PROBABILITA' ELEMENTARE E CENNI DI STATISTICA

Anno Accademico 2008-2009

versione del 13/11/2008

Alberto Gandolfi
Dipartimento di Matematica U. Dini,
Università di Firenze,
Viale Morgagni 67/A, 50134 Firenze, Italy
email: gandolfi@math.unifi.it

Indice

Capitolo 1. INTRODUZIONE 4

SPAZI DI PROBABILITÀ E VARIABILI ALEATORIE 7

Capitolo 2. PROBABILITÀ UNIFORMI 8
- 2.1. Spazi di probabilità uniformi 8
- 2.2. Calcolo combinatorio 9
- 2.3. Proprietà delle probabilità uniformi 11
- 2.4. Indipendenza 13
- 2.5. Utilizzo della probabilità 17
- 2.6. Stime di Probabilità e statistica 18
- 2.7. Teorema di De Moivre-Laplace 21

Capitolo 3. PROBABILITÀ SU INSIEMI FINITI 26
- 3.1. Spazi di probabilità su insiemi finiti 26
- 3.2. Indipendenza 27
- 3.3. Probabilità condizionate 28
- 3.4. Catene di Markov 32
- 3.5. Indipendenza condizionale 33
- 3.6. Formula di Bayes 33
- 3.7. Alcuni calcoli di probabilità in genetica 35

Capitolo 4. VARIABILI ALEATORIE FINITE 38
- 4.1. Variabili aleatorie 38
- 4.2. Valore atteso 40
- 4.3. Indipendenza tra variabili aleatorie 45
- 4.4. Deviazioni dalla media 50
- 4.5. Diseguaglianze e legge debole dei grandi numeri 53
- 4.6. Approssimazione di Poisson 56

Capitolo 5. PROBABILITÀ SU INSIEMI DISCRETI 57
- 5.1. Spazi di probabilità su insiemi discreti 57
- 5.2. Variabili aleatorie discrete 59
- 5.3. Vettori aleatori e variabili aleatorie congiunte 63
- 5.4. Covarianza e correlazione 66

Capitolo 6. PROBABILITÀ NEL CONTINUO 69
- 6.1. Variabili aleatorie continue 69
- 6.2. Funzione di distribuzione 71

6.3. Variabili esponenziali, beta e gamma 74
6.4. Trasformazioni di densità per variabili continue 75
6.5. Valori attesi per variabili aleatorie continue 78
6.6. Distribuzione congiunta di due variabili aleatorie 79
6.7. Vettori aleatori continui 84
6.8. Distribuzioni normali multivariate 87
6.9. Trasformazioni di densità di vettori continui 88

Capitolo 7. STATISTICA DESCRITTIVA 90
7.1. L'istogramma 90
7.2. Elementi di Statistica Descrittiva 91

Capitolo 8. STATISTICA INFERENZIALE 94
8.1. Stima di parametri 94
8.2. Il metodo dei momenti 94
8.3. Stima di parametri: il metodo della massima verosimiglianza 96
8.4. Test di ipotesi 96

CAPITOLO 1

INTRODUZIONE

Il calcolo delle probabilità e la statistica costituiscono quella parte della matematica e della scienza più in generale che si occupa dei fenomeni casuali; questo fatto crea talvolta qualche difficoltà nel comprenderne scopi e metodi. Per tentare di ovviare a queste incomprensioni, noi iniziamo qui presentando alcuni problemi, tratti da situazioni concrete del mondo reale, che, opportunamente formalizzati e predisposti per un approccio scientifico utile alla loro trattazione, faranno da guida per lo sviluppo di quasi tutta la teoria che presenteremo. Il lettore è invitato a tentare di fornire qualche risposta, sia inizialmente che durante lo sviluppo della materia, nonché a riferirsi a questi problemi quando stia perdendo di vista il senso di quanto viene discusso. Tutto ciò renderà la teoria molto semplice e naturale.

Esempio 1. *Supponiamo di partecipare ad un gioco in cui viene lanciata, da un addetto, una moneta* 1000 *volte, e supponiamo che noi si vinca 1$ se esce testa e se ne perda 1 se esce croce. Vogliamo valutare la nostra reazione ai vari possibili valori di* k*, il numero di teste sui* 1000 *lanci. Quale reazione sarebbe ragionevole avere se* k *fosse* 510*? E* 492*? Oppure* 459*,* 423*,* 397*,* 354*,* 299*,* 212*,* 154*,* 22*?*

Esempio 2. *Su una scatola di componenti elettronici contenente 100 pezzi è indicato che ognuno ha tempo medio di funzionamento di 1000 ore. Come nell'esempio precedente, proviamo ad immaginare quale sarebbe una reazione ragionevole se il tempo totale di funzionamento dell'intera scatola fosse 99800, oppure 95400 oppure 90200 oppure 84000 oppure 72000 ore.*

Esempio 3. *Supponiamo che in una popolazione vi siano solo due possibili forme di un certo carattere, per semplificare diciamo capelli biondi o castani, determinati dalla trasmissione genetica che ci è ora nota, con biondo carattere recessivo. Se nella popolazione i biondi sono il* 30%*, su* 100 *figli di genitori entrambi castani, quanti ce ne aspettiamo di biondi?*

Problemi come questi si riferiscono ad esperimenti di cui non si riesce a determinare con certezza l'esito. Ci sono di solito varie o anche

molte possibili alternative e le informazioni in nostro possesso non ci permettono di prevedere quale di queste si verificherà.

Non ci interessa discutere qui se questo sia solo un problema di mancanza di informazioni (come ad esempio nei problemi che si riferiscono alla meccanica classica) o se l'indeterminatezza sia intimamente connessa con la natura dell'esperimento (come si assume, per esempio, in meccanica quantistica o in una interpretazione soggettiva della materia qui esposta). Ci interessa soltanto che tale indeterminatezza renda per noi l'esperimento ad esito incerto. Chiameremo tali esperimenti casuali.

Poiché però spesso a noi interessa tentare qualche previsione dell'esito dell'esperimento descriveremo qui un metodo che è stato sviluppato a partire dal XVII secolo (e presumibilmente già nel mondo greco) per suggerirci gli esiti più frequenti. L'idea principale è quella di introdurre una nuova quantità, a cui si dà comunemente il nome di probabilità, che appunto serva come misura di quanto spesso o quanto facilmente si realizza un certo esito di un esperimento casuale.

Naturalmente non importa che tale quantità esista realmente, così come non è molto importante sapere se la lunghezza di un tavolo esiste davvero, quest'ultima cosa essendo difficilmente accertabile anche per il fatto che alcuni atomi si attaccano e si staccano continuamente alle estremità e che queste non sono affatto liscie quando osservate al microscopio. Quello che importa però, sia per le lunghezze che per le probabilità, è che queste quantità siano misurabili con relativa facilità in vari casi interessanti e che si sia poi in grado, sulla base di tali misurazioni, di dire qualcosa di utile per qualche problema di un certo rilievo.

In considerazione di queste finalità pratiche non deve sorprendere che, essendovi vari diversi ambiti di applicazione, vi siano definizioni di probabilità non uguali tra loro e che vi siano discussioni anche accese sui meriti delle varie definizioni; queste però spesso finiscono per essere argomentazioni su quale sia quella più 'vera'. La discussione filosofica è poi ulteriormente complicata dall'ovvia presenza di avvenimenti con esito non determinato a cui non è semplice associare una valutazione di probabilità, e dalla meno ovvia esistenza di altri il cui risultato è ben determinato, ma di complessità tale che conviene inventarsi una probabilità.

Noi presenteremo qui alcune di queste definizioni e poi discuteremo lo sviluppo di una teoria elementare ed alcune applicazioni di rilievo.

La caratteristica principale della probabilità è che, al contrario della lunghezza che si misura attraverso la comparazione fisica con un oggetto campione, essa si misura attraverso la riflessione astratta, usando logica e calcolo matematico. Queste note espongono vari metodi di calcolo di probabilità e varie applicazioni dei risultati.

Essendo un prodotto di calcoli logici, la probabilità verrà misurata tramite numeri puri; spesso nel linguaggio comune essa viene riportata in frazione di 100, ossia in percentuale: così si parla di 30% o 2%. E' però più comodo matematicamente esprimere la probabilità in frazione di 1, indicando quindi $0,3 = 30\%$ oppure $0,02 = 2\%$, principalmente perché la moltiplicazione di probabilità riesce così più comoda: il 2% del 30% è lo $0,6\%$, direttamente ottenibile da $0,02 \cdot 0,3 = 0,006$.

Si noti che anche le frequenze sono espresse tramite percentuali (come fatto nell'esempio 3), ma che si tratta in parte di una coincidenza essendo entrambi numeri puri.

Avendo deciso di misurare le probabilità con valori in $[0, 1]$ si vede che per due casi estremi è possibile determinare subito il valore della probabilità. Agli eventi logicamente impossibili da realizzare è naturale assegnare probabilità 0, mentre a quelli certi va assegnata probabilità 1. A tutte le altre situazioni andrà assegnata una probabilità nell'intervallo $[0, 1]$ chiuso (ciò signfica che non escludiamo l'esistenza di situazioni logicamente possibili ma di probabilità 0). Nei prossimi capitoli cominceremo a vedere vari metodi per tale assegnazione.

Esercizio 1. *Calcolare il* 3% *del* 25%.

Esercizio 2. *Calcolare il* 90% *del* 20% *dell'*80%.

Esercizio 3. *Calcolare l'*80% *del* 120%.

Esercizio 4. *Calcolare, se possibile, la probabilità che una asserzione falsa venga correttamente ritenuta vera.*

Esercizio 5. *Calcolare, se possibile, la probabilità di un evento la cui probabilità sia uguale ad* 1 *meno la probabilità stessa.*

SPAZI DI PROBABILITÀ E VARIABILI ALEATORIE

CAPITOLO 2

PROBABILITÀ UNIFORMI

2.1. Spazi di probabilità uniformi

Iniziamo ora a definire questa nuova quantità, la probabilità, in modo che serva per alcune situazioni semplici, come ad esempio:

Esempio 4. *Nel lancio di una moneta calcolare la probabilità che venga testa.*

Esempio 5. *Nel lancio di un dado calcolare la probabilità che il dado mostri la faccia 3.*

In questi esempi dobbiamo naturalmente fare un piccolo sforzo di astrazione. Assumiamo che il risultato di un lancio sia necessariamente una faccia (non una moneta verticale o la sparizione del dado), e che prima del lancio vi sia stato un adeguato mescolamento. Cosa questo sia non è ben determinato, ma a noi interessa l'esito di tale mescolamento, e cioè che, per quanto ne sappiamo, ognuna delle facce si comporta in modo equivalente a tutte le altre. Ossia, se dobbiamo assegnare una probabilità ad una, dobbiamo assegnare la stessa probabilità alle altre. Avendo già deciso che la probabilità che qualcosa avvenga è 1 ne consegue che per queste situazioni è adeguata la prossima definizione. E' chiaro che il discorso esposto finora è euristico, ossia non rigoroso ma fatto cercando di interpretare la realtà esterna, mentre da ora in poi si inizia a fare matematica partendo da una definizione precisa e sviluppandone le conseguenze. Per decidere a quali situazioni si potrà applicare si ritorna a fare discorsi euristici: a tutte quelle situazioni in cui vi siano un numero finito di alternative equivalenti dal punto di vista probabilistico. In molti casi non si potrà utilizzare questa definizione e ne svilupperemo altre nel seguito.

Visto che si parla di un numero finito di alternative conviene considerare un insieme finito ed adottare quindi la terminologia delle teoria degli insiemi.

Definizione 1. *(Spazi di probabilità uniformi). Sia S un insieme finito. I suoi sottinsiemi $A \subseteq S$ sono detti eventi, e la probabilità uniforme su S è una funzione P definita su ogni evento A da*

$$P(A) = \frac{|A|}{|S|},$$

avendo indicato con $|A|$ il numero di elementi dell'insieme A.

Si dice talvolta che queste probabilità sono definite come rapporto tra il numero di *casi favorevoli* ed il numero di *casi possibili*.

Gli elementi di S sono anche detti eventi elementari e per $s \in S$ utilizzeremo la notazione $P(s) = P(\{s\})$.

Esempio 6. *La probabilità di ottenere due teste lanciando una moneta blu ed una rossa è* $1/4$.

2.2. Calcolo combinatorio

La necessità di stabilire delle cardinalità di vari insiemi ha portato allo sviluppo del *calcolo combinatorio*, le cui formule principali sono le seguenti:

Il numero di campioni ordinati ossia delle k-ple ordinate con ripetizione da n elementi, dette disposizioni con ripetizione, è dato da

$$D_{n,k}^{(r)} = n^k.$$

Il numero dei campioni senza ripetizione ordinati ossia delle k-ple ordinate senza ripetizione da n elementi, dette disposizioni senza ripetizione, è dato da

$$D_{n,k} = n(n-1) \cdot \dots \cdot (n-k+1) = (n)_k.$$

Il numero delle sottopopolazioni, ossia delle k-ple non ordinate senza ripetizione da n elementi, dette combinazioni senza ripetizione, è dato dal coefficiente binomiale

$$C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Il numero delle sottopopolazioni con ripetizione, ossia delle k-ple non ordinate con ripetizione da n elementi, dette combinazioni con ripetizione, è dato da

$$C_{n,k}^{(r)} = \binom{n+k-1}{k-1}.$$

Il numero delle permutazioni di n elementi è dato da

$$P_n = D_{n,n} = n!.$$

Il numero di ripartizioni in sottopopolazioni di $k_1, k_2, \dots, k_r$ elementi di un insieme di n elementi, con $k_1 + k_2 + \dots + k_r = n$, è dato dal coefficiente multinomiale

$$C_{n,(k_1,k_2,\dots,k_r)} = \frac{n!}{k_1!k_2!\dots k_r!}.$$

Esempio 7. *Calcoliamo la probabilità di uscita di un numero (diciamo il* 23*) su una certa ruota nel lotto (in cui si estraggono senza reinserimento* 5 *numeri da* 90*). Se* A *è l'evento che esce il* 23 *Si può prendere come insieme* S *l'insieme delle disposizioni di* 90 *elementi a* 5 *a* 5*, e poi contare i casi favorevoli all'evento. Si ha:*

$$P(A) = \frac{5 \quad D_{89,4}}{D_{90,5}} = \frac{5 \quad 89! \quad 85!}{85! \quad 90!} = \frac{5}{90} = \frac{\binom{89}{4}}{\binom{90}{5}}.$$

Esempio 8. *La probabilità di due volte la faccia* 6 *lanciando due dadi è* 1/36*.*

Osservazione 1. *Ad un'osservazione più accurata, gli eventi descritti nei due esempi precedenti andrebbero meglio specificati; infatti non abbiamo chiarito se i numeri del lotto siano estratti insieme o uno dopo l'altro, così come non abbiamo detto se i dadi siano distinguibili (ad esempio colorati) o totalmente indistinguibili. Questo ha ripercussioni sulla scelta di* S*, per decidere se in esso occorra distinguere l'ordine o meno. Consideriamo ad esempio i due dadi: se li consideriamo distinguibili si avranno* $D^{(r)}_{6,2} = 36$ *possibili disposizioni con ripetizioni altrimenti dovremmo considerare le* $C^{(r)}_{6,2} = 21$ *combinazioni con ripetizione. Non vi è nessun ragionamento elementare che permette di scegliere logicamente tra queste due possibilità e quindi occorre sciogliere l'ambiguità con qualche ragionamento aggiuntivo. In primo luogo, la meccanica suggerisce che la colorazione dei dadi non dovrebbe influenzarne il moto, quindi nulla dovrebbe cambiare tra dadi con colori diversi o uguali; in secondo luogo, l'esperienza di molti esperimenti suggerisce che i due dadi vadano comunque considerati diversi, in modo che* 36 *è la cardinalità appropriata per* S*.*

Queste ed altre osservazioni ci portano a fare un'ipotesi che manterremo per tutto lo sviluppo successivo: **nel mondo macroscopico tutti gli oggetti si comportano come se fossero distinguibili**. *La stessa ipotesi non viene mantenuta nel mondo microscopico, ad esempio dalla statistica di Bose-Einstein, ma di questo non ci occupiamo per ora.*

Una conseguenza di questa ipotesi è che nella teoria per il mondo macroscopico non utilizzeremo le dispozioni con ripetizione.

Un'altra conseguenza è che anche i numeri della tombola o del lotto vanno considerati distinguibili, sia che questi siano estratti insieme sia che siano estratti uno dopo l'altro, e questo è il motivo per cui non è necessario specificare questi dettagli. Tuttavia, quando l'estrazione è senza reinserimento, ossia nel caso delle combinazioni, il numero di configurazioni che viene a coincidere quando si ignora l'ordine è un fattor comune, quindi qui si possono usare indifferentemente combinazioni o dispozioni, a patto di essere consistenti tra numeratore e denominatore (come visto nell'esempio 7).

.

Esempio 9. *In estrazioni senza reinserimento da un'urna con n palline che ne contiene k di un certo tipo calcolare la probabilità che la prima estratta sia del tipo dato e la probabilità che lo sia la seconda estratta. Siano A_1 ed A_2 tali eventi e si noti che nel secondo non abbiamo speficicato quale fosse stata la prima estratta (che quindi possiamo assumere di non aver visto). Naturalmente, con S tale che $|S| = n$ si ha $P(A_1) = k/n$. Inoltre, assumendo che le palline siano numerate da 1 ad n e che le prime k siano del tipo dato e considerando*

$$S = \{(a_1, a_2)|a_i \in \{1, \ldots, n\}, i = 1, \ldots, 2, a_1 \neq a_2\}$$

si ha $|S| = D_{n,2} = n(n-1)$. Infine, $A_i = \{(a_1, a_2) \in S|a_i \in \{1, \ldots, k\}\}$ per cui $|A_i| = |A_2| = k(n-1)$ e $P(A_1) = P(A_2) = k/n$.

E' immediato estendere questo ragionamento per concludere che in estrazioni senza reinserimento (come ovviamente avviene anche in quelle con reinserimento) la probabilità di un evento alla i-sima estrazione se non sono note le precedenti è sempre la stessa.

2.3. Proprietà delle probabilità uniformi

Dalla definizione di probabilità uniforme discendono alcune proprietà elementari, di verifica immediata dalle proprietà delle frazioni:

Lemma 1. *(i) $P(\emptyset) = 0, P(S) = 1$*
(ii) per ogni $A \subseteq S$, $0 \leq P(A) \leq 1$;
(iii) per ogni $s \in S$, $P(s) = \frac{1}{|S|}$;
(iv) per ogni $A \subseteq B \subseteq S$, $P(A) \leq P(B)$.

In taluni casi è più semplice calcolare la probabilità della negazione di un evento A, ossia del suo complemento insiemistico $A^c = S \backslash A$. Si ha

Corollario 1. *(I) Per ogni evento $A \subseteq S$ si ha $P(A^c) = 1 - P(A)$;*
(II) in generale, se $A \subseteq B \subseteq S$ si ha che $P(B \backslash A) = P(B) - P(A)$.

A volte è più agevole dedurre la probabilità di certi eventi da altri per i quali la probabilità si deriva più facilmente. Come si vede dalle proprietà della cardinalità:

Lemma 2. *Per ogni $A, B \subseteq S$ si ha*
(1) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
(2) se $A \cap B = \emptyset$ allora $P(A \cup B) = P(A) + P(B)$
(1) se $A_i \cap A_j = \emptyset$ per ogni $i, j = 1, \ldots, n$, $i \neq j$, allora $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$

Esempio 10. *Ricalcoliamo la probabilità di uscita del* 23 *su una certa ruota nel lotto. Se* A *è l'evento che esce il* 23 *ed* A_i *è l'evento che il* 23 *esce all'i-sima estrazione, per* $i = 1, \ldots, 5$*, si ha che gli* A_i *sono disgiunti e* $P(A_i) = 1/90$*, per cui*

$$P(A) = P(\cup_{i=1}^5 A_i) = \sum_{i=1}^5 P(A_i) = 5/90$$

Per la probabilità dell'unione di n eventi anche non disgiunti si ha:

Lemma 3. *(Formula di inclusione-esclusione) Per ogni* $A_i \subseteq S$*,* $i = 1, \ldots, n$

$$P(\cup_{i=1}^n A_i) = \sum_{k=1}^n \sum_{\{i_1,\ldots,i_k\}\in I_{n,k}} (-1)^{k+1} P(A_{i_1} \cap \cdots \cap A_{i_k}), \tag{2.1}$$

ove $I_{n,k} = \{\{i_1, \ldots, i_k\} | i_j \in \{1, \ldots, n\}$ *per ogni* j, $\quad i_j \neq i_{j'}$ *per* $j \neq j'\}$.

Dimostrazione. Segue dalla parte (1) del Lemma 2 per induzione. Per $n = 2$ essa è equivalente infatti alla tesi. Supponiamo quindi valida la conclusione per ogni famiglia di al più $n-1$ eventi. Di nuovo dalla parte (1) del Lemma 2 e dall'ipotesi di induzione si ha

$$\begin{aligned}
P(\cup_{i=1}^n A_i) &= P(\cup_{i=1}^{n-1} A_i \cup A_n) \\
&= P(\cup_{i=1}^{n-1} A_i) + P(A_n) - P(\cup_{i=1}^{n-1} A_i \cap A_n) \\
&= \sum_{k=1}^{n-1} \sum_{\{i_1,\ldots,i_k\}\in I_{n-1,k}} (-1)^{k+1} P(\cap_{j=1}^k A_{i_j}) + P(A_n) \\
&\quad - \sum_{k'=1}^{n-1} \sum_{\{i_1,\ldots,i_{k'}\}\in I_{n-1,k'}} (-1)^{k'+1} P(\cap_{j=1}^{k'} A_{i_j} \cap A_n) \\
&= \sum_{k=1}^{n} \sum_{\{i_1,\ldots,i_k\}\in I_{n,k}} (-1)^{k+1} P(A_{i_1} \cap \cdots \cap A_{i_k}).
\end{aligned}$$

L'ultima uguaglianza vale avendo posto $k' = k-1$, da cui $-(-1)^{k'+1} = (-1)^{k+1}$, in quanto i termini con $n \notin \{i_1, \ldots, i_k\}$ vengono dalla prima sommatoria, quelli con $k > 1$ ed $n \in \{i_1, \ldots, i_k\}$ vengono dalla seconda ed il termine con $k = 1$ ed $i_1 = n$ è $P(A_n)$. □

Si noti che la dimostrazione precedente è basata unicamente sulla parte (1) del Lemma 2.

Esempio 11. *In 3 lanci di una moneta, se* $A = \{$*esce almeno una testa*$\}$ *e* $A_i = \{$*esce testa all'i-simo lancio*$\}$*, si ha* $A = \cup_{i=1}^3 A_i$*. Si può applicare la* (2.1) *e per questo basta osservare che per* $i, j \in \{1, \ldots, 3\}$ *diversi*

tra loro

$$P(A_i) = 4/8, \quad P(A_i \cap A_j) = 2/8 \text{ e } P(A_1 \cap A_2 \cap A_3) = 1/8$$

come si ottiene facilmente dalla scelta di un appropriato spazio di probabilità finito per ogni evento. Si ottiene quindi $P(A) = 7/8$.

Esempio 12. *Collocando a caso* n *palline numerate da* 1 *ad* n *in* n *urne anch'esse numerate, una per urna, calcoliamo la probabilità che nessuna pallina sia al posto giusto. Indichiamo con* B *tale evento. Si può passare all'evento complementare ed usare la* (2.1). *Osservando poi che se* A_i *è l'evento che l'i-sima pallina è al suo posto allora* $P(\cap_{i=1}^k A_i) = \frac{1}{n(n-1)\dots(n-k+1)} = \frac{(n-k)!}{n!}$ *allora per la probabilità dell'evento* B *è possibile dare una formula esplicita:*

$$P(B) = 1 - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} = 1 - \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!}.$$

2.4. Indipendenza

La teoria della probabilità esposta finora, basata sulla definizione di probabilità uniformi, era costituita essenzialmente di conteggi di cardinalità di insiemi, e, a parte la terminologia, non si discostava molto dalla teoria degli insiemi. C'è però un concetto intuitivo la cui traduzione nell'ambito della teoria le conferisce uno sviluppo automono. Si tratta dell'indipendenza che noi percepiamo tra vari eventi, ad esempio tra i risultati di lanci successivi di una moneta o di un dado (a patto che siano stati opportunamente mescolati tra un lancio e l'altro).

Per capire come inserire tale concetto all'interno della teoria consideriamo un esempio semplice: in due lanci successivi di una moneta il conteggio ci dice che la probabilità di due teste è 1/4, che risulta quindi uguale a 1/2 moltiplicato per 1/2. In altre situazioni che riteniamo indipendenti si verifica la stessa proprietà per cui è naturale porre la definizione seguente. Come al solito, queste erano riflessioni euristiche e da qui comincia la teoria.

Definizione 2. *(i) due eventi* $A, B \subseteq S$ *si dicono indipendenti se*

$$P(A \cap B) = P(A)P(B); \tag{2.2}$$

(ii) n *eventi* $A_i \subseteq S$, $i = 1, \dots, n$ *si dicono (collettivamente) indipendenti se per ogni sottofamiglia* $J \subseteq \{1, \dots, n\}$ *di indici,*

$$P(\cap_{i \in J} A_i) = \prod_{i \in J} P(A_i).$$

Esempio 13. *Nel lancio di un dado, se* A *è l'evento che esce un pari,* B *l'evento che esce un numero minore o uguale a* 2 *e* C *l'evento che esce un numero minore o uguale a* 3, *allora* $P(A) = 1/2$, $P(B) =$

$1/3$, $P(C) = 1/2$ e $P(A \cap B) = P(A \cap C) = 1/6$ per cui A e B sono indipendenti e A e C non lo sono.

Il concetto di indipendenza è però molto utile non quando si deve verificare l'indipendenza dalla definizione, come nell'ultimo esempio, ma quando l'indipendenza si deriva da qualche altra informazione e si utilizza la formula (2.2), ossia $P(A \cap B) = P(A)P(B)$, per calcolare $P(A \cap B)$ dagli altri due valori.

In taluni casi l'indipendenza di certi eventi si può desumere da un risultato teorico, come per l'indipendenza dei complementi nel prossimo teorema, ma più spesso è la situazione che si sta modellizzando che suggerisce l'indipendenza di certi eventi: così per esempio eventi relativi a mazzi di carte mescolati diversi o a diverse estrazioni con reinserimento sono da ritenersi indipendenti e sarebbe utile poter sfruttare questa informazione senza dover verificare l'indipendenza all'interno del modello. In altre parole, sarebbe comodo sapere a priori che esiste uno spazio di probabilità in cui ci sono eventi indipendenti con probabilità qualsiasi, in modo che esso possa fare da modello per la nostra situazione. Dimostreremo un teorema abbastanza generale quando avremo esteso il concetto di probabilità (vedi Teorema 4).

Esempio 14. *Due lanci ripetuti di dado sono indipendenti, quindi se A indica l'uscita di due 3 e A_i, $i = 1, 2$, indica l'uscita del 3 all'i-simo dado, allora $A = A_1 \cap A_2$ e $P(A) = P(A_1)P(A_2) = \frac{1}{36}$.*

Esempio 15. *Anche n lanci ripetuti di dado sono collettivamente indipendenti, quindi se A indica l'uscita del 3 in tutti i dadi e A_i, $i = 1, \ldots, n$, indica l'uscita del 3 all'i-simo dado, allora $P(A) = \prod_{i=1}^{n} P(A_i) = \frac{1}{6^n}$.*

Vediamo ora che l'indipendenza collettiva di eventi implica l'indipendenza collettiva di parte degli eventi con i complementari degli altri.

Teorema 1. *Dati eventi $A_1, \ldots, A_n \subseteq S$ collettivamente indipendenti in uno spazio di probabilità (S, P), indicando con $A_i^1 = A_i$ e con $A_i^0 = A_i^c$, si ha che per ogni $\alpha = (\alpha_1, \ldots, \alpha_n) \in \{0, 1\}^n$, gli eventi $A_1^{\alpha_1}, \ldots, A_n^{\alpha_n}$ sono collettivamente indipendenti.*

Dimostrazione. Una dimostrazione si può fare per induzione su n.

Per $n = 2$ verifichiamo prima che l'affermazione è vera per $\alpha = (1, 0)$, ossia per $A_1 = A_1^1$ e $A_2^c = A_2^0$. Si ha che, essendo $A_1 \cap A_2 \subseteq A_1$, vale

$$\begin{aligned} P(A_1 \cap A_2^c) &= P(A_1 \backslash (A_1 \cap A_2)) \\ &= P(A_1) - P(A_1 \cap A_2) \\ &= P(A_1) - P(A_1)P(A_2) \\ &= P(A_1)(1 - P(A_2)) = P(A_1)P(A_2^c) \end{aligned}$$

ove si è usato che $P(A_1 \cap A_2) = P(A_1)P(A_2)$ per la supposta indipendenza.

Per il resto della dimostrazione basta utilizzare più volte quanto ora verificato. Se scambiamo il ruolo di A_1 e A_2 otteniamo la tesi per $\alpha = (0,1)$, e poi partendo dall'indipendenza di A_1 e A_2^c si ottiene quella per $\alpha = (0,0)$.

Ora supponiamo che la tesi sia vera fino ad $n-1$ e verifichiamola per n. Nella definizione di indipendenza collettiva si considerano anche i sottinsiemi di indici e se $k \leq n-1$ e $\{i_1, \ldots, i_k\} = J \subseteq \{1, \ldots, n\}$ dalla indipendenza collettiva di $A_{i_1}, \ldots, A_{i_k}$ discende, per l'ipotesi di induzione su n anche quella di $A_{i_1}^{\alpha_{i_1}}, \ldots, A_{i_k}^{\alpha_{i_k}}$ per ogni $(\alpha_{i_1}, \ldots, \alpha_{i_k}) \in \{0,1\}^k$. Quindi rimane solo da verificare la fattorizzazione per $A_1^{\alpha_1}, \ldots, A_n^{\alpha_n}$. Si può procedere con una seconda induzione sul numero m di zeri di α. Se $m = 0$ la fattorizzazione vale per ipotesi e supponiamo che valga quando ci sono al più $m-1$ zeri. Si consideri ora $\alpha = (\alpha_1, \ldots, \alpha_n)$ con m zeri e supponiamo, per semplicità di notazione e senza perdita di generalità , che $\alpha_n = 0$; si ha che la famiglia $A_1^{\alpha_1}, \ldots, A_{n-1}^{\alpha_{n-1}}, A_n^1$ è tale che il suo vettore α ha solo $m-1$ zeri e quindi per essa vale la fattorizzazione:

$$\begin{aligned} P(\cap_{i=1}^{n-1} A_i^{\alpha_i} \cap A_n^1) &= \prod_{i=1}^{n-1} P(A_i^{\alpha_i}) P(A_n^1) \\ &= P(\cap_{i=1}^{n-1} A_i^{\alpha_i}) P(A_n^1), \end{aligned}$$

ove la seconda uguaglianza è vera per l'ipotesi di induzione su n. Ma allora $\cap_{i=1}^{n-1} A_i^{\alpha_i}$ e A_n^1 sono indipendenti, per cui, per la verifica fatta per $n = 2$, anche $\cap_{i=1}^{n-1} A_i^{\alpha_i}$ e A_n^0 sono indipendenti. Ne segue che anche per m zeri si ha:

$$\begin{aligned} P(\cap_{i=1}^{n} A_i^{\alpha_i}) &= P(\cap_{i=1}^{n-1} A_i^{\alpha_i} \cap A_n^1) \\ &= P(\cap_{i=1}^{n-1} A_i^{\alpha_i}) P(A_n^1) \\ &= \prod_{i=1}^{n-1} P(A_i^{\alpha_i}) P(A_n^n) \\ &= \prod_{i=1}^{n} P(A_i^{\alpha_i}), \end{aligned}$$

ove la terza uguaglianza segue dall'ipotesi di induzione su n.

□

Esempio 16. *Se A indica l'uscita di almeno una testa in 10 lanci di una moneta ed A_i l'uscita di testa all'i-simo lancio, allora gli A_i sono collettivamente indipendenti e $A = \cup_{i=1}^{10} A_i$; ma gli eventi*

non sono disgiunti e quindi il calcolo con la formula di inclusione-esclusione è piuttosto laborioso. Tuttavia, $A^c = \cap_{i=1}^{10} A_i^c$ *e dall'indipendenza collettiva si ha* $P(A^c) = \prod_{i=1}^{10} P(A_i^c) = (\frac{1}{2})^{10}$*, così che* $P(A) = 1 - 1/2^{10} = 1023/1024$*.*

Facciamo una pausa per riassumere le regole che abbiamo visto per calcolare probabilità. Per calcolare la probabilità di un evento A si può provare a:

(1) contare gli elementi di A e di S;

(2) provare a passare a A^c;

(3) vedere A come unione, ossia $A = \cup_{i=1}^n B_i$ per certi B_i, e poi

(3_1) se B_i sono disgiunti si ha $P(A) = \sum_{i=1}^n P(B_i)$,

(3_2) se B_i non sono disgiunti si può provare ad usare la formula di inclusione-esclusione, in particolare per n piccolo;

(4) vedere A come intersezione, ossia $A = \cap_{i=1}^n B_i$ per certi B_i, e poi

(4_1) se B_i sono collettivamente indipendenti si ha $P(A) = \prod_{i=1}^n P(B_i)$,

(4_2) se B_i non sono collettivamente indipendenti non abbiamo ancora una formula adeguata.

Esercizio 6. *In una vicenda che dette origine a molti di questi calcoli, un incallito giocatore, Chevalier De Méré, scrisse nel 1654 a Pascal chiedendogli se l'uscita di almeno un* 6 *in* 4 *lanci di dado avesse la stessa probabilità dell'uscita di almeno un doppio* 6 *in* 24 *lanci di due dadi. Si racconta che il De Méré avesse perso una grossa somma assumendo che in entrambi i casi si trattasse di 'due probabilità su tre' di vincere. Pascal rispose al De Méré, poi scrisse a sua volta a Fermat e questo fu l'inizio della teoria che stiamo presentando. Calcolare le probabilità di questi eventi.*

Con i metodi precedenti si può anche scrivere in forma esplicita la probabilità che escano esattamente k teste in n lanci di una moneta. Scriviamo ora l'espressione in forma ancora più generale.

Esempio 17. *Supponiamo di effettuare* n *esperimenti indipendenti tali che in ciascuno la probabilità di successo sia un certo valore* $p \in [0,1]$*, ad esempio potrebbe essere* $p = 1/6$ *se per successo intendessimo l'uscita del* 3 *in un dado e così via. Se* $\mathcal{P}(k,n,p)$ *indica la probabilità di esattamente* k *successi su* n *prove indipendenti ognuna con probabilità di successo* p *ed* A_i *indica il successo alla i-sima prova si ha:*

$$\begin{aligned}
\mathcal{P}(k,n,p) &= P(\cup_{J\subseteq\{1,\dots,n\},|J|=k}(\cap_{j\in J} A_j \cap \cap_{j\in J^c} A_j^c)) \\
&= \sum_{J\subseteq\{1,\dots,n\},|J|=k} P(\cap_{j\in J} A_j \cap \cap_{j\in J^c} A_j^c) \\
&= \sum_{J\subseteq\{1,dots,n\},|J|=k} \prod_{j\in J} P(A_j) \prod_{j\in J^c} P(A_j^c) \\
&= \binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}$$

Ci si riferisce a queste probabilità chiamandole distribuzione binomiale o di Bernoulli di parametri n e p od usando il simbolo $B(n, p)$.

Nel caso delle monete siamo interessati alla distribuzione di Bernoulli $B(n, 1/2)$. L'espressione ora ottenuta permette di fare un passo formale verso la soluzione del problema 1 posto all'inizio, ma ancora calcolare queste probabilità nel caso di 1000 monete non sarebbe agevole e non sapremmo nemmeno bene con cosa confrontare i valori così ottenuti.

2.5. Utilizzo della probabilità

Cominciamo dalla seconda questione: abbiamo visto alcuni ragionamenti che conducono al calcolo di probabilità e forse è il momento di tentare una prima connessione tra i valori di tali probabilità e la realtà. Non abbiamo ancora abbastanza strumenti per una discussione approfondita, ma possiamo fare alcune osservazioni che comincino a dare l'idea che quanto abbiamo fatto ha un qualche senso.

1) Non possiamo aspettarci che un evento di probabilità piccola, ad esempio 1% si realizzi di sicuro esattamente una volta su 100, nè che si realizzi di sicuro circa una volta, e nemmeno che si realizzi molto probabilmente esattamente una volta. L'unica cosa che possiamo sperare e che comunque fornisce una certa consistenza a quanto fatto finora è che tale evento si realizzi molto probabilmente circa una volta.

In effetti la probabilità che si realizzi esattamente una volta, data dalla distribuzione di Bernoulli $B(100, \frac{1}{100})$, è

$$\mathcal{P}(1, 100, \frac{1}{100}) = 100\frac{1}{100^1}(\frac{99}{100})^{99} \approx \frac{1}{e} \approx 37\%$$

che è una probabilità intermedia (quindi si tratta di un evento nè nè poco probabile).

Esercizio 7. *Verificare che la probabilità che un evento di probabilità* 1% *si verifichi tra* 0 *e* 3 *volte su* 100 *prove indipendenti è circa* 96%*: se questa è una ragionevole interpretazione del fatto che si realizza 'circa' una volta abbiamo verificato che succede con probabilità abbastanza vicina al al* 100%*.*

Analogamente, eventi di probabilità $1/m$ si verificheranno circa 1 volta ogni m prove indipendenti con alta probabilità.

Alla luce di questo è ragionevole affermare che se individuiamo un evento A *a priori*, ossia prima che sia effettuato l'esperimento che ne verificherà il realizzarsi o meno, e se la probabilità di A è circa $1/m$ con m grande ($m = 50, 100, 1000$), allora sarà sorprendente (tanto più quanto minore è la probabilità) vedere A realizzato già alla prima di queste prove.

Si possono trarre due conclusioni. Innanzitutto, ci sono eventi di probabilità piccola che noi tendiamo a considerare difficilmente realizzabili. In effetti, ci sono costantemente eventi di probabilità talmente

bassa che vengono regolarmente ignorati (tipo l'arrivo di un grande meteorite o che si assembli spontaneamente un essere vivente). Poi si deve osservare che tale giudizio di irrilevanza va fatta non nel senso di irrealizzabilità o meno, ma nel senso di realizzabilità anch'essa con probabilità alta o bassa.

2) D'altra parte se abbiamo 100 eventi disgiunti di probabilità ognuno 1% segue dal Lemma 2 che la probabilità dell'unione è 100%, ossia uno di questi accade di sicuro. Una cosa di questo tipo accade per esempio nella tombola, in cui ogni numero ha probabilità 1/90 ma in ogni estrazione viene estratto un numero, o ancora più evidentemente in una lotteria.

Questa osservazione si può leggere così : *a posteriori*, ossia dopo che un esperimento è stato realizzato, possiamo selezionare eventi che avevano probabilità (a priori!) piccolissima e che si sono realizzati.

Esercizio 8. *Calcolare la probabilità che 10 lanci successivi di un dado (quindi indipendenti) risultino in una successione data di facce, per esempio* (1234565432).

I calcoli a priori vengono definiti **probabilità** ed in questo ambito siamo riusciti a dare un senso ai nostri calcoli, mentre quelli a posteriori (in cui rientrano i calcoli relativi all'esempio 1 sulle monete) vengono definiti **statistica** e per questo la nostra analisi è ancora insufficiente.

Osservazione 2. *Le riflessioni di questo paragrafo sembrano coincidere con una frase di Cicerone, che dice che gettando degli schizzi di colore a caso sul muro sarà molto facile osservare dei tratti che assomiglino ad una faccia, ma sarà assai difficile che questa possa essere quella della Venere di Milo.*

2.6. Stime di Probabilità e statistica

I risultati del capitolo precedente permettono di scrivere formalmente varie probabilità di interesse nel problema delle monete. Per esempio la probabilità di esattamente k teste su 1000 lanci di una moneta è

$$\begin{aligned}\mathcal{P}(k, 1000, 1/2) &= \binom{1000}{k}(\frac{1}{2})^{1000} \\ &= \frac{1000!}{k!(1000-k)!}.\end{aligned}$$

Con un calcolatore oggi è possibile ottenere una buona approssimazione di questo valore, ma nel '700 era necessario cercarne un'approssimazione analitica; oggi questa può comunque essere utile quando il numero di prove sia molto grande oppure per stime rapide quando non si abbia un computer a disposizione.

Il prossimo teorema introduce un'approssimazione analitica del fattoriale, detta di Stirling, esplicitata a meno di una costante che verrà determinata nel seguito.

TEOREMA 2. *Esiste una costante $c \in \mathbb{R}$ tale che per ogni $n \in \mathbb{N}$ si ha*

$$e^{\frac{1}{12(n+1)}} e^c n^{n+1/2} e^{-n} \leq n! \leq e^{\frac{1}{12n}} e^c n^{n+1/2} e^{-n}$$

per cui

$$\lim_{n\to\infty} \frac{n!}{e^c n^{n+1/2} e^{-n}} = 1.$$

DIMOSTRAZIONE. Studiamo

$$d_n = \log(n!) - [(n+1/2)\log(n) - n];$$

si ha

$$d_n - d_{n+1} = (n+1/2)\log(\frac{n+1}{n}) - 1.$$

Da

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \dots$$

e

$$\log(\frac{1}{1-x}) = x + \frac{x^2}{2} + \frac{x^3}{3} + \dots,$$

si ha che

$$\frac{1}{2}\log(\frac{1+x}{1-x}) = x + \frac{x^3}{3} + \frac{x^5}{5} + \dots$$

Per cui, ponendo $\frac{n+1}{n} = \frac{1+\frac{1}{2n+1}}{1-\frac{1}{2n+1}}$ si ha

$$\begin{aligned} d_n - d_{n+1} &= (n+1/2)(\frac{2}{2n+1} + \frac{2}{3(2n+1)^3} + \dots) - 1 \quad (2.3) \\ &= \frac{1}{3(2n+1)^2} + \frac{1}{5(2n+1)^4} + \dots \geq 0. \end{aligned}$$

Dalla (2.3) si ha che

$$\begin{aligned} d_n - d_{n+1} &\leq \sum_{k=1}^{\infty} (\frac{1}{(2n+1)^2})^k \\ &= \frac{1}{(2n+1)^2} \frac{1}{3(1-\frac{1}{(2n+1)^2})} \\ &= \frac{1}{3(4n^2+4n)} \quad (2.4) \\ &= \frac{1}{12n^2+12n} \\ &= \frac{1}{12n} - \frac{1}{12(n+1)}. \end{aligned}$$

Pertanto la successione $d_n - \frac{1}{12n}$ è crescente in n. Poiché da (2.3) la successione d_n è decrescente in n, quindi limitata, si ha che $d_n - \frac{1}{12n}$ è anch'essa limitata, ed essendo crescente, ha limite: esiste $c \in \mathbb{R}$ tale che

$$\begin{aligned} \lim_{n\to\infty} d_n - \frac{1}{12n} &= \lim_{n\to\infty} d_n \\ &= \sup_n d_n - \frac{1}{12n} = c; \end{aligned}$$

inoltre, $d_n \leq c + \frac{1}{12n}$ per ogni $n \in \mathbb{N}$.

D'altra parte, segue da da (2.3) che

$$\begin{aligned} d_n - d_{n+1} &\geq \frac{1}{3(2n+1)^2} \\ &= \frac{1}{12n^2+12n+3} \\ &\geq \frac{1}{12(n+1)} - \frac{1}{12(n+2)}. \end{aligned}$$

essendo $12(n+1)(n+2) \geq 12n^2 + 12n + 3$. Quindi $d_n - \frac{1}{12(n+1)}$ è decrescente e

$$\begin{aligned} c &= \lim_{n\to\infty} d_n - \frac{1}{12(n+1)} \\ &= \inf_n d_n - \frac{1}{12(n+1)}. \end{aligned}$$

Per cui $c + \frac{1}{12(n+1)} \leq d_n$. Passando agli esponenziali di queste quantità si ottiene la tesi. □

Ora possiamo fare qualche stima più esplicita, eccetto che per la costante c.

Esempio 18. *La probabilità di esattamente $n/2$ teste su n lanci di una moneta, assumendo n pari per semplicità, soddisfa:*

$$\begin{aligned} \mathcal{P}(n/2, n, 1/2) &= \frac{n!}{((n/2)!)^2}\frac{1}{2^n} \\ &= \frac{e^c n^{n+1/2} e^{-n}}{e^{2c}(n/2)^n e^{-n\frac{n}{2}}}\frac{1}{2^n} = \frac{2}{\sqrt{n}e^c} \end{aligned}$$

e, tenuto conto degli errori,

$$\mathcal{P}(n/2, n, 1/2) = \frac{2}{\sqrt{n}e^c} e^{S(n)}$$

con $|S(n) \leq \frac{1}{3(n+1)}$ per $n \geq 4$. Il lettore è invitato a verificare questa stima.

Non avendo ancora una stima per $\bar{c} = e^c$ possiamo principalmente concludere per ora che $\mathcal{P}(n/2, n, 1/2)$ tende a 0 in n ed è dell'ordine di $1/\sqrt{n}$.

Per p generico si semplificano altrettanto bene le stime quando $k = np$ è intero:

$$\mathcal{P}(np, n, p) = \frac{n!}{(np)!((1-p)n)!} p^{np}(1-p)^{n(1-p)} \approx \frac{1}{\sqrt{n}e^c\sqrt{p(1-p)}}$$

Le stime per altri varori di k non si semplificano altrettanto bene, ma si può almeno confrontare i vari termini con quelli che abbiamo appena stimato, che per motivi che vedremo vengono definiti 'centrali'.

Lemma 4. *Per ogni $n \in \mathbb{N}$ e per ogni $p \in [0,1]$ $\mathcal{P}(k, n, p)$ è strettamente crescente in k per $k < p(n+1) - 1$, è strettamente decrescente per $k > p(n+1) - 1$, così che*

$$\arg\max_k \mathcal{P}(k, n, p) = [p(n+1) - 1, p(n+1)] \cap \mathbb{N}$$

Dimostrazione. Si ha

$$\begin{aligned} \mathcal{P}(k+1, n, p) &= \binom{n}{k+1} p^{k+1}(1-p)^{n-k-1} \\ &= \frac{n-k}{k+1}\frac{p}{1-p}\mathcal{P}(k, n, p) \end{aligned}$$

quindi la monotonia di $\mathcal{P}(k, n, p)$ dipende da $\frac{n-k}{k+1}\frac{p}{1-p}$. Le monotonie discendono dal fatto che $\frac{n-k}{k+1}\frac{p}{1-p} > 1$, e questo a sua volta vale se e solo se $p(n+1)-1 > k$. In particolare, se $p(n+1) \in \mathbb{N}$, $\mathcal{P}(p(n+1)-1, n, p) = \mathcal{P}(p(n+1), n, p)$ così che $\arg\max_k \mathcal{P}(k, n, p) = \{p(n+1)-1, p(n+1)\}$, come richiesto. □

Esempio 19. *Nel caso delle monete l'andamento di $\mathcal{P}(k, n, 1/2)$ in k dipende solo dai coefficienti binomiali;* $\arg\max_k \mathcal{P}(k, n, 1/2)$ *dipende dalla parità di n e vale $n/2$ oppure $\{(n+1)/2 - 1, (n+1)/2\}$.*

2.7. Teorema di De Moivre-Laplace

Abbiamo visto che il termine centrale, ossia più probabile, di una distribuzione binomiale relativa ad n prove vale circa $c/\sqrt{n}$. poiché i termini vicini non saranno molto diversi con un numero di termini dell'ordine di $\sqrt{n}$ si otterrà una probabilità quasi piena. Questa osservazione è specificata molto meglio e resa rigorosa nel seguente teorema, che enunciamo e dimostriamo solo nel caso particolare di $p = 1/2$. Si utlizzerà la funzione gaussiana $e^{-x^2/2}$ definita per $x \in \mathbb{R}$.

Teorema 3 (Teorema di De Moivre-Laplace per $p = 1/2$). *Per ogni $a_1, a_2 \in \mathbb{R}$, $0 \le a_1 \le a_2$ si ha che*

$$Q_n(a_1, a_2) = \sum_{n/2 + \frac{a_1}{2}\sqrt{n} \le j \le n/2 + \frac{a_2}{2}\sqrt{n}} \mathcal{P}(j, n, 1/2)$$

soddisfa

$$\lim_{n\to\infty} Q_n(a_1,a_2) = \int_{a_1}^{a_2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Dimostrazione. Per semplicità esponiamo la dimostrazione per n pari; se n è dispari la dimostrazione richiede solo modifiche irrilevanti.

Poniamo $n = 2\nu$ e, per ogni $k = 0, 1, 2, \ldots$

$$\begin{aligned}
\alpha_k = \mathcal{P}(\nu+k, 2\nu, 1/2) &= \binom{2\nu}{\nu+k}\frac{1}{2^{2\nu}} \\
&= \frac{(2\nu)!}{(\nu+k)!(\nu-k)!}\frac{1}{2^{2\nu}} \\
&= \frac{(2\nu)!}{(\nu+k-1)!(\nu-k+1)!}\frac{1}{2^{2\nu}}\frac{\nu-k+1}{\nu+k} \\
&= \alpha_{k-1}\frac{\nu-k+1}{\nu+k} \\
&= \alpha_0\frac{(\nu-k+1)(\nu-k+2)\ldots\nu}{(\nu+k)(\nu+k-1)\ldots(\nu+1)} \\
&= \alpha_0\frac{(1-\frac{k-1}{\nu})(1-\frac{k-2}{\nu})\ldots 1}{(1+\frac{k}{\nu})(1+\frac{k-1}{\nu})\ldots(1+\frac{1}{\nu})}.
\end{aligned}$$

Ricordiamo che

$$\log(1+x) = x - x^2/2 + x^3/3 + \cdots = x + R(x)$$

con

$$|R(x)| \le x^2/2 + x^3/3 + \cdots \le 1/2\sum_{k=2}^{\infty} x^k = \frac{x^2}{2(1-x)} < x^2$$

per $2(1-x) > 1$, ossia per $x < 1/2$.

Quindi $1 + x = e^{x+R(x)}$, per cui

$$\begin{aligned}
\alpha_k &= \alpha_0\frac{e^{-(\frac{k-1}{\nu}+\cdots+\frac{1}{\nu})}}{e^{(\frac{k}{\nu}+\cdots+\frac{1}{\nu})}}e^{\bar{R}(k)} \\
&= \alpha_0 e^{-\frac{k^2}{\nu}}e^{\bar{R}(k)} \\
&= \alpha_0 e^{-\frac{2k^2}{n}}e^{\bar{R}(k)}, \qquad (2.5)
\end{aligned}$$

con $|\bar{R}(k)| \le 2kR(\frac{k}{\nu}) \le 2\frac{k^3}{\nu^2} = 2\frac{2^3k^3}{n^2}$. Dalla formula di Stirling

$$\alpha_0 = \frac{2}{\sqrt{n}e^c}e^{S(n)}$$

con $|S(n)| \le \frac{1}{4(n+1)}$.

Risulta quindi che $\alpha_k \approx c_1 e^{-c_2 k}$, con c_1 e c_2 costanti, così che possiamo stimarlo tramite l'integrale della funzione $c_1 e^{-c_2 x}$. Per far ciò si

noti che la funzione $e^{-\frac{x^2}{2}}$ è decrescente per $x \geq 0$. Dalla monotonia discende che per $k \geq 1$:

$$\int_{k-1}^{k} e^{-\frac{x^2}{\nu}} dx \geq e^{-\frac{k^2}{\nu}} \geq \int_{k}^{k+1} e^{-\frac{x^2}{\nu}} dx$$

. Per cui

$$\begin{aligned} Q_n(a_1, a_2) &= \sum_{n/2+\frac{a_1}{2}\sqrt{n} \leq j \leq n/2+\frac{a_2}{2}\sqrt{n}} \mathcal{P}(j, n, 1/2) \\ &= \sum_{\frac{a_1}{2}\sqrt{n} \leq k \leq \frac{a_2}{2}\sqrt{n}} \mathcal{P}(n/2+k, n, 1/2) \\ &\geq \frac{2}{\sqrt{n}} \int_{\lceil \frac{a_1}{2}\sqrt{n} \rceil}^{\lfloor \frac{a_2}{2}\sqrt{n} \rfloor + 1} e^{-c} e^{-\frac{2x^2}{n}} dx e^{S(n)} e^{-\max|\bar{R}(k)|} \end{aligned}$$

in cui $\lfloor t \rfloor$ indica la parte intera di t, $\lceil t \rceil = \lfloor t \rfloor + 1$, e il massimo è preso in k con $\frac{a_1}{2}\sqrt{n} \leq k \leq \frac{a_2}{2}\sqrt{n}$. poiché l'integrando è positivo si può stimare restringendo ulteriormente l'intervallo di integrazione a $[\frac{a_1}{2}\sqrt{n}+1, \frac{a_2}{2}\sqrt{n}]$ se $\frac{a_1}{2}\sqrt{n}+1 \leq \frac{a_2}{2}\sqrt{n}$, oppure all'insieme vuoto. Per ottenere la funzione $e^{-y^2/2}$ poniamo $y = \frac{2x}{\sqrt{n}}$ ottenendo

$$\begin{aligned} Q_n(a_1, a_2) &\geq \left(\int_{a_1}^{a_2} e^{-c} e^{-\frac{y^2}{2}} dy - \int_{a_1}^{a_1+2/\sqrt{n}} e^{-c} e^{-\frac{y^2}{2}} dy \right) e^{-\max|\bar{R}(k)|+S(n)} \\ &\geq \left(\int_{a_1}^{a_2} e^{-c} e^{-\frac{y^2}{2}} dy - e^{-a_1^2} e^{-c} \frac{2}{\sqrt{n}} \right) e^{\frac{-a_2^3}{\sqrt{n}} - \frac{1}{4(n+1)}} \end{aligned}$$

essendo

$$\max|\bar{R}(k)| \leq \max_{\frac{a_1}{2}\sqrt{n} \leq k \leq \frac{a_2}{2}\sqrt{n}} \frac{2^3 k^3}{n^2} = \frac{a_2^3}{\sqrt{n}}.$$

Nella corrispondente stima dall'alto occorre separare il termine centrale, che non è dominato dall'integrale se $j = n/2$:

$$\begin{aligned} Q_n(a_1, a_2) &\leq \left(\int_{\lfloor a_1\sqrt{n} \rfloor - 1}^{\lfloor a_2\sqrt{n} \rfloor} e^{-c} e^{-\frac{x^2}{\nu}} e^{\max|\bar{R}(k)|} dx + 1 \right) e^{-c} \frac{2}{\sqrt{n}} e^{-\frac{1}{4(n+1)}} \\ &\leq \left(e^{\frac{a_2^3}{\sqrt{n}}} \int_{a_1 - \frac{2}{\sqrt{n}}}^{a_2} e^{-c} e^{-\frac{y^2}{2}} dy + e^{-c} \frac{2}{\sqrt{n}} \right) e^{-\frac{1}{4(n+1)}}. \end{aligned}$$

Prendendo il limite per n che diverge si ottiene che

$$\lim_{n\to\infty} Q_n(a_1, a_2) = \int_{a_1}^{a_2} e^{-c} e^{-x^2/2} dx. \tag{2.6}$$

Rimane da determinare il valore di e^c. Si verificherà in modo probabilistico nel seguito, utilizzando il risultato (2.6) qui dimostrato, che $e^c = \sqrt{2\pi}$, da cui la tesi.

□

Osservazione 3. *Lo stesso risultato vale per $a_1 < a_2 < 0$ per la simmetria della funzione gaussiana e della distribuzione di Bernoulli per $p = 1/2$. Per cui, per l'additività dell'integrale rispetto al dominio di integrazione, il risultato di De Moivre Laplace vale per qualsiasi $a_1, a_2 \in \mathbb{R}$ con $a_1 \leq a_2$. Poiché $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$ segue per simmetria che $\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-c} e^{-x^2/2} dx = \frac{1}{2} = \sum_{n/2 \leq j} \mathcal{P}(j, n, 1/2) + \frac{c}{\sqrt{n}}$ per qualche $c \in \mathbb{R}$. Quindi*

$$\begin{aligned}
\lim_{n\to\infty} Q_n(a, +\infty) &= \lim_{n\to\infty} \sum_{n/2+a/2\sqrt{n} \leq j} \mathcal{P}(j, n, 1/2) \\
&= \lim_{n\to\infty} 1/2 - \frac{c}{\sqrt{n}} - \sum_{n/2 \leq j \leq n/2+a/2\sqrt{n}} \mathcal{P}(j, n, 1/2) \\
&= 1/2 - \int_0^a \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= \int_a^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx
\end{aligned}$$

Essendo valido questo ragionamento anche per valori negativi, si conclude che il Teorema di De Moivre Laplace vale per qualsiasi a_1, a_2 in $\bar{\mathbb{R}}$ esteso.

Osservazione 4. *Si noti che nella dimostrazione del teorema si è ottenuto non solo il limite richiesto, ma anche delle stime dall'alto e dal basso di $Q_n(a_1, a_2)$. Gli errori sono però solitamente piccoli, dell'ordine di $e^{\frac{a_2^3}{\sqrt{n}}}$ moltiplicativamente e $\frac{2}{\sqrt{2\pi n}}$ additivamente che in pratica è possibile trascurarli senza alterare sensibilmente l'approssimazione di Q. Pertanto abitualmente, occasionalmente senza prestare sufficiente cura, si approssima $Q_n(a_1, a_2)$ con $\int_{a_1}^{a_2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.*

Questa tecnica ci permette quindi di approssimare i valori relativi alla probabilità che il numero di successi in prove indipendenti con probabilità di successo 1/2 sia compresa in un certo intervallo di valori, che, come ora vedremo, può suggerire una risposta molto precisa al problema iniziale sulle monete.

Come si era visto, la probabilità di ogni singolo risultato è piccola se il numero di prove è grande, e quindi non permette di per sè di valutare la plausibilità di quel numero di teste. Invece la probabilità di un numero di teste minore o uguale ad un dato valore $\bar{k}$ cambia drasticamente, essendo attorno ad 1/2 per $\bar{k}$ vicino ad $n/2$ e vicino a 0 per $\bar{k}$ molto piccolo. Quindi questa probabilità ci può permettere di distinguere tra i risultati plausibili e quelli che non lo sono.

Esempio 20. *Se si sono ottenute 460 teste su 1000 lanci di una moneta, la probabilità di ottenere un numero di testi minore o uguale a quello ottenuto è stimabile come segue. Se S_{1000} è il numero di teste*

ottenute in 1000 monete, vogliamo stimare

$$P(S1000 \leq 460) = Q_n(-\infty, a)$$

per un certo a*. La determinazione di* a *va fatta imponendo* $1000/2 + \frac{a}{2}\sqrt{1000} = 460$*, ossia* $a = \frac{460-1000/2}{\sqrt{1000}/2} \approx -2,53$*. Ora si puo' ottenere dalle tavole della funzione gaussiana l'approssimazione*

$$P(S1000 \leq 460) \approx \int_{-\infty}^{-2,53} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 0,006 = 0,6\%.$$

Se ne conclude che 460 *è un numero di teste assolutamente inaccettabile su* 1000 *lanci di una moneta e che già per questo numero di teste avremmo dovuto considerare l'esperimento come truccato!*

Osservazione 5. *E' chiaro che per decidere delle nostre reazioni in base ai calcoli precedenti dovremmo fissare delle soglie sotto le quali consideriamo una probabilità troppo piccola e quindi il risultato anomalo. Storicamente, allo scopo di scrivere delle tavole utilizzabili in molte situazioni, si sono considerate i valori del* 5% *come soglia per il dubbio sulla correttezza di un esperimento e dell'*1% *per la ragionevole assunzione che l'esperimento non è corretto*

Esercizio 9. *Verificare che su* 1000 *monete avremmo dovuto avere dubbi (soglia* 5%*) con meno di* 474 *teste e ragionevole certezza dell'anomalia del risultato (soglia* 1%*) con meno di* 463 *teste.*

CAPITOLO 3

PROBABILITÀ SU INSIEMI FINITI

3.1. Spazi di probabilità su insiemi finiti

Nel caso delle probabilità uniformi ci trovavamo di fronte ad alternative tutte equivalenti. Tuttavia negli calcoli successivi abbiamo preso un valore p qualsiasi ed abbiamo ricavato formule (come la distribuzione di Bernoulli) in dipendenza di questo p.

Viene quindi naturale di considerare delle probabilità che non vengano da conteggi di insiemi, ma siano semplicemente dei valori in $[0, 1]$ soddisfacenti a certe regole, anche perché così si possono fare modelli per situazioni in cui non ci sono elementi da contare (tipo la probabilità che un tiro faccia centro o che una misura ecceda di una certa frazione il valore vero).

Rimaniamo comunque per ora su un insieme finito e richiediamo per la probabilità che soddisfi alcune delle proprietà che abbiamo verificato essere vere nel caso uniforme. In particolare, scorrendo quanto verificato nel caso uniforme, si vede che la proprietà principale da cui derivano tutte le altre è la (2) del lemma 2, ossia quella secondo cui se $A \cap B = \emptyset$ allora $P(A \cup B) = P(A) + P(B)$. Assumiamola quindi come ipotesi:

Definizione 3. *Dato un insieme finito S si dice probabilità (finita) su S ogni funzione P definita sui sottinsiemi di S tale che:*
(1) $P(S) = 1$
(2) per ogni $A \subseteq S$, $P(A) \in [0, 1]$;
(3) se $A \cap B = \emptyset$ allora $P(A \cup B) = P(A) + P(B)$.

Si possono ricavare tutte le proprietà viste per le probabilità uniformi valgono anche per le probabilità finite. Così in particolare vale la formula della distribuzione di Bernoulli, adesso definita per ogni p anche non razionale.

Anche se la definizione precedente sembra assai generale i modelli che determina sono facilmente identificabili:

Esercizio 10. *Verificare che P è una probabilità su un insieme finito S se e solo se esiste una funzione non negativa f su S tale che per ogni evento $A \subseteq S$, $P(A) = \frac{\sum_{s \in S} f(s)}{\sum_{s \in S} f(s)}$.*

Osservazione 6. *Finora abbiamo visto quindi due definizioni: prima la probabilità come* **rapporto tra numero di casi favorevoli e**

possibili, *ed ora una definizione* **assiomatica** *di probabilità. Discuteremo di altre due definizioni nel proseguimento.*

3.2. Indipendenza

Il prossimo risultato garantisce l'esistenza di uno spazio in cui si possono trovare eventi indipendenti di probabilità assegnata. Questo non garantisce l'esistenza contemporanea di altri eventi di probabilità data, ma già è sufficiente per molte situazioni.

Teorema 4. *(Esistenza di eventi collettivamente indipendenti). Dati valori $p_1, \ldots, p_n \in [0,1]$ esiste uno spazio di probabilità (S, P) ed eventi $A_1, \ldots, A_n \subseteq S$ collettivamente indipendenti e tali che $P(A_i) = p_i$.*

Dimostrazione. Consideriamo

$$S = \{(\alpha_1, \ldots, \alpha_n) : \alpha_j \in \{0,1\}, j = 1, \ldots, n\}$$

e poniamo

$$A_i = \{(\alpha_1, \ldots, \alpha_n) \in S : \alpha_i = 1\}$$

e

$$P((\alpha_1, \ldots, \alpha_n)) = \prod_{j=1}^{n} p_j^{\alpha_j}(1-p_j)^{1-\alpha_j}.$$

Chiaramente $P \geq 0$ e

$$\sum_{(\alpha_1,\ldots,\alpha_n)\in S} P((\alpha_1, \ldots, \alpha_n)) = \sum_{(\alpha_1,\ldots,\alpha_n)\in S} \prod_{j=1}^{n} p_j = 1$$

per cui segue dall'esercizio 10 che (S, P) è uno spazio di probabilità. Inoltre per ogni $J \subseteq \{1, \ldots, n\}$

$$P(\cap_{i\in J} A_i) = \sum_{(\alpha_1,\ldots,\alpha_n)\in \cap_{i\in J} A_i} \prod_{j=1}^{n} p_j = \prod_{i\in J} p_i$$

per cui $P(A_i) = p_i$ e $P(\cap_{i\in J} A_i) = \prod_{i\in J} p_i = \prod_{i\in J} P(A_i)$ come richiesto.

□

Per ulteriori eventi oltre a quelli indipendenti la cui esistenza è garantita dal teorema occorrono naturalmente ulteriori verifiche.

Esempio 21. *Un esempio di uno spazio di probabilità con tre eventi $A_i, i = 1, 2, 3$ collettivamente indipendenti, tali che $P(A_i) = p_i$ ed un evento A_4 che sia incompatibile con $\cap_{i=1}^{3} A_i$ ed indipendente da A_1 si può ottenere ponendo S come sopra con $n = 4$ e, per esempio,*

$$P((\alpha_1, \ldots, \alpha_4)) = \prod_{j=1}^{4} p_j^{\alpha_j}(1-p_j)^{1-\alpha_j}.$$

per $(\alpha_1, \ldots, \alpha_4) \neq (1,1,1,1), (1,1,0,1)$ *e*

$$P(1,1,1,1) = 0 \quad e \quad P(1,1,0,1) = p_1p_2(1-p_3)p_4 + p_1p_2p_3(1-p_4).$$

Può essere $P(A_4) = (1 - p_1p_2p_3)/2$ *?*

3.3. Probabilità condizionate

In alcune situazioni si vuol calcolare la probabilità di un evento quando si sa che un altro si è già realizzato. Un buon sistema per trattare questa situazione è restringere l'ambito dei possibili risultati all'evento che si sa essere realizzato rendendone 1 la probabilità. Questo risulta naturale se l'evento già realizzato aveva probabilità positiva (ché altrimenti ci si immette in una considerazione forse non priva di senso, ma che richiede certamente un dettagliato chiarimento):

Definizione 4. *Dato uno spazio di probabilità* (S, P) *e due eventi* $A, B \subseteq S$*, con* $P(B) \neq 0$*, si dice probablità condizionata di* A *dato* B *il valore*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.7}$$

Esempio 22. *Nel lancio di un dado, se* $A = \{3\}$ *e* $B = \{dispari\}$ *(nello stesso lancio),* $P(A|B) = \frac{1/6}{1/2} = \frac{1}{3}$ *è la probabilità che sia uscito il* 3 *sapendo che è uscito un dispari.*

Come già accaduto per l'indipendenza, talvolta è conveniente 'dedurre' il valore della probabilità condizionata dal problema ed usare (3.7) per ricavare uno degli altri termini. Questo pone gli stessi problemi già discussi trattando dell'indipendenza: se vogliamo procedere in questa direzione dovremmo dimostrare un risultato che garantisca l'esistenza di uno spazio di probabilità e di eventi opportuni aventi le probabilità condizionate, ed eventualmente anche alcune non condizionate, date. Tuttavia qui la situazione è più complessa di quella relativa all'indipendenza e a volte le probabilità, condizionate e non, che si vorrebbero modellizzare possono non essere compatibili, come si vede nell'esempio seguente

Esempio 23. *(da A. H-S. Ang, W. H. Tang, Probability concepts in engineering, Wiley ed., 2nd edition, es. 2.3 pag. 67) Si assume che gli eventi* E_2 *e* E_3*, che nel testo indicano elementi specifici di un processo di produzione, soddisfino:* $P(E_2) = 0.7$*,* $P(E_3) = 0.95$ *e* $P(E_3|E_2^c) = 0.6$*. Ma ciò non è possibile, in quanto risulterebbe* $P(E_3 \cap E_2^c) = P(E_3|E_2^c)P(E_2^c) = 0.18$ *e quindi*

$$P(E_3 \cap E_2) = P(E_3) - P(E_3 \cap E_2^c) = 0.77 > 0.7 = P(E_2)$$

che è una contraddizione.

Un risultato di esistenza, piuttosto intricato, si ottiene invertendo il teorema seguente.

Teorema 5. *Se per gli eventi A_i, $i = 1, \ldots, n$, vale che $P(\cap_{i=1}^k A_i) \neq 0$ per ogni $k = 1, \ldots, n$ allora*

$$P(\cap_{i=1}^n A_i) = \prod_{k=1}^n P(A_k | \cap_{i=1}^{k-1} A_i) \tag{3.8}$$

Dimostrazione. Si ha via via che $P(A_2|A_1)P(A_1) = P(\cap_{i=1}^2 A_i)$ e proseguendo $P(A_k| \cap_{i=1}^{k-1} A_i)P(\cap_{i=1}^{k-1} A_i) = P(\cap_{i=1}^k A_i)$ che per $k = n$ fornisce la tesi. □

Diciamo che una famiglia di insiemi $B_1, \ldots, B_n \subseteq S$ è una partizione di S se

(a) $B_i \cap B_j = \emptyset$ se i è diverso da j;

(b) $\cup_{i=1}^n B_i = S$.

Si noti che se $B_1, \ldots, B_n$ e $C_1, \ldots, C_k$ sono partizioni di S allora la famiglia $B_i \cap C_j$, al variare di i e j, è anch'essa una partizione.

Teorema 6. *(Esistenza di eventi con probabilità condizionate in serie). Se per ogni $j = 1, \ldots, k$ sono dati numeri $p(i_1, i_2, \ldots, i_j) \in [0, 1]$, $1 \leq i_j \leq n$, tali che*

$$\sum_{i_j=1}^n p(i_1, i_2, \ldots, i_j) = 1 \tag{3.9}$$

(per ogni valore degli indici $i_1, \ldots, i_{j-1}$) allora esistono uno spazio di probabilità S e k partizioni $B_1(j), \ldots, B_n(j)$, $j = 1, \ldots, k$ tali che

$$P(B_{i_j}(j)| \cap_{m=1}^{j-1} B_{i_m}(m)) = p(i_1, i_2, \ldots, i_j)$$

per qualsiasi valore di i_m, $m = 1, \ldots, j$.

Dimostrazione. Consideriamo

$$S = \{(\alpha_1, \ldots, \alpha_k) : \alpha_j \in \{1, \ldots, n\}, j = 1, \ldots, k\}$$

e poniamo

$$B_i(j) = \{(\alpha_1, \ldots, \alpha_k) \in S : \alpha_j = i\}$$

e

$$P((\alpha_1, \ldots, \alpha_k)) = \prod_{j=1}^k p(\alpha_1, \ldots, \alpha_j).$$

Chiaramente $P \geq 0$ e

$$\sum_{(\alpha_1,\ldots,\alpha_n)\in S} P((\alpha_1, \ldots, \alpha_n)) = \sum_{(\alpha_1,\ldots,\alpha_n)\in S} \prod_{j=1}^k p(\alpha_1, \ldots, \alpha_j) = 1$$

per la (3.9), per cui segue dall'esercizio 10 che (S, P) è uno spazio di probabilità. Inoltre

$$\begin{aligned} P(B_{i_j}(j)| \cap_{m=1}^{j-1} B_{i_m}(m)) &= \frac{P(\cap_{m=1}^{j} B_{i_m}(m))}{P(\cap_{m=1}^{j-1} B_{i_m}(m))} \\ &= \frac{P((i_1, \ldots, i_j))}{\sum_{i_j=1}^{n} P((i_1, \ldots, i_j))} \\ &= \frac{\prod_{m=1}^{j} p(i_1, \ldots, i_m)}{\sum_{i_j=1}^{n} \prod_{m=1}^{j} p(i_1, \ldots, i_m)} \\ &= \frac{\prod_{m=1}^{j} p(i_1, \ldots, i_m)}{\prod_{m=1}^{j-1} p(i_1, \ldots, i_m)} = p(i_1, \ldots, i_j) \end{aligned}$$

come richiesto. □

Naturalmente questi risultati sono interessanti quando si conoscono le probabilità condizionate nella parte destra di (3.8), e questo è proprio il caso nelle estrazioni successive senza reinserimento.

Esempio 24. *Estraendo senza reinserimento da un'urna contenente* 3 *palline bianche e* 5 *rosse, che indicheremo schematicamente come* $|3B \quad 2R|$, *ed indicando con* B_i *e* R_i *gli eventi che escano una pallina bianca o rossa, rispettivamente, alla i-sima estrazione si ha che* $P(B_1) = 3/8$, $P(R_2|B_1) = 5/7$ *da cui* $P(B_1 \cap R_2) = 18/56$.

Ci sono altre condizioni sulle probabilità condizionate che garantiscono l'esistenza di uno spazio di probabilità soddisfacente quelle condizioni (tipo le cosiddette catene di Markov), ma la loro trattazione richiede un'approfondimento eccessivo per questa trattazione.

Dalla conoscenza di tutte le probabilità condizionate a certi eventi è poi possibile risalire alle probabilità non condizionate degli eventi stessi.

Vale ora il seguente risultato, talvolta detto teorema delle probabilità totali o delle probabilità composte.

Teorema 7 (Teorema delle probabilità totali). *Se* B_i, $i = 1, \ldots, n$, *costituiscono una partizione di* S *con* $P(B_i)$ *diverso da* 0 *per ogni* i, *allora per ogni evento* A *si ha*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{3.10}$$

Dimostrazione. Dalla definizione di probabilità condizionata si ha

$$\begin{aligned}\sum_{i=1}^{n} P(A|B_i)P(B_i) &= \sum_{i=1}^{n} \frac{P(A\cap B_i)}{P(B_i)}P(B_i) \\ &= \sum_{i=1}^{n} P(A\cap B_i) \\ &= P(\cup_{i=1}^{n} A\cap B_i) = P(A) \qquad (3.11)\end{aligned}$$

in cui le due ultime uguaglianze seguono, la prima dal fatto che i B_i sono disgiunti e quindi così sono gli $A\cap B_i$, e la seconda dal fatto che $\cup_{i=1}^{n} B_i = S$.

□

Esempio 25. *Nella situazione e con le notazioni dell'ultimo esempio relativo alle estrazioni da* $|3B \quad 2R|$ *si ha:*

$$P(R_2) = P(R_2|B_1)P(B_1) + P(R_2|R_1)P(R_1) = \frac{5}{7}\cdot\frac{3}{8} + \frac{4}{7}\cdot\frac{5}{8} = \frac{5}{8}.$$

Si noti che $P(R_2) = P(R_1)$*, come già in precedenza argomentato.*

Questo teorema è molto utile, ma a volte bisogna saper trovare la partizione giusta a cui condizionare:

Esempio 26. *In un gioco ci sono tre porte e dietro ad una sola vi è un premio. Un concorrente sceglie una porta senza aprirla, poi dal conduttore gli viene aperta una porta senza premio ed offerta la possibilità di cambiare la sua scelta (necessariamente nell'unica porta finora non menzionata). E' conveniente cambiare o la cosa è irrilevante in quanto comunque aveva scelto una porta a caso?*

Diciamo di giudicare sulla base della probabilità di trovare il premio, che vogliamo massima, e fissiamo due strategie: cambiare o non cambiare (al momento della scelta). Trovandoci ad aver scelto la prima porta, per calcolare le probabilità di trovare il premio la cosa migliore sembra essere scegliere la seguente partizione: se a *è la porta scelta, siano* A_a *l'evento che il premio è in* a *ed* A_a^c *il suo complemento (che necessariamente formano una partizione dell'insieme dei possibili esiti, quali che essi siano). Ora, detto* V *l'evento che trovo il premio, si ha:* $P(V) = P(V|A_a)P(A_a) + P(V|A_a^c)P(A_a^c)$*. Naturalmente, visto che la posizione del premio è casuale (o quantomeno ignota al concorrente)* $P(A_a) = 1/3$*. Nella strategia di cambiare,* $P(V|A_a) = 0$ *e* $P(V|A_a^c) = 1$ *e viceversa nella strategia di non cambiare, per cui quando si cambia* $P(V) = 2/3$ *e non cambiando* $P(V) = 1/3$*. Quindi conviene cambiare.*

Si può usare una combinazione del teorema delle probabilità totali con la definizione di probabilità condizionata per calcolare probabilità condizionate non direttamente evidenti.

Esercizio 11. *Calcolare la probabilità che il secondo numero estratto nella tombola (quindi senza reinserimento) sia il 30 sapendo che il primo estratto è pari.*

3.4. Catene di Markov

E' chiaro quindi che una volta che si riesce a mostrare l'esistenza di spazi di probabilità che permettono di descrivere certe famiglie di probabilità condizionate, questi modelli vengono poi sistematicamente utilizzati. Un modello molto ricco, sia matematicamente che per le molteplici applicazioni, viene dal semplificare la dipendenza descritta nel teorema 6. Invece di assegnare le probabilità condizionali dati un certo numero di altri eventi (a decrescere) ipotizziamo che ci sia un ordine (per esempio temporale) e che la probabilità condizionale di ogni evento dati tutti i precedenti dipenda solo dall'ultimo. Questo è un buon modello per situazioni come il capitale posseduto dopo varie partite ad un gioco o la posizione di una particella che si muova a caso su un insieme finito decidendo ogni volta dove andare indipendentemente dal passato.

Più precisamente, supponiamo di voler trovare uno spazio di probabilità per k partizioni $B_1(j), \dots, B_n(j)$, $j = 1, \dots, k$, tali che

$$P(B_{i_j}(j)| \cap_{m=1}^{j-1} B_{i_m}(m)) = P(B_{i_j}(j)|B_{i_{j-1}}(j-1)).$$

Per semplicità consideriamo il caso in cui le probabilità condizionate dipendono solo dall'indice dell'evento e non dipendono dalla partizione: $P(B_r(j)|B_z(j-1)) = a_{z,r}$. Con gli opportuni vincoli esiste sempre uno spazio di probabilità con queste probabilità condizionali, e si possono anche assegnare le probabilità non condizionali della prima partizione.

Teorema 8. *Data una matrice stocastica $A = (a_{i,j})_{i,j=1,\dots,k}$, ossia una matrice tale che $\sum_{j=1}^k a_{i,j} = 1$ per ogni i, ed un vettore stocastico $\pi = (\pi_1, \dots, \pi_k)$ stocastico, ossia tale che $\sum_{i=1}^k \pi_i = 1$ allora esiste uno spazio di probabilità (S, P) e k partizioni $B_1(j), \dots, B_n(j)$, $j = 1, \dots, k$, tali che*

$$P(B_{i_j}(j)| \cap_{m=1}^{j-1} B_{i_m}(m)) = P(B_{i_j}(j)|B_{i_{j-1}}(j-1)) = a_{i_{j-1},i_j}$$

e

$$P(B_r(1)) = \pi_r.$$

Dimostrazione. Basta considerare $S = \{1, \dots, n\}^k$ e per ogni $s = (s_1, \dots, s_k) \in S$ porre

$$P(s) = \pi_{s_1} a_{s_1,s_2} \cdot \dots \cdot a_{s_{k-1},s_k}$$

definendo $B_r(j) = \{s \in S : s_j = r\}$. Si vede che P è una probabilità e valgono le proprietà richieste (esercizio). $\square$

La probabilità P determinata nel precedente teorema insieme con le relative partizioni è detta catena di Markov di matrice A e vettore iniziale π. La teoria delle catene di Markov è molto interessante per le molteplici applicazioni e per la connessione con le proprietà algebriche, di cui si vede nel prossimo esercizio un esempio.

Esercizio 12. *Dimostrare che in una catena di Markov*

$$P(B_r(j)|B_z(1)) = a_{z,r}^{(j)}$$

dove $a_{z,r}^{(j)}$ *indica l'elemento di posizione* (z, r) *della matrice* A^j *(ossia la matrice alla potenza* j*).*

3.5. Indipendenza condizionale

E' possibile definire anche l'indipendenza di due eventi condizionatamente ad un terzo ponendo:

Definizione 5. *Dati eventi* A, B *e* C *tali che* $P(C) \neq 0$*, si dice che* A *e* B *sono condizionatamente indipendenti dato* C *se*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Esercizio 13. *Se* A *e* B *sono condizionatamente indipendenti dato* C *allora* $P(A|B \cap C) = P(A|C)$.

3.6. Formula di Bayes

La formula presentata in questa sezione permette di tornare indietro, ossia calcolare la probabilità che avevano certi eventi già realizzati (ma del cui esito non abbiamo informazione).

Esempio 27. *Supponiamo di avere due urne,* $|3N \quad 2R|$ *e* $|2N \quad 8R|$*, e di usare il seguente procedimento di estrazione: scelgo un'urna a caso e poi da questa estraggo una pallina a caso. Se la pallina è rossa, qual è la probabilità che l'urna estratta sia la prima?*

In questa situazione conosciamo le probabilità iniziali di scelta dell'urna e poi le probabilità condizionate di prendere una certa pallina scelta l'urna, ma appunto il problema chiede di tornare indietro.

Teorema 9 (Formula di Bayes). *Data una partizione Se* B_i*,* $i = 1, \ldots, n$*, con* $P(B_i) > 0$*, ed un evento* A *sottinsieme di un insieme* S*, con* $P(A) \neq 0$*, si ha*

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \tag{3.12}$$

Dimostrazione. Dalla definizione di probabilità condizionata e dalla formula delle probabilità totali si ha

$$\frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} = \frac{P(A \cap B_1)}{P(A)} = P(B_1|A).$$

□

Esempio 28. *Nell'esempio precedente siano A l'evento che viene scelta l'urna $|3N \quad 2R|$ e B l'evento che viene scelta l'altra urna; A e B formano una partizione. Se R è l'evento che è stata estratta una pallina rossa si risponde al problema determinando $P(A|R)$. E dalla formula di Bayes:*

$$P(A|R) = \frac{\frac{2}{5} \cdot \frac{1}{2}}{\frac{2}{5} \cdot \frac{1}{2} + \frac{5}{10} \cdot \frac{1}{2}} = \frac{4}{9}.$$

Statistica Bayesiana

E' possibile utilizzare la formula di Bayes anche per un'analisi statistica diversa da quelle a cui si è accennato finora. Per distinguere i due approcci quest'ultimo viene detto frequentista o classico (e lo riprenderemo nella seconda parte di queste note) ed il primo, a cui accenneremo ora, Bayesiano.

Supponiamo che si faccia il seguente esperimento: le due scatole degli esempi precedenti sono scelte con probabilità a noi non note p_A e p_B, $p_A + p_B = 1$, poi viene estratta una pallina dalla scatola scelta e ci viene mostrata. L'idea della statistica Bayesiana è quella di determinare in qualche modo delle probabilità iniziali, o a priori, per la scelta di A e B e poi aggiornarle con la formula di Bayes. Visto che non sappiamo nulla del valore di p_A un'ipotesi sensata è di considerare inizialmente le due scatole come equiprobabili, per cui si parte dalla probabilità a priori P_0 tale che $P_0(A) = P_0(B) = \frac{1}{2}$. Poi avendo a disposizione un dato (che la pallina estratta è rossa) si aggiorna la nostra stima delle probabilità in

$$P_1(A) = P_0(A|R) = \frac{4}{9}$$

e $P_1(B) = P_0(B|R) = \frac{5}{9}$. Naturalmente la procedura può essere iterata: se viene ripetuto l'esperimento descritto sopra con le stesse probabilità a noi non note, quindi dopo il reinserimento delle palline estratte, e la pallina mostrata è di nuovo rossa (indichiamo con R_2 l'evento corrispondente) ora si può aggiornare la stima delle probabilità partendo

da P_1:

$$\begin{aligned} P_2(A) &= P_1(A|R_2) \\ &= \frac{P_1(R_2|A)P_1(A)}{P_1(R_2|A)P_1(A) + P_1(R_2|B)P_1(B)} \\ &= \frac{\frac{2}{5}\cdot\frac{4}{9}}{\frac{2}{5}\cdot\frac{4}{9}+\frac{5}{10}\cdot\frac{5}{9}} = \frac{16}{41} \end{aligned}$$

Correttamente $\frac{16}{41} < \frac{4}{9}$ perché l'uscita di un'altra pallina rossa rafforza la convinzione che sia scelta più frequentemente la scatola B che contiene più rosse e che quindi A abbia una probabilità minore di quanto pensavamo prima.

3.7. Alcuni calcoli di probabilità in genetica

Esempio 29. *Qual è la probabilità che nasca un figlio biondo da due genitori castani?*

Per poter procedere la questione deve essere formulata in termini scientifici e saranno poi necessari alcuni dati, che immaginiamo più facilmente reperibili della frequenza di figli biondi da genitori castani. La teoria rivelatasi più adeguata è quella genetica (che qui esponiamo in forma molto semplificata) in cui si assume:

1. ogni individuo ha, relativamente al carattere in oggetto, due alleli;
2. tali alleli possono essere o b per il carattere biondo o C per il carattere castano;
3. b è un allele recessivo, C dominante per cui il fenotipo (ossia la forma espressa del carattere da parte dell'individuo portatore di due alleli è sempre castano salvo quando i due alleli sono entrambi b);
4. (legge di Hardy-Weinberg) se p_b è la frequenza di alleli b in una popolazione (ossia il rapporto tra il numero di loci con allele b rispetto al numero totali di loci, quest'ultimo essendo due volte il numero di individui della popolazione) allora la frequenza di alleli C sarà $p_C = 1-p_b$ e la frequenza di individui con alleli bb sarà $p_{bb} = p_b \cdot p_b$ e di quelli con alleli CC sarà $p_{CC} = p_C \cdot p_C$.

Sulla base di questi elementi, ed osservando che se un individuo è scelto a caso allora la probabilità di un certo evento coincide con la sua frequenza nella popolazione, è possibile determinare molte probabilità relativamente ai genotipi ed ai fenotipi di una popolazione in un dato momento.

Esempio 30. *La probabilità che un individuo sia eterozigote (ossia abbia due alleli diversi) è* $p_{bC} = 1 - p_{bb} - p_{CC} = 2p_bp_C$.

La probabilità che un individuo castano sia eterozigote è data dalla probabilità condizionata che l'individuo sia eterozigote dato che è

castano: indicando con $\bar{C}$ quest'ultimo evento, la probabilità è quindi

$$P(bC|\bar{C}) = \frac{P(bC \cap \bar{C})}{P(\bar{C})} = \frac{2p_b p_C}{1-p_b^2} = \frac{2p_b(1-p_b)}{1-p_b^2} = \frac{2p_b}{1+p_b}.$$

Quest'ultima probabilità si può vedere come una applicazione, in un caso molto semplice, del teorema di Bayes in quanto è immediato desumere le probabilità di un certo fenotipo dato il genotipo (che sono peraltro 0 o 1) e qui si richiede la probabilità condizionata inversa (esercizio).

Tuttavia nel problema 2 dobbiamo anche prendere in considerazione la riproduzione in quanto siamo interessati a questioni di discendenza. Un'ipotesi ragionevole relativamente alla riproduzione è che

5. ogni figlio viene generato scegliendo due genitori scelti a caso (eventualmente lo stesso!) ed il suo genotipo è generato scegliendo a caso un allele da ogni genitore.

Questa ipotesi non è molto realistica per le popolazioni umane, ma sembra più ragionevole per animali inferiori e poi è molto semplice. Prima di procedere osserviamo che c'è però un problema di coerenza tra quest'ultima ipotesi e le precedenti (questo problema fu portato verso il 1920 al matematico Hardy dal biologo Punnet e dette in seguito luogo alla legge che porta il nome di Hardy): se ad una certa generazione valgono le ipotesi 1-4 e la riproduzione segue l'ipotesi 5 si manterranno le condizioni 1-4 anche alla generazione successiva?

Per questa verifica si usa il teorema delle probabilità totali. Se indichiamo con $P_{\alpha\beta}$, $M_{\alpha\beta}$ e $F_{\alpha\beta}$, gli eventi che il padre, la madre o il figlio rispettivamente, hanno genotipo $\alpha\beta$ si ha:

$$\begin{aligned}
P(F_{bb}) &= \sum_{\alpha,\beta,\alpha',\beta'=b,C} P(F_{bb}|P_{\alpha,\beta} \cap M_{\alpha',\beta'}) P(P_{\alpha,\beta} \cap M_{\alpha',\beta'}) \\
&= P(P_{b,b} \cap M_{b,b}) + \frac{1}{2} P(P_{b,C} \cap M_{b,b}) \\
&\qquad + \frac{1}{2} P(P_{b,b} \cap M_{b,C}) + \frac{1}{4} P(P_{b,C} \cap M_{b,C}) \\
&= p_b^4 + 2\frac{1}{2} 2p_b^3(1-p_b) + \frac{1}{4} 4p_b^2(1-p_b)^2 \\
&= p_b^2 = P(P_{bb}). \qquad (3.13)
\end{aligned}$$

La stessa cosa si può verificare per gli altri genotipi, per cui le condizioni 1-4 risultano in effetti stabili rispetto alla riproduzione modellizzata da 5.

Siamo ora in grado di fornire una risposta al problema 2. Se indichiamo con $F_{\bar{b}}$ ($F_{\bar{C}}$) l'evento che il figlio è biondo (o castano) ed analogamente denotiamo gli eventi che il padre o la madre hanno un certo fenotipo, una soluzione al problema viene data dal calcolo di

$$P(F_{\bar{b}}|P_{\bar{C}} \cap M_{\bar{C}}) = P(F_{bb}|P_{\bar{C}} \cap M_{\bar{C}}).$$

Dalla definizione di probabilità condizionate si ha:

$$\begin{aligned} P(F_{bb}|P_{\bar{C}} \cap M_{\bar{C}}) &= \frac{P(F_{bb} \cap P_{\bar{C}} \cap M_{\bar{C}})}{P(P_{\bar{C}} \cap M_{\bar{C}})} \\ &= \frac{\sum_{\alpha,\beta,\alpha',\beta'=b,C \text{ ma non uguali a bb}} P(F_{bb} \cap P_{\alpha,\beta} \cap M_{\alpha',\beta'})}{\sum_{\alpha,\beta,\alpha',\beta'=b,C \text{ ma non uguali a bb}} P(P_{\alpha,\beta} \cap M_{\alpha',\beta'})}. \end{aligned}$$

L'indipendenza nella scelta dei genitori implica che

$$\sum_{\alpha,\beta,\alpha',\beta'=b,C \text{ ma non uguali a bb}} P(P_{\alpha,\beta} \cap M_{\alpha',\beta'}) = (1-p_b^2)^2.$$

Per il calcolo del numeratore possiamo utilizzare di nuovo la definizione di probabilità condizionata, osservando che solo una delle probabilità condizionate è diversa da 0:

$$\begin{aligned} & \sum_{\alpha,\beta,\alpha',\beta'=b,C \text{ ma non uguali a bb}} P(F_{bb} \cap P_{\alpha,\beta} \cap M_{\alpha',\beta'}) \\ = & \sum_{\alpha,\beta,\alpha',\beta'=b,C \text{ ma non uguali a bb}} P(F_{bb}|P_{\alpha,\beta} \cap M_{\alpha',\beta'}) P(P_{\alpha,\beta} \cap M_{\alpha',\beta'}) \\ = & \frac{1}{4} 4(1-p_b)^2 p_b^2. \end{aligned}$$

Per cui si ha

$$P(F_{bb}|P_{\bar{C}} \cap M_{\bar{C}}) = \frac{(1-p_b)^2 p_b^2}{(1-p_b^2)^2} = \frac{p_b^2}{(1+p_b)^2}.$$

Ora rimane solo da determinare p_b. In effetti, sappiamo che la frequanza di (fenotipi) biondi è il 30% e che la probabilità di un genotipo puro bb è p_b^2. Assumendo che questi due valori siano uguali si ha $p_b = \sqrt{0.3}$.

Per cui $P(F_{bb}|P_{\bar{C}} \cap M_{\bar{C}}) = \frac{0,3}{(1+\sqrt{0.3})^2} \approx 0,125$.

Anche se era ovvio che la probabilità di un figlio biondo dovesse essere minore per due genitori castani rispetto a due qualunque genitori, quantificare tale riduzione è un risultato per nulla immediato.

CAPITOLO 4

VARIABILI ALEATORIE FINITE

4.1. Variabili aleatorie

In alcuni problemi si è portati a considerare funzioni definite su uno spazio di probabilità .

Esempio 31. *Il numero di teste in* 100 *lanci di una moneta può essere visto come una funzione* $T_{100} : S \to \mathbb{R}$ *con* $S = \{0,1\}^{100}$ *che indica l'insieme dei possibili risultati del lancio di* 100 *monete, indicando testa con* 1 *e croce con* 0*, definita da*

$$T_{100}(a_1, \ldots, a_{100}) = \sum_{i=1}^{100} a_i;$$

ove su S *si considera la probabilità uniforme.*

Per distinguerla dalla teoria delle funzioni e per ricordare che stiamo parlando di fenomeni casuali, queste funzioni vengono chiamate in modo diverso:

Definizione 6. *Dato uno spazio di probabilità finito* (S, P) *ogni funzione* $X : S \to \mathbb{R}$ *è detta* **variabile aleatoria**.

Le variabili aleatorie costituiscono una generalizzazione del concetto di evento, infatti per ogni evento A in uno spazio di probabilità (S, P) la funzione indicatrice di $\mathbb{I}_A$ di A definita da

$$\mathbb{I}_A(s) = \begin{cases} 1 & \text{se } s \in A \\ 0 & \text{se } s \notin A \end{cases}$$

è una variabile aleatoria.

Spesso delle variabili aleatorie ci interessano i valori che esse possono assumere e le relative probabilità : denotiamo con $S_X = X(S) \subseteq \mathbb{R}$ il codominio di una variabile aleatoria X e per ogni $B \subseteq S_X$ con

$$\begin{aligned} P_X(B) &:= P(s \in S : X(s) \in B) \\ &= P(X^{-1}(B)) \\ &=: P(X \in B) \end{aligned}$$

la probabilità che X assuma valori in B. Conseguentemente, $P_X(x) = P_X(\{x\})$ indica la probabilità che X assuma il valore $x \in S_X$.

Lemma 5. *Per ogni variabile aleatoria finita* X*, la coppia* (S_X, P_X) *è uno spazio di probabilità finito detto* **distribuzione di** X.

Dimostrazione. Si ha:
(i) $P_X(S_X) = P(X^{-1}S_X) = P(S) = 1$
(ii) Per ogni $B \subseteq S_X$, $0 \leq P_X(B) \leq P(S) = 1$
(iii) Per ogni $B, C \subseteq S_X$, $B \cap C = \emptyset$,

$$\begin{aligned} P_X(B \cup C) &= P(X^{-1}(B \cup C)) \\ &= P(X^{-1}(B) \cup X^{-1}(C)) \\ &= P(X^{-1}(B)) + P(X^{-1}(C)) \\ &= P(B) + P(C) \end{aligned}$$

in cui la terza uguaglianza deriva, come sempre, dal fatto che gli eventi sono disgiunti e che quindi sono disgiunte le immagini inverse.

Infine, $|S_X| \leq |S| < \infty$. □

Osservazione 7. *Si noti che P_X risulta essere una probabilità proprio perché utilizziamo X^{-1} per trasportare P in P_X. Se $Z : T \to S$ fosse una fuzione a valori in uno spazio di probabilità (S, P) il trasporto nella direzione opposta dato da $P \circ Z$ non dà in generale uno spazio di probabilità perché le immagini di insiemi disgiunti non è detto che siano disgiunte.*

Se D indica una certa distribuzione e X è una variabile aleatoria si scrive $X \sim D$ (letto X ha distribuzione D) per indicare che X ha la distribuzione D. Ad esempio, se X è la la funzione identità su uno spazio di probabilità uniforme su $\{1, 2, \ldots, n\}$ indichiamo con $X \sim U(\{1, 2, \ldots, n\}) = U(n)$ il fatto la distribuzione di X è appunto lo spazio uniforme di partenza.

Poiché (S_X, P_X) è uno spazio di probabilità gran parte delle informazioni su X possono essere dedotte direttamente dalla sua distribuzione, tant'è che se due variabili aleatorie hanno la stessa distribuzione condividono gran parte delle proprietà . Per cui si pone:

Definizione 7. *Si dice che due variabili aleatorie X definita su (S_1, P_1) ed Y definita su (S_2, P_2) sono uguali in distribuzione, e si denota con $X =^d Y$, se $(S_X, P_X) = (S_Y, P_Y)$*

Da questo punto di vista sembra perfino che non sia necessario introdurre le variabili aleatorie, in quanto il loro studio si riconduce a quello di spazi di probabilità . Tuttavia, solo 'gran parte' delle proprietà dipendono dalla distributione, non tutte. In particolare, in qualunque modello di fenomeni reali si è interessati al valore assunto da una variabile aleatoria e non solo alla sua distribuzione; in questa direzione si osservi che variabili aleatorie uguali in distribuzione non sono necessariamente uguali, anzi possono essere sempre diverse:

Esempio 32. *Nel lancio di una moneta, la funzione indicatrice di testa $\mathbb{I}_T$ e la funzione indicatrice di croce $\mathbb{I}_C$ soddisfano $\mathbb{I}_T =^d \mathbb{I}_C$ ma $\mathbb{I}_T(s) \neq \mathbb{I}_C(s)$ per ogni $s \in S$; in altre parole, i guadagni di chi punta su testa sono opposti a quelli di chi punta su croce*

Come accaduto nell'ultimo esempio, capita di considerare due o più variabili aleatorie $X_1, X_2. \dots, X_n$ definite su uno stesso spazio di probabilità . Anche ogni combinazione $\phi(X_1, X_2, \dots, X_n)$ di queste determinata da una funzione $\phi : B \to \mathbb{R}$, con $B \subseteq \mathbb{R}^n$ tale che $S_{X_1} \times S_{X_2} \times \cdots \times S_{X_n} \subseteq B$, è una variabile aleatoria definita da

$$\phi(X_1, X_2, \dots, X_n)(s) = \phi(X_1(s), X_2(s), \dots, X_n(s)).$$

Così $X_1 + X_2$, $4X_1$ e X_1^2 sono variabili aleatorie sullo stesso spazio di X_1 e X_2.

La determinazione della distribuzione di

$$Y = \phi(X_1(s), X_2(s), \dots, X_n(s))$$

può essere laboriosa, come si intravede dall'esempio seguente:

Esempio 33. *Se X_1 e X_2 sono i risultati del lancio di due dadi e $X_1 + X_2$ è la loro somma, con una certa pazienza si puo' ricavare la distribuzione di $X_1 + X_2$ dalle probabilità dei risultati dei lanci dei due dadi. In particolare, $S_{X_1+X_2} = \{2, 3, \dots, 12\}$ e*

$$P_{X_1+X_2}(s) = \begin{cases} \frac{s-1}{36} & se\ s \leq 6 \\ \frac{12-s+1}{36} & se\ s \geq 7 \end{cases}$$

Alcune variabili aleatorie notevoli sono elencate di seguito:

Esempio 34. *Su un insieme finito $S \in \mathbb{R}$ la variabile aleatoria X che può assumere ognuno dei valori di S con la stessa probabilità è detta uniforme e si indica $X \sim U(\{1, \dots, n\})$ se $|S| = n$, per cui $P(X = k) = 1/n$ per $k = 1, \dots, n$.*

Esempio 35. *Il numero di successi T_n in n prove indipendenti ognuna con probabilità di successo p è una variabile aleatoria che si dice Binomiale o Bernoulli di parametri (n, p) e si indica generalmente $T_n \sim B(n, p)$, per cui $P(T_n = k) = p^k(1-p)^{n-k}$ per $k = 0, \dots, n$.*

4.2. Valore atteso

Introduciamo ora una quantità , utilizzabile per l'analisi di variabili alatorie, che può essere giustificata in vari modi.

Esempio 36. *Storicamente fu descritta per la prima volta relativamente al gioco dei punti, in cui due giocatori A e B si disputano 24 gettoni a testa o croce: vince il primo che arriva a 2 vittorie. Naturalmente esiste un modo equo di dividersi i gettoni senza giocare: 12 ognuno. Se però A, che scommette su testa, ha vinto la prima partita, come ci si può dividere la posta senza giocare ulteriormente? Se si decidesse di fermarsi dopo un'ulteriore partita allora se in questa esce ancora testa A prende 24 gettoni, e se esce croce ci si puo' dividere 12 gettoni a testa; quindi dopo una vittoria di A è ragionevole dare i 12*

ad A e dividersi gli altri 6 ciascuno, per cui A ha 18 gettoni. Si nota però che $18 = \frac{1}{2}24 + \frac{1}{2}12$. *Se invece si decidesse di finire il gioco la probabilità di A di vincere tutta la posta dato che è uscita una testa sarebbe* $P(T \cup CT) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ *e si nota che di nuovo* $18 = \frac{3}{4} \times 24$.

Esempio 37. *Supponiamo di scommettere sul risultato di un dado, vincendo 3 se esce il 6 ed altrimenti perdendo 1. In ogni partita la nostra vincita sarà quindi una variabile aleatoria* X *tale che* $X(6) = 3$ *e* $X(i) = -1$ *per ogni* $i = 1, \ldots, 5$. *Dopo* 60 *partite ci aspettiamo di aver vinto circa* 10 *partite e perse* 50, *con una vincita totale di* -20. *Si noti che*

$$-20 = 30 - 50 = 3 \times 10 - 1 \times 50 = 3\frac{60}{6} - 1\frac{60 \times 5}{6} = 60(3\frac{1}{6} - 1\frac{5}{6})$$

.

In questi ed in molti altri esempi compare quindi la quantità somma dei valori vinti moltiplicati per la probabilità di vincerli e questa quantità può essere quindi un modo di valutare l'esito di una variabile aleatoria. Per cui si pone:

Definizione 8. *Data una variabile aleatoria* X *definita su uno spazio di probabilità* (S, P) *si dice* **valore atteso** *o* **speranza matematica** *o aspettazione o valor medio di* X *il valore*

$$E(X) = M(X) = \sum_{s \in S} X(s)P(s). \tag{4.14}$$

Il valore atteso dipende solo dalla distribuzione di X:

Lemma 6.

$$E(X) = \sum_{x \in S_X} xP_X(x)$$

Dimostrazione.

$$\begin{aligned} E(X) &= \sum_{s \in S} X(s)P(s) \\ &= \sum_{x \in S_X} \sum_{s \in S : X(s) = x} xP(s) \\ &= \sum_{x \in S_X} xP_X(x) \end{aligned}$$

□

Esempio 38. *Il valore atteso del risultato* X *del lancio di un dado è :*

$$E(X) = \sum_{i=1}^{6} i\frac{1}{6} = 3,5.$$

L'esempio mostra che non sempre il valore atteso è un valore che può essere assunto dalla variabile aleatoria e mostra quindi che l'idea che rappresenti il valore 'che ci aspettiamo' è soltanto approssimativa e va specificata nei sensi che abbiamo indicato in precedenza.

Esempio 39. *Il valore atteso di una variabile aleatoria* $T_n \sim B(n,p)$ *è :*

$$E(T_n) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np.$$

Alcune proprietà elementari del valore atteso sono:

Lemma 7. *Se* $X_1, X_2, \ldots, X_n$ *sono variabili aleatorie su uno spazio di probabilità* (S,P) *e* $a_1, a_2, \ldots, a_n \in \mathbb{R}$ *sono costanti si ha che:*
(i) $E(a) = a$
(ii) $E(a_1X_1 + a_2X_2) = E(X_1) + E(X_2)$ *e pertanto*
(iii) $E(\sum_{i=1}^n a_iX_i) = \sum_{i=1}^n a_iE(X_i)$

Dimostrazione. (i) $E(a) = \sum_{s\in S} aP(s) = a$.
(ii)

$$\begin{aligned} E(a_1X_1 + a_2X_2) &= \sum_{s\in S}(a_1X_1 + a_2X_2)(s)P(s) \\ &= \sum_{s\in S}(a_1X_1(s) + a_2X_2(s))P(s) \\ &= \sum_{s\in S} a_1X_1(s)P(s) + \sum_{s\in S} a_2X_2(s)P(s) \\ &= E(X_1) + E(X_2). \end{aligned}$$

(iii) segue per induzione su n □

Esprimere una variabile aleatoria come somma di altre variabili può quindi condurre ad una drastica semplificazione del calcolo del valore atteso, come si vede dall'esempio seguente.

Esempio 40. *Se* $T_{1000} \sim B(1000, 1/2)$ *è il numero di teste in* 1000 *lanci di una moneta allora* $T_{1000} = \sum_{i=1}^{1000} X_i$ *con* X_i *la funzione indicatrice di testa all'i-simo lancio. Si ha che* $E(X_i) = 1/2$ *per ogni* i *e*

$$E(T_n) = E(\sum_{i=1}^{1000} X_i) = \sum_{i=1}^{1000} E(X_i) = 500.$$

Pertanto il valore atteso riproduce in questo caso correttamente l'idea di quello che 'ci aspettiamo', così come era stato assunto nelle discussioni iniziali ed anche utilizzato nel teorema di De Moivre-Laplace. Generalizzando si ottiene che se $T_n \sim B(n,p)$ *allora* $T_n = \sum_{i=1}^n X_i$ *con* $X_i \sim B(1,p)$. *Per cui* $E(X_i) = 1p - 0(1-p) = p$ *e* $E(T_n) = np$.

In questo calcolo la proprietà di indipendenza della distribuzione di Bernoulli non è stata utilizzata ed infatti il Lemma vale in generale per tutte le variabili aleatorie.

Esempio 41. *Se T_n è il totale di n estrazioni dalla tombola senza reinserimento allora $T_n = \sum_{i=1}^n X_i$, con X_i i-simo numero estratto. poiché*

$$E(X_i) = \sum_{j=1}^{90} \frac{j}{90} = \frac{91}{2}$$

per ogni $i \leq 90$, e $E(X_i) = 0$ per ogni $i > 90$, si ha

$$E(T_n) = \begin{cases} n\frac{91}{2} & se\ n \leq 90 \\ 90\frac{91}{2} & se\ n \geq 90. \end{cases}$$

Si noti il caso $n = 90$, in cui tutte le palline sono estratte e $T_{90} \equiv 90\frac{91}{2}$ con probabilità uno.

Ci sono altri modi per indicare un 'valore medio' di una variabile aleatoria, utili in particolari contesti. Tra questi vi è il centro o i centri delle probabilità :

Definizione 9. *Si dice* **mediana** *di una variabile aleatoria X definita su (S, P) ogni valore $m(X)$ tale che*

$$P(X \leq m(X)) \geq \frac{1}{2} \ e\ P(X \geq m(X)) \geq \frac{1}{2}.$$

Esempio 42. *Tutti i valori in $[3, 4]$ sono mediane del risultato del lancio di un dado.*

Esempio 43. *La mediana del numero T_n di teste in 100 lanci di una moneta è 50, come si vede per simmetria, mentre per 101 lanci è un qualunque numero in $[50, 51]$*

Come si è visto dagli esempi, la mediana non è necessariamente unica e può coincidere o meno con il valore atteso.

Definizione 10. *Si dice* **moda** *di una variabile aleatoria X ogni valore $mo(X)$ tale che per ogni $x \in S_X$*

$$P(X = mo(X)) \geq P(X = x).$$

E' moda di una variabile aleatoria $T_n \sim B(n, p)$ ogni valore in $[p(n+1) - 1, p(n+1)] \cap \mathbb{N}$ mentre tutti i valori possibili sono moda per le variabili aleatorie uniformi. La moda di una variabile aleatoria è utile solo in alcune situazioni: ad esempio, se si è obbligati a scommettere su uno solo dei risultati possibili di un esperimento casuale la moda permette di massimizzare le probabilità di vittoria.

Quando si osservano dei dati $\{x_1, \ldots, x_n\}$ dopo aver condotto un esperimento casuale è possibile definire lo spazio di probabilità $(S_{\mathbb{O}}, P_{\mathbb{O}})$,

che chiameremo empirico o campionario, in cui $S_{\mathbb{O}}$ è costituito dai valori osservati e $P_{\mathbb{O}}$ è determinato dalla frequenze, ossia

$$P_{\mathbb{O}}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{x_i=x}.$$

In tale spazio il valore atteso della variabile aleatoria identità I è la media empirica dei dati:

$$\begin{aligned} \mu = E(I) &= \sum_{x\in\{x_1,\ldots,x_n\}} xP_{\mathbb{O}}(x) \\ &= \sum_{x\in\{x_1,\ldots,x_n\}} x\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{x_i=x} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i. \end{aligned}$$

La mediana di I è la mediana empirica $m = m(I)$ tale che $|\{i : x_i \leq m\}| \geq n/2$ e $|\{i : x_i \geq m\}| \geq n/2$. Si noti che le mediane empiriche si possono determinare ordinando (debolmente) le osservazioni x_i, che vengono in questo caso denotate con $\{x_{(1)}, \ldots, x_{(n)}\}$ con $x_{(1)} \leq \cdots \leq x_{(n)}$, e scegliendo poi un valore nell'intervallo (banale se n è dispari) $[x_{(\lfloor \frac{n+1}{2} \rfloor)}, x_{(\lceil \frac{n+1}{2} \rceil)}]$. La moda di I è la moda dei dati, ossia ogni valore $mo(I)$ che massimizza la frequenza di osservazione.

Esempio 44. *Con dati* $5.0, 3.5, 2.3, 5.0, 1.2, 1.6$ *risultano:* $\mu = 3.1, m \in [2.3, 3.5], mo = 5.0$.

Prima di condurre un esperimento è possibile valutare a priori le quantità empiriche appena descritte considerandole come funzioni di variabili aleatorie $X_1, \ldots, X_n$. In particolare la media empirica delle variabili aleatorie diviene

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

di cui si discuteranno varie proprietà nel seguito.

Per concludere la discussione sul valore atteso osserviamo che data una variabile aleatoria X ed una funzione $\Phi : S_X \to \mathbb{R}$, il risultato seguente sul cambiamento di variabili permette di semplificare il calcolo del valore atteso della variabile aleatoria $\Phi(X)$.

Teorema 10. *Data una variabile aleatoria X su (S, P) e $\Phi : B \to \mathbb{R}$ con $S_X \subseteq B$ si ha*

$$\begin{aligned} E(\Phi(X)) &= \sum_{\phi\in S_{\Phi(X)}} \phi P(\Phi(X) = \phi) \\ &= \sum_{x\in S_X} \Phi(x)P_X(x) \end{aligned}$$

DIMOSTRAZIONE. la prima uguaglianza discende dal Lemma 6. Per la seconda si ha

$$\begin{aligned} E(\Phi(X)) &= \sum_{s\in S} \Phi(X(s))P(s) \\ &= \sum_{x\in S_X} \sum_{s\in S:X(s)=x} \Phi(x)P(s) \\ &= \sum_{x\in S_X} \Phi(x)P_X(x) \end{aligned}$$

□

La prima espressione di $E(\Phi(X))$ data nel teorema richiede la determinazione di $(S_{\Phi(X)}, P_{\Phi(X)})$, mentre per la seconda espressione basta la distribuzione di X che può essere utilizzata per tutte le Φ.

Il valore atteso di una potenza di una variabile aleatoria X viene definito **momento di ordine** k di X ed in accordo con l'ultimo risultato vale $\sum_{x\in S_X} x^k P_X(x)$.

4.3. Indipendenza tra variabili aleatorie

E' stata sollevata nell'ultimo capitolo la questione della dipendenza tra variabili aleatorie. Per definire la mancanza di dipendenza estendiamo il concetto di indipendenza tra eventi. Per comprendere come realizzare questa estensione partiamo dall'esempio seguente.

Per semplificare la notazione, per variabili aleatorie $X_1, , \ldots, X_n$ definite su uno stesso spazio di probabilità (S, P) indicheremo con

$$\begin{aligned} &\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \\ &\qquad = \{s \in S : X_1(s) = x_1 \text{ e } X_2(s) = x_2 \text{ e } \ldots \text{ e } X_n(s) = x_n\}. \end{aligned}$$

ESEMPIO 45. *Se $X_i = \mathbb{I}_{A_i}$, con A_i eventi indipendenti in uno spazio di probabilità (S, P), allora*

$$\begin{aligned} P(\mathbb{I}_{A_1} = 1)P(\mathbb{I}_{A_2} = 1) &= P(A_1)P(A_2) \\ &= P(A_1 \cap A_2) \\ &= P(\mathbb{I}_{A_1}\mathbb{I}_{A_2} = 1) \\ &= P(\mathbb{I}_{A_1} = 1, \mathbb{I}_{A_2} = 1) \end{aligned} \tag{4.15}$$

ma abbiamo anche visto che

$$\begin{aligned} P(\mathbb{I}_{A_1} = 1)P(\mathbb{I}_{A_2} = 0) &= P(A_1)P(A_2^c) \\ &= P(A_1 \cap A_2^c) \\ &= P(\mathbb{I}_{A_1} = 1, \mathbb{I}_{A_2} = 0) \end{aligned} \tag{4.16}$$

e così via, di modo che

$$P(\mathbb{I}_{A_1} = i)P(\mathbb{I}_{A_2} = j) = P(\mathbb{I}_{A_1} = i, \mathbb{I}_{A_2} = j)$$

per ogni $i, j = 0, 1$.

Per definire l'indipendenza si può quindi generalizzare questa proprietà :

DEFINIZIONE 11. *n variabili aleatorie $X_1, , \ldots, X_n$ definite su uno spazio di probabilità (S, P) sono indipendenti se*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \tag{4.17}$$

Per $n = 2$, quindi, X_1, X_2 sono indipendenti se $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$ per ogni possibile valore x_1, x_2. Si noti che nella definizione di indipendenza per variabili aleatorie non è stata richiesta la fattorizzazione delle probabilità per sottofamiglie: mostreremo tra poco che questa proprietà è conseguenza della definizione data.

ESEMPIO 46. *Due estrazioni successive X_1 e X_2 dalla tombola sono indipendenti se effettuate con reinserimento e dipendenti se non c'è reinserimento; infatti, nel primo caso $P(X_1 = k, X_2 = m) = \frac{1}{90^2} = P(X_1 = k)P(X_2 = m)$ per ogni $k, m = 1, \ldots, 90$, mentre nel secondo caso $P(X_1 = 1, X_2 = 1) = 0 \neq \frac{1}{90^2} = P(X_1 = 1)P(X_2 = 1)$.*

LEMMA 8. *Se $X_1, \ldots, X_n$ sono variabili aleatorie indipendenti allora*
(I) per ogni $A_1, \ldots, A_n$, $A_i \subseteq S_{X_i}$ si ha

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

(II) per ogni $J \subseteq \{1, \ldots, n\}$ le variabili aleatorie X_i, $i \in J$, sono indipendenti,
(III) per ogni $A_1, \ldots, A_n$, $A_i \subseteq S_{X_i}$, gli eventi $(X_i \in A_i) = \{s \in S : X_i(s) \in A_i\}$ sono collettivamente indipendenti.

DIMOSTRAZIONE. Indichiamo con $\mathbf{x} = (x_1, \ldots, x_n)$ i vettori n dimensionali:

(I) Fissati $A_1, \ldots, A_n$ come nell'ipotesi si ha

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(\cup_{x_1 \in A_1, \ldots, x_n \in A_n} \{\mathbf{X} = \mathbf{x}\}).$$

Gli eventi $\{\mathbf{X} = \mathbf{x}\}$ sono disgiunti per definizione di funzione, quindi per l'indipendenza degli X_i

$$\begin{aligned} P(X_1 \in A_1, \ldots, X_n \in A_n) &= \sum_{x_1 \in A_1, \ldots, x_n \in A_n} P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{x_1 \in A_1, \ldots, x_n \in A_n} \prod_{i=1}^{n} P(X_i = x_i) \\ &= \prod_{i=1}^{n} \sum_{x_1 \in A_1, \ldots, x_n \in A_n} P(X_i = x_i) \\ &= \prod_{i=1}^{n} P(X_i \in A_i) \end{aligned}$$

(II) segue da (I) con

$$A_i = \begin{cases} x_i & \text{se } i \in J \\ S_{X_i} & \text{se } i \notin J \end{cases}$$

essendo

$$\begin{aligned} P(X_i = x_i \text{ se } i \in J) &= P(X_i = x_i \text{ se } i \in J \text{ e } X_j \in S_{X_j} \text{ se } j \notin J) \\ &= \prod_{i \in J} P(X_i = x_i) \prod_{j \notin J} P(X_j \in S_{X_j}) \\ &= \prod_{i \in J} P(X_i = x_i) \end{aligned}$$

(III) segue da (II) prendendo $X_i = \mathbb{I}_{A_i}$. □

Quindi la richiesta di fattorizzazione delle probabilità per tutti i valori del codominio delle variabili aleatorie include già la fattorizzazione delle stesse espressioni per sottinsiemi di funzioni.

Vediamo ora che funzioni di variabili aleatorie indipendenti sono ancora indipendenti, nel senso che

Teorema 11. *Date variabili aleatorie indipendenti $X_1, \ldots, X_n$ definite su uno spazio di probabilità (S, P) e due funzioni $\phi : C \to \mathbb{R}$ e $\psi : D \to \mathbb{R}$ tali che $S_{X_1} \times S_{X_2} \times \cdots \times S_{X_k} \subseteq C$ e $S_{X_{k+1}} \times \cdots \times S_{X_n} \subseteq D$ si ha che $T = \phi(X_1, \ldots, X_k)$ e $Z = \psi(X_{k+1}, \ldots, X_n)$ sono indipendenti.*

Dimostrazione. poiché gli eventi nell'unione che segue sono, come al solito, disgiunti, ed essendo le variabili aleatorie indipendenti si

ha:

$$
\begin{aligned}
&P(T=t, Z=z)\\
&\quad = P(s : \phi(X_1(s),\dots,X_k(s)) = t,\\
&\qquad\qquad \psi(X_{k+1}(s),\dots,X_n(s)) = z)\\
&\quad = P\Big(\bigcup_{\substack{\boldsymbol{x}=(x_1,\dots,x_n):\phi(x_1,\dots,x_k)=t,\\ \psi(x_{k+1},\dots,x_n)=z}} \{s : (X_1(s),\dots,X_n(s)) = \boldsymbol{x}\}\\
&\quad = \sum_{\substack{\boldsymbol{x}=(x_1,\dots,x_n):\phi(x_1,\dots,x_k)=t,\\ \psi(x_{k+1},\dots,x_n)=z}} P((X_1,\dots,X_n) = \boldsymbol{x})\\
&\quad = \sum_{\substack{\boldsymbol{x}=(x_1,\dots,x_n):\phi(x_1,\dots,x_k)=t,\\ \psi(x_{k+1},\dots,x_n)=z}} \prod_{i=1}^{n} P(X_i = x_i)\\
&\quad = \sum_{\substack{(x_1,\dots,x_k):\\ \phi(x_1,\dots,x_k)=t}} \prod_{i=1}^{k} P(X_i = x_i) \sum_{\substack{(x_{k+1},\dots,x_n):\\ \psi(x_{k+1},\dots,x_n)=z}} \prod_{i=k+1}^{n} P(X_i = x_i)\\
&\quad = \sum_{\substack{\boldsymbol{x}=(x_1,\dots,x_k):\\ \phi(x_1,\dots,x_k)=t}} P((X_1,\dots,X_k) = \boldsymbol{x})\\
&\qquad\qquad \times \sum_{\substack{\boldsymbol{x'}=(x_{k+1},\dots,x_n):\\ \psi(x_{k+1},\dots,x_n)=z}} P((X_{k+1},\dots,X_n) = \boldsymbol{x'})\\
&\quad = P(T=t)P(Z=z).
\end{aligned}
$$

□

Questo risultato ha molte conseguenze interessanti, ad esempio la seguente:

Esempio 47. *Giocando $k+1$ partite alla roulette, se X_i è il numero che esce nell'i-sima partita si ha che le variabili $X_1,\dots,X_k,X_{k+1}$ sono indipendenti. Ogni strategia che, a k fissato, cerchi di determinare su quale numero scommettere alla $k+1$-sima partita osservando i risultati delle partite precedenti è equivalente ad una funzione $\phi(X_1,\dots,X_k)$ e, per il teorema appena provato, X_{k+1} ne risulta quindi indipendente. Per cui il teorema porta a concludere che non esiste nessuna strategia che dalle prime k partite dia suggerimenti sulla successiva partita.*

Si può ovviamente pensare una strategia che non fissi k a priori (tipo aspettare la prima uscita del 3), ma ne rimandiamo l'analisi in quanto richiede un modello con una infinità di possibili risultati, che tratteremo più avanti.

Una delle proprietà fondamentali delle variabili aleatorie indipendenti si riferisce al valore atteso del prodotto di variabili aleatorie.

TEOREMA 12. *Se X_i, $i = 1, \ldots, n$ sono variabili aleatorie indipendenti su (S, P) allora*

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i).$$

DIMOSTRAZIONE. Cominciamo da $n = 2$:

$$\begin{aligned}
E(X_1X_2) &= \sum_{t \in S_{X_1X_2}} tP(X_1X_2 = t) \\
&= \sum_{t \in S_{X_1X_2}} \sum_{x_1 \in S_{X_1}, x_2 \in X_2} x_1x_2P(X_1 = x_1, X_2 = x_2) \\
&= \sum_{x_1 \in S_{X_1}} x_1P(X_1 = x_1) \sum_{t \in S_{X_1X_2}} \sum_{x_2 \in S_{X_2}: x_1x_2 = t} x_2P(X_2 = x_2) \\
&= \sum_{x_1 \in S_{X_1}} x_1P(X_1 = x_1) \sum_{x_2 \in S_{X_2}} x_2P(X_2 = x_2) \\
&= E(X_1)E(X_2)
\end{aligned}$$

poiché la seconda uguaglianza discende dall'additività della probabilità dell'unione disgiunta di insiemi.

Il risultato per n generico discende per induzione su n essendo $\prod_{i=1}^{n-1} X_i$ indipendente da X_n. □

ESERCIZIO 14. *Due variabili aleatorie dipendenti tali che $E(X_1X_2) = E(X_1)E(X_2)$ si possono ottenere assegnando le probablità seguenti. Siano X_1, X_2 a valori in $\{0, 1\}$ e siano $P(X_1 = 0, Y_1 = 0) = \frac{2}{12}$, $P(X_1 = , Y_1 = 0) = \frac{6}{12}$, $P(X_1 = 0, Y_1 = 1) = \frac{1}{12}$ e di conseguenza $P(X_1 = 1, Y_1 = 1) = \frac{3}{12}$. Verificare quanto affermato.*

DEFINIZIONE 12. *Variabili aleatorie $X_1, \ldots, X_n$ su uno spazio di probabilità che siano indipendenti e tali che $X_i =^d X_j$ per ogni $i, j = 1, \ldots, n$ si dicono* **indipendenti ed identicamente distribuite** *o* **i.i.d.**.

4.4. Deviazioni dalla media

Il valore atteso $E(X)$ di una variabile aleatoria X dà un'idea di quanto ci possiamo aspettare, ma abbiamo già visto che non è esattamente il valore $E(X)$ che ci possiamo attendere come risultato di una realizzazione di X. Il teorema di De Moivre-Laplace fornisce una stima molto precisa di quanto i valori che si realizzano più spesso differiscono dal valore atteso di una distribuzione di Bernoulli, ma per il caso generale discutiamo ora una prima stima. Ricordiamo inoltre che non abbiamo ancora chiarito il significato del termine $\sqrt{p(1-p)}$.

Lo scarto massimo $\max(X-E(X))$ tende a sopravvalutare lo scarto tipico di una realizzazione e chiaramente lo scarto medio $E(X-E(X))$ non ha significato, essendo identicamente nullo per la linearità del valore atteso. Lo scarto assoluto medio $E(|X-E(X)|)$ è una soluzione migliore.

Esempio 48. *Per $X \sim B(1,p)$ si ha $E(X)=p$ e lo scarto assoluto medio risulta essere $2p(1-p)$; per il risultato del lancio di un dado Y lo scarto assoluto medio è $1,5$.*

Si osservi però che il calcolo dello scarto assoluto medio per variabili con distribuzione $B(n,p)$ risulta laborioso e che comunque non risulta spiegato il termine $\sqrt{p(1-p)}$. Pensandoci, non c'è una ragione evidente per non considerare lo scarto quadratico medio di X

$$SD(X) = \sqrt{E(X-E(X))^2}$$

detta anche **deviazione standard** di X. O anche qualche altra potenza. Tuttavia risulta che

Esempio 49. *Per $X \sim B(1,p)$ lo scarto quadratico medio risulta essere $\sqrt{p(1-p)}$ e per il risultato del lancio di un dado lo scarto quadratico medio è circa $1,70$.*

Si vede da questi esempi che lo scarto quadratico medio non differisce sostanzialmente dallo scarto assoluto medio, ma soprattutto 'spiega' il termine che appare nel teorema di De Moivre-Laplace. L'espressione $Var(X) = E(X-E(X))^2$ è detta **varianza** di X. La varianza può essere anche calcolata come segue.

Lemma 9. *Per ogni variabile aleatoria X si ha*

$$Var(X) = E(X^2) - (E(X))^2.$$

Dimostrazione.

$$\begin{aligned} Var(X) &= E(X-E(X))^2 \\ &= E(X^2 - 2XE(X) + (E(X))^2) \\ &= E(X^2) - (E(X))^2 \end{aligned} \tag{4.18}$$

□

Un importante risultato relativo alla varianza è che risulta additiva per somme di variabili aleatorie indipendenti.

TEOREMA 13. *Per ogni $X_1, \dots X_n$ variabili aleatorie indipendenti su uno spazio di probabilità (S, P) si ha*

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

e

$$SD(\sum_{i=1}^{n} X_i) = \sqrt{\sum_{i=1}^{n} Var(X_i)}.$$

DIMOSTRAZIONE. Iniziamo da $n = 2$.

$$\begin{aligned} Var(X_1 + X_2) &= E(X_1 + X_2)^2 - (E(X_1 + X_2))^2 \\ &= E(X_1)^2 + E(X_2)^2 + 2E(X_1X_2) \\ &\qquad -(E(X_1))^2 - (E(X_2))^2 - 2E(X_1)E(X_2) \\ &= Var(X_1) + Var(X_2) + 2(E(X_1X_2) - E(X_1)E(X_2)) \\ &= Var(X_1) + Var(X_2) \end{aligned}$$

poiché per variabili indipendenti $E(X_1X_2) = E(X_1)E(X_2)$ dal Teorema 12.

Il risultato per n generico si ottiene per induzione essendo $\sum_{i=1}^{n-1} X_i$ indipendente da X_n.

Il risultato per la deviazione standard segue dalla definizione. □

Quindi per variabili i.i.d. $X_1, \dots, X_n$ si ha che $SD(\sum_{i=1}^{n} X_i) = nSD(X_1)$. In particolare se $X_i \sim B(1, p)$ allora $SD(\sum_{i=1}^{n} X_i) = \sqrt{np(1-p)}$.

Quindi la deviazione standard è contemporaneamente facile da calcolare per somme di variabili indipendenti ed assegna un significato all'espressione $\sqrt{np(1-p)}$ che appare nel teorema di De Moivre-Laplace. Questo ci permette di congetturare un'estensione di questo risultato a tutte le variabili aleatorie finite. Rileggendo infatti il risultato per

$T_n \sim B(n,p)$ ossia tale che $T_n = \sum_{i=1}^n X_i$, $X_i \sim B(1,p)$, si ha

$$\begin{aligned}
\sum_{j \geq np + a\sqrt{np(1-p)}} \mathcal{P}(j,n,p) &= \sum_{j \geq np + a\sqrt{np(1-p)}} P(\sum_{i=1}^n X_i = j) \\
&= P\left(\sum_{i=1}^n X_i \geq np + a\sqrt{np(1-p)}\right) \\
&= P\left(\sum_{i=1}^n X_i \geq nE(X_1) + a\sqrt{nVar(X_1)}\right) \\
&= P\left(\frac{\sum_{i=1}^n X_i - nE(X_1)}{\sqrt{nVar(X_1)}} \geq a\right) \\
&= P\left(\frac{(\sum_{i=1}^n X_i) - E(\sum_{i=1}^n X_i)}{\sqrt{Var(\sum_{i=1}^n X_i)}} \geq a\right).
\end{aligned}$$

L'espressione $\frac{(\sum_{i=1}^n X_i) - E(\sum_{i=1}^n X_i)}{\sqrt{Var(\sum_{i=1}^n X_i)}}$ può essere interpretata in un modo diretto: $\sum_{i=1}^n X_i$ è una variabile aleatoria che dà i possibili valori della media empirica, a questa viene sottratto il suo valore atteso dividendo poi per la sua deviazione standard.

Definizione 13. *Data una variabile aleatoria X, si dice* **variabile standardizzata** *o* **versione standardizzata di** X *l'espressione*

$$\frac{X - E(X)}{SD(X)}$$

.

Il Teorema di De Moivre-Laplace ci dice quindi che la versione standardizzata della somma di variabili aleatorie indipendenti (binomiali) ha probabilità descritte asintoticamente dalla gaussiana. La versione standardizzata della somma è ben definita per ogni variabile aleatoria finita, quindi è ragionevole congetturare che

Teorema 14 (Teorema Centrale del Limite). *Per variabili aleatorie i.i.d. $X_1, \ldots, X_n$ su uno spazio di probabilità finito vale per ogni $a \in \mathbb{R}$*

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n X_i - nE(X_1)}{\sqrt{nVar(X_1)}} \geq a\right) = \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Per ora non dimostreremo questo teorema poiché ne vale in realtà una versione ancora più generale.

Ci si potrebbe porre il problema di stimare la deviazione della deviazione dal valor medio dalla sua media, per esempio con $E(|X - E(X) -$

$\sqrt{E(X-E(X))^2}|)$ oppure $E(\sqrt{(X-E(X)-\sqrt{E(X-E(X))^2})^2})$. Tuttavia questo non viene generalmente sviluppato perché (1), come illustrato dal Teorema centrale del limite, spesso è il secondo momento, e quindi la deviazione standard, che descrive in dettaglio la distribuzione della deviazione dalla media, (2) non introduce nessuna novità teorica perché si tratta sempre di un valore atteso di una deviazione e (3) diventa più chiaro porsi direttamente il problema della ricostruzione di una variabile aleatoria dalla conoscenza dei suoi momenti, un problema che non trattiamo in queste note.

4.5. Diseguaglianze e legge debole dei grandi numeri

La varianza dà un'idea della deviazione tipica dalla media, ma ora vediamo qualche risultato che dia una stima di questa deviazione.

Lemma 10 (Diseguaglianza di Markov). *Per ogni variabile aleatoria $X \geq 0$ non negativa e per ogni $a > 0$ si ha*

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

.

Dimostrazione. Essendo $X \geq 0$, per ogni $a > 0$ si ha:

$$\begin{aligned} E(X) &= \sum_{x \in S_X} x P_X(x) \\ &\geq \sum_{x \in S_X, x \geq a} x P_X(x) \geq a P(X \geq a) \end{aligned}$$

□

Da questo segue

Corollario 2 (Diseguaglianza di Chebyshev). *Per ogni variabile aleatoria X finita e per ogni $a > 0$ si ha:*

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2},$$

ossia

$$P(\left|\frac{X - E(X)}{SD(X)}\right| \geq a) \leq \frac{1}{a^2}.$$

Dimostrazione. Essendo $(|X - E(X)|)^2 = (X - E(X))^2 \geq 0$, dal Lemma 10 segue che per ogni $a > 0$ si ha:

$$P(|X - E(X)| \geq a) = P((X - E(X))^2 \geq a^2) \leq \frac{Var(X)}{a^2}$$

□

Queste diseguaglianze non sono troppo accurate, come si vede dall'esempio seguente, anzi talvolta sono banali in quanto il maggiorante è maggiore di 1.

Esempio 50. *Se X è il risultato del lancio di un dado allora*

$$\frac{1}{3} = P(X \geq 5) \leq \frac{1}{5}3,5 = 0,7$$

dal Lemma 10 e

$$\frac{1}{3} = P(|X - 3,5| \geq 2,5) \leq \frac{1}{(2,5)^2}Var(X) \approx 0.47$$

dal Corollario 2.

Si possono tuttavia dedurre due cose.

Lemma 11. *Se $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$ allora la costante nell'approssimazione di Stirling soddisfa $e^c = \sqrt{2\pi}$.*

Dimostrazione. Dalla parte dimostrata della legge di De Moivre-Laplace scritta in termini di variabili aleatorie sappiamo che per ogni $a_1, a_2 \geq 0$

$$Q_n(a_1, a_2) = P(\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \in [a_1, a_2]) \to_{n\to\infty} \int_{a_1}^{a_2} \frac{1}{e^c} e^{-x^2/2} dx$$

e per la simmetria della distribuzione di Bernoulli e della gaussiana, questo risultato vale anche quando $a_1, a_2 \leq 0$ e dall'additività per ogni $a_1, a_2 \in \mathbb{R}$. A noi interessa qui per $a_1 = -a = -a_2 < 0$.
1. Assumendo $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$ si ha che per ogni $a > 0$ vale $1 \geq Q_n(-a, a)$ per ogni n; pertanto anche

$$\begin{aligned} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 &\geq \lim_{n\to\infty} Q_n(-a, a) \\ &= \int_{-a}^{a} \frac{1}{e^c} e^{-x^2/2} dx \end{aligned} \tag{4.19}$$

per ogni $a > 0$. Questo implica $e^c \geq \sqrt{2\pi}$.
2. D'altra parte, dalla definizione di limite, per ogni $\epsilon > 0$ esiste $N > 0$ tale che per ogni $n \geq N$ vale la prima diseguaglianza in

$$\begin{aligned} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx &= 1 = P(\sum_{i=1}^n X_i \in [0, n]) \\ &= P(\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \in [-a, a]) \\ &\quad -P(|\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}}| \geq a) \\ &\leq \int_{-a}^{a} \frac{1}{e^c} e^{-x^2/2} dx + \epsilon + P(|\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}}| \geq a) \\ &\leq \int_{-\infty}^{\infty} \frac{1}{e^c} e^{-x^2/2} dx + \epsilon + \frac{1}{a^2}, \end{aligned}$$

in cui l'ultima diseguaglianza segue dalla diseguaglianza di Chebyshev. Quindi se $a > \epsilon^{-1/2}$ si ha

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \leq \int_{-\infty}^{\infty} \frac{1}{e^c} e^{-x^2/2} dx + 2\epsilon$$

da cui $e^c \leq \sqrt{2\pi}$. □

La seconda conseguenza riguarda la probabilità di una deviazione dal valore atteso dell'ordine di n:

Teorema 15 (Legge (debole) dei grandi numeri). *Per variabili aleatorie finite i.i.d. $X_1, X_2, \ldots$ su si ha che per ogni $\alpha > 0$*

$$\lim_{n\to\infty} P(\left|\sum_{i=1}^{n} X_i - E(X_1)\right| > \alpha) = 0$$

nel senso che fissata una distribuzione finita (S_X, P_X) per ogni $\epsilon > 0$ esiste N tale che per ogni $n \geq N$ se si prendono n variabili aleatorie indipendenti ognuna con distribuzione (S_X, P_X) allora $P(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_1)\right| > \alpha) < \epsilon$.

L'esistenza di variabili indipendenti con una distribuzione data verrà verificata più avanti.

Dimostrazione. Dalla diseguaglianza di Chebyshev si ha:

$$\begin{aligned} P(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_1)\right| > \alpha) &= P(\left|\sum_{i=1}^{n} X_i - nE(X_1)\right| > n\alpha) \\ &\leq \frac{1}{n^2\alpha^2} Var(\sum_{i=1}^{n} X_i) \\ &= \frac{nVar(X_1)}{n^2\alpha^2} \to_{n\to\infty} 0 \end{aligned}$$

□

Esempio 51. *Se X_i sono i risultati di lanci indipendenti di un dado allora*

$$P(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_1)\right| > 10^{-10}) \to_{n\to\infty} 0$$

anzi

$$P(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_1)\right| > 10^{-10}) \leq \frac{2,9 \times 10^{20}}{n}.$$

Si noti che quest'ultima stima ha senso solo per $n \geq 2,9 \times 10^{20}$; tuttavia ha poco senso dare troppa importanza al valore quantitativo di queste stime a causa della loro scarsa accuratezza.

4.6. Approssimazione di Poisson

Vediamo ora un'approssimazione per la distribuzione binomiale quando la probabilità di successo p ed il numero di prove n sono tali che pn è dell'ordine di 1:

Esempio 52. *Sorteggiando con reinserimento dalla tombola* 180 *volte la probabilità che l'*1 *esca esattamente* 2 *volte è :*

$$\mathcal{P}(2, 180, \frac{1}{90}) = \binom{180}{2} \frac{1}{90^2} (\frac{89}{90})^{178}.$$

Teorema 16 (Approssimazione di Poisson). *Se* $p = p_n$ *è tale che* $\lim_{n\to\infty} np_n = \lambda > 0$ *si ha*

$$\lim_{n\to\infty} \mathcal{P}(k, n, p_n) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Dimostrazione. poiché $\lim_{n\to\infty}(1 - p_n)^n = e^{-\lambda}$ si ha

$$\begin{aligned} \mathcal{P}(k, n, p_n) &= \binom{n}{k} (p_n)^k (1 - p_n)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{k!} (p_n^k)(1 - p_n)^n (1 - p_n)^{-k} \\ &\to_{n\to\infty} \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

□

CAPITOLO 5

PROBABILITÀ SU INSIEMI DISCRETI

5.1. Spazi di probabilità su insiemi discreti

Contemporaneamente allo sviluppo delle probabilità finite alcuni studiosi si resero conto che taluni problemi non si potevano formalizzare con un numero finito di possibilità. I primi problemi di questo tipo appaiono in un libro di Huygens del 1657.

Esempio 53. *Se giocando a dadi A vince se esce prima il 6 di 1 e 2, e viceversa per B, qual è la probabilità che A vinca? Chiaramente è naturale considerare la probabilità che vinca A al lancio k, per $k \in \mathbb{N}$.*

Ciò motiva l'introduzione di spazi di probabilità con S numerabile; ma questo introduce una nuova scelta: se $\{A_i\}_{i\in\mathbb{N}}$ è una famiglia numerabile di eventi disgiunti, si dovrà richiedere la numerabile additività della probabilità o solo quella finita? La questione non è risolta in modo univoco: una condizione più stringente limita il campo di applicazione della teoria ma ne semplifica gli sviluppi ed è quindi ragionevole richiederla quando il campo di applicazione resti comunque sufficientemente ampio. In generale, l'imposizione dell'additività numerabile non pone restrizioni di rilievo alla applicazioni fisiche (trattandosi perlopiù di esperimenti ripetibili a piacere) mentre ne pone in ambiti economici, trattandosi spesso in quel caso di situazioni solo occasionalmente ripetibili. Noi qui lo adotteremo sia per semplicità sia perché comunque la teoria qui esposta ha una sufficiente ampiezza di applicazioni anche in campo socio-economico-finanziario.

Definizione 14. *Uno* **spazio di probabilità discreto** *è una coppia (S, P) in cui S è un insieme al più numerabile e P è una funzione definita sulle parti di S tale che:*
(1) $P(S) = 1$
(2) per ogni $A \subseteq S$, $P(A) \in [0, 1]$;
(3) se A_i, $i = 1, 2, \ldots$, $A_i \subseteq S$ sono insiemi disgiunti allora

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

in cui si intende che a destra c'è una serie a termini positivi convergente.

La richiesta di additività numerabile rende in realtà il modello più semplice di quanto l'assiomatica appena descritta sembrasse prospettare.

LEMMA 12. *Tutti e soli gli spazi di probabilità discreti sono ottenuti da un insieme al più numerabile S e da una funzione $q : S \to \mathbb{R}^+$ tale che $\sum_{s\in S} q(s) < \infty$ ponendo per ogni $A \subseteq S$*

$$P(A) = \frac{\sum_{s\in A} q(s)}{\sum_{s\in S} q(s)}.$$

DIMOSTRAZIONE. Ogni spazio di probabilità discreto (S, P) si può rappresentare come detto con $q(s) = P(s)$. Viceversa, dato q come nell'asserzione basta porre $P(s) = \frac{q(s)}{\sum_{s\in S} q(s)}$; la verifica che P è una probabilità è lasciata per esercizio. □

OSSERVAZIONE 8. *Per gli spazi di probabilità discreti vale l'additività finita della probabilità e quindi valgono tutti i risultati dei capitoli 2 e 3 nella forma in cui sono enunciati, ossia riferiti ad un numero finito di eventi. E' solo quando è coinvolta una famiglia numerabile di eventi che dobbiamo dedurre i risultati dall'additività numerabile (e dalle proprietà delle serie).*

ESEMPIO 54. *Nell'esempio 53 se A è l'evento che vince A; A_i è l'evento che vince A alla i-sima prova; N_j è l'evento che alla j-sima prova non escono nessuno di $\{1, 2, 6\}$ e S_j è l'evento che esce il 6 alla j-sima prova per l'indipendenza della prove si ha*

$$P(A_i) = P(\cap_{j=1}^{i-1} N_j \cap S_i) = (\frac{1}{2})^{i-1}\frac{1}{6}$$

e quindi

$$\begin{aligned} P(A) &= P(\cup_{i=1}^{\infty} A_i) \\ &= \sum_{i=1}^{\infty} P(A_i) \\ &= \sum_{i=1}^{\infty} (\frac{1}{2})^{i-1}\frac{1}{6} = \frac{1}{3}. \end{aligned}$$

La probabilità che il primo successo di prove indipendenti ognuna con probabilità di successo p avvenga alla n-sima prova si può calcolare considerando, come nell'esempio, l'evento A_i che il primo successo sia alla i-sima prova e l'evento B_j che indica successo alla j-sima prova; per l'indipendenza della prove si ha

$$P(A_i) = P(\cap_{j=1}^{i-1} B_j^c \cap B_i) = (1-p)^{i-1}p.$$

Ora questo suggerisce di definire uno spazio di probabilità discrete $(\overline{S}, \overline{P})$ con $\overline{S} = \mathbb{N}$ e $\overline{P}$ data da

$$\overline{P}(i) = (1-p)^{i-1}p.$$

Infatti correttamente si ha;

$$\sum_{i=1}^{\infty} \overline{P}(i) = \sum_{i=1}^{\infty} (1-p)^{i-1} p = 1$$

.

Definizione 15. *Lo spazio di probabilità* $(\overline{S}, \overline{P})$ *relativo al primo successo in prove indipendenti ognuna con probabilità di successo* p *è detto* **distribuzione geometrica** *di parametro* p.

Si noti che anche l'approssimazione di Poisson generava una funzione di k per $k \in \mathbb{N}$; inoltre poiché

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = 1$$

si possono prendere questi valori come probabilità:

Definizione 16. *Per ogni* $\lambda \in \mathbb{R}$, *uno spazio di probabilità* $(\overline{S}, \overline{P})$ *con* $\overline{S} = \mathbb{N}$ *e* $\overline{P}$ *data da*

$$\overline{P}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

è detto **distribuzione di Poisson** *di parametro* λ.

5.2. Variabili aleatorie discrete

Passiamo ora allo studio delle variabili aleatorie definite su uno spazio di probabilità discreto. Non c'è nessuna difficoltà a porre

Definizione 17. *Dato uno spazio di probabilità* (S, P), *una variabile aleatoria discreta* X *è una funzione* $X : S \to \mathbb{R}$.

Anche per una variabile aleatoria discreta X è possibile definire la distribuzione (S_X, P_X) come in (4.14).

Esercizio 15. *Verificare che* (S_X, P_X) *è uno spazio di probabilità discreto.*

Si può poi ripetere la Definizione 7 di uguaglianza in distribuzione.

Esempio 55. *Se* (S, P), $S = \mathbb{N}$ *è uno spazio di probabilità che descrive la distribuzione del primo successo in prove indipendenti con probabilità di successo* p, *la variabile aleatoria* $Y =$ *tempo del primo successo è definita da* $X(k) = k$ *ha distribuzione* $P(X = j) = (1-p)^{j-1}p$ *é detta variabile geometrica*(p), *mentre la variabile aleatoria* $Y =$ *tempo di attesa del primo successo è definita da* $Y(k) = k - 1$ *ha distribuzione* $P(Y = j) = (1-p)^j p$.

Si noti che $Y =^d X - 1$.

Esempio 56. *Se* (S_λ, P_λ) *è uno spazio di probabilità di Poisson e una variabile aleatoria* N *ha distribuzione* (S_λ, P_λ) *allora si dice che* N *ha distribuzione di Poisson*(λ).

C'è invece qualche problema nel definire il valore atteso:

Esempio 57. *In un gioco, se la prima uscita del* 6 *in un dado è alla k-sima prova si vince (o si perde se negativo) l'importo* x_k. *Giochereste se*

(a) $x_k = (-1)^k k$*?*

(b) $x_k = (-1)^k (6/5)^k$*?*

Si consideri uno spazio di probabilità geometrico(p) *e la variabile aleatoria* $X(k) = (x_k)$. *Per valutare il nostro vantaggio nel gioco verrebbe di calcolare* $E(X)$, *ma una ragionevole espressione sarebbe* $\sum_{k=1}^{\infty} x_k (5/6)^k$ *che nel caso (a) dà* $11/375 - 6/11 < 0$, *ma nel caso (b) dà* $1/5 \sum_{k=1}^{\infty} (-1)^k$ *che è una serie indeterminata.*

Per non rischiare di incontrare situazioni come questa e per garantire che le principali proprietà del valore atteso siano conservate si pone:

Definizione 18. *Dato uno spazio di probabilità* (S, P), *ed una variabile aleatoria discreta* X *si dice valore atteso, o speranza matematica, di* X *il valore*

$$E(X) = \sum_{s \in S} X(s) P(s)$$

se

$$E(|X|) = \sum_{s \in S} |X(s)| P(s) < \infty;$$

in altre parole si richiede la convergenza assoluta della serie che definisce il valore atteso.

Esempio 58. *Per il caso (b) dell'esempio precedente il valore atteso non esiste e si dovranno sviluppare altri metodi.*

Esempio 59. *Se* $X \sim$ *geometrica(p) allora* $X \geq 0$ *e quindi basta che sia finita* $E(X)$ *stessa: derivando per serie, come lecito all'interno del raggio di convergenza di una serie di potenze, si ha:*

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} kp(1-p)^{k-1} \qquad (5.20) \\ &= -\sum_{k=1}^{\infty} p \frac{d}{dp} (1-p)^k \\ &= -\frac{d}{dp} \sum_{k=1}^{\infty} p(1-p)^k \\ &= -p \frac{d}{dp} \frac{1}{p} = \frac{1}{p}. \end{aligned}$$

Intuitivamente, se la probabilità di successo è $p = \frac{1}{m}$, allora il primo successo arriverà in media alla m-sima prova.

Esempio 60. *Se $X \sim$ Poisson(λ) allora $X \geq 0$ e*

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda. \end{aligned} \tag{5.21}$$

Avendo quindi assunto la convergenza assoluta della serie che definisce il valore atteso, valgono tutti i risultati relativi alle proprietà del valore atteso. In particolare valgono i Lemmi 6 e 7. Il lettore è invitato a verificare che le dimostrazioni dei Lemmi suddetti possono essere adattate anche al caso presente.

Come esempio mostriamo come si adatta la dimostrazione della parte (ii) del Lemma 7 nel caso $a_1 = a_2 = 1, X_1 = X, X_2 = Y$. Assumendo che $E(X)$ ed $E(Y)$ esistano, prima si dimostra che $E(X+Y)$ esiste e poi se ne calcola il valore. Per la diseguaglianza triangolare, poichè $X \leq Y$ implica $E(X) \leq E(Y)$ e poiché si può cambiare a piacere l'ordine di sommazione di una serie a termini positivi convergente, si ha:

$$\begin{aligned} E(|X+Y|) &\leq E(|X|+|Y|) \\ &= \sum_{s \in S} (|X(s)|+|Y(s)|)P(s) \\ &= \sum_{s \in S} (|X(s)|P(s) + \sum_{s \in S} (|Y(s)|P(s) \\ &= E(|X|)+E(|Y|) < \infty \end{aligned} \tag{5.22}$$

dall'ipotesi; la finitezza del risultato giustifica a posteriori la riorganizzazione delle somme. Ora

$$\begin{aligned} E(X+Y) &= \sum_{s \in S} (X(s)+Y(s))P(s) \\ &= \sum_{s \in S} (X(s)P(s) + \sum_{s \in S} (Y(s)P(s) \\ &= E(X)+E(Y) < \infty \end{aligned} \tag{5.23}$$

in cui abbiamo riorganizzato nuovamente l'ordine di sommazione in quanto le serie coinvolte sono assolutamente convergenti.

Esercizio 16. *Verificare che le altre dimostrazioni citate si estendono al caso delle variabili discrete.*

Aggiungiamo un'altra semplice conseguenza della diseguaglianza triangolare:

Esercizio 17. *Mostrare che $|E(X)| \leq E(|X|)$.*

Considerando le difficoltà relative ai prodotti infiniti la definizione di indipendenza di una famiglia di variabili aleatorie discrete può essere data riferendosi a sottofamiglie finite:

Definizione 19. *Le variabili di una famiglia al più numerabile di variabili aleatorie $X_1, \dots X_n, \dots$ si dicono indipendenti se sono indipendenti le variabili aleatorie in ogni sottofamiglia finita.*

Esercizio 18. *Verificare che valgono il Lemma 8 ed il teorema 11 nel caso delle variabili discrete.*

Anche per le variabili aleatorie discrete il valore atteso del prodotto di variabili indipendenti è uguale al prodotto dei valori attesi, questa volta però occorre introdurre una condizione sufficiente affinché il valore atteso del prodotto esista. Per questo bisogna conviene premettere la discussione sul valore atteso di funzioni delle variabili aleatorie.

Infatti, se ϕ è una funzione definita su S_X allora $\phi(X)$ è una variabile aleatoria discreta, ma non è detto che anche se X ha valore atteso ci sia il valore atteso di $\phi(X)$.

Esercizio 19. *Mostrare che esistono una variabile aleatoria X ed una funzione ϕ definita su S_X tali che $E(X)$ esiste ma $E(\phi(X))$ non esiste.*

Tuttavia, se il valore atteso di $\phi(X)$ esiste, allora si può calcolare con il cambiamento di variabili:

Esercizio 20. *Verificare che se X è una variabile aleatoria discreta e ϕ è una funzione definita su S_X e se $E(\phi(X))$ ammette valore atteso, allora vale il Teorema 10.*

Ora torniamo alla questione del valore atteso di variabili aleatorie indipendenti. Analogamente a prima, dalla sola esistenza del valore atteso di X ed Y non si può dedurre l'esistenza del valore atteso di XY o viceversa.

Esercizio 21. *Mostrare che esistono variabili aleatorie X ed Y tali che $E(X)$ ed $E(Y)$ esistono ma non esiste $E(XY)$. Viceversa, mostrare che esistono variabili aleatorie X ed Y tali che $E(XY)$ esiste ma $E(X)$ non esiste.*

Occorre quindi condizioni per garantire l'esistenza di altri valori attesi. Un primo risultato dice che i momenti successivi implicano l'esistenza dei momenti precedenti.

Lemma 13. *Se X è una variabile aleatoria discreta tale che $E(X^k)$ esiste, con questo intendendo che $X \geq 0$ oppure $k \in \mathbb{N}$, per qualche k, allora esiste $E(X^h)$ per ogni $h \leq k$, sempre intendendo $X \geq 0$ oppure h intero.*

Dimostrazione.

$$\begin{aligned}
E(|X^h|) &= \sum_{s\in S} |X(s)^h| P(s) \\
&= \sum_{s\in S:|X(s)^h|\le 1} |X(s)^h| P(s) + \sum_{s\in S:|X(s)^h|>1} |X(s)^h| P(s) \\
&\le P(|X^h| \le 1) + \sum_{s\in S:|X(s)^h|>1} |X(s)^k| P(s) \\
&\le P(|X^h| \le 1) + \sum_{s\in S} |X(s)^k| P(s) < \infty
\end{aligned}$$

□

Lemma 14. *Se X ed Y sono variabili aleatorie discrete tali che $E(X^2)$ ed $E(Y^2)$ esistono, allora esistono $E(X)$, $E(Y)$ ed $E(XY)$.*

Dimostrazione. L'esistenza di $E(X)$ ed $E(Y)$ segue dal lemma precedente. Da $0 \le (a-b)^2 = a^2 + b^2 - 2ab$ si ha che per qualsiasi coppia di numeri reali a e b vale che $ab \le (a^2+b^2)/2$, per cui

$$\begin{aligned}
E(|XY|) &= \sum_{s\in S} |X(s)||Y(s)| P(s) \\
&= \frac{1}{2} \sum_{s\in S} (|X(s)|^2 + |Y(s)|^2) P(s) \\
&\le \frac{1}{2} (\sum_{s\in S} |X(s)|^2 P(s) + \sum_{s\in S} |Y(s)|^2 P(s)) < \infty
\end{aligned}$$

□

Ora abbiamo condizioni sufficienti per generalizzare i risultati sui valori attesi di variabili indipendenti.

Esercizio 22. *Verificare che se X_i sono variabili aleatorie discrete tali che $E(X_i^2)$ esiste per ogni i allora vale il Teorema 12.*

Con il secondo momento è quindi possibile definire la varianza e la deviazione standard di X e vale anche in questo caso l'additività delle varianze per variabili indipendenti.

Esercizio 23. *Verificare che se X è una variabile aleatoria discreta tale che $E(X^2)$ esiste allora la varianza e la deviazione standard di X esistono e vale il Teorema 13.*

5.3. Vettori aleatori e variabili aleatorie congiunte

In molti problemi, come quando abbiamo parlato di indipendenza di variabili aleatorie o linearità del valore atteso, si considerano più variabili aleatorie contemporaneamente. In questo caso conviene pensarle come un vettore aleatorio.

Definizione 20. *Dato uno spazio di probabilità (discreta)* (S, P), *un* **vettore aleatorio** $\boldsymbol{X} = (X_1, \ldots, X_n)$ *n-dimensionale è una funzione* $\boldsymbol{X}\colon S \to \mathbb{R}^n$ *tale che* $s \to \boldsymbol{X}(s) = ((X_1(s), \ldots, X_n(s))$.

Si possono ripetere ora molte definizioni e proprietà delle variabili aleatorie, che si ottengono semplicemente sostituendo $\boldsymbol{X}$ ad X.

Definizione 21. *La distribuzione di un vettore aleatorio* $\boldsymbol{X} = (X_1, \ldots, X_n)$, *detta anche* **distribuzione congiunta** *delle* X_i, *è la coppia* $(S_{\boldsymbol{X}}, P_{\boldsymbol{X}})$ *con*

$$S_{\boldsymbol{X}} = \{\boldsymbol{x} \in \mathbb{R}^n |\ \textit{esiste } s \in S : \boldsymbol{X}(s) = \boldsymbol{x}\}$$

e

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P\{s \in S : \boldsymbol{X}(s) = \boldsymbol{x}\}$$

Si noti che per l'usuale proprietà delle funzioni, per ogni $B \subseteq S_{\boldsymbol{X}}$

$$P_{\boldsymbol{X}}(B) = \sum_{\boldsymbol{x} \in B} P_{\boldsymbol{X}}(\boldsymbol{x}).$$

Esercizio 24. *Se* $\boldsymbol{X}$ *è un vettore aleatorio discreto, verificare che* $(S_{\boldsymbol{X}}, P_{\boldsymbol{X}})$ *è uno spazio di probabilità discreto.*

In un vettore aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)$ ciascuna delle componenti X_i è una variabile aleatoria, con una sua distribuzione (S_{X_i}, P_{X_i}), ed è interessante ed utile in vari problemi studiare la relazione tra queste e la distribuzione congiunta $(S_{\boldsymbol{X}}, P_{\boldsymbol{X}})$. Le (S_{X_i}, P_{X_i}) sono dette **distribuzioni marginali**.

Per semplificare la notazione assumeremo che $S_{\boldsymbol{X}} = \mathbb{R}^n$ ponendo $P_{\boldsymbol{X}}(\boldsymbol{x}) = 0$ se $\boldsymbol{x}$ non era originalmente appartenente al codominio di $\boldsymbol{X}$.

Lemma 15. *Per ogni un vettore aleatorio* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *vale*

$$P_{X_1}(x_1) = \sum_{(x_2, \ldots, x_n) \in \mathbb{R}^{n-1}} P_{\boldsymbol{X}}(x_1, \ldots, x_n).$$

Dimostrazione. Per ogni $x_1 \in \mathbb{R}$, gli eventi

$$A_{(x_2, \ldots, x_n)} = \{s \in S : \boldsymbol{x}(s) = (x_1, \ldots, x_n)\} \subseteq S$$

sono disgiunti, per cui

$$\begin{aligned} \sum_{(x_2, \ldots, x_n) \in \mathbb{R}^{n-1}} P_{\boldsymbol{X}}(x_1, \ldots, x_n) &= \sum_{(x_2, \ldots, x_n) \in \mathbb{R}^{n-1}} P(A_{(x_2, \ldots, x_n)}) \\ &= P(\cup_{(x_2, \ldots, x_n) \in \mathbb{R}^{n-1}} A_{(x_2, \ldots, x_n)}) \\ &= P(s : X_1(s) = x_1) = P_{X_1}(x_1) \end{aligned}$$

□

Naturalmente il Lemma precedente si applica ad ogni proiezione o marginale X_i, per cui dalle congiunte si determinano le marginali.

Esercizio 25. *Dimostrare tramite un esempio che l'opposto non è vero e vi sono distribuzioni congiunte diverse che danno luogo alle stesse marginali.*

L'indipendenza delle variabili aleatorie si può esprimere in termini del rapporto tra le distribuzioni congiunta e marginali, nel senso che le componenti di un vettore aleatorio $\boldsymbol{X}= (X_1, \ldots, X_n)$ sono indipendenti se e solo se

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} P_{X_i}(x_i)$$

per tutti gli $\boldsymbol{x}= (x_1, \ldots, x_n) \in \mathbb{R}^n$. Quindi nel caso indipendente le marginali permettono la ricostruzione della distribuzione congiunta.

Se di una o più variabili conosciamo il valore assunto abbiamo delle distribuzioni condizionali.

Definizione 22. *Dato un vettore aleatorio* $\boldsymbol{X}= (X_1, \ldots, X_n)$*, per* $k = 1, \ldots, n$*, se* $(x_1, \ldots, x_k)$ *è tale che* $P(X_1 = x_1, \ldots, X_k = x_k) \neq 0$*, si dice* **distribuzione condizionata o condizionale** *di* $X_{k+1}, \ldots, X_n$ *dato che* $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ *lo spazio di probabilità costituito da*

$$\begin{aligned} &S_{X_{k+1},\ldots,X_n|(X_1,\ldots,X_k)=(x_1,\ldots,x_k)} \\ &\quad = \{(x_{k+1}, \ldots, x_n) : (x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) \in S_{\boldsymbol{X}}\} \end{aligned}$$

e

$$\begin{aligned} &P_{X_{k+1},\ldots,X_n|X_1,\ldots,X_k}((x_{k+1}, \ldots, x_n)|(x_1, \ldots, x_k)) \\ &\quad = \frac{P_{\boldsymbol{X}}(\boldsymbol{x})}{P(X_1 = x_1, \ldots, X_k = x_k)} \end{aligned}$$

Naturalmente si può prendere $S_{X_{k+1},\ldots,X_n|(X_1,\ldots,X_k)=(x_1,\ldots,x_k)} = \mathbb{R}^{n-k}$.

Esercizio 26. *Verificare che la coppia definita nella definizione precedente è uno spazio di probabilità .*

Naturalmente la medesima definizione si poteva dare permutando gli indici. Dal teorema delle probabilità totali si vede che

$$\begin{aligned} &P(X_1 = x_1, \ldots, X_n = x_n) \\ &\quad = \sum_{(x_1,\ldots,x_k)\in S_{X_1,dots,X_k}} P_{X_{k+1},\ldots,X_n|X_1,\ldots,X_k}((x_{k+1}, \ldots, x_n)|(x_1, \ldots, x_k)) \\ &\qquad\qquad \cdot P(X_1 = x_1, \ldots, X_k = x_k). \end{aligned}$$

Per $k = 1$ questa osservazione indica come ricostruire la distribuzione congiunta dalla conoscenza delle distribuzioni condizionate e della relativa marginale.

Definizione 23. *Il valore atteso calcolato rispetto alla probabilità condizionale di una variabile aleatoria date le altre si chiama* **valore atteso condizionale** *e si denota*

$$E(X_n|(X_1,\ldots,X_{n-1}) = (x_1,\ldots,x_{n-1}).$$

5.4. Covarianza e correlazione

Vogliamo introdurre ora una misura della dipendenza di variabili aleatorie meno stringente dell'indipendenza, che richiede la verifica di molte condizioni relative alla fattorizzazione di probabilità. Abbiamo visto che nel caso di variabili aleatorie indipendenti il valore atteso del prodotto si fattorizza, per cui viene naturale di studiare la quantità seguente:

Definizione 24. *Dato un vettore aleatorio* (X,Y)*, ossia due variabili aleatorie e la loro distribuzione congiunta, tali che* $E(X^2), E(Y^2) < \infty$*, si dice* **covarianza** *di* X *ed* Y *il valore*

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Si noti che per il Lemma 14 la definizione è ben posta e che vale $Cov(X,X) = Var(X)$.

Lemma 16.

$$Cov(X,Y) = E((X - E(X)(Y - E(Y)).$$

Dimostrazione. Se $Cov(X,Y)$ esiste allora $E(X^2), E(Y^2) < \infty$ ed anche $E(|XY|), E(|X|), E(|Y|) < \infty$, il che implica che

$$\begin{aligned} E(|X - E(X)||Y - E(Y)|) &\leq E(|XY| + |X||E(Y)| + |Y||E(X)| \\ &\quad +|E(X)||E(Y)|) \\ &\leq E(|XY|) + 3E(|X|)E(|Y|) < \infty \end{aligned}$$

per cui $E((X - E(X))(Y - E(Y))$ esiste. Ora

$$\begin{aligned} E((X - E(X))(Y - E(Y)) &\leq E(XY) - XE(Y) - YE(X) \\ &\quad +E(X)E(Y) \\ &\leq E(XY) + E(X)E(Y) < \infty \end{aligned}$$

□

Confrontando tutte le componenti di un vettore aleatorio si ottiene la seguente:

Definizione 25. *Dato un vettore aleatorio* $\boldsymbol{X} = (X_1,\ldots,X_n)$ *si dice* **matrice di covarianza** *di* $\boldsymbol{X}$ *la matrice* $\Sigma = (\sigma_{i,j})$ *con* $\sigma_{i,j} = Cov(X_i, X_j)$.

Si noti che $\sigma_{i,i} = \sigma_i^2 = Var(X_i)$.

Vediamo ora come cambia la matrice di covarianza per trasformazioni lineari. Dato un vettore aleatorio $\boldsymbol{Z}= (Z_1, \ldots, Z_n)$ con una matrice di covarianza Σ e dati una matrice A che sia $(n \times n)$ ed un vettore ad n componenti b, allora

$$\boldsymbol{X} = A\boldsymbol{Z} + b$$

soddisfa

$$E(\boldsymbol{X}) = \sum_j a_{i,j} + b_j$$

e

$$Cov(X_i, X_j) = \sum_{m,r} a_{i,m} a_{j,r} Cov(Z_m, Z_r).$$

Per cui $E(\boldsymbol{X}) = AE(\boldsymbol{Z}) + b$ e $Cov(\boldsymbol{X}) = ACov(\boldsymbol{Z})A^T$.

Limitiamoci ora al caso di due variabili e consideriamo più in dettaglio il significato della covarianza in relazione alla dipendenza. Dimensionalmente $Cov(X, Y)$ è il prodotto delle dimensioni di X ed Y. Per ottenere un numero puro dividiamo per il prodotto delle deviazioni standard, o equivalentemente, consideriamo la correlazione tra le variabili standardizzate.

Definizione 26. *Dato un vettore aleatorio (X, Y) tale che $E(X^2), E(Y^2) < \infty$, si dice* **correlazione** *di X ed Y il valore*

$$r = r(X, Y) = \frac{Cov(X, Y)}{SD(X)SD(Y)}.$$

Si ricorda che la variabili standardizzata è definita, per una variabile aleatoria discreta X con secondo momento finito, da $\bar{X} = \frac{X-E(X)}{SD(X)}$ e che valgono $E(\bar{X}) = 0$ e $Var(\bar{X}) = E(\bar{X}) = 1 = SD(\bar{X})$

Lemma 17. $r = E(\bar{X}\bar{Y}) = Cov(\bar{X}\bar{Y})$

Dimostrazione. Dalle proprietà di $\bar{X}$ e $\bar{Y}$ si ha

$$\begin{aligned} Cov(\bar{X}\bar{Y}) &= E(\bar{X}\bar{Y}) \\ &= E((\frac{X - E(X)}{SD(X)})(\frac{Y - E(Y)}{SD(Y)})) \\ &= \frac{Cov(X, Y)}{SD(X)SD(Y)} \end{aligned}$$

□

Oltre ad essere un numero puro la correlazione è limitata:

Teorema 17. *Per ogni vettore aleatorio (X, Y) con secondo momento finito si ha*

$$-1 \leq r(X, Y) \leq 1;$$

inoltre $r = 1$ se e solo se $Y = cX + d$ con $c > 0$ ed $r = -1$ se e solo se $Y = -cX + d$ con $c > 0$.

Dimostrazione. La dimostrazione che segue può essere letta in due modi, prima considerando solo i simboli in alto poi quelli in basso nelle coppie $\pm$ e $\mp$. Si ha che

$$\begin{aligned} 0 \leq E(\bar{X} \pm \bar{Y})^2 &= E(\bar{X}^2) + E(\bar{Y}^2) \pm 2E(\bar{X}\bar{Y}) \\ &= 2 \pm 2r(X,Y). \end{aligned} \tag{5.24}$$

Questo implica che

$$-1 \leq r(X,Y) \leq 1;$$

inoltre, se $r(X,Y) = \mp 1$ vale il segno di uguaglianza in (5.24) per cui $E(\bar{X} \pm \bar{Y})^2 = 0$. Poiché $(\bar{X} \pm \bar{Y})^2 \geq 0$ ne discende che $\bar{X} = \mp \bar{Y}$ e di conseguenza

$$Y = \mp \frac{SD(Y)}{SD(X)} X - \frac{SD(Y)}{SD(X)} E(X) + E(Y) = \mp cX + d$$

con $c = \frac{SD(Y)}{SD(X)} \geq 0$. $\square$

CAPITOLO 6

PROBABILITÀ NEL CONTINUO

6.1. Variabili aleatorie continue

In molti problemi l'insieme dei valori che possono essere assunti da una variabile aleatoria può avere la cardinalità del continuo. In questo caso la formalizzazione di spazi di probabilità adeguati non è elementare e qui ci limitiamo a trattare un caso particolare in cui l'insieme è $\mathbb{R}$ e la probabilità è determinata da funzioni continue a tratti; questi spazi di probabilità costituiscono la distribuzione di parecchie variabili aleatorie continue che permettono di studiare numerosi problemi. In un certo senso seguiamo la linea storica di studiare prima le variabili aleatorie continue definite dalla loro distribuzione e rimandare la chiarificazione dei fondamenti ad un'analisi successiva.

Sarebbe possibile utilizzare funzioni anche con una infinità numerabile di punti di discontinuità, ma per semplicità consideriamo il caso in cui questi sono al più un numero finito.

Definizione 27. *Uno* **spazio di probabilità sul continuo** *è una coppia* $(\mathbb{R}, f)$ *con* $\mathbb{R}$ *insieme dei numeri reali ed* $f : \mathbb{R} \to \mathbb{R}$ *una funzione continua tranne al più un numero finito di punti di discontinuità tale che*

(i) $f(x) \geq 0$ *per ogni* $x \in \mathbb{R}$
(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

Si noti che l'integrale nella definizione è ben definito, come integrale improprio. Infatti l'integrale esiste in ogni intervallo $[a, b]$ in cui f è continua; inoltre, detto $D(f) = \{x_1, \ldots, x_n\}$, $x_1 < \cdots < x_n$, l'insieme finito dei punti di discontinuità di f e considerate successioni $\bar{x}_j(0) \searrow_j -\infty$, $x_j(i) \nearrow_j x_i$, $\bar{x}_j(i) \searrow_j x_i$ e $x_j(n) \nearrow_j \infty$ esistono gli integrali in $[\bar{x}_j(i), x_j(i+1)]$, $i = 0, \ldots, n-1$. Per definizione di integrale improprio, (ii) dice che il limite seguente esiste, vale quanto indicato:

$$\lim_{j\to\infty} \sum_{i=0}^{n-1} \int_{\bar{x}_j(i)}^{x_j(i+1)} f(x)dx = 1$$

e non dipende dalle successioni scelte. Allo stesso modo sono definiti gli integrali per ogni intervallo $[a, b]$. Conviene in realtà considerare gli intervalli semiaperti della forma $]a, b]$, $a < b$, $a, b \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ perchè si puo' esprimere un ulteriore intervallo come unione di due disgiunti.

Definizione 28. *In uno spazio di probabilità continuo* $(\mathbb{R}, f)$ *si definiscono* **eventi** *le unioni finite di intervallini semiaperti, aperti o chiusi, e si indica con* $\mathcal{C}$ *la famiglia di tali insiemi. Per ogni* $A = \cup_{i=1}^{n}(a_i, b_i)$ *si pone*

$$P_f(A) = \sum_{i=1}^{n} \int_{a_i}^{b_i} f(x)dx,$$

quali che fossero gli estremi degli intervallini aperti o chiusi.

Proposizione 1. *(i) Per ogni* $a \in \mathbb{R}$, $P(\{a\}) = 0$
(ii) $P([a, b]) = P(]a, b[) = P(]a, b])$.

Dimostrazione. (i) si ottiene dalla monotonia della probabilità che implica che per ogni n

$$P(\{a\}) \leq P(]a - 1/n, a]) = 1/n$$

. La parte (ii) si dimostra dalla (i) poiché le differenze tra gli eventi coinvolti sono costituite da uno o due punti al più . □

Si noti che (i) ed (ii) della definizione di spazio di probabilità continuo sono l'analogo di (1) e (2) nella definizione 14. L'additività numerabile invece è qui una conseguenza delle proprietà di additività dell'integrazione rispetto al dominio. Verificheremo per semplicità solo l'additività finita.

Lemma 18. *In uno spazio di probabilità* $(\mathbb{R}, f)$ *se* $A, B \in \mathcal{C}$ *sono tali che* $\cup_{I \in A} I = \cup_{J \in B} J$ *allora* $P_f(A) = P_f(B)$.

Dimostrazione. Si noti che, in virtù della Proposizione precedente, possiamo utilizzare il tipo di intervallini più conveniente e che l'intersezione di due intervallini semiaperti non disgiunti è un intervallino semiaperto. Se $A = \{I_1, \ldots, I_n\}$ e $B = \{J_1, \ldots, J_k\}$ allora $\{L_{i,j} = I_i \cap J_j, i = 1, \ldots, n, j = 1, \ldots, k\} \in \mathcal{C}$ e per l'additività dell'integrale rispetto al dominio

$$P_f(A) = \sum_{i=1}^{n} \int_{I_i} f(x)dx = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_{L_{i,j}} f(x)dx = \sum_{j=1}^{k} \int_{J_j} f(x)dx = P_f(B)$$

in cui la seconda e la terza uguaglianza dipendono dal fatto che se un intervallo $]a, b]$ è unione finita di intervallini $]a_i, b_i]$ disgiunti allora questi si possono ordinare in modo che $b_i = a_{i+1}$ e

$$\sum_{i=1}^{n} \int_{a_i}^{b_i} f(x)dx = \int_{a}^{b} f(x)dx.$$

□

Uno spazio di probabilità $(\mathbb{R}, f)$ può essere utilizzato per definire la distribuzione di una variabile aleatoria continua.

Definizione 29. *Si dice che una variabile aleatoria continua X ha distribuzione $(\mathbb{R}, f)$ se per ogni $A \in \mathcal{C}$ vale*

$$P(X \in A) = \int_A f(x)dx.$$

In tal caso f è detta la densità di X.

Si noti che l'espressione di sinistra è soltanto formale in quanto non è stato definito lo spazio di probabilità su cui è definito X e su cui agisce P.

Esempio 61. *Una variabile aleatoria con densità $f(x) = \mathbb{I}_{[0,1]}$ è detta uniforme in $[0,1]$. In generale, la variabile aleatoria con densità $f(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}$ è detta* **uniforme** *in $[a,b]$ e si indica $X \sim U([a,b])$.*

Esempio 62. *Una variabile aleatoria con densità $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ è detta Gaussiana o* **normale standard**. *In generale, la variabile aleatoria con densità $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ è detta* **gaussiana o normale** *di parametri μ e σ^2 (ne vedremo nel prossimo capitolo il significato) e si usa la notazione $X \sim N(\mu, \sigma^2)$. Si noti che non esiste una forma elementare per la primitiva della Gaussiana e che solo l'integrale su tutto $\mathbb{R}$ si può calcolare mediante il cambio di variabili in coordinate polari: detto $I = \int_{-\infty}^{+\infty} e^{-x^2/2}dx$ si ha*

$$I^2 = \int_{-\infty}^{+\infty} e^{-x^2/2}dx \int_{-\infty}^{+\infty} e^{-y^2/2}dx = \int_0^{2\pi}\int_0^{+\infty} re^{-r^2/2}drd\theta = 2\pi.$$

6.2. Funzione di distribuzione

Una descrizione delle variabili aleatorie che inizia ad unificare il trattamento delle variabili discrete e continue è la funzione cumulativa o funzione di distribuzione.

Definizione 30. *Si dice funzione* **funzione cumulativa** *o* **funzione di distribuzione** *F_X di una variabile aleatoria X la funzione $F_X : \mathbb{R} \to \mathbb{R}$ tale che*

$$F_X(t) = P(X \leq t).$$

Tale definizione è valida sia per le variabili aleatorie discrete che continue.

Esempio 63. *La funzione di distribuzione di una variabile aleatoria $Y \sim U(\{1, \ldots, 6\})$ vale*

$$F_Y(t) = \mathbb{I}_{t\geq 0}\frac{\lfloor t \rfloor \wedge 6}{6}.$$

Esempio 64. *Se $Z \sim U([a,b])$ allora*

$$F_Z(t) = \mathbb{I}_{t\geq a}\frac{t \wedge b - a}{b-a}.$$

Esempio 65. *In generale i valori della funzione di distribuzione della normale standard* $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx$ *non si calcolano attraverso un integrale esplicito e sono approssimati numericamente: si possono trovare in tutti i programmi statistici per computer o tabulati nei testi di probabilità e statistica.*

Descriviamo ora alcune proprietà delle funzioni di distribuzione, una delle quali differenzia le variabili discrete da quelle continue.

Teorema 18. *Se* X *è una variabile aleatoria (discreta o continua) allora*
(1) F_X *è non decrescente in* t.
(2) $\lim_{t\to-\infty} F_X(t) = 0$
(3) $\lim_{t\to\infty} F_X(t) = 1$
(4) F_X *è continua da destra.*

Inoltre, se X *è una variabile aleatoria discreta allora*
(5a) se $(a, b]$ *non contiene punti di discontinuità di* F_X *allora* F_X *è costante in* $(a, b]$.

Se invece X *è una variabile aleatoria continua con densità* f_X *allora*
(5b) F_X *è continua e* $F'_X(t)$ *esiste per tutti i punti* t *tranne al più un numero finito ed è continua.*

In tutti i punti t *di continuità di* f_X *vale* $F'_X(t) = f_X(t)$

Dimostrazione.

(1) Chiaramente $\{s : X(s) \le t\} \subseteq \{s : X(s) \le u\}$ per tutti i $t \le u$ da cui la diseguaglianza sulle probabilità .

(2) Per le variabili discrete

$$\begin{aligned} \lim_{t\to-\infty} F_X(t) &= \lim_{t\to-\infty} P(X \le t) \\ &= \lim_{t\to-\infty} \sum_{x\in S_X : x\le t} P_X(x) = 0 \end{aligned}$$

in quanto la sommatoria nell'ultima espressione è il resto di una serie convergente; mentre per le variabili continue

$$\lim_{t\to-\infty} F_X(t) = \lim_{t\to-\infty} \int_{-\infty}^{t} f_X(x)dx = 0$$

per definizione di integrale improprio.

(3) Si ha

$$\lim_{t\to\infty} F_X(t) = 1 - \lim_{t\to-\infty} P(X > t) = 1$$

per gli stessi motivi del punto precedente.

(4) Vale $\lim_{u\searrow t} F_X(u) - F_X(t) = \lim_{u\searrow t} P(t < X \le u) = 0$; per variabili discrete in quanto anche in questo caso, dopo aver riordinato la serie, si tratta del resto di una serie convergente. Per le variabili

continue, se f è continua in t allora

$$P(t < X \leq u) = \int_t^u f_X(x)dx = \max\{f(x) : x \in [t,u]\}(t-u) \to 0;$$

altrimenti il risultato vale per definizione di integrale improprio in t.

(5) Poiché F_X è non decrescente esiste $\lim_{u \nearrow t} F_X(u) \leq F_X(t)$ e

$$\begin{aligned} F_X(t) - \lim_{u \nearrow t} F_X(u) &= \lim_{u \nearrow t} P(u < X \leq t) \\ &= P(X = t) - \lim_{u \nearrow t} P(u < X < t) = P(X = t) \end{aligned}$$

per gli stessi motivi di (4).

(5a) Quindi se X è discreta F_X è discontinua in t se e solo se $t \in S_X$ e se $(a,b]$ non contiene punti di discontinuità di F_X allora

$$0 = P(X \in (a,b]) = F_X(b) - F_X(a).$$

(5b) Se invece X è continua allora per il teorema fondamentale del calcolo integrale $F_X(t) = \int_{-\infty}^t f_X(x)dx$ è derivabile nei punti in cui f_X è continua e vale $F'_X(t) = f_X(t)$. □

Vediamo ora che queste proprietà caratterizzano le funzioni che possono essere funzioni di distribuzione per qualche variabile aleatoria.

Lemma 19. *Se una funzione soddisfa (1)-(4) e (5a) allora è la funzione di distribuzione di una variabile aleatoria discreta; se invece soddisfa (1)-(4) e (5b) è la funzione di distribuzione di una variabile continua.*

Dimostrazione. Da (1) F è non decrescente quindi in ogni punto esiste il limite sinistro $F(t^-)$ e destro $F(t^+)$ e da (2) e (3) è limitata. Per questo l'insieme S dei punti di discontinuità di F è al più numerabile: infatti un punto di discontinuità è caratterizzato dal fatto che $F(t^-) \neq F(t^+)$ e di punti tali che $F(t^+) - F(t^-) \geq 1/n$ ce ne sono al più n; l'unione su n di tali insiemi dà tutti i punti di discontinuità . Inoltre segue da (4) che $F(t) = F(t^+)$.

Se vale (5a) possiamo porre $P(x) = F(t) - F(t^-) > 0$ per ogni $x \in S$. Se indichiamo con $\ldots, x_{-n}, \ldots, x_0, \ldots, x_n, \ldots$ i punti di S ordinati, allora dalla (5a) $F(x_{n-1}) = F(x_{n-1}^+) = F(x_n^-)$ per cui da (2) e (3) segue che:

$$\begin{aligned} \sum_{x \in S} P(x) &= \sum_{N \in \mathbb{Z}} F(x_n) - F(x_n^-) \\ &= \sum_{N \in \mathbb{Z}} F(x_n) - F(x_{n-1}) \\ &= \lim_{n \to \infty} (F(x_n) - F(x_{-n})) = 1. \end{aligned} \tag{6.25}$$

Risulta poi

$$\sum_{x \in S, x \leq t} P(x) = \lim_{n \to \infty} (F(x) - F(x_{-n})) = F(x). \tag{6.26}$$

Se invece vale (5b) si può porre $f(t) = F'(t)$ per tutti i t in cui F è derivabile ed un valore qualunque altrove. Dalla (1) $f \geq 0$ e dalle (2) e (3)

$$\int_{-\infty}^{\infty} f(t)dt = \lim_{t\to\infty}(F(t) - F(-t)) = 1.$$

Infine,

$$\int_{-\infty}^{x} f(t)dt = F(x) - \lim_{t\to\infty} -F(-t) = F(x)$$

per cui F è la funzione di distribuzione di una variabile aleatoria di densità f.

□

Definizione 31. *Si dice che due variabili aleatorie continue X ed Y sono uguali in distribuzione, e si indica $X =^d Y$, se $F_X(t) = F_Y(t)$ per ogni t.*

6.3. Variabili esponenziali, beta e gamma

Oltre alle variabili uniformi e gaussiane, vi sono altre distribuzioni che sono risultate rilevanti in vari contesti applicativi.

Distribuzione esponenziale. Un esempio importante di variabile aleatoria continua ha densità

$$f(t) = \lambda e^{-\lambda t}\mathbb{I}_{t\geq 0}$$

detta **distribuzione esponenziale** $exp(\lambda)$. Si ha

$$F_X(t) = 1 - e^{-\lambda t}$$

e

$$P(X > t + T|X > t) = e^{-\lambda T}$$

una proprietà che si esprime come perdita di memoria. Per questa caratteristica la distribuzione esponenziale è usata come modello per il tempo di vita di componenti elettronici.

In particolare questo modello potrebbe essere usato per analizzare i tempi di funzionamento dei componenti elettronici di cui all'esempio 2. Per completare l'analisi di questo esempio si devono però considerare le somme di tempi di vita indipendenti, ossia somme di variabili aleatorie continue indipendenti e questo verrà sviluppato nei prossimi capitoli.

Vedremo in seguito che gli intertempi tra arrivi modellizzati da una distribuzione di Poisson sono esponenziali.

Distribuzione gamma Una variabile aleatoria X con densità

$$f_X(x) = \frac{\lambda(\lambda x)^{k-1}}{\Gamma(k)}e^{-\lambda x}$$

per $x \geq 0$, $\lambda > 0$ e $k \geq 1$, è detta avere distribuzione gamma e si indica $X \sim \Gamma(k, \lambda)$. Si è preso

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$$

per cui la densità gamma è ben definita. Si noti che nel caso $k = 1$ si ottiene la distribuzione esponenziale e in generale si vedrà che nel caso k intero ha distribuzione $\Gamma(k, \lambda)$ la somma

$$T_k = X_1 + \ldots X_k$$

di k variabili aleatorie esponenziali indipendenti di parametro λ, mentre $\Gamma(k) = (k-1)!$. Tale somma si può interpretare come il tempo di attesa di k eventi indipendenti in serie.

Distribuzione Beta Si dice che una variabile aleatoria ha **distribuzione Beta** $B(q, r)$ di parametri q ed r se ha densità

$$f_B(x) = cx^{q-1}(1-x)^{r-1}$$

per $x \in [0, 1]$. Determiniamo ora il valore della costante c. Calcoliamo l'integrale seguente, nel quale operiamo il cambiamento di variabili $w = tx$, $z = t(1-x)$ con inversa $t = w + z$ e $x = \frac{w}{w+z}$ e determinante Jacobiano della trasformazione $\frac{1}{z+w}$:

$$\begin{aligned}
\Gamma(q+r) \int_0^1 f_B(x) dx &= \int_0^\infty t^{q+r-1} e^{-t} dt \int_0^\infty cx^{q-1}(1-x)^{r-1} dx \\
&= c \int_0^\infty \int_0^\infty e^{-(w+z)} w^{q-1} (z)^{r-1} dw dz \\
&= c \quad \Gamma(q)\Gamma(r).
\end{aligned}$$

Per cui $c = \frac{\Gamma(q+r)}{\Gamma(q)\Gamma(r)}$. Ora è facile notare che

$$m_1 = \frac{\Gamma(q+r)}{\Gamma(q)\Gamma(r)} \frac{\Gamma(q+1)\Gamma(r)}{\Gamma(q+1+r)} = \frac{q}{q+r}$$

ed analogamente $m_2 = \frac{(q+1)q}{(q+r+1)(q+r)}$.

6.4. Trasformazioni di densità per variabili continue

Descriviamo come cambiano le densità di variabili aleatorie prendendo una funzione di forma semplice di una data variabile aleatoria.

Teorema 19 (Trasformazione di densità). *Consideriamo la funzione $g(X)$ dove X è una variabile aleatoria discreta e $g : \mathbb{R} \to \mathbb{R}$ monotona (crescente o decrescente). Anche $Y = g(X)$ è una variabile aleatoria e la sua densità di probabilità è*

$$p_Y(y) = P(Y = y) = P(g^{-1}(Y) = g^{-1}(y)) = p_X(g^{-1}(y)).$$

Se poi X è continua con densità f_X ed una funzione $\phi : \mathbb{R} \to \mathbb{R}$ continua e strettamente monotona con inversa $x = \phi^{-1}(y)$, si ha:

$$f_Y(y) = f_X(\phi^{-1}(y))|\frac{d}{dy}\phi^{-1}(y)|$$

per ogni y tranne al più i punti in cui f_Y e $f_X \circ \frac{d}{dy}\phi^{-1}$ sono discontinui, che sono al più un numero finito di punti.

Dimostrazione. Supponiamo ϕ crescente. Per ogni $a, b \in \mathbb{R}$, $a < b$ si ha, effettuando nell'integrazione della terza riga il cambiamento di variabili $y = \phi(x)$ con inversa $x = \phi^{-1}(y)$ e quindi $dx = \frac{d}{dy}\phi^{-1}(y)dy$:

$$\begin{aligned} \int_a^b f_Y(y)dy &= P(Y \in]a, b]) \\ &= P(\phi^{-1}(Y) \in]\phi^{-1}(a), \phi^{-1}(b)]) \qquad (6.27) \\ &= \int_{\phi^{-1}(a)}^{\phi^{-1}(b)} f_X(x)dx \\ &= \int_a^b f_X(\phi^{-1}(y))\frac{d}{dy}\phi^{-1}(y)dy. \end{aligned}$$

Se a è un punto di continuità per f_Y e per $f_X \circ \phi^{-1}$ allora per il teorema del valor medio, se $[a, b]$ è piccolo a sufficienza da non contenere punti di discontinuità di f_Y e per $f_X \circ \phi^{-1}$:

$$\frac{1}{(b-a)}\int_a^b f_Y(y)dy = f_Y(\alpha(a, b))$$

e

$$\frac{1}{(b-a)}\int_a^b f_X(\phi^{-1}(y))\frac{d}{dy}\phi^{-1}(y) = f_X(\phi^{-1}(\beta(a, b)))\frac{d}{dy}\phi^{-1}(\beta(a, b))$$

con $\alpha(a, b), \beta(a, b) \in [a, b]$; poiché l'uguaglianza (6.27) vale per ogni intervallo $]a, b]$ si ha:

$$\begin{aligned} f_Y(a) &= \lim_{b \to a^+} f_Y(\alpha(a, b)) \\ &= \lim_{b \to a^+} f_X(\phi^{-1}(\beta(a, b)))\frac{d}{dy}\phi^{-1}(\beta(a, b)) \\ &= f_X(\phi^{-1}(a)\frac{d}{dy}\phi^{-1}(a) \end{aligned}$$

quindi $f_Y(y) = f_X(\phi^{-1}(y))\frac{d}{dy}\phi^{-1}(y)$ per ogni y tranne al più i punti in cui f_Y e $f_X \circ \frac{d}{dy}\phi^{-1}$ sono discontinui, che sono al più un numero finito di punti.

Se invece ϕ è decrescente è facile verificare che tutti gli intervalli di integrazione risultano invertiti e quindi per riportare l'usuale espressione degli estremi di integrazione dal minore al maggiore si cambia il segno in $\frac{d}{dy}\phi^{-1}$. □

Osservazione 9. *La trasformazione delle densità è facile da ricordare pensando alla densità di X come $f_X(x)dx$ per cui se $Y = \phi(X)$ con inversa $X = \phi^{-1}(Y)$ allora si considera il cambiamento di variabili $y = \phi(x)$ con inversa $x = \phi^{-1}(y)$ e quindi $dx = \frac{d}{dy}\phi^{-1}(y)dy$ per cui il risultato precedente si ottiene sostituendo formalmente i valori x e dx così ottenuti.*

Esempio 66. *Se $X \sim N(0,1)$ e $Y = \sigma X + \mu$ allora $X = \phi^{-1}(Y) = \frac{Y-\mu}{\sigma}$ e*

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Distribuzione log normale Si dice log normale una variabile aleatoria Y tale che $\ln Y$ sia $N(\lambda, \zeta)$ (con ζ deviazione standard). Passando alla variabile standardizzata, $\frac{\ln Y - \lambda}{\zeta} \sim N(0,1)$ per cui

$$P(a < Y \leq b) = \Phi(\frac{\ln b - \lambda}{\zeta}) - \Phi(\frac{\ln a - \lambda}{\zeta}).$$

Questa variabile aleatoria è un buon modello per tutte quelle situazioni in cui si vengono a moltiplicare, anziché sommare, variabili i.i.d., per esempio nelle variazioni dei valori o della quotazione delle azioni in borsa. Questo perché $\ln Y$ diviene in questo caso una somma ed un meccanismo tipo il TCL porta alla normalità di $\ln Y$. La densità di una variabile aleatoria log normale si può calcolare attraverso la trasformazione di variabili descritta sopra. Infatti con $x = \ln(y) = \phi^{-1}(y)$, ossia $\phi(x) = e^x$, e $X \sim N(\lambda, \zeta)$ si ha

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\zeta} e^{-\frac{(\ln y - \lambda)^2}{2\zeta^2}} y^{-1}.$$

Se g non è monotona e $g^{-1}(y) = \{x_1, \ldots, x_k\}$ assume più valori allora si devono sommare le formule precedenti sui valori assunti:

$$p_Y(y) = \sum_{i=1}^{k} p_X(x_i)$$

e

$$f_Y(y) = \sum_{i=1}^{k} f_X(x_i) |\frac{d}{dy}\phi_i^{-1}(y)| = \sum_{i=1}^{k} f_X(x_i) |\frac{1}{(\frac{d\phi}{dx}(x_i))}|,$$

dove ϕ_i è l'i-sima componente dell'inversa di ϕ.

Esempio 67. *Se $X \sim N(0,1)$, la densità di $Y = X^2$ si calcola osservando che $g^{-1}(y) = \sqrt{y}$ per cui la derivata vale $\frac{d}{dy}g^{-1}(y) = \pm\frac{1}{2y^{1/2}}$ e si ha:*

$$\begin{aligned} f_Y(y) &= \sum x = \sqrt{y}, -\sqrt{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{2y^{1/2}} \\ &= \frac{1}{\sqrt{2\pi} y^{1/2}} e^{-y/2}. \end{aligned}$$

Si noti che questa deve necessariamente essere la densità di una $\Gamma(1/2, 1/2)$ *che formalmente è* $f(t) = \frac{(1/2)^{1/2}t^{1/2-1}}{\Gamma(1/2)}e^{-t/2}$ *visto che hanno la stessa forma della dipendenza funzionale ed entrambe le densità hanno integrale* 1*; questo ci ha condotto alla scoperta del valore* $\Gamma(1/2) = \sqrt{\pi}$.

6.5. Valori attesi per variabili aleatorie continue

Definiamo ora il valore atteso di una variabile aleatoria continua.

Definizione 32. *Data una variabile aleatoria continua* X *con densità* f *si dice* **valore atteso** *di* f*:*

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

se esiste

$$E(|X|) = \int_{-\infty}^{\infty} |x| f(x) dx < \infty.$$

Per il valore atteso di una funzione di una variabile aleatoria continua si vede immediatamente che vale l'espressione analoga a quella usata nel caso discreto.

Teorema 20. *Data una variabile aleatoria continua* X *con densità* f_X *ed una funzione reale di variabile reale* g *monotona si ha*

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

se esiste

$$E(|g(X)|) = \int_{-\infty}^{\infty} |g(x)| f(x) dx < \infty.$$

Dimostrazione. Dalla definizione di valore atteso, posto $Y = g(X)$, basta fare nuovamente il cambiamento di variabile $x = g^{-1}(y)$ con $dy = dx/|\frac{d}{dy}g^{-1}(y)|$ nell'ultimo integrale

$$\begin{aligned} E(g(X)) &= E(Y) = \int_{\mathbb{R}} y f_Y(y) dy \\ &= \int_{\mathbb{R}} y f_X(g^{-1}(y)) |\frac{d}{dy} g^{-1}(y)| dy \\ &= \int_{\mathbb{R}} g(x) f_X(x) dx \end{aligned}$$

con la condizione sull'esistenza di $E(|g(X)|)$ che si ottiene applicando quanto visto alla funzione $|g(x)|$ che è monotona quando lo è g. □

Con questa definizione valgono tutte le proprietà viste per il valore atteso, la varianza e la deviazione standard di variabili aleatorie.

Esempio 68. *Calcolare valore atteso e SD della distribuzione uniforme su* $[a, b]$.

Esempio 69. *Se* $X \sim N(0,1)$ *allora* $E(X) = 0$ *e* $SD(X) = 1$. *Infatti*

$$E(X) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x e^{-x^2/2} dx = 0$$

essendo la funzione dispari; inoltre integrando per parti questo integrale noto si ha:

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} ([x e^{-x^2/2}]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} x^2 e^{-x^2/2} dx) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2/2} dx) \end{aligned}$$

da cui, essendo $E(X) = 0$,

$$Var(X) = E(X^2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-x^2/2} dx = 1.$$

Esercizio 27. *Verificare che se* $X \sim exp(\lambda)$ *allora*

$$E(X) = \int_{-\infty}^{\infty} t\lambda e^{-\lambda t} dt = 1/\lambda,$$

$$Var(X) = \int_{-\infty}^{\infty} t^2 \lambda e^{-\lambda t} dt - E(X)^2 = 1/(\lambda)^2,$$

e $SD(X) = 1/\lambda$.

6.6. Distribuzione congiunta di due variabili aleatorie

Finora abbiamo considerato variabili aleatorie continue isolatamente, ora iniziamo a trattare il caso di due variabili considerate congiuntamente. Il caso di n variabili non conterrà differenze sostanziali tranne la più complessa formalizzazione e la necessità di trattazione algebrica. Per semplicità limiteremo i possibili insiemi di discontinuità nella prossima definizione a rette.

Definizione 33. *La* **densità congiunta** *di due variabili aleatorie* X *ed* Y *è una funzione* $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$, *continua tranne al più un insieme costituito da un numero finito di rette, tale che*

(i) $f_{X,Y}(x, y) \geq 0$ *per ogni* $(x, y) \in \mathbb{R}^2$

e

(ii) $\int_{\mathbb{R}^2} f_{X,Y}(x, y) dx dy = 1$.

Per $f = f_{X,Y}$ $(\mathbb{R}^2, f)$ *è detto spazio di probabilià continua in* $\mathbb{R}^2$ *e anche* **distribuzione congiunta** *di* X *ed* Y; *per ogni rettangolo* $A = (a_1, b_1) \times (a_2, b_2)$ *aperto o chiuso con* a_i *e* b_i *finiti o infiniti si ha*

$$P((X, Y) \in A) = \int_A f_{X,Y}(x, y) dx dy.$$

La funzione $F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = P((X,Y) \in (-\infty, x] \times (-\infty, y])$ *è detta* **funzione di distribuzione congiunta**.

Si noti che gli integrali utilizzati in questa definizione sono integrali di Riemann (eventualmente impropri) in dimensione 2 e sono ben definiti essendo le funzioni coinvolte continue su insiemi normali. Inoltre, nei punti in cui la funzione di distribuzione $F_{X,Y}$ è differenziabile, dal teorema fondamentale del calcolo integrale si ha

$$fX, Y(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}.$$

Si possono ripetere ora molte definizioni e proprietà delle distribuzioni congiunte discrete.

Definizione 34. *Data la distribuzione congiunta di due variabili aleatorie* X *ed* Y*, la distribuzione di* X *si ottiene considerando la* **distribuzione marginale** *la cui densità* f_X *soddisfa*

$$f_X(x) = \int_{\mathbb{R}} f(X,Y)dy.$$

Da questo punto di vista f_X *e l'analoga* f_Y *sono dette* **densità marginali***di* X *ed* Y *rispettivamente.*

Per cui anche nel caso continuo dalle congiunte si determinano le distribuzioni marginali.

Esempio 70. *Sia* $f = c \cdot \mathbb{I}_T$ *in cui* T *è il triangolo*

$$\{\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}.$$

Dalla (ii) della definizione di spazio continuo si ha

$$\begin{aligned} 1 &= \int_{\mathbb{R}^2} f(x,y)dxdy \\ &= c\int_{\mathbb{R}} (\int_{\mathbb{R}} \mathbb{I}_T(x,y)dy)dx \\ &= c\int_0^1 (\int_0^{1-x} dy)dx \\ &= c\int_0^1 (1-x)dx = c/2 \end{aligned} \tag{6.28}$$

da cui $c = 2$*, poichè essendo i domini di integrazione normali e l'integranda continua l'integrale si può calcolare tramite integrali ripetuti. La marginale di* X *si ottiene da*

$$f_X(x) = \int_{\mathbb{R}} f(x,y)dy = 2(1 - x_1).$$

Data una funzione $\Phi : \mathbb{R}^2 \to \mathbb{R}$ ed una coppia di variabili aleatorie X ed Y di cui sia nota la densità congiunta $f_{X,Y}$, la composizione $\Phi(X,Y)$ è una variabile aleatoria. Analogamente a quanto provato per

una singola variabile aleatoria continua si prova, con le proprietà della trasformazione di densità discusse più avanti, che il valore atteso di una funzione di X ed Y vale

$$E(\Phi(X,Y)) = \int_{\mathbb{R}^2} \Phi(x,y) f_{X,Y}(x,y)dxdy. \tag{6.29}$$

Anche qui vale l'additività del valore atteso

Lemma 20. *Data una coppia di variabili aleatorie X ed Y si ha*

$$E(X+Y) = E(X) + E(Y).$$

Dimostrazione.

$$\begin{aligned} E(X+Y) &= \int_{\mathbb{R}^2} (x+y) f_{X,Y}(x,y)dxdy \\ &= \int_{\mathbb{R}} (\int_{\mathbb{R}} x f_{X,Y}(x,y)dy)dx + \int_{\mathbb{R}} (\int_{\mathbb{R}} y f_{X,Y}(x,y)dx)dy \\ &= \int_{\mathbb{R}} x f_X(x)dx + \int_{\mathbb{R}} x f_Y(y)dy = E(X) + E(Y) \end{aligned}$$

□

L'indipendenza delle variabili aleatorie continue si può esprimere in termini del rapporto tra densità congiunta e marginali, nel senso che due variabili aleatorie X ed Y sono **indipendenti** se e solo se

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

per tutti $(x,y) \in \mathbb{R}^2$. Quindi anche per le variabili continue le marginali permettono la ricostruzione della distribuzione congiunta quando le variabili siano indipendenti.

Lemma 21. *Se X ed Y sono variabili aleatorie indipendenti allora*

$$E(XY) = E(X)E(Y).$$

Dimostrazione.

$$\begin{aligned} E(XY) &= \int_{\mathbb{R}^2} xy f_{X,Y}(x,y)dxdy \\ &= \int_{\mathbb{R}^2} x f_X(x) y f_Y(y)dxdy \\ &= \int_{\mathbb{R}} x f_X(x)dx \int_{\mathbb{R}} y f_Y(y)dy = E(X)E(Y) \end{aligned}$$

□

Questo implica che anche per le variabili aleatorie continue indipendenti vale che $Var(X+Y) = Var(X) + Var(Y)$.

Esempio 71. *Se $X \sim N(0,1)$ e $Y = \sigma X + \mu$ allora per le proprietà del valore atteso*

$$E(Y) = E(\sigma X + \mu) = \mu$$

e

$$Var(Y) = E((\sigma X + \mu - \mu)^2) = \sigma^2 E(X) = \sigma^2$$

così che $SD(Y) = \sigma$. Una tale variabile aleatoria Y è detta normale con media μ e varianza σ^2, che si indica con $Y \sim N(\mu, \sigma^2)$.

Anche per le variabili continue è possibile definire le distribuzioni condizionali.

Definizione 35. *Date variabili aleatorie continue X ed Y con densità congiunta $f_{X,Y}$ si definisce* **densità condizionata** *o condizionale di X dato Y la funzione*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

per tutti i valori $x \in \mathbb{R}$ e $y \in \mathbb{R}$ tali che $f_Y(y) \neq 0$.

Per un generico evento ammissibile A scriveremo

$$P(X \in A|Y = y) = \int_{\mathbb{A}} f_{X|Y}(x|y)dy.$$

Definizione 36. *Il valore atteso calcolato rispetto alla probabilità condizionale di una variabile aleatoria date le altre si chiama* **valore atteso condizionale** *o condizionato e si denota*

$$E(X|Y = y) = \int_{\mathbb{R}} x f_{X|Y}(x|y)dy.$$

Dalla definizione di densità condizionali si ottiene subito la seguente versione del teorema delle probabilità totali:

$$f_X(x) = \int_{\mathbb{R}} f_{X|Y}(x|y) f_Y(y)dy$$

e quindi

$$P(X \in A) = \int_{\mathbb{R}} \int_A f_{X|Y}(x|y) f_Y(y)dydx = \int_{\mathbb{R}} P(X \in A|Y = y) f_Y(y)dy.$$

Esempio 72. *Verifichiamo che se X ed Y sono $Exp(\lambda)$ e indipendenti allora $X + Y \sim \Gamma(2, \lambda)$. Infatti,*

$$\begin{aligned} f_{X+Y}(t) &= \int_{-\infty}^{+\infty} \lambda e^{-\lambda x} \mathbb{I}_{x>0} \lambda e^{-\lambda(t-x)} \mathbb{I}_{t-x>0} dx \\ &= \lambda^2 e^{-\lambda t} \int_0^t dx = t\lambda^2 e^{-\lambda t}. \end{aligned}$$

In generale, se $X \sim \Gamma(k,\lambda)$ ed $Y \sim \Gamma(h,\lambda)$ sono indipendenti allora $X+Y \sim \Gamma(k+h,\lambda)$, integrando per parti si ha

$$\begin{aligned}
f_{X+Y}(t) &= \int_{-\infty}^{+\infty} \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x} \mathbb{I}_{x>0} \frac{\lambda^h (t-x)^{h-1}}{(h-1)!} e^{-\lambda(t-x)} \mathbb{I}_{t-x>0} dx \\
&= \lambda^{k+h} e^{-\lambda t} \int_0^t \frac{x^{k-1}(t-x)^{h-1}}{(k-1)!(h-1)!} dx \\
&= \lambda^{k+h} e^{-\lambda t} ([\frac{x^k (t-x)^{h-1}}{k!(h-1)!}]_0^t + \int_0^t \frac{x^k (t-x)^{h-2}}{(k)!(h-2)!} dx) \\
&= \lambda^{k+h} e^{-\lambda t} \int_0^t \frac{x^{k-1+h-1}}{(k-1+h-1)!} dx \\
&= \frac{t^{k+h-1} \lambda^{k+h}}{(k+h-1)!} e^{-\lambda t}.
\end{aligned}$$

Esempio 73. *(Distribuzione $\chi^2(n)$)Dalle proprietà di somme di variabili aleatorie indipendenti con distribuzione gamma, e da quanto visto sulla distribuzione di X^2 se $X \sim N[0,1]$ abbiamo che se $X_i \sim N(0,1)$ sono i.i.d. allora $T = \sum_{i=1}^n X_i^2 \sim \Gamma(n/2, 1/2)$. Questa distribuzione è detta* **chi-quadro con n gradi di libertà**, *si indica con $T \sim \chi^2(n)$ e si è rivelata molto utile in varie analisi statistiche.*

Esempio 74. *Verifichiamo che se $X, Y \sim N(0,1)$ indipendenti allora $X+Y \sim N(0,2)$, indicando le varianze (oppure $X+Y \sim N(0,\sqrt{2})$ indicando le deviazioni standard). Infatti, separando l'esponente in un quadrato con resto e cambiando poi la variabile d'integrazione in $\sqrt{2}x - t/\sqrt{2}$, si ha:*

$$\begin{aligned}
f_{X+Y}(t) &= \int_{-\infty}^{+\infty} \frac{1}{2\pi} e^{-x^2/2} e^{-(t-x)^2/2} dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-(\sqrt{2}x - t/\sqrt{2})^2/2 - t^2/4} dx = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-t^2/4}.
\end{aligned}$$

In generale, se $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ indipendenti, allora $X+Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ come si vede ora. Nel prossimo calcolo completeremo il quadrato all'esponente ottenendo delle costanti (ossia funzioni di σ_X ecc. ma non della variabile di integrazione x e della variabile della funzione t), ma il loro valore sarà irrilevante, quindi non lo calcoliamo. Potremo calcolare quanto sarà il valore di tali costanti direttamente nell'ultima formula in quanto sappiamo che $\int_{-\infty}^{+\infty} f_{X+Y}(t)dt = 1$. Per calcolare l'integrale nella terza riga si effettua il cambiamento di variabili $z = ax + bt$ ed il risultante integrale si

calcola essendo l'integranda una gaussiana.

$$\begin{aligned} f_{X+Y}(t) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_X} \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-(x-\mu_X)^2/(2\sigma_X)} e^{-(t-x-\mu_Y)^2/(2\sigma_Y)} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-(\alpha x^2+\beta xt+\gamma t^2)/2-\gamma' t^2} dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-(ax+bt)^2/2-\gamma' t^2} dx \\ &= e^{-\gamma' t^2} \frac{1}{2\pi} \int_{-\infty}^{+\infty} c e^{-z^2/2} dz = d e^{-\gamma' t^2}. \end{aligned}$$

Abbiamo quindi concluso che la densità di $X+Y$ è gaussiana. Ora, poiché sappiamo che $E(X+Y) = E(X)+E(Y) = \mu_X+\mu_Y$ e $Var(X+Y) = Var(X) + Var(Y) = \sigma_X^2 + \sigma_Y^2$, quest'ultima uguaglianza discendendo dall'indipendenza, necessariamente $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ e la forma delle costanti d e γ' è determinata.

Relazioni con la distribuzione di Poisson. Date variabili aleatorie i.i.d. $X_i \sim Exp(\lambda)$ ed un valore $T > 0$ consideriamo la probabilità che

$$\{N_T = k\} = \{\sum_{i=1}^{k} X_i < T < \sum_{i=1}^{k+1} X_i\}$$

che si può interpretare dicendo che il $k+1$-simo tempo di vita è quello che permette al totale di superare T, ossia che il numero N_T di interruzioni prima di T è esattamente k. Dalla formula delle probabilità totali per le densità condizionate continue si ha

$$P(N_T = k) = \int_0^T \frac{\lambda(\lambda x)^{k-1}}{\Gamma(k)} e^{-\lambda x} e^{-\lambda(T-x)} dx = \frac{(\lambda T)^k}{k!} e^{-\lambda(T)}$$

ossia N_T ha distribuzione di Poisson(λT).

6.7. Vettori aleatori continui

Anche per il caso continuo introduciamo i vettori aleatori, naturalmente attraverso le loro densità congiunta. Per semplicità ammettiamo insiemi di possibili discontinuità lineari.

Definizione 37. *Uno spazio di probabilità continua in $\mathbb{R}^n$ è una coppia $(\mathbb{R}^n, f)$ in cui $f : \mathbb{R}^n \to \mathbb{R}$ è una funzione continua tranne al più un insieme costituito da un numero finito di iperpiani tale che*

(i) $f(\boldsymbol{x}) \geq 0$ per ogni $\boldsymbol{x} \in \mathbb{R}^n$.

(ii) $\int_{\mathbb{R}^n} f(\boldsymbol{x}) d\boldsymbol{x} = 1$

Si dice che un vettore aleatorio continuo $\boldsymbol{X} = (X_1, \ldots, X_n)$ n-dimensionale ha distribuzione $(\mathbb{R}^n, f_{\boldsymbol{X}})$ se per ogni plurirettangolo $A = \prod_{i=1}^n (a_i, b_i]$ si ha

$$P(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}.$$

La distribuzione di un vettore aleatorio $\boldsymbol{X} = (X_1, \ldots, X_n)$ *è detta anche* **distribuzione congiunta** *delle* X_i *e la funzione* $f_{\boldsymbol{X}}(\boldsymbol{x})$ *è detta* **densità congiunta**.

Si noti che gli integrali utilizzati in questa definizione sono integrali di Riemann (eventualmente impropri) in dimensione n e sono ben definiti essendo le funzioni coinvolte continue su insiemi normali. Si possono ripetere ora molte definizioni e proprietà dei vettori aleatori discreti.

DEFINIZIONE 38. *In un vettore aleatorio continuo* $\boldsymbol{X} = (X_1, \ldots, X_n)$, *ciascuna delle sue componenti* X_i *è una variabile aleatoria continua, con una sua distribuzione* $(\mathbb{R}, f_{X_i})$ *ottenuta come* **distribuzione marginale** *da*

$$f_{X_1}(x_1) = \int_{\mathbb{R}^{n-1}} f(\boldsymbol{x}) dx_2 \cdot \ldots \cdot dx_{n-1}.$$

Naturalmente la definizione precedente si applica ad ogni proiezione o marginale X_i, per cui anche nel caso continuo dalle congiunte si determinano le distribuzioni marginali.

ESEMPIO 75. *Sia* $f = c \cdot \mathbb{I}_T$ *in cui* T *è il triangolo*

$$\{\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2 : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}.$$

Dalla (ii) della definizione di spazio continuo si ha

$$\begin{aligned} 1 &= \int_{\mathbb{R}^2} f(\boldsymbol{x}) d\boldsymbol{x} \\ &= c \int_{\mathbb{R}} (\int_{\mathbb{R}} \mathbb{I}_T(\boldsymbol{x}) dx_2) dx_1 \\ &= c \int_0^1 (\int_0^{1-x_1} dx_2) dx_1 \\ &= c \int_0^1 (1 - x_1) dx_1 = c/2 \end{aligned} \tag{6.30}$$

da cui $c = 2$, *poichè essendo i domini di integrazione normali e l'integranda continua l'integrale si può calcolare tramite integrali ripetuti. La marginale di* X_1 *si ottiene da*

$$f_{X_i}(x_1) = \int_{\mathbb{R}^{n-1}} f(\boldsymbol{x}) dx_2 = 2(1 - x_1).$$

Data una funzione $\Phi : \mathbb{R}^n \to \mathbb{R}$ ed un vettore aleatorio $\boldsymbol{X}$, la composizione $\Phi(\boldsymbol{X})$ è una variabile aleatoria. Analogamente a quanto provato per una singola variabile aleatoria continua si prova, con le proprietà della trasformazione di densità discusse più avanti, che il valore atteso di una funzione di $\boldsymbol{X}$ si calcola

$$E(\Phi(\boldsymbol{X})) = \int_{\mathbb{R}^n} \Phi(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}.$$

LEMMA 22. *Dato il vettore aleatorio* $\boldsymbol{X}= (X_1, \ldots, X_n)$ *si ha*

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i).$$

DIMOSTRAZIONE. Utilizzando l'induzione su n, è sufficiente dimostrare l'asserzione per $n = 2$:

$$\begin{aligned} E(X_1 + X_2) &= \int_{\mathbb{R}^2} (x_1 + x_2) f(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathbb{R}} (\int_{\mathbb{R}} x_1 f(\boldsymbol{x}) dx_2) dx_1 + \int_{\mathbb{R}} (\int_{\mathbb{R}} x_2 f(\boldsymbol{x}) dx_1) dx_2 \\ &= \int_{\mathbb{R}} x_1 f_{X_1}(x_1) dx_1 + \int_{\mathbb{R}} x_2 f_{X_2}(x_2) dx_2 = E(X_1) + E(X_2) \end{aligned}$$

□

L'indipendenza delle variabili aleatorie continue si può esprimere in termini del rapporto tra densità congiunta e marginali, nel senso che le componenti di un vettore aleatorio $\boldsymbol{X}= (X_1, \ldots, X_n)$ sono indipendenti se e solo se

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

per tutti $\boldsymbol{x}= (x_1, \ldots, x_n) \in \mathbb{R}^n$. Anche per le variabili continue le marginali permettono la ricostruzione della distribuzione congiunta quando le variabili siano indipendenti.

LEMMA 23. *Se le variabili aleatorie continue* X_i, $i = 1, \ldots, n$ *sono indipendenti allora*

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i).$$

DIMOSTRAZIONE.

$$\begin{aligned} E(\prod_{i=1}^{n} X_i) &= \int_{\mathbb{R}^n} \prod_{i=1}^{n} x_i f(\boldsymbol{x}) d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \prod_{i=1}^{n} (x_i f_{X_i}(x_i)) d\boldsymbol{x} \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}^n} x_i f_{X_i}(x_i) dx_i = \prod_{i=1}^{n} E(X_i) \end{aligned}$$

□

Anche per le variabili continue è possibile definire le distribuzioni condizionali.

DEFINIZIONE 39. *Dato un vettore aleatorio* $\boldsymbol{X}= (X_1, \ldots, X_n)$, *per* $k = 1, \ldots, n$, *se* $(x_1, \ldots, x_k)$ *è tale che la densità marginale soddisfa* $f_{(X_1,\ldots,X_k)}(x_1, \ldots .x_k) \neq 0$, *si dice* **distribuzione condizionata** *di* $X_{k+1}, \ldots, X_n$ *dato che* $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ *lo spazio di probabilità costituito da* $\mathbb{R}^{n-k}$ *con densità*

$$f_{(X_{k+1},\ldots,X_n)|(X_1,\ldots,X_k)}((x_{k+1}, \ldots, x_n)|(x_1, \ldots, x_k)) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{(X_1,\ldots,X_k)}(x_1, \ldots .x_k)}$$

DEFINIZIONE 40. *Il valore atteso calcolato rispetto alla probabilità condizionale di una variabile aleatoria date le altre si chiama* **valore atteso condizionale** *e si denota*

$$E(X_n|(X_1, \ldots, X_{n-1}) = (x_1, \ldots, x_{n-1})).$$

Le definizioni di $Cov(X, Y)$ e $r(X, Y)$ sono le stesse del caso discreto e vale che se X ed Y sono indipendenti allora $Cov(X, Y) = 0$. In generale definiamo

DEFINIZIONE 41. *Dato un vettore aleatorio* $\boldsymbol{X}= (X_1, \ldots, X_n)$, *per* $k = 1, \ldots, n$, *si definisce (il vettore del) valore atteso di* $\boldsymbol{X}$: $E(\boldsymbol{X}) = (E(X_1), \ldots E(X_n))$, *e la matrice di covarianza* $\Sigma = Cov(\boldsymbol{X}) = (a_{i,j}) = (Cov(X_i, X_j))$

6.8. Distribuzioni normali multivariate

Abbiamo visto che se X ed Y sono normali indipendenti allora $X + Y$ è normale e naturalmente lo è anche aX per ogni costante a (esercizio). Quindi se Z_i, $i = 1, \ldots, n$ sono normali indipendenti allora lo è anche qualsiasi combinazione lineare $\sum_{i=1}^{n} a_i Z_i$. Definiamo quindi come normali multivariati i vettori aleatori che risultano da una tale trasformazione:

DEFINIZIONE 42. *Un vettore aleatorio* $\boldsymbol{X}$ *si dice* **normale multivariato** *se esiste un vettore aleatorio* $\boldsymbol{Z} = (Z_i, \ldots, Z_n)$ *tale che le* Z_i *sono i.i.d.* $N(0, 1)$, *una matrice* A *ed un vettore* b *tali che* $\boldsymbol{X} = A\boldsymbol{Z}+b$. *In questo caso* $\mu = E(\boldsymbol{X}) = b$ *e* $\Sigma = Cov(\boldsymbol{X}) = AA^T$ *(una matrice* $(n \times n)$*) e si indica* $\boldsymbol{X} \sim N(\mu, \Sigma) = N(b, AA^T)$.

Quindi i vettori con distribuzione normale multivariata sono idenfiticati dalla matrice A oppure dalla matrice di covarianza Σ, ma le due rappresentazioni sono equivalenti, in quanto $\Sigma = AA^T$ è simmetrica e semidefinita positiva e data una matrice σ con queste caratteristiche esiste sempre una matrice A tale che $\sigma = AA^T$.

E' quindi possibile considerare anche nel caso continuo variabili i.i.d. tali che esistono valore atteso e varianza, e quindi formulare il Teorema Centrale del Limite. Enunciamo senza dimostrazione il teorema, che vale anche in questo caso.

TEOREMA 21 (Teorema Centrale del Limite per variabili continue). *Per variabili aleatorie continue i.i.d.* $X_1, \ldots, X_n$ *tali che* $E(X_i^2) < \infty$ *vale per ogni* $a \in \mathbb{R}$

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n X_i - nE(X_1)}{\sqrt{nVar(X_1)}} \leq a\right) = \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

6.9. Trasformazioni di densità di vettori continui

Iniziamo discutendo della densità di una funzione di due variabili aleatorie X ed Y in un caso particolare:

TEOREMA 22. *Date due variabili aleatorie* X *ed* Y *con densità congiunta continua* $f_{X,Y}$*, ed una funzione* $g : \mathbb{R}^2 \to \mathbb{R}$ *che per ogni* $y \in \mathbb{R}$ *sia derivabile con continuità in* x *con derivata di segno costante allora la densità di* $g(X, Y)$ *soddisfa:*

$$\begin{aligned} f_{g(X,Y)} &= \int_{\mathbb{R}} f_{X,Y}(g^{-1}(t,y), y) \left| \frac{\partial g^{-1}(t,y)}{\partial t} \right| dy \\ &= \int_{\mathbb{R}} f_{X,Y}(g^{-1}(t,y), y) \frac{1}{\left| \frac{\partial g(x,y)}{\partial x} \right|} dy \end{aligned}$$

DIMOSTRAZIONE. Assumiamo inizialmente che $\frac{\partial g(x,y)}{\partial x} > 0$ e calcoliamo la funzione di distribuzione di $g(X, Y)$:

$$\begin{aligned} F_{g(X,Y)}(t) &= P(g(X,Y) \leq t) = \int\int_{x,y:g(x,y)\leq t} f_{X,Y}(x,y) dx dy \\ &= \int_{\mathbb{R}} dy \int_{-\infty}^{g^{-1}(t,y)} f_{X,Y}(x,y) dx \end{aligned}$$

in cui nell'ultima riga si è utilizzata l'integrazione ripetuta in quanto g è invertibile per ogni valore fissato di y. Derivando si ha:

$$f_{g(X,Y)} = \frac{d}{dt} F_{g(X,Y)}(t) = \int_{\mathbb{R}} dy f_{X,Y}(g^{-1}(t,y), y) \frac{\partial g^{-1}(t,y)}{\partial t}.$$

L'asserzione si ottiene dalle proprietà della derivata della funzione inversa e da un semplice cambiamento di segno nel caso la derivata sia negativa. □

ESERCIZIO 28. *Osservare che per ogni coppia di variabili aleatorie* X *ed* Y *la densità di* $X + Y$ *è* $\int_{\mathbb{R}} f_{X,Y}(t-y, y) dy$*, che nel caso che* X *ed* Y *siano indipendenti diventa* $\int_{\mathbb{R}} f_X(t-y) f_Y(y) dy$*. Confrontare con il calcolo che abbiamo fatto tramite la versione continua del teorema delle probabilità totali.*

ESERCIZIO 29. *Osservare che per ogni coppia di variabili aleatorie* X *ed* Y *la densità di* XY *è* $\int_{\mathbb{R}} f_{X,Y}(t/y, y) \frac{1}{|y|} dy$ *e quella di* X/Y *è* $\int_{\mathbb{R}} f_{X,Y}(ty, y)|y| dy$*.*

Concludiamo con il risultato relativo alla trasformazione di densità per i vettori aleatori.

TEOREMA 23 (Trasformazione di densità per vettori aleatori). *Data un vettore aleatorio* $\boldsymbol{X}$ *n dimensionale continuo con densità* $f_{\boldsymbol{X}}$ *ed un'applicazione* $\phi : \mathbb{R}^n \to \mathbb{R}^n$ *continua e tale che il determinante dello Jacobiano* $\det(J(\phi(\boldsymbol{X})) \neq 0$, *con* $J(\phi(\boldsymbol{X})) = (\frac{\partial \phi_i}{\partial x_j})_{i,j=1,\ldots,n}$, *quindi invertibile con inversa* $\boldsymbol{x} = \phi^{-1}(\boldsymbol{y})$ *si ha:*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\phi^{-1}(\boldsymbol{y}))|\det(J(\phi^{-1}(\boldsymbol{Y}))|$$

in tutti i punti di continuità di $f_{\boldsymbol{Y}}$.

DIMOSTRAZIONE. Per ogni plurirettangolo D si ha, effettuando nell'integrazione della terza riga il cambiamento di variabili $\boldsymbol{y} = \phi(\boldsymbol{x})$ con inversa $\boldsymbol{x} = \phi^{-1}(\boldsymbol{y})$:

$$\begin{aligned} \int_D f_{\boldsymbol{Y}}(\boldsymbol{y})d\boldsymbol{y} &= P(\boldsymbol{Y} \in D) \\ &= P(\phi^{-1}(\boldsymbol{Y}) \in \phi^{-1}(D)) \\ &= \int_{\phi^{-1}(D)} f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_D f_{\boldsymbol{X}}(\phi^{-1}(\boldsymbol{y}))|\det(J(\phi^{-1}(\boldsymbol{Y}))|d\boldsymbol{y}. \end{aligned} \tag{6.31}$$

La conclusione segue per il teorema del valor medio come nel caso delle variabili aleatorie continue. □

ESERCIZIO 30. *Date matrici normali multivariate, se la matrice* A, *o equivalentemente* Σ, *è invertibile allora si può scrivere la densità congiunta.*

CAPITOLO 7

STATISTICA DESCRITTIVA

La statistica descrittiva si occupa del rapporto tra gli avvenimenti con risultati non esattamente predicibili del mondo reale e la loro rappresentazione in termini scientificamente utilizzabili. Ad esempio, quando si si fanno osservazioni o si registrano i risultati di esperimenti la rappresentazione delle informazioni ottenute in termini trasmissibili ed analizzabili richiede il prendere molte decisioni spesso altamente opinabili. Per esempio, quando si analizzano i dati relativi a migliaia di pazienti per studiare certe patologie o certe terapie, spesso questi sono scarabocchiati su fogli da medici diversi e vanno letti (e qui già ci vorrebbe l'aiuto di un medico), interpretati, trasferiti in forma unificata indovinando possibili differenze tra i diversi redattori e spesso rendendo numeriche delle valutazioni qualitative. Poi i dati vanno accuratamente trascritti in un data base, a quel punto analizzati per cercare delle indicazioni significative; una volta che ci siano delle osservazioni apparentemente sensate queste vanno giustificate e successivamente presentate.

E' chiaro che tutto questo è un processo lungo e delicato e che vi sono innumerevoli articoli e trattati che lo studiano. Noi qui ci occuperemo brevemente solo di un aspetto particolare: supponendo di avere dati già in forma numerica ci poniamo il problema di generare un grafico o alcuni valori rappresentativi.

Supponiamo quindi di avere un *campione* $x_1, \ldots, x_n \in \mathbb{R}$ e poniamo

$$x_{min} = \min\{x_i, i = 1, \ldots, n\}$$

e

$$x_{max} = \max\{x_i, i = 1, \ldots, n\}$$

così che $[x_{min}, x_{max}]$ è il range del campione.

7.1. L'istogramma

L'*istrogramma* dei dati, detto usualmente istogramma, è una rappresentazione grafica dei dati che rappresenta frequenze attraverso le aree. Descriveremo ora una modalità di realizzazione degli istogrammi, che come si vedrà prevede comunque numerose scelte arbitrarie.

Dividiamo $[\lfloor x_{min} \rfloor, \lfloor x_{max} \rfloor + 1]$ in intervallini scegliendo

$$\lfloor x_{min} \rfloor = d_0 < d_1 < \cdots < d_m = \lfloor x_{max} \rfloor + 1$$

e formando intervallini semiaperti a destra (scegliendoli semiaperti a destra si semplificano leggermente le definizioni):

$$[d_k, d_{k+1}[$$

per $k = 0, \ldots, m-1$.

Determinati gli intervallini, o separati comunque i dati in classi, per ogni intervallo o classe k si dice:

- **frequenza assoluta** $|\{i : x_i \in [d_k, d_{k+1}[\}|$,
- **frequenza relativa**$|\{i : x_i \in [d_k, d_{k+1}[\}|/n$

e **istogramma** la funzione che per $x \in [d_k, d_{k+1}[$ vale

$$I(x) = \frac{|\{i : x_i \in [d_k, d_{k+1}[\}|}{n(d_{k+1} - d_k)}.$$

Inoltre si chiama **frequenza cumulativa** o **funzione di ripartizione empirica** in x la funzione $F(x) = |\{i : x_i \leq x\}|/n$

Esercizio 31. *Con dati* $0, 1, 2, 2, 1, 1, 0, 2, 1$ *si disegnino gli istogrammi ottenuti con la scelta* $d_0 = 0, d_1 = 1, d_2 = 3$ *e con la scelta* $d_0 = 0, d_1 = 1, d_2 = 2, d_3 = 3$.

Osservazione 10. *Attraverso l'istogramma possiamo capire e mostrare dove i dati si concentrano di più di meno, anche se l'arbitrarietà della scelta degli intervalli e del comportamento agli estremi degli intervalli lascia spazio a molti fraintendimenti e possibili rappresentazioni falsate.*

Osservazione 11. *Se cominciamo ad immaginare, come faremo nella statistica inferenziale, che i dati siano realizzazioni di un qualche meccanismo casuale, ad esempio variabili aleatorie continue indipendenti, allora una delle cose che possono essere suggerite dall'istogramma è la forma della densità di tali variabili aleatorie.*

7.2. Elementi di Statistica Descrittiva

Per descrivere sinteticamente i dati si utilizzano le definizioni seguenti. Per dati univariati o comunque per ogni singolo tipo di dato:

(1) **valor medio empirico** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
(2) **mediana** ogni m tale $|\{i : x_i \leq m\}|/n \geq 50\%$ e $|\{i : x_i \geq m\}|/n \geq 50\%$
(3) **quantile-p** con $p \in [0, 1]$ ogni y_p tale che $|\{i : x_i \leq y_p\}|/n \geq p$ e $|\{i : x_i \geq y_p\}|/n \geq 1 - p$
(4) **1° (o 3°) quartile** il quantile-0.25 (o 0.75)
(5) **varianza empirica** $\sigma_n^2 = Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$;
(6) **SD empirica** $SD = \sigma_n = \sigma_n(x) = \sqrt{\sigma_n^2}$;
(7) **scarto interquartile** $y_{0.75} - y_{0.25}$;
(8) **campo di variazione** $x_{max} - x_{min}$;
(9) **MAD** la mediana di $|x_i - \bar{x}|$;
(10) **momento empirico k-simo** $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$;

(11) **dati standardizzati** $z_i = \frac{x_i - \bar{x}}{\sigma_n}$;
per motivi che vedremo si utilizzano spesso σ_{n-1}^2 e σ_{n-1}, definiti come i precedenti con $n-1$ al posto di n. Si noti che la $m = y_{0.5}$ e che

LEMMA 24. *Media e mediana minimizzano rispettivamente lo scarto quadratico medio e lo scarto assoluto medio.*

DIMOSTRAZIONE. Si ha che

$$\frac{\partial}{\partial t}\sum_{i=1}^{n}(x_i - t)^2 = 2(\sum_{i=1}^{n} x_i - nt) = 0$$

implica $t = \bar{x}$, che è un punto di minimo essendo la funzione nonnegativa e divergente asintoticamente. Invece se m è una mediana

$$\sum_{i=1}^{n} |x_i - t| - \sum_{i=1}^{n} |x_i - m| = \sum_{i=1,\dots,n:x_i \le m} (t-m)$$
$$+ \sum_{i=1,\dots,n:x_i \ge t} (m-t) + \sum_{i=1,\dots,n:m<x_i<t} (t - x_i - |x_i - m|) \ge 0$$

essendo che per $x_i \ge m$ si ha $t - m \ge |x_i - m|$. □

Per coppie di dati (x_i, y_i):

(1) **diagramma di dispersione**: il grafico dei punti (x_i, y_i) in $\mathbb{R}^2$;
(2) **correlazione** si considerano: $r = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sigma_n(x)} \frac{y_i - \bar{y}}{\sigma_n(y)}$

e la regressione che ora descriviamo.
Ipotizziamo che $y_i = a + bx_i + r_i$ con r_i un errore relativamente piccolo e tale abbia media nulla per ogni valore fissato di x siano circa nulli; questo significa che detto $\bar{y}_i|x_i = \frac{\sum_{j=1,\dots,n:x_j=x_i} y_j}{\{j:x_j=x_i\}}$ sia ha $\bar{y}_i|x_i \approx a + bx_i$. Determinando a e b in modo che minimizzino lo scarto quadrato medio $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ si ottiene: minimizzando prima in a a b fissato: $a = \frac{1}{n}\sum_{i=1}^{n}(y_i - bx_i) = \bar{y} - b\bar{x}$ per quanto visto sopra sulla media empirica. Sostituendo e derivando:

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - \bar{y} + b(\bar{x} - x_i))^2 = \sum_{i=1}^{n}(\bar{x} - x_i)(y_i - \bar{y} + b(\bar{x} - x_i)) = 0$$

implica

$$b = \frac{\sum_{i=1}^{n}(\bar{x} - x_i)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)},$$

che è un punto di minimo essendo la funzione nonnegativa e divergente asintoticamente, e quindi $a = \bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}$. Per cui il modello lineare prevede

$$y_i = \bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x} + \frac{Cov(x,y)}{Var(x)}x_i$$

o anche

$$\frac{y_i - \bar{y}}{SD(y)} = r\frac{x_i - \bar{x}}{SD(x)}$$

che si memorizza dicendo che in variabili standardizzate il coefficiente della retta interpolatrice è r. Quindi una rappresentazione dei dati si ottiene dalla **retta di regressione** di equazione:

$$y = \bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x} + \frac{Cov(x,y)}{Var(x)}x$$

oppure in variabili standardizzate

$$y = rx.$$

Si dicono poi **residui** i valori $r_i = y_i - a - bx_i$ e, ponendo $C_{x,y} = Cov(x,y)$ e $\sigma_x^2 = Var(x) = \sigma_n(x)$, **varianza residua del modello lineare**:

$$\begin{aligned}
V_M &= \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 \\
&= \sum_{i=1}^{n}\big((y_i - \bar{y})^2 + (\bar{y} - a - bx_i)^2 + 2(y_i - \bar{y})(\bar{y} - a - bx_i)\big) \\
&= \sum_{i=1}^{n}\left((y_i - \bar{y})^2 + (\frac{C_{x,y}}{\sigma_x^2}(\bar{x} - x_i))^2 + 2(y_i - \bar{y})(\frac{C_{x,y}}{\sigma_x^2}(\bar{x} - x_i))\right) \\
&= \sigma_y^2 + \frac{C_{x,y}^2}{\sigma_x^2} - 2\frac{C_{x,y}^2}{\sigma_x^2} = \sigma_y^2 - \frac{C_{x,y}^2}{\sigma_x^2} \leq \sigma_y^2.
\end{aligned}$$

Per cui la varianza residua del modello lineare è sempre minore della varianza originale delle y: $\sigma_y^2 - V_M$ è detta **varianza spiegata** e può essere una misura dell'efficiacia del modello lineare, per cui si definisce **coefficiente di determinazione** il valore:

$$R^2 = \frac{(\sigma_y^2 - V_M)}{\sigma_y^2} = \frac{C_{x,y}^2}{\sigma_x^2\sigma_y^2} = r^2.$$

Per cui si trova un nuovo significato di r come misura della capacità del modello lineare di predire la variabile dipendente.
Si chiama poi **regressione non lineare** l'uso della regressione lineare non tra le variabili osservate x ed y, ma tra loro funzioni. Se si sostituisce x con $g(x)$ il modello si adatta immediatamente facendo la regressione $y = a + bg(x)$. Mentre è più interessante il caso in cui si sostituisce y con $f(y)$, per una funzione f invertibile. Si può allora considerare la regressione $f(y) = a + bx$ e stimare a e b, ottenendo la predizione $y_i = f^{-1}(a + bx_i)$ e la varianza del modello $V_M = \sum_{i=1}^{n}(y_i - f^{-1}(a + bx_i))^2$. Si noti che ora non è detto che $V_M \leq \sigma_y^2$, ossia che $R^2 \geq 0$, anche se naturalmente un valore negativo vuol dire che il modello è peggio di predire y_i con $\bar{y}$.

CAPITOLO 8

STATISTICA INFERENZIALE

Immaginiamo ora che i dati siano realizzazioni di un qualche meccanismo casuale descrivibile nell'ambito del calcolo delle probabilità. Consideriamo un caso semplice, in cui i dati del campione che stiamo considerando siano realizzazioni di variabili aleatorie indipendenti ed identicamente distribuite, discrete o continue a seconda che i dati osservati siano numeri interi o meno.

Esempio 76. *Possiamo immaginare che i dati dell'esempio 31 di cui sono stati fatti gli istogrammi nel paragrafo precedente siano realizzazioni di variabili i.i.d.* $B(2, p)$*, ossia ognuno dei dati corrisponda al numero di successi in 2 prove indipendenti ognuna con probabilità di successo* p*; oppure che siano* $B(n, p)$*.*

8.1. Stima di parametri

Una volta decisa la distribuzione delle variabili aleatorie (vedremo poi qualche metodo per prendere meglio questa decisione) spesso rimangono dei parametri da determinare, come ad esempio la probabilità di successo p nell'ultimo esempio o uno o più parametri di una distribuzione continua. A questo scopo sono stati sviluppati vari metodi: presentiamo ora quello detto metodo dei momenti.
In generale consideriamo che la distribuzione dipenda da l parametri $\theta_1, \ldots, \theta_l$. Vogliamo determinare dei valori

$$(\hat{\theta}_1, \ldots, \hat{\theta}_l) = (\hat{\theta}_1(x_1, \ldots, x_n), \ldots, \hat{\theta}_l(x_1, \ldots, x_n))$$

che siano stime ragionevoli dei valori veri dei parametri.
Ci sono poi vari metodi di valutazione degli stimatori basati sul fatto che le x_i sono pensate come realizzazioni di variabili aleatorie. Si dice **non distorto** uno stimatore se

$$E((\hat{\theta}_1(X_1, \ldots, X_n), \ldots, \hat{\theta}_l(X_1, \ldots, X_n)) = (\theta_1, \ldots, \theta_l).$$

8.2. Il metodo dei momenti

Si considerano allora i momenti k-simi della distribuzione $E(X^k) = E_{\theta_1,\ldots,\theta_l}(X^k)$ e si uguagliano ai momenti k-simi empirici corrispondenti, formando un numero di equazioni sufficiente a permettere di determinare la stima dei parametri ponendo $(\hat{\theta}_1, \ldots, \hat{\theta}_l)$ uguale alle soluzioni del sistema composto di tali equazioni $m_k = E(X^k)$.

Esempio 77. *Con i dati dell'esercizio 31 si può stimare la probabilità p della binomiale con il metodo dei momenti. Ora se $X \sim B(2,p)$ allora $E(X) = 2p$ ed $m_1 = \bar{x} = 10/9$, per cui $\hat{p} = 10/18$.*

Esempio 78. *Se invece interpretiamo i dati $-2, 3, 1$ come realizzazioni di variabili aleatorie i.i.d. continue $X_i \sim U([-a,a])$ allora $f_{X_i}(x) = \frac{1}{2a}\mathbb{I}_{[-a,a]}(x)$ e $E(X_i) = 0$ per simmetria e*

$$E(X_i^2) = \frac{1}{2a}\int_{-a}^{a} x^2 dx = \frac{a^2}{3}$$

per cui $\hat{a} = \sqrt{14} \approx 3.74$. Se $\hat{a}$ fosse risultato minore di x_{max} ne avremmo concluso che o il metodo di stima dei parametri non era adeguato o le assunzioni fatte non costituivano un buon modello per i dati.

Vediamo ora le principali distribuzioni, i loro parametri e la relazione con alcuni momenti idonei alla stima.

Distribuzione	Densità o funzione di distribuzione	parametri	relazione con i momenti
Binomiale $B(n,p)$	$P_{n,p}(k) = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0,\dots,n$	p	$m_1 = np$
Geometrica $Geom(p)$	$P_p(k) = p(1-p)^{k-1}$, $k = 0,\dots$	p	$m_1 = 1/p$
Poisson $Poiss(\lambda)$	$P_\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k = 0,\dots$	λ	$m_1 = \lambda$
Esponenziale $Exp(\lambda)$	$f_\lambda(t) = \lambda e^{-\lambda t}$, $t \geq 0$	λ	$m_1 = 1/\lambda$
Gamma $\Gamma(k,\lambda)$	$f_{k,\lambda}(t) = \frac{\lambda(\lambda t)^{k-1}}{\Gamma(k)}e^{-\lambda t}$, $t \geq 0$	k, λ	$m_1 = k/\lambda$ $m_2 = \frac{k(k+1)}{\lambda^2}$
Gaussiana $N(\mu,\sigma)$	$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $t \in \mathbb{R}$	μ, σ	$m_1 = \mu$ $m_2 = \sigma^2 + \mu^2$
Lognormale $LogN(\lambda,\zeta)$	$f_{\lambda,\zeta}(t) = \frac{1}{\sqrt{2\pi}\zeta t}e^{-\frac{(\ln t-\lambda)^2}{2\zeta^2}}$, $t \geq 0$	λ, ζ	$m_1 = e^{\lambda+\zeta^2/2}$ $m_2 = e^{2(\lambda+\zeta^2)}$
Uniforme $U(a,b)$	$f_{U(a,b)}(t) = \frac{1}{b-a}$, $t \in [a,b]$	a, b	$m_1 = \frac{a+b}{2}$ $m_2 = \frac{(a+b)^2}{3}$
Beta $B(q,r)$	$f_{B(q,r)}(x)$ $= \frac{\Gamma(q+r)}{\Gamma(q)\Gamma(r)}x^{q-1}(1-x)^{r-1}$, $t \geq 0$	q, r	$m_1 = \frac{q}{q+r}$ $m_2 = \frac{(q+1)q}{(q+r+1)(q+r)}$

Tabella 1. Principali distribuzioni, parametri e momenti.

8.3. Stima di parametri: il metodo della massima verosimiglianza

Un altro metodo di stima puntuale dei parametri si basa sul determinare i valori dei parametri che massimizzano la probabilità o la densità di probabilità delle osservazioni: tale probabilità è detta verosimiglianza.

DEFINIZIONE 43. *Date osservazioni $x_1, \ldots, x_n$ ed un modello basato su una distribuzione discreta $P_{\theta_1,\ldots,\theta_k}$ o continua con densità $f_{(X_1,\ldots,X_n),(\theta_1,\ldots,\theta_k)}$ con parametri $\theta_1, \ldots, \theta_k$ si dice* **verosimiglianza** *il valore:*

$$L_{\theta_1,\ldots,\theta_k}(x_1, \ldots, x_n) = P_{\theta_1,\ldots,\theta_k}(X_1 = x_1, \ldots, X_n = x_n)$$

nel caso discreto o

$$L_{\theta_1,\ldots,\theta_k}(x_1, \ldots, x_n) = f_{(X_1,\ldots,X_n),(\theta_1,\ldots,\theta_k)}(x_1, \ldots, x_n)$$

nel caso continuo.

ESEMPIO 79. *Se le X_i sono ipotizzate i.i.d. continue con densità $f_{\theta_1,\ldots,\theta_k}$ allora $L_{\theta_1,\ldots,\theta_k}(x_1, \ldots, x_n) = \prod_{i=1}^n f_{\theta_1,\ldots,\theta_k}(x_i)$.*

DEFINIZIONE 44. *Date osservazioni $x_1, \ldots, x_n$ si dice stima di massima versomiglianza dei parametri $\theta_1, \ldots, \theta_k$ l'argomento di $max_{\theta_1,\ldots,\theta_k} L_{\theta_1,\ldots,\theta_k}(x_1, \ldots, x_n)$.*

Poiché la verosimiglianza è spesso un prodotto si utilizza spesso il logaritmo, per cui in generale la stima di massima verosimiglianza si ottiene risolvendo

$$\frac{\partial}{\partial \theta_j} \log L_{\theta_1,\ldots,\theta_k}(x_1, \ldots, x_n) = 0.$$

ESEMPIO 80. *Determinare nel caso di distribuzioni esponentiali (λ) la stima di massima verosimiglianza di λ e confrontare con la stima del metodo dei momenti.*

ESEMPIO 81. *Determinare nel caso di distribuzioni lognormali (λ, ζ) la stima di massima verosimiglianza di λ e ζ e confrontare con la stima del metodo dei momenti.*

8.4. Test di ipotesi

Ci poniamo ora in una prospettiva diversa. Immaginiamo di avere un modello teorico per un esperimento con risultati casuali e dei dati raccolti dopo la realizzazione dell'esperimento. Ci potremmo chiedere se i dati confermino o meno il modello teorico, ma da un punto di vista scientifico (si veda per esempio K. Popper, Logica della scoperta scientifica, Einaudi, Torino, 1970) è più sensato chiedersi all'opposto se i dati contraddicano o meno l'ipotesi teorica. Questo è il senso del test di ipotesi.

In generale, un modello teorico contiene molte ipotesi, anche non esplicitate: per esempio, quando discutiamo del risultato del lancio di dadi stiamo assumendo che in ogni esperimento riusciamo a leggere il risultato e che i risultati stessi siano riportati in modo fedele e non per esempio alterati dopo la realizzazione dell'esperimento, ecc. ecc.
Qui presenteremo una versione semplificata di questa problematica, in cui si testa semplicemente il valore dei parametri di una distribuzione o, tutt'al più , la scelta della distribuzione stessa.
Lo schema che presentiamo è detto test di ipotesi e si realizza come segue:

- (0) Si fissa un **livello del test**, detto α, tipo $\alpha = 0.05$ oppure $\alpha = 0.01$.
- (1) Si descrive l'ipotesi che si vuole testare, detta **Ipotesi Nulla** H_0, ed un'**ipotesi alternativa**, con cui l'ipotesi nulla viene confrontata, indicata con H_1. Per semplificare assumiamo che le due ipotesi identifichino una famiglia di distribuzioni ed una partizione (Θ_0, Θ_1) dell'insieme dei possibili parametri Θ di queste distribuzioni. Si noti che, in generale, H_1 sarà molto più vaga di H_0, riflettendo l'impostazione filosofica precedentemente esposta.
- (2) Si determina una certa funzione $Q = Q(x_1, \ldots, x_n; \theta) \in \mathbb{R}$ dei dati e dei parametri con la proprietà che per ogni distribuzione identificata dall'ipotesi nulla e per ogni $\theta \in \Theta_0 \quad Q(X_1, \ldots, X_n; \theta)$, con X_i variabili aleatorie i.i.d. distribuite secondo la distribuzione scelta, ha una distribuzione nota, almeno approssimativamente. Si determina poi per ogni $\theta \in \Theta_0$ la **statistica del test** $a(\theta) = Q(x_1, \ldots, x_n; \theta)$. La statistica del test viene indicata con lettere tipo z, t, χ^2 ecc.
- (3) Su questa base siamo ora in grado di calcolare la probabilità o la densità di probabilità dell'evento $Q = a$. Tuttavia, abbiamo visto che questa (densità di) probabilità non è utile ai fini di valutare H_0. Si cercano quindi degli insiemi $I(a) \subset \mathbb{R}$ (generalmente intervalli o unioni di intervalli) con la proprietà che
  - $P(Q \in I(a))$ è monotono in a;
  - $P(Q \in I(a)) >> 0$ per certi valori di a;
  - per ogni distribuzione identificata dall'ipotesi nulla e per ogni $\theta \in \theta_0$, $P(Q \in I(a))$ tende a 0 quando a tende ad un estremo del condominio $S_{Q(X_1,\ldots,X_n;\theta)}$ di Q.

  e si definisce **valore** p **del test** o p-value il valore

  $$p = sup_{\theta \in \Theta_0} P_\theta(Q(X_1, \ldots, X_n; \theta) \in I(a(\theta)))$$

  dove $a(\theta)$ è la statistica del test.

- (4) Si confrontano p e α. Se $p < \alpha$ si rigetta H_0 e si assume H_1 ed altrimenti non si rigetta.
- (3') Si può ottenere una procedura alternativa al calcolo del valore p invertendo la parte finale del ragionamento precedente. Una volta fissato il livello del test α ogni possibile realizzazione dei dati porta ad un valore di a tale che H_0 si rigetta o meno, e quindi si può definire la **regione di rigetto**

  $D = D(\alpha) = \{(x_1, \ldots, x_n) : \ H_0 \text{ si rigetta al livello } \alpha\}$
- (4') Per determinare se rigettare o meno H_0 si verifica semplicemente se per i dati osservati vale $(x_1, \ldots, x_n) \in D(\alpha)$

Si osservi che in un test si può

- Erroneamente rigettare H_0, detto errore di prima specie;
- Erroneamente accettare H_0, detto errore di seconda specie.

Per ridurre la probabilità di errore di prima specie occorre ridurre α e per controllare l'errore di seconda specie si deve aumentare α.
Vediamo ora vari esempi di test.

Osservazione 12. *Qualche testo costruisce il test di ipotesi partendo senza definire preventivamente il livello del test e quindi direttamente dalla regione di rigetto D. A quel punto il livello del test dipende da D ed è definita come*

$$\alpha(D) = sup_{\theta \in \Theta_0} P_\theta((X_1, \ldots, X_n) \in D).$$

Test per la media di osservazioni gaussiane a varianza nota. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)$ con σ^2 nota e vogliamo testare il valore di μ che assumiamo essere un certo valore μ_0 mentre i dati sembrano suggerire un valore più piccolo; vogliamo limitare l'errore di prima specie. Allora ha senso porre

- $\alpha = 0.1$
- H_0: $\mu = \mu_0$ e H_1: $\mu < \mu_0$;
- $z = \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma}$ con $Z = \frac{\sum X_i - \mu_0}{\sqrt{n}\sigma} \sim N(0, 1)$;
- $p = P(Z \leq z)$ e confrontare con 0.01;
- oppure $D = \{x_i : \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma} \leq -\phi_{0.99}\}$.

Test per la media di osservazioni gaussiane a varianza non nota. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)$ con σ^2 non nota e vogliamo testare il valore di μ che assumiamo essere un certo valore μ_0 mentre i dati sembrano suggerire un valore più piccolo; vogliamo limitare l'errore di seconda specie. Allora ha senso porre

- $\alpha = 0.5$
- H_0: $\mu = \mu_0$ e H_1: $\mu < \mu_0$;

- $t = \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma}$ con $T = \frac{\sum X_i - \mu_0}{\sqrt{n}\sigma} \sim Stud(n-1)$;
- $p = P(T \leq t)$ e confrontare con 0.05;
- oppure $D = \{x_i : \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma} \leq -t_{n-1,0.95}\}$ con $t_{n-1,x}$ quantile di ordine x della distribuzione di Student con $(n-1)$ gradi di libertà .

Test per la varianza di osservazioni gaussiane . Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)$ e vogliamo testare il valore di σ^2 che assumiamo essere un certo valore σ_0 mentre i dati sembrano suggerire un valore più grande; vogliamo limitare l'errore di seconda specie. Allora ha senso porre

- $\alpha = 0.5$
- H_0: $\sigma = \sigma_0$ e H_1: $\sigma > \sigma_0$;
- $\chi^2 = \frac{(n-1)\sum(x_i - \bar{x})^2}{\sigma^2}$ con $K = \frac{(n-1)\sum(X_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$;
- $p = P(K \geq \chi^2)$ e confrontare con 0.05;
- oppure $D = \{x_i : \frac{(n-1)\sum(x_i - \bar{x})^2}{\sigma^2} \geq \chi^2_{n-1,0.95}\}$ con $\chi^2_{n-1,x}$ quantile di ordine x della distribuzione del Chi-quadro con $n-1$ gradi di libertà o $\Gamma((n-1)/2, 1/2)$.

Test per la probabilità di successo a due code. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)B(1, p)$ e vogliamo testare il valore di p che assumiamo essere un certo valore p_0 mentre i dati potrebbero suggerire un valore diverso; vogliamo limitare l'errore di prima specie. Allora ha senso porre

- $\alpha = 0.1$
- H_0: $p = p_0$ e H_1: $p \neq p_0$;
- $z = \frac{\sum x_i - p_0}{\sqrt{np_0(1-p_0)}}$ con $Z = \frac{\sum X_i - p_0}{\sqrt{np_0(1-p_0)}} \sim N(0, 1)$ approssimativamente;
- $p = P(|Z| \geq |z|)$ e confrontare con 0.01;
- oppure $D = \{x_i : |\frac{\sum x_i - p_0}{\sqrt{np_0(1-p_0)}}| \geq \phi_{0.995}\}$.

Test a una coda per il confronto delle medie in due popolazioni. Assumiamo che vi siano due popolazioni con dati x_i, $i = 1, \ldots, n_1$ e y_j, $j = 1, \ldots, n_2$ generati da variabili aleatorie X_i ed Y_j, rispettivamente, entrambe i.i.d. ed independenti tra loro. Vogliamo testare se $E(X) = E(Y)$ mentre i dati suggeriscono che sia minore. Il test si realizza nel modo seguente:

- $\alpha = 0.1$
- H_0: $E(X) = E(Y)$ e H_1: $E(X) \leq E(Y)$;

- $t = \frac{(\sum x_i - \sum y_j)\sqrt{n_1+n_2-2}}{\sqrt{((n_1-1)S_X^2+(n_2-1)S_Y^2)(1/n_1+1/n_2)}}$ con $T = \frac{(\sum X_i - \sum X_j)\sqrt{n_1+n_2-2}}{\sqrt{((n_1-1)S_X^2+(n_2-1)S_Y^2)(1/n_1+1/n_2)}} \sim Stud(n_1+n_2-2)$ approssimativamente;
- $p = P(T \leq t)$ e confrontare con 0.01;
- oppure $D = \{x_i : t \leq -\phi_{0.99}\}$.

Test del chi-quadro per l'adattamento di una variabile aleatoria ad una distribuzione finita. Assumiamo che le variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. con una data distribuzione finita
$P = P(\theta_1, \ldots, \theta_r) = (p_1(\theta_1, \ldots, \theta_r), \ldots, p_k(\theta_1, \ldots, \theta_r))$ nota a meno di r parametri incogniti sullo spazio $S = \{s_1, \ldots, s_k\}$. Vogliamo testare se la distribuzione sia adeguata a rappresentare i dati limitando l'errore di seconda specie. Poniamo n_j il numero di osservazioni di j nel campione. Allora ha senso porre

- $\alpha = 0.5$
- H_0: la distribuzione è $P(\theta_1, \ldots, \theta_r)$ per qualche valore dei parametri e H_1: la negazione di H_0;
- $\chi^2 = \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j}$ con $K = \sum_{j=1}^k \frac{(Y_j - np_j)^2}{np_j} \sim \chi^2(k-r-1)$ approssimativamente, avendo posto Y_j variabile aleatoria che dà il numero di osservazioni di j in un campione di dimensione n;
- $p = P(K \geq \chi^2)$ e confrontare con 0.05;
- oppure $D = \{x_i : \sum_{j=1}^k \frac{(n_j - np_j)^2}{np_j} \geq \chi^2_{k-r-1,0.95}\}$ con $\chi^2_{m,x}$ quantile di ordine x della distribuzione del Chi-quadro con m gradi di libertà o $\Gamma(m/2, 1/2)$.

Test del chi-quadro per l'adattamento di una variabile aleatoria ad una distribuzione continua. Se si assume che le variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. con una data distribuzione continua con densità f_X allora ci si può rifare al caso discreto dividendo in k classi I_j, $j = 1, \ldots, k$ e prendendo $p_j(\theta_1, \ldots, \theta_r) = P(X \in I_j) = \int_{I_j} f_X(t)dt$ sempre a meno di r parametri incogniti. Ci si può rifare quindi al caso precedente, sempre utilizzando la distribuzione χ^2 con $k - r + 1$ gradi di libertà . Si noti che i gradi di libertà dipendono dal numero di intervalli scelti, ma naturalmente la soglia determinata dal χ^2 cresce al crescere del numero di intervallini (quindi si ottiene un fitting più difficile confrontato con una soglia maggiore).

Test di Kolmogorov-Smirnov per l'adattamento di variabili aleatorie qualsiasi. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. con una data funzione di distribuzione F_X. Vogliamo testare se la distribuzione sia adeguata a rappresentare i dati limitando l'errore di seconda specie. Sia F_n la

funzione di distribuzione empirica (che vale k/n in $[x_{(k)}, x_{(k+1)}]$)
Allora ha senso porre

- $\alpha = 0.5$
- H_0: la distribuzione è $P(\theta_1, \ldots, \theta_r)$ per qualche valore dei parametri e H_1: la negazione di H_0;
- $d_n = \max_t |F_n(t) - F_X(t)|$ con D_n una distribuzione tabulata:
- $p = P(D_n \geq d_n)$ e confrontare con 0.05;
- oppure $D = \{x_i : \max_t |F_n(t) - F_X(t)| \geq D_n(0.95)\}$ con $D_n(x)$ quantile di ordine x.

Come esempi numerici consideriamo lo z-test ad una coda da cui eravamo partiti.

Esempio 82. *Ritornando all'esempio 1, se esce testa 459 volte su 1000 lanci di una moneta, allora si esegue un test di ipotesi relativamente alla correttezza della moneta ponendo:*

- $\alpha = 0.2$
- H_0*:* $p = 1/2$ *e* H_1*:* $p \leq 1/2$*;*
- $z = \frac{\sum_{i=1}^{1000} x_i - 1/2}{\sqrt{1000}/2} \approx -2.59$*;*
- *dal TCL si ha* $p = P(Z \leq -2.59) \approx \int_{-\infty}^{-2.59} \frac{\sqrt{2\pi}}{e^{-\frac{x^2}{2}}} \approx 0.0048 < 0.01$ *quindi si rigetta* H_0*, ossia si conclude che l'insieme delle ipotesi fatte non è consistente con i dati osservati.*

Osservazione 13. *Avevamo già visto che questo numero di teste non era accettabile; la terminlogia del test di ipotesi ha permesso di chiarire bene la procedura e le conclusioni.*

Ora possiamo concludere l'analisi dell'esempio 2 attraverso uno z-test approssimato.

Esempio 83. *Si ha*

- $\alpha = 0.2$
- H_0*: I tempi di vita sono realizzazioni di variabili aleatorie i.i.d. esponenziali di parametro* λ *tale che il tempo medio di funzionamento previsto corrisponde al valor medio della distribuzione, ossia* $\frac{1}{\lambda} = 1000$*, quindi* $\lambda = \frac{1}{1000}$ *e* H_1*:* $\lambda \leq \frac{1}{1000}$*;*
- $z = \frac{\sum_{i=1}^{1000} x_i - 100000}{\sqrt{100} \cdot 1000} = \frac{90200 - 100 \cdot 1000}{\sqrt{100} \cdot 1000} \approx -0.98$*;*
- *dal TCL si ha* $p \approx \int_{-\infty}^{-0,98} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 16,35\% > 0.01$ *quindi non si rigetta* H_0*, ossia si conclude che i dati non contraddicono l'insieme delle ipotesi fatte.*

Quindi un tempo totale di funzionamento di 90000 ore era perfettamente in linea con la durata media prevista.

Esercizio 32. *Verificare che con una soglia dell'*1% *la regione di rigetto* D *corrisponde ad un funzionamento totale minore di 76700*

ore e quindi solo sotto tale soglia si dovrebbe considerare falsa la dichiarazione sulla durata media.

Esercizio 33. *Perchè questa volta si tende a considerare inappropriato un risultato ancora perfettamente adeguato?*