81} |X_i| > 95.8 | \theta = \frac{2}{3}\right) \\ &= 1 - \mathbb{P}\left(\sum_{i=1}^{81} |X_i| < 95.8 | \theta = \frac{2}{3}\right) \\ &= 1 - \Phi_{\mathrm{N}(121.5, 182.2)}(95.8) \\ &= 1 - \texttt{pnorm}(95.8, 121.5, \texttt{sqrt}(182.2)) \\ &= 0.9715 \end{aligned}$$

which is very good.

13.3 Generalized likelihood ratio test (GLRT)

Osservazione 337. In this last part regarding hypotheses testing we look at complex (no more simple) hypotheses. The test has this structure

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \boldsymbol{\Theta}$, that is the two set form a bipartition of the parameter space. What should we do in the general situation?

13.3.1 Basic setup

Osservazione importante 115 (GLRT construction and variants). We use the generalized likelihood ratio test. We base our decision on a ratio that involves likelihoods; we can't use the likelihood in a point since both the null and the alternative have a set of points. Then we take the maximum likelihood under different parameter sets:

- at the numerator the maximum likelihood in the null parameter space Θ_0;
- at the denominator some books use the likelihood in the alternative parameter space Θ_1, others (and we'll do this way because it simplifies some aspects) in the full parameter space $\boldsymbol{\Theta}$, including θs which belong to the null; so at the denominator we'll have the likelihood of the maximum likelihood estimates. This latter is better for two reason:

1. it's not a constrainted optimization problem at the denominator (so that simplifies);
2. in this manner we have a ratio which is for sure between 0 and 1, and it is interpretable in terms of "likelihood of the null". We have the maximum likelihood under the null divided by the maximum likelihood overall: the more the ratio is close to 1, the more the null likely

Finally note there's no unique formulation: some put the alternative/full at numerator and null at the denominator so here be dragons.

Definizione 13.3.1 (GLRT)**.** Our test will be:

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\theta\in\Theta_0} f(\mathbf{x}|\theta)}{\sup\limits_{\theta\in\mathbf{\Theta}} f(\mathbf{x}|\theta)} = \frac{\mathrm{L}\left(\hat{\theta}_0\right)}{\mathrm{L}\left(\hat{\theta}\right)}$$

and as said it will be $0 \leq \lambda(\mathbf{x}) \leq 1$.

Osservazione importante 116 (α and cutoff point). We can define rejection region similarly to what we've done so far such as $R = \{\mathbf{x} : \lambda(\mathbf{x}) < k_\alpha\}$ (we reject the null when the ratio is low); the cutoff point k_α depends on α (to control the first type error) and is chosen such as the maximum probability of rejection in the null parameter space is at most α

$$\sup_{\theta\in\Theta_0} \mathbb{P}\left(\lambda(\mathbf{x}) < k_\alpha; \theta\right) = \alpha$$

Osservazione importante 117 (pros and cons). We cannot say that this test is optimal/UMP in the sense of Neyman-Pearson, it's just for a particular system of hypotheses. However we have that:

- **pros**: GLRT is consistent in the sense that

  $$\lim_{n\to\infty} \mathbb{P}\left(\lambda_n(\mathbf{x}) < k_\alpha | H_1\right) = 1$$

  that is: the probability of rejecting the null, in case the alternative is true, tends to 1 as $n \to \infty$. Consistency here means that the power goes to 1 (we reject correctly) when n increases;

- **cons**: most of times it's very difficult to find the distribution of the ratio $\lambda(\mathbf{x})$; we should have the distribution of $\lambda(\mathbf{x})$ because by using it we can set α and determine k_α.

Osservazione 338. What can we do when we don't know the distribution/the ratio cannot be used to determine k_α? In this situation we have another important result which helps us in order to determine k_α.
The following is one of the most important result for testing hypotheses.

13.3.2 Wilks theorem

13.3.2.1 Basic version

Teorema 13.3.1 (Wilks theorem (1938)). *For a system of hypotheses (special case) such as:*

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

under the null hypothesis H_0, it's possible to prove that the limit of transformation $-2\log\lambda(\mathbf{x})$ is a chi square with one degree of freedom:

$$\lim_{n\to\infty} -2\log\lambda_n(\mathbf{x}) \xrightarrow{d} \chi^2(1)$$

Osservazione 339. It's so important because if we don't know the distribution of $\lambda(\mathbf{x})$ we can use this approximation.
The convergence is distribution, weak, and the rate of convergence is lower than other kind of convergences (need more n), but in any case it's a very important result.

NB: L'ha skippata

Proof. Let $\ell(\theta_0)$ be the log-likelihood at θ_0 and $\ell\left(\hat{\theta}\right)$ the MLE loglik

$$-2\log\lambda = -2\ell(\theta_0) + 2\ell\left(\hat{\theta}\right)$$

with

$$\lambda = \frac{\mathrm{L}(\theta_0)}{\mathrm{L}\left(\hat{\theta}\right)}$$

Applying Taylor expansion to $\ell(\theta_0)$ around $\theta = \hat{\theta}$

$$\begin{aligned} \ell(\theta_0) &\approxeq \ell\left(\hat{\theta}\right) + (\theta_0 - \hat{\theta})\underbrace{\ell'(\hat{\theta})}_{=0} + \frac{1}{2}(\theta_0 - \hat{\theta})^2 \underbrace{\ell''(\hat{\theta})}_{-i_n(\hat{\theta})} \\ &\approxeq \ell\left(\hat{\theta}\right) - \frac{1}{2}(\theta_0 - \hat{\theta})^2 i_n(\hat{\theta}) \end{aligned}$$

Now consider that

$$\frac{i_n(\hat{\theta})}{n} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2}\log f(x_i|\hat{\theta})$$

where $\frac{\partial^2}{\partial\theta^2}\log f(x_i|\hat{\theta})$ is a random variable because

$$\frac{\partial^2}{\partial\theta^2}\ell\left(\hat{\theta}\right) = \frac{\partial^2}{\partial\theta^2}\log\prod_{i=1}^{n} f(x_i|\hat{\theta})$$

According to the weak law of large numbers (WLLN)

$$\begin{aligned} \frac{i_n(\hat{\theta})}{n} &\xrightarrow{p} \mathbb{E}\left[-\ell''(\hat{\theta})\right] = I_1(\hat{\theta}) \\ i_n(\hat{\theta}) &\xrightarrow{p} n\,\mathbb{E}\left[-\ell''(\hat{\theta})\right] = n \cdot I_1(\hat{\theta}) = I_n(\hat{\theta}) \end{aligned}$$

so the observed Fisher information $\xrightarrow{p}$ expected Fisher information

$$\ell(\theta_0) \approxeq \ell\left(\hat{\theta}\right) - \frac{1}{2} nI(\hat{\theta})(\theta_0 - \hat{\theta})^2$$

Moreover by ML $\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}\right] = \theta_0$ under the H_0. Therefore

$$\ell(\theta_0) \approxeq \ell\left(\hat{\theta}\right) - \frac{1}{2} nI(\theta_0)(\theta_0 - \hat{\theta})^2$$

poi così a random

$$\begin{aligned} -2\log\lambda &= -2\ell\left(\hat{\theta}\right) + nI(\theta_0)(\theta_0 - \hat{\theta})^2 - 2\ell\left(\hat{\theta}\right) \\ &\approxeq nI(\theta_0)(\theta_0 - \hat{\theta})^2 \end{aligned}$$

under the H_0 and if $n \to \infty$.
By CLT then

$$\frac{\hat{\theta} - \theta_0}{\sigma^2_{\hat{\theta}}} = \frac{1}{nI(\theta_0)}$$

therefore

$$nI(\theta_0)(\hat{\theta} - \theta_0)^2 \xrightarrow{d} \chi^2(1)$$

□

13.3.2.2 General formulation of Wilks theorem

Osservazione 340. Problem with the base version is that is very specific to a single hypotheses setup. Therefore a more general version is needed.

Teorema 13.3.2 (Wilks theorem)**.** *Let our system of hypotheses be general*

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \quad (\theta \notin \Theta_0) \end{cases}$$

Then we can construct the transformation $-2\log\lambda(\mathbf{x})$*. It is possible to show that even in this situation it has a limiting distribution of a chi-square with* k *degrees of freedom:*

$$W = -2\log\lambda(\mathbf{x}) \xrightarrow{d} \chi^2(k)$$

where k *is the difference between the dimensionality (number of parameters) of* Θ_1 *and* Θ_0*.*

Esempio 13.3.1 (Haemoglobin repeated measurement)**.** Let $X_{ij} \sim \mathrm{N}\left(\mu_i, \sigma_i^2\right)$ be the haemoglobin of an athlete i (with $i = 1, \ldots, n$) measured at time j (with $j = 1, \ldots, T_i$): different individuals can have different number of observations. Each athlete has an own distribution in terms of μ_i and σ_i^2.

We wish to test if the haemoglobin measurement across athletes have the same variance (haemoglobin varies equally in all people): the hypotheses are

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 \\ H_1 : \sigma_i^2 \text{ are not equal} \end{cases}$$

To calculate the ratio we need the maximum likelihood under the null and the general maximum likelihood. The likelihood under the null (uses a common σ^2 as assumption) involve a double product for both athletes and times:

TODO: Non chiaro ché sotto calcoli la ver per H_1 invece di tutta

$$\mathrm{L}\left(\mathbf{x}|\mu_i, \sigma^2\right) = \prod_{i=1}^{n}\prod_{j=1}^{T_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_{ij}-\mu_i)^2}{\sigma^2}\right)$$

The likelihood under the alternative H_1 encompasses different σ_i^2, and is:

$$\mathrm{L}\left(\mathbf{x}|\mu_i, \sigma_i^2\right) = \prod_{i=1}^{n}\prod_{j=1}^{T_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(x_{ij}-\mu_i)^2}{\sigma_i^2}\right)$$

We compute not the ratio but directly the logarithm used for the $-2\log\lambda(\mathbf{x})$ transformation Then

$$\begin{aligned}\log\lambda(\mathbf{x}) &= \log\left(\frac{\prod_{i=1}^{n}\prod_{j=1}^{T_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x_{ij}-\mu_i)^2}{\sigma^2}\right)}{\prod_{i=1}^{n}\prod_{j=1}^{T_i}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{1}{2}\frac{(x_{ij}-\mu_i)^2}{\sigma_i^2}\right)}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{T_i}\log\sigma - \sum_{i=1}^{n}\sum_{j=1}^{T_i}\log\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\frac{(x_{ij}-\mu_i)^2}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\frac{(x_{ij}-\mu_i)^2}{\sigma_i^2}\end{aligned}$$

Looking at this last formula we don't have any clue of which probabilistic distribution could it be. However thanks to Wilks theorem we know that if we multiply it by -2 then we have that $-2\log\lambda(\mathbf{x}) \sim \chi^2(n-1)$. The number of degrees of freedom is obtained by:

- the dimensionality of the denominator: under the null H_0 we need to estimate n mus and 1 sigma, so it's $n+1$;
- under the alternative H_1 we have to estimate is $n+n=2n$ parameters (n mu and n sigma)
- therefore, overall, the number of degrees of freedom of the Chi square is $2n-(n+1)=n-1$

Now if we had data we could finish by calculating the test. The next example we have the data so we see the complete solution of the problem.

Osservazione 341. This second example is another way to resolve the so called ANOVA. Fisher developed it in 1921 before Wilks theorem; first a refresher of the classic way and second how to do the same using the generalized testing procedure.

NB: Soluzione giusto accennata, facciamo con il test

Esempio 13.3.2 (Analysis of variance (Fisher 1921))**.** Let X_{ij} be the length of newborn from mothers divided in $k=3$ groups:

- $i = 1$: no cocaine
- $i = 2$: usage only during the first trimester of pregnancy
- $i = 3$: usage during the 9 months

with groups $i = 1, \dots, k = 3$, and mother (we consider one baby per mother) inside group denoted by $j = 1, \dots, n_i$.
The model is that $X_{ij} \sim \mathrm{N}\left(\mu_i, \sigma^2\right)$ (so each cocaine group has a common mean but we make the homoschedasticity hypothesis with a common σ^2). The data is as follows ($n = 94$) showing a diminishing length for cocaine use:

$$\begin{aligned} \overline{x}_1 &= 51.1, & n_1 &= 39 \\ \overline{x}_2 &= 49.3, & n_2 &= 19 \\ \overline{x}_3 &= 48.0, & n_1 &= 36 \end{aligned}$$

By ANOVA we assume $X_{ij} = \mu_i + \varepsilon_{ij}$, with $\varepsilon_{ij} \sim \mathrm{N}\left(0, \sigma^2\right)$; the aim is to check

$$\begin{cases} H_0 := \mu_1 = \mu_2 = \mu_3 \\ H_1 : \text{not equal} \end{cases}$$

The Fisher solution was, under H_0, homoschedasticity, Gaussianity of error:

$$F = \frac{\text{var between}}{\text{var within}} \sim \mathrm{F}\left(k-1, n-k\right)$$

where

$$\begin{aligned} \text{var between} &= \frac{SSE^b}{k-1} = \frac{\sum_{i=1}^n n_i(\overline{x}_i - \overline{x})^2}{k-1} \\ \text{var within} &= \frac{SSE^w}{n-k} = \frac{\sum_{i=1}^n n_i \sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}{n-k} \end{aligned}$$

Moreover

$$SSE^b + SSE^w = SSE^T = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})^2$$

In our case:

$$SSE^b = 181.375 \quad SSE^w = 885.58 \quad SSE^T = 1066.955$$

Then

$$F = \frac{181.375/2}{885.58/91} = 9.31$$

and $p = 1 - \mathrm{F}\left(2, 91, 9.31\right) = 0.0002$ so we reject H_0.

Esempio 13.3.3. We solve the same problem with GLRT, since it's an equivalent alternative. We know that $-2\log\lambda(\mathbf{x}) \xrightarrow{d} \chi^2\left(k-1\right)$ with $k - 1 = 2$. To compute the test we need both the likelihoods:

- the likelihood under H_0 is (with single unique μ; σ^2 is common by setup of anova)

$$f(\mathbf{x}|\mu,\sigma^2) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{1}{2}\frac{(x_{ij}-\mu)^2}{\sigma^2}\right)$$

- Likelihood under H_1 (we have different μ_1, μ_2, μ_3 pointed by μ_i):

$$f(\mathbf{x}|\mu_i,\sigma^2) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{1}{2}\frac{(x_{ij}-\mu_i)^2}{\sigma^2}\right)$$

In order to compute the likelihoods and their ratio, we need to have estimates of the parameters involved, that is μ_i, σ^2; we can estimate μ_i through $\overline{x}_i$ (the sample equivalent):

- under H_0 the MLE estimate of μ is the sample mean $\hat{\mu} = \overline{x}$ (gaussian distribution) that is:

$$\overline{x} = \frac{\sum_i \sum_j x_{ij}}{n}$$

  while the MLE estimation of σ^2 is given by the sample

$$\hat{\sigma}^2 = S_0^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2}{n}$$

  Here S_0^2 means sample variance under the null hypothesis. The denominator is not $n-1$ because the MLE estimator of the variance it's biased, it's not correct

- under H_1 the MLE of μ_i and σ^2 are:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} = \overline{x}_i$$
$$\hat{\sigma}^2 = S_1^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{n}$$

  where $S_1^2 = SSE^W/n$ and S_1^2 means sample variance under the alternative hypothesis.

We now want to have the ingredients to compute the likelihoods (substituting parameters with their estimates) and leaving the function only data dependent. The ratio of the likelihood is then

$$\lambda(\mathbf{x}) = \frac{\prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi S_0^2}}\exp\left(-\frac{1}{2}\frac{(x_{ij}-\overline{x})^2}{\sigma^2}\right)}{\prod_{i=1}^{k}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi S_1^2}}\exp\left(-\frac{1}{2}\frac{(x_{ij}-\overline{x}_i)^2}{\sigma^2}\right)} \overset{(1)}{=} \left(\frac{S_0^2}{S_1^2}\right)^{-\frac{n}{2}} \frac{\exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2}{S_0^2}\right)}{\exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{S_1^2}\right)}$$

where taking out of the double product we elevate the S n times, the /2is given by square root and the minus considered that the S are at the denominator.

Now consider just the numerator of the last equation: we see that by multiplying and dividing by n the red part below is equal to S_0^2:

$$\exp\left(-\frac{1}{2}\frac{n}{n}\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x})^2}{S_0^2}\right) = \exp\left(-\frac{1}{2}n\right)$$

if we do this trick similarly at the denominator we get the same except noting the definition of S_1^2.

$$\lambda(\mathbf{x}) = \left(\frac{S_0^2}{S_1^2}\right)^{-\frac{n}{2}} \cdot \frac{\exp\left(-\frac{1}{2}n\right)}{\exp\left(-\frac{1}{2}n\right)} = \left(\frac{S_0^2}{S_1^2}\right)^{-\frac{n}{2}} = \left(\frac{S_1^2}{S_0^2}\right)^{\frac{n}{2}}$$

Now calculating with data it turns out that $S_0^2 = 885.58$, $S_1^2 = 1066.995$: therefore

$$\lambda(\mathbf{x}) = \left(\frac{885.58}{1066.995}\right)^{\frac{94}{2}} = 0.0001654$$

and to use the Wilks general test

$$w_c = -2\log(0.0001654) = 17.41386 \xrightarrow{d} \chi^2\,(k-1)$$

the degrees of freedom are $k-1$, and in our case 2:

- under the null are k means + 1 variance $(k+1)$
- under the alternative 1 mean and 1 variance (2)
- so $k+1-2 = k-1$

Finally we could compare equivalently

- pvalue vs α: $p < \alpha \implies R$
- w_c vs q_α: $w_c > q_\alpha \implies R$

In our example

- we could compare the test with the reject value, obtained using `qchisq(0.95, 2) = 5.991`, and since 17.4> 5.9 we reject the null.
- we could otherwise calculate the p-value of the test

  $$\mathbb{P}\,(W \geq w_c) = 1 - \texttt{pchisq}(17.41, 2) = 0.0002$$

  and since $p < \alpha = 0.05$ we reject as well

13.3.3 Asymptotic equivalent test: Wilks, Wald, Score

They are

1. Wilks test (1958)

   $$W = -2\log\lambda(\mathbf{x})$$

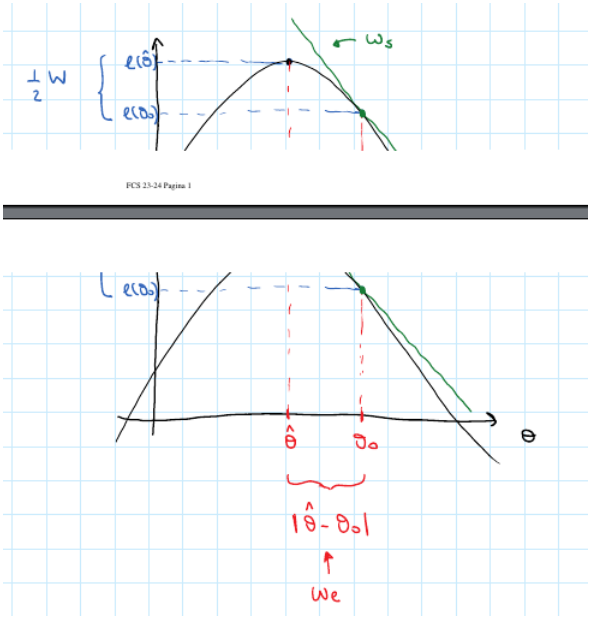


Figure 13.3: Asymptotic equivalent tests

2. Wald test (1943):

$$W_e = n \cdot i_1(\hat{\theta}) \cdot (\hat{\theta} - \theta_0)^2$$

   the difference between $\hat{\theta}$ and θ_0 is an important part of the Wald text

3. Score test (Rao, 1947):

$$W_s = \frac{\ell'(\theta_0)^2}{I_n(\theta_0)} = \frac{\ell'(\theta_0)^2}{n \cdot I_1(\theta_0)}$$

   where $\ell'(\theta_0)$ is the first derivative of log-likelihood evaluated at θ_0. The green line in the image is the tangent (first derivative) of the log likelihood in the point θ_0. The test is proportional to the inclination of the loglikelihood in the point

They describe different aspect of the likelihood (see fig 13.3), but they are all asymptotic equivalent. One can compute one of them and the numeric value obtained is the same as the one obtained with the other (as $n \to \infty$)

Esempio 13.3.4. Suppose we have observed $n = 50$ values from $\text{Exp}(\theta)$ with $\overline{x} = 0.19$ (credo oppure 0.13 non so).
Check the hypothesis

$$\begin{cases} H_0 : \theta = 5.5 \\ H_1 : \theta \neq 5.5 \end{cases}$$

by using the three asymptotic tests.
Given the exponential distribution the density $f(x) = \theta \exp(-\theta x)$ we have that

likelihood and loglikelihood

$$\mathrm{L}\left(\theta\right)=\prod_{i=1}^{n}\theta e^{-\theta x_i}=\theta^n e^{-\theta\sum_{i=1}^{n}x_i}$$

$$\ell(\theta)=n\log\theta-\theta\sum_{i=1}^{n}x_i$$

The maximum likelihood estimate is

$$\hat{\theta}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\overline{x}}=\frac{1}{0.19}=5.26$$

For

1. the Wilks tests:

$$\begin{aligned}W&=-2\log\left(\frac{\mathrm{L}\left(\theta_0\right)}{\mathrm{L}\left(\hat{\theta}\right)}\right)=-2\Big(\ell(\theta_0)-\ell\Big(\hat{\theta}\Big)\Big)\\&=-2\left(n\log\frac{\theta_0}{\hat{\theta}}-(\theta_0-\hat{\theta})\sum_{i=1}^{n}x_i\right)\overset{(1)}{=}-2\left(n\log\frac{\theta_0}{\hat{\theta}}-(\theta_0-\hat{\theta})\frac{n}{\hat{\theta}}\right)\\&=-2n\left(\log\frac{\theta_0}{\hat{\theta}}-\frac{\theta_0-\hat{\theta}}{\hat{\theta}}\right)=-2n\left(\log\frac{\theta_0}{\hat{\theta}}+1-\frac{\hat{\theta}}{\hat{\theta}}\right)\end{aligned}$$

   where in (1) we merely noted that $\sum_{i=1}^{n}x_i=n/\hat{\theta}$. For our data we have

$$W=-100\left(\log\frac{5.5}{5.26}+1-\frac{5.5}{5.26}\right)=0.101$$

2. the Wald test $W_e=n\cdot i(\hat{\theta})\cdot(\hat{\theta}-\theta_0)^2$ we need to evaluate

$$\ell''(\theta)\big|_{\theta=\hat{\theta}}=\frac{\partial}{\partial\theta}\underbrace{\left(n\cdot\frac{1}{\theta}-\sum_{i=1}^{n}x_i\right)}_{\ell'(\theta)}\big|_{\theta=\hat{\theta}}=-n\frac{1}{\theta^2}\big|_{\theta=\hat{\theta}}$$

   so

$$w_e=\frac{n}{\hat{\theta}^2}(\hat{\theta}-\theta_0)^2=n\left(1-\frac{\theta_0}{\hat{\theta}}\right)^2=50\left(1-\frac{5.5}{5.26}\right)^2=0.1045$$

3. the Score test

$$\begin{aligned}w_s&=\frac{\ell'(\theta_0)^2}{nI_1(\theta_0)}=\frac{\ell'(\theta_0)^2}{n\cdot\theta_0^{-2}}=\frac{\left(\frac{n}{\theta_0}-\sum_{i=1}^{n}x_i\right)^2}{\frac{n}{\theta_0^2}}=\frac{\left(\frac{n}{\theta_0}-\frac{n}{\hat{\theta}}\right)^2}{\frac{n}{\theta_0^2}}\\&=\frac{n^2\left(\frac{1}{\theta_0}-\frac{1}{\hat{\theta}}\right)}{n\cdot\theta_0^{-2}}=n\left(1-\frac{\theta_0}{\hat{\theta}}\right)^2=0.1045\end{aligned}$$

Esempio 13.3.5 (Esame vecchio viroli)**.** Given the system of hypotheses for a geometric random variable $X \sim p(1-p)^{x-1}$ with parameter p

$$\begin{cases} H_0 : p = 1/2 \\ H_1 : p \neq 1/2 \end{cases}$$

derive the **generalized likelihood ratio test**.
Per glrt lei intende $\lambda(x)$ (senza il $-2\log$) quindi calcoliamo la verosimiglianza e facciamo il rapporto con al numeratore quella sotto nulla ($p = 1/2$) mentre al denominatore quella basata su stima di massima verosimiglianza. Si ha

$$\begin{aligned} \mathrm{L}(p) &= \prod_{i=1}^{n} p \cdot (1-p)^{x_i - 1} = p^n (1-p)^{\sum_{i=1}^n (x_i - 1)} \\ &= p^n (1-p)^{(\sum_{i=1}^n x_i) - n)} = \left(\frac{p}{1-p}\right)^n \cdot (1-p)^{\sum_{i=1}^n x_i} \end{aligned}$$

Pertanto si ha

$$\lambda(x) = \frac{\mathrm{L}(\theta_0)}{\mathrm{L}\left(\hat{\theta}\right)} = \frac{1^n \cdot \left(\frac{1}{2}\right)^{\sum_{i=1}^n x_i}}{\left(\frac{\hat{p}}{1-\hat{p}}\right)^n \cdot (1-\hat{p})^{\sum_{i=1}^n x_i}} = \left(\frac{1-\hat{p}}{\hat{p}}\right)^n \cdot \left[\frac{1}{2(1-\hat{p})}\right]^{\sum_{i=1}^n x_i}$$

Esempio 13.3.6 (Esame vecchio viroli)**.** Given the system of hypothese for a gamma random variable $X \sim \mathrm{Gamma}(2, \beta)$ with parameter β

$$\begin{aligned} H_0 &: \beta = 1 \\ H_1 &: \beta \neq 1 \end{aligned}$$

derive the **genralized likelihood ratio test**

- $\lambda = \hat{\beta}^{-2n} e^{-\sum_{i=1}^n x_i (1-\hat{\beta})}$
- $\lambda = \hat{\beta}^{-2n} e^{-\sum_{i=1}^n x_i (1+\hat{\beta})}$
- $\lambda = \hat{\beta}^{2n} e^{-\sum_{i=1}^n x_i (1-\hat{\beta})}$
- $\lambda = \hat{\beta}^{-2n} e^{-\sum_{i=1}^n x_i (n+\hat{\beta})}$

Dovrebbe essere la prima. Si ha che la densità di una siffatta gamma è

$$f(x) = \frac{\beta^2 x^{2-1} e^{-\beta x}}{\Gamma(2)} = \frac{\beta^2 x e^{-\beta x}}{1} = \beta^2 x e^{-\beta x}$$

La verosimiglianza

$$\mathrm{L}(\beta) = \prod_{i=1}^{n} \beta^2 \cdot x \cdot e^{-\beta x} = \beta^{2n} \cdot \left(\prod_{i=1}^{n} x_i\right) \cdot e^{-\beta \sum_{i=1}^n x_i}$$

Da cui il GLRT è

$$\begin{aligned} \lambda = \frac{\mathrm{L}(\beta_0)}{\mathrm{L}\left(\hat{\beta}\right)} &= \frac{1 \cdot (\prod_{i=1}^n x_i) \cdot e^{-\sum_{i=1}^n x_i}}{\hat{\beta}^{2n} \cdot (\prod_{i=1}^n x_i) \cdot e^{-\hat{\beta} \cdot \sum_{i=1}^n x_i}} = \hat{\beta}^{-2n} \cdot e^{-\sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i} \\ &= \hat{\beta}^{-2n} \cdot e^{-\sum_{i=1}^n x_i (1-\hat{\beta})} \end{aligned}$$

Esempio 13.3.7 (Esame vecchio viroli)**.** Given the system of hypotheses for a Poisson random variable X with parameter θ

$$\begin{cases} H_0 : \theta = 10 \\ H_1 : \theta \neq 10 \end{cases}$$

derive the **wilks statistics**

1. $W = -2n(\hat{\theta} + \theta_0 + \hat{\theta} \log \frac{\theta_0}{\hat{\theta}})$
2. $W = -2n(\hat{\theta} - \theta_0 + \log \frac{\theta_0}{\hat{\theta}})$
3. $W = -2n(\hat{\theta} - \theta_0 + \hat{\theta} \log \frac{\theta_0}{\hat{\theta}})$ credo sia questa
4. $W = 2n(\hat{\theta} - \theta_0 + \hat{\theta} \log \frac{\theta_0}{\hat{\theta}})$

Si ha che per la poisson la density è

$$f(\theta) = e^{-\theta} \cdot \frac{\theta^x}{x!}$$

Per cui la verosimiglianza

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} e^{-\theta} \cdot \frac{\theta^x}{x!} = e^{-n\theta} \cdot \frac{\prod_{i=1}^{n} \theta^{x_i}}{\prod_{i=1}^{n} x_i!} = e^{-n\theta} \cdot \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Il GLRT è

$$\lambda(x) = \frac{\mathrm{L}(\theta_0)}{\mathrm{L}\left(\hat{\theta}\right)} = e^{-\theta_0 n + \hat{\theta} n} \cdot \left(\frac{\theta_0}{\hat{\theta}}\right)^{\sum_{i=1}^{n} x_i}$$

Mentre lo score di wilk

$$\begin{aligned} W = -2 \log \lambda(x) &= -2 \left(\log e^{-\theta_0 n + \hat{\theta} n} + \log \left(\frac{\theta_0}{\hat{\theta}}\right)^{\sum_{i=1}^{n} x_i} \right) \\ &= -2 \left(-\theta_0 n + \hat{\theta} + \underbrace{\sum_{i=1}^{n} x_i}_{n\overline{x}} \cdot \left(\log \frac{\theta_0}{\hat{\theta}} \right) \right) \\ &= -2n \left(\hat{\theta} - \theta_0 + \overline{x} \cdot \log \left(\frac{\theta_0}{\hat{\theta}} \right) \right) \end{aligned}$$

Esempio 13.3.8 (Esame vecchio viroli)**.** Le $X_1, \ldots, X_n$ be a random sample from the density function

$$f(x) = \theta^2 x e^{-\theta x}$$

with $x > 0$ and $\theta > 0$. Find the **Wald test** for the system of hypoteses

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

1. $W_I = \frac{\theta_0^2\left(2n/\theta_0 - \sum_{i=1}^n x_i\right)^2}{2n}$
2. $W_I = \left(\frac{2n}{\sum_{i=1}^n x_i} - \theta_0\right)^2$
3. it cannot be derived
4. $W_I = \frac{\left(\sum_{i=1}^n x_i\right)^2}{2n}\left(\frac{2n}{\sum_{i=1}^n x_i} - \theta_0\right)^2$ dovrebbe essere questa, vedi sotto

Partiamo derivando la verosimiglianza, log verosimiglianza e sue derivate sino alla seconda. Si ha

$$\begin{aligned}
\mathrm{L}(\theta) &= \prod_{i=1}^n \theta^2 x_i e^{-\theta x_i} = \theta^{2n} \cdot \prod_{i=1}^n x_i \cdot e^{-\theta \sum_{i=1}^n x_i} \\
\ell(\theta) &= 2n \log\theta + \sum_{i=1}^n \log(x_i) - \theta \sum_{i=1}^n x_i \\
\ell(\theta)' &= \frac{2n}{\theta} + 0 - \sum_{i=1}^n x_i \\
\ell(\theta)'' &= -\frac{2}{\theta^2}
\end{aligned}$$

Per cui

$$W = \frac{2n}{\hat\theta^2}(\hat\theta - \theta_0)^2$$

Ora per avere/sostituire lo stimatore $\hat\theta$ annulliamo la derivata prima e risolviamo per θ

$$\begin{aligned}
\ell(\theta)' &= 0 \\
\frac{2n}{\theta} - \sum_{i=1}^n x_i &= 0 \\
\hat\theta &= \frac{2n}{\sum_{i=1}^n x_i}
\end{aligned}$$

Quindi per concludere

$$W = 2n \cdot \frac{\left(\sum_{i=1}^n x_i\right)^2}{(2n)^2}\left(\frac{2n}{\sum_{i=1}^n x_i} - \theta_0\right) = \frac{\left(\sum_{i=1}^n x_i\right)^2}{(2n)}\left(\frac{2n}{\sum_{i=1}^n x_i} - \theta_0\right)$$

Esempio 13.3.9 (Esame vecchio viroli)**.** Given the system of hyoptheses for a poisson random variable X with parameter θ

$$\begin{aligned}
H_0 &: \theta = 10 \\
H_1 &: \theta \neq 10
\end{aligned}$$

derive the **score test**

1. $W_s = n\theta_0(1 + \frac{\hat\theta}{\theta_0})^2$

2. $W_s = n\theta_0(1 - \frac{\hat{\theta}}{\theta_0})^2$ risposta corretta (assunta dopo)
3. $W_s = n\hat{\theta}_0(1 - \frac{\hat{\theta}}{\theta_0})^2$
4. $W_s = n\theta_0(1 - \frac{\theta_0}{\hat{\theta}})^2$

Per una variabile poissoniana si ha che

$$\begin{aligned}
\mathrm{L}(\theta) &= \prod_{i=1}^{n} e^{-\theta} \cdot \frac{\theta^{x_i}}{x_i!} = e^{-n\theta} \cdot \frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \\
\ell(\theta) &= -n\theta + \left(\sum_{i=1}^{n} x_i\right) \log\theta - \sum_{i=1}^{n} \log(x_i!) \\
\ell(\theta)' &= -n + \frac{\sum_{i=1}^n x_i}{\theta} \\
\ell(\theta)'' &= -\frac{\sum_{i=1}^n x_i}{\theta^2} \\
i(\theta) &= \frac{\sum_{i=1}^n x_i}{\theta^2} \\
I(\theta) &= \mathbb{E}\left[\frac{\sum_{i=1}^n x_i}{\theta^2}\right] = \frac{\mathbb{E}\left[\sum_{i=1}^n x_i\right]}{\theta^2} = \frac{n\,\mathbb{E}\left[X_i\right]}{\theta^2} = \frac{n}{\theta}
\end{aligned}$$

Poi per il calcolo del test

$$W = \frac{\left[\ell'(\theta_0)\right]^2}{I(\theta_0)}$$

Si ha che

$$\begin{aligned}
\left[\ell'(\theta_0)\right]^2 &= \left[\frac{\sum_{i=1}^n x_i}{\theta_0} - n\right]^2 = \left[n\frac{\frac{\sum_{i=1}^n x_i}{n}}{\theta_0} - n\right]^2 \\
&= \left[n\left(\frac{\hat{\theta}}{\theta_0} - 1\right)\right]^2 = n^2\left(1 - \frac{\hat{\theta}}{\theta_0}\right)^2
\end{aligned}$$

E infine

$$W = \frac{\left[\ell'(\theta_0)\right]^2}{I(\theta_0)} = n^2\left(1 - \frac{\hat{\theta}}{\theta_0}\right)^2 \cdot \frac{\theta_0}{n} = n\theta_0\left(1 - \frac{\hat{\theta}}{\theta_0}\right)^2$$

Esempio 13.3.10 (Esame vecchio viroli)**.** We want to check the system of hypotheses for the parameter θ of a poisson random variable

$$\begin{cases} H_0 : \theta = 10 \\ H_0 : \theta \neq 10 \end{cases}$$

Given the **score test** $W_s = n\theta_0\left(1 - \frac{\hat{\theta}}{\theta_0}\right)^2$ and the observed sample $\{17, 14, 16, 9, 10, 12\}$ compute the p value of the score test in order to decide about the null hypothesis

- 0.62
- 0.05
- 0.02 (taluni suggeriscon questa ma check)
- 0.98

Ricordando che la stima di max verosimiglianza del parametro della poisson è la media campionaria

$$W = n\theta_0\left(1 - \frac{\hat{\theta}}{\theta_0}\right)^2 = 6 \cdot 10 \cdot \left(1 - \frac{\text{sample mean}}{10}\right)^2$$

il valore del test e il p (chi quadrato con 1 grado di libertà) sono

```
(W = 6 * 10 *( (1 - mean(c(17, 14, 16, 9, 10, 12))/10)^2))
```

```
## [1] 5.4
```

```
pchisq(q = W, df = 1, lower.tail = FALSE)
```

```
## [1] 0.02013675
```

Esempio 13.3.11 (Esame vecchio viroli)**.** Let $X_1, \dots, X_n$ be a random sample from the density function

$$f(x) = \frac{2}{\theta} x e^{-x^2/\theta}$$

with $X > 0$ and $\theta > 0$ and $\mathbb{E}[X] = \theta$. Find the **score test** for the system of simple hypotheses

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

under the assumption that the mle for θ is unbiased

1. $W_s = n\frac{(\hat{\theta}-\theta_0)^2}{\theta_0^2}$
2. $W_s = n\frac{\hat{\theta}-\theta_0}{\theta_0^2}$
3. $W_s = n\frac{(\hat{\theta}-\theta_0)^2}{\theta_0}$
4. $W_s = \frac{\theta_0^2}{(\hat{\theta}-\theta_0)^2}$

Il punto a cui sono arrivato

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \frac{2}{\theta} \cdot x_i \cdot e^{-\frac{x_i^2}{\theta}} = \frac{2}{\theta} \cdot \prod_{i=1}^{n} x_i \cdot e^{-\frac{\sum_{i=1}^{n} x_i^2}{\theta}}$$

$$\ell(\theta) = n(\log 2 - \log \theta) + \sum_{i=1}^{n} \log x_i - \frac{\sum_{i=1}^{n} x_i^2}{\theta}$$

$$\ell(\theta)' = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i^2}{\theta^2}$$

$$\ell(\theta)'' = \frac{n}{\theta^2} - 2\frac{\sum_{i=1}^{n} x_i^2}{\theta^3}$$

$$i(\theta) = 2\frac{\sum_{i=1}^{n} x_i^2}{\theta^3} - \frac{n}{\theta^2}$$

Per $I(\theta)$ serve $\mathbb{E}\left[X^2\right]$ come si vedrà poi quindi lo determiniamo

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_0^{+\infty} \frac{2}{\theta} x^3 e^{-x^2/\theta} = -\int x^2 \left(-\frac{2x}{\theta}\right) e^{-x^2/\theta} \\
&= -\left[x^2 e^{-x^2/\theta} - \int 2x \cdot e^{-x^2/\theta}\right] = -\left[x^2 e^{-x^2/\theta} + \theta \int \frac{(-2x)}{\theta} \cdot e^{-x^2/\theta}\right] \\
&= -\left[x^2 e^{-x^2/\theta} + \theta x^2 e^{-x^2/\theta}\right] = \left[-e^{-x^2/\theta}\left[x^2 + \theta\right]\right]_0^{+\infty} = -\theta
\end{aligned}$$

Poi per:

$$\begin{aligned}
I(\theta) &= \mathbb{E}\left[2\frac{\sum_{i=1}^{n} x_i^2}{\theta^3} - \frac{n}{\theta^2}\right] = \frac{2}{\theta^3}\,\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] - \frac{n}{\theta^2} \\
&= \frac{2n}{\theta^3}\,\mathbb{E}\left[X_i^2\right] - \frac{n}{\theta^2} = \frac{2n}{\theta^3}(-\theta) - \frac{n}{\theta^2} = -\frac{2n}{\theta^2} - \frac{n}{\theta^2} = -\frac{3n}{\theta^2}
\end{aligned}$$

Infine

$$\frac{\left[\ell'(\theta_0)\right]^2}{I(\theta_0)} = \left(\frac{\sum_{i=1}^{n} x_i^2}{\theta_0^2} - \frac{n}{\theta_0}\right)^2 \cdot \left(\frac{\theta_0^2}{-3n}\right)$$

Chapter 14

Confidence intervals

Here we switch the focus from point estimation to interval estimation: a set of possible value with a degree of confidence about our estimate. We want to estimate a set of possible values for parameter: a confidence interval.

Esempio 14.0.1. Lets suppose we have $X_i \sim \mathrm{N}\left(\mu, \sigma_0^2\right)$ IID rvs $i = 1, \ldots, n$ with μ unknown and σ_0^2 known. We are interested in estimating μ and accompany that estimate with a range of plausible alternate values.
The log likelihood function is

$$\ell(\mu; \mathbf{x}) = -\frac{n}{2} \log 2\pi\sigma_0^2 - \frac{1}{2} \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_0^2} = \frac{-n\mu^2 + 2\mu \sum_{i=1}^n x_i}{2\sigma_0^2} + c$$

The log likelihood function is depicted in figure 14.1; the parabola is concave with respect to μ taken as incognita (coeff di secondo grado negativo) so there's a maximum. This will be obtained for the MLE estimate $\hat{\mu}$; other than that we are interested in other estimate that are close to ML estimate.

Osservazione importante 118. Interval estimation can provide information about the uncertainty (sample variability) around $\hat{\mu}$.

Definizione 14.0.1 (Confidence interval)**.** It's an interval $[L(\mathbf{x}), U(\mathbf{x})] \subseteq \boldsymbol{\Theta}$ estimate for θ derived from the observed sample $\mathbf{x} = (x_1, \ldots, x_n)$ where, for notation

- we call $L(\mathbf{x}) = \hat{\theta}_L$ the *lower bound*
- we call $U(\mathbf{x}) = \hat{\theta}_U$ the *upper bound*
- $\hat{\theta}_U - \hat{\theta}_L$ is the interval width

Osservazione importante 119. Now the questions are:

- how can we derive $\hat{\theta}_L, \hat{\theta}_U$?
- what is the probability that the true value θ_0 is in $[\hat{\theta}_L, \hat{\theta}_U]$, or differently said, the probability that $[\hat{\theta}_L, \hat{\theta}_U]$ *covers* the true parameter θ_0?

Definizione 14.0.2 (Confidence interval/set)**.** When

$$\mathbb{P}\left(\theta_0 \in [\hat{\theta}_L, \hat{\theta}_U]\right) = 1 - \alpha$$

a confidence interval/set at α

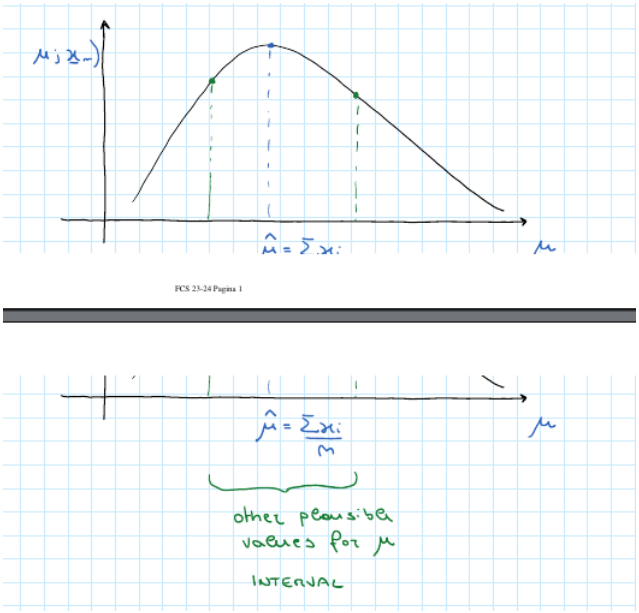


Figure 14.1: Loglikelihood

14.1 Methods of finding interval estimators

14.1.1 Pivotal quantity for θ

Definizione 14.1.1 (Pivotal quantity)**.** A random variable $Q(\mathbf{x}, \theta)$ is a pivotal quantity if its pdf is independent of θ, $\forall \theta \in \mathbf{\Theta}$.

Esempio 14.1.1. Let $X_i \sim \mathrm{N}\left(\mu, \sigma_0^2\right)$ iid rvs with μ unknown and σ_0^2 known. Our sample is $\mathbf{x} = (x_1, \ldots, x_n)$ we have that the sample mean $\overline{x} \sim \mathrm{N}\left(\mu, \sigma_0^2/n\right)$ and that implies that the standardized is independent from μ and therefore pivotal

$$Q(\mathbf{x}, \mu) = \frac{\overline{x} - \mu}{\sigma_0/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

Given $a, b \in \mathbb{R}$ as usual we have that the integral between them is the probability of the variable assuming the value between a and b. Now for a confidence interval we set $a = -z_{\alpha/2}$, $b = z_{\alpha/2}$ so that the area between a and b will be $1 - \alpha$. So we have:

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\overline{x} - \mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$
$$\mathbb{P}\left(-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \overline{x} - \mu \leq z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$
$$\mathbb{P}\left(\overline{x} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$

so in this case

- $\overline{x} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$ is $\hat{\mu}_L$

- $\overline{x} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$ is $\hat{\mu}_U$
- $1-\alpha$ is the probability that $[\hat{\mu}_L, \hat{\mu}_U]$ contains the true value μ_0
- sample size n: the larger n, the lower the width
- value of σ_0^2: the larger σ_0^2, the larger the width
- confidence level $(1-\alpha)$, the larger $(1-\alpha)$ the larger the width

Esempio 14.1.2. Let $X_i \sim \text{Exp}(\lambda)$ be $i = 1, \dots, n$ iid rvs. Let's find a confidence set for λ.
For the point estimation we have that:

$$\begin{aligned}
\text{L}(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\sum_{i=1}^{n} x_i \lambda\right) \\
\ell(\lambda) &= n \log \lambda - \lambda \sum_{i=1}^{n} x_i \\
\frac{\partial \ell(\lambda)}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \\
\hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned}$$

Now we know that

- if $X_i \sim \text{Exp}(\lambda)$ then $\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda)$ (shape and rate formulation);
- if $X \sim \text{Gamma}(\alpha, \beta)$, then $cX \sim \text{Gamma}(\alpha, \beta/c)$.

Therefore we have that

$$\begin{aligned}
&\frac{\sum_{i=1}^{n} x_i}{n} \sim \text{Gamma}(n, n\lambda) \\
&\underbrace{\frac{\lambda \sum_{i=1}^{n} x_i}{n}}_{\text{pivotal}} \sim \text{Gamma}(n, n) \\
&\underbrace{\lambda \sum_{i=1}^{n} x_i}_{\text{pivotal}} \sim \text{Gamma}(n, 1)
\end{aligned}$$

Then the confidence level will be derived choosing (see fig 14.2) a and b from the theoretical distribution to ensure $1-\alpha$ probability is between them, and developing as follows

$$\begin{aligned}
&\mathbb{P}\left(a \leq \lambda \sum_{i=1}^{n} x_i \leq b\right) = 1-\alpha \\
&\mathbb{P}\left(\frac{a}{\sum_{i=1}^{n} x_i} \leq \lambda \leq \frac{b}{\sum_{i=1}^{n} x_i}\right) = 1-\alpha
\end{aligned}$$

so it will be $\frac{a}{\sum_{i=1}^{n} x_i}$ the lower margin and $\frac{b}{\sum_{i=1}^{n} x_i}$ the upper one.

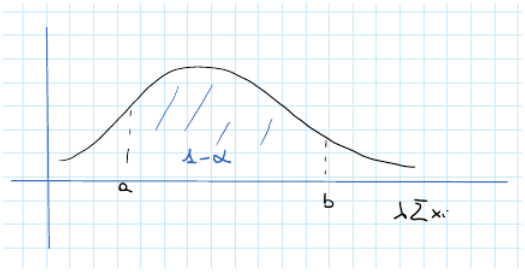


Figure 14.2: Confint gamma

14.1.2 Asymptotic confidence intervals

NB: vaneggiamenti vari sull'applicabilità anche se non cè media a 0:33 del 21/11

Proposizione 14.1.1. *Imagine that:*

- *the sampling is iid;*
- *the maximum likelihood estimator for* θ *is* $\sum_{i=1}^{n} x_i$ *or* $\overline{x}$ *(many estimator have such shape);*
- *assume that* $\overline{x}$ *is* unbiased.

Then the standardized version is asymptotic pivot *(parameters free when* $n \to \infty$*) since:*

$$\frac{\overline{x} - \theta}{\frac{\hat{\sigma}(\mathbf{x})}{\sqrt{n}}} \xrightarrow{d} \mathrm{N}(0,1)$$

Esempio 14.1.3 (Bernoulli)**.** Let $X_i \sim \mathrm{Bern}(\theta)$ be iid. We have that the estimator for θ is currently a mean

$$\begin{aligned}
\mathrm{L}(\theta) &= \theta^{\sum_{i=1}^{n} x_i} \cdot (1-\theta)^{n - \sum_{i=1}^{n} x_i} \\
\ell(\theta) &= \sum_{i=1}^{n} x_i \log\theta + (n - \sum_{i=1}^{n} x_i) \log(1-\theta) \\
\hat{\theta}_{ML} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned}$$

Remembering that for the Bernoulli

$$\begin{aligned}
\mathrm{Var}[X_i] &= \theta(1-\theta) \\
\mathrm{Var}[\overline{x}] &= \frac{\theta(1-\theta)}{n}
\end{aligned}$$

if we standardize, we have that for CLT:

$$\frac{\overline{x} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} = \sqrt{\frac{n}{\theta(1-\theta)}} \cdot (\overline{x} - \theta) \xrightarrow{d} \mathrm{N}(0,1)$$

Now notice that $\frac{n}{\theta(1-\theta)} = I(\theta)$. This because of

$$\ell'(\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}$$
$$\ell''(\theta) = -\frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2}$$
$$I(\theta) = \mathbb{E}\left[-\ell''(\theta)\right]$$

We furthermore have that:

$$i(\hat{\theta}) = -\ell''(\theta)\big|_{\theta=\hat{\theta}}$$

Now we want to show that

$$i(\hat{\theta}) \xrightarrow{d} I(\theta)$$

that is the expected information is the convergence point of the observed information. This is a general properties which holds, we verify it here for the example at hand.
Let's start by the observed information:

$$i(\hat{\theta}) = -\ell''(\hat{\theta}) = \frac{\sum_{i=1}^n x_i}{\hat{\theta}^2} + \frac{n - \sum_{i=1}^n x_i}{(1-\hat{\theta})^2} \overset{(1)}{=} n \cdot \left(\frac{\hat{\theta}}{\hat{\theta}^2} + \frac{1-\hat{\theta}}{(1-\hat{\theta})^2}\right)$$
$$= n\left(\frac{1}{\hat{\theta}} + \frac{1}{(1-\hat{\theta})}\right) = \frac{n}{\hat{\theta}(1-\hat{\theta})}$$

where in (1) we substituted remembering that $\hat{\theta} = \sum_{i=1}^n x_i/n$. The latter is similar to the variance we used to standardize the mean (with $\hat{\theta}$ instead of θ).
So we have checked somehow that $i(\hat{\theta}) \xrightarrow{d} I(\theta)$: now this holds in general and is useful because in order to construct our interval, we start from a standardization that assure us the gaussianity; in the standardization we can replace the variance computed with an estimate for the parameter ($\hat{\theta}$) with the unknown parameter (θ). We standardize with what we have in the sample.
Therefore we go with

$$\sqrt{\frac{n}{\overline{x}(1-\overline{x})}} \cdot (\overline{x} - \theta) \xrightarrow{d} \mathrm{N}(0,1)$$

with $\frac{n}{\overline{x}(1-\overline{x})} = i(\theta)$. Finally to construct the interval when n increases

$$\lim_{n\to\infty} \mathbb{P}\left(-z_{\alpha/2} \leq \sqrt{\frac{n}{\overline{x}(1-\overline{x})}} \cdot (\overline{x} - \theta) \leq z_{\alpha/2}\right) = 1 - \alpha$$

so we can conclude that when n is large

$$\mathbb{P}\left(\overline{x} - z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} \leq \theta \leq \overline{x} + z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right) = 1 - \alpha$$

and $\overline{x} - z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$ is the lower limit while $\overline{x} + z_{\alpha/2}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$ the upper one.

14.1.3 Wald asymptotic confidence intervals

Osservazione 342. We construct an asymptotic confidence interval based on the asymptotic wald test. This is one of the most used for simplicity.
This is a general procedure for all the MLEstimators. It does not need the estimator be expressed in terms of sum or mean as above

Osservazione importante 120 (Idea of the method). If we remember, under the regularity conditions of Cramer-Rao inequality, the MLE are asymptotically gaussian distributed (they are BAN - best asymptotic gaussian). This fact can be used to construct intervals.

For all the probabilistic models, provided we have a maximum likelihood estimator (even if it's obtained not in closed form) we are sure that the square root of the observed information (which converges to the expected information) times the the difference between MLEstimate and theta (its' a standardization as above men) converges to a standard gaussian:

$$\sqrt{i(\hat{\theta}_{ML})} \cdot (\hat{\theta}_{ML} - \theta) \xrightarrow{d} \mathrm{N}(0,1), \quad \forall \theta \in \boldsymbol{\Theta}$$

This is a very general result that can be used to costruct a confidence interval; therefore to do it

$$\lim_{n\to\infty} \mathbb{P}\left(-z_{\alpha/2} \leq \sqrt{i(\hat{\theta}_{ML})} \cdot (\hat{\theta}_{ML} - \theta) \leq z_{\alpha/2}\right) = 1 - \alpha$$

and when as n gets larger we get the confidence interval by isolating theta:

$$\mathbb{P}\left(\hat{\theta}_{ML} - z_{\alpha/2} \cdot \frac{1}{\sqrt{i(\hat{\theta}_{ML})}} \leq \theta \leq \hat{\theta}_{ML} + z_{\alpha/2} \cdot \frac{1}{\sqrt{i(\hat{\theta}_{ML})}}\right) = 1 - \alpha$$

14.1.4 Exercises on confidence intervals

Esempio 14.1.4. Let $X_i \sim \mathrm{Exp}(\theta)$ be iid rvs. We have that maximum likelihood estimator for θ $\hat{\theta}_{ML} = \frac{1}{\overline{x}}$.
We could use the second method (asymptotic confidence interval) to find a confidence interval for $\frac{1}{\theta}$ or we could use method three (wald) to find a confint directly for theta.
According to Wald what we need is the observed information; for the expo distribution we derived it many times and it's $i(\hat{\theta}) = \frac{n}{\hat{\theta}^2} = n\overline{x}^2$. So just according to these information we can say that:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{\overline{x}} - z_{\alpha/2}\frac{1}{\sqrt{n\overline{x}}} \leq \theta \leq \frac{1}{\overline{x}} + z_{\alpha/2}\frac{1}{\sqrt{n\overline{x}}}\right) = 1 - \alpha$$

Esempio 14.1.5. Let $X_1, \ldots, X_n$ iid rvs from Unif $(0, \theta)$. We know (did in past exercises) that an estimator for θ is the maximum $X_{(n)} = \max\{X_i\}_{i=1,\ldots,n}$:

1. starting from the distribution of $X_{(n)}$ try to find a pivotal quantity for θ;
2. compute the probability that θ is between the maximum and the maximum times 1.1:

   $$\mathbb{P}\left(X_{(n)} \leq \theta \leq 1.1X_{(n)}\right)$$

   for $n = 30$

We have that

1. we remember that the density that $f_{(n)} = n \cdot x^{n-1} \cdot \frac{1}{\theta^n}$ (from previous notes) and notice that if the maximum resides between 0 and θ, $0 \leq x_{(n)} \leq \theta$, then $0 \leq \frac{x_{(n)}}{\theta} \leq 1$.
   Can we get the distribution of $\frac{x_{(n)}}{\theta}$? because in this manner we remove the effect of theta (having a pivot). Then what is the pdf of $Y = \frac{X_{(n)}}{\theta}$? Can Y be a pivot?
   In order to find it we can apply the formula for transformation:

$$f_Y(y) = \left| \frac{\partial g^{-1}(y)}{\partial y} \right| \cdot f_X(g^{-1}(y))$$

   if $Y = g(X) = \frac{X_{(n)}}{\theta}$ then the inverse transformation $g^{-1}(y) = \theta Y = X_{(n)}$ and the first derivative is $\frac{\partial g^{-1}(y)}{\partial y} = \theta$. So applying the formula to get the density, it is:

$$f_Y(y) = \theta \cdot n \cdot (\theta y)^{n-1} \frac{1}{\theta^n} = n y^{n-1}$$

   where we avoided the absolute value since $\theta > 0$. So the density doesn't depending anymore on θ, so $Y = \frac{X_{(n)}}{\theta}$ is a pivotal quantity/rv. We use this to solve the second point.

2. in order to solve, we try to get $\frac{X_{(n)}}{\theta}$ inside to use the pivot stuff:

$$\begin{aligned}
\mathbb{P}\left(X_{(n)} \leq \theta \leq 1.1 X_{(n)}\right) &= \mathbb{P}\left(1 \leq \frac{\theta}{X_{(n)}} \leq 1.1\right) \\
&= \mathbb{P}\left(\frac{1}{1.1} \leq \frac{X_{(n)}}{\theta} \leq \frac{1}{1}\right) = \mathbb{P}\left(0.9091 \leq Y \leq 1\right) \\
&= \int_{0.9091}^{1} n y^{n-1} \, \mathrm{d}y = [y^n]_{0.9091}^{1} = 1 - (0.9091)^3 0 \\
&= 0.9427
\end{aligned}$$

Esempio 14.1.6 (Assignment 4 Viroli, Exercise 2 (hypothesis testing and confidence intervals))**.** Let $(Y_1, x_1), \ldots, (Y_n, x_n)$ be the data, where $\{Y_i\}_i$ are independently and normally distributed random variables in the following way

$$Y_i \sim \mathrm{N}\left(\theta x_i, 1\right), \quad i = 1, \ldots, n$$

and where $\{x_i\}_i$ are known constants (i.e., non-random).

1. Show that the maximum likelihood estimator $\hat{\theta}$ of θ is given by

$$\hat{\theta} = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2$$

2. Find a 95% confidence interval for θ based on inverting the test statistic $\hat{\theta}$ (hint: see section 9.2.1. Casella Berger)

3. We are interested in testing the following hypotheses:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq theta_0 \end{cases}$$

We use the log likelihood ratio statistic as test-statistic.

- derive and simplify

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} \mathrm{L}(\theta)}{\sup_{\theta \in \Theta} \mathrm{L}(\theta)}$$

where L is the likelihood of θ

- determine the rejection region $[0; c]$ associated with λ for a test with significance level α.

4. Galton in 1885 studied the relationship between the heights of parents (x_i) and their (adult) children (Y_i). We present the data (i = 1,...,500) in the figure below, including the line $y = x$. We have adjusted the unit scale slightly (of both parental and children heights) so that $Var(Y_i) = 1$ (immagine omessa).
The question that Galton wanted to answer was whether $\theta = 1$. use the fact that

$$\sum_{i=1}^{500} x_i y_i = 372628$$
$$\sum_{i=1}^{500} x_i^2 = 371321$$

- Determine whether Galton would reject the null hypothesis based on the likelihood ratio test at significance level $\alpha = 0.05$. [Hint: use (3)]
- Determine a 95% confidence interval based on the MLE of θ. [Hint: use (d)]

Respectively:

1. the likelihood is

$$\mathrm{L}\left(\theta; \mathbf{y}, \mathbf{x}, \sigma^2 = 1\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - x_i\theta)^2}$$

the loglikelihood

$$\begin{aligned} \ell\left(\theta; \mathbf{y}, \mathbf{x}, \sigma^2 = 1\right) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - x_i\theta)^2} \\ &= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_i - x_i\theta)^2 \right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\theta)^2 \end{aligned}$$

its first derivative with respect to θ

$$\frac{\partial \ell\left(\theta; \mathbf{y}, \mathbf{x}, \sigma^2 = 1\right)}{\partial \theta} = 0 - \frac{1}{2}\sum_{i=1}^{n}\left[2(y_i - x_i\theta)(-x_i)\right] = \sum_{i=1}^{n}\left[(y_i - x_i\theta)(x_i)\right]$$
$$= \sum_{i=1}^{n} x_i y_i - \theta \sum_{i=1}^{n} x_i^2$$

And by equating to 0 we find the MLE as

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

2. the MLE estimator $\hat{\theta}$ is asymptotically gaussian by properties of ML estimators and precisely

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2(\theta)\right)$$

where $\sigma^2(\theta) = \frac{1}{I(\theta)}$. Therefore we have that

$$\frac{\hat{\theta} - \theta}{\frac{\sigma(\theta)}{\sqrt{n}}} \xrightarrow{d} \mathrm{N}\left(0, 1\right)$$

with $\sigma(\theta) = \frac{1}{\sqrt{I(\theta)}}$. We thus need $I(\theta)$

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[-\frac{\partial^2 \ell\left(\theta; \mathbf{y}, \mathbf{x}, \sigma^2 = 1\right)}{\partial^2 \theta}\right]$$

we have

$$\frac{\partial^2 \ell\left(\theta; \mathbf{y}, \mathbf{x}, \sigma^2 = 1\right)}{\partial^2 \theta} = \left(\sum_{i=1}^{n} x_i y_i - \theta \sum_{i=1}^{n} x_i^2\right)' = -\sum_{i=1}^{n} x_i^2$$

which is a constant. So

$$I(\theta) = \mathbb{E}\left[i(\theta)\right] = \mathbb{E}\left[-\frac{\partial^2 \ell\left(\theta; \mathbf{y}, \mathbf{x}, \sigma^2 = 1\right)}{\partial^2 \theta}\right] = \mathbb{E}\left[\sum_{i=1}^{n} x_i^2\right] = \sum_{i=1}^{n} x_i^2$$

and

$$\sigma(\theta) = \frac{1}{\sqrt{I(\theta)}} = \frac{1}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Finally

$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\frac{1/\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$\dots$$

$$\mathbb{P}\left(\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{n}\sqrt{\sum_{i=1}^{n} x_i^2}} < \theta < \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{n}\sqrt{\sum_{i=1}^{n} x_i^2}}\right) = 1 - \alpha \tag{14.1}$$

3. respectively:

   - to derive and simplify the likelihood ratio test we have that

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\theta\in\Theta_0} \mathrm{L}(\theta)}{\sup_{\theta\in\boldsymbol{\Theta}} \mathrm{L}(\theta)} = \frac{\mathrm{L}(\theta_0)}{\mathrm{L}\left(\hat{\theta}\right)} \\
&= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - x_i\theta_0)^2}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - x_i\hat{\theta})^2}} = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\theta_0)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\hat{\theta})^2}} \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\theta_0)^2 + \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\hat{\theta})^2\right) \\
&= \exp\left[\frac{1}{2}\left(\sum_{i=1}^{n}(y_i^2 + x_i^2\hat{\theta}^2 - 2x_iy_i\hat{\theta}) - \sum_{i=1}^{n}(y_i^2 + x_i^2\theta_0^2 - 2x_iy_i\theta_0)\right)\right] \\
&= \dots \\
&= \exp\left(\frac{(\hat{\theta}^2 - \theta_0^2)\left(\sum_{i=1}^{n} x_i^2\right)}{2} + \left(\theta_0 - \hat{\theta}\right)\sum_{i=1}^{n} x_iy_i\right) \qquad (14.2)
\end{aligned}$$

   - to determine the rejection region $[0; c]$ for $\lambda(\mathbf{x})$ with significance level α, we can use Wilks theorem. In our case, for a system of hypotheses such as:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

     under the null hypothesis H_0, for $n \to \infty$:

$$W = -2\log\lambda(\mathbf{x}) \xrightarrow{d} \chi^2(1)$$

     so we'll need to compare $-2\log\lambda(\mathbf{x})$ with the $1-\alpha$ percentile (q_α) obtained using `qchisq`. If $-2\log\lambda(\mathbf{x}) > q_\alpha$ we reject the null. Eg for 1 degree of freedom and $\alpha = 0.05$ we have that

$$q_\alpha = \texttt{qchisq}(\texttt{0.95}, \texttt{df} = \texttt{1}) = 3.84$$

     If for some reason we want to have a rule to reject on the $\lambda(\mathbf{x})$ unit/scale, as the text seems to suggest, then we have to translate the rejection rule there:

$$\begin{aligned}
-2\log(\lambda(\mathbf{x})) &> q_\alpha \\
\log(\lambda(\mathbf{x})) &< \frac{q_\alpha}{-2} \\
\lambda(\mathbf{x}) &< \exp\left(-\frac{q_\alpha}{2}\right)
\end{aligned}$$

     so the c requested by the text is

$$c = \exp\left(-\frac{q_\alpha}{2}\right)$$

     and being $\lambda(\mathbf{x}) \geq 0$, the rejection region will be $[0, \exp\left(-\frac{q_\alpha}{2}\right))$.

4. we have that the maximum likelihood estimate for θ is

$$\hat{\theta} = \frac{\sum_{i=1}^{500} x_i y_i}{\sum_{i=1}^{500} x_i^2} = \frac{372628}{371321} = 1.0035199$$

so descriptively seems that mean height tends to increase with $+0.35\%$ after a generation. Now:

- for testing the value, by substituting all the components in equation 14.2 we have

```
sumxiyi = 372628
sumxi2 = 371321
hat_theta = sumxiyi/sumxi2
theta_0 = 1
lambda_x = exp(
(hat_theta^2 - theta_0^2)*sumxi2/2 +
(theta_0 - hat_theta)*sumxiyi
)
test = -2*log(lambda_x)

## results
lambda_x
```

```
## [1] 0.1002356
```

```
test
```

```
## [1] 4.600464
```

Therefore we have that

$$\lambda(\mathbf{x}) = 0.1002356 \qquad W = -2\log\lambda(\mathbf{x}) = 4.6004643$$

and being $W > 3.84$ we reject the $H_0 : \theta = 1$ (so there's evidence of increased mean height by generation in generation).
Same results are obtained using the cutoff on the $\lambda(\mathbf{x})$ scale being

$$0.10 < \exp\left(-\frac{3.84}{2}\right) = 0.1466$$

- finally for the confidence interval we substitute some values in equation 14.1:

```
alpha = 0.05
n = 500
z_alpha2 = qnorm(1 - alpha/2)
half_width <- z_alpha2 * (1/(sqrt(n) * sqrt(sumxi2)))
ci <- c(hat_theta - half_width, hat_theta + half_width)
ci
```

```
## [1] 1.003376 1.003664
```

So the 95% confidence interval is 1.003376 - 1.0036637 and coherently with the test it does not include values below 1.

Chapter 15

Multiple testing

15.1 Introduction

Topic particular useful for big data: sometimes we need to perform several test (eg in a regression model test on several predictors)

We want a procedure for multiple testing, we coauld repeat the test many times but there are problems with the control of first type error

in each row we have a gene and we measure the expression the gene in several individuals. In this kind of unit we change, in this cas we have patients in column and variable in row

In this kind of application several observation for few subject. So in this type of data p (number of variables, up to G_p) is definitely larger than n (number of patients). In this application patients were $n = 6$ and genes were $p = 2000$

when $p \gg n$ we have a BIG DATA baby

In the example, the aim is fo find/identify the genes with a different expression level between healthy and tumor tissues/patients (to find genetic predictors useful for clinics).

So formally it would be a system of 2000 null hypotheses where the single is formulated like this:

$$H_0 : \mu_i^H = \mu_i^T, \quad i = 1, \dots, p$$

we measure averages and compare the two means for the i-th gene. If we refuse we identify different expression. We have to repeat the same testing procedure many many times, so this is a typical situation of multiple testing.

Osservazione 343. Basically the H_0 here is that all the genes are equally expressed.

Osservazione importante 121. The problem with this approach is the inflation of the type I error.

We define a new quantity called family-wise error rate (FWER):

Osservazione importante 122. It's the probability of having at least one rejection under hypotheses of 2000 null hypotheses

$$\overline{\alpha} = \mathbb{P}\left(\text{at least one rejection} | H_0\right)$$

Ideally we would like that $\overline{\alpha} \approx \alpha$, that is the FWER is equal to the first type error we can fix in advance. But in developing it we find that:

$$\begin{aligned}
\overline{\alpha} &= \mathbb{P}\left(\text{at least one rejection} | H_0\right) \\
&= 1 - \mathbb{P}\left(\text{no rejections} | H_0\right) \\
&= 1 - \mathbb{P}\left(\overline{R}^1 \cap \ldots \cap \overline{R}^p | H_0\right) \\
&\overset{(\perp\!\!\!\perp)}{=} 1 - \prod_{i=1}^{p} \mathbb{P}\left(\overline{R}^i | H_0\right) \\
&= 1 - \mathbb{P}\left(\overline{R}^i | H_0\right)^p \\
&= 1 - (1-\alpha)^p
\end{aligned}$$

were we assumed independent experiments regarding different genes. Therefore it's not equal to α but is related to it and number of tests: if $\alpha = 0.05$ and $p = 20$ then $\overline{\alpha} = 0.64$ which is very high/severely inflated error. If p increase then $\overline{\alpha}$ with it.

We found some manner to keep the FWER very low, at the level we want

15.2 Methods

15.2.1 Bonferroni correction

Definizione 15.2.1 (Bonferroni procedure)**.** When one have to perform p test we use not the standard α to reject all the nulls, we reject using a corrected alpha α_i^*

$$\alpha_i^* = \frac{\alpha_i}{p}, i = 1, \ldots, p$$

The effect is that in this manner the FWER corrected according to bonferroni $\overline{\alpha}^B$ becomes

$$\overline{\alpha}^B = 1 - \left(1 - \frac{\alpha_i}{p}\right)^p$$

The second quantity has a very precise limit. If we fix as often is a common $\alpha_1 = \ldots = \alpha_p = \alpha$ we have that

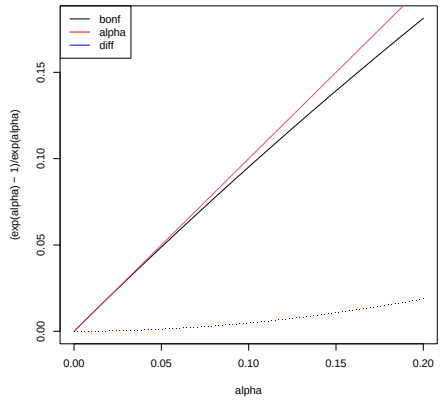
$$\lim_{p \to \infty} \left(1 - \frac{\alpha}{p}\right)^p = e^{-\alpha}$$

and so

$$\overline{\alpha}^B = 1 - \frac{1}{e^\alpha} = \frac{e^\alpha - 1}{e^\alpha} \approx \alpha, \quad \text{if } 0 < \alpha < 0.10$$

as plotted in **??**.

```
alpha <- seq(0, 0.2, length.out = 100)
plot(alpha, (exp(alpha) - 1)/exp(alpha), type = 'l', col = 'black')
lines(alpha, alpha, col = 2)
points(alpha, alpha - (exp(alpha) - 1)/exp(alpha), col = 'blue', pch='.')
legend("topleft", col = c("black", "red", "blue"), legend = c("bonf", "alpha", "diff"
```



Osservazione 344. An important property is that the correction according to bonferroni assures that the alpha corrected is always below the nominal (as we have seen in the plot as well)

Proposizione 15.2.1 (Properties)**.** *We have that* $\overline{\alpha}^B \leq \alpha$

Proof.

$$\overline{\alpha}^B \overset{(1)}{=} \mathbb{P}\left(\bigcup_{i=1}^{p} R^{(i)}|H_0\right) \leq \sum_{i=1}^{p} \mathbb{P}\left(R^{(i)}|H_0\right) = \sum_{i=1}^{p} \alpha/p = \alpha$$

(1) the fwer is the probability of the union event of rejections (it's the event "at least one rejection"), which turns out to be less than or equal to the sum of single probabilities. □

Osservazione importante 123. Problems of bonferroni:

- it controls false positives only (not the false negatives)
- it increases the probability of false negatives $\implies$ reduced POWER
- it's too conservative! we don't reject

15.2.2 Sidak correction

Not very used but better than Bonferroni

Definizione 15.2.2. Starting from the equation of FWER $\overline{\alpha}$ we want to choose α^S to make the first exactly 0.05 (let's say).

$$\begin{aligned} \overline{\alpha} &= 1 - (1 - \alpha^s)^p \\ (1 - \alpha^s)^p &= 1 - \overline{\alpha} \\ 1 - \alpha^s &= (1 - \overline{\alpha})^{1/p} \\ \alpha^s &= 1 - (1 - \overline{\alpha})^{1/p} \end{aligned}$$

Esempio 15.2.1 (Comparison with bonferroni)**.** Fix FWER as we want $\overline{\alpha} = 0.05$ and $p = 20$ then:

$$\alpha^B = \frac{0.05}{20} = 0.025$$
$$\alpha^S = 1 - (1 - 0.05)^{1/20} = 0.0256$$
$$\alpha^B \leq \alpha^S$$

So in general:

- Sidak is larger and less conservative
- assure us to have the level we want

15.2.3 FDR

Osservazione importante 124. In DNA experiments (and other contexts) we care more on β rather than α because controlling the false negatives is more important: if a gene is important and we don't detect it is more severe than declaring a unuseful gene as useful.
Our attention should be focused on the percentage of false significant genes among all the significant outcomes (FALSE NEGATIVE). It's quite complicate so we will see the procedure instead of the mechanism

Definizione 15.2.3 (False discovery rate)**.** It's defined as

$$FDR = \frac{\text{n of false rejections}}{\text{n of rejections}}$$

Osservazione 345. Ofcourse it's a math/theoretical quantity; in reality we don't know if a rejetion is false or true, we just know the number of rejection

Looking at table 15.1:

- let U, V, T, S be the number of times we find in each situation;
- we have that $U+V+T+S = p$ (which is not a percentage but the number of test performed)
- we define p_0 as the number of null hypotheses that are true (variables not relevant)
- and p_1 those that are false (number of relevant variables).
- we don't know the rows totals, only the columns totals

According to this table we can define

- the previous quantity, family-wise error rate, probability to reject at least one time given the null:

$$FWER = \mathbb{P}(V \geq 1)$$

	H_0 not rejected	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	$p-R$	R	p

Table 15.1: FWER tab

- we can define **False Discovery Proportion** is

$$FDP = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

(is not observed since we don't know the rows totals, only the columns totals).

- the **false discovery rate** is again a theoretical quantity and it's

$$FDR = \mathbb{E}[FDP]$$

we want to keep the theoretical quantity FDR as low as possible

Osservazione 346. In practical situation we cannot compute the FDR (dont know the false discovery proportion and cannot compute its expectation) but we have some complicate procedure (the article originating this stuff is very complicate) in order to estimate it.

Now:

- if *all the nulls are true*, $S = 0$ and $V = R$, therefore FDP becomes

$$FDP = \begin{cases} 1 & \text{if } R > 0 \ (V > 0) \\ 0 & \text{if } R = 0 \ (V = 0) \end{cases}$$

and

$$\begin{aligned} FDR &= \mathbb{E}[FDP] \\ &= \mathbb{P}(R > 0) \cdot \mathbb{E}\left[\frac{V}{R} | R > 0\right] + \mathbb{P}(R = 0) \cdot \underbrace{\mathbb{E}\left[\frac{V}{R} | R = 0\right]}_{=0} \\ &= \mathbb{P}(R > 0) \cdot \mathbb{E}\left[\frac{V}{R} | R > 0\right] \\ &= \mathbb{P}(V \geq 1) \cdot 1 \\ &= FWER = \overline{\alpha} \end{aligned}$$

so if all the nulls are true the $FDR = FWER$;

- if *not all the nulls are true*, then $V < R$ and:

$$\begin{aligned} FDR &= \mathbb{E}[FDP] \\ &= \mathbb{P}(R > 0) \cdot \mathbb{E}\left[\frac{V}{R} | R > 0\right] \\ &= \mathbb{P}(V \geq 0) \cdot \mathbb{E}\left[\frac{V}{R} | V \geq 0\right] \end{aligned}$$

but when $V = 0$ the right quantity is zero and thus we can write

$$FDR = \mathbb{E}[FDP] = \mathbb{E}\left[\frac{V}{R}|V \geq 1\right] \cdot \mathbb{P}(V \geq 1)$$
$$< FWER$$

so if not all the nulls are true the $FDR < FWER$ because $V < R$.

Osservazione 347. According to this quantity Benjamini and Hockberg developed a procedure able to estimate the quantities in order to improve the power of multiple testing and at the same time to control the first type error.

15.2.4 Benjamini and Hockberg (1995)

The steps are:

1. Perform a test on each variable ($i = 1, \ldots, p$)
2. we reorder the variables according to the p-values to ν_i we get in each single test so that the lower p-value are first, the larger are latter:

$$\nu_1 < \ldots < \nu_p$$

3. we define

$$I_i = \frac{i \cdot \alpha}{C_p \cdot p}$$

where

- i refers to the number in the ordered variables sequence
- $C_p = 1$ if the hypotheses are independent (most of times it is, so often it will be $I_i = i \cdot \alpha/p$) or $C_p = \sum_{i=1}^{p}(1/i)$ otherwise (if hypotheses are somewhat dependent we use this)

4. we define $R = \max\{i : \nu_i < l_i\}$ (that is the number of variable for which we reject)
5. let the BH rejection threshold be the largest p-value among the rejected ones (se ho ben capito), that is the rejection tresh is $t = \nu_R$
6. reject all the null hypotheses for which $\nu_i \leq t$

Teorema 15.2.2 (Important result)**.** *If we apply the procedure above, the false discovery rate we get is contained below α since*

$$FDR = \mathbb{E}[FDP] = \frac{p_0}{p}\alpha \leq \alpha \tag{15.1}$$

and so we keep under control the number of false rejections.

Esempio 15.2.2. Suppose $p = 6$ and $\nu_1 = 0.0012$, $\nu_2 = 0.0046$, $\nu_3 = 0.0084$, $\nu_4 = 0.0136$, $\nu_5 = 0.0466$, $\nu_6 = 0.2319$ with independent hypotheses and $\alpha = 0.05$.

- no correction: we reject the first 5 hypotheses
- Bonferroni: $\alpha^B = 0.05/6 = 0.00833$, we reject only the first 2 hypotheses
- Sidak $\alpha^S = 1 - (1 - 0.05)^{1/6} = 0.0085$ we reject the first 3 hypotheses
- BH procedure: $l_i = i \cdot \alpha / p$. We have $l_1 = 1 \cdot 0.05/6 = 0.00833$, $l_2 = 0.0167$, $l_3 = 0.0250$, $l_4 = 0.0333$, $l_5 = 0.0416$ and $l_6 = 0.05$ from which $R = 4$ and $t = 0.0136$. We reject the first 4 hypotheses

```
# bh (already ordered)
pis <- c(0.0012, 0.0046, 0.0084, 0.0136, 0.0466, 0.2319)
(i <- rank(pis))

## [1] 1 2 3 4 5 6

alpha <- 0.05
(l <- i*alpha/6)

## [1] 0.008333333 0.016666667 0.025000000 0.033333333 0.041666667 0.050000000

(R <- sum(pis < l))

## [1] 4

(t <- pis[R])

## [1] 0.0136

(pis <= t)

## [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE
```

This example show how BH procedure is a kind of tradeoff between a very strong correction according to bonferroni, or even milder correction according to Sidak (which improves the first type error but decreases again the power as in bonf); BH is an intermediate way between those two and the situation of no correction. In situation where p is larger than 6 it works very well.

15.2.5 q-values (Storey, 2002)

Another tool with the aim to control both types of error (α, β). The idea of q-values is to adjust p-values in order to keep under control FDR instead of FWER. Again a complicate procedure.

What is important to know is that given this quantity

$$pFDR = \mathbb{E}\left[\frac{V}{R}|R > 0\right]$$

then the:

- p-value: is minimum probability under the null that $T_n \in R$ (our statistics belong to the rejection area)
- q-value: is the greatest lower bound of the $pFDR$ when $T_n \in R$ (we reject)

These are very complicate, no time to develop, but there are many R packages that gives q-values. Instead of comparing p-values with α we compare q-values.

Chapter 16

Non parametric inference

Osservazione 348. So far we ve seen methods two methods for testing hypotheses: the Fisher approach (we mentioned in an example) and more deeply the Neyman-Pearson approach.
If we work in a bayesian framework we have other tools that will be developed in other courses.
In both frequentist and bayesian framework we base our theory on a specific probabilistic distribution: we assume X is distributed according a distribution.
Another approach is nonparametric: the idea was to develop a procedure which were not based on probabilistic distribution/formulation. We see the most important tools here.

16.1 Intro

Osservazione 349. **Why non parametric tests**?

- Imagine we do not have/know the probabilistic model for our data (if we assume probabilistic model we make a restriction to a specific case);
- The aim is to check an assumption on the shape of the distribution, which is not expressed as a parameter;
- we can apply these strategy to ordinal and categorical variables or ranks: for this kind of data is more difficult to find a distribution, so we don't have a probabilistic model
- this approach is a faster alternative to nonparametric bootstrap (which will see)

When using them?

- the null hypothesis is not about a parameter;
- data are expressed in different forms than those required by a parametric test;
- the distributional form of data is not known or data are low-dimensional that n is small (asymptotic results can not be used). EG Wilks test works

if $n \to \infty$; it will be distributed according to a chi square. But if n is small we don't have a distribution, so we can use it when asymptotic results can be used.

However some limitations are:

- using nonparametric techniques when you can use parametric ones you have a loss of information;
- for large samples, some nonparametric techniques can be computational intensive.

16.2 Sign test (one sample)

Osservazione 350. Let's start from the simple one: it's called sign test (for one sample) because it's based on the number of + and -.
Let's start with an example.

Esempio 16.2.1. A credit institution wants to open a new branch in a suburb of a big city; however a necessary condition is that the median income of the local families has to be at least equal to 25000 euros.
In order to verify this condition, the income of a sample of $n = 15$ families has been examined, yielding the following results (in k/thousands of euros):

$$24, 26, 32, 40, 22, 25, 30, 44, 21, 18, 26, 27, 28, 28, 27$$

Do the data suggest the opening of the new branch? ($\alpha = 0.01$)
We answer this question by a sign test.

- The **data** are: n observations, *at least ordinal* (not necessarily numerical, but we should be able to order the data from the smallest to the largest), that corresponds to the n sample units; let $X_1, \ldots, X_n$ be the random variables that describe the sampling of these units from the population (we don't make any assumption of the distribution of X_i).
- **Assumptions**: the rvs $X_j, i = 1, \ldots n$ are independent (not iid)
- **Aim**: Test the null hypothesis on the unknown population *median*, X_{MED} is equal to a certain value

  $$H_0 : X_{MED} = \theta^*$$

  being θ^* a specific value (not a parameter).

How can we develop these test? Let's consider the following transformation of random variable. First we construct an indicator function 1 if each single value is larger that θ^* set by H_0 or 0 if lower; so we transform the data to a sequence of 0 and 1.

$$\Psi_i = \begin{cases} 1 & X_i > \theta^* \\ 0 & X_i < \theta^* \end{cases}, \quad i = 1, \ldots, n$$

Note that under the null we expect that the number of values larger than the median are circa 0.5 and 0.5 below, so a single value is distributed with a Bernoulli(0.5):

$$\Psi_i | H_0 \sim \text{Bern}\,(0.5), \quad i = 1, \ldots, n$$

From these two considerations we can deduce that, as a consequence if we sum the number of 1 we have that we have a binomial distribution (with parameter n, total obs, and 0.5):

$$Y = \sum_{i=1}^{n} \Psi_i | H_0 \sim \text{Bin}\,(n, 0.5)$$

So good we have a distribution under the null: we observe some data and we compare the results with the distribution under the null; if the observation is inconsistent with the distribution under the null we have observed a very rare event or the null is wrong. The logic is the same of Fisher approach, using the null only, typical Fisher reasoning.
Therefore the sample random variable Y, that counts how many sample observations are larger than the median θ^* (or, in other words, how many differences from θ^* have positive sign), has a known distribution under H_0.

The test statistic Y depends on the alternative hypothesis H_1 (in the sense we have to choose wwhere to reject the null):

TODO: ma non doveva essere fisheriana?

1. for one sided hypotheses:

   - for the one-sided hypotheses of the type:

     $$\begin{cases} H_0 : X_{MED} = \theta^* \\ H_0 : X_{MED} > \theta^* \end{cases}$$

     we compute the test statistic by $y_c = S_1 =$ number of observations *greater than* θ^*
   - for the one-sided hypotheses of the type:

     $$\begin{cases} H_0 : X_{MED} = \theta^* \\ H_0 : X_{MED} < \theta^* \end{cases}$$

     we compute the test statistic by $y_c = S_2 =$ number of observations *less than* θ^*.
   - in both cases, under the null, $Y \sim \text{Bin}\,(n, 1/2)$ and the p-value is defined by $p = \mathbb{P}\,(Y \geq y_c)$. We reject when $p < \alpha$.
     By symmetry notice that we have the equivalency $p = \mathbb{P}\,(Y \geq S_2) = \mathbb{P}\,(Y < S_1)$.

2. For a two-sided hypotheses:

   $$\begin{cases} H_0 : X_{MED} = \theta^* \\ H_0 : X_{MED} \neq \theta^* \end{cases}$$

   we compute the test statistic by $y_c = \max\{S_1, S_2\}$ where S_1 and S_2 are the counts of the number of observations greater than and less than θ^*

x_i	24	26	32	40	22	25	30	44	21	18	26	27	28	28	27
ψ_i	0	1	1	1	0	nd	1	1	0	0	1	1	1	1	1

Table 16.1: Dati sign test

respectively.
The p-value is defined by $p = 2\,\mathbb{P}(Y \geq y_c)$ (we multiply by 2 because its a double sided test). As usual we compare it with α.

Esempio 16.2.2. In our example the institute open only if the income is larger: so we adopt a one sided hypothesis, we are interest in one direction.

$$H_0 : X_{MED} = \theta^*$$
$$H_1 : X_{MED} > \theta^*$$

with $\theta^* = 25$, and refuse if $p < \alpha = 0.01$. Data are reported in table 16.1, with the dummification as well (note, if we encounter the median we want to test, we skip it, we don't consider this value, this the reason of n.d). Being interested to the eventual branch opening, which correspond the values of Y that play 'against' the null hypothesis. In this case we are interested in S_2 :

$$\begin{aligned} p &= \mathbb{P}(Y \geq y_c | H_0 : \theta = \theta^*) \overset{(1)}{=} \mathbb{P}(Y \geq 4 | H_0) \\ &= \mathbb{P}(\text{Bin}(14, 0.5) \geq 4) \\ &= \binom{14}{4} 0.5^4 0.5^{10} + \binom{14}{5} 0.5^5 0.5^9 + \ldots + \binom{14}{14} 0.5^{14} \\ &= 0.9713 \end{aligned}$$

where in (1) $P(Y > 4)$ perchè 4 sono gli 0.
So pvalue $0.9713 > \alpha$ and the null hypothesis survives. this was clear in data we had many 1 rather than 0

TODO: non convintissimo ne

```
mu0 <- 25
x <- c( 24 , 26,32,40,22,25,30,44,21,18,26,27,28,28,27)
phi <- as.integer(x < mu0) # qui l'alternativa è ad una coda greater
phi[x==mu0] <- NA
phi <- na.omit(phi)
s2 <- sum(phi, na.rm=TRUE)
## under the null a number equal or greater than s2
1 - pbinom(s2 - 1, size = length(phi), prob = 0.5)

## [1] 0.9713135
```

16.3 Sign test (two paired samples)

Osservazione 351. We can apply the sign test also for two sample. The idea is similar in this case we are interested in comparing two sequences of observation. Let's see an example

before	36	35	42	38	47	45	34	40
after	41	38	43	52	54	41	45	50

Table 16.2: Dati sign test two samples

Esempio 16.3.1. The head of an advertising firm aims to evaluate the effect of the radio advertising. For this reason he decides to launch, on a local station, a two-weeks campaign to promote a new iron.
The sales of the promoted iron are monitored for 4 months (2 months preceding and 2 following the advertising launch) on a random sample of 8 shops located in the area reached by the radio station.
The obtained results are reported in table 16.4. Evaluate the effectiveness of the ad campaign ($\alpha = 0.05$).

Osservazione 352. If there is effect the average/median should changed before-after.
We have here two problem:

1. The first problem is that these are *paired* data (same unit in different times) so *not independent* observation. There is a way to make data independent and is to compute the difference if we compite the difference we reduce, they can be assumed to be independent. So first we transform our data creating 8 observation like $d_i = after_i - before_i$

2. the second problem is that here n is small (it's 8 before and 8 after) so we cannot use testing based on fisher/neyman pearson which relies on larger numbers.
   In a typical course we would use student t-test. Problem here is that 8 is low; if the data is gaussian ok we use t-test, otherwise we use a sign test.

Finally: the null is that in classical t-test the *mean* of $d = 0$, in the sign test we consider the *median* of the difference instead. In our case the alternative of interest is that the difference is > 0 (we hope to refuse the null)

Osservazione importante 125 (Choosing null and alternative). How to decide the system of hypothesis? Think about the clinical problem and think about what we want to reject (null) and which condition/direction should move to rejection (alternative).
Here we want to check if the campain produce an increase, so we are interested in an increase and unilateral testing system.
In some problem this is not the only system of hypotheses we could construct. We have to choose and we should interpret the result in a coherent manner to the system of hypotheses.

Osservazione importante 126 (Ripasso). Paired Data: Student's t-test:

- The **data** are dependent (same shops) and they are called paired. Two samples are paired when the same n individuals are observed twice.

- The effect of the experimental factor is investigated by studying the differences within couples of subjects, $d_i (i = 1, \ldots, n)$: it resorts to a single sample composed of the n differences.

x_{1i}	36	35	42	38	47	45	34	40
x_{2i}	41	38	43	52	54	41	45	50
$d_i = x_{2i} - x_{1i}$	5	3	1	14	7	-4	11	10
ψ_i	1	1	1	1	1	0	1	1

Table 16.3: Dati sign test two samples

- The null hypothesis states the irrelevance of the experimental factor: given the mean value of the differences, $\overline{d}$, the question is whether the sample comes from a population with zero mean: $H_0 : \mu_D = 0$.
- Assuming that the random variable D is normally distributed in the population, we can use the following result:

$$T|H_0 = \frac{\overline{D} - 0}{s_D/\sqrt{n}} \sim \mathrm{T}\,(n-1)$$

  where s_D is the sample standard deviation computed on the set $\{di(i = 1, \ldots, n)\}$. For n large enough t_{n-1} converges to $\mathrm{N}\,(0, 1)$.

Osservazione importante 127. Paired Data: sign test

- **data**: $2n$ independent observations at least ordered, each couple corresponding to one of the n sample units. Let (X_{1i}, X_{2i}) be the generic random vector associated to the sampling of the i-th unit from the population and let $D_i = X_{2i} - X_{1i}$ be the corresponding difference random variables.
- Aim: Test the hypothesis that the unknown median of D, D_{MED}, also said treatment effect, is null:

$$H_0 : D_{MED} = 0$$

- with reference to the example the system of hypotheses is:

$$\begin{cases} H_0 : D_{MED} = 0 \\ H_1 : D_{MED} > 0 \end{cases}$$

  and the test statistic is $y_c = S_1 =$ number of positive differences

Esempio 16.3.2. Going back to the example $n = 8$, our null is $H_0 : D_{MED} = 0$, $\alpha = 0.05$. The completed data are reported in table 16.3; we apply the sign test to the univariate differences.

Since $Y = \sum_{i=1}^{n} \psi_i = 7$ we have

$$\begin{aligned} p &= \mathbb{P}\,(Y \geq 7|H_0) = \mathbb{P}\,(\mathrm{Bin}\,(8, 0.5) \geq 7) \\ &= 1 - \texttt{pbinom}(6, 8, 0.5) \\ &= 0.035 \end{aligned}$$

For the given α, both tests suggest that the median number of iron sales is significantly increased, i.e. the ad campaign has been effective.
In R we don't need a specific procedure/function, we use `pbinom`.

Osservazione importante 128. Some remarks on the sign test:

1. the Sign test can be seen as the binomial test for the control of hypothesis $H_0 : p = 0.5$, where p is the probability of the success event 'positive difference';
2. if we invert of the choice of what is labeled 'success' and 'failure' (increase or decrease in post-pre) is equivalent to invert the sign of the alternative in one-sided hypotheses and consequently S_1 or S_2;
3. according to the test, observations that coincide with the median do not carry any information and are discarded (with consequent reduction of the sample size). In the delta example if we have difference of exactly 0 we don't consider them;
4. it's possible to prove that when n increases the sign test is equivalent to the z test for the hypothesis $H_0 : p = 0.5$, that is the following transformation

$$z = \frac{\frac{y}{n} - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{n}}} = \frac{2y - n}{2n} 2\sqrt{n} = \frac{y - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$$

is distributed according to a standard normal (where y is the sum above, i guess). However this is not needed (if n is large we don't need this method).

Esempio 16.3.3 (Esame vecchio viroli)**.** At the beginning of a tutorial 8 names are read out in random order to 10 medical students foru are names of prominent sporting personalities (group A) and four are national and international politicials (group b). At the end of the session students are asked to recall as many of the names as possible. The number recalled were

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
A	3	2	2	4	3	2	3	3	1	4
B	2	1	3	3	1	0	3	2	2	3

Is there evidence of a difference between recall rates for the two groups ($\alpha = 0.05$)? consider the sign test and decide wheter a one or two tailed test is appropriate

1. no because p = 0.17 (taluni suggeriscono questa e io la confermo sotto)
2. no because = p = 0.92
3. yes because p = 0.04
4. no because p = 0.05468

Allora direi test a due code sicur

```
a <- c(3, 2, 2,   4,   3,   2,   3,   3,   1,    4)
b <- c(2,   1,   3,   3,   1,   0,  3,   2,   2,    3)
## differenze
d <- b - a
d2 <- d[d != 0 ]
d2 <- as.integer(d2 > 0)
sum(d2)
```

```
## [1] 2

2*pbinom(2, length(d2), prob = 0.5) # 2 per avere il test a due code (calcolo p sotto

## [1] 0.1796875
```

Esempio 16.3.4 (Esame vecchio viroli)**.** Eleven children are given an arithmetic test; after 3 weeks special tuition they are given a further test of equal difficulty. Their marks in each test (out of 90) are given in the table

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11
before	76	62	66	69	76	73	68	79	71	68	70
after 3 weeks	74	74	62	77	88	75	74	87	76	70	87

considere a two tail sign test with $\alpha = 0.05$ in order to decide about the claim that the average improvement due to extra tuition is 0

1. null is not rejected p = 0.059
2. null is not rejected p = 0.065 (taluni suggeriscono questa)
3. null rejected p = 0.0117
4. null not rejected p = 0.0993

Qua invece test ad una coda after > before

```
bef <- c(76,  62,  66,  69,  76,  73,  68,  79,  71,   68,  70)
aft <- c(74,  74,  62,  77,  88,  75,  74,  87,  76,   70,  87)
## differenze
d <- aft - bef
d2 <- d[d != 0 ]
d2 <- as.integer(d2 > 0)
sum(d2)

## [1] 9

## qui ci fermiamo un tick prima a 8 invece di 9 e prendiamo sino a
## 8 incluso, facendo il complementare; per due perché a due code
2*(1 - pbinom(8, length(d2), prob = 0.5))

## [1] 0.06542969

2 * pbinom(8, length(d2), prob = 0.5, lower.tail = FALSE)

## [1] 0.06542969
```

16.4 Wilcoxon (or signed rank) test for one sample

Osservazione 353. One of the most used: Wilcoxon test for one sample

Osservazione importante 129. We have:

- **Data**: n observations, *at least ordinal*, that corresponds to the n sample units; let $X_1, \dots, X_n$ be the random variables that describe the sampling of these units from the population.
- **Assumptions**:
  1. The r.v.s $\{X_i, i = 1, \dots, n\}$ are independent
  2. The r.v.s $\{X_i, i = 1, \dots, n\}$ are *simmetrically* distributed around the median (we should see simmetry in the data)
- **Aim**: Test the hypothesis on the unknown population median, θ:

  $$H_0 : X_{MED} = \theta^*$$

  being θ^* a specific value.
- we transform the data again by dichotomizing if above or below the median, then we them compute the rank of the difference in absolute values (not considering the sign, eg 1 is the closest to the median in absolute value):

  $$\Psi_i = \begin{cases} 1 & X_i > \theta^* \\ 0 & X_i < \theta^* \end{cases}$$
  $$R_i = \text{rank of } |X_i - \theta^*| \quad i = 1, \dots, n$$

  the definition of rank requires that a nondecreasing ranking list of the absolute differences from θ^* is produced.
- Wilcoxon showed that, under the listed assumptions, the sample random variable

  $$T^+ = \sum_{i=1}^{n} \Psi_i R_i$$

  (that cumulates the ranks of the differences with positive sign) has distribution completely deducible from combinatory calculus, under H_0. So we have it by tables or computer programs and we can procede in testing
- T^+ is distribution-free test-statistic for the control of $H_0 : X_{MED} = \theta^*$. If if the median is larger than that given by null, that is $\theta > \theta^*$, we expect that the sum of positive ranks is larger than the sum of negative ranks, that is $T^+ > T^-$. These two (T^+, T^-) are linked by:

  $$T^+ + T^- = \sum_{i=1}^{n} \Psi_i R_i + \sum_{i=1}^{n} (1 - \Psi_i) R_i = \sum_{i=1}^{n} R_i \overset{(1)}{=} \frac{n(n+1)}{2}$$

  where in (1) sum of first n integer is the gauss trick; so if T^+ increases T^- decrease

x_i	$X_i - 25$	ranks R_i	sum R_i (T-)	sum R_i (T+)
24	-1	2	2	0
26	1	2	0	2
32	7	11,5	0	11,5
40	15	13	0	13
22	-3	7	7	0
25	0		0	0
30	5	10	0	10
44	19	14	0	14
21	-4	9	9	0
18	-7	11,5	11,5	0
26	1	2	0	2
27	2	4,5	0	4,5
28	3	7	0	7
28	3	7	0	7
27	2	4,5	0	4,5
Sum			29,5	75,5

Table 16.4: Dati sign test two samples

Esempio 16.4.1. The calculation needed to analyze the data for the credit institution example are in table 16.4: first column is the data, second is the difference between them and the median under null hypothesis to check; then are the ranks (when we have observation with the same absolute difference how to handle the ranks?. We give the intermediate rank to all of them (eg 1,2,3 observation are equal and we assign2)) and their sums. We see that the sum of negative ranks is lower than that of the positive ranks.
To reproduce the test in R, we have simply:

```
x <- c(24 , 26, 32, 40, 22, 25, 30, 44, 21, 18, 26, 27, 28, 28, 27)
wilcox.test(x, mu = 25, alternative = "less") # if we reject the null, we don't open

## Warning in wilcox.test.default(x, mu = 25, alternative = "less"):
non è possibile calcolare p-value esatto in presenza di ties
## Warning in wilcox.test.default(x, mu = 25, alternative = "less"):
non è possibile calcolare p-valu esatti in presenza di zeri

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x
## V = 75.5, p-value = 0.9304
## alternative hypothesis: true location is less than 25
```

We see that the null is not rejected and we open (p-value is similar to what obtained before).

x_{1i}	36	35	42	38	47	45	34	40
x_{2i}	41	38	43	52	54	41	45	50
$d_i = x_{2i} - x_{1i}$	5	3	1	14	7	-4	11	10
ψ_i	1	1	1	1	1	0	1	1
$\lvert d_i - 0 \rvert$	5	3	1	14	7	4	11	10
r_i	4	2	1	8	5	3	7	6
v_i	4	2	1	8	5	0	7	6

Table 16.5: Dati sign test two samples paired

16.5 Wilcoxon's test for two (paired) samples

Osservazione 354. The procedure is the same but the difference is that we compute d_i as before and the median to check 0.

Esempio 16.5.1. Going back to the advertising example (One-tailed test) $n = 8$, $H_0 : D_{MED} = 0$, $\alpha = 0.05$.
Data with computation is reported in table 16.5. We have that $T^+ = \sum_{i=1}^{n} \psi_i R_i = 33$, p-value $= \mathbb{P}(T^+ \geq 33 | H_0) = 0.01953$.
In R we have

```
x2 <- c(41,38,43,52,54,41,45,50)
x1 <- c(36,35,42,38,47,45,34,40)
wilcox.test(x2, x1, paired = TRUE, alternative = 'greater')

##
##  Wilcoxon signed rank exact test
##
## data:  x2 and x1
## V = 33, p-value = 0.01953
## alternative hypothesis: true location shift is greater than 0
```

The last line of 16.5 can be described by the following random variables:

$$V_i = \begin{cases} i & \text{if rank } i \text{ is associated to a positive difference} \\ 0 & \text{if rank } i \text{ is associated to a negative difference} \end{cases}, \quad i = 1, \ldots, n$$

Such random variables are dichotomous, each associated to a different rank value, so that $T^+ = \sum_{i=1}^{n} V_i$.
About these random variables, it is possible to notice that we have the ingredients to standardize since:

- The expected value:

$$\mathbb{E}\left[T^+ | H_0\right] = \sum_{i=1}^{n} \mathbb{E}\left[V_i | H_0\right] = \sum_{i=1}^{n} (i \cdot 0.5 + 0 \cdot 0.5) = \frac{n(n+1)}{4}$$

- The variance

$$\begin{aligned}\text{Var}\left[T^+|H_0\right] &= \sum_{i=1}^{n}\text{Var}\left[V_i|H_0\right] = \sum_{i=1}^{n}\left(\mathbb{E}\left[V_i^2|H_0\right] - \mathbb{E}\left[V_i|H_0\right]^2\right)\\ &= \sum_{i=1}^{n}\left[(i^2\cdot 0.5 + 0^2\cdot 0.5) - \left(\frac{i}{2}\right)^2\right]\\ &= \sum_{i=1}^{n}\frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}\end{aligned}$$

Osservazione importante 130. Some remarks on Wilcoxon's test

1. According to the test, observations that coincide with θ^* do not carry any information and are discarded (with consequent reduction of the sample size). For information about the treatment of ties see HWC page 42 and page 50 (pt.11).

2. If we standardize, the large-sample version is a Normal test, thanks to CLT (Liapounov's[1] version for random variables that are independent but not identically distributed): when n diverges

$$\frac{T^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\Big|H_0 \xrightarrow{d} \text{N}(0,1)$$

   so if we can standardize when n increases it converges in distribution to a normal

Esempio 16.5.2 (Esame vecchio viroli)**.** on the day of the third round of the open golf championship before play started a television commentator said that conditions were such that the average scores of players were likely to be higher than those for the second round. for a random sample of 10 of the 77 players partecipating in both rounds the yielded scores are given in the table

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
second round	34	38	31	31	39	31	35	41	40	25
third round	29	25	34	43	37	42	46	43	35	29

Consider a wilcoxon test with appropriate alternative and $\alpha = 0.05$ to decide about the claim

1. the claim cannot be accepted becuse the p-value is 0.7128
2. the claim cannot be accepted becuse the p-value is 0.6458
3. the claim cannot be accepted becuse the p-value is 0.3239
4. the claim is accepted becuse the p-value is 0.7128

[1] See Randles e Wolfe (1979), Introduction to the theory of nonparametric statistics, Wiley

```
second <- c(34,  38,  31,  31,  39,  31,  35,  41,  40,   25)
third <- c(29,  25,  34,  43,  37,  42,  46,  43,  35,   29)
dif <- third - second
wilcox.test(dif, alt = 'great')$p.val

## Warning in wilcox.test.default(dif, alt = "great"):  non è possibile
calcolare p-value esatto in presenza di ties

## [1] 0.3229094
```

16.6 Wilcoxon Mann-Whitney (or rank sum) test

Esempio 16.6.1. Consider the cancer study where we observed the number of foci (the term focus indicates a mass of tumor cells positioned within a glandular space) in 22 histological sections of mastectomies of two different breast cancer (A and B):

```
Type A: 15 26 10 32 12 88 18 38 10 22 27 51
Type B: 68 100 100 76 25 93 28 54 87 100
```

Is there a significant difference in terms of number of foci between the two samples? ($\alpha = 0.01$).

Osservazione 355. In order to compare the central location in two independent samples according to a parametric perspective, we can use the Student's t-test

Osservazione importante 131 (Ripasso). We have that

- The test is

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  where

$$s_p = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

- assumptions
  - Homoscedasticity
  - $X \sim N$ in both populations

$$X_1 \sim \mathrm{N}(\mu_1 \sigma_1), X_2 \sim \mathrm{N}(\mu_2 \sigma_2), \sigma_1^2 = \sigma_2^2 = \sigma^2$$

Osservazione 356. What is the corresponding nonparametric test?

Osservazione importante 132 (Mann-Whitney test). We have that:

- **Data**: Two samples, C_1 and C_2 , composed of n_1 and n_2 (at least ordered) observations, respectively; let $X_{11}, \ldots, X_{1i}, ..., X_{1n_1}$ and $X_{21}, \ldots, X_{2j}, \ldots, X_{2n_2}$ be the random variables associated to the random operation of unit sampling from the corresponding populations.

- **Assumptions**:
  1. the random variables X_{1i}, $i = 1, \ldots, n_1$ are IID with unknown distribution function $F(\cdot)$;
  2. the random variables X_{2i}, $i = 1, \ldots, n_2$ are IID with unknown distribution function $G(\cdot)$;
  3. They are distinguished by a 'location shift' model that is: $G(x) = F(x - \Delta)$, $\forall x$, with $\Delta \in \mathbb{R}$
  4. The random variables X_{1i} and X_{2j} are mutually independent, that is, the two samples are independent.
- **Aim**: Test the hypothesis that the two populations have the same location

$$H_0 : G(x) = F(x), \forall x$$

that is

$$H_0 : \Delta = 0$$

Such hypothesis can be tested by the Wilcoxon Mann-Whitney test and its finite sample distribution can be obtained by ranks and combinatorial probability (similarly to the Wilcoxon's procedure).

- From asymptotic theory we can derive the large sample version of the test. For $\min n_1, n_2$ diverging,

$$\frac{S_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \Big| H_0 \xrightarrow{d} \mathrm{N}(0,1)$$

Esempio 16.6.2. In R coming back to the example

```
x1 <- c(15, 26, 10, 32, 12, 88, 18, 38, 10, 22, 27, 51)
x2 <-c(68, 100, 100, 76, 25, 93, 28, 54, 87, 100)
wilcox.test(x1, x2, alternative = 'two.sided')

## Warning in wilcox.test.default(x1, x2, alternative = "two.sided"):
non è possibile calcolare p-value esatto in presenza di ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x1 and x2
## W = 14, p-value = 0.002661
## alternative hypothesis: true location shift is not equal to 0
```

Chapter 17

Bootstrap

```
eval <- TRUE
```

17.1 Introduction

Osservazione 357. Bootstrap

- is an inferential computer-based method based re-sampling and simulation. The idea is solve complicated inferential problem by using the sample only
- The term bootstrap was coined by the saying "to pull oneself up by one's bootstrap", which comes almost certainly from the book The Surprising Adventures of Baron Munchhausen by Rudolph Erich Raspe (1785).
- was introduced by Bradley Efron in the late 1970s, and therefore quite recently (it requires intensive computational resources). The first time Efron proposed the method it was rejected from many journals (too strange/innovative for that time).
- is a widely used technique (see Efron and Tibshirani, 1993 for further details) used to solve various inferential problems:
  - point estimation
  - interval estimation
  - testing hypothesis

We start by a motivating example.

Esempio 17.1.1. A teacher want to summarize how many times a day students pick up their smartphone in the lab with totally 100 students. Instead of summarizing the pickups in the whole lab, the teacher makes a online survey which also provided the pickup-counting APP. In the next few days, she received 30 students responses with their number of pickups in a given day. The mean of the number of pickups is 20.06 and that the estimated standard deviation in the sample is 10.

- The common measure of accuracy is the standard error of the estimate. The standard error of the sample mean is $\sigma/\sqrt{n}$, so 1.64.

- Since n = 30, the Central Limit Theorem tells us that the sampling distribution of $\overline{X}$ so closely approximated by a normal distribution. We can then compute the C.I. of μ $(\overline{x} \pm z_{\alpha/2}\sigma/\sqrt{n})$ that is [17.38,23.81].

Here we assumed a gaussian distribution for a count number and its quite small. Even if we have a conficence interval it is acccurate? How accurate is this estimate result?

Osservazione 358. There are potential general problems:

- sample size could be low and we couldn't rely on asymptotic;

- we may not know the population distribution of the variable of interest and we cannot derive the distribution of the estimator.

- we may not have precise formula for the standard error of an estimator. If we want to make a inference about the median, what is the standard error of sample median?

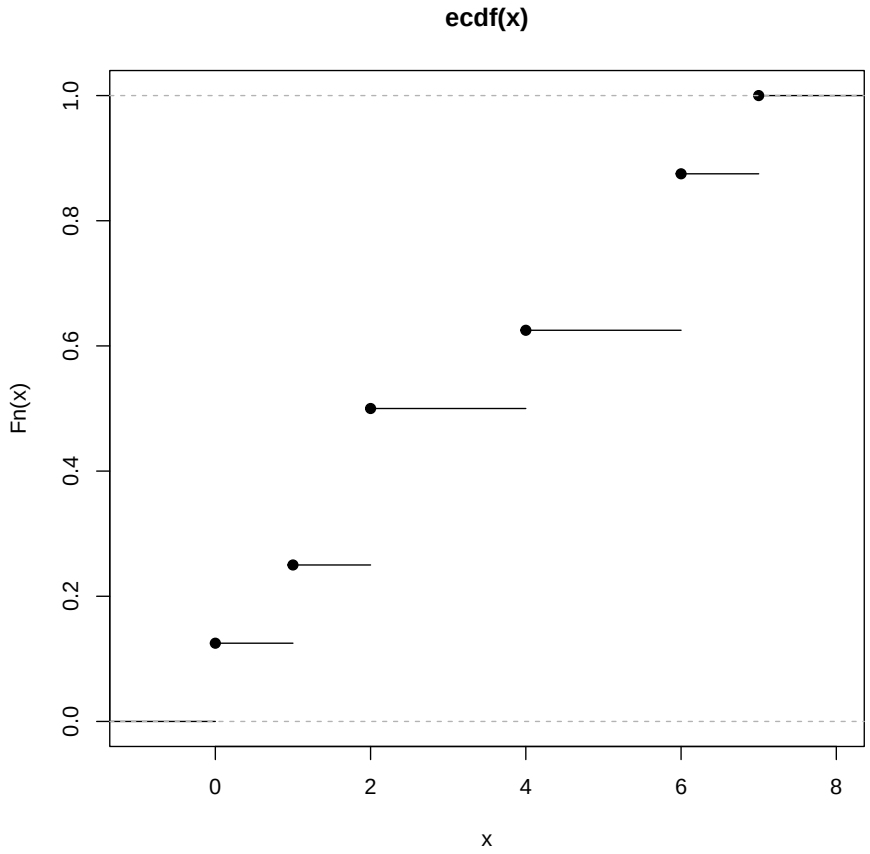
Osservazione 359. Since bootstrap is based on the concept of ecdf we revise it briefly.

Esempio 17.1.2. A random sample of $n = 8$ people yields the following (ordered) counts of the number of times they swam in the past month: $0, 1, 2, 2, 4, 6, 6, 7$. X is the number of swam. The distribution function F of X is unknown. The empirical distribution function $\hat{F}$ estimated on the sample of observations is:

$$\hat{F}(x) = \begin{cases} 0 & x < 0 \\ 1/8 & 0 \leq x < 1 \\ 2/8 & 1 \leq x < 2 \\ 4/8 & 2 \leq x < 4 \\ 5/8 & 4 \leq x < 6 \\ 7/8 & 6 \leq x < 7 \\ 1 & x \geq 7 \end{cases}$$

Point is: the unknown distribution function F can be estimated by $\hat{F}$.

```
## Ecdf in R
x = c(0,1,2,2,4,6,6,7)
plot(ecdf(x))
```



Osservazione importante 133 (Main bootstrap ideas/notation). Bootstrap matra is: a bootstrap sample relates to our sample as the latter relates to the population. All the high level bootstrap functioning can be summarized as follow (figure 17.1):

- Imagine X is the feature of interest in the population, with a distribution not known (otherwise we would use other inferential tool, not bootstrap). What we do is obtain a random sample $\mathbf{x} = \{x_1, \ldots, x_n\}$ which is a single element coming from the sample space (one of the possible sample we could extract). The sample space is not known since we dont know F the distribution in the population
- then bootstrap is based on the plug-in idea: a theorical distribution for our data F is replaced by its estimate, say $\hat{F}$, which is given by the empirical distribution function obtained with data from our sample.

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\cdot}(x_i \leq x)$$

in the formula, with a sample size n, we sum the number of times an

observation is lower or equal to a specific point of the ECDF. We then create the bootstrap world starting from $\hat{F}$

- in the bootstrap world the feature X^* (which is our sample and is completely known) with distribution $\hat{F}$ is "the population" used to generate pseudo/bootstrap samples $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$.
  The sample space of X^* (all the possible samples we could generate) is known since we know X^* entirely via its $\hat{F}$.
  It is important to note that $\hat{F}$ is a random function: for every sample of the 'real world' we will have a specific $\hat{F}$, and therefore many possible 'bootstrap worlds' exist (which depend on the empirical observation).

- when it comes to estimation our quantity of interest is something on the population X, say $T(F)$ (which is an estimator applied to the theorical distribution).
  In the real world we use a statistics applied to the sample $s(\mathbf{x})$ (eg this is the sample mean on our data) which is realization of the estimator
  The estimator itself has a distribution function G

  $$G(x) = P_F(s(\mathbf{x}) < x)$$

  but generally speaking we don't unknown unless we are in special cases (eg mean when n becomes large is gaussian). We don't have all the possible sample and can't compute the estimator on all of them. Therefore not knowing the distribution we don't know some of its feature such as expected value $E_F[s(\mathbf{x})]$, variance $var_F[s(\mathbf{x})]$

- on the contrary in the bootstrap world we have our estimator and can construct it's distribution as well.
  Let $\mathbf{x}^*$ be a bootstrap sample generated from $\hat{F}$. In the boostrap world we have the distribution function of the estimator

  $$\hat{G}(x) = \mathbb{P}_{\hat{F}}(s(\mathbf{x}^*) \leq x),$$

  which is known, and conditional to the true sample $\mathbf{x}$ we have observed.
  So in the bootstrap world we can compute estimates such as expected value of the estimator or its variance and plug in them back in the normal world

- one cool thing is that we can use $\hat{G}(x)$ in order to evaluate the goodness of our estimator.
  In this step, the assumption is that $\hat{G}(x)$ should represent a a 'good' approximation of $G(x)$. Equivalently, this means that the sample properties of $s(\mathbf{x}^*)$ should be the same of $s(\mathbf{x})$.

Osservazione 360. Bootstrap is divided in parametric and non parametric; non parametric was the first developed and it's simpler, so we start there.

17.1.1 Nonparametric bootstrap

Esempio 17.1.3 (Nonparametric bootstrap sample generation in R)**.** Imagine $0, 1, 2, 2, 4, 6, 6, 7$ is our sample $\mathbf{x}$; to construct a bootstrap sample $\mathbf{x}^*$ starting from $\mathbf{x}$ we extract with replacement for the same number of items

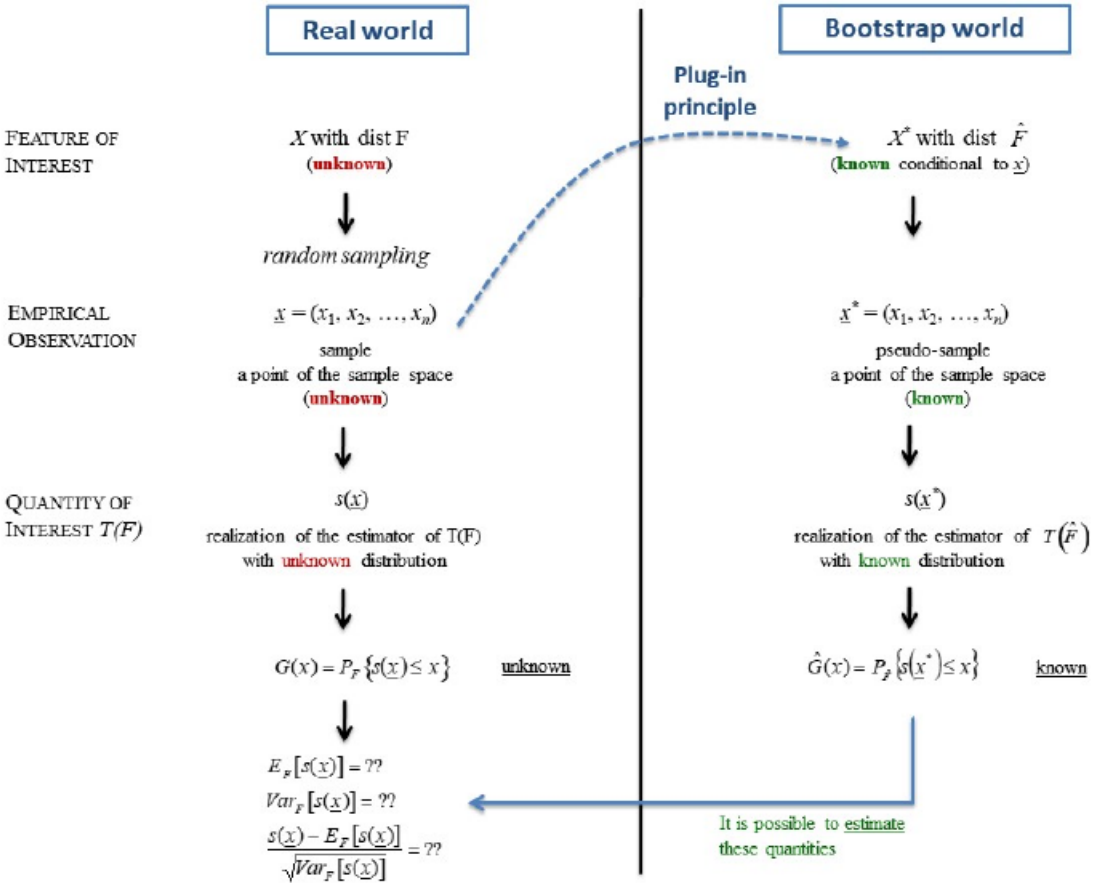


Figure 17.1: The bootstrap MF image

```
x  = c(0,1,2,2,4,6,6,7)
set.seed(1)
(xb = sample(x, size = length(x), replace=TRUE))

## [1] 0 2 6 0 1 4 6 2

# if we run this many times we get different  bootstrap samples
```

Osservazione importante 134. Points:

- let M be the number of all the possible samples of dimension n which can be drawn from $\hat{F}$;
- when n is finite we can count the total boostrap sample we can generate: that is if $\hat{F}$ is discrete, $M = n^n$. For instance if $n = 3$ and $x = \{5, 10, 2\}$ then $M = 27$.
- having all the possible samples, we can construct the distribution function of the estimator by applying to the samples and taking the frequency. So we can calculate the expectation of the estimator, the variance and the mse as well. we compute M distinct values $s(\mathbf{x}^*_{(m)})$ with $m = 1, \dots, M$ and we can get:

$$\hat{G}(x) = \frac{\sum_{m=1}^{M} \mathbb{1}_{\cdot}\left(s(\mathbf{x}^*_{(m)}) \leq x\right)}{M}$$
$$E_{\hat{F}}\left[s(\mathbf{x}^*)\right] = \frac{\sum_{m=1}^{M} s(\mathbf{x}^*_{(m)})}{M} = \overline{s}(\mathbf{x}^*_{(m)})$$
$$Var_{\hat{F}}\left[s(\mathbf{x}^*)\right] = \frac{\sum_{m=1}^{M}(s(\mathbf{x}^*_{(m)}) - \overline{s}(\mathbf{x}^*_{(m)}))^2}{M}$$
$$MSE_{\hat{F}}\left[s(\mathbf{x}^*)\right] = \frac{\sum_{m=1}^{M}(s(\mathbf{x}^*_{(m)}) - T(\hat{F}))^2}{M}$$

- problem is that M increase very quickly when n increases and it becomes difficult/cumbersome to reconstruct the total number of possible sample (eg even 8^8 is a big number).
  So instead of reconstructing all the sample space, instead of considering all the possible samples, we can take a reasonable subset: we take B samples randomly selected, $B < M$, that is $\mathbf{x}^*_{(b)}$ with $b = 1, \dots, B$.
  If we use B we can compute the previous quantities straightforwardly:

$$\hat{G}(x) = \frac{\sum_{b=1}^{B} \mathbb{1}_{\cdot}\left(s(\mathbf{x}^*_{(b)}) \leq x\right)}{B}$$
$$E_{\hat{F}}\left[s(\mathbf{x}^*)\right] = \frac{\sum_{b=1}^{B} s(\mathbf{x}^*_{(b)})}{B} = \hat{\theta}^*$$
$$Var^*_{\hat{F}}\left[s(\mathbf{x}^*)\right] = \frac{\sum_{b=1}^{B}(s(\mathbf{x}^*_{(b)}) - \hat{\theta}^*)^2}{B-1}$$
$$Bias^*_{\hat{F}}\left[s(\mathbf{x}^*)\right] = \hat{\theta}^* - T(\hat{F})$$

  where $\hat{\theta}^*$ represents the estimate of $T(\hat{F})$

Esempio 17.1.4 (Mouse)**.** Dataset: mouse (Efron, B. and Tibshirani, R., 1993) A small randomized experiment were done with 16 mouse, 7 to treatment group and 9 to control group. Treatment was intended to prolong survival after a test surgery. Measurements are in days as follows

```
ctrl <- c( 52, 104, 146, 10, 50, 31, 40, 27, 46)
trt <- c(94, 197, 16, 38, 99, 141, 23)
```

Imagine that we want to estimate the expected survival time in the group of *treated*. It's a small sample (n = 7)! Non Gaussian distribution! Why not classical theory? n is small, 7 is not enough.
We now create the samples in the bootstrap world: first approach is to work with all the possible samples. All samples/possible combinations are created using `expand.grid`:these are the position/indexes of our original sample:

```
x <- trt
## ------------------------------------------------------------------
## 1) Approccio a tappeto with all the possible samples (here 7^7)
## ------------------------------------------------------------------
## generate all the indexes combinations in samples
(unit <- seq_along(x))

## [1] 1 2 3 4 5 6 7

all_samples <- as.matrix(
    expand.grid(unit, unit, unit, unit, unit, unit, unit) # n = 7
)
head(all_samples) # all possible combinations

##      Var1 Var2 Var3 Var4 Var5 Var6 Var7
## [1,]    1    1    1    1    1    1    1
## [2,]    2    1    1    1    1    1    1
## [3,]    3    1    1    1    1    1    1
## [4,]    4    1    1    1    1    1    1
## [5,]    5    1    1    1    1    1    1
## [6,]    6    1    1    1    1    1    1

(M <- nrow(all_samples)) # 7^7

## [1] 823543

# vector in which we put the computed bootstrap values
mean.M <- rep(0,M)
for (i in 1:M){
    ## extract the bootstrap samples and calculate the estimator
    ## (expected survival time)
    mean.M[i] <- mean(x[all_samples[i, ]])
}

## sample estimate: this is the estimator in the real world
(est.media <- mean(x))
```

```
## [1] 86.85714

## expected value in the bootstrap world
(EB.mean <- mean(mean.M))

## [1] 86.85714

## surprisingly we have the same value (this is one of the property of
## the mean)
## parameter in the boostrap world is still the mean of our sample
## (which is used as population)
TF.mean <- mean(x)
## estimate of distorsion, variance and mean square error in the
## boostrap world
bias.mean <- EB.mean - TF.mean
variance.mean <- 1/M*sum((mean.M-EB.mean)^2)
MSE.mean <- 1/M*sum((mean.M-TF.mean)^2)

## Median
## ------------------------------------------------------------
## we use the same a tappeto approach here but we want to investigate
## the median estimator (s(x)) for the median of the population (T(F))

## We are not forced to do so, we could use the mean to estimate the
## median or viceversa. Good to know: we can compare the performace
## with bootstrap

median.M <- rep(0,M)
for (i in 1:M){
    ## one can change the estimator here and compare with other for
    ## the median by evaluating bias etc
    median.M[i] <- median(x[all_samples[i,]])
}
## sample estimate
est.median <- median(x)
## TF
TF.median <- median(x)
## Expected value in the bootstrap world (we use mean here!)
EB.median <- mean(median.M)
## Estimate of bias, variance and mean square error in the boostrap world
(bias.median <- EB.median - TF.median)

## [1] -14.27116

(variance.median <- 1/M*sum((median.M-EB.median)^2))

## [1] 1431.463

(MSE.median <- 1/M*sum((median.M-TF.median)^2))

## [1] 1635.128
```

```
## we discover that using the median to estimate the population median
## is biased: the median is a biased estimator for the median
## (differently from mean-mean)
```

Now back to the mean problem but using the **montecarlo approach**

```
## instead of using all the M, we use a subset B of 10000
## here we use sample with replace 10000 from the data to extract the bootstrap samples
## we use sample with replace = TRUE
B = 10000
mean.B <- rep(0,B)
for (i in 1:B)
    mean.B[i] <- mean(x[sample(1:7, 7, replace=TRUE)]) # <- here the news

## comparison of the three: it's can be not equal because B is lower
## than M. if increase B the approximation will become better
c(mean(mean.B), mean(mean.M), mean(x))

## [1] 86.41440 86.85714 86.85714

## the remaining evaluation is similar (remembering to substitute B
## where appropriate)

## sample estimate
est.mean <- mean(x)
## parameter to estimate
TF.mean <- mean(x)
## expected value in the bootstrap world
EB.mean <- mean(mean.B)
## estimate of bias, variance and mean square error in the boostrap world
(bias.mean2 <- EB.mean - TF.mean)

## [1] -0.4427429

(variance.mean2 <- 1/B*sum((mean.B-EB.mean)^2))

## [1] 535.1076

(MSE.mean2 <- 1/B*sum((mean.B-TF.mean)^2))

## [1] 535.3037
```

Osservazione 361. We have seen estimator mean and median, now we use bootstrap to compare estimator.

Esempio 17.1.5 (Estimator comparison by bootstrap)**.** We want to construct and compare 4 different estimators for the variance of a population. Sample variance is the most common choice but we could construct other estimator as well.
In general, there's no need to be correspondance for the quantity we want to estimate and the quantity we calculate in the sample: eg we have seen that the

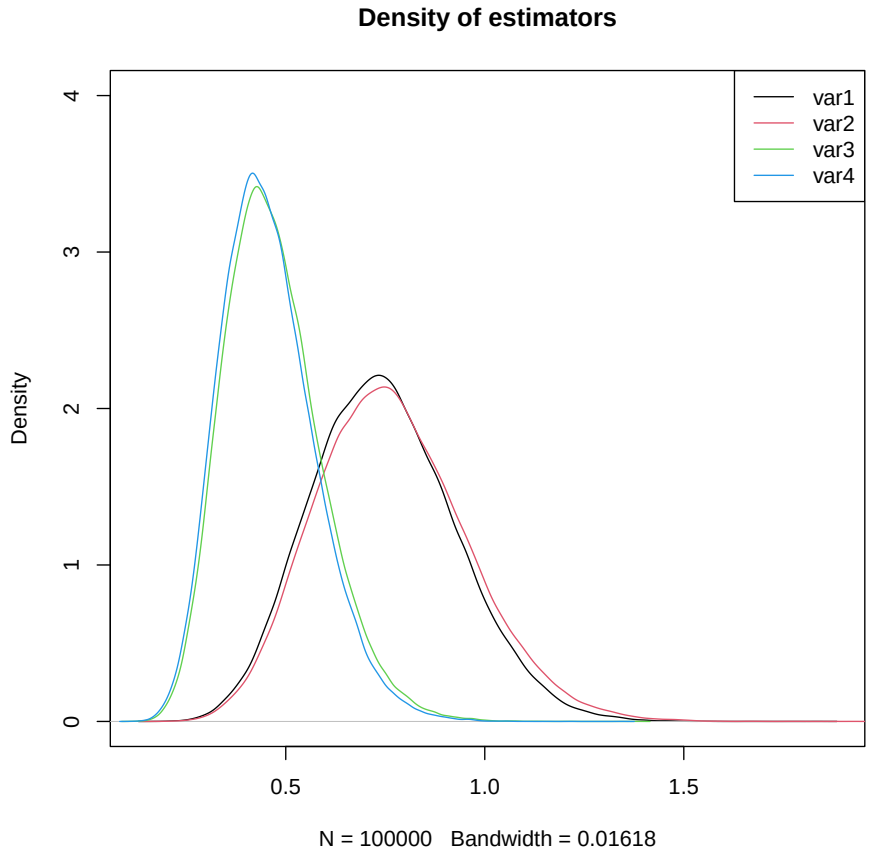
median is a bad estimator for the median (because of bias).
So we then compare them in terms of bias etc in a sample of 40 statistical units.
The estimator we'll compare are:

1. sample variance
$$v_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$
2. sample variance with respect to the median $x_{0.5}$: (instead of sample mean we put the median)
$$v_2 = \frac{\sum_{i=1}^{n}(x_i - x_{0.5})^2}{n}$$
3. the square of the mean of the absolute differences from the mean
$$v_3 = \left(\frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n}\right)^2$$
4. the square of the mean of the absolute differences from the median
$$v_4 = \left(\frac{\sum_{i=1}^{n}|x_i - x_{0.5}|}{n}\right)^2$$

The only tool we studied in classical statistics is the sample variance: its consistent, its biased (not (n-1) at denominator) however it's asymptotically correct. When n increases it's a good estimator.
However we don't know the properties of the remaining: we use bootstrap to evaluate them.

```
## we generate some data
set.seed(1)
n = 40
x = rnorm(n, mean = 5, sd = 1)

## we construct the 4 function, one for each estimator
var1 <- function(x) mean((x - mean(x))^2) # here we implement the biased one
var2 <- function(x) mean((x - median(x))^2)
var3 <- function(x) (mean(abs(x - mean(x))))^2
var4 <- function(x) (mean(abs(x - median(x))))^2

c(var1(x), var2(x), var3(x), var4(x))

## [1] 0.7665239 0.7678987 0.4763340 0.4763340

## 1,2 and 3,4 are similar (this because mean and median are similar)

### Now we estimate of the bias with nonparametric bootstrap using
### montecarlo approach (because 40^40 is too large). We use a
## limited number of pesudo sample (here 100000)
```

```
B = 100000

## we construct four object which will contans results of applying the
## estimators to the bootstrap samples; initialize them to 0
var.boot.1 = var.boot.2 = var.boot.3 = var.boot.4 = numeric(B)

## save  all the variances calculated on the samples
for (i in 1:B) {
    ## all the samples will be n = 40, with indexes as
    index = sample(1:n, n, replace = TRUE)
    ## we sample the index above, extract data below and apply the
    ## estimators
    var.boot.1[i] <- var1(x[index])
    var.boot.2[i] <- var2(x[index])
    var.boot.3[i] <- var3(x[index])
    var.boot.4[i] <- var4(x[index])
}

## how can we study the distribution of the estimators? First by
## plotting, eg histogram or density
## hist(var.boot.1, 50)
## par(mfrow = c(1, 2))
plot(density(var.boot.1), ylim = c(0,4),
     main = 'Density of estimators')
lines(density(var.boot.2), col = 2)
lines(density(var.boot.3), col = 3)
lines(density(var.boot.4), col = 4)
legend('topright', legend = sprintf("var%d", 1:4),
       col = 1:4, lty = 'solid')
```




```
## they are similar by couples as expected

## Estimate of the variance according to the four estimators
## is given by the mean of the boostrap
est <- c(mean(var.boot.1), # biased, but we knew it
         mean(var.boot.2), # this is better (more close to 0.78 which
                           # is the sample var 0.78)
         mean(var.boot.3), # 3 and 4 are the worst
         mean(var.boot.4))

## now we want to compute the bias. it's a numerical bias but a good
## estimator of the real bias of the formulas.  We need to compare the
## estimates with the "population value" which is the uncorrected
## variance in the sample (which is population in our bootstrap world)
## The variance in the "population" is the uncorrected variance
## applied to the sample/population
TF.theta = mean((x-mean(x))^2)
(bias <- est - TF.theta)

## [1] -0.020057141  0.001107954 -0.296744700 -0.309587939
```

```
## regarding the bias for the first estimator, we know from theory
## that bias is known and it is -1/n*var(x), in this case -1/n*1
## (variance of x is 1 because using rnorm)
# (if we increase B to infinity we get the same result as theory.
## bias of the second is lower surprisingly.
## third bias is larger.
## fourth is the largest one

## what we do with the bias? we can use it to correct our estimator!
## so we can use bootstrap to increase property of the estimator.
var1(x) - bias[1]

## [1] 0.7865811

var2(x) - bias[2]

## [1] 0.7667907

var3(x) - bias[3]

## [1] 0.7730787

var4(x) - bias[4]

## [1] 0.7859219

## Here we don't have 1 the true variance, because we started from a
## sample with 0.78. However this four estimator have been corrected
## and are better than the standard one

## here we can compute the MSE not only the bias: from yesterday
## lesson do something with this

## bias.median <- EB.median-TF.median
## variance.median <- 1/M*sum((median.M-EB.median)^2)
## MSE.median <- 1/M*sum((median.M-TF.median)^2)

## to do as exercise the mse of the estimator and choose the best one
```

17.1.2 Parametric bootstrap

A recap: our aim is to estimate the expected survival time (μ). A common choice is $\overline{X}$, the mean in the sample. Regarding inference we have several alternative solution on possible tools we can use.
Possible solutions: to study the distribution of $\overline{X}$ as estimator for $\mathbb{E}[X]$ in the treatment group we have different alternatives:

- if we dont know the distribution of X, we know that the sample mean is asymptotically distriuted as a gaussian with E_F and Var_F/n by the

Central Limit Theorem: the sample mean has the following distribution

$$\overline{X} \xrightarrow{d} \mathrm{N}\left(E_F\left[X\right], \frac{Var_F[X]}{n}\right)$$

- otherwise we can use non parametric bootstrap (example with mean/median as above). By generating subsamples with replacment (function `sample` with `replacement=TRUE`) we reconstruct the distribution of the estimator $\hat{G}$:

$$\hat{G}(x) = Pr_{\hat{F}}(s(\mathbf{x}^*) < x)$$

We can even compare what happens with bootstrap and CLT (`dnorm` with the density)

Osservazione importante 135 (Possible solutions: parametric approaches). In non parametric bootstrap, in the bootstrap world we used the empirical distribution function of our original sample.
Now if for some reason we assume a model or know the data have a specific distribution (we can check it by overlapping theorical distribution function to the empirical distribution function; if the overlap is good we can continue) we can introducetheorical models as explained below.
Eg For the duration/survival time X we could reasonably assume that

- it is distributed according to an exponential random variable with unknown μ: $X \sim \mathrm{Exp}\left(\mu\right)$ with $f(x,\mu) = \frac{1}{\mu}\exp(-x/\mu)$ and $\mathbb{E}\left[X\right] = \mu$, $\mathrm{Var}\left[X\right] = \mu^2$
- we can estimate μ with the sample mean $\hat{\mu} = \overline{x}$

Above we have seen that the assumption of an exponential distribution is reasonable since the ecdf overlap quite well with the theorical distribution.

Osservazione importante 136 (Possible solutions: toward parametric bootstrap). If we want to use parametric strategies, we can have 3 different solutions:

1. analytical solution (using probabilistic tools): if $X \sim \mathrm{Exp}\left(\mu\right)$ then it's possible to show that the sum and the mean of exponential is a Gamma and especially $\overline{X} \sim \mathrm{Gamma}\left(n, \frac{\mu}{n}\right)$ with $\mathbb{E}\left[\overline{X}\right] = \mu$ and $\mathrm{Var}\left[\overline{X}\right] = \frac{\mu^2}{n}$
2. if we don't remember the first solution, we can still use the central limit theorem. Also in this case (even if X is exponential) the sum/mean converges to a normal

$$\overline{X} \xrightarrow{d} \mathrm{N}\left(\hat{\mu}, \frac{\hat{\mu}^2}{n}\right)$$

3. **parametric bootstrap**: once we have found the theorical distribution by estimating the parameters $\hat{\mu}$ we can construct the bootstrap world is the distribution function of an exponential distribution of parameter $\hat{\mu}$

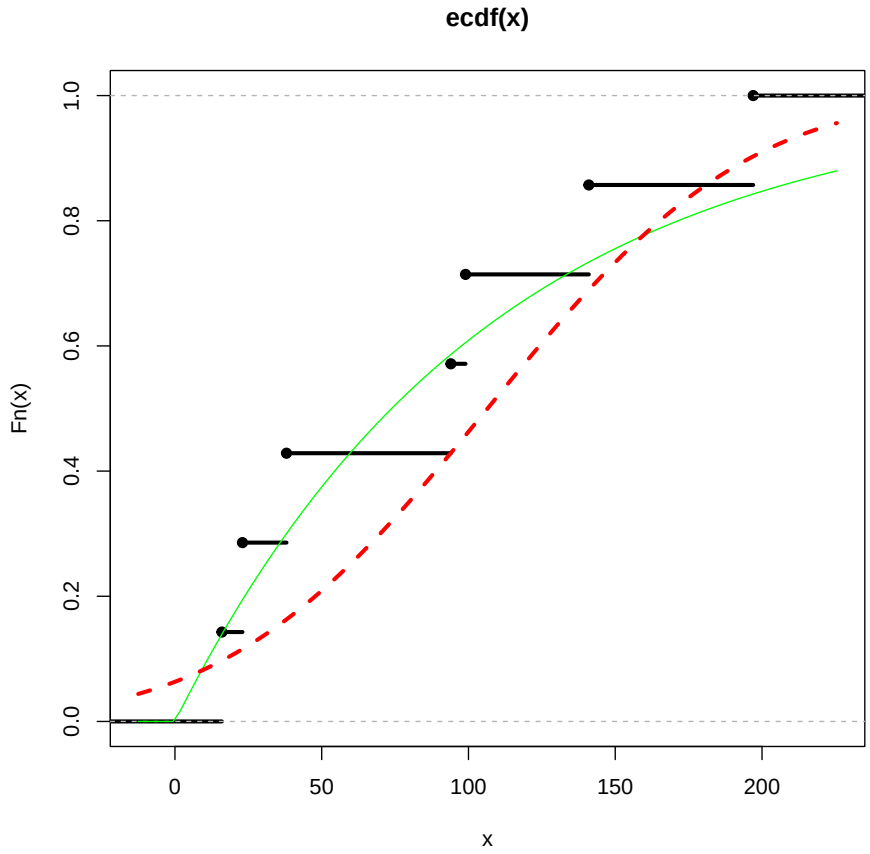
$$f(\underline{x}) = \frac{1}{\mu}\exp\left(-\frac{\underline{x}}{\hat{\mu}}\right)$$

Instead of resampling our initial values using `sample` with `replace=TRUE` we sample from according to a exponential distribution using `rexp` with the proper parameters in this manner

```
mu = mean(x)  ## mean of the sample/parameter estimation
rexp(7, 1/mu) ## if we run it this a lot of times we get the parametric bootstrap samples

## [1] 12.9854060  1.4618649  5.3537551  2.2713512  5.3148767  0.9285539  3.4747200
```

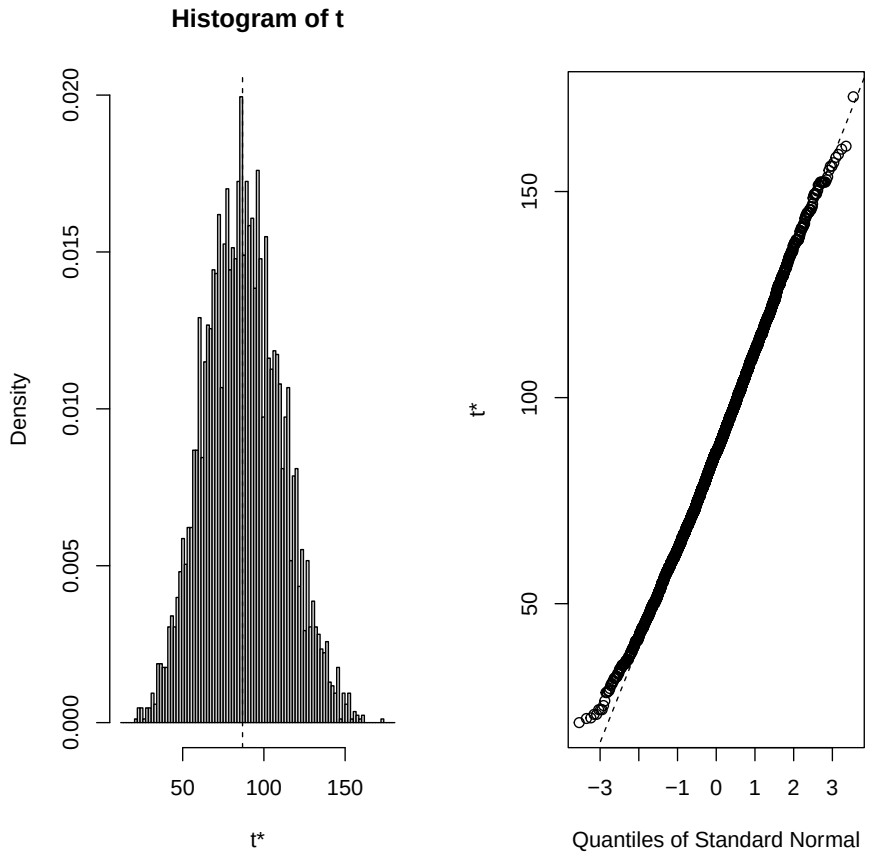
surpringly enough for the same problem we have 3 solution from the parametric approach and 2 from non-parametric solutions. We can compare all the possible 5 different manner and compare the possible solutions (compare the distribution of the estimator with different solutions) . last distribution of xbar according to the three parametric solutions. the continupis dark like is given by parametric bootstrap, the dashed blue by the exact solution (gamma based solution).. if we compare parametric with exact they almost overlap... this indication parametric bootstrap works... in red teh solution from CLT (tis a symmetric and gaussian) its not good because n = 7 (bad gaussian approximation)

```
## -------------------------------------
## comparison of the bootstrap solutions
## -------------------------------------

# sample
x <- trt
plot(ecdf(x), lwd = 3)
## empirical distribution function and compare with an exponential and
## a normal. 1/mean because in R the exponential distribution has the
## classical way
curve(pexp(x, 1/mean(x)), add=TRUE, col = "green")
curve(pnorm(x, mean = mean(x), sd=sd(x)),
      add = TRUE,
      lwd = 3,
      lty = 2,
      col = 'red')
```

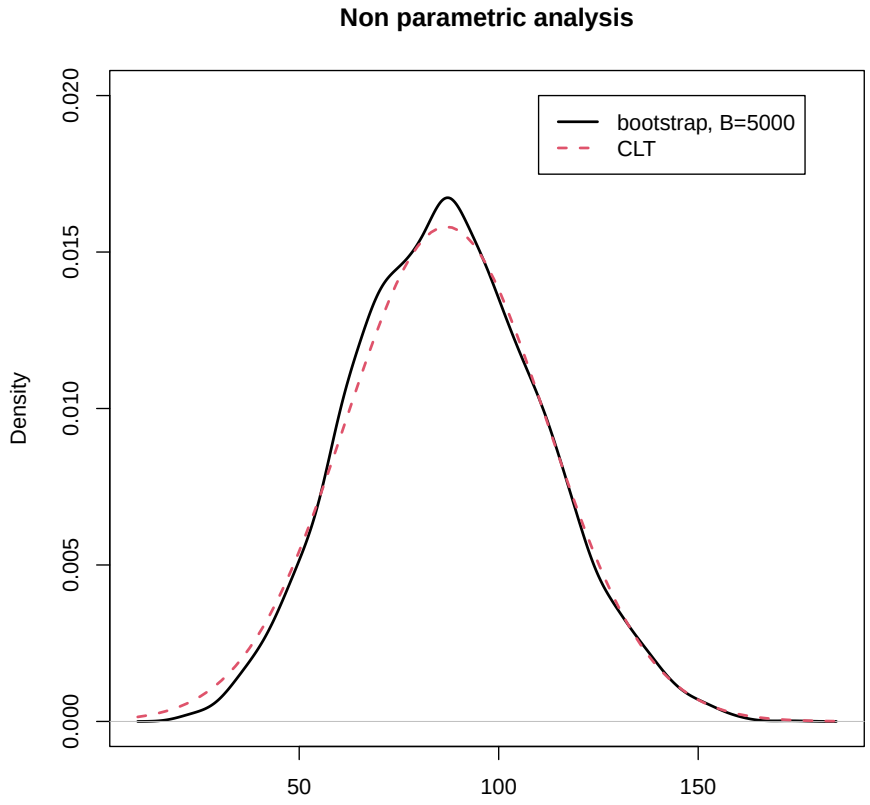



```
## red curve is cumulative function of a gaussian distribution
## compared to our data. Not very good, red is far from. This suggest
## approximating with a gaussian could not be a good idea
## OTOH approximating with exp seems better (o almeno così dice lei)

## non parametric bootstrap
## ------------------------
## instead of sample we use function from boot and botstrap libraries

## first we need a function that computes the statistics we are
## interested on. In order to estimate the expected survival time we
## use the mean of the data. np suffix for non parametric
## the function has to take a dataframe and an index of selection
## for the data which is run by the bootstrap machinery
media.f.np <- function(data, i) mean(data[i])
## the function boot want 3 stuff: data, the function we have created,
## number of iteration/subsamples (B for us)
bootstrap.np <- boot::boot(x, media.f.np, R = 5000)
bootstrap.np  ## results. in this results we have the bias and the
```

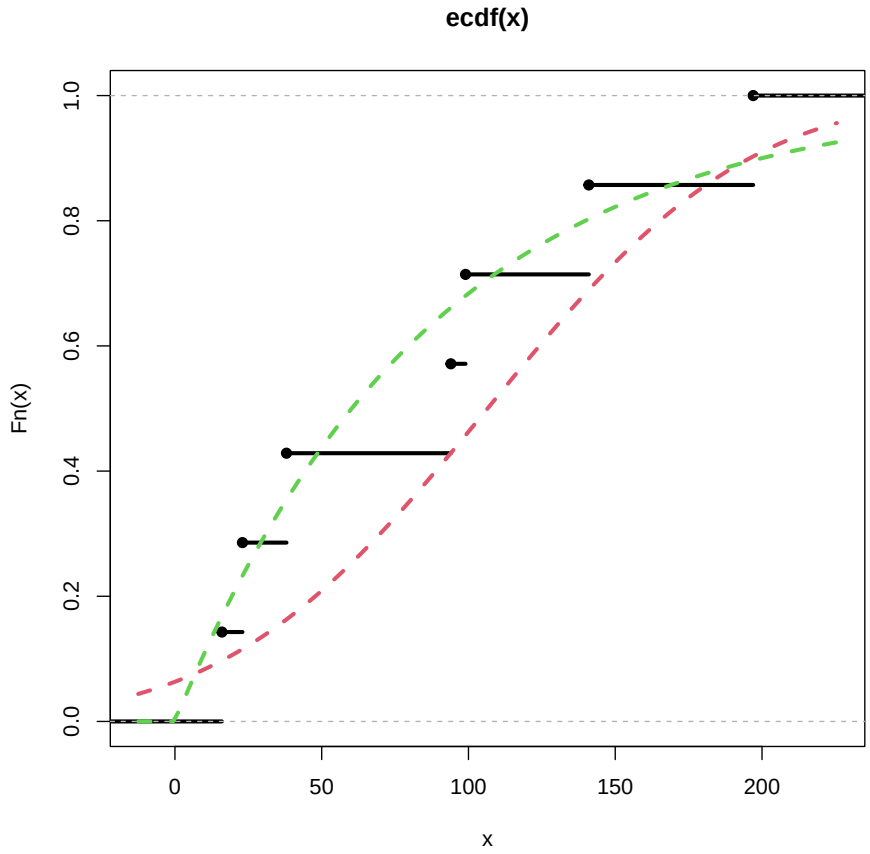
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = x, statistic = media.f.np, R = 5000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1* 86.85714 0.3830286    23.63797

              ## stderr of our statistics computed on pseudo
              ## samples. the expected survival time is 86 the with a
              ## bias (not high) and a stderr of our statistics
              ## computed on the pseudo samples
plot(bootstrap.np)
```



```
## we have two plots here. On the left is the distribution of
## statistics (the mean) in each pseudosample (the mean of this
## distribution is 86 which is the vertical line); on the right we
## have a comparison of the mean in each boot sample with theoretical
## quantiles (if correct it should be at 45 degrees)


## Comparison with CLT
## -------------------
## below plot of all the stats in the bootstrap world (they are
## contained in $t of bootstrap.np which containst all the means in
## the boottrap samples)
plot(density(bootstrap.np$t),
     main = "Non parametric analysis",
     lwd = 2,
     xlab = " ",
     ylim = c(0,0.02))
## we add the theoretical curve (n = 7) for which we need to estimate
## the true value of the param in the boostrap world. we need above
## because with CLT X sim E(mean(sample), var/n)
media = mean(x)
varianza = var(x)
curve(dnorm(x, mean = media, sd = sqrt(varianza/7)),
      add = TRUE, col = 2, lwd = 2, lty = 2)
legend(110, 0.020,
       legend = c("bootstrap, B=5000","CLT"),
       lwd = 2, lty = c(1:2), col = c(1:2))
```

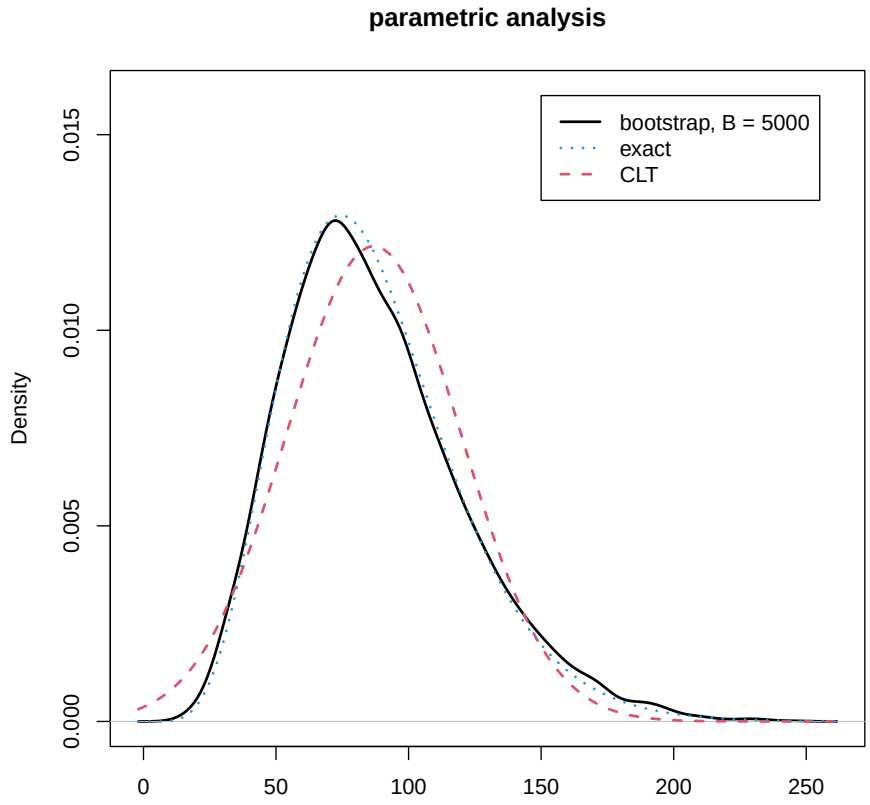



```
## we can see that there's no perfect match between CLT approximation
## and our data, there's distorsion and CLT is not so good because (n
## = 7) so in this case non parametric boostrap is better


## --------------------
## parametric solutions
## --------------------

## ecdf data
plot(ecdf(x), lwd = 3)
## theorical gamma in the parametric solution: if X is an exponential,
## then we have that xbar is a gamma
curve(pgamma(x, shape = 1, scale = media),
      add = TRUE, lwd = 3, lty = 2, col = 3)
## if we compare with the gaussian the green/gamma is better as
## parametric approximation of data
curve(pnorm(x, mean = mean(x), sd = sd(x)),
      add = TRUE, lwd = 3, lty = 2, col = 2)
```

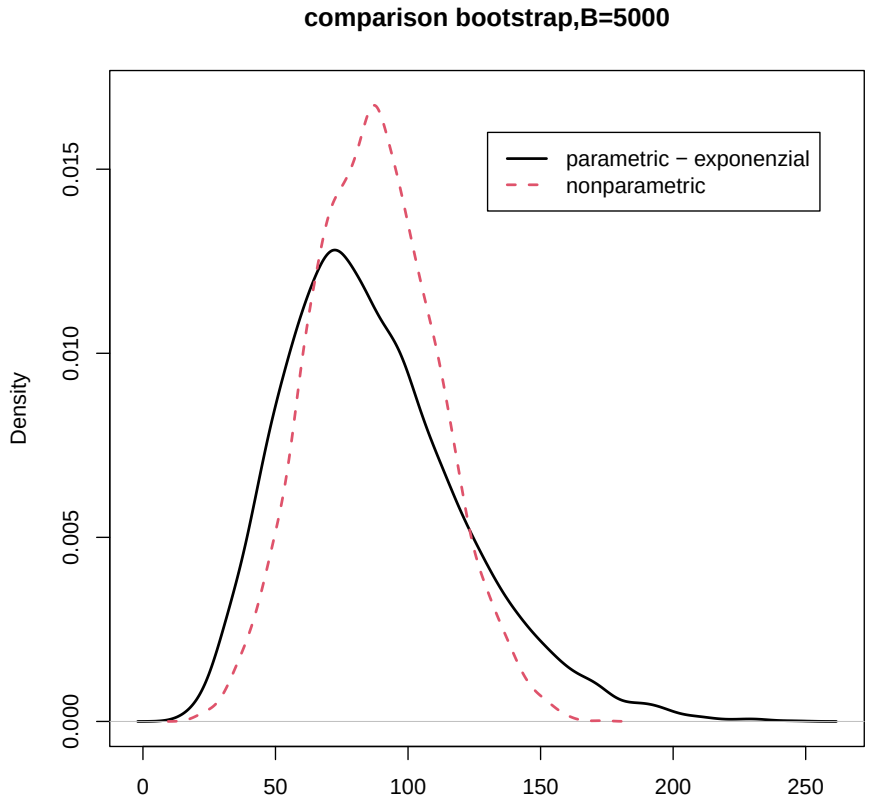



```
## Parametric Analysis using boot
## ------------------------------

## we don't have index in the function below, differently from nonpar
## boot otherwise the function is the statistics in the bootstrap
## world as above
media.f.p <- function(data) mean(data)

## another function which is the data generating function: we generate
## data according to a gamma (a special case, it's an exponential
## distribution credo dato che shape = 1 and second parameter the
## parameter of exponential)
rg <- function(data, mle) rgamma(length(data), shape = 1, scale = mle)
## length(data) = 7, mle is given as parameter... in the possible
## solutions, we assume data is generated according to a 1 as the
## first parameter and 19

## now we 're ready to call boot
bootstrap.p <-
    boot::boot(x, # data
```

```
                media.f.p, # stats in the boot world
                R = 5000,
                sim = "parametric", ## specify this or non-paramatric
                                    ## is default
                ran.gen = rg, ## if boot is parametric we add the
                              ## function that generates data
                mle = mean(x) ## parameter of the function generating
                              ## data other than data above
                )
## results
bootstrap.p

##
## PARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = x, statistic = media.f.p, R = 5000, sim = "parametric",
##     ran.gen = rg, mle = mean(x))
##
##
## Bootstrap Statistics :
##     original   bias    std. error
## t1* 86.85714 0.375214    33.89069

##  here our expected survival time is 86 again and we have different
##  bias and stderr

## Comparison among parametric bootstrap (exponential), analytical
## solution and CLT.
## 1) param boot
plot(density(bootstrap.p$t), main = "parametric analysis",
     lwd = 2, xlab = " ", ylim = c(0, 0.016))
## 2) clt (point B solutions)
curve(dnorm(x, mean=media, sd = sqrt(media^2/7)),
      add=TRUE, col=2, lwd=2, lty=2)
## 3) analytical gamma (point A solutions)
curve(dgamma(x, shape=7, scale=media/7), add=TRUE,
      col = 4, lwd = 2, lty = 3)
legend(150, 0.016, legend = c("bootstrap, B = 5000", "exact", "CLT"),
       lwd = 2, lty = c(1,3,2), col = c(1,4,2))
```

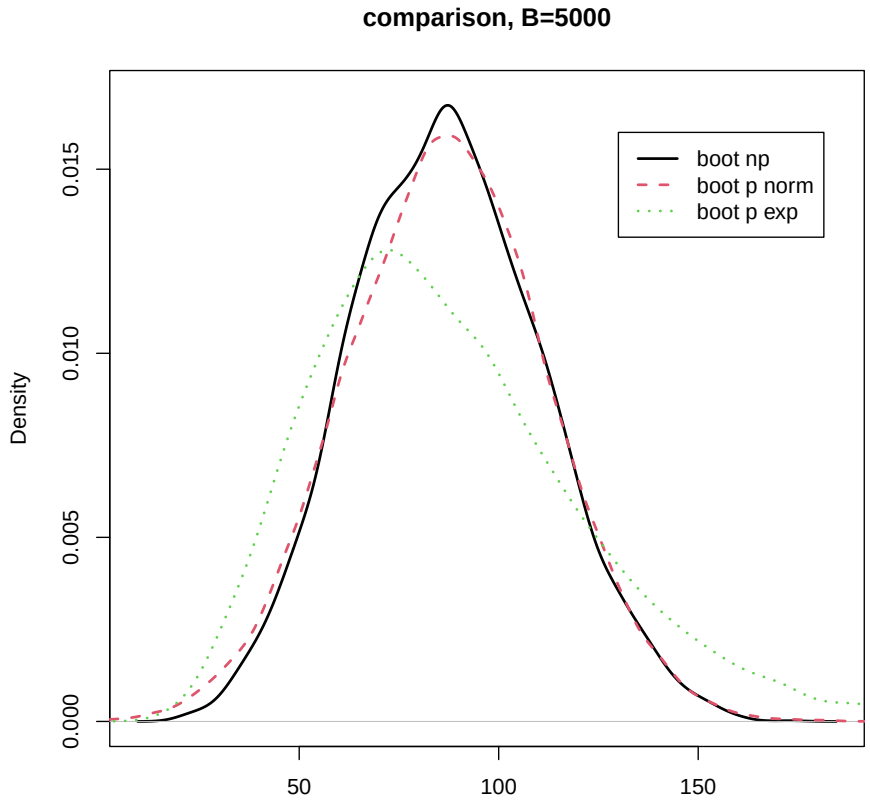



```
## boot and exat are similar clt is not

## Comparison between parametric bootstrap (exponential) and
## nonparametric bootstrap parametric
plot(density(bootstrap.p$t), main = "comparison bootstrap,B=5000",
     lwd = 2, # para
     xlab = " ",
     ylim = c(0,0.017))
## non parametric
lines(density(bootstrap.np$t), col=2, lwd=2, lty=2)
legend(130,0.016,legend=c("parametric - exponenzial","nonparametric"),
       lwd=2,lty=c(1,2),col=c(1,2))
```




```
## there are many difference reflecting assuming a gamma distribution
## on parametric boot

## Suggestion: first use the ecdf choose between non par and par using
## the overlapping (there exponential distrib where better) (here
## parametric is preferable)


############ parametric bootstrap normal
## here a different way to generate data: whats'happen if instead
## parametric-exponential we use parametric gaussian (which is not good
## but we can try)

## construct function for the boot
media.f.p <- function(data) mean(data)

## here is different (we use rnorm instead of rgamma), parameters are
## passed as vector of two elements
rg <- function(data, theta){
    media = theta[1]
```

```
    varianza = theta[2]
    rnorm(length(data), mean=media, sd=sqrt(varianza))
}

## we compute mean and variance of the two estimates used for data
## generation
theta = c(mean(x), var(x))
bootstrap.p.norm <-
    boot::boot(x,   # here we change only the data generating function
               media.f.p,
               R = 50000,
               sim = "parametric",
               ran.gen = rg,
               mle = theta) # mle is passed as second argument to
                            # ran.gen

## comparison of nonpar boot with param-gaussian and param-gamma
plot(density(bootstrap.np$t), main="comparison, B=5000",lwd=2,
     xlab=" ",ylim=c(0,0.017))
lines(density(bootstrap.p.norm$t), col=2, lwd=2, lty=2)
lines(density(bootstrap.p$t), col=3, lwd=2, lty=3)
legend(130, 0.016, legend = c("boot np","boot p norm","boot p exp"),
       lwd=2, lty=c(1,2,3), col=c(1,2,3))
```




```
## the non param boot is closed to the param with the gaussian
## distribution
```

17.2 Confidence intervals

Osservazione 362. Statistical literature is rich regardingly; we have several method to construct cis (we will see all of them and try to say which is better).

17.2.1 Methods

17.2.1.1 Boostrap Gaussian Intervals

Osservazione 363. Imagine we have an estimator that is approximately gaussian: we can check the histogram of the estimator (`boostrap$t`) or the qqplot.
In the case it's centered on $T(F)$ which is the *true value* we want to estimate ($T(F)$ is the mean of `bootstrap$t`, the statistics adopted) and eventually the *bias* (computing according to the bootstrap), and having the variance of the estimator.
So we have that $s(x) \sim \mathrm{N}\left(T(F) + \mathrm{Bias}\left(s(x)\right), \mathrm{Var}\left[s(x)\right]\right)$

Then $s(x) \sim \mathrm{N}\left(T(F) + \mathrm{Bias}\left(s(x)\right), \mathrm{Var}\left[s(x)\right]\right))$ from which its standardization is distributed according to a standard normal, so we have that

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{s(x) - T(F) - \mathrm{Bias}\left(s(x)\right)}{std\left[s(x)\right]} \leq z_{\alpha/2}\right) = 1 - \alpha \tag{17.1}$$

where $\alpha/2$ is the confidence vaue we choose and $std(s(x)) = \sqrt{\mathrm{Var}\left[s(x)\right]}$. From the previous equation we can construct the confidence interval for $T(F)$ as usual developing as follows:

$$s(x) - \mathrm{Bias}\left(s(x)\right) \pm z_{\alpha/2} std(s(x)) \tag{17.2}$$

If in the previous we estimate $Bias_F[s(x)]$ and $std_F(s(x))$ with the bootstrap method (star values below) we get the interval:

$$s(x) - Bias^*[s(x)] \pm z_{\alpha/2} std^*[s(x)] \tag{17.3}$$

with an approximate confidence level of $1 - \alpha$ (the interval has a probability of $1 - \alpha$ to include the true value $T(F)$).

17.2.1.2 Percentile intervals

Imagine that the formulation of statistics in the bootstrap world is the same as the functional: eg we use mean to estimate the mean, the median to estimate the median, the variance for the variance and so on etc, that is $s(x) = T(\hat{F}) = \hat{\theta}$. Then in this case the sequence of the statistics computed at each boot sample is a sequence of estimations of the parameter of interest.
Then the sequence:

$$\left\{s(x^*_{(1)}), s(x^*_{(2)}), \ldots, s(x^*_{(B)}), \right\} = \left\{\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^* B\right\} \sim \hat{G}(x) \tag{17.4}$$

So in order to have a CI we order the values of our estimates and take two quantiles. The bootstrap percentile interval is:

$$\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2} \tag{17.5}$$

In R its a matter of `quantile(bootstrap$t, probs = c(0.025, 0.975))`.

Finally it's possible to show an interesting property that the interval fullfill. These intervals are equivariant with respect to monotone transformations, i.e. if $\phi = m(\theta)$ with m monotone transform we will have:

$$\left[\hat{\phi}^*_{\alpha/2}, \hat{\phi}^*_{1-\alpha/2}\right] = m(\hat{\theta}^*_{\alpha/2}), m(\hat{\theta}^*_{1-\alpha/2}) \tag{17.6}$$

eg if we have an interval for theta, by taking the log of the lower and upper bound we have automatically a CI for log(theta); this is not true for all kind of interval.

17.2.1.3 Basic Intervals

Based on quantity called *root* given by the statistics - true parameter:

$$R(s(x), T(F)) = s(x) - T(F)$$

it's similar to the pivot, a trasformatio of data not any more depending on parameters. The root has distribution function:

$$H(x) = \mathbb{P}_F(s(x) - T(F) \leq x)$$

let $h_{\alpha/2}$ and $h_{1-\alpha/2}$ be the quantiles of such function, such that $H(h_{\alpha/2}) = \alpha/2$ and $H(h_{1-\alpha/2}) = 1 - \alpha/2$. Then we can claim that

$$\mathbb{P}\left(h_{\alpha/2} \leq s(x) - T(F) \leq h_{1-\alpha/2}\right) = 1 - \alpha$$

from which we derive the confidence interval, again for $T(F)$

$$s(x) - h_{1-\alpha/2}; \quad s(x) - h_{\alpha/2};$$

where $h_{\alpha/2}$ and $h_{1-\alpha/2}$ can be estimated with the bootstrap. Infact $s(x) - T(F)$ has a distribution that can be derived by bootstrap. It is possible to show that the distribution $\hat{H}$ is the same distribution $\hat{G}$ translated by the constant $T(\hat{F})$

$$\hat{H}(x) = \hat{G}(x + T(\hat{F})) = \frac{\sum_{b=1}^{B} \mathbb{1}\left(s(x^*_{(b)} - T(\hat{F}) \leq x)\right)}{B}$$

This means that the quantiles of H can be obtained by translating the quantiles of G according to $\hat{h}_{\alpha/2} = \hat{g}_{\alpha/2} - T(\hat{F})$. In other terms, if we know G (which is derived with bootstrap, the basic bootstrap interval is finally:

$$s(x) - \hat{g}_{1-\alpha/2} + T(\hat{F}); \quad s(x) - \hat{g}_{\alpha/2} + T(\hat{F});$$

17.2.1.4 t bootstrap Intervals

Similar somewhat to the previous, we consider a root quantity which is a standardization with

$$R(s(x) - T(F)) = \frac{s(x) - T(F)}{\sqrt{Var^*[s(x)]}} \tag{17.7}$$

Problem is that the denominator should be the variability of each s, s applied to the boot sample, which we don't know. We have just one overall variability given by our distribution of the results.
So idea is to introduce a second level of bootstrap to estimate the variance of each single s, to standardize the s obtained: the pseudo sample become the new population for a second level of sampling. Since computationally can be proibitive it has been proven that this work well if we take $B = 1000$ ath the first level and 25 resampling at the second level. The expression 17.7 represents a kind of generalization of the t statistics test of Student, from which the names has been taken.
If $s(x)$ is biased the 17.7 can be rewritten as

$$R(s(x) - T(F)) = \frac{s(x) - Bias^*[s(x)] - T(F)}{\sqrt{Var^*[s(x)]}} \tag{17.8}$$

The application of 17.7 or 17.8 or the derivation of confidence intervals requires a bootstrap estimate of the denominator quantity. This leads to a double level

of bootstrap: for each bootstrap sample $x^*_{(b)}$ we resample a subset in order to get an estimate of its variance.

The b-th ratio becomes

$$R(s(x) - T(F)) = \frac{s(x^*_{(b)}) - T(\hat{F})}{\sqrt{Var^*[s(x^*_{(b)})]}} \tag{17.9}$$

Usually, we consider at least $B = 1000$ for the first level and $B = 25$ for the variance estimated at the second level. At the end of this double level, the aim is to estimate the empirical distribution function the empirical distribution function of H:

$$\hat{H}(x) = \frac{\sum_{b=1}^{B} \mathbb{1}\left(\frac{s(x^*_{(b)}) - T(\hat{F})}{\sqrt{Var^*[s(x^*_{(b)})]}} \leq x\right)}{B}$$

needed to obtain $\hat{h}_{1-\alpha/2}$ and $\hat{h}_{\alpha/2}$ used for the interval from which:

$$s(x) - \hat{h}_{1-\alpha/2}\sqrt{Var^*[s(x)]}; \quad s(x) - \hat{h}_{\alpha/2}\sqrt{Var^*[s(x)]};$$

17.2.1.5 BCa Intervals

Bca bootstrap (bias corrected accelerated) intervals start from the percentile intervals and generaize/correct them in order to deal with the possible bias of the estimator adopting a transformation on percentiles that depends on two parameters: an *acceleration* and a parameter of *bias-correction*.
The analytical derivation is complex, but the idea is to start from the ordered sequence of 17.4 and to get the percentiles, in order to deal with the possible bias and the eventual dependence on the variance $T(\theta)$ from θ.
The BCa intervals are equivariant with respect to monotone transformations.

17.2.2 Comparison among the different approaches

In order to evaluate the different methods for constructing bootstrap-based confidence intervals, we consider the criterion of *accuracy*.

Without loss of generality, we consider one-side confidence intervals of the type $[-\infty, q_\alpha]$ such as $\mathbb{P}(\theta \leq q_\alpha) = \alpha$: suppose that $\hat{q}_\alpha$ is the estimated upper value. A confidence interval is said accurate for θ if $\mathbb{P}(\theta \leq \hat{q}_\alpha) \approxeq \alpha$.
There are different degrees of accuracy:

- accuracy of the first order $\mathbb{P}(\theta \leq \hat{q}_\alpha) = \alpha + O(n^{-1/2})$: quantity with the same order/magnitude of $n^-1/2$ is $1/sqrt(n)$
- accuracy of the second order $\mathbb{P}(\theta \leq \hat{q}_\alpha) = \alpha + O(n^{-1})$. Accuracy of order 2 is $1/n$; the second is better because $1/n < 1/sqrt(n)$
- accuracy of the order p-th $\mathbb{P}(\theta \leq \hat{q}_\alpha) = \alpha + O(n^{-p/2})$

where $O(n^{-p/2})$ denotes a quantity $r(n)$ function of n such that if $n \to \infty$ then $|r(n)| < M\left|n^{-p/2}\right|$ where M is a positive constant.
It is possible to show that:

- the gaussian intervals are accurate of order one. They are accurate of second order if X is Gaussian distributed.

- the basic bootstrap intervals are accurate to the first order.

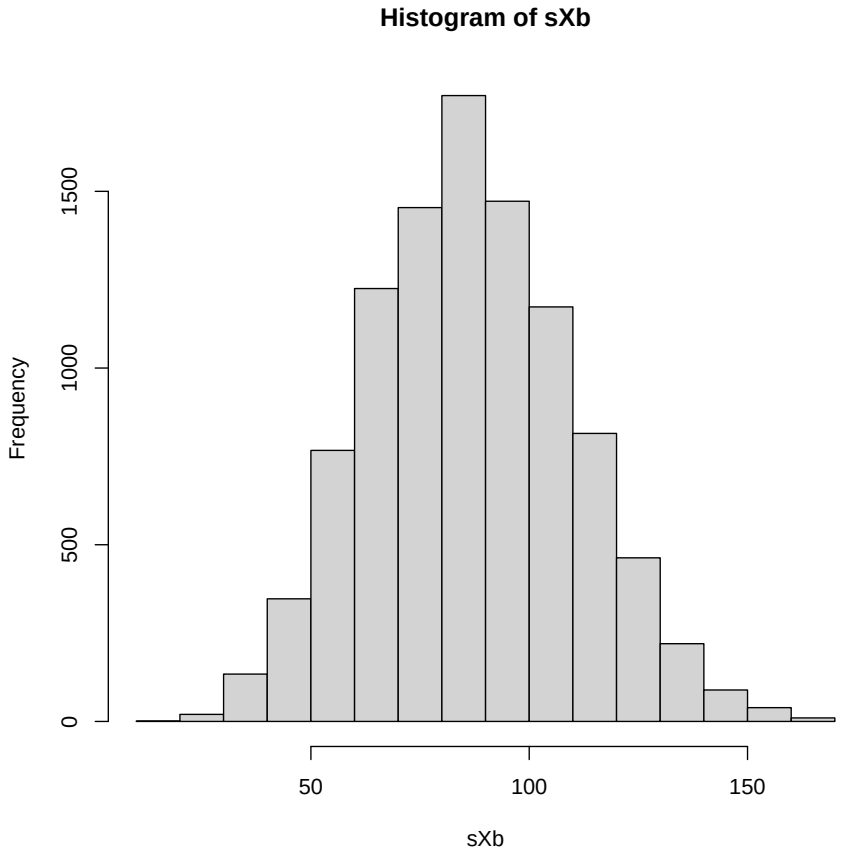
- the percentile intervals are accurate to the first order.

- the t bootstrap intervals are accurate to the second order.

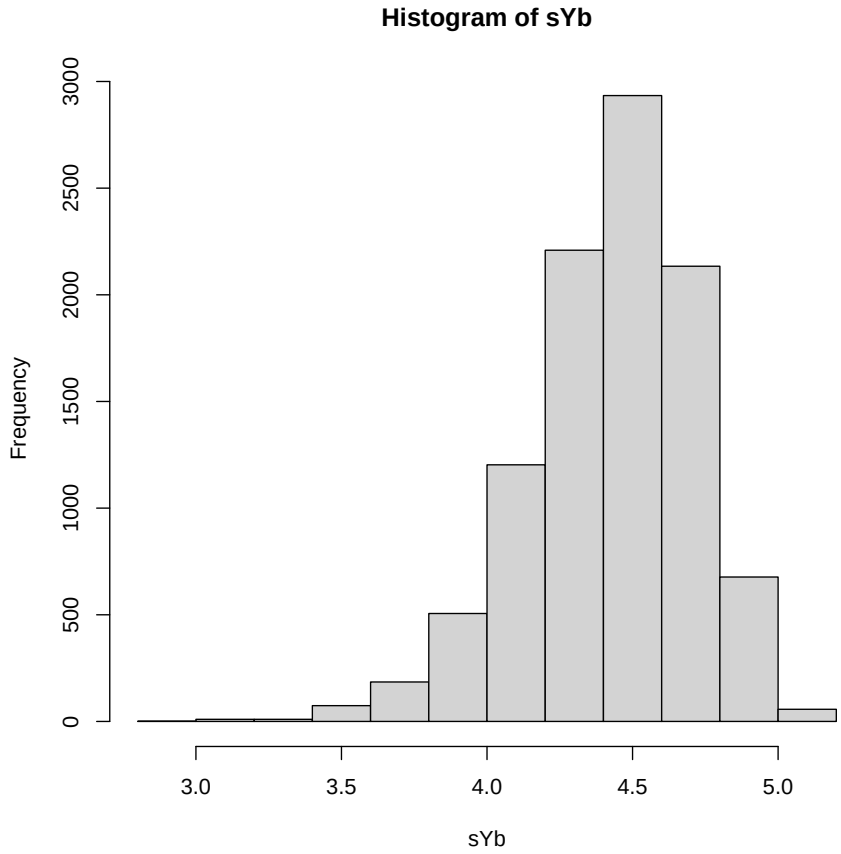
- the BCa intervals are accurate to the second order.

We can also define a property of **coherence**: a confidence interval is coherence when it can vary only within its definition domain. Upper and lower bound are consistent on the parameter space: eg if the parameter is positive then the CI must be positive, if it must be between 0 and 1 so is the CI, and so on (not all are coherent).
Among the considered methods, the percentile and BCa intervals are always coherent. They are generally also more stable and robust.

```
# sample
set.seed(21)
x = c(94,197,16,38,99,141,23)
n = 7

mean(x)

## [1] 86.85714

library(boot)

## again nonparametric bootstrap. we see the montonic property
B = 10000
Xb = matrix(0,B,n)
sXb = matrix(0,B)
sYb = matrix(0,B)

## monotomic property: for each pseudosample we compute
for (i in 1:B) {
    Xb[i,] = sample(x, n, replace=TRUE) # each row contains the pseudosample
    # we compute two estimator
    sXb[i] = mean(Xb[i,])  # the classical mean estimator
    sYb[i] = log(sXb[i])   # its transformation here the logarithm
}

hist(sXb) # histogram of the mean
```




```
hist(sYb) # histogram of the log of the mean
```




```
## We start computing gaussian intervals for Xb and Yb (remember that
## gaussian intervals are not equivariant).
## in order to have them we compute the standard deviation estimate
std.X = sd(sXb)
## interval: we dont use bias here becaus mean is unbiased. TF is the mean
c(mean(x)-1.96*std.X, mean(x)+1.96*std.X)
```

```
## [1]  41.35963 132.35465
```

```
## we repeat for the logarithm to check equivariance
## standard error estimate
std.Y=sd(sYb)
## interval
c(log(mean(x))-1.96*std.Y, log(mean(x))+1.96*std.Y)
```

```
## [1] 3.903327 5.025203
```

```
### percentile intervals
quantile(sXb,c(0.025,0.975))
```

```
##      2.5%     97.5%
##  44.14286 134.42857
```

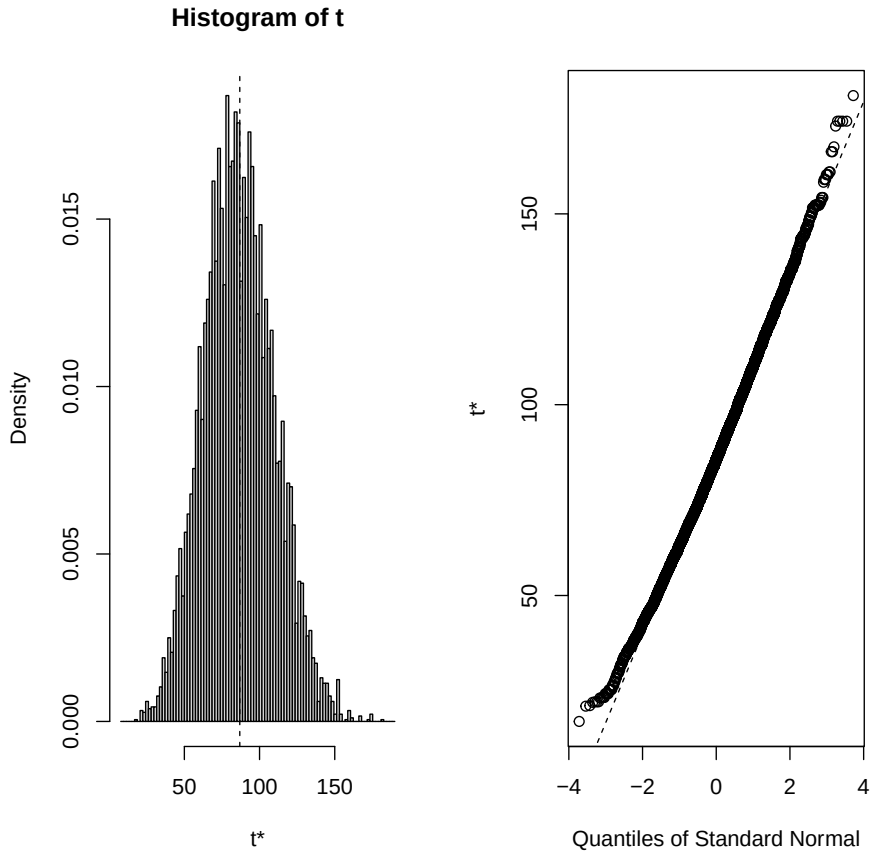
```r
quantile(sYb,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 3.787431 4.901033
```

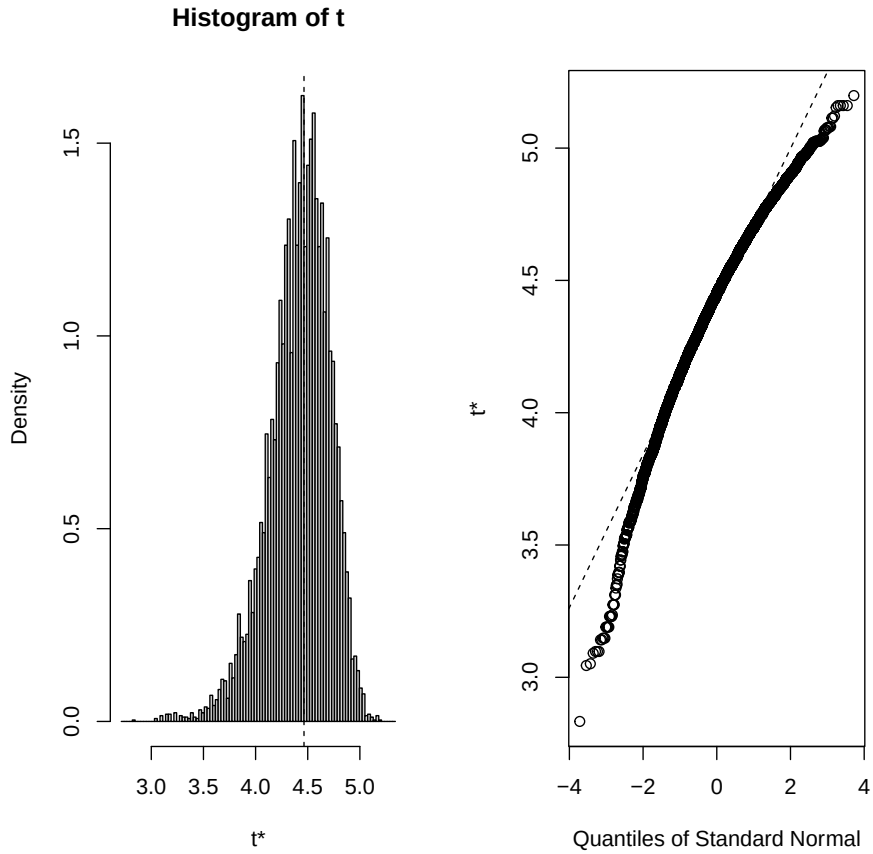
```r
## check of the TRANSFORMING RESPECTING (equivariance) property
exp(quantile(sYb, c(0.025,0.975)))
```

```
##      2.5%     97.5%
##  44.14286 134.42857
```

```r
## check equivariance with the other methods using the function boot::boot (because b
## and other are there automatically implemented)
stima.1 <- function(data) mean(data)
stima.2 <- function(data) log(mean(data))
theta1 <-function(data, i) stima.1(data[i])
theta2 <-function(data, i) stima.2(data[i])

### non parametric BOOTSTRAP
set.seed(1)
stima1.boot <- boot::boot(x, theta1, R=10000)
## reset the seed to have the same extracted indexes and
set.seed(1)
stima2.boot <- boot::boot(x, theta2, R=10000)
plot(stima1.boot)
```




```
plot(stima2.boot)
```




```
## confidence intervals: boot.ci is used to obtain confidence intervals
## it requires the output of boot::boot as input
## the types of ci are in order the gaussian, percentile, the basic and the
## bca. T-student is not implemented here
int.conf.stima1 <- boot::boot.ci(stima1.boot, conf=0.95, type=c("norm","perc","basic"
int.conf.stima2 <- boot::boot.ci(stima2.boot, conf=0.95, type=c("norm","perc","basic"

## if gaussian would be equivariant: the following should be equal
int.conf.stima1$normal[-1]

## [1]  41.72732 132.87400

exp(int.conf.stima2$normal[-1])

## [1]  51.51822 160.01986

## doing the same for bca is definitely better: they are slightly
## different (it depends on resampling error), when n increases they are equal
int.conf.stima1$bca[-c(1,2,3)]

## [1]  46.57143 138.28571
```

```r
exp(int.conf.stima2$bca[-c(1,2,3)])

## [1]  46.28571 137.57143

## all exponentials are as follow and if we check only percentile and
## bca does respect equivariance
exp(int.conf.stima2$normal[-1]) #

## [1]  51.51822 160.01986

exp(int.conf.stima2$basic[-c(1,2,3)])

## [1]  56.54084 173.71429

exp(int.conf.stima2$percent[-c(1,2,3)])

## [1]  43.42857 133.42857

exp(int.conf.stima2$bca[-c(1,2,3)])

## [1]  46.28571 137.57143

## ----------------------------
## t bootstrap: ?bootstrap::boott
## ----------------------------
## we give:
## x data
## stima.1 is the function of the previous bootstrap,
## nbootsd is the number of second level bootrap (n = 25 is enough but here we use 200)
## nboott is the number if first level bootstrap (1000 repetition)
(int.conf.stima1.tboot <- bootstrap::boott(x, stima.1, nbootsd=200, nboott=1000))

## $confpoints
##          0.001      0.01    0.025     0.05      0.1      0.5      0.9    0.95
## [1,] -27.20032 -1.525099 17.21612 37.50721 51.18583 86.85714 130.0624 150.621
##         0.975     0.99    0.999
## [1,] 177.9451 218.4972 681.2087
##
## $theta
## NULL
##
## $g
## NULL
##
## $call
## bootstrap::boott(x = x, theta = stima.1, nbootsd = 200, nboott = 1000)

(int.conf.stima2.tboot <- bootstrap::boott(x, stima.2, nbootsd=200, nboott=1000))

## $confpoints
##         0.001     0.01    0.025     0.05     0.1      0.5      0.9     0.95
## [1,] 2.771402 3.542624 3.758855 3.870192 4.00266 4.464265 4.886389 5.018471
```

```
##         0.975    0.99    0.999
## [1,] 5.129019 5.34457 6.905304
##
## $theta
## NULL
##
## $g
## NULL
##
## $call
## bootstrap::bootT(x = x, theta = stima.2, nbootsd = 200, nboott = 1000)

## it prints confidence bounds according to different levels (eg for
## 95 we take 0.025 and 0.975 values)
```

17.3 Hypothesis testing

Osservazione 364. Idea: various tool to test hypothesis in situation where classical theory cannot be applied. in many strange (maybe n is small or distribution is complicate) situations bootstrap can be useful.

Start immagine we are interested to evaluate an hypotesis on the distribution F of X:

$$H_0 : F = F_0$$

Esempio 17.3.1. We want to check that the expectation of a cumulative distribution function is a certain value

$$H_0 : E_F[X] = \mu_0$$

here we put the F as underscore to say its a characteristics of distribution function of X.
As usual we have observed a sample $\{x_1, \dots, x_n\}$.
We can then construct a test statistic of the data: in order to use bootstrap for hypothesis testing we should construct the statistics under to the null hypothesis. So we construct $T(x_1, \dots, x_n|H_0)$ which depends on the sample and on the value of the null hypothesis (we say under the null hypothesis).
Here bootstrap is used to obtain the distribution of the estimator T under the null via resampling: in the classical approach we use other asymptotic stuff, with bootstrap we don't need it, we reconstruct everything with resampling. If the results we obtained applying the estimator to our sample is rare this can be because of two reason: either we were unlucky or the null is not true.
Generally, $T(x_1, \dots, x_n|H_0)$ is defined so as extremes values (far from 0) will be against H_0 . For instance the statistics could be the standardized sample mean

$$T(x_1, \dots, x_n|H_0) = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Do we know the distribution of this test statistics? No unless in special cases. We don't know the distribution of this statistics unless x is gaussian; if x is

gaussian the statistics is gaussian otherwise we can rely only on clt. But if n is low we can't rely on it either so here it comes the bootstrap, to construct it distribution under the null.

For the rejection we can adopt two strategies as usual:

1. given the size α, the domain space of $T(x_1, \ldots, x_n|H_0)$ is split in two parts: a rejection region and an acceptance region (or not rejection region). Below a two-sided test, uses absolute value:

$$\begin{cases} \text{if } |T(x_1, \ldots, x_n|H_0)| \leq k_\alpha, \implies \overline{R} \\ \text{if } |T(x_1, \ldots, x_n|H_0)| > k_\alpha, \implies R \end{cases}$$

   where k_α is chosen such that $\mathbb{P}(\text{Reject}|H_0) = \alpha$.

2. alternatively the bootstrap p-value will be the probability of tail above the test under the null as usual: given the value of the test statistic on a sample $t_c = T(x_1, \ldots, x_n|H0)$, we can compute the p-value and compare it with the nominal significance value α:

$$p = \mathbb{P}_F \{|T(X_1, \ldots, X_n|H_0)| > |t_c| \, |H_0\}$$

   Therefore

$$\begin{cases} \text{if } p < \alpha, \implies R \\ \text{if } p \geq \alpha, \implies \overline{R} \end{cases}$$

   Typical issues:

- if we know how to construct $T(x_1, \ldots, x_n|H0)$ and if we know its distribution under the null then we can apply Fisher significance theory (we don't need bootstrap)
- if we do not know how to construct $T(x1, \ldots, x_n|H0)$ we can use the simple or the generalized likelihood ratio and we can have as a consequence a known distribution of the likelihood ratio test or an asymptotic known distribution of the LRT (according to Wilks) and we end up with a known distribution under the null (so we are in the context of Neyman-Pearson theory)
- finally if we do not know how to construct $T(x1, \ldots, x_n|H0)$ so we use the generalized likelihood ratio test, we don't have a known distribution of the null. Then we may use asymptotic theory (Wilks theorem, Score test, Wald test)

   Consider

$$T(x_1, \ldots, x_n|H_0) = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

for $H_0 : E_F[X] = \mu_0$ What is the p-value $\mathbb{P}(|T(X_1, \ldots, X_n|F_0)| > |t_c| \, |E_F[X] = \mu_0)$? Imagine 3 classical situations

- $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ then $T(x_1, \ldots, x_n|H_0)$ is distributed as a gaussian if we know the denominator, or a t of Student with $n-1$ degrees of freedom if it's estimated from the sample;
- if X is not Gaussian, but $n \to \infty$ then it is an approximate standard Gaussian thanks to the CLT;
- if X is not Gaussian and n is small we cannot use the CLT (even if you try to work with the Neyman-Pearson approach you will not have a solution, distribution of LRT is not known).
  Well in this case we can use bootstrap

The bootstrap solution procedure. Imagine that the assumptions in assumption regarding F. We cam use parametric bootstrap if F is a probabilistic distribution or a non-parametric bootstrap if F is estimated according to ecdf. We can use both strategies. In both cases:

- generate B bootstrap samples/replications starting from F *under the null* that is we generate $\hat{F}_0 : x^*_{(b)}$. Under the idea that the null is true we replicate data B times using F (wheter its a theorical or empirical cdf)
- then we compute the estimator to obtain B bootstrap values $t^*_{c(b)} = T(x^*_{(b)}|H_0)$
- finally the bootstrap p-value is the number of times the t we have computed in the bootstrap samples is larger than t_c we have computed in the real sample. Once we have p-value we choose to reject the null hypothesis.
  That is compute the bootstrap p-value as following:

$$p = \frac{\sum_{b=1}^{B} \mathbb{1}_{\cdot}\left(\left|t^*_{c(b)}\right| \geq |t_c|\right)}{B}$$

  where $t_c = T(x_1, \ldots, x_n|H_0)$.

The challenge here is how to properly define $\hat{F}_0$ and sampling from it.

Osservazione 365. Parametric bootstrap is easier than non parametric for testing hypotheses

Nonparametric bootstrap

Osservazione 366. In non param bootstrap we should imagine the situation under the null. The way to solve it has not a procedure written in books but depends on the null. eg if its the null is hypothesiss on mean, we should transform our sample so that the mean of the sample is the mean under the null. Once done that the sample respect the null and so we can start bootstrap sampling

If we use a nonparametric boostrap we can start from the empirical distribution function $\hat{F}$. $\hat{F}$ has not the expected value under the null, but it has a different expectation, say μ. We can then introduce a centering transformation (a translation) in the following way:

$$\tilde{x}_1 = x_1 - \mu + \mu_0, \ldots, \tilde{x}_n = x_n - \mu + \mu_0$$

so we change all data according to this, where μ is the mean of our sample and μ_0 is the mean under the null we want check.
Now, the new empirical distribution function $\hat{F}_0$ of $(\hat{x}_1, \dots, \hat{x}_n)$ has $E_{\hat{F}_0}[X] = \mu_0$.
If the hypothesis is on variance we standardize by dividing or multiplying so that the variance of the sample will be the variance of the null: eg if we want to check that the variance is 1 we divide for the sample variance (credo) and start subsampling.
Same if the hypothesis is second moment: we have to find the transformation that make our sample like that, then we start sampling.

Esempio 17.3.2 (Data mouse example)**.** Immagine we want compare the means of two groups (mouses under treatment or no treatment):

$$H_0 : E_{F_A}[Z] = E_{F_B}[Y]$$

where our data is the union of the two groups $X = (Z, Y)$, $Z \sim F_A$ and $Y \sim F_B$.
In this case classical theory suggests to use this test statistic:

$$T(x_1, \dots, x_n | H_0) = \frac{\overline{z} - \overline{y}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

At the numerator we have the difference between means in sample (and difference under null which is zero) while at the denominator the standard error. But how is it distributed? Imagine different situations:

1. in classical inference
   (a)
   (b) If Z and Y are distributed as a Gaussian with known variances σ_A^2 and σ_B^2, then it is a standard normal
   (c) If Z and Y are distributed as a Gaussian with unknown variances but equal (homoschedasticity), then it is distributed according to a t of Student with $n_A + n_b - 2$ degrees of freedom.
   (d) If they are not Gaussian but $n \to \infty$, then it is approximatively standard Gaussian thanks to the CLT.
2. What about the other cases (ex. mouse case)? here we can use bootstrap

With a nonparametric bootstrap we have two alternative solutions:

- stronger hpotheses here: we imagine that the two groups have the same distributon (part of the same population) we reconstruct the F by unifying the two sample and generating data with dimensionality equal to the dimensionality of the sum of dimensionality two groups (eg if the data is composed of 5 A and 7 B we generate data of 12 drawing from the pooled group and assign the first 5 to A and the second 7 to B).
  Ridetto più formalmente Z and Y are considered as a unique sample. Then we construct the empirical distribution function of $\mathbf{x} = z_1, \dots, z_{n_A}, y_1, \dots, y_{n_B}$, from which we get a unique estimate $\hat{F}_0$.
  This is a strong assumption. We are indeed assuming that $F_A = F_B$, from which the null hypothesis is a direct consequence.

- if our the hypothesis is only on the mean (eg we check means are equal), we can apply centering as we did before, so the two samples have the same mean. The strategy is to compute an overall mean first

$$\overline{x} = \frac{n_A \overline{z} + n_B \overline{y}}{n_A + n_B}$$

we normalize both samples subtracting their mean and adding the common mean; in this way the two sample will have the same mean

$$\begin{cases} \tilde{z}_i = z_i - \overline{z} + \overline{x} \\ \tilde{y}_i = y_i - \overline{y} + \overline{x} \end{cases}$$

These two samples will be our population in the bootstrap world and we will start subsampling to reconstruct the null hypothesis distribution.
Then we can apply a balanced nonparametric bootstrap, in order to consider the different sizes of the two samples (when $n_A \neq n_B$). This is a less restrictive solution (but applicable only on hypotheses on mean): we are just assuming that $E_{F_A}[Z] = E_{F_B}[Y]$.

Likelihood ratio test

Osservazione 367. Here something about a more general procedure. Reason why bootstrap is so used everywhere is that it's very general and can be used to recreate distribution even in very complex models (eg neural network).

Osservazione importante 137. The bootstrap is often used to reconstruct the distribution of the log-likelihood ratio when we can not use the Wilks theorem because the regularity conditions do not hold or the sample size is small. In this case it is a parametric bootstrap.

The strategy: if the null hypo is that a parameter of our model is equal to a certain value (eg logistic model), one repeat the estimation of the model B times under the null (with the value of the parameter = the value we want to check

We should re-estimate the model many times by fixing the parameter in the sample equal to the value we want to check before to start subsampling; then we subsample and take something like likelihood ratio test, or likelihood or whatever and at the end we compare the value obtained in the starting sample with the distribution reconstructed under the null.

Below an example for the likelihood ratio test.
If i have an idea for a null hypothesis but no idea about the test statistics, i can use likelihood ratio test: eg in mixture model one could make an hypothesis regarding the number of components.
Anyway how to use likelihood ratio test with bootstrap?

Given the null hypothesis:

$$H_0 : F = F_0(\Theta_0)$$

and the alternative hypothesis (with likelihood ratio test we are in the neyman-pearson approach so we have alternative)

$$H_1 : F = F_1(\Theta_1)$$

with Θ_0, Θ_1 two sets of parameters.
Within our hypotheses we can comput the log ratio is:

$$-2\log\left[\frac{\mathrm{L}\left(\hat{\theta}_0\right)}{\mathrm{L}\left(\hat{\theta}\right)}\right]$$

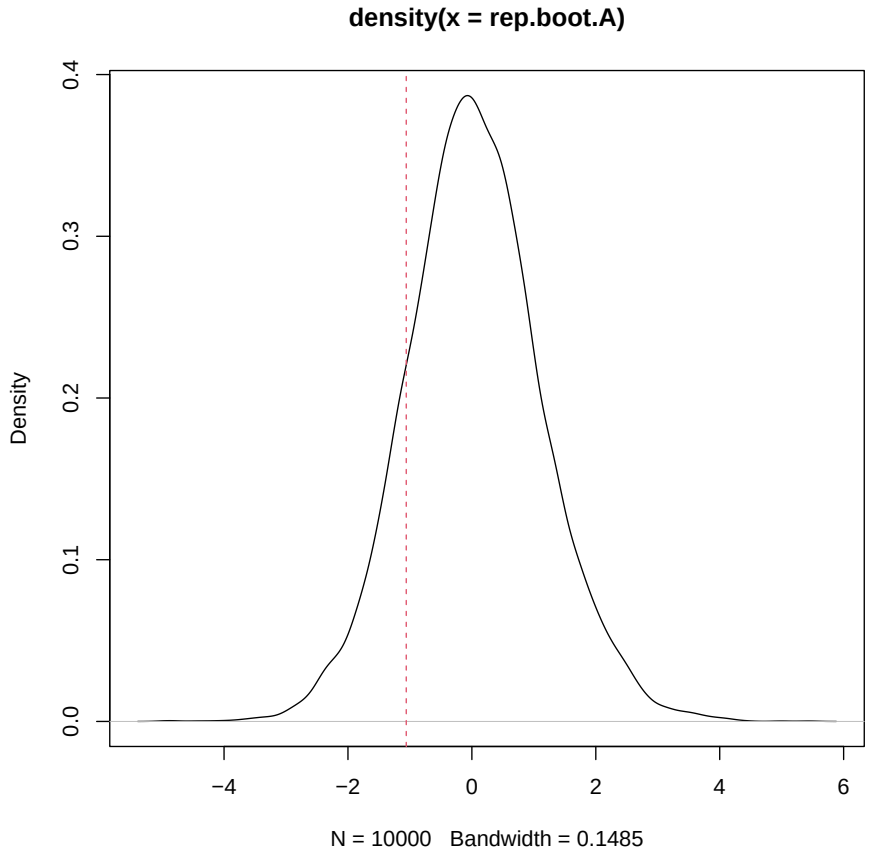
by getting the maximum likelihood estimate for the parameters $\hat{\theta}_0$ (under H_0) and $\hat{\theta}$ (unconstrained model, usually the alternative distribution coincides with the unconstrained model).
The idea is then reconstructing the distribution of the likelihood ratio test:

- the numerator will be different: *parametric bootstrap* will allow us to re-sampling B values under the null $\hat{F}_0 = F_0(\hat{\theta}_0)$ from which we recompute B replications of the numerator $\mathrm{L}\left(()\hat{\theta}_0\right)$.
- the denominator remains the same

At the end we have a distribution for our log-ratio and we collocate our real log-ratio (the log-ratio obtained on our data) in this distribution in order to check if it is rare enough to be considered rejectable.

```
#########################################################################
#
# INSTALL THE PACKAGES BOOT AND BOOTSTRAP
#
#########################################################################

library(boot)
library(bootstrap)


################################################
# part 1: comparison between the mean of 2 groups
################################################

## data of the mouse example are contained in the library bootstrap
## ?mouse.c
## ?mouse.t

## we create a unique dataframe which contain the data
survival <- c(mouse.c, mouse.t)
group <- factor(c(rep("C", 9), rep("T", 7)))
dati.mouse <- data.frame(survival, group)

## Now we want to check if the two groups have or not the same expected
## survival time; we have two strategy

## ------------------------------------------------
## A) the two groups are considered a unique sample
## resample B times from it to create the distribution of the test
```

```r
## below the capital T of the data mouse example.
statistica.test <- function(data){
    media.T <- mean(data$survival[-(1:9)]) # mean in the treated group
    media.C <- mean(data$survival[1:9])    # mean in the control group
    var.T <- var(data$survival[-(1:9)])
    var.C <- var(data$survival[1:9])
    (media.C - media.T)/sqrt(var.T/7 + var.C/9)
}

## the test observed in our sample; t_c in slides.
## since data is in the right order
t.oss <- statistica.test(dati.mouse)

## a glue function to use the boot machinery: here data will be
## shuffled and then put under null hypothesis
theta1 <- function(data, i) statistica.test(data[i,])

## reconstruct the statistic distribution we start with the bootstrap
set.seed(100)
diff.mean.mouse.bootstrap.A <- boot::boot(dati.mouse, theta1, R = 10000)
## extract the B t test computed
rep.boot.A <- diff.mean.mouse.bootstrap.A$t

### grafico: histogram and our sample value t_c as abline
plot(density(rep.boot.A))
abline(v = t.oss, col = 2,lty = 2) # we think the null hyp will be refused
```




```
## p-value of the bootstrap below: here is the formula based on
## indicator functions. here it's a double tailed test
n <- length(rep.boot.A)
(p.value.boot.A <- (sum(abs(rep.boot.A)>=abs(t.oss)))/n)
```

```
## [1] 0.3111
```

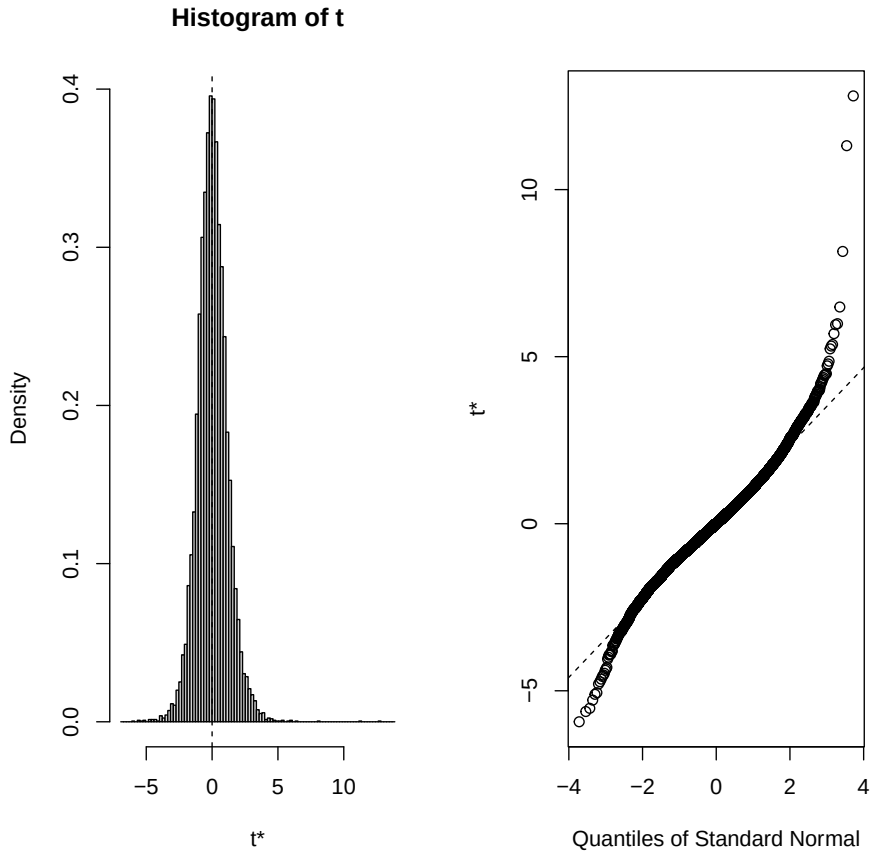
```
## the above is the area before the red line and nella seconda parte
## della curva specularmente.
## essendo p altino we don't reject the null: no effetto survival time

## So the following stats are apparenly different, but the difference
## is not significant
c(mean(mouse.c), mean(mouse.t))
```

```
## [1] 56.22222 86.85714
```

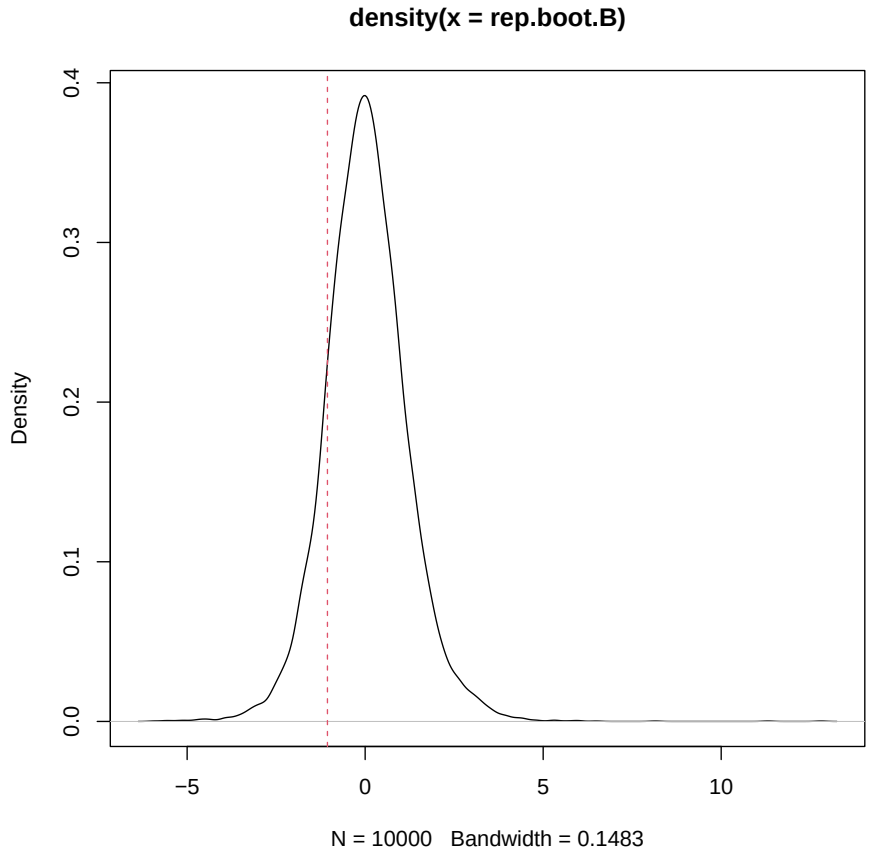
```
## ------------------------------------------------------------
## CASE B:
##
## we recenter the two groups and apply stratification (that is when
```

```r
## we resample we consider that the controls were 9 and treated 7:
## take this number in the bootstrap sampling)
##
## it is necessary to define again the funtion of the statistics of
## interest (by using the groups) case different variances
statistica.test <- function(data){
    n.T <- sum(data$group=="T")
    n.C <- sum(data$group=="C")
    media.T <- mean(data$survival[data$group=="T"])
    media.C <- mean(data$survival[data$group=="C"])
    var.T <- var(data$survival[data$group=="T"])
    var.C <- var(data$survival[data$group=="C"])
    (media.C-media.T)/sqrt(var.T/n.T + var.C/n.C)

    ## or equivalently: since this is exactly the t test, we could
    ## use the already implementad function
    ## t.test(survival ~ group, data = dati.mouse)$statistic
}
theta1 <- function(data, i) statistica.test(data[i,])

## testing hypothesis: first of all the sampling. we create a new
## survival which will be union of normalized data under null hypothesis
survival <- c(mouse.c - mean(mouse.c) + mean(dati.mouse$survival),
              mouse.t - mean(mouse.t) + mean(dati.mouse$survival))
## check della standardizzazione
mean(survival)

## [1] 69.625

mean(dati.mouse$survival)

## [1] 69.625

mean(survival[1:9]) # mean of the first group

## [1] 69.625

mean(survival[10:16])

## [1] 69.625

## new dataframe, under the null, to be used for sampling in bootstrap
dati.mouse.H0 <- data.frame(survival, group)

## bootstrap: to resample in a stratified way (that is resample 9 from
## controls and 7 from treated specify the variable under strata
set.seed(100)
diff.mean.mouse.bootstrap.B <- boot(
    dati.mouse.H0,
    theta1,
    R = 10000,
```

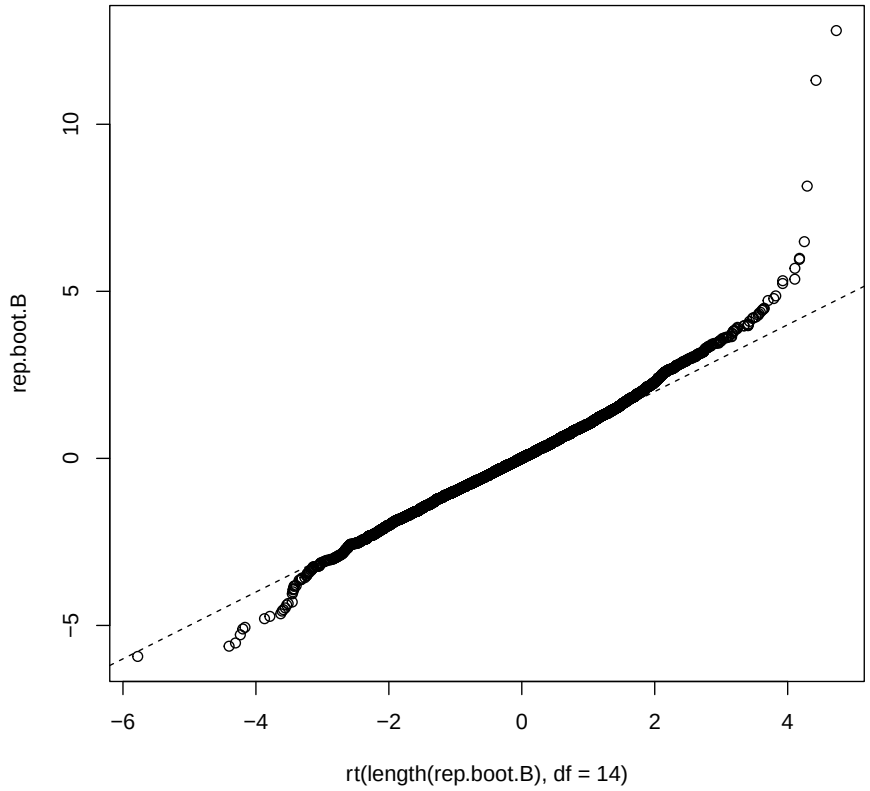
```
    strata = dati.mouse.H0$group)

## distribution of the test statistics under H0
plot(diff.mean.mouse.bootstrap.B)
```




```
rep.boot.B <- diff.mean.mouse.bootstrap.B$t

## grafical representation (observed value is the same as before). we
## have similar results, which is normal in normal circumstances
## p-value follows
plot(density(rep.boot.B))
abline(v = t.oss, col = 2, lty = 2)
```




```
n = length(rep.boot.B)
(p.value.boot.B <- (sum(abs(rep.boot.B) >= abs(t.oss)))/n)

## [1] 0.3088

## comparison with a t student (often done, however t assumes a
## gaussian which is not true): a t of Student with n1+n2-2 d.o.f.
qqplot(rt(length(rep.boot.B), df=14), rep.boot.B)
abline(a=0,b=1,lty=2)
```




```
2*(1-pt(abs(t.oss),df=14)) # t-test value very close but it's by

## [1] 0.3075027

                           # chance

## # p-value
## p.value.boot.B<-(sum(abs(rep.boot.B)>=abs(t.oss)))/10000
## p.value.boot.B
## # comparison:
## p.value.boot.A
## p.value.boot.B
```

17.4 Assignment

Esempio 17.4.1 (Assignment 5 Viroli, bootstrap stuff)**.** The number of volcanic eruptions from 15 active volcanoes in the past year is as follows:

```
[0, 0, 4, 3, 0, 5, 2, 0, 1, 4, 3, 9, 0, 3, 0]
```

1. Compute the 95% bootstrap confidence interval for the eruptions of the volcanoes, taking into account the nature of the data.
2. We want to estimate the variability in the number of eruptions (in terms of variance) using the corrected sample variance. Reconstruct the distribution of the variance through a non-parametric bootstrap. In particular, obtain a bootstrap estimate of the variance and its bias (B=10000).
3. Assuming that the number of eruptions follows a Poisson distribution, repeat procedure b) with a parametric bootstrap.
4. By comparing the biases and the distribution plots of variances, can we conclude whether a parametric or non-parametric bootstrap is more reliable for estimating the population variance through the corrected sample variance?
5. Now, we want to check if the Poisson probabilistic model is suitable for describing the data (distributional conformity test). One way to do this is by testing the null hypothesis:

$$H_0 : h(Y) = \mathrm{Var}[Y] - \mathbb{E}[Y] = 0$$

   To achieve this, reconstruct the distribution of the function h(Y) under the null hypothesis and calculate the bootstrap p-value (where sample mean and corrected sample variance can be used as estimators for the two quantities).

```r
x <- c(0, 0, 4, 3, 0, 5, 2, 0, 1, 4, 3, 9, 0, 3, 0)
B <- 10000

## ------------------------------------------------------------------------
## A) given the count nature of data (positive or null) we opt for a
##    coherent confidence interval (that is: can't be negative). Available
##    methods are percentile and bca; I apply a bca one which is a better
##    generalization of the first.

stat_a <- function(data, i) mean(data[i])
set.seed(1)
boot_a <- boot::boot(data = x, statistic = stat_a, R = B)
boot::boot.ci(boot_a, conf = 0.95, type = 'bca')

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot::boot.ci(boot.out = boot_a, conf = 0.95, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 1.2,  3.8 )
## Calculations and Intervals on Original Scale
```

```r
## ------------------------------------------------------------------------
## B)

stat_b <- function(data, i) var(data[i])
set.seed(1)
boot_b <- boot::boot(data = x, statistic = stat_b, R = B)
bstats_b <- boot_b$t
## estimate and bias
estimate_b <- mean(bstats_b)
TF.variance <- var(x) # DUBBI QUI, forse dovrebbe essere la versione
                      # non corretta, con n al denominatore
bias_b <- estimate_b - TF.variance
c('Estimate' = estimate_b, 'Bias' = bias_b)
```

```
##   Estimate       Bias
##  6.2085238 -0.4295714
```

```r
## ------------------------------------------------------------------------
## C)

lambda_c <- mean(x)
stat_c <- function(data) var(data)
rg_c <- function(data, lambda) rpois(length(data), lambda = lambda)
set.seed(1)
boot_c <- boot::boot(data = x,
                     statistic = stat_c,
                     R = B,
                     sim = 'parametric',
                     ran.gen = rg_c,
                     mle = lambda_c)
bstats_c <- boot_c$t
## estimate and bias
estimate_c <- mean(bstats_c)
bias_c <- estimate_c - TF.variance
c('Estimate' = estimate_c, 'Bias' = bias_c)
```

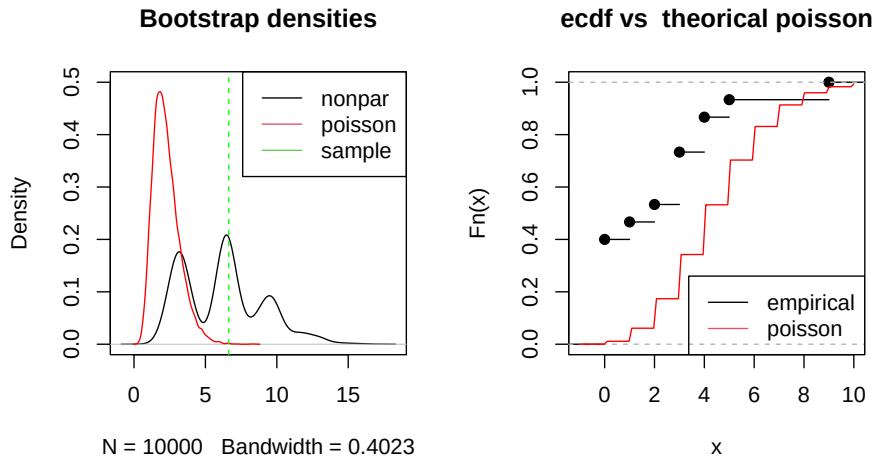
```
##  Estimate      Bias
##  2.264464 -4.373631
```

```r
## ------------------------------------------------------------------------
## D) We can conclude a non parametric bootstrap is better for these reasons:
##
##    1) bias comparison: bias is 10x higher for the poisson based estimate
c('bias_np' = bias_b, 'bias_pois' = bias_c)
```

```
##    bias_np  bias_pois
## -0.4295714 -4.3736314
```

```r
##    2) the distribution of bootstrap estimated variances is better
##       centered (as coherent with bias) with our sample value
par(mfrow = c(1,2))
```

```
plot(density(bstats_b), col = 'black', ylim = c(0,0.5), main = 'Bootstrap densities')
lines(density(bstats_c), col = 'red')
abline(v = var(x), col = 'green', lty = 'dashed')
legend('topright', col = 1:3, lty = 1,
       legend = c("nonpar", "poisson", "sample"))

##    3) ecdf and theorical poisson distributions are quite different ..
plot(ecdf(x), main = 'ecdf vs  theorical poisson')
curve(ppois(x, lambda = mean(x)), add = TRUE, col = 'red')
legend('bottomright', col = 1:2, lty = 1, legend = c("empirical", "poisson"))
```




```
## ------------------------------------------------------------
## E) the null hypothesis is satisfied if var(y) = E(y). So eg
##    let's force the expected value to be like the variance

h_obs <- var(x) - mean(x)
stat_e <- function(data, i){
    sel <- data[i]
    var(sel) - mean(sel)
}

## null-ification of the sample data
c('mean(x)' = mean(x), 'var(x)' = var(x))

##  mean(x)   var(x)
## 2.266667 6.638095

x2 <- x + (var(x) - mean(x))
c('mean(x2)' = mean(x2), 'var(x2)' = var(x2))

## mean(x2)  var(x2)
## 6.638095 6.638095
```

```r
## now lets bootstrap the statistic
set.seed(1)
boot_e <- boot::boot(data = x2, statistic = stat_e, R = B)
bstats_e <- boot_e$t

## basically we're interested in a two tail test in order to detect
## any difference between mean and var (in any direction)
n <- length(bstats_e)
(pval_e <- sum(abs(bstats_e) >= abs(h_obs))/n)

## [1] 0.0428

## we reject the null (so mean != var) despite being not to far from
## alpha = 0.05
```

Esempio 17.4.2 (Esame vecchio viroli)**.** The dataset women in R (`data(women)`) contains weight andheight of 12 american women aged 30-39. Compuite teh body mass index as

$$bm = 703 \cdot \frac{w}{h^2}$$

where w is weight and h is the height. Use a non parametric bootstrap with $B = 50000$ to estimate the bias of the mean when it is used to estimate the population median of bm

1. about 0
2. about 0.445
3. about 0.263 direi questo
4. about 0.210

```r
bm <- 703 * (women$weight/ (women$height)^2)

## Error in women$weight:  $ operator is invalid for atomic vectors

e <- median(bm)

## Error in eval(expr, envir, enclos):  oggetto 'bm' non trovato

boot_f <- function(data, i) mean(data[i])
res <- boot::boot(data = bm, statistic = boot_f, R=50000)

## Error in eval(expr, envir, enclos):  oggetto 'bm' non trovato

mean(res$t)

## Error in eval(expr, envir, enclos):  oggetto 'res' non trovato

(bias <- mean(res$t) - e)

## Error in eval(expr, envir, enclos):  oggetto 'res' non trovato
```

Esempio 17.4.3 (Esame vecchio viroli)**.** The dataset women in R (`data(women)`) contains weight andheight of 12 american women aged 30-39. Compuite teh body mass index as

$$bm = 703 \cdot \frac{w}{h^2}$$

where w is weight and h is the height. Use a non parametric bootstrap with $B = 50000$ to estimate the MSE of the median when it is used to estimate the population median of

1. about 0.075 suggerita da taluni, confermata sotto

```
bm <- 703 * (women$weight/ (women$height)^2)

## Error in women$weight:  $ operator is invalid for atomic vectors

e <- median(bm)

## Error in eval(expr, envir, enclos):  oggetto 'bm' non trovato

boot_f <- function(data, i) median(data[i])
res <- boot::boot(data = bm, statistic = boot_f, R=50000)

## Error in eval(expr, envir, enclos):  oggetto 'bm' non trovato

MSE <- sum((res$t-e)^2)/50000

## Error in eval(expr, envir, enclos):  oggetto 'res' non trovato
```