Università degli Studi di Roma
"Tor Vergata"

Appunti del corso di

PROBABILITÀ e STATISTICA

Barbara Pacchiarotti

ultimo aggiornamento marzo 2020

Indice

Introduzione

In questo corso tratteremo argomenti che appartengono a tre discipline distinte.

1. **STATISTICA DESCRITTIVA**
2. **CALCOLO DELLE PROBABILITÀ**
3. **STATISTICA INFERENZIALE**

Scopo di questa introduzione è dare una prima idea di cosa siano e che relazioni abbiano tra loro.

Tutti abbiamo un'idea di cosa sia *un'indagine statistica:*

- censimento decennale della popolazione da parte dell'ISTAT;
- sondaggio d'opinione;
- previsioni e proiezioni di risultati elettorali;
- ispezione di un campione di pezzi da un lotto numeroso per avere un controllo della qualità media di un prodotto;
- sperimentazione di un nuovo prodotto su un campione di casi (nuovo farmaco su pazienti, nuovo carburante su automobili, etc...).

In breve, in Statistica, vengono rilevate *grandezze* o *caratteri* relative ad una *popolazione* intesa in senso lato come collezioni di individui o oggetti, meglio ancora di *misure.*

Veniamo ora alle differenze tra Statistica Descrittiva e Inferenziale.
Ad esempio volendo vedere come i cittadini di un paese ripartiscono i voti tra i vari partiti vi sono due modi:

1. si chiede a ciascun individuo di esprimere il suo voto, quindi si elaborano i dati (percentuali varie). Ci si troverà di fronte ad una mole ingenti di dati da elaborare che daranno esattamente la ripartizione cercata.

2. si interroga un numero limitato di cittadini (sondaggio). Una volta, però, che si hanno i dati (molti meno che nel caso precedente) occorrerà domandarsi quanto i dati relativi al sondaggio siano significativi e che cosa a partire da essi si possa dire (inferire) sul voto dell'intera popolazione.

Il primo è una caso di **Statistica Descrittiva,** che quindi si occupa di elaborare, ordinare e sistemare un insieme di dati. L'altro un caso di **Statistica Inferenziale** e pone una

questione più delicata: cioè in che modo i risultati possono estendersi all'intera popolazione. Si osservi che due sondaggi distinti darebbero, probabilmente, risultati diversi, in altre parole il risultato del sondaggio è *casuale:* ecco perché prima di affrontare i problemi di tipo inferenziale sarà necessario analizzare la struttura dei fenomeni casuali (o aleatori). Ciò è l'oggetto del **Calcolo delle Probabilità.**

Capitolo 1

Distribuzioni di frequenze

1.1 Variabili e dati

La Statistica riguarda i metodi scientifici per raccogliere, ordinare, riassumere e presentare i dati, per trarre valide conclusioni ed eventualmente prendere ragionevoli decisioni sulla base di tale analisi.

Definizione 1.1.1. Le variabili oggetto di osservazione statistica si classificano in tre tipi, a seconda del tipo di valori che esse assumono.

$$variabili \begin{cases} numeriche & \begin{cases} discrete \\ continue \end{cases} \\ categoriche \end{cases}$$

Una variabile si dice numerica se i valori che essa assume sono numeri, categorica altrimenti. Una variabile numerica si dice discreta se l'insieme dei valori che essa a priori può assumere è finito o numerabile, continua se l'insieme dei valori che essa a priori può assumere è l'insieme dei numeri reali $\mathbb{R}$ o un intervallo $I \subset \mathbb{R}$.

Esempio 1.1.2. **[Variabile discreta]** N, numero di nati in una famiglia. $N = 0, 1, \ldots$

Esempio 1.1.3. **[Variabile continua]** H, altezza in centimetri di un individuo. $H \in \mathbb{R}$.

Esempio 1.1.4. **[Variabile categorica]** C, Colore degli occhi di un individuo. C=*marrone, blu, verde,...*

Ci occuperemo per il momento di dati rappresentati da variabili numeriche. Si dicono *grezzi* i dati che non sono stati ordinati numericamente. Una *serie* è un ordinamento di dati grezzi in ordine crescente o decrescente. La differenza tra il più grande e il più piccolo si dice *campo di variazione.* Per esempio se il peso maggiore tra 100 studenti è $74kg$ e il peso minore $60kg$ allora il campo di variazione è $14kg$.

1.2 Distribuzioni di frequenze. Classi

Per studiare i dati a disposizione occorre costruire una *distribuzione di frequenze:* ovvero una tabella dove in una colonna si mettono i valori assunti dalla variabile e in un'altra

il numero delle volte che tali valori vengono assunti (*frequenze*). Ancora più interessanti sono le *frequenze relative* ovvero il numero delle volte in cui un certo valore compare diviso il totale dei dati a disposizione oppure le *frequenze percentuali* ottenute dalle frequenze relative moltiplicando per 100. Pertanto la somma delle frequenze dà il totale delle osservazioni, la somma delle frequenze relative dà come somma 1 e la somma delle frequenze relative percentuali dà come somma 100. Vediamo quest'esempio relativo al peso in chilogrammi di 10 studenti.

Quando si hanno a disposizione un gran numero di dati si può costruire una distribuzione di frequenze in *classi* e determinare il numero di individui appartenenti a ciascuna classe, tale numero è detto *frequenza della classe.* Consideriamo la variabile P peso di un gruppo di 100 studenti.

P Peso in kg	Numero di studenti
$64 < P \leq 66$	5
$66 < P \leq 68$	18
$68 < P \leq 70$	42
$70 < P \leq 72$	27
$72 < P \leq 74$	8
	100

In questo caso si parla di *dati raggruppati.* Se avessimo considerato tutti e cento i pesi avremmo avuto maggiori informazioni, ma avremmo avuto più difficoltà a maneggiare la tabella. Benché il procedimento distrugga molte delle informazioni contenute nei dati originari, tuttavia si trae un importante vantaggio dalla visione più sintetica che si ottiene. Si chiama *ampiezza della classe* la differenza tra il valore massimo e il valore minimo. Si chiama *valore centrale della classe* la semisomma degli estremi. Si noti, che le classi sono state prese *aperte* a sinistra e *chiuse* a destra. Questo non è un caso, ma un modo abbastanza standard di procedere ed il motivo esula dallo scopo di queste note. Per scopi di ulteriore analisi matematica tutte le osservazioni di una classe verranno fatte coincidere con il valore centrale della classe. Per esempio tutti i dati della classe 64-66 saranno considerati pari a $65kg$.

Riassumendo, date un certo numero di osservazioni grezze per formare una distribuzione di frequenze occorre:

- determinare il campo di variazione, dopo aver ordinati tutti i dati;
- dividere il campo di variazione in classi, eventualmente di ampiezza nulla, il che equivale a considerare tutti i valori senza raggrupparli.

Data l'importanza che, vedremo, rivestono i valori centrali fare in modo che questi coincidano quanto più possibile con valori assunti realmente.

1.3 Rappresentazione grafica

Come rappresentare una distribuzione di frequenze? I modi standard sono gli *istogrammi* per le variabili numeriche, i *diagrammi a barre* per le variabili categoriche.

1.3.1 Istogrammi

Un **istogramma** consiste in un insieme di rettangoli adiacenti (ognuno relativo ad una classe) aventi base sull'asse x con punto medio nel valore centrale della classe e altezza proporzionale alla frequenza della classe e tale che l'area del rettangolo sia pari alla frequenza relativa o percentuale della classe.

In questo modo se si sommano le aree di tutti i rettangoli ottenuti si ottiene un valore fisso. Precisamente 1 se si sta costruendo l'istogramma con le frequenze relative (si ricorda che la somma delle frequenze relative è pari a 1), 100 se si sta costruendo l'istogramma con le frequenze percentuali (si ricorda che la somma delle frequenze percentuali è pari a 100). Quindi le altezze dei vari rettangoli si ottengono dividendo le frequenze relative o percentuali per l'ampiezza della classe. Se i dati sono interi, in genere, si prendono classi di ampiezza unitaria, centrate nel valore intero. In tal caso l'altezza dei rettangoli coincide con le frequenze (dato che l'ampiezza della classe è 1)! Vediamo alcuni esempi.

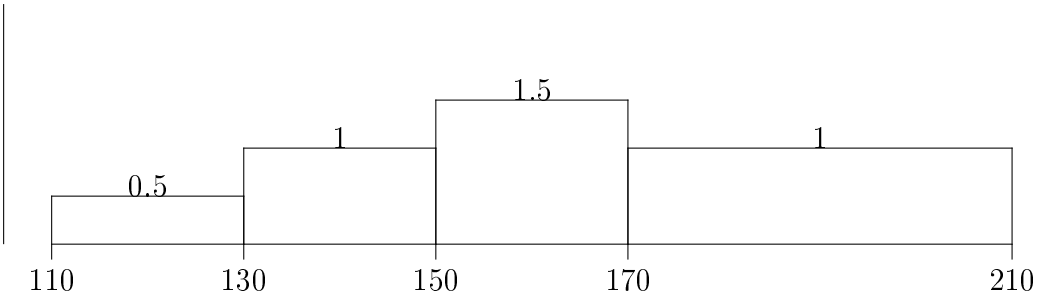
Esempio 1.3.1. Si consideri la seguente distribuzione di frequenze:

D	Frequenze
$110 < D \leq 130$	20
$130 < D \leq 150$	40
$150 < D \leq 170$	60
$170 < D \leq 210$	80
	Tot. 200

Riferendosi alla tabella in questione vogliamo determinare le frequenze percentuali e costruire il relativo istogramma.
Dapprima completiamo la tabella con le frequenze richieste.

D	Frequenze	Freq. percentuali
$110 < D \leq 130$	20	10%
$130 < D \leq 150$	40	20%
$150 < D \leq 170$	60	30%
$170 < D \leq 210$	80	40%
	Tot. 200	100%

Quindi riportiamo sull'asse x gli estremi delle classi; poi ricordando che l'area di ogni rettangolo deve dare le frequenze percentuali, si ha: per la prima classe l'altezza è $10/(130-110) = 0.5$, per la seconda $20/(150-130) = 1.0$, per la terza $30/(170-150) = 1.5$ e per la quarta $40/(210-170) = 1.0$. Da cui,



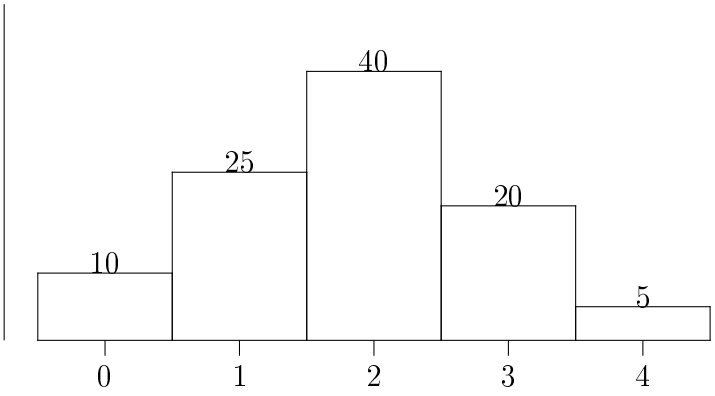
Esempio 1.3.2. La tabella mostra la distribuzione di frequenze per il dato X =numero dei figli in 200 famiglie.

X	frequenze
0	20
1	50
2	80
3	40
4	10
	Tot. 200

Riferendosi alla tabella in questione vogliamo determinare le frequenze relative percentuali e costruire il relativo istogramma.
Dapprima completiamo la tabella con le frequenze richieste.

X	frequenze	freq. percentuali
0	20	10%
1	50	25%
2	80	40%
3	40	20%
4	10	5%
	Tot. 200	100%

Quindi riportiamo sull'asse x i valori. Qui abbiamo dati interi (non si possono avere 2.5 figli!) non raggruppati in classi. Come abbiamo detto in precedenza, scegliamo classi di ampiezza 1. Così le altezze coincidono con le frequenze. Si ottiene,



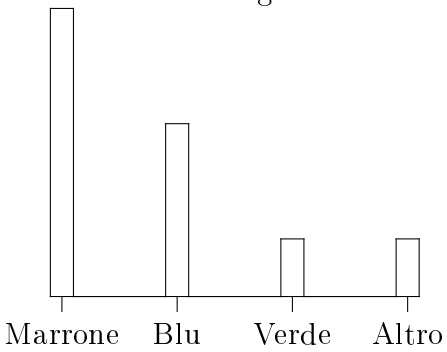
1.3.2 Diagrammi a barre

I **diagrammi a barre** somigliano agli istogrammi, ma sono diversi, data la natura diversa dei dati che rappresentano. Sono sempre dei rettangoli, non adiacenti, in cui l'altezza rappresenta la frequenza relativa o percentuale di quella classe. Sull'asse x si riportano i *tipi,* in un ordine deciso dall'osservatore stesso.

Esempio 1.3.3. Sia C la variabile colore degli occhi di 300 persone.

C	frequenze	freq. percentuali
Marrone	150	50%
Blu	90	30%
Verde	30	10%
Altro	30	10%
	Tot. 300	100%

Un relativo diagramma a barre è il seguente.



1.4 Frequenze cumulate e loro rappresentazione

Per variabili numeriche (non ha senso per variabili categoriche), si chiamano *frequenze cumulate (relative o percentuali),* la somma di tutte le frequenze (relative o percentuali) inferiori o uguali al confine di una fissata classe. Questo è il motivo per cui le classi, come abbiamo visto in precedenza, vengono prese aperte a sinistra e chiuse a destra. Il grafico delle frequenze cumulate è detto *ogiva.* Per meglio chiarire calcoliamo le frequenze cumulate degli esempi visti in precedenza. Ci limiteremo al caso dei dati continui suddivisi in classi.

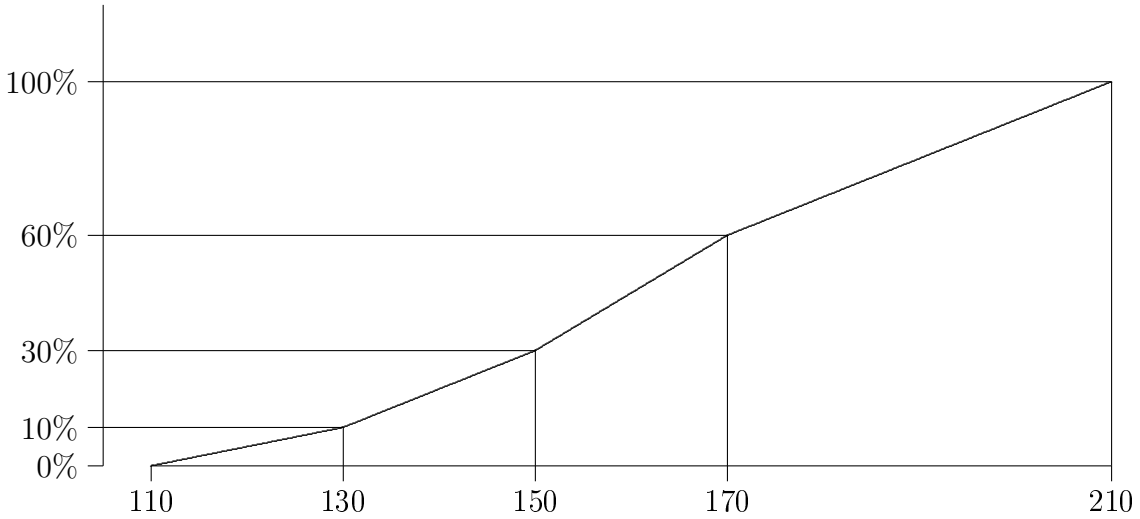
Esempio 1.4.1. Si consideri la seguente distribuzione di frequenze:

D	Frequenze
$110 < D \leq 130$	20
$130 < D \leq 150$	40
$150 < D \leq 170$	60
$170 < D \leq 210$	80
	Tot. 200

Riferendosi alla tabella in questione vogliamo determinare le frequenze percentuali cumulate e costruire il relativo grafico.
Dapprima completiamo la tabella con le frequenze richieste. Non ha senso sommare la colonna delle frequenze cumulate, mentre deve essere sempre l'ultimo valore uguale al totale, 1 se sono frequenze relative, 100 se sono percentuali. Per il 100% della popolazione si ha $D \leq 210$.

D	Frequenze	Freq. percentuali	Freq. cum. percentuali
$110 < D \leq 130$	20	10%	10%
$130 < D \leq 150$	40	20%	30%
$150 < D \leq 170$	60	30%	60%
$170 < D \leq 210$	80	40%	100%
	Tot. 200	100%	

Questa tabella ci dice che $D \leq 130$ per il 10% della popolazione; $D \leq 150$ per il 30% della popolazione; $D \leq 170$ per il 60% della popolazione; e come già osservato $D \leq 210$ per il 100% della popolazione. Il grafico che si ottiene è il seguente.



Si osservi che l'ogiva si fa sempre partire da zero. Ovvero, in questo caso, il numero di individui per cui il dato è minore o uguale di 110 è zero.

Capitolo 2

Indici di posizione e di dispersione

Per variabili numeriche ha senso calcolare alcuni indici, quali la media, la mediana, la varianza, ecc... Vediamo in dettaglio cosa rappresentano.

2.1 Indici di posizione

Si chiamano *indici di posizione* quegli indici che aiutano a capire dove è posizionata, ovvero quali sono i valori che assume una certa distribuzione.

2.1.1 Media

La media è un indice di posizione. Come si calcola la media di una distribuzione?

Cominciamo dal caso di dati *non* raggruppati e supponiamo di avere la seguente distribuzione. N rappresenta la numerosità della popolazione, n il numero di classi. $N = n$ se e soltanto se c'è un solo individuo in ogni classe ovvero $f_i = 1$ per ogni $i = 1, 2, \ldots, n$..

X	frequenze	freq. rel
x_1	f_1	p_1
x_2	f_2	p_2
.	.	.
.	.	.
x_n	f_n	p_n
	Tot. N	1

Definizione 2.1.1. Si chiama **media** di X e si indica con $\overline{X}$, la quantità:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i = \sum_{i=1}^{n} x_i p_i. \tag{2.1}$$

Per chi non ha simpatia per il simbolo di sommatoria possiamo riscrivere, per esteso

$$\overline{X} = \frac{1}{N}(x_1\, f_1 + x_2\, f_2 + \ldots + x_n\, f_n) = (x_1\, p_1 + x_2\, p_2 + \ldots + x_n\, p_n).$$

Chiariamo con un esempio.

Esempio 2.1.2. Riprendiamo la distribuzione del peso di alcuni studenti già vista in precedenza. Quanto vale la media di P?

P Peso in kg	Frequenze	Frequenze relative
65	2	0.2
68	1	0.1
69	1	0.1
70	2	0.2
72	1	0.1
74	3	0.3
Totali	10	1.0

Applicando la (2.1), si ha

$$\begin{aligned} \overline{P} &= \frac{1}{10}(65 \cdot 2 + 68 \cdot 1 + 69 \cdot 1 + 70 \cdot 2 + 72 \cdot 1 + 74 \cdot 3) \\ &= (65 \cdot 0.2 + 68 \cdot 0.1 + 69 \cdot 0.1 + 70 \cdot 0.2 + 72 \cdot 0.1 + 74 \cdot 0.3) = 70.1 \end{aligned}$$

Cosa succede se si dispone di dati raggruppati? Semplicemente che tutti i valori di una classe vengono identificati con il valore centrale di quella classe, che è quindi il valore utilizzato per il calcolo della media. Anche qui chiariamo con un esempio.

Esempio 2.1.3. Si consideri la seguente distribuzione di frequenze:

D	Frequenze
$110 < D \leq 130$	20
$130 < D \leq 150$	40
$150 < D \leq 170$	60
$170 < D \leq 210$	80
	Tot. 200

Quanto vale la media di D? Si ha,

$$\overline{D} = \frac{1}{200}(120 \cdot 20 + 140 \cdot 40 + 160 \cdot 60 + 190 \cdot 80) = 164.$$

Attenzione! La media, o valore medio, può anche essere una valore diverso da quelli assunti... Anzi in generale lo è.

2.1.2 Mediana, quartili, percentili

La **mediana** è un altro indice di posizione. La mediana è un valore che provoca la ripartizione della popolazione in esame in due parti ugualmente numerose: per il 50% della popolazione il dato è minore della mediana, per il restante 50% il dato è maggiore della mediana. Per chiarire, se diciamo che il *reddito mediano* dei lavoratori di una certa città è 1500 euro, stiamo dicendo che la metà dei lavoratori percepisce meno di 1500 euro e la restante metà più di 1500 euro. Come si calcola la mediana? Anche qui ci limiteremo al caso di dati continui. Se abbiamo i dati non raggruppati, possiamo pensarli sotto forma di fila ordinata; allora la mediana è il valore centrale, se sono in numero dispari, la semisomma dei valori centrali se sono in numero pari. Vediamo un esempio.

Esempio 2.1.4. Supponiamo che per il dato X si siano osservati i valori 67, 72, 78, 78, 84, 85, 87, 91. Si tratta di un campione di numerosità 8 (pari), quindi $\text{Med}(X) = \frac{1}{2}(78+84) = 81$.

Supponiamo invece che per il dato X si siano osservati i valori 65, 67, 72, 78, 78, 84, 85, 87, 91. Si tratta di un campione di numerosità 9 (dispari), quindi $\text{Med}(X) = 78$.

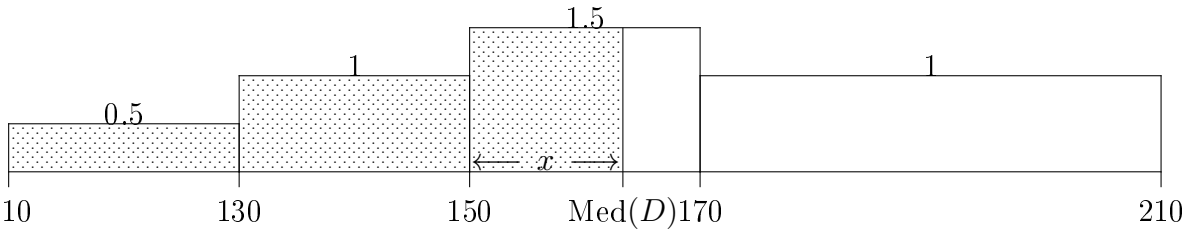
In questo caso, si può calcolare la mediana attraverso l'istogramma o il grafico delle frequenze cumulate. I due metodi sono equivalenti, dipende un po' dai gusti personali.

Partiamo dal caso in cui si voglia utilizzare l'istogramma. Occorre trovare quel valore sull'asse x tale che divida esattamente a metà *l'area* delimitata dall'istogramma. Si ricorda che per come viene costruito l'istogramma l'area totale sottesa ha un valore fissato: vale 1 se si stanno utilizzando le frequenze relative, 100 se si stanno utilizzando le frequenze percentuali. Chiariamo anche qui con un esempio.

Esempio 2.1.5. Riprendiamo una distribuzione già vista.

D	Frequenze	Freq. rel %
$110 < D \leq 130$	20	10%
$130 < D \leq 150$	40	20%
$150 < D \leq 170$	60	30%
$170 < D \leq 210$	80	40%
	Tot. 200	100%

Il relativo istogramma è:



Dato che è costruito con le frequenze percentuali l'area racchiusa dall'istogramma è 100. La mediana è quel valore che ripartisce l'area in due parti uguali. Nel nostro caso 50 prima (ombreggiata) 50 dopo (vedi la Figura). Indicando con x la quantità $\text{Med}(D) - 150$, deve essere:

$$20 \cdot 0.5 + 20 \cdot 1 + x \cdot 1.5 = 50, \quad x \cdot 1.5 = 20, \quad x = 13.3,$$

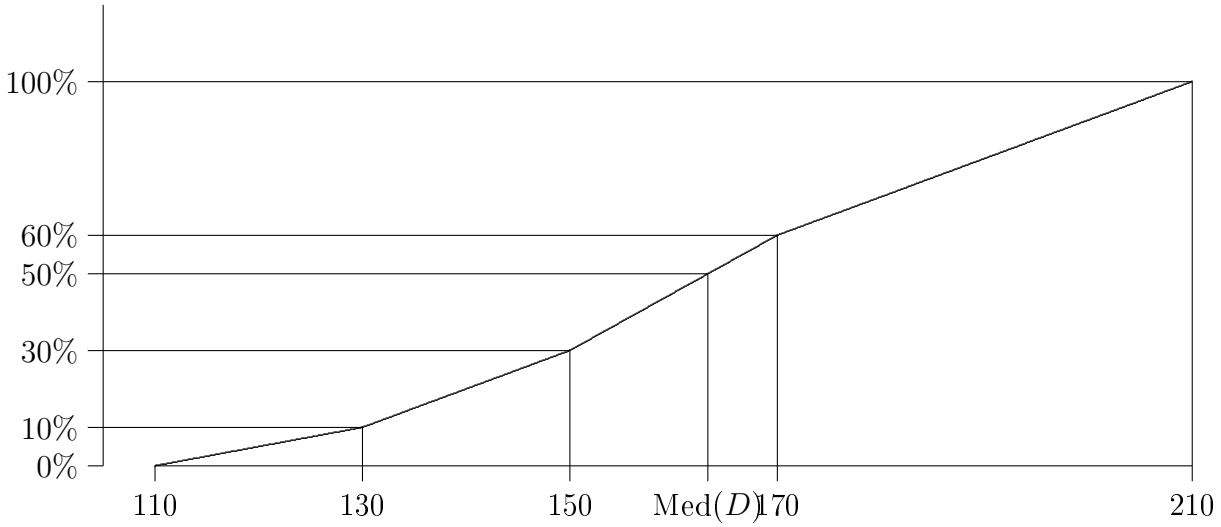
Quindi $\text{Med}(D) = 150 + x = 150 + 13.3 = 163.3$.

Passiamo al caso in cui si voglia utilizzare il grafico delle frequenze cumulate. In questo caso la mediana è quel valore sull'asse x cui corrisponde sulle ordinate il 50% se si stanno utilizzando frequenze percentuali, 0.5 se si stanno utilizzando frequenze relative. Si utilizza per calcolarla la condizione di allineamento di tre punti. Anche in questo caso chiariamo con un esempio.

Esempio 2.1.6. Riprendiamo sempre la stessa distribuzione di frequenze.

D	Frequenze	Freq. rel %	Freq. rel cum. %
$110 < D \leq 130$	20	10%	10%
$130 < D \leq 150$	40	20%	30%
$150 < D \leq 170$	60	30%	60%
$170 < D \leq 210$	80	40%	100%
	Tot. 200	100%	

Il grafico delle frequenze percentuali cumulate che si ottiene è il seguente.



Ricordiamo che tre punti di coordinate $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ e $P_3(x_3, y_3)$, sono allineati se e solo se le coordinate soddisfano la seguente condizione:

$$(y_3 - y_1)(x_2 - x_1) = (y_2 - y_1)(x_3 - x_1).$$

Nel nostro caso i punti che devono essere allineati sono $(150, 30)$, $(\text{Med}(D), 50)$ e $(170, 60)$. La condizione di allineamento diventa:

$$(60 - 30)(\text{Med}(D) - 150) = (50 - 30)(170 - 150),$$

da cui

$$30(\text{Med}(D) - 150) = 400, \quad 3\text{Med}(D) = 490, \quad \text{Med}(D) = 163.3.$$

Ovvero lo stesso risultato raggiunto con il metodo dell'istogramma.

In modo analogo alla mediana si possono definire i quartili e i percentili. I **quartili** sono quei valori che ripartiscono la popolazione, pensata sempre come una fila ordinata, in quattro parti ugualmente numerose (pari ciascuna al 25% del totale). Il primo quartile q_1, lascia alla sua sinistra il 25% della popolazione (a destra quindi il 75%), il secondo quartile q_2 lascia a sinistra il 50% (a destra quindi il 50%). Esso chiaramente coincide con la mediana. Il terzo quartile lascia a sinistra il 75% della popolazione (a destra quindi il 25%).

Come si calcolano i quartili?

Se abbiamo i dati sotto forma di fila ordinata, $(x_1, x_2, \ldots, x_N)$, allora si procede in modo analogo a quanto fatto per la mediana. Più precisamente se vogliamo calcolare q_1 si moltiplica $p = 0.25$ (la percentuale che lascia a sinistra) per la numerosità del campione N. Ci sono

due possibilità pN è un intero, diciamolo k. In tal caso $q_1 = \frac{1}{2}(x_k + x_{k+1})$. pN non è un intero. Sia allora $k = [pN]$. In tal caso $q_1 = x_{k+1}$. Dovendo calcolare gli altri quartili basta mettere al posto di p il valore giusto. Riprendiamo l'Esempio 2.1.4.

Esempio 2.1.7. Supponiamo che per il dato X si siano osservati i valori 67, 72, 78, 78, 84, 85, 87, 91. Si tratta di un campione di numerosità 8, quanto valgono i quartili? Qui $N = 8$.

Per il primo quartile dobbiamo considerare la quantità $0.25 \cdot 8 = 2$, intero. Quindi $q_1 = \frac{1}{2}(x_2 + x_3) = \frac{1}{2}(72 + 78) = 75$. $q_2 = \text{Med}(X)$ (vista nell'Esempio precedente). Per il terzo quartile dobbiamo considerare la quantità $0.75 \cdot 8 = 6$, intero. Quindi $q_3 = \frac{1}{2}(x_6 + x_7) = \frac{1}{2}(85 + 87) = 86$.

Supponiamo invece che per il dato X si siano osservati i valori 65, 67, 72, 78, 78, 84, 85, 87, 91. Si tratta di un campione di numerosità 9, quanto valgono i quartili? Qui $N = 9$.

Per il primo quartile dobbiamo considerare la quantità $0.25 \cdot 9 = 2.25$, non intero. Quindi $q_1 = x_3 = 72$. $q_2 = \text{Med}(X)$ (vista nell'Esempio precedente). Per il terzo quartile dobbiamo considerare la quantità $0.75 \cdot 9 = 6.75$, non intero. Quindi $q_3 = x_7 = 85$.

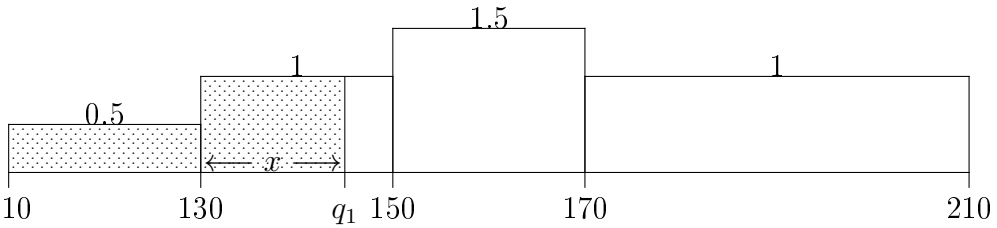
Come si procede nel caso in cui si hanno dati sotto forma di distribuzione di frequenze? Come visto per la mediana si può utilizzare l'istogramma o il grafico delle frequenze cumulate.

Supponiamo di voler calcolare il primo quartile (gli altri casi si trattano in modo analogo). Partiamo dal caso in cui si voglia utilizzare l'istogramma. Occorre trovare quel valore sull'asse x tale che divida *l'area* delimitata dall'istogramma in due parti, quella a sinistra pari al 25% la restante pari al 75%. Si ricorda che per come viene costruito l'istogramma l'area totale sottesa ha un valore fissato: vale 1 se si stanno utilizzando le frequenze relative. 100 se si stanno utilizzando le frequenze percentuali. Chiariamo anche qui con un esempio.

Esempio 2.1.8. Riprendiamo una distribuzione già vista.

D	Frequenze	Freq. rel %
$110 < D \leq 130$	20	10%
$130 < D \leq 150$	40	20%
$150 < D \leq 170$	60	30%
$170 < D \leq 210$	80	40%
	Tot. 200	100%

Il relativo istogramma è:



Dato che è costruito con le frequenze percentuali l'area racchiusa dall'istogramma è 100. Il primo quartile è quel valore che ripartisce l'area in due parti: 25% (ombreggiata), 75% (vedi la Figura). Indicando con x la quantità $q_1 - 130$, deve essere:

$$20 \cdot 0.5 + x \cdot 1 = 25, \quad x = 15,$$

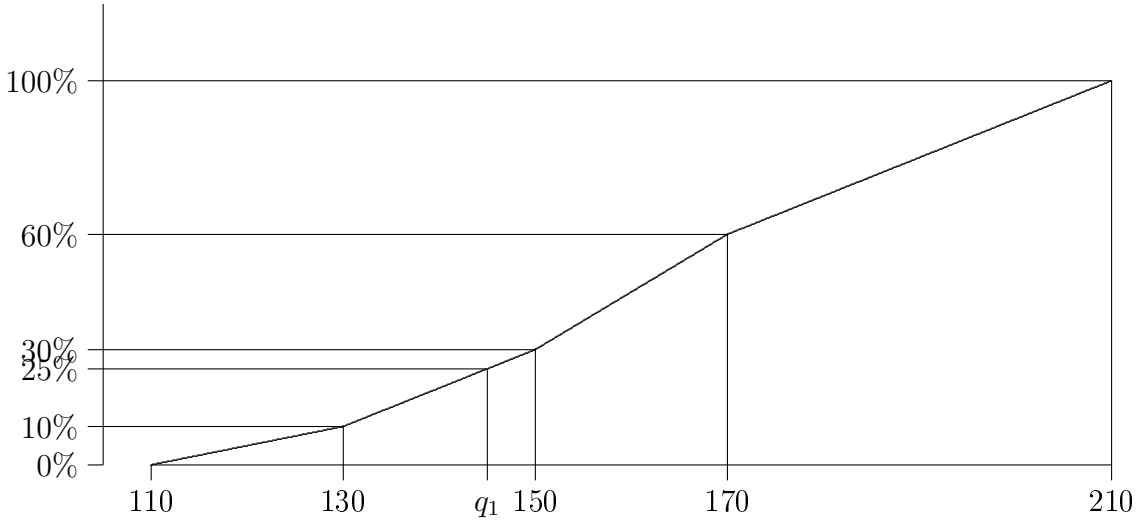
Quindi $q_1 = 130 + x = 130 + 15 = 145$.

Passiamo al caso in cui si voglia utilizzare il grafico delle frequenze cumulate. In questo caso la mediana è quel valore sull'asse x cui corrisponde sulle ordinate il 25% se si stanno utilizzando frequenze percentuali, 0.25 se si stanno utilizzando frequenze relative. Si utilizza per calcolarla la condizione di allineamento di tre punti.Anche in questo caso chiariamo con un esempio.

Esempio 2.1.9. Riprendiamo sempre la stessa distribuzione di frequenze.

D	Frequenze	Freq. rel %	Freq. rel cum. %
$110 < D \leq 130$	20	10%	10%
$130 < D \leq 150$	40	20%	30%
$150 < D \leq 170$	60	30%	60%
$170 < D \leq 210$	80	40%	100%
	Tot. 200	100%	

Il grafico delle frequenze percentuali cumulate che si ottiene è il seguente.



Ricordiamo che tre punti di coordinate $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ e $P_3(x_3, y_3)$, sono allineati se e solo se le coordinate soddisfano la seguente equazione

$$(y_3 - y_1)(x_2 - x_1) = (y_2 - y_1)(x_3 - x_1).$$

Nel nostro caso i punti che devono essere allineati hanno coordinate $(130, 10)$, $(q_1, 25)$ e $(150, 30)$. La condizione di allineamento diventa

$$(30 - 10)(q_1 - 130) = (25 - 10)(150 - 130),$$

da cui

$$q_1 - 130 = 15, \quad q_1 = 145.$$

Ovvero lo stesso risultato raggiunto con il metodo dell'istogramma.

In modo analogo si possono definire i **percentili,** ovvero quei valori che ripartiscono la popolazione in 100 parti ugualmente numerose (ciascuna pari pertanto all'1%). Per calcolare i percentili si utilizzano i metodi già visti per la mediana ed i quartili, ovviamente utilizzando i dovuti aggiustamenti.

2.2 Indici di dispersione

2.2.1 Varianza e scarto quadratico medio

Media e mediana, abbiamo visto essere degli indici di posizione (perché dicono accanto a quale valore il campione di dati è "posizionato") e sono tanto più significative quanto più i dati sono concentrati vicino ad esse. È interessante misurare quindi il grado di *dispersione* dei dati rispetto, ad esempio, alla media. Si osservi che la *somma di tutte le deviazioni* dalla media è sempre zero, ovvero

$$\sum_{i=1}^{n}(x_i - \overline{X})f_i = 0,$$

perciò, per misurare in modo significativo la dispersione dei dati rispetto alla media, si può considerare, ad esempio, la somma dei moduli delle deviazioni, oppure la somma dei quadrati delle deviazioni. Ci occuperemo di quest'ultimo indice, che, per motivi qui non facilmente spiegabili, occupa un posto decisamente più importante nell'ambito di tutta la Statistica.

Cominciamo dal caso di dati *non* raggruppati e supponiamo di avere la seguente distribuzione. N rappresenta la numerosità della popolazione.

X	frequenze	freq. rel
x_1	f_1	p_1
x_2	f_2	p_2
.	.	.
.	.	.
x_n	f_n	p_n
	Tot. N	1

Definizione 2.2.1. Si chiama **varianza** di X e si indica con s_X^2, la media degli scarti al quadrato, ovvero la quantità

$$s_X^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i = \sum_{i=1}^{n}(x_i - \overline{X})^2 p_i. \tag{2.2}$$

Definizione 2.2.2. Si chiama **scarto quadratico medio** o anche **deviazione standard** di X e si indica con s_X, la radice della varianza, ovvero

$$s_X = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2 p_i}. \tag{2.3}$$

Per visualizzare meglio si può aggiungere una colonna alla distribuzione, quella degli scarti al quadrato, e quindi fare la media di quella colonna, ovvero si costruisce la tabelle seguente.

X	$(X-\overline{X})^2$	frequenze	freq. rel
x_1	$(x_1-\overline{X})^2$	f_1	p_1
x_2	$(x_2-\overline{X})^2$	f_2	p_2
.	.	.	.
.	.	.	.
x_n	$(x_n-\overline{X})^2$	f_n	p_n
		Tot. N	1

Si osservi che se i dati x_i rappresentano, ad esempio, lunghezze misurate in metri, la media, tutti gli altri indici di posizione e la deviazione standard sono misurate in metri, mentre la varianza è misurata in metri quadri. La varianza e la deviazione standard sono grandezza *non* negative, che si annullano solo quando gli x_i sono tutti uguali tra loro e quindi uguali alla loro media. È utile per il calcolo esplicito della varianza (la cui dimostrazione è lasciata per esercizio agli studenti più interessati e volenterosi!) la seguente formula

$$s_X^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 f_i - \overline{X}^2 = \sum_{i=1}^{n} x_i^2 p_i - \overline{X}^2 = \overline{X^2} - \overline{X}^2. \tag{2.4}$$

Ovvero la varianza di una distribuzione è pari alla media del quadrato della variabile meno la media della variabile al quadrato.

Per visualizzare meglio si può aggiungere una colonna alla distribuzione, quella della variabile al quadrato, e quindi fare la media di quella colonna, ovvero si costruisce la tabelle seguente.

X	X^2	frequenze	freq. rel
x_1	x_1^2	f_1	p_1
x_2	x_2^2	f_2	p_2
.	.	.	.
.	.	.	.
x_n	x_n^2	f_n	p_n
		Tot. N	1

Cosa succede se si dispone di dati raggruppati? Semplicemente che tutti i valori di una classe vengono identificati con il valore centrale di quella classe che è quindi il valore utilizzato per il calcolo della varianza. Anche qui chiariamo con un esempio.

Esempio 2.2.3. Si consideri la seguente distribuzione di frequenze:

D	Frequenze
$110 < D \leq 130$	20
$130 < D \leq 150$	40
$150 < D \leq 170$	60
$170 < D \leq 210$	80
	Tot. 200

Abbiamo già trovato la media di D. La ricordiamo per completezza.

$$\overline{D} = \frac{1}{200}(120 \cdot 20 + 140 \cdot 40 + 160 \cdot 60 + 190 \cdot 80) = 164.$$

Per il calcolo della varianza possiamo procedere utilizzando la definizione oppure la (2.4). Mostriamo che giungiamo allo stesso risultato. Completiamo la tabella con la colonna degli scarti al quadrato e con quella della variabile al quadrato.

D	$(D-\overline{D})^2$	D^2	Frequenze
$110 < D \leq 130$	$(120-164)^2$	120^2	20
$130 < D \leq 150$	$(140-164)^2$	140^2	40
$150 < D \leq 170$	$(160-164)^2$	160^2	60
$170 < D \leq 210$	$(190-164)^2$	190^2	80
			Tot. 200

Proviamo ad applicare la definizione (faccio la media della colonna degli scarti al quadrato)

$$s_D^2 = \frac{1}{200}[(120-164)^2 \cdot 20 + (140-164)^2 \cdot 40 + (160-164)^4 \cdot 60 + (190-164)^2 \cdot 80] = 584$$

Invece applicando la (2.4) dobbiamo calcolare $\overline{D^2}$,

$$\overline{D^2} = \frac{1}{200}(120^2 \cdot 20 + 140^2 \cdot 40 + 160^2 \cdot 60 + 190^2 \cdot 80) = 27480,$$

e quindi

$$s_D^2 = \overline{D^2} - \overline{D}^2 = 27480 - 164^2 = 584.$$

Pertanto

$$s_D = \sqrt{584} = 24.17.$$

Capitolo 3

Correlazione tra variabili e regressione lineare

3.1 Correlazione tra variabili. Scatterplot

Talvolta più caratteri vengono misurati per ogni individuo: peso, altezza, sesso, reddito, ecc. Si vuole vedere se c'è una qualche relazione tra essi. Noi considereremo il caso di due caratteri quantitativi e supporremo che i dati (sempre non raggruppati per questo tipo di analisi) siano sotto forma di coppie (x_1, y_1), (x_2, y_2),...,(x_N, y_N) in cui la prima coordinata rappresenta il primo carattere X e la seconda coordinata il secondo carattere Y. Ogni coppia è relativa ad un individuo. Possono esserci più coppie coincidenti. In un primo approccio grafico si possono disegnare sul piano tutti i punti di coordinate (x_i, y_i) e vedere se essi tendono a disporsi secondo un andamento regolare. Si fa quello che di chiama lo *scatterplot.*

Si possono presentare varie situazioni.

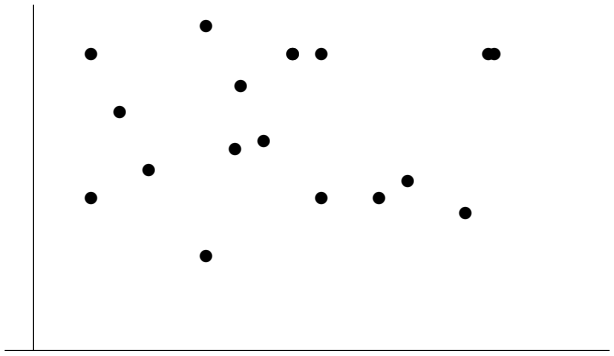


Figure 3.1

I punti della figura 3.1 sembrano non avere alcuna correlazione, mentre quelli delle altre figure manifestano una certa *tendenza.* Precisamente i punti della figura 3.2 sembrano avere un andamento quadratico (il grafico si accosta a quello di una parabola) e quelli delle ultime due (Figure 3.3 e 3.4) sembrano avere un andamento lineare, ovvero si avvicinano ad una retta.

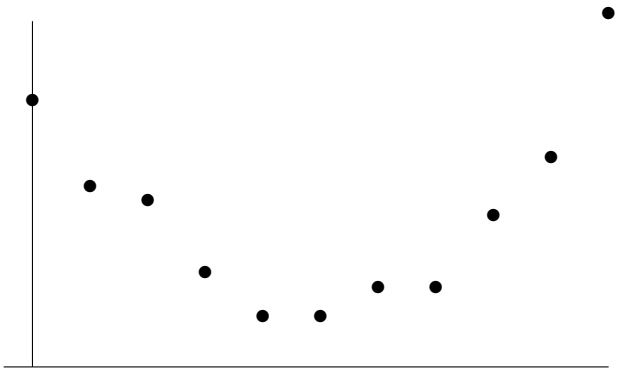


Figure 3.2

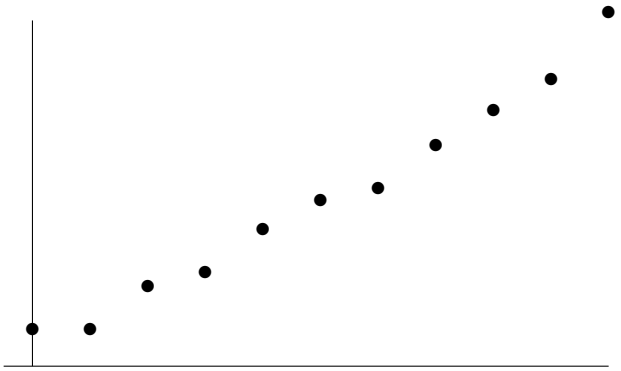


Figure 3.3

L'idea è quella di trovare la curva (retta, parabola o altro), se esiste, che meglio descriva l'andamento dei dati per poi utilizzarla per "stimare" il valore di un carattere conoscendo l'altro. Analiticamente ci occuperemo solo dei dati che tendono a disporsi secondo una retta. Molti altri casi poi si possono ricondurre a questo.

Puntualizziamo ora alcuni concetti emersi da questi esempi.

Definizione 3.1.1. Supponiamo di avere N osservazioni congiunte di due variabili X e Y, ovvero di avere $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. Si dice **covarianza** tra X e Y e si indica con s_{XY} la quantità

$$s_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y}) = \overline{XY} - \overline{X} \cdot \overline{Y}. \tag{3.1}$$

L'uguaglianza tra le due diverse espressioni di s_{XY} si dimostra facilmente svolgendo i conti (anche questo è lasciato per esercizio agli studenti più interessati e volenterosi!). Si noti anche che è immediato verificare che $s_{XY} = s_{YX}$.

Dalla definizione, si vede che la covarianza può avere segno positivo o negativo. Se $s_{XY} > 0$, in base alla definizione significa che, mediamente, a valori grandi (o piccoli di X) corrispondono valori grandi (o piccoli, rispettivamente) di Y. Se invece $s_{XY} < 0$ significa che, mediamente, a valori grandi (o piccoli di X) corrispondono valori piccoli (o grandi, rispettivamente) di Y. questo giustifica la seguente definizione.

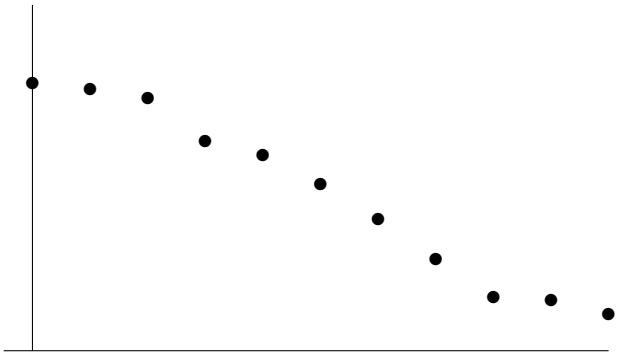


Figure 3.4

Definizione 3.1.2. Si dice che le variabili X e Y sono *direttamente correlate* se $s_{XY} > 0$. Si dice che le variabili X e Y sono *inversamente correlate* se $s_{XY} < 0$. Si dice che le variabili X e Y sono *incorrelate* se $s_{XY} = 0$.

Un altro importante strumento per studiare la correlazione tra due variabili è il coefficiente di correlazione.

Definizione 3.1.3. Si chiama **coefficiente di correlazione** di X e Y e si indica con ρ_{XY} la quantità

$$\rho_{XY} = \frac{s_{XY}}{s_X\, s_Y}, \tag{3.2}$$

ovvero la covarianza divisa per il prodotto delle deviazioni standard. Si osservi che il coefficiente di correlazione è definito solo se $s_X \neq 0$ e $s_Y \neq 0$, ovvero se X ed Y non sono costanti.

L'importanza del coefficiente di correlazione (rispetto alla covarianza, che è un concetto simile) dipende dal fatto, che non dimostriamo (perché non abbiamo gli strumenti!) che esso risulta sempre in modulo minore o uguale a 1:

$$-1 \leq \rho_{XY} \leq 1.$$

Pertanto esso è un indice *normalizzato,* la cui grandezza ha un significato assoluto.

Inoltre $\rho_{XY} = \pm 1$ se e soltanto se i punti dati già sono allineati. Questo vuol dire che tanto più $|\rho_{XY}|$ si avvicina a 1 tanto più l'idea di approssimare i punti con una retta è buona.

3.2 Metodo dei Minimi Quadrati. Regressione Lineare

Ci occuperemo ora proprio del problema di ricercare, in generale, una relazione del tipo $Y = aX + b$ tra le due variabili. In base a quanto detto nel paragrafo precedente, se il coefficiente di correlazione non vale ± 1, non esiste una retta che passi per tutti i punti dati. Tuttavia possiamo ugualmente cercare una retta che passi "abbastanza vicino" a tutti i punti.

L'idea è questa. Abbiamo una "nuvola" di punti nel piano (x_1, y_1),. . .,(x_N, y_N) e cerchiamo due numeri a e b (i coefficienti che individuano la retta) per cui la retta$Y = aX + b$ passi il

più possibile vicino a questi punti. Consideriamo allora l'espressione

$$\sum_{i=1}^{N}[y_i - (ax_i + b)]^2,$$

che dà per a, b fissati, la somma dei quadrati delle distanze tra il punto originale (x_i, y_i) e il punto di uguale ascissa che si trova sulla retta $Y = aX + b$, ovvero il punto di coordinate $(x_i, \hat{y}_i)$, con $\hat{y}_i = ax_i + b$.

Cerchiamo ora i valori di a e b che rendono minima questa quantità (metodo dei minimi quadrati o regressione). Si trovano (anche questo conto non rientra nelle nostre competenze!) i seguenti valori

$$\begin{aligned} \hat{a} &= \frac{s_{XY}}{s_X^2} \\ \hat{b} &= \overline{Y} - \hat{a}\overline{X}. \end{aligned}$$

Pertanto la retta dei minimi quadrati o di regressione è

$$Y = \hat{a}X + \hat{b} = \frac{s_{XY}}{s_X^2}X + \overline{Y} - \frac{s_{XY}}{s_X^2}\overline{X}.$$

Osservazione 3.2.1. Notare che il coefficiente angolare della retta ha il segno della covarianza, coerentemente alla definizione data di correlazione diretta e inversa: se tra X e Y c'è una correlazione diretta (o inversa), la retta di regressione, sarà una retta crescente (rispettivamente decrescente). Se X e Y sono incorrelate la retta di regressione è la retta orizzontale di equazione $Y = \overline{Y}$. Questo vuol dire che nessuna previsione può essere fatta su Y se si conosce X.
Si osservi anche che la retta di regressione passa per il punto di coordinate $(\overline{X}, \overline{Y})$.

Attenzione! Aver determinato la retta di regressione non significa affatto che la variabile Y sia (in modo pur approssimato) una funzione del tipo $Y = aX + b$. Le equazioni che determinano i coefficienti $\hat{a}$ e $\hat{b}$ servono a determinare una relazione di tipo affine presupponendo che questa ci sia. Per capire se è ragionevole che sussista una relazione affine tra le due variabili abbiamo due strumenti:

- il calcolo del coefficiente di correlazione lineare (che deve essere vicino a ± 1);
- l'esame visivo dello scatterplot.

Anche qui chiudiamo il Capitolo con un esempio chiarificatore.

Esempio 3.2.2. Con riferimento a due fenomeni sono state annotate le seguenti osservazioni:

X	1	4	5	10
Y	14	8	4	2

a) determinare il grado di correlazione lineare tra X e Y;

b) trovare la retta di regressione di Y su X;

c) disegnare lo scatter plot dei dati e la retta trovata

d) stimare il valore di Y quando X vale 8.

Riempiendo la tabella si ottiene

X	Y	XY	X^2	Y^2
1	14	14	1	196
4	8	32	16	64
5	4	20	25	16
10	2	20	100	4

da cui

$$\begin{aligned}
\overline{X} &= \frac{1}{4}(1+4+5+10) = 5 \\
\overline{Y} &= \frac{1}{4}(14+8+4+2) = 7 \\
\overline{XY} &= \frac{1}{4}(14+32+20+20) = 21.5 \\
\overline{X^2} &= \frac{1}{4}(1+16+25+100) = 35.5 \\
\overline{Y^2} &= \frac{1}{4}(196+64+16+4) = 70
\end{aligned}$$

Pertanto,

$$\begin{aligned}
s_{XY} &= \overline{XY} - \overline{X}\,\overline{Y} = 21.5 - 7 \cdot 5 = -13.5 \\
s_X^2 &= \overline{X^2} - \overline{X}^2 = 35.5 - 5^2 = 10.5 \\
s_Y^2 &= \overline{Y^2} - \overline{Y}^2 = 70 - 7^2 = 21
\end{aligned}$$

e quindi, il coefficiente di correlazione lineare è

$$\rho_{XY} = \frac{s_{XY}}{\sqrt{s_X^2}\sqrt{s_Y^2}} = -\frac{13.5}{\sqrt{10.5}\sqrt{21}} = -0.91,$$

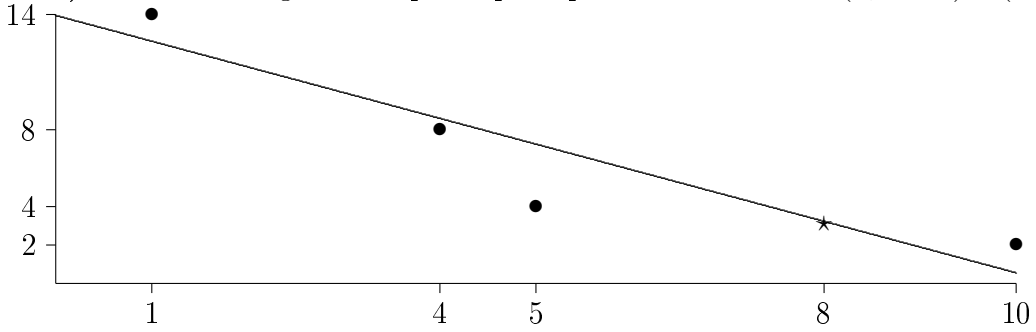
un buon livello di accettabilità.

c) La retta di regressione di Y su X è $Y = aX + b$, dove

$$a = \frac{s_{XY}}{s_X^2} = -\frac{13.5}{10.5} = -1.29 \quad \text{e} \quad b = \overline{Y} - a\overline{X} = 7 - (-1.29) \cdot 5 = 13.45,$$

quindi la retta richiesta è $Y = -1.29X + 13.45$.

c) La retta di regressione passa per i punti di coordinate $(0, 13.45)$ e $(10, 0.55)$.



d) Il valore stimato quando $X = 8$ è $Y(8) = -1.29 \cdot 8 + 13.45 = 3.13$. La stellina sul grafico lo rappresenta.

Osservazione 3.2.3. Facendo riferimento a due caratteri abbiamo sin qui considerato Y come variabile dipendente ed X come variabile indipendente. Ma, se ha senso logico, si può scambiare la X con la Y. Nel primo caso si parla di regressione di Y rispetto a X (o su X), nel secondo di regressione di X rispetto a Y (o su X). In particolare l'equazione della retta di regressione di Y rispetto a X ha equazione, $Y = \hat{a}X + \hat{b}$, dove

$$\begin{aligned} \hat{a} &= \frac{s_{XY}}{s_X^2} \\ \hat{b} &= \overline{Y} - \hat{a}\overline{X}, \end{aligned}$$

mentre scambiando il ruolo di X con Y si ottiene che la retta di regressione di X su Y ha equazione $X = \hat{c}Y + \hat{d}$, dove

$$\begin{aligned} \hat{c} &= \frac{s_{XY}}{s_Y^2} \\ \hat{d} &= \overline{X} - \hat{c}\overline{Y}. \end{aligned}$$

Si osservi che le due rette *non* sono la stessa retta! Anzi sono la stessa retta se e solo se $\rho_{XY} = \pm 1$, ovvero se i punti dati sono già allineati. Esse passano entrambe per il punto di coordinate $(\overline{X}, \overline{Y})$, hanno coefficiente dello stesso segno (ovvero sono entrambe crescenti o entrambe decrescenti), ma non si sovrappongono. In caso di X e Y non correlate ($s_{XY} = 0$), allora le due rette di regressione sono parallele agli assi, quindi perpendicolari. Quella di Y su X ha (come visto) equazione $Y = \overline{Y}$, quella di X su Y ha equazione $X = \overline{X}$. Ma in questo caso la retta di regressione non è interessante.

Capitolo 4

Introduzione alla probabilità

4.1 Spazi di probabilità

Il calcolo delle probabilità si occupa di studiare i fenomeni *casuali* o *aleatori* ovvero i fenomeni dei quali non si può prevedere con certezza l'esito. Se si lancia un dado o una moneta non c'è modo di sapere con esattezza quale sarà il risultato del nostro esperimento. Tuttavia si possono fare delle previsioni su quello che accadrà. Cominciamo proprio da questo esempio.

Esempio 4.1.1. Si lancia un dado. I possibili risultati di questo esperimento sono 1, 2, 3, 4, 5, 6. Che vuol dire probabilità di avere 1? E probabilità di avere un numero dispari? Se indichiamo con $\Omega = \{1, 2, 3, 4, 5, 6\}$ l'insieme dei possibili risultati ogni "evento" (esce 1, esce un numero dispari, ecc) può essere identificato con un sottoinsieme di Ω. Per esempio:

- *"esce 1"* $=\{1\}$;
- *"esce un numero dispari"* $=\{1, 3, 5\}$.

Vediamo di formalizzare quanto detto.
Sia Ω l'insieme di tutti i possibili risultati di un esperimento. Ω è detto *spazio campionario.* Un *evento* A è un *particolare* sottoinsieme di Ω di cui è possibile calcolare la probabilità (non sempre si può calcolare la probabilità di tutti i sottoinsiemi di Ω). Ω è detto *evento certo*, $\emptyset$ è detto *evento impossibile.*

Data l'identificazione di un evento con un sottoinsieme di Ω gli eventi si possono "combinare" per formarne degli altri. Dati A, B eventi in Ω,

i) $A \cup B$ è l'evento che si verifica se si verifica A oppure B;

ii) $A \cap B$ è l'evento che si verifica se si verificano sia A che B;

iii) A^c è l'evento che si verifica se non si verifica A.

Dunque una *buona famiglia* di eventi deve essere tale da garantire che tutti gli insiemi ottenuti componendo eventi (con le operazioni classiche: unione, intersezione, complementare) sia ancora un evento (ovvero devo poterne calcolare ancora la probabilità)! Una buona famiglia di eventi si chiama σ-algebra. Vediamone la definizione.

Definizione 4.1.2. Una famiglia di sottoinsiemi di Ω si dice una *σ-algebra,* se soddisfa le seguenti proprietà:

- $\Omega \in \mathcal{F}$;
- dato $A \in \mathcal{F}$, allora $A^c \in \mathcal{F}$;
- dati $A_1, A_2, \ldots$ in $\mathcal{F}$ allora $A_1 \cup A_2 \cup \ldots \in \mathcal{F}$.
- dati $A_1, A_2, \ldots$ in $\mathcal{F}$ allora $A_1 \cap A_2 \cap \ldots \in \mathcal{F}$.

Esempio 4.1.3. Torniamo al lancio del dado. Siano A l'evento *"esce un numero pari"*, B l'evento *"esce un numero dispari"* e C l'evento *"esce un multiplo di 3"*. Determiniamo $A \cup B$, $A \cap B$ e C^c. Si ha

$$\begin{aligned} A &= \text{"esce un numero pari"} = \{2,4,6\} \\ B &= \text{"esce un numero dispari"} = \{1,3,5\} \\ C &= \text{"esce un multiplo di 3"} = \{3,6\}, \end{aligned}$$

allora

$$A \cup B = \Omega, \quad A \cap B = \emptyset, \quad C^c = \{1,2,4,5\}.$$

Esempio 4.1.4. Si lancia tre volte una moneta. In questo caso lo spazio campionario può essere descritto dal seguente insieme.

$$\Omega = \{(TTT), (TTC), (TCT), (TCC), (CTT), (CTC), (CCT), (CCC)\}.$$

Siano Siano A l'evento *"due o più teste"* e B l'evento *"tutti i lanci stesso risultato"*. Determiniamo $A \cup B$ e $A \cap B$.

$$\begin{aligned} A &= \text{"due o più teste"} = \{(TTT), (TTC), (CTT), (TCT)\} \\ B &= \text{"tutti i lanci stesso risultato"} = \{(TTT), (CCC)\} \end{aligned}$$

allora

$$A \cup B = \{(TTT), (TTC), (CTT), (TCT), (CCC)\}, \quad A \cap B = \{(TTT)\}.$$

Definizione 4.1.5. Siano Ω uno spazio campionario e $\mathcal{F}$ una famiglia di eventi che sia una σ-algebra. Una (misura di) probabilità su Ω è una funzione

$$\mathbb{P} : \mathcal{F} \longrightarrow [0,1],$$

tale che

i) $\mathbb{P}(\Omega) = 1$;

ii) Dati $A_1, A_2, \ldots$ eventi in $\mathcal{F}$ disgiunti $\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots$.

L'Osservazione che segue contiene alcune importanti proprietà di una probabilità.

Osservazione 4.1.6. 1. Si osservi che per ogni evento A, si ha $A \cap A^c = \emptyset$ e $A \cup A^c = \Omega$, quindi

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c),$$

da cui

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

2. Se A e B sono due eventi in Ω allora

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

ATTENZIONE! Una probabilità su Ω si intende nota quando è possibile calcolare la probabilità di un qualunque evento in $\mathcal{F}$.

Definizione 4.1.7. La terna spazio campionario, famiglia di eventi, probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ prende il nome di spazio di probabilità.

4.2 Spazi di probabilità finiti

Sia Ω uno spazio campionario finito, ovvero $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. In questo caso la *buona famiglia,* ovvero la σ-algebra $\mathcal{F}$, è quella formata da tutti i sottoinsiemi di Ω. Ovvero, per chi lo ricorda, dall'insieme delle *parti* di Ω. Per definire una probabilità su Ω occorre e basta assegnare la probabilità degli eventi elementari, ovvero dei sottoinsiemi formati dai singoli punti. Precisamente occorre conoscere

$$\mathbb{P}(\{\omega_i\}) = p_i \quad \text{per ogni} \quad i = 1, 2 \ldots, n.$$

Si ha,

- $0 \leq p_i \leq 1$ perché le p_i sono delle probabilità;
- $p_1 + p_2 + \ldots p_n = 1$, in quanto la somma delle p_i dà la probabilità di Ω.

In questo modo è ben definita $\mathbb{P}(A)$ per ogni evento. Infatti $\mathbb{P}(A)$ si ottiene sommando la probabilità degli eventi elementari che compongono A. Come caso particolare abbiamo quello in cui tutti gli eventi elementari hanno la stessa probabilità (lancio del dado: tutti gli esiti sono equiprobabili). Tali spazi si chiamano *uniformi* o *equiprobabili.* In tal caso

$$\mathbb{P}(\{\omega_i\}) = \frac{1}{n} \quad \text{per ogni} \quad i = 1, 2 \ldots, n,$$

e

$$\mathbb{P}(A) = \frac{r}{n},$$

essendo r pari alla cardinalità di A.

Esempio 4.2.1. Torniamo al dado. Ω contiene 6 elementi, quindi qui $n = 6$ allora dovendo calcolare la probabilità dell'evento$A$ =*"esce un pari"* basta osservare che A contiene tre elementi, quindi qui $r = 3$ e $\mathbb{P}(A) = \frac{3}{6} = \frac{1}{2}$.

ATTENZIONE! Si può calcolare la probabilità di un evento contando quanti elementi contiene, se solo se, tutti gli eventi elementari hanno la stessa probabilità. Ribadiamo che in generale $\mathbb{P}(A)$ si calcola sommando la probabilità degli eventi elementari che compongono A.

Vediamo di chiarire con qualche altro esempio.

Esempio 4.2.2. [Spazio **non** uniforme] Tre cavalli a, b e c sono in gara. La probabilità che a vinca è doppia di quella che vinca b, che a sua volta è doppia di quella che vinca c. Quali sono le probabilità di vittoria dei tre cavalli? Qual è la probabilità che non vinca a?

Dunque $\Omega = \{a, b, c\}$, e detta p, la probabilità che vinca c, si ha

$$\begin{aligned} \mathbb{P}(\{c\}) &= p \\ \mathbb{P}(\{b\}) &= 2p \\ \mathbb{P}(\{a\}) &= 4p. \end{aligned}$$

Dovendo essere $\mathbb{P}(\Omega) = \mathbb{P}(\{a, b, c\}) = 4p + 2p + p = 1$, si ricava $p = \frac{1}{7}$. Pertanto,

$$\begin{aligned} \mathbb{P}(\{c\}) &= 1/7 \\ \mathbb{P}(\{b\}) &= 2/7 \\ \mathbb{P}(\{a\}) &= 4/7. \end{aligned}$$

La Probabilità che non vinca a è $\mathbb{P}(\{a\}^c) = \mathbb{P}(\{b, c\}) = \frac{2}{7} + \frac{1}{7} = \frac{3}{7}$, o anche $\mathbb{P}(\{a\}^c) = 1 - \mathbb{P}(\{a\}) = 1 - \frac{4}{7} = \frac{3}{7}$.

Esempio 4.2.3. [Spazio uniforme] Si sceglie a caso una carta da una mazzo standard da 52. Siano $A =$*"esce quadri"* e $B =$*"esce una figura"* calcoliamo $\mathbb{P}(A)$, $\mathbb{P}(B)$ e $\mathbb{P}(A \cap B)$.

Intanto

$$\Omega = \{1_\heartsuit, 2_\heartsuit, \ldots, K_\heartsuit, 1_\diamondsuit, 2_\diamondsuit, \ldots, K_\diamondsuit, 1_\spadesuit, 2_\spadesuit, \ldots, K_\spadesuit, 1_\clubsuit, 2_\clubsuit, \ldots, K_\clubsuit\}.$$

Ω contiene 52 eventi elementari tutti aventi stessa probabilità, pertanto per ogni $\omega \in \Omega$, $\mathbb{P}(\{\omega\}) = 1/52$. Inoltre

$$\begin{aligned} A &= \{1_\diamondsuit, 2_\diamondsuit, \ldots, K_\diamondsuit\} \\ B &= \{J_\heartsuit, Q_\heartsuit, K_\heartsuit, J_\diamondsuit, Q_\diamondsuit, K_\diamondsuit, J_\spadesuit, Q_\spadesuit, K_\spadesuit, J_\clubsuit, Q_\clubsuit, K_\clubsuit\} \\ A \cap B &= \{J_\diamondsuit, Q_\diamondsuit, K_\diamondsuit\}. \end{aligned}$$

Ora basta contare quanti elementi ha ognuno di questi insiemi: A ha 13 elementi, B 12 elementi e $A \cap B$ 3 elementi, pertanto

$$\mathbb{P}(A) = \frac{13}{52}, \quad \mathbb{P}(B) = \frac{12}{52} \quad \text{e} \quad \mathbb{P}(A \cap B) = \frac{3}{52}.$$

4.3 Spazi di probabilità infiniti

Gli spazi di probabilità infiniti possono essere divisi in due grandi categorie molto diverse tra loro: numerabili e continui

Spazi numerabili

In questo caso

$$\Omega = \{\omega_1, \omega_2, \ldots\}.$$

Questi sono una generalizzazione degli spazi finiti (che infatti sono inclusi in quelli numerabili). Si procede come nel caso finito. In questo caso la *buona famiglia,* ovvero la σ-algebra $\mathcal{F}$, è quella formata da tutti i sottoinsiemi di Ω. Ovvero, per chi lo ricorda, dall'insieme delle *parti* di Ω. Occorre e basta assegnare la probabilità degli eventi elementari. Precisamente occorre conoscere

$$\mathbb{P}(\{\omega_i\}) = p_i \quad \text{per ogni} \quad i = 1, 2 \ldots$$

Si ha,

- $0 \leq p_i \leq 1$ perché le p_i sono delle probabilità;
- $\sum_{i=1}^{+\infty} p_i = p_1 + p_2 + \ldots = 1$, in quanto la somma delle p_i dà la probabilità di Ω.

Si noti che siamo dinanzi a somme infinite, più precisamente a *serie.* Esse richiedono tecniche molto più sofisticate che non le somme finite. Anche in questo caso la probabilità di un qualsiasi evento A è la somma finita o infinita delle probabilità dei singoli eventi elementari contenuti in A.

Spazi continui

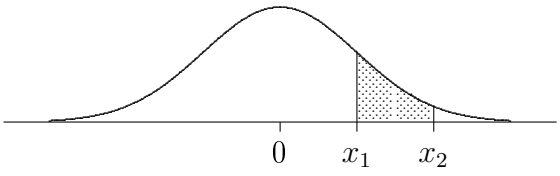
In questo caso $\Omega = (a, b)$ è un intervallo di $\mathbb{R}$, eventualmente tutto $\mathbb{R}$. Fate attenzione: in questo caso *non* si può procedere assegnando la probabilità di tutti gli eventi elementari! Si procede assegnando una funzione $f \geq 0$, definita su (a, b) con

$$\int_a^b f(x)\, dx = 1,$$

tale che per ogni $x_1 < x_2$ si abbia

$$\mathbb{P}((x_1, x_2)) = \int_{x_1}^{x_2} f(x)\, dx.$$

Quindi la σ-algebra $\mathcal{F}$ qui è quella *generata* dagli intervalli di $\mathbb{R}$. Ovvero $\mathcal{F}$ contiene tutti gli intervalli di $\mathbb{R}$, tutti i complementari, tutte le intersezioni numerabili, tutte le unioni numerabili.



Esempio 4.3.1. Per esempio se f è la funzione rappresentata in figura, allora $\Omega = \mathbb{R}$ e la probabilità dell'intervallo (x_1, x_2) è pari all'area ombreggiata.

Capitolo 5

Probabilità condizionata, indipendenza

5.1 Probabilità condizionata

Sia E un evento arbitrario in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, con $\mathbb{P}(E) > 0$. La probabilità di un evento A sapendo che E si è verificato, si chiama probabilità *condizionata di A dato E* e si indica con $\mathbb{P}(A|E)$. Per definizione

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)}. \tag{5.1}$$

Esempio 5.1.1. Si lancia una coppia di dadi. Si sa (qualcuno lo ha visto) che la somma è sei. Calcoliamo la probabilità che uno dei due dadi abbia dato come esito due.

Costruiamo lo spazio campionario. Si tratta di tutti i possibili risultati,

$$\Omega = \{(1,1), \ldots, (1,6), (2,1), \ldots, (2,6), \ldots, (6,1), \ldots, (6,6)\},$$

si tratta di uno spazio con 36 elementi, tutti equiprobabili, pertanto ciascuna coppia ha probabilità $1/36$.

Sappiamo che la somma è sei, dunque l'evento E, noto è

$$E = \{(1,5), (5,1), (2,4), (4,2), (3,3)\}.$$

Mentre l'evento A di cui dobbiamo calcolare la probabilità condizionata è

$$A = \{(1,2), (2,1), (2,2), (3,2), (2,3), (4,2), (2,4), (5,2), (2,5), (6,2), (2,6)\}.$$

Ora

$$A \cap E = \{(2,4), (4,2)\},$$

Pertanto

$$\mathbb{P}(E) = \frac{5}{36}, \quad \mathbb{P}(A \cap E) = \frac{2}{36}$$

e quindi applicando la (5.1) si ha,

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)} = \frac{\frac{2}{36}}{\frac{5}{36}} = \frac{2}{5} = 0.4.$$

La probabilità dell'evento A, non condizionata, era $\mathbb{P}(A) = \frac{11}{36} \sim 0.31$, quindi la conoscenza dell'evento E ha *alterato* la probabilità di A. Precisamente l'ha fatta aumentare (ha dato un'indicazione utile). Potrebbe anche accadere che la faccia diminuire o anche che la lasci invariata. Quest'ultimo caso risulterà piuttosto interessante.

5.1.1 Intersezione di eventi. Regola del prodotto

La definizione di probabilità condizionata fornisce un metodo di calcolo per la probabilità dell'intersezione di due eventi,

$$\mathbb{P}(A \cap E) = \mathbb{P}(A|E)\mathbb{P}(E). \tag{5.2}$$

Questa formula (che può essere estesa all'intersezione di n eventi), permette di calcolare la probabilità di un evento che sia il risultato di una successione finita di esperimenti aleatori. Vediamo un esempio chiarificatore.

Esempio 5.1.2. Sono date tre scatole.

- La scatola A contiene 10 lampade: 4 difettose;
- La scatola B contiene 6 lampade: 1 difettosa;
- La scatola C contiene 8 lampade: 3 difettose.

Una scatola viene scelta a caso, quindi da essa scegliamo una lampada a caso. In questo caso la successione è formata da due esperimenti:

i) si sceglie la scatola a caso;

ii) si sceglie la lampada dalla scatola.

Sia $D =$*"lampada difettosa"*. Come posso calcolare $\mathbb{P}(D)$?

La lampada difettosa può essere pescata dalla scatola A (evento $D \cap A$), dalla B (evento $D \cap B$) o dalla C (evento $D \cap C$). Pertanto,

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

e per la (5.2) si ha

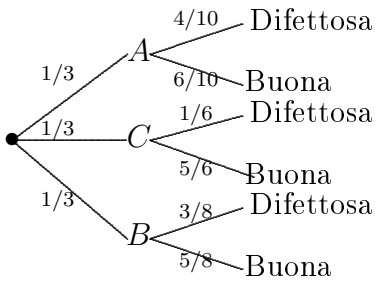
$$\mathbb{P}(D) = \mathbb{P}(D|A)\mathbb{P}(A) + \mathbb{P}(D|B)\mathbb{P}(B) + \mathbb{P}(D|C)\mathbb{P}(C) = \frac{4}{10} \cdot \frac{1}{3} + \frac{1}{6} \cdot \frac{1}{3} + \frac{3}{8} \cdot \frac{1}{3} = \frac{113}{360}.$$

La situazione si può rappresentare con il seguente **diagramma ad albero.**

Considero i possibili cammini per prendere una lampada difettosa: posso prendere la lampada difettosa da A, in tal caso $\mathbb{P}(A \cap D)$ è il prodotto delle probabilità segnate sul cammino che porta alla lampada difettosa passando per A, ovvero $\frac{1}{3}\frac{4}{10}$. Ripetendo lo stesso ragionamento per la lampada difettosa presa da B e poi da C ed essendo i tre cammini incompatibili, si ottiene

$$\mathbb{P}(D) = \frac{4}{10} \cdot \frac{1}{3} + \frac{1}{6} \cdot \frac{1}{3} + \frac{3}{8} \cdot \frac{1}{3} = \frac{113}{360},$$

come già trovato in precedenza.



5.2 Formula di Bayes

La formula di Bayes serve per calcolare la probabilità condizionata di un evento B dato un evento A, quando è nota la probabilità di A dato B. Essa si ottiene facilmente osservando che $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$ in quanto l'intersezione è commutativa, pertanto applicando la (5.2), si ha

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

da cui, essendo $\mathbb{P}(A) \neq 0$ (si ricorda chè si può condizionare solo rispetto ad eventi di probabilità non nulla), si ha

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}. \tag{5.3}$$

Nella pratica quando un esperimento si compone di più esperimenti aleatori in successione temporale (come nel caso del paragrafo precedente: prima scelgo la scatola, poi scelgo la lampada), è facile calcolare la probabilità di un evento successivo dato uno precedente. Sempre riferendosi al caso precedente è facile calcolare la probabilità di ottenere una lampada difettosa se già so che scatola ho scelto! Più complicato, è conoscendo l'esito del risultato finale, calcolare la probabilità di un evento precedente. Sempre riferendosi all'esempio precedente, mi posso domandare qual è la probabilità di aver scelto la scatola A sapendo che la lampada è difettosa. Qui ci aiuta la formula di Bayes! La probabilità richiesta è allora $\mathbb{P}(A|D)$. Per la formula di Bayes si ha

$$\mathbb{P}(A|D) = \frac{\mathbb{P}(D|A)\mathbb{P}(A)}{\mathbb{P}(D)} = \frac{\frac{4}{10} \cdot \frac{1}{3}}{\frac{113}{360}} = \frac{48}{113}.$$

Si osservi che, guardando l'albero, $\mathbb{P}(D|A)$ si calcola facendo il rapporto tra la probabilità del cammino che porta ad una lampada difettosa passando per A diviso la somma delle probabilità di tutti i cammini che portano ad una lampada difettosa. Quindi dà la misura di quanto il cammino passante per A *pesi* rispetto a tutti cammini che portano all'esito finale del nostro esperimento (nel caso specifico avere una lampada difettosa).

5.3 Indipendenza

All'inizio di questo capitolo abbiamo visto che se vengono lanciati due dadi la probabilità di avere un 2 è *condizionata* dal fatto di sapere che la somma dei due dadi è 6. Precisamente

avevamo visto che la probabilità era aumentata. Ci sono dei casi in cui la conoscenza di un evento non altera la probabilità di un altro. In tal caso diremo che gli eventi sono *indipendenti.* Più precisamente possiamo dare la definizione seguente.

Definizione 5.3.1. Siano A e B due eventi di probabilità non nulla. A si dice **indipendente** da B se

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

o equivalentemente che

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

Si osservi che se A è indipendente da B allora B è indipendente da A. La definizione precedente formalizza un concetto intuitivo, pertanto è chiaro che il concetto d'indipendenza sia simmetrico. È bene anche osservare che alcune volte l'indipendenza di due eventi è ovvia, altre volte assolutamente no.

Esempio 5.3.2. Si lanci tre volte una moneta equa. Scriviamo lo spazio di probabilità che descrive questo esperimento.

$$\Omega = \{(TTT), (TTC), (TCT), (TCC), (CTT), (CTC), (CCT), (CCC)\}.$$

Questo è formato da otto elementi tutti equiprobabili, quindi la probabilità di ciascuna terna è $1/8$.

Si considerino gli eventi. A=*"primo lancio testa"*, B =*"secondo lancio croce"*, C =*"testa si presenta esattamente due volte consecutive"*. Vediamo se a due a due questi eventi sono o no indipendenti.

Abbiamo

$$\begin{aligned} A &= \{(TTT), (TTC), (TCT), (TCC)\} \\ B &= \{(TCT), (TCC), (CCT), (CCC)\} \\ C &= \{(TTC), (CTT)\}, \end{aligned}$$

pertanto

$$\mathbb{P}(A) = \frac{4}{8} = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{4}{8} = \frac{1}{2}, \quad \mathbb{P}(C) = \frac{2}{8} = \frac{1}{4}.$$

Passiamo a calcolare la probabilità delle intersezioni fatte due a due. Abbiamo

$$\begin{aligned} A \cap B &= \{(TCT), (TCC)\} \\ A \cap C &= \{(TTC)\} \\ B \cap C &= \emptyset, \end{aligned}$$

pertanto

$$\begin{aligned} \mathbb{P}(A \cap B) &= \frac{2}{8} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B) \\ \mathbb{P}(A \cap C) &= \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C) \\ \mathbb{P}(B \cap C) &= 0 \neq \frac{1}{2} \cdot \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(C). \end{aligned}$$

Dunque sono risultati indipendenti gli insiemi A e B (ovvio l'evento A si riferisce al primo lancio e l'evento B al secondo lancio) ma anche A e C sono indipendenti, il che non era ovvio sin dall'inizio. B e C sono invece dipendenti e si osservi che due eventi disgiunti (ovvero con intersezione vuota) sono sempre dipendenti. Fate attenzione! Molti studenti confondono eventi disgiunti con eventi indipendenti, mentre le due cose sono sempre incompatibili. Infatti se A e B sono eventi di probabilità non nulla e disgiunti il verificarsi di uno rende l'altro impossibile pertanto lo condiziona pesantemente.

Capitolo 6

Variabili aleatorie

6.1 Generalità

Definizione 6.1.1. Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Una **variabile aleatoria** su Ω è una funzione che ad ogni $\omega \in \Omega$ associa un numero reale,

$$X : \Omega \longrightarrow E \subset \mathbb{R},$$

tale che, per ogni $a \leq b \in \mathbb{R}$, sia possibile calcolare la probabilità dell'insieme $\{\omega \in \Omega : a < X(\omega) \leq b\}$, che quindi deve essere un evento in $\mathcal{F}$. Al posto di $\{\omega \in \Omega : a < X(\omega) \leq b\}$ scriveremo $\{a < X \leq b\}$, lasciando sottointesa la dipendenza da $\omega \in \Omega$.

L'insieme E dei valori assunti da X è *l'immagine* di X ed è assai importante per la caratterizzazione della variabile aleatoria.

Definizione 6.1.2. Una variabile aleatoria X si dice *discreta,* se l'insieme E è discreto. In particolare si dice *finita* se l'insieme E è un insieme finito, *numerabile* se l'insieme E è un insieme numerabile, (per esempio i numeri interi). Una variabile aleatoria si dice *continua,* se l'insieme E è un insieme continuo (per esempio un intervallo di $\mathbb{R}$).

In realtà la definizione di variabile aleatoria è più complessa, tuttavia questa può bastare per i nostri scopi. Cominciamo col caso più semplice di variabili aleatorie finite.

6.2 Variabili aleatorie finite

6.2.1 Distribuzione

Esempio 6.2.1. Si lancia un dado. Sia X la variabile che vale zero se esce un pari ed uno se esce un numero dispari.

In questo caso $\Omega = \{1, 2, 3, 4, 5, 6\}$ e l'insieme dei valori assunti da X è $E = \{0, 1\}$. Si ha

$$X(1) = X(3) = X(5) = 1$$
$$X(2) = X(4) = X(6) = 0$$

Esempio 6.2.2. Si lanciano due dadi. Sia Y la somma dei due numeri usciti. Lo spazio campionario che descrive tutte le possibilità che si hanno nelle due estrazioni è

$$\Omega = \{(1,1),(1,2),\ldots,(1,6),\ldots,(6,1),(6,2),\ldots,(6,6)\}.$$

In questo caso l'insieme dei valori assunti da X è $E = \{2,3,\ldots,12\}$. Inoltre si ha, per esempio,

$$Y((1,1)) = 2, \quad Y((2,1)) = 3 \quad Y((3,2)) = 5.$$

È chiaro che essendo una variabile aleatoria funzione di esperimenti aleatori, essa assume i suoi valori con una certa probabilità. Più precisamente tornando agli esempi precedenti ci si può fare domande del tipo: qual è la probabilità che X sia zero? Qual è la la probabilità che Y sia cinque? E così via. Vediamo di rispondere.

Esempio 6.2.3. Torniamo alla variabile X dell'Esempio 6.2.1. X, come già detto può assumere valori in $E = \{0,1\}$. Che vuol dire che $X = 0$? Vuol dire che il lancio ha dato un risultato pari. Formalizzando abbiamo che l'evento $\{X = 0\}$ si può scrivere come segue,

$$\{X = 0\} = \{2,4,6\},$$

quindi

$$\mathbb{P}(X = 0) = \mathbb{P}(\{2,4,6\}) = \frac{3}{6} = \frac{1}{2}.$$

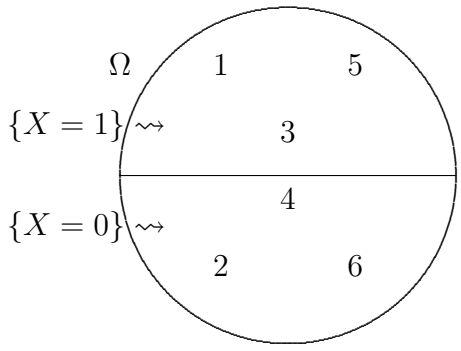
Analogamente, che vuol dire $X = 1$? Vuol dire che il lancio ha dato un risultato dispari. Formalizzando abbiamo che l'evento $\{X = 1\}$ si può scrivere come segue,

$$\{X = 1\} = \{1,3,5\},$$

quindi

$$\mathbb{P}(X = 1) = \mathbb{P}(\{1,3,5\}) = \frac{3}{6} = \frac{1}{2}.$$

Volendo rappresentare graficamente abbiamo:



Abbiamo così *partizionato* Ω attraverso X. Possiamo riassumere quello che abbiamo trovato nella seguente tabella, in cui in una colonna riportiamo i valori assunti da X e nell'altra le probabilità con cui questi valori vengono assunti.

X	Prob.
0	1/2
1	1/2

Tutto ciò vi ricorda qualcosa??!! Quanto fa la somma della colonna delle probabilità? È un caso?

Passiamo all'altra variabile aleatoria vista sino ad ora.

Esempio 6.2.4. Torniamo alla variabile Y dell'Esempio 6.2.2. Y, come già detto può assumere valori in $E = \{2, 3, \ldots, 12\}$. Che vuol dire che $Y = 2$? Vuol dire che i due dadi hanno dato come risultato $(1, 1)$. Formalizzando abbiamo che l'evento $\{Y = 2\}$ si può scrivere come segue,

$$\{Y = 2\} = \{(1, 1)\},$$

quindi

$$\mathbb{P}(Y = 2) = \mathbb{P}(\{(1, 1)\}) = \frac{1}{36}.$$

Cerchiamo anche qui di scrivere $\mathbb{P}(Y = k)$ per $k = 2, 3, \ldots, 12$. Sono un po' più di valori, ma con un po' di pazienza dovremmo farcela...

$$\begin{aligned}
\mathbb{P}(Y = 2) &= \mathbb{P}(\{(1, 1)\}) = \frac{1}{36} \\
\mathbb{P}(Y = 3) &= \mathbb{P}(\{(1, 2), (2, 1)\}) = \frac{2}{36} \\
\mathbb{P}(Y = 4) &= \mathbb{P}(\{(1, 3), (3, 1), (2, 2)\}) = \frac{3}{36} \\
\mathbb{P}(Y = 5) &= \mathbb{P}(\{(1, 4), (4, 1), (2, 3), (3, 2)\}) = \frac{4}{36} \\
\mathbb{P}(Y = 6) &= \mathbb{P}(\{(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)\}) = \frac{5}{36} \\
\mathbb{P}(Y = 7) &= \mathbb{P}(\{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}) = \frac{6}{36} \\
\mathbb{P}(Y = 8) &= \mathbb{P}(\{(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)\}) = \frac{5}{36} \\
\mathbb{P}(Y = 9) &= \mathbb{P}(\{(3, 6), (6, 3), (4, 5), (5, 4)\}) = \frac{4}{36} \\
\mathbb{P}(Y = 10) &= \mathbb{P}(\{(4, 6), (6, 4), (5, 5)\}) = \frac{3}{36} \\
\mathbb{P}(Y = 11) &= \mathbb{P}(\{(5, 6), (6, 5))\}) = \frac{2}{36} \\
\mathbb{P}(Y = 12) &= \mathbb{P}(\{(6, 6)\}) = \frac{1}{36}
\end{aligned}$$

Possiamo riassumere quello che abbiamo trovato nella seguente tabella (fatta in orizzontale per motivi di spazio!), in cui in una riga riportiamo i valori assunti da X e nell'altra le probabilità con cui questi valori vengono assunti.

Y	2	3	4	5	6	7	8	9	10	11	12
Prob	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Tutto ciò vi ricorda qualcosa??!! Quanto fa la somma della colonna delle probabilità? È di nuovo un caso? Forse no.

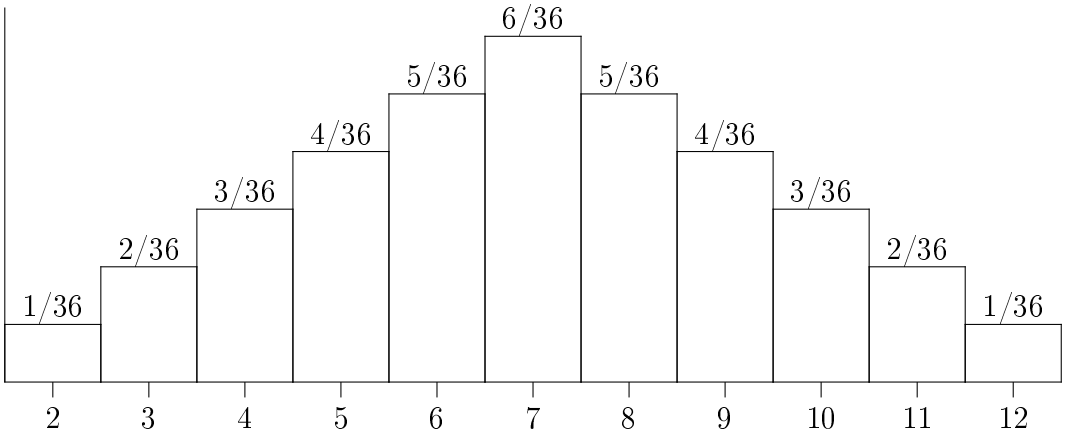
Vediamo di generalizzare. Cominciamo dal caso (visto negli esempi precedenti) in cui la variabile aleatoria assume un numero finito di valori. Abbiamo quindi una variabile aleatoria X che assume valori in $E = \{x_1, x_2, \ldots, x_n\}$, ciascuno con una certa probabilità: $\mathbb{P}(X = x_i) = p_i$, $i = 1, 2, \ldots, n$. Riportando in una tabella si ha

X	x_1	x_2	$\ldots$	x_n
Prob	p_1	p_2	$\ldots$	p_n

Si osservi che:

- $0 \leq p_i \leq 1$, in quanto sono probabilità;
- $\sum_{i=1}^{n} = p_1 + p_2 + \ldots p_n = \mathbb{P}(\Omega) = 1$.

L'insieme dei valori assunti da X e le probabilità con cui vengono assunti si chiama **legge** o **distribuzione di** X**.** La distribuzione di una variabile aleatoria ha molte analogie con una distribuzione di frequenze in cui in una colonna ci sono i valori assunti nell'altra le frequenze relative. Anche graficamente la distribuzione di una variabile aleatoria si può rappresentare graficamente con un istogramma. Vediamo la rappresentazione grafica della variabile Y dell'Esempio 6.2.2.



Osservazione 6.2.5. Si osservi che le probabilità relative ai valori assunti dalla variabile Y hanno a che fare con le aree individuate dall'istogramma che descrive la sua distribuzione. Più precisamente sempre riferito alla variabile Y somma di dadi, se vogliamo, per esempio, calcolare $\mathbb{P}(Y = 5)$ basterà andare a vedere qual è l'area del rettangolo centrato in 5 (4/36). Se invece vogliamo calcolare una probabilità più complicata, per esempio $\mathbb{P}(4 < Y \leq 6)$ basterà andare a sommare le aree interessate: $4 < Y \leq 6$ vuol dire $Y = 5$ oppure $Y = 6$ Quindi la probabilità richiesta è la somma delle aree dei due rettangoli interessati,

$$\mathbb{P}(4 < Y \leq 6) = \mathbb{P}(Y = 5) + \mathbb{P}(Y = 6) = \frac{4}{36} + \frac{5}{36} = \frac{9}{36} = \frac{1}{4}.$$

6.2.2 Media e varianza

In analogia a quanto fatto per le distribuzioni di frequenze possiamo introdurre la media e la varianza di una variabile aleatoria.

Definizione 6.2.6. Si chiama **media** di X e si indica con μ_X o $\mathbb{E}[X]$ (la $\mathbb{E}$ sta per expectation, in inglese aspettazione o media), la quantità

$$\mu_X = \mathbb{E}[X] = \sum_{i=1}^{n} x_i p_i = x_1 \cdot p_1 + x_2 \cdot p_2 + \dots x_n \cdot p_n. \tag{6.1}$$

Come per le distribuzioni di frequenze la media è un indice di posizione. Dà indicazioni intorno a quali valori la variabile aleatoria è posizionata.

Definizione 6.2.7. Si chiama **varianza** di X e si indica con σ_X^2 o $\mathrm{Var}[X]$, la quantità

$$\mathrm{Var}[X] = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i = (x_1 - \mu_X)^2 \cdot p_1 + (x_2 - \mu_X)^2 \cdot p_2 + \dots (x_n - \mu_X)^2 \cdot p_n. \tag{6.2}$$

Definizione 6.2.8. Si chiama **scarto quadratico medio** o anche **deviazione standard** di X e si indica con σ_X, la radice della varianza, ovvero

$$\sigma_X = \sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2 p_i}. \tag{6.3}$$

Si osservi che se i dati x_i rappresentano, ad esempio, lunghezze misurate in metri, la media (e tutti gli altri indici di posizione) e la deviazione standard sono misurate in metri, mentre la varianza è misurata in metri quadri. La varianza e la deviazione standard sono grandezza *non* negative, che si annullano solo quando gli x_i sono tutti uguali tra loro e quindi uguali alla loro media. È utile per il calcolo esplicito della varianza (la cui dimostrazione è lasciata per esercizio agli studenti più interessati e volenterosi!)

$$\sigma_X^2 = \sum_{i=1}^{n} x_i^2 p_i - \mu_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{6.4}$$

Vediamo un esempio.

Esempio 6.2.9. Sia Y la variabile aleatoria dell'Esempio 6.2.2. Si ha

$$\mathbb{E}[Y] = 2\cdot\frac{1}{36}+3\cdot\frac{2}{36}+4\cdot\frac{3}{36}+5\cdot\frac{4}{36}+6\cdot\frac{5}{36}+7\cdot\frac{6}{36}+8\cdot\frac{5}{36}+9\cdot\frac{4}{36}+10\cdot\frac{3}{36}+11\cdot\frac{2}{36}+12\cdot\frac{1}{36} = 7.$$

Inoltre

$$\mathbb{E}[Y^2] = 2^2 \cdot \frac{1}{36} + 3^2 \cdot \frac{2}{36} + 4^2 \cdot \frac{3}{36} + 5^2 \cdot \frac{4}{36} + \dots + 12^2 \cdot \frac{1}{36} = \frac{1974}{36}$$

pertanto

$$\mathrm{Var}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \frac{1974}{36} - 7^2 = 5.83$$

6.3 Variabili aleatorie numerabili

6.3.1 Distribuzione

Supponiamo ora che X sia una variabile casuale su Ω con un insieme immagine infinitamente numerabile $E = \{x_1, x_2, \ldots\}$. In questo caso la distribuzione di X si definisce come nel caso precedente (a parte accorgimenti tecnici). La variabile X assume valori in E, ciascuno con una certa probabilità: $\mathbb{P}(X = x_i) = p_i$, $i = 1, 2, \ldots$ Riportando in una tabella si ha

X	x_1	x_2	$\ldots$
Prob	p_1	p_2	$\ldots$

Si osservi che anche in questo caso:

- $0 \leq p_i \leq 1$, in quanto sono probabilità;
- $\sum_{i=1}^{+\infty} p_i = p_1 + p_2 + \ldots = \mathbb{P}(\Omega) = 1$.

Esempio 6.3.1. Sia X la prima volta che esce sei in lanci ripetuti di un dado (tempo di primo successo). X può assumere valori in $E = \{1, 2, 3, \ldots\}$. Cerchiamo di trovare la distribuzione di questa variabile aleatoria. Indichiamo con A_1=*"esce 6 al primo lancio"*, A_2=*"esce 6 al secondo lancio"*, ecc... Questi eventi sono chiaramente indipendenti (ognuno relativo ad un lancio diverso), ed ognuno ha ovviamente probabilità $1/6$. Passiamo a calcolare la distribuzione di X.
Che vuol dire $X = 1$? Vuol dire che al primo lancio è uscito 6, quindi

$$\mathbb{P}(X = 1) = \mathbb{P}(A_1) = \frac{1}{6}.$$

Che vuol dire $X = 2$? Vuol dire che al primo lancio *non* è uscito 6 ed è invece uscito 6 al secondo lancio, ovvero, data l'indipendenza dei lanci,

$$\mathbb{P}(X = 2) = \mathbb{P}(A_1^c \cap A_2) = \mathbb{P}(A_1^c)\mathbb{P}(A_2) = \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36}.$$

Che vuol dire $X = 3$? Vuol dire che ai primi due lanci *non* è uscito 6 ed è invece uscito 6 al terzo lancio, ovvero, data l'indipendenza dei lanci,

$$\mathbb{P}(X = 3) = \mathbb{P}(A_1^c \cap A_2^2 \cap A_3) = \mathbb{P}(A_1^c)\mathbb{P}(A_2^c)\mathbb{P}(A_3) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^2 \frac{1}{6} = \frac{26}{216}.$$

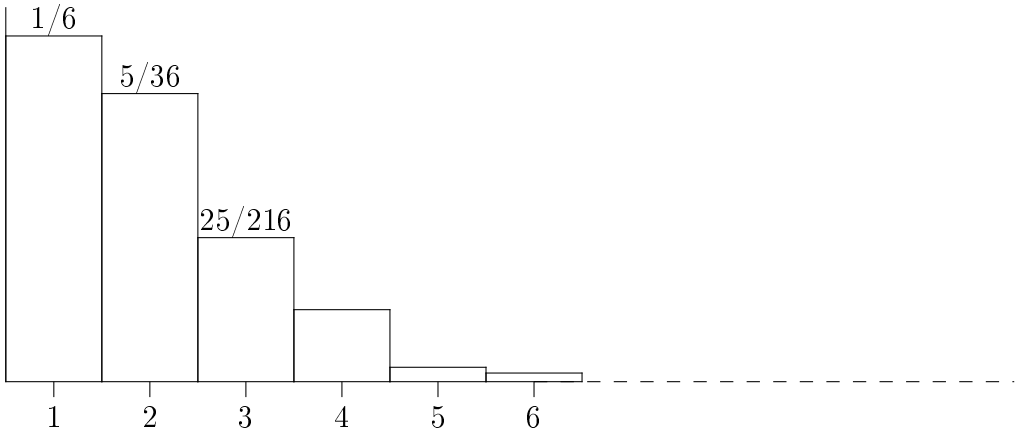
Proviamo a generalizzare.
Che vuol dire $X = k$? Vuol dire che i primi $k-1$ lanci *non* hanno dato 6 ed il k-esimo lancio ha dato 6. In formule, per $k = 1, 2, \ldots$,

$$\mathbb{P}(X = k) = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}.$$

Volendo riportare i risultati in una tabella (infinita!), abbiamo:

X	1	2	3	$\ldots$
Prob	1/6	5/36	25/216	$\ldots$

Volendo fare una rappresentazione grafica, abbiamo il seguente istogramma (infinito!)



Osservazione 6.3.2. Si osservi che, anche in questo caso, le probabilità relative ai valori assunti dalla variabile hanno a che fare con le aree individuate dall'istogramma che descrive la sua distribuzione. Più precisamente sempre riferito alla variabile X prima volta che esce il numero 6, se vogliamo, per esempio, calcolare $\mathbb{P}(X = 1)$ basterà andare a vedere qual è l'area del rettangolo centrato in 1 (1/6). Se invece vogliamo calcolare una probabilità più complicata, per esempio $\mathbb{P}(X < 4)$ basterà andare a sommare le aree interessate: $X < 4$ vuol dire $X = 1$ oppure $X = 2$ oppure $X = 3$ Quindi la probabilità richiesta è la somma delle aree dei tre rettangoli interessati,

$$\mathbb{P}(X < 4) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = \frac{1}{6} + \frac{5}{36} + \frac{25}{216} = \frac{91}{216}.$$

6.3.2 Media e Varianza

Anche in questo caso possiamo definire media e varianza di una variabile aleatoria numerabile come nel caso precedente.

La media e la varianza di X sono definite da

$$\mu_X = \mathbb{E}[X] = \sum_{i=1}^{+\infty} x_i p_i = x_1 p_1 + x_2 p_2 + \dots$$

e

$$\sigma_X^2 = \mathrm{Var}[X] = \sum_{i=1}^{+\infty} (x_i - \mu_X)^2 p_i = (x_1 - \mu_X)^2 \cdot p_1 + (x_2 - \mu_X)^2 \cdot p_2 + \dots,$$

rispettivamente, quando le rispettive serie sono assolutamente convergenti. Si può dimostrare che $\mathrm{Var}[X]$ esiste se e solo se esiste $\mathbb{E}[X^2] = \sum_{i=1}^{+\infty} x_i^2 p_i$ ed in tal caso, come nel caso finito,

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Quando la varianza di X esiste, lo scarto quadratico medio è definito da

$$\sigma_X = \sqrt{\mathrm{Var}[X]}.$$

Nel caso di variabili numerabili ci limiteremo al caso di variabili aleatorie in cui media e varianza sono note. Insomma non sarà richiesto di sommare delle serie. Nel capitolo successivo vedremo qualche esempio.

6.4 Variabili aleatorie continue

6.4.1 Distribuzione

Supponiamo ora che l'insieme E sia un insieme continuo. Per esempio un intervallo di $\mathbb{R}$. Rammentiamo che, dalla definizione di variabile aleatoria, deve essere possibile calcolare $\mathbb{P}(a < X \leq b)$ per ogni $a \leq b$. In tal caso allora, dato che non è possibile assegnare la probabilità dei singoli punti in E, supporremo che esista una funzione $f : \mathbb{R} \to \mathbb{R}$, tale che $\mathbb{P}(a < X < b)$ sia uguale all'area sottesa dal grafico di f tra $x = a$ e $x = b$. Nel linguaggio del calcolo

$$\mathbb{P}(a < X \leq b) = \int_a^b f(x)\, dx. \tag{6.5}$$

In questo caso la funzione f è detta distribuzione o densità di X. Essa è non nulla nell'intervallo in cui la X prende valori (quindi in E) e soddisfa le seguenti proprietà:

- $f(x) \geq 0$ per ogni $x \in \mathbb{R}$;
- $\int_{\mathbb{R}} f(x)dx = 1$.

Ovvero, f è *non* negativa e l'area totale sottesa dal suo grafico è pari ad 1. Il grafico di f è l'analogo dell'istogramma visto per variabili discrete. Intuitivamente è come se, dato che la variabile può assumere tutti i valori in un certo intervallo, si prendesse un istogramma con classi sempre più piccole, al limite otterremo il profilo di una funzione.

Dalla (6.5) segue immediatamente che

$$\mathbb{P}(X = a) = \mathbb{P}(a < X \leq a) = \int_a^a f(x)\, dx = 0 \quad \forall\, a \in \mathbb{R},$$

pertanto, per una variabile aleatoria continua

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b).$$

Attenzione!!! Per una variabile finita o numerabile questo è falso, in quanto la probabilità che $X = a$ non è sempre zero.

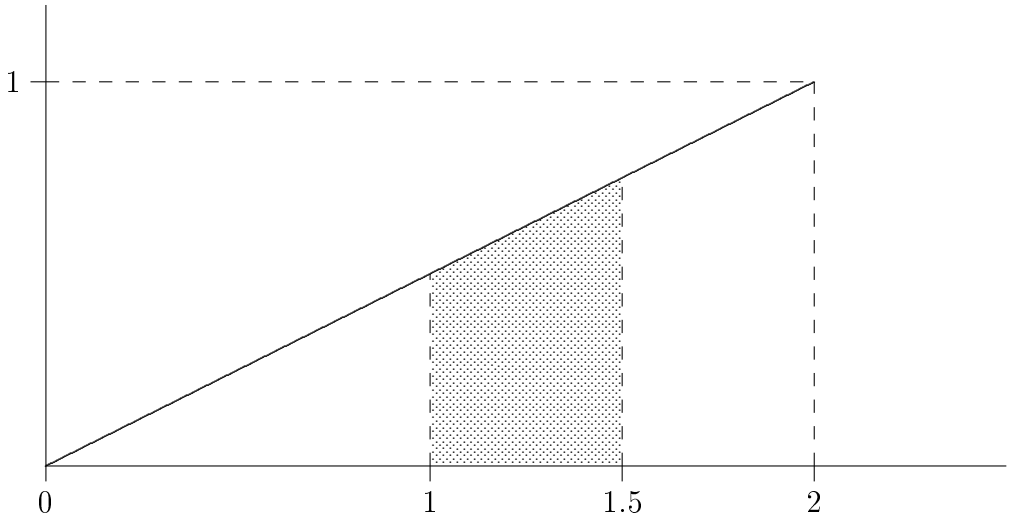
Vediamo anche qui un esempio.

Esempio 6.4.1. Sia X la variabile aleatoria continua con funzione densità f data da

$$f(x) = \begin{cases} \frac{1}{2}x & \text{se} \quad 0 \leq x \leq 2 \\ 0 & \text{altrimenti,} \end{cases}$$

Verifichiamo che si tratta veramente di una densità e calcoliamo $\mathbb{P}(1 < X < 1.5)$.

Tracciamo il grafico della funzione f.



Intanto verifichiamo che f sia veramente una densità. Che $f \geq 0$ è immediato. Mostriamo che l'area sottesa dal suo grafico è uno.

$$\int_{\mathbb{R}} f(x)\,dx = \int_0^2 \frac{1}{2}x\,dx = \frac{1}{4}x^2\Big|_0^2 = 1.$$

Dunque f è una densità. Per calcolare la probabilità richiesta occorre calcolare l'area ombreggiata, ovvero

$$\mathbb{P}(1 \leq X \leq 1.5) = \int_1^{1.5} \frac{1}{2}x\,dx = \frac{1}{4}x^2\Big|_1^{3/2} = \frac{5}{16}.$$

Analogamente a quanto fatto per le distribuzioni di frequenze possiamo qui introdurre, *mediana, quartili, percentili.* Più in generale data la distribuzione di una variabile aleatoria X possiamo dare la seguente definizione.

Definizione 6.4.2. Si definisce **quantile** di ordine $\alpha \in (0,1)$ quel valore q_α sull'asse x tale

$$\mathbb{P}(X \leq q_\alpha) = \alpha.$$

Si osservi, che questa definizione può essere data, ovviamente, per tutte le variabili aleatorie (discrete e continue). La poniamo qui perché noi la utilizzeremo quasi esclusivamente per variabili aleatorie continue. Torniamo all' Esempio 6.4.1.

Esempio 6.4.3. Calcoliamo la mediana, $q_{.5}$ della distribuzione dell'esempio precedente.

Qui $\alpha = 0.5$, pertanto deve essere

$$\mathbb{P}(X \leq q_\alpha) = \int_0^{q_{.5}} \frac{1}{2}x\,dx = 0.5.$$

Da cui

$$\frac{1}{4}x^2\Big|_0^{q_{.5}} = \frac{1}{4}q_{.5}^2 = 0.5, \quad q_{.5} = \sqrt{2} = 1.41.$$

6.4.2 Media e varianza

La media e la varianza di una variabile aleatoria X continua sono definite da

$$\mu_X = \mathbb{E}[X] = \int_{\mathbb{R}} x\, f(x)\, dx$$

e

$$\sigma_X^2 = \mathrm{Var}[X] = \int_{\mathbb{R}} (x - \mu_X)^2\, dx,$$

rispettivamente, quando i rispettivi integrali sono assolutamente convergenti. Si può dimostrare che $\mathrm{Var}[X]$ esiste se e solo se esiste $\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2\, f(x)\, dx$ ed in tal caso, come nel caso finito,

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Quando la varianza di X esiste, lo scarto quadratico medio è definito da

$$\sigma_X = \sqrt{\mathrm{Var}[X]}.$$

Esempio 6.4.4. Calcoliamo media e varianza della variabile aleatoria X dell'Esempio 6.4.1.

$$\mathbb{E}[X] = \int_0^2 x \cdot \frac{1}{2} x\, dx = \frac{1}{2} \int_0^2 x^2\, dx = \frac{1}{6} x^3 \Big|_0^2 = \frac{4}{3}.$$

Per il calcolo della varianza abbiamo bisogno di $\mathbb{E}[X^2]$.

$$\mathbb{E}[X^2] = \int_0^2 x^2 \cdot \frac{1}{2} x\, dx = \frac{1}{2} \int_0^2 x^3\, dx = \frac{1}{8} x^4 \Big|_0^2 = 2,$$

quindi

$$\mathrm{Var}[X] = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}.$$

6.5 Variabili aleatorie indipendenti [cenni]

Partiamo da un esempio.

Esempio 6.5.1. Si lancia una coppia di dadi. Sia X la variabile che fornisce il massimo dei valori usciti e Y la variabile che fornisce la somma dei valori usciti. Vogliamo studiare come sono *collegati* i valori assunti da X e da Y. In questo caso abbiamo già visto che

$$\Omega = \{(1,1), (1,2), \ldots, (1,6), \ldots, (6,1), (6,2), \ldots, (6,6)\}.$$

Si tratta di una spazio equiprobabile che contiene 36 elementi e quindi ciascun elemento (coppia) ha probabilità $1/36$. Per precisione cerchiamo $\mathbb{P}(X = h, Y = k)$ per $h = 1, 2, 3, 4, 5, 6$ e $k = 2, 3, \ldots, 12$..

Calcoliamo la distribuzione di X. Il massimo assume valori in $E = \{1,2,3,4,5,6\}$. Precisamente si ha:

$$\begin{aligned}
\{X=1\} &= \{(1,1)\}\\
\{X=2\} &= \{(1,2),(2,1),(2,2)\}\\
\{X=3\} &= \{(1,3),(3,1),(2,3),(3,2),(3,3)\}\\
\{X=4\} &= \{(1,4),(4,1),(2,4),(4,2),(3,4),(4,3),(4,4)\}\\
\{X=5\} &= \{(1,5),(5,1),(2,5),(5,2),(3,5),(5,3),(4,5),(5,4),(5,5)\}\\
\{X=6\} &= \{(1,6),(6,1),(2,6),(6,2),(3,6),(6,3),(4,6),(6,4),(5,6),(6,5),(6,6)\}
\end{aligned}$$

Da cui si ha:

X	1	2	3	4	5	6
Prob	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

La distribuzione di Y l'abbiamo già calcolata nell'Esempio 6.2.2. La riportiamo qui per completezza.

Y	2	3	4	5	6	7	8	9	10	11	12
Prob	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Adesso cerchiamo di studiare se c'è una qualche relazione tra i valori assunti a X e quelli assunti da Y.

Per esempio: che vuol dire $\{X=1, Y=2\}$? Vuol dire che è uscita la coppia (1,1)! Pertanto $\mathbb{P}(\{X=1,Y=2\}) = 1/36$. Che vuol dire $\{X=1,Y=3\}$? È impossibile!! Pertanto $\mathbb{P}(\{X=1,Y=3\}) = 0$. Rappresentiamo con una tabella tutte le probabilità.

$X \diagdown Y$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	0	0	0	0	0	0	0	0	0	0
2	0	2/36	1/36	0	0	0	0	0	0	0	0
3	0	0	2/36	2/36	1/36	0	0	0	0	0	0
4	0	0	0	2/36	2/36	2/36	1/36	0	0	0	0
5	0	0	0	0	2/36	2/36	2/36	2/36	1/36	0	0
6	0	0	0	0	0	2/36	2/36	2/36	2/36	2/36	1/36

La domanda ora è: le due variabili sono indipendenti? Ovvero il risultato di una variabile è indipendente dal risultato dell'altra? In formule, mi chiedo se

$$\mathbb{P}(X=h|Y=k) = \mathbb{P}(X=h),$$

o equivalentemente

$$\mathbb{P}(X=h, Y=k) = \mathbb{P}(X=h)\mathbb{P}(Y=k).$$

La risposta qui è chiaramente no. Infatti riprendiamo la distribuzione di X e di Y. Si ha

$$\mathbb{P}(X=1,Y=2) = \frac{1}{36} \neq \mathbb{P}(X=1)\mathbb{P}(Y=2) = \frac{1}{36}\cdot\frac{1}{36}.$$

Alla luce di questo esempio possiamo dire che due variabili aleatorie sono indipendenti, se il risultato di una è indipendente dal risultato dell'altra. Per esempio se X è una variabile

che riguarda il primo lancio di un dado e Y riguarda il secondo lancio le due variabili sono ovviamente indipendenti perché i due lanci lo sono. Nell'esempio precedente sia la X che la Y sono variabili legate contemporaneamente al risultato di entrambi i lanci ed era abbastanza presumibile, anche se non necessario, che fossero dipendenti.

6.6 Proprietà della media e della varianza

Siano X e Y due variabili aleatorie sullo spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$, a un numero reale. Vogliamo qui elencare, senza dimostrarle, delle importanti proprietà della media e della varianza che ci risulteranno molto utili in seguito.

- MEDIA

$$\begin{aligned} \mathbb{E}[X+Y] &= \mathbb{E}[X]+\mathbb{E}[Y] \\ \mathbb{E}[a] &= a \\ \mathbb{E}[X+a] &= \mathbb{E}[X]+a \\ \mathbb{E}[aX] &= a\mathbb{E}[X] \end{aligned}$$

- VARIANZA Se X e Y sono indipendenti, si ha

$$\begin{aligned} \mathrm{Var}[X+Y] &= \mathrm{Var}[X]+\mathrm{Var}[Y] \\ \mathrm{Var}[a] &= 0 \\ \mathrm{Var}[X+a] &= \mathrm{Var}[X] \\ \mathrm{Var}[aX] &= a^2\mathrm{Var}[X] \end{aligned}$$

Capitolo 7

Alcune distribuzioni "famose"

7.1 Distribuzione di Bernoulli

La variabile aleatoria con distribuzione di Bernoulli è, forse, la più semplice, ma anche una delle più importanti. Vediamo a cosa corrisponde. Consideriamo un esperimento con due soli esiti; definiamo *successo* uno dei due esiti, *insuccesso* l'altro. Sia p la probabilità di successo, quindi $1-p$ è la probabilità di insuccesso. A questo punto sia X la variabile aleatoria che vale 1 se ottengo il successo 0 se ottengo l'insuccesso. Si ha, ovviamente

$$\mathbb{P}(X=1)=p \qquad \mathbb{P}(X=0)=1-p. \tag{7.1}$$

Sotto forma di tabella

X	0	1
Prob	$1-p$	p

Diremo che X ha distribuzione di Bernoulli di parametro p ($p \in [0,1]$) e scriveremo $X \sim \mathrm{Be}(p)$.

Esempio 7.1.1. Si lanci una moneta equa e sia *testa* il successo. Se X vale 1 quando esce testa e zero quando esce croce, si ha $X \sim \mathrm{Be}(1/2)$.

Calcoliamo media e varianza di una legge di Bernoulli. Applicando la (6.1), si ha

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p),$$

ed anche

$$\mathbb{E}[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p).$$

Pertanto

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

7.2 Distribuzione Binomiale

La variabile aleatoria con distribuzione binomiale è un esempio molto importante di variabile aleatoria finita. Vediamo a cosa corrisponde.

Consideriamo n prove ripetute ed indipendenti di un esperimento con due soli esiti (prove bernoulliane); come nel caso precedente definiamo *successo* uno dei due esiti, *insuccesso* l'altro. Sia p la probabilità di successo (in ogni singola prova), $1-p$ la probabilità di insuccesso. A questo punto sia X la variabile aleatoria che *conta* quanti successi ottenuti nelle prove fatte. X si dice variabile binomiale di parametri n (numero delle prove) e p probabilità di successo (su ogni singola prova). Tale variabile aleatoria si può vedere come la somma di n variabili indipendenti bernoulliane, ciascuna corrispondente ad una prova. In formule, se $Y_1, Y_2, \ldots, Y_n$ sono bernoulliane indipendenti di parametro p, $X = Y_1 + Y_2 + \ldots Y_n$. Scriveremo $X \sim \mathrm{Bi}(n,p)$.

Vediamo qual è la sua distribuzione. Intanto X può assumere valori da 0 (nessun successo) a n (tutti successi), ovvero $E = \{0, 1, 2, \ldots, n\}$. Si può dimostrare che si ha

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2 \ldots, n. \tag{7.2}$$

Chiariamo con degli esempi.

Esempio 7.2.1. Viene lanciata 6 volte una moneta equa. Sia testa il successo. Calcolare la probabilità che

1. testa non esca affatto;
2. esca esattamente una testa;
3. escano almeno due teste.

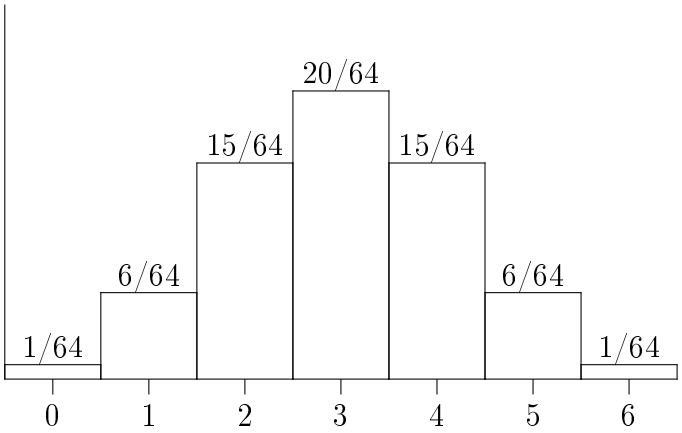
Sia X la variabile aleatoria che conta il numero di teste su 6 lanci. $X \sim \mathrm{Bi}(6, 1/2)$. Si tratta allora di tradurre in termini di X le probabilità richieste.

1. Si richiede $P(X = 0)$, per la (7.2), si ha
$$\mathbb{P}(X = 0) = \binom{6}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^6 = \frac{6!}{0!6!} \frac{1}{2^6} = \frac{1}{64}.$$
2. Si richiede $P(X = 1)$, per la (7.2), si ha
$$\mathbb{P}(X = 1) = \binom{6}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^5 = \frac{6!}{1!5!} \frac{1}{2^6} = \frac{6}{64}.$$
3. Si richiede $\mathbb{P}(X \geq 2)$. Utilizzando quanto sin qui trovato, si ha
$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X < 2) = 1 - (\mathbb{P}(X = 0) + \mathbb{P}(X = 1)) = 1 - \frac{7}{64} = \frac{57}{64}.$$

Per completezza possiamo calcolare la distribuzione di X. Applicando la (7.2) otteniamo la tabella seguente:

X	0	1	2	3	4	5	6
Prob	1/64	6/64	15/64	20/64	15/64	6/64	1/64

Vediamo anche come è fatto l'istogramma di questa distribuzione.



Questa distribuzione è *simmetrica* ed il valore più probabile è chiaramente 3.

Esempio 7.2.2. Un dado viene lanciato 7 volte; un lancio è un successo se esce il numero 1. Calcolare la probabilità

1. di avere 3 successi;
2. di non avere successi.

Anche qui il problema è modellare. Sia X il numero dei lanci che hanno dato 1. $X \sim \mathrm{Bi}(7, 1/6)$. Si tratta di nuovo di tradurre in termini di X le probabilità richieste.

1. Si richiede $P(X = 3)$, per la (7.2), si ha
$$\mathbb{P}(X = 3) = \binom{7}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^4.$$
2. Si richiede $P(X = 0)$, per la (7.2), si ha
$$\mathbb{P}(X = 0) = \binom{7}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^7.$$

Per il calcolo della media e della varianza di una variabile binomiale utilizziamo il fatto che si scrive come somma di variabili bernoulliane indipendenti, pertanto

$$\mathbb{E}[X] = \mathbb{E}[Y_1 + Y_2 + \ldots Y_n] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \ldots + \mathbb{E}[Y_n] = p + p + \ldots + p = np,$$

ed anche

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{Var}[Y_1 + Y_2 + \ldots Y_n] = \mathrm{Var}[Y_1] + \mathrm{Var}[Y_2] + \ldots + \mathrm{Var}[Y_n] \\ &= p(1-p) + p(1-p) + \ldots + p(1-p) = np(1-p).\end{aligned}$$

Riassumendo, se $X \sim \mathrm{Bi}(n, p)$:

$$\mathbb{E}[X] = np, \tag{7.3}$$
$$\mathrm{Var}[X] = np(1-p). \tag{7.4}$$

Osservazione 7.2.3. Estrazioni con reimmissione. Consideriamo un insieme di N oggetti di cui K neri e i restanti $N-K$ bianchi. Supponiamo di estrarre n oggetti con reimmissione. Allora ad ogni estrazione la probabilità di estrarre un oggetto nero è la stessa ed è pari a $p = \frac{K}{N}$ ed ogni estrazione è indipendente dalla precedente. In questo caso detto X il numero di oggetti neri estratti X ha legge $\mathrm{Bi}(n,p)$. Lo stesso risultato si ottiene se le n estrazioni vengono fatte da una popolazione infinita ciascuna con probabilità di successo pari a p.

7.3 Distribuzione Ipergeometrica

Esercizio 1. In una partita di 500 pezzi, il 10% sono difettosi. Un ispettore ne osserva un campione di 20. Determinare la legge della variabile aleatoria X che conta il numero di pezzi difettosi trovati nel campione. Questo esempio apparentemente simile al precedente introduce in realtà una situazione nuova. Il primo pezzo estratto ha probabilità 0.1 di essere estratto. Ma una volta estratto il primo, l'estrazione del secondo non è indipendente dalla prima. Il problema si può risolvere con tecniche di calcolo combinatorio standard. Estraiamo 20 oggetti da un insieme di 500. Questo si può fare in $\binom{500}{20}$ modi. Quindi

$$\text{casi possibili} = \binom{500}{20}.$$

Gli oggetti sono solo di due tipi: 450 *insuccesso* (non difettosi) e 50 *successo* (difettosi). Quante sono le scelte di 20 oggetti che ne contengono k difettosi e quindi $20-k$ non difettosi?

$$\text{casi favorevoli} = \binom{50}{k}\binom{450}{20-k}.$$

Quindi la probabilità che ci siano esattamente k difettosi è:

$$\frac{\binom{50}{k}\binom{450}{20-k}}{\binom{500}{20}}, \quad k = 0, 1, \ldots, 20.$$

La situazione dell'esempio precedente si generalizza nel modo seguente. Supponiamo di estrarre *senza reimmissione* n oggetti da un insieme che ne contiene N, di cui K *successi* e $N-K$ *insuccessi*. La variabile aleatoria X che conta il numero di successi nel campione estratto si dice avere **legge ipergeometrica di parametri** (N,K,n). Si noti che $N \geq K$ e $N \geq n$. Scriveremo $X \sim IG(N,K,n)$.

Vediamo qual è la sua distribuzione. Intanto X può assumere valori da 0 (nessun successo) a n (tutti successi), ovvero $E = \{0,1,2,\ldots,n\}$. Si può dimostrare che si ha

$$\mathbb{P}(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = 0,1,2\ldots n;\ k \leq K;\ (n-k) \leq (N-K). \tag{7.5}$$

Si può inoltre dimostrare che, se $X \sim\sim IG(N,K,n)$:

$$\mathbb{E}[X] = n\frac{K}{N}, \tag{7.6}$$

$$\mathrm{Var}[X] = n\frac{K}{N}\Big(1-\frac{K}{N}\Big)\Big(\frac{N-n}{N-1}\Big). \tag{7.7}$$

Osservazione 7.3.1. Estrazioni con e senza reimmissione. Analogie tra legge binomiale e ipergeometrica C'è una certa analogia tra legge binomiale e legge ipergeometrica: entrambe contano il numero di successi presenti in un campione di ampiezza fissata estratto da un insieme che contiene successi e insuccessi in proporzioni fissate. Quello che cambia è la modalità dell'estrazione: con reimmissione (*binomiale*) senza reimmissione (*ipergeometrica*).

Intuitivamente, i due metodi di estrazione non dovrebbero dare risultati molto diversi quando il numero N di individui della popolazione è grande rispetto all'ampiezza n del campione estratto. In questo caso infatti, anche estraendo con reimmissione è molto improbabile che càpiti di estrarre due volte lo stesso oggetto. In effetti su può dimostrare la seguente

Proposizione 7.3.2. *Sia $X \sim \mathrm{IG}(N, K, n)$. Se facciamo tendere N e K all'infinito in modo che sia costante $p = \frac{K}{N}$ si ha:*

$$\mathbb{P}(X = k) \longrightarrow \binom{n}{k} p^k (1-p)^{n-k}.$$

7.4 Distribuzione Geometrica

Riprendiamo e generalizziamo la variabile aleatoria dell'Esempio 6.3.1. Sia X il tempo di primo successo in prove indipendenti bernoulliane di parametro p, ovvero X è la prima volta in cui compare il successo (il valore 1) in una serie di prove indipendenti. Qual è la distribuzione di X? Intanto siano $Y_1, Y_2, \ldots$ bernoulliane di parametro p corrispondenti alle singole prove. X può assumere valori in $E = \{1, 2, \ldots\}$, quindi è una variabile discrete. Calcoliamo la sua distribuzione.
Che vuol dire $X = 1$? Vuol dire che la prima prova è un successo, quindi

$$\mathbb{P}(X = 1) = \mathbb{P}(Y_1 = 1) = p.$$

Che vuol dire $X = 2$? Vuol dire che la prima prova *non* è un successo ma la seconda prova sì, ovvero, data l'indipendenza delle prove,

$$\mathbb{P}(X = 2) = \mathbb{P}(Y_1 = 0, Y_1 = 1) = \mathbb{P}(Y_1 = 0)\mathbb{P}(Y_2 = 1) = (1-p) \cdot p.$$

Che vuol dire $X = 3$? Vuol dire che le prime due prove sono un insuccesso ed è invece un successo la terza prova, ovvero, data l'indipendenza delle prove,

$$\begin{aligned}\mathbb{P}(X = 3) &= \mathbb{P}(Y_1 = 0, Y_2 = 0, Y_3 = 1) = \mathbb{P}(Y_1 = 0)\mathbb{P}(Y_2 = 0)\mathbb{P}(Y_3 = 1) \\ &= (1-p) \cdot (1-p) \cdot p = (1-p)^2 p.\end{aligned}$$

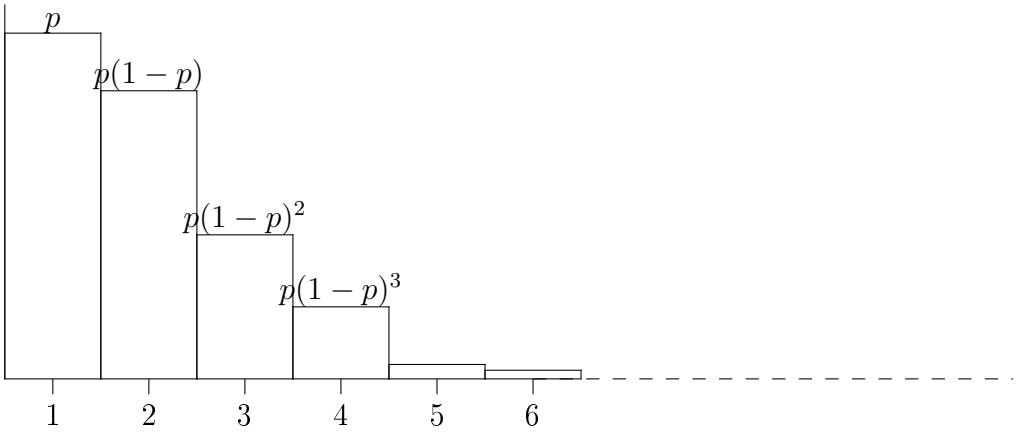
Proviamo a generalizzare.
Che vuol dire $X = k$? Vuol dire che le prime $k-1$ prove *non* hanno dato il successo e la k-esima prova ha dato il successo. In formule, per $k = 1, 2, \ldots$,

$$\mathbb{P}(X = k) = \mathbb{P}(Y_1 = 0, \ldots, Y_{k-1} = 0, Y_k = 1) = (1-p)^{k-1} p.$$

Volendo riportare i risultati in una tabella (infinita!), abbiamo:

X	1	2	3	...
Prob	p	$(1-p)p$	$(1-p)^2p$	...

Volendo fare una rappresentazione grafica, abbiamo il seguente istogramma (infinito!)



La variabile aleatoria X qui descritta, corrispondente al tempo di primo successo in prove indipendenti bernoulliane di parametro p è detta *geometrica* di parametro p. Scriveremo $X \sim \mathrm{Ge}(p)$.

Si può dimostrare che, se $X \sim \mathrm{Ge}(p)$:

$$\mathbb{E}[X] = \frac{1}{p}, \tag{7.8}$$

$$\mathrm{Var}[X] = \frac{1-p}{p^2}. \tag{7.9}$$

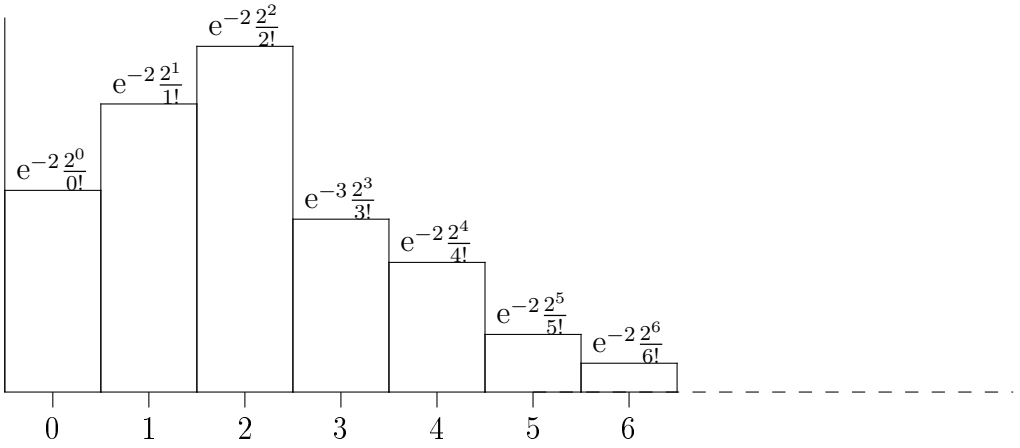
7.5 Distribuzione di Poisson

Diremo che X ha legge di *Poisson* di parametro $\lambda > 0$ e scriveremo $X \sim \mathrm{Po}(\lambda)$, se X ha la seguente distribuzione

$$\mathbb{P}(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{per} \quad k = 0, 1, \ldots.$$

Questa distribuzione infinitamente numerabile si manifesta in molti fenomeni naturali, come il numero di chiamate ad un centralino in un'unità di tempo (minuto, ora, giorno, ecc), il numero di auto che transitano in un certo incrocio sempre in un'unità di tempo e altri fenomeni simili.

Segue il diagramma della distribuzione di Poisson per $\lambda = 2$



Si può dimostrare che, se $X \sim \mathrm{Po}(\lambda)$:

$$\mathbb{E}[X] = \lambda, \tag{7.10}$$

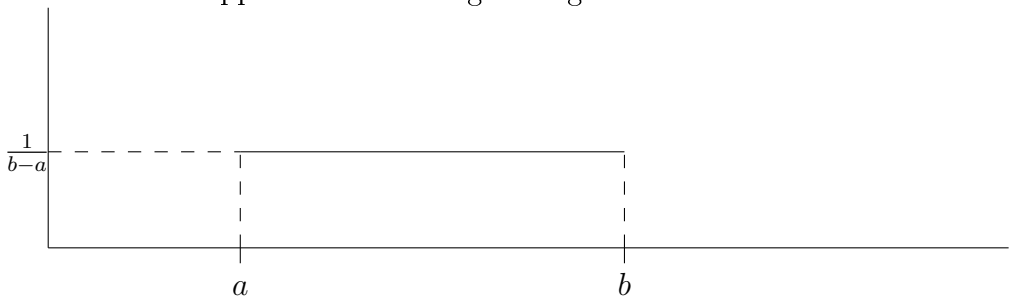
$$\mathrm{Var}[X] = \lambda. \tag{7.11}$$

7.6 Distribuzione uniforme (continua)

Diremo che X ha distribuzione uniforme su (a, b), se ha densità costante data da

$$f(x) = \frac{1}{b-a} \quad \text{per} \quad x \in (a, b).$$

Scriveremo $X \sim \mathrm{Un}(a, b)$. Questa variabile aleatoria assegna stessa probabilità ad intervalli della stessa lunghezza ovunque siano posizionati. È appunto *uniforme*. Se $X \sim \mathrm{Un}(a, b)$, la sua densità è rappresentata dal seguente grafico.



Calcoliamo media e varianza della legge uniforme. Per la media si ha,

$$\mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\, \frac{x^2}{2}\bigg|_a^b = \frac{1}{b-a}\, \frac{b^2-a^2}{2} = \frac{a+b}{2}.$$

Per calcolare la varianza abbiamo bisogno di $\mathbb{E}[X^2]$. Si ha,

$$\mathbb{E}[X^2] = \int_a^b x^2 \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\, \frac{x^3}{3}\bigg|_a^b = \frac{1}{b-a}\, \frac{b^3-a^3}{3} = \frac{a^2+ab+b^2}{3}.$$

Pertanto,

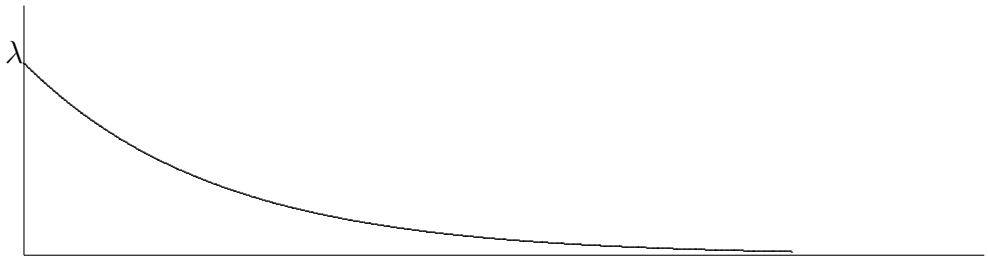
$$\mathrm{Var}[X] = \frac{a^2+ab+b^2}{3} - \Big(\frac{a+b}{2}\Big)^2 = \frac{(b-a)^2}{12}.$$

7.7 Distribuzione esponenziale

Diremo che X ha distribuzione esponenziale di parametro $\lambda > 0$, se ha densità data da

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \quad \text{per} \quad x > 0.$$

Scriveremo $X \sim \mathrm{Exp}(\lambda)$. Questa variabile aleatoria viene usata per modellare tempi aleatori. Il tempo di durata di un apparecchio elettronico, il tempo di attesa per un arrivo in una coda ecc... Se $X \sim \mathrm{Exp}(\lambda)$, la sua densità è rappresentata dal seguente grafico.



Si può dimostrare (ciò è lasciato per esercizio agli studenti più volenterosi che si ha,

$$\mathbb{E}[X] = \frac{1}{\lambda}.$$

$$\mathrm{Var}[X] = \frac{1}{\lambda^2}.$$

Capitolo 8

Il modello Normale

8.1 Distribuzione Normale o Gaussiana

La distribuzione *Normale* o *Gaussiana* è una delle più importanti distribuzioni utilizzate in statistica per diversi motivi. Essa è una distribuzione continua a valori su tutto $\mathbb{R}$. Molte variabili casuali reali (la quantità di pioggia che cade in una certa regione, misurazioni varie, ecc...) seguono una distribuzione Normale. Inoltre la distribuzione Normale serve per approssimare (vedremo in che senso) molte altre distribuzioni. La distribuzione Normale è una distribuzione continua, pertanto essa è caratterizzata dalla sua densità.

Definizione 8.1.1. Diremo che X è una variabile aleatoria Normale di parametri $\mu \in \mathbb{R}$, $\sigma^2 > 0$, e scriveremo $X \sim \mathrm{N}(\mu, \sigma^2)$, se la sua densità è

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R}.$$

I due diagrammi sottostanti mostrano le variazioni di f al variare di μ e σ^2. Si noti, in particolare che queste curve campaniformi sono simmetriche rispetto alla retta $x = \mu$.

Si può dimostrare che, se $X \sim \mathrm{N}(\mu, \sigma^2)$:

$$\mathbb{E}[X] = \mu, \tag{8.1}$$

$$\mathrm{Var}[X] = \sigma^2. \tag{8.2}$$

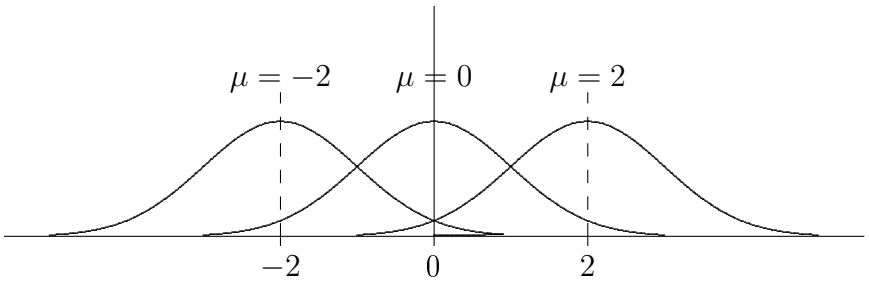


Figura 8.1 Distribuzioni normali con $\sigma = 1$ fisso, al variare di μ.

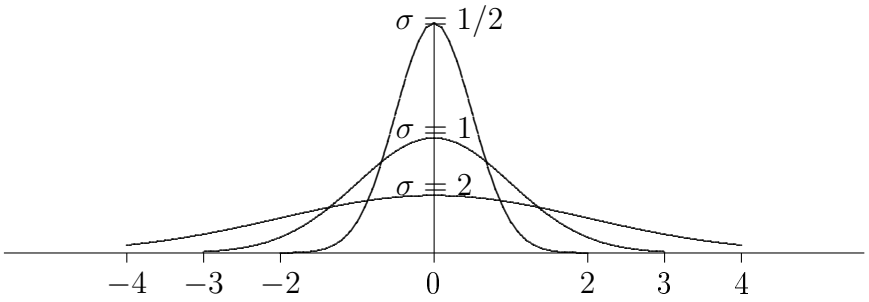


Figura 8.2 Distribuzioni normali con $\mu = 0$ fisso, al variare di σ.

Quindi i parametri della normale sono proprio la media e la varianza. C'era da aspettarselo visto che erano stati chiamati μ e σ^2! Il caso $\mu = 0$ e $\sigma^2 = 1$ è un caso speciale.

Una variabile aleatoria $X \sim \mathrm{N}(0, 1)$, si chiama *Normale standard* e come vedremo sarà per noi di grande importanza da qui in seguito. Si può osservare dalle figure sopra, che al variare di μ la campana viene soltanto traslata mantenendo la stessa forma, sempre simmetrica rispetto alla retta $x = \mu$. Al variare di σ, invece la campana si modifica, precisamente è più *concentrata* vicino alla media per valori piccoli di σ e molto più *dispersa* per valori grandi di σ, il che è ovvio se si pensa al significato del parametro σ.

Se $X \sim \mathrm{N}(\mu, \sigma^2)$, si può dimostrare che

$$X^* = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1).$$

Il processo che permette di passare da una Gaussiana qualunque ad una gaussiana standard si chiama, appunto, *standardizzazione.*

Vale anche un viceversa. Se $X^* \sim \mathrm{N}(0, 1)$, allora

$$X = \sigma X^* + \mu \sim \mathrm{N}(\mu, \sigma^2).$$

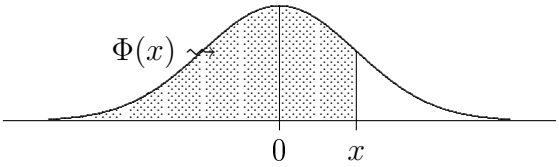
Sia $X \sim \mathrm{N}(0, 1)$. Supponiamo di voler calcolare $\mathbb{P}(1 < X < 2)$. Per quanto ne sappiamo sino ad ora questo è pari all'area sottesa dalla densità gaussiana tra $x = 1$ e $x = 2$, quindi è

$$\mathbb{P}(1 < X < 2) = \int_1^2 \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, dx.$$

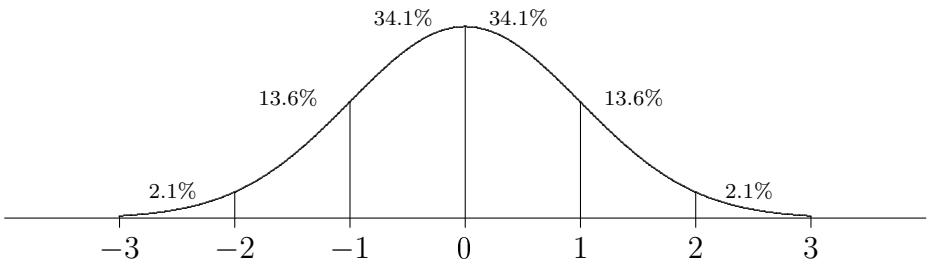
Ma quanto fa questo integrale? La risposta è che questo integrale non si può risolvere analiticamente. Pertanto ci sono delle tabelle che forniscono l'area sottesa dalla densità di una Gaussiana standard. Noi utilizzeremo la seguente tabella che fornisce la probabilità che una gaussiana standard sia minore di x, ovvero l'ara ombreggiata.

Tabella 1

Area sottesa dalla Gaussiana Standard ($\Phi(x)$ è l'area ombreggiata)



x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56750	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76731	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84850	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92786	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95819	.95907	.95994	.96080	.96160	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97933	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99745	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000



Esaminando un po' più da vicino il grafico della normale standard, possiamo vedere che per $-1 < x < 1$ si ha il 68.2% della distribuzione e per $-2 < x < 2$ ben il 95.4%, e per $-3 < x < 3$ praticamente il 100%.

La Tabella della pagina precedente fornisce l'area sottesa dalla curva normale standard minore di un x positivo. La simmetria della curva rispetto all'asse ad $x = 0$ consente di ricavare l'area compresa tra due valori qualunque di x. Se indichiamo con $\Phi(x)$ il valore dell'area sottesa sino ad x (ovvero il valore fornito dalla tabella), possiamo allora rispondere alla questione prima posta, ovvero quanto fa $\mathbb{P}(1 < X < 2)$? Semplice:

$$\mathbb{P}(1 < X < 2) = \Phi(2) - \Phi(1) = 0.97725 - 0.84134 = 0.13591.$$

Si osservi che mezza campana ha area 0.5 proprio per motivi di simmetria.

Per chiarire vediamo qualche altro esempio.

Esempio 8.1.2. Sia $X \sim \mathrm{N}(0, 1)$. Determiniamo

1. $\mathbb{P}(0 \leq X \leq 1.42)$;
2. $\mathbb{P}(-0.73 \leq X \leq 0)$;
3. $\mathbb{P}(-1.37 < X \leq 2.01)$;
4. $\mathbb{P}(-1.79 \leq X \leq -0.54)$;
5. $\mathbb{P}(X \geq 1.13)$;
6. $\mathbb{P}(X \leq 1.42)$;
7. $\mathbb{P}(|X| \leq 0.50)$;
8. $\mathbb{P}(|X| \geq 0.30)$.

Per risolvere queste questioni useremo la funzione $\Phi(x) = \mathbb{P}(X \leq x)$, dove $X \sim \mathrm{N}(0, 1)$, i cui valori possono essere trovati sulla Tabella 1.

1. $\mathbb{P}(0 \leq X \leq 1.42) = \Phi(1.42) - 0.5000 = 0.42220$.
2. $\mathbb{P}(-0.73 \leq X \leq 0) = \mathbb{P}(0 \leq X \leq 0.73) = \Phi(0.73) - 0.5000 = 0.26731$.

3.

$$\begin{aligned}\mathbb{P}(-1.37 < X \leq 2.01) &= \mathbb{P}(-1.37 < X \leq 0) + \mathbb{P}(0 < X \leq 2.01)\\ &= \mathbb{P}(0 < X \leq 1.37) + \mathbb{P}(0 < X \leq 2.01)\\ &= \Phi(1.37) + \Phi(2.01) - 1 = 0.9147 + 0.9778 - 1 = 0.8925;\end{aligned}$$

4.

$$\mathbb{P}(0.65 \leq X < 1.26) = \Phi(1.26) - \Phi(0.65) = 0.8962 - 0.7422 = 0.1540;$$

5.

$$\begin{aligned}\mathbb{P}(-1.79 \leq X \leq -0.54) &= \mathbb{P}(0.54 \leq X \leq 1.79)\\ &= 0.9633 - 0.7054 = 0.2579;\end{aligned}$$

6.

$$\begin{aligned}\mathbb{P}(X \geq 1.13) &= 1.0000 - \mathbb{P}(X \leq 1.13)\\ &= 1.0000 - \Phi(1.13) = 1.000 - 0.8708 = 0.1292;\end{aligned}$$

7.

$$\mathbb{P}(X \leq 1.42) = \Phi(1.42) = 0.9222;$$

8.

$$\begin{aligned}\mathbb{P}(|X| \leq 0.50) &= \mathbb{P}(-0.50 \leq X \leq 0.50)\\ &= 2\,\mathbb{P}(0 \leq X \leq 0.50) = 2(\Phi(0.50) - 0.5000)\\ &= 2 \cdot (0.6915 - 0.5000) = 0.3830;\end{aligned}$$

9.

$$\begin{aligned}\mathbb{P}(|X| \geq 0.30) &= \mathbb{P}(X \leq -0.30) + \mathbb{P}((X \geq 0.30)\\ &= 2\mathbb{P}(X \geq 0.30) = 2(1 - \mathbb{P}(X \leq 0.30))\\ &= 2(1.0000 - 0.6179) = 0.7643.\end{aligned}$$

Tramite la Tabella 1 è possibile calcolare la probabilità che una gaussiana qualunque sia in un fissato intervallo. Come si fa? Si procede attraverso la standardizzazione. Precisamente sia $X \sim \mathrm{N}(\mu, \sigma^2)$. Vogliamo calcolare $\mathbb{P}(a < X < b)$. Sappiamo che $X^* = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$, pertanto si ha

$$\mathbb{P}(a < X < b) = \mathbb{P}\Big(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\Big) = \mathbb{P}\Big(\frac{a-\mu}{\sigma} < X^* < \frac{b-\mu}{\sigma}\Big).$$

Quindi la probabilità che una gaussiana qualunque si trovi in un certo intervallo è uguale alla probabilità che una gaussiana standard sia in un intervallo di estremi opportunamente modificati, quindi è calcolabile come fatto precedentemente.

Esempio 8.1.3.

Supponiamo che la temperatura T nel mese di giugno sia distribuita normalmente con media $\mu = 20^oC$ e scarto quadratico medio $\sigma = 5^oC$. Vogliamo determinare la probabilità che la temperatura sia

1. minore di 15^oC;

2. maggiore di 30^oC.

Se T è la temperatura nel mese di giugno, allora $T \sim \mathrm{N}(\mu, \sigma^2)$, dove $\mu = 20^0C$ e $\sigma = 5^0C$. Quindi $T^* = \frac{T-20}{5} \sim \mathrm{N}(0,1)$.

1. $\mathbb{P}(T < 15) = \mathbb{P}\Big(\dfrac{T-20}{5} < \dfrac{15-20}{5}\Big) = \mathbb{P}(T^* < -1) = 1.000 - \Phi(1) = 1.000 - 0.8413 = 0.1587$

2. $\mathbb{P}(T > 30) = \mathbb{P}\Big(\dfrac{T-20}{5} > \dfrac{30-20}{5}\Big) = \mathbb{P}(T^* > 2) = 1.000 - \Phi(2) = 1.000 - 0.9772 = 0.0228$

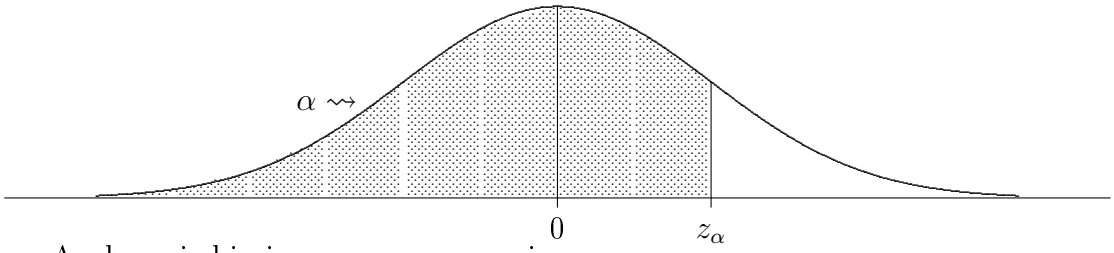
Veniamo ora ad un altro problema importante in statistica. Determinare i quantili di una distribuzione normale. Dalla tabella si può anche risolvere il problema inverso. Ovvero dato il valore della probabilità (dell'area $\Phi(x)$) trovare il valore di x. Per esempio, sia $X \sim \mathrm{N}(0,1)$ se $\mathbb{P}(X \leq x) = 0.85083$, quanto vale x? Si cerca 0.85083 nella tabella e si va a vedere quale x corrisponde. Si trova: $x = 1.04$. E se abbiamo un valore della probabilità che non c'è nella tabella? Per esempio, se $X \sim \mathrm{N}(0,1)$ e $\mathbb{P}(X \leq x) = 0.97$ quanto vale x? Se andiamo nella tabella 0.97 non si trova! Allora si cerca il valore più vicino a 0.97 è 0.96995. Pertanto $x = 1.88$.

In generale, chiameremo z_α il quantile di ordine α di una una distribuzione normale standard, ovvero quel valore sull'asse delle x tale che

$$\mathbb{P}(X \leq z_\alpha) = \mathbb{P}(X < z_\alpha) = \alpha, \quad X \sim \mathrm{N}(0,1), \quad \alpha \in (0,1). \tag{8.3}$$

Riportiamo qui, dato che li useremo spesso, alcuni quantili "famosi" della gaussiana standard.

$\alpha = 0.95$ allora $z_\alpha = z_{.95} = 1.64$;
$\alpha = 0.975$ allora $z_\alpha = z_{.975} = 1.96$;
$\alpha = 0.99$ allora $z_\alpha = z_{.99} = 2.29$;
$\alpha = 0.995$ allora $z_\alpha = z_{.995} = 2.58$.



Anche qui chiariamo con un esempio.

Esempio 8.1.4. Sia $X \sim \mathrm{N}(0,1)$. Si determini $t \in \mathbb{R}$ tale che:

1. $\mathbb{P}(0 \leq X \leq t) = 0.4332$;

2. $\mathbb{P}(t \leq X \leq 0) = 0.3461$;

3. $\mathbb{P}(t < X \leq 2.01) = 0.0440$;

4. $\mathbb{P}(X \geq t) = 0.5120$;

5. $\mathbb{P}(X \leq t) = 0.6700$;

6. $\mathbb{P}(|X| \leq t) = 0.9750$;

7. $\mathbb{P}(|X| \geq t) = 0.1836$.

Anche per risolvere questo esercizio useremo la funzione $\Phi(t) = \mathbb{P}(X \leq t)$, dove $X \sim \mathrm{N}(0,1)$, i cui valori possono essere trovati sulla Tabella 1.

1. $\mathbb{P}(0 \leq X \leq t) = \Phi(t) - 0.5000 = 0.4332$, da cui $\Phi(t) = 0.9332$ e quindi $t = 1.50$;

2. Osserviamo che affinché la probabilità $\mathbb{P}(t \leq X \leq 0)$ abbia senso deve essere $t < 0$. In tal caso si ha,
$$\mathbb{P}(t \leq X \leq 0) = \mathbb{P}(0 \leq X \leq -t) = \Phi(-t) - 0.5000 = 0.3461,$$
da cui $\Phi(-t) = 0.8461$, $-t = 1.02$ e quindi $t = -1.02$;

3. Osserviamo che occorre distinguere le due possibilità $t < 0$ e $t > 0$. Se fosse $t < 0$ allora si avrebbe
$$\begin{aligned} \mathbb{P}(t < X \leq 2.01) &= \mathbb{P}(t < X \leq 0) + \mathbb{P}(0 < X \leq 2.01) \\ &= \Phi(-t) + \Phi(2.01) - 1 = \\ &= \Phi(-t) + 0.9778 - 1 = 0.0440, \end{aligned}$$
da cui $\Phi(-t) = -0.4778 + 0.0440 < 0$. Il che è impossibile perchè Φ è una funzione sempre positiva. Dunque $t > 0$ ed in tal caso si ha,
$$\begin{aligned} &\mathbb{P}(t < X \leq 2.01) \\ &\qquad = \Phi(2.01) - \Phi(t) = 0.9778 - \Phi(t) = 0.0440, \end{aligned}$$
da cui $\Phi(t) = 0.9778 - 0.0440 = 0.9338$ e quindi $t = 1.50$.

4. Essendo la probabilità richiesta > 0.5, segue che $t < 0$. Si ha,
$$\mathbb{P}(X \geq t) = \Phi(-t) = 0.5120,$$
da cui $-t = 0.03$ e $t = -0.03$;

5. Essendo la probabilità richiesta > 0.5, segue che $t > 0$. Si ha,
$$\mathbb{P}(X \leq t) = \Phi(t) = 0.6700,$$
da cui $t = 0.44$;

6. $$\mathbb{P}(|X| \leq t) = \mathbb{P}(-t \leq X \leq t) = 2\mathbb{P}(0 \leq X \leq t) = 2(\Phi(t) - 0.5000) = 0.9750,$$
da cui $\Phi(t) = 0.98750$ e $t = 2.24$;

7. $$\mathbb{P}(|X| \geq t) = \mathbb{P}(X \leq -t) + \mathbb{P}(X \geq t) = 2\mathbb{P}(X \geq t) = 2(1 - \Phi(t)) = 0.1836,$$
da cui $\Phi(t) = 0.9082$ e $t = 1.33$.

8.2 Il Teorema Limite Centrale

In questa sezione affrontiamo uno dei risultati più importanti del calcolo delle probabilità e parte delle sue applicazioni in statistica: il *Teorema Limite Centrale.*
In termini semplicistici esso afferma che la somma di un *gran* numero di variabili aleatorie tutte con la stessa distribuzione *tende* ad avere una distribuzione normale. L'importanza di ciò sta nel fatto che siamo in grado di ottenere stime della probabilità che riguardano la somma di variabili aleatorie indipendenti ed identicamente distribuite (i.i.d.), a prescindere da quale sia la distribuzione di ciascuna.

Precisamente siamo in presenza di variabili aleatorie $X_1, X_2, \ldots, X_n$ indipendenti e con stessa legge, quindi con stessa media μ e stessa varianza σ^2. Posto $S_n = X_1 + X_2 + \ldots X_n$, S_n si comporta quasi come una variabile normale. Di che parametri? Per sapere i parametri di una legge normale occorre sapere la sua media e la sua varianza. Ricordando le proprietà della media e della varianza si ha:

$$\mathbb{E}[S_n] = \mathbb{E}[X_1 + X_2 + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_n] = n\mu, \tag{8.4}$$

$$\mathrm{Var}[S_n] = \mathrm{Var}[X_1 + X_2 + X_n] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_n] = n\sigma^2. \tag{8.5}$$

Riassumendo S_n ha circa una distribuzione $\mathrm{N}(n\mu, n\sigma^2)$. Scriveremo $S_n \simeq X \sim \mathrm{N}(n\mu, n\sigma^2)$, dove $\simeq$ vuol dire che è approssimativamente uguale a.

L'enunciato presentato nella sua forma più generale è il seguente.

Teorema 8.2.1. *Siano $X_1, X_2, \ldots,$ variabili aleatorie indipendenti ed identicamente distribuite tutte con media μ e varianza σ^2. Per n grande,*

$$S_n \simeq Z \sim \mathrm{N}(n\mu, n\sigma^2), \tag{8.6}$$

o anche standardizzando,

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \simeq Z^* \sim \mathrm{N}(0, 1). \tag{8.7}$$

Che vuol dire tutto ciò? Chiariamo con un esempio.

Esempio 8.2.2. Una compagnia americana di assicurazioni ha 10000 (10^4) polizze attive. Immaginando che il risarcimento annuale medio per ogni assicurato si possa modellare con una variabile aleatoria di media 260\$ e scarto quadratico medio di 800\$ vogliamo calcolare la probabilità (approssimata) che il risarcimento annuale superi i 2.8 milioni di dollari.

Numeriamo gli assicurati in modo che X_i sia il risarcimento annuale richiesto dall'i-esimo assicurato, dove $i = 1, 2, \ldots, 10^4$. Allora il risarcimento annuale totale è $S_{10^4} = \sum_{i=1}^{10^4} X_i$. Seguendo i dati $\mu = \mathbb{E}[X_i] = 260\$$ e $\sigma = \sqrt{\mathrm{Var}[X_i]} = 800\$$. Inoltre possiamo supporre che le variabili X_i siano indipendenti, dato che ciascun assicurato chiede il risarcimento in modo indipendente dagli altri. Per il TLC la variabile aleatoria X ha una distribuzione approssimativamente normale $S_{10^4} \simeq Z \sim \mathrm{N}(n\mu, n\sigma^2)$ con $n\mu = 10^4 \cdot 260\$ = 2.6 \cdot 10^6\$$ e $n\sigma^2 = 10^4 \cdot 800^2\$^2 = 64 \cdot 10^8\$^2$, quindi

$$\mathbb{P}(S_{10^4} > 2.8 \cdot 10^6) \simeq \mathbb{P}(Z > 2.8 \cdot 10^6) =$$
$$\mathbb{P}\Big(\frac{Z - 2.6 \cdot 10^6}{\sqrt{64 \cdot 10^8}} > \frac{2.8 \cdot 10^6 - 2.6 \cdot 10^6}{\sqrt{64 \cdot 10^8}}\Big) = \mathbb{P}(Z^* > 2.5) = 0.0062$$

Osservazione 8.2.3. È importante osservare che nel TLC parliamo di n *grande.* Ma che vuol dire n *grande*? Quante variabili aleatorie i.i.d. dobbiamo sommare per avere una buona approssimazione? Purtroppo non esiste una regola generale. Il valore di n dipende ogni volta dalle distribuzioni di partenza. In ogni caso si tende ad applicarlo quando $n > 30$.

È altresì molto importante osservare che se le variabili di partenza $X_1, X_2, \ldots, X_n$ sono esse stesse Gaussiane il TLC è **esatto.** Ovvero per ogni n,

$$S_n = Z \sim \mathrm{N}(n\mu, n\sigma^2), \tag{8.8}$$

o anche standardizzando,

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = Z^* \sim \mathrm{N}(0,1). \tag{8.9}$$

8.3 Applicazioni del TLC

Vediamo alcune importanti applicazioni del TLC che utilizzeremo spesso nel seguito.

8.3.1 Approssimazione della binomiale

Abbiamo visto nella sezione riguardante la distribuzione binomiale che questa può essere pensata come somma di variabili aleatorie bernoulliane. Precisamente se $X \sim \mathrm{Bi}(n,p)$ allora

$$X = Y_1 + Y_2 + \ldots Y_n,$$

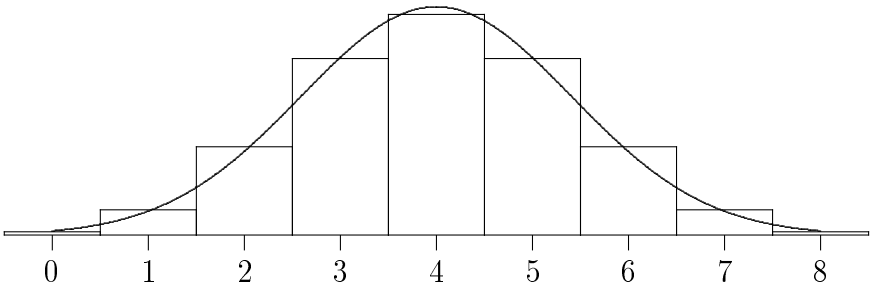
con le $Y_i \sim \mathrm{Be}(p)$ indipendenti. Allora se n è *grande,* possiamo applicare il TLC ed ottenere che

$$X \simeq Z \sim \mathrm{N}(np, np(1-p)), \tag{8.10}$$

o ancora,

$$\frac{X - np}{\sqrt{np(1-p)}} \simeq Z^* \sim \mathrm{N}(0,1). \tag{8.11}$$

Questa proprietà risulta forse più chiara se si guarda alla figura che segue che mostra il raffronto tra una variabile $X \sim \mathrm{Bi}(8, 1/2)$ ed una $Z \sim \mathrm{N}(4,2)$.



Osservazione 8.3.1. La densità binomiale è simmetrica per $p = 0.5$ ed è tanto più asimmetrica quanto più p è lontano da 0.5 (quindi vicino a 0 oppure a 1). È buona norma quella di applicare l'approssimazione normale della binomiale $\mathrm{Bi}(n,p)$, solo se

$$np > 5, \quad n(1-p) > 5.$$

L'idea del TLC è quella che le aree sottese dalla distribuzione di partenza (in questo caso le aree sottese dall'istogramma) vengono approssimate con le aree sottese dalla curva gaussiana. L'approssimazione si può render più precisa tramite l'introduzione della *correzione di continuità.* Per capire di cosa si tratta vediamo un esempio.

Esempio 8.3.2.

Un dado equilibrato viene lanciato 900 volte. Sia X il numero di volte che esce il sei. Utilizzando l'approssimazione normale della binomiale calcolare:

1. $\mathbb{P}(X > 180)$;
2. $\mathbb{P}(X \geq 160)$.

Se X è il numero di volte che esce il 6 in 900 lanci di un dado, $X \sim \text{Bi}(n, p)$, con $n = 900$ e $p = 1/6$. $\mathbb{E}[X] = 150$, $\text{Var}[X] = 125$. Allora per il TLC si ha

$$X \simeq Z \sim \text{N}(150, 125),$$

oppure

$$\frac{X - 150}{\sqrt{125}} \simeq Z^* \sim \text{N}(0, 1).$$

1.

$$\begin{aligned} \mathbb{P}(X > 180) &\simeq P(Z > 180) \\ &= \mathbb{P}\Big(\frac{Z - 150}{11.18} > \frac{180 - 150}{11.18}\Big) = \mathbb{P}(Z^* > 2.68) \\ &= 1 - \Phi(2.68) = 0.0368. \end{aligned}$$

Se avessimo usato la correzione di continuità, indicando con Z una gaussiana con stessa media e stessa varianza di X, avremmo ottenuto,

$$\begin{aligned} \mathbb{P}(X > 180) &\simeq \mathbb{P}(Z > 180.5) = \mathbb{P}\Big(\frac{Z - 150}{11.18} > \frac{180.5 - 150}{11.18}\Big) \\ &= \mathbb{P}(Z^* > 2.73) = 1 - \Phi(2.73) = 0.0317. \end{aligned}$$

2.

$$\begin{aligned} \mathbb{P}(X \geq 160) &\simeq \mathbb{P}(Z \geq 160) = \mathbb{P}\Big(\frac{Z - 150}{11.18} \geq \frac{160 - 150}{11.18}\Big) = \mathbb{P}(Z^* \geq 0.89) \\ &= 1 - \Phi(0.89) = 0.18673 \end{aligned}$$

Usando la correzione di continuità, indicando con Z una gaussiana con stessa media e stessa varianza di X, avremmo ottenuto,

$$\begin{aligned} \mathbb{P}(X \geq 160) &\simeq \mathbb{P}(Z \geq 159.5) = \mathbb{P}\Big(\frac{Z - 150}{11.18} \geq \frac{159.5 - 150}{11.18}\Big) \\ &= \mathbb{P}(Z^* \geq 0.85) = 1 - \Phi(0.85) = 0.1977. \end{aligned}$$

ATTENZIONE! La correzione di continuità si usa per avere una approssimazione migliore quando si passa da una distribuzione discreta ad una continua! Mai negli altri casi!

8.3.2 Approssimazione della media campionaria

Un altro caso importante in cui si può applicare il TLC è quello della *media campionaria.* Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie i.i.d. con media μ e varianza σ^2. Si chiama media campionaria e si indica con $\overline{X}_n$ la quantità

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots X_n) = \frac{S_n}{n}. \tag{8.12}$$

Anche qui siamo in presenza di una somma (a parte il fattore $1/n$) di variabili aleatorie i.i.d. ed anche qui possiamo applicare il TLC. Vediamo come diventa in questo caso specifico. Sappiamo che $\overline{X}_n$ per n *grande* si comporta come una gaussiana. Di che parametri? Occorre pertanto calcolare media e varianza della variabile aleatoria $\overline{X}_n$. Ricordando le proprietà della media e della varianza, si ha

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\Big[\frac{1}{n}(X_1 + X_2 + \ldots X_n)\Big] = \frac{1}{n}(\mu + \mu + \ldots \mu) = \mu,$$

ed inoltre

$$\mathrm{Var}[\overline{X}_n] = \mathrm{Var}\Big[\frac{1}{n}(X_1 + X_2 + \ldots X_n)\Big] = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots \sigma^2) = \frac{1}{n}\sigma^2.$$

Quindi

$$\overline{X}_n \simeq Z \sim \mathrm{N}(\mu, \sigma^2/n), \tag{8.13}$$

o ancora,

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \simeq Z^* \sim \mathrm{N}(0,1). \tag{8.14}$$

Si osservi che se le variabili $X_1, X_2, \ldots, X_n$ sono gaussiane allora anche in questo caso il TLC è **esatto,** ovvero

$$\overline{X}_n = Z \sim \mathrm{N}(\mu, \sigma^2/n), \tag{8.15}$$

o ancora,

$$\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} = Z^* \sim \mathrm{N}(0,1). \tag{8.16}$$

8.4 Alcune distribuzioni legate alla normale

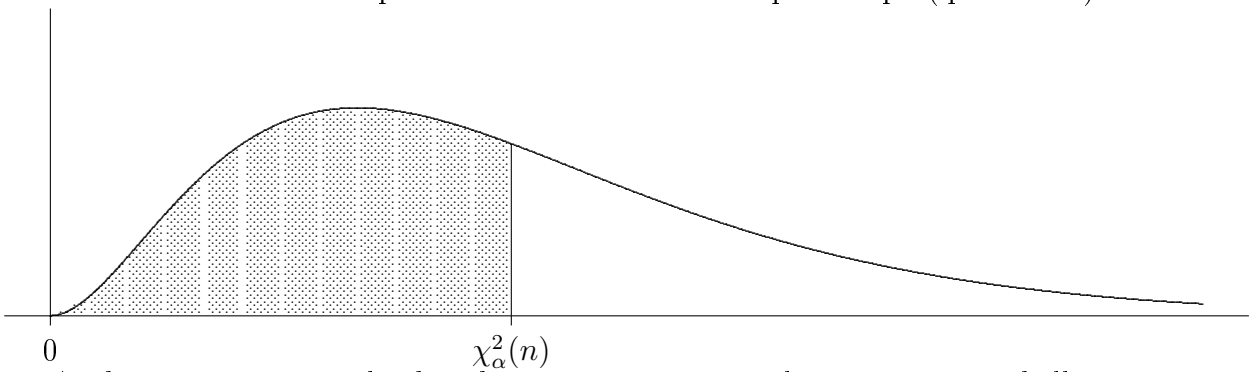
8.4.1 La distribuzione χ^2 (chi quadro)

Definizione 8.4.1. Diremo che la variabile aleatoria X ha distribuzione χ^2 con n gradi di libertà e scriveremo $X \sim \chi^2(n)$ se X ha densità data da

$$f(x) = c_n x^{n/2-1} \mathrm{e}^{-x/2}, \quad \text{per} \quad x > 0,$$

dove c_n è una costante opportuna.

Vediamo l'andamento tipico di una distribuzione di questo tipo (qui $n = 15$):



Anche qui, come per la distribuzione gaussiana utilizzeremo una tabella opportuna. Chiameremo $\chi^2_\alpha(n)$ quel valore sull'asse x tale che:

$$\mathbb{P}(X < \chi^2_\alpha(n)) = \alpha, \quad se \quad X \sim \chi^2(n), \quad \alpha \in (0,1). \tag{8.17}$$

Per maggiore chiarezza si veda la figura precedente.

Si può dimostrare che se $X \sim \chi^2(n)$, allora

$$\mathbb{E}[X] = n, \qquad \mathrm{Var}[X] = 2n.$$

Inoltre se n è *grande*

$$X \simeq Z \sim \mathrm{N}(n, 2n). \tag{8.18}$$

I valori dei quantili di un $\chi^2(n)$ sono tabulati per i primi valori di n e per qualche valore tipico di α. Per n grande, grazie alla (8.18) possiamo usare la seguente approssimazione per i quantili delle distribuzioni χ^2,

$$\chi^2_\alpha(n) \simeq z_\alpha \sqrt{2n} + n.$$

Infatti, proprio grazie alla (8.18), si ha

$$\alpha = \mathbb{P}(X \le \chi^2_\alpha(n)) \simeq \mathbb{P}(Z \le \chi^2_\alpha(n)) = \mathbb{P}\Big(\frac{Z-n}{\sqrt{2n}} \le \frac{\chi^2_\alpha(n)-n}{\sqrt{2n}}\Big) = \mathbb{P}\Big(Z^* \le \frac{\chi^2_\alpha(n)-n}{\sqrt{2n}}\Big),$$

quindi ricordando la definizione di quantile di una gaussiana standard,

$$z_\alpha \simeq \frac{\chi^2_\alpha(n)-n}{\sqrt{2n}},$$

e quindi

$$\chi^2_\alpha(n) \simeq z_\alpha \sqrt{2n} + n.$$

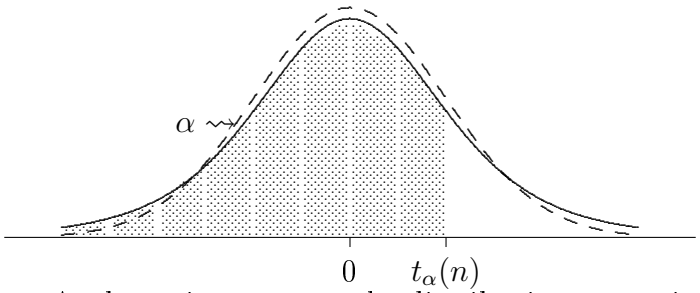
8.4.2 La distribuzione di Student

Definizione 8.4.2. Diremo che la variabile aleatoria X ha distribuzione di Student con n gradi di libertà e scriveremo $X \sim t(n)$ se X ha densità data da

$$f(x) = c_n \Big(1 + \frac{x^2}{n}\Big)^{-\frac{n+1}{2}}, \quad \text{per} \quad x \in \mathbb{R},$$

dove c_n è una costante opportuna.

Vediamo l'andamento tipico di una distribuzione di questo tipo (qui $n = 5$). Nella figura, *tratteggiata,* è riportata anche la densità di una gaussiana standard.



Anche qui, come per la distribuzione gaussiana utilizzeremo una tabella opportuna. Chiameremo $t_\alpha(n)$ quel valore sull'asse

$$\mathbb{P}(X < t_\alpha(n)) = \alpha, \quad se \quad X \sim t(n), \quad \alpha \in (0,1). \tag{8.19}$$

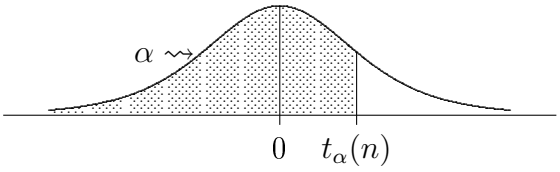
Per maggiore chiarezza si veda la Figura precedente.

I valori dei quantili di una distribuzione $t(n)$ sono tabulati per i primi valori di n e per qualche valore tipico di α. Si fa presente (ma non possiamo dimostrarlo!) che per n grande si ha che la distribuzione di una variabile aleatoria con distribuzione $t(n)$ può essere approssimata con una $\mathrm{N}(0,1)$, quindi per n grande possiamo usare la seguente approssimazione per i quantili delle distribuzioni $t(n)$,

$$t_\alpha(n) \simeq z_\alpha.$$

Tabella 2

Tabella dei quantili $t_\alpha(n)$ della legge di Student $t(n)$

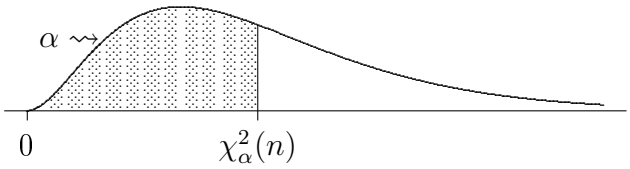


$n \setminus \alpha$	.900	.950	.975	.990	.995
1	3.07768	6.31375	12.70615	31.82096	63.65590
2	1.88562	2.91999	4.30266	6.96455	9.92499
3	1.63775	2.35336	3.18245	4.54071	5.48085
4	1.53321	2.13185	2.77645	3.74694	4.60408
5	1.47588	2.01505	2.57058	3.36493	4.03212
6	1.43976	1.94318	2.44691	3.14267	3.70743
7	1.41492	1.89458	2.36462	2.99795	3.49948
8	1.39682	1.85955	2.30601	2.89647	3.35538
9	1.38303	1.83311	2.26216	2.82143	3.24984
10	1.37218	1.81246	2.22814	2.76377	3.16926
11	1.36343	1.79588	2.20099	2.71808	3.10582
12	1.35622	1.78229	2.17881	2.68099	3.05454
13	1.35017	1.77093	2.16037	2.65030	3.01228
14	1.34503	1.76131	2.14479	2.62449	2.97685
15	1.34061	1.75305	2.13145	2.60248	2.94673
16	1.33676	1.74588	2.11990	2.58349	2.92070
17	1.33338	1.73961	2.10982	2.56694	2.89823
18	1.33039	1.73406	2.10092	2.55238	2.87844
19	1.32773	1.7213	2.09302	2.53948	2.86094
20	1.32534	1.72472	2.08596	2.52798	2.84534
21	1.32319	1.72074	2.07961	2.51765	2.83137
22	1.32124	1.71714	2.07388	2.50832	2.81876
23	1.31946	1.71387	2.06865	2.49987	2.80734
24	1.31784	1.71088	2.06390	2.49216	2.79695
25	1.31635	1.70814	2.05954	2.48510	2.78744
26	1.31497	1.70562	2.05553	2.47863	2.77872
27	1.31370	1.70329	2.05183	2.47266	2.77068
28	1.31253	1.70113	2.04841	2.46714	2.76326
29	1.31143	1.69913	2.04523	2.46202	2.75639
30	1.31042	1.69726	2.04227	2.45726	2.74998
40	1.30308	1.68385	2.02107	2.42326	2.70446
50	1.29871	1.67591	2.00856	2.40327	2.67779
60	1.29582	1.67065	2.00030	2.39012	2.66027
70	1.29376	1.66692	1.99444	2.38080	2.64790
80	1.29222	1.66413	1.99007	2.37387	2.63870
90	1.29103	1.66196	1.98667	2.36850	2.63157
100	1.29008	1.66023	1.98397	2.36421	2.62589
110	1.28930	1.65882	1.98177	2.36072	2.62127
120	1.28865	1.65765	1.97993	2.35783	2.61742

Per $n > 120$ si può utilizzare l'approssimazione $t_\alpha(n) \simeq z_\alpha$

Tabella 3

Tabella dei quantili $\chi^2_\alpha(n)$ della legge chi-quadro $\chi^2(n)$



$n \setminus \alpha$	.010	.025	.050	.950	.975	.990
1	0.00016	0.00098	0.00393	3.84146	5.02390	6.63489
2	0.02010	0.05064	0.10259	5.99148	7.37778	9.21035
3	0.11483	0.21579	0.35185	7.81472	9.34840	11.34488
4	0.29711	0.48442	0.71072	9.48773	11.14326	13.27670
5	0.55430	0.83121	1.14548	11.07048	12.83249	15.08632
6	0.87208	1.23734	1.63538	12.59158	14.44935	16.81187
7	1.23903	1.68986	2.16735	14.06713	16.01277	18.47532
8	1.64651	2.17972	2.73263	15.50731	17.53454	20.09016
9	2.08789	2.70039	3.32512	16.91896	19.02278	21.66605
10	2.55820	3.24696	3.94030	18.30703	20.48320	23.20929
11	3.05350	3.81574	4.57481	19.67515	21.92002	24.72502
12	3.57055	4.40378	5.22603	21.02606	23.33666	26.21696
13	4.10690	5.00874	5.89186	22.36203	24.73558	27.68818
14	4.66042	5.62872	6.57063	23.68478	26.11893	29.14116
15	5.22936	6.26212	7.26093	24.99580	27.48836	30.57795
16	5.81220	6.90766	7.96164	26.29622	28.84532	31.99986
17	6.40774	7.56418	8.67175	27.58710	30.19098	33.40872
18	7.01490	8.23074	9.23045	28.86932	31.52641	34.80524
19	7.63270	8.90651	10.11701	30.14351	32.85234	36.19077
20	8.26037	9.59077	10.85080	31.41042	34.16958	37.56627
21	8.89717	10.28291	11.59132	32.67056	35.47886	38.93223
22	9.54249	10.98233	12.33801	33.92446	36.78068	40.28945
23	10.19569	11.68853	13.09051	35.17246	38.07561	41.63833
24	10.85635	12.40115	13.84842	36.41503	39.36406	42.97978
25	11.52395	13.11971	14.61140	3765249	40.64650	44.31401
26	12.19818	13.84388	15.37916	38.88513	41.92314	45.64164
27	12.87847	14.57337	16.15139	40.11327	43.19452	46.96284
28	13.56467	15.30785	16.92788	41.33715	44.46.79	48.27817
29	14.25641	16.04075	17.70838	42.55695	45.72228	49.58783
30	14.95346	16.79076	18.49267	43.77295	46.97992	50.89218

Per $n \geq 30$ si può utilizzare l'approssimazione
$\chi^2_\alpha(n) \simeq z_\alpha\sqrt{2n} + n$

Capitolo 9

Stima dei parametri

9.1 Modelli statistici

In questo paragrafo introdurremo le prime nozioni di *statistica inferenziale.* Cominciamo con un esempio.

Esempio 9.1.1. Consideriamo il seguente problema. Una macchina produce in serie componenti meccanici di dimensioni specificate. Naturalmente, la macchina sarà soggetta a piccole imprecisioni casuali, che faranno oscillare le dimensioni reali dei pezzi prodotti. Ciò che conta è che essi si mantengano entro dei prefissati limiti di tolleranza. Al di fuori di questi limiti il pezzo è inutilizzabile. Si pone dunque un problema di controllo di qualità. Il produttore, ad esempio, deve essere in grado di garantire a chi compra, che solo lo 0.5% dei pezzi sia difettoso. Per fare ciò, occorre anzitutto stimare qual è attualmente la frazione di pezzi difettosi prodotti, per intervenire sulla macchina, qualora questa frazione non rientrasse nei limiti desiderati.

Modelliamo la situazione. Supponiamo che ogni pezzo abbia probabilità p (piccola) di essere difettoso e che il fatto che un pezzo sia difettoso non renda né più né meno probabile il fatto che un altro pezzo sia difettoso. Sotto queste ipotesi (ovvero nei limiti in cui queste considerazioni sono ragionevoli) il fenomeno può essere modellato da una successione di variabili aleatorie di Bernoulli di parametro p. $X_i \sim \mathrm{Be}(p)$, vale 1 se il pezzo i-esimo è difettoso, 0 altrimenti. Il punto fondamentale è che il parametro p è la quantità da *stimare.*

Abbiamo una popolazione, quella dei pezzi via via prodotti; questa popolazione è distribuita secondo una legge di tipo noto (Bernoulli), ma contenente un parametro incognito.

Per stimare il valore vero del parametro p, il modo naturale di procedere è estrarre un *campione casuale* dalla popolazione: scegliamo a caso n pezzi prodotti, e guardiamo se sono difettosi. L'esito di questa ispezione è descritto da una n-upla di variabili aleatorie indipendenti $X_1, X_2, \ldots, X_n$ con distribuzione $\mathrm{Be}(p)$.

Fissiamo ora le idee emerse da questo esempio in qualche definizione di carattere generale.

Definizione 9.1.2. *Un modello statistico* è una famiglia di leggi di variabili aleatorie dipendenti da uno o più parametri incogniti.

Un *campione casuale* di ampiezza n è una n-upla di variabili aleatorie i.i.d. $X_1, X_2, \ldots, X_n$ estratte dalla famiglia.

Riassumendo uno dei problemi fondamentali della statistica inferenziale è quello di scoprire la vera distribuzione della popolazione, a partire dalle informazioni contenute in un campione casuale estratto da essa. Spesso la natura del problema (come quello appena visto) ci consente di formulare un modello statistico, per cui la distribuzione della popolazione non è completamente incognita ma piuttosto è di tipo noto (bernoulliana nel problema precedente) ma con parametri incogniti. Quindi il problema è ricondotto a stimare il valore vero del parametro a partire dal campione casuale. Questa operazione prende il nome di **stima dei parametri.**

9.2 Stima puntuale

9.2.1 Stimatori e stima puntuale della media

Torniamo all'Esempio 9.1.1 del controllo qualità. Il nostro obiettivo è stimare il parametro p della popolazione bernoulliana da cui estraiamo un campione casuale $X_1, X_2, \ldots, X_n$. Si ricordi che essendo $X_i \sim \mathrm{Be}(p)$, $\mathbb{E}[X_i] = p$, quindi stimare il valore di p vuol dire stimare la *media* della popolazione. Siano ora $x_1, x_2, \ldots, x_n$ i *valori osservati* sul campione casuale. Si noti la differenza: *prima* di eseguire il campionamento, il campione casuale è una n-upla di variabili aleatorie $(X_1, X_2, \ldots, X_n)$; *dopo* aver eseguito il campionamento, cioè l'estrazione degli n individui, le n variabili aleatorie assumono valori numerici $(x_1, x_2, \ldots, x_n)$. Per cogliere bene la differenza tra questi oggetti, cosa che sarà fondamentale nel seguito, si pensi al seguente esempio. Gli stessi 10 biglietti della lotteria sono oggetti ben diversi *prima* e *dopo* l'estrazione dei numeri vincenti: *prima* sono "possibilità di vincere una cifra variabile", *dopo* sono una somma di denaro (certa) oppure carta straccia. Tenendo presente l'analogia tra le media di n numeri e valore atteso di una variabile aleatoria, è naturale scegliere $\overline{x}_n$ come *stima* di p, ovvero

$$\hat{p} = \overline{x}_n,$$

per indicare che il valore stimato di p, $\hat{p}$ è il valore della media campionaria. Naturalmente potremmo essere stati sfortunati ed avere selezionato per caso, un campione di pezzi su cui la media campionaria $\overline{x}_n$ è lontana dal valore vero del parametro p.

Per ora, a conforto della scelta fatta, possiamo osservare che se consideriamo la *variabile aleatoria* "media campionaria"

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

sappiamo che

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}[X_i] = p.$$

Ovvero il valore atteso della variabile aleatoria $\overline{X}_n$ è il vero valore del parametro p. Questo è il motivo, o meglio un motivo, perché scegliamo il valore $\overline{x}_n$, calcolato dopo il campionamento come stima del parametro p.

Per quanto la sfortuna possa giocare a nostro sfavore, è ragionevole aspettarsi che il valore $\overline{x}_n$ risenta tanto meno delle oscillazioni casuali quanto più è grande n. Questo fatto può essere dimostrato in modo rigoroso, noi ci accontenteremo dell'intuizione per dire che,

in qualche senso, si ha

$$\overline{X}_n \overset{n\to+\infty}{\rightsquigarrow} \mathbb{E}[X_i],$$

nel nostro caso

$$\overline{X}_n \overset{n\to+\infty}{\rightsquigarrow} p.$$

Si ricordi che $\mathrm{Var}[\overline{X}_n] = \sigma^2/n$ tende a zero per $n \to +\infty$, quindi la media campionaria *tende* a diventare costante (quindi uguale alla media).

Fissiamo ora qualche altra definizione generale.

Definizione 9.2.1. Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale di ampiezza n estratto da una popolazione con legge dipendente da un parametro $\vartheta \in \mathbb{R}$ incognito.

Si chiama **statistica** una variabile aleatoria che sia funzione *solo* del campione, ovvero sia calcolabile esattamente dopo il campionamento,

$$T = f(X_1, X_2, \ldots, X_n).$$

Si chiama **stimatore** di ϑ una statistica T

$$T = f(X_1, X_2, \ldots, X_n),$$

usata per stimare il valore del parametro ϑ.

Si chiama **stima** di ϑ il valore numerico, $\hat{\vartheta}$

$$\hat{\vartheta} = f(x_1, x_2, \ldots, x_n),$$

calcolato a campionamento eseguito.

Definizione 9.2.2. Uno stimatore si dice **non distorto** o **corretto,** se

$$\mathbb{E}[T] = \mathbb{E}[f(X_1, X_2, \ldots, X_n)] = \vartheta,$$

ovvero se la sua media coincide con il parametro da stimare.

Definizione 9.2.3. Uno stimatore si dice **consistente** se *tende* al parametro da stimare quando l'ampiezza del campione $n \to +\infty$.

Esempio 9.2.4. Riprendendo l'esempio del controllo qualità, la variabile aleatoria $\overline{X}_n$ è uno stimatore non distorto e consistente di p. Il valore numerico, calcolato su un particolare campione è una stima di p.

Ricapitoliamo: per stimare il "valore vero " del parametro incognito di una distribuzione a partire da un campione casuale, si costruisce un opportuno stimatore T, ossia una variabile aleatoria funzione del campione casuale. A campionamento eseguito, il valore di T viene preso come stima del valore del parametro incognito. Criteri (non gli unici) per valutare la bontà di uno stimatore sono la correttezza e la consistenza. Questo metodo di stima prende il nome di **stima puntuale.**

Vediamo alcune considerazioni importanti.

Osservazione 9.2.5. Cambiando campione lo stimatore rimane lo stesso, la stima no!!!

È chiaro che stimando il parametro con il valore dello stimatore si commette un errore e che questo, se lo stimatore è corretto e consistente, diminuisce all'aumentare dell'ampiezza del campione

L'esempio visto in precedenza rientra in un caso più generale in cui il parametro incognito da stimare è la media della popolazione. In questo caso parliamo di*stima puntuale della media.* Lo stimatore *naturale* in questo caso è sempre la media empirica, che risulta essere un *buon* stimatore dato che è non distorto e consistente.

9.2.2 Stima puntuale della varianza

In questo caso supponiamo che il parametro incognito della distribuzione da stimare sia la varianza. Vogliamo trovare un *buon* stimatore della varianza. Partiamo da un caso concreto (che è poi quello per noi di maggiore interesse). Supponiamo di voler stimare la varianza σ^2 di una popolazione $\mathrm{N}(\mu, \sigma^2)$ in base a delle osservazioni $X_1, X_2, \ldots, X_n$ estratte da questa popolazione. In statistica descrittiva abbiamo introdotto la varianza campionaria di un dato X, che supponendo tutte le frequenze uguali a 1, è data da,

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Perciò sembra *naturale* stimare σ^2 con

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Si osservi che, quando si ha a che fare con una legge che dipende da due parametri ϑ_1, ϑ_2, (come la normale, appunto), la stima di uno , diciamo di ϑ_1 è un problema diverso a seconda che si conosca o no il valore dell'altro parametro ϑ_2. Ad esempio, T_1 permette di stimare σ^2 senza conoscere nemmeno μ (implicitamente μ viene stimato con $\overline{X}_n$). Se però conoscessimo il vero valore di μ potremmo usare lo stimatore

$$T_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

che, intuitivamente, dovrebbe dare una stima migliore di σ^2, visto che utilizza un'informazione in più (il vero valore di μ). Si noti che se μ è incognito, T_2 *non* è uno stimatore, in quanto dipende da quantità incognite e quindi non è calcolabile dal campione.

Sono dei buoni stimatori T_1 e T_2?

Si può dimostrare (non lo facciamo!) che sono entrambi consistenti, ovvero per n grande tendono al parametro da stimare. Tuttavia mentre T_2 è non distorto ($\mathbb{E}[T_2] = \sigma^2$), T_1 risulta distorto! Precisamente si ha $\mathbb{E}[T_1] = \frac{n-1}{n}\sigma^2$. Per averne uno corretto occorre prendere $\frac{n}{n-1}T_1$, cioè

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Riassumendo se stiamo campionando da una popolazione con due parametri, media e varianza, per la stima puntuale della varianza si utilizza come stimatore S_X^2 in caso di media incognita, T_2 in caso di media nota (caso abbastanza poco frequente).

9.3 Stima per intervalli. Intervalli di confidenza per la media

Entriamo ora nel vivo dei metodi della statistica inferenziale, parlando di *stima per intervalli,* ovvero di *intervalli di confidenza.*

9.3.1 Stima della media di una popolazione normale con varianza nota

Esempio 9.3.1. Nella progettazione della cabina di pilotaggio di un aereo (disposizione della strumentazione, dimensioni dei comandi, ecc.) occorre anche tenere conto dei valori antropometrici medi dei piloti (statura, lunghezza delle braccia, ecc.). Supponiamo che la statura dei piloti sia distribuita secondo una legge normale $\mathrm{N}(\mu, \sigma^2)$. Ciò che interessa è stimare la media μ, a partire, ad esempio, dalle stature di 100 piloti dell'aviazione civile. In prima approssimazione supponiamo che la varianza (quindi la deviazione standard) sia nota e che risulti $\sigma = 6.1cm$.

Dunque abbiamo $X_1, X_2, \ldots, X_{100}$ variabili aleatorie i.i.d. $\mathrm{N}(\mu, 6.1^2)$. La stima puntuale della media sarà data da

$$\hat{\mu} = \overline{x}_n,$$

ossia utilizziamo la statistica $\overline{X}_n$ per stimare μ. Ricordiamo che se le X_i sono gaussiane allora la media campionaria è gaussiana, precisamente

$$\overline{X}_n \sim \mathrm{N}\Big(\mu, \frac{\sigma^2}{n}\Big) = \mathrm{N}\Big(\mu, \frac{6.1^2}{100}\Big),$$

e quindi

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_{100} - \mu}{0.61} \sim \mathrm{N}(0,1).$$

Pertanto, fissato $\alpha \in (0,1)$ per la (8.3), abbiamo:

$$\mathbb{P}\Big(\frac{|\overline{X}_{100} - \mu|}{0.61} < z_{(1+\alpha)/2}\Big) = \mathbb{P}\Big(- z_{(1+\alpha)/2} < \frac{\overline{X}_{100} - \mu}{0.61} < z_{(1+\alpha)/2}\Big) = \alpha.$$

Ad esempio, per $\alpha = 0.95$, $z_{(1+\alpha)/2} = z_{.975} = 1.96$, quindi

$$\mathbb{P}\Big(- 1.96 < \frac{\overline{X}_{100} - \mu}{0.61} < 1.96\Big) = 0.95,$$

ovvero

$$\mathbb{P}(\mu - 1.1956 < \overline{X}_{100} < \mu + 1.1956) = 0.95.$$

Risolvendo la disequazione rispetto a μ, anziché rispetto a $\overline{X}_{100}$,

$$\mathbb{P}(\overline{X}_{100} - 1.1956 < \mu < \overline{X}_{100} + 1.1956) = 0.95.$$

Questo significa che *prima* di eseguire il campionamento valutiamo pari a 0.95 la probabilità che

$$\overline{X}_{100} - 1.1956 < \mu < \overline{X}_{100} + 1.1956.$$

Si noti che l'intervallo

$$(\overline{X}_{100} - 1.1956, \overline{X}_{100} + 1.1956)$$

è, prima di eseguire il campionamento, un intervallo aleatorio (i suoi estremi sono variabili aleatorie) che, con probabilità 0.95, contiene il valore vero del parametro μ.

Eseguiamo ora il campionamento, e supponiamo di trovare, dalla misurazione delle stature dei 100 piloti,

$$\overline{x}_{100} = 178.5cm.$$

L'intervallo di confidenza diventa ora un *intervallo numerico,* non più aleatorio:

$$(\overline{x}_{100} - 1.1956, \overline{x}_{100} + 1.1956) = (178.5 - 1.1956, 178.5 + 1.1956) \simeq (177.3, 179.7).$$

Qual è il significato di questo intervallo trovato? Possiamo dire, ancora, che, con probabilità 0.95, $\mu \in (177.3, 179.7)$? La risposta è **NO,** perché μ non è una variabile aleatoria, ma un numero a noi incognito: il fatto che μ appartenga all'intervallo numerico (177.3,179.7) è semplicemente *vero* o *falso,* ma non dipende da alcun fenomeno aleatorio, e non ha senso, di conseguenza parlare di probabilità a questo riguardo. Si dice invece che, **con una confidenza del 95% il parametro μ appartiene all'intervallo,** e chiameremo questo intervallo aleatorio **intervallo di confidenza per μ al livello del 95%.**

Sintetizziamo ora la discussione fatta su questo esempio con una definizione più generale.

Definizione 9.3.2. Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione con distribuzione avente un parametro incognito $\vartheta \in \mathbb{R}$ e siano $T_1 = f_1(X_1, X_2, \ldots, X_n)$ e $T_2 = f_2(X_1, X_2, \ldots, X_n)$ due statistiche (variabili aleatorie funzione solo del campione) tali che

$$\mathbb{P}\Big(f_1(X_1, X_2, \ldots, X_n) < \vartheta < f_2(X_1, X_2, \ldots, X_n)\Big) = \alpha.$$

A campionamento eseguito l'intervallo numerico

$$\Big(f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n)\Big)$$

si chiama intervallo di confidenza al livello α per ϑ.

I numeri $f_1(x_1, x_2, \ldots, x_n)$ e $f_2(x_1, x_2, \ldots, x_n)$ vengono detti *limiti di confidenza.*

Si osservi il diverso uso che si fa dei termini *probabilità* e *confidenza:* prima di eseguire il campionamento, parliamo di probabilità che una variabile aleatoria assuma certi valori, dopo aver eseguito il campionamento, parliamo di confidenza che un certo parametro appartenga o meno a un intervallo numerico. Si può anche dire, meno rigorosamente, che la probabilità a che fare con avvenimenti futuri, la confidenza ha a che fare con fatti già accaduti.

Dall'Esempio 9.3.1 estraiamo il procedimento con cui, si determina, in generale, un intervallo di confidenza al livello α per la media μ di una popolazione $\mathrm{N}(\mu, \sigma^2)$ nel caso in cui σ^2 sia nota. Se $X_1, X_2, \ldots, X_n$ è un campione casuale di ampiezza n estratto da una popolazione $\mathrm{N}(\mu, \sigma^2)$, dato che,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1),$$

si ha

$$\mathbb{P}\Big(\frac{|\overline{X}_n - \mu|}{\sigma/\sqrt{n}} < z_{(1+\alpha)/2}\Big) = \mathbb{P}\Big(- z_{(1+\alpha)/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z_{(1+\alpha)/2}\Big) = \alpha,$$

o anche

$$\mathbb{P}\Big(\overline{X}_n - z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}}\Big) = \alpha.$$

Quindi a campionamento avvenuto l'intervallo numerico

$$\Big(\overline{x}_n - z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}}\Big)$$

è un intervallo di confidenza al livello α per μ.

Osservazione 9.3.3. Si osservi che l'intervallo di confidenza per μ è simmetrico ed è centrato nella stima puntuale e che tale intervallo è effettivamente calcolabile solo se σ è nota. Vedremo nel paragrafo successivo cosa accade se σ non è nota.

Osservazione 9.3.4. Si osservi che la "bontà" della stima dipende da due fattori:

- il livello di confidenza: più grande è α, più affidabile è la stima;
- l'ampiezza dell'intervallo: più è piccola, più è precisa la stima.

Dato che al crescere di α cresce (ovviamente) anche $z_{(1+\alpha)/2}$, fissato n, maggiore è il livello di confidenza, maggiore sarà l'ampiezza dell'intervallo. Pertanto affidabilità e precisione della stima sono due obiettivi tra loro antagonisti: migliorando uno si peggiora l'altro. Se si vuole aumentare la precisione della stima senza diminuire l'affidabilità (in genere 95% o 99%), occorre aumentare l'ampiezza del campione.

Per chiarire riprendiamo l'Esempio 9.3.1.

Esempio 9.3.5. Se, con gli stessi dati campionari, volessimo un intervallo di confidenza al 99% cosa cambierebbe? Con $n = 100$, $\alpha = 0.99$, $\sigma = 6.1cm$, $\overline{x}_n = 178.5cm$, avremmo che i limiti di confidenza sarebbero

$$\overline{x}_n \pm z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}} = (178.5 \pm 2.57 \cdot 0.61)cm,$$

quindi l'intervallo, in centimetri, sarebbe

$$(176.9, 180.1).$$

Abbiamo ottenuto un intervallo più ampio, quindi una stima più imprecisa: questo è il prezzo da pagare per avere una stima più affidabile.

Supponiamo ora di aver estratto (nello stesso esempio) un campione di 1000 individui e di aver trovato ancora $\overline{x}_n = 178.5cm$. L'intervallo di confidenza al livello 95% sarebbe ora quello di estremi

$$\overline{x}_n \pm z_{(1+\alpha)/2}\frac{\sigma}{\sqrt{n}} = (178.5 \pm 1.96 \cdot 0.193)cm,$$

quindi l'intervallo, in centimetri, sarebbe

$$(178.1, 178.9).$$

Abbiamo mantenuto la stessa affidabilità dell'Esempio iniziale ed abbiamo aumentato anche la precisione (l'intervallo è più stretto) ma abbiamo dovuto aumentare l'ampiezza del campione.

9.3.2 Stima della media di una popolazione normale con varianza incognita

Consideriamo ancora il problema di determinare un intervallo di confidenza per la media di una popolazione normale, mettendoci ora nell'ipotesi (più realistica) che anche la varianza sia incognita. La linea del discorso è sempre la stessa: cercheremo un intervallo simmetrico di estremi (aleatori) $\overline{X}_n \pm E$ che contenga il parametro da stimare μ, con probabilità α. Nel caso precedente abbiamo usato il fatto che

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1),$$

ma qui σ non la conosciamo! L'idea è quella di sostituire σ con la sua stima non distorta, ovvero S_n. La fortuna è che la quantità che si ottiene ha anch'essa una distribuzione nota (altrimenti sarebbe del tutto inutile), precisamente

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1).$$

Allora ricordando la definizione di t_α data nella (8.19) si ha:

$$\mathbb{P}\Big(\frac{|\overline{X}_n - \mu|}{S_n/\sqrt{n}} < t_{(1+\alpha)/2}(n-1)\Big) = \mathbb{P}\Big(-t_{(1+\alpha)/2}(n-1) < \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} < t_{(1+\alpha)/2}(n-1)\Big) = \alpha,$$

o anche

$$\mathbb{P}\Big(\overline{X}_n - t_{(1+\alpha)/2}(n-1)\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + t_{(1+\alpha)/2}(n-1)\frac{S_n}{\sqrt{n}}\Big) = \alpha.$$

Quindi a campionamento avvenuto l'intervallo numerico

$$\Big(\overline{x}_n - t_{(1+\alpha)/2}(n-1)\frac{s_n}{\sqrt{n}}, \overline{x}_n + t_{(1+\alpha)/2}(n-1)\frac{s_n}{\sqrt{n}}\Big)$$

è un intervallo di confidenza al livello α per μ.

Osservazione 9.3.6. A parità di dati l'intervallo di confidenza quando σ è noto è più stretto, ovvero la stima è più precisa. Questo è ovvio se si pensa che c'è un parametro in meno da stimare e quindi un'approssimazione in meno! Ma se $n > 120$, si ha che $t_\alpha(n-1) \simeq z_\alpha$. Quindi otteniamo numericamente lo stesso intervallo sia in caso di varianza nota che incognita, malgrado si utilizzino statistiche diverse per costruire gli intervalli.

9.3.3 Stima della media di una popolazione qualsiasi per grandi campioni

Nel paragrafo precedente abbiamo determinato l'intervallo di confidenza per la media di una popolazione normale, con varianza incognita. L'ipotesi di normalità era stata usata per affermare che

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1).$$

Se la popolazione non è normale, ma n è grande si ha

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \simeq T_n \sim t(n-1).$$

Ovvero la quantità utilizzata per costruire l'intervallo di confidenza ha ancora una distribuzione approssimativamente $t(n-1)$. Di conseguenza, ripetendo lo stesso identico ragionamento del paragrafo precedente, si ha che a campionamento avvenuto, l'intervallo numerico

$$\left(\overline{x}_n - t_{(1+\alpha)/2}(n-1)\frac{s_n}{\sqrt{n}}, \overline{x}_n + t_{(1+\alpha)/2}(n-1)\frac{s_n}{\sqrt{n}}\right)$$

è un intervallo di confidenza *approssimato* al livello α per μ.

Si ricordi anche che, se n è molto grande $n > 120$ si può fare l'ulteriore approssimazione $t_\alpha(n-1) \simeq z_\alpha$ ed in questo caso l'intervallo numerico

$$\left(\overline{x}_n - z_{(1+\alpha)/2}\frac{s_n}{\sqrt{n}}, \overline{x}_n + z_{(1+\alpha)/2}\frac{s_n}{\sqrt{n}}\right)$$

è un intervallo di confidenza *approssimato* al livello α per μ.

Esempio 9.3.7. Supponiamo che il tempo, misurato in giorni, tra due guasti successivi di un impianto segua una legge esponenziale di parametro λ incognito. Si decide di misurare un campione di 30 intertempi $X_1, X_2, \ldots, X_{30}$ per stimare il tempo medio tra due guasti. Si trova, sul campione osservato,

$$\overline{x}_{30} = 9.07 \quad s_{30} = 9.45.$$

Sulla base di queste osservazioni un intervallo di confidenza (approssimato) al livello α per la media di questa popolazione è quello di estremi

$$\overline{x}_{30} \pm t_{(1+\alpha)/2}(29)\frac{s_{30}}{\sqrt{n}},$$

quindi al livello 95%, gli estremi sono

$$9.07 \pm t_{.975}(29)\frac{9.45}{\sqrt{30}}.$$

Essendo $t_{.975}(29) = 2.0452$, l'intervallo richiesto è

$$(5.54, 12.6).$$

Ricordiamo che la media di una legge $\mathrm{Exp}(\lambda)$ è $1/\lambda$ pertanto quello trovato è un intervallo di confidenza per $1/\lambda$! Dunque un intervallo di confidenza per λ è

$$\left(\frac{1}{12.6}, \frac{1}{5.54}\right) = (0.079, 0.18).$$

9.3.4 Stima di una proporzione per grandi campioni

Un caso particolarmente interessante di stima della media per una popolazione non normale, sempre per grandi campioni, è quello in cui la popolazione è bernoulliana. Esempio tipico di questa situazione è il problema del sondaggio d'opinione: si vuole stimare la proporzione complessiva che è d'accordo con l'opinione A, per esempio vota a favore di un certo partito, osservando il valore che questa proporzione ha su un campione di n individui. Un altro esempio di questa situazione è il seguente. Supponiamo che un prodotto venga venduto in lotti numerosi; il produttore vuole garantire che la proporzione di pezzi difettosi sia in un certo intervallo prefissato.

Sappiamo che se $X_1, X_2, \ldots, X_n$ è un campione casuale estratto da una popolazione bernoulliana $\mathrm{Be}(p)$, nelle ipotesi in cui vale l'approssimazione normale, vedi Osservazione 8.3.1, è

$$\overline{X}_n \simeq Z \sim \mathrm{N}\Big(p, \frac{p(1-p)}{n}\Big),$$

o anche

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \simeq Z^* \sim \mathrm{N}(0,1),$$

Perciò fissato α, risulta

$$\mathbb{P}\Big(- z_{(1+\alpha)/2} < \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} < z_{(1+\alpha)/2}\Big) \simeq \mathbb{P}(-z_{(1+\alpha)/2} < Z^* < z_{(1+\alpha)/2}) = \alpha.$$

Per calcolare l'intervallo di confidenza per p risolviamo rispetto a p, si ottiene

$$\mathbb{P}\Big(\overline{X}_n - z_{(1+\alpha)/2}\sqrt{\frac{p(1-p)}{n}} < p < \overline{X}_n + z_{(1+\alpha)/2}\sqrt{\frac{p(1-p)}{n}}\Big) \simeq \alpha.$$

C'è ancora un problema! Gli estremi dell'intervallo dipendono ancora da p che è incognito. Quindi questo *non* è un intervallo di confidenza! In genere si fa la seguente (ulteriore) approssimazione. La quantità

$$\sqrt{\frac{p(1-p)}{n}},$$

si stima con

$$\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}.$$

Di conseguenza, si ha che a campionamento avvenuto, l'intervallo numerico

$$\Big(\overline{x}_n - z_{(1+\alpha)/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}, \overline{x}_n + z_{(1+\alpha)/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}\Big)$$

è un intervallo di confidenza *approssimato* al livello α per p.

Esempio 9.3.8. Supponiamo si voglia stimare la proporzione di elettori che approva l'operato del capo del governo. Su un campione di 130 persone intervistate nel mese di maggio, 75 erano favorevoli. Su un secondo campione di 1056 persone intervistate a ottobre 642 erano

favorevoli. Per ciascuno dei due campioni si vuole costruire un intervallo di confidenza al 95% per la proporzione degli elettori favorevoli al capo del governo. Si vuole inoltre confrontare la precisione delle due stime (ovvero confrontare le ampiezze dei due intervalli).

Nel primo caso $n = 130$ e $\hat{p} = \overline{x}_n = \frac{75}{130} = 0.57692$, quindi l'intervallo al livello 95% ha estremi

$$\overline{x}_n \pm z_{(1+\alpha)/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}} = \frac{75}{130} \pm 1.96\sqrt{\frac{\frac{75}{130}(1-\frac{75}{130})}{130}}.$$

Quindi l'intervallo richiesto è

$$(0.492, 0.662),$$

di ampiezza 0.170.

Nel secondo caso $n = 1056$ e $\hat{p} = \overline{x}_n = \frac{642}{1056} = 0.60795$, quindi l'intervallo al livello 95% ha estremi

$$\overline{x}_n \pm z_{(1+\alpha)/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}} = \frac{642}{1056} \pm 1.96\sqrt{\frac{\frac{642}{1056}(1-\frac{642}{1056})}{1056}}.$$

Quindi l'intervallo richiesto è

$$(0.579, 0.637),$$

di ampiezza 0.059, questa stima è molto più precisa!

9.4 Stima per intervalli. Intervalli di confidenza per la differenza di due medie di popolazioni normali

In molti studi siamo più interessati alla differenza tra due parametri piuttosto che al loro valore assoluto. Ci occuperemo soltanto della differenza tra due medie (si potrebbe, per esempio, considerare la differenza tra due varianze). In questo paragrafo tutte le osservazioni saranno estratte da popolazioni normali. Quindi anche se non indicato saremo in queste ipotesi.

9.4.1 Varianze note

Siano $X_1, X_2, \ldots, X_n$ v.a. i.i.d. $\mathrm{N}(\mu_X, \sigma_X^2)$ e $Y_1, Y_2, \ldots, Y_m$ v.a. i.i.d. $\mathrm{N}(\mu_Y, \sigma_Y^2)$. Supponiamo che i due campioni siano tra loro indipendenti e tentiamo di stimare $\mu_X - \mu_Y$. Dato che $\bar{X}_n$ e $\bar{Y}_m$ sono gli stimatori puntuali di μ_X e μ_Y rispettivamente sembra ragionevole prendere come stimatore puntuale di $\mu_X - \mu_Y$, la differenza $\bar{X}_n - \bar{Y}_n$. Per ottenere un intervallo di confidenza occorre conoscere la distribuzione di $\bar{X}_n - \bar{Y}_n$. Sappiamo che

$$\bar{X}_n \sim \mathrm{N}(\mu_X, \sigma_X^2/n), \quad \text{e} \quad \bar{Y}_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2/m),$$

si può dimostrare che

$$\bar{X}_n - \bar{Y}_m \sim \mathrm{N}(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m).$$

Pertanto

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim \mathrm{N}(0,1). \tag{9.1}$$

Pertanto si ha

$$\mathbb{P}\Big(\frac{|(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)|}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} < z_{(1+\alpha)/2}\Big) =$$
$$\mathbb{P}\Big(- z_{(1+\alpha)/2} < \frac{|(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)|}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} < z_{(1+\alpha)/2}\Big) = \alpha,$$

o anche

$$\mathbb{P}\Big((\bar{X}_n - \bar{Y}_m) - z_{(1+\alpha)/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < (\bar{X}_n - \bar{Y}_m) + z_{(1+\alpha)/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\Big) = \alpha.$$

Quindi a campionamento avvenuto l'intervallo numerico

$$(\bar{x}_n - \bar{y}_m) - z_{(1+\alpha)/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < (\bar{x}_n - \bar{y}_m) + z_{(1+\alpha)/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \qquad (9.2)$$

è un intervallo di confidenza al livello α per $\mu_X - \mu_Y$.

Osservazione 9.4.1. Si osservi che l'intervallo di confidenza per $\mu_X - \mu_Y$ è simmetrico ed è centrato nella stima puntuale e che tale intervallo è effettivamente calcolabile solo se σ_X e σ_Y sono note. Vedremo nel paragrafo successivo cosa accade se σ_X e σ_Y non sono note.

Osservazione 9.4.2. Si che la "bontà" della stima dipende da due fattori:

- il livello di confidenza: più grande è α, più affidabile è la stima;
- l'ampiezza dell'intervallo: più è piccola, più è precisa la stima.

Dato che al crescere di α cresce (ovviamente) anche $z_{(1+\alpha)/2}$, fissati n e m, maggiore è il livello di confidenza, maggiore sarà l'ampiezza dell'intervallo. Pertanto affidabilità e precisione della stima sono due obiettivi tra loro antagonisti: migliorando uno si peggiora l'altro. Se si vuole aumentare la precisione della stima senza diminuire l'affidabilità (in genere 95% o 99%), occorre aumentare l'ampiezza dei campioni.

Esempio 9.4.3. Due diversi tipi di guaine isolanti per cavi elettrici vengono testati per determinare a che voltaggio cominciano a rovinarsi. Sottoponendo gli esemplari a livelli crescenti di tensione si registrano guasti alle tensioni seguenti:

Tipo A	36	44	41	53	38	36	34	54	52	37	51	44	35	44
Tipo B	52	64	38	68	66	52	60	44	48	46	70	62		

Supponiamo di sapere che il voltaggio tollerato dai cavi abbia distribuzione normale: per quelli di tipo A, con media incognita μ_A e varianza $\sigma_A^2 = 40$, mentre per quelli di tipo B i parametri sono μ_B (incognito) e $\sigma_B^2 = 100$. Si determini un intervallo di confidenza al 95% per $\mu_A - \mu_B$.

Per il tipo A abbiamo le osservazioni $X_1, \ldots, X_{14} \sim \mathrm{N}(\mu_A, \sigma_A^2)$ e per il tipo B le osservazioni $Y_1, \ldots, Y_{12} \sim \mathrm{N}(\mu_B, \sigma_B^2)$. Per la (9.1)

$$\frac{\bar{X}_{14} - \bar{Y}_{12} - (\mu_A - \mu_B)}{\sqrt{\sigma_A^2/14 + \sigma_B^2/12}} \sim \mathrm{N}(0,1).$$

Il valore osservato della media empirica per i due campioni è

$$\bar{x}_{14} = \frac{1}{14}(36+44+41+53+38+36+34+54+52+37+51+44+35+44) = 42.7857,$$

$$\bar{y}_{12} = \frac{1}{12}(52+64+38+68+66+52+60+44+48+46+70+62) = 55.8333.$$

Pertanto l'intervallo di confidenza richiesto al 95%, dalla (9.2), è

$$(\bar{x}_n - \bar{y}_m) - z_{(1+\alpha)/2}\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}} < \mu_A - \mu_B < (\bar{x}_n - \bar{y}_m) + z_{(1+\alpha)/2}\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}},$$

sostituendo $n = 14$, $m = 12$, $z_{(1+\alpha)/2} = z_{0.975} = 1.96$, $\sigma_A^2 = 40$, $\sigma_B^2 = 100$ si ottiene l'intervallo $(-19.60; -6.49)$.

9.4.2 Varianze incognite, ma uguali

Supponiamo ore $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (ma σ incognito). Possiamo ancora dire che,

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} = \frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{N}(0,1). \tag{9.3}$$

Ora però la quantità non può essere utilizzata perché *non* calcolabile (contiene l'incognita σ)! Occorre stimarla. Utilizziamo le varianze campionarie dei due campioni.

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2.$$

Definiamo la **varianza campionaria pesata**

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Questa è la quantità più naturale da introdurre per stimare σ^2 in quanto tiene conto di entrambe le varianze e del peso di ciascuna di esse. Tuttavia la cosa importante che ci permette di poter costruire un test è che sostituendo S al posto di σ nella (9.3) si ottiene una variabile aleatoria che ha ancora una legge nota (se questo non fosse vero la quantità nella (9.4) sarebbe inutile!!!). Precisamente

$$\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \sim t(n+m-2). \tag{9.4}$$

Con un ragionamento analogo a quello fatto in tutti gli altri casi possiamo allora costruire l'intervallo di confidenza per $\mu_X - \mu_Y$. Si ha

$$\mathbb{P}\Big(\frac{|(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)|}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} < t_{(1+\alpha)/2}(n+m-2)\Big) =$$

$$\mathbb{P}\Big(-t_{(1+\alpha)/2}(n+m-2) < \frac{|(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)|}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} < t_{(1+\alpha)/2}(n+m-2)\Big) = \alpha,$$

o anche, detta per brevità

$$D = \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}},$$

$$\mathbb{P}\Big((\bar{X}_n - \bar{Y}_m) - t_{(1+\alpha)/2}(n+m-2)D < \mu_X - \mu_Y < (\bar{X}_n - \bar{Y}_m) + t_{(1+\alpha)/2}(n+m-2)D \Big) = \alpha.$$

Quindi a campionamento avvenuto, detta

$$d = \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}},$$

l'intervallo numerico

$$(\bar{x}_n - \bar{y}_m) - t_{(1+\alpha)/2}(n+m-2)d < \mu_X - \mu_Y < (\bar{x}_n - \bar{y}_m) + t_{(1+\alpha)/2}(n+m-2)d \qquad (9.5)$$

è un intervallo di confidenza al livello α per $\mu_X - \mu_Y$.

Esempio 9.4.4. Un produttore di batterie dispone di due tecniche di fabbricazione differenti. Supponiamo di sapere che la capacità delle batterie abbia distribuzione normale: per quelle di tipo A, con media incognita μ_X e varianza incognita σ_X^2, mentre per quelle di tipo B con media incognita μ_Y e varianza incognita σ_Y^2. Supponiamo inoltre $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Vengono osservati i valori della capacità di 14 batterie di tipo A e 12 di tipo B. Si determini un intervallo di confidenza al 90% per $\mu_X - \mu_Y$, nell'ipotesi che il valore osservato della media empirica per i due campioni sia

$$\bar{x}_{14} = 135.79 ampère/ora, \quad \bar{y}_{12} = 143 ampère/ora,$$

e i valori osservati per le varianze campionarie (corrette) S_X^2 e S_Y^2 siano

$$s_X^2 = 44.38, \quad s_Y^2 = 46.33$$

Per il tipo A abbiamo le osservazioni $X_1, \ldots, X_{14} \sim \mathrm{N}(\mu_X, \sigma^2)$ e per il tipo B le osservazioni $Y_1, \ldots, Y_{12} \sim \mathrm{N}(\mu_Y, \sigma^2)$. Per la (9.4)

$$\frac{\bar{X}_{14} - \bar{Y}_{12}}{\sqrt{\frac{1}{14} + \frac{1}{12}} \sqrt{\frac{(14-1)S_X^2 + (12-1)S_Y^2}{14+12-2}}} \sim t(14+12-2),$$

ovvero

$$\frac{\bar{X}_{14} - \bar{Y}_{12}}{\sqrt{\frac{13}{84}} \sqrt{\frac{13S_X^2 + 11S_Y^2}{24}}} \sim t(24),$$

quindi

$$d = \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = \sqrt{\frac{13}{84}} \sqrt{\frac{13 \cdot 44.38 + 11 \cdot 46.33}{24}} = 2.65$$

Pertanto l'intervallo di confidenza richiesto al 90%, dalla (9.5), è

$$(\bar{x}_n - \bar{y}_m) - t_{(1+\alpha)/2}(n+m-2)d < \mu_X - \mu_Y < (\bar{x}_n - \bar{y}_m) + t_{(1+\alpha)/2}(n+m-2)d,$$

sostituendo $n = 14$, $m = 12$, $t_{(1+\alpha)/2}(24) = t_{0.95}(24) = 1.71$, si ottengono gli estremi

$$(\bar{x}_n - \bar{y}_m) \pm t_{(1+\alpha)/2}(n+m-2)d = (-7.21 \pm 4.53)ampère/ora,$$

quindi l'intervallo è (-11.74,-2.68).

Osservazione 9.4.5. *Abbiamo visto fin qui, come si trova un intervallo di confidenza per la differenza di medie di due popolazioni normali indipendenti in due casi particolari. Varianze note (***a.***) o varianze incognite, ma uguali (***b.***). Vediamo ora di renderci conto, brevemente, di quale sia il significato concreto di queste ipotesi.*

a. Oltre al caso in cui le varianze siano effettivamente note (per esempio per studi precedenti) si tratta in questo modo anche il caso in cui i campioni sono entrambi molto numerosi ($n, m > 30$). In questo caso è prassi trattare le varianze campionarie come fossero le varianze effettive.

b. Questa seconda ipotesi può essere sensata quando le due popolazioni in esame sono, in qualche senso, entrambe parte di una popolazione più vasta, per cui possiamo pensare che condividano la varianza della popolazione più vasta, pur avendo magari medie diverse. Ad esempio se X e Y rappresentano la statura di un individuo adulto scelto a caso in due regioni differenti. Potranno avere medie differenti, ma è ragionevole supporre che abbiamo stessa varianza. Questa ipotesi può essere sensata anche quando, nel caso di grandi campioni, si trovano varianze campionarie *vicine* tra loro.

Capitolo 10

Test d'ipotesi

10.1 Generalità

Ci sono molte situazioni in cui un'indagine campionaria viene eseguita per *prendere una decisione* su un'intera popolazione. Per esempio:

- si vaccinano alcune persone per decidere se un vaccino è efficace o no;
- si lancia un certo numero di volte una moneta per decidere se è equa o no;
- si misura il grado di impurità in un certo numero di campioni di acqua per decidere se è potabile o no.

Tale decisione viene detta *decisione statistica.* Quando si tratta di raggiungere una decisione statistica occorre fare delle *ipotesi statistiche.* Se vogliamo testare una moneta per vedere se è equa faremo le ipotesi "la moneta è equa" e "la moneta non è equa". Analogamente se vogliamo testare un vaccino faremo le ipotesi "vaccino efficace" e "vaccino non efficace". Vedremo poi come si modellano queste ipotesi. Le ipotesi fatte devono comprendere tutte le possibilità di interesse. Chiameremo *ipotesi nulla* e la indicheremo con H_0 una delle due ipotesi, in genere quella che si vuole rifiutare; *ipotesi alternativa* e la indicheremo con H_1 ogni altra ipotesi. I procedimenti o regole che permettono poi di accettare o respingere un'ipotesi vengono detti **test d'ipotesi.**

Esempio 10.1.1. Devo decidere se una moneta è buona o no. La lancio 100 volte. Facciamo le ipotesi:
H_0: la moneta è equa;
H_1: la moneta non è equa.
Se esce un numero di teste compreso tra 40 e 60 dico che la moneta è buona, altrimenti dico che è truccata. Questo è un test d'ipotesi.

Ovviamente si può sbagliare! Sulla base delle osservazioni campionarie posso commettere due tipi di errore.
Errore del I tipo: Rifiuto H_0, invece H_0 è vera;
Errore del II tipo: Accetto H_0, invece H_0 è falsa.

Riassumendo un test statistico è una procedura con cui, a partire dai dati campionari si decide se rifiutare H_0 o non rifiutarla. Tutte le possibilità sono raccolte nella seguente tabella.

	Se H_0 è vera	Se H_0 è falsa
e noi rifiutiamo H_0	**Errore del I tipo**	Decisione corretta
e noi non rifiutiamo H_0	Decisione corretta	**Errore del II tipo**

Esempio 10.1.2. Nell'Esempio 10.1.1 gli errori sono i seguenti.

Errore del I tipo : Esce un numero di teste minore di 40 o maggiore di 60 ma in realtà la moneta è equa.

Errore del II tipo : Esce un numero di teste compreso tra 40 e 60 ma in realtà la moneta è truccata.

Minimizzare entrambi gli errori è impossibile!! Anzi si può dimostrare (difficile per noi...) che minimizzando uno, l'altro diventa maggiore. Pertanto occorre *scegliere* quale errore sia più grave e cercare di contenere quello. La regola vuole che sia più grave l'errore di I tipo quindi si cerca di tenere quello basso senza far sfuggire fuori controllo l'altro. La scelta delle ipotesi viene fatta (se possibile) in modo tale che sia più grave rifiutare H_0 quando invece H_0 è vera. Detto in altri termini l'ipotesi nulla è quella che vogliamo rifiutare solo di fronte a "prove schiaccianti". Si chiama *livello di significatività* del test la massima probabilità di rifiutare H_0 quando H_0 è vera, ovvero il massimo della probabilità dell'errore di I specie. Il livello di significatività va stabilito a priori, cioè prima di eseguire il test. Valori tipici per il livello di significatività sono 1%, 5%.

Esempio 10.1.3. Viene somministrato un nuovo vaccino. Occorre decidere se sia efficace o no. È più grave decidere che sia efficace quando non lo è o che non sia efficace quando lo è? Sicuramente è più grave la prima eventualità. Allora, in questo caso, si pone

H_0: vaccino non efficace;

H_1: vaccino efficace.

	Se il vaccino non è efficace	Se il vaccino è efficace
e noi lo riteniamo efficace	**Errore del I tipo**	Decisione corretta
e noi lo riteniamo non efficace	Decisione corretta	**Errore del II tipo**

Anche questa situazione può aiutare a chiarire. Si processa un imputato. È più grave decidere che è colpevole quando non lo è o che non è colpevole quando lo è? Sicuramente è più grave la prima eventualità. Allora, in questo caso,

H_0: imputato non colpevole;

H_1: imputato colpevole.

	Se l'imputato non è colpevole	Se l'imputato è colpevole
e noi lo riteniamo colpevole	**Errore del I tipo**	Decisione corretta
e noi lo riteniamo non colpevole	Decisione corretta	**Errore del II tipo**

Nei casi precedenti era molto semplice decidere quale fosse l'eventualità più grave. Vediamo situazioni in cui è meno evidente qual è la regola generale con cui vengono scelte la ipotesi. Anche qui è più conveniente dedurre un criterio da alcuni esempi.

Esempio 10.1.4. Vediamo alcune situazione classiche che si possono presentare.

1. Il contenuto dichiarato delle bottiglie di acqua minerale di una certa marca è $990ml$. Un'associazione di consumatori sostiene che in realtà le bottiglie contengono, in media, una quantità inferiore d'acqua.

2. Due amici giocano a testa o croce; uno dei due ha il sospetto che la moneta sia truccata e decide di registrare l'esito di un certo numero di lanci (come nell'Esempio 10.1.1).

3. Un ingegnere suggerisce alcune modifiche che si potrebbero apportare ad una linea produttiva per aumentare il numero di pezzi prodotti giornalmente. Si decide di sperimentare queste modifiche su una macchina: se i risultati saranno buoni verranno applicati alle altre macchine.

In questi casi si possono fare le seguenti considerazioni.

1. Supponiamo che la quantità d'acqua contenuta in ciascuna bottiglia si possa modellare con una variabile aleatoria $X \sim \mathrm{N}(\mu, \sigma^2)$. Dobbiamo eseguire un test sulla media. Qui il criterio è quello *innocentista:* ci vuole una forte evidenza per accusare il produttore di vendere bottiglie irregolari. Quindi:

   $H_0 : \mu = 990ml$ o anche $H_0 : \mu \geq 990ml$ (il produttore non imbroglia);

   $H_1 : \mu < 990ml$ (il produttore imbroglia).

2. Qui il risultato di un lancio è una variabile aleatoria $X \sim \mathrm{Be}(p)$. Dobbiamo eseguire un test sul parametro p (che ricordiamo è anche la media di X). Qui anche seguiamo una ipotesi *innocentista.* Supponiamo pertanto che la moneta sia equa, quindi

   $H_0 : p = 0.5$;

   $H_1 : p \neq 0.5$.

3. Il numero dei pezzi prodotti dalla macchina prima della modifica si può modellare con una variabile aleatoria X (con legge non nota) con media μ_0, nota. L'idea è che, poiché ogni cambiamento ha un costo, si seguirà il suggerimento dell'ingegnere solo se i risultati sperimentali forniranno una forte evidenza del fatto che la macchina modificata sia più produttiva di quella originaria, ovvero che ora il numero dei pezzi prodotti sia una variabile aleatoria X con media $\mu > \mu_0$. Perciò:

   $H_0 : \mu = \mu_0$ o anche $\mu \leq \mu_0$;

   $H_1 : \mu > \mu_0$.

Vediamo di formalizzare e di generalizzare. Abbiamo un campione $X_1, X_2, \ldots, X_n$ estratto da una popolazione avente una distribuzione dipendente da un parametro $\vartheta \in \Theta$ sul quale vogliamo fare delle ipotesi. Dal punto di vista matematico le ipotesi possono essere così viste:

$$H_0 : \vartheta \in \Theta_0, \quad H_1 : \vartheta \in \Theta_1,$$

dove $\Theta_0 \cap \Theta_1 = \emptyset$ e $\Theta_0 \cup \Theta_1 = \Theta$. Ora dobbiamo decidere il test ovvero la regola per accettare o rifiutare H_0. Si sceglie una statistica, diciamo $T(X_1, X_2, \ldots, X_n)$, quindi si rifiuta H_0 se

$$T(x_1, x_2, \ldots, x_n) \in I,$$

ovvero se la statistica scelta calcolata sulle osservazioni campionarie cade in una certa regione. L'insieme R dei possibili risultati campionari che portano a rifiutare H_0 è detta **regione critica** o **regione di rifiuto** del test:

$$R = \{(x_1, x_2, \ldots, x_n) : T(x_1, x_2, \ldots, x_n) \in I\}.$$

Cerchiamo di completare la formalizzazione dei casi visti nell'Esempio 10.1.4.

Esempio 10.1.5. 1. Supponiamo di misurare il contenuto di 100 bottiglie di acqua minerale. Come costruisco un test per decidere se contengono, in media, la quantità d'acqua desiderata? Dunque abbiamo un campione $X_1, X_2, \ldots, X_n$ (ciascuna X_i rappresenta la quantità d'acqua contenuta in una delle bottiglie) e supponiamo che $X_i \sim \mathrm{N}(\mu, \sigma^2)$. Abbiamo visto che le ipotesi sono:

$$H_0 : \mu = (\geq) 990ml \quad H_1 : \mu < 990ml.$$

Ora devo decidere la regola per rifiutare H_0. Una regola "ragionevole" sembra la seguente: rifiuto H_0 se $\overline{X}_n$ sul campione assume un valore troppo più piccolo di $990ml$, ovvero se $\overline{x}_n < k$, dove k è un valore da determinare, vedremo come. Quindi la statistica da utilizzare per il test è

$$T(X_1, X_2, \ldots, X_n) = \overline{X}_n$$

e la regione critica è

$$R = \{(x_1, x_2, \ldots, x_n) : \overline{x}_n < k\}.$$

2. Torniamo all'Esempio 10.1.1. Ricordiamo che lanciamo la moneta 100 volte e decidiamo che la moneta non è equa se esce un numero di teste minore di 40 o maggiore di 60. Vediamo di formalizzare. Abbiamo un campione $X_1, X_2, \ldots, X_{100}$ dove $X_i \sim Be(p)$ (X_i vale 1 se se esce testa al lancio i, 0 se esce croce) e vogliamo prendere una decisione sul parametro p.

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5.$$

Qui la statistica da utilizzare per il test è

$$T(X_1, X_2, \ldots, X_n) = X_1 + X_2 + \ldots + X_n,$$

mentre la regione di rifiuto è

$$R = \{(x_1, x_2, \ldots, x_n) : x_1 + x_2 + \ldots + x_n < 40 \text{ o } x_1 + x_2 + \ldots + x_n > 60\}.$$

3. Applichiamo la modifica ad una delle macchine e per 100 giorni andiamo a vedere quanti pezzi produce. Abbiamo un campione $X_1, X_2, \ldots, X_{100}$ dove ciascuna X_i è una variabile aleatoria con media μ (attenzione: la macchina è stata modificata, la sua media ora non so quanto vale, è proprio questo il problema!). Ricordiamo che prima

della modifica il numero dei pezzi prodotti aveva media μ_0. Le ipotesi che facciamo, abbiamo visto, sono

$$H_0 : \mu = (\leq)\mu_0 \quad H_1 : \mu > \mu_0.$$

Ora devo decidere la regola per rifiutare H_0. Una regola "ragionevole" sembra la seguente: rifiuto H_0 se $\overline{X}_n$ sul campione assume un valore molto grande, ovvero se $\overline{x}_n > k$, dove k è un valore da determinare, vedremo come. Quindi qui la statistica da utilizzare per il test è ancora la media campionaria,

$$T(X_1, X_2, \ldots, X_n) = \overline{X}_n$$

e la regione critica è

$$R = \{(x_1, x_2, \ldots, x_n) : \overline{x}_n > k\}.$$

La forma della regione critica si decide sulla base di osservazioni "ragionevoli". Per esempio $\{\overline{X}_n > k\}$ o $\{\overline{X}_n < k\}$. Mentre l'esatta regione critica, che nei casi precedenti equivale a determinare il valore di k (cioè quanto grande o quanto piccola deve essere la media empirica per rifiutare H_0) si decide in base al livello fissato del test. Come la regione critica ed il livello del test sono legati lo vedremo volta per volta nei casi che andremo a trattare.

Consideriamo un campione $X_1, X_2, \ldots, X_n$ estratto da una popolazione con distribuzione dipendente da un parametro $\vartheta \in \Theta$ e **ricapitoliamo i passi in cui si articola un test statistico:**

1. Si scelgono l'ipotesi nulla e l'ipotesi alternativa (questo comporta un giudizio su quale delle due ipotesi sia quella da rifiutare solo in caso di forte evidenza);

2. si sceglie una statistica (per esempio $\overline{X}_n$) su cui basare il test, e si decide la *forma* della regione critica (per esempio $\{\overline{X}_n > k\}$), questo in base a considerazioni *ragionevoli*;

3. si sceglie il livello (per esempio $\alpha = 0.05$) e quindi si determina esattamente la regione critica (per esempio $\{\overline{X}_n > 15.8\}$). Come già detto torneremo su questo punto;

4. si esegue il campionamento e si calcola la statistica coinvolta nel test e si vede se appartiene o no alla regione di rifiuto. Quindi si prende la decisione se rifiutare o no l'ipotesi nulla.

10.2 Test sulla media per una popolazione normale

10.2.1 Varianza nota

Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una popolazione normale con media incognita (parametro su cui vogliamo fare un'ipotesi) e varianza nota. Quindi X_i sono variabili i.i.d. con distribuzione $\mathrm{N}(\mu, \sigma^2)$.

L'ipotesi nulla e l'ipotesi alternativa, rifacendoci agli esempi precedenti sono del tipo:

H_0	H_1
$\mu = \mu_0$	$\mu \neq \mu_0$
$\mu = \mu_0$	$\mu > \mu_0$
$\mu = \mu_0$	$\mu < \mu_0$

dove μ_0 è un valore fissato: NOTO! In questo caso si dice che H_0 è una ipotesi semplice, ovvero Θ_0 è formato da un unico punto.

Nei tre casi la statistica che sembra ragionevole utilizzare, dato che devo fare ipotesi sulla media, è la media campionaria, $\overline{X}_n$. Inoltre la regione critica adatta per rifiutare H_0 sarà, rispettivamente, del tipo:

si rifiuti H_0 se $|\overline{X}_n - \mu_0| > k$, ovvero si rifiuta H_0 se la media campionaria è lontana dal valore che ci si attende μ_0;

si rifiuti H_0 se $\overline{X}_n > k$, ovvero si rifiuta H_0 se la media campionaria è molto maggiore del valore che ci si attende μ_0;

si rifiuti H_0 se $\overline{X}_n < k$, ovvero si rifiuta H_0 se la media campionaria è molto minore del valore che ci si attende μ_0.

Fissiamo ora il livello α della regione di rifiuto. Dobbiamo fare in modo che l'errore di prima specie sia α. Ma ricordiamo che l'errore di prima specie è la probabilità di rifiutare H_0 quando H_0 è vera. Quindi è la probabilità della regione critica, quando H_0 è vera.

Quindi deve essere, nel primo caso,

$$\mathbb{P}(|\overline{X}_n - \mu_0| > k) = \alpha.$$

Come fare? Ricordiamo che, nel nostro caso, se l'ipotesi è vera,

$$\overline{X}_n \sim \mathrm{N}(\mu_0, \sigma^2/n),$$

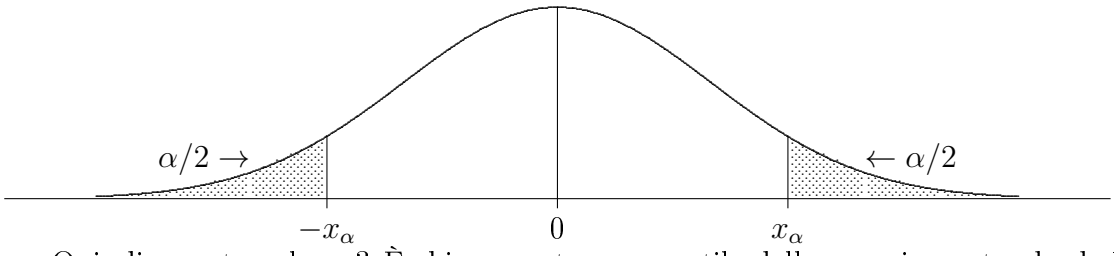
oppure, che è lo stesso, standardizzando

$$\frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} = Z^* \sim \mathrm{N}(0,1).$$

Quindi,

$$\mathbb{P}(|\overline{X}_n - \mu_0| > k) = \mathbb{P}\Big(\frac{|\overline{X}_n - \mu_0|}{\sigma}\sqrt{n} > \frac{k}{\sigma}\sqrt{n}\Big) = \mathbb{P}\Big(|Z^*| > \frac{k}{\sigma}\sqrt{n}\Big) = \alpha.$$

Dunque deve essere uguale ad α l'area ombreggiata in figura:



Quindi quanto vale x_α? È chiaramente un quantile della gaussiana standard. Ma quale? Quanta area lascia x_α alla sua sinistra? Facile! L'area *non ombreggiata* è $1-\alpha$ pertanto l'area a sinistra è di x_α è $1-\alpha+\alpha/2 = 1-\alpha/2$. Pertanto, ricordando che abbiamo indicato con z_α i quantili della gaussiana standard, abbiamo $x_\alpha = z_{1-\alpha/2}$. Pertanto deve essere

$$\frac{k}{\sigma}\sqrt{n} = z_{1-\alpha/2} \quad \Rightarrow \quad k = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Quindi la regione critica di livello α è,

$$\{|\overline{X}_n - \mu_0| > z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\Big\},$$

oppure, che è lo stesso,

$$\Big\{\frac{|\overline{X}_n - \mu_0|}{\sigma}\sqrt{n} > z_{1-\alpha/2}\Big\}.$$

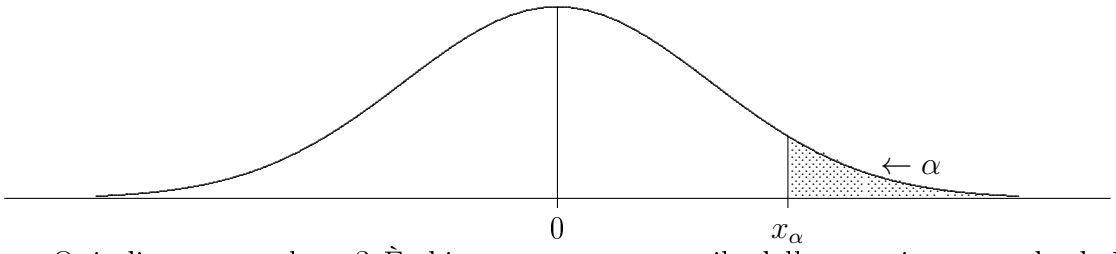
Passiamo al secondo caso. Deve essere

$$\mathbb{P}(\overline{X}_n > k) = \alpha.$$

Procediamo esattamente come nel caso precedente.

$$\mathbb{P}(\overline{X}_n - \mu_0 > k - \mu_0) = \mathbb{P}\Big(\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} > \frac{k-\mu_0}{\sigma}\sqrt{n}\Big) = \mathbb{P}\Big(Z^* > \frac{k-\mu_0}{\sigma}\sqrt{n}\Big) = \alpha.$$

Dunque deve essere uguale ad α l'area ombreggiata in figura:



Quindi quanto vale x_α? È chiaramente un quantile della gaussiana standard. Ma quale? Quanta area lascia x_α alla qua sinistra? Facile! L'area *non ombreggiata* è $1-\alpha$ pertanto l'area a sinistra è di x_α è $1-\alpha$. Pertanto, ricordando che abbiamo indicato con z_α i quantili della gaussiana standard, abbiamo $x_\alpha = z_{1-\alpha}$. Pertanto deve essere

$$\frac{k-\mu_0}{\sigma}\sqrt{n} = z_{1-\alpha} \quad \Rightarrow \quad k = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

Quindi la regione critica di livello α è,

$$\{\overline{X}_n > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\Big\},$$

oppure, che è lo stesso,

$$\Big\{\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} > z_{1-\alpha}\Big\}.$$

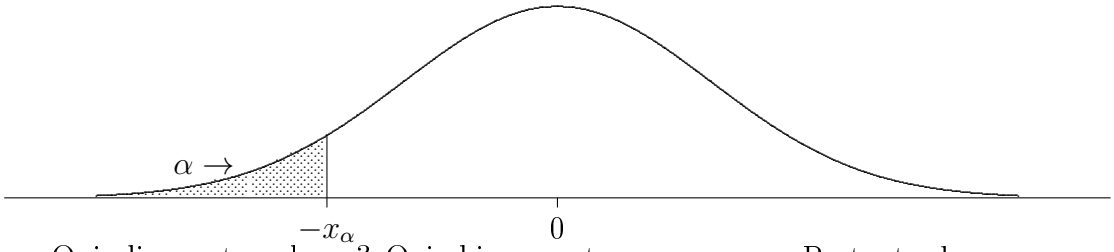
Veniamo al terzo ed ultimo caso. Deve essere

$$\mathbb{P}(\overline{X}_n < k) = \alpha.$$

Procediamo esattamente come nel caso precedente.

$$\mathbb{P}(\overline{X}_n - \mu_0 < k - \mu_0) = \mathbb{P}\Big(\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} < \frac{k-\mu_0}{\sigma}\sqrt{n}\Big) = \mathbb{P}\Big(Z^* < \frac{k-\mu_0}{\sigma}\sqrt{n}\Big) = \alpha.$$

Dunque deve essere uguale ad α l'area ombreggiata in figura:



Quindi quanto vale x_α? Qui chiaramente $x_\alpha = -z_{1-\alpha}$. Pertanto deve essere

$$\frac{k-\mu_0}{\sigma}\sqrt{n} = -z_{1-\alpha} \quad \Rightarrow \quad k = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

Quindi la regione critica di livello α è,

$$\{\overline{X}_n < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\},$$

oppure, che è lo stesso,

$$\left\{\frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} < -z_{1-\alpha}\right\}.$$

Riassumendo. Supponiamo di voler eseguire un test sulla media di una popolazione normale di varianza σ^2 nota, estraendo un campione casuale di ampiezza n. Se poniamo

$$z^* = \frac{\overline{x}_n - \mu_0}{\sigma}\sqrt{n},$$

possiamo esprimere la regola di decisione del test, in dipendenza dall'ipotesi nulla e dal livello di significatività che abbiamo scelto, nel modo seguente:

H_0	H_1	Rifiuto H_0 se
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z^*\rvert > z_{1-\alpha/2}$
$\mu = \mu_0$	$\mu > \mu_0$	$z^* > z_{1-\alpha}$
$\mu = \mu_0$	$\mu < \mu_0$	$z^* < -z_{1-\alpha}$

Questo test che utilizza la distribuzione normale si chiama *z-test.*

ATTENZIONE! Come si procede nel caso in cui H_0 non sia semplice? Risulta più complicato il calcolo di k (in base al livello) tuttavia si trova che il test è esattamente lo stesso che abbiamo ottenuto nel caso in cui H_0 è semplice. Precisamente si ha,

H_0	H_1	Rifiuto H_0 se
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z^*\rvert > z_{1-\alpha/2}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$z^* > z_{1-\alpha}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$z^* < -z_{1-\alpha}$

Esempio 10.2.1. Da una popolazione normale di media incognita e deviazione standard $\sigma = 2$ si estrae un campione di ampiezza 10, per sottoporre a test l'ipotesi nulla $H_0 : \mu = 20$, contro l'alternativa $\mu \neq 20$.

1. Qual è la regione critica, ai livelli 1%, 5%, 10%, per questo test?
2. Supponendo di aver estratto un campione per cui $\overline{x}_n = 18.58$, si tragga una conclusione, a ciascuno dei tre livelli di significatività.

1. La regione critica del test è
$$\left\{\frac{|\overline{X}_n - \mu_0|}{\sigma}\sqrt{n} > z_{1-\alpha/2}\right\} = \left\{\frac{|\overline{X}_{10} - 20|}{2}\sqrt{10} > z_{1-\alpha/2}\right\},$$
con
$$z_{1-\alpha/2} = \begin{cases} 2.57 & \text{per } \alpha = 0.01 \\ 1.96 & \text{per } \alpha = 0.05 \\ 1.64 & \text{per } \alpha = 0.10 \end{cases}$$
2. Se $\overline{x}_{10} = 18.58$, allora
$$z^* = \frac{|\overline{x}_{10} - 20|}{2}\sqrt{10} = 2.25,$$
pertanto la conclusione che si trae in ciascuno dei tre casi è:

   i dati campionari *non consentono* di rifiutare l'ipotesi nulla, al livello di significatività dell' 1%.

   i dati campionari *consentono* di rifiutare l'ipotesi nulla, al livello di significatività del 5%.

   i dati campionari *consentono* di rifiutare l'ipotesi nulla, al livello di significatività del 10%.

Come si vede, la decisione che si prende non dipende solo dai dati campionari, ma anche dal livello di significatività fissato. In questo caso la discrepanza tra il valore della media osservato (18.58) e quello ipotizzato (20) viene ritenuto statisticamente significativo al livello del 5% e del 10%, ma non al livello dell'1%. *Questo significa che se il valore vero del parametro è 20, la probabilità di ottenere, per effetto delle oscillazioni casuali, uno scostamento della media campionaria dal valore 20 pari a quello osservato, è inferiore al 5%, ma superiore all'1%.*

10.2.2 Varianza incognita

Consideriamo ancora il problema di determinare un test sulla media di una popolazione normale, mettendoci ora nell'ipotesi (più realistica) che anche la varianza sia incognita. La linea del discorso è sempre la stessa: cercheremo una regione critica "ragionevole" sempre basata sulla media campionaria. Nel caso precedente abbiamo usato il fatto che, se $X_1, X_2, \ldots, X_n$ è un campione estratto da una popolazione $\mathrm{N}(\mu, \sigma^2)$, allora
$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1),$$

ma qui σ non la conosciamo! L'idea è quella di sostituire σ con la sua stima non distorta, ovvero S_n. La fortuna è che la quantità che si ottiene ha anch'essa una distribuzione nota (altrimenti sarebbe del tutto inutile), precisamente

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1).$$

Allora ricordando la definizione di t_α data nella (8.19) possiamo subito scrivere come verrà il test.

Riassumendo. Supponiamo di voler eseguire un test sulla media di una popolazione normale di varianza σ^2 incognita, estraendo un campione casuale di ampiezza n. Se poniamo

$$t = \frac{\overline{x}_n - \mu_0}{s_n}\sqrt{n},$$

possiamo esprimere la regola di decisione del test, in dipendenza dall'ipotesi nulla e dal livello di significatività che abbiamo scelto, nel modo seguente:

H_0	H_1	Rifiuto H_0 se
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert > t_{1-\alpha/2}(n-1)$
$\mu = \mu_0$ $(\mu \leq \mu_0)$	$\mu > \mu_0$	$t > t_{1-\alpha}(n-1)$
$\mu = \mu_0$ $(\mu \geq \mu_0)$	$\mu < \mu_0$	$t < -t_{1-\alpha}(n-1)$

Questo test, è detto *t-test.*

10.3 Test sulla media di una popolazione qualsiasi per grandi campioni

- Nel caso di un campione numeroso ($n \geq 30$) estratto da una popolazione **qualsiasi** come già visto per gli intervalli di confidenza, possiamo considerare la relazione

$$\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \simeq T_n \sim t(n-1),$$

ovvero $\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}$ approssimativamente una distribuzione $t(n-1)$. Pertanto si può fare esattamente lo stesso test visto per campioni gaussiani in caso di varianza incognita.

Esempio 10.3.1. Dall'esperienza passata è noto che il numero di rapine che ogni settimana avvengono in una certa città è una variabile aleatoria di media 1. Nel corso dell'anno passato ci sono state 85 rapine (quindi una media di 85/52 rapine alla settimana) con una deviazione standard (campionaria) pari a 1.5. Si può affermare che l'entità del fenomeno sia cresciuta in modo significativo? Per rispondere si faccia un test, al livello dell'1%, sull'ipotesi che il parametro non sia cresciuto.

Questa volta abbiamo un campione numeroso estratto da una popolazione non normale (di legge sconosciuta). Le ipotesi sono $H_0 : \mu = 1$ e $H_1 : \mu > 1$. pertanto la regione critica è della forma $\{\overline{X}_n > k\}$. Possiamo fare un t-test. Calcoliamo

$$t = \frac{\overline{x}_n - \mu_0}{\sqrt{s_n^2/n}} = \frac{85/52 - 1}{\sqrt{1.5^2/52}} = 3.05.$$

I gradi di libertà sono 51, il livello di significatività 0.01 perciò la regola di decisione è si rifiuti H_0 se $t > t_{.99}(51) = 2.37$. Perciò si può rifiutare l'ipotesi al livello dell'1% e concludere che il numero di rapine è aumentato.

10.4 Test su una frequenza per grandi campioni

Possiamo ripetere per il test d'ipotesi gran parte dei ragionamenti visti nel Paragrafo 9.3.4 per il calcolo degli intervalli di confidenza per la frequenza di una popolazione bernoulliana. Volendo fare un test per le ipotesi seguenti,

H_0	H_1
$p = p_0$	$p \neq p_0$
$p = p_0$ $(p \leq p_0)$	$p > p_0$
$p = p_0$ $(p \geq p_0)$	$p < p_0$

utilizzeremo il fatto che, per l'approssimazione normale, se il campione è sufficientemente numeroso, si ha, nell'ipotesi $H_0 : p = p_0$ (se l'ipotesi non è semplice, si dimostra in modo più complicato, ma si ottiene lo stesso test),

$$\frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \simeq Z^* \sim \mathrm{N}(0,1),$$

pertanto calcoleremo questa quantità in base ai dati del campione e la confronteremo con l'opportuno quantile della legge normale standard. Per la verifica delle condizioni di applicabilità dell'approssimazione normale ricordiamo che deve essere

$$n\overline{x}_n > 5 \quad \text{e} \quad n(1-\overline{x}_n) > 5.$$

Si ottiene il test seguente.

Riassumendo. Supponiamo di voler eseguire un test sulla frequenza di una popolazione bernoulliana, estraendo un campione casuale di ampiezza n. Se poniamo

$$z = \frac{\overline{x}_n - p_0}{\sqrt{p_0(1-p_0)/n}},$$

possiamo esprimere la regola di decisione del test, in dipendenza dall'ipotesi nulla e dal livello di significatività che abbiamo scelto, nel modo seguente:

H_0	H_1	Rifiuto H_0 se
$p = p_0$	$p \neq p_0$	$\lvert z\rvert > z_{1-\alpha/2}$
$p = p_0$ $(p \leq p_0)$	$p > p_0$	$z > z_{1-\alpha}$
$p = p_0$ $(p \geq p_0)$	$p < p_0$	$z < -z_{1-\alpha}$

Esempio 10.4.1. Una partita di pezzi viene ritenuta inaccettabile se contiene (almeno) l'8% di pezzi difettosi. Per decidere se accettare o no il lotto, si esamina un campione di 100 pezzi. Se tra questi si trovano 9 pezzi difettosi cosa si può dire?

Si tratta qui di eseguire un test sul parametro p di una popolazione bernoulliana. L'ipotesi nulla è $H_0 : p \leq 0.08$ e quindi $H_1 : p > 0.08$ dal punto di vista del produttore mentre è $H_0 : p \geq 0.08$ e quindi $H_1 : p < 0.08$ dal punto di vista dell'acquirente. Si rifletta su questo fatto!

Poniamoci dal punto di vista del produttore ed eseguiamo un test al livello del 5%. Osserviamo che il campione è abbastanza numeroso da consentire l'uso dell'approssimazione normale, infatti:

$$n\overline{x}_n = 9 > 5, \quad n(1 - \overline{x}_n) = 91 > 5.$$

Le ipotesi sono, come osservato, $H_0 : p \leq 0.08$ e $H_1 : p > 0.08$, pertanto la regione critica è della forma $\{\overline{X}_n > k\}$. Calcoliamo

$$z = \frac{\overline{x}_n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.09 - 0.08}{\sqrt{0.08 \cdot 0.92/100}} = 0.37.$$

Poichè $z_{.95} = 1.64$ e $z < z_{.95}$ l'ipotesi nulla non è rigettata ed il lotto non può essere rigettato.

10.5 Test sulla differenza di due medie di popolazioni normali

Consideriamo ora un problema leggermente diverso dai precedenti. Supponiamo di voler confrontare le medie di due popolazioni diverse, estraendo un campione casuale da ciascuna di esse. Questa situazione si può verificare in molte *indagini comparative* : si vuole confrontare la produttività di una macchina con quella di un'altra; si vuole sapere se la popolazione di una città ha reddito medio superiore a quello di un'altra città, ecc.

Cominciamo a considerare il caso di **due popolazioni normali indipendenti,** $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. Supponiamo di estrarre da ciascuna popolazione un campione casuale; i due campioni possono anche non avere la stessa ampiezza:

$$(X_1, \ldots, X_n), \quad (Y_1, \ldots, Y_m).$$

Vogliamo confrontare le medie delle due popolazioni. L'ipotesi nulla di solito è una delle seguenti.

$$H_0 : \mu_X = \mu_Y, \quad H_0 : \mu_X \geq \mu_Y, \quad H_0 : \mu_X \leq \mu_Y.$$

Più in generale, potrebbe essere del tipo:

$$H_0 : \mu_X = \mu_Y + a, \quad H_0 : \mu_X \geq \mu_Y + a, \quad H_0 : \mu_X \leq \mu_Y + a,$$

con $a \in \mathbb{R}$ costante fissata. Poiché lo stimatore naturale di una media è la media campionaria, è ragionevole considerare la statistica $\bar{X}_n - \bar{Y}_m$ per stimare la differenza $\mu_X - \mu_Y$. Con considerazioni analoghe a a quelle fatte per i test su una sola media, ad esempio, se $H_0 : \mu_X = \mu_Y + a$, il test sarà del tipo "si rifiuti H_0 se $|\bar{X}_n - \bar{Y}_m - a| > k$"; se invece $H_0 : \mu_X \geq \mu_Y + a$, il test sarà del tipo "si rifiuti H_0 se $\bar{X}_n - \bar{Y}_m - a < k$" ed ancora se $H_0 : \mu_X \leq \mu_Y + a$, il test sarà del tipo "si rifiuti H_0 se $\bar{X}_n - \bar{Y}_m - a > k$". Il problema, come al solito, è costruire una variabile aleatoria di legge nota a partire da $\bar{X}_n - \bar{Y}_m$. Si noti che non abbiamo ancora

fatto alcuna ipotesi sulle varianze delle due popolazioni: abbiamo quindi quattro parametri incogniti. Tratteremo solo due situazioni particolari.

a. Le varianze σ_X^2, σ_Y^2 sono note;

b. le varianze σ_X^2, σ_Y^2 sono incognite, ma uguali fra loro.

Torneremo poi sul significato di queste ipotesi.

10.5.1 Varianze note

Ragioniamo, per fissare le idee, nel caso in cui l'ipotesi nulla sia $H_0 : \mu_X = \mu_Y + a$. Sappiamo che

$$\bar{X}_n \sim \mathrm{N}(\mu_X, \sigma_X^2/n), \quad \bar{Y}_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2/m),$$

da cui se è vera l'ipotesi nulla si può dimostrare che risulta

$$\bar{X}_n - \bar{Y}_m - a \sim \mathrm{N}\Big(0, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\Big),$$

ovvero

$$\frac{\bar{X}_n - \bar{Y}_m - a}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathrm{N}(0,1).$$

Fissiamo ora il livello α della regione di rifiuto. Dobbiamo fare in modo che l'errore di prima specie sia α. Ma ricordiamo che l'errore di prima specie è la probabilità di rifiutare H_0 quando H_0 è vera. Quindi è la probabilità della regione critica, quando H_0 è vera. Quindi deve essere,

$$\mathbb{P}(|\bar{X}_n - \bar{Y}_m - a| > k) = \alpha.$$

Come fare? Ricordiamo che, nel nostro caso, se l'ipotesi è vera,

$$\frac{\bar{X}_n - \bar{Y}_m - a}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = Z^* \sim \mathrm{N}(0,1).$$

Quindi,

$$\begin{aligned}
\mathbb{P}(|\bar{X}_n - \bar{Y}_m - a| > k) &= \mathbb{P}\left(\frac{|\bar{X}_n - \bar{Y}_m - a|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} > \frac{k}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right) \\
&= \mathbb{P}\left(|Z^*| > \frac{k}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right) = \alpha.
\end{aligned}$$

Perché ciò sia vero deve essere (per i dettagli si veda il caso del test su una media),

$$\frac{k}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = z_{1-\alpha/2},$$

pertanto si rifiuta H_0 se

$$z^* = \frac{|\bar{x}_n - \bar{y}_m - a|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} > z_{1-\alpha/2}.$$

In modo analogo ri ricavano le ragioni critiche quando H_0 ha un'altra forma.
Riassumendo. Supponiamo di voler eseguire un test sulla differenza delle medie di due popolazione normali di varianze σ_X^2 e σ_Y^2 note, estraendo un campione casuale di ampiezza n dalla prima e un campione di ampiezza m dalla seconda. Se poniamo

$$z^* = \frac{|\bar{x}_n - \bar{y}_m - a|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}},$$

possiamo esprimere la regola di decisione del test, in dipendenza dall'ipotesi nulla e dal livello di significatività che abbiamo scelto, nel modo seguente:

H_0	H_1	Rifiuto H_0 se
$\mu_X = \mu_Y + a$	$\mu_X \neq \mu_Y + a$	$\lvert z^* \rvert > z_{1-\alpha/2}$
$\mu_X \leq \mu_Y + a$	$\mu_X > \mu_Y + a$	$z^* > z_{1-\alpha}$
$\mu_X \geq \mu_Y + a$	$\mu_X < \mu_Y + a$	$z^* < -z_{1-\alpha}$

10.5.2 Varianza incognite ma uguali

Stiamo ora supponendo $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (ma σ incognito). Possiamo ancora dire che se è vera l'ipotesi $H_0 : \mu_X = \mu_Y + a$,

$$\frac{\bar{X}_n - \bar{Y}_m - a}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} = \frac{\bar{X}_n - \bar{Y}_m - a}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathrm{N}(0,1). \tag{10.1}$$

Ora però la quantità non può essere utilizzata perché *non* calcolabile (contiene l'incognita σ)! Occorre stimarla. Utilizziamo le varianze campionarie dei due campioni.

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2.$$

Definiamo la **varianza campionaria pesata**

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Questa è la quantità più naturale da introdurre per stimare σ^2. Tuttavia la cosa importante che ci permette di poter costruire un test è che sostituendo S al posto di σ nella (10.1) si ottiene una variabile aleatoria che ha ancora una legge nota. Precisamente

$$\frac{\bar{X}_n - \bar{Y}_m - a}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}} \sim t(n+m-2).$$

Con un ragionamento analogo a quello fatto in tutti gli altri casi possiamo allora costruire il test.
Riassumendo. Supponiamo di voler eseguire un incognite, estraendo un campione casuale di ampiezza n dalla prima e un campione di ampiezza m dalla seconda. Se poniamo

$$t = \frac{\bar{x}_n - \bar{y}_m - a}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}},$$

possiamo esprimere la regola di decisione del test, in dipendenza dall'ipotesi nulla e dal livello di significatività che abbiamo scelto, nel modo seguente:

H_0	H_1	Rifiuto H_0 se
$\mu_X = \mu_Y + a$	$\mu_X \neq \mu_Y + a$	$\lvert t\rvert > t_{1-\alpha/2}(n+m-2)$
$\mu_X \leq \mu_Y + a$	$\mu_X > \mu_Y + a$	$t > t_{1-\alpha}(n+m-2)$
$\mu_X \geq \mu_Y + a$	$\mu_X < \mu_Y + a$	$t < -t_{1-\alpha}(n+m-2)$

Osservazione 10.5.1. *Abbiamo visto fin qui, come si esegue un test sulla differenza di medie di due popolazioni normali indipendenti in due casi particolari. Varianze note o varianze incognite, ma uguali. Vediamo ora di renderci conto, brevemente, di quale sia il significato concreto di queste ipotesi.*

a. Oltre al caso in cui le varianze siano effettivamente note (per esempio per studi precedenti) si tratta in questo modo anche il caso in cui i campioni sono entrambi molto numerosi ($n, m > 30$). In questo caso è prassi trattare le varianze campionarie come fossero le varianze effettive.

b. Questa seconda ipotesi può essere sensata quando le due popolazioni in esame sono, in qualche senso, entrambe parte di una popolazione più vasta, per cui possiamo pensare che condividano la varianza della popolazione più vasta, pur avendo magari medie diverse. Ad esempio se X e Y rappresentano la statura di un individuo adulto scelto a caso in due regioni differenti. Potranno avere medie differenti, ma è ragionevole supporre che abbiamo stessa varianza. Praticamente si considerano varianze uguali quando il rapporto S_X^2/S_Y^2 tra le varianze campionarie è *abbastanza* vicino a 1.

Esempio 10.5.2. L'osservazione dei guasti occorsi a due tipi di macchine fotocopiatrici ha mostrato che: 71 guasti della macchina A hanno richiesto un tempo medio di riparazione di 83.2 minuti, con una deviazione standard di 19.3 minuti, mentre 75 guasti della macchina B hanno richiesto un tempo medio di riparazione di 90.8 minuti, con una deviazione standard di 21.4 minuti. Si esegua un test, al livello del 5% sull'ipotesi di uguaglianza dei tempi medi di riparazione. Si supponga che i tempi di riparazione abbiano legge normale.

I dati a nostra disposizione sono:

$$\bar{x}_n = 83.2; \quad s_X = 19.3; \quad n = 71;$$
$$\bar{y}_n = 90.8; \quad s_Y = 21.4; \quad m = 75.$$

Possiamo procedere in due modi. Poiché i campioni sono numerosi, si possono considerare le varianza note (sostituendo a σ_X la sua stima s_X, e analogamente per Y) e applicare il test descritto nel Paragrafo 10.5.1 con $a = 0$. In questo caso, calcoliamo

$$z^* = \frac{|\bar{x}_n - \bar{y}_m|}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} = \frac{83.2 - 90.8}{\sqrt{\frac{19.3^2}{71} + \frac{21.4^2}{75}}} = -22556.$$

Il test al livello del 5% è: si rifiuti H_0 se $|z^*| > z_{0.975} = 1.96$. Dato che 2.2556>1.96, si può rifiutare l'ipotesi.

Il secondo modo di procedere è questo. Poiché le varianze campionarie hanno rapporto $S_Y^2/S_X^2 = 21.4^2/19.3^2 = 1.23$, abbastanza vicino a 1, possiamo ritenere le varianze incognite

ma uguali ed applicare il test descritto nel Paragrafo 10.5.2. In questo caso calcoliamo

$$t = \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}}$$
$$= \frac{83.2 - 90.8}{\sqrt{\frac{1}{71}+\frac{1}{75}}\sqrt{\frac{7019.3^2+7421.4^2}{71+75-2}}} = -2.249.$$

Il test al livello del 5% è: si rifiuti H_0 se $|t| > t_{0.975}(144) \sim z_{0.975} = 1.96$ (**RICORDA!** dato che i gradi di libertà della t sono più di 120, i quantili della t si approssimano con quelli della gaussiana). Dato che 2.2249>1.96, si può rifiutare l'ipotesi.

10.6 Il test chi quadro (χ^2)

10.6.1 Il test chi quadro di adattamento

Ci occupiamo ora di un'importante procedura statistica che ha lo scopo di verificare se certi dati empirici si adattino bene ad una distribuzione teorica assegnata. Il significato di questo problema sarà illustrato dai prossimi esempi che costituiranno la guida del discorso.

Esempio 10.6.1. Negli esperimenti di Mendel con i piselli si rilevarono i dati seguenti.

Tipologia	N^o di casi osservati
Lisci-gialli	315
Lisci-verdi	108
Rugosi-gialli	101
Rugosi-verdi	32

Secondo la sua teoria sull'ereditarietà, i numeri avrebbero dovuto essere nella proporzione 9:3:3:1. Esiste qualche ragione di dubitare della sua teoria?

Esempio 10.6.2. In base ad una ricerca condotta due anni fa, si può ritenere che il numero di incidenti automobilistici per settimana, in un certo tratto di autostrada, segua una legge di Poisson di parametro $\lambda = 0.4$. Se nelle ultime 85 settimane si sono rilevati i seguenti dati

N^o incidenti per settimana	0	1	2	3 o più	Totale
N^o di settimane in cui si è verificato	50	32	3	0	85

si può affermare che il modello sia ancora applicabile alla descrizione del fenomeno, o qualcosa è cambiato?

Esempio 10.6.3. I tempi di vita di 100 lampadine estratte casualmente da un lotto sono stati misurati, e i dati raggruppati come segue.

Tempo di vita (in mesi)	N^o di lampadine
meno di 1	24
da 1 a 2	16
da 2 a 3	20
da 3 a 4	14
da 4 a 5	10
da 5 a 10	16
più di 10	0
Totale	100

In base questi dati si può ritenere che il tempo di vita segua una legge esponenziale di parametro $\lambda = 0.33$?

Per arrivare a rispondere a questi problemi, cominciamo a descrivere la situazione generale di cui quelle precedenti sono esemplificazioni concreta. Supponiamo di avere una tabella che rappresenta n osservazioni di una variabile raggruppate in k classi (qui k *deve* essere finito!!). Le classi possono rappresentare:

a) caratteristiche qualitative (piselli lisci-verdi, lisci-gialli, ecc);

b) valori assunti da una variabile discreta (ogni classe un singolo valore, oppure una classe raggruppa le code, ecc);

c) intervalli di valori assunti da una variabile continua.

Per ciascuna classe A_i, $i = 1, 2, \ldots, k$ supponiamo di avere oltre la *frequenza osservata* anche la *frequenza attesa* con cui vogliamo confrontare la frequenza osservata e dedurre se la discrepanza tra le due possa essere giustificata dal caso oppure debba essere attribuita ad un errore nel modello scelto. Torniamo ai nostri esempi e cerchiamo di capire in quei casi quali siano le frequenze attese.

Esempio 10.6.4. Riprendiamo qui la situazione vista nell'Esempio 10.6.1. Il numero totale dei piselli è 315+108+101+32=556. Poiché i numeri sono attesi nella proporzione 9:3:3:1 e 9+3+3+1=16, avremmo dovuto aspettarci
$\frac{9}{16} \cdot 556 = 312.75$, lisci-gialli;
$\frac{3}{16} \cdot 556 = 104.25$, lisci-verdi;
$\frac{3}{16} \cdot 556 = 104.25$, rugosi-gialli;
$\frac{1}{16} \cdot 556 = \ \ 34.75$, rugosi-verdi.
Riassumendo, si ha

Tipologia	N^o di casi osservati	N^o di casi aspettati
Lisci-gialli	315	312.75
Lisci-verdi	108	104.25
Rugosi-gialli	101	104.25
Rugosi-verdi	32	34.75

Esempio 10.6.5. Riprendiamo qui la situazione vista nell'Esempio 10.6.2. La distribuzione teorica con cui si vogliono confrontare i dati è la legge di Poisson di parametro 0.4. Più

precisamente se X è il numero degli incidenti per settimana, vogliamo vedere, se $X \sim \text{Po}(0.4)$. Se $X \sim \text{Po}(0.4)$, si avrebbe:

$$\begin{aligned}
\mathbb{P}(X=0) &= \mathrm{e}^{-0.4} = 0.670 \\
\mathbb{P}(X=1) &= 0.4\mathrm{e}^{-0.4} = 0.268 \\
\mathbb{P}(X=2) &= \frac{0.4^2}{2}\mathrm{e}^{-0.4} = 0.054 \\
\mathbb{P}(X\geq 3) &= 1-(0.670+0.268+0.054) = 0.008.
\end{aligned}$$

Pertanto, in 85 settimane, avremmo dovuto aspettarci
$0.670 \cdot 85 = 56.95$ settimane in cui $X = 0$;
$0.268 \cdot 85 = 22.78$ settimane in cui $X = 1$;
$0.054 \cdot 85 = \ \ 4.59$ settimane in cui $X = 2$;
$0.008 \cdot 85 = \ \ 0.68$ settimane in cui $X \geq 3$.
Riassumendo, si ha

N^o incidenti per settimana	0	1	2	3 o più	Totale
N^o di settimane in cui si è verificato	50	32	3	0	85
N^o di settimane atteso	56.95	22.78	4.59	0.68	85

Esempio 10.6.6. Riprendiamo qui la situazione vista nell'Esempio 10.6.3. La distribuzione teorica con cui si vogliono confrontare i dati è la legge esponenziale di parametro 0.33. Se $X \sim \text{Exp}(0.33)$, si avrebbe:

$$\begin{aligned}
\mathbb{P}(0<X\leq 1) &= \int_0^1 \lambda\mathrm{e}^{-\lambda x}\,dx = 1-\mathrm{e}^{-0.33} = 0.2811 \\
\mathbb{P}(1<X\leq 2) &= \int_1^2 \lambda\mathrm{e}^{-\lambda x}\,dx = \mathrm{e}^{-0.33}-\mathrm{e}^{-0.33\cdot 2} = 0.2021 \\
\mathbb{P}(2<X\leq 3) &= \int_2^3 \lambda\mathrm{e}^{-\lambda x}\,dx = \mathrm{e}^{-0.33\cdot 2}-\mathrm{e}^{-0.33\cdot 3} = 0.1453 \\
\mathbb{P}(3<X\leq 4) &= \int_3^4 \lambda\mathrm{e}^{-\lambda x}\,dx = \mathrm{e}^{-0.33\cdot 3}-\mathrm{e}^{-0.33\cdot 4} = 0.1044 \\
\mathbb{P}(4<X\leq 5) &= \int_4^5 \lambda\mathrm{e}^{-\lambda x}\,dx = \mathrm{e}^{-0.33\cdot 4}-\mathrm{e}^{-0.33\cdot 5} = 0.0.0751 \\
\mathbb{P}(5<X\leq 10) &= \int_5^{10} \lambda\mathrm{e}^{-\lambda x}\,dx = \mathrm{e}^{-0.33\cdot 5}-\mathrm{e}^{-0.33\cdot 10} = 0.1552 \\
\mathbb{P}(X>10) &= \int_{10}^{+\infty} \lambda\mathrm{e}^{-\lambda x}\,dx = \mathrm{e}^{-0.33\cdot 10} = 0.0369
\end{aligned}$$

Pertanto, su 100 lampadine, avremmo dovuto aspettarci
$0.2811 \cdot 100 = 28.11$ lampade per cui $0 < X \leq 1$;
$0.2021 \cdot 100 = 20.21$ lampade per cui $1 < X \leq 2$;
$0.1453 \cdot 100 = 14.53$ lampade per cui $2 < X \leq 3$;
$0.1044 \cdot 100 = 10.44$ lampade per cui $3 < X \leq 4$;

$0.0751 \cdot 100 = \ 7.51$ lampade per cui $4 < X \leq 5$;
$0.1552 \cdot 100 = 15.52$ lampade per cui $5 < X \leq 10$;
$0.0369 \cdot 100 = \ 3.69$ lampade per cui $X > 10$.

Tempo di vita (in mesi)	N^o di lampadine	N^o di lampadine atteso
meno di 1	24	28.11
da 1 a 2	16	20.21
da 2 a 3	20	14.53
da 3 a 4	14	10.44
da 4 a 5	10	7.51
da 5 a 10	16	15.52
più di 10	0	3.69
Totale	100	100.00

Veniamo ora al punto fondamentale: come valutare la bontà dell'adattamento delle frequenze assolute osservate alle frequenza assolute attese?

Supponiamo di avere in generale, n osservazioni raggruppate in k classi, $A_1, A_2, \ldots, A_k$; siano p_i le frequenze relative attese di ciascuna classe ($p_1 + p_2 + \ldots + p_k = 1$) e quindi $np_1, np_2, \ldots, np_k$ le frequenze assolute attese. Siano poi $N_1, N_2, \ldots, N_k$ le frequenze assolute osservate. Calcoliamo in base a questi dati la seguente statistica:

$$Q = \sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i}. \tag{10.2}$$

Si osservi che ogni addendo di Q ha a numeratore lo scarto quadratico tra le frequenza attesa e quella osservata, e a denominatore la frequenza attesa, che fa "pesare" diversamente i vari addendi. La Q sarà tanto più piccola quanto migliore è l'adattamento delle frequenze osservate a quelle attese. Inoltre la discrepanza tra frequenze osservate e attese è pesata più o meno a seconda della frequenza attesa. A parità di discrepanza pesa di più quella relativa a frequenze attese più piccole. La quantità Q è pertanto una *buona* statistica per valutare l'adattamento. Se l'ipotesi nulla è,

H_0 : le osservazioni si adattano ai dati teorici,

il test sarà del tipo

"Si rifiuti H_0 se $Q > k$" con k opportuno.

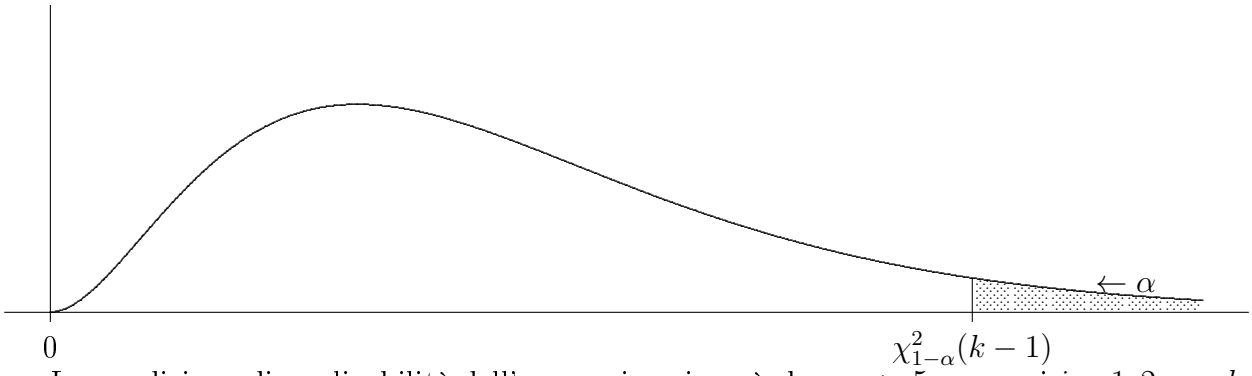
Il risultato fondamentale che permette di determinare il k opportuno è dato dal fatto che se n è grande allora la statistica Q ha una distribuzione che tende ad una legge $\chi^2(k-1)$, ovvero

$$\sum_{i=1}^{k} \frac{(np_i - N_i)^2}{np_i} \simeq W \sim \chi^2(k-1).$$

Questo, come già visto in precedenza permette di calcolare in modo completo la regione di rifiuto. Se vogliamo un test al livello α, la regione di rifiuto è

$$\{Q > \chi^2_{1-\alpha}(k-1)\}.$$

Graficamente,



La condizione di applicabilità dell'approssimazione è che $np_i > 5$ per ogni $i = 1, 2, \ldots, k$.
Torniamo agli esempi fatti sinora e vediamo che conclusione possiamo trarre.

Esempio 10.6.7. Nel caso esposto nell'Esempio 10.6.4 le frequenze attese sono tutte maggiori di 5, pertanto possiamo procedere al calcolo della quantità Q.

$$Q = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.470.$$

Poiché ci sono 4 modalità il numero dei gradi di libertà è 3.
Ora,
$\chi^2_{.99}(3) = 11.3$, così che non possiamo rifiutare la teoria al livello dello 0.01;
$\chi^2_{.95}(3) = 7.81$, così che non possiamo rifiutare la teoria al livello dello 0.05.
Concludiamo che la teoria concorda con l'esperimento.

Esempio 10.6.8. Guardando la tabella che compare nell'Esempio 10.6.5 si osserva che le ultime due classi hanno frequenze attese <5, perciò non possiamo utilizzare la tabella così com'è per effettuare il test; se uniamo le ultime due classi otteniamo una nuova classe $\{X \geq 2\}$ con frequenza attesa 4.59+0.68=5.27, e frequenza osservata pari a 3+0=3 (si ricordi che è solo la frequenza attesa che deve essere ≥ 5). Perciò consideriamo la nuova tabella:

N^o incidenti per settimana	0	1	2 o più	Totale
N^o di settimane in cui si è verificato	50	32	3	85
N^o di settimane atteso	56.95	22.78	5.27	85

Il valore della statistica Q è

$$Q = \frac{(50 - 56.95)^2}{56.95} + \frac{(32 - 22.78)^2}{22.78} + \frac{(3 - 5.27)^2}{5.27} = 5.56.$$

Poiché ci sono 3 modalità il numero dei gradi di libertà è 2.
Ora,
$\chi^2_{.99}(2) = 9.21$, cosìche non possiamo rifiutare l'ipotesi al livello dello 0.01;
$\chi^2_{.95}(2) = 5.99$, cosìche non possiamo rifiutare l'ipotesi al livello dello 0.05.
Concludiamo che non abbiamo reali motivi per credere che le cose siano cambiate.

Esempio 10.6.9. Guardando la tabella che compare nell'Esempio 10.6.6 si osserva che per poter effettuare il test basta unire le ultime due classi. Introduciamo la nuova classe $X > 5$ con frequenza attesa 15.52+3.69=19.21 e frequenza osservata 16. Abbiamo

Tempo di vita (in mesi)	N^o di lampadine	N^o di lampadine atteso
meno di 1	24	28.11
da 1 a 2	16	20.21
da 2 a 3	20	14.53
da 3 a 4	14	10.44
da 4 a 5	10	7.51
più di 5	16	19.21
Totale	100	100.00

Il valore della statistica Q è

$$Q = \frac{(24-28.11)^2}{28.11} + \ldots + \frac{(10-7.51)^2}{7.51} + \frac{(16-19.21)^2}{19.21} = 6.11.$$

Poiché ci sono 6 modalità il numero dei gradi di libertà è 5.
Ora,

$\chi^2_{.99}(5) = 15.09$, così che non possiamo rifiutare l'ipotesi al livello dello 0.01;
$\chi^2_{.95}(5) = 11.07$, così che non possiamo rifiutare l'ipotesi al livello dello 0.05.

Concludiamo che i dati statistici quindi confermano che il tempo di vita delle lampadine segue effettivamente una legge Exp(0.33).

10.6.2 Il test chi quadro di indipendenza

Il test chi-quadro può essere utilizzato anche per verificare l'indipendenza o meno di due variabili. È questo un altro problema che si presenta spesso nelle applicazioni. Disponiamo di n osservazioni **congiunte** di due variabili e ci chiediamo: esiste una qualche dipendenza tra le variabili o no? Nel Capitolo 2, abbiamo visto come si possa valutare la correlazione di due variabili numeriche: utilizzo di scatterplot, coefficiente di correlazione, retta dei minimi quadrati... Ora vedremo un metodo diverso che permette di trattare sia variabili numeriche che variabili categoriche, valutando quantitativamente l'indipendenza (o la dipendenza) di queste. Al solito, introduciamo il problema con alcuni esempi.

Esempio 10.6.10. Un certo corso universitario è impartito a studenti del terzo anno di tre diversi indirizzi. Gli studenti frequentano le medesime lezioni di un professore che registra il numero di studenti di ogni indirizzo che hanno superato l'esame. I dati sono i seguenti:

	Indirizzo A	Indirizzo B	Indirizzo C	Tot. esami
esame superato	30	15	50	95
esame non superato	40	8	37	85
Tot Studenti	70	23	87	180

Il rendimento degli studenti, relativamente all'esame in questione, si può ritenere sostanzialmente equivalente, oppure le differenze sono statisticamente significative? Questo equivale a chiedersi se le due variabili (categoriche) "indirizzo" e "rendimento" sono tra loro indipendenti o no.

Esempio 10.6.11. Sono state effettuate delle prove di resistenza su pneumatici di 4 diverse marche, e si è registrata la durata X, di questi pneumatici (in chilometri percorsi prima dell'usura). I dati sono i seguenti:

	$X \leq 30000$	$30000 < X \leq 45000$	$X > 45000$	Tot.
Marca A	26	118	56	200
Marca B	23	93	84	200
Marca C	15	116	69	200
Marca D	32	121	47	200
Tot	96	448	256	800

ci chiediamo se le 4 marche si possano ritenere equivalenti, quanto alla durata degli pneumatici, oppure no. In altre parole, questo equivale a chiedersi se la variabile numerica "durata" sia indipendente o no dalla variabile categorica "marca".

Cominciamo ad introdurre un po' di terminologia e notazioni. Una tabella come quelle riportate nei due esempi precedenti si chiama **tabella di contingenza.** In una tabella di questo tipo n osservazioni sono classificate secondo un certo criterio X in r classi $A_1, A_2, \ldots, A_r$ e, contemporaneamente sono classificate secondo un altro criterio Y in s classi $B_1, B_2, \ldots, B_s$. Ogni osservazione appartiene così ad una ed una sola classe A_i e ad una e una sola classe B_j. L'insieme delle n osservazioni è così ripartito in $r \cdot s$ classi $(X \in A_i, Y \in B_j)$. La tabella riporta all'incrocio della colonna A_i con la riga B_j la frequenza n_{ij} della classe $(X \in A_i, Y \in B_j)$. Si calcolano poi i totali di riga e di colonna e si ottiene la tabella seguente:

	A_1	A_2	$\ldots$	A_r	Tot.
B_1	n_{11}	n_{12}	$\ldots$	n_{1r}	$n_{1.}$
B_2	n_{21}	n_{22}	$\ldots$	n_{2r}	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
B_s	n_{s1}	n_{s2}	$\ldots$	n_{sr}	$n_{s.}$
Tot	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.r}$	n

Si osservi che abbiamo indicato con $n_{j.}$ il totale della riga j, quindi la frequenza della classe B_j, per $j = 1, 2, \ldots, s$ e con $n_{i.}$ il totale della colonna i, quindi la frequenza della classe A_i, per $i = 1, 2, \ldots, r$.

Vogliamo trovare una regola per testare l'ipotesi

H_0 : Le variabili sono indipendenti.

Come si deve procedere? Quali dovrebbero essere le frequenze in ipotesi di indipendenza?

Poiché i valori di X sono raggruppati nelle classi A_i e i valori di Y nelle classi B_j, se X e Y sono indipendenti, per definizione di variabili indipendenti, deve essere

$$\mathbb{P}(X \in A_i, Y \in B_j) = \mathbb{P}(X \in A_i)\mathbb{P}(Y \in B_j).$$

Se stimiamo $\mathbb{P}(A_i)$ e $\mathbb{P}(B_j)$ con le frequenze relative di ciascuna classe, ovvero

$$\mathbb{P}(A_i) = \frac{n_{.i}}{n} \quad \text{e} \quad \mathbb{P}(B_j) = \frac{n_{j.}}{n},$$

allora la frequenza relativa attesa di $(X \in A_i, Y \in B_j)$, cioè quella che si avrebbe in ipotesi di indipendenza, è

$$\hat{p}_{ij} = \frac{n_{.i}}{n}\frac{n_{j.}}{n},$$

quindi la frequenza della classe attesa è

$$n\hat{p}_{ij} = \frac{n_{.i}n_{j.}}{n}.$$

La situazione si può ora descrivere in termini simili a quelli usati per il test χ^2 di adattamento: abbiamo $k = rs$ classi; di ogni classe conosciamo la osservata n_{ij} e la frequenza attesa frequenza $\frac{n_{.i}n_{j.}}{n}$. Indipendenza delle variabili X e Y significa allora adattamento delle frequenze osservate alle frequenze relative attese (che sono quelle calcolate in ipotesi di indipendenza). Costruiamo anche qui la statistica Q:

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{.i}n_{j.}}{n}\right)^2}{\frac{n_{.i}n_{j.}}{n}}.$$

Il test sull'ipotesi di indipendenza sarà del tipo: rifiutare l'ipotesi se $Q > k$, con k opportuno da calcolare in base al livello.

Il risultato fondamentale che permette di determinare il k opportuno è dato dal fatto che se n è grande allora la statistica Q ha una distribuzione che tende ad una legge $\chi^2(r-1)(s-1)$, ovvero

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{.i}n_{j.}}{n}\right)^2}{\frac{n_{.i}n_{j.}}{n}} \simeq W \sim \chi^2((r-1)(s-1)).$$

Questo, come già visto in precedenza permette di calcolare in modo completo la regione di rifiuto. Se vogliamo un test al livello α, la regione di rifiuto è

$$\{Q > \chi^2_{1-\alpha}((r-1)(s-1))\}.$$

Anche in questo caso l'approssimazione si può utilizzare se le frequenze attese sono maggiori di 5 ovvero $n\hat{p}_{ij} = \frac{n_{.i}n_{j.}}{n} > 5.$ per ogni $i = 1, \ldots, r$, $J = 1 \ldots, s$.

Illustriamo ora il procedimento sui due esempi visti in precedenza.

Esempio 10.6.12. Riprendiamo la tabella di contingenza dell'Esempio 10.6.10.

	Indirizzo A	Indirizzo B	Indirizzo C	Tot. esami
esame superato	30	15	50	95
esame non superato	40	8	37	85
Tot Studenti	70	23	87	180

Costruiamo a partire da questa la tabella con le frequenze attese in ipotesi di indipendenza. Si ha,

	Indirizzo A	Indirizzo B	Indirizzo C	Tot. esami
esame superato	95· 70/180	95· 23/180	95· 87/180	95
esame non superato	85· 70/180	85· 23/180	85· 87/180	85
Tot Studenti	70	23	87	180

ovvero

	Indirizzo A	Indirizzo B	Indirizzo C	Tot. esami
esame superato	36.94	12.14	45.92	95
esame non superato	33.06	10.86	41.08	85
Tot Studenti	70	23	87	180

Osserviamo che tutti i numeri che compaiono in quest'ultima tabella sono maggiori di 5. Ciò permette di poter applicare l'approssimazione con il chi-quadro. Calcoliamo Q.

$$Q = \frac{(30-36.94)^2}{36.94} + \frac{(15-12.14)^2}{12.14} + \frac{(50-45.92)^2}{45.92} + \\ + \frac{(40-33.06)^2}{33.06} + \frac{(8-10.86)^2}{10.86} + \frac{(37-41.08)^2}{41.08} = 4.96.$$

I gradi di libertà sono (3-1)(2-1)=2. Il test al livello del 5% è: si rifiuti l'ipotesi di indipendenza se $Q > \chi^2_{.95}(2) = 5.991$. Perciò al livello del 5% i dati *non* consentono di rifiutare l'ipotesi di indipendenza.

Esempio 10.6.13. Riprendiamo la tabella di contingenza dell'Esempio 10.6.11.

	$X \leq 30000$	$30000 < X \leq 45000$	$X > 45000$	Tot.
Marca A	26	118	56	200
Marca B	23	93	84	200
Marca C	15	116	69	200
Marca D	32	121	47	200
Tot	96	448	256	800

Costruiamo a partire da questa la tabella con le frequenze attese in ipotesi di indipendenza. In questo caso i calcoli sono semplificati dal fatto che i 4 totali di riga sono tutti uguali tra loro: perciò sulla prima colonna compare sempre 96· 200/800=24; sulla seconda colonna sempre 448· 200/800=112; sulla terza colonna sempre 256· 200/800=64.

	$X \leq 30000$	$30000 < X \leq 45000$	$X > 45000$	Tot.
Marca A	24	112	64	200
Marca B	24	112	64	200
Marca C	24	112	64	200
Marca D	24	112	64	200
Tot	96	448	256	800

Osserviamo che tutti i numeri che compaiono in quest'ultima tabella sono maggiori di 5. Ciò permette di poter applicare l'approssimazione con il chi-quadro. Calcoliamo Q.

$$Q = \frac{(26-24)^2}{24} + \frac{(23-24)^2}{24} + \frac{(15-24)^2}{24} + \frac{(32-24)^2}{24} + \\ + \frac{(118-112)^2}{112} + \frac{(93-112)^2}{112} + \frac{(116-112)^2}{112} + \frac{(121-112)^2}{112} + \\ + \frac{(56-64)^2}{64} + \frac{(84-64)^2}{64} + \frac{(69-64)^2}{64} + \frac{(47-64)^2}{64} = 22.82.$$

I gradi di libertà sono (3-1)(4-1)=6. Il test al livello del 5% è: si rifiuti l'ipotesi di indipendenza se $Q > \chi^2_{.95}(6) = 12.592$. Perciò al livello del 5% i dati consentono di rifiutare l'ipotesi di indipendenza.