Calcolo di Probabilità e Statistica

A.A. 2021-2022

Docente: A. Buonocore

Dispense tratte dal libro consigliato a cura dello studente **S. Cerrone**
e integrate con gli appunti delle lezioni dello studente **V. Manno**

Programma

N.B.: Questi appunti sono praticamente i capitoli del libro senza esempi con alcune integrazioni. Se volete appunti del corso, esercizi svolti o altro, trovate tutto al link seguente: https://urly.it/3jsff.

1. Elementi di probabilità

Introduzione

Il concetto di probabilità di un evento, quando si effettua un esperimento, è passabile di due interpretazioni: interpretazione frequentista e interpretazione soggettivistica (o personale).

Nell'*interpretazione frequentista* la probabilità di un esito è considerata una proprietà dell'esito stesso. In particolare, si pensa che essa possa essere determinata operativamente ripetendo in continuazione l'esperimento, come rapporto tra il numero di casi in cui si è registrato l'esito sul totale.

Nell'*interpretazione soggettivistica*, non si crede che la probabilità di un esito sia una proprietà oggettiva, ma piuttosto la precisazione del livello di fiducia che lo studioso ripone nel verificarsi dell'esito.

Spazio degli esiti ed eventi

Si consideri un esperimento il cui esito non sia prevedibile con certezza. Quello che normalmente si può fare comunque, è individuare la rosa degli esiti plausibili. L'insieme di tutti gli esiti si dice spazio degli esiti o **spazio campione** e normalmente si denota con S o con $\boldsymbol{\Omega}$.

Di seguito alcuni esempi:

1) Se l'esito dell'esperimento consiste nella determinazione del sesso di un neonato, allora poniamo $\Omega = \{f, m\}$ dove si intende che l'esito f rappresenta la nascita di una femmina, e l'esito m quella di un maschio.
2) Se l'esperimento consiste in una gara tra sette cavalli denotati dai numeri $1,2,3,4,5,6$ e 7, allora $\Omega = \{\text{tutti gli ordinamenti di } (1,2,3,4,5,6,7)\}$. In questo caso l'esito $(2,3,7,6,5,4,1)$ è quello in cui il cavallo 2 arriva primo, il 3 secondo, il 7 terzo, e così via.

I sottoinsiemi dello spazio degli esiti si dicono **eventi**, quindi un evento E è un insieme i cui elementi sono esiti possibili. Se l'esito dell'esperimento è contenuto in E, diciamo che l'evento E si è verificato.

La **unione $\boldsymbol{A \cup B}$** di due eventi dello stesso spazio campione, è definita come l'insieme degli esiti che stanno o in A o in B. Quindi l'evento $A \cup B$ si verifica se *almeno uno* tra A e B si verifica.

- Nell'esempio 1, se $A = \{f\}$ e $B = \{m\}$, allora $A \cup B = \{f, m\}$ coincide con l'intero spazio campione

L'**intersezione $\boldsymbol{A \cap B}$** di due eventi è l'insieme formato dagli esiti che sono presenti sia in A che in B. Come evento rappresenta il verificarsi di *entrambi* gli eventi.

- Nell'esempio 2, se $A = \{\text{tutti gli esiti che terminano con } 5\}$ e $B = \{(5,1,2,3,4,6,7)\}$ allora chiaramente l'evento non contiene esiti possibili e non può avvenire mai. *Evento vuoto*.

Se $A \cap B = \emptyset$, ovvero se A e B non possono verificarsi entrambi, li diremo **eventi mutuamente esclusivi** o **eventi disgiunti**.

Per ogni evento E, definiamo l'**evento complementare** di E (in simboli E^c) come l'insieme formato dagli esiti di Ω che non stanno in E. Quindi $E^c = \Omega \setminus E$.

Se, per una coppia di eventi A e B accade che tutti gli esiti di A appartengono anche a B, si dice che A è contenuto in B, e si scrive $A \subset B$ (o, equivalentemente $B \supset A$).

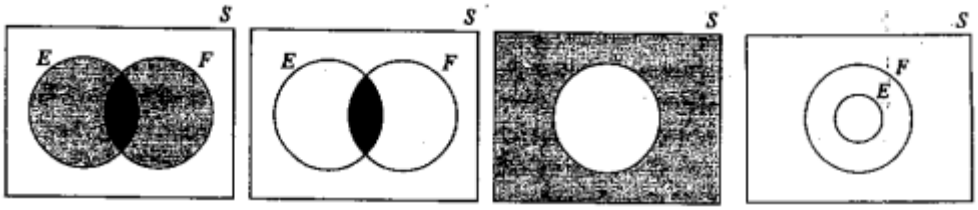
È anche possibile definire l'unione o l'intersezione di più di due eventi:

$$\bigcup_{i=1}^{n} E_i = E_1 \cup E_2 \cup \ldots \cup E_n \qquad \bigcap_{i=1}^{n} E_i = E_1 \cap E_2 \cap \ldots \cap E_n$$

In altre parole, l'unione degli E_i si verifica se *almeno uno* degli eventi E_i si verifica, mentre l'intersezione degli E_i si verifica solo se *tutti* gli eventi E_i si verificano.

I diagrammi di Venn e l'algebra degli eventi

Un tipo di rappresentazione grafica degli eventi, molto utile per illustrare le relazioni logiche che li legano, sono i **diagrammi di Venn**. Lo spazio degli esiti S è rappresentato da un grande rettangolo che contiene il resto della figura, oppure dal foglio stesso. Gli eventi da prendere in considerazione, invece, sono rappresentati da cerchi. A questo punto, tutti gli eventi complessi di nostro interesse possono essere evidenziati colorando opportune regioni del diagramma. Di seguito rappresentiamo in ordine gli insiemi $E \cup F, E \cap F, E^c$ e $E \subset F$:



Elenchiamo alcune proprietà sugli operatori unione e intersezione e complementare:

- Commutativa: $A \cup B = B \cup A$ $\quad A \cap B = B \cap A$
- Associativa: $(A \cup B) \cup C = A \cup (B \cup C)$ $\quad (A \cap B) \cap C = A \cap (B \cap C)$
- Distributiva: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ $\quad (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$
- **Leggi di De Morgan**: $(A \cup B)^c = A^c \cap B^c$ $\quad (A \cap B)^c = A^c \cup B^c$

Per dimostrare queste proprietà basta un approccio intuitivo che sfrutti i diagrammi di Eulero Venn. Infatti, basta disegnare i diagrammi sia a sinistra che a destra dell'equivalenza e compararli.

Assiomi della probabilità

Si associa ad ogni evento E sullo spazio campione Ω, un numero che si denota con $\mathcal{P}(E)$ e che si dice **probabilità** dell'evento E. Le probabilità dei vari eventi devono rispettare alcuni assiomi:

1) Ogni probabilità è un numero compreso tra 0 e 1: $\boldsymbol{0 \leq \mathcal{P}(E) \leq 1}$

2) L'evento Ω si verifica con probabilità 1. Ovvero, vi è assoluta certezza che si realizzi un esito contenuto nello spazio campione: $\boldsymbol{\mathcal{P}(\Omega) = 1}$ (tale evento viene denominato *evento certo*)

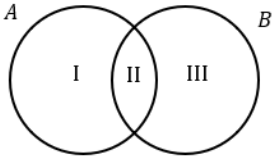
3) Preso un insieme finito o numerabile di eventi mutuamente esclusivi, la probabilità che se ne verifichi almeno uno è uguale alla somma delle loro probabilità:

$$\boldsymbol{\mathcal{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathcal{P}(E_i)\,, \forall n \in \mathbb{N}}$$

Tale proprietà prende il nome di *finita additività;* nel caso di una successione numerabile di elementi di una σ-algebra si parla di σ-additività (vedi "Assiomi di Kolmogorov").

$\mathcal{P}(E)$ può essere interpretato come la frequenza relativa dell'evento E quando l'esperimento è ripetuto un gran numero di volte. Inoltre, i precedenti assiomi permettono di dedurre un gran numero di proprietà delle probabilità degli eventi:

- $\boldsymbol{1 = \mathcal{P}(\Omega) = \mathcal{P}(E \cup E^c) = \mathcal{P}(E) + \mathcal{P}(E^c)}$
  Essendo E ed E^c eventi disgiunti, possiamo applicare gli assiomi 2 e 3.
- $\boldsymbol{\mathcal{P}(E^c) = 1 - \mathcal{P}(E)}$
  La probabilità che un evento qualsiasi non si verifichi è pari a uno meno la probabilità che si verifichi.
- $\boldsymbol{\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)}$
  Si dimostra facilmente con i diagrammi di Venn. Poiché le regioni I, II e III in figura sono disgiunte, si può applicare tre volte l'assioma 3 per ottenere $\mathcal{P}(A \cup B) = \mathcal{P}(\mathrm{I}) + \mathcal{P}(\mathrm{II}) + \mathcal{P}(\mathrm{III})$, $\mathcal{P}(A) = \mathcal{P}(\mathrm{I}) + \mathcal{P}(\mathrm{II})$ e $\mathcal{P}(B) = \mathcal{P}(\mathrm{II}) + \mathcal{P}(\mathrm{III})$; ora è evidente che sommare e sottrarre la stessa quantità non cambi il contributo. Dunque:



$$\mathcal{P}(A \cup B) = \mathcal{P}(\mathrm{I}) + \mathcal{P}(\mathrm{II}) + \mathcal{P}(\mathrm{III}) + \mathcal{P}(\mathrm{II}) - \mathcal{P}(\mathrm{II}) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$$

- $\boldsymbol{\mathcal{P}(\emptyset) = 0}$
  Scriviamo l'insieme $\emptyset$ come unione di un numero infinito di insiemi $\emptyset$. Applicando l'assioma 3 risulta $\mathcal{P}\left(\bigcup_{i=1}^{\infty} \emptyset_i\right) = \sum_{i=1}^{\infty} \mathcal{P}(\emptyset_i)$ ma allora è evidente che ogni $\mathcal{P}(\emptyset_i) = 0$ altrimenti per qualsiasi altro valore la serie divergerebbe contraddicendo l'assioma 1
- $\boldsymbol{\mathcal{P}(B) = \mathcal{P}(B \cap A) + \mathcal{P}(B \cap A^c)}$
  Essendo banalmente $B = B \cap \Omega = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c)$. Questa proprietà rende possibile calcolare la probabilità di un evento B tramite l'ausilio di un evento ausiliario A.
- $\boldsymbol{A \cup B = A \cup (A^c \cap B) \Rightarrow \mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(A^c \cap B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)}$
  Questa formula permette di calcolare la probabilità dell'unione di due eventi A e B, a patto di conoscere già le probabilità degli eventi A, B e della loro intersezione. Per tre eventi:
  $\boldsymbol{\mathcal{P}(A \cup B \cup C)} = \mathcal{P}\big((A \cup B) \cup C\big) = \mathcal{P}(A \cup B) + \mathcal{P}(C) - \mathcal{P}\big((A \cup B) \cap C\big) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B) + \mathcal{P}(C) - \mathcal{P}\big((A \cap C) \cup (B \cap C)\big) = \mathcal{P}(A) + \mathcal{P}(B) + \mathcal{P}(C) - \mathcal{P}(A \cap B) - [\mathcal{P}(A \cap C) + \mathcal{P}(B \cap C) - \mathcal{P}(A \cap B \cap C)] =$
  $$= \boldsymbol{\mathcal{P}(A) + \mathcal{P}(B) + \mathcal{P}(C) - \mathcal{P}(A \cap B) - \mathcal{P}(A \cap C) - \mathcal{P}(B \cap C) + \mathcal{P}(A \cap B \cap C)}$$
- $\boldsymbol{\mathcal{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \mathcal{P}\left(\bigcap_{i=1}^{n} A_i^c\right)}$
  Sia $\{A_n\}_n$ una successione numerabile di eventi. Poiché $\bigcup_{i=1}^{n} A_i = \left(\bigcap_{i=1}^{n} A_i^c\right)^c$, evidentemente: $\mathcal{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathcal{P}\left(\left(\bigcap_{i=1}^{n} A_i^c\right)^c\right) = 1 - \mathcal{P}\left(\bigcap_{i=1}^{n} A_i^c\right)$. Grazie a questa proprietà possiamo calcolare la probabilità di un unione come la probabilità di un'intersezione (più facilmente calcolabile).

Assiomi di Kolmogorov

Una famiglia $\mathcal{A}$ di sottoinsiemi di Ω è detta una **σ-algebra degli eventi** se verifica le seguenti proprietà:

1)	$\Omega \in \mathcal{A}$	l'*evento certo* è un evento ($A \in \mathcal{A}$ = "A è un evento")
2)	$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$	L'*evento contrario* è un evento
3)	Se $\{A_n\}_{n\in\mathbb{N}}$ è una successione di elementi di $\mathcal{A}$, allora: $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$	$\mathcal{A}$ è stabile rispetto all'unione numerabile

Dalla precedente definizione si possono effettuare le seguenti osservazioni sugli assiomi di Kolmogorov:

- Dagli assiomi 1 e 2 segue che $\emptyset = \Omega^c \in \mathcal{A}$ (l'evento impossibile è un evento)
- Dagli assiomi 2 e 3 e dalle leggi di De Morgan si dimostra anche la stabilità rispetto l'intersezione numerabile: sia $\{A_n\}_{n\in\mathbb{N}}$ una successione numerabile di eventi in $\mathcal{A}$ risulta: $\bigcap A_n = (\bigcup A_n^c)^c$

Diamo due proprietà che si basano sulla σ-additività:

- **La σ-additività comporta anche la finita additività**
  Siano $B_1 = A_1, B_2 = A_2, \dots, B_n = A_n, B_{n+1} = B_{n+2} = \dots = \emptyset$ sequenze di eventi disgiunti. Si ha:
  $$\mathcal{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathcal{P}(B_i) = \sum_{i=1}^{n} \mathcal{P}(B_i) + \sum_{i=n+1}^{\infty} \mathcal{P}(B_i) = \sum_{i=1}^{n} \mathcal{P}(B_i) + 0 = \mathcal{P}\left(\bigcup_{i=1}^{n} B_i\right)$$
- **Proprietà di passaggio al limite sulle successioni crescenti**: Sia $\{A_n\}_{n\in\mathbb{N}}$ una successione crescente di eventi (cioè, $\forall n \in \mathbb{N}$, risulta $A_n \subseteq A_{n+1}$). Si ha allora $\mathcal{P}(\bigcup A_n) = \mathcal{P}\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \mathcal{P}(A_n)$
- **Proprietà di passaggio al limite sulle successioni decrescenti**: Sia $\{A_n\}_{n\in\mathbb{N}}$ una successione decrescente di eventi ($\forall n \in \mathbb{N}$, $A_n \supseteq A_{n+1}$). Si ha allora $\mathcal{P}(\bigcap A_n) = \mathcal{P}\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \mathcal{P}(A_n)$

Spazi di esiti equiprobabili

Se Ω è un insieme finito è naturale assumere che ogni evento abbia la stessa probabilità di realizzarsi. Dunque, se $\Omega = \{1,2,\dots,N\}$, l'equiprobabilità degli esiti si scrive $\mathcal{P}(\{1\}) = \mathcal{P}(\{2\}) = \dots = \mathcal{P}(\{N\}) =: p$. Dagli assiomi 2 e 3 segue che $1 = \mathcal{P}(\Omega) = \mathcal{P}(\{1\}) + \mathcal{P}(\{2\}) + \dots + \mathcal{P}(\{N\}) = \sum_{i=1}^{N} p$; da cui si deduce che $\mathcal{P}(\{i\}) = p = \frac{1}{N}$. Da questo risultato e ancora dall'assioma 3 si conclude che per ogni evento E, $\mathcal{P}(E) = \frac{|E|}{|\Omega|}$.

In altre parole, se si assume che ogni esito di Ω abbia la medesima probabilità, allora la probabilità di un qualunque evento E è pari al rapporto tra il numero di esiti contenuti in E e il numero totale di esiti di Ω.

Principio di enumerazione

Consideriamo la realizzazione di due diversi esperimenti (detti 1 e 2), che possono avere rispettivamente m e n esiti differenti. Allora complessivamente vi sono mn diversi risultati se si considerano entrambi gli esperimenti contemporaneamente.

L'enunciato si dimostra enumerando tutte le possibili coppie di risultati dei due esperimenti, che sono:

$$\begin{matrix} (1,1) & \cdots & (1,n) \\ \vdots & \ddots & \vdots \\ (m,1) & \cdots & (m,n) \end{matrix}$$

dove si intende che si ottiene il risultato (i,j) se nell'esperimento 1 si realizza l'esito i-esimo tra gli m possibili, e nell'esperimento 2 quello j-esimo tra gli n possibili. Siccome la tabella ottenuta ha m righe e n colonne, vi sono complessivamente mn esiti possibili.

Generalizzazione del principio di enumerazione: Se si eseguono r esperimenti, ed è noto che il primo esperimento ammette n_1 esiti possibili, per ognuno dei quali il secondo esperimento ammette n_2 esiti diversi, inoltre se per ogni combinazione di esiti dei primi due esperimenti il terzo ammette n_3 esiti diversi, e così via, allora vi sono un totale di $n_1 \cdot n_2 \cdot \ldots \cdot n_r$ combinazioni di esiti degli r esperimenti considerati tutti insieme.

Per illustrarne un'applicazione, proviamo a determinare il numero di modi diversi in cui si possono ordinare n oggetti. Per esempio, il numero di modi in cui si possono ordinare i tre simboli a, b e c sono sei; ovvero, abc, acb, bac, bca, cab e cba. Ciascuno di questi ordinamenti prende il nome di ***permutazione*** dei tre simboli considerati. Questo risultato è facilmente deducibile dal principio di enumerazione generalizzato; infatti, il primo simbolo può essere scelto in tre modi diversi e per ogni scelta del primo abbiamo due modi per scegliere il secondo ed una sola scelta per il terzo (quest'ultimo è scelto per esclusione). Quindi, vi sono $3 \cdot 2 \cdot 1 = 6$ possibili permutazioni.

In modo analogo, supponendo di avere n oggetti scopriamo che vi sono $n(n-1)(n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = n!$ diverse permutazioni degli n oggetti.

Il coefficiente binomiale

Vogliamo determinare il numero di diversi gruppi di r oggetti che si possono formare scegliendoli da un insieme di n. Ad esempio, quanti diversi gruppi di tre lettere si possono formare usando le cinque lettere A, B, C, D, E? Si può ragionare nel modo seguente. Vi sono 5 scelte per la prima lettera, 4 per la seconda e 3 per la terza, vi sono quindi $5 \cdot 4 \cdot 3$ modi di scegliere tre lettere su cinque, tenendo conto dell'ordine. Tuttavia, ogni gruppo di tre lettere viene contato più volte, perché stiamo tenendo conto dell'ordine. Ad esempio, la tripletta A, C, D compare in tutte le sue 6 permutazioni. Poiché stiamo contando $3!$ volte ogni gruppo di tre lettere, se ne deduce che il numero di gruppi diversi di tre lettere può essere ricavato come

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$$

Più in generale, poiché il numero di modi diversi di scegliere r oggetti su n tenendo conto dell'ordine è dato da $n(n-1) \ldots (n-r+1)$, e poiché ogni gruppo di lettere fissato viene contato $r!$ volte (una per ogni sua permutazione), il numero di diversi gruppi di r elementi, scelti in un insieme di n oggetti è dato dalla formula del *coefficiente binomiale*:

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!} = \frac{n(n-1) \ldots (n-r+1)}{r!}$$

Questo valore si dice il numero di ***combinazioni*** di n elementi presi r alla volta.

Proprietà del binomiale: Siccome $0! = 1$, si noti che vale $\binom{n}{0} = 1 = \binom{n}{n}$, inoltre $\binom{n}{r} = \binom{n}{n-r}$ per $0 \leq r \leq n$

Caso particolare del coefficiente binomiale si ha quando le combinazioni sono con *ripetizione*. Il numero di quest'ultimo è dato da $\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$

Introduciamo anche il **coefficiente multinomiale** indicato con il simbolo $\binom{m}{n_1,n_2,\dots,n_r} = \frac{m!}{n_1!\cdot n_2!\cdot\dots\cdot n_r!}$
Questo è utile per calcolare in quanti modi si possono formare r gruppi di un insieme di m elementi in modo che il gruppo i-esimo contenga esattamente n_i elementi. Ciò significa che $n_1 + n_2 + \cdots n_r = m$

Probabilità condizionata

L'importanza che ha il concetto di probabilità condizionata è duplice. In primo luogo, accade spesso di volere calcolare delle probabilità quando si è in possesso di informazioni parziali sull'esito dell'esperimento, o di volerle ricalcolare una volta ottenute nuove informazioni. Quelle di questo tipo sono probabilità condizionate. Secondariamente vi è una sorta di bonus nel fatto che a volte il modo più semplice di determinare la probabilità di un evento complesso, consiste nel condizionarlo al realizzarsi o meno di un evento accessorio.

Per illustrare questo concetto, immaginiamo di tirare due dadi. Lo spazio campione di questo esperimento può essere descritto da $\Omega = \{(i,j)|i = 1,2,\dots,6 \wedge j = 1,2,\dots,6\}$. Supponiamo che ciascuno dei 36 esiti di Ω abbia la stessa probabilità, ovvero $1/36$ (in queste ipotesi si dice che i due dadi siano *onesti*). Supponiamo infine che il primo dado sia risultato in un 3. Allora, possedendo questa informazione, qual è la probabilità che la somma dei due dadi valga 8? Dato che il primo dado ha totalizzato un 3, vi sono solo 6 risultati possibili per l'esperimento, che sono $(3,1)$, $(3,2)$, $(3,4)$, $(3,5)$ e $(3,6)$. Inoltre, siccome in origine ciascuno di questi esiti aveva la stessa probabilità di realizzarsi, essi dovrebbero essere ancora equiprobabili. Ciò significa che, se il primo dado ha dato un 3, allora la probabilità (condizionata) di ciascuno degli esiti possibili $(3,j)$ è $1/6$, mentre la probabilità (condizionata) degli altri 30 elementi di Ω è 0 . Se ne conclude che la probabilità cercata è $1/6$.

Se denotiamo con E e F rispettivamente l'evento che la somma dei due dadi valga 8 e l'evento che il primo dado risulti in un 3, allora la probabilità che abbiamo appena calcolato si dice *probabilità condizionata di E dato F*, e si denota con $\boldsymbol{\mathcal{P}(E|F)}$.

Con un ragionamento analogo a quello dell'esempio è possibile trovare una formula generale per $\mathcal{P}(E|F)$, valida per qualunque coppia di eventi. Infatti, se si è verificato l'evento F, affinché si verifichi anche E, il caso deve appartenere all'intersezione $E \cap F$, ovvero favorire un elemento che sta sia in E sia in F. In secondo luogo, essendosi verificato F, questo evento diviene il nuovo (ridotto) spazio campione e per questo la probabilità condizionata dell'evento $E \cap F$ sarà pari al rapporto tra la sua probabilità e quella di F

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

Si noti che tale equazione ha senso solo se $\mathcal{P}(F) > 0$ e infatti in caso contrario $\mathcal{P}(E|F)$ non si definisce.

La precedente equazione può essere scritta anche come $\boldsymbol{\mathcal{P}(E \cap F) = \mathcal{P}(E|F)\mathcal{P}(F)}$. Parafrasandola, dice che la probabilità che E e F si verifichino entrambi è pari quella che si verifichi F per la probabilità condizionata di E dato che si è verificato F. Questa formula mostra la sua utilità quando si vuole calcolare la probabilità di una intersezione.

Compatibilità con l'interpretazione frequentista della probabilità degli eventi

Supponiamo di realizzare un numero molto elevato n di ripetizioni di un esperimento. Poiché $\mathcal{P}(F)$ è il limite della frazione di prove in cui si verifica F, su un numero elevato n di tentativi, saranno circa $n\mathcal{P}(F)$ quelli in cui si realizza F. Analogamente saranno approssimativamente $n\mathcal{P}(E \cap F)$ quelli in cui si realizzano sia E sia F. Perciò limitatamente agli esperimenti che hanno visto la realizzazione F, la frazione di quelli per i quali ha avuto luogo anche l'evento E è circa uguale a

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

Le approssimazioni fatte divengono esatte quando n tende all'infinito, e quindi la formula della probabilità condizionata è la corretta definizione di probabilità di E qualora si sia verificato F.

Teorema sulla probabilità condizionale

Sia B un evento, con $\mathcal{P}(B) > 0$. L'applicazione $Q: \mathcal{A} \to \mathbb{R}^+$ definita $Q(A) = \mathcal{P}(A|B)$ è una probabilità.

Dimostriamolo sfruttando la definizione di probabilità. L'applicazione Q è una probabilità (su $\mathcal{A}$) se verifica le seguenti due condizioni:

1) $\boldsymbol{Q(\Omega) = 1}$
   Sfruttando la probabilità condizionata, risulta
   $$Q(\Omega) = \mathcal{P}(\Omega|B) = \frac{\mathcal{P}(\Omega \cap B)}{\mathcal{P}(B)} = \frac{\mathcal{P}(B)}{\mathcal{P}(B)} = 1$$
2) **Se $\{\boldsymbol{A_n}\}_n$ è una successione di elementi di $\boldsymbol{\mathcal{A}}$ due a due disgiunti risulta $\boldsymbol{Q(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Q(A_i)}$**
   $$Q\left(\bigcup_{i=1}^{n} A_i\right) = \mathcal{P}(\cup_{i=1}^{n} A_i\,|B) = \mathcal{P}\big((A_1 \cup \ldots \cup A_n)\big|B\big) = \frac{\mathcal{P}\big((A_1 \cup \ldots \cup A_n) \cap B\big)}{\mathcal{P}(B)}$$
   A tale risultato basta applicare la probabilità distributiva e notare che gli eventi $A_i \cap B$ sono anch'essi disgiunti; dunque, è possibile applicare l'assioma 3 della probabilità:
   $$\frac{\mathcal{P}\big((A_1 \cap B) \cup \ldots \cup (A_n \cap B)\big)}{\mathcal{P}(B)} = \frac{\mathcal{P}(A_1 \cap B) + \cdots + \mathcal{P}(A_n \cap B)}{\mathcal{P}(B)} =$$
   $$= \sum_{i=1}^{n} \frac{\mathcal{P}(A_1 \cap B) + \cdots + \mathcal{P}(A_n \cap B)}{\mathcal{P}(B)} = \sum_{i=1}^{n} \mathcal{P}(A_i|B) = \sum_{i=1}^{n} Q(A_i)$$

Fattorizzazione di un evento e formula di Bayes

Siano E ed F due eventi qualsiasi. È possibile esprimere E come $E = (E \cap F) \cup (E \cap F^c)$; infatti ogni punto che appartiene all'evento E, o sta sia in E sia in F, oppure sta in E ma non in F. Inoltre, $E \cap F$ ed $E \cap F^c$ sono eventi disgiunti; dunque: $\boldsymbol{\mathcal{P}(E)} = \mathcal{P}(E \cap F) + \mathcal{P}(E \cap F^c) = \mathcal{P}(E|F)\mathcal{P}(F) + \mathcal{P}(E|F^c)\mathcal{P}(F^c) =$

$$= \boldsymbol{\mathcal{P}(E|F)\mathcal{P}(F) + \mathcal{P}(E|F^c)\big(1 - \mathcal{P}(F)\big)}$$

Questa equazione afferma che la probabilità dell'evento E si può ricavare come media pesata delle probabilità condizionali di E sapendo: (**1**) che F si è verificato e (**2**) che F non si è verificato.
I pesi corretti sono le probabilità degli eventi rispetto a cui si condiziona.

Questa formula è estremamente utile, in quanto in molte situazioni non è possibile calcolare una probabilità complessa direttamente, mentre essa è facilmente ricavabile dalla formula appena descritta, condizionando al verificarsi o meno di un secondo evento. L'evento accessorio va scelto in modo che, una volta che si sappia se esso si è verificato o meno, risulti evidente la probabilità dell'evento complesso di partenza, tenendo conto di questa informazione.

Siano assegnati una quantità finita (o numerabile) di eventi mutuamente esclusivi $F_1, F_2, \ldots, F_n$ tali che

$$\bigcup_{i=1}^{n} F_i = \Omega$$

Questa proprietà si cita dicendo che gli eventi F_i *ricoprono* Ω e significa che si verifica sempre almeno uno di essi (esattamente uno, se gli eventi sono anche disgiunti). Consideriamo un ulteriore evento E, che riscriviamo come

$$E = \bigcup_{i=1}^{n} (E \cap F_i)$$

notando che anche gli eventi $E \cap F_i$ sono mutuamente esclusivi. Si ottiene dall'assioma 3:

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

Questa formula, detta **formula di fattorizzazione** (o di disintegrazione), mostra che è possibile calcolare la probabilità di un evento E condizionando rispetto a quale si verifichi tra un gruppo di eventi accessori mutuamente esclusivi e che ricoprono Ω. Di nuovo $\mathcal{P}(E)$ può essere vista come la media pesata delle probabilità condizionate $\mathcal{P}(E|F_i)$, usando come pesi le corrispondenti $\mathcal{P}(F_i)$.

Si immagini ora di disporre dell'ulteriore informazione che si sia effettivamente verificato l'evento E. Che probabilità avranno gli eventi F_j tenendone conto?

$$\boldsymbol{\mathcal{P}(F_j|E) = \frac{\mathcal{P}(F_j \cap E)}{\mathcal{P}(E)} = \frac{\mathcal{P}(E|F_j)\mathcal{P}(F_j)}{\sum_{i=1}^{n} \mathcal{P}(E|F_i)\mathcal{P}(F_i)}}$$

Tale equazione prende il nome di **formula di Bayes**. Se pensiamo agli eventi F_j come a possibili "ipotesi" alternative che abbiano influenza su un qualche esperimento, si può immaginare che la formula di Bayes ci mostri come è necessario modificare le opinioni su tali ipotesi da prima a dopo l'esperimento stesso, con le loro probabilità che passano da $\mathcal{P}(F_j)$ a $\mathcal{P}(F_j|E)$.

Dimostrazione della formula di Bayes

Teorema: Sia $(\Omega, \mathcal{A}, \mathcal{P})$ uno spazio di probabilità. Sia $A_1, \dots, A_n$ una partizione di Ω, tale che $\mathcal{P}(A_i) > 0$ per ogni $i = 1,2, \dots, n$. Sia infine B un evento $\mathcal{P}(B) > 0$. Allora per ogni $j = 1,2, \dots, n$ risulta:

$$\mathcal{P}(A_j|B) = \frac{\mathcal{P}(B|A_j)\mathcal{P}(A_j)}{\sum_{i=1}^{n} \mathcal{P}(B|A_i)\mathcal{P}(A_i)}$$

Per la probabilità condizionale di A_j dato B abbiamo che $\mathcal{P}(A_j|B) = \mathcal{P}(A_j|B) = \frac{\mathcal{P}(A_j \cap B)}{\mathcal{P}(B)}$.
Sappiamo che $\mathcal{P}(A_j \cap B) = \mathcal{P}(B \cap A_j) = \mathcal{P}(B|A_j)\mathcal{P}(A_j)$.
Mentre per quanto riguarda il denominatore possiamo scrivere:

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{n} A_i\right) = \bigcup_{i=1}^{n} (B \cap A_i) \Rightarrow P(B) = \sum_{i=1}^{n} \mathcal{P}(B \cap A_i) = \sum_{i=1}^{n} \mathcal{P}(B|A_i)\mathcal{P}(A_i)$$

Da cui segue la formula di Bayes precedentemente descritta.

Nella formula di Bayes abbiamo le seguenti diciture:

1) $\mathcal{P}(A_j|B)$ viene detta **probabilità a posteriori**
2) $\mathcal{P}(B|A_j)$ è detta **probabilità di verosimiglianza**
3) $\mathcal{P}(A_j)$ prende il nome di **probabilità a priori**

Eventi indipendenti

La probabilità di E condizionata ad F, è generalmente diversa dalla probabilità non condizionata, $\mathcal{P}(E)$. Insomma, sapere che l'evento F si è verificato, modifica di solito la probabilità che si sia verificato E. Nel caso particolare in cui invece $\mathcal{P}(E|F)$ e $\mathcal{P}(E)$ siano uguali, diciamo che E è indipendente da F. Quindi E è indipendente da F se la conoscenza che F si è avverato non cambia la probabilità di E.
Siccome $\mathcal{P}(E|F) = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)}$, si vede che E è indipendente da F se $\boldsymbol{\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F)}$.

Poiché questa equazione è simmetrica in E e F, quando E è indipendente da F, è anche vero che F è indipendente da E. Si dà allora la seguente definizione.
Definizione 1: Due eventi E e F si dicono *indipendenti* se vale l'equazione $\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F)$, altrimenti si dicono *dipendenti*.

Diamo ora un utile risultato sull'indipendenza di eventi.
Proposizione: Se E e F sono indipendenti, lo sono anche E e F^c.
Dobbiamo dimostrare che $\mathcal{P}(E \cap F^c) = \mathcal{P}(E)\mathcal{P}(F^c)$. Siccome E è l'unione disgiunta di $E \cap F$ e $E \cap F^c$:

$$\mathcal{P}(E \cap F^c) = \mathcal{P}(E) - \mathcal{P}(E \cap F) = \mathcal{P}(E) - \mathcal{P}(E)\mathcal{P}(F) = \mathcal{P}(E)\big(1 - \mathcal{P}(F)\big) = \mathcal{P}(E)\mathcal{P}(F^c)$$

Quindi, se E e F sono indipendenti, la probabilità che E si realizzi non è modificata dall'informazione se F sia verificato oppure no.

Si noti che se E è indipendente sia da F sia da G, non possiamo concludere che E sia indipendente anche da $F \cap G$. Ad esempio, si lancino due dadi onesti. Siano $E = \{\text{la somma dei due punteggi è pari a } 7\}$, $F = \{\text{il primo dado totalizza un } 4\}$ e $G = \{\text{il secondo dado totalizza un } 3\}$. Si può dimostrare che E è indipendente da F come pure da G. Tuttavia, $\mathcal{P}(E|F \cap G) = 1$. Diamo allora la seguente definizione.

Definizione 2: I tre eventi E, F e G si dicono indipendenti se valgono tutte e quattro le equazioni seguenti:

- $\mathcal{P}(E \cap F \cap G) = \mathcal{P}(E)\mathcal{P}(F)\mathcal{P}(G)$
- $\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F)$
- $\mathcal{P}(E \cap G) = \mathcal{P}(E)\mathcal{P}(G)$
- $\mathcal{P}(G \cap G) = \mathcal{P}(F)\mathcal{P}(G)$

Si noti che se tre eventi E, F e G sono indipendenti, allora ciascuno di essi è indipendente da qualunque evento si possa costruire con gli altri due. Ad esempio E risulta indipendente da $F \cup G$, infatti:

$$\mathcal{P}\big(E \cap (F \cup G)\big) = \mathcal{P}\big((E \cap F) \cup (E \cap G)\big) = \mathcal{P}(E \cap F) + \mathcal{P}(E \cap G) - \mathcal{P}\big(E \cap (F \cap G)\big) =$$
$$= \mathcal{P}(E)\mathcal{P}(F) + \mathcal{P}(E)\mathcal{P}(G) - \mathcal{P}(E)\mathcal{P}(F \cap G) = \mathcal{P}(E)[\mathcal{P}(F) + \mathcal{P}(G) - \mathcal{P}(F \cap G)] = \mathcal{P}(E)\mathcal{P}(F \cup G)$$

Chiaramente la definizione precedente si può estendere senza sforzo ad un numero finito arbitrario di eventi. Gli eventi $E_1, E_2, \dots, E_n$ si dicono indipendenti se per ogni loro sottogruppo $E_{\alpha_1}, E_{\alpha_2}, \dots, E_{\alpha_r}$, con $1 \le a_1 < a_2 < \cdots < a_{r-1} < a_r \le n$, vale l'equazione:

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

Accade spesso che un esperimento casuale (in particolare quelli di interesse statistico) consista di una successione di prove, come il lancio ripetuto di una moneta. In molte di tali situazioni è ragionevole assumere che gli esiti di qualunque gruppo di queste prove non influenzino quelli delle altre. In questi casi gli eventi che dipendono dai singoli sottoesperimenti sono indipendenti, e l'intero ambito prende il nome di *schemi delle prove indipendenti*.

Schema delle prove indipendenti

Supponiamo di dover eseguire n volte (in condizioni di indipendenza) un certo esperimento che produce due soli possibili risultati, che chiameremo convenzionalmente *successo* e *insuccesso* e che indicheremo rispettivamente con i simboli 1 e 0. Supponiamo inoltre di sapere che il successo si presenta con probabilità p nella generica ripetizione dell'esperimento; analogamente l'insuccesso si presenta con probabilità $p - 1$ (con $0 < p < 1$).

Lo spazio campione è $\Omega = \{0,1\}^n$. Sia dunque $\omega = (\omega_1, \omega_2, \dots, \omega_n)$ un elemento di Ω; si può osservare che la quantità $\sum_{i=1}^{n} \omega_i$ e $n - \sum_{i=1}^{n} \omega_i$ non sono altro che il numero di simboli uguali a 1 e il numero di simboli uguali a 0 che formano la sequenza ω. Pertanto, avremo:

$$\mathcal{P}(\omega) = p^{\sum_{i=1}^{n} \omega_i}(1-p)^{\left(n - \sum_{i=1}^{n} \omega_i\right)}$$

Il gioco della zara

L'esempio è esemplificativo della descrizione formale di un *esperimento aleatorio* e dei *numeri non deterministici*.
Si consideri l'esperimento $\mathcal{E}_2$ che consiste nel lanciare due dadi onesti di colore differenti in maniera tale

che si può parlare di primo e secondo dado. Lo *spazio campione*, ovvero l'insieme dei possibili risultati di $\mathcal{E}_2$, è: $\Omega_2 \coloneqq \{(i,j): i,j = 1,2,\dots,6\}$.

Ne risulta che $|\Omega_2| = 6^2$, ovvero che in Ω_2 ci sono 36 coppie ordinate ognuna delle quali è un *punto campione*. D'altra parte, l'ipotesi dell'onestà dei due dadi porta a ritenere che per $i,j = 1,2,\dots,6$ si ha $\mathcal{P}_2(\{(i,j)\}) = \frac{1}{6^2} = \frac{1}{36}$; ovvero, che in questo specifico contesto è possibile utilizzare la definizione di probabilità introdotta da Laplace (rapporto tra gli esiti favorevoli e gli esiti possibili). Allora, considerando *evento* ogni sottoinsieme di Ω_2, si ha: $A \in \mathscr{P}(\Omega_2)$ (l'insieme della parti di Ω_2), $\mathcal{P}_2(A) = \frac{|A|}{36}$.

In definitiva, la terna $(\Omega_2, \mathscr{P}(\Omega_2), \mathcal{P}_2)$ con $\mathcal{P}_2$ definita in $\mathcal{P}_2(A)$ descrive completamente e in maniera formale l'esperimento $\mathcal{E}_2$.
Raramente, però, si è interessati al risultato (esito, punto campione) dell'esperimento ma, piuttosto, il valore rilevante è una sua funzione. In questo contesto, ad esempio, ha senso considerare la somma dei punteggi dei due dadi: $Z_2 : (i,j) \in \Omega_2 \to (i+j) \in \mathbb{R}$.

La funzione Z_2 assume con probabilità positiva i valori: $2,3,\dots,11,12$. Per ognuno di tali valori è possibile determinare agevolmente la sua *controimmagine* tramite la funzione Z_2:

$$\{Z_2 = 2\} = \{(1,1)\}, \qquad \{Z_2 = 12\} = \{(6,6)\},$$
$$\{Z_2 = 3\} = \{(1,2),(2,1)\}, \qquad \{Z_2 = 11\} = \{(5,6),(6,5)\},$$
$$\{Z_2 = 4\} = \{(1,3),(2,2),(3,1)\}, \qquad \{Z_2 = 10\} = \{(4,6),(5,5),(6,4)\},$$
$$\{Z_2 = 5\} = \{(1,4),(2,3),(3,2),(4,1)\}, \qquad \{Z_2 = 9\} = \{(3,6),(4,5),(5,4),(6,3)\},$$
$$\{Z_2 = 6\} = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}, \qquad \{Z_2 = 8\} = \{(2,6),(3,5),(4,4),(5,3),(6,2)\},$$
$$\{Z_2 = 7\} = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

Si osservi che la classe costituita dai sottoinsiemi così individuati costituisce una *partizione* di Ω_2.

Da $\mathcal{P}_2(A) = \frac{|A|}{36}$ e dalle partizioni precedentemente descritte se ne deduce che

$$\mathcal{P}_2(Z_2 = 2) = \mathcal{P}_2(Z_2 = 12) = \frac{1}{36}, \qquad \mathcal{P}_2(Z_2 = 3) = \mathcal{P}_2(Z_2 = 11) = \frac{2}{36},$$
$$\mathcal{P}_2(Z_2 = 4) = \mathcal{P}_2(Z_2 = 10) = \frac{3}{36}, \qquad \mathcal{P}_2(Z_2 = 5) = \mathcal{P}_2(Z_2 = 9) = \frac{4}{36},$$
$$\mathcal{P}_2(Z_2 = 6) = \mathcal{P}_2(Z_2 = 8) = \frac{5}{36}, \qquad \mathcal{P}_2(Z_2 = 7) = \frac{6}{36}.$$

Inoltre, risulta:

$$\sum_{s=2}^{12} \mathcal{P}_2(Z_2 = s) = \frac{2(1+2+3+4+5)+6}{36} = 1$$

Estensione al caso di tre dadi

Si consideri l'esperimento $\mathcal{E}_3$ che consiste nel lanciare tre dadi onesti di colore differenti in maniera tale che si può parlare di primo, secondo e terzo dado. Lo spazio campione, ovvero l'insieme dei possibili risultati di $\mathcal{E}_3$, è: $\Omega_3 \coloneqq \{(i,j,k): i,j,k = 1,2,\dots,6\}$.

Ne risulta che $|\Omega_3| = 6^3$, ovvero che in Ω_3 ci sono 216 coppie ordinate ognuna delle quali è un punto campione. D'altra parte, l'ipotesi dell'onestà dei tre dadi porta a ritenere che per $i,j,k = 1,2,\dots,6$ si ha $\mathcal{P}_3(\{(i,j,k)\}) = \frac{1}{6^3} = \frac{1}{216}$; ovvero, che in questo specifico contesto è possibile utilizzare la definizione di probabilità introdotta da Laplace (rapporto tra gli esiti favorevoli e gli esiti possibili). Allora, considerando evento ogni sottoinsieme di Ω_3, si ha: $A \in \mathscr{P}(\Omega_3)$, $\mathcal{P}_3(A) = \frac{|A|}{216}$.

In definitiva, la terna $(\Omega_3, \mathscr{P}(\Omega_3), \mathcal{P}_3)$ descrive completamente e in maniera formale l'esperimento $\mathcal{E}_3$.
In questo contesto ha senso considerare la somma dei punteggi dei tre dadi:

$$Z_3: (i,j.k) \in \Omega_3 \to (i+j+k) \in \mathbb{R}$$

La funzione Z_3 assume con probabilità positiva i valori: $3,4,\dots,17,18$. Per ognuno di tali valori è possibile determinare agevolmente la sua controimmagine tramite la funzione Z_3. A tale scopo, per sintetizzare le formule, è utile individuare le classi di terne compatibili con la comma in questione e che sono diverse per almeno un punteggio con le rispettive dimensioni. Ad esempio, per $\{Z_3 = 9\}$ ci sono le seguenti sei classi di terne: $[(1,2,6); \mathbf{6}], [(1,3,5); \mathbf{6}], [(1,4,4); \mathbf{3}], [(2,2,5); \mathbf{3}], [(2,3,4); \mathbf{6}], [(3,3,3); \mathbf{1}]$ (con $[(a,b,c); \boldsymbol{x}]$ indichiamo le $\boldsymbol{x}$ terne che si ottengono permutando i punteggi a, b, c in tutti i modi possibili).

Dopo di ciò, si ha: $\mathcal{P}_3(Z_3 = 9) = \frac{\mathbf{6+6+3+3+6+1}}{6^3} = \frac{25}{216}$.

Allo stesso modo, in $\{Z_3 = 10\}$ ci sono le seguenti sei classi di terne:
$[(1,3,6); \mathbf{6}], [(1,4,5); \mathbf{6}], [(2,2,6); \mathbf{3}], [(2,3,5); \mathbf{6}], [(2,4,4); \mathbf{3}], [(3,3,4); \mathbf{4}] \Rightarrow \mathcal{P}_3(Z_3 = 10) = \frac{27}{216}$.

Le precedenti formule forniscono la soluzione al problema posto dai gentiluomini fiorentini a Galileo Galilei: in totale, ci sono 25 terne di punteggi che fanno realizzare il 9 e 27 terne di punteggi che fanno realizzare il 10. Si verifichi per esercizio che $\sum_{s=3}^{18} \mathcal{P}_2(Z_2 = s) = 1$.

Osservazione: il numero aleatorio Z_2 può essere considerato anche nello spazio di probabilità $(\Omega_3, \mathscr{P}(\Omega_3), \mathcal{P}_3)$: $\quad Z_2 : (i,j,k) \in \Omega_3 \to (i+j+0k) \in \mathbb{R}$.
Ovviamente, questo comporta un aggravo nelle formule in quanto, ad esempio, $\{Z_2 = 2\} = \{(1,1,1), (1,1,2), (1,1,3), (1,1,4), (1,1,5), (1,1,6)\} \Rightarrow |Z_2 = 2| = 6$. Questo fatto non altera i risultati individuati a inizio paragrafo; infatti, relativamente all'evento $\{Z_2 = 2\}$, tenendo presente la prima delle formule si ha: $\mathcal{P}_3(Z_2 = 2) = \frac{6}{216} = \frac{1}{36} = \mathcal{P}_2(Z_2 = 2)$.

Il problema delle Concordanze

Sia n un numero intero e si abbia a disposizione un mazzetto di n cartoncini. Essi vengono numerati da 1 a n sul fronte inizialmente bianco; sul dorso i cartoncini presentano la stessa immagine. L'esperimento $\mathcal{E}$ consiste nel (i) mischiare i cartoncini, (ii) appoggiare su un tavolo il mazzetto mostrandone il dorso, (iii) mostrare il fronte del cartoncino più in alto "chiamando" il numero "uno" e riporre accanto al mazzetto il cartoncino mostrandone il fronte, (iv) procedere allo stesso modo fino all'ultimo cartoncino.
Per $i = 1,2,\dots,n$, si designa con C_i (concordanza) l'evento che si verifica quando si presenta il cartoncino con numero i alla i-esima chiamata.

Probabilità di osservare almeno una concordanza nelle $\boldsymbol{n}$ chiamate

L'evento "almeno una concordanza nelle n chiamate" risulta essere l'unione di tutte le concordanze: $\bigcup_{i=1}^{n} C_i$. Ovviamente, nulla vieta che nella stessa effettuazione dell'esperimento si possono verificare due o più concordanze; pertanto, è necessario applicare la formula di inclusione-esclusione:

$$\mathcal{P}\left(\bigcup_{i=1}^{n} C_i\right) = \sum_{i=1}^{n} \mathcal{P}(C_i) - \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathcal{P}(C_i \cap C_j) + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n} \mathcal{P}(C_i \cap C_j \cap C_k) + \cdots + (-1)^{n-1}\mathcal{P}\left(\bigcap_{i=1}^{n} C_i\right)$$

Si osservi che ci sono $n!$ possibili mischiate; risulta: $i = 1,2,\dots,n,\ \mathcal{P}(C_i) = \frac{(n-1)!}{n!}$

Infatti, bloccando solo il cartoncino numerato con i all'i-esimo posto dall'alto si verifica C_i, ma tutti gli altri possono permutare in tutti i modi possibili. Con lo stesso ragionamento, si ha:

$$i,j = 1,2,\dots,n : i < j, \mathcal{P}(C_i \cap C_j) = \frac{(n-2)!}{n!} \quad \text{e} \quad i,j,k = 1,2,\dots,n : i < j < k, \mathcal{P}(C_i \cap C_j \cap C_k) = \frac{(n-3)!}{n!}$$

Allo stesso modo si procede per i successivi addendi. In particolare, dal momento che $(n-n)! = 1$, si ha:

$$\mathcal{P}\left(\bigcap_{i=1}^{n} C_i\right) = \frac{1}{n!}$$

La formula di inclusione-esclusione diventa:

$$\mathcal{P}\left(\bigcup_{i=1}^{n} C_i\right) = \frac{(n-1)!}{n!}\sum_{i=1}^{n} 1 - \frac{(n-2)!}{n!}\sum_{i=1}^{n}\sum_{j=i+1}^{n} 1 + \frac{(n-3)!}{n!}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n} 1 + \cdots + (-1)^{n-1}\frac{1}{n!} =$$
$$= \frac{(n-1)!}{n!}\binom{n}{1} - \frac{(n-2)!}{n!}\binom{n}{2} + \frac{(n-3)!}{n!}\binom{n}{3} + \cdots + (-1)^{n-1}\frac{1}{n!} =$$
$$= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} + \cdots + (-1)^{n-1}\frac{1}{n!}$$

Il passaggio intermedio nell'espressione precedente segue da questo ragionamento (valido per il secondo addendo ma facilmente generalizzabile agli altri): il numero degli addendi 1 è uguale al numero delle combinazioni semplici di lunghezza 2 ottenute dall'insieme $S_n = \{1,2,\dots,n\}$ in quanto il valore dell'indice j è strettamente maggiore del valore dell'indice i. In conclusione, si ha:

$$\mathcal{P}\left(\bigcup_{i=1}^{n} C_i\right) = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i!}$$

Probabilità di osservare zero concordanze nelle n chiamate

Si indichi con $E_{0,n}$ l'evento che si presenta quando in una esecuzione dell'esperimento $\mathcal{E}$ non si osserva alcuna concordanza (ovvero, 0 concordanze); ovviamente, risulta che:

$$E_{0,n} = \left(\bigcup_{i=1}^{n} C_i\right)^{c} \Rightarrow \mathcal{P}(E_{0,n}) = 1 - \mathcal{P}\left(\bigcup_{i=1}^{n} C_i\right) = \sum_{i=1}^{n} \frac{(-1)^{i}}{i!}$$

Prima di terminare questo sottoparagrafo si osservi che, indicato con $N_{0,n}$ il numero delle mischiate che non fanno verificare alcuna concordanza, dalla formula precedente si ricava:

$$\mathcal{P}(E_{0,n}) = \frac{N_{0,n}}{n!} \Rightarrow N_{0,n} = n! \cdot \mathcal{P}(E_{0,n})$$

Probabilità di osservare SOLO una concordanza nelle n chiamate

Si indichi con $F_{1,1,n}$ l'evento che si presenta quando in una esecuzione dell'esperimento $\mathcal{E}$ si osserva esattamente una concordanza e questa avviene alla prima chiamata. Ovviamente, indicato con $M_{1,1,n}$ il numero delle mischiate che fanno verificare solo la concordanza nella prima chiamata e zero concordanze nelle successive, risulta che $M_{1,1,n} = N_{0,n-1}$: infatti, dopo la concordanza alla prima chiamata, dalla seconda alla n-esima chiamata non si verificano altre concordanze. Allora, in virtù dell'espressione del sottoparagrafo precedente si ha:

$$\mathcal{P}(F_{1,1,n}) = \frac{M_{1,1,n}}{n!} = \frac{N_{0,n-1}}{n!} = \frac{(n!-1)\cdot\mathcal{P}(E_{0,n-1})}{n!} = \frac{1}{n!}\mathcal{P}(E_{0,n-1})$$

Si osservi che il risultato così ottenuto non dipende dal fatto che è stato richiesto di osservare la concordanza alla prima chiamata in quanto lo stesso ragionamento è valido per la richiesta di osservare un'unica concordanza nella generica chiamata. In altri termini, indicato con $F_{1,i,n}$ l'evento che si presenta quando in una esecuzione dell'esperimento $\mathcal{E}$ si osserva esattamente una concordanza e questa avviene alla i-esima chiamata, allora si ha:

$$i = 1,2,\dots,n, \qquad \mathcal{P}(F_{1,i,n}) = \mathcal{P}(F_{1,1,n}) = \frac{1}{n!}\mathcal{P}(E_{0,n-1})$$

Dopo di ciò, se $E_{1,n}$ è l'evento che si presenta quando in una esecuzione dell'esperimento $\mathcal{E}$ si osserva una sola concordanza, risulta

$$\mathcal{P}(E_{1,n}) = \sum_{i=1}^{n} \mathcal{P}(F_{1,i,n}) = \frac{1}{n}\mathcal{P}(E_{0,n-1})\sum_{i=1}^{n} 1 = \mathcal{P}(E_{0,n-1}) = \sum_{i=1}^{n-1} \frac{(-1)^{i}}{i!}$$

Probabilità di osservare esattamente due concordanze nelle n chiamate

Si procede sulla falsariga del paragrafo precedente. Si lascia come esercizio.

2. Variabili aleatorie e valore atteso

Variabili aleatorie

Quando si realizza un esperimento causale, non sempre si è interessati in ugual modo a tutte le informazioni ricavabili dal suo esito. Spesso si può individuare una singola quantità numerica (ricavabile dall'esito stesso) che racchiude tutto ciò che in realtà vogliamo sapere. Se tiriamo due dadi, ad esempio, può accadere che ci interessi solamente il valore della loro somma e non ciascuno dei punteggi. Potremmo volere registrare che il totale realizzato è 7, senza dare importanza a quale sia l'esito vero e proprio dell'esperimento. Un ingegnere civile che segue il livello di un bacino idrico, allo stesso modo, potrebbe decidere di prendere delle misurazioni solo alla fine di ogni stagione delle piogge, perché magari le oscillazioni giornaliere non aggiungono informazioni rilevanti.

Quantità di interesse che, come queste, sono determinate dal risultato di un esperimento casuale sono dette **variabili aleatorie**. Siccome il valore di una variabile aleatorie è determinato dall'esito dell'esperimento, possiamo assegnare delle probabilità ai suoi valori possibili.

Variabili aleatorie con un numero finito o numerabile di valori possibili sono dette **discrete**. Esistono comunque anche variabili aleatorie dette appunto **continue**, che possono assumere un insieme continuo di valori possibili, come può essere un intervallo di numeri reali. Un esempio è il tempo di vita di una automobile, che può assumere qualunque valore di un qualche intervallo (a, b).

Definizione: La *funzione di ripartizione* (o di distribuzione) F di una varibile aleatoria X, è definita, per ogni numero reale x, tramite $F(x) \coloneqq \mathcal{P}(X \leq x)$.

N.B.: la notazione corretta è $\mathcal{P}(\{X \leq x\})$ ma con leggero abuso di notazione possiamo omettere le graffe. Quindi $F(x)$ esprime la probabilità che la variabile aleatoria X assuma un valore minore o uguale a x. Useremo la notazione $X \sim F$ per indicare che F è la funzione di ripartizione di X.

Tutte le questioni di probabilità che si possano sollevare su una variabile aleatoria, ammettono una risposta in termini della sua funzione di ripartizione. Ad esempio, volendo calcolare $\mathcal{P}(a < X \leq b)$, basta notare che $\{X \leq b\}$ è l'unione dei due eventi disgiunti $\{X \leq a\}$ e $\{a < X \leq b\}$. Quindi, $\mathcal{P}(X \leq b) = \mathcal{P}(X \leq a) + \mathcal{P}(a < X \leq b)$ da cui $\mathcal{P}(a < X \leq b) = \mathcal{P}(X \leq b) - \mathcal{P}(X \leq a) = F(b) - F(a)$.

Proprietà della funzione di distribuzione

Siano X una variabile aleatoria, F la sua funzione di ripartizione. Valgono le seguenti proprietà:

i. $\boldsymbol{0 \leq F(x) \leq 1, \forall x \in \mathbb{R}}$
   $F(x)$ rappresenta la probabilità di un evento (l'evento $\{X \leq x\}$) dunque ne rispetta gli assiomi.

ii. F è non decrescente: $\forall x, y \in \mathbb{R} : x \leq y \Rightarrow F(x) \leq F(y)$
   Essendo $x \leq y$, si deduce che la probabilità dell'evento rappresentata da $F(x)$ è contenuta in quella rappresentata da $F(y)$; ovvero: $\{X \leq x\} \subseteq \{X \leq y\}$ e quindi, per la proprietà di *isotonia* ($A \subseteq B \Rightarrow \mathcal{P}(B) = \mathcal{P}(A) + \mathcal{P}(A^c \cap B) \Rightarrow \mathcal{P}(A) \leq \mathcal{P}(B)$), si ha $F(x) \leq F(y)$.

iii. $\boldsymbol{\lim_{n \to +\infty} F(n) = 1 \wedge \lim_{n \to -\infty} F(n) = 0}$
   Cominciamo con la prima osservando che $F(n)$ è la probabilità dell'evento $A_n = \{X \leq n\}$. Sappiamo che la successione di eventi $\{A_n\}_{n \in \mathbb{N}}$ è *crescente* e $\bigcup A_n = \Omega$. Dunque, per la proprietà di passaggio al limite delle successioni (vedi Assiomi di Kolmogorov) si ha $\lim_{n \to +\infty} F(n) = \lim_{n \to +\infty} \mathcal{P}(A_n) = \mathcal{P}(\bigcup A_n) = \mathcal{P}(\Omega) = 1$. Ragionamento analogo per la seconda: $F(n)$ è la probabilità dell'evento $A_n = \{X \leq -n\}$ ma essendo la successione *decrescente* si ha $\bigcap A_n = \emptyset$. Dunque, sempre per la suddetta proprietà, si ha $\lim_{n \to -\infty} F(n) = \lim_{n \to -\infty} \mathcal{P}(A_n) = \mathcal{P}(\bigcap A_n) = \mathcal{P}(\emptyset) = 0$.

iv. F è continua a destra; ovvero: $\forall \boldsymbol{x} \in \mathbb{R}, \boldsymbol{F(x^+) = \lim_{n\to+\infty} F\left(x+\frac{1}{n}\right) = \lim_{t\to x^+} F(t) = F(x)}$

Attraverso la definizione di funzione di ripartizione e scomponendo l'evento in un'unione di due eventi disgiunti si ottiene: $F(x^+) = \lim_{n\to+\infty} F\left(x+\frac{1}{n}\right) = \lim_{n\to+\infty} \mathcal{P}\left(X \le x+\frac{1}{n}\right) =$

$= \lim_{n\to+\infty} \mathcal{P}\left((X \le x) \cup \left(x < X \le x+\frac{1}{n}\right)\right) = \lim_{n\to+\infty} \left(\mathcal{P}(X \le x) + \mathcal{P}\left(x < X \le x+\frac{1}{n}\right)\right) =$

$= \lim_{n\to+\infty} \mathcal{P}(X \le x) + \lim_{n\to+\infty} \mathcal{P}\left(x < X \le x+\frac{1}{n}\right) = \lim_{n\to+\infty} F(x) + 0 = F(x).$

N.B: $\lim_{n\to+\infty} \mathcal{P}\left(x < X \le x+\frac{1}{n}\right) = 0$ poiché la probabilità dell'evento $\left\{X \in \left(x, x+\frac{1}{n}\right]\right\}$ tende a 0 al crescere di n.

Teorema 1: Sia $F: \mathbb{R} \to \mathbb{R}$ una funzione, che gode delle proprietà (i), (ii), (iii) e (iv) contemporaneamente. Allora esiste una variabile aleatoria X definita su un opportuno spazio di probabilità, che ammette F come funzione di ripartizione.

Tale teorema è utile per dimostrare che una determinata funzione è di ripartizione (non lo sarà se almeno una delle proprietà sopra descritte non è verificata).

Teorema 2: La funzione di ripartizione individua univocamente la legge della variabile aleatoria. In altre parole, se due variabili aleatorie hanno la stessa funzione di ripartizione, allora hanno la stessa legge (le due variabili si dicono **somiglianti** o anche **equamente distribuite**).

Calcolo della quantità $\mathcal{P}(X \in I)$ per casi particolari di insiemi misurabili I:

- $I = (a,b] \Rightarrow \mathcal{P}(a < X \le b) = \mathcal{P}(X \le b) - \mathcal{P}(X \le a) = F(b) - F(a)$
- $I = (-\infty, b] \Rightarrow \mathcal{P}(X < b) = \lim_{n\to+\infty} \mathcal{P}\left(X \le b - \frac{1}{n}\right) = \lim_{t\to b^-} F(t) = F(b^-)$
- $I = \{x\} \Rightarrow \mathcal{P}(X = x) = \mathcal{P}(X \le x) - \mathcal{P}(X < x) = F(x) - F(x^-)$
- $I = [a,b) \Rightarrow \mathcal{P}(a \le X < b) = \mathcal{P}(X < b) - \mathcal{P}(X < a) = F(b^-) - F(a^-)$
- $I = (a,b) \Rightarrow \mathcal{P}(a < x < b) = \mathcal{P}(X < b) - \mathcal{P}(X \le a) = F(b^-) - F(a)$
- $I = [a,b] \Rightarrow \mathcal{P}(a \le X \le b) = \mathcal{P}(X \le b) - \mathcal{P}(X < a) = F(b) - F(a^-)$
- $I = (a,+\infty) \Rightarrow \mathcal{P}(X > a) = 1 - \mathcal{P}(X \le a) = 1 - F(a)$
- $I = [a,+\infty) \Rightarrow \mathcal{P}(X \ge a) = 1 - \mathcal{P}(X < a) = 1 - F(a^-)$

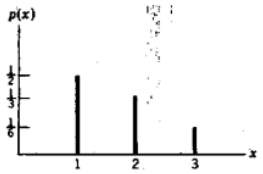
Variabili aleatorie discrete e continue

Definzione 1: Se X è una variabile aleatoria discreta, la sua *funzione di massa (di probabilità)* si definisce nel modo seguente, $\boldsymbol{p(a) := \mathcal{P}(X = a)}$

La funzione $p(a)$ è non nulla su un insieme al più numerabile di valori. Infatti, se $x_1, x_2, \dots$ sono i valori possibili di X, allora $p(x_i) > 0$, per $i = 1,2, \dots$ mentre $p(x) = 0$ per tutti gli altri valori di x.

Siccome X deve assumere uno dei valori $x_1, x_2, \dots$, necessariamente la funzione di massa deve soddisfare la seguente equazione: $\sum_{i=1}^{\infty} \boldsymbol{p(x_i) = 1}$.

Esempio 1: Consideriamo una variabile aleatoria X che può assumere i valori 1,2 o 3. Se sappiamo che $p(1) = \frac{1}{2}$ e $p(2) = \frac{1}{3}$ allora, dato che $p(1) + p(2) + p(3) = 1$, ne segue che $p(3) = \frac{1}{6}$ (a destra il grafico di questa funzione di massa).



Per una variabile aleatoria discreta, la funzione di ripartizione F può essere espressa in funzione della funzione di massa di probabilità p, tramite $\boldsymbol{F(a) = \sum_{x \le a} p(x)}$ dove si intende che la serie è limitata ai soli valori possibili di X minori o uguali ad a. Si noti che la F che ne risulta è una funzione a gradini, e più precisamente, se $x_1 < x_2 < \cdots$ sono i valori possibili di X, allora F è costante su ciascuno degli intervalli $[x_{i-1}, x_i)$ e in x_i fa un salto di ampiezza $p(x_i)$, passando da $p(x_1) + p(x_2) + \cdots + p(x_{i-1})$ a $p(x_1) + p(x_2) + \cdots + p(x_{i-1}) + p(x_i)$.

Supponendo che X abbia la stessa funzione di massa dell'Esempio 1, con $p(1) = \frac{1}{2}, p(2) = \frac{1}{3}, p(3) = \frac{1}{6}$ la funzione di ripartizione F di X è data da

$$F(a) = \begin{cases} 0 & se\ a < 1 \\ \frac{1}{2} & se\ 1 \leq a < 2 \\ \frac{5}{6} & se\ 2 \leq a < 3 \\ 1 & se\ 3 \leq a \end{cases}$$

Una variabile aleatoria che possa assumere una infinità non numerabile di valori non potrà essere discreta. Si dirà invece **continua** se esiste una funzione non negativa f, definita su tutto $\mathbb{R}$, avente la proprietà che per ogni insieme B di numeri reali, $\mathcal{P}(X \in B) = \int_B f(x)dx$.

Definizione 2: la funzione f che compare nell'equazione $\boldsymbol{\mathcal{P}(X \in B) = \int_B f(x)dx}$ è la *funzione di densità (di probabilità)* della variabile aleatoria X.

Suddetta equazione dice che la probabilità che una variabile aleatoria continua X appartenga a un insieme B si può trovare integrando la sua densità su tale insieme. Poiché X deve assumere un qualche valore di $\mathbb{R}$, la sua densità deve soddisfare la seguente equazione:

$$\boldsymbol{1 = \mathcal{P}(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)dx}$$

Tutte le probabilità che riguardano una variabile aleatoria continua possono essere espresse in termini di integrali della sua densità. Ad esempio, se poniamo $B = [a, b]$, ricaviamo da $\mathcal{P}(X \in B) = \int_B f(x)dx$ che $\mathcal{P}(a \leq X \leq b) = \int_a^b f(x)dx$ e se in quest'ultima equazione poniamo $b = a$, troviamo che $\mathcal{P}(X = a) = \int_a^a f(x)dx = 0$; ovvero, la probabilità che una variabile aleatoria continua assuma un qualunque valore particolare a è nulla (si veda figura).

Una relazione che lega la funzione di ripartizione F alla densità f è la seguente:

$$F(a) \coloneqq P(X \in (-\infty, a]) = \int_{-\infty}^{a} f(x)dx$$

Derivando entrambi i membri si ottiene allora la relazione fondamentale: $\boldsymbol{\frac{d}{da}F(a) = f(a)}$

La densità è la derivata della funzione di ripartizione. Una interpretazione forse meno astrtta della funzione di densità si può ricavare dall'equaizone $\mathcal{P}(a \leq X \leq b) = \int_a^b f(x)dx$ nel modo che segue: se $\varepsilon > 0$ è piccolo si può approssimare l'integrale con il teorema del valore medio:

$$\mathcal{P}\left(a - \frac{\varepsilon}{2} \leq X \leq a + \frac{\varepsilon}{2}\right) = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)dx \approx \varepsilon f(a)$$

Si scopre così che la probabilità che X stia in un intorno di a di ampiezza ε è approssimativamente uguale a $\varepsilon f(a)$, e quindi $f(a)$ rappresenta una indicazione di quanto è probabile che X cada "vicino" ad a (si rammenti che $\{X = a\}$ ha probabilità nulla).

Osservazione: quando conosciamo la funzione di massa di una varaibile aleatoria discreta, oppure la funzione di densità di probabilità di una continua, oppure ancora quando conosciamo la funzione di ripartizione di una variabile aleatoria qualsiasi, abbiamo abbastanza informazioni da poter calcolare la probabilità di ogni evento che dipenda solo da tale variabile aleatoria. Si dice in questo caso che

conosciamo la **distribuzione** o **legge** della variabile aleatoria considerata. Perciò, affermare ad esempio che X e Y hanno la stessa distribuzione, vuole dire che le rispettive funzioni sono identiche, $X \sim F_X \equiv F_Y \sim Y$, e quindi anche $\mathcal{P}(X \in A) = \mathcal{P}(Y \in A)$ per ogni insieme di valori $A \subset \mathbb{R}$.

Coppie e vettori di variabili aleatorie

Ci sono situazioni in cui la scelta di ridurre un esperimento casuale allo studio di una sola variabile aleatoria, è destinata a fallire a priori, perché l'oggetto di interesse sono proprio le relazioni presenti tra due o più grandezze numeriche. Ad esempio, in un esperimento sulle possibili cause di tumore, potremmo voler indagare il rapporto tra il numero medio di sigarette fumate quotidianamente e l'età in cui vinee riscontrata questa patologia. Analogamente, un ingegnere meccanico che si occupi di montaggio di laminati in acciaio, potrebbe volere conoscere la relazione tra il diametro dei punti di saldatura e la loro sollecitazione di taglio.
Per specificare la relazione tra due variabili aleatorie X e Y, il punto di partenza è estendere il concetto di funzione di ripartizione.

Definizione 1: Siano X e Y due variabili aleatorie che riguardano lo stesso esperimento casuale. Si dice *funzione di ripartizione congiunta* di X e Y, e si indica normalmente con la lettera F, la funzione di due variabili seguente: $\boldsymbol{F(x,y) \coloneqq \mathcal{P}(X \leq x, Y \leq y)}$ dove la virgola nell'argomento di $\mathcal{P}$ denota l'intersezione tra eventi.

La conoscenza di questa funzione permette, almeno in teoria, di calcolare le probabilità di tutti gli eventi che dipendono, singolarmente o congiuntamente, da X e Y.
Ad esempio la funzione di ripartizione di X, che denotiamo questa volta con F_X, può essere ottenuta dalla funzione di ripartizione congiunta F così: $F_X(x) \coloneqq \mathcal{P}(X \leq x) = \mathcal{P}(X \leq x, Y < \infty) = F(x, \infty)$ perché $Y < \infty$ sempre nel senso del limite $\lim_{y \to \infty} F(x,y)$. E analogamente $F_Y(y) = F(\infty, y)$.

Distribuzione congiunta per variabili aleatorie discrete

Come nel caso scalare, se sappiamo che un vettore aleatorio è di tipo discreto, possiamo definire e utilizzare la funzione di massa di probabilità.

Definizione 2: Se X e Y sono variabili aleatorie discrete che assumono i valori $x_1, x_2, \ldots$ e $y_1, y_2, \ldots$ rispettivamente, la funzione $\boldsymbol{p(x_i, y_j) \coloneqq \mathcal{P}(X = x_i, Y = y_j)}$, $i, j = 1,2, \ldots$ è la loro *funzione di massa di probabilità congiunta*.

Le funzioni di massa individuali di X e Y si possono ricavare da quella congiunta notando che, siccome Y deve assumere uno dei valori y_i, l'evento $\{X = x_i\}$ può essere visto come l'unione al variare di j degli eventi $\{X = x_i, Y = y_j\}$, che sono mutuamente esclusivi; in formule, $\{X = x_i\} = \cup_j \{X = x_i, Y = y_j\}$ da cui:

$$p_X(x_i) \coloneqq \mathcal{P}(X = x_i) = \mathcal{P}\left(\bigcup_j \{X = x_i, Y = y_j\}\right) = \sum_j \mathcal{P}(X = x_i, Y = y_j) = \sum_j p(x_i, y_j)$$

Analogamente $p_Y(y_j) = \sum_i p(x_i, y_j)$.

Anche se abbiamo mostrato che le funzioni di massa individuali (o *marginali*) si possono sempre ricavare da quella congiunta, il viceversa è falso. Quindi, conoscere $\mathcal{P}(X = x_i)$ e $\mathcal{P}(Y = y_j)$non permette di ricavare $\mathcal{P}(X = x_i, Y = y_j)$.

Distribuzione congiunta per variabili aleatorie continue

Due variabili aleatorie X e Y sono *congiuntamente continue* se esiste una funzione non negativa $f(x,y)$, definita per tutti gli x e y, avente la proprietà che per ogni sottoinsieme C del piano cartesiano

$$\boldsymbol{\mathcal{P}((X,Y) \in C) = \iint_{(x,y) \in C} f(x,y)\, dx\, dy}$$

Definizione 3: La funzione di due variabili f, che compare nell'equazione precedente è la *densità congiunta* delle variabili aleatorie X e Y.

Se A e B sono sottoinsiemi qualsiasi di $\mathbb{R}$, e se si denota con $C \coloneqq A \times B$ il loro prodotto cartesiano su $\mathbb{R}^2$, ovvero $C \coloneqq \{(x,y) \in \mathbb{R} : x \in A, y \in B\}$ si vede dall'equazione della definizione 3 che la densità congiunta f soddisfa $\mathcal{P}(X \in A, Y \in B) = \int_B \int_A f(x,y)dxdy$ e quindi, ponendo $A = (-\infty, a\}, B = (-\infty, b]$, si può riscrivere la funzione di ripartizione congiunta di X e Y come segue:

$$F(a,b) \coloneqq \mathcal{P}(X \leq a, Y \leq b) = \mathcal{P}(X \in A, Y \in B) = \int_B \int_A f(x,y)dxdy = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)dxdy$$

da cui derivando, nelle due direzioni $f(a,b) = \frac{d^2}{dadb} F(a,b)$ in tutti i punti in cui le derivate parziali sono definite. Anche qui, come nel caso scalare (vedi Variabili aleatorie discrete e continue), è possibile ottenere una formula approssimata che motiva la scelta del nome di *densità di probabilità*:

$$\mathcal{P}(a \leq X \leq a + da, b \leq Y \leq b + db) = \int_b^{b+db} \int_a^{a+da} f(x,y)dxdy \approx f(a,b)\, da\, db$$

L'approssimazione finale è valida (per il teorema del valore medio) se gli incrementi da e db sono piccoli e f è continua nel punto (a,b). Se ne deduce che $f)a,b)$ è circa pari al rapporto tra la probabilità di un rettangolo attorno al punto (a,b), e l'area $da\ db$ del rettangolino stesso, è insomma una densità di probabilità nel senso comune che questo termina assume, e una indicazione di quanto è probabile che (X,Y) cada vicino ad (a,b).

Se X e Y sono congiuntamente continue, allora prese individualmente, sono variabili aleatorie continue nel senso usuale; inoltre le loro *densità marginali* si ricavano come segue. Per ogni insieme A di numeri reali, $\int_A f(x,y)dx = \mathcal{P}(X \in A) = \mathcal{P}(X \in A, Y \in \mathbb{R}) = \int_A \int_{-\infty}^{\infty} f(x,y)dydx$. Da questa equazione, visto che A è un insieme arbitrario deve valere per forza l'uguaglianza degli integrandi: $\boldsymbol{f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy}$.
Analogamente, si può ricavare la funzione di densità marginale di Y che è $\boldsymbol{f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx}$.

Variabili aleatorie indipendenti

Due variabili aleatorie sono indipendenti se tutti gli eventi relativi alla primo sono indipendenti da tutti quelli relativi alla seconda.

Definizione 4: Due variabili aleatorie che riguardano lo stesso esperimento casuale si dicono *indipendenti* se, per ogni coppia di insiemi di numeri reali A e B, è soddisfatta l'equazione $\boldsymbol{\mathcal{P}(X \in A, Y \in B) = \mathcal{P}(X \in A)\mathcal{P}(Y \in B)}$; ovvero, se per ogni scelta di A e B, gli eventi $\{X \in A\}$ e $\{Y \in B\}$ risulta indipendenti. In caso contrario X e Y si dicono *dipendenti*.

Usando gli assiomi della probabilità è possibile dimostrare che questa definizione è equivalente alla richiesta che per ogni coppia di reali a e b, $\boldsymbol{\mathcal{P}(X \leq a, Y \leq b) = \mathcal{P}(X \leq a)\mathcal{P}(Y \leq b)}$; ovvero che la funzione di ripartizione congiunta sia il prodotto delle marginali: $\boldsymbol{F(a,b) = F_X(a)F_Y(b)}, \forall a, b \in \mathbb{R}$ dove si intendeche $F_X \sim X, F_Y \sim Y$ e F è la funzione di ripartizione congiunta di X e Y.

Se le variabili considerate sono discrete, l'indipendenza è anche equivalente a chiedere che la funzione di massa congiunta sia il prodotto delle marginali: $\boldsymbol{p(x,y) = p_X(x)p_Y(y)}, \forall x, y \in \mathbb{R}$.

Tale equivalenza si prova facilmente. Per una direzione basta notare che l'equazione nella Definizione 4 implica $p(x,y) = p_X(x)p_Y(y)$ non appena si pone $A = \{x\}$ e $B = \{y\}$. Per l'alta direzione è necessario dimostrare che l'equazione $\mathcal{P}(X \in A, Y \in B) = \mathcal{P}(X \in A)\mathcal{P}(Y \in B)$ è soddisfatta per ogni scelta di insiemi reali A e B. A tal scopo, supposta vera l'equazione $p(x,y) = p_X(x)p_Y(y)$, si ha:

$$P(X \in A, Y \in B) = \sum_{x \in A} \sum_{y \in B} p(x,y) = \sum_{x \in A} \sum_{y \in B} p_X(x)p_Y(y) = \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y) = \mathcal{P}(X \in A)\mathcal{P}(Y \in B)$$

Nel caso di variabili aleatorie congiuntamente continue invece, X e Y sono indipendenti se e solo se la densità congiunta è il prodotto delle marginali: $f(x,y) = f_X(x)f_Y(y), \forall x,y \in \mathbb{R}$. Questa ulteriore equivalenza può essere provata con passaggi simili a quelli qui sopra, sfruttando le equazioni seguenti $P(X \in A, Y \in B) = \int_B \int_A f(x,y)\,dxdy$ e $f(x,y) = \frac{d^2}{dxdy} P(X \in A, Y \in B)$.

Il senso della definizione e delle molte forme equivalenti che abbiamo dato è che due variabili aleatorie sono indipendenti se conoscere il valore di una non cambia la distribuzione dell'altra.

Generalizzazione a più di due variabili aleatorie

Tutti gli argomenti visti in questa sezione si possono estendere in maniera più o meno naturale ad un numero arbitrario n di variabili aleatorie. La funzione di ripartizione congiunta di $X_1, X_2, \dots, X_n$ è la funzione di n variabili F, definita da $F(a_1, a_2, \dots, a_n) \coloneqq \mathcal{P}(X_1 \leq a, X_2 \leq a_2, \dots, X_n \leq a_n)$. Se queste variabili aleatorie sono discrete, è possibile definire la funzione di massa di probabilità congiunta p, che è data da $p(x_1, x_2, \dots, x_n) \coloneqq \mathcal{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$.

Altrimenti, le variabili aleatorie $X_1, X_2, \dots, X_n$ sono congiuntamente continue, se esite una densità di probabilità congiunta f; funzioen di n variabili a valori positivi tale che, per ogni sottoinsieme C di $\mathbb{R}^n$, $\mathcal{P}\big((X_1, X_2, \dots, X_n) \in C\big) = \iint \dots \int_{x_1,x_2,\dots,x_n \in C} f(x_1, x_2, \dots, x_n)dx_1dx_2 \dots dx_n$.
Ciò significa che se $A_1, A_2, \dots, A_n$ sono insiemi di numeri reali, allora

$$\mathcal{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \int_{A_1}\int_{A_2}\dots\int_{A_n} f(x_1, x_2, \dots, x_n)dx_1dx_2 \dots dx_n$$

Anche il concetto di indipendenza si estende a più di due dimensioni. In generale le n variabili aleatorie $X_1, X_2, \dots, X_n$ si dicono indipendenti se per ogni n-upla $A_1, A_2, \dots, A_n$ di sottoinsiemi di $\mathbb{R}$, è soddisfatta l'equazione $\mathcal{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \prod_{i=1}^n \mathcal{P}(X_i \in A_i) \Rightarrow F(a_1, a_2, \dots, a_n) = \prod_{i=1}^n F_{X_i}(a_i)$.

Per concludere, collezioni *infinite* di variabili aleatorie si dicono indipendenti se ogni loro sottogruppo finito è formato da variabili aleatorie tutte indipendenti.

Distribuzioni condizionali

Le relazioni esistenti tra due variabili aleatorie possono essere chiarite dallo studio della distribuzione condizionale di una delle due, dato il valore dell'altra. Si ricorda che presi comunque de eventi E e F con $\mathcal{P}(F) > 0$, la probabilità di E condizionata a F è data dall'espressione $\mathcal{P}(E|F) \coloneqq \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)}$. È naturale applicare questo schema alle variabili aleatorie *discrete*.

Definizione 5: Siano X e Y due variabili aleatorie discrete con funzione di massa congiunta $p(\cdot,\cdot)$. Si dice *funzione di massa di probabilità condizionata* di X dato Y, e si indica con $p_{X|Y}(\cdot \mid \cdot)$, la funzione di due variabili così definita: $\forall x, y$ con $p_Y(y) > 0$

$$\boldsymbol{p_{X|Y}(x|y) \coloneqq \mathcal{P}(X = x|Y = y) = \frac{\mathcal{P}(X = x, Y = y)}{\mathcal{P}(Y = y)} = \frac{p(x,y)}{p_Y(y)}}$$

Se y non è un valore possibile di Y, ovvero se $\mathcal{P}(Y = y) = 0$, la quantità $p_{X|Y}(x|y)$ non è definita.

N.B.: Se X e Y sono variabili congiuntamente continue, non è possibile utilizzare la definizione di distribuzione condizionata valida per quelle discrete.

Definizione 6: Siano X e Y due variabili aleatorie con funzione di densità congiunta f. Si dice *densità condizionale* di X rispetto a Y, e si indica con $f_{X|Y}(\cdot \mid \cdot)$, la funzione di due variabili seguente, che è definita per ogni x e per tutte le y per le quali $f_Y(y) > 0$: $\boldsymbol{f_{X|Y}(x,y) \coloneqq \frac{f(x,y)}{f_Y(y)}}$

Tale definizione è giustificata dalle equazioni $\mathcal{P}\left(a-\frac{\varepsilon}{2} \leq X \leq a+\frac{\varepsilon}{2}\right)=\int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)dx \approx \varepsilon f(a)$ e $\mathcal{P}(a \leq X \leq a+da, b \leq Y \leq b+db)=\int_b^{b+db}\int_a^{a+da} f(x,y)dxdy \approx f(a,b)\,da\,db$. Infatti, moltiplicando il lato sinistro dell'equazione nella Definizione 6 per dx e quello destro per $\frac{dxdy}{dy}$, si ottiene:

$$f_{X|Y}(x,y)dx=\frac{f(x,y)}{f_Y(y)}\frac{dxdy}{dy} \approx \frac{\mathcal{P}(x \leq X \leq x+dx, y \leq Y \leq y+dy)}{\mathcal{P}(y \leq Y \leq y+dy)}= \\ =\mathcal{P}(x \leq X \leq x+dx | y \leq Y \leq y+dy)$$

In altre parole, per valori piccoli di dx e di dy, $f_{X|Y}dx$ rappresenta la probabilità condizionata che X stia nell'intervallo $[x, x+dx]$, sapendo che Y appartiene all'intervallo $[y, y+dy]$.

La densità condizionale ci permette di definire la probabilità di eventi relativi a una variabile aleatoria quando conosciamo il valore di una seconda. Più precisamente se X e Y sono congiuntamente continue e A è un sottoinsieme dei numeri reali, per ogni y si può definire $\boldsymbol{\mathcal{P}(X \in A|Y=y) := \int_A f_{X|Y}(x|y)dx}$.
La grandezza $\mathcal{P}(X \in A|Y=y)$ non è una probabilità condizionata nel senso usuale del termine, in quanto l'evento $\{Y=y\}$ ha sempre probabilità zero. Cionondimeno, sfruttando la densità condizionale di X rispetto a Y siamo riusciti a dare un senso persino un valore numerico a questo oggetto di sicuro interesse pratico.

Si noti che se X e Y sono indipendenti, allora $f_{X|Y}(x,y)=f_X(x),\ \mathcal{P}(X \in A|Y=y)=\mathcal{P}(X \in A)$ e quindi l'indipendenza si comporta nei confronti del condizionamento rispetto a variabili aleatorie continue, esattamente come nel caso più semplice di condizionamento rispetto a eventi di probabilità positiva.

Valore atteso

Uno dei concetti più importanti di tutta la teoria della probabilità è quello di valore atteso.

Definizione 1: Sia X una variabile aleatoria discreta che può assumere i valori $x_1, x_2, \ldots$; il *valore atteso* di X, che si indica con $E[X]$, è (se esiste■) il numero $\boldsymbol{E[X] := \sum_i x_i \mathcal{P}(X=x_i)}$

In altri termini, si tratta della media pesata dei valori possibili di X, usando come pesi le probabilità che tali valori vengano assunti da X. Per questo $E[X]$ è anche detta *media* di X (anche se questo termine è poco consigliato perché può assumere anche altri significati), oppure *aspettazione*.

- Il valore atteso di X è definito solo se la serie $E[X]$ converge in valore assoluto, ovvero deve valere $\sum_i |x_i|\mathcal{P}(X=x_i) < \infty$. In caso contrario si dice che X non ha valore atteso. Tutte le variabili aleatorie che tratteremo nel seguito si supporranno dotate di valore atteso finito. Esempi di distribuzioni per le quali il valore atteso non ha senso sono dati dalle funzioni di massa seguenti:

$$p_1(k)=\begin{cases}2^{-n-1} & se\ k=\pm 2, n=1,2,\ldots \\ 0 & altrimenti\end{cases}, \qquad p_2(k)=\begin{cases}\dfrac{1}{k^2+k} & se\ k=1,2,\ldots \\ 0 & altrimenti\end{cases}$$

Per illustrare il concetto di media pesata, facciamo un semplice esempio. Se X è una variabile aleatoria con funzione di massa $p(0)=\frac{1}{2}=p(1)$ allora $E[X]=0\cdot\frac{1}{2}+1\cdot\frac{1}{2}=\frac{1}{2}$ è semplicemente la media aritmetica dei valori che X può assumere. Però, se $p(0)=\frac{1}{3}, p(1)=\frac{2}{3}$ allora, $E[X]=0\cdot\frac{1}{3}+1\cdot\frac{2}{3}=\frac{2}{3}$ è una media pesata degli stessi valori 0 e 1, dove al secondo è stato dato un peso che è il doppio di quello del primo.

L'interpretazione frequentista della probabilità fornisce una importante giustificazione del concetto di valore atteso. Da tale punto di vista la probabilità di un evento è definita come il limite a cui tende, empiricamente, il rapporto tra il numero di ripetizioni in cui si è realizzato l'evento e il numero totale di ripetizioni di un esperimento. Consideriamo una variabile aleatoria X che può assumere i valori $x_1, x_2, \ldots, x_n$, con funzione di massa di probabilità p. Immaginando che X sia la vincita in una singola mano

di un gioco casuale, qual è la vincita media (nel senso comune del termine) se giochiamo molte mani? Su un numero N di ripetizioni dell'esperimento, ciascuno dei valori x_i si verificherà un certo numero N_i di volte. L'interpretazione frequentista afferma che se N è molto grande, $N_i \approx Np(x_i)$. D'altronde la vincita media è data da $\frac{x_1N_1+x_2N_2+\cdots+x_nN_n}{N} = \sum_{i=1}^{n} x_i \frac{N_i}{N} \approx \sum_{i=1}^{n} x_i p(x_i) = E[X]$ e quindi coincide approssimativamente con la definizione di valore atteso di X.

N.B.: il valore atteso di X non è uno dei valori che X può assumere. Perciò, anche se $E[X]$ è chiamato *valore atteso* di X, non vuole affatto dire che noi ci attendiamo di vedere questo valore, ma piuttosto che ci aspettiamo che sia il **limite** a cui tende il punteggio medio su un numero crescente di ripetizioni.

Definizione 2: Sia X una variabile aleatoria continua con funzione di densità f; il *valore atteso*, o *aspettazione* o anche *media* di X, che si indica con $E[X]$, è (se esiste) la quantità $\boldsymbol{E[X] := \int_{-\infty}^{\infty} xf(x)dx}$

Osservazione: $E[X]$ ha le stesse unità di misura della variabile aleatoria X.

Proprietà del valore atteso

Consideriamo una variabile aleatoria X di cui conosciamo la distribuzione. Se anziché volere calcolare il valore atteso di X, ci interessasse determinare quello di una sua qualche funzione $g(X)$, come potremmo fare? Una prima strada è notare che $g(X)$ stessa è una variabile aleatoria, e quindi ha una sua distribuzione che in qualche modo si può ricavare; dopo averla ottenuta, il valore atteso $E[g(X)]$ si calcola con la definizione usuale applicata alla nuova variabile aleatoria.

Supponiamo di volere determinare il valore atteso di $g(X)$: siccome questa variabile aleatoria assume il valore $g(x)$ quando $X = x$, sembra intuitivo che $E[g(X)]$ coincida con la media pesata dei valori possibili di $g(X)$, usando come peso da dare a $g(x)$ la probabilità (o densità nel caso continuo) che X sia pari a x.

Proposizione 1: Valore atteso di una funzione di variabile aleatoria

1) Se X è una variabile aleatoria discreta con funzione di massa di probabilità p, allora, per ogni funzione reale g, $\boldsymbol{E[g(X)] = \sum_x g(x)p(x)}$
2) Se X è una variabile aleatoria continua con funzione di densità di probabilità f, allora, per ogni funzione reale g, $\boldsymbol{E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx}$

Anche in questo caso si richiede, affinché $E[g(X)]$ abbia senso, che la serie in 1 e l'integrale in 2 convergano in valore assoluto. In ogni caso, nella pratica sono poche (ma non assenti) le variabili aleatorie che non soddisfano tali verifiche.

Corollario 2: Per ogni coppia di costanti reali a e b, $\boldsymbol{E[aX + b] = aE[X] + b}$
Dimostrazione: Nel caso discreto, $E[aX + b] = \sum_x (ax + b)p(x) = a\sum_x xp(x) + b\sum_x p(x) = aE[X] + b$.
Nel caso continuo, $E[aX + b] = \int_{-\infty}^{\infty}(ax + b)f(x)dx = a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx = aE[X] + b$

Se nel precedente corollario si pone $a = 0$, risulta $E[b] = b$. Se invece si pone $b = 0$ allora $E[aX] = aE[X]$

Come già accennato, il termine *valore atteso* fa tra i suoi sinonimi *aspettazione* e *media*. Un'ulteriore denominazione è quella di *momento primo*, con riferimento alla definizione seguente.

Definizione 1: Se $n = 1,2, \ldots$, la quantità $E[X^n]$, quando esiste, è detta *momento n-esimo* della variabile aleatoria X.

Volendo essere più espliciti, si può applicare il Corollario 2 per ricavare,

$$E[X^n] = \begin{cases} \sum_x x^n p(x) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{\infty} x^n f(x)dx & \text{se } X \text{ è continua} \end{cases}$$

Valore atteso della somma di variabili aleatorie

La versione in due dimensioni della Proposizione 1 afferma che se X e Y sono due variabili aleatorie e g è una qualunque funzione di due variabili, allora, se $E[g(X,Y)]$ esiste,

$$E[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y)p(x,y) & \text{nel caso discreto} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy & \text{nel caso continuo} \end{cases}$$

Si può applicare questo enunciato a $g(X,Y) = X + Y$, ottenendo che $\boldsymbol{E[X+Y] = E[X] + E[Y]}$.

Tale risultato è valido sia nel caso discreto (che si lascia come esercizio), sia in quello continuo, come è dimostrato dai passaggi seguenti.

$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy = \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f(x,y)dy\right)dx + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f(x,y)dx\right)dy =$$
$$= \int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y f_Y(y)dy = E[X] + E[Y]$$

Applicando ricorsivamente l'equazione $E[X+Y] = E[X] + E[Y]$ si può estendere la portata alla somma di un numero finito di variabili aleatorie portandoci alla seguente formula generica valida per ogni n:

$$\boldsymbol{E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]}$$

Osservazione 1: Vi è un interessante proprietà della media che emerge quando si vuole *predire* con il minore errore possibile il valore che verrà assunto da una variabile aleatoria. Supponiamo di voler predire il valore di X. Se scegliamo un numero reale c e diciamo che X sarà uguale a c, il quadrato dell'errore che commetteremo è $(X-c)^2$. Mostriamo di seguito che la media dell'errore al quadrato (o errore quadratico medio) è minimizzata se per c scegliamo il valore della media di X. Infatti, detta $\mu := E[X]$,

$$E[(X-c)^2] = E[(X-\mu+\mu-c)^2] = E[(X-\mu)^2 + 2(X-\mu)(\mu-c) + (\mu-c)^2] =$$
$$= E[(X-\mu)^2] + 2(\mu-c)E[X-\mu] + (\mu-c)^2 = E[(X-\mu)^2] + (\mu-c)^2$$

infatti $E[X-\mu] = E[X] - \mu = 0$, allora $E[(X-c)^2] \geq E[(X-\mu)^2]$. Perciò la migliore previsione di X, in termini di minimizzazione dell'errore quadratico medio, è la sua aspettazione.

Varianza

Data una variabile aleatoria X, di cui sia nota la distribuzione, sarebbe molto utile se si potessero riassumere le caratteristiche fondamentali della sua distribuzione con quantità sintetiche come è la media $E[X]$. Tuttavia $E[X]$ è il "baricentro" dei valori possibili di X, e non coglie la vriabilità, la dispersione di questi valori. Ad esempio, se W, Y e Z sono definite come segue: $W := 0$ con probabilità 1,

$Y := \begin{cases} -1 \text{ con probabilità } \frac{1}{2} \\ 1 \quad \text{con probabilità } \frac{1}{2} \end{cases}$, $Z := \begin{cases} -100 \text{ con probabilità } \frac{1}{2} \\ 100 \quad \text{con probabilità } \frac{1}{2} \end{cases}$; allora tutte hanno medai nulla, ma vi è molta più variabilità in Y che non in W (che è addirittura costante), e ancora di più in Z.

Siccome i valori di X sono distribuiti comunque attorno alla sua media $\mu := E[X]$, un approccio per misurare la loro variabilità potrebbe essere quantificare la loro distanza da μ, ad esempio calcolando quanto vale $E[|X-\mu|]$. Questo metodo in linea di principio funziona, nel senso che variabili aleatorie che assumono valori sparsi su un supporto più largo, sono associate a valori più elevati di questa grandezza, tuttavia le difficoltà matematiche che sorgono a causa del valore assoluto sono notevoli, e in realtà se lo si sostituisce con un elevamento al quadrato, si ottiene una definizione molto più fruttuosa.

Definizione 1: Sia X una variabile aleatoria con media μ. La *varianza* di X, che si denota con $\mathrm{Var}(X)$, è (se esiste) la quantità $\boldsymbol{\mathrm{Var}(X) := E[(X-\mu)^2]}$

Esiste una formula alternativa per la varianza, che si ricava in questo modo:

$$\mathrm{Var}(X) \coloneqq E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = \boldsymbol{E[X^2] - E[X]^2}$$

In altri termini, la varianza di X è uguale al valore atteso del quadtrado di X (anche detto il *momento secondo*), meno il quadrato della media di X. Nella pratica questa formula è spesso il miglior modo id calcolare $\mathrm{Var}(X)$

Esempio (Varianza della funzione indicatrice di un evento):
Sia I la funzione indicatrice di un evento A: $I \coloneqq \begin{cases} 1 \text{ se } A \text{ si verifica} \\ 0 \text{ se } A \text{ non si verifica} \end{cases}$.
Allora, notando che $I^2 = I$ sempre (infatti i valori possibili di I sono solamente 0 e 1, che soddisfano $1^2 = 1$ e $0^2 = 0$), $\mathrm{Var}(I) = E[I^2] - E[I]^2 = E[I] - E[I]^2 = E[I](1 - E[I]) = \mathcal{P}(A)\big(1 - \mathcal{P}(A)\big)$.

Una utile identità che riguarda la varianza è la seguente. Per ogni coppia di costanti reali a e b,

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

Per dimostrarla, poniamo $\mu \coloneqq E[X]$ e ricordiamo che $E[aX + b] = aE[X] + b = a\mu + b$, in modo tale che

$$\mathrm{Var}(aX + b) \coloneqq E[(aX + b - E[aX + b])^2] = E[(aX + b - a\mu - b)^2] = E[a^2(X-\mu)^2] = \\ = a^2 E[(X-\mu)^2] = a^2\mathrm{Var}(X)$$

Definizione 2: La quantità $\sqrt{\mathrm{Var}(X)}$ è detta *deviazione standard* della variabile aleatoria X.

La covarianza e la varianza della somma di variabili aleatorie

Come abbiamo visto nella sezione del Valore atteso della somma di variabili aleatorie, la media della somma di variabili aleatorie coincide con la somma delle loro medie. Per la varianza questo in generale non è vero. Ad esempio, $\mathrm{Var}(X + X) = \mathrm{Var}(2X) = 2^2\mathrm{Var}(X) = 4\mathrm{Var}(X) \neq \mathrm{Var}(X) + \mathrm{Var}(X)$.

Vi è tuttavia un caso importante in cui la varianza della somma di due variabili aleatorie è pari alla somma delle loro varianze, ovvero quando le variabili aleatorie sono indipendenti. Prima di dimostrare questo risultato, però dobbiamo definire il concetto di covarianza di due variabili aleatorie.

Definizione 1: Siano assegnate due variabili aleatorie X e Y di media μ_X e μ_Y rispettivamente. La loro *covarianza*, che si indica con $\mathrm{Cov}(X, Y)$ è (se esiste) la quantità $\boldsymbol{\mathrm{Cov}(X, Y) \coloneqq E[(X - \mu_X)(Y - \mu_Y)]}$

Si può ottenere anche una formula alternativa più semplice che si trova espandendo il prodotto al secondo membro: $\mathrm{Cov}(X, Y) = E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] = E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y = \boldsymbol{E[XY] - E[X]E[Y]}$.

Dalla definizione 1 si deducono alcune semplici proprietà, quali la simmetria, $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ e il fatto che la covarianza generalizza il concetto di varianza, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.

Un altro enunciato interessante, la cui semplice dimostrazione la si lascia come esercizio, è che per ogni costante a si ha che $\mathrm{Cov}(aX, Y) = a\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, aY)$.

Lemma 1: Se X, Y e Z sono variabili aleatorie qualsiasi, $\boldsymbol{\mathrm{Cov}(X + y) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)}$
Dimostrazione: $\mathrm{Cov}(X + Y, Z) = E[(X + Y)Z] - E[X + Y]E[Z] = E[XZ + YZ] - (E[X] + E[Y])E[Z] = E[XZ] - E[X]E[Z] + E[YZ] - E[Y]E[Z] = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z)$

Tele lemma può essere generalizzato come segue: $\mathrm{Cov}(\sum_{i=1}^n X_i, Y) = \sum_{i=1}^n \mathrm{Cov}(X_i, Y)$

Proposizione 2: Se $X_1, \dots, X_n$ e $Y_1, \dots, Y_m$ sono variabili aleatorie qualsiasi,

$$\mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}(X_i, Y_j)$$

Dimostrazione: $\mathrm{Cov}\big(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\big) = \sum_{i=1}^n \mathrm{Cov}\big(X_i, \sum_{j=1}^m Y_j\big) = \sum_{i=1}^n \mathrm{Cov}\big(\sum_{j=1}^m Y_j, X_i\big) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}\big(Y_j, X_i\big) = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}\big(X_i, Y_j\big)$.

Utilizzando a questo punto l'equazione $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ sulla variabile aleatoria $\sum_i X_i$, si ottinene finalmente la formula che fornisce la varianza di una somma di variabili aleatorie:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{n} X_i\,, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j)$$

Nel caso in cui $n = 2$, la formula precedente si riduce a $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$

Teorema 3: Se X e Y sono variabili aleatorie indipendenti, allora $\boldsymbol{E[XY] = E[X]E[Y]}$.
Questo inoltre implica che $\boldsymbol{\mathrm{Cov}(X,Y) = 0}$ e quindi che, se $X_1, \dots, X_n$ sono indipendenti

$$\mathbf{Var}\left(\sum_{i=1}^{n} \boldsymbol{X_i}\right) = \sum_{i=1}^{n} \mathbf{Var}(\boldsymbol{X_i})$$

Dimostrazione: Proviamolo che $E[XY] = E[X]E[Y]$ nel caso in cui X e Y sono entrambe discrete (il caso in cui siano continue si prova in maniera analoga): $E[XY] = \sum_i \sum_j x_i y_j \mathcal{P}(X = x_i, Y = y_j) = \sum_i \sum_j x_i y_j \mathcal{P}(X = x_i)\mathcal{P}(Y = y_j) = \sum_i x_i \mathcal{P}(X = x_i) \sum_j y_i \mathcal{P}(Y = y_j) = E[X]E[Y]$. Che la covarianza di X e Y sia nulla segue dall'equazione $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y]$, mentre l'ultima parte dell'enunciato è conseguenza dell'equazione $\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j)$.

In generale si può mostrare che un valore positivo di $\mathrm{Cov}(X,Y)$ indica che X e Y tendenzialmente assumono valori grandi o piccoli contemporaneamente. La forza della relazione tra X eY è misurata più propriamente dal *coefficiente di correlazione lineare*, un numero puro (senza unità di misura) che tiene conto anche delle derivazioni standard di X e Y. Esso si indica con $\mathrm{Corr}(X,Y)$ ed è definito come

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Si può dimostrare che questa quantità è sempre compresa tra -1 e $+1$.

La funzione generatrice dei momenti

Definizione 1: La *funzione generatrice dei momenti*, o più semplicemente *funzione generatrice* ϕ, di una variabile aleatoria X. È definita, per tutti i t reali per i quali il valore atteso di e^{tX} ha senso, dall'espressione

$$\boldsymbol{\phi(t) \coloneqq E[e^{tX}] = \begin{cases} \sum_{x} e^{tx} p(x) & \textbf{se } \boldsymbol{X} \textbf{ è discreta} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,dx & \textbf{se } \boldsymbol{X} \textbf{ è continua} \end{cases}}$$

Il nome adottato deriva dal fatto che tutti i momenti di cui è dotata X possono essere ottenuti derivando più volte nell'origine la funzione $\phi(t)$. Ad esempio, $\phi'(t) = \frac{d}{dt}E[e^{tX}] = E\left[\frac{d}{dt}e^{tX}\right] = E[Xe^{tX}]$ da cui $\phi'(0) = E[X]$. Analogamente, $\phi''(t) = \frac{d^2}{dt^2}E[e^{tX}] = E\left[\frac{d^2}{dt^2}e^{tX}\right] = E[X^2e^{tX}]$ da cui $\phi''(0) = E[X^2]$, è il momento secondo di X. Più in generale, la derivata n-esima di $\phi(t)$ calcolata in 0 fornisce il momento n-esimo di X: $\boldsymbol{\phi^{(n)}(0) = E[X^n], n \geq 1}$.

Un'altra importante proprietà di ϕ è che la funzione generatrice dei momenti della somma di variabili aleatorie indipendenti è il prodotto delle funzioni generatrici delle singole variabili aleatorie.

Proposizione 1: Se X e Y sono variabili aleatorie indipendenti con funzioni generatrici ϕ_X e ϕ_Y rispettivamente, e se ϕ_{X+Y} è la funzione generatrice dei momenti di $X+Y$, allora $\boldsymbol{\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)}$

Dimostrazione: Si noti intanto che se X e Y sono indipendenti, lo sono anche le variabili aleatorie e^{tX} ed e^{tY}. Infatti, comunque si scelgano A e B, $\mathcal{P}(e^{tX} \in A, e^{tY} \in B) = \mathcal{P}(e^{tX} \in A)\mathcal{P}(e^{tY} \in B)$ (questa equazione che mostra l'indipendenza va verificata). D'altra parte, se A' è l'insieme formato dai numeri z tali

che $e^{tz} \in A$, allora $e^{tX} \in A \Leftrightarrow X \in A'$. Se si definisce analogamente B', si vede che $\mathcal{P}(e^{tX} \in A, e^{tY} \in B) = \mathcal{P}(X \in A', Y \in B') = \mathcal{P}(X \in A')\mathcal{P}(Y \in B') = \mathcal{P}(e^{tX} \in A)\mathcal{P}(e^{tY} \in B)$. A questo punto, basta sfruttare il fatto che l'indipendenza implica che la media del prodotto è il prodotto delle medie, per concludere che $\phi_{X+Y}(t) \coloneqq E\left[e^{t(X+Y)}\right] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = \phi_X(t)\phi_Y(t)$

Osservazione: Un ulteriore risultato che mostra l'importanza della funzione generatrice dei momenti è che essa *determina la distribuzione*, nel senso che due variabili aleatorie con identica funzione generatrice hanno necessariamente la stessa legge (e quindi la stessa funzione di ripartizione, e la stessa funzione di massa, ovvero la stessa densità).

La legge debole dei grandi numeri

Cominciamo con un risultato preliminare.

Proposizione 1 (Disuguaglianza di Markov): Se X è una variabile aleatoria che non è mai negativa, allora per ogni $a > 0$, $\boldsymbol{\mathcal{P}(X \geq a) \leq \frac{E[X]}{a}}$

Dimostrazione: Diamo la dimostrazione nel caso che X sia continua con densità f.

$$E[X] = \int_0^\infty xf(x)dx = \underbrace{\int_0^a xf(x)dx}_{\text{quantità positiva}} + \int_a^\infty xf(x)dx \geq \underbrace{\int_a^\infty xf(x)dx}_{x \geq a} \geq \int_a^\infty af(x)dx = a\int_a^\infty f(x)dx$$
$$= a\mathcal{P}(X \geq a)$$

E l'enunciato segue dividendo entrambi i termini per a.

Come corollario, ricaviamo la proposizione seguente.

Proposizione 2 (Disuguaglianza di Chebyshev): Se X è una variabile aleatoria con media μ e varianza σ^2, allora per ogni $r > 0$, $\boldsymbol{\mathcal{P}(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2}}$.

Dimostrazione: Gli eventi $\{|X - \mu| \geq r\}$ e $\{(X - \mu)^2 \geq r^2\}$ coincidono e sono quindi equiprobabili. Visto che $(X - \mu)^2$ è una variabile aleatoria non negativa, possiamo applicare la disuguaglianza di Markov con $a = r^2$ ottenendo: $\mathcal{P}(|X - \mu| \geq r) = \mathcal{P}((X - \mu)^2 \geq r^2) \leq \frac{E[(X-\mu)^2]}{r^2} = \frac{\sigma^2}{r^2}$.

L'importanza delle disuguaglianze di Markov e di Chebyshev, sta nel fatto che permettono di limitare le probabilità di eventi rari che riguardano variabili aleatorie di cui conosciamo solo la media, oppure la media e la varianza. Naturalmente, quando la distribuzione è nota, tali probabilità possono essere calcolata esattamente e non vi è necessità di ridursi all'utilizzo di maggiorazioni.

Concludiamo questo paragrafo provando, grazie alla disuguaglianza di Chebyshev, la legge debole dei grandi numeri, un enunciato che afferma che la media aritmetica di n copie indipendenti di una variabile aleatoria tende al valore atteso di quest'ultima per n che tende all'infinito. Tale convergenza si precisa dicendo che scelto un ε comunque piccolo, la media aritmetica si discosta dal valore atteso per più di ε con probabilità che tende a zero, quando n tende all'infinito.

Teorema 3 (Legge debole dei grandi numeri): Sia $X_1, X_2, \ldots$ una successione di variabili aleatorie *indipendenti e identicamente distribuite* (i.i.d.), tutte con media $E[X_i] = \mu$. Allora per ogni $\varepsilon > 0$,

$$\boldsymbol{\mathcal{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad \text{quando } n \to \infty}$$

Dimostrazione: Proveremo il risultato solo sotto l'ipotesi aggiuntiva che le X_i abbiano varianza finita σ^2. Dalle proprietà di media e varianza segue che

$$E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \quad \text{e} \quad \text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

La seconda ad esempio si prova in questo modo:

$$\mathrm{Var}\left(\frac{X_1+\cdots+X_n}{n}\right)=\frac{1}{n^2}\mathrm{Var}(X_1+\cdots+X_n)=\frac{\mathrm{Var}(X_1)+\cdots+\mathrm{Var}(X_n)}{n^2}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

Segue allora dalla disuguaglianza di Chebyshev applicata alla variable aleatoria $\frac{(X_1+\cdots+X_n)}{n}$, che

$$\mathcal{P}\left(\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|>\varepsilon\right)\leq\frac{\sigma^2}{n\varepsilon^2}$$

Poiché il secondo membro tende a zero per n che tende all'infinito, l'enunciato è provato.

Una applicazione di questo teorema è la seguente, che permette anche di giustificare l'interpretazione frequentista della probabilità di un evento. Supponiamo di ripetere in successione molte copie indipendenti di un esperimento, in ciascuna delle quali può verificarsi un certo evento E. Ponendo

$$X_i=\begin{cases}1 & \text{se } E \text{ si realizza nell'esperimento } i\text{-esimo}\\ 0 & \text{se } E \text{ non si realizza nell'esperimento } i\text{-esimo}\end{cases}$$

la sommatoria $X_1+X_2+\cdots+X_n$ rappresenta il numero di prove, tra le prime n, in cui si è verificato l'evento E. Poiché $E[X_i]=\mathcal{P}(X_i=1)=\mathcal{P}(E)$ si deduce che la frazione delle n prove nelle quali si realizza E, tende (nel senso della legge debole dei grandi numeri) alla probabilità $\mathcal{P}(E)$.

Sul concetto di misurabilità

Corrispondenze e Applicazioni

Siano S e T due insiemi. Un qualsiasi sottoinsieme del prodotto cartesiano $S\times T$ è una **corrispondenza** tra S e T. Quindi, gli elementi di una corrispondenza sono coppie ordinate in quanto il primo elemento della coppia deve appartenere a S e il secondo a T. Ad esempio, se $S=\{a,b,c,d,e\}$ e $T=\{1,2\}$, alcune corrispondenze tra S e T sono elencate di seguito: $G_1=\{(a,1),(b,1),(c,2),(d,2)\}, G_2=\{(b,1),(b,2)\}$, $G_3=\{(a,1),(b,1),(c,1),(d,1),(e,1)\}, G_4=\{(a,1),(b,1),(c,1),(d,1),(e,2)\}$.

Si osservi che in G_1 l'elemento $e\in S$ non compare come primo elemento di una coppia. In G_2 appare (due volte) un solo un elemento di S. Invece, in G_2e in G_4 ci sono tutti gli elementi di S (una sola volta).

Nel caso in cui in una corrispondenza ogni elemento S appare una ed una sola volta la corrispondenza è chiamata **applicazione**. Per le applicazioni, l'insieme S è il *dominio*, l'insieme T è il *codominio* e il secondo elemento di ciascuna sua coppia si dice essere l'*immagine* del primo. Quindi, negli esempi elencati G_3 è un'applicazione (l'applicazione costante) e 1 è l'immagine di tutti gli elementi di S. Anche G_4 è un'applicazione e 2 è l'immagine di $e\in S$.
Di solito, per un'applicazione si usa la notazione $f\colon S\to T$ e se $s\in S$ allora $f(s)\in T$ è la sua immagine.

Le applicazioni possono essere *iniettive*: non ci sono due o più coppie con il secondo elemento uguale. In altri termini, un'applicazione è iniettiva quando elementi diversi del dominio, $s_1\neq s_2$, hanno immagini diverse nel codominio: $f(s_1)\neq f(s_2)$. L'applicazione costante non è iniettiva.

Le applicazioni possono essere *suriettive*: ogni elemento di T è presente nella corrispondenza. In altri termini, per ogni elemento di $t\in T$ esiste almeno un elemento $s\in S$ per il quale $f(s)=t$. Quindi un'applicazione costante può essere suriettiva solo nel caso di un codominio costituito da un singoletto. Negli esempi elencati sopra, G_3 non è applicazione suriettiva mentre lo è G_4.

Un'applicazione che è sia iniettiva che suriettiva è detta *biettiva*. Infine, se tra due insiemi S e T si può definire un'applicazione biettiva allora essi hanno la stessa cardinalità (oppure, si dice che S e T sono *equipotenti*). Un insieme che è equipotente ad una sua parte propria si dice *infinito*.

Controimmagine di un'applicazione

Siano S e T due insiemi e f un'applicazione di S in T. La controimmagine tramite f di un qualsiasi sottoinsieme B di T, indicata con $f^{-1}(B)$ è l'insieme A costituito dagli elementi di S la cui immagine è in B: $B\subseteq T,\ A=f^{-1}(B)=\{s\in S\colon f(s)\in B\}\subseteq S$.

Per la controimmagine sussistono i seguenti risultati:

1) La controimmagine del vuoto è il vuoto.
2) La controimmagine del codominio è il dominio.
3) Le controimmagini di insiemi disgiunti in T sono insiemi disgiunti in S.
4) La controimmagine di una unione è l'unione delle controimmagini.
5) La controimmagine di una intersezione è l'intersezione delle controimmagini.
6) La controimmagine di una unione fatta da sottoinsiemi di T a due a due disgiunti è unione di sottoinsiemi di S a due a due disgiunti.
7) La controimmagine della differenza di due sottoinsiemi di T è la differenza delle rispettive controimmagini.
8) La controimmagine del complementare di un sottoinsieme di T è il complementare della controimmagine.

Nel seguito per indicare la controimmagine di un sottoinsieme B si userà la notazione abbreviata: $f^{-1}(B)$.

Applicazioni misurabili e numeri aleatori

Se $\mathcal{F}$ è una σ-algebra sull'insieme S, la coppia $(S, \mathcal{F})$ si dice uno **spazio misurabile** e gli elementi di $\mathcal{F}$ sono chiamati *insiemi misurabili* o solo *misurabili*.

Siano $(S_1, \mathcal{F}_1)$ e $(S_2, \mathcal{F}_2)$ due spazi misurabili. Un'applicazione f di S_1 in S_2 si dice essere $\mathcal{F}_1/\mathcal{F}_2$*-misurabile* se la controimmagine di un qualsiasi elemento di $\mathcal{F}_2$ appartiene a $\mathcal{F}_1$: $B \in \mathcal{F}_2,\ f^{-1}(B) \in \mathcal{F}_1$.

All'insieme $\mathbb{R}$ è associata, di norma, la σ-algebra generata dagli intervalli aperti e limitati che si designa con il simbolo $\mathcal{B}$. Gli elementi di $\mathcal{B}$ sono chiamati *insiemi di Borel* oppure ***boreliani***.

Si consideri ora uno spazio di probabilità $(\Omega, \mathcal{F}, \mathcal{P})$ e lo spazio misurabile $(\mathbb{R}, \mathcal{B})$. Un *numero aleatorio* è un'applicazione X di Ω in $\mathbb{R}$ con la condizione che essa sai $\mathcal{F}/\mathcal{A}$-misurabile (oppure, dato che $\mathcal{A}$ è fissata, che essa sia $\mathcal{F}$-misurabile). In altri termini, la controimmagine tramite X di un boreliano deve essere un evento: $B \in \mathcal{B},\ X^{-1}(B) \in \mathcal{F}$.

Con il requisito della $\mathcal{F}$-misurabilità di X si può definire una misura di probabilità sulla σ- algebra $\mathcal{B}$: $B \in \mathcal{B},\ \mathcal{P}_X(B) \coloneqq \mathcal{P}(X \in B)$. Infatti,

- $B \in \mathcal{B},\ \mathcal{P}_X(B) = \mathcal{P}(X \in B) \geq 0$;
- Per il secondo punto del precedente elenco si ha che $\mathcal{P}_X(\mathbb{R}) = \mathcal{P}(X \in \mathbb{R}) = \mathcal{P}(\Omega) = 1$;
- Per il terzo e il quarto punto del precedente elenco si ha che $(B_n)_{n\in\mathbb{N}} \subseteq \mathcal{B} : B_i \cap B_j = \emptyset\ (i \neq j)$,

$$\mathcal{P}_X\left(\bigcup_{n\in\mathbb{N}} B_n\right) = \mathcal{P}\left(X \in \bigcup_{n\in\mathbb{N}} B_n\right) = \mathcal{P}\left(\bigcup_{n\in\mathbb{N}} \{X \in B_n\}\right) = \sum_{n\in\mathbb{N}} \mathcal{P}(X \in B_n) = \sum_{n\in\mathbb{N}} \mathcal{P}_X(B_n)$$

La misura di probabilità $\mathcal{P}_X$ è la *distribuzione* oppure la *legge* del numero aleatorio X. Essa fornisce tutte le informazioni probabilistiche relative al numero aleatorio X ed esse si basano sulla misura di probabilità $\mathcal{P}$ della descrizione matematica dell'esperimento aleatorio sottostante.

3. Modelli di variabili aleatorie

Alcuni tipi di variabili aleatorie compaiono molto frequentemente in natura o negli studi tecnologici. In questo capitolo, presentiamo dei modelli di variabili aleatorie particolari, che sono caratterizzati dalla grande generalità dei campi applicativi nei quali compaiono.

Variabili aleatorie di Bernoulli e binomiali

Supponiamo che venga realizzata una prova, o un esperimento, il cui esito può essere solo un "successo" o un "fallimento". Se definiamo la variabile aleatoria X in modo che sia $X = 1$ nel primo caso e $X = 0$ nel secondo, la funzione di massa di probabilità di X è data da $\mathcal{P}(X = 0) = 1 - p$ e $\mathcal{P}(X = 1) = p$ dove con p abbiamo indicato la probabilità che l'esperimento registri un "successo". Ovviamente si ha $0 \leq p \leq 1$.

Definizione 1: Una variabile aleatoria X si dice di *Bernoulli* o *bernoulliana* se la sua funzione di massa di probabilità è del tipo $\boldsymbol{\mathcal{P}(X = 0) = 1 - p, \mathcal{P}(X = 1) = p}$, per una scelta opportuna del parametro p.

In altri termini; una variabile aleatorie è bernoulliana se può assumere solo i valori 0 e 1. Il suo valore atteso è dato da $\boldsymbol{E[X] \coloneqq 1 \cdot \mathcal{P}(X = 1) + 0 \cdot \mathcal{P}(X = 0) = p}$ ed è quindi pari alla probabilità che la variabile aleatoria assuma il valore 1.

Definizione 2: Supponiamo di realizzare n ripetizioni indipendenti di un esperimento, ciascuna delle quali può concludersi in un "successo" con probabilità p, o in un "fallimento" con probabilità $1 - p$. Se X denota il numero totale di successi, X si dice variabile aleatoria *binomiale* di parametri (n, p).

La funzione di massa di probabilità per una variabile aleatoria binomiale di parametri (n, p) è data da

$$\boldsymbol{\mathcal{P}(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, \qquad i = 0, 1, \dots, n}$$

dove il coefficiente binomiale: $\binom{n}{i} = \frac{n!}{i!(n-i)!}$, rappresenta il numero di combinazioni differenti che possiamo ottenere scegliendo i elementi da un insieme di n oggetti.

La correttezza dell'equazione precedente può essere verificata nel modo seguente: innanzitutto, fissata una qualunque sequenza di esiti con i successi e $n - i$ fallimenti, la probabilità che si realizzi esattamente tale sequenza è $p^i(1 - p)^{n-i}$ per l'indipendenza delle ripetizioni. L'equazione segue quindi dal contare quante sono le diverse sequenze di esiti con questa caratteristica. Esse sono $\binom{n}{i}$ perché corrispondono a tutti i modi in cui si possono scegliere gli i esperimenti che hanno dato esito positivo sugli n in totale.

Si noti che la somma delle probabilità di tutti i valori possibili di una variabile aleatoria binomiale, è pari a 1 per la formula delle potenze del binomio:

$$\sum_i \mathcal{P}(X = i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1 - p)^{n-i} = [p + (1 - p)]^n = 1$$

Per come è stata definita la variabile aleatoria binomiale di parametri (n, p) (il numero di esperimenti con esito positivo, su n ripetizioni indipendenti, ciascuna con probabilità di successo p), essa può essere rappresentata come somma di bernoulliane. Più precisamente, se X è binomiale di parametri (n, p), si può scrivere $X = \sum_{i=1}^{n} X_i$ dove X_i è la funzione indicatrice del successo dell'i-esimo esperimento:

$$X_i = \begin{cases} 1 \text{ se la prova } i\text{-esima ha successo} \\ 0 \text{ altrimenti} \end{cases}$$

È evidente che le X_i sono tutte bernoulliane di parametro p, quindi abbiamo che $E[X_i] = p$ per la formula successiva alla definizione 1; $E\left[X_i^2\right] = p$ (infatti $X_i \equiv X_i^2$); dunque,

$$\boldsymbol{\mathrm{Var}(X_i) = E\left[X_i^2\right] - E[X_i]^2 = p - p^2 = p(1 - p)}$$

Per quanto riguarda X, poi, dalle proprietà di media e varianza e dalla rappresentazione fornita dall'equazione $X = \sum_{i=1}^{n} X_i$, otteniamo che

$$E[X] = \sum_{i=1}^{n} E[X_i] = np, \qquad \mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p) \text{ per l'indipendenza delle } X_i$$

Osservazione 1: Se X_1 e X_2 sono binomiali di parametri (n_1, p) e (n_2, p) e sono indipendenti, la loro somma $X_1 + X_2$ è binomiale di parametri $(n_1 + n_2, p)$. Questo può essere facilmente dedotto dal fatto che si effettuano n_1 e poi ancora n_2 prove indipendenti dello stesso esperimento con probabilità di successo p, se X_1 e X_2 rappresentano il numero di successi nelle due *tranche* di prove, $X_1 + X_2$ rappresenta il numero di successi sul totale delle $n_1 + n_2$ prove. È quindi binomiale con i parametri precedentemente citati per costruzione.

Calcolo esplicito della distribuzione binomiale

Supponiamo che X sia binomiale di parametri (n, p). Per potere calcolare operativamente la funzione di ripartizione

$$\mathcal{P}(X \leq i) = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \qquad i = 0,1, \dots, n$$

o la funzione di massa

$$\mathcal{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad i = 0,1, \dots, n$$

è molto utile la seguente relazione tra $\mathcal{P}(X = k+1)$ e $\mathcal{P}(X = k)$:

$$\mathcal{P}(X = k+1) = \frac{p}{1-p} \frac{n-k}{k+1} \mathcal{P}(X = k)$$

la cui dimostrazione è lasciata come esercizio.

Variabili aleatorie di Poisson

Proseguiamo la panoramica con un'altra importante variabile aleatoria discreta che assume solo i valori interi non negativi.

Definizione 1: Una variabile aleatoria X che assuma i valori $0,1,2, \dots$, è una variabile aleatoria di *Poisson* o *poissoniana* di parametro $\lambda > 0$, se la sua funzione di massa di probabilità è data da

$$\boldsymbol{\mathcal{P}(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \qquad i = 0, 1, 2, \dots}$$

È immediato verificare che la precedente equazione rappresenta una funzione di massa accettabile, infatti

$$\sum_{i=0}^{\infty} \mathcal{P}(X = i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1$$

Sia X una variabile aleatoria di Poisson. Per determinare la media e la varianza, calcoliamo la sua funzione generatrice dei momenti.

$$\phi(t) = E[e^{tX}] = \sum_{i=0}^{\infty} e^{ti} \mathcal{P}(X = i) = e^{-\lambda} \sum_{i=0}^{\infty} e^{ti} \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

Derivando si trova allora $\phi'(t) = \lambda e^t e^{\lambda(e^t-1)}$ e $\phi''(t) = (\lambda e^t)^2 e^{\lambda(e^t-1)} + \lambda e^t e^{\lambda(e^t-1)}$ e valutando le due espressioni in $t = 0$, si ottiene $E[X] = \phi'(0) = \lambda$ e $\mathrm{Var}(X) = \phi''(0) - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.
Quindi, sia il valore atteso, sia la varianza delle poissoniane coincidono con il parametro λ.

La variabile aleatoria di Poisson ha un vasto campo di applicazioni, in aree numerose e diverse, anche perché può esser utilizzata come approssimazione di una binomiale di parametri (n, p), quando n è molto grande e p molto piccolo. Per convincerci di questo fatto, sia X una variabile aleatoria binomiale di parametri (n, p), e si ponga $\lambda = np$. Allora

$$\mathcal{P}(X=i)=\frac{n!}{(n-i)!\,i!}p^i(1-p)^{n-i}=\frac{n(n-1)\dots(n-i+1)}{i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}=$$
$$=\frac{n(n-1)\dots(n-i+1)}{n^i}\cdot\frac{\lambda^i}{i!}\cdot\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^i}$$

Se si suppone che n sia molto grande e p molto piccolo, valgono le seguenti approssimazioni,

$$\left(1-\frac{\lambda}{n}\right)^n\approx e^{-\lambda},\qquad \frac{n}{n}\cdot\frac{n-1}{n}\cdot\dots\cdot\frac{n-i+1}{n}\approx 1,\qquad \left(1-\frac{\lambda}{n}\right)^i\approx 1$$

E quindi, se n è grande, p è piccolo e $\lambda=np$: $\boldsymbol{\mathcal{P}(X=i)\approx\frac{\lambda^i}{i!}e^{-\lambda}}$

In altri termini, il totale dei "successi" in un gran numero n di ripetizioni indipendenti di un esperimento che ha una piccola probabilità di riuscita p, è una variabile aleatoria con distribuzione approssimativamente di Poisson, con media $\lambda=np$.

Quelli che seguono sono alcuni esempi di variabili aleatorie che seguono con buona approssimazione la legge di Poisson (ovvero che rispettano approssimativamente l'equazione della Definizione 1, per una qualche scelta di λ):

1) Il numero di refusi in una pagina (o un insieme di pagine) di un libro.
2) Il numero di individui, all'interno di una certa categoria di persone, che raggiungono i cento anni di età.
3) La quantità di numeri telefonici errati che vengono composti in una giornata.
4) Il numero di transistor che si guastano nel loro primo giorno di utilizzo.
5) Il numero di clienti che entrano in un ufficio postale nell'arco di una giornata.
6) La quantità di particelle alfa emesse in un periodo di tempo fissato da un campione di materiale radioattivo.

Ciascuna delle variabili aleatorie dei precedenti, come di numerosi altri esempio, è approssimativamente di Poisson per lo stesso motivo; ovvero, perché alcune variabili aleatorie binomiali si possono approssimare con poissoniane. Ad esempio, possiamo supporre che ciascuna lettera tipografata nella pagina di un libro abbia una probabilità p molto piccola di essere sbagliata, e così il numero totale di refusi è circa poissoniano con media $\lambda=np$, dove n è il (presumibilmente elevato) numero di lettere in una pagina di testo. Analogamente, possiamo immaginare che all'interno di una certa categoria di persone, ciascuno indipendentemente dagli altri abbia una piccola probabilità p di superare i cento anni di età, e quindi il numero di individui ai quali capiterà è approssimativamente una variabile aleatoria di Poisson di media $\lambda=np$, dove n è il numero (elevato) di persone di quel gruppo. Si ragioni sul perché anche le restanti variabili aleatorie degli esempi dal 2 al 6 debbano avere distribuzione approssimativamente poissoniana.

La distribuzione di Poisson è *riproducibile*, nel senso che la somma di due poissoniane indipendenti è ancora una poissoniana. Per dimostrarlo, siano assegnate due variabili aleatorie di Poisson e indipendenti, X_1 e X_2, di parametri rispettivamente λ_1 e λ_2, e calcoliamo la funzione generatrice dei momenti della loro somma: $\phi_{X_1+X_2}(t)=\phi_{X_1}(t)\phi_{X_2}(t)=e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)}=e^{(\lambda_1+\lambda_2)(e^t-1)}$. Siccome $e^{(\lambda_1+\lambda_2)(e^t-1)}$ è la funzione generatrice di una poissoniana di media $\lambda_1+\lambda_2$, e $\phi_{X_1+X_2}$ determina la distribuzione di X_1+X_2, si deduce che X_1+X_2 è una variabile aleatoria di Poisson media $\lambda_1+\lambda_2$.

Calcolo esplicito della distribuzione di Poisson

Se X è una variabile aleatoria di Poisson di media λ, allora

$$\frac{\boldsymbol{\mathcal{P}(X=i+1)}}{\boldsymbol{\mathcal{P}(X=i)}}=\frac{\boldsymbol{\lambda^{i+1}e^{-\lambda}}}{\boldsymbol{(i+1)!}}\frac{\boldsymbol{i!}}{\boldsymbol{\lambda^i e^{-\lambda}}}=\frac{\boldsymbol{\lambda}}{\boldsymbol{i+1}}$$

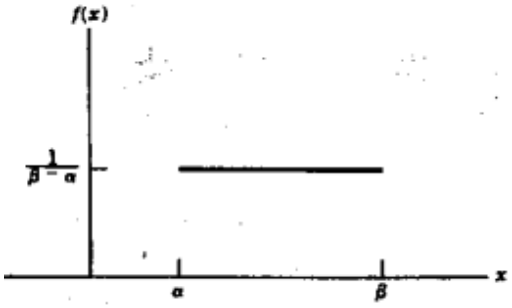
È possibile utilizzare la precedente equazione ricorsivamente, a partire da $\mathcal{P}(X=0)=e^{-\lambda}$, per calcolare successivamente $\mathcal{P}(X=1)=\lambda\mathcal{P}(X=0),\ \mathcal{P}(X=2)=\frac{\lambda}{2}\mathcal{P}(X=1),\dots,\mathcal{P}(X=i+1)=\frac{\lambda}{i+1}\mathcal{P}(X=i)$

Variabili aleatorie uniformi

Definizione 1: Una variabile aleatoria continua si dice *uniforme* sull'intervallo $[\alpha,\beta]$, se ha funzione di densità data da

$$\boldsymbol{f(x)=\begin{cases}\dfrac{1}{\beta-\alpha} & \text{se } \alpha\le x\le\beta\\ 0 & \text{altrimenti}\end{cases}}$$

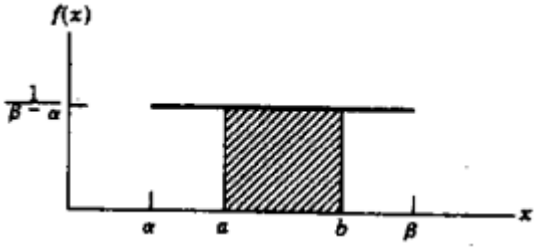
Il grafico di una densità di questo è illustrato nella figura a destra. Si noti che essa soddisfa le condizioni per essere una densità di probabilità, in quanto



$$\int_{-\infty}^{\infty} f(x)dx=\frac{1}{\beta-\alpha}\int_{\alpha}^{\beta}dx=1$$

Per potere assumere la distribuzione uniforme, nella pratica, occorre che lav ariabile aleatoria abbia come valori possibili i punti di un intervallo limitato $[\alpha,\beta]$; inoltre si deve poter supporre che essa abbia le stese probabilità di cadere vicino ad un qualunque punto dell'intervallo.

La probabiità che una variabile aleatoria X, uniforme su $[\alpha,\beta]$, appartenga ad un dato intervallo contenuto in $[\alpha,\beta]$ è pari al rapporto tra le lunghezze dei due intervalli. Infatti, se $[a,b]$ è contenuto in $[\alpha,\beta]$ (come si nota dalla figura a destra),



$$\mathcal{P}(a<X<b)=\frac{1}{\beta-\alpha}\int_{\alpha}^{\beta}dx=\frac{b-a}{\beta-\alpha}$$

Determiniamo ora la media di una variabile aleatoria X, uniforme su $[\alpha,\beta]$:

$$\boldsymbol{E[X]:=\int_{\alpha}^{\beta}\frac{x dx}{\beta-\alpha}=\frac{\beta^2-\alpha^2}{2(\beta-\alpha)}=\frac{(\beta-\alpha)(\beta+\alpha)}{2(\beta-\alpha)}=\frac{\alpha+\beta}{2}}$$

Perciò il valore atteso di una variabile aleatoria uniforme è il punto medio del suo intervallo di definizione, come si poteva intuire direttamente senza fare i calcoli.

Per ottenere la varianza ci serve il momento secondo. $E[X^2]=\int_{\alpha}^{\beta}\frac{x^2dx}{\beta-\alpha}=\frac{\beta^3-\alpha^3}{3(\beta-\alpha)}=\frac{\alpha^3+\alpha\beta+\beta^2}{3}$ quindi

$$\boldsymbol{\mathrm{Var}(X)=\frac{\alpha^3+\alpha\beta+\beta^2}{3}-\left(\frac{\alpha+\beta}{2}\right)^2=\frac{\alpha^2-2\alpha\beta+\beta^2}{12}=\frac{(\beta-\alpha)^2}{12}}$$

L'esempio che segue fornisce un'illustrazione di come si possano usare i semplici numeri generati dal calcolatore per simulare esperimenti probabilistici anche complessi. Consideriamo una clinica sperimentale che desidera testare l'efficacia di un nuovo farmaco per ridurre il livello di colesterolo nel sangue. Vengono assunti 1000 volontari che si sottoporranno al test. Per non trascurare la possibilità che il livello di colesterolo durante il periodo di somministrazione possa cambiare per fattori esterni (come i cambiamenti climatici), si decide di dividere i volontari in 2 gruppi di 500: quello di *trattamento*, a cui viene somministrato i farmaco e quello di *controllo*, a cui viene dato un placebo. Sia ai volontari, sia a coloro che somministrano il farmaco non viene rivelata la composizione dei gruppi, per evitare reazioni emotive.
È chiaramente di fondamentale importanza il modo in cui vengono formati i due gruppi. Si desidera infatti che essi siano più simili possibile in tutti gli aspetti tranne la composizione della sostanza somministrata: in questo modo si può senz'altro concludere che ogni differenza significativa nella risposta dei due gruppi sia realmente dovuta al farmaco. Vi è accordo in generale sul fatto che il miglior modo per ottenere questo risultato sia quello di scegliere i 500 volontari di un gruppo in maniera completamente casuale, ovvero la

scelta dovrebbe essere fatta in modo che ciascuno dei $\binom{1000}{500}$ sottoinsiemi di 500 volontari abbia la stessa probabilità di essere scelto come gruppo di trattamento. Come si può realizzare questo esperimento casuale?

Esempio 1 (Scelta di un sottoinsieme casuale): Consideriamo un insieme di n elementi, numerati con gli interi $1,2,\dots,n$. Si vuole scegliere a caso uno dei suoi $\binom{n}{k}$ sottoinsiemi di cardinalità k, in modo che abbiano tutti la medesima probabilità di essere selezionati. Per risolvere questo problema a prima vista complesso, partiamo dalla fine, e supponiamo di avere effettivamente generato nel modo richiesto uno dei sottoinsiemi di k elementi. Per $j = 1,2,\dots,n$, poniamo $I_j \coloneqq \begin{cases}1 \text{ se l'elemento } j\text{-esimo è nel sottoinsieme}\\ 0 \text{ altrimenti}\end{cases}$
e calcoliamo la distribuzione condizionata di I_j dati $I_1, I_2, \dots, I_{j-1}$.
Per prima cosa notiamo ce la probabilità che l'elemento 1 stia nel sottoinsieme è $\frac{k}{n}$ (lo si può vedere (**1**) o perché vi è una probabilità di $\frac{1}{n}$ che l'elemento 1 sia il j-esimo elemento estratto, per $j = 1,2,\dots,k$; (**2**) o perché la frazione di esiti della selezione casuale che contengono l'elemento 1 è data da $\binom{1}{1}\binom{n-1}{k-1}\Big/\binom{n}{k} = \frac{k}{n}$.
Per questo abbiamo che $\mathcal{P}(I_1 = 1) = \frac{k}{n}$. Calcoliamo adesso la probabilità che l'elemento 2 appartenga al sottoinsieme, condizionata ad $I_1 = 1$, a parte il primo, i restanti $k-1$ elementi del sottoinsieme vengono scelti a caso tra gli $n-1$ elementi disponibili dell'insieme di partenza. Perciò in analogia con quanto già detto per l'elemento 1, otteniamo che $\mathcal{P}(I_2 = 1|I_1 = 1) = \frac{k-1}{n-1}$.
Similmente, se $I_1 = 0$, allora il primo elemento non appartiene al sottoinsieme, e i k elementi di quest'ultimo vengono scelti a caso tra gli altri $n-1$ elementi, così che $\mathcal{P}(I_2 = 1|I_1 = 0) = \frac{k}{n-1}$.
Mettendo assieme le ultime due equazioni, si può dire che $\mathcal{P}(I_2 = 1|I_1) = \frac{k-I_1}{n-1}$ e, generalizzando questo procedimento, si arriva a scoprire che

$$\mathcal{P}\left(I_{j+1}\middle|I_1, I_2, \dots, I_j\right) = \frac{k - \sum_{i=1}^{j} I_i}{n-j}, \qquad j = 1, \dots, n$$

infatti $\sum_{i=1}^{j} I_i$ rappresenta il numero di elementi tra i primi j che appartengono al sottoinsieme, così che condizionando ai valori di $+I_1, I_2, \dots, I_j$, restano $k - \sum_{i=1}^{j} I_i$ elementi del sottoinsieme che devono essere scelti tra gli $n-j$ che rimangono dell'insieme di partenza.

Riconsiderando il problema dall'inizio. Se U è una variabile aleatoria uniforme su $[0,1]$ e $0 \le a \le 1$, allora $\mathcal{P}(U < a) = a$. Si possono perciò utilizzare le equazioni $\mathcal{P}(I_1 = 1) = \frac{k}{n}$ e $\mathcal{P}\left(I_{j+1}\middle|I_1, I_2, \dots, I_j\right) = \frac{k-\sum_{i=1}^{j} I_i}{n-j}$ per costruire un sottoinsieme casuale con le caratteristiche richieste: si genera una successione $U_1, U_2, \dots$ di (al più n) variabili aleatorie uniformi su $[0,1]$ e indipendenti, e quindi si pone

$$I_1 \coloneqq \begin{cases}1 \text{ se } U_1 < \frac{k}{n}\\ 0 \text{ altrimenti}\end{cases} \quad \text{e, per } j = 1,2,\dots \quad I_{j+1} \coloneqq I_1 \coloneqq \begin{cases}1 \text{ se } U_{j+1} < \frac{k - I_1 - \dots - I_j}{n-j}\\ 0 \text{ altrimenti}\end{cases}$$

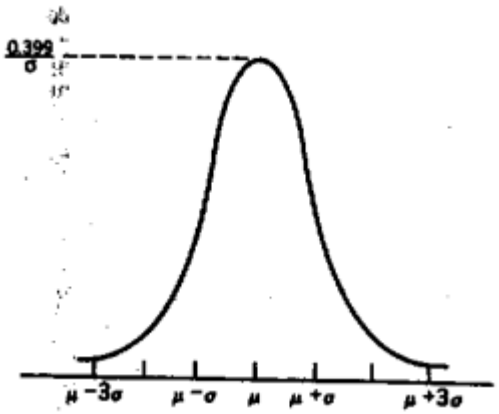
Il procedimento termina non appena $I_1 + \dots + I_j = k$, e a quel punto il sottoinsieme casuale consiste dei k elementi le cui corrispondenti funzioni indicatrici I sono pari a 1. In formule $S = \{i : I_i = 1\}$

Variabili aleatorie normali o gaussiane

Definizione 1: Una variabile aleatoria X si dice *normale* oppure *gaussiana* di parametri μ e σ^2, e si scrive $X \sim \mathcal{N}(\mu, \sigma^2)$, se X ha funzione di densità data da

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}, \qquad \forall \boldsymbol{x} \in \mathbb{R}$$

La densità normale è una curva a campana simmetrica rispetto all'asse $x = \mu$, dove ha il massimo pari a $\left(\sigma\sqrt{2\pi}\right)^{-1} \approx \frac{0.399}{\sigma}$ (si veda figura di fianco).



La distribuzione normale venne introdotta da un matematico francese che la utilizzò per approssimare le probabilità associate a variabili aleatorie binomiali quando il parametro n è grande. Il suo risultato fu poi esteso da Laplace e altri, fino ad essere incluso nel teorema del limite centrale. Quest'ultimo fornisce la giustificazione teorica di un fatto evidente dall'esperienza empirica, ovvero che molti fenomeni casuali seguono una legge approssimativamente normale.
Alcuni esempi di tale comportamento sono la statura delle persone, la velocità in ciascuna direzione di una molecola di gas, gli errori di misurazione delle grandezze fisiche.

La funzione generatrice dei momenti di una variabile aleatoria gaussiana di parametri μ e σ^2 si deduce come segue:

$$\boldsymbol{\phi(t)} \coloneqq E[e^{tX}] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{t(\sigma y+\mu)} e^{-\frac{y^2}{2}} dx =$$

ponendo $y = \frac{x-\mu}{\sigma}$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{\frac{2\sigma t y - y^2}{2}} dx = \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} e^{\frac{\sigma^2 t^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{y^2-2\sigma t y+\sigma^2 t^2}{2}} dx = e^{\mu t+\frac{\sigma^2 t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\sigma t)^2}{2}} dx$$

$$= \boldsymbol{e^{\mu t+\frac{\sigma^2 t^2}{2}}}$$

dove l'ultima uguaglianza segue perché l'espressione dentro l'integrale rappresenta la densità di probabilità di una variabile aleatoria normale di parametri σt e 1, e come tale il suo integrale su tutto $\mathbb{R}$ è pari a1.

Derivando l'espressione della funzione generatrice data da $\phi(t) = e^{\mu t+\frac{\sigma^2 t^2}{2}}$ si ottiene

$$\phi'(t) = (\mu + \sigma^2 t) e^{\mu t+\frac{\sigma^2 t^2}{2}}$$
$$\phi''(t) = [\sigma^2 + (\mu + \sigma^2 t)^2] e^{\mu t+\frac{\sigma^2 t^2}{2}}$$

Da cui ricaviamo i primi due momenti e la varianza di una variabile aleatoria gaussiana:

$$E[X] = \phi'(0) = \mu$$
$$E[X^2] = \phi''(0) = \sigma^2 + \mu^2$$
$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \sigma^2$$

Così che i parametri μ e σ^2 rappresentano rispettivamente la media e la varianza della distribuzione normale.

Un risultato importante riguardo questo tipo di variabili aleatorie è che se X è gaussiana e Y è una trasformazione lineare di X, allora Y è a sua volta gaussiana. L'enunciato seguente precisa quanto detto.

Proposizione 1: Sia $X \sim \mathcal{N}(\mu, \sigma^2)$, e sia $Y = \alpha X + \beta$, dove α e β sono due costanti reali e $\alpha \neq 0$. Allora Y è una variabile aleatoria normale con media $\sigma\mu + \beta$ e varainza $\alpha^2\sigma^2$.
Dimostrazione: Calcoliamo la funzione generatrice di Y:

$$E\left[e^{t(\alpha X+\beta)}\right] = e^{\beta t} E[e^{\alpha t X}] = e^{\beta t}\phi_X(\alpha t) = e^{\beta t} e^{\mu\alpha t+\frac{\sigma^2\alpha^2 t^2}{2}} \text{ per la } \phi(t) = e^{\mu t+\frac{\sigma^2 t^2}{2}}$$
$$= e^{(\alpha\mu+\beta)t+\frac{(\alpha^2\sigma^2)t^2}{2}}$$

L'equazione $\phi(t) = e^{\mu t+\frac{\sigma^2 t^2}{2}}$ afferma che l'espressione ottenuta è la funzione generatrice di una variabile

aleatoria gaussiana di media $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$. Siccome la funzione generatrice di Y ne determina la distribuzione, quanto detto dimostra l'enunciato.

Un corollario della precedente proposizione è che se $X \sim \mathcal{N}(\mu, \sigma^2)$, allora $Z = \frac{X-\mu}{\sigma}$ è una variabile aleatoria normale con media 0 e varianza 1. Una tale variabile aleatoria si dice *normale standard*; la sua funzione di ripartizione riveste un ruolo importante in statistica ed è normalmente indicata con il simbolo Φ:

$$\boldsymbol{\Phi(x) \coloneqq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy}, \qquad \forall \boldsymbol{x} \in \mathbb{R}$$

Il fatto che $Z = \frac{(X-\mu)}{\sigma}$ abbia distribuzione normale standard quando X è gaussiana di media μ e varianza σ^2 ci permette di esprimere le probabilità relative a X in termini di probabilità su Z. Ad esempio per trovare $\mathcal{P}(X < b)$, notiamo che $X < b$ se e solo se $\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}$ così che

$$\mathcal{P}(X < b) = \mathcal{P}\left(\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = \mathcal{P}\left(Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Analogamente, per ogni $a < b$, si ha che

$$\mathcal{P}(a < X < b) = \mathcal{P}\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = \mathcal{P}\left(Z < \frac{b-\mu}{\sigma}\right) - \mathcal{P}\left(Z < \frac{a-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

In entrambi i casi ci siamo ricondotti a determinare un valore di $\Phi(x)$ (c'è una tabella per le approssimazioni poiché il risultato non si può calcolare analiticamente).

Variabili aleatorie esponenziali

Definizione 1: Una variabile aleatoria continua la cui funzione di densità di probabilità è data da

$$\boldsymbol{f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{se } x \leq 0 \end{cases}}$$

per un opportuno valore della costante $\lambda > 0$, si dice *esponenziale* con parametro (o *intensità*) λ.

La funzione di ripartizione di una tale variabile aleatoria è data da

$$F(x) = \mathcal{P}(X \leq x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}, \qquad x \geq 0$$

Nella pratica, la distribuzione esponenziale può rappresentare il tempo di attesa prima che si verifichi un certo evento casuale. Ad esempio il tempo che trascorrerà (a partire da questo momento) fino al verificarsi di un terremoto, o allo scoppiare id un nuovo conflitto, o al giungere della prossima telefonata di qualcuno che ha sbagliato numero, sono tutte variabili aleatorie che in pratica tendono ad avere distribuzioni esponenziali.

La funzione generatrice dei momenti di una variabile aleatoria esponenziale di intensità λ è data da

$$\phi(t) \coloneqq E[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda - t)x} dx = \frac{\lambda}{\lambda - t}, \qquad t < \lambda$$

Derivando si trova che $\phi'(t) = \frac{\lambda}{(\lambda-t)^2}$, $\phi''(t) = \frac{2\lambda}{(\lambda-t)^3}$ e da cui è facile ottenere i momenti e la varianza.

$$E[X] = \phi'(0) = \frac{1}{\lambda}$$
$$E[X^2] = \phi''(0) = \frac{2}{\lambda^2}$$
$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{1}{\lambda^2}$$

Per una variabile aleatoria esponenziale, λ è il reciproco del valore atteso, e la varianza è il quadrato di quest'ultimo.

La proprietà centrale della distribuzione esponenziale è la sua *assenza di memoria*. Con questa espressione, riferita ad una variabile aleatoria positiva X si intende che $\boldsymbol{\mathcal{P}(X > s + t|X > t) = \mathcal{P}(X > s)}$, $\forall s, t \geq 0$

Per capire perché la precedente equazione è detta proprietà di assenza di memoria, si immagini che X rappresenti il tempo di vita di un certo oggetto prima di guastarsi. Sapendo che tale oggetto è già in funzione da un tempo t e non si è ancora rotto, qual è la probabilità che esso continui a funzionare almeno per un ulteriore intervallo di tempo s? Chiaramente la probabilità richiesta è quella espressa dal membro sinistro dell'equazione, ovvero $\mathcal{P}(X > s + t|X > t)$. Infatti dire che l'oggetto non si è ancora guastato al tempo t equivale a dire che il tempo in cui avverrà la rottura (X), è superiore a t, mentre affermare che l'oggetto funzionerà per un ulteriore tempo s a partire dal tempo t, significa che il tempo X dovrà essere maggiore di $t + s$. In questo senso, l'equazione afferma che la distribuzione del tempo di vita rimanente dell'oggetto considerato, è la medesima sia nel caso in cui esso stia funzionando da un tempo t, sia nel caso in cui esso sia nuovo, o, in altri termini, se $\mathcal{P}(X > s + t|X > t) = \mathcal{P}(X > s)$ è soddisfatta, non vi è alcun bisogno di tenere presente l'età dell'oggetto, perché fino a che esso funzione, si comporta esattamente come se fosse "nuovo di zecca".

La condizione di assenza di memoria è equivalente a chiedere che

$$\frac{\mathcal{P}(X > s + t, X > t)}{\mathcal{P}(X > t)} = \mathcal{P}(X > s) \Rightarrow \mathcal{P}(X > s + t) = \mathcal{P}(X > s)\mathcal{P}(X > t)$$

Quest'ultima formulazione è facilmente verificabile se X è esponenziale, visto che, per $x > 0, \mathcal{P}(X > x) = e^{-\lambda x}$ e ovviamente, $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$. Abbiamo quindi provato che le variabili aleatorie esponenziali sono prive di memoria (in realtà esse sono le *uniche* ad avere questa proprietà).

Proposizione 1: Se $X_1, X_2, \dots, X_n$ sono variabili aleatorie esponenziali e indipendenti, di parametri $\lambda_1, \lambda_2, \dots, \lambda_n$ rispettivamente, allora la variabile aleatoria $Y \coloneqq \min(X_1, X_2, \dots, X_n)$ è esponenziale di parametro $\sum_{i=1}^{n} \lambda_i$
Dimostrazione: Basta dimostrare che $\mathcal{P}(Y \leq x) = 1 - e^{-x\sum_{i=1}^{n}\lambda_i}$, ovvero che $\mathcal{P}(T > x) = e^{-x\sum_{i=1}^{n}\lambda_i}$.
Siccome il minore di un insieme di numeri è più grande di x se e solo se ciascuno dei numeri in questione è maggiore di x, abbiamo che $\mathcal{P}(Y > x) = \mathcal{P}(\min(X_1, X_2, \dots, X_n) > x) = \mathcal{P}(X_1 > x, X_2 > x, \dots, X_n > x)$ per l'indipendenza degli X_i si ha

$$\mathcal{P}(X_1 > x, X_2 > x, \dots, X_n > x) = \prod_{i=1}^{n} \mathcal{P}(X_i > x) = \prod_{i=1}^{n} \left(1 - F_{X_i}(x)\right) = \prod_{i=1}^{n} e^{-\lambda_i x} = e^{-x\prod_{i=1}^{n}\lambda_i}$$

Le distribuzioni chi-quadro

Definizione 1: Siano $X_1, X_2, \dots, X_n$ n variabili aleatorie normali standard e indipendenti, allora la somma dei loro quadrati, $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ è una variabile che prende il nome di *chi-quadro a n gradi di liberò*. La notazione che useremo per indicare questo fatto è la seguente: $\mathcal{X}^2(n)$.

Calcolo della speranza di $\mathcal{X}^2(n)$:
Per come abbiamo definito sopra la Y abbiamo $E[Y] = E[X_1^2 + X_2^2 + \cdots + X_n^2] = E[X_1^2] + E[X_2^2] + \cdots + E[X_n^2]$ ricordando che la varianza di una variabile aleatoria standard coincide con la media del quadrato della variabile, otteniamo $E[Y] = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}[X_n] = n$

4. La distribuzione delle statistiche campionarie

Introduzione

La statistica è la scienza che si occupa di trarre conclusioni dai dati sperimentali. Una situazione tipica con la quale bisogna spesso confrontarsi negli ambienti tecnologici, è quella in cui si studia un insieme molto grande, detto *popolazione*, di oggetti a cui sono associate delle quantità misurabili. L'approccio statistico consiste nel selezionare un sottoinsieme ridotto di oggetti, che viene detto *campione*, e analizzarlo sperando di essere in grado di trarre da esso delle conclusioni valide per la popolazione nel suo insieme.

Per basare sui dati del campione delle inferenze che riguardino l'intera popolazione, è necessario assumere qualche condizione sulle relazioni che legano questi due insiemi. Un'ipotesi fondamentale, in molti casi del tutto ragionevole, è che vi sia una (implicita) distribuzione di probabilità della popolazione, nel senso che da essa si estraggono degli oggetti in maniera causale, le quantità numeriche loro associate possono essere pensate come variabili aleatorie indipendenti, tutte con tale distribuzione. Se tutto il campione viene selezionato in maniera causale, sembra ragionevole supporre che i suoi dati siano valori indipendenti provenienti da tale distribuzione.

Definizione 1: Un insieme $X_1, X_2, \dots, X_n$ di variabili aleatorie indipendenti, tutte con la stessa distribuzione F, si dice *campione* o *campione aleatorio* della distribuzione F.

In pratica la distribuzione F non è mai completamente nota, però è possibile usare i dati per fare dell'*inferenza* su F. In alcuni casi è possibile che F sia nota eccetto che per dei parametri incogniti (si potrebbe ad esempio sapere che F è una distribuzione normale, ma non conoscerne la media e la varianza; oppure F potrebbe essere di Poisson, ma con parametro incognito); in altri casi potremmo non sapere praticamente nulla di F (tranne forse assumere che essa sia continua, oppure discreta). I problemi in cui la distribuzione F è nota a meno di un insieme di parametri incogniti sono detti problemi di inferenza *parametrica*; quelli in cui nulla si sa sulla distribuzione F sono invece problemi di inferenza *non parametrica*.

In questo capitolo ci occupiamo delle distribuzioni di probabilità di alcune statistiche. Il termine *statistica* indica una *variabile aleatoria* che è semplicemente una funzione dei dati di un campione; i due principali esempi di statistiche che affrontiamo, sono la media campionaria e la varianza campionaria.

La media campionaria

Consideriamo una popolazione di elementi, a ciascuno dei quali è associata una grandezza numerica. La popolazione potrebbe ad esempio essere costituita dagli individui adulti facenti parte di una qualche categoria di persone, e la grandezza numerica di interesse potrebbe essere il reddito annuale, la statura, l'età o altro. Sia $X_1, X_2, \dots, X_n$ un campione di dati estratto da questa popolazione. È comune supporre che i valori numerici associati a ciascuno degli elementi del campione, siano variabili aleatorie indipendenti e identicamente distribuite. Denotiamo con μ e σ^2 la loto media e la loro varianza, che prendono il nome di *media* e *varianza della popolazione*. Definiamo la *media campionaria* come

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

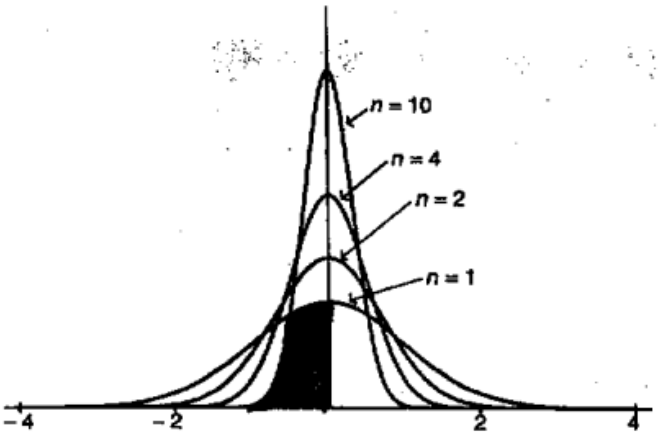
Si noti che $\bar{X}$ è una funzione delle variabili aleatorie $X_1, X_2, \dots, X_n$. In quanto tale è una *statistica*, e in particolare è a sua volta una variabile aleatoria. Ha senso quindi domandarsi quanto valgano il valore atteso della media campionaria e la sua varianza. È facile vedere che

$$\boldsymbol{E}[\bar{\boldsymbol{X}}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} = \frac{n\mu}{n} = \boldsymbol{\mu}$$

e, per la varianza (si ricordi che vale l'indipendenza):

$$\mathbf{Var}(\bar{\boldsymbol{X}}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

La media campionaria ha quindi lo stesso valore atteso della distribuzione da stimare, mentre la sua varianza risulta ridotta di un fattore n. Da questo possiamo dedurre che $\bar{X}$ è centrata attorno a μ, e la sua variabilità si riduce sempre di più con l'aumentare di n.
Una esemplificazione di questo comportamento è illustrata nella figura a destra, che riporta, per diversi valori di n, le densità di probabilità per le medie campionarie di una popolazione normale standard.



Il teorema del limite centrale

In questa sezione affrontiamo uno dei risultati più notevoli della teoria della probabilità, il *teorema del limite centrale*. In termini semplicistici, esso afferma che la somma di un numero elevato di variabili aleatorie indipendenti, tende ad avere distribuzione approssimativamente normale. L'importanza è duplice: da un lato siamo in grado di ottenere stime approssimative delle probabilità che riguardano la somma di variabili aleatorie indipendenti, dall'altro abbiamo giustificato il fatto notevole che la distribuzione empirica delle frequenza di un gran numero di popolazioni naturali esibisca forme a campana (in realtà, gaussiane).

Teorema 1 (Teorema del limite centrale): Siano $X_1, X_2, \dots, X_n$ delle variabili aleatorie indipendenti e identicamente distribuite, tutte con media μ e varianza σ^2. Allora se n è grande, la somma $X_1 + X_2 + \cdots + X_n$ è approssimativamente normale con media $n\mu$ e varianza $n\sigma^2$.

Si può anche normalizzare la somma precedente in modo da ottenere una distribuzione approssimativamente normale *standard*. Si ha infatti che

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \dot{\sim} \mathcal{N}(0,1)$$

dove con il simbolo $\dot{\sim}$ si intende "è approssimativamente distribuito come". Ciò significa che per n grande e x qualsiasi vale l'approssimazione

$$\mathcal{P}\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right) \approx \Phi(x)$$

dove Φ denota la funzione di ripartizione della normale standard introdotta nel capitolo delle variabili aleatorie normali o gaussiane.

Una delle più dirette applicazioni del teorema del limite centrale riguarda le variabili aleatorie binomiali. Siccome una binomiale X di parametri (n, p) rappresenta il numero di successi in n prove indipendenti, ciascuna con probabilità p di riuscita, possiamo scrivere $X = X_1 + X_2 + \cdots + X_n$ dove

$$X_i \coloneqq \begin{cases} 1 \text{ se l'}i\text{-esima prova ha successo} \\ 0 \text{ altrimenti} \end{cases}$$

Poiché, come sappiamo, $E[X_i] = p$ e $\mathrm{Var}(X_i) = p(1-p)$, segue dal teorema del limite centrale che, per n grande, $\frac{X-np}{\sqrt{np(1-p)}} \dot{\sim} \mathcal{N}(0,1)$; ovvero, vale una *approssimazione normale* delle variabili aleatorie binomiali.
In altri termini, la funzione di massa di una variabile aleatoria binomiale di parametri (n, p) tende a divenire gaussiana al crescere di n.

È bene notare che a questo punto disponiamo di due diverse approssimazioni per le variabili aleatorie binomiali: quella di Poisson, che è valida quando n è grande e p piccolo, e quella normale, che è valida quando $np(1-p)$ è grande (in effetti, per ottenere risultati accettabili, basta che $np(1-p)$ sia almeno 10).

Distribuzione approssimata della media campionaria

Sia $X_1, X_2, \dots, X_n$ un campione proveniente da una popolazione di media μ e varianza σ^2. Vediamo come il teorema del limite centrale ci permette di approssimare la distribuzione della media campionaria,

$$\bar{X} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i$$

Siccome il prodotto di una variabile aleatoria normale per una costante è ancora normale, ne segue che, quando n è grande, $\bar{X}$ è approssimativamente gaussiana. Poiché inoltre la media campionaria ha valore atteso μ e deviazione standard $\frac{\sigma}{\sqrt{n}}$, otteniamo che

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \dot{\sim} \mathcal{N}(0,1)$$

Quando un campione è abbastanza numeroso?

Il teorema del limite centrale lascia aperta la questione di quanto grande debba essere la numerosità del campione n, affinché l'approssimazione normale sia valida. In effetti la risposta dipende dalla distribuzione da cui vengono campionati i dati. Ad esempio, se la distribuzione della popolazione è normale, allora $\bar{X}$ sarà a sua volta normale indipendentemente dall'ampiezza del campione (questo perché la distribuzione normale è riproducibile). Una buona regola empirica è che si può essere confidenti nella validità dell'approssimazione se n è almeno 30. Questo vuol dire che, per quanto "poco gaussiana" sia la distribuzione considerata, la media campionaria di un gruppo di dati di numerosità 30 risulta comunque approssimativamente normale. Si tenga presente, comunque, che in molti casi è possibile che questo accada anche per n molto più piccolo, e in effetti spesso $n = 5$ è sufficiente ad ottenere approssimazioni non troppo sbagliate.

La varianza campionaria

Sia $X_1, X_2, \dots, X_n$ un campione aleatorio, proveniente da una distribuzione di media μ e varianza σ^2. Sia $\bar{X}$ la sua media campionaria.

Definizione 1: La statistica S^2, definita da

$$\boldsymbol{S^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

si dice *varianza campionaria*. La sua radice quadrata $S = \sqrt{S^2}$ prende invece il nome di *deviazione standard campionaria*.

Volendo calcolare $E[S^2]$, sfruttiamo il fatto che per una qualsiasi n-upla di numeri $x_1, x_2, \dots, x_n$,

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

dove $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Applicando a $X_1, X_2, \dots, X_n$ questo enunciato implica che

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \Rightarrow (n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Prendendo il valore atteso di entrambi i membri di quest'ultima equazione, e ricordando che il momento secondo di una qualunque variabile aleatoria W si può ottenere come $E[W^2] = \mathrm{Var}(W) + E[W]^2$, deduciamo che

$$(\boldsymbol{n-1})\boldsymbol{E}[\boldsymbol{S^2}] = E\left[\sum_{i=1}^{n} X_i^2\right] - E[n\bar{X}] = nE[X_1^2] - nE[\bar{X}^2] = n\mathrm{Var}(X_1) + nE[X_1]^2 - n\mathrm{Var}(\bar{X}) - nE[\bar{X}]^2 =$$

$$= n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2 = (\boldsymbol{n-1})\boldsymbol{\sigma^2} \Rightarrow E[S^2] = \sigma^2$$

Il valore atteso della varianza campionaria coincide con la varianza della popolazione.

La distribuzione della media campionaria

Siccome la somma di variabili aleatorie normali e indipendenti ha ancora distribuzione gaussiana, anche $\bar{X}$ è normale. La sua media e la sua varianza, come nel caso generale, sono μ e $\frac{\sigma^2}{n}$ rispettivamente, e quindi

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

è una variabile aleatoria normale standard.

5. Stima parametrica

Introduzione

Consideriamo un campione aleatorio $X_1, X_2, \dots, X_n$ estratto da una distribuzione F_θ che dipende da un vettore di parametri incogniti θ. Potrebbe ad esempio trattarsi di variabili aleatorie di Poisson, delle quali ignoriamo il valore di λ; oppure potremmo avere a che fare con un campione normale, della cui distribuzione ignoriamo media e varianza. Mentre quando si fa della probabilità è normale supporre che le distribuzioni in gioco siano completamente note, in statistica è vero il contrario, e il problema centrale è quello di dire qualcosa (ovvero *fare dell'inferenza*) sui parametri sconosciuti, usando i dati osservati.

Stimatori di massima verosimiglianza

Una qualunque statistica il cui scopo sia quello di dare una stima di un parametro θ si dice *stimatore* di θ; gli stimatori sono quindi variabili aleatorie. Il valore deterministico assunto da uno stimatore è detto invece *stima*. Ad esempio, come avremo modo di vedere, la media campionaria $\bar{X} \coloneqq \sum_i \frac{X_i}{n}$ di un campione normale $X_1, X_2, \dots, X_n$ costituisce lo stimatore abituale della media μ della distribuzione.

Consideriamo delle variabili aleatorie $X_1, X_2, .., X_n$, la cui distribuzione congiunta sia nota a meno di un parametro incognito θ. Un problema di interesse consiste nello stimare θ usando i valori che vengono assunti da queste variabili aleatorie. Per esemplificare, potremmo immaginare che le X_i siano variabili aleatorie esponenziali e indipendenti, tutte di media θ incognita. In questo caso la loro densità congiunta sarebbe data da ($x_i > 0, i = 1, \dots, n$):

$$f(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_n}(x_n) = \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdot \dots \cdot \frac{1}{\theta} e^{-\frac{x_n}{\theta}} = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^n x_i}$$

e il nostro obiettivo consisterebbe nello stimare θ partendo dai valori osservati $X_1, X_2, \dots, X_n$.

Vi è una classe particolare di stimatori, detti *stimatori di massima verosimiglianza*, che è largamente utilizzata in statistica. Uno stimatore di questo tipo si ottiene con il ragionamento seguente. Denotiamo con $f(x_1, x_2, \dots, x_n|\theta)$ la funzione di massa congiunta di $X_1, X_2, \dots, X_n$ oppure la loro densità congiunta, a seconda che siano variabili aleatorie discrete o continue. Poiché stiamo supponendo che θ sia un incognita, mostriamo esplicitamente che f dipende da θ. Se interpretiamo $f(x_1, x_2, \dots, x_n|\theta)$ come la verosimiglianza (o plausibilità, o credibilità, in un italiano più diretto) che si realizzi la n-upla di dati $x_1, x_2, \dots, x_n$ quando θ è il vero valore assunto dal parametro, sembra ragionevole adottare come stima di θ quel valore che rende massima la verosimiglianza per i dati osservati. In altri termini, la stima di massima verosimiglianza $\hat{\theta}$ è definita come il valore di θ che rende massima $f(x_1, x_2, \dots, x_n|\theta)$, quando i valori osservati sono $x_1, x_2, \dots, x_n$. La funzione $f(x_1, x_2, \dots, x_n|\theta)$ è detta funzione di *likelihood* (verosimiglianza in inglese).

Nel calcolare il valore di θ che massimizza f, conviene spesso usare il fatto che le due funzioni $f(x_1, x_2, \dots, x_n|\theta)$ e $\log[f(x_1, x_2, \dots, x_n|\theta)]$ assumono il massimo in corrispondenza dello stesso valore di θ. Quindi è possibile ottenere $\hat{\theta}$ anche massimizzando $\log[f(x_1, x_2, \dots, x_n|\theta)]$, che è detta funzione di log-*likelihood*.

Confronto stimatori

Fissato n in $\mathbb{N}$, si ricordi che

a) Un campione casuale semplice $\bar{X} = (X_1, X_2, \dots, X_n)$ è una n-upla di numeri aleatori congiuntamente distribuiti, indipendenti e somiglianti; le componenti di $\bar{X}$ sono chiamate *osservazioni*
b) Una variabile aleatoria X avente la stessa legge (distribuzione) delle osservazioni è detta essere *genitrice* del campione casuale semplice $\bar{X}$;
c) In altri termini le osservazioni sono repliche indipendenti della genitrice;
d) Nella legge di X, in generale, compare un *parametro* incognito θ di cui è noto l'insieme Θ dei suoi valori; Θ è la *regione parametrica*; nulla vieta che θ abbia dimensione maggiore di 1;

e) Una *statistica* T è una funzione misurabile e calcolabile di un campione casuale semplice: $T = g(\bar{X})$; nulla vieta che T abbia dimensione maggiore di 1;
f) Uno stimatore è una statistica che viene utilizzata ai fini della stima di una funzione $\psi(\theta)$ del parametro incognito θ.

Quando si vuole mettere in evidenza la dipendenza dal parametro incognito della media della genitrice si può usare la notazione $E_X(\theta)$: in altri termini, la media deve essere vista non come un numero ma come una funzione definita nella regione parametrica. Allo stesso modo per la varianza, $D_X^2(\theta)$ e per gli altri momenti teorici. Ovviamente, la stessa notazione e la stessa interpretazione si può usare per la media e la varianza di una qualsiasi statistica.

Si ricordi, infine, che uno stimatore T è corretto ai fini della stima di $\psi(\theta)$ se e solo se: $\theta \in \Theta$ e $E_T(\theta) = \psi(\theta)$. Nel caso in cui lo stimatore T non è corretto, la funzione $\theta \in \Theta, d_T(\theta) \coloneqq E_T(\theta) - \psi(\Theta)$ è la sua *distorsione*.

Esempio 1: Sia $X \sim B(1,p)$ con $p \in (0,1)$ e $n = 3$. La media campionaria $\bar{X}$ è uno stimatore corretto per p. Infatti, si ha: $E_{\bar{X}}(p) = \frac{1}{3}E_{X_1+X_2+X_2}(p) = \frac{1}{3}\left[E_{X_1}(p) + E_{X_2}(p) + E_{X_3}(p)\right] = \frac{1}{3}(p+p+p) = \frac{1}{3}\cdot 3p = p$.
Invece, lo stimatore $V = \frac{X_1+2X_2+X_3}{5}$ non è corretto; infatti, risulta:

$$E_V(p) = \frac{1}{5}E_{X_1+2X_2+X_3}(p) = \frac{1}{5}\left[E_{X_1}(p) + 2E_{X_2}(p) + E_{X_3}(p)\right] = \frac{1}{5}(p+2p+p) = \frac{1}{5}\cdot 4p = \frac{4}{5}p \neq p$$

Quindi, la distorsione dello stimatore V vale: $p \in (0,1),\ d_V(p) = \frac{4}{5}p - p = -\frac{1}{5}p$.

Ai fini del confronto di due stimatori T e S per la stessa funzione $\psi(\theta)$ si considera il *rischio quadratico medio* che per lo stimatore T è la seguente funzione: $\theta \in \Theta,\ R_T(\theta) = E\{[T - \psi(\theta)]^2\}$. Posto $\mu_T(\theta) \coloneqq E_T(\theta)$ e ricordando che è nulla la media di una variabile aleatoria scartata rispetto alla sua media, risulta:

$$\begin{aligned} R_T(\theta) = E\{[T-\psi(\theta)]^2\} &= E\{\{[T-\mu_T(\theta)] + [\mu_T(\theta) - \psi(\theta)]\}^2\} = \\ &= E\{[T-\mu_T(\theta)]^2\} + 2[\mu_T(\theta) - \psi(\theta)]E[T-\mu_T(\theta)] + [\mu_T(\theta) - \psi(\theta)]^2 = \\ &= D_T^2(\theta) + d_T^2(\theta) \end{aligned}$$

In altri termini, il rischio quadratico medio è la somma della varianza dello stimatore con il quadrato della sua distorsione. Ne discende che per gli stimatori corretti il rischio coincide con la varianza.

Esempio 2: In riferimento all'Esempio 1, si ha:

$$\theta \in \Theta, \qquad R_{\bar{X}}(p) = D_{\bar{X}}(p) = \frac{1}{9}\left[D_{X_1}(p) + D_{X_2}(p) + D_{X_3}(p)\right] = \frac{1}{9}\cdot 3p(1-p) = \frac{p(1-p)}{3}$$

Inoltre, procedendo allo stesso modo, si ha:

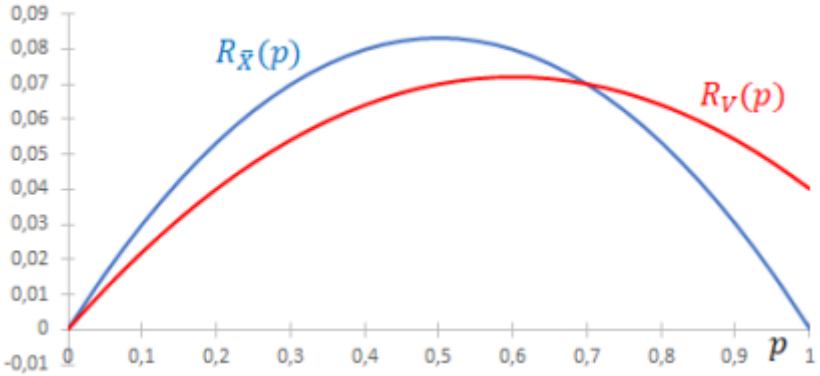
$$D_V(p) = \frac{1}{25}\left[D_{X_1}(p) + D_{2X_2}(p) + D_{X_3}(p)\right] = \frac{1}{25}\cdot 6p(1-p) = \frac{6p(1-p)}{25}$$

Tenendo conto di quest'ultima e del fatto che $d_v(p) = -\frac{1}{5}p$, in definitiva si ha:

$$R_v(p) = D_V(p) + d_V^2(p) = \frac{6p(1-p)}{25} + \frac{1}{25}p^2$$

Definizione 1: Siano T e S due stimatori per $\psi(\theta)$. Si dice che S è preferibile a T e si scrive $S \leq T$ se e solo se, in tutta la regione parametrica, il rischio quadratico medio di S è non maggiore del rischio quadratico medio di T. In simboli: $S < T \Leftrightarrow S \leq T$ e $\exists \theta_0 \in \Theta : R_S(\theta_0) < R_T(\theta_0)$

Esempio 3: In relazione all'Esempio 2 si presenta il grafico del rischio quadratico medio in funzione di p degli stimatori $\bar{X}$ e V.



Nell'intervallo da 0 a $0{,}7$ il rischio quadratico medio dello stimatore V è non maggiore di quello dello stimatore $\bar{X}$ ma nell'intervallo da $0{,}7$ a 1 è il rischio dello stimatore $\bar{X}$ a essere non maggiore di quello di V; se ne evince che i due stimatori considerati non sono confrontabili rispetto al rischio quadratico medio.

Si consiglia come esercizio di verificare che, con riferimento all'Esempio 1 e ai fini della stima di θ, lo stimatore $\bar{X}$ è strettamente preferibile allo stimatore corretto X_1.

Definizione 3: Si dice che uno stimatore T è ammissibile ai fini della stima di $\psi(\theta)$ se non esiste un altro stimatore strettamente preferibile a T.

Stimatori dei momenti

Si ricordi che, in generale la legge di probabilità di un numero aleatorio X è caratterizzata da uno o più parametri. Quindi ogni momento teorico di X, se esistente, è funzione dei parametri. Ad esempio, nel caso $X \sim B(n,p)$ risulta $\mu_1' = E(X) = np$ e $\mu_2 = D^2(X) = np(1-p)$. Nel caso in cui il numero aleatorio X sia dotato di funzione generatrice dei momenti allora si può affermare che esistono finiti sia il momento che il momento centrale di qualsiasi ordine di X.

Il metodo dei momenti

Siano n, k e $p \geq k$ numeri interi positivi. Si supponga che la genitrice X di un campione casuale semplice $\bar{X} = (X_1, X_2, \dots, X_n)$ abbia la legge di probabilità caratterizzata da k parametri incogniti $\theta_1, \theta_2, \dots, \theta_k$. Allora, per quanto detto in premessa, scelti opportunamente p indici interi positivi, $j_1, j_2, \dots, j_p$ risulta:

$$r = 1,2,\dots,p, \qquad \mu'_{j_r} = f_j(\theta_1, \theta_2, \dots, \theta_k)$$

La precedente equazione rappresenta un sistema di p equazioni nei k parametri incogniti.
L'obiettivo è quello di risolvere tale sistema per ottenere:

$$s = 1,2,\dots,k, \qquad \theta_S = g_s\left(\mu'_{j_1}, \mu'_{j_2}, \dots, \mu'_{j_p}\right)$$

Se si riesce ad ottenere suddetta equazione allora lo stimatore fornito dal metodo dei momenti consiste nella sostituzione dei p momenti teorici considerati con i corrispondenti momenti campionari.
Pertanto, si ha:

$$s = 1,2,\dots,k, \qquad \widehat{\Theta}_{s,MM} \coloneqq g_s\left(\bar{X}^{(j_1)}, \bar{X}^{(j_2)}, \dots, \bar{X}^{(j_p)}\right)$$

Sussistono i seguenti risultati.

Proposizione 1: Per le genitrici dotate di funzioni generatrici dei momenti gli stimatori campionari sono stimatori consistenti dei rispettivi momenti teorici.

Proposizione 2: Siano consistenti gli stimatori campionari in $\widehat{\Theta}_{s,MM} \coloneqq g_s\left(\bar{X}^{(j_1)}, \bar{X}^{(j_2)}, \dots, \bar{X}^{(j_p)}\right)$. Se, per $s = 1,2,\dots,k$, la funzione g_s è continua allora $\widehat{\Theta}_{s,MM}$ è uno stimatore consistente per θ_s

Alcuni esempi

1) Sia $b > 0$ e $X \sim U(0, b)$ con b parametro incognito. Scegliendo $p = 1$ e $j_1 = 1$, si ha:

$$\mu_1' = \frac{b}{2} \Leftrightarrow b = 2\mu_1'$$

Ne discende che $\hat{B}_{MM} \coloneqq 2\bar{X}^{(1)} = 2\bar{X}$. Ovviamente, lo stimatore ottenuto è consistente per b.

2) Sia $b > 0$ e $X \sim U(-b, b)$ con b parametro incognito. Scegliendo $p = 1$ e $j_1 = 1$, si ha:

$$\mu_1' = 0$$

Quindi la scelta $j_1 = 1$ non conduce ad una funzione di p. Scegliendo $j_1 = 2$ e tenendo conto che $b > 0$, si ha:

$$\mu_2' = \frac{b^2}{3} \Leftrightarrow b = \sqrt{3\mu_2'}$$

Ne discende che $\hat{B}_{MM} \coloneqq \sqrt{3\bar{X}^{(2)}}$. Lo stimatore ottenuto è consistente per b in quanto la radice quadrata è una funzione continua.

3) Siano $a < b$ due numeri reali tali che $a + b \neq 0$ e $a^2 + ab + b^2 \neq 0$. Sia $X \sim U(a, b)$ con a, b parametri incogniti. Risulta:

$$\begin{cases} \mu_1' = \dfrac{a+b}{2} \\ \mu_2' = \dfrac{a^2+ab+b^2}{12} \end{cases} \Leftrightarrow \begin{cases} a = \mu_1' - \sqrt{3[\mu_2' - (\mu_1')^2]} \\ b = \mu_1' + \sqrt{3[\mu_2' - (\mu_1')^2]} \end{cases}$$

Ne discende che

$$\begin{cases} \hat{A}_{MM} \coloneqq \bar{X} - \sqrt{3[\bar{X}^{(2)} - (\bar{X})^2]} \\ \hat{B}_{MM} \coloneqq \bar{X} + \sqrt{3[\bar{X}^{(2)} - (\bar{X})^2]} \end{cases}$$

Lo stimatore ottenuto è consistente per a (per b) in quanto la radice quadrata è una funzione continua e tali sono le operazioni di somma e prodotto di due funzioni continue.

6. Statistica descrittiva

Definizione e classificazione dei caratteri

Con il termine statistica descrittiva si intende un insieme di tecniche e strumenti finalizzati ad assolvere uno dei principali compiti assegnati alla Statistica: *descrivere, rappresentare e sintetizzare in maniera opportuna un insieme di dati relativamente ad una o più caratteristiche di una* ***popolazione*** *di interesse*.

Per *popolazione* (o collettivo statistico) si intende la totalità dei casi, ovvero dei *membri* (o unità statistiche) sui quali è possibile *rilevare* uno o più *caratteri* che rivestono particolare importanza per il fenomeno che si sta studiando.

Esempio 1: "Durante il semestre viene proposto agli studenti dei corsi di Laurea offerti in Ateneo il questionario sulla valutazione della struttura didattica nella sua complessità."
Qui la popolazione è costituita dagli studenti con o senza obbligo di frequenza ciascuno dei quali è un membro. Per lo studio in questione ci si serve di un questionario avente un certo numero di domande raggruppate per sezioni. In questo contesto ciascuna domanda del questionario corrisponde ad un *carattere*. Per la maggior parte delle domande lo studente può rispondere scegliendo una tra 4 possibili risposte tra loro alternative: *"Decisamente no"*, *"Più no che sì"*, *"Più sì che no"*, *"Decisamente sì"*.

Esempio 2: "In un allevamento di bufale da latte si vuole mettere in relazione la produzione giornaliera di latte con la grandezza delle mammelle e con lo stato di salute delle stesse."
Qui la popolazione è costituita da tutte le bufale dell'allevamento ciascuna delle quali è un membro. Per lo studio in questione i caratteri da rilevare su ciascun membro sono: la produzione di latte, una misura lineare delle mammelle, lo stato di salute delle mammelle. Per la rilevazione sono necessari due strumenti di misura e la diagnosi di un veterinario determinata su una prestabilita *scala* di valori.

Le diverse espressioni con le quali si manifesta un carattere si chiamano *modalità*.

Nell'Esempio 1 le modalità, per ciascun carattere, sono rappresentate dalle 4 risposte alternative tra loro. Ciascuno dei caratteri, ovvero una domanda proposta nel questionario, è di tipo *qualitativo ordinale*. Infatti, si osservi che esse sono delle etichette e che in relazione al gradimento espresso dallo studente risulta: $\text{Decisamente no} < \text{Più no che sì} < \text{Più sì che no} < \text{Decisamente sì}$.

Invece nell'Esempio 2, i caratteri "produzione giornaliera di latte" e "grandezza delle mammelle" sono *quantitativi continui* in quanto per la loro determinazione sono necessari uno strumento di misurazione di una capacità (volume) e uno strumento per la misurazione di una lunghezza e pertanto i dati ottenuti sono numeri decimali appartenenti ad un conveniente intervallo. Nello stesso esempio, il carattere "stato di salute delle mammelle" è invece di tipo qualitativo ordinale e le modalità sono i diversi valori della scala prescelta (ad esempio, mammelle sane, infiammazione lieve, infiammazione moderata, infiammazione grave). Il fatto che si parla di scala comporta la presenza di un ordinamento tra le etichette.

Per un esempio di carattere *qualitativo nominale* si pensi al gruppo sanguigno che, prescindendo dal fattore Rh, si manifesta con le modalità: A, B, AB e 0.

Per un esempio di carattere *quantitativo discreto* si pensi al numero delle persone presenti nello stato di famiglia dei residenti nel comune di Napoli alla data del più recente censimento ISTAT.

Distribuzioni di frequenza

Si può senz'altro pensare ad una popolazione come ad un insieme. La cardinalità della popolazione rilevata è detta *taglia*; essa, di solito, si designa con la lettera N. Un carattere viene designato con una lettera latina maiuscola mentre i valori rilevati vengono rappresentati con la stessa lettera ma in minuscolo. Se, allora, Y rappresenta il carattere sotto studio non continuo, la sequenza $\bar{y} = (y_1, y_2, \dots, y_N)$ rappresenta l'intera rilevazione dei dati. Denotiamo ora con $x_1, x_2, \dots, x_k$ (con $k \leq N$) le modalità del carattere Y. Per $j =$

$(1,2,\dots,k)$, il numero n_j che rappresenta quante volte è presente la modalità x_j in $\bar{y}$ è detto *frequenza assoluta* (o semplicemente *frequenza*) della modalità x_j. In aggiunta, $j = (1,2,\dots,k)$, $\boldsymbol{f_j = \frac{n_j}{N}}$ è detto *frequenza relativa* della modalità x_j.

La frequenza relativa è più informativa della frequenza assoluta in quanto tiene conto anche della taglia. È del tutto ovvio che la somma delle k frequenze assolute vale N mentre la somma delle k frequenze relative vale 1.

Infine, $j = (1,2,\dots,k)$, $\boldsymbol{F_j = f_1 + f_2 + \cdots + f_j = F_{j-1} + f_j}$ è detta *frequenza assoluta cumulata* di x_j. È del tutto ovvio che $F_k = 1$.

La rappresentazione tabellare di quanto appena esposto è detta *distribuzione di frequenza* della rilevazione dati di $\bar{y} = (y_1, y_2, \dots, y_N)$:

modalità del carattere	frequenza assoluta	frequenza relativa	frequenza relativa cumulata
x_1	n_1	$f_1 = n_1 / N$	$F_1 = f_1$
x_2	n_2	$f_1 = n_1 / N$	$F_2 = F_1 + f_2$
x_{k-1}	n_{k-1}	$f_{k-1} = n_{k-1} / N$	$F_{k-1} = F_{k-2} + f_{k-1}$
x_k	n_k	$f_k = n_k / N$	1
	———	———	
	N	1	

Per un carattere quantitativo continuo è necessario dapprima procedere alla suddivisione in *classi di modalità* dell'intervallo nel quale si manifesta il carattere stesso. Qui bisogna fare attenzione a rendere le classi contigue ma senza ingenerare dubbi di collocazione di un dato nelle classi stesse. Allo scopo, se i dati $(y_1, y_2, \dots, y_N)$ sono riportati con s cifre decimali, è sufficiente tenere conto dell'operazione di arrotondamento e rappresentare gli estremi delle classi con $s + 1$ cifre decimali. Ad esempio, supponiamo che una rilevazione di un peso è effettuata con bilancia digitale precisa all'ettogrammo. Sia $3{,}2\ kg$ il peso minore rilevato: $3{,}2$ rappresenta tutte le misurazioni comprese nell'intervallo $[3{,}15;\ 3{,}25[$. Allo stesso modo sia $4{,}4\ kg$ il peso maggiore rilevato: $4{,}4$ rappresenta tutte le misurazioni comprese nell'intervallo $[4{,}35;\ 4{,}45[$. Quindi se è vero che l'intervallo nel quale si osservano i dati è $[3{,}2;\ 4{,}4]$ è a maggior ragione vero che senza l'operazione di arrotondamento esso sarebbe stato $[3{,}15;\ 4{,}45[$. Allora, è quest'ultimo intervallo che deve essere suddiviso nel numero desiderato di classi e queste devono avere come estremi dei numeri aventi due cifre decimali. Dopo di ciò, per un carattere quantitativo, la prima colonna contiene le classi di modalità così individuate.

Moda e quartili

Definizione 1: Si consideri un carattere (di qualsiasi tipo). La modalità corrispondente alla frequenza (assoluta o relativa) più grande viene detta *moda* M_0 della rilevazione dati.

Definizione 2: Si consideri un carattere (non qualitativo nominale). La modalità corrispondente alla più piccola frequenza relativa cumulata maggiore o uguale a $0{,}5$ viene detta *mediana* M_1 oppure *secondo quartile* Q_2 della rilevazione dati. La mediana suddivide la rilevazione dati ordinata $\left(y_{(1)}, y_{(2)}, \dots, y_{(N)}\right)$ con

$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$ in due parti: i dati minori della mediana sono nello stesso numero dei dati maggiori della mediana.

Definizione 3: Si consideri un carattere (non qualitativo nominale). La modalità corrispondente alla più piccola frequenza relativa cumulata maggiore o uguale a $0{,}25$ viene detta *primo quartile* Q_1 della rilevazione dati. Il primo quartile corrisponde anche alla mediana dei dati minori della mediana.

Definizione 4: Si consideri un carattere (non qualitativo nominale). La modalità corrispondente alla più piccola frequenza relativa cumulata maggiore o uguale a $0{,}75$ viene detto *terzo quartile* Q_3 della rilevazione dati. Il terzo quartile corrisponde anche alla mediana dei dati maggiori della mediana.

Definizione 5: Si consideri un carattere (non qualitativo nominale). Il dato più piccolo $y_{(1)}$ è il *quartile di ordine zero* Q_0 della rilevazione dati.

Definizione 6: Si consideri un carattere (non qualitativo nominale). Il dato più grande $y_{(N)}$ è il *quarto quartile* Q_4 della rilevazione dati.

Rappresentazioni grafiche

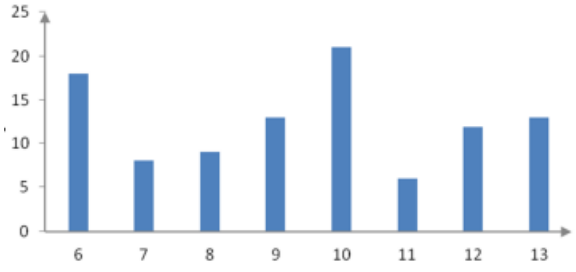
Le distribuzioni di frequenza possono essere rappresentate in forma grafica con scelta eseguita opportunamente rispetto al tipo di carattere. Per i caratteri qualitativi e quantitativi discreti, oltre al *diagramma circolare*, la rappresentazione grafica più usata è quella del *diagramma a barre verticali*: ogni barra verticale è centrata attorno ad una modalità ed ha altezza pari alla frequenza assoluta (oppure relativa).

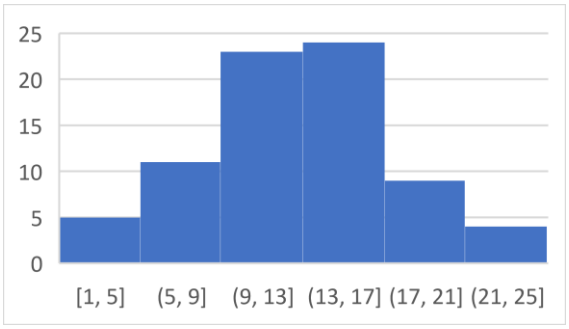
Per i soli caratteri quantitativi continui è opportuno utilizzare la rappresentazione grafica detta *istogramma*.

La differenza qualitativa tra un istogramma ed un diagramma a barre verticali è che nell'istogramma le barre verticali devono essere contigue (in effetti sono dei rettangoli e, quindi, dotati di base e altezza). Ma la differenza sostanziale è che nell'istogramma la frequenza della classe di modalità rappresenta l'area del rettangolo (e non la sua altezza come nel diagramma a barre).

Pertanto, per qualsiasi $j \in \{1,2,\dots,k\}$, se b_j rappresenta l'ampiezza della j-esima classe di modalità, l'altezza h_j del relativo rettangolo è ottenuto mediante la formula inversa per l'area e quindi: $h_j = \frac{n_j}{bj}$.

Esempio grafico di un diagramma a barre e di un istrogramma rispettivamente:





Indici di posizione

Si vuole ora considerare il problema di *sintetizzare* la rilevazione dati $\bar{y} = (y_1, y_2, \dots, y_N)$ con un unico valore di *tendenza centrale*, ossia un valore che fornisca un'indicazione di massima sulla localizzazione di $\bar{y}$. Ciò è utile non solo per una più immediata comprensione dei risultati dell'indagine ma anche per istituire un confronto del fenomeno studiato con altri fenomeni dello stesso tipo.

Per i caratteri quantitativi è possibile fare ricorso ai due indici media e mediana. Nel caso di una taglia N dispari, un modo pratico per ottenere la mediana senza costruire la distribuzione di frequenza è quello di ordinare i dati dal più piccolo al più grande e poi, ricorsivamente, depennare il minimo e il massimo fino a quando resta un unico elemento che è la mediana. Se invece N è pari, alla fine di tutti i depennamenti restano due elementi. In tal caso, bisogna separare il caso (i) i due elementi restanti sono uguali (il loro valore comune coincide con la mediana) dal caso (ii) i due elementi restanti sono diversi. Nel caso (ii) con un carattere quantitativo la mediana è la semisomma dei due elementi restanti. Nel caso (ii) con un carattere qualitativo ordinale la mediana è indeterminata.

Per i caratteri quantitativi ci sono almeno due approcci teorici in grado di far ottenere indici di tendenza centrale: le medie analitiche; i centri.

Medie analitiche

Sia $\bar{y} = (y_1, y_2, \dots, y_N)$ una rilevazione dati su un carattere quantitativo. Per ottenere una *media analitica* bisogna dapprima specificare un criterio C (o *funzione di circostanza*) rispetto al quale si vuole ottenere la valutazione di tendenza centrale. Dopo di ciò bisogna determinare un numero reale y per il quale, indicata con $\bar{y}^* = (y, y, \dots, y)$ una rilevazione dati (fittizia) di taglia N aventi tutti gli elementi uguali a y, la valutazione della funzione di circostanza C su $\bar{y}$ deve coincidere con la valutazione di C su $\bar{y}^*$. In simboli: $C(y_1, y_2, \dots, y_N) = C(y, y, \dots, y)$.

Esempio 1 (media aritmetica): Si scelga come funzione di circostanza C la somma dei dati: $C(y_1, y_2, \dots, y_N) = y_1 + y_2 + \cdots + y_N$. Dopo di ciò $C(y_1, y_2, \dots, y_N) = C(y, y, \dots, y) \Leftrightarrow y_1 + y_2 + \cdots + y_N = y + y + \cdots + y = N \cdot y$. In definitiva, la soluzione dell'equazione di circostanza è la *media aritmetica* dei dati:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Esempio 2 (media geometrica): Si scelga come funzione di circostanza C il prodotto dei dati (che devono essere rilevati da un carattere positivo): $C(y_1, y_2, \dots, y_n) = y_1 \cdot y_2 \cdot \dots \cdot y_N$. Dopo di ciò, $C(y_1, y_2, \dots, y_N) = C(y, y, \dots, y) \Leftrightarrow y_1 \cdot y_2 \cdot \dots \cdot y_N = y \cdot y \cdot \dots \cdot y = y^N$. In definitiva, la soluzione dell'equazione di circostanza è la *media geometrica* dei dati:

$$M_g = \sqrt[N]{y_1 \cdot y_2 \cdot \dots \cdot y_N} = \sqrt[N]{\prod_{i=1}^{N} y_i}$$

Esempio 3 (media armonica): Si scelga come funzione di circostanza C la somma dei reciproci dei dati (che devono essere rilevati da un carattere non nullo): $C(y_1, y_2, \dots, y_N) = \frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}$. Dopo di ciò,

$$C(y_1, y_2, \dots, y_N) = C(y, y, \dots, y) \Leftrightarrow \frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N} = \frac{1}{y} + \frac{1}{y} + \cdots + \frac{1}{y} = \frac{N}{y}$$

In definitiva, la soluzione dell'equazione di circostanza è la *media armonica* dei dati:

$$M_a = \frac{N}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}} = \left(\frac{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_N}}{N}\right)^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{y_i}\right)^{-1}$$

Dall'Esempio 1 e dall'Esempio 3 si può dire che la media armonica M_a di dati non nulli è uguale al reciproco della media aritmetica dei loro reciproci.

Centri

Sia $\bar{y} = (y_1, y_2, \dots, y_N)$ una rilevazione dati su un carattere quantitativo Y e sia $\forall x \in \mathbb{R}, d(x, \bar{y}) \geq 0$ una funzione che si ritiene adatta a rappresentare la *distanza* di un generico valore reale x da tutti gli elementi della rilevazione dati $\bar{y}$. Si definisce *centro* di una rilevazione dati $\bar{y}$, e lo si indica con $\xi(\bar{y})$, il punto minimo assoluto della funzione $d(x, \bar{y})$; in simboli: $\xi(\bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}(x, \bar{y})$.

In particolare, sono molto spesso considerate le seguenti funzioni di distanza *di tipo potenze*:

$$\forall x \in \mathbb{R}, d_0(x, \bar{y}) = \frac{1}{N}\sum_{i=1}^{N}|x - y_i|^0 \qquad \text{e} \qquad \forall r \in \mathbb{N}, \forall x \in \mathbb{R}, d_r(x, \bar{y}) = \sqrt[N]{\frac{1}{N}\sum_{i=1}^{N}|x - y_i|^r}$$

La funzione $d_r(x, \bar{y})$ è detta anche *distanza di ordine* r e il suo punto di minimo assoluto è detto *centro di ordine* r. Quindi il centro di ordine r di una rilevazione dati $\bar{y}$ è il numero reale $\xi_r(\bar{y})$ che rende minima la distanza di ordine r. In simboli: $\xi_r(\bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, d_r(x, \bar{y})$.

Teorema 1: Il centro di ordine 0 della rilevazione dati $\bar{y}$, ovvero $\xi_{00}(\bar{y})$, coincide con la moda della rilevazione dati.

Dimostrazione: Per definizione $d_0(x, \bar{y}) = \frac{1}{N}\sum_{i=1}^{N}|x - y_i|^0$ rappresenta la distanza di ordine 0 di x dall'intera rilevazione dati $\bar{y}$, mentre $|x - y_i|^0$ rappresenta la distanza tra x e il generico dato y_i.
Se ne ricava che quando x coincide con y_i la distanza è nulla ovvero l'addendo i-esimo non porta contributo alla distanza complessiva. Pertanto, si ha:

$$d_0(x, \bar{y}) = \begin{cases} 1, & \text{se } x \notin \{y_1, y_2, \dots, y_N\}, \\ 1 - \dfrac{n_x}{N} < 1, & \text{se } x \in \{y_1, y_2, \dots, y_n\}. \end{cases}$$

Nella precedente formula n_x rappresenta il numero delle volte che si presenta il dato x. Allora, il minimo si trova tra gli elementi di $\bar{y}$ e precisamente è quel dato al quale compete la frequenza maggiore che per definizione è la moda della rilevazione dati:

$$\xi_0(\bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}|x - y_i|^0 = M_0$$

Teorema 2: il centro di ordine 1 della rilevazione dati $\bar{y}$, ovvero $\xi_1(\bar{y})$, coincide con la mediana:

$$\xi_1(\bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, d_1(x, \bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}|x - y_i| = M_1 \equiv Q_2$$

Teorema 3: il centro di ordine 2 della rilevazione dati $\bar{y}$, $\xi_2(\bar{y})$, coincide con la media aritmetica.

Dimostrazione: Per definizione, $d_2(x, \bar{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|x - y_i|^2}$. D'altra parte, dal momento che la funzione radice quadrata è strettamente crescente nel suo dominio $[0, +\infty[$, il minimo della distanza di ordine 2 viene raggiunto in corrispondenza del minimo del radicando, ovvero della funzione:

$$x \in \mathbb{R}, \qquad f(x) = \frac{1}{N}\sum_{i=1}^{N}(x - y_i)^2$$

Pertanto, $\xi_2(\bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, d_2(x, \bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, f(x)$.

Ricerca dei punti stazionari di $f(x)$*:*

$$x \in \mathbb{R}, \qquad f'(x) = \frac{2}{N}\sum_{i=1}^{N} x - y_i \Rightarrow f'(x) = 0 \Leftrightarrow \sum_{i=1}^{N}(x - y_i) = 0 \Leftrightarrow x = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

Ricerca dei punti di minimo e massimo relativo di $f(x)$:

$$f''(x) = \frac{2}{N}\sum_{i=1}^{N} 1 = 2 \Rightarrow f''(\bar{y}) = 2 > 0$$

la qual cosa implica che $\bar{y}$ è un punto di minimo relativo per $f(x)$.
Ricerca del minimo assoluto di $f(x)$: $\bar{y} = \underset{x\in\mathbb{R}}{\operatorname{argmin}} f(x)$ in quanto $\lim_{x\to+\infty} f(x) = \lim_{x\to-\infty} f(x) = +\infty$, la derivata prima di $f(x)$ è definita in $\mathbb{R}$, e $f(x)$ ammette un unico punto di minimo relativo. In definitiva, come volevasi dimostrare: $\xi_2(\bar{y}) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\, d_2(x, \bar{y}) = \underset{x\in\mathbb{R}}{\operatorname{argmin}} f(x) = \bar{y}$.

Definizione 1: Il *centro di ordine infinito* di una rilevazione di dati $\bar{y}$ si indica con il simbolo $\xi_\infty(\bar{y})$ ed è definito dalla posizione: $\xi_\infty(\bar{y}) \coloneqq \lim_{r\to+\infty} \xi_r(\bar{y})$.

Definizione 2: Il valore centrale di una rilevazione dati $\bar{y}$ è la semisomma tra il dato minimo e il dato massimo:

$$\frac{y_{(1)} + y_{(N)}}{2} = \frac{Q_{(0)} + Q_{(4)}}{2}$$

Teorema 4: Il centro di ordine infinito della rilevazione dati $\bar{y}$ coincide con il valore centrale:

$$\xi_\infty(\bar{y}) = \frac{y_{(1)} + y_{(N)}}{2}$$

Indici di dispersione

Un *indice di dispersione* (o indicatore di dispersione o indice di variabilità o indice di variazione) serve per descrivere sinteticamente la misura con la quale una rilevazione dati di un carattere quantitativo è distante da una sua tendenza centrale. La dispersione esprime la bontà o la inadeguatezza di un indice di tendenza centrale quale descrittore di una distribuzione di frequenza.

Per i caratteri qualitativi si usano gli *indici di diversità* dei quali quello maggiormente usato è l'*indice di ricchezza* che opera un semplice conteggio del numero delle modalità presenti nella rilevazione dati.

Sia $\bar{y} = (y_1, y_2, \dots, y_N)$ una rilevazione dati su un carattere quantitativo. Sono indici di dispersione i seguenti:

a) La differenza tra il dato più grande da quello più piccolo (*campo o intervallo di variazione*): $\Gamma = y_{(N)} - y_{(1)} = Q_4 - Q_0$. Il campo di variazione è associato al valore centrale $\frac{y_{(1)}+y_{(N)}}{2}$.

b) La differenza tra il terzo e il primo quartile (*differenza interquartilica*): $\gamma = Q_3 - Q_1$. La differenza interquartilica è associata alla mediana Q_2.

c) La media del valore assoluto delle differenze dei dati dalla loro mediana Q_2 (*scarto mediano assoluto)*: $S_{Q_2} = \frac{1}{N}\sum_{i=1}^{N}|y_i - Q_2|$. Lo scarto mediano assoluto è associato alla mediana Q_2.

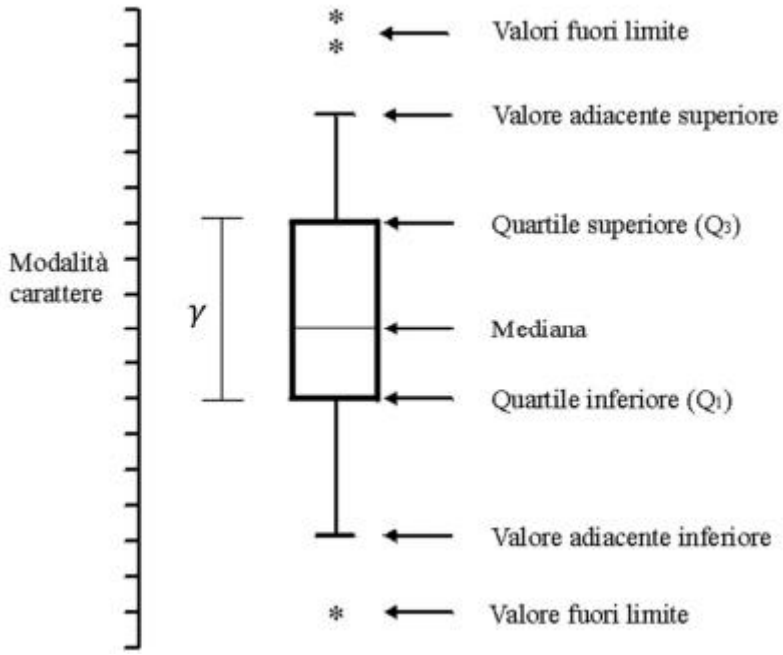
d) La media del valore assoluto delle differenze dei dati dalla loro media aritmetica $\bar{y}$ (*scarto medio assoluto*): $S_{\bar{y}} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}|$. Lo scarto medio assoluto è associato alla media aritmetica $\bar{y}$.

e) La media del quadrato delle differenze dei dati dalla loro media aritmetica M_2 (*varianza*): $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$. La varianza è associata alla media aritmetica $\bar{y}$.

f) La radice quadrata della varianza (*scarto tipo* o *deviazione standard*): $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$. La deviazione standard è associata alla media aritmetica $\bar{y}$.

Diagramma scatola con baffi

Un metodo grafico per rappresentare una distribuzione di frequenze che mette in risalto anche la dispersione intorno alla mediana è il *grafico della scatola con baffi* (Box Plot):



La linea interna alla scatola rappresenta la *Mediana* della distribuzione. Le linee estreme della scatola rappresentano il primo ed il terzo quartile.

La distanza interquartilica y, è una misura della **dispersione** della distribuzione. Il 50% dei dati si trovano comprese tra questi due valori. Se l'intervallo interquartilico è piccolo, tale metà delle osservazioni si trova fortemente concentrata intorno alla mediana; all'aumentare della distanza interquartilica aumenta la dispersione del 50% dei dati centrali intorno alla mediana.

Le distanze tra ciascun quartile e la mediana forniscono informazioni relativamente alla **forma** della distribuzione. Se una distanza è diversa dall'altra allora la distribuzione è asimmetrica.

Le linee che si allungano dai bordi della scatola (*baffi*) individuano gli intervalli in cui sono posizionati i valori rispettivamente minori di Q_1 e maggiori di Q_2; i punti estremi dei "baffi" evidenziano i *valori adiacenti*. Il *valore adiacente inferiore* (VAI) è il valore più piccolo tra i dati che risulta maggiore o uguale a $Q_1 - 1{,}5y$. Il *valore adiacente superiore* (VAS), invece, è il valore più grande tra i dati che risulta minore o uguale a $Q_3 + 1{,}5y$.

I valori esterni ai valori adiacenti (chiamati in genere *valori fuori limiti* oppure *valori anomali*), vengono segnalati individualmente nel box-plot per meglio evidenziarne la presenza e la posizione. Questi valori infatti costituiscono una "anomalia" rispetto alla maggior parte dei valori osservati e pertanto è necessario identificarli per poterne analizzare le caratteristiche e le eventuali cause che li hanno determinati. Essi forniscono informazioni ulteriori sulla dispersione e sulla forma della distribuzione.

Quando il valore adiacente superiore coincide con il dato più grande e il valore adiacente inferiore coincide con il dato più piccolo, allora non comparirà alcun valore anomalo.

7. Formulario

Elementi di probabilità

- Finita additività (gli eventi E_i sono mutuamente esclusivi):
  $$\mathcal{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathcal{P}(E_i)\,, \forall n \in \mathbb{N}$$
- Formula di inclusione-esclusione: $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B)$, caso generale:
  $$\mathcal{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathcal{P}(E_i) - \sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathcal{P}(E_i \cap E_j) + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n} \mathcal{P}(E_i \cap E_j \cap E_k) + \cdots + (-1)^{n-1}\mathcal{P}\left(\bigcap_{i=1}^{n} E_i\right)$$
- Probabilità condizionata:
  $$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}$$
- Probabilità dell'evento E come media pesata delle probabilità condizionali di E (F si è verificato o no):
  $$\mathcal{P}(E) = \mathcal{P}(E|F)\mathcal{P}(F) + \mathcal{P}(E|F^c)\big(1 - \mathcal{P}(F)\big)$$
- Formula di fattorizzazione (E ed F_i sono mutuamente esclusivi):
  $$\mathcal{P}(E) = \sum_{i=1}^{n} \mathcal{P}(E|F_i)\mathcal{P}(F_i)$$
- Formula di Bayes:
  $$\mathcal{P}\big(A_j\big|B\big) = \frac{\mathcal{P}\big(B\big|A_j\big)\mathcal{P}\big(A_j\big)}{\sum_{i=1}^{n} \mathcal{P}(B|A_i)\mathcal{P}(A_i)}$$
- Eventi indipendenti: $\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F)$. Mentre per i eventi si ha:
  $$\mathcal{P}\left(\bigcap_{i=1}^{r} E_{\alpha_i}\right) = \prod_{i=1}^{r} \mathcal{P}\big(E_{\alpha_i}\big)$$
  Dove $E_{\alpha_1}, E_{\alpha_2}, \dots, E_{\alpha_r}$, con $1 \le a_1 < a_2 < \cdots < a_{r-1} < a_r \le n$, sono sottogruppi di $E_1, E_2, \dots, E_n$.

Richiesta	Proprietà	Formula
sequenze con ripetizione	L'ordine è rilevante (due sequenza con stessi simboli in ordine diverso sono considerate differenti) ed è possibile che uno o più elementi vengano ripetuti una o più volte	n^m dove n è la cardinalità dell'insieme e m la lunghezza della sequenza
disposizioni *(sequenze senza ripetizione)*	L'ordine è rilevante ed ogni elemento è presente una ed una sola volta nella sequenza.	$\frac{n!}{(n-r)!}$ dove n è la cardinalità dell'insieme e r la lunghezza della sequenza
permutazioni	Il numero di modi diversi in cui posso disporre n oggetti	$n!$
permutazioni con ripetizione	In quanti modi si possono formare r gruppi di un insieme di m elementi in modo che il gruppo i-esimo contenga esattamente n_i elementi. N.B.: $n_1 + n_2 + \cdots + n_r = m$	$\binom{m}{n_1, n_2, \dots, n_r} = \frac{m!}{n_1! \cdot n_2! \cdot \dots \cdot n_r!}$
combinazioni	Il numero di gruppi costituiti da r elementi a partire da un insieme di n oggetti	$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$
combinazioni con ripetizione	Combinazioni dove è possibile che uno o più elementi vengano ripetuti una o più volte	$\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!\,(n-1)!}$

La tabella sopra riporta degli elementi di calcolo combinatorio.

Variabili aleatorie e valore attero

- Proprietà della funzione di massa $p(a) \coloneqq \mathcal{P}(X = a)$
  $\sum_{i=1}^{\infty} p(x_i) = 1$
- Funzione di densità:
  $1 = \mathcal{P}(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)dx \qquad \mathcal{P}(a \leq X \leq b) = \int_{a}^{b} f(x)dx$
- Funzione di ripartizione congiunta:
  $F(x,y) \coloneqq \mathcal{P}(X \leq x, Y \leq y) \Rightarrow F_X(x) = F(x,\infty) \wedge F_y(y) = F(\infty, y)$
- Funzione di massa di probabilità congiunta
  $p(x_i, y_j) \coloneqq \mathcal{P}(X = x_i, Y = y_j) \Rightarrow p_X(x_i) = \sum_j p(x_i, y_j) \wedge p_Y(y_j) = \sum_i p(x_i, y_j)$
- Densità congiunta
  $\mathcal{P}((X,Y) \in C) = \iint_{(x,y)\in C} f(x,y)dxdy \Rightarrow f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \wedge f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx$
- Variabili aleatorie indipendenti:
  $\mathcal{P}(X \leq a, Y \leq b) = \mathcal{P}(X \leq a)\mathcal{P}(Y \leq b), \qquad F(a,b) = F_X(a)F_Y(b), \qquad p(x,y) = p_X(x)p_Y(y)$
- Valore atteso di una variabile:
  discreta: $E[X] \coloneqq \sum_i x_i \mathcal{P}(X = x_i)$, continua: $E[X] \coloneqq \int_{-\infty}^{\infty} xf(x)dx$
- Valore atteso di una funzione di variabile aleatoria
  discreta: $E[g(X)] = \sum_x g(x)p(x)$, continua: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$
- Momento n-esimo della variabile aleatoria X:
  $E[X^n] = \sum_x x^n p(x)$ se X è discreta, $E[X^n] = \int_{-\infty}^{\infty} x^n f(x)dx$ se X è continua
- Valore atteso della somma di variabili aleatorie:
  $E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$
- Varianza:
  $\mathrm{Var}(X) \coloneqq E[(X-\mu)^2] = E[X^2] - E[X]^2$
- Identità per la varianza
  $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
- Covarianza
  $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] \Rightarrow \mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- Proprietà della covarianza:
  $\mathrm{Cov}(aX,Y) = a\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,aY), \quad \mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$
- Varianza di una somma di variabili aleatorie:
  $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j)$
  Per $n = 2$: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$
- Varianza per due variabili aleatorie indipendenti:
  $E[XY] = E[X]E[Y] \Rightarrow \mathrm{Cov}(X,Y) = 0$
- Coefficiente di Correlazione:
  $\mathrm{Corr}(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$

- Funzione generatrice dei momenti:
  $\phi(t) = E[e^{tX}] = \sum_x e^{tx}p(x)$ se X è discreta, $\quad \phi(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx}f(x)dx$ se X è continua
- Funzione generatrice per variabili indipendenti:
  $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$
- Legge debole dei grandi numeri:
  $\mathcal{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad$ quando $n \to \infty$

Modelli di variabili aleatorie

- Variabili aleatorie uniformi
  $f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{se } \alpha \le x \le \beta \\ 0 & \text{altrimenti} \end{cases}, \quad E[X] = \frac{\alpha + \beta}{2}, \quad Var(X) = \frac{(\beta - \alpha)^2}{12}$

Binomiale $X \sim B(n,p)$	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ con $k = 0,1,\dots,n$	Esperimento di n prove ripetute (indipendenti); p rappresenta la probabilità di successo in una singola prova.
Bernoulliana $X \sim B(1,p)$	$p(k) = p^k (1-p)^{n-k}$ con $k = 0,1$	Caso particolare del binomiale in cui il numero di prove è pari a 1
Geometrica $X \sim G(p)$	$p(k) = (1-p)^{k-1}p$ con $k = 1,2,\dots$	Esperimento di prove (indipendenti) ripetute fino ad ottenere il successo; p rappresenta la probabilità di successo in una singola prova
Poisson $X \sim Poisson(\lambda)$	$p(k) = \frac{a^k}{k!} e^{-\lambda}$ con $k = 0,1,\dots,n$	Caso particolare della binomiale in cui: n, il numero di prove, viene fatto tendere a ∞; mentre p, la probabilità di successo di una singola prova, a 0; il parametro della densità λ è pari a np
Ipergeometrica $X \sim IG(a,b,n)$	$p(k) = \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}$ con $k = \max\{0, n-b\}, \dots, \min\{a,n\}$	Esperimento di n prove ripetute (non indipendenti)

La precedente tabella riporta le principali densità discrete.

La distribuzione delle statistiche campionarie

- La media campionaria con campione di dati X_i, loro media μ e loro varianza σ^2:
  $\bar{X} \coloneqq \frac{X_1 + X_2 + \cdots + X_n}{n}, \quad E[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- Per il teorema del limite centrale:
  $\mathcal{P}\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right) \approx \Phi(x)$
- La varianza campionaria:
  $S^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Stima parametrica

- Stimatori di massima verosimiglianza:
  $\text{massimizzare } \theta \text{ in} \log[f(x_1, x_2, \dots, x_n|\theta)] \text{ con } f(x_1, x_2, \dots, x_n) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^{n} x_i}$

Statistica descrittiva

- Frequenza relativa di x_j (n_j è la frequenza assoluta: quante volte è presente la modalità x_j in $\bar{y}$):
  $f_j = \frac{n_j}{N} \text{ con } j = (1,2, \dots, k), \text{ modalità del carattere } Y\text{: } (x_1, \dots, x_k), \bar{y} = (y_1, \dots, y_N)$
- Frequenza assoluta cumulata di x_j:
  $F_j = f_1 + f_2 + \cdots + f_j = F_{j-1} + f_j, \qquad F_k = 1$
- Distanza di ordine r
  $$d_r(x, \bar{y}) = \sqrt[n]{\frac{1}{N}\sum_{i=1}^{N} |x - y_i|^r}$$
- Campo di variazione:
  $\Gamma = y_{(N)} - y_{(1)} = Q_4 - Q_0$
- Differenza interquartilica:
  $\gamma = Q_3 - Q_1$
- Scarto mediano assoluto:
  $$S_{Q_2} = \frac{1}{N}\sum_{i=1}^{N} |y_i - Q_2|$$
- Scarto medio assoluto:
  $$S_{\bar{y}} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \bar{y}|$$
- Varianza
  $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2$$
- Scaro tipo (o deviazione standard)
  $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

8. Bonus: risposte alle domande orali più frequenti

Le seguenti risposte sono state affettuosamente scopiazzate dagli appunti di **L. Brando**

L'impostazione assiomatica di Kolmogorov.

Una famiglia $\mathcal{A}$ di sottoinsiemi di Ω è detta una ***σ*-algebra degli eventi** se verifica le seguenti proprietà:

4)	$\Omega \in \mathcal{A}$	l'*evento certo* è un evento ($A \in \mathcal{A}$ = "A è un evento")
5)	$A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$	L'*evento contrario* è un evento
6)	Se $\{A_n\}_{n\in\mathbb{N}}$ è una successione di elementi di $\mathcal{A}$, allora: $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$	$\mathcal{A}$ è stabile rispetto all'unione numerabile

Dalla precedente definizione si possono effettuare le seguenti osservazioni sugli assiomi di Kolmogorov:

- Dagli assiomi 1 e 2 segue che $\emptyset = \Omega^c \in \mathcal{A}$ (l'evento impossibile è un evento)
- Dagli assiomi 2 e 3 e dalle leggi di De Morgan si dimostra anche la stabilità rispetto l'intersezione numerabile: sia $\{A_n\}_{n\in\mathbb{N}}$ una successione numerabile di eventi in $\mathcal{A}$ risulta: $\bigcap A_n = (\bigcup A_n^c)^c$

La formula delle alternative e il teorema di Bayes.

Ho un insieme di eventi $A_1, A_2, \dots, A_n$ si dice "costruire" un insieme di alternative se gode di 3 proprietà:

1) $i = 1,2,\dots,n,\ \mathcal{P}(A_i) > 0$
2) $i,j = 1,2,\dots,n$ con $i \neq j,\ A_i \cap A_j = \emptyset$
3) $\mathcal{P}(\bigcup_{i=1}^{n} A_i) = \Omega = \sum_{i=1}^{n} \mathcal{P}(A_i)$

Da queste 3 proprietà si ricava la formula $\mathcal{P}(B) = \sum_{i=1}^{n} \mathcal{P}(B|A_i)\mathcal{P}(A_i)$. Infatti, $B = B \cap \Omega = B \cap (\bigcup A_i) = \bigcup(B \cap A_i) \Rightarrow\ \mathcal{P}(B) = \mathcal{P}(\bigcup(B \cap A_i)) = \sum_{i=1}^{n} \mathcal{P}(B \cap A_i) = \sum_{i=1}^{n} \mathcal{P}(B|A_i)\mathcal{P}(A_i)$.

Se ho $A_1, \dots, A_n$ insieme di alternative che partizionano lo spazio Ω degli eventi ed un evento $B \subseteq \Omega$ tale che $\mathcal{P}(B) = \emptyset$, si trova l'espressione per

$$\mathcal{P}(A_i|B) = \frac{\mathcal{P}(B|A_i)\mathcal{P}(A_i)}{\sum_{j=1}^{n} \mathcal{P}\big(B\big|A_j\big)\mathcal{P}\big(A_j\big)}$$

N.B.: $\mathcal{P}(A)$ si chiama probabilità a priori mentre $\mathcal{P}(A|B) = \frac{\mathcal{P}(A\cap B)}{\mathcal{P}(B)}$ è la probabilità condizionata di A dato B. Mentre $\mathcal{P}(B|A)$ si dice verosimiglianza.

Teorema di Bayes: Sia $(\Omega, \mathcal{A}, \mathcal{P})$ uno spazio di probabilità. Sia $A_1, \dots, A_n$ una partizione di Ω, tale che $\mathcal{P}(A_i) > 0$ per ogni $i = 1,2,\dots,n$. Sia infine B un evento $\mathcal{P}(B) > 0$. Allora per ogni $j = 1,2,\dots,n$ risulta:

$$\mathcal{P}\big(A_j\big|B\big) = \frac{\mathcal{P}\big(B\big|A_j\big)\mathcal{P}\big(A_j\big)}{\sum_{i=1}^{n} \mathcal{P}(B|A_i)\mathcal{P}(A_i)}$$

Tale teorema è impiegato per calcolare la probabilità di una causa che ha provocato l'evento verificato $E = (E \cap F) \cup (E \cap F^c)$. Infatti, $\mathcal{P}(E) = \mathcal{P}(E \cap F) + \mathcal{P}(E \cap F^c) = \mathcal{P}(E|F)\mathcal{P}(F) + \mathcal{P}(E|F^c)\mathcal{P}(F^c) = \mathcal{P}(E|F)\mathcal{P}(F) + \mathcal{P}(E|F^c)\big(1 - \mathcal{P}(F)\big)$.

Gli elementi fondamentali nella definizione di una variabile aleatoria.

Se $\mathcal{F}$ è una σ-algebra sull'insieme S, la coppia $(S, \mathcal{F})$ si dice uno **spazio misurabile** e gli elementi di $\mathcal{F}$ sono chiamati *insiemi misurabili* o solo *misurabili*. All'insieme $\mathbb{R}$ è associata, di norma, la σ-algebra generata dagli *insiemi di Borel* (intervalli aperti e limitati).

Si consideri ora uno spazio di probabilità $(\Omega, \mathcal{F}, \mathcal{P})$ e lo spazio misurabile $(\mathbb{R}, \mathcal{B})$. Un *numero aleatorio* è un'applicazione X di Ω in $\mathbb{R}$ con la condizione che la controimmagine tramite X di un boreliano deve essere un evento: $B \in \mathcal{B},\ X^{-1}(B) \in \mathcal{F}$.

Definiamo ora una misura di probabilità sulla σ- algebra $\mathcal{B}$:
$B \in \mathcal{B},\ \mathcal{P}_X(B) \coloneqq \mathcal{P}[X^{-1}(B)] = \mathcal{P}(X \in B)\ \forall B \in \mathcal{B}.$

La misura di probabilità $\mathcal{P}_X$ è la *distribuzione* oppure la *legge* del numero aleatorio X. Essa fornisce tutte le informazioni probabilistiche relative al numero aleatorio X ed esse si basano sulla misura di probabilità $\mathcal{P}$ della descrizione matematica dell'esperimento aleatorio $\mathcal{E}$ sottostante.

Le proprietà della funzione di distribuzione.

La funzione di distribuzione, data una variabile aleatoria X è la funzione che fa corrispondere ai valori di X la probabilità $\mathcal{P}(X \leq x)$: $F_X: \mathbb{R} \to [0,1];\ \ F_X(x) \coloneqq \mathcal{P}(X \leq x)$ (quindi la funzione di distribuzione è la probabilità che X assuma un valore minore od uguale a x)

Proprietà:

1) F_X è non decrescente.
Sia $x_1 < x_2 \Rightarrow F_X(x_2) = \mathcal{P}(X \leq x_2) = \mathcal{P}\big((X \leq x_1) \cup (x_1 < X < x_2)\big) = \mathcal{P}(X \leq x_1) + \mathcal{P}(x_1 < X < x_2) = F_X(x_1) + \mathcal{P}(x_1 < X < x_2) \Rightarrow F_X(x_2) \geq F_X(x_1)$

2) F è continua a destra; ovvero: $\forall x \in \mathbb{R}, F(x^+) = \lim_{n\to+\infty} F\left(x + \frac{1}{n}\right) = \lim_{t\to x^+} F(t) = F(x)$
Attraverso la definizione di funzione di ripartizione e scomponendo l'evento in un'unione di due eventi disgiunti si ottiene: $F(x^+) = \lim_{n\to+\infty} F\left(x + \frac{1}{n}\right) = \lim_{n\to+\infty} \mathcal{P}\left(X \leq x + \frac{1}{n}\right) =$
$= \lim_{n\to+\infty} \mathcal{P}\left((X \leq x) \cup \left(x < X \leq x + \frac{1}{n}\right)\right) = \lim_{n\to+\infty} \left(\mathcal{P}(X \leq x) + \mathcal{P}\left(x < X \leq x + \frac{1}{n}\right)\right) =$
$= \lim_{n\to+\infty} \mathcal{P}(X \leq x) + \lim_{n\to+\infty} \mathcal{P}\left(x < X \leq x + \frac{1}{n}\right) = \lim_{n\to+\infty} F(x) + 0 = F(x)$

3) $\lim_{n\to+\infty} F(n) = 1 \ \wedge\ \lim_{n\to-\infty} F(n) = 0$
Cominciamo con la prima osservando che $F(n)$ è la probabilità dell'evento $A_n = \{X \leq n\}$. Sappiamo che la successione di eventi $\{A_n\}_{n\in\mathbb{N}}$ è *crescente* e $\cup A_n = \Omega$. Dunque, per la proprietà di passaggio al limite delle successioni si ha $\lim_{n\to+\infty} F(n) = \lim_{n\to+\infty} \mathcal{P}(A_n) = \mathcal{P}(\cup A_n) = \mathcal{P}(\Omega) = 1$. Ragionamento analogo per la seconda: $F(n)$ è la probabilità dell'evento $A_n = \{X \leq -n\}$ ma essendo la successione *decrescente* si ha $\cap A_n = \emptyset$. Dunque, sempre per la suddetta proprietà, si ha $\lim_{n\to-\infty} F(n) = \lim_{n\to-\infty} \mathcal{P}(A_n) = \mathcal{P}(\cap A_n) = \mathcal{P}(\emptyset) = 1$.

Collegamento tra la legge binomiale e quella di Poisson.

Siamo in un esperimento il cui esito è "successo" $X = 1$ o "fallimento" $X = 0$ e la cui funzione di massa di probabilità di X è definita come segue: **(1)** $\mathcal{P}(X = 0) = 1 - p,\ \mathcal{P}(X = 1) = p$ con $0 \leq p \leq 1$.

Definizione 1: una variabile aleatoria si dice bernoulliana se la sua funzione di massa di probabilità è come la (1). Può assumere solo valori nell'intervallo $[0,1]$ e la sua media è $E[X] \coloneqq 1\mathcal{P}(X = 1) + 0\mathcal{P}(X = 0) = p$

Definizione 2: n ripetizioni indipendenti di un esperimento che può concludersi in un successo di probabilità p o in un fallimento con probabilità $1 - p$. Indico con X il numero totali di successi, ovvero la *variabile aleatoria binomiale di probabilità* (n, p)**: (2)** $\mathcal{P}(X = i) = \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i}$ con $i = 0,1,..,n$

Fissata una qualunque sequenza di esiti con i successi e $n - i$ fallimenti, la probabilità che si realizzi esattamente tale sequenza è $p^i(1-p)^{n-i}$ per l'indipendenza delle ripetizione. Le sequenze di esiti sono $\binom{n}{i}$. Allora $\sum \mathcal{P}(X = i) = \sum_{i=0}^{n} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} = [p + (1-p)]^n = 1$.

La variabile aleatoria binomiale di parametri (n, p) viene rappresentata come somma di bernoulliane. Se X è binomiale di parametri (n, p) allora **(3)** $X = \sum_{i=1}^{n} x_i$ dove $x_i = 1$ se la i-esima prova ha successo, $x_i = 0$ altrimenti.

x_i bernoulliana di parametro p: $E[x_i] = p, E[x_i^2] = p$ (essendo $x_i = x_i^2$); $D^2[x_i] = p - p^2 = p(1-p)$. Dalle proprietà di media e varianza, e dalla (3) segue che la media $E[X] = \sum_{i=1}^{n} E[x_i] = np$, mentre la varianza $D^2[X] = \sum_{i=1}^{n} D^2[x_i] = np(1-p)$.

Definizione 3: Una variabile aleatoria X di valori 1,2, … si dice di Poisson di parametro $\lambda > 0$ se la sua funzione di massa di probabilità è data da **(4)** $\mathcal{P}(X = i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!}$.

Dalla (4): $\sum_{i=0}^{\infty} \mathcal{P}(x = i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda} \cdot e^{-\lambda} = 1$

Sia X una variabile aleatoria di Poisson. Per determinare la media e la varianza, calcoliamo la sua funzione generatrice dei momenti: $\phi(t) = E[e^{tX}] = \sum_{i=0}^{\infty} e^{ti} \mathcal{P}(X = i) = e^{-\lambda} \sum_{i=0}^{\infty} e^{ti} \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$. Derivando si trova allora $\phi'(t) = \lambda e^t e^{\lambda(e^t-1)}$ e $\phi''(t) = (\lambda e^t)^2 e^{\lambda(e^t-1)} + \lambda e^t e^{\lambda(e^t-1)}$ e valutando le due espressioni in $t = 0$, si ottiene $E[X] = \phi'(0) = \lambda$ e$D^2[X] = \phi''(0) - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Reale risposta: Sia X una variabile aleatoria binomiale di parametri (n, p) e $\lambda = np$ si ha:

$$\mathcal{P}(X = i) = \frac{n!}{(n-i)!\, i!} p^i (1-p)^{n-i} = \frac{n(n-1)\dots(n-i+1)}{i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} =$$

$$= \frac{n(n-1)\dots(n-i+1)}{n^i} \cdot \frac{\lambda^i}{i!} \cdot \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^i}$$

Se si suppone che n sia molto grande e p molto piccolo, valgono le seguenti approssimazioni,

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \qquad \frac{n}{n} \cdot \frac{n-1}{n} \cdot \dots \cdot \frac{n-i+1}{n} \approx 1, \qquad \left(1 - \frac{\lambda}{n}\right)^i \approx 1$$

E quindi, se n è grande, p è piccolo e $\lambda = np$: $\mathcal{P}(X = i) \approx \frac{\lambda^i}{i!} e^{-\lambda}$

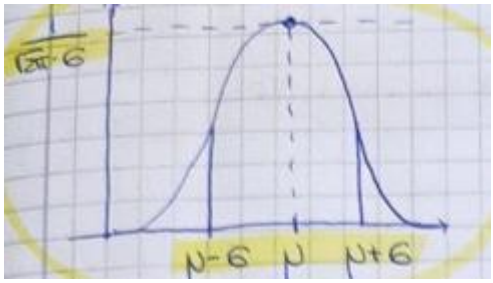
La legge normale e l'uso della tavola relativa alla funzione di distribuzione standard.

Una variabile aleatoria ha *legge normale* o *gaussiana* di parametri μ (media) e σ^2 (scarto quadratico medio) se ha la seguente densità di probabilità:

$$f(X, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow X \sim N(\mu, \sigma^2)$$

μ indica la distribuzione: valor a cui corrisponde il valore massimo mentre σ^2 caratterizza la forma della curva in quanto misura di dispersione dei valori attorno al valore medio

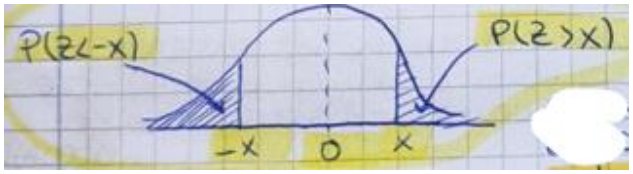
1) Simmetrico rispetto all'asse verticale di $x = \mu$ (massimo assoluto)
2) $\mu \pm \sigma$ sono i punti di flesso
3) La tavola della distribuzione naturale standard ti permette di calcolare la probabilità di un quartile o trovare il valore del quartile che delimita l'area alla sua sinistra (della curva gaussiana)



Definizione: Sia $X \sim N(\mu, \sigma^2)$ allora $Z \coloneqq \frac{x-\mu}{\sigma}$ è una variabile aleatoria con $\mu = 1$ e $\sigma^2 = 1$. Si dice *normale standard* la funzione di ripartizione $\phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{y}{z}} dy \quad \forall x \in \mathbb{R}$

N.B.: Z ha distribuzione normale standard qunado x è gaussiana con media μ e varianza σ^2. Questo ci permette di esprimere la probabilità relativa a x in termini di probabilità su Z.

Se ho $x > 0$ e Z variabile aleatoria normale standard allora $\phi(-x) = \mathcal{P}(Z < -x) = \mathcal{P}(Z > x) = 1 - \mathcal{P}(Z < x) = 1 - \phi(x)$



Proprietà della media e della varianza di una variabile aleatoria.

Sia X una variabile aleatoria discreta, definiamo la media $E[X] = \sum x_i \mathcal{P}(X = x_i)$.
Sia X una variabile aleatoria continua con funzione di densità f, definiamo la media $E[X] \coloneqq \int_{-\infty}^{\infty} xf(x)dx$.

Proprietà: Sia X una variabile aleatoria di cui conosciamo la distribuzione; posso considerare una sua funzione $g(X)$. Ne consegue che $g(X)$ è una variabile aleatoria di cui conosco la distribuzione e posso ottenere $E[g(X)]$.

Proposizione 1: Sia X una variabile aleatoria discreta con funzione di massa di probabilità p. Allora $E[g(X)] = \sum_x g(x)p(x)$.

Proposizione 2: Sia X una variabile aleatoria discreta con funzione di densità di probabilità f. Allora $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

Corollario: $\forall a, b,\ E[aX + b] = aE[X] + b$

- Caso discreto: $E[aX + b] = \sum(ax + b)p(x) = a\sum xp(x) + b\sum p(x) = aE[X] + b$
- Caso continuo: $E[aX + b] = \int(ax + b)f(x)dx = \underbrace{a\int xf(x)dx}_{a=0, E(b)=b} + \underbrace{b\int f(x)dx}_{b=0, E(ax)=aE(x)} = aE[X] + b$

Definizione: Sia $E[X^n]$ con $n = 1,2, \dots, n$. Allora $E[X^n] = \begin{cases} \sum_x x^n p(x) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{\infty} x^n f(x)dx & \text{se } X \text{ è continua} \end{cases}$

Sia X una variabile aleatoria di cui conosco la distribuzione e con valori distribuiti attorno alla media $\mu \coloneqq E[X]$ (con $E[|X - \mu|]$ quantifico la loro distanza da μ).

Sia X una variabile aleatoria con media μ. Definisco la **varianza** come il rapporto tra il valore atteso del quadrato di X e il quadrato della media. In simboli: $D^2[X] = E[(X - \mu)^2] = E[X^2] - E^2[X]$.

$$\forall a, b, \quad D^2[aX + b] = E[(aX + b - E[aX + b])^2] = E[(aX + b - a\mu - b)^2] = E[a^2(X - \mu)^2]$$
$$= a^2 E[X - \mu]^2 = a^2 D^2[X] \rightarrow \begin{cases} a = 0 \Rightarrow D^2[b] = 0 \\ a = 1 \Rightarrow D^2[X + b] = D^2[X] \\ b = 0 \Rightarrow D^2[aX] = a^2 D^2[X] \end{cases}$$

Deviazione standard: $D[X] = \sqrt{D^2[X]} = \sqrt{E[X^2] - E^2[X]}$

Enunciato e utilizzo della legge debole dei grandi numeri.

Markov: Sia X una variabile aleatoria, $\forall a > 0,\ \mathcal{P}(X \geq a) \leq \frac{E[X]}{a}$
Dim.: $E[X] \coloneqq \int_0^{\infty} xf(x)dx = \int_0^a xf(x)dx + \int_a^{\infty} xf(x)dx \geq \int_a^{\infty} xf(x)dx \geq \int_a^{\infty} af(x)dx =$
$= a\int_a^{\infty} f(x)dx = a\mathcal{P}(X \geq a)$. Da cui segue la tesi dividendo entrambi i fattori per a.

Chebyshev: : Sia X una variabile aleatoria con media μ e varianza σ^2, $\forall r > 0,\ \mathcal{P}(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2}$
Dim.: $\{|X - \mu| \geq r\}$ è equiprobabile con $\{(X - \mu)^2 \geq r^2\}$ da cui segue la tesi per il teorema di Markov con $a = r^2$. Infatti: $\mathcal{P}((X - \mu)^2 \geq r^2) \leq \frac{E[(X-\mu)^2]}{r^2} = \frac{\sigma^2}{r^2}$.

N.B.: limitare la probabilità di eventi rari che riguardano variabili aleatorie di cui conosciamo solo media e varianza.

Legge debole dei grandi numeri: Sia $X_1, X_2, \dots$ una successione di variabili aleatorie *indipendenti e identicamente distribuite* (i.i.d.), tutte con media $E[X_i] = \mu$. Allora per ogni $\varepsilon > 0$,

$$\mathcal{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \to 0 \quad \text{quando } n \to \infty$$

Dim.: Sappiamo che $E\left(\frac{X_1+\cdots+X_n}{n}\right) = \mu$ e $D^2\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{\sigma^2}{n}$. Applicando Chebyshev su $\frac{X_1+\cdots+X_n}{n}$ risulta $\mathcal{P}\left(\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$. Ma allora il $\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \varepsilon\right) = \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$.

Enunciato e utilizzo del Teorema centrale di convergenza.

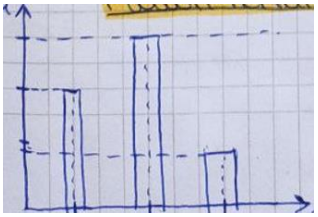
Siano $X_1, \dots, X_n$ delle variabili aleatorie indipendenti ed identicamente distribuite tutte con media μ e varianza σ^2. Se n è grande segue che $X_1 + \cdots + X_n$ ha media $n\mu$ e varianza $n\sigma^2$.

Normalizziamo la somma precedente; quindi, ho una distribuzione approssimativamente normale standard: $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \sim N(0,1)$. Per n grande vale $\forall x : \mathcal{P}\left(\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}} < x\right) \approx \phi(X)$ (formulazione di ripartizione della normale standard).

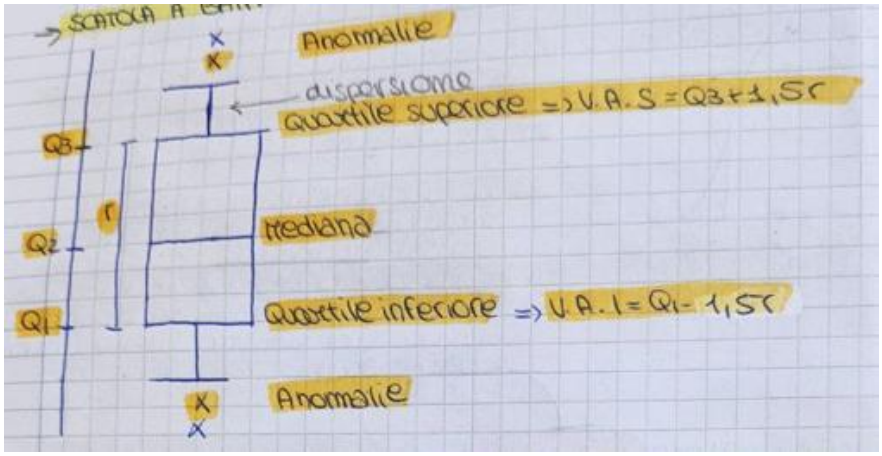
Trova diretta applicazione con le variabili aleatorie binomiali: Sia X di parametri (n, p) il numero di successi in n prove indipendenti con probabilità p di riuscita. Sia ha $X = X_1 + X_2 + \cdots X_n$ con $X_i = 1$ se l'i-esima prova ha successo, $X_i = 0$ altrimenti. Poiché $E[X_i] = p$ e $D^2[X_i] = p(1-p)$ si ha $\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$.

La descrizione dell'istogramma e del diagramma "scatola e baffi".

Istogrammi: mostrano la forma dei dati. L'asse orizzontale rappresenta i valori dei dati e ciascuna barra include un range di valori: l'asse verticale indica quanti sono i punti nei dati ad avere i valori nel range specificati dalla barra.



Scatola a baffi:



- $r = Q_3 - Q_1$: misura la dispersione della distribuzione (da forma alla distribuzione)
- I baffi sono gli intervalli che hanno valori minori di Q_1 e maggiori di Q_3 (punti estremi = valori adiacenti)
- $Q_3 + 1{,}5r$ è il valore più grande tra le osservazioni
- $Q_1 - 1{,}5r$ è il valore più piccolo tra le osservazioni
- I valori adiacenti danno informazioni sulla dispersione e sulla forma della dispersione (se coincidono con gli estremi allora non ci sono anomalia)

La scatola a baffi rappresenta la distribuzione di un campione tramite semplici indici di dispersione e di posizione.

I centri e le medie analitiche.

Sia $\bar{y} = (y_1, \dots, y_n)$ una rilevazioni dati su un criterio quantitativo. Per avere una media analitica bisogna specificare un criterio; ovvero, una *funzione di circostanza*.

Indico con $\bar{y}^* = (y, \dots, y)$ una falsa rilevazione dati di taglia n avente tutti gli elementi uguali a y, allora la risoluzione della media di y è data da: $C(y_1, \dots, y_n) = C(y, \dots, y)$ (la valutazione della funzione di circostanza C su $\bar{y}$ deve coincidere con la valutazione di C su $\bar{y}^*$.

Media aritmetica: $C(y_1, y_2, \dots y_n) = y_1 + y_2 + \cdots + y_n \Rightarrow C(y_1, y_2, \dots y_n) = C(y, \dots, y) = ny$:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{N} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

Media geometrica: $C(y_1, y_2, \dots y_n) = y_1 \cdot y_2 \cdot \dots \cdot y_n \Rightarrow C(y_1, y_2, \dots y_n) = C(y, \dots, y) = y^n$:

$$M_g = \sqrt[n]{y_1 \cdot y_2 \cdot \dots \cdot y_n} = \sqrt[n]{\prod_{i=1}^{n} y_i}$$

Media armonica: $C(y_1, y_2, \dots y_n) = \frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_n} \Rightarrow C(y_1, y_2, \dots y_n) = C(y, \dots, y) = \frac{n}{y}$:

$$M_a = \frac{n}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_n}} = \left(\frac{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_n}}{n}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{y_i}\right)^{-1}$$

Il concetto di *centro* è importante in quanto essi vanno a riassumere tramite dei valori il fenomeno descritto nella distribuzione. Si vuole trovare una funzione che rappresenta la distanza di x da tutti i punti della sua rilevazione dati.

$\forall x \in \mathbb{R}, d(x, \bar{y}) \geq 0$ (funzione che rappresenta la distribuzione x dei dati $\bar{y}$.

Centro di una rilevazione dati $\bar{y}$: $d_r(x, \bar{y}) = \begin{cases} \sqrt[n]{\frac{1}{N}\sum_{i=1}^{N} |x - y_i|^r} & \text{se } r > 0 \\ \sum_{i=1}^{N} |x - y_i| & \text{se } r = 0 \end{cases}$

$\forall r \in \mathbb{R}, \quad \xi_r(\bar{y}) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, d_r(x, \bar{y})$

- Il centro di ordine 0 della rilevazione dati $\bar{y}$, ovvero $\xi_0(\bar{y})$, coincide con la *moda*
- Il centro di ordine 1 della rilevazione dati $\bar{y}$, ovvero $\xi_1(\bar{y})$, coincide con la *mediana*
- Il centro di ordine 2 della rilevazione dati $\bar{y}$, ovvero $\xi_2(\bar{y})$, coincide con la *media* aritmetica

Stimatori e loro proprietà.

Uno stimatore è una funzione che associa ad ogni possibile campione un valore del parametro da stimare. È una funzione di un campione di dati estratti casualmente di una popolazione.

Proprietà di correttezza: lo stimatore $T = C(x_1, \dots, x_n)$ di $\varphi(\theta)$ è *corretto* quando $E^{\theta}[T] = \varphi(\theta)$. Se non è corretto allora è *distorto* con indice di distorsione $d(\theta) = E^{\theta}(T) - \varphi(\theta)$

Proprietà di consistenza: uno stimatore T è *consistente* quando $\varepsilon > 0, \lim_{n} \mathcal{P}(|T - \varphi(\theta)| < 0) = 1$

Confronto di stimatori in base al rischio quadratico medio.

Confrontando due stimatori T e S per $\varphi(\theta)$ indico il rischio quadratico medio la funzione $\theta \in \Theta, R_T(\theta) = E\{[T - \varphi(\theta)]^2\}$ con $[T - \varphi(\theta)]^2$ variabile aleatoria data dalla somma della varianza dello stimatore con il quadrato della sua distorsione.

T e S sono stimatori per $\varphi(\theta)$ se

1) S è *preferibile* a $T \Leftrightarrow R_S(\theta) \leq R_T(\theta)$
2) S è *strettamente preferibile* a $T \Leftrightarrow R_S(\theta) < R_T(\theta)$
3) S è *ammissibile* per la stima di $\varphi(\theta)$ quando non esiste altro stimatore a lui strettamente preferibile a T

Descrizione e fondamento teorico del metodo dei momenti.

La legge di probabilità di un numero aleatorio X è caratterizzata da uno o più parametri. Ogni momento teorico di X, se esistente, è funzione dei parametri. Nel caso in cui il numero aleatorio X sia dotato di funzione geometrica dei momenti allora si può affermare che esistono finiti sia il momento che il momento centrale di qualsiasi ordine di X.

Metodo dei momenti: Siano n, k e $p \geq k$ numeri interi positivi. Si supponga che la genitrice X di un campione casuale semplice $\bar{X} = (X_1, X_2, \dots, X_n)$ abbia la legge di probabilità caratterizzata da k parametri incogniti $\theta_1, \theta_2, \dots, \theta_k$. Allora scelti opportunamente p indici interi positivi, $j_1, j_2, \dots, j_p$ risulta:

$$r = 1,2, \dots, p, \qquad \mu'_{j_r} = f_j(\theta_1, \theta_2, \dots, \theta_k)$$

La precedente equazione rappresenta un sistema di p equazioni nei k parametri incogniti.
L'obiettivo è quello di risolvere tale sistema per ottenere:

$$s = 1,2, \dots, k, \qquad \theta_S = g_s\left(\mu'_{j_1}, \mu'_{j_2}, \dots, \mu'_{j_p}\right)$$

Se si riesce ad ottenere suddetta equazione allora lo stimatore fornito dal metodo dei momenti consiste nella sostituzione dei p momenti teorici considerati con i corrispondenti momenti campionari.
Pertanto, si ha:

$$s = 1,2, \dots, k, \qquad \widehat{\Theta}_{s,MM} \coloneqq g_s\left(\bar{X}^{(j_1)}, \bar{X}^{(j_2)}, \dots, \bar{X}^{(j_p)}\right)$$

Proposizione 1: Per le genitrici dotate di funzioni generatrici dei momenti gli stimatori campionari sono consistenti dei rispettivi momenti teorici

Proposizione 2: Siano consistenti gli stimatori campionari di $\widehat{\Theta}_{s,MM}$. Se per $s = 1, \dots, k$ la funzione g_s è continua allora $\widehat{\Theta}_{s,MM}$ è uno stimatore consistente per θ_S

Descrizione e fondamento teorico del metodo della massima verosimiglianza.

Una qualunque statistica il cui scopo sia quello di dare una stima di un parametro θ si dice *stimatore* di θ; gli stimatori sono quindi variabili aleatorie.
Siano $X_1, \dots, X_n$ variabili aleatorie note a meno di un parametro incognito θ ed andiamo a stimare θ usando i valori che vengono assunti da queste variabili aleatorie.

Sia $f(X_1, \dots, X_n|\theta)$ funzione di massima congiunta di $X_1, \dots, X_n$ oppure la loro densità congiunta a seconda che siano variabili aleatorie discrete o continue.

θ incognita $\Rightarrow$ f dipende da θ
$f(X_1, \dots X_n|\theta)$ *verosimiglianza* $\Rightarrow$ quando θ è il vero valore assunto dal parametro (adotto una stima di θ che rende massima la verosimiglianza)

Dunque: la stima di massima verosimiglianza θ è definita come il valore θ che rende massima la funzione $f(X_1, \dots X_n|\theta)$ (detta funzione *likelihood*) quando i valori assegnati sono $X_1, \dots X_n$.
Calcolare il valore di θ che massimizza f: $f(X_1, \dots X_n|\theta)$ e $\log[f(X_1, \dots X_n|\theta)]$ assumono il massimo in corrispondenza dello stesso valore di θ.

- $\bar{X} = (X_1, X_2, \dots, X_n)$ campione casuale semplice di valori da 0 a $+\infty$
- $\bar{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n) \in (0; +\infty) \times (0; +\infty) \times \dots \times (0; +\infty) \Rightarrow f_X(\bar{\mathcal{X}}) = F_{\mathcal{X}_1}(X_1; \theta) \cdot \dots \cdot F_{\mathcal{X}_n}(X_n; \theta)$
- Funzione di verosimiglianza: $\prod_{i=1}^{n} f_{\mathcal{X}}(\mathcal{X}_i; \theta) = F_{\bar{\mathcal{X}}}(\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n; \theta)$