

Introduzione alla Statistica Matematica

Note del Corso

Francesco Grotto, Mario Maurelli, Maurizio Pratelli



Front Page Photo ©Filippo Poli,
CC BY-SA 4.0 <<https://creativecommons.org/licenses/by-sa/4.0/>>, via Wikimedia Commons

LaTeX template ©Mathias Legrand (legrand.mathias@gmail.com),
CC BY-NC-SA 3.0 <<http://creativecommons.org/licenses/by-nc-sa/3.0/>>,
via <http://www.LaTeXTemplates.com>

The following notes are licensed under
<https://creativecommons.org/licenses/by-nc-sa/4.0/> (CC BY-NC-SA 4.0)
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International



Non è consentito distribuire o utilizzare a fini commerciali i contenuti che seguono.

Indice

I	Statistica Descrittiva	
1	Statistiche Riassuntive e Grafici	7
1.1	Campioni, grafici a barre, istogramma	7
1.2	Indici statistici	10
1.3	Funzione di ripartizione empirica	13
1.4	Percentili, Quantili e Boxplot	13
2	Dati Multivariati	17
2.1	Covarianza e correlazione campionarie	17
2.2	La retta di regressione	18
II	Probabilità	
3	Probabilità e (In)dipendenza	23
3.1	Spazi di Probabilità	24
3.2	Calcolo Combinatorio	27
3.3	Probabilità Condizionata e formula di Bayes	28
3.4	Indipendenza	31
3.5	Probabilità sulla retta reale	33
3.6	Esercizi	35

4	Variabili Aleatorie	39
4.1	La legge di una Variabile Aleatoria	39
4.2	Funzione di Ripartizione e Quantili	41
4.3	Variabili Aleatorie Notevoli	43
4.4	Valore Atteso, Varianza e Momenti	50
4.5	Esercizi	55
5	Variabili Aleatorie Multivariate	61
5.1	Variabili Doppie	61
5.2	Indipendenza di Variabili Aleatorie	63
5.3	Covarianza e Correlazione	67
5.4	Teoremi Limite in Probabilità	69
5.5	Variabili Chi-Quadro e di Student	74
5.6	Esercizi	77

III

Inferenza Statistica

6	Campioni di Variabili Aleatorie	83
6.1	Campioni Statistici e Stimatori	83
6.2	Stima parametrica	85
6.3	Statistiche campionarie di variabili Gaussiane	87
6.4	Esercizi	88
7	Intervalli di Fiducia	89
7.1	Definizione	89
7.2	Intervalli di fiducia per la media di un campione Gaussiano	90
7.3	Intervalli di fiducia per la media di un campione Bernoulli	92
7.4	Intervalli di fiducia per la media di un campione di taglia grande	94
7.5	Intervalli di fiducia per la varianza di un campione Gaussiano	94
7.6	Esercizi	95
8	Test Statistici	97
8.1	Concetti generali	97
8.2	Z-test	101
8.3	Test sulla media di un campione Gaussiano con varianza sconosciuta, o T-test	105
8.4	Test approssimato su un campione di Bernoulli	107
8.5	Test approssimato sulla media di un campione di taglia grande	108
8.6	Test sulla varianza di un campione Gaussiano	108
8.7	Confronto tra due campioni statistici	109
8.8	Esercizi	111



Statistica Descrittiva

1	Statistiche Riassuntive e Grafici	7
1.1	Campioni, grafici a barre, istogramma	
1.2	Indici statistici	
1.3	Funzione di ripartizione empirica	
1.4	Percentili, Quantili e Boxplot	
2	Dati Multivariati	17
2.1	Covarianza e correlazione campionarie	
2.2	La retta di regressione	

1. Statistiche Riassuntive e Grafici

La statistica si occupa dello studio di dati, ovvero della loro raccolta –questione di cui queste note non si occupano– e della loro analisi ed interpretazione. Le risposte di tale analisi, tramite i metodi matematici di cui il seguito dà una introduzione elementare, non dipendono solo dai dati in se, ma dalla conoscenza pregressa del problema studiato e quindi dalle eventuali ipotesi ed assunzioni che entrano a far parte del modello matematico usato.

Si parla di **statistica descrittiva** quando i dati vengono analizzati, seppur con tecniche anche raffinate, senza fare assunzioni esterne all'insieme di dati considerati. Lo scopo è quindi l'organizzazione dei dati in modo da evidenziarne la struttura, e di rappresentarli in modo efficace. Di questo ci occuperemo brevemente in questa prima parte delle note.

L'**inferenza statistica** invece studia i dati utilizzando un modello probabilistico, cioè suppone che i dati siano valori assunti da **variabili aleatorie** aventi una certa **distribuzione di probabilità** dipendente da dei parametri non noti. In questo caso, l'analisi statistica ha come scopo la stima di questi parametri, ed il modello matematico che risulta da tale stima può essere utilizzato per fare delle previsioni ed informare delle decisioni. La definizione matematicamente rigorosa di questi concetti è il tema principale di questi appunti. Chiaramente conclusioni più forti dell'analisi dei dati seguono a prezzo di assunzioni più stringenti e dettagliate, che quindi rischiano di essere meno affidabili e non sempre verificate.

Non ci occuperemo della **statistica bayesiana**, un terzo approccio al problema dell'analisi dei dati, che però citiamo per l'importanza nelle applicazioni.

1.1 Campioni, grafici a barre, istogramma

I dati che ci proponiamo di analizzare altro non sono che risultati numerici di misure ripetute. Si considera cioè una **popolazione**, l'insieme di oggetti o fenomeni che si vuole studiare, su ognuno dei quali è possibile effettuare la stessa misura, ovvero considerarne un **carattere**. La popolazione può essere reale, ad esempio la popolazione italiana, un'urna con 10 biglie, oppure può essere ideale, ad esempio, per un esperimento fisico ripetuto più volte, la popolazione è data da “tutte le infinite possibili ripetizioni dell'esperimento”). Un **campione statistico** (*statistical sample*) è un sottoinsieme della popolazione scelto per rappresentarla. I nostri **dati** consistono nelle

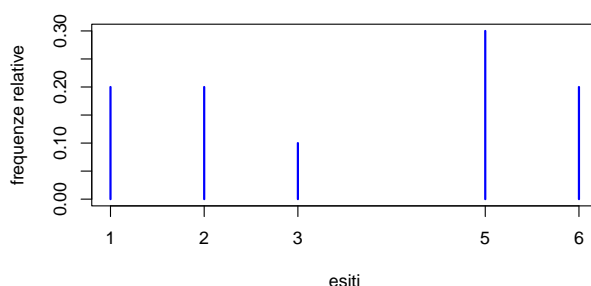
misure effettuate su di esso (spesso nell'impossibilità di effettuare la misura sulla totalità della popolazione).

Il carattere la cui misura sul campione costituisce l'insieme dei dati può essere quantitativo o qualitativo. Un carattere **quantitativo** è una quantità numerica, ovvero *gli esiti della misura sono numeri paragonabili tra loro*, come ad esempio l'altezza di una persona o il prezzo di un oggetto. Un carattere **qualitativo** invece produce misure che non possiamo paragonare tra loro in modo significativo con dei numeri: ad esempio il carattere “nome proprio di una persona” può anche essere riassunto numerando tutti i nomi (ad esempio in ordine alfabetico), ma questi numeri non misurano il modo in cui i diversi nomi differiscono.

Dato un possibile esito della misura (qualitativo o quantitativo), la sua *frequenza* (assoluta) è il numero di volte in cui questo esito compare nei dati, mentre la *frequenza relativa* è la frazione (di solito percentuale, per comodità) di volte in cui questo esito compare sul totale dei dati.

La situazione forse più semplice da rappresentare è quella in cui l'esito della misura di un carattere può assumere una quantità finita e relativamente piccola di valori (qualitativi o quantitativi); parliamo in questo caso di **carattere discreto**. In questo caso, per rappresentare la distribuzione dei dati possiamo usare un **diagramma a barre**, che consiste in una serie di barre verticali, ciascuna delle quali ha altezza pari (o proporzionale) alla frequenza relativa di un possibile esito della misura. Un altro modo di rappresentare questi dati è il diagramma a torta (*piechart*).

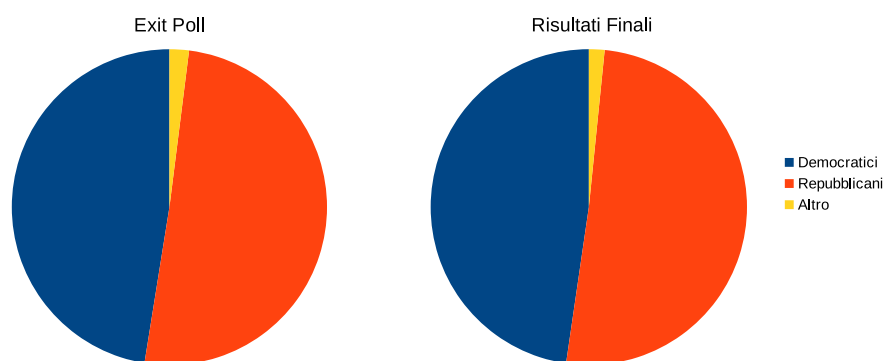
■ **Esempio 1.1 — Caratteri Qualitativi, Diagramma a Barre.** Lanciamo un dado (non necessariamente equo) 10 volte. Gli esiti possibili sono $1, 2, \dots, 6$, la popolazione (ideale) è costituita da tutti i possibili lanci del dado, mentre il campione è dato dai 10 lanci effettuati. Supponiamo di ottenere i seguenti esiti: 6, 2, 3, 1, 2, 5, 6, 5, 5, 1. Il diagramma a barre è il seguente:



Notiamo che le frequenze e quindi anche la loro rappresentazione dipendono dai dati del campione scelto e non coincidono in generale con le frequenze su tutta la popolazione.

■ **Esempio 1.2 — Caratteri Qualitativi, Campione vs. Popolazione.** Un *exit poll* è un sondaggio sul voto espresso condotto tra votanti che hanno appena lasciato il seggio elettorale. Gli esiti possibili sono i partiti o i candidati in corsa, e il campione è una parte di solito molto ridotta della popolazione (che è l'insieme di tutti i votanti), tanto che spesso se ne discute la rappresentatività.

Alle *midterm elections* statunitensi, il 9 Novembre 2022, un *exit poll* della CNN ha intervistato 18571 persone, che alla Camera hanno votato i Democratici per il 47.47% oppure i Repubblicani per il 50.53% (il 2% altro). Abbiamo in questo caso il vantaggio di poter confrontare una misura su (quasi) tutta la popolazione: il risultato generale ha visto i Democratici al 47.7% e i Repubblicani al 50.7% (1.6% altro).



Vediamo dunque che la frequenza di un carattere in un campione non coincide con quella su tutta la popolazione. ■

Si pone quindi il problema di come scegliere il campione in modo che sia rappresentativo: benchè sia una questione molto importante e complessa nelle applicazioni, non rientra negli obiettivi di queste note. Ci limitiamo a menzionare che nella statistica inferenziale, oggetto principale di questi appunti, è sottointesa l'assunzione di avere a disposizione un campione "scelto a caso uniformemente" nella popolazione: vedremo quanto questo potrà essere informativo, cioè quanto questa ipotesi permetta di dedurre, assieme al dato del campione, informazioni sulle frequenze di tutta la popolazione.

Tornando alla rappresentazione dei dati, vi sono casi in cui i risultati possono assumere un gran numero di valori quantitativi: ad esempio se misuriamo la temperatura (in Kelvin) di un certo gas, questa può assumere un qualunque valore positivo. In casi come questo parliamo di **carattere continuo**. Poiché i dati possono essere tutti diversi tra loro, ha poco senso considerare la frequenza di un singolo esito. La rappresentazione grafica più immediata usata in questo caso è l'**istogramma**¹. Quest'ultimo consiste in una serie di colonne, ognuna delle quali ha per base un intervallo numerico e per *area* la frequenza relativa dei dati contenuti in quell'intervallo².

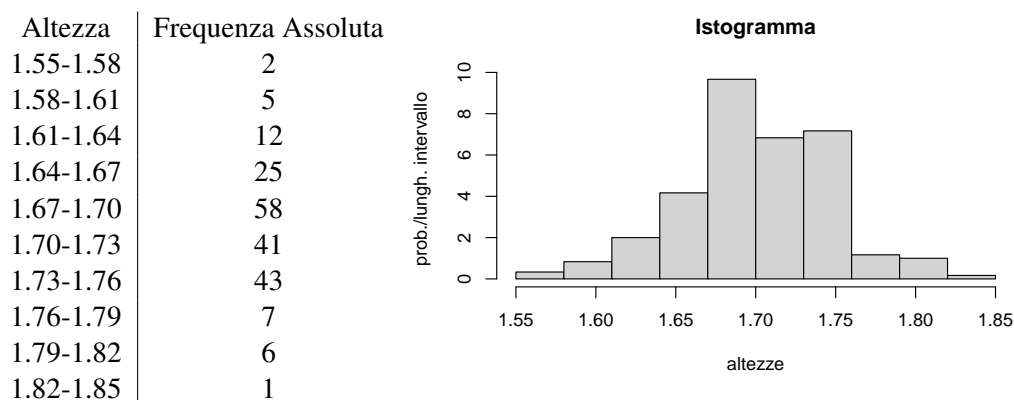


Figura 1.1: Le frequenze assolute delle altezze (in metri) di 200 persone (dati inventati), e il relativo istogramma.

¹Parola grecizzante coniata da Karl Pearson per indicare questi "grafici a bastoni".

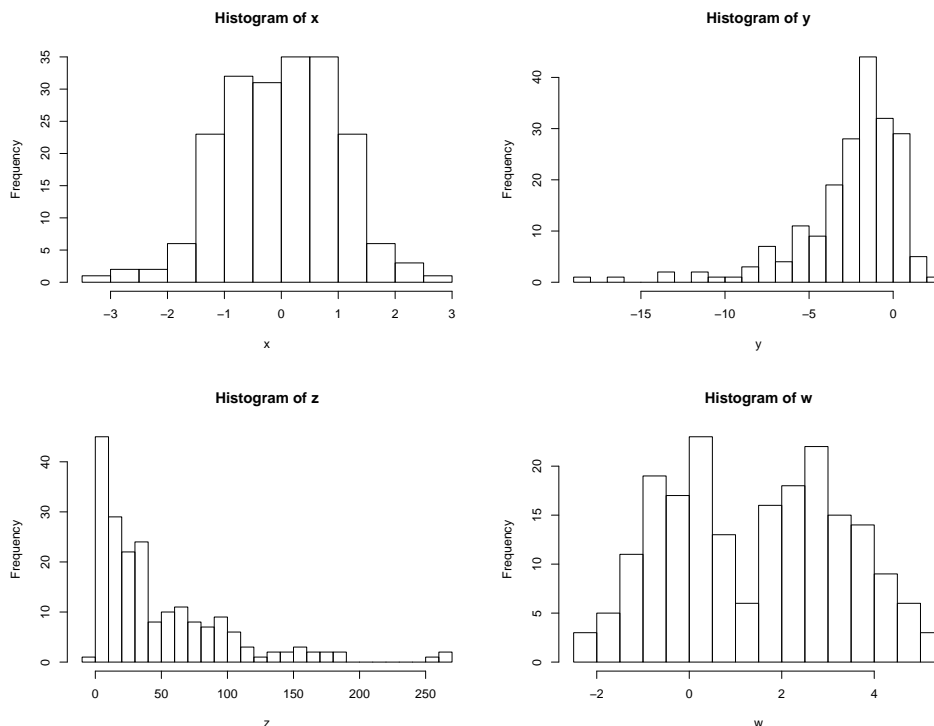
²In alcuni libri di testo, si prende come altezza, e non come area, la frequenza relativa dei dati nell'intervallo considerato. In questo caso privilegiamo la definizione con l'area, per due motivi: 1) è più adatta in contesti in cui gli intervalli non hanno la stessa ampiezza; 2) si confronta bene con la densità di probabilità della popolazione, come vedremo.

Per rappresentare un istogramma, è necessario scegliere gli intervalli da usare come basi dei rettangoli. Non tutte le scelte sono buone: nell'esempio precedente (delle altezze di 200 lampadine), scegliere 2 intervalli ci farebbe perdere molta informazione, mentre scegliere 100 intervalli renderebbe l'istogramma troppo dipendente da piccole variazioni dei dati. Esistono alcune regole empiriche sul numero di intervalli da scegliere (generalmente dipendenti dalla numerosità del campione), ma non c'è una regola generale e ogni scelta deve essere confrontata con i dati effettivi a disposizione, ed eventualmente modificata per rendere l'istogramma sufficiente informativo e robusto.

La forma dell'istogramma, o del grafico a barre, viene descritta con una serie di termini relativi a varie caratteristiche:

- “normale” se la forma assomiglia a quella di una campana simmetrica;
- “unimodale” o “bimodale” a seconda che l'istogramma si concentri rispettivamente attorno a una singola o a due colonne più alte;
- nel caso unimodale, asimmetrica (*skewed*) a destra o a sinistra se i dati sono più concentrati a destra o a sinistra del picco;
- “platicurtica” o “leptocurtica” a seconda che, rispettivamente, i dati siano concentrati quasi totalmente in un certo intervallo (anche grande), oppure siano composti da un gruppo centrale e da molti *outliers* (dati isolati distanti dal corpo della distribuzione).

Sono riportati qua sotto 4 istogrammi, rispettivamente normale, asimmetrico a sinistra, a destra, e bimodale. Nella prossima sezione mostriamo come delle quantità matematiche derivate dai dati quantificano questo tipo di descrizione.



1.2 Indici statistici

Supponiamo di avere un vettore $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ di dati numerici. Gli indici statistici sono quantità numeriche che riassumono alcune proprietà significative della distribuzione dei dati (x_1, \dots, x_n) .

La media campionaria e la mediana campionaria descrivono dove **si trova il “centro” della distribuzione dei dati**:

Definizione 1.2.1 La **media campionaria** (*sample average*) è la media aritmetica dei dati,

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

Definizione 1.2.2 La **mediana** (*median*) è il dato x_i tale che metà degli altri valori è minore o uguale a x_i e l'altra metà maggiore o uguale (**nel caso n sia pari si può prendere la media aritmetica dei due valori centrali in questo senso**).

Nel seguito preferiremo perlopiù la media campionaria alla mediana. Tuttavia la mediana è utile nel caso di dati molto asimmetrici (*skewed*) e soprattutto è robusta rispetto alle code della distribuzione: in altre parole la media campionaria viene facilmente spostata da singoli dati molto piccoli o molto grandi, e questo non succede alla mediana.

■ **Esempio 1.3** Supponiamo di aver misurato le altezze di 10 persone, che

La media campionaria si può calcolare a partire dalle frequenze relative dei possibili esiti: se questi ultimi sono a_1, \dots, a_M (ovvero ogni singolo dato x_i è uguale a uno degli a_j),

$$\bar{x} = \sum_{j=1}^M a_j p(a_j, x), \quad p(a_j, x) = \frac{\#\{i : x_i = a_j\}}{n},$$

dove appunto $p(a_j, x)$ è la frequenza relativa di a_j nel campione x .

La varianza campionaria si usa per misurare la dispersione dei dati, cioè quanto i dati sono “sparsi” attorno al “centro” della distribuzione. Essa è infatti la media degli scarti quadratici dalla media campionaria \bar{x} :

Definizione 1.2.3 La **varianza campionaria** (*sample variance*) oppure la **varianza empirica** (*empirical variance*), rispettivamente

$$\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \text{var}_e(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

La differenza tra le due quantità suddette è importante in inferenza statistica, e sarà chiara in seguito. Con semplici passaggi si verifica che

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2, \quad \text{var}_e(x) = \sum_{i=1}^n \frac{x_i^2}{n} - \bar{x}^2.$$

Inoltre, come per la media campionaria, la varianza empirica (e quindi quella campionaria) può essere espressa in termini delle frequenze relative: detti sempre a_j , $j = 1 \dots M$, i possibili esiti e $p(a_j, x)$ la frequenza relativa di a_j nel campione x , si ha

$$\text{var}_e(x) = \sum_{j=1}^M a_j^2 p(a_j, x) - \bar{x}^2.$$

Definizione 1.2.4 La radice della varianza è chiamata scarto quadratico medio o anche **deviazione standard** (esiste dunque sia campionaria che empirica): la indichiamo **$\sigma(x)$** (o $\sigma_e(x)$ quando è empirica) e si ha dunque $\sigma(x) = \sqrt{\text{var}(x)}$.

La varianza, indicata quindi con $\sigma^2(x)$, è evidentemente eguale a 0 se e solo se i dati sono tutti uguali. Più in generale vale:

Proposizione 1.2.1 Preso un **campione di dati** x ed un **numero positivo** d si ha

$$\frac{\#(x_i : |x_i - \bar{x}| > d)}{n} \leq \frac{\text{var}_e(x)}{d^2}$$

dove $\#A$ indica il numero degli elementi di A .

Dimostrazione. Si ha che

$$\sum_{i=1}^n (x_i - \bar{x})^2 \geq \sum_{i: |x_i - \bar{x}| > d} (x_i - \bar{x})^2 \geq \sum_{i: |x_i - \bar{x}| > d} d^2 = d^2 \cdot \#(x_i : |x_i - \bar{x}| > d),$$

da cui

$$\#(x_i : |x_i - \bar{x}| > d) \leq \frac{\sum_i (x_i - \bar{x})^2}{d^2},$$

e dividendo per n si ha la tesi. ■

Questa formula è un caso particolare della disuguaglianza di Chebyshev che vedremo più avanti, notiamo che il termine di sinistra è la frazione di dati che differiscono da \bar{x} più di d , ed è quindi evidente l'idea di varianza come misura della dispersione dei dati.

■ **Esempio 1.4** Calcoliamo media e varianza campionaria dei dati nell'Esempio 1.1 (10 lanci di un dado) a partire dalle frequenze relative:

$$\bar{x} = \sum_{j=1}^6 j \cdot p(j, x) = 3.6, \quad \text{var}(x) = \frac{10}{9} \left(\sum_{j=1}^6 j^2 \cdot p(j, x) - \bar{x}^2 \right) \simeq 4.04$$

(le frequenze sono quelle riportate nel grafico a barre). ■



Come nel precedente esempio, *indicheremo sempre l'inizio della parte decimale di un numero non intero con il punto* (e non con la virgola), come è uso diffuso nei linguaggi di programmazione al calcolatore.

Se la varianza fa uso della funzione $x \mapsto x^2$ per misurare la distanza media dei punti dalla media campionaria, tramite la funzione $x \mapsto x^3$, che assume valori grandi man mano che x si allontana dall'origine *ma mantenendo il segno di x* , possiamo misurare l'asimmetria della distribuzione. Questo perchè se ci sono molti dati a sinistra della media campionaria, nella somma di termini $(\bar{x} - x_i)^3$ prevarranno quelli con segno negativo, e viceversa.

Definizione 1.2.5 La *sample skewness* (misura campionaria di asimmetria) è

$$b = \frac{1}{\sigma^3} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3.$$

Senza entrare nei dettagli, menzioniamo che usando la funzione $x \mapsto x^4$ si può misurare in modo simile a sopra “quanto piatta” (o anche “quanto normale”) è la distribuzione dei dati, definendo la **curtosi** (*kurtosis*). Nel caso di $x \mapsto x^4$ consideriamo una funzione che restituisce numeri di segno sempre positivo, che crescono molto velocemente all'allontanarsi dall'origine (“pesano le code” della distribuzione). Questa idea chiaramente si può generalizzare ulteriormente per definire varie altre quantità interessanti, su cui non ci soffermiamo.

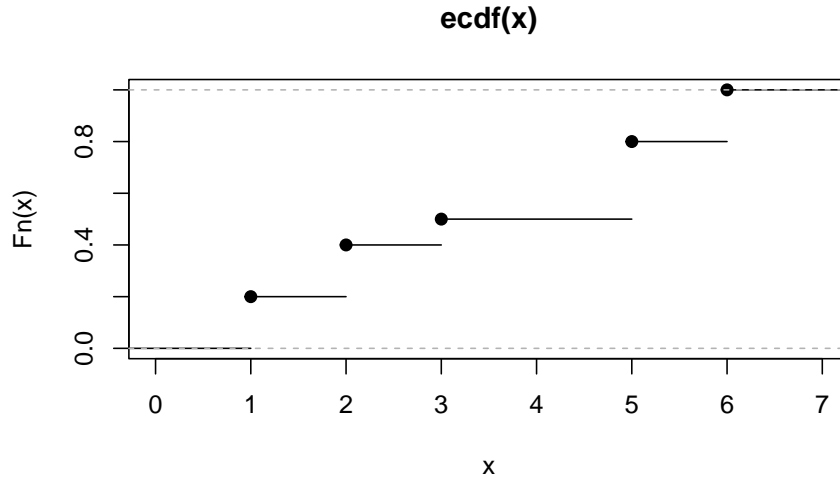


Figura 1.2: Le e.c.d.f. dei dati dell'Esempio 1.1.

1.3 Funzione di ripartizione empirica

Sia $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ un vettore di dati numerici. Introduciamo ora uno strumento che risulterà utile nel caratterizzare la distribuzione dei dati: la funzione di ripartizione empirica, indicata anche come e.c.d.f. (*empirical cumulative distribution function*).

Prima di introdurre la definizione formale, vediamo come si ottiene in pratica la e.c.d.f.: ordiniamo i numeri x_1, \dots, x_n in ordine crescente chiamandoli $x_{(1)}, \dots, x_{(n)}$ (naturalmente è possibile che due o più siano eguali). La funzione $F_e(t)$ è eguale a 0 per $t < x_{(1)}$, poi fa un salto di ampiezza $1/n$ in corrispondenza di ognuno dei valori $x_{(i)}$ (se due dati sono eguali il salto è di ampiezza $2/n$, se tre dati sono eguali il salto è di ampiezza $3/n$ e così via) e poi è definitivamente eguale a 1 per $t \geq x_{(n)}$.

■ **Esempio 1.5** I seguenti dati sono gli esiti di 36 misurazioni di rumore (in dB, in giorni diversi alla stessa ora) all'esterno della Great Central Station di Manhattan:

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85, 69,
94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65.

La e.c.d.f. corrispondente è riportata nella Figura 1.3 che mostra come, quando i dati sono più numerosi, i “gradini” siano meno accentuati. ■

Vediamo ora la definizione precisa:

Definizione 1.3.1 La funzione di ripartizione empirica $F_e : \mathbb{R} \rightarrow \mathbb{R}$ è definita da

$$F_e(t) = \frac{\#(i \mid x_i \leq t)}{n}.$$

In altre parole, per ogni $t \in \mathbb{R}$, $F_e(t)$ restituisce la frequenza relativa dei dati minori o uguali a t . Osserviamo che (banalmente) la funzione F_e è sempre non-decrescente, e $F_e(-\infty) = 0$, $F_e(+\infty) = 1$.

1.4 Percentili, Quantili e Boxplot

Consideriamo ancora un vettore di dati $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Dato un numero $0 < k < 100$, è naturale chiedersi come scegliere una soglia che divida in due parti i dati, in modo che una di queste

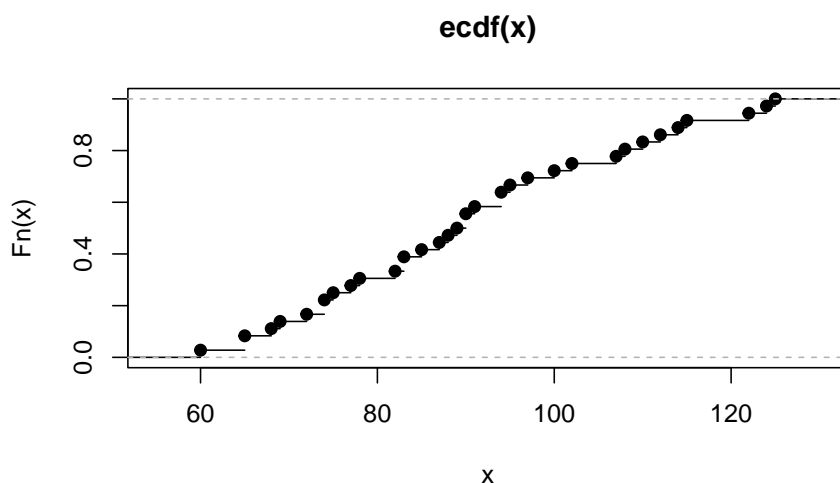


Figura 1.3: e.c.d.f. dei dati dell'Esempio 1.5.

contenga dati più piccoli di ogni dato dell'altra, e al contempo contenga in totale il $k\%$ dei dati. Ad esempio, se i dati misurano la qualità di un prodotto o l'esito di un test, questa è la procedura con cui si sceglie una porzione predeterminata del campione in modo da selezionare gli esiti migliori, o peggiori.

In effetti, abbiamo già affrontato la questione definendo la mediana: in quel caso $k = 50$, ovvero la mediana divide il campione in due parti con la stessa numerosità. Come in quel caso, definire in generale il k -mo percentile come il più piccolo dato che supera il $k\%$ dei dati non è una buona scelta, perchè (essendo i dati discreti) non sempre uno degli x_i soddisfa esattamente la condizione. Si dà quindi una definizione più elaborata:

Definizione 1.4.1 Si chiama **k -mo percentile** il dato x_i tale che:

- almeno $\frac{k}{100} \cdot n$ dati siano inferiori o eguali a x_i ,
- almeno $\frac{100-k}{100} \cdot n$ siano superiori o eguali a x_i ,

e se due o più dati soddisfano questa condizione, si prende come k -mo percentile la media aritmetica tra il più piccolo e il più grande di questi ultimi.

Esercizio 1.1 Si mostri che la seguente definizione è equivalente a quella proposta sopra: il k -mo percentile è il numero t definito da

$$t = \frac{1}{2} \max\{x_i : F_e(x_i) \leq k/100\} + \frac{1}{2} \min\{x_i : F_e(x_i) \geq (100-k)/100\}.$$

Si osservi che con i dati dell'Esempio 1.1 (vedi la e.c.d.f. in Figura 1.2), per $k = 50$ il massimo e minimo che appaiono nella formula coincidono, mentre per $k = 60$ no. Qual è il k -mo percentile nei due casi? ■

A livello operativo, il k -mo percentile si calcola in questo modo: detti n la numerosità del campione, $x_{(1)}, \dots, x_{(n)}$ i dati ordinati in senso crescente, $\beta = k/100$,

- se βn non è intero, si prende come k -mo percentile $x_{(\lceil \beta n \rceil)}$, dove $\lceil \beta n \rceil$ è l'arrotondamento di βn all'intero successivo;
- se βn è intero, si prende come k -mo percentile la media aritmetica tra $x_{(\beta n)}$ e $x_{(\beta n+1)}$.

In effetti, il parametro $\beta = k/100$ è spesso usato al posto di k , per cui diamo la seguente definizione, che prevarrà quando studieremo le variabili aleatorie.

Definizione 1.4.2 Dato $k \in (0, 100)$, e posto $\beta = \frac{k}{100} \in (0, 1)$, il k -mo percentile è anche detto β -quantile.

In particolare il 25-mo percentile è chiamato primo quartile, il 50-mo percentile (la mediana) è chiamato secondo quartile, il 75-mo percentile è detto terzo quartile.

Il *box-plot* dei dati è una rappresentazione grafica che evidenzia l'intervallo in cui sono concentrati i dati: si ottiene sovrapponendo a una linea che va dal minimo al massimo dei dati un rettangolo che va dal primo al terzo quartile, con una linea che lo divide al livello della mediana. Qui sotto è riportato un esempio di box-plot. Si noti che il rettangolo centrale contiene il 50% dei dati centrali.

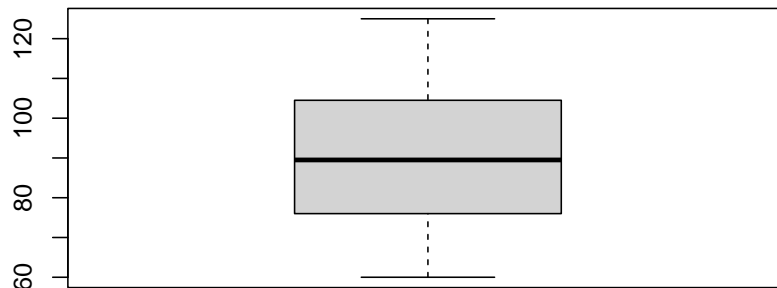
■ **Esempio 1.6** Nell'Esempio 1.5, i 36 dati ordinati sono i seguenti:

60 65 65 68 69 72 74 74 75 77 78 82 83 83 85 87 88 89
90 90 91 94 94 95 97 100 102 107 108 110 112 114 115 122 124 125

Il primo, secondo e terzo quartile sono quindi:

$$Q_1 = \frac{75 + 77}{2} = 76, \quad Q_2 = \frac{89 + 90}{2} = 89.5, \quad Q_3 = \frac{102 + 107}{2} = 104.5.$$

Il corrispondente boxplot è



■

In statistica si parla di valore anomalo o *outlier* per indicare un valore che differisce in modo significativo dalla “grande maggioranza” dei dati. In alcuni casi si pone il problema di individuare tali valori, per almeno due motivi: gli outliers possono risultare da errori di misurazione, oppure possono indicare code della distribuzione particolarmente pesanti (cioè la distribuzione assume valori estremi con probabilità non troppo bassa). Come osservato nel caso della media campionaria, gli outliers possono influenzare in modo significativo alcuni indici statistici. Non esiste una regola generale per individuare gli outliers, ma un criterio possibile è il seguente: si considerano outlier quelli non compresi nell'intervallo $[Q_1 - 1.5L, Q_3 + 1.5L]$, dove Q_1 e Q_3 sono rispettivamente il primo e il terzo quartile e $L = Q_3 - Q_1$ (detto anche scarto interquartile). In alcune rappresentazioni del boxplot, la linea ha per estremi il più basso e il più alto valore non outlier, mentre gli eventuali dati outlier vengono rappresentati con dei puntini all'esterno di questa linea.

Quando l'istogramma dei dati è (ragionevolmente) normale, vale la seguente regola empirica sulla concentrazione dei dati:

- circa il 68% dei dati appartiene all'intervallo $[\bar{x} - \sigma(x), \bar{x} + \sigma(x)]$;
- circa il 95% dei dati appartiene all'intervallo $[\bar{x} - 2\sigma(x), \bar{x} + 2\sigma(x)]$;
- circa il 99.7% dei dati appartiene all'intervallo $[\bar{x} - 3\sigma(x), \bar{x} + 3\sigma(x)]$.

Si tratta appunto di una regola empirica che come tale non può essere dimostrata, ma quando studieremo più avanti le variabili Gaussiane (chiamate anche normali) forniremo una dimostrazione precisa di queste proprietà.

2. Dati Multivariati

Ci limitiamo in questo capitolo al caso di coppie di dati (o dati bivariati): ciò che diremo si adatta facilmente al caso generale. I dati dell'indagine statistica sono in questo caso un insieme di coppie di numeri $(x, y) = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathbb{R}^{2 \times n}$. Ad esempio, (x_i, y_i) può essere la misurazione di temperatura e pressione arteriosa per l' i -esimo membro di un campione di pazienti.

2.1 Covarianza e correlazione campionarie

Consideriamo un insieme di n coppie di numeri $(x, y) = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathbb{R}^{2 \times n}$. Ricordiamo che \bar{x} e \bar{y} sono le medie campionarie di x e y e $\sigma(x)$, $\sigma(y)$ sono le deviazioni standard campionarie di x e y .

Definizione 2.1.1 Si chiamano **covarianza campionaria** e **covarianza empirica** i numeri

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

Si verifica facilmente che:

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

Definizione 2.1.2 Supponiamo $\sigma(x) \neq 0$ e $\sigma(y) \neq 0$: si chiama **coefficiente di correlazione tra x e y** il numero

$$r(x, y) = \frac{\text{cov}(x, y)}{\sigma(x)\sigma(y)}.$$

Si può osservare che la definizione non cambia se si sostituiscono covarianza e deviazione standard con covarianza empirica e deviazione standard empirica, ed in realtà vale la formula

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Dalla **disuguaglianza di Schwartz** in \mathbb{R}^n si deduce

$$\sum_{i=1}^n |(x_i - \bar{x}) \cdot (y_i - \bar{y})| \leq \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}$$

e di conseguenza $|r(x, y)| \leq 1$.

2.2 La retta di regressione

Intuitivamente, il coefficiente di correlazione misura il legame di natura lineare tra i dati x e y : introducendo **la retta di regressione** quantifichiamo questa affermazione. L'idea è di approssimare nel modo migliore i dati y_i con una combinazione lineare affine $a + bx_i$, per farlo misuriamo la distanza dei dati da tale retta con i quadrati degli scarti: siamo dunque condotti a cercare i parametri a, b calcolando

$$\inf_{a, b \in \mathbb{R}} \sum_{i=1}^n (y_i - a - bx_i)^2.$$

Questa quantità ha in effetti un unico minimo, e vale il seguente risultato.

Teorema 2.2.1 Supponiamo che $\sigma(x) \neq 0$ e $\sigma(y) \neq 0$. Il minimo al variare di $a, b \in \mathbb{R}$ della quantità $\sum_{i=1}^n (y_i - a - bx_i)^2$ si ottiene scegliendo

$$b^* = \frac{\text{cov}(x, y)}{\text{var}(x)}, \quad a^* = -b^* \bar{x} + \bar{y},$$

e vale

$$\min_{a, b \in \mathbb{R}} \sum_{i=1}^n (y_i - a - bx_i)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 (1 - r(x, y)^2).$$

La retta $y = a^* + b^*x$ è chiamata **retta di regressione**: quanto più $r(x, y)$ è vicino a 1 in valore assoluto, tanto più i punti tendono ad essere allineati e vicino a tale retta. In particolare $|r(x, y)| = 1$ se e solo se i punti si trovano tutti su una stessa retta. Inoltre $r(x, y)$ è positivo o negativo se il coefficiente angolare della retta è rispettivamente positivo o negativo. Se $r(x, y)$ è prossimo a zero il Teorema sopra enunciato mostra che non è possibile dare una buona approssimazione dei dati con una retta, nel senso che per ogni retta considerata esisteranno dei dati distanti da essa. Questo non esclude che possano esistere altre relazioni (approssimate) tra x e y , ad esempio quadratiche o più in generale polinomiali.

Dimostrazione. Consideriamo la funzione

$$Q(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2,$$

che si osserva facilmente essere continua (in effetti è infinitamente differenziabile). La funzione Q tende a $+\infty$ quando $|a|, |b| \rightarrow \infty$ e pertanto ha un minimo. Nel punto di minimo si devono annullare le derivate parziali, ovvero $\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial b} = 0$, condizione da cui otteniamo il sistema di equazioni

$$\begin{cases} \sum_i y_i - na - b \sum_i x_i = 0, \\ \sum_i x_i y_i - a \sum_i x_i - b \sum_i x_i^2 = 0. \end{cases}$$

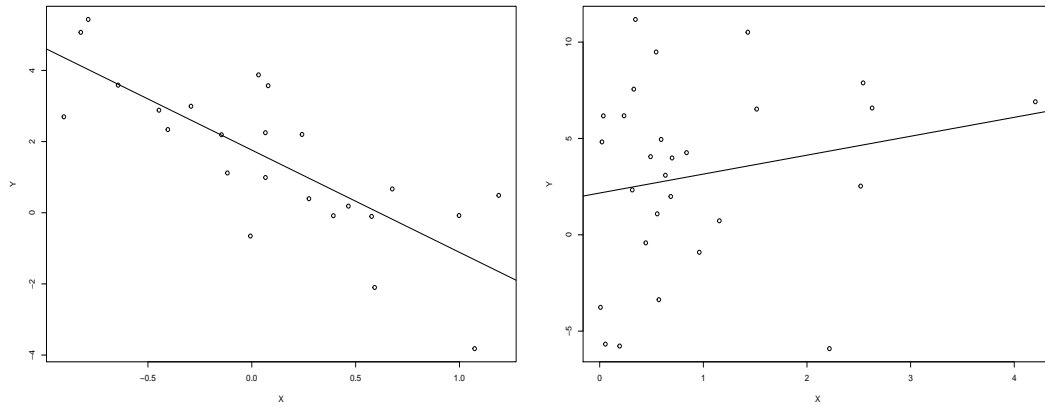


Figura 2.1: Due insiemi di coppie di dati rappresentati come punti sul piano, con relativa retta di regressione. In quello di sinistra, $r(x,y) \approx -0.78$, $a^* \approx 1.76$ e $b^* \approx -2.87$, mentre in quello di destra i dati sono molto più sparpagliati e il coefficiente di correlazione è ≈ 0.2422 .

Dividendo per n si ottiene

$$\begin{cases} a + b\bar{x} = \bar{y}, \\ a\bar{x} + b\sum_i \frac{x_i^2}{n} = \sum_i \frac{x_i y_i}{n}, \end{cases}$$

che è un sistema di due equazioni nelle due incognite a, b . L'unica soluzione è

$$a^* = -b^*\bar{x} + \bar{y}, \quad b^* = \frac{\sum_i \frac{x_i y_i}{n} - \bar{x}\bar{y}}{\sum_i \frac{x_i^2}{n} - \bar{x}^2} = \frac{\text{cov}_e(x,y)}{\text{var}_e(x)}.$$

Essendo l'unico punto stazionario di Q questo coincide con il suo minimo, che quindi è unico. Il valore di Q nel minimo, $Q(a^*, b^*) = \sum_{i=1}^n (y_i - \bar{y})^2 (1 - r(x,y)^2)$, si ottiene sostituendo i valori trovati e sviluppando i calcoli. ■



Probabilità

3	Probabilità e (In)dipendenza	23
3.1	Spazi di Probabilità	
3.2	Calcolo Combinatorio	
3.3	Probabilità Condizionata e formula di Bayes	
3.4	Indipendenza	
3.5	Probabilità sulla retta reale	
3.6	Esercizi	
4	Variabili Aleatorie	39
4.1	La legge di una Variabile Aleatoria	
4.2	Funzione di Ripartizione e Quantili	
4.3	Variabili Aleatorie Notevoli	
4.4	Valore Atteso, Varianza e Momenti	
4.5	Esercizi	
5	Variabili Aleatorie Multivariate	61
5.1	Variabili Doppie	
5.2	Indipendenza di Variabili Aleatorie	
5.3	Covarianza e Correlazione	
5.4	Teoremi Limite in Probabilità	
5.5	Variabili Chi-Quadro e di Student	
5.6	Esercizi	

3. Probabilità e (In)dipendenza

Non possiamo conoscere con certezza il futuro, ovvero ogni affermazione che possiamo fare su eventi futuri può realizzarsi o meno. La teoria della Probabilità nasce per tentare di quantificare questa incertezza, misurando in qualche modo la fiducia che un evento possa accadere. I primi due obiettivi della Probabilità in quanto parte della Matematica sono: rappresentare con oggetti matematici gli eventi in analisi e la loro probabilità; determinare le proprietà che questi oggetti soddisfano mostrando da un lato che costituiscono un modello della realtà, e dall'altro che di questa possono fornire informazioni ulteriori rispetto a quelle osservate.

La definizione matematica di probabilità è stata data negli anni Trenta del Novecento dal grande matematico Kolmogorov ed è di tipo assiomatico, cioè definisce la probabilità in modo rigoroso tramite le sue proprietà elementari. Storicamente, tale definizione fu preceduta da altre non rigorose, rifacentesi piuttosto a interpretazioni del concetto di probabilità, e che possono essere utili per collegare il fenomeno aleatorio reale con un modello probabilistico associato. Ne menzioniamo brevemente due: la definizione classica e quella frequentista, tralasciandone altre comunque importanti come la definizione soggettivista.

La definizione classica considera la probabilità come rapporto tra numero di casi favorevoli e numero di casi possibili (si veda il *modello uniforme* nel seguito). Questa interpretazione è adatta a descrivere problemi come le estrazioni da una popolazione ed è conveniente sia per alcuni calcoli elementari (riducendosi essenzialmente a contare gli elementi di un insieme) sia per l'intuizione probabilistica (se la probabilità di un evento A è $1/4$, possiamo immaginare una popolazione di casi equiprobabili in cui A si realizzi in un quarto dei casi). È meno adatta, a descrivere fenomeni in cui la popolazione è ideale, e possibilmente infinita, come l'esecuzione di un esperimento fisico (affetto da errore) o l'effetto della somministrazione di un farmaco (affetto dalla variabilità dei soggetti cui è somministrato).

La definizione frequentista considera la probabilità di un evento come il limite, per $n \rightarrow \infty$, delle frequenze relative dell'evento in n prove ripetute dell'esperimento. Questa interpretazione descrive meglio fenomeni con popolazione ideale, come appunto l'esecuzione di un esperimento fisico. Richiede però di conoscere a priori che il limite esista (e la definizione rigrosa del limite non è chiara). Al fine dell'intuizione, può essere utile osservare che, idealmente, la definizione

frequentista è il limite della definizione classica per grande numero di prove.

3.1 Spazi di Probabilità

Consideriamo un certo esperimento aleatorio, cioè un fenomeno (fisico, biologico, sociale, ...) il cui esito non è determinabile con certezza a priori. La costruzione delle probabilità come oggetto matematico inizia da una semplice osservazione: possiamo rappresentare tutti gli esiti possibili dell'esperimento con gli elementi di un insieme astratto Ω , detto *spazio campionario* (*sample space*). I sottoinsiemi di Ω , che *definiremo* eventi, rappresentano affermazioni che possiamo fare sull'esito dell'esperimento. Di conseguenza, combinazioni di affermazioni diverse ottenute tramite i connettivi logici ("o", "e", "non", "implica") corrispondono come è noto ad operazioni insiemistiche. I singoli esiti, cioè gli elementi di Ω , sono detti eventi elementari (con abuso di notazione, parliamo di eventi elementari per intendere sia gli elementi ω in Ω , sia gli insiemi singoletto $\{\omega\}$).

■ **Esempio 3.1** Per rappresentare il risultato del lancio di un dado a sei facce l'insieme degli esiti è $\Omega = \{1, 2, 3, 4, 5, 6\}$. Le affermazioni “è uscito un numero pari” oppure “è uscito un numero maggiore di 3” si traducono con i sottoinsiemi $A = \{2, 4, 6\}$ e $B = \{4, 5, 6\}$; l'affermazione “non maggiore di 3” diventa $B^c = \{1, 2, 3\}$, l'affermazione “pari e maggiore di 3” diventa $A \cap B = \{4, 6\}$. Allo stesso modo “pari oppure maggiore di 3” diventa $A \cup B$. L'intero spazio Ω diventa l'evento che sicuramente si è realizzato, “è uscito un numero intero tra 1 e 6”, l'insieme vuoto \emptyset rappresenta un evento impossibile, come “è uscito 7”. ■

Esercizio 3.1 Verificare le seguenti corrispondenze tra operazioni logiche e operazioni insiemistiche:

$$\begin{aligned} \cap_i A_i &\leftrightarrow \text{si verificano tutti gli } A_i, \\ \cup_i A_i &\leftrightarrow \text{si verifica almeno un } A_i, \\ A^c = \Omega \setminus A &\leftrightarrow \text{non si verifica } A, \\ A \setminus B &\leftrightarrow \text{si verifica } A \text{ ma non } B, \\ A \Delta B := (A \setminus B) \cup (B \setminus A) &\leftrightarrow \text{si verifica o } A \text{ o } B. \end{aligned}$$

Verificare inoltre la seguente corrispondenza tra relazioni logiche e relazioni insiemistiche:

$$\begin{aligned} A \subseteq B &\leftrightarrow \text{se accade } A, \text{ allora accade } B, \\ A \cap B = \emptyset &\leftrightarrow \text{non possono accadere sia } A \text{ sia } B \text{ (eventi incompatibili)}. \end{aligned}$$

Alla luce di queste considerazioni, risulta chiara la necessità di operare con famiglie di eventi chiuse per le operazioni insiemistiche sopra menzionate, ovvero tali che le operazioni insiemistiche tra eventi producano ancora degli eventi. Questo è banale se la famiglia di eventi che consideriamo consiste di *tutti* i sottoinsiemi di Ω , $\mathcal{P}(\Omega)$, ma in generale questo non è possibile (fatto su cui torniamo più sotto). Siamo stati dunque condotti a dare la seguente definizione:

Definizione 3.1.1 Sia Ω un insieme e $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ una famiglia di sottoinsiemi di Ω . Diciamo che \mathcal{F} è un'algebra di parti se

- contiene l'insieme vuoto \emptyset e Ω , $\emptyset, \Omega \in \mathcal{F}$;
- è chiusa per complementari, ovvero se $A \in \mathcal{F}$ anche $A^c = \Omega \setminus A \in \mathcal{F}$;
- è chiusa per unioni (finite), ovvero se $A_1, \dots, A_n \in \mathcal{F}$ anche $A_1 \cup \dots \cup A_n \in \mathcal{F}$.

Se inoltre Ω e una sua algebra di parti \mathcal{F} hanno infiniti elementi, diciamo che \mathcal{F} è una σ -algebra

se soddisfa l'ipotesi addizionale:

- è chiusa per unioni numerabili, ovvero se esiste una successione di sottoinsiemi (eventi) $A_1, \dots, A_n, \dots \in \mathcal{F}$ anche la loro unione $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Esercizio 3.2 Usando le leggi di de Morgan, $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$, si mostri che un'algebra di parti è anche chiusa per intersezioni finite, ovvero se $A_1, \dots, A_n \in \mathcal{F}$ anche $A_1 \cap \dots \cap A_n \in \mathcal{F}$. Nel caso di una σ -algebra, si mostri che essa è chiusa per intersezioni numerabili. ■

La definizione di σ -algebra permette di eseguire tutte le operazioni insiemistiche che ci servono anche nel caso in cui Ω sia infinito, come accadrà diffusamente nel seguito.

■ **Esempio 3.2** Se vogliamo descrivere le nascite del prossimo anno in Italia dovremo considerare $\Omega = \mathbb{N}_0 = \{1, 2, \dots\}$, perchè sebbene l'esito sia certamente finito esso può assumere valori arbitrariamente grandi, anche se più grande è l'esito considerato, meno esso sarà probabile. ■

■ **Esempio 3.3** Lo spazio campionario Ω associato a due lanci (ordinati) di un dado è dato da

$$\{(i, j) \mid i, j \in \{1, 2, \dots, 6\}\} = \{1, 2, \dots, 6\}^2,$$

ovvero dall'insieme di tutte le coppie di possibili esiti. ■

Più in generale, se un esperimento è costituito da una successione (ordinata) di n sotto-esperimenti, allora il suo spazio campionario è

$$\Omega = \{(\omega_1, \omega_2, \dots, \omega_n) \mid \omega_1 \in \Omega_1, \dots, \omega_n \in \Omega_n\}$$

dove Ω_i è l'insieme degli esiti dell' i -simo sotto-esperimento.

Esercizio 3.3 In 3 lanci di dado, a quali insiemi corrispondono gli eventi “esce un numero pari al primo lancio” ed “esce 3 al primo lancio e un numero pari al terzo lancio”? ■

Il grado di fiducia che un evento si realizzi, chiamato probabilità, è rappresentato da un numero compreso tra 0 e 1; inoltre è intuitivo supporre che se due eventi sono incompatibili (ovvero sono disgiunti, hanno intersezione vuota) la probabilità che si realizzi uno qualsiasi dei due debba essere la somma delle probabilità dei singoli eventi. Questo equivale a dire che la probabilità è una funzione d'insieme (finitamente) additiva. Come sopra, la definizione generale deve tenere debitamente conto del caso in cui dobbiamo trattare infiniti eventi.

Definizione 3.1.2 Dato Ω un insieme e \mathcal{F} una σ -algebra di parti di Ω , una *misura di probabilità* (*probability measure*, anche semplicemente “una probabilità”) è una funzione $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ tale che

- l'evento certo ha probabilità unitaria, $\mathbb{P}(\Omega) = 1$;
- (σ -additività) se $(A_n)_{n=1,2,\dots}$ è una successione di elementi di \mathcal{F} a due a due disgiunti, si ha

$$\mathbb{P}\left(\bigcup_{n=1}^{+\infty} A_n\right) = \sum_{n=1}^{+\infty} \mathbb{P}(A_n).$$

Si dice *trascurabile* (*negligible*) un evento A tale che $\mathbb{P}(A) = 0$, e si dice *quasi certo* (*almost sure*) un evento A tale che $\mathbb{P}(A) = 1$.

Una terna $(\Omega, \mathcal{F}, \mathbb{P})$ formata da un insieme Ω , una σ -algebra \mathcal{F} di parti di Ω ed una probabilità \mathbb{P} definita su \mathcal{F} viene chiamata spazio di Probabilità.

Naturalmente la σ -additività implica l'additività su finiti sottoinsiemi disgiunti, cioè, dati A_1, \dots, A_N elementi di \mathcal{F} a due a due disgiunti,

$$\mathbb{P}\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N \mathbb{P}(A_n);$$

infatti, è sufficiente scegliere $A_n = \emptyset$ da un certo n in poi per specializzare la definizione generale al caso finito.

Ogni misura di probabilità verifica le seguenti proprietà, la cui dimostrazione è lasciata per esercizio:

Proposizione 3.1.1 In uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, abbiamo:

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ e di conseguenza $\mathbb{P}(\emptyset) = 0$;
- se $B \subset A$, $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B)$, dove ricordiamo che la differenza insiemistica corrisponde a $A \setminus B = A \cap B^c$;
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
- $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$.

Esercizio 3.4 Dimostrare questa proposizione. Come si estendono le ultime due formule nel caso di 4 o più eventi? ■

La σ -additività permette di “passare al limite”, nel senso reso preciso dal seguente risultato

Proposizione 3.1.2 Data una successione di insiemi $A_1, \dots, A_n, \dots \in \mathcal{F}$, assumiamo (alternativamente) che

- la successione sia crescente, ovvero $A_n \subseteq A_{n+1}$, e poniamo $A = \bigcup_{n=1}^{+\infty} A_n = \lim_{n \rightarrow \infty} A_n$;
- la successione sia decrescente, ovvero $A_n \supseteq A_{n+1}$, e poniamo $A = \bigcap_{n=1}^{+\infty} A_n = \lim_{n \rightarrow \infty} A_n$.

In entrambi i casi vale

$$\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n).$$

Dimostrazione. Consideriamo dapprima il caso in cui la successione è crescente. Poniamo $B_1 = A_1, B_n = A_n \setminus A_{n-1}$ per $n > 1$. Gli insiemi $(B_n)_{n \geq 1}$ sono a due a due disgiunti e per l'additività finita si ha $\mathbb{P}(B_n) = \mathbb{P}(A_n) - \mathbb{P}(A_{n-1})$. Poiché $\bigcup_{n \geq 1} A_n = \bigcup_{n \geq 1} B_n$, si ha $\mathbb{P}(A) = \sum_{n=1}^{+\infty} \mathbb{P}(B_n) = \lim_{n \rightarrow \infty} \sum_{h=1}^n \mathbb{P}(B_h) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$.

Nel caso in cui la successione è decrescente la tesi segue dal caso crescente semplicemente passando al complementare. ■

In effetti si può dimostrare che se una funzione $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ definita su una σ -algebra è finitamente additiva (vale l'additività su famiglie finite di eventi) ed inoltre soddisfa la tesi della Proposizione precedente, allora è anche σ -additiva.

Prima di procedere oltre, torniamo sulla questione del perché in alcune situazioni la probabilità è assegnata solo su alcuni e non tutti i sottoinsiemi dello spazio Ω , ovvero perché abbiamo ammesso che \mathcal{F} non coincida con $\mathcal{P}(\Omega)$, l'insieme delle parti di Ω (che è sempre una σ -algebra). Il motivo di questo è una difficoltà esclusivamente di ordine matematico, ovvero l'impossibilità, in alcuni casi, di estendere la probabilità a tutti i sottoinsiemi dell'insieme Ω .

■ **Esempio 3.4** Cerchiamo di rappresentare la scelta casuale di un punto del segmento $[0, 1]$, in modo che tale scelta sia “uniforme”, ovvero che tutti i punti abbiano la stessa probabilità di essere scelti (su questi concetti torneremo diffusamente in seguito). Scegliamo dunque $\Omega = [0, 1]$. Se ogni punto avesse la stessa probabilità non nulla, la probabilità dell'intero Ω sarebbe infinita, quindi non possiamo costruire così il nostro modello. Invece, è ragionevole assegnare a un evento (sottoinsieme) $A \subset \Omega = [0, 1]$ una probabilità uguale alla sua lunghezza: in effetti così facendo la

probabilità di tutto Ω è 1, e dovrebbero essere soddisfatte anche le altre richieste che facciamo alla funzione \mathbb{P} . Tuttavia, si può dimostrare (ed è una questione alquanto delicata) che *non è possibile definire la lunghezza di un qualsiasi sottoinsieme del segmento*, quindi in questo caso la σ -algebra non può essere $\mathcal{P}(\Omega)$, e la sua scelta non è banale. ■

Questo problema non si pone se Ω è numerabile, ovvero se si possono indicizzare i suoi elementi con (un sottoinsieme possibilmente finito de)i numeri naturali \mathbb{N} ,

$$\Omega = (\omega_1, \omega_2, \dots, \omega_n, \dots).$$

In tal caso la probabilità è univocamente determinata dai numeri

$$p_i = \mathbb{P}(\omega_i) = \mathbb{P}((\omega_i)) \in [0, 1]$$

(ometteremo d'ora in avanti le doppie parentesi nelle probabilità di insiemi con un singolo elemento) e per ogni evento $A \subset \Omega$ si ha $\mathbb{P}(A) = \sum_{\omega_i \in A} p_i$. Il simbolo di somma qui indica una somma finita se A ha finiti elementi (ovvero sempre nel caso in cui Ω stesso sia finito), e la somma di una serie¹ se A ha infiniti elementi.

3.2 Calcolo Combinatorio

Un esempio significativo di spazio di probabilità è il modello uniforme:

Definizione 3.2.1 Una distribuzione di probabilità \mathbb{P} si dice uniforme su Ω se Ω è un insieme finito, $\mathcal{F} = \mathcal{P}(\Omega)$ e gli eventi elementari ω_i sono equiprobabili.

Ne sono esempi i modelli associati all'estrazione da un'urna e al lancio di un dado equilibrato. Come accennato sopra, non esiste una distribuzione uniforme di probabilità su un insieme Ω numerabile ma infinito.

Per una probabilità uniforme \mathbb{P} su uno spazio finito Ω , si ottiene la formula

$$\mathbb{P}(A) = \frac{\#A}{\#\Omega} = \frac{\text{"casi favorevoli"}}{\text{"casi possibili"}}, \quad A \subseteq \Omega,$$

dove con $\#A$ si indica la cardinalità (o numero degli elementi) dell'insieme A .

In questo ambito, i problemi diventano molto spesso calcoli combinatori: riportiamo qui (ed useremo nel seguito) solamente tre formule notevoli. Dato un intero n , anziché dire “un insieme di n elementi”, scriveremo per brevità $\{1, \dots, n\}$.

Proposizione 3.2.1 Siano k ed n due interi.

- il numero di sequenze ordinate, possibilmente con ripetizione, di k numeri da 1 a n , cioè il numero di funzioni da $\{1, \dots, k\}$ a $\{1, \dots, n\}$ è n^k ;
- il numero di modi in cui si possono ordinare gli elementi di $\{1, \dots, n\}$ (ovvero il numero di funzioni biettive dall'insieme a se stesso, o di permutazioni di n elementi) è

$$n! = 1 \cdot 2 \cdots (n-1) \cdot n;$$

- se $0 \leq k \leq n$, il numero di sottinsiemi di $\{1, \dots, n\}$ formati da k elementi (coefficiente binomiale) è

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

¹ Si può dimostrare che non esistono serie convergenti di termini non-negativi con una quantità più che numerabile di addendi, da cui l'impossibilità di estendere queste affermazioni al caso in cui Ω è più che numerabile.

dove $n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$ per $n \in \mathbb{N}_0$ e $0! = 1$.

Questi enunciati si dimostrano facilmente per induzione su n . Il coefficiente binomiale è così detto perché interviene nella formula del binomio di Newton:

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}, \quad a, b \in \mathbb{R}, n \in \mathbb{N}.$$

A titolo d'esempio, e lasciando la dimostrazione per esercizio, riportiamo due formule che si possono dedurre dalle precedenti.

Esercizio 3.5 Siano $0 \leq k \leq n$: il numero di sequenze ordinate senza ripetizione di k numeri di $\{1, \dots, n\}$ è $\frac{n!}{(n-k)!}$. Si noti che questo numero coincide anche con il numero delle applicazioni iniettive da $\{1, \dots, k\}$ in $\{1, \dots, n\}$. ■

Esercizio 3.6 Siano k_1, \dots, k_h interi con $k_1 + \dots + k_h = n$: il numero di modi in cui si possono scegliere h sottinsiemi di $\{1, \dots, n\}$ formati rispettivamente da k_1, \dots, k_h elementi è

$$\binom{n}{k_1, \dots, k_h} = \frac{n!}{k_1! \dots k_h!}.$$

■ **Esempio 3.5** Consideriamo l'esperimento dato da k estrazioni ordinate, con rimpiazzo, da una popolazione S di n oggetti, ad esempio un'urna di n biglie (per rimpiazzo si intende che l'oggetto estratto viene nuovamente inserito nell'urna e quindi può essere selezionato nelle estrazioni successive). Questo esperimento è descritto dallo spazio di probabilità $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ con $\Omega = S^n$ e \mathbb{P} uniforme: infatti tutte le possibili sequenze di k oggetti estratti sono equiprobabili. In particolare $P(A) = \#A/\#\Omega$ con $\#\Omega = n^k$.

Lo stesso modello si può usare per k lanci di un dado equilibrato (qui $S = \{1, 2, \dots, 6\}$) e k lanci di moneta equilibrata (qui $S = \{\text{testa}, \text{croce}\}$). ■

■ **Esempio 3.6** Consideriamo l'esperimento dato da k estrazioni ordinate, questa volta senza rimpiazzo, da una popolazione S di n oggetti, con $k \leq n$. Anche in questo caso il modello è uniforme, ma con un diverso spazio campionario: $\Omega = \{(x_1, \dots, x_k) \mid x_i \in S, x_i \text{ tutti distinti}\}$, con $\#\Omega = n!/(n-k)!$. ■

Esercizio 3.7 Consideriamo ora l'estrazione senza ordine e senza rimpiazzo di 2 oggetti da un'urna di 10 oggetti (che indichiamo con $\{1, 2, \dots, 10\}$); qui “senza ordine” indica che ignoriamo l'ordine di estrazione, quindi ad esempio consideriamo uguali gli esiti $(3, 7)$ e $(7, 3)$. Scrivere lo spazio campionario corrispondente e la sua cardinalità; la misura di probabilità è uniforme?

Se si considera l'estrazione senza ordine ma con rimpiazzo? ■

3.3 Probabilità Condizionata e formula di Bayes

Quando si è a conoscenza della realizzazione di un evento, cambia la valutazione di probabilità di ogni altro evento: ad esempio se si sa che il numero uscito dal lancio di un dado è pari, la probabilità che sia uscito il numero 6 non è più $\frac{1}{6}$, ma $\frac{1}{3}$. Infatti, se si è realizzato l'evento $B = \{2, 4, 6\}$ (cioè è uscito un numero pari) sono rimasti 3 casi possibili dei quali uno è favorevole: se indichiamo con $A = \{6\}$, notiamo che la nuova probabilità che è stata attribuita ad A verifica la formula $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Si possono fornire diversi esempi simili che sempre verificano la formula sopra riportata: queste considerazioni sono all'origine della definizione che segue.

Definizione 3.3.1 Assegnato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ ed un evento B non trascurabile, si chiama probabilità condizionata di A rispetto a B il numero

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

Essa indica la probabilità che viene associata all'evento A , coerentemente con la valutazione precedentemente assegnata, in seguito all'informazione che si è realizzato l'evento B . È facile verificare che, fissato B non trascurabile, la funzione

$$A \mapsto \mathbb{P}(A | B)$$

è effettivamente una probabilità.

Esercizio 3.8 Verificare con un esempio che la funzione $B \mapsto \mathbb{P}(A | B)$ non è una probabilità. ■

Dati due eventi A e B non trascurabili, è immediato constatare che

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

Più in generale vale la seguente:

Proposizione 3.3.1 — Condizionamento ripetuto. Siano A_1, \dots, A_n eventi, e supponiamo che $A_1 \cap \dots \cap A_{n-1}$ sia non trascurabile: vale la formula

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 | A_1) \dots \mathbb{P}(A_n | A_1 \cap \dots \cap A_{n-1}).$$

La dimostrazione si ottiene (per induzione) esplicitando i vari termini; si noti che, se $1 \leq k < n-1$, anche $A_1 \cap \dots \cap A_k$ è non trascurabile.

Definizione 3.3.2 Una partizione di Ω è una collezione di n eventi B_1, \dots, B_n a due a due disgiunti, tali che $B_1 \cup \dots \cup B_n = \Omega$. Un sistema di alternative è una partizione di Ω in eventi non trascurabili.

Teorema 3.3.2 — Formula della probabilità totale o formula di fattorizzazione. Sia B_1, \dots, B_n un sistema di alternative. Allora, per un qualunque evento A , vale

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

Dimostrazione. Si noti che $A = (A \cap B_1) \cup \dots \cup (A \cap B_n)$, come unione di eventi a due a due disgiunti. Si ha pertanto

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

■

La formula della probabilità totale si usa in esempi in cui non è nota la probabilità \mathbb{P} ma sono note le probabilità condizionate a un sistema di alternative, cioè $\mathbb{P}(\cdot | B_i)$, come si vede da queste due classiche applicazioni:

■ **Esempio 3.7** Siano u_1 e u_2 due urne, la prima contenente 5 biglie rosse e 5 blu, la seconda 8 biglie rosse e 2 blu. L'esperimento in esame consiste nello scegliere casualmente una delle due

urne e poi nell'estrarre una biglia dall'urna scelta e osservarne il colore. Vogliamo calcolare la probabilità di estrarre una biglia rossa. Lo spazio campionario è dato da

$$\Omega = \{(x, y) \mid x \in \{u_1, u_2\}, y \in \{r, b\}\} = \{u_1, u_2\} \times \{r, b\}$$

(dove r, b stanno per “rosso” e “blu”)². Sulla probabilità sappiamo che la scelta dell'urna è equiprobabile, cioè

$$\mathbb{P}(\text{urna 1}) = \mathbb{P}(\text{urna 2}) = 1/2,$$

e che, data la scelta dell'urna 1, rispettivamente dell'urna 2, la probabilità di una biglia rossa è $5/10$, rispettivamente $8/10$, cioè

$$\mathbb{P}(\text{biglia rossa} \mid \text{urna 1}) = 5/10,$$

$$\mathbb{P}(\text{biglia rossa} \mid \text{urna 2}) = 8/10.$$

Grazie alla formula della probabilità totale abbiamo

$$\mathbb{P}(\text{biglia rossa}) = \mathbb{P}(\text{biglia rossa} \mid \text{urna 1})\mathbb{P}(\text{urna 1}) + \mathbb{P}(\text{biglia rossa} \mid \text{urna 2})\mathbb{P}(\text{urna 2}) = 13/20.$$

■

Esercizio 3.9 Un test diagnostico per una data malattia ha indice di sensibilità 0.99, cioè, se una persona è malata, il test dà esito positivo con probabilità 0.99; inoltre il test ha indice di specificità 0.97, cioè, se una persona è sana, il test dà esito negativo con probabilità 0.97. Supponiamo che l'1% della popolazione soffra di tale malattia. Qual è la probabilità che il test, effettuato su un individuo scelto a caso, dia esito positivo? ■

Teorema 3.3.3 — Formula di Bayes. Siano A e B due eventi non trascurabili. Allora vale la formula

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

Come conseguenza, se B_1, \dots, B_n è un sistema di alternative e A è un evento non trascurabile, vale

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}{\sum_{j=1}^n \mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}.$$

Dimostrazione. La prima formula segue dalla definizione di probabilità condizionata:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

La seconda formula segue dalla prima e dalla formula della probabilità totale. ■

La formula di Bayes si applica per “invertire il condizionamento”, tipicamente in contesti in cui accade un evento A riferito a un “osservabile” (negli esempi che seguono, colore della biglia, esito del test) e vogliamo dedurre la probabilità di un'alternativa B_i di eventi “causa” (urna scelta, persona sana o malata).

²Nel seguito, chiamiamo l'evento “viene scelta l'urna 1”, cioè $\{u_1\} \times \{r, b\}$ (da non confondersi con u_1), e anche “biglia rossa” l'evento “viene estratta una biglia rossa”, cioè $\{u_1, u_2\} \times \{r\}$ (da non confondersi con r), e così via.

■ **Esempio 3.8** Nel contesto dell'Esempio 3.7, supponiamo di estrarre una biglia rossa, vogliamo calcolare la probabilità che la biglia provenga dall'urna 1. Per il teorema di Bayes abbiamo

$$\mathbb{P}(\text{urna 1} \mid \text{biglia rossa}) = \frac{\mathbb{P}(\text{biglia rossa} \mid \text{urna 1})\mathbb{P}(\text{urna 1})}{\mathbb{P}(\text{biglia rossa})} = \frac{5/10 \cdot 1/2}{13/20} \approx 0.385.$$

Esercizio 3.10 Nel contesto dell'Esercizio 3.9, supponiamo che il test dia risultato positivo. Qual è la probabilità che la persona sia effettivamente malata? piuttosto bassa! ■

La formula della probabilità totale e il teorema di Bayes sono valide anche se il sistema di alternative anziché essere finito è numerabile, naturalmente sostituendo alle somme finite le somme di una serie.

Esercizio 3.11 Qual è la probabilità che, in una estrazione del lotto, tutti e 5 i numeri estratti siano inferiori o eguali a 20? (Lo si calcoli dapprima con un ragionamento combinatorio, e quindi con la formula di Bayes). ■

3.4 Indipendenza

Vogliamo codificare nel formalismo matematico l'idea che la conoscenza che si è realizzato un certo evento non modifica la valutazione di probabilità di un altro evento, e viceversa. Consideriamo quindi due eventi A e B (non trascurabili) e imponiamo che

$$\mathbb{P}(A) = \mathbb{P}(A \mid B), \quad \mathbb{P}(B) = \mathbb{P}(B \mid A).$$

Un esame immediato mostra che queste richieste sono equivalenti tra loro, ed equivalenti all'eguaglianza $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. A differenza delle due precedenti, quest'ultima è simmetrica rispetto ai due eventi ed ha senso anche se uno dei due (o anche tutti e due) sono trascurabili: viene scelta dunque come definizione di indipendenza.

Definizione 3.4.1 Due eventi A e B sono indipendenti se vale

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

Lasciamo come (facile) esercizio le seguenti proprietà:

Esercizio 3.12

- Se A e B sono indipendenti, lo sono anche A^c e B , A e B^c , A^c e B^c ;
- se $\mathbb{P}(A) = 0$ oppure $\mathbb{P}(A) = 1$, A è indipendente da qualsiasi altro evento;
- due eventi disgiunti non possono essere indipendenti, a meno che almeno uno dei due sia trascurabile.

Esercizio 3.13 Un mazzo di carte Italiane è composto da 40 carte, divise in 4 semi –bastoni, coppe, denari, spade– e ciascun seme ha carte numerate da 1 a 10 (al lettore la scelta della variante grafica dei semi). Si estrae una carta a caso: gli eventi “esce un asso” (cioè un 1) ed “esce una carta di denari” sono indipendenti? ■

Vediamo ora come si estende la nozione di indipendenza al caso di più eventi, cominciando con tre. Affinché A, B e C siano indipendenti, occorre anzitutto che siano soddisfatte le tre eguaglianze

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

cioè che siano a due a due indipendenti, ma deve anche essere soddisfatta l'eguaglianza

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

Non basta infatti che siano a due a due indipendenti, come è mostrato nell'esempio seguente.

■ **Esempio 3.9** Sull'insieme $\Omega = \{1, 2, 3, 4\}$ munito della distribuzione uniforme di probabilità, gli eventi $A = \{1, 2\}$, $B = \{1, 3\}$ e $C = \{2, 3\}$ sono a due a due indipendenti. Tuttavia, valgono anche

$$\mathbb{P}(A \cap B \mid C) = 0 \neq \frac{1}{4} = \mathbb{P}(A \cap B), \quad \mathbb{P}(C \mid A \cap B) = 0 \neq \frac{1}{2} = \mathbb{P}(C),$$

ovvero conoscere l'evento C fornisce informazione su $A \cup B$ e viceversa. Non è ragionevole dunque dire che A, B, C sono globalmente indipendenti. ■

Per un numero arbitrario di eventi, si dà la seguente:

Definizione 3.4.2 Assegnati n eventi A_1, \dots, A_n , questi si dicono indipendenti se per ogni intero k con $2 \leq k \leq n$ e per ogni scelta di interi $1 \leq i_1 < i_2 < \dots < i_k \leq n$, vale l'eguaglianza

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \dots \mathbb{P}(A_{i_k})$$

Quando il numero di eventi cresce, il numero di eguaglianze da verificare sale notevolmente³. A priori questo appare essere un problema teorico, soprattutto nella descrizione probabilistica di un esperimento ripetuto più volte nelle medesime condizioni (che sarà uno dei nostri obiettivi principali): ci attendiamo infatti che gli eventi associati a ripetizioni distinte siano indipendenti. La complicazione è però solo apparente: eventi (ed in seguito variabili aleatorie) indipendenti sono facilmente costruiti considerando spazi prodotto. Ne vediamo un esempio paradigmatico con il risultato che segue.

Proposizione 3.4.1 Si consideri

$$\Omega = \{a = (a_1, \dots, a_n) \mid a_i = 0, 1\} = \{0, 1\}^n, \quad \mathcal{F} = \mathcal{P}(\Omega),$$

su cui definiamo, per ogni singolo $a = (a_1, \dots, a_n) \in \Omega$,

$$\mathbb{P}(\{a\}) = p^{\#\{i: a_i=1\}} (1-p)^{\#\{i: a_i=0\}} = p^{\sum_{i=1}^n a_i} (1-p)^{n - \sum_{i=1}^n a_i}.$$

Lo spazio $(\Omega, \mathcal{F}, \mathbb{P})$ è di probabilità, e gli eventi

$$A_i = \{a \in \Omega : a_i = 1\}, \quad i = 1, \dots, n,$$

sono tutti indipendenti tra di loro, così come i loro complementari A_i^c .

Lo spazio di probabilità suddetto descrive ad esempio n lanci di una moneta, in generale non equilibrata se $p \neq 1/2$. Nel caso di una moneta equilibrata, $p = 1/2$, siccome i due esiti di un lancio hanno la stessa probabilità, ogni stringa di risultati ha la stessa probabilità di accadere, e \mathbb{P} è la distribuzione uniforme su $\Omega = \{0, 1\}^n$.

Dimostrazione. È sufficiente mostrare che vale, per ogni $k = 0, 1, \dots, n$,

$$\mathbb{P}(B_k) = p^k, \quad B_k = \{a \in \Omega : a_{n-k} = \dots = a_n = 1\}, \quad B_0 = \Omega$$

(sono gli eventi che fissano le ultime k coordinate). Scegliendo $k = 0$ si mostra che $\mathbb{P}(\Omega) = 1$, quindi lo spazio $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ è di probabilità. Inoltre, per verificare l'indipendenza degli eventi A_i dobbiamo considerare le intersezioni

$$A_{i_1} \cap \dots \cap A_{i_k} = \{a \in \Omega : a_{i_1} = \dots = a_{i_k} = 1\}$$

³Per n eventi le eguaglianze da verificare sono $2^n - n - 1$.

per ogni scelta della famiglia di indici i_1, \dots, i_k . Non è difficile convincersi che riordinando gli indici (questo non cambia le probabilità degli eventi come si vede dalla definizione di \mathbb{P}) si ottiene $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(B_k)$. Vale quindi

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(B_k) = p^k = \mathbb{P}(B_1)^k = \mathbb{P}(A_{i_1}) \dots \mathbb{P}(A_{i_k}),$$

(usando a destra il ragionamento per $k = 1$) da cui la tesi.

Calcoliamo quindi la probabilità di B_k :

$$\begin{aligned} \mathbb{P}(B_k) &= \sum_{a \in B_k} \mathbb{P}(\{a\}) = \sum_{a_1, \dots, a_{n-k}=0,1} \mathbb{P}(\{(1, \dots, 1, a_{k+1}, \dots, a_n)\}) \\ &= \sum_{a_1, \dots, a_{n-k}=0,1} p^k p^{\sum_{i=1}^{n-k} a_i} (1-p)^{n-\sum_{i=1}^{n-k} a_i} = p^k \sum_{h=0}^{n-k} \binom{n-k}{h} p^h (1-p)^{n-k-h} = p^k, \end{aligned}$$

in cui passando dalla prima alla seconda riga abbiamo usato il fatto che ci sono $\binom{n-k}{h}$ scelte delle entrate a_1, \dots, a_{n-k} in cui esattamente h sono uguali a 1. ■

Esercizio 3.14 Si lanciano due monete equilibrate, e poi se ne lancia una truccata come segue: esce testa se i primi due lanci sono concordi, croce altrimenti. Introdurre uno spazio di probabilità per descrivere l'esperimento, e verificare che gli eventi $A = \{\text{testa al primo lancio}\}$, $B = \{\text{testa al secondo lancio}\}$, $C = \{\text{testa al terzo lancio}\}$ sono a due a due indipendenti ma non sono globalmente indipendenti. ■

NB La terminologia può *erroneamente* indurre a pensare che l'indipendenza probabilistica sia legata al concetto di causalità. Quest'ultimo, in effetti, nell'ambito della Statistica o del calcolo delle Probabilità non ha una definizione univoca nè una generalmente accettata.

Come mostra l'esercizio precedente, due eventi possono essere indipendenti anche in presenza di una qualche relazione causale (il primo lancio determina assieme al secondo l'esito del terzo). Viceversa, due eventi possono essere dipendenti anche in assenza di una relazione causale. Ad esempio, i dinosauri sono estinti e, per quanto ne sappiamo, nessuno di loro sapeva leggere: dato un certo fossile di dinosauro gli eventi "appartiene a una specie estinta" e "appartiene a una creatura che non sapeva leggere" si verificano sempre assieme, ovvero non sono indipendenti, e tuttavia l'analfabetismo non si annovera tra le cause dell'estinzione dei dinosauri.

3.5 Probabilità sulla retta reale

Introduciamo in questo paragrafo due classi di esempi di probabilità sulla retta reale $\Omega = \mathbb{R}$. Questi esempi *non sono esaustivi* (come vedremo nel prossimo capitolo), però nelle applicazioni ci si riduce praticamente sempre a questi casi.

Probabilità discreta

Abbiamo già sostanzialmente discusso la prima classe di esempi: una *probabilità discreta* (o atomica) è una probabilità sullo spazio $\Omega = \mathbb{R}$ con la σ -algebra $\mathcal{F} = \mathcal{P}(\mathbb{R})$, che sia concentrata su una successione (finita o numerabile) di punti $x_1, x_2, \dots \in \mathbb{R}^4$. Posto $p_i = p(x_i) = \mathbb{P}(x_i)$, vale

$$\mathbb{P}(A) = \sum_{i: x_i \in A} p(x_i), \quad \forall A \subseteq \mathbb{R}. \quad (3.1)$$

⁴In questo caso in effetti potremmo semplicemente ridurci a $\mathcal{F} = \mathcal{P}(\{x_1, x_2, \dots\}) \subset \mathcal{P}(\mathbb{R})$, ma non c'è problema a considerare tutti i sottoinsiemi di \mathbb{R} , a differenza dell'esempio successivo.

Definizione 3.5.1 Si chiama *funzione di massa* o anche *densità discreta* della probabilità discreta \mathbb{P} la funzione $p(x_i) = \mathbb{P}(x_i)$.

La densità discreta soddisfa $p(x_i) \geq 0$ e $\sum_{i=1,2,\dots} p(x_i) = 1$. Viceversa, date una successione di punti x_1, x_2, \dots in \mathbb{R} e una funzione $p(x_i)$ di tali punti che soddisfa $p(x_i) \geq 0$ e $\sum_{i=1,2,\dots} p(x_i) = 1$, la formula (3.1) definisce un'unica probabilità discreta \mathbb{P} su $\Omega = \mathbb{R}$ avente $p(x_i)$ per densità discreta. In particolare, grazie alla formula (3.1), per conoscere una probabilità discreta basta conoscere i valori sui quali è concentrata (cioè x_1, x_2, \dots) e i numeri $p_i = p(x_i)$.

Si può anche pensare di definire la funzione di massa $p(x) = \mathbb{P}(x)$ su tutta la retta (naturalmente sarà eguale a 0 fuori dei punti x_1, x_2, \dots) e allora la sola funzione di massa identifica la probabilità.

Probabilità definite da densità

Prima di affrontare la seconda classe di esempi, riportiamo il solo enunciato del modo in cui matematicamente viene risolto il problema sollevato nell'esempio 3.4.

Teorema 3.5.1 Esiste una σ -algebra $\mathcal{F} \subset \mathcal{P}(\mathbb{R})$ (che *non* coincide con $\mathcal{P}(\mathbb{R})$) tale che tutti gli intervalli (a, b) , $-\infty \leq a \leq b \leq +\infty$ (incluse quindi le semirette) appartengono ad \mathcal{F} . Inoltre, esiste una funzione $\lambda : \mathcal{F} \rightarrow \mathbb{R} \cup \{\infty\}$ che è σ -additiva (nel senso della Definizione 3.1.2) e tale che, per ogni $-\infty \leq a \leq b \leq +\infty$,

$$\lambda((a, b)) = \lambda((a, b]) = \lambda([a, b)) = \lambda([a, b]) = b - a.$$

In effetti, esistono più σ -algebre con le suddette proprietà, ma questo ai nostri fini non è rilevante: quello che conta è poter misurare tutti i segmenti (incluse le semirette) e sottoinsiemi prodotti da essi con semplici operazioni insiemistiche (quelle consentite dalla definizione di σ -additività). Chiameremo gli elementi di \mathcal{F} *insiemi misurabili*. Intersecando tutti gli elementi di \mathcal{F} con l'intervallo $[0, 1]$ otteniamo i sottoinsiemi misurabili di quest'ultimo, che sono una σ -algebra sulla quale possiamo definire la probabilità $\mathbb{P}(A) = \lambda(A)$ per ogni $A \subset [0, 1]$ misurabile, in modo che

$$\mathbb{P}((a, b)) = |b - a|, \quad \forall a, b \in [0, 1].$$

Vale la pena notare che λ (come di conseguenza la probabilità definita dall'ultima equazione) assegna lunghezza nulla ai singoli punti, ovvero $\lambda(x) = \lambda([x, x]) = x - x = 0$. Lo stesso vale per ogni sottoinsieme $A \subset \mathbb{R}$ con al più numerabili elementi: questa è una conseguenza immediata della σ -additività.

NB **Una premessa importante.** Nel seguito faremo diffusamente uso di integrali di funzioni sulla retta reale. Ai fini del rigore matematico è necessario in questo contesto introdurre la nozione di integrale secondo Lebesgue e le sue proprietà. Ciò va oltre lo scopo di queste note, per cui nel seguito ogni integrale e l'integrabilità delle funzioni possono essere intesi nel senso di Riemann. Ometteremo tutti i dettagli che richiedono ulteriore discussione teorica, facendone però debitamente menzione. Essi non sono infatti dettagli trascurabili nella Teoria della Probabilità: il lettore sia conscio che tali omissioni comportano precise limitazioni nei nostri argomenti che vanno tenute a mente.

Definizione 3.5.2 Si chiama densità di probabilità sulla retta reale una funzione non-negativa $f : \mathbb{R} \rightarrow [0, +\infty)$, integrabile e tale che $\int_{-\infty}^{+\infty} f(x)dx = 1$. Ad ogni densità di probabilità si associa un'unica probabilità sulla σ -algebra degli insiemi misurabili di \mathbb{R} definita da:

$$\forall A \in \mathcal{F}, \quad \mathbb{P}(A) = \int_A f(x)dx.$$

Perchè questa sia una buona definizione va controllato che \mathbb{P} così definita sia una probabilità nel senso della Definizione 3.1.2. Anzitutto, $\mathbb{P}(\mathbb{R}) = \int_{\mathbb{R}} f(x)dx = 1$ per ipotesi, ed inoltre se $A \cap B = \emptyset$ si ha

$$\mathbb{P}(A \cup B) = \int_{A \cup B} f(x)dx = \int_A f(x)dx + \int_B f(x)dx = \mathbb{P}(A) + \mathbb{P}(B).$$

Omettiamo il controllo della σ -additività, per il quale sono necessari risultati sull'integrazione secondo Lebesgue (in particolare, il *Lemma di Beppo-Levi* o di convergenza monotona).

L'Esempio 3.4 (un numero scelto a caso tra 0 e 1) è un caso particolare di probabilità associata a una densità, nel quale la densità è definita da

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1, \\ 0 & \text{altrove.} \end{cases}$$

NB Come discusso in quel caso particolare, la probabilità definita da una densità non può essere costruita su tutti i sottinsiemi di \mathbb{R} , ma solo sugli insiemi misurabili. Nelle applicazioni e negli esempi non capiterà mai di incontrare un insieme non misurabile: d'ora innanzi scriveremo $A, B, \dots \subseteq \mathbb{R}$ assumendo sempre che siano misurabili.

Vale in generale per le probabilità definite da densità l'osservazione fatta nell'Esempio 3.4 sulla probabilità di un singolo punto: se X ha densità f ,

$$\mathbb{P}(X = t) = \mathbb{P}_X(t) = \int_{\{t\}} f(x)dx = 0, \quad \forall t \in \mathbb{R},$$

e vale più in generale $\mathbb{P}(A) = 0$ per ogni $A \subset \mathbb{R}$ con al più numerabili elementi.

Esercizio 3.15 — Una probabilità non discreta e non definita da densità. Introduciamo la seguente notazione: dato un insieme $S \subset \mathbb{R}$ la sua *funzione indicatrice* è

$$\chi : \mathbb{R} \rightarrow \{0, 1\}, \quad \chi_S(x) = \begin{cases} 1 & x \in S, \\ 0 & x \notin S. \end{cases}$$

Si consideri lo spazio di probabilità così definito: $\Omega = [0, 1]$, \mathcal{F} i sottoinsiemi misurabili e, per ogni $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \frac{1}{2}|A| + \frac{1}{2}\chi_{\{2\}}(A).$$

Si mostri che questa è una misura di probabilità, e si osservi che non è discreta, ma non è nemmeno definita tramite una densità. ■

3.6 Esercizi

Esercizio 3.16 Da un mazzo di 40 carte napoletane (quattro semi: bastoni, coppe, denari, spade; ciascuno seme ha carte da 1 a 10) vengono estratte tre carte, senza rimpiazzo. Qual è la probabilità che le prime due siano di denaro e la terza di coppe? Qual è la probabilità che escano esattamente due carte di denaro (in qualunque ordine)? ■

Esercizio 3.17 Tre turisti (tra loro estranei) arrivano in un paese con cinque alberghi, avendo prenotato ciascuno un albergo in modo del tutto casuale. a) Qual è la probabilità che si trovino tutti in alberghi differenti? b) Qual è la probabilità che si trovino tutti nello stesso albergo? ■

Esercizio 3.18 In una classe di 10 alunni, di cui 6 donne e 4 uomini, vengono estratte a sorte 5 persone. Qual è la probabilità che, tra le persone estratte, ci siano esattamente 3 donne? Qual è la probabilità che ci siano almeno 4 donne? ■

Esercizio 3.19 Una ditta classifica le sue consegne come in città o fuori città e come urgenti o non urgenti. Sappiamo che il 50% delle consegne è urgente, il 40% è in città e il 20% è urgente e fuori città. Calcolare le probabilità che una consegna sia rispettivamente: a) urgente e in città; b) urgente e fuori città; c) urgente, sapendo che è in città. Gli eventi “la consegna è urgente” e “la consegna è in città” sono indipendenti? ■

Esercizio 3.20 Secondo le previsioni meteo, domani a Pisa ci sarà: pioggia con probabilità del 70%, vento con probabilità del 50%, pioggia ma non vento con probabilità del 40%. Calcolare le probabilità di: a) pioggia o vento; b) pioggia e vento; c) né pioggia né vento; d) vento, sapendo che poverà. ■

Esercizio 3.21 Un'urna contiene 10 biglie, di cui 5 rosse e 5 blu.

- Effettuiamo 4 estrazioni con reimmissione. Qual è la probabilità che almeno 3 biglie su 4 siano rosse?

Una seconda urna contiene 8 biglie rosse e 2 blu. Scegliamo un'urna a caso ed estraiamo da questa 4 biglie con reimmissione.

- Qual è la probabilità che almeno 3 siano rosse?
 - Se almeno 3 biglie estratte sono rosse, qual è la probabilità che abbiamo scelto la prima urna?
 - Gli eventi “rossa alla prima estrazione” e “rossa alla seconda estrazione” sono indipendenti?
-

Probabilità Condizionata

Esercizio 3.22 In una famiglia ci sono due figli. Se almeno uno dei due è femmina, qual è la probabilità che siano entrambi femmine? ■

Esercizio 3.23 In due lanci di un dado equilibrato, gli eventi “1 al primo lancio” e “somma degli esiti dei due lanci = 3” sono indipendenti? ■

Esercizio 3.24 Nella città di Urbopoli, si stima che, tra coloro che abitualmente utilizzano gli autobus pubblici, il 20% lo faccia sistematicamente senza pagare il biglietto. Tra coloro che non pagano il biglietto, il 70% sono individui di età inferiore a 15 anni, mentre tra coloro che lo pagano chi ha meno di 15 anni rappresenta il 40%.

1. Si sceglie a caso un individuo (nella popolazione di coloro che utilizzano gli autobus); quanto vale la probabilità che si tratti di un individuo di età inferiore a 15 anni?
2. Se l'individuo scelto ha meno di 15 anni, quanto vale la probabilità che egli paghi

abituamente il biglietto?

3. Se l'individuo scelto ha 15 anni o più, quanto vale la probabilità che egli abitualmente non paghi il biglietto?

Esercizio 3.25 Un'assicurazione auto classifica i guidatori come prudenti o imprudenti. Un guidatore prudente ha almeno un incidente l'anno con probabilità dell'1%, un guidatore imprudente con probabilità del 5%. Il 60% dei guidatori assicurati è classificato come prudente.

- Qual è la probabilità che un assicurato abbia almeno un incidente in un dato anno?
- Se un assicurato ha avuto almeno un incidente, qual è la probabilità che sia classificato come prudente?

Esercizio 3.26 Un furto è stato commesso a Urbopoli. Sulla base degli indizi raccolti, il commissario è convinto che il sig. Rossi sia colpevole di tale furto con probabilità del 40% (cioè, su 100 casi simili, il sig. Rossi sarebbe colpevole in 40 casi). Non ci sono altri sospettati. Il commissario viene ora a conoscenza di un nuovo indizio: il colpevole ha i capelli biondi, come il sig. Rossi. Da una statistica, emerge che il 20% degli abitanti di Urbopoli ha i capelli biondi. Come cambia la valutazione del commissario circa la probabilità di colpevolezza del sig. Rossi?

Esercizio 3.27 Un'urna contiene 5 palline, di cui 4 rosse e 1 blu. Si consideri il seguente esperimento. Si estrae una prima pallina dall'urna. Se questa prima pallina estratta è blu, l'esperimento termina. Se invece la prima pallina estratta è rossa, la pallina non viene reinserita nell'urna, ma se ne aggiunge un'altra che è rossa con probabilità $3/4$ e blu con probabilità $1/4$, e successivamente si estrae nuovamente una pallina dall'urna. Dopo la seconda estrazione l'esperimento termina in ogni caso.

- Qual è la probabilità che durante l'esperimento venga estratta una pallina blu? Ripetendo 10 volte l'esperimento dall'inizio, determinare la probabilità di estrarre una pallina blu per 3 volte.
- Sapendo che la pallina blu è stata estratta alla seconda estrazione, qual è la probabilità che la pallina inserita nell'urna dopo la prima estrazione fosse blu?

Esercizio 3.28 Un sacchetto contiene 10 monete, di cui 8 "oneste" (ossia una faccia è Testa e una faccia Croce, e nel lancio della moneta le due facce sono equiprobabili) e 2 con entrambe le facce uguali a Testa. L'esperimento consiste nell'estrarre una moneta dal sacchetto e lanciarla: se esce Testa l'esperimento è concluso, altrimenti si reinserisce la moneta e si ripete l'operazione.

1. Determinare la probabilità che durante l'esperimento vengano eseguiti k lanci, con $k \in \mathbb{N}$.
2. Sapendo che l'esito Testa non si è avuto nei primi 10 lanci, determinare la probabilità di successo in meno di 13 lanci.
3. Supponiamo che al primo lancio sia uscita Testa, qual è la probabilità che sia stata estratta dal sacchetto una moneta onesta?

Esercizio 3.29 La distribuzione in Italia dei gruppi sanguigni è la seguente:

	0	A	B	AB
+	40	36	7.5	2.5
-	7	6	1.5	0.5

(ad esempio, il 40% degli individui è di tipo 0+).

- Qual è la probabilità che l'individuo estratto abbia Rh+?
- Qual è la probabilità che abbia Rh+, sapendo che ha gruppo AB?
- Gli eventi “Rh+” e “gruppo sanguigno AB” sono indipendenti?

■

Esercizio 3.30 Vengono lanciati $n \geq 1$ dadi a sei facce: qual è la probabilità che la somma dei punti di tutti i dadi sia divisibile per 7?

Suggerimento: i casi $n = 1, 2$ sono facili, per il caso n -esimo si condiziona sul risultato dei primi $n - 1$ dadi, ottenendo una formula ricorsiva.

■

Esercizio 3.31 — Difficile. Si elegge il rappresentante degli studenti al corso di laurea di Scienze Scientifiche. Il candidato Avogadro vince con a voti, mentre il candidato Bohr viene sconfitto ottenendo $b < a$ voti. Si mostri che la probabilità che durante lo spoglio dei voti Avogadro sia stato in testa a Bohr in ogni momento è data da $\frac{a-b}{a+b}$.

Suggerimento: Condizionare sull'ultimo voto scrutinato e procedere per ricorsione.

■

4. Variabili Aleatorie

Consideriamo il lancio di n monete. In un esperimento di questo tipo ha spesso poco interesse chiedersi l'esito di un particolare lancio: l'esito del lancio della quinta moneta non ha per noi nulla di speciale, mentre quasi tutte le domande rilevanti coinvolgono il numero delle teste (o delle croci). In termini più astratti, l'esito dell'esperimento è soggetto a una grande variabilità –gli esiti dei singoli lanci delle monete– ma a noi interessa una quantità che da questa variabilità dipende senza però contenerne tutta l'informazione.

Siamo condotti quindi a studiare una *funzione* dello spazio di probabilità associato all'esperimento: a ogni esito degli n lanci la funzione assocerà il numero di teste (o croci) uscite, ed esiti diversi possono avere chiaramente lo stesso numero di teste. Funzioni come questa, definite su uno spazio di probabilità, sono dette *Variabili Aleatorie* (*Random Variables*).

Questa idea diventa fondamentale per descrivere *osservazioni diverse fatte su uno stesso spazio di probabilità* $(\Omega, \mathcal{F}, \mathbb{P})$, corrispondenti a diverse funzioni di Ω , ovvero variabili aleatorie diverse. In effetti, è proprio nel caso in cui vogliamo confrontare diverse misurazioni di uno stesso esperimento che il semplice formalismo degli spazi di probabilità diventa scomodo e limitante, mentre introducendo la nozione di variabile aleatoria una gran quantità di domande e operazioni rientrano in modo naturale nel formalismo matematico.

4.1 La legge di una Variabile Aleatoria

Definizione 4.1.1 Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Una *Variabile Aleatoria* è una funzione $X : \Omega \rightarrow \mathbb{R}$ tale che per ogni $A \subseteq \mathbb{R}$ (misurabile) vale $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$.

La definizione viene data in questo modo per tenere conto del fatto che non sempre la σ -algebra \mathcal{F} degli eventi coincide con tutti i sottoinsiemi di Ω , e dunque dobbiamo assicurarci di poter calcolare la probabilità di eventi associati a una variabile aleatoria, per i quali introduciamo la seguente notazione: se $A \subseteq \mathbb{R}$ scriveremo equivalentemente

$$\{X \in A\} = X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}, \quad \mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$$

(in cui l'ultima espressione fornisce la definizione delle precedenti).

■ **Esempio 4.1** Come visto nella Proposizione 3.4.1, n lanci indipendenti di una moneta non truccata si descrivono con lo spazio di probabilità

$$\Omega = \{0, 1\}^n = \{(a_1, \dots, a_n) : a_i = 0, 1 \quad \forall i = 1, \dots, n\}, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}(a) = (1/2)^n.$$

Indicando con 0 l'esito "croce" e con 1 "testa", le variabili aleatorie

$$X = X(a) = a_1 + \dots + a_n, \quad Y = Y(a) = n - X(a) = n - a_1 - \dots - a_n$$

rappresentano rispettivamente il numero di teste e di croci. Se $A = \{0, 1, 2\}$, l'evento $\{X \in A\}$ corrisponde a "testa compare al massimo 2 volte (negli n lanci)". Si possono naturalmente definire variabili aleatorie più complicate: ad esempio la variabile

$$Z = a_1 a_2 + a_2 a_3 + \dots + a_{n-1} a_n$$

conta quante volte lanci successivi hanno dato lo stesso esito (rimandiamo agli esercizi per altri esempi di questo tipo). ■

NB

In Probabilità, come abbiamo fatto sopra, una volta specificato qual è lo spazio Ω sottostante si è soliti omettere la dipendenza di una variabile aleatoria $X : \Omega \rightarrow \mathbb{R}$ dal suo argomento: questo non causa confusione grazie a una serie di altri accorgimenti notazionali, invitiamo quindi il lettore ad osservare attentamente i simboli che usiamo e ad attenersi alle convenzioni stabilite.

Esercizio 4.1 Gli eventi $\{X \in A\} \subseteq \Omega$ al variare di $A \subseteq \mathbb{R}$ formano una σ -algebra su Ω (contenuta in quella di partenza \mathcal{F}). ■

Proposizione 4.1.1 La funzione di insiemi \mathbb{P}_X è una probabilità su \mathbb{R} (con la σ -algebra dei misurabili), detta *legge (di probabilità) di X* .

Dimostrazione. La parte non banale è il controllo della σ -additività: se $(A_n)_{n \geq 1}$ è una successione di sottinsiemi di \mathbb{R} a due a due disgiunti, anche le immagini inverse sono disgiunte e si ha

$$\mathbb{P}_X \left(\bigcup_{n=1}^{+\infty} A_n \right) = \mathbb{P} \left(X^{-1} \left(\bigcup_{n=1}^{+\infty} A_n \right) \right) = \sum_{n=1}^{+\infty} \mathbb{P} (X^{-1} (A_n)) = \sum_{n=1}^{+\infty} \mathbb{P}_X (A_n).$$

Lasciamo al lettore i restanti dettagli. ■

Quando due variabili aleatorie hanno la stessa legge di probabilità sono dette *equidistribuite* (o anche *isonome*). Non a caso indichiamo (e indicheremo sempre) le variabili aleatorie con lettere maiuscole X, Y, T, \dots e le variabili reali (come quelle di integrazione) con lettere minuscole: questo è importante per evitare confusioni!

NB

Nelle applicazioni, spesso non vengono definiti lo spazio di probabilità Ω e la variabile $X : \Omega \rightarrow \mathbb{R}$, ma solo la legge di probabilità \mathbb{P}_X . Si può mostrare che, assegnata una probabilità \mathbb{Q} sui sottinsiemi di \mathbb{R} , è possibile costruire un insieme Ω , una probabilità \mathbb{P} sui sottinsiemi di Ω ed una variabile $X : \Omega \rightarrow \mathbb{R}$ tale che $\mathbb{P}_X = \mathbb{Q}$. Siamo dunque sempre ricondotti alla situazione definita sopra.

In particolare, quando si sta considerando una singola variabile aleatoria, \mathbb{P}_X riassume tutte l'informazione in gioco: è equivalente avere una variabile aleatoria (cioè una funzione $X : \Omega \rightarrow \mathbb{R}$) oppure una distribuzione di probabilità sulla retta reale \mathbb{R} . In effetti la costruzione stessa delle variabili aleatorie come funzioni di uno spazio Ω sottostante diventa rilevante solo quando ce n'è più di una e si è interessati alle relazioni tra di loro.

Ai due esempi di probabilità su \mathbb{R} introdotti nel capitolo precedente corrispondono due tipi di variabili aleatorie: anche qui è importante chiarire subito che non tutte le v.a. appartengono a uno di questi due tipi, ma sono i casi che appaiono più frequentemente.

Definizione 4.1.2 Una variabile aleatoria è detta discreta se la sua immagine $X(\Omega) \subset \mathbb{R}$ è un sottinsieme al più numerabile di \mathbb{R} , o equivalentemente se la sua legge di probabilità è discreta.

Conoscere la legge di probabilità di una v.a. discreta equivale a conoscere quali sono i valori (x_1, x_2, \dots) che prende la variabile, ed i numeri $p(x_i) = \mathbb{P}_X(x_i) = \mathbb{P}(X = x_i)$ (cioè la sua funzione di massa o densità discreta). Preso $A \subseteq \mathbb{R}$ vale

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \sum_{x_i \in A} p_X(x_i).$$

Il seguente esempio è utile per collegare (al momento solo intuitivamente) i modelli probabilistici alla statistica descrittiva e poi a quella inferenziale.

■ **Esempio 4.2** Scegliamo come Ω l'insieme degli n individui della popolazione di un Paese, $\mathcal{F} = \mathcal{P}(\Omega)$, e come \mathbb{P} la probabilità uniforme. Definiamo la variabile aleatoria $X = X(\omega)$ come “il numero di figli di ω ”, $\omega \in \Omega$ un individuo. La corrispondente densità discreta p_X è data da

$$p_X(x_i) = \frac{\#\{\omega \mid X(\omega) = x_i\}}{\#\Omega},$$

cioè la frequenza relativa di $X = x_i$ su tutta la popolazione (non su un particolare campione statistico); ad esempio, se X = numero di figli, $p_X(2)$ è la percentuale, su tutta la popolazione italiana, degli individui che hanno esattamente 2 figli. ■

Definizione 4.1.3 Una variabile aleatoria è detta con densità (o *assolutamente continua*) se la sua legge di probabilità è definita da una densità f , cioè se esiste una densità di probabilità f tale che valga

$$\mathbb{P}_X(A) = \mathbb{P}\{X \in A\} = \int_A f(x)dx.$$

In particolare, se $A = [a, b]$ è un segmento, per una variabile X con densità f vale

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

4.2 Funzione di Ripartizione e Quantili

Data una singola variabile aleatoria $X : \Omega \rightarrow \mathbb{R}$, al fine di studiare la sua legge \mathbb{P}_X (che di per se è un oggetto matematico abbastanza complicato, quella che si chiama una *misura su \mathbb{R}*), si introduce un oggetto più familiare che comunque continua a codificare tutte le proprietà di \mathbb{P}_X , ovvero la *funzione di ripartizione* (*Cumulative Distribution Function*, per brevità *c.d.f.*).

Definizione 4.2.1 Si chiama Funzione di ripartizione della v.a. X la funzione

$$F_X : \mathbb{R} \rightarrow [0, 1], \quad F_X(x) = \mathbb{P}\{X \leq x\}.$$

È evidente dalla definizione che in realtà F_X dipende solo dalla legge di probabilità della v.a. X ; quando non c'è pericolo di confusione scriviamo più semplicemente F . Le seguenti proprietà di base delle funzioni di ripartizione costituiscono un esercizio istruttivo.

Esercizio 4.2 Sia $F = F_X$ la c.d.f di una variabile aleatoria X . Si mostri che

- F è non decrescente, ovvero se $x < y$ allora $F(x) \leq F(y)$;
- valgono

$$\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1;$$

- F è continua a destra, ossia per ogni $x \in \mathbb{R}$ vale $F(x_n) \rightarrow F(x)$ per ogni successione $x_n \rightarrow x$ con $x_n \geq x$.

In effetti, queste proprietà caratterizzano le funzioni di ripartizione, come chiarisce il seguente risultato che riportiamo senza dimostrazione.

Proposizione 4.2.1 Data una funzione $F : \mathbb{R} \rightarrow [0, 1]$ con le proprietà sopra elencate, esiste una ed una sola probabilità \mathbb{P} sui sottinsiemi misurabili di \mathbb{R} tale che $F(t) = \mathbb{P}((-\infty, t])$, ovvero che F è la c.d.f. $F = F_X$ di una variabile X la cui legge di probabilità sia $\mathbb{P}_X = \mathbb{P}$.

Di conseguenza due v.a. che hanno la stessa funzione di ripartizione sono equidistribuite e dalla funzione di ripartizione si può ricostruire la legge di probabilità: questo è in generale non banale (come la dimostrazione della Proposizione), ma in esempi specifici si riduce spesso ad un esercizio.

Dalla funzione di ripartizione F di una v.a. X , si ricava la probabilità che X cada in un dato intervallo: per $a < b$ vale

$$\mathbb{P}\{a < X \leq b\} = F(b) - F(a),$$

come si verifica facilmente a partire da $\{a < X \leq b\} = \{X \leq b\} \setminus \{X \leq a\}$. Tale formula si estende ai casi $a = -\infty$ (cioè per $\mathbb{P}\{X \leq b\}$) e $b = +\infty$ (cioè per $\mathbb{P}\{X > a\}$), ponendo $F(-\infty) = 0$, $F(+\infty) = 1$ (coerentemente con i limiti visti nel precedente esercizio). Questa formula è particolarmente utile in contesti, come sarà per le v.a. Gaussiane, in cui non si sa calcolare la probabilità $\mathbb{P}\{a < X \leq b\}$ direttamente, ma sono noti almeno in modo approssimato i valori assunti dalla funzione di ripartizione.

La funzione di ripartizione è ben definita per qualsiasi variabile aleatoria, e possiamo usarla per distinguere vari tipi di variabili aleatorie. La c.d.f. di una variabile aleatoria X discreta che assume valori x_1, x_2, \dots è una funzione costante a tratti (“a gradini”),

$$F_X(t) = \sum_{x_i \leq t} p(x_i).$$

La funzione esibisce un salto in corrispondenza di ogni punto x tale che $\mathbb{P}\{X = x\} > 0$ e la probabilità di quel punto è proprio l’ampiezza del salto: si ha pertanto

$$\mathbb{P}\{X = x\} = F(x) - F_-(x),$$

dove con $F_-(x) = \lim_{y \nearrow x} F(y)$ si intende il limite sinistro di F nel punto x .

Quando la variabile ha densità f —e solamente in quel caso!— la sua funzione di ripartizione diventa $F(x) = \int_{-\infty}^x f(t)dt$ ed è ovviamente continua. Nel caso in cui la densità f sia continua a tratti (come sarà in tutti gli esempi che vedremo), o equivalentemente la c.d.f. F sia continua su \mathbb{R} e C^1 a tratti, partendo dalla formula $F(x) = \int_{-\infty}^x f(t)dt$, si ottiene (in tutti i punti in cui f è continua)

$$f(x) = \frac{dF(x)}{dx};$$

quindi f può essere ottenuta derivando F .



Si possono definire “variabili continue” le variabili aleatorie la cui c.d.f. è continua. Come appena detto, le variabili con densità –che abbiamo chiamato anche assolutamente continue– hanno c.d.f. continua, ma **non tutte le variabili con c.d.f. continua sono variabili con densità**. I controesempi (cosiddette “distribuzioni singolari”) sono purtroppo fuori dalla portata di queste note, per cui ci limitiamo a raccomandare di non confondere le variabili con densità, o assolutamente continue, con quelle continue nel senso suddetto.

Inoltre, non tutte le densità f sono continue a tratti, e in tal caso la formula $f(x) = \frac{dF(x)}{dx}$ può non essere valida; tuttavia negli esempi che tratteremo vedremo sempre densità continue a tratti (se non proprio continue).

Il fatto che la c.d.f. descriva completamente una singola variabile aleatoria X (vedremo che nel caso di due o più variabili aleatorie la situazione non è così semplice) suggerisce che tra tutti gli eventi associati ad una variabile aleatoria i più interessanti siano quelli della forma $\{X \leq r\}$.

È quindi naturale chiedersi, dato $0 < \beta < 1$, quale sia il numero r_β tale che $\mathbb{P}(X \leq r_\beta) = F(r_\beta) = \beta$, ovvero qual è la semiretta di valori di X che ha una probabilità assegnata. Questa **non** può essere una definizione per r_β , per due motivi: non sempre esiste un numero r tale che $F(r) = \beta$ e non è detto che tale numero sia univocamente determinato. Il primo problema è risolto dando la seguente definizione modificata, che però rimane non univoca.

Definizione 4.2.2 Assegnata una v.a. X ed un numero β con $0 < \beta < 1$, si chiama β -quantile un numero r tale che si abbia $\mathbb{P}(X \leq r) \geq \beta$ e $\mathbb{P}(X \geq r) \geq 1 - \beta$.

Se F non ha un intervallo nel quale è costante, e dunque è strettamente crescente, il quantile r_β è unico per ogni $\beta \in (0, 1)$ e $r_\beta = F^{-1}(\beta)$. Naturalmente, questo non può accadere per le variabili aleatorie discrete.

Il quantile (o meglio, la scelta del β -quantile nel caso ci siano più numeri r_β che soddisfano la soluzione) dipende solo dalla c.d.f. della variabile aleatoria o, equivalentemente, dalla sua legge di probabilità. Esiste un modo per rendere univoca la definizione di β -quantile per ogni variabile aleatoria: la discussione viene lasciata come esercizio guidato alla fine del capitolo.

Vediamo invece come si determina un quantile in pratica nel caso delle v.a. discrete.

■ **Esempio 4.3** Prendiamo una v.a. X discreta, siano x_1, x_2, \dots i valori che assume e supponiamo di poterli ordinare $x_1 < x_2 < \dots$; questo non sempre è possibile, perchè può esistere una successione di punti x_i che ha limite $-\infty$, ma nella maggior parte degli esempi che consideriamo esiste un x_i più piccolo di tutti. Se non esiste alcun x_j tale che $F(x_j) = \beta$ allora r_β è il più piccolo x_j tale che $F(x_j \geq \beta)$, ma se esiste un x_j tale che $F(x_j) = \beta$, ogni numero r con $x_j \leq r \leq x_{j+1}$ soddisfa la definizione precedente. Di solito per convenzione si prende $r_\beta = \frac{x_j + x_{j+1}}{2}$. ■

Riprendiamo qui l'Esempio 4.2 per vedere il collegamento tra c.d.f. ed e.c.d.f.:

■ **Esempio 4.4** Nel contesto dell'Esempio 4.2, la funzione di ripartizione (c.d.f.) coincide con la funzione di ripartizione empirica (e.c.d.f.) sostituendo al campione *l'intera popolazione*. Analogamente vale per i β -quantili. ■

4.3 Variabili Aleatorie Notevoli

Cataloghiamo di seguito le distribuzioni di probabilità su \mathbb{R} che è imprescindibile conoscere in ogni applicazione. Sono tutte variabili discrete o definite tramite densità.

Variabili Binomiali

Torniamo all'esempio con cui abbiamo aperto il capitolo: consideriamo n prove ripetute di un esperimento che ha solo due esiti (o di cui ci interessano solo due esiti), chiamiamo successo uno di questi e sia p con $0 < p < 1$ la probabilità del successo: ricordiamo che un esempio di spazio di probabilità che modella questa situazione è stato dato nella Proposizione 3.4.1.

Sia X la variabile che conta il numero dei successi. I valori possibili sono $0, 1, \dots, n$ e vale

$$\mathbb{P}(X = h) = \binom{n}{h} p^h (1-p)^{n-h}, \quad 0 \leq h \leq n.$$

Infatti, per l'indipendenza delle prove, ogni sequenza di risultati con h successi ed $(n-h)$ insuccessi ha probabilità $p^h \cdot (1-p)^{n-h}$, e il numero di tali sequenze, cioè il numero di modi di disporre gli h successi, è $\binom{n}{h}$. Si noti che la probabilità totale è 1 come conseguenza della formula del binomio di Newton. In generale, una variabile aleatoria $X : \Omega \rightarrow \{1, \dots, n\}$ per cui $\mathbb{P}\{X = h\}$ è data dalla suddetta formula per un qualche $p \in (0, 1)$ è detta *variabile Binomiale* di parametri n, p , o anche –per brevità– variabile di tipo $B(n, p)$. Quando $n = 1$ la variabile è detta *di Bernoulli* di parametro p (talvolta indicata con $B(p)$).

Sono variabili di Bernoulli il numero di teste in n lanci di una moneta, il numero di “6” in n lanci di un dado, il numero di biglie rosse in n estrazioni con rimpiazzo da un'urna contenente biglie di diversi colori, e così via.

Il seguente esempio mostra ancora una volta come l'informazione codificata da una variabile aleatoria $X : \Omega \rightarrow \mathbb{R}$ non sia contenuta nel particolare spazio di probabilità Ω sottostante, quanto nella legge \mathbb{P}_X .

■ **Esempio 4.5** Dati $n \in \mathbb{N}_0$ e $p \in [0, 1]$, si considerino due spazi di probabilità:

$$\Omega = \{0, 1\}^n, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}(\{a\}) = p^{\sum_{i=1}^n a_i} (1-p)^{n-\sum_{i=1}^n a_i}, \quad \forall a = (a_1, \dots, a_n) \in \Omega$$

(definito nella Proposizione 3.4.1) e

$$\Omega' = \{0, 1, \dots, n\}, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Definiamo le variabili aleatorie

$$X : \Omega \rightarrow \{0, 1, \dots, n\}, X(a) = \sum_{i=1}^n a_i, \quad Y : \Omega' \rightarrow \{0, 1, \dots, n\}, Y(k) = k.$$

Sappiamo già che la legge di X è binomiale $B(n, p)$, ma dalla definizione si vede immediatamente che anche Y ha legge binomiale $B(n, p)$. In altre parole, $\mathbb{P}_X = \mathbb{P}_Y$ (come uguaglianza tra probabilità sulla retta \mathbb{R}) anche se le due variabili sono costruite su spazi di probabilità distinti e alquanto diversi tra di loro. ■



Più in generale, data una successione $x_1, x_2, \dots \in \mathbb{R}$ e una successione di numeri non-negativi $p_1, p_2, \dots \in [0, \infty)$ tali che $\sum_{i=1}^{\infty} p_i = 1$, possiamo definire una variabile aleatoria discreta tramite

$$\Omega = \mathbb{N}, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}(\{k\}) = p_k, \quad X(k) = x_k.$$

Questo è un modo di costruire una variabile aleatoria X con una data legge di probabilità discreta $\mathbb{P}_X(k) = \mathbb{P}(X = x_k) = p_k$, ma come appena visto di gran lunga non l'unico.

Esercizio 4.3 Su 5 lanci di un dado equilibrato, qual è la probabilità che il “6” appaia almeno due volte? ■

Variabili Geometriche

Consideriamo nuovamente un esperimento che ha solo due esiti, chiamiamo successo uno di questi e sia p con $0 < p < 1$ la probabilità del successo, e ripetiamo l'esperimento fino a quando non avviene un successo. La variabile X , detta geometrica di parametro p , conta l'istante del primo successo (cioè il numero h tale che alla prova h -sima si verifichi il primo successo): i valori possibili di X sono i naturali positivi $1, 2, 3, \dots$ e si ha

$$\mathbb{P}(X = h) = (1 - p)^{h-1} p, \quad h \in \mathbb{N}_0.$$

Infatti, detto A_i l'evento "successo alla prova i -sima", per l'indipendenza delle prove abbiamo

$$\mathbb{P}(X = h) = \mathbb{P}(A_1^c \cap A_2^c \cap \dots \cap A_{h-1}^c \cap A_h) = \mathbb{P}(A_1^c) \cdot \mathbb{P}(A_2^c) \cdots \mathbb{P}(A_{h-1}^c) \cdot \mathbb{P}(A_h) = (1 - p)^{h-1} p.$$

Le variabili geometriche sono caratterizzate da una interessante proprietà detta *assenza di memoria*, la cui discussione viene lasciata all'Esercizio 4.4 qui di seguito.

Esercizio 4.4 Data una variabile geometrica X di parametro p , mostrare che per ogni coppia di interi positivi n, h si ha

$$\mathbb{P}\{X = n + h \mid X > n\} = \mathbb{P}\{X = h\}.$$

Viceversa, assumiamo ora sia data una variabile discreta $X : \Omega \rightarrow \mathbb{N}_0$ che soddisfa questa ultima equazione (si dice appunto che X è una variabile senza memoria, perchè sapere che siamo oltre l' n -esimo passo è come ripartire da quest'ultimo). Si mostri che X è geometrica per un certo parametro p .

Suggerimento: si usi l'equazione dell'assenza di memoria con $n = 1$ per ottenere una formula ricorsiva per $p(h) = \mathbb{P}\{X = h\}$. ■

Esercizio 4.5 Lanciamo ripetutamente un dado equilibrato. Qual è la probabilità che il "6" appaia per la prima volta al quarto lancio? Se il "6" non è apparso nei primi due lanci, qual è la probabilità che appaia per la prima volta al sesto lancio? ■

Variabili di Poisson

Dato $\lambda > 0$, si dice che una variabile $X : \Omega \rightarrow \mathbb{N}$ a valori naturali è una variabile di Poisson di parametro $\lambda > 0$ se

$$\mathbb{P}(X = h) = e^{-\lambda} \frac{\lambda^h}{h!}, \quad h \in \mathbb{N}.$$

Il fatto che $\sum_{h=0}^{+\infty} \mathbb{P}(X = h) = 1$ (ossia la probabilità totale è 1) è una conseguenza dello sviluppo esponenziale $e^\lambda = \sum_{h=0}^{+\infty} \lambda^h / h!$.

Si può mostrare (si veda il prossimo esercizio) che la distribuzione di Poisson di parametro λ è una buona approssimazione della distribuzione binomiale di parametri n e p , quando n è grande, p è piccolo e np vale circa λ . In altre parole, la distribuzione di Poisson conta il numero di successi in prove ripetute, quando il numero n di prove è elevato e la probabilità p di successo è bassa. Per questo, la distribuzione di Poisson viene anche detta distribuzione degli eventi rari, e si usa per modellizzare, ad esempio:

- il numero di particelle α emesse da una sorgente radioattiva in un dato intervallo di tempo; in questo caso, la singola prova è l'emissione della particella α da un singolo nucleo della sorgente;

- il numero di clienti a un certo sportello (o a un certo server) in un dato intervallo di tempo; in questo caso, la singola prova è il recarsi o meno allo sportello da parte di una singola persona della popolazione;
- (in prima approssimazione) il numero di grandi terremoti in una data area in un certo intervallo di tempo; in questo caso, la singola prova è il verificarsi di un grande terremoto in un intervallo di tempo molto piccolo.

Esercizio 4.6 Sia $p^{(n)}(k) = \binom{n}{p_n} p_n^k (1 - p_n)^{n-k}$ la densità discreta binomiale $B(n, p_n)$, con $p_n = \lambda/n$. Mostrare che per ogni $k \in \mathbb{N}$,

$$\lim_{n \rightarrow \infty} p^{(n)}(k) = p_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

ovvero alla densità discreta $P(\lambda)$. ■

Esercizio 4.7 Il numero di clienti delle poste di Urbopoli, in una data ora, è descritto da una v.a. di Poisson di parametro 2.3. Qual è la probabilità che, in quell'ora, vi siano al massimo 2 clienti? ■

Esercizio 4.8 Costruire esplicitamente uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e una variabile aleatoria $X : \Omega \rightarrow \mathbb{R}$ avente come legge una variabile geometrica oppure una variabile di Poisson. *Suggerimento:* in analogia con lo spazio di probabilità costruito nella Proposizione 3.4.1, considerare $\Omega = \{0, 1\}^{\mathbb{N}}$ (dobbiamo avere a disposizione infinite prove). ■

Variabili Uniformi su Intervalli

Dati due numeri reali $a < b$, la densità uniforme sull'intervallo $[a, b]$ è costante sull'intervallo e nulla fuori da esso, ovvero

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{altrove} \end{cases}.$$

L'esempio precedentemente visto del numero preso a caso tra 0 e 1 corrisponde quindi alla densità uniforme su $[0, 1]$.

Variabili Esponenziali

La densità esponenziale di parametro λ ($\lambda > 0$) è così definita:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & x \leq 0. \end{cases}$$

Si controlla facilmente che è una densità di probabilità:

$$\int_0^{+\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x} \Big|_0^{+\infty} = 1.$$

Poiché la densità è diversa da 0 solo per x positivo, la variabile X prende solo valori positivi (nel senso che $\mathbb{P}\{X \leq 0\} = 0$).

La v.a. esponenziale descrive ad esempio il tempo di attesa tra due eventi aleatori, come ad esempio due chiamate ad un call center.

Esercizio 4.9 Mostrare che una variabile esponenziale X di parametro λ è senza memoria, ovvero vale l'eguaglianza (per s, t positivi)

$$\mathbb{P}\{X \leq s+t \mid X > s\} = \mathbb{P}\{X \leq t\}.$$

Viceversa, data una variabile X con densità che assume solo valori positivi e che soddisfa l'equazione suddetta (è senza memoria), si mostri che è una variabile esponenziale per un certo parametro λ .

Suggerimento: usare la formula dell'assenza di memoria per ottenere l'equazione

$$S(t+s) = S(t)S(s), \quad S(t) = \mathbb{P}\{X > s\},$$

e usare il fatto (che diamo per noto) che le uniche soluzioni continue di questa equazione funzionale sono della forma $S(t) = e^{-\lambda t}$ con $\lambda \geq 0$. ■

Esercizio 4.10 Il tempo di vita (in giorni) di un certo macchinario (cioè il tempo di funzionamento prima del primo guasto) è descritto da una v.a. esponenziale di parametro $1/8$. Qual è la probabilità che il primo guasto si verifichi dopo 6 giorni? ■

Trasformazioni di Variabili con Densità

L'ultima classe di variabili aleatorie che presenteremo è quella Gaussiana, di importanza fondamentale in inferenza statistica. Premettiamo però una breve discussione sulle trasformazioni di variabili aleatorie con densità.

Supponiamo che $X : \Omega \rightarrow \mathbb{R}$ abbia densità f , e data una funzione $h : \mathbb{R} \rightarrow \mathbb{R}$ ci domandiamo se la variabile aleatoria

$$Y : \Omega \rightarrow \mathbb{R}, \quad Y = h \circ X$$

abbia densità e come calcolarla. Anzitutto non è detto che Y abbia densità per ogni funzione h , e anche quando Y ha densità non esiste una regola generale per calcolarla. Possiamo solo osservare che se in qualche modo si riesce a calcolare la c.d.f. di Y ,

$$F_Y(y) = \mathbb{P}\{Y \leq y\} = \mathbb{P}\{h(X) \leq y\},$$

ed essa risulta essere continua su tutto \mathbb{R} e C^1 a tratti, derivando si ottiene la densità di Y .

Vediamo come questo risulta possibile assumendo che h abbia delle buone proprietà. Nel risultato seguente, col termine generico di intervallo aperto intendiamo sia un intervallo limitato del tipo $]a, b[$ sia una semiretta $] -\infty, b[$ oppure $]a, +\infty[$, sia tutta la retta $] -\infty, +\infty[$.

Proposizione 4.3.1 — Cambio di Variabile. Supponiamo che X abbia densità f_X che sia supportata su un intervallo aperto A (cioè tale che f sia nulla su A^c), sia $h : A \rightarrow B$, con B un altro intervallo aperto, e assumiamo che h sia biunivoca, C^1 e con inversa C^1 . La variabile $Y = h \circ X$ ha densità f_Y data dalla formula

$$f_Y(y) = \begin{cases} f_X(h^{-1}(y)) \cdot \left| \frac{dh^{-1}(y)}{dy} \right| & y \in B, \\ 0 & y \notin B. \end{cases}$$

Dimostrazione. Diamo una dimostrazione nel caso in cui f sia continua a tratti (cioè F_X sia continua su \mathbb{R} e C^1 a tratti). Osserviamo che sotto le ipotesi assunte su h , essa può essere solo strettamente crescente o strettamente decrescente. Se h è crescente, allora

$$F_Y(y) = \mathbb{P}\{Y \leq y\} = \mathbb{P}\{h(X) \leq y\} = \mathbb{P}\{X \leq h^{-1}(y)\} = F_X(h^{-1}(y))$$

e derivando si ottiene la formula $f_Y(y) = f_X(h^{-1}(y)) \cdot \frac{dh^{-1}(y)}{dy}$. Se invece h è decrescente,

$$\mathbb{P}\{h(X) \leq y\} = \mathbb{P}(X \geq h^{-1}(y)) = 1 - F_X(h^{-1}(y)),$$

per cui ora derivando appare un segno meno, che però è compensato dal fatto che la derivata di h^{-1} è negativa: per tenere assieme i due casi basta prendere il valore assoluto come nell'enunciato. ■

La formula proposta sopra è molto comoda, ma bisogna essere sicuri che possa essere applicata: è istruttivo al riguardo svolgere l'esercizio che segue curando di controllare se può essere utilizzata la formula proposta sopra.

Esercizio 4.11 Sia X una variabile con densità uniforme su $[-1, 2]$, si determini se $Z = X^2$ ha densità e se sì la si calcoli.

Sia poi Y una variabile con densità esponenziale di parametro 2, si determini se $W = Y^2$ ha densità e se sì la si calcoli. ■

Variabili Gaussiane

La funzione $f(x) = e^{-x^2/2}$ è regolarissima (è infinitamente derivabile), tende a 0 molto velocemente per $|x| \rightarrow \infty$ e quindi è integrabile, ma non è possibile scrivere la sua primitiva in termini di funzioni elementari. Non possiamo chiarire qui il significato preciso di questa affermazione: per il lettore può significare semplicemente che l'unico modo di rappresentare il valore di $\int_0^t e^{-x^2/2} dx$ per un generico t è ricorrere ad approssimazioni numeriche. Tuttavia, per alcuni particolari valori di t tale integrale assume valori semplici: ad esempio vale

$$\int_{-\infty}^{+\infty} e^{-x^2/2} dx = \sqrt{2\pi}$$

(due metodi per mostrarlo sono proposti come esercizi alla fine del capitolo). Di conseguenza, dividendo la funzione $e^{-x^2/2}$ per $\sqrt{2\pi}$ si ottiene una densità di probabilità: si chiama *densità Gaussiana (o Normale) standard*, indicata $N(0, 1)$, la funzione

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

La sua funzione di ripartizione (di cui come detto non possiamo dare una rappresentazione alternativa) è

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

Queste funzioni sono talmente centrali in Probabilità e Statistica che i simboli φ, Φ e q_α sono riservati ad indicare rispettivamente la densità, la c.d.f. e lo α -quantile della variabile $N(0, 1)$.

Le seguenti proprietà di base della c.d.f. Gaussiana sono lasciate come semplice esercizio.

Esercizio 4.12 Si mostri che la densità φ è una funzione pari, ovvero $\varphi(x) = \varphi(-x)$, e di conseguenza dati $x \in \mathbb{R}$ e $0 < \alpha < 1$, si ha

$$\Phi(-x) = 1 - \Phi(x) \quad q_{1-\alpha} = -q_\alpha.$$

(Questa uguaglianza non è specifica della Gaussiana standard, ma conseguenza del solo fatto che la densità è una funzione pari). Di conseguenza, se X è una v.a. Gaussiana standard, allora

$$\mathbb{P}\{-t \leq X \leq t\} = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1$$

per ogni $t \in \mathbb{R}$. Inoltre $\Phi(0) = \mathbb{P}\{X \geq 0\} = \mathbb{P}\{X \leq 0\} = 1/2$. ■

Come abbiamo detto, detta X una v.a. Gaussiana standard, non è possibile (tranne in pochissimi casi particolari) rappresentare l'integrale

$$\mathbb{P}\{a < X < b\} = \int_a^b \varphi(x) dx = \Phi(b) - \Phi(a)$$

tramite funzioni più semplici, e quindi ricorriamo direttamente ad approssimazioni numeriche. Ne riportiamo alcune significative: se X è Gaussiana standard,

$$\mathbb{P}\{-1 \leq X \leq 1\} \approx 0.68, \quad \mathbb{P}\{-2 \leq X \leq 2\} \approx 0.94, \quad \mathbb{P}\{-3 \leq X \leq 3\} \approx 0.997.$$

NB Nelle vecchie tavole di approssimazioni per $\Phi(t)$ ne venivano riportati i valori per $0 < t < 4$. In effetti, per $t \geq 4$, $\Phi(t)$ è talmente vicina a 1 da poterla approssimare direttamente con 1. Per i valori negativi di t ci si riporta al caso positivo tramite $\Phi(t) = 1 - \Phi(-t)$.

Data X Gaussiana standard, siano $\sigma > 0$ e $m \in \mathbb{R}$ e consideriamo la v.a. $Y = \sigma X + m$. La c.d.f. di Y è data da

$$F_Y(t) = \mathbb{P}\{Y \leq t\} = \mathbb{P}\{\sigma X + m \leq t\} = \mathbb{P}\left\{X \leq \frac{t-m}{\sigma}\right\} = \Phi\left(\frac{t-m}{\sigma}\right),$$

e la sua densità (ottenuta derivando F_Y o applicando la formula di cambio di variabile) è

$$f_Y(t) = \frac{1}{\sigma} f_X\left(\frac{t-m}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-m)^2}{2\sigma^2}}.$$

Chiamiamo *densità Gaussiana (o normale)* $N(m, \sigma^2)$ quest'ultima funzione, ovvero diciamo che Y è Gaussiana $N(m, \sigma^2)$ se la sua densità ha questa forma.

Si possono ricondurre tutti i calcoli relativi alla c.d.f. di una variabile Gaussiana generica alla c.d.f. standard Φ : è sufficiente sostituire a Y Gaussiana $N(m, \sigma^2)$ la rappresentazione $\sigma X + m$, con X Gaussiana standard. Ad esempio, in particolare,

$$\mathbb{P}\{a < Y < b\} = \mathbb{P}\{(a-m)/\sigma < X < (b-m)/\sigma\}.$$

Le approssimazioni riportate sopra per una variabile X Gaussiana standard diventano, nel caso in cui Y è Gaussiana $N(m, \sigma^2)$,

$$\mathbb{P}\{m - \sigma \leq Y \leq m + \sigma\} \approx 0.68,$$

$$\mathbb{P}\{m - 2\sigma \leq Y \leq m + 2\sigma\} \approx 0.94,$$

$$\mathbb{P}\{m - 3\sigma \leq Y \leq m + 3\sigma\} \approx 0.997.$$

Queste forniscono una argomentazione precisa alla regola empirica sulla dispersione dei dati introdotta al primo capitolo quando l'istogramma dei dati è (abbastanza) normale.

Esercizio 4.13 Dimostrare la seguente *proprietà di riproducibilità* delle v.a. Gaussiane: Sia Y una v.a. Gaussiana $N(m, \sigma^2)$, sia $V = \alpha Y + \beta$, con $\alpha, \beta \in \mathbb{R}$, $\alpha \neq 0$. Allora V è una v.a. Gaussiana $N(\alpha m + \beta, \alpha^2 \sigma^2)$. ■

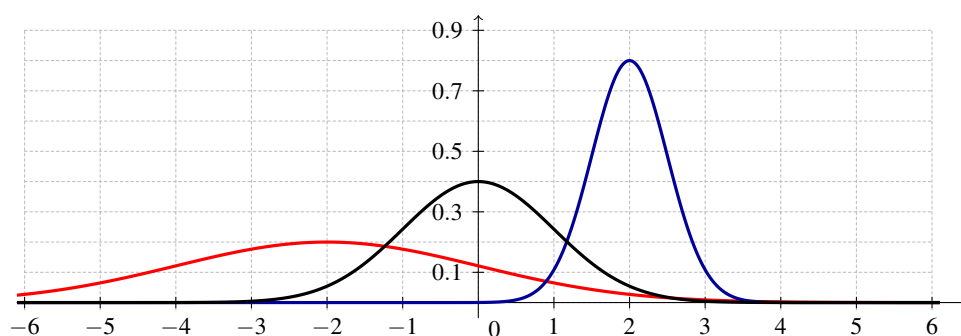


Figura 4.1: Grafici della densità Gaussiana con diversi parametri: rispettivamente in nero, rosso e blu le Gaussiane di parametri $m = 0, -2, 2$ e $\sigma^2 = 1, 2, 0.5$ (in particolare in nero la Gaussiana standard). Si noti che gli assi hanno scale diverse, in proporzione 1:1 le campane risultano molto più schiacciate.

Esercizio 4.14 Una fabbrica produce viti la cui lunghezza (in cm) ha distribuzione Gaussiana $N(1.7, 0.6^2)$. Qual è la probabilità che una vite abbia lunghezza inferiore a 1.6 cm? Qual è il valore x tale che il 95% delle viti abbia lunghezza almeno x (in cm)? ■

4.4 Valore Atteso, Varianza e Momenti

Abbiamo definito la media campionaria di un insieme di dati $x = (x_1, \dots, x_n)$ come la loro media aritmetica, $\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$, per dare una misura del “centro” della distribuzione dei dati, e di seguito la varianza campionaria per misurare quanto i dati sono “dispersi”. Le stesse idee sono alla base delle definizioni di valore atteso e varianza di una variabile aleatoria, oggetto matematico che si propone di descrivere la casualità dell’esito di un esperimento.

Iniziamo definendo il valore atteso, quantità che indica appunto “attorno a che valore è centrata” la variabile aleatoria. Per una variabile discreta che assume valori x_1, x_2, \dots è naturale considerare come valore atteso una media di questi valori, ma una media *pesata*, tenendo conto che a ogni punto è associata una probabilità, e punti con una probabilità più alta di essere estratti devono contare di più nella media. Per le variabili con densità, semplicemente rimpiazziamo la somma pesata dalla funzione di massa con un integrale pesato dalla densità. Naturalmente, nel caso di variabili discrete che assumono infiniti valori la media pesata diventa una serie, di cui bisogna controllare la convergenza, e nel caso di variabili con densità lo stesso vale per l’integrale con cui definiamo il valore atteso.

Definizione 4.4.1 — Valore Atteso. Sia X una variabile discreta con funzione di massa p : Si dice che X ha valore atteso se $\sum_i |x_i| p(x_i) < +\infty$, e in tal caso si chiama valore atteso il numero

$$\mathbb{E}[X] = \sum_i x_i p(x_i).$$

Sia poi X con densità f : essa ha valore atteso se $\int_{-\infty}^{+\infty} |x| f(x) dx < +\infty$ e in tal caso si chiama valore atteso il numero

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx.$$

In effetti, esiste una definizione più generale valida per tutte le variabili aleatorie che contiene quelle sopra scritte come casi particolari, ma necessita di elementi di teoria dell’integrazione generale. Il valore atteso (*expectation*) è chiamato anche speranza matematica o momento primo.

Notiamo che il valore atteso dipende solo dalla funzione di massa (nel caso discreto) o dalla densità (nel caso con densità); quindi il valore atteso di X dipende solo dalla legge P_X di X .

NB Se X prende solo valori positivi, ha sempre senso scrivere $\mathbb{E}[X]$ ammettendo che possa assumere il valore $+\infty$. Per le variabili discrete questo significa che i valori x_1, x_2, \dots sono tutti positivi e quindi ha comunque senso $\sum_{i=1}^{+\infty} x_i p(x_i)$; per le variabili con densità questo significa $f(x) = 0$ per $x < 0$ e quindi ha senso $\mathbb{E}[X] = \int_0^{+\infty} x f(x) dx \in [0, +\infty]$. Come conseguenza immediata, presa una generica v.a. X questa ha valore atteso se $\mathbb{E}[|X|] < +\infty$.

Esercizio 4.15 Calcolare il valore atteso dell'esito del lancio di un dado equo.

Supponiamo di ricevere due euro se esce 6, un euro se esce 4 o 5, zero altrimenti. Qual è il valore atteso del denaro che riceviamo? ■

Vediamo ora come calcolare il valore atteso di trasformazioni di una variabile aleatoria X , ossia di una nuova variabile Y della forma $g(X)$, $g: \mathbb{R} \rightarrow \mathbb{R}$.

Proposizione 4.4.1 Sia X discreta: la variabile $g(X)$ ammette valore atteso se $\sum_i |g(x_i)| p(x_i) < +\infty$, e in tal caso

$$\mathbb{E}[g(X)] = \sum_i g(x_i) p(x_i).$$

Sia X con densità f : la variabile $g(X)$ ha valore atteso se $\int_{-\infty}^{+\infty} |g(x)| f(x) dx < +\infty$, e in tal caso

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx.$$

Dimostrazione. Riportiamo solo la dimostrazione nel caso discreto: cominciamo a supporre che i numeri $g(x_i)$ siano tutti diversi, allora la variabile $g(X)$ ha come valori $g(x_1), g(x_2), \dots$ con probabilità $p(x_1), p(x_2), \dots$ e la formula è dunque evidente. In generale però alcuni dei numeri $g(x_i)$ possono essere eguali e questo costringe a modificare la dimostrazione.

Supponiamo dapprima che la funzione g sia a valori positivi: la variabile $Y = g(X)$ assume un insieme (finito o) numerabile di valori (y_1, y_2, \dots) . Consideriamo gli insiemi $A_i = \{j \mid g(x_j) = y_i\}$ e osserviamo che $p_Y(y_i) = \sum_{j \in A_i} \mathbb{P}(X = x_j) = \sum_{j \in A_i} p_X(x_j)$. Poiché quelle che seguono sono somme di serie a termini positivi, possiamo usare la proprietà associativa della somma: si ottiene pertanto

$$\mathbb{E}[Y] = \sum_i y_i p_Y(y_i) = \sum_i y_i \left(\sum_{j \in A_i} p_X(x_j) \right) = \sum_i \left(\sum_{j \in A_i} g(x_j) p_X(x_j) \right) = \sum_j g(x_j) p_X(x_j),$$

cioè l'eguaglianza desiderata. Il caso generale si ottiene scrivendo la funzione g nella forma $g = g^+ - g^-$ e sommando le relative serie (ricordiamo che con $g^+(x) = \max(g(x), 0)$ e $g^-(x) = -\min(g(x), 0)$ intendiamo la parte positiva e la parte negativa della funzione g). ■

Esercizio 4.16 Calcolare il valore atteso del quadrato dell'esito del lancio di un dado equo. ■

Quando si passa alle variabili con densità, una dimostrazione rigorosa è impegnativa ma può essere istruttivo vedere qualche facile esempio.

Esercizio 4.17 Sia X con densità uniforme su $[0, 1]$ e $Y = X^2$: mostrare che la densità di Y è data da $f_Y(y) = \frac{1}{2\sqrt{y}}$ per $y \in [0, 1]$ e nulla altrove, e che le due formule

$$\mathbb{E}[X^2] = \int_0^1 x^2 dx = 1/3, \quad \mathbb{E}[Y] = \int_0^1 \frac{y}{2\sqrt{y}} dy = 1/3,$$

(la prima dalla Proposizione appena illustrata, la seconda dalla definizione di valore atteso) danno lo stesso risultato. ■

Discutiamo ora alcune proprietà generali del valore atteso.

Proposizione 4.4.2 Se la v.a. X (discreta o con densità) ha valore atteso, valgono le seguenti proprietà:

- per ogni $a, b \in \mathbb{R}$, $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$; in particolare, $\mathbb{E}[b] = b$;
- $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$;
- se $\mathbb{P}(X \geq 0) = 1$, allora $\mathbb{E}[X] \geq 0$.

Sono tutte conseguenze abbastanza dirette della definizione, che lasciamo per esercizio.

Definizione 4.4.2 Si dice che la v.a. X (discreta o con densità) ammette momento di ordine $n = 1, 2, \dots$ se $\mathbb{E}[|X|^n] < +\infty$ e in tal caso si chiama momento di ordine n il numero $\mathbb{E}[X^n]$.

Esercizio 4.18 Sia $\beta > 1$ un parametro e sia X una v.a. avente densità $f(x) = (\beta - 1)x^{-\beta}$ per $x \in (1, +\infty)$, $f(x) = 0$ per $x \notin (1, +\infty)$. Determinare, in funzione di β e n , se X ammette momento di ordine n e, se sì, calcolarlo. ■

Naturalmente il momento primo è il valore atteso. Il risultato che segue mostra che se una v.a. possiede momento n -simo, possiede anche tutti i momenti di ordine inferiore.

Proposizione 4.4.3 Siano $1 \leq m < n$: se $\mathbb{E}[|X|^n] < +\infty$, anche $\mathbb{E}[|X|^m] < +\infty$.

La dimostrazione più generale (valida per ogni tipo di variabile aleatoria) segue dalla importante *disuguaglianza di Jensen*, dalla quale si ricava un risultato più preciso,

$$\mathbb{E}[|X|^m]^{1/m} \leq \mathbb{E}[|X|^n]^{1/n}.$$

Dimostrazione. Limitandoci al caso discreto (quello continuo è del tutto analogo), partendo dalla disuguaglianza valida per ogni numero t , $|t|^m \leq |t|^n + 1$, si ha

$$\sum_{x_i} |x_i|^m p(x_i) \leq \sum_{x_i} (|x_i|^n + 1) p(x_i) = \sum_{x_i} |x_i|^n p(x_i) + \sum_{x_i} p(x_i) < +\infty,$$

da cui $\mathbb{E}[|X|^n] < +\infty \Rightarrow \mathbb{E}[|X|^m] < +\infty$. ■

Proposizione 4.4.4 — Disuguaglianza di Markov. Se X è una variabile aleatoria (discreta o con densità) a valori positivi e $a > 0$ vale

$$a\mathbb{P}\{X \geq a\} \leq \mathbb{E}[X].$$

Dimostrazione. Consideriamo dapprima il caso discreto: vale

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i) \geq \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) \cdot \begin{cases} a & \text{se } x_i \geq a, \\ 0 & \text{se } x_i < a \end{cases} \\ &= \sum_{i: x_i \geq a} a \mathbb{P}(X = x_i) = a\mathbb{P}(X \geq a). \end{aligned}$$

Nel caso con densità,

$$\mathbb{E}[X] = \int_0^{\infty} t f_X(t) dt \geq \int_a^{\infty} t f_X(t) dt \geq \int_a^{\infty} a f_X(t) dt = a\mathbb{P}(X \geq a),$$

usando la stessa idea del caso discreto. ■



Nell'ultima Proposizione non si fa alcuna ipotesi sull'esistenza dei momenti: le disuguaglianze hanno senso e sono verificate anche se qualcuno dei valori attesi è infinito.

Riprendiamo il nostro esempio di estrazione per comprendere la relazione tra valore atteso e media campionaria:

■ **Esempio 4.6** Sia S una popolazione e sia $X : S \rightarrow \mathbb{R}$ una v.a. che rappresenta un carattere (ad esempio S = popolazione italiana e $X(\omega)$ = numero di figli di ω). Supponiamo che X sia discreta, a valori in x_1, x_2, \dots . L'estrazione di un individuo dalla popolazione viene modellizzata prendendo $\Omega = S$, \mathbb{P} uniforme. Il valore atteso di X è quindi

$$\mathbb{E}[X] = \sum_i x_i \mathbb{P}\{X = x_i\} = \sum_i x_i \frac{\#\{\omega \in S \mid X(\omega) = x_i\}}{\#S},$$

dove $\#\{\omega \in S \mid X(\omega) = x_i\}/\#S$ è la frequenza relativa di $X = x_i$ su tutta la popolazione. Ricordando la formula per la media campionaria in termini della frequenza relativa, notiamo che il valore atteso è la media campionaria *su tutta la popolazione* (e non su un singolo campione). ■

Per tutto il resto di questo paragrafo supponiamo che le variabili X, Y, \dots considerate abbiano momento secondo. Definiamo ora la varianza, che indica quanto è dispersa (attorno al valore atteso) una variabile aleatoria. Come per i dati la varianza empirica è la media campionaria dei quadrati degli scarti dei dati dalla media, la varianza di una v.a. è la media pesata (cioè il valore atteso) dei quadrati degli scarti di una v.a. X dal suo valore atteso.

Definizione 4.4.3 Si chiama varianza di una v.a. X il numero

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

Si chiama scarto quadratico medio o anche deviazione standard la radice quadrata della varianza di X , cioè

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

Come per il valore atteso, la varianza di X dipende solo dalla legge P_X di X .

Sviluppando i calcoli si trova l'eguaglianza

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

(di solito è più comodo calcolare la varianza in questo modo). Notiamo anche che si ha

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

■ **Esercizio 4.19** Calcolare la varianza dell'esito del lancio di un dado equilibrato.

In analogia al caso discreto, la seguente disuguaglianza quantifica la relazione tra la varianza e la dispersione dei dati:

Proposizione 4.4.5 — Disuguaglianza di Chebyshev. Se X è una variabile aleatoria e $d > 0$ vale

$$\mathbb{P}\{|X - \mathbb{E}[X]| > d\} \leq \frac{\text{Var}(X)}{d^2}.$$

Dimostrazione. Questa disuguaglianza deriva immediatamente dalla disuguaglianza di Markov prendendo $Y = (X - \mathbb{E}[X])^2$, $a = d^2$ e tenendo presente che

$$\mathbb{P}\{|X - \mathbb{E}[X]| > d\} = \mathbb{P}((X - \mathbb{E}[X])^2 > d^2)$$

(i due membri sono la probabilità dello stesso evento). ■

Il caso limite si ha quando la varianza è 0 : la varianza della variabile X è eguale a 0 se e solo se X è costante eccetto che su un insieme trascurabile, cioè $\mathbb{P}\{|X - \mathbb{E}[X]| \neq 0\} = 0$.

Consideriamo ancora l'esempio di estrazione:

■ **Esempio 4.7** Sia S una popolazione e sia $X : S \rightarrow \mathbb{R}$ una v.a. discreta che rappresenta un carattere (ad esempio S = popolazione italiana e $X(\omega)$ = numero di figli di ω). Come per il valore atteso, si può mostrare che la varianza $\text{Var}(X)$ è la varianza empirica del carattere *su tutta la popolazione*. ■

Concludiamo questa sezione tornando alle variabili aleatorie notevoli introdotte precedentemente, per discuterne i momenti.

■ **Esempio 4.8** (Variabile Binomiale). Cominciamo con la variabile di Bernoulli, $n = 1$: essa assume il valore 1 con probabilità p ed il valore 0 con probabilità $(1 - p)$, ed è dunque immediato constatare che si ha $\mathbb{E}[X] = p$, $\mathbb{E}[X^2] = p$ e quindi $\text{Var}(X) = p - p^2 = p(1 - p)$.

Poiché una variabile X Binomiale di parametri n e p può essere vista come somma di n variabili di Bernoulli di parametro p indipendenti, si ha $\mathbb{E}[X] = np$ e $\text{Var}(X) = np(1 - p)$. ■

■ **Esempio 4.9** (Variabile di Poisson). Presa X di Poisson di parametro λ , a priori non sappiamo se ha momento primo ma poiché prende solo valori positivi ha sicuramente senso scrivere

$$\mathbb{E}[X] = \sum_{h=0}^{+\infty} h \mathbb{E}^{-\lambda} \frac{\lambda^h}{h!} = \lambda \mathbb{E}^{-\lambda} \sum_{h=1}^{+\infty} \frac{\lambda^{h-1}}{(h-1)!} = \lambda \mathbb{E}^{-\lambda} \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = \lambda.$$

Con conti simili si prova $\mathbb{E}[X^2] = \lambda + \lambda^2$ e quindi $\text{Var}(X) = \lambda$. ■

■ **Esempio 4.10** (Densità uniforme). Premettiamo una osservazione: se la densità di una v.a. X è diversa da 0 solo su un intervallo limitato, sicuramente la variabile possiede tutti i momenti. Prendiamo ora X con densità uniforme su $[a, b]$:

$$\begin{aligned} \mathbb{E}[X] &= \int_a^b \frac{x}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}, \\ \mathbb{E}[X^2] &= \int_a^b \frac{x^2}{b-a} dx = \frac{x^3}{3(b-a)} \Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}. \end{aligned}$$

e quindi $\text{Var}(X) = \frac{(b-a)^2}{12}$. ■

■ **Esempio 4.11** (Densità esponenziale). Anche di questa variabile a priori non sappiamo quali momenti possieda, tuttavia (essendo la densità diversa da 0 solo per x positivo) possiamo calcolare (con un cambio di variabili)

$$\mathbb{E}[X] = \int_0^{+\infty} \lambda x e^{-\lambda x} dx = \frac{1}{\lambda} \int_0^{+\infty} t e^{-t} dt = \frac{1}{\lambda}$$

Con calcoli simili si trova $\mathbb{E}[X^2] = \frac{2}{\lambda^2}$ e quindi $\text{Var}(X) = \frac{1}{\lambda^2}$. ■

■ **Esempio 4.12** (Variabili Gaussiane). Cominciamo con X Gaussiana standard: poichè l'esponenziale cresce più velocemente di ogni potenza, vale per ogni n

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} |x|^n e^{-\frac{x^2}{2}} dx < +\infty,$$

quindi la variabile possiede tutti i momenti. Inoltre se $n = 2h + 1$ è dispari vale

$$\mathbb{E}[X^n] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^n e^{-\frac{x^2}{2}} dx = \lim_{M \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{-M}^{+M} x^n e^{-\frac{x^2}{2}} dx = 0.$$

Rimangono da calcolare gli n -momenti con n pari, cominciamo col momento secondo. Con una integrazione per parti (considerando che $-xe^{-\frac{x^2}{2}}$ è la derivata di $e^{-\frac{x^2}{2}}$) si ottiene

$$\mathbb{E}[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} dx = \left. \frac{-xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \right|_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = 0 + 1 = 1,$$

e dunque anche $\text{Var}(X) = 1$. Con conti analoghi si prova l'eguaglianza $\mathbb{E}[X^{2h+2}] = (2h+1)\mathbb{E}[X^{2h}]$ e quindi ad esempio $\mathbb{E}[X^4] = 3$, $\mathbb{E}[X^6] = 5 \times 3 = 15$ e così via.

Per quanto riguarda le variabili Gaussiane più generali, una variabile Y con densità $N(m, \sigma^2)$ si scrive nella forma $(\sigma X + m)$ con X standard e quindi $\mathbb{E}[Y] = \mathbb{E}[\sigma X + m] = m$ e $\text{Var}(Y) = \text{Var}(\sigma X + m) = \sigma^2 \text{Var}(X) = \sigma^2$. In modo analogo si procede con i momenti di ordine superiore. ■

4.5 Esercizi

Proprietà Generali delle Variabili Aleatorie

Esercizio 4.20 Si introducano uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e una variabile aleatoria $X : \Omega \rightarrow \mathbb{R}$ che descrivono “il massimo tra i valori di due lanci di un dado a sei facce”. Si calcoli la legge di X e si dica se corrisponde a uno degli esempi di questo capitolo. Si calcolino poi valore atteso e varianza di X . ■

Esercizio 4.21 — Difficile. Si generalizzino i risultati dell'esercizio precedente al caso di n lanci di un dado a k facce. Per farlo, ricordiamo il *principio di inclusione-esclusione*: dati insiemi A_1, \dots, A_n possibilmente non disgiunti, la cardinalità della loro unione è data da:

$$\begin{aligned} |A_1 \cup A_2| &= |A_1| + |A_2| - |A_1 \cap A_2|, \\ |A_1 \cup A_3| &= |A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| + |A_1 \cap A_2 \cap A_3|, \\ \left| \bigcup_{i=1}^n A_i \right| &= \sum_{i=1}^n (-1)^{i+1} \sum_{1 \leq j_1 \leq \dots \leq j_i \leq n} \left| \bigcap_{h=1}^i A_{j_h} \right|. \end{aligned}$$

Esercizio 4.22 Un test a risposta multipla prevede 10 domande con 4 possibili risposte per ogni domanda. Il punteggio viene assegnato con le seguenti regole: +1 punto per ogni risposta esatta; -0.25 punti per ogni risposta sbagliata o non data. Supponendo di rispondere a tutte le domande del test in maniera casuale, con uguale probabilità per ogni risposta, si determini:

- la probabilità di ottenere almeno 8 risposte esatte;
- il valore atteso della variabile aleatoria che descrive il punteggio ottenuto al test.

Esercizio 4.23 In un'urna ci sono 6 biglie blu e 4 rosse. Vengono effettuate 4 estrazioni casuali con reinserimento (cioè, ad ogni estrazione, la biglia estratta viene rimessa nell'urna). Qual è la probabilità che venga estratta al massimo una biglia rossa? ■

Esercizio 4.24 Sappiamo dalla letteratura scientifica che un certo farmaco è efficace contro una data malattia nell'80% dei pazienti. Presi 10 pazienti a caso, calcolare:

- la probabilità che il farmaco sia efficace su almeno 9 di loro;
- il valore atteso e la deviazione standard del numero di pazienti (tra i 10 scelti) su cui il

farmaco è efficace.

Le risposte ai punti precedenti cambiano se sappiamo che i pazienti scelti sono parenti tra loro?

■

Esercizio 4.25 Sia $c \in \mathbb{R}$ un parametro e sia $p : \mathbb{N} \rightarrow \mathbb{R}$ data da $p(k) = c2^{-k}$.

- Determinare l'unico valore di c per cui p sia una funzione di massa.

Prendiamo ora c pari a tale valore e sia X una v.a. discreta con funzione di massa p .

- Dire per quali valori di $\alpha \in \mathbb{R}$ la v.a. $e^{\alpha X}$ ammette momento primo.

■

Esercizio 4.26 I 42 studenti di un corso di calligrafia, dei quali 13 sono mancini, sono divisi per le esercitazioni in due gruppi di eguale numero per sorteggio. Si indichino rispettivamente con X e Y le v.a. il cui valore è il numero di mancini presenti nel primo e nel secondo gruppo.

- Calcolare la funzione di probabilità (o densità discreta) della v.a. X .
- Provare che le v.a. X e Y sono equidistribuite, ma non sono indipendenti.
- È possibile determinare i valori attesi $E[X]$ e $E[Y]$ senza usare le leggi di X e Y ?

■

Esercizio 4.27 Un'auto va in panne in un punto a caso di una strada di 60km. Non avendo altre informazioni, con quale probabilità l'auto si è fermata tra il km 20 e il km 40?

■

Esercizio 4.28 Sia $p \in [0, 1]$, e si consideri la funzione

$$f(x) = \begin{cases} p, & 0 < x \leq 1 \\ (1-p), & 1 < x \leq 2 \\ 0, & x \leq 0, x > 2. \end{cases}$$

- Dopo aver riconosciuto che la funzione sopra scritta è una densità di probabilità, scrivere l'espressione della funzione di ripartizione e la formula per il β -quantile (con $0 < \beta < 1$) per una v.a. X che abbia quella densità.
- Calcolare i momenti primo e secondo di una v.a. X che abbia densità $f(x)$. Esaminare quando, al variare di p in $[0, 1]$, la varianza è massima e quando è minima.

■

Esercizio 4.29 Sia $c \in \mathbb{R}$ un parametro, si consideri la funzione

$$f(x) = \begin{cases} cx^2 & -1 < x < 1 \\ 0 & |x| \geq 1 \end{cases}.$$

- Dimostrare che f è densità se e solo se $c = 3/2$.

D'ora in poi consideriamo $c = 3/2$, sia X una v.a. con densità f .

- Determinare i momenti, se esistono, di X e la varianza, se esiste, di X .
- Determinare la densità, se esiste, di $Y = X^3$ e riconoscerla.

■

Esercizio 4.30 Dati $a, b \in \mathbb{R}$ parametri reali, si consideri la funzione

$$f(x) = \begin{cases} -x & -1 \leq x < 0 \\ ax + b & 0 \leq x \leq 1 \\ 0 & |x| > 1 \end{cases},$$

- Trovare tutti i valori di a e b per i quali f è una densità di probabilità.
- Sia X una v.a. avente densità f (con a e b che soddisfano alle condizioni trovate nel punto precedente): calcolare a e b in modo che si abbia $E[X] = 0$.
- In funzione di a e b , calcolare $\mathbf{P}(|X| \geq \frac{1}{2} \mid X \leq 0)$.

■

Esercizio 4.31 Dati $a, b \in \mathbb{R}$ parametri reali, si consideri la funzione

$$F_{a,b}(x) = \begin{cases} 0, & \text{se } x \leq 0 \\ a - e^{-x}, & \text{se } 0 < x < 1 \\ a - e^{-1} + b(x-1), & \text{se } 1 \leq x \leq 2 \\ 1, & \text{se } x > 2 \end{cases}$$

- Si determinino i valori di a e b affinché $F_{a,b}$ sia la funzione di ripartizione di una variabile aleatoria X con densità.
- Usando i valori di a e b trovati in (i), scrivere la formula per il calcolo dei β -quantili della variabile aleatoria X per ogni $\beta \in (0, 1)$.
- Sia X la variabile aleatoria al primo punto, scrivere la densità della variabile aleatoria $Y = (X + 1)^2$, e calcolare il valore atteso di Y .

■

Esercizio 4.32 Mostrare che la seguente funzione è la densità di probabilità di una variabile aleatoria:

$$f(x) = \begin{cases} 3x^{-4}, & \text{se } x > 1 \\ 0, & \text{se } x \leq 1 \end{cases}.$$

Se X ha densità $f_X = f$, si determini quali momenti possiede X e li si calcoli. Si determini se la variabile $Y = \log(X)$ ha densità e momento primo (ed in tal caso li si calcoli).

■

Esercizio 4.33 Per quali valori di $a \in \mathbb{R}$ la funzione

$$F(x) = \begin{cases} 0 & x < 0 \\ ax^2 & 0 \leq x < 1 \\ 1 & x \geq 1, \end{cases}$$

è la c.d.f. di una variabile aleatoria? Per quali la c.d.f. di una variabile con densità?

■

Esercizio 4.34 — Difficile. Si mostri che la seguente è una densità di probabilità,

$$f(x) = \begin{cases} \frac{1}{1+\pi \sin(x)} & x \in [-1, 1] \\ 0 & x \notin [-1, 1]. \end{cases}$$

e si calcolino tutti i momenti pari di una variabile aleatoria con tale densità. ■

Variabili Gaussiane

Esercizio 4.35 — Integrale Gaussiano, metodo di Laplace. Si consideri il quadrato dell'integrale Gaussiano,

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right)^2 = \left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right) \cdot \left(\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy \right) = \iint_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} dx dy,$$

e si mostri passando in coordinate polari $x = r \cos(\theta)$, $y = r \sin(\theta)$ che tale integrale doppio ha valore 2π . ■

Esercizio 4.36 — Integrale Gaussiano, metodo di Feynman. Sia $G = \int_0^\infty e^{-x^2} dx$. Si consideri la funzione di $a \geq 0$ data da

$$g(a) = \int_0^\infty \frac{e^{-a^2(x^2+1)}}{x^2+1} dx; \quad g(0) = \frac{\pi}{2}, \quad g(\infty) = 0,$$

(il valore in $a = 0$ ricordando la derivata dell'arcotangente), e differenziando sotto il segno di integrale,

$$g'(a) = -2Ge^{-a^2},$$

da cui, integrando ambo i membri in da si ottiene una equazione chiusa per G . ■

Esercizio 4.37 Sia X una v.a. gaussiana di media 2 e varianza 9. Si calcolino $\mathbb{P}\{1 \leq X \leq 3\}$, $\mathbb{P}\{X \geq 2\}$, $\mathbb{P}\{X \geq 3\}$, $\mathbb{P}\{|X - 2| \leq 3\}$. ■

Esercizio 4.38 Una fabbrica produce viti la cui lunghezza ha distribuzione gaussiana di media 1.7cm e deviazione standard 0.6cm.

- Calcolare la probabilità che una vite abbia lunghezza inferiore a 1.6cm.
 - Quanto deve valere x (in cm) in modo che il 95% delle viti abbia lunghezza almeno x ?
 - Qual è la probabilità che, estratte 3 viti a caso, esattamente 2 abbiano lunghezza inferiore a 1.6cm?
-

Esercizio 4.39 L'esplosione di un'autocisterna in autostrada ha provocato l'emissione di un inquinante. Si stima che la quantità relativa di inquinante riversato sull'autostrada segua una distribuzione gaussiana centrata nel punto dell'esplosione e di deviazione standard 0.5km. Le autorità vogliano chiudere l'autostrada per un tratto assicurandosi che la quantità di inquinante riversato al di fuori di quel tratto sia inferiore allo 0.5% di tutto l'inquinante riversato. Quale tratto bisogna chiudere? ■

Generazione di Numeri Casuali

Data una variabile aleatoria X e fissato $\beta \in (0, 1)$, abbiamo visto che in generale il β -quantile non è determinato univocamente. Un modo per risolvere questa ambiguità è scegliere tra i possibili β -quantili il più piccolo, ovvero definire

$$r_\beta = \inf\{r \in \mathbb{R} : F(r) \geq \beta\}, \quad \beta \in (0, 1).$$

La funzione a destra si chiama *inversa generalizzata* di F .

Esercizio 4.40 Data $F : \mathbb{R} \rightarrow [0, 1]$ non decrescente e continua a destra, con $F(-\infty) = 0$ e $F(+\infty) = 1$, si definisca

$$F^{\leftarrow}(t) = \inf\{s \in \mathbb{R} : F(s) \geq t\}.$$

Si mostri che:

- F^{\leftarrow} è non decrescente;
- $F^{\leftarrow}(F(t)) \leq t$ per ogni $t \in \mathbb{R}$;
- $F(F^{\leftarrow}(t)) \geq t$ per ogni $t \in \mathbb{R}$;
- $F^{\leftarrow}(t) \leq s$ se e solo se $F(s) \geq t$.

■

Il problema di generare (con una macchina) numeri *veramente* casuali è complesso, esula dallo scopo di queste note e richiederebbe una discussione approfondita sul significato di “esito casuale” di un esperimento. Assumendo però di poter generare un numero casuale con distribuzione uniforme sull’intervallo $[0, 1]$ (compito implementato come funzione di base in praticamente ogni software per il calcolo numerico), possiamo mostrare come generare un numero casuale la cui legge di probabilità è data da una c.d.f. F assegnata.

Esercizio 4.41 Sia data una funzione di ripartizione $F : \mathbb{R} \rightarrow \mathbb{R}$, ovvero F soddisfa le proprietà che caratterizzano le c.d.f. (che abbiamo preso come ipotesi nell’esercizio precedente). Sia poi $U : \Omega \rightarrow [0, 1]$ una variabile con densità uniforme definita su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. Si mostri che

$$X : \Omega \rightarrow \mathbb{R}, \quad X = F^{\leftarrow}(U),$$

è una variabile aleatoria avente c.d.f. $F_X = F$.

■

5. Variabili Aleatorie Multivariate

Iniziamo ora la discussione delle distribuzioni multivariate, ovvero delle variabili aleatorie multiple. Per semplicità di esposizione ci limitiamo in molti casi a discutere solamente variabili doppie, l'estensione al caso generale non comporta complicazioni aggiuntive.

5.1 Variabili Doppie

Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità, e consideriamo due variabili aleatorie $X, Y : \Omega \rightarrow \mathbb{R}$. Il vettore (X, Y) può essere dunque visto come una funzione

$$(X, Y) : \Omega \rightarrow \mathbb{R}^2, \quad \Omega \ni \omega \mapsto (X(\omega), Y(\omega)) \in \mathbb{R}^2.$$

Perciò, in analogia al caso univariato, definiamo la legge della variabile aleatoria doppia (X, Y) come una probabilità sui sottoinsiemi (misurabili) di \mathbb{R}^2 ,

$$\mathbb{P}_{(X,Y)}(A) = \mathbb{P}((X, Y) \in A) = \mathbb{P}\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}, \quad A \subseteq \mathbb{R}^2.$$

Se $A = A_1 \times A_2$ è un sottoinsieme rettangolare, notiamo che $\{(X, Y) \in A\} = \{X \in A_1, Y \in A_2\}$. Al fine di evitare confusioni, precisiamo che la virgola indica l'intersezione di due condizioni: nel caso precedente

$$\{X \in A_1, Y \in A_2\} = \{X \in A_1\} \cap \{Y \in A_2\} = X^{-1}(A_1) \cap Y^{-1}(A_2) = (X, Y)^{-1}(A_1 \times A_2).$$

Se X e Y rappresentano due caratteri (quantitativi) di un dato esperimento, la variabile doppia (X, Y) rappresenta quindi la coppia dei due caratteri; ad esempio, in un'estrazione di un individuo da una popolazione, se X è l'altezza di un individuo e Y è il suo peso, la coppia (X, Y) rappresenta la coppia (altezza, peso) dell'individuo estratto. L'evento $\{X \in A_1, Y \in A_2\}$ corrisponde a “si verificano congiuntamente $X \in A_1$ e $Y \in A_2$ ”; nell'esempio (altezza, peso), $\{X \in [1.70, 1.80], Y \in [80, 90]\}$ sta per “l'individuo ha altezza tra 1.70 e 1.80 e peso tra 80 e 90”.

Data una variabile doppia (X, Y) possiamo considerare separatamente le leggi \mathbb{P}_X e \mathbb{P}_Y (su \mathbb{R}) delle due componenti: tali leggi sono dette *distribuzioni marginali* della variabile doppia. Si vede

facilmente dalla definizione che, per ogni $A \subseteq \mathbb{R}$ (misurabile),

$$\mathbb{P}_X(A) = \mathbb{P}_{(X,Y)}(A \times \mathbb{R}), \quad \mathbb{P}_Y(A) = \mathbb{P}_{(X,Y)}(\mathbb{R} \times A).$$

In particolare la legge congiunta $\mathbb{P}_{(X,Y)}$ determina univocamente le leggi marginali \mathbb{P}_X e \mathbb{P}_Y .

NB

Le distribuzioni marginali non contengono tutta l'informazione della legge $\mathbb{P}_{(X,Y)}$, ossia non si può ricostruire univocamente quest'ultima a partire dalle marginali, come risulterà evidente nella discussione sulla (in)dipendenza di variabili aleatorie a seguire. L'idea intuitiva è che $\mathbb{P}_{(X,Y)}$ codifica anche le relazioni tra le v.a. X e Y , cosa che invece non fanno le leggi marginali \mathbb{P}_X e \mathbb{P}_Y .

Definizione 5.1.1 Una variabile doppia (X, Y) è detta discreta se la sua immagine è concentrata in un insieme finito o numerabile di punti (x_i, y_j) . In questo caso la sua distribuzione di probabilità $\mathbb{P}_{(X,Y)}$ è identificata dalla funzione di massa $p(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j)$ e si ha, per $A \subseteq \mathbb{R}^2$,

$$\mathbb{P}_{(X,Y)}(A) = \mathbb{P}\{(X, Y) \in A\} = \sum_{(x_i, y_j) \in A} p(x_i, y_j).$$

Proposizione 5.1.1 Data una variabile doppia (X, Y) discreta con funzione di massa $p(x_i, y_j)$, le sue componenti sono variabili aleatorie discrete con funzioni di massa

$$p_X(x_i) = \sum_{y_j} p(x_i, y_j) \quad p_Y(y_j) = \sum_{x_i} p(x_i, y_j).$$

Dimostrazione. Notiamo che $(X = x_i) = \bigcup_j (X = x_i, Y = y_j)$ e questi ultimi insiemi sono a due a due disgiunti. Si ha pertanto

$$p_X(x_i) = \mathbb{P}(X = x_i) = \sum_{y_j} \mathbb{P}(X = x_i, Y = y_j) = \sum_{y_j} p(x_i, y_j),$$

l'altra uguaglianza è analoga. ■

Esercizio 5.1 Dimostrare che, se X e Y sono v.a. reali discrete, allora (X, Y) è una v.a. doppia discreta. ■

Definizione 5.1.2 Una variabile doppia (X, Y) ha densità se esiste una funzione $f: \mathbb{R}^2 \rightarrow [0, \infty)$ integrabile e con $\iint_{\mathbb{R}^2} f(x, y) dx dy = 1$ tale che valga, per $A \subseteq \mathbb{R}^2$

$$\mathbb{P}_{(X,Y)}(A) = \mathbb{P}\{(X, Y) \in A\} = \iint_A f(x, y) dx dy.$$

NB

Non ci addentriamo nei dettagli della definizione degli integrali doppi: ai fini di queste note basta assumere di poter integrare funzioni di due variabili su sottoinsiemi (misurabili) di \mathbb{R}^2 , e che per funzioni non-negative integrate su insiemi rettangolari $A = A_1 \times A_2$ vale:

$$\iint_A f(x, y) dx dy = \int_{A_1} \left(\int_{A_2} f(x, y) dy \right) dx = \int_{A_2} \left(\int_{A_1} f(x, y) dx \right) dy$$

(formula nota come Teorema di Fubini-Tonelli), cioè l'integrale doppio coincide con un integrale iterato e l'ordine in cui si integrano le variabili non conta.

Proposizione 5.1.2 Se (X, Y) ha densità $f(x, y)$, anche X e Y hanno densità rispettivamente date da

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx.$$

Dimostrazione. Notiamo che, per ogni insieme (misurabile) $A \subseteq \mathbb{R}$, $\{X \in A\} = \{X \in A, Y \in \mathbb{R}\}$. Si ha pertanto

$$\mathbb{P}\{X \in A\} = \mathbb{P}\{X \in A, Y \in \mathbb{R}\} = \int_A \left(\int_{\mathbb{R}} f(x, y) dy \right) dx.$$

Quindi $f_X(x) = \int_{\mathbb{R}} f(x, y) dy$ è la densità della v.a. X . Analogamente si ottiene la formula per la densità f_Y . ■

A differenza del caso discreto, se X e Y sono v.a. con densità, non è detto che (X, Y) abbia densità. Ad esempio, si può dimostrare (con tecniche più avanzate) che (X, X) non ha densità.

5.2 Indipendenza di Variabili Aleatorie

La seguente definizione codifica l'idea intuitiva che ogni informazione legata ad X è indipendente da ogni informazione legata ad Y (cioè non modifica le probabilità di eventi legati a Y).

Definizione 5.2.1 — Indipendenza di variabili aleatorie. Due variabili aleatorie X e Y sono dette indipendenti se, presi comunque A, B sottinsiemi di \mathbb{R} , gli eventi $X^{-1}(A)$ e $Y^{-1}(B)$ sono indipendenti, cioè se vale

$$\mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{Y \in B\}.$$

Più in generale, le variabili aleatorie $X_1, \dots, X_n : \Omega \rightarrow \mathbb{R}$ si dicono indipendenti se presi comunque $A_1, \dots, A_n \subseteq \mathbb{R}$ vale

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

Facciamo subito alcuni esempi per chiarire il concetto.

■ **Esempio 5.1** Lanciamo n volte una moneta: le v.a. $X_i = 1$ se esce testa all' i -simo lancio, $= 0$ altrimenti, sono indipendenti, come facile conseguenza del fatto che gli eventi “il k -esimo lancio ha dato testa” al variare di k sono tutti indipendenti (cf. il Capitolo 3). ■

■ **Esempio 5.2** Data una variabile X (non costante), le variabili X e $-X$ non sono indipendenti: intuitivamente se conosciamo una abbiamo immediatamente l'altra cambiando di segno, e infatti scegliendo ad esempio $B = -A$ (lo stesso sottoinsieme di \mathbb{R} ma con i segni cambiati) nella definizione vale

$$\mathbb{P}\{X \in A, -X \in -A\} = \mathbb{P}\{X \in A\}, \quad \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{-X \in -A\} = \mathbb{P}\{X \in A\}^2,$$

che sono due numeri diversi se scegliamo un qualsiasi A tale che $\mathbb{P}\{X \in A\} \neq 0, 1$ (questo è sempre possibile se X non è costante). ■

NB Come per gli eventi, l'indipendenza a due a due di variabili aleatorie X_1, \dots, X_n non implica l'indipendenza della famiglia X_1, \dots, X_n .

È importante avere una caratterizzazione pratica dell'indipendenza, e questa è fornita dal seguente risultato.

Proposizione 5.2.1 Date due variabili discrete X e Y , con immagine rispettivamente nei punti x_i, y_j , queste sono indipendenti se e solo se vale l'eguaglianza tra le funzioni di massa

$$p(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j), \quad \forall (x_i, y_j).$$

Date due variabili X, Y tali che (X, Y) abbia densità, le variabili sono indipendenti se e solo se vale l'eguaglianza tra densità

$$f(x, y) = f_X(x) \cdot f_Y(y), \quad \forall (x, y).$$

Dimostrazione. Ci limitiamo al caso di variabili discrete. Da una parte, prendendo $A = (x_i)$ e $B = (y_j)$, si ha $p(x_i, y_j) = \mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i) \mathbb{P}(Y = y_j) = p_X(x_i) p_Y(y_j)$. Viceversa, presi due sottinsiemi A e B di \mathbb{R} si ha:

$$\begin{aligned} \mathbb{P}\{X \in A, Y \in B\} &= \sum_{x_i \in A, y_j \in B} p(x_i, y_j) = \sum_{x_i \in A} \sum_{y_j \in B} p_X(x_i) p_Y(y_j) \\ &= \left(\sum_{x_i \in A} p_X(x_i) \right) \left(\sum_{y_j \in B} p_Y(y_j) \right) = \mathbb{P}\{X \in A\} \mathbb{P}\{Y \in B\}, \end{aligned}$$

da cui la tesi. ■

I prossimi due esempi mostrano –come anticipato– che due variabili aleatorie possono avere le stesse distribuzioni marginali ma essere diverse: ad esempio perché in un caso le componenti sono indipendenti e nell'altro no.

■ **Esempio 5.3** Consideriamo la variabile doppia discreta (X, Y) con funzione di massa

$$p(1, 1) = p(1, -1) = p(-1, 1) = p(-1, -1) = \frac{1}{4}.$$

Per quanto visto sopra, le due componenti X e Y hanno la stessa funzione di massa, $p_X(1) = p_X(-1) = 1/2$, $p_X = p_Y$. Le variabili X, Y sono inoltre indipendenti: si verifica facilmente che $p(x, y) = p_X(x) p_Y(y)$. ■

■ **Esempio 5.4** Consideriamo una variabile doppia discreta (X, Y) con una funzione di massa leggermente diversa dall'esempio precedente:

$$p(1, 1) = p(-1, -1) = \frac{1}{2}.$$

Le due componenti X e Y hanno ancora la stessa funzione di massa $p_X(1) = p_X(-1) = 1/2$, $p_X = p_Y$. Tuttavia X e Y non sono indipendenti, perché valgono

$$p(1, -1) = 0 \neq p_X(1) p_Y(-1) = \frac{1}{4}, \quad p(-1, 1) = 0 \neq p_X(-1) p_Y(1) = \frac{1}{4},$$

contraddicendo la caratterizzazione data sopra. ■

Se in generale date le distribuzioni marginali $\mathbb{P}_X, \mathbb{P}_Y$ non conosciamo la legge della coppia $\mathbb{P}_{(X, Y)}$, possiamo però sempre costruire artificialmente uno spazio di probabilità in cui due variabili X, Y hanno marginali assegnate $\mathbb{P}_X, \mathbb{P}_Y$ e sono indipendenti.

Esercizio 5.2 Siano date due leggi di probabilità \mathbb{P}_1 e \mathbb{P}_2 su \mathbb{R} (con la σ -algebra dei misurabili), e si definisca sugli insiemi rettangolari

$$\Omega = \mathbb{R}^2, \quad \mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1) \mathbb{P}_2(A_2), \quad \forall A_1, A_2 \subseteq \mathbb{R}.$$

Si assuma di poter estendere la definizione agli insiemi misurabili di \mathbb{R}^2 , ovvero in questo caso

quelli per i quali è definita l'area. Si introducano poi due variabili aleatorie

$$X : \Omega = \mathbb{R}^2 \rightarrow \mathbb{R}, X(x, y) = x, \quad Y : \Omega = \mathbb{R}^2 \rightarrow \mathbb{R}, Y(x, y) = y.$$

Si verifichi che $\mathbb{P}_X = \mathbb{P}_1$, $\mathbb{P}_Y = \mathbb{P}_2$ e che le variabili X, Y sono indipendenti. ■

La seguente osservazione, che lasciamo come ulteriore ed istruttivo esercizio, è spesso utile per mostrare che due variabili sono indipendenti.

Esercizio 5.3 Date X e Y variabili indipendenti, e due funzioni $h, k : \mathbb{R} \rightarrow \mathbb{R}$, mostrare che la variabili $h(X)$ e $k(Y)$ sono indipendenti. ■

Più in generale, funzioni di più variabili indipendenti sono indipendenti se la stessa variabile non compare in due funzioni diverse: ad esempio se X, Y e Z sono indipendenti, lo sono anche $\sqrt{X^2 + Y^2}$ e Z^3 mentre non sono in generale indipendenti $\sqrt{X^2 + Y^2}$ e $\sqrt{X^2 + Z^2}$ (questo ultimo fatto è più difficile da dimostrare e non lo discutiamo in dettaglio).

Come per gli eventi, un esempio rilevante di v.a. indipendenti è legato alle prove ripetute: in un esperimento ripetuto più volte nelle medesime condizioni, le v.a. associate a ripetizioni distinte (o gruppi disgiunti di ripetizioni distinte) sono indipendenti. A questo proposito, riprendiamo il nostro esempio paradigmatico:

■ **Esempio 5.5** Consideriamo n prove ripetute, in cui ciascuna prova abbia come esiti il successo con probabilità p e l'insuccesso con probabilità $1 - p$. Abbiamo visto che l'esperimento è descritto da

$$\Omega = \{0, 1\}^n, \quad \mathcal{F} = \mathcal{P}(\Omega), \quad \mathbb{P}\{a\} = p^{\sum_{i=1}^n a_i} (1 - p)^{n - \sum_{i=1}^n a_i}.$$

Per $i = 1, \dots, n$, sia X_i la v.a. che rappresenta l'esito dell' i -simo esperimento, cioè $X_i(a) = a_i$. Le v.a. X_i sono tutte indipendenti, come si può dimostrare osservando che $\{X_i = 1\} = A_i$, l'evento "successo all' i -sima prova". Osserviamo che le X_i hanno tutte la stessa legge di Bernoulli di parametro p . Osserviamo inoltre che $X = X_1 + \dots + X_n$ conta il numero di successi in n lanci ed è quindi v.a. binomiale; in altre parole, la somma di n v.a. di Bernoulli di parametro p indipendenti è una v.a. binomiale di parametri n e p .

Supponendo n pari, le v.a. $Y_1 =$ numero di teste nei primi $n/2$ lanci, $Y_2 =$ numero di teste negli ultimi $n/2$ lanci, Y_1 e Y_2 sono indipendenti, in quanto $Y_1 = X_1 + \dots + X_{n/2}$, $Y_2 = X_{n/2+1} + \dots + X_n$ sono funzioni di gruppi disgiunti di v.a. indipendenti. ■

Un altro esempio, rilevante per la statistica, è il seguente:

■ **Esempio 5.6** Sia S una popolazione e sia $X : S \rightarrow \mathbb{R}$ una v.a. che rappresenta un carattere (ad esempio $S =$ popolazione italiana e $X(\omega) =$ numero di figli di ω). Consideriamo l'estrazione di un campione di n elementi da una popolazione, sia X_i l'esito della misurazione del carattere per l' i -simo elemento. Supponiamo che ogni elemento del campione sia scelto in modo casuale tra la popolazione ed indipendentemente dalla scelta degli altri elementi. L'estrazione si può così modellizzare:

$$\Omega = S^n, \quad \mathbb{P} \text{ uniforme su } \Omega,$$

(ciascun elemento $(\omega_1, \dots, \omega_n)$ è un possibile campione). Il carattere X_i misurato sull' i -simo elemento è dato dalla v.a.

$$X_i(\omega_1, \dots, \omega_n) = X(\omega_i).$$

Si può dimostrare che le X_i sono indipendenti e hanno tutte la stessa legge, che è la legge di X . ■

Data una variabile doppia (X, Y) e una funzione $h : \mathbb{R}^2 \rightarrow \mathbb{R}$, è in generale difficile calcolare la legge della variabile aleatoria $h(X, Y)$. In alcuni casi specifici possiamo però procedere con un calcolo esplicito adatto alla situazione: vediamo un esempio.

Proposizione 5.2.2 Se X e Y sono rispettivamente Binomiali $B(n, p)$ e $B(m, p)$ e sono indipendenti, $Z = X + Y$ è binomiale $B(n + m, p)$.

Dimostrazione. Supponiamo dapprima che Y sia di Bernoulli : in tal caso

$$\begin{aligned} \mathbb{P}\{Z = h\} &= \mathbb{P}\{X = h, Y = 0\} + \mathbb{P}\{X = h - 1, Y = 1\} = \\ &= \binom{n}{h} p^h (1 - p)^{n-h} (1 - p) + \binom{n}{h-1} p^{h-1} (1 - p)^{n-h+1} p = \binom{n+1}{h} p^h (1 - p)^{n+1-h}, \end{aligned}$$

ovvero Z è di tipo $B(n + 1, p)$. La dimostrazione si completa poi per induzione su m . ■

Questo risultato suggerisce una formula generale nel caso in cui le variabili X, Y di cui consideriamo la somma sono indipendenti.

Proposizione 5.2.3 Supponiamo che $X, Y : \Omega \rightarrow \mathbb{N}$ siano variabili discrete a valori naturali e indipendenti, e sia $Z = X + Y$: dette p_X, p_Y e p_Z le funzioni di massa rispettivamente di X, Y e Z , si ha

$$p_Z(n) = \sum_{h=0}^n p_X(h) \cdot p_Y(n-h).$$

Dimostrazione. Questa formula è una conseguenza immediata dell'eguaglianza insiemistica

$$\{Z = n\} = \bigcup_{h=0}^n \{X = h, Y = n - h\}$$

e di conseguenza

$$p_Z(n) = \mathbb{P}\{Z = n\} = \sum_{h=0}^n \mathbb{P}\{X = h\} \mathbb{P}\{Y = n - h\} = \sum_{h=0}^n p_X(h) p_Y(n - h),$$

in cui il secondo passaggio fa uso essenziale dell'indipendenza. ■

Esercizio 5.4 Mostrare che se X è di Poisson di parametro λ , Y di Poisson di parametro μ , e X, Y sono indipendenti, allora $X + Y$ è di Poisson di parametro $(\lambda + \mu)$. ■

Vale una analoga formula nel caso di variabili con densità.

Proposizione 5.2.4 — Formula della convoluzione. Siano X e Y indipendenti, con densità rispettive f_X e f_Y , e sia $Z = X + Y$: la variabile Z ha densità data da

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx = \int_{-\infty}^{+\infty} f_Y(y) f_X(z-y) dy.$$

Dimostrazione. Osserviamo che vale

$$A_z = \{(x, y) \in \mathbb{R}^2 : x + y \leq z\} = \{(x, y) \in \mathbb{R}^2 : x \leq z - y\},$$

e dunque la funzione di ripartizione di Z è data da (usando la sostituzione $x' = x + y$ nell'ultimo passaggio)

$$\begin{aligned} F_Z(z) &= \mathbb{P}(X + Y \leq z) = \iint_{A_z} f_X(x) f_Y(y) dx dy = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{z-y} f_X(x) dx \right) f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^z f_X(x' - y) dx' \right) f_Y(y) dy \end{aligned}$$

da cui la tesi discende immediatamente derivando in z sotto il segno di integrale (o applicando il teorema fondamentale del calcolo volendo procedere in modo più elementare). ■

Usando quest'ultimo risultato (e svolgendo pazientemente i calcoli) si ottiene il seguente fatto, che incoraggiamo il lettore a dimostrare.

Esercizio 5.5 Se X e Y sono indipendenti e Gaussiane rispettivamente $N(m_1, \sigma_1^2)$ e $N(m_2, \sigma_2^2)$, allora $Z = X + Y$ è Gaussiana $N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$. ■

5.3 Covarianza e Correlazione

Proposizione 5.3.1 Supponiamo che X ed Y abbiano valore atteso, allora $X + Y$ ha valore atteso e valgono le seguenti proprietà:

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$;
- se $X \geq Y$, allora $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

Dimostrazione. Riportiamo la dimostrazione della prima proprietà solo nel caso delle variabili discrete. Anzitutto $X + Y$ ha valore atteso se $\sum_{x_i, y_j} |x_i + y_j| p(x_i, y_j) < +\infty$. Ma

$$\begin{aligned} \sum_{x_i, y_j} |x_i + y_j| p(x_i, y_j) &\leq \sum_{x_i, y_j} (|x_i| + |y_j|) p(x_i, y_j) = \sum_{x_i} |x_i| \sum_{y_j} p(x_i, y_j) + \sum_{y_j} |y_j| \sum_{x_i} p(x_i, y_j) \\ &= \sum_{x_i} |x_i| p_X(x_i) + \sum_{y_j} |y_j| p_Y(y_j) < +\infty. \end{aligned}$$

Poiché le serie convergono assolutamente, possiamo ripetere i passaggi togliendo i valori assoluti (la diseuguaglianza diventa eguaglianza) trovando esattamente $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

La seconda proprietà si mostra considerando la variabile $X - Y$, e applicando una proprietà del valore atteso vista nel capitolo precedente. ■

In generale, se X e Y hanno valore atteso, non è detto che il prodotto $X \cdot Y$ lo abbia, ma c'è un caso particolare nel quale siamo sicuri che esista $\mathbb{E}[XY]$ e si ha anche una regola del prodotto.

Proposizione 5.3.2 Supponiamo che X e Y abbiano valore atteso e che siano indipendenti: anche XY ha valore atteso e vale la formula

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Dimostrazione. Di nuovo ci limitiamo al caso discreto: vale

$$\begin{aligned} \sum_{x_i, y_j} |x_i y_j| p(x_i, y_j) &= \sum_{x_i, y_j} |x_i| |y_j| p_X(x_i) p_Y(y_j) \\ &= \left(\sum_{x_i} |x_i| p_X(x_i) \right) \cdot \left(\sum_{y_j} |y_j| p_Y(y_j) \right) < +\infty. \end{aligned}$$

Abbiamo usato la proprietà distributiva del prodotto rispetto alla somma, che vale anche per una serie di termini positivi (o se la serie è assolutamente convergente), ed è così provato che XY ha valore atteso. Ripetendo le stesse eguaglianze senza valore assoluto, si ottiene la formula $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$. ■

Se X e Y sono v.a. indipendenti, allora, per tutte le funzioni $h, k : \mathbb{R} \rightarrow \mathbb{R}$, sono anche $h(X)$ e $k(Y)$ sono indipendenti, in particolare, se $h(X)$ e $k(Y)$ hanno valore atteso, vale

$$\mathbb{E}[h(X)k(Y)] = \mathbb{E}[h(X)] \cdot \mathbb{E}[k(Y)].$$

Esercizio 5.6 Lanciamo due volte un dado equilibrato. Calcolare il valore atteso della somma degli esiti e il valore atteso del prodotto degli esiti. ■

Proposizione 5.3.3 — Disuguaglianza di Schwartz. Siano X e Y due variabili aleatorie: vale la disuguaglianza

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]} \cdot \sqrt{\mathbb{E}[Y^2]},$$

(in cui, se una delle due variabili non ha momento secondo, il membro destro è infinito).

Il principio della dimostrazione è simile alla disuguaglianza classica di Schwartz su \mathbb{R}^n , sorvoliamo sui dettagli ma è importante sottolineare una conseguenza. Abbiamo detto che se X e Y hanno valore atteso non è detto che il prodotto XY abbia valore atteso (a parte il caso in cui sono indipendenti), tuttavia se X e Y hanno momento secondo, il prodotto XY ha valore atteso.

Consideriamo ora due variabili aleatorie (con momento secondo) X e Y e introduciamo la covarianza e il coefficiente di correlazione, come misura della dipendenza lineare tra X e Y .

Definizione 5.3.1 Si chiama covarianza tra X e Y il numero

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

Se $\text{Var}(X) \neq 0$ e $\text{Var}(Y) \neq 0$, si chiama coefficiente di correlazione il numero

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Quando $\text{Cov}(X, Y) = 0$, quindi $\rho(X, Y) = 0$, le variabili sono dette scorrelate.

La covarianza soddisfa le proprietà sotto elencate, che è molto facile verificare direttamente.

Esercizio 5.7 Usando la definizione di covarianza si mostri che:

- $\text{Cov}(aX + bY + c, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$;
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ $\text{Var}(X) = \text{Cov}(X, X)$;
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$;
- sono equivalenti le seguenti tre proprietà:
 1. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$,
 2. $\text{Cov}(X, Y) = 0$,
 3. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$;

Proposizione 5.3.4 Date due variabili aleatorie X, Y valgono:

- $|\rho(X, Y)| \leq 1$;
- $\min_{a, b \in \mathbb{R}^2} \mathbb{E}[(Y - a - bX)^2] = \text{Var}(Y) \cdot (1 - \rho(X, Y)^2)$.

Dimostrazione. La prima proprietà segue dalla disuguaglianza di Schwarz,

$$|\text{Cov}(X, Y)| \leq \mathbb{E}[|X - \mathbb{E}[X]| |Y - \mathbb{E}[Y]|] \leq \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]} \sqrt{\mathbb{E}[(Y - \mathbb{E}[Y])^2]} = \sigma(X)\sigma(Y).$$

La seconda proprietà invece richiede un procedimento sostanzialmente identico a quello usato nel primo capitolo per trovare la retta di regressione, omettiamo i dettagli. ■

Osserviamo che, presi a_*, b_* che realizzano il minimo $\min_{a, b \in \mathbb{R}^2} \mathbb{E}[(Y - a - bX)^2]$ della precedente proposizione, la retta $y = a_* + b_*x$ è la migliore approssimazione lineare tra X e Y . Poiché il valore del minimo è proporzionale a $1 - \rho^2$, la relazione tra X e Y è tanto meglio approssimata da una retta, quanto più $|\rho|$ è vicino a 1: in questo senso, il coefficiente di correlazione è una misura della dipendenza *lineare* tra X e Y .

Esercizio 5.8 Due variabili indipendenti sono scorrelate, ma *non* è vero il viceversa: si usi come controesempio una coppia di variabili X, Y in cui X è di tipo uniforme sull'intervallo $[0, 1]$ e $Y = X^2$. ■

NB È importante sottolineare che la non-correlazione non implica l'indipendenza: come per i dati bivariati, e come si vede dal suggerimento dato nell'esercizio, possono esistere altre forme di dipendenza non lineare tra X e Y .

■ **Esempio 5.7** Sia S una popolazione e siano $X, Y : S \rightarrow \mathbb{R}$ due v.a. discrete che rappresentano due caratteri (ad esempio S = popolazione italiana, $X(\omega)$ = numero di figli di ω , $Y(\omega)$ = numero di case di proprietà). Come per il valore atteso, si può mostrare che il coefficiente di correlazione $\rho(X, Y)$ è il coefficiente di correlazione empirico tra i due caratteri *su tutta la popolazione*. ■

5.4 Teoremi Limite in Probabilità

Lanciamo 3 volte una moneta equilibrata: non possiamo fare previsioni precise, nemmeno con probabilità alta, su quale sarà il numero di teste (tra 0 e 3). Lanciamo ora la stessa moneta 1000 volte: benché ancora il numero di teste non possa essere determinato con precisione, è esperienza comune attendersi che esso si aggiri intorno a 500; in altre parole, la frequenza relativa dell'evento "testa" è vicina a $1/2$. I teoremi limite si occupano proprio di verificare e quantificare questo tipo di fenomeni. Precisamente, nell'esempio precedente, la legge dei grandi numeri afferma che, su un grande numero di lanci, la frequenza dell'evento "testa" è vicina a $1/2$, mentre il teorema centrale del limite quantifica quanto possa oscillare il numero di teste attorno al valore atteso 500.

I teoremi limite che vedremo qui sono relativi a variabili aleatorie i.i.d., come dalla seguente definizione:

Definizione 5.4.1 Consideriamo una famiglia di variabili aleatorie X_1, \dots, X_n, \dots possibilmente infinita. Scriviamo per brevità che esse sono *i.i.d.* (da *independent and identically distributed*) se sono indipendenti e equidistribuite (nel caso di famiglia infinita, diciamo che X_1, \dots, X_n, \dots sono indipendenti se, per ogni n , X_1, \dots, X_n sono indipendenti). Equivalentemente, X_1, \dots, X_n, \dots sono i.i.d. se hanno tutte la stessa funzione di ripartizione,

$$\mathbb{P}_{X_n}(t) = \mathbb{P}(X_n \leq t) = F(t), \quad t \in \mathbb{R}, n = 1, 2, \dots,$$

e se vale

$$\mathbb{P}(X_{k_1} \leq t_1, \dots, X_{k_n} \leq t_n) = F_{X_1}(t_1) \cdots F_{X_n}(t_n)$$

per ogni scelta di indici $k_1, \dots, k_n \in \mathbb{N}$ e valori $t_1, \dots, t_n \in \mathbb{R}$.

Il tipico esempio di v.a. i.i.d. è legato alle ripetizioni di un esperimento:

■ **Esempio 5.8** Consideriamo n (o anche infinite) ripetizioni di un esperimento, nelle stesse condizioni. Sia X la v.a. che descrive un carattere dell'esperimento, cioè una caratteristica quantitativa dell'esito dell'esperimento (ad esempio, nel lancio di una moneta, $X = 1$ se esce testa, $= 0$ altrimenti). Per ogni i intero positivo, sia X_i il valore del carattere dell'esito dell' i -simo esperimento (ad esempio $X_i = 1$ se esce testa all' i -simo lancio, $= 0$ altrimenti). Allora le X_i sono indipendenti (l'esito di una ripetizione esperimento non influenza le altre ripetizioni) ed equidistribuite, con la stessa distribuzione di X . ■

Come caso particolare del precedente esperimento, ritroviamo l'estrazione di un campione da una popolazione:

■ **Esempio 5.9** Consideriamo una popolazione, sia X la v.a. che rappresenta un carattere degli individui della popolazione. Supponiamo di estrarre un campione dalla popolazione, in questo modo: estraiamo casualmente n individui, in modo indipendente l'uno dall'altro, e chiamiamo X_i il carattere dell' i -simo individuo estratto (stiamo cioè ripetendo n volte l'esperimento di estrazione dalla popolazione). Allora gli X_i sono indipendenti e hanno la stessa distribuzione di X . ■

Introduciamo ora la media aritmetica di n v.a. i.i.d., che studieremo nei teoremi limite:

$$\bar{X} := \bar{X}_n := \frac{X_1 + \dots + X_n}{n}.$$

Notiamo che \bar{X} è essa stessa una variabile aleatoria. Inoltre, se X_1, \dots, X_n rappresenta un campione, \bar{X} rappresenta la media campionaria di tale campione; infatti, come vedremo, \bar{X} gioca un ruolo importante nella statistica inferenziale.

■ **Esempio 5.10** Consideriamo n ripetizioni di un esperimento, con esito successo, che chiamiamo A e che ha probabilità p , o insuccesso. Sia X_i la v.a. di Bernoulli associata all' i -sima ripetizione, cioè $X_i = 1$ se si verifica A all' i -sima ripetizione, $= 0$ altrimenti. Allora $X_1 + \dots + X_n$, v.a. binomiale di parametri n e p , rappresenta la frequenza assoluta del successo, mentre \bar{X} rappresenta la frequenza relativa del successo. ■

I teoremi limite studiano limiti delle v.a.. È quindi necessario dapprima capire che cosa intendiamo per limite di una successione di v.a.. La legge dei grandi numeri riguarda il limite in probabilità, che definiamo ora:

Definizione 5.4.2 Date variabili aleatorie X e X_1, X_2, \dots sullo stesso spazio di probabilità, si dice che X_n converge in probabilità a X per $n \rightarrow \infty$ se vale:

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

Il limite può anche essere una costante $X = c$.

Intuitivamente, la successione $(X_n)_n$ tende in probabilità a X se, per n grande, X_n è vicina a X con probabilità alta. Non ci dilunghiamo su questa nozione di convergenza, e ci limitiamo al seguente risultato che offre una condizione sufficiente per la convergenza in probabilità a una costante.

Proposizione 5.4.1 Supponiamo che si abbia

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = c \in \mathbb{R}, \quad \lim_{n \rightarrow \infty} \text{Var}(X_n) = 0,$$

allora la successione $(X_n)_{n \geq 1}$ converge in probabilità alla costante c .

Dimostrazione. Dalla disuguaglianza di Markov applicata alla variabile $(X_n - c)^2$ si ha

$$\mathbb{P}(|X_n - c| > \varepsilon) \leq \frac{1}{\varepsilon^2} \mathbb{E}[(X_n - c)^2],$$

di cui espandiamo il membro destro come segue:

$$\begin{aligned} \mathbb{E}[(X_n - c)^2] &= \mathbb{E}[(X_n - \mathbb{E}[X_n] + \mathbb{E}[X_n] - c)^2] \\ &= \mathbb{E}[(X_n - \mathbb{E}[X_n])^2] + 2\mathbb{E}[X_n - \mathbb{E}[X_n]] \cdot (\mathbb{E}[X_n] - c) + (\mathbb{E}[X_n] - c)^2 \\ &= \text{Var}(X_n) + (\mathbb{E}[X_n] - c)^2. \end{aligned}$$

Di conseguenza $\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - c)^2] = 0$, il che conclude la dimostrazione. ■

Teorema 5.4.2 — Legge debole dei Grandi Numeri. Sia X_1, X_2, \dots una successione di variabili i.i.d. dotate di momento secondo finito, e sia $\mu = \mathbb{E}[X_i]$ il loro valore atteso. Allora \bar{X}_n converge in probabilità a μ per $n \rightarrow \infty$, cioè, per ogni $\varepsilon > 0$ si ha

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

In linguaggio discorsivo si può dire che la media campionaria dei primi n termini converge alla media teorica (che è la media sulla popolazione in caso di estrazione).

Dimostrazione. Posto $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$, si ha

$$\begin{aligned} \mathbb{E}[\bar{X}_n] &= \frac{\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]}{n} = \mu, \\ \text{Var}(\bar{X}_n) &= \frac{\text{Var}(X_1) + \dots + \text{Var}(X_n)}{n^2} = \frac{\text{Var}(X_1)}{n} \xrightarrow{n \rightarrow \infty} 0, \end{aligned}$$

dove abbiamo usato $\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n)$ per l'indipendenza delle X_i . La tesi segue grazie alla Proposizione precedente. ■

In realtà le ipotesi possono essere indebolite: è sufficiente che le variabili siano scorrelate, che abbiano tutte valore atteso μ e varianze equilimitate (cioè esiste C tale che $\text{Var}(X_i) \leq C$ per ogni i).

Il fatto che si chiami legge debole dei grandi numeri fa pensare che esistano delle versioni più generali: in effetti è così ma questo va al di là degli scopi di questo corso.

Una conseguenza particolarmente importante della L.G.N. è che, per un grande numero di prove ripetute, la frequenza relativa di un evento approssima la probabilità teorica dell'evento stesso:

■ **Esempio 5.11** Riprendiamo l'esempio 5.10 di n ripetizioni di un esperimento di Bernoulli. La L.G.N. ci dice che, per $n \rightarrow \infty$, la frequenza relativa \bar{X} di un evento A tende in probabilità al valore atteso $\mathbb{E}[X_1] = p$, cioè alla probabilità teorica di A . Ad esempio, la frequenza relativa di “testa” su n lanci di moneta tende in probabilità a $1/2$ per $n \rightarrow \infty$. ■

In particolare, ritroviamo qui, in modo rigoroso, il limite delle frequenze relative che era postulato nella definizione frequentista della probabilità.

Esercizio 5.9 Lanciamo un dado 1000 volte. Fornire un valore approssimato per il numero di 6 nei 1000 lanci. ■

Una conseguenza della legge dei grandi numeri è il comportamento, per n grande, della seguente v.a.:

$$S_n^2 := \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}.$$

Notiamo che, se (X_1, \dots, X_n) rappresenta un campione, S^2 rappresenta la varianza campionaria di tale campione.

Proposizione 5.4.3 Sia X_1, X_2, \dots una successione di variabili i.i.d. dotate di momento quarto, e sia $\sigma^2 = \text{Var}(X_i)$ la loro varianza. Per $n \rightarrow \infty$, S_n converge in probabilità a σ , ovvero

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} \mathbb{P}(|S_n^2 - \sigma^2| > \varepsilon) = 0.$$

Dimostrazione. Premettiamo questo fatto (senza dimostrazione), che useremo nel seguito: se $(Y_n)_n$ tende a Y in probabilità e $(Z_n)_n$ tende a Z in probabilità, allora $(Y_n Z_n)_n$, rispettivamente $(Y_n + Z_n)_n$, tende in probabilità a YZ , rispettivamente $Y + Z$. Sviluppando i quadrati nella definizione di S_n^2 , si vede che

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

Per $n \rightarrow \infty$, $n/(n-1)$ tende a 1. Poiché le X_i sono i.i.d. (con momento quarto), le X_i^2 sono i.i.d. (con momento secondo), quindi per la L.G.N. abbiamo $\frac{1}{n} \sum_{i=1}^n X_i^2 \rightarrow \mathbb{E}[X_1^2]$, dove la convergenza è in probabilità. Sempre per la L.G.N. abbiamo $\bar{X}_n \rightarrow \mathbb{E}[X_1]$ e quindi $\bar{X}_n^2 \rightarrow \mathbb{E}[X_1]^2$, sempre con convergenza in probabilità. Quindi vale la convergenza in probabilità

$$S_n^2 \rightarrow \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \sigma^2.$$

■

Per il teorema centrale del limite, useremo la convergenza in distribuzione. La definizione di convergenza in distribuzione non è affatto semplice, si può esprimere facilmente se la variabile limite X ha funzione di ripartizione continua.

Definizione 5.4.3 Siano $(X_n)_{n \geq 1}$ un successione di v.a. ed X una variabile aleatoria, siano rispettivamente F_n ed F le funzioni di ripartizione di X_n ed X e supponiamo che $F(\cdot)$ sia continua: si dice che la successione $(X_n)_{n \geq 1}$ converge ad X in distribuzione se per ogni t si ha

$$\lim_{n \rightarrow \infty} F_n(t) = F(t).$$

Nel caso generale questa definizione deve essere modificata, ma a noi interessa in realtà il caso in cui la variabile limite è Gaussiana oppure di Student (quindi con funzione di ripartizione continua). Si noti che la convergenza in distribuzione riguarda solo le leggi delle variabili aleatorie coinvolte.

L'enunciato del Teorema Centrale del Limite¹ è il seguente:

Teorema 5.4.4 — Teorema Centrale Del Limite. Sia X_1, X_2, \dots una successione di variabili i.i.d. dotate di momento secondo finito, con valore atteso $\mathbb{E}[X_i] = \mu$ e varianza $\sigma^2(X_i) = \sigma^2 > 0$.

¹Il nome più comune dato in italiano (e anche in francese) a questo risultato è “teorema del limite centrale”. In realtà tale nome sembra essere frutto di un'errata traduzione, dal tedesco “zentraler Grenzwertsatz”, termine introdotto da George Pólya: l'aggettivo “centrale” (zentraler) si riferisce al teorema (Satz), non al limite (Grenzwert).

Presi $-\infty \leq a < b \leq +\infty$ si ha

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(a \leq \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx = \Phi(b) - \Phi(a)$$

equivalentemente, $\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$ converge in distribuzione a una v.a. Gaussiana standard.

Agli effetti pratici, possiamo dire che nelle ipotesi del teorema 4.2.2 la variabile

$$\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$$

quando n è grande, è approssimativamente Gaussiana standard. Naturalmente n grande non è un termine preciso: deve essere almeno $n \geq 50$, ma per una buona approssimazione almeno $n \geq 80$.

Questo teorema è stato dimostrato da Laplace, con contributi precedenti di De Moivre, limitatamente al caso in cui le X_i sono di Bernoulli di parametro p , e dunque $X_1 + \dots + X_n$ risulta Binomiale di parametri n e p . In quel caso la dimostrazione segue dalla Formula di Stirling,

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e} \right)^n.$$

La versione più generale del teorema che abbiamo enunciato è dovuta a Paul Lévy, e la sua dimostrazione richiede strumenti più avanzati, ma chiarisce bene il posto assolutamente centrale che le variabili Gaussiane ricoprono in statistica.

Uno degli usi più frequenti è l'approssimazione di variabili Binomiali con n grande: poiché possiamo pensare una variabile X Binomiale di parametri n e p come somma di n variabili indipendenti di Bernoulli di parametro p , la variabile $\frac{X - np}{\sqrt{np(1-p)}}$ è approssimativamente $N(0, 1)$. In questo caso poi ci sono valutazioni più precise della numerosità n : perché l'approssimazione sia buona, si deve avere $np(1-p) \geq 15$, perché sia ottima si deve avere $np(1-p) \geq 20$

■ **Esempio 5.12** Gli aerei della compagnia Airfly hanno 180 posti, ma la compagnia sa che in media si presentano all'imbarco solo nove passeggeri su dieci, e per questo mette in vendita 195 biglietti (*overbooking*). Ci domandiamo qual è la probabilità che qualche cliente rimanga a terra e di conseguenza la compagnia sia costretta a un risarcimento.

Il numero aleatorio X di clienti che si presentano è una variabile Binomiale di parametri 195 e 0.9, e bisogna calcolare $\mathbb{P}(X \geq 181)$: in base al Teorema Centrale del Limite, la variabile

$$\frac{X - 195 \times 0.9}{\sqrt{195 \times 0.1 \times 0.9}} = \frac{X - 175.5}{4.18}$$

è approssimativamente Gaussiana standard. Di conseguenza

$$\begin{aligned} \mathbb{P}(X \geq 181) &= \mathbb{P} \left(\frac{X - 175.5}{4.18} \geq \frac{181 - 175.5}{4.18} \right) = \mathbb{P} \left(\frac{X - 175.5}{4.18} \geq 1.31 \right) \\ &\approx 1 - \Phi(1.31) \approx 0.095 \end{aligned}$$

è una buona approssimazione della risposta. ■

Esercizio 5.10 Calcolare in modo approssimato la probabilità che, su 1000 lanci di moneta equilibrata, ci siano almeno 480 teste. Calcolare poi, sempre in modo approssimato, il valore k tale che, con probabilità del 95%, testa compaia almeno k volte (su 1000). ■

Il seguente risultato (che si dimostra come conseguenza del T.C.L.) mostra come si possa rimpiazzare la varianza teorica σ^2 con la varianza campionaria S_n^2 :

Proposizione 5.4.5 Sia X_1, X_2, \dots una successione di variabili i.i.d. con valore atteso $\mathbb{E}[X_i] = \mu$ e varianza $\text{Var}(X_i) = \sigma^2 > 0$. Presi $-\infty \leq a < b \leq +\infty$ si ha

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(a \leq \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx = \Phi(b) - \Phi(a)$$

5.5 Variabili Chi-Quadro e di Student

Raccogliamo in quest'ultima sezione alcuni esempi di variabili aleatorie che emergono in modo naturale nello studio di famiglie di variabili Gaussiane indipendenti.

Cominciamo col definire una delle cosiddette *funzioni speciali*: la funzione Gamma di Eulero,

$$\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx, \quad r > 0.$$

Per un generico r questo integrale non ha una rappresentazione in termini di funzioni elementari. Se $r > 1$ vale la relazione $\Gamma(r) = (r-1)\Gamma(r-1)$: infatti integrando per parti si ottiene

$$\int_0^{+\infty} x^{r-1} e^{-x} dx = -x^{r-1} e^{-x} \Big|_0^{+\infty} + (r-1) \int_0^{+\infty} x^{r-2} e^{-x} dx.$$

Poiché $\Gamma(1) = 1$, ne segue che per n intero positivo si ha $\Gamma(n) = (n-1)!$ e dunque la funzione Gamma può essere vista come una estensione a numeri non interi del fattoriale.

Definizione 5.5.1 — Densità Gamma. Si chiama densità Gamma di parametri $r, \lambda > 0$, indicata con $\Gamma(r, \lambda)$, la funzione così definita:

$$f(x) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x} & x > 0, \\ 0 & x \leq 0. \end{cases}$$

Si tratta effettivamente di una densità, come si vede con un cambio di variabili:

$$\int_0^{+\infty} (\lambda x)^{r-1} e^{-\lambda x} \lambda dx = \int_0^{+\infty} t^{r-1} e^{-t} dt = \Gamma(r), \quad (\lambda x = t).$$

Si può osservare che la densità esponenziale di parametro λ corrisponde alla densità $\Gamma(1, \lambda)$.

Proposizione 5.5.1 Una variabile X con densità $\Gamma(r, \lambda)$ ha tutti i momenti, e per ogni $\beta > 0$ vale

$$\mathbb{E}[X^\beta] = \frac{\Gamma(r+\beta)}{\Gamma(r)\lambda^\beta}.$$

In particolare,

$$\mathbb{E}[X] = \frac{\Gamma(r+1)}{\Gamma(r)\lambda} = \frac{r}{\lambda}, \quad \mathbb{E}[X^2] = \frac{\Gamma(r+2)}{\Gamma(r)\lambda^2} = \frac{(r+1)r}{\lambda^2}, \quad \text{Var}(X) = \frac{r}{\lambda^2}.$$

Dimostrazione. Poiché X prende solo valori positivi, vale

$$\mathbb{E}[X^\beta] = \frac{1}{\Gamma(r)} \int_0^{+\infty} x^\beta \lambda^r x^{r-1} e^{-\lambda x} dx = \frac{1}{\Gamma(r)\lambda^\beta} \int_0^{+\infty} \lambda^{r+\beta} x^{r+\beta-1} e^{-\lambda x} dx = \frac{\Gamma(r+\beta)}{\Gamma(r)\lambda^\beta},$$

da cui segue l'esistenza di tutti i momenti e la formula proposta. Si noti che non abbiamo calcolato i valori di $\Gamma(r)$, ma solamente usato l'eguaglianza $\Gamma(r) = (r-1)\Gamma(r-1)$. ■

NB

Ci si riferisce a r, λ rispettivamente come *parametri di forma e tasso di decadimento (shape e decay rate)*. Questa è la parametrizzazione per softwares quali **R**. Alcuni autori preferiscono mantenere il parametro *shape* r , ma usare come secondo parametro $s = 1/\lambda$, detto *di scala (scale)*. Con questa parametrizzazione la densità diventa

$$f(x) = \frac{1}{\Gamma(r)} \frac{1}{s} \left(\frac{x}{s}\right)^{r-1} e^{-\frac{x}{s}}, \quad x > 0$$

(e nulla per $x \leq 0$).

Proposizione 5.5.2 Date due variabili indipendenti X, Y con densità rispettivamente $\Gamma(r, \lambda)$ e $\Gamma(s, \lambda)$ (con **lo stesso** parametro λ), allora la variabile $(X + Y)$ ha densità $\Gamma(r + s, \lambda)$.

Dimostrazione. Applichiamo la formula della convoluzione vista nella Proposizione 5.2.4: se $Z = X + Y$ si ha

$$f_Z(z) = \frac{\lambda^{r+s} e^{-\lambda z}}{\Gamma(r)\Gamma(s)} \int_0^z x^{r-1} (z-x)^{s-1} dx = \frac{\lambda^{r+s} z^{r+s-1} e^{-\lambda z}}{\Gamma(r)\Gamma(s)} \int_0^1 y^{r-1} (1-y)^{s-1} dy$$

in cui il primo passaggio è una semplice sostituzione della definizione di densità Gamma nella formula di convoluzione, e il secondo segue dalla sostituzione $x = zy$. L'integrale che appare a destra è una funzione dei parametri r, s detta *funzione Beta*, e vale:

$$\begin{aligned} \Gamma(r)\Gamma(s) &= \left(\int_0^\infty x^{r-1} e^{-x} dx \right) \cdot \left(\int_0^\infty y^{s-1} e^{-y} dy \right) = \int_0^\infty \int_0^\infty x^{r-1} y^{s-1} e^{-x-y} dx dy \\ &= \int_0^\infty \int_0^1 e^{-a} (ab)^{r-1} a^{s-1} (1-b)^{s-1} a db da \\ &= \left(\int_0^\infty e^{-a} a^{r+s-1} da \right) \cdot \int_0^1 b^{r-1} (1-b)^{s-1} db = \Gamma(r+s) \int_0^1 b^{r-1} (1-b)^{s-1} db, \end{aligned}$$

in cui per passare alla seconda riga abbiamo usato il cambio di variabili $(x, y) = (ab, a(1-b))$. Da questo calcolo, dividendo per $\Gamma(r)\Gamma(s)$, segue la tesi. ■

Esercizio 5.11 Scegliendo nell'ultimo calcolo della dimostrazione $r = s = 1/2$, e grazie a

$$\int_0^1 \frac{dx}{\sqrt{x(1-x)}} = \pi$$

(la si dimostri) si ottenga il valore speciale $\Gamma(1/2) = \sqrt{\pi}$. ■

Proposizione 5.5.3 — Densità χ^2 (Chi-quadro). Siano X_1, \dots, X_n variabili Gaussiane standard indipendenti: la variabile $(X_1^2 + \dots + X_n^2)$ ha densità $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$. Tale densità è anche detta densità Chi-quadro con n gradi di libertà ($n \in \mathbb{N}_0$), e indicata con $\chi^2(n)$.

Dimostrazione. Alla luce della Proposizione 5.5.2 è sufficiente provare che, se X è Gaussiana standard, X^2 è $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$ (si procede poi per induzione su n). Sembra naturale cercare di applicare la formula di cambio di variabile della Proposizione 4.3.1 con $h: \mathbb{R} \rightarrow [0, \infty)$, $h(x) = x^2$, ma questa funzione *non soddisfa le ipotesi necessarie perchè non è bigettiva*.

Il problema si può aggirare, ma per il nostro caso specifico basta in effetti differenziare sotto il segno di integrale. Vale infatti, per $Z = X^2$,

$$F_Z(t) = \mathbb{P}(X^2 \leq t) = \begin{cases} 0 & t < 0, \\ \mathbb{P}(-\sqrt{t} \leq X \leq \sqrt{t}) = 2F_X(\sqrt{t}) - 1 & t \geq 0, \end{cases}$$

dunque la densità di probabilità si ricava calcolando $f_Z(t) = \frac{d}{dt}F_Z(t)$, ed essendo (per $t > 0$)

$$2 \frac{d}{dt} F_X(\sqrt{t}) = \frac{2}{\sqrt{2\pi}} \frac{d}{dt} \int_{-\infty}^{\sqrt{t}} e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \cdot \frac{e^{-t/2}}{2\sqrt{t}} = \frac{e^{-t/2}(t)^{-1/2}}{2^{1/2}\Gamma(1/2)},$$

dove abbiamo usato che $\Gamma(1/2) = \sqrt{\pi}$ (vedi l'esercizio sopra), si ha la tesi. ■

La terminologia gradi di libertà è per il momento misteriosa, diventerà chiara quando ne vedremo le applicazioni statistiche. Lo α -quantile della variabile chi-quadro (che comparirà diverse volte più avanti) è indicato $\chi^2_{\alpha,n}$. Poiché la variabile $\chi^2(n)$ è in realtà una $\Gamma(\frac{n}{2}, \frac{1}{2})$, non c'è bisogno di nuovi conti per calcolare i momenti di questa variabile, così come è evidente che somma di variabili chi-quadro indipendenti è ancora una variabile chi-quadro: per essere più precisi, se X ha densità $\chi^2(n)$, Y ha densità $\chi^2(m)$ e sono indipendenti, allora $(X + Y)$ ha densità $\chi^2(n + m)$.

Vediamo ora utili approssimazioni della variabile $\chi^2(n)$ quando n è grande (diciamo $n \geq 80$): partiamo dall'osservazione che, presa C_n con densità $\chi^2(n)$, si può pensare $C_n = X_1^2 + \dots + X_n^2$ dove le X_i sono Gaussiane standard indipendenti, inoltre $\mathbb{E}[X_i^2] = 1$ e $\text{Var}(X_i^2) = 2$. Valgono i seguenti risultati:

- per la legge dei Grandi Numeri la successione $\frac{C_n}{n}$ converge a 1 in probabilità (per $n \rightarrow \infty$), e quindi $\frac{C_n}{n} \approx 1$;
- per il teorema Centrale del Limite la successione $\frac{C_n - n}{\sqrt{2n}}$ converge in distribuzione alla variabile $N(0, 1)$, e quindi $\frac{C_n - n}{\sqrt{2n}}$ è approssimativamente Gaussiana standard.

Proposizione 5.5.4 — Densità di Student. Siano X, C_n due variabili indipendenti con densità rispettivamente $N(0, 1)$ e $\chi^2(n)$, $n \geq 1$. La variabile

$$T_n = \frac{X}{\sqrt{\frac{C_n}{n}}} = \sqrt{n} \frac{X}{\sqrt{C_n}} \quad \text{ha densità} \quad f_{T_n}(t) = \frac{\Gamma(n/2 + 1/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-n/2-1/2},$$

detta densità di Student a n gradi di libertà.

Questa definizione apparentemente misteriosa segue da una precisa applicazione statistica. Omettiamo la dimostrazione, che richiede calcoli un po' laboriosi di cui non c'è bisogno nelle applicazioni: bastano alcune proprietà elementari della funzione di ripartizione di T_n perchè ci affidiamo al software per valori approssimati dei quantili di queste variabili.

Proposizione 5.5.5 La densità di T_n è una funzione pari. Di conseguenza indicati con $F_n(x)$ e $\tau_{(\alpha,n)}$ rispettivamente la Funzione di ripartizione e lo α -quantile della variabile T_n , valgono

$$F_n(-x) = 1 - F_n(x), \quad \tau_{(\alpha,n)} = -\tau_{(1-\alpha,n)}.$$

La variabile T_n ha momenti fino all'ordine $(n - 1)$, e i momenti di ordine dispari, quando esistono, sono nulli.

In effetti le proprietà elencate dipendono solo dal fatto che f_{T_n} è pari, non dalla sua espressione specifica: lo lasciamo come facile esercizio.

Abbiamo visto poco fa che, se n è grande e C_n è una variabile $\chi^2(n)$ allora $\frac{C_n}{n} \approx 1$. Ci aspettiamo di conseguenza che T_n sia approssimativamente Gaussiana standard, come rende preciso il seguente:

Teorema 5.5.6 Consideriamo, per ogni n , una variabile T_n di Student a n gradi di libertà: la successione $(T_n)_{n \geq 1}$ converge in distribuzione alla variabile $N(0, 1)$.

Anche di questo fatto omettiamo la dimostrazione. Per quello che riguarda le applicazioni pratiche, quanto n è alto, nei conti la variabile di Student è sostituita dalla variabile Gaussiana

standard. Nelle applicazioni statistiche, della variabile di Student serve avere solo la funzione di ripartizione o il quantile, e questi usualmente vengono calcolati con un opportuno software; in passato esistevano le tavole della variabile di Student ma adesso sono superate.

Rispetto alla densità $N(0, 1)$, la densità di Student va a 0 più lentamente per $|x| \rightarrow \infty$ (si parla di "code pesanti" o "fat tails"); sono tracciate sotto, sovrapposte, la densità $N(0, 1)$ in grassetto e le densità di Student a 2 e a 4 gradi di libertà.

5.6 Esercizi

Esercizio 5.12 Data una variabile Gaussiana standard X (ovvero $N(0, 1)$) e una funzione $f: \mathbb{R} \rightarrow \mathbb{R}$ pari (ovvero $f(x) = f(-x)$) si mostri che le variabili aleatorie $X, f(X)$ sono scorrelate ma (ovviamente) non indipendenti.

Si consideri poi una coppia di variabili aleatorie X, Y con densità congiunta data da

$$f(x, y) = f_{(X, Y)}(x, y) = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{x^2 - 2rxy + y^2}{2(1-r^2)}}$$

con $r \in [-1, 1]$. Si mostri che le variabili X, Y sono entrambe Gaussiane $N(0, 1)$, e che le seguenti affermazioni sono equivalenti:

- X, Y sono scorrelate;
- X, Y sono indipendenti;
- $r = 0$.

(ovvero ognuna implica ambo le altre). ■

Esercizio 5.13 — Difficile. Si considerino due variabili X, Y indipendenti equidistribuite con legge uniforme su $[0, 1]$. Ricordando la formula di Leibniz per π ,

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1},$$

si mostri che la probabilità che l'intero più vicino a X/Y sia pari è data da $\frac{5-\pi}{4}$.

Suggerimento: Si rappresenti Ω come il quadrato $[0, 1]^2$, X, Y come le due coordinate, e si osservi che l'evento considerato è dato da una successione di sottoinsiemi triangolari di $[0, 1]^2$, di cui è sufficiente sommare le aree. ■

Esercizio 5.14 In una fabbrica, il numero di guasti in un giorno ha distribuzione di Poisson di parametro 1.5; inoltre supponiamo che il numero di guasti in un giorno non influenzi il numero di guasti in altri giorni.

- Qual è la probabilità che, in un dato giorno, ci siano almeno 2 guasti?
 - Qual è la probabilità che, in 2 giorni consecutivi, ci siano almeno 3 guasti?
 - Qual è la probabilità che, in una data settimana lavorativa (cioè 5 giorni), in almeno 4 giorni si verifichi al massimo un guasto?
-

Esercizio 5.15 Una ditta produce certi componenti elettronici dei quali circa il 20% sono difettosi: questi componenti sono esportati in scatole da 400 pezzi e la ditta si impegna a sostituire integralmente la scatola se il numero di pezzi difettosi è superiore a 90. (i) Qual è (approssimativamente) la probabilità che la ditta debba sostituire una scatola? (ii) Se si vuole

che la probabilità di dover sostituire una scatola sia inferiore a 0.05, come deve migliorare la produzione (cioè di quanto deve -approssimativamente- scendere la percentuale di pezzi difettosi)? ■

Esercizio 5.16 L'altezza degli italiani adulti segue una distribuzione Gaussiana di media 1.77 cm e deviazione standard 10.6cm (dati inventati).

- Qual è la probabilità che l'altezza di un italiano (adulto) sia maggiore di 180 cm?
- Qual è la probabilità che, scelti due italiani a caso, la media delle loro altezze sia maggiore di 180 cm?
- Qual è la probabilità che, scelti 100 italiani a caso, la media delle loro altezze sia maggiore di 180 cm?
- Come cambia la risposta al punto precedente, se la distribuzione dell'altezza ha sempre media 1.77 e deviazione standard 10.6 ma non è necessariamente Gaussiana? ■

Esercizio 5.17 Un server riceve email la cui dimensione ha valore atteso 4.73 MB e deviazione standard 0.53 MB. In un giorno il server riceve 70 email.

- Qual è la probabilità che la dimensione totale delle 70 email superi 340 MB?
- Quanto deve valere y in MB affinché, con probabilità del 95%, la dimensione totale delle email sia inferiore a y ? ■

Esercizio 5.18 La temperatura corporea in una persona sana ha distribuzione gaussiana di media 98.2 gradi Fahrenheit (F) e deviazione standard 0.62 gradi F. Un certo ospedale indica 100.6 come temperatura minima per la febbre.

- Quale percentuale di persone sane sono considerate con la febbre dall'ospedale?
- In quale intervallo centrale si colloca il 60% delle temperature delle persone sane?
- Se misuriamo la temperatura a 10 persone scelte a caso, con quale probabilità la loro temperatura media sarà superiore a 98.7 gradi F?
- Se misuriamo la temperatura a 10 persone scelte a caso, con quale probabilità almeno 6 di loro avranno temperatura superiore ai 98.5 gradi F? ■

Esercizio 5.19 Il numero giornaliero di clienti a uno sportello segue una distribuzione di Poisson di parametro 3. Supponiamo che il numero di clienti in un dato giorno sia indipendenti dal numero di clienti in altri giorni.

- Qual è la distribuzione del numero di clienti in una settimana lavorativa (= 5 giorni)?
- Stimare, tramite la disuguaglianza di Chebyshev, la probabilità che in una settimana ci siano almeno 20 clienti. ■

Esercizio 5.20 Il numero giornaliero di errori commessi da un server segue una distribuzione di Poisson di parametro 2.5. Supponiamo che il numero di errori commessi in un dato giorno sia indipendenti dal numero di errori in altri giorni.

- Qual è la probabilità che, in un giorno, avvengano almeno 3 errori?
- Qual è la probabilità che, in due dati giorni consecutivi, avvengano almeno 3 errori?
- Qual è la probabilità che, in un anno (= 365 giorni), avvengano almeno 900 errori?

- Qual è la probabilità che, in almeno 2 giorni di una data settimana, il server commetta almeno 3 errori?

Esercizio 5.21 Un test diagnostico per una certa malattia ha specificità del 97% (cioè è negativo su una data persona sana con probabilità del 97%) e sensibilità del 99% (cioè è positivo su una data persona malata con probabilità del 99%). Supponiamo che l'1% della popolazione soffra di questa malattia. Supponiamo inoltre che gli esiti di ripetizioni del test su una data persona siano indipendenti. Estraiamo una persona a caso ed eseguiamo su questa il test 2 volte.

- Se la persona è sana, qual è la probabilità che il test dia esito positivo in entrambe le esecuzioni?
- Se il test dà esito positivo in entrambe le esecuzioni, qual è la probabilità che la persona sia sana?
- Gli esiti delle due esecuzioni sono indipendenti?



Inferenza Statistica

6	Campioni di Variabili Aleatorie	83
6.1	Campioni Statistici e Stimatori	
6.2	Stima parametrica	
6.3	Statistiche campionarie di variabili Gaussiane	
6.4	Esercizi	
7	Intervalli di Fiducia	89
7.1	Definizione	
7.2	Intervalli di fiducia per la media di un campione Gaussiano	
7.3	Intervalli di fiducia per la media di un campione Bernoulli	
7.4	Intervalli di fiducia per la media di un campione di taglia grande	
7.5	Intervalli di fiducia per la varianza di un campione Gaussiano	
7.6	Esercizi	
8	Test Statistici	97
8.1	Concetti generali	
8.2	Z-test	
8.3	Test sulla media di un campione Gaussiano con varianza sconosciuta, o T-test	
8.4	Test approssimato su un campione di Bernoulli	
8.5	Test approssimato sulla media di un campione di taglia grande	
8.6	Test sulla varianza di un campione Gaussiano	
8.7	Confronto tra due campioni statistici	
8.8	Esercizi	

6. Campioni di Variabili Aleatorie

Lo scopo dell'inferenza statistica è l'analisi di un campione per ricavare informazioni su (un carattere di) una intera popolazione. Ad esempio, un sondaggio sulle intenzioni di voto raccoglie le risposte di un certo numero di intervistati per ricostruire gli orientamenti elettorali dell'intera popolazione, in questo caso l'intero corpo elettorale).

L'assunzione di base della statistica inferenziale è che si possa descrivere la misura del carattere desiderato su un individuo scelto casualmente con una variabile aleatoria, di cui vogliamo determinare la legge, o informazioni su di essa. La casualità risiede quindi nell'estrazione dell'individuo dalla popolazione, idea anticipata nell'Esempio 4.2, dunque le ipotesi che faremo sul modello probabilistico corrispondono ad assunzioni sul modo in cui il campione viene scelto.

Ipotizziamo quindi che:

- ogni nuova estrazione non è condizionata dalla precedente (ad esempio la popolazione deve essere molto grande, in modo che scegliere un individuo non cambi la statistica su quelli rimanenti): questo corrisponderà all'ipotesi di *indipendenza*;
- l'estrazione di ogni nuovo individuo per il campione sia effettuata in modo uguale alle precedenti: questo corrisponderà all'ipotesi di *equidistribuzione*.

In termini matematici, il nostro studio della statistica inizia quindi dall'analisi di famiglie finite di variabili aleatorie i.i.d., di cui vogliamo determinare informazioni sulla legge a partire dagli esiti.

6.1 Campioni Statistici e Stimatori

Assumiamo nel seguito di avere a disposizione uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ su cui sono definite tutte le variabili aleatorie che consideriamo.

Come anticipato sopra, il carattere che vogliamo studiare viene rappresentato con una variabile aleatoria X , la cui distribuzione \mathbb{P}_X è secondo i casi parzialmente o del tutto sconosciuta. Per determinarla, assumiamo di avere a disposizione n variabili aleatorie X_1, \dots, X_n , indipendenti e aventi la stessa legge di X , che rappresentano il campione estratto dalla popolazione.

Definizione 6.1.1 Data $F = F_X$ c.d.f. di una variabile aleatoria X , una famiglia finita X_1, \dots, X_n di variabili aleatorie i.i.d. con legge data dalla c.d.f. F si dice *campione statistico* o *campione*

I aleatorio della v.a. X di numerosità (o *taglia*) n .

NB È doveroso insistere ancora una volta sull'importanza di usare *lettere minuscole* x_1, x_2, \dots per indicare dati numerici (ad esempio risultati di alcune misurazioni) e *lettere maiuscole* X_1, X_2, \dots per indicare *variabili aleatorie*: le variabili aleatorie esistono solo nel modello matematico, i dati numerici nella realtà.

Se i numeri x_1, \dots, x_n sono gli esiti di n misurazioni, questi possono essere *interpretati* come gli esiti $X_1(\omega), \dots, X_n(\omega)$ di n variabili aleatorie, ma *non sono variabili aleatorie essi stessi*.

Sia (X_1, \dots, X_n) un campione i.i.d. di una v.a. X . Se non abbiamo ulteriori informazioni sulla distribuzione \mathbb{P}_X di X (equivalentemente sulla sua c.d.f. $F(\cdot)$), non si può dire molto di più di quello che è stato fatto; in verità è possibile dare risultati ulteriori ma ci si addentra nella *statistica non parametrica* che è al di là degli obiettivi di questo corso.

A volte però la distribuzione di probabilità \mathbb{P}_X è *parzialmente specificata* nel senso che appartiene ad una famiglia di probabilità (equivalentemente c.d.f.) dipendenti da un opportuno *parametro* usualmente indicato θ , o da più parametri indicati $\theta_1, \dots, \theta_m$.

Ad esempio è ragionevole supporre che l'altezza di una popolazione (ad esempio gli abitanti di una zona) sia rappresentata da una variabile Gaussiana, della quale però non si conosce né media né varianza. Oppure la densità esponenziale è usata per modellizzare i tempi di decadimento radioattivo o la durata di vita di certe apparecchiature (apparecchiature "che non invecchiano"), però non è specificato il parametro: questo parametro deve essere valutato a partire dalle osservazioni.

L'obiettivo della **stima parametrica** è ricostruire il parametro incognito, o i parametri incogniti a partire dalle osservazioni, cioè in funzione del campione.

Una funzione $g(X_1, \dots, X_n)$ di un campione statistico è chiamata **statistica campionaria**. Tali sono ad esempio la **media campionaria** e la **varianza campionaria**, rispettivamente definite come

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}, \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Uno **stimatore** di un parametro θ della distribuzione è una statistica, cioè una funzione del campione $g(X_1, \dots, X_n)$, atta a stimare θ . Come vedremo, la media campionaria \bar{X} è lo stimatore abituale del valore atteso $\mathbb{E}[X]$, mentre la varianza campionaria S^2 è lo stimatore abituale della varianza $\text{Var}(X)$.

NB Poiché uno stimatore è una funzione del campione, anch'esso è una variabile aleatoria: realizzazioni diverse x_1, \dots, x_n del campione X_1, \dots, X_n portano a valori diversi dello stimatore.

A questo punto, si pone il problema di come scegliere un buon stimatore per un parametro e di come valutare la bontà di un dato stimatore. Per quanto riguarda la bontà di un dato stimatore, vediamo alcuni criteri comunemente usati.

Definizione 6.1.2 Dato un parametro θ della distribuzione e un campione X_1, \dots, X_n , una statistica $g(X_1, \dots, X_n)$ si dice *stimatore corretto* (o *non distorto*, *unbiased*) del parametro θ se ammette momento primo e

$$\mathbb{E}[g(X_1, \dots, X_n)] = \theta,$$

cioè la media dello stimatore è il parametro θ .

La seguente proposizione garantisce che media campionaria e varianza campionaria sono stimatori *corretti* di valore atteso e varianza, rispettivamente:

Proposizione 6.1.1 Sia X_1, \dots, X_n un campione statistico, supponiamo che le variabili possiedano momento secondo e siano $\mu = \mathbb{E}[X_i]$ e $\sigma^2 = \text{Var}(X_i)$: indicando $\bar{X} = \frac{X_1 + \dots + X_n}{n}$, si ha

$$\mathbb{E}[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \sigma^2/n, \quad \mathbb{E}\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}\right] = \sigma^2.$$

Dimostrazione. Le prime due uguaglianze sono immediate dalla proprietà di linearità del valore atteso e dall'ipotesi di indipendenza. Per quanto riguarda la seconda, osserviamo che vale

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2;$$

per ogni X_i si ha $\mathbb{E}[X_i^2] = \text{Var}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$, per cui ricordando anche le prime due formule enunciate si conclude che

$$\mathbb{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \sum_{i=1}^n \mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}^2] = n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\sigma^2,$$

da cui la tesi. ■

Definizione 6.1.3 Dato un parametro θ della distribuzione e un campione X_1, \dots, X_n, \dots (di infinite v.a. i.i.d.) di X , la successione di statistiche $g_n(X_1, \dots, X_n)$ si dice uno stimatore *consistente* di θ se, per $n \rightarrow \infty$, $g_n(X_1, \dots, X_n)$ tende a θ in probabilità, cioè se, per ogni $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\{|g_n(X_1, \dots, X_n) - \theta| > \varepsilon\} = 0.$$

In altre parole, quando la taglia del campione diventa molto grande, $g_n(X_1, \dots, X_n)$ si avvicina, con alta probabilità, al valore del parametro θ .

La L.G.N. e il suo corollario ci dicono che media campionaria e varianza campionaria sono (successioni di) stimatori *consistenti* di valore atteso e varianza, rispettivamente.

Definizione 6.1.4 Dato un parametro θ della distribuzione e un campione X_1, \dots, X_{n_0} , con $n_0 \geq n, m$, dati due stimatori *corretti* $g(X_1, \dots, X_n)$ e $h(X_1, \dots, X_m)$ e che ammettono momento secondo, diciamo che $g(X_1, \dots, X_n)$ è più efficiente di $h(X_1, \dots, X_m)$ se

$$\text{Var}(g(X_1, \dots, X_n)) \leq \text{Var}(h(X_1, \dots, X_m)).$$

In altre parole, la dispersione di $g(X_1, \dots, X_n)$ attorno al valor medio θ è minore o uguale alla dispersione di $h(X_1, \dots, X_m)$ attorno a θ .

Dato un campione i.i.d. X_1, \dots, X_n con momento secondo finito, la media campionaria \bar{X}_n è sempre più efficiente al crescere di n : infatti $\text{Var}(\bar{X}_n) = \text{Var}(X_1)/n$ è una funzione decrescente di n . Si può dimostrare che anche la varianza campionaria S_n^2 è sempre più efficiente al crescere di n .

6.2 Stima parametrica

Ci occupiamo ora di come scegliere un buon stimatore. Esponiamo due metodi: il *metodo della massima verosimiglianza* ed il *metodo dei momenti*.

Supponiamo dunque di avere un *campione statistico* la cui legge di probabilità dipende da un parametro $\theta \in \Theta$, nel quale le variabili possono o essere discrete con funzione di massa $p_\theta(x)$, oppure con *densità* $f_\theta(x)$.

Definizione 6.2.1 Si chiama **funzione di verosimiglianza** la funzione $L(\theta; \dots)$ definita, nel caso delle *variabili discrete*, da

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n p_\theta(x_i)$$

e nel caso delle *variabili con densità* da

$$L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

Si noti che, nel caso di variabili discrete, la verosimiglianza è la *densità congiunta* di (X_1, \dots, X_n) . Analoga è l'interpretazione nel caso delle variabili con densità.

Definizione 6.2.2 Si chiama **stima di massima verosimiglianza** (*maximum likelihood estimation*), se esiste, una statistica campionaria, usualmente indicata $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, tale che valga l'eguaglianza

$$L(\hat{\theta}; x_1, \dots, x_n) = \max_{\theta \in \Theta} L(\theta; x_1, \dots, x_n) \quad \text{per ogni } (x_1, \dots, x_n).$$

Nel caso discreto, se x_1, \dots, x_n sono gli esiti del campione, la stima di massima verosimiglianza sceglie il (o un) parametro $\hat{\theta}$ che massimizza la probabilità degli esiti x_1, \dots, x_n effettivamente ottenuti. Nel caso con densità, l'idea è simile, anche se la densità congiunta valutata in x_1, \dots, x_n non è la probabilità di x_1, \dots, x_n .

L'idea del metodo dei momenti è di confrontare i momenti teorici (i valori attesi di X^k)

$$m_k(\theta) = \mathbb{E}_\theta [X^k]$$

con i *momenti empirici* (le medie campionarie di X_1^k, \dots, X_n^k)

$$\sum_{i=1}^n \frac{x_i^k}{n}.$$

Poiché la media campionaria è un buon stimatore del valore atteso, è ragionevole prendere, come stima di θ , un valore $\tilde{\theta}$ che realizzi l'uguaglianza tra i momenti teorici e quelli empirici.

Definizione 6.2.3 si chiama **stima col metodo dei momenti**^a, se esiste, una statistica campionaria $\tilde{\theta} = \tilde{\theta}(x_1, \dots, x_n)$ che permetta di eguagliare alcuni momenti teorici con i corrispondenti momenti empirici, cioè di scrivere, per uno o più interi positivi k ,

$$\mathbb{E}_{\tilde{\theta}} [X^k] = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad \text{per ogni } (x_1, \dots, x_n).$$

^aNon esiste una notazione consolidata per la stima col metodo dei momenti, qui scegliamo di indicarla con $\tilde{\theta}$.

Le definizioni della stima di massima verosimiglianza e della stima col metodo dei momenti si generalizzano in modo naturale al caso di più parametri, prendendo $\theta = (\theta_1, \dots, \theta_m)$ multidimensionale.

Precisiamo subito che queste stime non sempre esistono, a volte coincidono e a volte sono sostanzialmente diverse. Per quanto riguarda la stima col metodo dei momenti, supponendo ad esempio che ci sia un solo parametro, a volte è sufficiente il momento primo, a volte bisogna passare a un momento di ordine superiore.

Osserviamo che *la coppia di parametri media campionaria e varianza empirica è la stima della coppia media e varianza col metodo dei momenti, per $k = 1, 2$.*

Proponiamo ora alcuni esempi per comprendere come applicare questi metodi.

■ **Esempio 6.1 — Campione con densità esponenziale.** Supponiamo che X_1, \dots, X_n abbiano densità esponenziale di parametro θ con $0 < \theta < +\infty$.

Innanzitutto noi adatteremo questo modello ad un campione di osservazioni x_1, \dots, x_n formato da numeri *positivi*: poiché $\mathbb{E}_\theta[X_i] = 1/\theta$, la stima col metodo dei momenti si ottiene imponendo $\frac{1}{\theta} = \sum_i \frac{x_i}{n}$, cioè $\tilde{\theta} = \frac{1}{\bar{x}}$.

Allo stesso modo (sempre se gli x_i sono positivi) la verosimiglianza coincide con $\theta^n e^{-\theta(\sum_i x_i)}$: questa funzione tende a 0 per $\theta \rightarrow 0$ e per $\theta \rightarrow \infty$ (quindi ha massimo) e annullando la derivata si ottiene di nuovo $\hat{\theta} = \frac{n}{\sum_i x_i} = \frac{1}{\bar{x}}$. ■

■ **Esempio 6.2 — Densità uniformi su un intervallo variabile.** Supponiamo dunque che, per $0 < \theta < +\infty$, la densità sia

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & \text{altrove} \end{cases}.$$

Partiamo dunque dalle osservazioni x_1, \dots, x_n che di nuovo supponiamo positive: poiché $\mathbb{E}_\theta[X_i] = \theta/2$, anche questa volta la stima col metodo dei momenti è facile e si ottiene $\tilde{\theta} = 2 \left(\sum_i \frac{x_i}{n} \right) = 2\bar{x}$.

È invece più complicata la stima di massima verosimiglianza: notiamo che la verosimiglianza è

$$L(\theta; x_1, \dots, x_n) = \begin{cases} \theta^{-n} & x_i \leq \theta \quad \forall i \\ 0 & \text{altrimenti} \end{cases}$$

cioè la verosimiglianza è diversa da 0 solo se $\theta \geq \max(x_1, \dots, x_n)$, dopodiché è una funzione decrescente. Ne segue che si ha $\hat{\theta}(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$. ■

Esercizio 6.1 Si consideri un campione di *variabili di Poisson* di parametro θ con $0 < \theta < +\infty$. Si determinino lo stimatore di massima verosimiglianza e la stima con il metodo del momento primo per θ . Si osservi che coincidono e che sono stimatori corretti e consistenti. ■

6.3 Statistiche campionarie di variabili Gaussiane

Assumendo più ipotesi sul campione aleatorio che modella i dati reali chiaramente si possono inferire molte più informazioni, al costo di doversi assicurare che tali ipotesi siano motivate, ma questa è una preoccupazione sperimentale e non matematica.

Grazie all'universalità del Teorema Centrale del Limite, molti fenomeni sono bene descritti da campioni di variabili Gaussiane, delle cui medie e varianze campionarie possiamo studiare in dettaglio le proprietà. I risultati esposti in questo paragrafo saranno fondamentali nel capitolo successivo.

Teorema 6.3.1 Siano X_1, \dots, X_n un campione aleatorio Gaussiano $N(m, \sigma^2)$, e poniamo come sopra:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Valgono i seguenti risultati:

1. le variabili \bar{X}_n e S_n^2 sono indipendenti;
2. la variabile \bar{X} ha densità $N(m, \sigma^2/n)$;

3. la variabile $\frac{n-1}{\sigma^2} S_n^2$ ha densità $\chi^2(n-1)$;
4. la variabile $T = \sqrt{n} \frac{(\bar{X}_n - m)}{S}$ ha densità di Student $T(n-1)$.

Non diamo la dimostrazione di questo risultato, tranne che per la distribuzione di \bar{X} : essa è Gaussiana per la proprietà di riproducibilità della distribuzione Gaussiana (combinazione lineare di v.a. Gaussiane indipendenti è Gaussiana), la media e la varianza di \bar{X} sono state determinate in precedenza (anche nel caso non Gaussiano).

6.4 Esercizi

7. Intervalli di Fiducia

Gli intervalli di fiducia (*confidence interval*¹) spesso sono trattati insieme alla stima parametrica. Tuttavia, come vedremo nel prossimo capitolo, essi possono essere visti come un esempio particolare di test statistico.

La quotidianità ci ha ormai abituato all'uso degli intervalli di fiducia. Ad esempio, durante le proiezioni dei risultati di un referendum, dopo un'ora di spoglio il SÌ viene dato al 26.2 ± 2.3 , dopo due ore e mezza al 25.9 ± 1.2 e infine dopo quattro ore al 26.1 ± 0.4 . All'aumentare della numerosità del campione (che in questo caso si avvicina durante lo spoglio a quella della popolazione totale) la misurazione diventa più precisa, viene cioè proposto un intervallo sempre più stretto.

Gli estremi di questi intervalli sono determinati a partire dagli esiti della misurazione del campione, dunque dal punto di vista matematico dobbiamo definirli come variabili aleatorie ottenute da funzioni del campione statistico.

7.1 Definizione

Consideriamo un campione statistico X_1, \dots, X_n la cui legge (equivalentemente, la c.d.f. delle variabili aleatorie) dipende da un parametro $\theta \in \Theta \subseteq \mathbb{R}$,

$$\mathbb{P}_\theta(X_1 \leq t) = \dots = \mathbb{P}_\theta(X_n \leq t) = F_\theta(t), \quad t \in \mathbb{R}.$$

In altre parole, assumiamo che sullo spazio di probabilità Ω sia definita una famiglia $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ di probabilità tali che per ogni θ fissato le variabili X_1, \dots, X_n siano i.i.d. con c.d.f. F_θ .

L'obiettivo dell'analisi statistica è quindi determinare il più precisamente possibile il valore del parametro θ partendo dagli esiti del campione. Si noti che questo corrisponde a fare una assunzione piuttosto restrittiva sul campione: stiamo ipotizzando di conoscere la legge a meno di un singolo parametro incognito.

■ **Esempio 7.1** Tornando all'esempio delle proiezioni durante lo spoglio dei voti. Ogni voto può essere solamente SÌ oppure NO (la scheda nulle conta come astensione, e non entra nel computo

¹ Il termine *intervallo di confidenza* è una cattiva traduzione entrata nell'uso comune.

della popolazione votante), dunque la situazione è descritta da un campione statistico di variabili di Bernoulli di parametro p . L'obiettivo è quindi determinare p , perchè ci aspettiamo che esso approssimi la percentuale della popolazione totale che ha votato SI. ■

Un intervallo di fiducia è dunque un intervallo i cui estremi sono calcolati a partire dai valori assunti da X_1, \dots, X_n , dunque un *intervallo casuale*, nel quale ci aspettiamo sia contenuto il parametro θ . Proprio perchè l'intervallo è casuale, *non è detto che θ appartenga all'intervallo*, e quindi la definizione matematica deve limitarsi a chiedere che questo avvenga con probabilità alta.

Definizione 7.1.1 Dato un campione statistico X_1, \dots, X_n di legge \mathbb{P}_θ , $\theta \in \Theta \subseteq \mathbb{R}$, e un numero $\alpha \in (0, 1)$, un *intervallo aleatorio*

$$I = [a(X_1, \dots, X_n), b(X_1, \dots, X_n)]$$

(dove $a, b: \mathbb{R}^n \rightarrow \mathbb{R}$ sono funzioni misurabili, e quindi $a(X_1, \dots, X_n)$ e $b(X_1, \dots, X_n)$ sono a loro volta variabili aleatorie) si dice essere un **intervallo di fiducia** per θ al livello $(1 - \alpha)$ se,

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(\theta \in I) = \mathbb{P}_\theta(a(X_1, \dots, X_n) \leq \theta, b(X_1, \dots, X_n) \geq \theta) \geq 1 - \alpha.$$

Tipicamente α è un numero piccolo (ad esempio 0.05 o 0.02) in modo che il *livello di fiducia* $1 - \alpha$ sia vicino a 1. La scelta dell'intervallo di fiducia è però oggetto di un compromesso a volte difficile, infatti osserviamo che:

- da un lato vogliamo intervalli di fiducia più piccoli possibile per identificare θ con precisione, ma al ridursi dell'intervallo deve diminuire il livello di fiducia $1 - \alpha$;
- dall'altro vogliamo un livello di fiducia alto, ovvero vogliamo sia molto probabile che il parametro θ stia nell'intervallo (altrimenti la nostra predizione è fuorviante), ma questo richiede intervalli grandi oppure estremi a, b molto precisi.

È di fatto impossibile cercare risultati generali: il resto del capitolo analizza quindi nel dettaglio esempi particolari, in cui il campione ha leggi di probabilità dei tipi più comuni esposti nei capitoli precedenti, con particolare attenzione ai campioni Gaussiani.

7.2 Intervalli di fiducia per la media di un campione Gaussiano

Sia X_1, \dots, X_n un campione statistico con legge Gaussiana $N(m, \sigma^2)$. Tale legge ha in effetti *due* parametri, corrispondenti a media e varianza: in questo paragrafo cerchiamo intervalli di fiducia per il parametro $\theta = m \in \mathbb{R}$.

Abbiamo visto nel capitolo precedente che $\bar{X}_n = (X_1 + \dots + X_n)/n$ è uno stimatore per la media m del campione, per cui è naturale cercare un intervallo di fiducia della forma

$$I = [\bar{X}_n - d, \bar{X}_n + d] =: [\bar{X}_n \pm d]$$

(l'ultima espressione è una abbreviazione definita da quella precedente), in cui $d > 0$ è un numero da determinare, in modo che il livello di fiducia sia alto ma senza che d sia troppo grande.

Definizione 7.2.1 Se $[\bar{X}_n \pm d]$, con $d > 0$ una variabile aleatoria possibilmente costante, è un intervallo di fiducia per la media m , d è detta **precisione della stima** e d/\bar{X}_n è detta **precisione relativa della stima**.

Proposizione 7.2.1 — Intervallo di fiducia per la media, varianza nota. Dato $\alpha \in (0, 1)$ e assumendo sia nota $\sigma > 0$, l'intervallo aleatorio

$$\left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right]$$

è un intervallo di fiducia per la media m del campione X_1, \dots, X_n con livello di fiducia $1 - \alpha$.

Dimostrazione. Dobbiamo imporre la condizione

$$1 - \alpha \leq \mathbb{P}_m(m \in I) = \mathbb{P}_\theta(|\bar{X}_n - m| \leq d) = \mathbb{P}_\theta\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - m| \leq \frac{d\sqrt{n}}{\sigma}\right),$$

in cui l'ultimo passaggio segue ricordando che la variabile $\sqrt{n}(\bar{X}_n - m)/\sigma$ è Gaussiana standard. Stiamo cercando il valore di d più piccolo possibile, per cui scegliamo quello per cui è verificata l'uguaglianza, il che corrisponde a scegliere esattamente il quantile $\frac{d\sqrt{n}}{\sigma} = q_{1-\frac{\alpha}{2}}$, da cui l'intervallo nell'enunciato. ■

Osserviamo che la precisione d della stima (cioè la semi-ampiezza dell'intervallo):

- cresce al crescere del livello di fiducia $1 - \alpha$ (livello di fiducia più alto implica minore precisione);
- cresce al crescere di σ^2 (dispersione maggiore implica minore precisione);
- decresce al crescere di n (maggiore numerosità del campione implica maggiore precisione).

In particolare, dalla formula per d è possibile determinare la numerosità del campione n per cui l'intervallo abbia una data precisione e un dato livello di fiducia.

■ **Esempio 7.2** Il peso (in kg) di una specie di salmoni segue una distribuzione Gaussiana di media m non nota e deviazione standard $\sigma = 0.4$. Su un campione di 10 salmoni, la media aritmetica dei pesi risulta essere 3.58 kg.

Seguendo la procedura sopra descritta, determiniamo un intervallo di fiducia per la media al livello $95\% = 1 - \alpha$:

$$\left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}}q_{0.975}\right] = \left[\bar{X}_{10} \pm \frac{0.4}{\sqrt{10}} \cdot 1.96\right] = [\bar{X}_{10} \pm 0.248].$$

Inserendo il dato $\bar{x}_n = 3.58$, otteniamo l'intervallo numerico $[3.332, 3.828]$.

Mantenendo il livello di fiducia al 95% , quanto grande deve essere n affinché la precisione della stima sia 0.1 ? Per trovare il minimo n che soddisfi $d \leq 0.1$, $1 - \alpha = 0.95$, imponiamo

$$d = \frac{\sigma}{\sqrt{n}}q_{1-\alpha/2} \leq 0.1$$

da cui $n \geq \frac{\sigma^2 q_{0.975}^2}{0.1^2} \approx 61.46$, quindi il minimo n è 62. ■

Vediamo ora il caso più realistico in cui la varianza σ^2 è sconosciuta. Non possiamo in questo caso usare il numero σ per definire gli estremi dell'intervallo, ma questo problema si aggira sostituendo a σ^2 lo stimatore

$$S_n^2 = \frac{\sum_{i=1}^n (X_i^2 - \bar{X})}{n-1}.$$

In questo caso, la variabile $\sqrt{n}(\bar{X}_n - m)/S_n^2$ è di Student² $T(n-1)$, per cui per determinare il miglior parametro d dobbiamo considerare il β -quantile $\tau_{\beta, n-1}$ della variabile $T(n-1)$. Con passaggi analoghi a quelli visti sopra si ottiene:

Proposizione 7.2.2 — Intervallo di fiducia per la media, varianza non nota. Dato $\alpha \in (0, 1)$, l'intervallo aleatorio

$$\left[\bar{X}_n \pm \frac{S_n}{\sqrt{n}}\tau_{1-\frac{\alpha}{2}, n-1}\right]$$

è un intervallo di fiducia per la media m del campione X_1, \dots, X_n con livello di fiducia $1 - \alpha$.

²Questo è il *nom de plume* scelto dallo statistico William Sealy Gosset, che ha introdotto l'idea sopra esposta per determinare intervalli di fiducia.

Quando n è grande ($n \geq 60$) si può approssimare il quantile della variabile di Student con quello della Gaussiana standard.

■ **Esempio 7.3** Nell'esempio precedente, supponiamo ora che anche σ sia non nota e che la deviazione standard sul campione sia $S_n = 0.4$. Come cambiano le risposte?

In questo caso, l'intervallo di fiducia di livello $1 - \alpha = 0.95$ è dato da

$$\left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} t_{0.975,9} \right] = \left[\bar{X}_{10} \pm \frac{S_n}{\sqrt{10}} \cdot 2.262 \right] = \left[\bar{X}_{10} \pm 0.715 \cdot S_n \right].$$

Inserendo i dati $\bar{x}_n = 3.58$, $S = 0.4$ otteniamo la precisione $d = 0.715 \cdot 0.4 = 0.286$, e l'intervallo numerico $[3.294, 3.866]$. Notiamo che, anche se il valore numerico assunto da S_n è lo stesso di σ , la precisione è minore: questo perché la distribuzione T di Student ha code più pesanti rispetto alla Gaussiana standard, cioè $t_{\beta,n} > q_{\beta}$ per ogni $\beta > 1/2$.

Per quanto riguarda il minimo n che soddisfi $d \leq 0.1$, $1 - \alpha = 0.05$, purtroppo in questo caso non è possibile dare una risposta a priori, poiché d dipende da S_n , che dipende dai dati e in particolare dalla loro taglia. Tuttavia si può dare una stima ragionevole su S_n , ad esempio non ci attendiamo che essa sia superiore a $2 \cdot 0.4$, quindi calcolare n per tale stima, infine, dopo aver raccolto i dati su un campione di taglia n , verificare che il valore di S_n fornito dai dati rispetti la stima che avevamo imposto, ad esempio $S_n \leq 2 \cdot 0.4$. ■

Gli intervalli di fiducia appena descritti sono detti *bilateri*, perchè entrambi gli estremi sono variabili aleatorie. A volte può essere interessante considerare un intervallo *unilatero*: ad esempio se ci chiediamo soltanto se la media non è troppo alta si considera un intervallo *unilatero sinistro* della forma $I = (-\infty, \bar{X}_n + d]$. Con passaggi simili a quelli visti sopra si ottiene:

Proposizione 7.2.3 — Intervalli di fiducia unilateri per la media, varianza nota. Dato $\alpha \in (0, 1)$, gli intervalli aleatori

$$\left(-\infty, \bar{X}_n + \frac{\sigma}{\sqrt{n}} q_{1-\alpha} \right], \quad \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} q_{1-\alpha} + \infty \right),$$

sono intervalli di fiducia per la media m del campione X_1, \dots, X_n con livello di fiducia $1 - \alpha$.

Gli intervalli di fiducia unilateri con varianza sconosciuta sono perfettamente analoghi, sostituendo a σ la variabile S_n e ai quantili della variabile Gaussiana standard quelli della variabile di Student.

7.3 Intervalli di fiducia per la media di un campione Bernoulli

Consideriamo un campione X_1, \dots, X_n di variabili di Bernoulli di parametro $p \in (0, 1)$, corrispondente al valore atteso delle singole variabili. Cerchiamo quindi intervalli di fiducia per il parametro p nella forma $[\bar{X}_n \pm d]$, essendo \bar{X}_n uno stimatore per p .

La determinazione di intervalli di fiducia per la media nel caso Gaussiano si basava sul fatto che la legge della variabile $\sqrt{n}(\bar{X}_n - m)/\sigma$ è Gaussiana standard. Nel caso delle variabili di Bernoulli sappiamo che $X_1 + \dots + X_n$ è di tipo binomiale $B(n, p)$, e dunque i relativi quantili, che appariranno nella determinazione dell'intervallo di fiducia, sono complicati e dipendono anche da n .

Quando n è grande (almeno $n \geq 80$), possiamo fare tuttavia affidamento sul Teorema Centrale del Limite, che afferma che la variabile

$$\frac{X_1 + \dots + X_n - np}{\sqrt{np(1-p)}} = \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$$

è approssimativamente Gaussiana standard. Questo ancora non ci permette di concludere agevolmente, perché la varianza $\sigma^2 = p(1-p)$ non è nota. Tuttavia σ^2 è funzione del parametro

incognito p e quindi è ragionevole stimarla con $\bar{X}_n(1 - \bar{X}_n)$. Precisamente, come conseguenza del Teorema Centrale del Limite, si dimostra che

$$\frac{X_1 + \cdots + X_n - np}{\sqrt{n\bar{X}_n(1 - \bar{X}_n)}} = \sqrt{n} \frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}}$$

converge in distribuzione a una Gaussiana standard, per $n \rightarrow \infty$. Quindi, con passaggi analoghi a quelli del paragrafo precedente si ottiene:

Proposizione 7.3.1 — Intervallo di fiducia per la media. Dato $\alpha \in (0, 1)$, l'intervallo aleatorio

$$\left[\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} q_{1-\alpha/2} \right]$$

è un intervallo di fiducia per la media p del campione X_1, \dots, X_n con livello di fiducia **approssimativamente** $1 - \alpha$. Precisamente, si ha

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(p \in \left[\bar{X}_n \pm \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} q_{1-\alpha/2} \right] \right) = 1 - \alpha.$$

Il termine “approssimativamente” significa che ad α fissato non è detto che il livello di fiducia sia $1 - \alpha$, ma per n grande abbastanza l'errore è trascurabile. Anche la precisione della stima,

$$d = \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} q_{1-\alpha/2},$$

è di conseguenza approssimata.

■ **Esempio 7.4** Si vuole condurre un sondaggio telefonico per determinare la percentuale di gradimento del governo. Qual è il numero minimo di telefonate che bisogna effettuare per avere una precisione della stima inferiore a 1% con fiducia al 95 % ?

Assumiamo che la risposta al sondaggio possa essere solo SÌ oppure NO, dunque stiamo considerando un campione aleatorio di variabili Bernoulli di cui vogliamo stimare il parametro p . Dalla definizione di livello di fiducia abbiamo $0.95 = 1 - \alpha$, dunque il quantile Gaussiano che ci interessa è $q_{1-\alpha/2} = q_{0.975} \approx 1.96$. La condizione sulla precisione (approssimata) risulta essere

$$\sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} q_{0.975} \leq 0.01, \quad \text{ovvero } \sqrt{n} \geq \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n)} \times 1.96}{0.01}$$

Non conosciamo a priori le risposte alle telefonate: vogliamo che n sia il più piccolo possibile tale che la condizione sia verificata per qualsiasi esito possibile. Scegliamo dunque l'esito che massimizza il membro di destra nell'ultima disuguaglianza: siccome $\max_{0 < x < 1} x(1 - x) = \frac{1}{4}$, ovvero il caso peggiore è quello in cui $\bar{X}_n = 0.5$, otteniamo la condizione

$$n \geq \left(\frac{0.5 \cdot 1.96}{0.01} \right)^2 = 9604.$$

Osserviamo che per n così grande, l'approssimazione del livello di fiducia dell'intervallo è molto buona. ■

7.4 Intervalli di fiducia per la media di un campione di taglia grande

La procedura sopra esposta per individuare intervalli di fiducia approssimati per la media usando il Teorema Centrale del Limite si può applicare esattamente nello stesso modo a campioni di variabili aleatorie di taglia grande (e di momento secondo finito):

Proposizione 7.4.1 — Intervallo di fiducia per la media, campioni grandi, varianza nota.

Sia X_1, \dots, X_n un campione i.i.d., la cui legge ha momento secondo finito, con n grande. Dato $\alpha \in (0, 1)$, l'intervallo aleatorio

$$\left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right]$$

è un intervallo di fiducia per la media m del campione con livello di fiducia **approssimativamente** $1 - \alpha$. Precisamente, si ha

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(m \in \left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right] \right) = 1 - \alpha.$$

Proposizione 7.4.2 — Intervallo di fiducia per la media, campioni grandi, varianza non nota. Sia X_1, \dots, X_n un campione i.i.d., la cui legge ha momento secondo finito, con n grande. Dato $\alpha \in (0, 1)$, l'intervallo aleatorio

$$\left[\bar{X}_n \pm \frac{S_n}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right]$$

è un intervallo di fiducia per la media m del campione con livello di fiducia **approssimativamente** $1 - \alpha$. Precisamente, si ha

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(m \in \left[\bar{X}_n \pm \frac{S_n}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right] \right) = 1 - \alpha.$$

Rimandiamo agli Esercizi alla fine del Capitolo per esempi di questo tipo.

7.5 Intervalli di fiducia per la varianza di un campione Gaussiano

Se vogliamo trovare intervalli di fiducia per la *varianza* di un campione Gaussiano $N(m, \sigma^2)$, non c'è sostanziale differenza tra i casi in cui m è nota oppure no: consideriamo direttamente il secondo.

Il calcolo di intervalli di fiducia bilateri in questo caso non è agevole, ci concentriamo su intervalli unilateri. Del resto, questi ultimi sono di maggiore interesse, essendo la varianza una quantità positiva di cui ci interessa capire se supera o meno una certa soglia. Ad esempio, in una misurazione di prodotti industriali standardizzati, siamo interessati a controllare che la varianza delle misure nel controllo qualità sia piccola, ovvero inferiore a una soglia fissata, ad indicare che la produzione è di buona qualità perché tutti i pezzi hanno misura simile.

Proposizione 7.5.1 — Intervalli di fiducia per la varianza. Dato $\alpha \in (0, 1)$, gli intervalli aleatori

$$\left(0, \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\chi_{\alpha, n-1}^2} \right] = \left(0, \frac{(n-1)S_n^2}{\chi_{\alpha, n-1}^2} \right], \quad \left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi_{(1-\alpha), n-1}^2}, +\infty \right) = \left[\frac{(n-1)S^2}{\chi_{1-\alpha, n-1}^2}, +\infty \right),$$

sono intervalli di fiducia per σ^2 (rispettivamente unilateri sinistri e destri) con livello di fiducia $1 - \alpha$.

Ricordiamo che con $\chi^2_{\beta, n-1}$ indichiamo il β -quantile di una variabile di tipo $\chi^2(n-1)$. In effetti, il punto cruciale in questo caso è il fatto che, se X_1, \dots, X_n è un campione di variabili $N(m, \sigma^2)$, la variabile

$$\sum_{i=1}^n \frac{(X_i - \bar{X}_n)^2}{\sigma^2} = (n-1) \frac{S_n^2}{\sigma^2}$$

è di tipo $\chi^2(n-1)$. Ricordando questo, la dimostrazione diventa una facile verifica della definizione che lasciamo al lettore.

7.6 Esercizi

Esercizio 7.1 Un certo metodo per misurare il pH di una soluzione fornisce un risultato distribuito come una Gaussiana, di media il valore autentico del pH della soluzione e deviazione standard 0.1. Vengono effettuate 50 misurazioni di una soluzione; la media degli esiti di tali misurazioni risulta 8.19.

- Fornire un intervallo di fiducia di livello 0.95 per il valore autentico del pH della soluzione.
- Quante misurazioni è necessario eseguire affinché, con livello di fiducia del 95%, la precisione dell'intervallo sia 0.01?

■

Esercizio 7.2 Un server che elabora un certo tipo di richiesta può commettere un errore. Su 1000 richieste, sono stati osservati 50 errori.

- Fornire un intervallo di fiducia di livello 0.99 per la probabilità di errore su una singola richiesta.
- Quante richieste dobbiamo analizzare, per ridurre la precisione a 0.003 (mantenendo lo stesso livello di fiducia)?

■

Esercizio 7.3 Si misurano i livelli di emoglobina di 80 persone adulte di sesso femminile, ottenendo una media di 14.1 g/dl e una deviazione standard di 0.7 g/dl. Fornire un intervallo di fiducia del 99% per il livello medio di emoglobina (in una persona adulta di sesso femminile). ■

Esercizio 7.4 Su un campione di 80 individui di una certa popolazione, 48 presentano un certo gene. (i) Trovare un intervallo di fiducia di livello 95% per la frequenza del gene nella popolazione. (ii) Quanto grande deve essere la taglia del campione, affinché la precisione della stima sia inferiore al 5%, mantenendo lo stesso livello 95%? ■

Esercizio 7.5 Per ogni anno dal 2016 al 2021 viene misurata la temperatura media (in gradi centigradi) a Milano nel mese di settembre; la media e la deviazione standard di tali dati risultano rispettivamente 20.83 e 1.35. Supponendo che la temperatura media di settembre sia distribuita come una gaussiana, fornire un intervallo di fiducia a livello 95% per la media delle temperature medie di settembre a Milano. ■

Esercizio 7.6 Si consideri un campione X_1, \dots, X_n di variabili con densità esponenziale di

parametro $\lambda > 0$. Si mostri che la variabile

$$\frac{\bar{X}_n}{\lambda} = \frac{X_1 + \dots + X_n}{n\lambda}$$

ha densità $\Gamma(n, 1/n)$. Si usi questo fatto per determinare intervalli di fiducia bilateri per λ , con estremi che dipendono dal β -quantile $\gamma_{\beta,n}$ della densità $\Gamma(n, 1/n)$. ■

Esercizio 7.7 Utilizzando la stessa procedura vista per campioni statistici di Bernoulli, si determini un intervallo di fiducia bilatero per il parametro λ di un campione X_1, \dots, X_n di variabili aleatorie di Poisson. ■

Esercizio 7.8 Siano X_1, X_2 un campione di variabili con densità uniforme sull'intervallo $[\theta - 1/2, \theta + 1/2]$, con $\theta \in \mathbb{R}$ il parametro da stimare. Posto $\bar{X} = \frac{X_1 + X_2}{2}$, si considerino gli intervalli aleatori

$$I_1 = \left[\bar{X} \pm \frac{\min\{|X_1 - X_2|, 1 - |X_1 - X_2|\}}{2} \right], \quad I_2 = \left[\bar{X} \pm \frac{1 - |X_1 - X_2|}{4} \right].$$

Si mostri che:

- sono entrambi intervalli di fiducia per θ con livello di affidabilità 50%;
- l'intervallo I_1 è mediamente più piccolo di I_2 , nel senso che se $\theta' \neq \theta$ vale

$$\mathbb{P}_\theta(\theta' \in I_1) \leq \mathbb{P}_\theta(\theta' \in I_2);$$

- se $|X_1 - X_2| \geq 1/2$ l'intervallo I_1 contiene *sempre* il vero parametro θ , ma questo non è vero per I_2 .

Si osservi infine che se $|X_1 - X_2|$ è molto piccolo, I_1 è quasi ridotto a un punto, ed è praticamente impossibile che contenga θ , per questo il livello di affidabilità dell'intervallo risulta essere così basso. ■

8. Test Statistici

Riprendendo la discussione metodologica avviata nei capitoli precedenti, ricordiamo quale sia lo scopo della statistica parametrica, e come la teoria della Probabilità permetta di perseguirlo.

Si considera un fenomeno di cui abbiamo a disposizione numerose osservazioni, situazione schematicamente descritta sopra come misura di un carattere su un campione di individui di una popolazione. Si propone un modello matematico, più precisamente probabilistico, con cui si vuole dare un'approssimazione della varietà di esiti delle misurazioni effettuate. Tale modello dipende da uno o più parametri, variando i quali ci aspettiamo di poter ottenere una buona approssimazione: l'obiettivo è determinare una procedura per scegliere tali parametri a partire dalle osservazioni effettuate.

Pianificare un test statistico significa per prima cosa *formulare una ipotesi* sui parametri incogniti e quindi *pianificare un esperimento* per decidere se è ragionevole accettare o respingere l'ipotesi in base ai dati disponibili. Prima di procedere con la discussione rigorosa, affianchiamo a questa premessa due importanti considerazioni.

In primo luogo, ci stiamo restringendo a considerare modelli matematici dati da leggi di variabili aleatorie dipendenti da pochi parametri –uno o due negli esempi più frequenti– sulle ragioni della cui scelta abbiamo proposto semplici considerazioni euristiche, perlopiù basate sul tipo di esiti numerici nei singoli esempi. Nonostante l'efficacia della teoria già a questo livello, va ricordato che essa è stata oggetto di ampie generalizzazioni nel corso di decenni, pur mantenendo le stesse basi concettuali.

In secondo luogo, sottolineiamo il fatto che le conclusioni della statistica –quali la determinazione di intervalli di fiducia o i risultati di test statistici– *non devono essere considerati come verità* concernenti il fenomeno osservato. Questa sarebbe infatti una interpretazione scorretta del metodo, che invece restituisce (internamente al modello matematico) una certa *approssimazione* degli esiti che sarà buona con un opportuno *grado di fiducia*.

8.1 Concetti generali

In termini generici, una ipotesi statistica è una affermazione sul parametro θ che governa la legge di un campione statistico in considerazione. Rigorosamente, si dà la seguente:

Definizione 8.1.1 — Ipotesi Nulla e Alternativa. Sia X_1, \dots, X_n un campione statistico la cui legge (ovvero la funzione di ripartizione di ogni singola variabile $F_\theta = F_{X_i}$) dipende da un parametro $\theta \in \Theta \subseteq \mathbb{R}$.

Data una partizione $\Theta = \Theta_0 \cup \Theta_1$ in due sottoinsiemi disgiunti, l'*ipotesi nulla* \mathcal{H}_0 è l'affermazione logica $\theta \in \Theta_0$, e Θ_0 è detto l'insieme dei parametri dell'ipotesi nulla. L'*ipotesi alternativa* \mathcal{H}_1 è l'affermazione complementare $\theta \in \Theta_1$, e Θ_1 è l'insieme dei parametri dell'ipotesi alternativa.

A questo livello non vi è differenza tra i ruoli delle ipotesi \mathcal{H}_0 e \mathcal{H}_1 , ma vi è una distinzione fondamentale contenuta invece nella definizione di test statistico, che introduciamo dapprima in termini euristici. Un **test statistico** è una procedura per decidere se accettare o rifiutare l'ipotesi nulla \mathcal{H}_0 a partire dai valori assunti dal campione X_1, \dots, X_n :

- si accetta l'ipotesi nulla \mathcal{H}_0 se i valori assunti dal campione sono con essa compatibili (in un senso al momento da rendere preciso): in questo caso si parla di **esito negativo** del test;
- si rifiuta \mathcal{H}_0 in favore di \mathcal{H}_1 se, con alto grado di fiducia, i valori non sono compatibili con \mathcal{H}_0 (e sono in favore di \mathcal{H}_1), cioè se c'è evidenza statistica contro \mathcal{H}_0 (e per \mathcal{H}_1): in questo caso si parla di **esito positivo** del test.

Notiamo qui l'asimmetria tra le ipotesi: rifiutare \mathcal{H}_0 vuol dire che i dati forniscono una evidenza contro l'ipotesi nulla \mathcal{H}_0 ; accettare \mathcal{H}_0 invece non implica un'evidenza per \mathcal{H}_0 , ma ci dice solo che i dati non forniscono evidenza contro \mathcal{H}_0 .

■ **Esempio 8.1** Per verificare se è in corso una infezione di un virus¹ si preleva un campione di materiale biologico, lo si sottopone a una procedura che trasforma il RNA in DNA (ed amplifica il numero delle molecole) e si introduce una sostanza che rende fluorescenti le parti di DNA virali. Si contano poi queste parti e si determina che l'infezione è in corso (ovvero il test è positivo) se tale numero supera una certa soglia, che a titolo esemplificativo diciamo essere l'1%.

Il modello probabilistico è un campione di variabili di Bernoulli di parametro $\theta \in \Theta = (0, 1)$, a rappresentare il fatto che singole sezione di DNA siano fluorescente (quindi virale) o meno. L'ipotesi nulla "la percentuale di sezioni fluorescenti di DNA non supera il 1%" (corrispondente all'esito negativo del test) è rappresentata dalla scelta $\Theta_0 = (0, 0.01]$ e $\Theta_1 = (0.01, 1)$. ■

■ **Esempio 8.2** In un controllo di qualità si desidera valutare la percentuale di pezzi difettosi in una produzione. In modo del tutto analogo all'esempio precedente, il modello probabilistico è un campione di variabili di Bernoulli di parametro $\theta \in \Theta = (0, 1)$, e l'ipotesi nulla "la percentuale di pezzi difettosi non supera il 2%" corrisponde a considerare $\Theta_0 = (0, 0.02]$ e $\Theta_1 = (0.02, 1)$.

Consideriamo invece la situazione di due fabbriche concorrenti, con l'obiettivo di valutare se una produca più pezzi difettosi dell'altra o viceversa. Possiamo a tal fine considerare come modello due campioni indipendenti X_1, \dots, X_n e Y_1, \dots, Y_n di variabili di Bernoulli di parametri rispettivamente $\theta_1, \theta_2 \in (0, 1)$. Essendo però interessati al confronto, possiamo in effetti ridurci a considerare il singolo parametro $\theta = \theta_1 - \theta_2 \in (-1, 1) = \Theta$. L'ipotesi nulla "non c'è differenza tra le due produzioni" corrisponde quindi a considerare $\Theta_0 = \{0\}$. ■



Il termine "ipotesi nulla" è ancora una volta una traduzione impropria di una locuzione inglese, *null hypothesis*, che dovrebbe essere resa invece con "ipotesi zero". La scelta del termine si riferisce al fatto che solitamente \mathcal{H}_0 è l'**ipotesi sotto cui la variabilità degli esiti dei dati è spiegata dalla sola casualità dell'estrazione** e non invece da una cattiva scelta dei parametri del modello. Nello specifico caso in cui il parametro θ misuri la differenza tra due situazioni, come nell'ultimo esempio proposto, \mathcal{H}_0 corrisponde all'ipotesi in cui tale differenza è appunto "nulla", e le piccole differenze nei dati sono dovute all'aleatorietà inerente alla misurazione.

¹L'esempio è naturalmente una descrizione semplificata della procedura reale con cui si esegue questo tipo di test.

Fissata l'ipotesi, dobbiamo quindi determinare un insieme di risultati che portano a rifiutarla: in altre parole, all'interno del modello probabilistico, dobbiamo scegliere un evento che se si realizza conduce a rifiutare \mathcal{H}_0 . Questo corrisponde a scegliere un sottoinsieme dello spazio di probabilità: vogliamo un insieme di elementi $\omega \in \Omega$ tale che l'osservazione di esiti $X_1(\omega), \dots, X_n(\omega)$ conduca a rifiutare \mathcal{H}_0 .

Definizione 8.1.2 — Regione critica e Test. La *regione critica* o *regione di rifiuto* per l'ipotesi nulla \mathcal{H}_0 è un evento $C \subset \Omega$; il complementare $A = \Omega \setminus C$ è detto invece *regione di accettazione*.

Data una realizzazione (x_1, \dots, x_n) del campione (X_1, \dots, X_n) , l'ipotesi nulla viene:

- *respinta* se il vettore di variabili aleatorie $X_1(\omega), \dots, X_n(\omega)$ assume valore (x_1, \dots, x_n) per un qualche $\omega \in C$, ovvero *se si verifica l'evento C*;
- *accettata* se invece all'interno della regione critica $\omega \in C$ il vettore $X_1(\omega), \dots, X_n(\omega)$ non coincide mai con i dati (x_1, \dots, x_n) , ovvero *se non si verifica l'evento C*.

NB

Si potrebbe pensare che sia più intuitivo e comodo definire la regione critica come un sottoinsieme dello spazio dei valori del campione X_1, \dots, X_n , ovvero di \mathbb{R}^n : date delle osservazioni $(x_1, \dots, x_n) \in \mathbb{R}^n$ vorremmo decidere se sono coerenti o meno con l'ipotesi nulla. Nella pratica, tuttavia, la regione critica è praticamente sempre definita come un evento, come ad esempio $\{\bar{X}_n \in I\}$ per qualche regione $I \subset \mathbb{R}$. In altre parole, la regione critica viene scelta come un evento associato a una statistica campionaria, dunque la definizione naturale è quella data. In effetti, trattando dati numerosi non è ragionevole né conveniente considerare valori di singole misure e si opta invece per indici statistici: la definizione matematica rispecchia questa esigenza.

Il risultato del test (accettare o rifiutare \mathcal{H}_0) è soggetto a due tipi di errore:

- un *errore di prima specie* consiste nel rifiutare l'ipotesi \mathcal{H}_0 quando questa è soddisfatta (*falso positivo*): se $\theta \in \Theta_0$, la probabilità di commettere un errore di prima specie è dunque $\mathbb{P}_\theta(C)$;
- un *errore di seconda specie* consiste nell'accettare l'ipotesi \mathcal{H}_0 quando questa non è soddisfatta (*falso negativo*): se $\theta \in \Theta_1$ la probabilità di commettere un errore di prima specie è $\mathbb{P}_\theta(A)$.

■ **Esempio 8.3** Tornando all'esempio del controllo di qualità in una singola produzione, si rifiuterà l'ipotesi se la percentuale di pezzi difettosi riscontrata è superiore a una certa soglia d . Scriveremo sinteticamente $C = \{\omega \in \Omega \mid \bar{X}_n(\omega) > d\} = \{\bar{X}_n > d\}$. Minimizzando l'errore di prima specie, la scelta di d dovrà essere tale che la probabilità $\mathbb{P}_\theta(C) = \mathbb{P}_\theta\{\bar{X} > d\}$ sia piccola per ogni $\theta \leq 0.02$. ■

Definizione 8.1.3 — Livello e potenza di un test. Dato $0 < \alpha < 1$, si dice che il test è di livello α se

$$\forall \theta \in \Theta_0, \quad \mathbb{P}_\theta(C) \leq \alpha.$$

Si chiama *potenza del test* (*statistical power*) la funzione, definita sui parametri dell'alternativa,

$$\Theta_1 \ni \theta \mapsto \mathbb{P}_\theta(C) \in [0, 1].$$

Fissare un livello significa pertanto fissare *un limite superiore per la probabilità dell'errore di prima specie*, scegliendo opportunamente la regione critica. I valori tipici del livello sono piccoli, ad esempio 0.05 o 0.02. Il livello del test è legato al grado di fiducia che chiediamo per rifiutare: più piccolo è il livello α , maggiore è l'evidenza che chiediamo per rifiutare \mathcal{H}_0 .

La potenza è la probabilità \mathbb{P}_θ di rifiutare correttamente \mathcal{H}_0 quando questa è falsa, quindi rappresenta la *capacità di accorgersi che l'ipotesi \mathcal{H}_0 non è soddisfatta*.

NB

È desiderabile avere test con livello basso (bassa probabilità di errore di prima specie) e potenza alta (bassa probabilità di errore di seconda specie). In genere, tuttavia, più basso è il livello, più bassa è la potenza, per cui le due richieste sono in contrapposizione. Questo è intuitivo: più evidenza chiediamo per rifiutare, più aumenta il rischio di accettare \mathcal{H}_0 quando questa non è soddisfatta. Per avere sia livello basso sia potenza alta, si può agire sulla taglia del campione: come vedremo, aumentare la taglia del campione diminuisce il livello e aumenta la potenza.

Introduciamo anche una nomenclatura alternativa, del tutto equivalente, comune a certe applicazioni, come ad esempio la teoria dell'affidabilità (*reliability theory*) nell'Ingegneria e nel campo assicurativo.

Definizione 8.1.4 — Curva operativa. Si chiama curva operativa la funzione, definita sull'intero insieme dei parametri,

$$\Theta \ni \theta \mapsto \beta(\theta) = \mathbb{P}_\theta(C^c) = 1 - \mathbb{P}_\theta(C) \in [0, 1],$$

ossia la probabilità, in funzione del parametro θ , di accettare l'ipotesi nulla. Si dice che:

- se $\theta \in \Theta_0$, $\beta(\theta)$ è il tasso di falso positivo (*false positive rate*);
- se $\theta \in \Theta_1$, $1 - \beta(\theta)$ è il tasso di falso negativo (*false negative rate*).

Si noti che il tasso di falso negativo coincide con la potenza del test, mentre il livello è il minimo valore assunto dal tasso di falso positivo su Θ_0 .

Il metodo classico di impostazione di un test prevede di fissare un livello α a priori e di scegliere la regione critica C in modo da soddisfare $\mathbb{P}_\theta(C) \leq \alpha$ per ogni $\theta \in \Theta_0$ (cosicché il livello del test sia appunto α). La scelta del livello α influisce sulla scelta della regione critica C e quindi sul risultato del test: più piccolo è α (cioè maggiore è l'evidenza che chiediamo per rifiutare), più piccola sarà la regione critica C . In particolare, se rifiutiamo \mathcal{H}_0 ad un livello α , allora rifiutiamo \mathcal{H}_0 a ogni livello $\alpha' \geq \alpha$, mentre potremmo accettare \mathcal{H}_0 per un qualche livello minore di α .

Per ovviare alla dipendenza del risultato del test da α , nelle applicazioni si fa praticamente sempre uso di un'altra formulazione, centrata sulla nozione di *p-value*². A differenza della regione critica, che è fissata a priori indipendentemente dai valori del campione, il *p-value* dipende dalla realizzazione (x_1, \dots, x_n) del campione (X_1, \dots, X_n) . Informalmente, si dice che

il p-value dei dati (x_1, \dots, x_n) è la probabilità, calcolata accettando l'ipotesi nulla \mathcal{H}_0 , di ottenere “dati più estremi”, rispetto all'ipotesi \mathcal{H}_0 , di quelli osservati.

Naturalmente questa non è una definizione matematica: anzitutto non è chiaro il significato di “estremo”, ed inoltre in generale l'ipotesi nulla è data da un insieme di valori Θ_0 , ma per calcolare una probabilità dobbiamo scegliere un singolo $\theta \in \Theta$. In effetti, la scarsa conoscenza dei fondamenti della statistica porta spesso nelle applicazioni ad applicare male il metodo del *p-value*, basandosi solo sulla nozione intuitiva e portando a invalidare i risultati dell'analisi statistica.

■ **Esempio 8.4** Torniamo al controllo di qualità con ipotesi nulla “la percentuale di pezzi difettosi non supera il 2%”. Se su 1000 pezzi ne risultano 30 difettosi, il *p-value* corrisponde alla probabilità di trovare *almeno* 30 pezzi difettosi su 1000, nell'ipotesi \mathcal{H}_0 che θ sia 2%. Si osservi che l'ipotesi nulla è data da un intervallo, $\Theta_0 = (0, 0.02)$, ma per calcolare il *p-value* noi scegliamo l'estremo $\theta = 0.02$. ■

Sebbene la nozione informale di *p-value* sopra descritta possa essere un utile strumento mnemonico, dobbiamo ora dare una definizione matematicamente rigorosa.

²Il termine anglosassone non ha un corrispettivo convenzionale in Italiano: alcune scelte diffuse sono “*valore-p*” e “*p-dei-dati*”.

Definizione 8.1.5 — p -value. Si assuma data una famiglia di regioni critiche $\{C(\alpha)\}_{\alpha \in (0,1)}$ tali che il test (per il parametro θ con ipotesi nulla $\mathcal{H}_0: \theta \in \Theta_0$) con regione critica $C(\alpha)$ abbia livello α .

Data una realizzazione (x_1, \dots, x_n) del campione (X_1, \dots, X_n) , il p -value è il numero $\bar{\alpha} = \bar{\alpha}(x_1, \dots, x_n)$ tale che:

- se $\alpha < \bar{\alpha}$, l'ipotesi nulla viene accettata dal test di regione critica $C(\alpha)$ (diremo che *viene accettata a livello α*),
- se invece $\alpha > \bar{\alpha}$, l'ipotesi viene rifiutata dal test di regione critica $C(\alpha)$.

Il calcolo del p -value è molto importante sia perché sintetizza in un solo numero la plausibilità di una ipotesi, sia perché, come esito quantitativo tra 0 e 1, fornisce un'informazione più ricca della semplice dicotomia accettazione o rifiuto, peraltro indipendente dal livello α . In pratica, se il p -value è molto basso, ad esempio inferiore a 0.001, l'ipotesi \mathcal{H}_0 è decisamente poco plausibile, se il p -value è alto, ad esempio superiore a 0.3, l'ipotesi \mathcal{H}_0 è molto plausibile, mentre per valori intermedi, il p -value ci dice quanto forte è l'indicazione contro \mathcal{H}_0 (più il p -value è basso, più forte è l'indicazione contro \mathcal{H}_0).

Proprio perché si rinuncia a fissare il livello α , non è possibile fissare un valore standard del p -value per cui si rifiuta l'ipotesi nulla, nello stesso modo in cui non ha senso fissare arbitrariamente un livello α sempre valido. La scelta di α o equivalentemente della soglia del p -value sotto cui rifiutare l'ipotesi nulla dipende dalla situazione. Ad esempio, se modificare la produzione per avere meno pezzi difettosi comporta un grande costo per l'azienda (non supportato dai maggiori ricavi), sceglieremo un livello α molto basso, per cui rifiuteremo l'ipotesi \mathcal{H}_0 (e conseguentemente modificheremo la produzione) solo in presenza di una forte evidenza di più del 2% di pezzi difettosi.

8.2 Z-test

Si chiama **Z-test** il test sulla media di un campione Gaussiano con varianza nota. Iniziamo discutendo il caso particolare in cui l'ipotesi nulla è data da un singolo valore per la media.

Consideriamo dunque un campione X_1, \dots, X_n di legge $N(m, \sigma^2)$, e supponiamo di conoscere il valore di $\sigma > 0$. Vogliamo effettuare un test per decidere se la media m coincide o meno con un certo valore m_0 , ovvero nella terminologia introdotta sopra scegliamo $m = \theta \in \Theta = \mathbb{R}$, e $\Theta_0 = \{m_0\}$, $\Theta_1 = \mathbb{R} \setminus \{m_0\}$. In altre parole, le ipotesi alternative sono

$$\mathcal{H}_0) m = m_0 \quad \text{contro} \quad \mathcal{H}_1) m \neq m_0$$

Come per la determinazione dell'intervallo di fiducia, in presenza di un campione Gaussiano con varianza nota i calcoli sono basati sul fatto che la variabile $Z = \sqrt{n}(\bar{X}_n - m)/\sigma$ è Gaussiana standard.

Formulazione del test

Poiché \bar{X}_n è uno stimatore corretto e consistente di m , l'intuizione ci porta a rifiutare l'ipotesi se \bar{X}_n si scosta troppo da m_0 , cioè a scegliere una regione critica delle forma

$$C = \{|\bar{X}_n - m_0| > d\},$$

dove il numero $d > 0$ deve essere determinato in funzione del livello α scelto. Si deve cioè avere $\mathbb{P}_{m_0}\{|\bar{X}_n - m_0| > d\} \leq \alpha$. Per ottenere una regione critica più grande possibile (allo scopo di aumentare la potenza del test) richiederemo che valga $\mathbb{P}_{m_0}\{|\bar{X}_n - m_0| > d\} = \alpha$.

Abbiamo dunque

$$\mathbb{P}_{m_0}(|\bar{X}_n - m_0| > d) = \mathbb{P}_{m_0}\left(\frac{\sqrt{n}}{\sigma}|\bar{X}_n - m_0| > \frac{d\sqrt{n}}{\sigma}\right) = \mathbb{P}\left(|Z| > \frac{d\sqrt{n}}{\sigma}\right) = \alpha,$$

e questo ci porta a definire la regione critica di livello α tramite:

$$\frac{d\sqrt{n}}{\sigma} = q_{1-\frac{\alpha}{2}}, \quad C = \left\{ \sqrt{n} \frac{|\bar{X}_n - m_0|}{\sigma} > q_{1-\frac{\alpha}{2}} \right\} = \left\{ |\bar{X}_n - m_0| > \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right\}.$$

In pratica di fronte a dati concreti x_1, \dots, x_n , si calcola la media empirica \bar{x}_n e si rifiuta l'ipotesi \mathcal{H}_0 se $|\bar{x}_n - m_0| > \sigma q_{1-\alpha/2}/\sqrt{n}$, si accetta \mathcal{H}_0 in caso contrario.

Notiamo che, analogamente al caso degli intervalli di fiducia, l'ampiezza della regione di accettazione C^c , cioè $2\sigma q_{1-\alpha/2}/\sqrt{n}$:

- cresce al crescere del livello del test $1 - \alpha$;
- cresce al crescere di σ^2 ;
- decresce al crescere di n .

NB Si nota facilmente che l'ipotesi $\mathcal{H}_0: m = m_0$ è accettata al livello α se e solo se m_0 appartiene all'intervallo di fiducia per la media con livello di fiducia $(1 - \alpha)$. Questa è una proprietà generale: si può dimostrare che è equivalente identificare un intervallo di fiducia al livello $(1 - \alpha)$ oppure un test nel quale l'ipotesi è semplice, cioè ridotta a un solo parametro.

Calcolo del p-value

Calcoliamo ora il p -value dei dati (x_1, \dots, x_n) . Procediamo prima usando la definizione informale di p -value come probabilità, sotto \mathcal{H}_0 , di ottenere “dati più estremi”, rispetto all'ipotesi \mathcal{H}_0 , di (x_1, \dots, x_n) . Poiché \bar{X} è la stima del valore atteso m e l'ipotesi nulla \mathcal{H}_0 è $m = m_0$, è intuitivo considerare come “dati più estremi” quelli (y_1, \dots, y_n) che verificano $|\bar{y}_n - m_0| > |\bar{x}_n - m_0|$ (dove \bar{x}_n, \bar{y}_n sono le medie campionarie di (x_1, \dots, x_n) , (y_1, \dots, y_n) rispettivamente). Ci aspettiamo quindi che il p -value di (x_1, \dots, x_n) sia

$$\bar{\alpha} = \bar{\alpha}(x_1, \dots, x_n) = \mathbb{P}_{m_0} \left(\sqrt{n} \frac{|\bar{X}_n - m_0|}{\sigma} > \frac{\sqrt{n}}{\sigma} |\bar{x}_n - m_0| \right) = 2 \left[1 - \Phi \left(\frac{\sqrt{n}}{\sigma} |\bar{x}_n - m_0| \right) \right],$$

in cui l'ultimo passaggio segue ricordando che $Z = \sqrt{n}(\bar{X} - m_0)/\sigma$ è una Gaussiana standard. Verifichiamo ora che il valore $\bar{\alpha}$ trovato sopra soddisfa la definizione rigorosa di p -value.

Dobbiamo verificare che, nel test di livello α , si rifiuta \mathcal{H}_0 se e solo se $\alpha > \bar{\alpha}$. Abbiamo visto che \mathcal{H}_0 è rifiutata al livello α se e solo se $\frac{\sqrt{n}}{\sigma} |\bar{x}_n - m_0| > q_{1-\frac{\alpha}{2}}$. Quindi, in caso di rifiuto, abbiamo

$$\bar{\alpha} = \mathbb{P}_{m_0} \left(\sqrt{n} \frac{|\bar{X}_n - m_0|}{\sigma} > \frac{\sqrt{n}}{\sigma} |\bar{x}_n - m_0| \right) < \mathbb{P}_{m_0} \left(\sqrt{n} \frac{|\bar{X}_n - m_0|}{\sigma} > q_{1-\frac{\alpha}{2}} \right) = \alpha,$$

mentre, in caso di accettazione, vale la disuguaglianza opposta. Dunque $\bar{\alpha}$ soddisfa la definizione rigorosa di p -value.

Calcolo della Curva Operativa

Calcoliamo ora la curva operativa $\beta(m)$ (probabilità di accettare quando m è il valore del parametro) al livello α :

$$\begin{aligned} \beta(m) &= \mathbb{P}_m \left(\frac{\sqrt{n}}{\sigma} |\bar{X}_n - m_0| \leq q_{1-\frac{\alpha}{2}} \right) = \mathbb{P}_m \left(-q_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}}{\sigma} (\bar{X}_n - m_0) \leq q_{1-\frac{\alpha}{2}} \right) \\ &= \mathbb{P}_m \left(\sqrt{n} \frac{(m_0 - m)}{\sigma} - q_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}}{\sigma} (\bar{X}_n - m) \leq \sqrt{n} \frac{(m_0 - m)}{\sigma} + q_{1-\frac{\alpha}{2}} \right) \\ &= \Phi \left(\sqrt{n} \frac{(m_0 - m)}{\sigma} + q_{1-\frac{\alpha}{2}} \right) - \Phi \left(\sqrt{n} \frac{(m_0 - m)}{\sigma} - q_{1-\frac{\alpha}{2}} \right) =: h \left(\sqrt{n} \frac{m - m_0}{\sigma} \right). \end{aligned}$$

Si può dimostrare che la funzione h è pari, cioè

$$h(x) = h(|x|) = \Phi(|x| + q_{1-\alpha/2}) - \Phi(|x| - q_{1-\alpha/2}),$$

ed è decrescente in $x \geq 0$ (ad α fissato). Quindi più $\sqrt{n} \frac{|m-m_0|}{\sigma}$ è grande, più piccola è la curva operativa $\beta(m)$, e quindi più grande è la potenza $1 - \beta(m)$. Ne segue che, per aumentare la potenza lasciando fisso il livello α del test, si può aumentare la taglia n del campione. Più precisamente, fissati $m \neq m_0$ e $\beta_0 \in (0, 1)$, usando la semplice stima

$$\beta(m) = \Phi\left(\sqrt{n} \frac{|m_0 - m|}{\sigma} + q_{1-\frac{\alpha}{2}}\right) - \Phi\left(\sqrt{n} \frac{|m_0 - m|}{\sigma} - q_{1-\frac{\alpha}{2}}\right) \leq 1 - \Phi\left(\sqrt{n} \frac{|m_0 - m|}{\sigma} - q_{1-\frac{\alpha}{2}}\right)$$

e imponendo che

$$1 - \Phi\left(\sqrt{n} \frac{|m_0 - m|}{\sigma} - q_{1-\frac{\alpha}{2}}\right) \leq \beta_0,$$

si trova il valore n per cui la curva operativa del test è inferiore a β_0 (cioè la potenza in m è superiore a $1 - \beta_0$). Si può infine dimostrare che, a m fissato, la curva operativa aumenta all'aumentare di α .

Esercizio 8.1 Il peso (in kg) di una popolazione di salmoni segue una distribuzione gaussiana di media m non nota e deviazione standard 0.5. Viene rilevato il peso su un campione di 16 esemplari. Si richiede di:

- formulare un test, di livello 5%, per decidere se l'ipotesi \mathcal{H}_0 “peso medio pari a 3” è plausibile oppure no, e applicarlo nel caso in cui la media sul campione sia 3.2;
- determinare il p -value per la media campionaria 3.2;
- calcolare la probabilità dell'errore di seconda specie (accettare \mathcal{H}_0 quando questa è falsa) assumendo $m = 3.4$;
- determinare la minima taglia n del campione per cui il test abbia potenza almeno 95% nel punto 3.4.

Soluzione. Sia X il peso di un esemplare della popolazione e siano X_1, \dots, X_n i pesi degli esemplari del campione ($n = 16$), sappiamo che X ha distribuzione $N(m, \sigma^2)$ con $\sigma = 0.5$. L'ipotesi è $\mathcal{H}_0 : m = 3$ contro $\mathcal{H}_1 : m \neq 3$, il livello α è 0.05 ($q_{0.975} = 1.96$) e la regione critica C è

$$C = \left\{ |\bar{X} - 3| > \frac{\sigma}{\sqrt{n}} q_{1-\alpha/2} \right\} = \{ |\bar{X} - 3| > 0.245 \}.$$

Dunque, se la media campionaria vale $\bar{x} = 3.2$, i dati cadono nella regione di accettazione C^c e quindi l'ipotesi \mathcal{H}_0 : “peso medio pari a 3” viene accettata a livello $\alpha = 0.05$: \mathcal{H}_0 è plausibile a questo livello.

Il p -value corrispondente a $\bar{x} = 3.2$ è

$$\bar{\alpha} = \mathbb{P}\{|Z| > \frac{\sqrt{n}}{\sigma} |\bar{x} - 3|\} = 2 \left(1 - \Phi\left(\frac{\sqrt{16}}{0.5} |3.2 - 3|\right) \right) = 2(1 - 0.9452) = 0.1096,$$

dunque, per $\bar{x} = 3.2$, rifiutiamo per $\alpha > 0.1096$ e accettiamo per $\alpha < 0.1096$.

La probabilità dell'errore di seconda specie per $m = 3.4$ è il valore della curva operativa del test nel punto 3.4: questo è

$$\begin{aligned} \beta(3.4) &= \Phi\left(\sqrt{n} \frac{|m_0 - m|}{\sigma} + q_{1-\frac{\alpha}{2}}\right) - \Phi\left(\sqrt{n} \frac{|m_0 - m|}{\sigma} - q_{1-\frac{\alpha}{2}}\right) \\ &= \Phi\left(\sqrt{16} \frac{|3 - 3.4|}{0.5} + 1.96\right) - \Phi\left(\sqrt{16} \frac{|3 - 3.4|}{0.5} - 1.96\right) \\ &= \Phi(5.16) - \Phi(1.24) = 1 - 0.8925 = 0.1075. \end{aligned}$$

In altre parole, se il peso medio è 3.4, c'è una probabilità di quasi l'11% di accettare l'ipotesi "peso medio pari a 3".

Perché il test abbia potenza almeno 0.95, imponiamo

$$1 - \Phi\left(\sqrt{n}\frac{|m_0 - m|}{\sigma} - q_{1-\frac{\alpha}{2}}\right) = 1 - \Phi(\sqrt{n} \cdot 0.8 - 1.96) \leq \beta_0 = 1 - 0.95 = 0.05,$$

cioè $\sqrt{n} \cdot 0.8 - 1.96 \geq q_{0.95} = 1.64$, cioè $n \geq 20.25$, quindi 21 è la minima taglia del campione richiesta. ■

Z-test unilatero

Occupiamoci ora, questa volta senza ripetere tutti i dettagli, dello Z-test per la media di un campione Gaussiano a varianza nota con ipotesi alternative

$$\mathcal{H}_0)m \leq m_0 \quad \text{contro} \quad \mathcal{H}_1)m > m_0.$$

Lo sviluppo del test è analogo al caso bilatero, e ometteremo i dettagli.

- **Formulazione del test:** L'intuizione ci spinge a rifiutare l'ipotesi se $(\bar{X}_n - m_0)$ è troppo grande, cioè a considerare una regione critica della forma $C = ((\bar{X}_n - m_0) > d)$, e la condizione sul livello diventa

$$\forall m \leq m_0, \quad \mathbb{P}_m((\bar{X}_n - m_0) > d) \leq \alpha.$$

Si vede facilmente che la probabilità sopra scritta cresce al crescere di m e si arriva (procedendo come nel caso precedente) a

$$\mathbb{P}_{m_0}((\bar{X}_n - m_0) > d) = \mathbb{P}_{m_0}\left(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - m_0) > \frac{d\sqrt{n}}{\sigma}\right) = \alpha,$$

da cui segue $\frac{d\sqrt{n}}{\sigma} = q_{1-\alpha}$: la *regione critica* al livello α diventa pertanto

$$C = \left\{ \sqrt{n}\frac{(\bar{X}_n - m_0)}{\sigma} > q_{1-\alpha} \right\} = \left\{ (\bar{X}_n - m_0) > \frac{\sigma}{\sqrt{n}}q_{1-\alpha} \right\}.$$

La dipendenza dell'ampiezza della regione di accettazione C^c dai parametri α , σ , n è la stessa del caso bilatero.

- **Calcolo del p-value:** Calcoliamo il p -value dei dati (x_1, \dots, x_n) . In questo caso, i dati (y_1, \dots, y_n) sono "più estremi" di (x_1, \dots, x_n) rispetto ad $\mathcal{H}_0 : m \leq m_0$ se $\bar{y}_n - m_0 > \bar{x}_n - m_0$. La probabilità, per $m \leq m_0$ di avere "dati più estremi" di (x_1, \dots, x_n) è

$$\mathbb{P}_m\left(\sqrt{n}\frac{(\bar{X}_n - m_0)}{\sigma} > \frac{\sqrt{n}}{\sigma}(\bar{x}_n - m_0)\right)$$

Come p -value si prende il massimo di tali probabilità per $m \in \mathcal{H}_0 : m \leq m_0$, che si realizza in m_0 : il p -value diventa

$$\bar{\alpha}(x_1, \dots, x_n) = \mathbb{P}_{m_0}\left(\sqrt{n}\frac{(\bar{X}_n - m_0)}{\sigma} > \frac{\sqrt{n}}{\sigma}(\bar{x}_n - m_0)\right) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(\bar{x}_n - m_0)\right)$$

e si può verificare che tale valore verifica la definizione rigorosa di p -value.

- **Curva operativa:** La curva operativa è la funzione

$$\beta(m) = \mathbb{P}_m \left(\frac{\sqrt{n}}{\sigma} (\bar{X}_n - m_0) \leq q_{1-\alpha} \right) = \Phi \left(\sqrt{n} \frac{(m_0 - m)}{\sigma} + q_{1-\alpha} \right).$$

Poiché Φ è crescente, per $m > m_0$ la curva operativa è decrescente in $\sqrt{n} \frac{(m_0 - m)}{\sigma}$. In particolare, a m e β_0 fissati, imponento $\beta(m) = \beta_0$ si può trovare n tale che il test abbia potenza $1 - \beta_0$ per il valore m fissato.

Se l'ipotesi è della forma $\mathcal{H}_0: m \geq m_0$, si invertono tutte le disequaglianze e si sostituisce $q_{1-\alpha}$ con q_α . Precisamente, la regione critica è

$$C = \left\{ \sqrt{n} \frac{(\bar{X}_n - m_0)}{\sigma} < q_\alpha \right\},$$

il p -value è

$$\bar{\alpha}(x_1, \dots, x_n) = \mathbb{P}_{m_0} \left(\sqrt{n} \frac{(\bar{X}_n - m_0)}{\sigma} < \frac{\sqrt{n}}{\sigma} (\bar{x}_n - m_0) \right) = \Phi \left(\frac{\sqrt{n}}{\sigma} (\bar{x}_n - m_0) \right)$$

e la curva operativa è

$$\beta(m) = \mathbb{P}_m \left(\frac{\sqrt{n}}{\sigma} (\bar{X}_n - m_0) \geq q_\alpha \right) = 1 - \Phi \left(\sqrt{n} \frac{(m_0 - m)}{\sigma} + q_\alpha \right).$$

Esercizio 8.2 Ripetere l'esercizio svolto di cui sopra (peso di un campione di salmone), nel caso del test unilatero con \mathcal{H}_0 : “peso medio non superiore a 3”.

8.3 Test sulla media di un campione Gaussiano con varianza sconosciuta, o T-test

Consideriamo ora il caso (più realistico) di campione gaussiano di cui non conosciamo la deviazione standard: si chiama **T-test** il test sulla media di un campione Gaussiano con varianza sconosciuta.

Sia X_1, \dots, X_n un campione aleatorio di legge gaussiana $N(m, \sigma^2)$, e ricordiamo la notazione

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}, \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Il procedimento per formulare il test è analogo a quanto visto per lo Z-test, ma questa volta il punto di partenza è il fatto che, se la media è m , la variabile $T = \sqrt{n}(\bar{X}_n - m)/S_n$ ha densità di Student $T(n-1)$. Consideriamo il test **bilatero** dell'ipotesi

$$\mathcal{H}_0: m = m_0 \quad \text{contro} \quad \mathcal{H}_1: m \neq m_0.$$

- **Formulazione del test:** Anche in questo caso cerchiamo una regione critica della forma $C = \{|\bar{X}_n - m_0| > d\}$. Imponiamo $\mathbb{P}_{m_0}(C) = \alpha$ e usiamo la statistica $T = \sqrt{n}(\bar{X}_n - m)/S$:

$$\mathbb{P}_{m_0}(|\bar{X}_n - m_0| > d) = \mathbb{P}_{m_0} \left(\frac{\sqrt{n}}{S} |\bar{X}_n - m_0| > \frac{d\sqrt{n}}{S} \right) = \mathbb{P} \left(|T| > \frac{d\sqrt{n}}{S} \right) = \alpha$$

e quindi $d\sqrt{n}/S_n = \tau_{1-\frac{\alpha}{2}, n-1}$, dove $\tau_{\beta, k}$ è il β -quantile della densità di Student $T(k)$. La regione critica a livello α è quindi

$$C = \left\{ \sqrt{n} \frac{|\bar{X}_n - m_0|}{S} > \tau_{(1-\frac{\alpha}{2}, n-1)} \right\}.$$

- **p-value:** Procedendo in modo analogo al caso σ nota, ma usando la statistica $T = \sqrt{n}(\bar{X}_n - m)/S$, otteniamo la formula per il p -value dei dati (x_1, \dots, x_n) :

$$\mathbb{P}_{m_0} \left(\sqrt{n} \frac{|\bar{X}_n - m_0|}{S} > \frac{\sqrt{n}}{s} |\bar{x}_n - m_0| \right) = 2 \left[1 - F_{n-1} \left(\frac{\sqrt{n}}{s} |\bar{x}_n - m_0| \right) \right],$$

dove \bar{x} e \bar{s} sono rispettivamente la media campionaria e la varianza campionaria di (x_1, \dots, x_n) e dove F_k è la c.d.f. della variabile $T(k)$ (che i software sono in grado di calcolare con ottima approssimazione).

- **Curva operativa:** In questo caso, *la curva operativa non è ben definita*. Il motivo è il seguente: se il valore atteso delle variabili è m_0 , qualunque sia σ la variabile $T = \sqrt{n}(\bar{X}_n - m_0)/S$ ha densità di Student $T(n-1)$, ma se la media è diversa da m_0 la densità di T non dipende solo da m ma anche da σ , dunque non possiamo applicare la definizione.

In modo simile si ottengono le formule per il test unilatero. Nel caso

$$\mathcal{H}_0) m \leq m_0 \quad \text{contro} \quad \mathcal{H}_1) m > m_0$$

la regione critica a livello α e il p -value sono dati rispettivamente da

$$C = \left\{ \sqrt{n} \frac{(\bar{X}_n - m_0)}{S_n} > \tau_{1-\alpha, n-1} \right\},$$

$$\bar{\alpha}(x_1, \dots, x_n) = \mathbb{P}_{m_0} \left(\sqrt{n} \frac{(\bar{X}_n - m_0)}{S_n} > \frac{\sqrt{n}}{s} (\bar{x}_n - m_0) \right) = 1 - F_{n-1} \left(\frac{\sqrt{n}}{s} (\bar{x}_n - m_0) \right).$$

Discorso analogo per $\mathcal{H}_0) m \geq m_0$.

■ **Esempio 8.5** Riprendiamo l'esempio del peso (in kg) di una popolazione di salmoni, supponendo sempre una distribuzione gaussiana ma questa volta di varianza non nota (e anche di media m non nota). Viene rilevato il peso su un campione di 16 esemplari. Vogliamo formulare un test, di livello 5%, per decidere se l'ipotesi \mathcal{H}_0 "peso medio pari a 3" è plausibile oppure no, e applicarlo nel caso in cui le misurazioni abbiano media campionaria 3.2 e deviazione standard campionaria 0.5.

Consideriamo un campione aleatorio X_1, \dots, X_{16} di legge $N(m, \sigma^2)$ con m e σ^2 entrambi non noti. L'ipotesi è $\mathcal{H}_0 : m = 3$ contro $\mathcal{H}_1 : m \neq 3$, il livello α è 0.05 ($\tau_{0.975, 15} = 2.13$) e la regione critica C è

$$C = \left\{ \sqrt{n} \frac{|\bar{X} - 3|}{S} > \tau_{1-\alpha/2, n-1} \right\} = \left\{ \frac{|\bar{X} - 3|}{S} > 0.5325 \right\}.$$

Dunque, se media campionaria e varianza campionaria valgono $\bar{x} = 3.2$ e $s = 0.5$, anche in questo caso i dati cadono nella regione di accettazione C^c e quindi l'ipotesi \mathcal{H}_0 : "peso medio pari a 3" viene accettata a livello $\alpha = 0.05$. Il p -value vale:

$$\mathbb{P}_3 \left(\sqrt{n} \frac{|\bar{X}_n - 3|}{S} > \frac{\sqrt{16}}{0.5} |3.2 - 3| \right) = 2 \left[1 - F_{15} \left(\frac{\sqrt{16}}{0.5} |3.2 - 3| \right) \right] = 2(1 - 0.9347) = 0.1306.$$

Anche se il valore di s è uguale al valore $\sigma = 0.5$ dell'esempio con varianza nota, il p -value è più alto nel caso nel caso varianza non nota: questo perché, come per gli intervalli di fiducia, l'incertezza sulla varianza si traduce in un'ampiezza maggiore della regione di accettazione. ■

8.4 Test approssimato su un campione di Bernoulli

Sia X_1, \dots, X_n un campione i.i.d. di variabili di Bernoulli di parametro $p \in (0, 1)$: vogliamo formulare un test relativo al parametro p . Esattamente come per gli intervalli di fiducia approssimati, usiamo il fatto che, *per n grande*, la variabile $\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{p(1-p)}}$ è approssimativamente Gaussiana standard.

Consideriamo ad esempio il test **bilatero** dell'ipotesi

$$\mathcal{H}_0) p = p_0 \quad \text{contro} \quad \mathcal{H}_1) p \neq p_0$$

al livello α , assumendo che n sia grande.

- **Formulazione del test:** La regione critica approssimata è

$$C = \left\{ \frac{\sqrt{n} |\bar{X}_n - p_0|}{\sqrt{p_0(1-p_0)}} > q_{1-\frac{\alpha}{2}} \right\}.$$

- **p-value:** Il p -value è

$$\mathbb{P}_{p_0} \left(\frac{\sqrt{n} |\bar{X}_n - p_0|}{\sqrt{p_0(1-p_0)}} > \frac{\sqrt{n} |\bar{x} - p_0|}{\sqrt{p_0(1-p_0)}} \right) \approx 2 \left[1 - \Phi \left(\frac{\sqrt{n} |\bar{x} - p_0|}{\sqrt{p_0(1-p_0)}} \right) \right]$$

- **Curva Operativa:** A differenza del caso varianza sconosciuta (nel quale la curva operativa *non era ben definita*) questa volta la curva operativa può essere calcolata approssimativamente. Tuttavia i calcoli sono più laboriosi di quelli fatti per il caso del test sulla media di un campione Gaussiano con varianza nota, poiché quando cambia p cambia anche la varianza. Per questo non li riportiamo.

Anche per quanto riguarda i test *unilateri* si procede in analogia con il caso Gaussiano, lasciamo i dettagli per esercizio.

■ **Esempio 8.6** In un controllo qualità su un certo tipo di pezzi, vengono estratti e verificati 1000 pezzi. Vogliamo formulare un test, di livello 5%, per decidere se l'ipotesi \mathcal{H}_0 “percentuale di pezzi difettosi non superiore al 2%” è plausibile o no, e applicarlo nel caso in cui vengono rilevati 25 pezzi difettosi (tra i 1000 estratti).

Consideriamo un campione aleatorio X_1, \dots, X_{1000} di variabili Bernoulli di parametro p (con esito 1 che indica un pezzo estratto difettoso, 0 altrimenti). L'ipotesi è $\mathcal{H}_0 : p \leq 0.02$ contro $\mathcal{H}_1 : p > 0.02$, il livello α è 0.05 ($q_{0.95} = 1.64$). La regione critica approssimata è

$$C = \left\{ \sqrt{n} \frac{\bar{X} - 0.02}{\sqrt{0.02(1-0.02)}} > q_{1-\alpha} \right\} = (\bar{X} - 0.02 > 0.0073).$$

Se il numero di pezzi difettosi trovati è 25, allora la media campionaria \bar{x} (frequenza relativa campionaria) vale $25/1000 = 0.025$ e quindi i dati cadono nella regione (approssimata) di accettazione C^c : l'ipotesi $\mathcal{H}_0 : p \leq 0.02$ viene accettata (è plausibile) a livello $\alpha = 0.05$. Per $\bar{x} = 0.025$,

$$\mathbb{P}_{0.02} \left(\frac{\sqrt{n}(\bar{X} - 0.02)}{\sqrt{0.02 \cdot 0.98}} > \frac{\sqrt{n}(\bar{x} - 0.02)}{\sqrt{0.02 \cdot 0.98}} \right) \approx 1 - \Phi \left(\frac{\sqrt{n}\bar{x} - 0.02}{\sqrt{0.02 \cdot 0.98}} \right) = (1 - \Phi(1.13)) = 0.129$$

restituisce il p -value approssimato. ■

8.5 Test approssimato sulla media di un campione di taglia grande

Come per gli intervalli di fiducia, l'approssimazione tramite il Teorema Centrale del Limite si può usare in generale per formulare test sulla media di una popolazione non necessariamente gaussiana, se il campione ha *taglia n grande*.

Sia X una v.a. con media m (non nota) e varianza σ^2 nota, sia X_1, \dots, X_n un campione i.i.d. con *n grande*. Per il Teorema Centrale del Limite, la statistica $Z = \sqrt{n}(\bar{X} - m)/\sigma$ ha distribuzione approssimativamente gaussiana standard. Quindi valgono le stesse formule viste per il caso di campioni gaussiani a varianza nota. Ad esempio, per il test bilatero

$$\mathcal{H}_0)m = m_0, \quad \text{contro} \quad \mathcal{H}_1)m \neq m_0$$

al livello α , la regione critica approssimata è

$$C = \left\{ \sqrt{n} \frac{|\bar{X}_n - m_0|}{\sigma} > q_{1-\frac{\alpha}{2}} \right\}$$

e il p -value di (x_1, \dots, x_n) è

$$\mathbb{P} \left(|Z| > \frac{\sqrt{n}}{\sigma} |\bar{x}_n - m_0| \right) \approx 2 \left[1 - \Phi \left(\frac{\sqrt{n}}{\sigma} |\bar{x}_n - m_0| \right) \right].$$

Nel caso di varianza non nota, (sempre con *n grande*), per il corollario al Teorema Centrale del Limite, la statistica $Z = \sqrt{n}(\bar{X} - m)/S$ ha distribuzione approssimativamente gaussiana standard. Quindi valgono ancora le formule viste per il caso di campioni gaussiani a varianza nota, ma rimpiazzando σ con S .

8.6 Test sulla varianza di un campione Gaussiano

Come per gli intervalli di fiducia, ci occupiamo solo del caso unilatero,

$$\mathcal{H}_0)\sigma^2 \leq \sigma_0^2 \quad \text{contro} \quad \mathcal{H}_1)\sigma^2 > \sigma_0^2,$$

ricordando che, se il campione X_1, \dots, X_n è formato da variabili Gaussiane con varianza σ^2 , la variabile $\frac{n-1}{\sigma^2} S_n^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ ha densità $\chi^2(n-1)$.

Formulazione del test: Risulta naturale considerare una regione critica della forma $C = \left\{ \sum_{i=1}^n (X_i - \bar{X}_n)^2 > d \right\}$, in cui d viene scelto imponendo livello α , ovvero chiedendo che per $\sigma \leq \sigma_0$,

$$\mathbb{P}_\sigma \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 > d \right) = \mathbb{P}_\sigma \left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2} > \frac{d}{\sigma_0^2} \right) \leq \alpha.$$

È facile provare che la probabilità sopra scritta cresce con σ e, cercando una regione critica più piccola possibile (purché di livello α) si ottiene

$$\mathbb{P}_{\sigma_0} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 > d \right) = \mathbb{P}_{\sigma_0} \left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2} > \frac{d}{\sigma_0^2} \right) = \alpha,$$

e questo ci porta a scegliere $\frac{d}{\sigma_0^2} = \chi_{1-\alpha, n-1}^2$. Si ottiene quindi la regione critica di livello α

$$C = \left\{ \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2} > \chi_{1-\alpha, n-1}^2 \right\} = \left\{ \frac{n-1}{\sigma_0^2} S_n^2 > \chi_{1-\alpha, n-1}^2 \right\}.$$

p-value: Il p -value dei dati (x_1, \dots, x_n) è

$$\mathbb{P}_{\sigma_0} \left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2} > \frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sigma_0^2} \right) = 1 - G_{n-1} \left(\frac{\sum_{i=1}^n (x_i - \bar{x}_n)^2}{\sigma_0^2} \right),$$

dove con G_n si indica la c.d.f. della variabile $\chi^2(n)$.

Curva operativa: La curva operativa è

$$\begin{aligned} \beta(\sigma) &= \mathbb{P}_{\sigma} \left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma_0^2} \leq \chi_{1-\alpha, n-1}^2 \right) \\ &= \mathbb{P}_{\sigma} \left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \leq \frac{\sigma_0^2}{\sigma^2} \chi_{1-\alpha, n-1}^2 \right) = G_{n-1} \left(\frac{\sigma_0^2}{\sigma^2} \chi_{1-\alpha, n-1}^2 \right). \end{aligned}$$

Se si considera un test unilatero nell'altra direzione, cioè dell'ipotesi $\mathcal{H}_0) \sigma^2 \geq \sigma_0^2$, la discussione è adattata invertendo tutte le disuguaglianze ed cambiando il quantile $\chi_{1-\alpha, n-1}^2$ con $\chi_{\alpha, n-1}^2$.

8.7 Confronto tra due campioni statistici

Introduciamo l'argomento di questo paragrafo con due esercizi, apparentemente simili.

Esercizio 8.3 Viene propagandata una cura dimagrante che promette *7 chili in 7 giorni*: vengono misurati “prima e dopo” 7 pazienti ottenendo i valori che seguono:

- “prima”: 72 89 94 77 86 91 83,
- “dopo”: 68 83 90 71 80 88 74.

Si può accettare l'affermazione dell'istituto che offre la cura? Quale sarebbe stata la conclusione se invece di promettere 7 chili in 7 giorni avessero promesso solo 6 chili in 7 giorni? ■

Esercizio 8.4 Vengono misurate le lunghezze delle tibie di uomini adulti provenienti da reperti di tombe etrusche: dal sito di Cerveteri vengono effettuate 13 misurazioni ottenendo un valore medio di 47.2 ed una varianza campionaria di 7.98, mentre dal sito di Ladispoli si ottengono 8 misurazioni con un valore medio di 44.9 ed una varianza campionaria di 8.85.

Si può affermare che la differenza sia una semplice fluttuazione statistica oppure si deve concludere che gli abitanti di Cerveteri erano veramente più alti? ■

Fermo restando che questi campioni sono troppo poco numerosi per trarre delle conclusioni veramente significative, si tratta di due situazioni *radicalmente diverse*. Infatti, nel primo caso (quello della cura dimagrante) i campioni sono *accoppiati*: se chiamiamo X_i, Y_i il peso della i -sima persona del campione rispettivamente prima e dopo la cura, X_i e Y_i non sono indipendenti (se sappiamo ad esempio che X_i è molto alto, la probabilità di avere anche Y_i alto sarà più alta). Invece, nel secondo caso (quello delle tibie degli etruschi), i campioni sono *indipendenti*: se chiamiamo X_i la lunghezza dell' i -simo elemento del campione del sito di Cerveteri, Y_j la lunghezza del j -simo elemento del campione del sito di Ladispoli, X_i e Y_j sono indipendenti (informazioni su X_i non cambiano le probabilità relative agli Y_j).

Nel caso dei campioni *accoppiati* non c'è sostanzialmente nulla di nuovo da aggiungere dal punto di vista teorico: si considera il campione $V_i = X_i - Y_i$ dato dalle differenze delle coppie e, se possibile, si esegue su questo uno dei test già visti (a seconda delle ipotesi nulla e alternativa e della distribuzione del campione).

Soluzione dell'Esercizio 8.3. Siano X_i, Y_i il peso dell' i -sima persona del campione rispettivamente prima e dopo la cura. Chiamiamo $V_i = X_i - Y_i$ e supponiamo che V_i formino un campione di legge gaussiana, con varianza sconosciuta. Si noti che il valore atteso m di V_i è la differenza media (sulla popolazione) del peso prima e dopo la cura, cioè la diminuzione media del peso con la cura. L'ipotesi nulla è $\mathcal{H}_0)m \geq 7$ contro l'ipotesi alternativa $\mathcal{H}_1)m < 7$. Usiamo quindi un t -test unilatero per il campione gaussiano V_i con varianza sconosciuta. Prendiamo la statistica $T = \sqrt{7} \frac{(\bar{V}-7)}{S}$, dove \bar{V} e S sono rispettivamente media campionaria e varianza campionaria. Il p -value del test è

$$\mathbb{P}\{T \leq \sqrt{7} \frac{(\bar{v}-7)}{s}\} = F_6 \left(\sqrt{7} \frac{(\bar{v}-7)}{s} \right),$$

con \bar{v}, s media e varianza campionaria dei dati v_i del campione V_i . Con dei conti elementari,

$$\sqrt{7} \frac{(\bar{v}-7)}{s} = -2.09, \quad F_6 \left(\sqrt{7} \frac{(\bar{v}-7)}{s} \right) = 0.04,$$

cioè l'ipotesi \mathcal{H}_0 è molto poco credibile. Se si considera l'ipotesi $\mathcal{H}_0)m \geq 6$, risulta

$$\sqrt{7} \frac{(\bar{v}-6)}{s} = -0.76, \quad F_6 \left(\sqrt{7} \frac{(\bar{v}-6)}{s} \right) = 0.23,$$

e questa volta l'ipotesi è molto più verosimile. ■

Quando invece i due campioni sono *indipendenti* (tra l'altro in genere di numerosità diverse), la situazione è radicalmente diversa. Il confronto tra due o più campioni aleatori diversi rientra nell'ambito dell'inferenza statistica detto *analisi della varianza*, argomento ampio di cui ci limitiamo a illustrare un primo passo. Ci occupiamo qui del caso di campioni gaussiani indipendenti in cui le varianze teoriche dei due campioni sono sconosciute ma uguali. Il risultato teorico che è alla base del confronto tra le medie di due campioni è il risultato seguente:

Teorema 8.7.1 Siano X_1, \dots, X_n un campione con densità $N(m_1, \sigma^2)$ e Y_1, \dots, Y_k un campione indipendente dal primo con densità $N(m_2, \sigma^2)$ (si suppone cioè che le varianze siano sconosciute ma eguali): la variabile

$$T_{n,k} = \frac{(\bar{X}_n - \bar{Y}) - (m_1 - m_2)}{\sqrt{\sum_i (X_i - \bar{X}_n)^2 + \sum_j (Y_j - \bar{Y})^2}} \frac{\sqrt{n+k-2}}{\sqrt{\frac{1}{n} + \frac{1}{k}}}$$

ha densità di Student $T_{(n+k-2)}$ (cioè a $(n+k-2)$ gradi di libertà).

Si può quindi formulare un test t su $m_1 - m_2$, la differenza delle medie nelle due popolazioni.

Naturalmente rimane aperto il problema di decidere se è verosimile che i due campioni abbiano la stessa varianza: per questo scopo esiste il cosiddetto F-test (o test di Fisher) che però non illustriamo.

Soluzione esercizio 8.4. Sia X_i la lunghezza della tibia dell' i -simo elemento del campione di Cerveteri, $i = 1, \dots, n = 13$, e sia Y_j la lunghezza della tibia del j -simo elemento del campione di Ladispoli, $j = 1, \dots, k = 8$. Supponiamo che i due campioni siano gaussiani e che le loro varianze siano uguali (le varianze campionarie sono abbastanza simili). Chiamiamo m_1, m_2 i valori attesi per la lunghezza della tibia per gli uomini di Cerveteri e di Ladispoli rispettivamente, poniamo $m = m_1 - m_2$. Consideriamo il test dell'ipotesi $\mathcal{H}_0)m = 0$ (cioè stessa media) contro $\mathcal{H}_1)m > 0$ (cioè la media a Cerveteri è superiore alla media a Ladispoli).

Formulazione del test: Per il teorema precedente, sotto l'ipotesi $\mathcal{H}_0 : m = 0$, la variabile

$$T = \frac{\bar{X}_n - \bar{Y}}{\sqrt{\sum_i (X_i - \bar{X}_n)^2 + \sum_j (Y_j - \bar{Y})^2}} \frac{\sqrt{19}}{\sqrt{\frac{1}{13} + \frac{1}{8}}}$$

ha densità di Student $T(19)$. Usiamo quindi un t-test unilatero: la regione critica di livello α è

$$C = \{T > \tau_{1-\alpha, 19}\}.$$

p-value: Il valore assunto da T per i dati raccolti è $t = \frac{47.2 - 44.9}{\sqrt{7.98 \cdot 12 + 8.85 \cdot 7}} \frac{\sqrt{19}}{\sqrt{\frac{1}{13} + \frac{1}{8}}} = 1.776$. Il p -value che risulta è quindi

$$\mathbb{P}(T > t) = 1 - F_{19}(1.776) = 0.045$$

con F_{19} cdf della distribuzione $T(19)$. Si tratta di un valore piuttosto basso (anche se non trascurabile). Ad esempio a livello $\alpha = 0.07$, si conclude che gli abitanti di Cerveteri erano effettivamente più alti (mentre non si può avallare tale conclusione per $\alpha = 0.01$). ■

8.8 Esercizi

Esercizio 8.5 Una ditta produce aste e dichiara una lunghezza media di 2.3m. Viene rilevata la lunghezza per 100 aste, ottenendo una media campionaria di 2.317m e deviazione standard campionaria di 0.1m. L'affermazione della ditta sulla lunghezza media è plausibile o no? a quale livello? ■

Esercizio 8.6 Un certo tipo di misurazioni del pH di una sostanza produce misure distribuite in modo gaussiano, con media pari al valore autentico del pH e deviazione standard $\sigma = 0.02$. Dieci misurazioni forniscono una media campionaria di 8.179. (i) Formulare un test di livello 0.05 per decidere se l'ipotesi "valore del pH pari o superiore a 8.2" sia plausibile o no e applicarlo al valore 8.179 assunto dalla media campionaria. (ii) Qual è il p -value corrispondente? (iii) Qual è la potenza del test in 8.1? ■

Esercizio 8.7 Una ditta produce chiodi di lunghezza dichiarata uguale a 5 cm, e il proprietario della ditta afferma che la deviazione standard delle lunghezze dei chiodi prodotti non supera 0.2 cm. Analizzando la lunghezza di un campione di 16 pezzi si trova media campionaria di 4.935 cm e varianza campionaria 0.06 cm. (i) Supponendo che le lunghezze dei chiodi possano essere rappresentate con variabili aleatorie Gaussiane, si può accettare al livello 0.05 l'affermazione del proprietario della ditta sulla deviazione standard della lunghezza dei chiodi prodotti? (ii) Descrivere un test da utilizzare per verificare l'ipotesi che la lunghezza dei chiodi prodotti sia di 5 cm, e usare i dati del campione analizzato per determinare la plausibilità di questa ipotesi. ■

Esercizio 8.8 Una ditta produce una lozione per la ricrescita dei capelli ed afferma che si nota una ricrescita in almeno il 60% dei casi: tuttavia l'unione consumatori ha effettuato un'indagine ed ha rilevato che su 137 persone che hanno usato quella lozione solo 70 hanno notato una ricrescita dei capelli ed afferma che questa indagine contraddice l'affermazione della ditta. In termini di un modello statistico, se p è la percentuale (sconosciuta) delle persone sulle quali la lozione ha effetti positivi, l'affermazione della ditta si traduce nell'ipotesi

$$H_0) p \geq 0.6 \text{ contro } H_1) p < 0.6.$$

(i) Si può accettare, al livello 0.05, l'ipotesi sopra indicata (cioè l'affermazione della ditta)? (ii) A quale livello (approssimativamente) si può accettare l'affermazione della ditta? ■

Esercizio 8.9 Si hanno a disposizione 16 osservazioni indipendenti di una v.a. Gaussiana con media μ sconosciuta e varianza nota eguale a 36; con queste osservazioni si vuole effettuare il test dell'ipotesi $H_0) \mu = 30$ contro $H_1) \mu \neq 30$.

Si decide di respingere l'ipotesi H_0 se la media campionaria delle 16 osservazioni cade al di fuori dell'intervallo (26.91, 33.09). (i) A quale livello viene effettuato il test? (ii) Se le osservazioni fossero 25 e il test venisse effettuato ancora allo stesso livello del punto (i), quale sarebbe la regione critica? ■

Esercizio 8.10 Il responsabile di una ditta petrolifera afferma che il contenuto medio di zolfo per litro, nella benzina prodotta da quella ditta, non supera 0.15mg/l; tuttavia l'unione consumatori contesta questa affermazione perché sono stati prelevati 41 campioni che hanno dato valori x_1, \dots, x_{41} dai quali si ottiene

$$\bar{x} = \frac{x_1 + \dots + x_{41}}{41} = 0.2.$$

Il responsabile afferma che questo dato non è significativo poiché la variabilità era alta: si è infatti ottenuto il valore $\sum_{i=1}^{41} (x_i - \bar{x})^2 = 1$. (Si interpretino i valori delle misurazioni come variabili Gaussiane con media e varianza ignote.) (i) Si può accettare l'affermazione del responsabile della ditta (cioè l'ipotesi che il contenuto medio di zolfo non superi 0.15mg/l)? Impostare un opportuno test e calcolare il relativo p -value. (ii) Scrivere l'intervallo di fiducia unilatero destro (della forma $[a, +\infty)$) per il contenuto medio di zolfo al livello del 95%. ■

Esercizio 8.11 Vengono effettuate 25 misurazioni del peso di un piccolo mammifero africano, e si ottengono i risultati (x_1, \dots, x_{25}) (espressi in grammi) dai quali si ricava una media campionaria eguale a 648.6 ed una varianza campionaria eguale a 457.66.

Si vuole verificare l'ipotesi che il peso medio di questo mammifero sia di 640 grammi, e a tal fine si suppone che i dati possano essere rappresentati con 25 variabili Gaussiane indipendenti. (i) Effettuare il test dell'ipotesi

$$H_0) m = 640 \quad \text{contro} \quad H_1) m \neq 640$$

supponendo che la varianza sia nota ed eguale a 400: dopo aver calcolato il p -value, che cosa si conclude? (ii) Effettuare lo stesso test sopra scritto, supponendo però che la varianza sia sconosciuta: che cosa si conclude? ■

Esercizio 8.12 Si consideri un monitoraggio sulla presenza di sostanze tossiche nell'aria, effettuato in 10 stazioni di monitoraggio vicine. I valori ottenuti restituiscono una concentrazione media di 4.8mg/dm³ con una varianza campionaria di 0.49mg/dm³. Supponiamo che la distribuzione della concentrazione sia Gaussiana. (i) Fornire una stima della concentrazione delle sostanze tossiche con una fiducia del 90% mediante un intervallo bilatero. Con quale fiducia si ottiene una precisione relativa di $5 \cdot 10^{-2}$? (ii) Dire se l'ipotesi che la concentrazione non sia superiore a 4.3mg/dm³ è plausibile. ■

Esercizio 8.13 Un certo farmaco viene testato per controllare possibili effetti collaterali. Su 863 pazienti cui viene somministrato il farmaco, 19 manifestano sintomi influenzali. Sappiamo (da precedenti studi) che il tasso di insorgenza di sintomi influenzali, in pazienti non trattati con il farmaco in esame, è 1.9%. (i) Vi è evidenza a livello 0.05 che il farmaco induca sintomi influenzali? (ii) Calcolare il p -value dei dati. ■

Esercizio 8.14 Uno studio afferma che la pressione sanguigna (sistolica) nell'uomo sia distribuita come una normale di media $\mu = 120$ e deviazione standard $\sigma = 7$ (dati inventati). Da una rilevazione su 100 individui, risulta una deviazione standard campionaria pari a 8.5. L'ipotesi "deviazione standard pari a 7" è plausibile a livello 0.05? ■

Esercizio 8.15 Si vuole testare un nuovo farmaco contro il colesterolo alto. Per questo, il farmaco viene somministrato a 100 persone con un alto livello di colesterolo. Al termine del periodo di somministrazione, si registra una riduzione media di 7.8, con una deviazione standard campionaria di 6.4. Che conclusioni possiamo trarne sull'efficacia del farmaco? ■

Esercizio 8.16 Due campioni di due soluzioni vengono analizzati per confrontarne il pH, con lo stesso metodo di misurazione. L'errore di misurazione ha distribuzione gaussiana. Vengono effettuate 10 misurazioni (indipendenti) per ciascuna soluzione, ottenendo per la prima soluzione una media di 6.267 e una deviazione standard di 0.0295, mentre per la seconda soluzione una media di 6.285 e una deviazione standard di 0.0327. (i) Formulare un test di livello 0.01 per decidere se l'ipotesi "le due soluzioni hanno lo stesso pH" è plausibile o no e applicarlo ai valori assunti dai campioni. (ii) Calcolare il p -value dei dati. ■