Appunti di Probabilità e Statistica

A.A. 2009/10

Fabio Zucca
e-mail: fabio.zucca@polimi.it

Dispense a cura di Lorenzo Valdettaro e Fabio Zucca

Libri consigliati:

1. D.C. Montgomery, G.C. Runger e N.F. Hubele: Engineering Statistics. Ed. John Wiley & Sons.
2. D.C. Montgomery, G.C. Runger e N.F. Hubele: Statistica per Ingegneria (a cura di A. Barchielli e M. Verri). Ed. EGEA.
3. A.M. Mood, F.A. Graybill e D.C. Boes: Introduzione alla statistica. Ed. McGraw-Hill
4. G. Cicchitelli: Probabilità e statistica. Maggioli Editore.
5. G. Cicchitelli: Complementi ed esercizi di statistica descrittiva ed inferenziale. Maggioli Editore. Ingegneria. Ed. Esculapio

Indice

Cap. 1. Introduzione

Scopo delle scienze fisiche e naturali in genere è quello di dare una descrizione in termini matematici (leggi) dei fenomeni naturali e, se possibile di effettuare previsioni attendibili. In alcuni casi la complessità del fenomeno o la sua natura caotica o, ancora, la sua natura *quantistica*, trascendono la nostra capacità di descrizione in termini deterministici.

Con la teoria della probabilità si cercano di fornire gli strumenti adatti al controllo di ciò che non riusciamo a descrivere completamente. Questo ha senso sia se la natura del fenomeno fosse completamente deterministica (si pensi alla natura caotica dei fenomeni atmosferici) che, a maggior ragione, se si ipotizza una *casualità intrinseca* (si pensi ai fenomeni descritti dalla meccanica quantistica).

Le leggi si riducono a relazioni miste del tipo

$$y = f(x, \epsilon)$$

dove x ed y possono essere pensate in spazi multidimensionali ed ϵ rappresenta la parte casuale.

Con la statistica si stimano le grandezze necessarie alla descrizione casuale del fenomeno. Più in dettaglio, lo scopo della statistica matematica è lo studio di popolazioni in senso lato.

Esempi: presentazione di risultati elettorali, proiezioni di risultati elettorali, indagini statistiche, distribuzione degli errori nella produzione di dispositivi meccanici, ecc.

I dati devono essere **raccolti**, **presentati**, **analizzati**, **interpretati**.

Due approcci fondamentali: la **statistica descrittiva** e la **statistica inferenziale**

Statistica descrittiva: si propone di

1. raccogliere e presentare i dati in forma sintetica, grafica e/o tabulare;

2. caratterizzare alcuni aspetti in modo sintetico: indici di posizione (es. valore medio), di dispersione (es. varianza), e di forma (es. simmetria);

3. studiare le relazioni tra i dati riguardanti variabili diverse.

Esempio: studio della altezza e del peso di una popolazione: grafico della distribuzione dei valori, media e varianza, relazione tra peso e altezza, ecc.

Statistica inferenziale: si cerca di far rientrare la collezione dei dati in categorie (distribuzioni) matematiche prestabilite. Si cerca di determinare le distribuzioni e i parametri che meglio si adattano ai dati: test di ipotesi e stima dei parametri. Si cerca quindi di costruire un modello per ottenere, in seguito, delle previsioni.

Esempio 1.0.1. Sondaggio a campione riguardo alle intenzioni di voto: quale conclusione trarre sull'insieme della popolazione?

Esempio 1.0.2. Si misurano i diametri di un campione di bulloni prodotti da una linea di produzione. Quale sarà il diametro medio e la variabilità nei diametri della produzione totale? Quanti bulloni risulteranno difettosi (diametri troppo larghi o troppo stretti)?

Si introduce il concetto di casualità: sondaggi diversi danno *probabilmente* risultati diversi. La probabilità *misura* il grado di attendibilità di un evento (matematicamente la probabilità sarà una *misura* nel senso dell'Analisi Matematica).

Il corso si articola in 3 parti:

1. Statistica descrittiva
2. Calcolo delle probabilità e variabili aleatorie
3. Statistica inferenziale

È utile correlare il materiale delle dispense con la lettura dei testi consigliati. Alcuni testi utilizzano notazioni o definizioni leggermente differenti, in tal caso per uniformità, ove siano presenti delle differenze ci riferiremo sempre a quelle delle presenti dispense.

Approfondimento

Alcune parti del testo saranno bordate in questo modo. Tali parti si possono intendere come approfondimenti *facoltativi*: chiunque voglia saperne di piu su questi argomenti è invitato a contattare gli autori.

Cap. 2. Statistica Descrittiva

Scopo: Introdurre gli strumenti basilari per l'analisi di un certo insieme di **dati**.

1. Raccogliere e di presentare i dati in forma sintetica, grafica e/o tabulare: istogrammi, diagrammi a barre, grafici di frequenza cumulativa, boxplots, scatterplots.
2. Caratterizzare alcuni aspetti in modo sintetico:
   (a) indici di posizione: valore medio, mediana, moda,
   (b) indici di dispersione: varianza, deviazione standard, quantile, quartile, differenza interquartile (IQR),
   (c) indice di forma: skewness, curtosi.
3. Studiare le relazioni tra i dati riguardanti variabili diverse: covarianza, coefficiente di correlazione, regressione lineare.

2.1 Tipi di Dati

Supponiamo di avere una successione di dati $\{x_i\}_{i=1}^n$ detto **campione**; n viene detta **ampiezza del campione**. Possiamo dividere i dati in due categorie principali:

1. Dati di tipo **numerico**
   (a) Variabili numeriche **discrete**, se la grandezza osservata appartiene, *a priori*, ad un insieme numerico finito o numerabile (ad esempio ad $\mathbb{N}$).
   (b) Variabili numeriche **continue**, se la grandezza osservata appartiene, *a priori*, ad un insieme non numerabile come, ad esempio $\mathbb{R}$ od un suo intervallo finito o meno.
2. Dati di tipo **categorico** se non sono numerici

Persona	Età	Altezza (metri)	Peso (Kg)	Genere musicale preferito
1	34	1.755	75.838	Lirica
2	43	1.752	77.713	Classica
3	35	1.747	76.448	Classica
4	33	1.831	85.514	Rap
5	51	1.748	74.241	Nessuna
6	29	1.754	78.706	Rap
7	47	1.752	77.295	Rock
8	51	1.696	65.507	Rock
9	59	1.784	85.392	Rock
10	24	1.743	80.905	Rap
	Var. num. **discreta**	Var. num. **continua**	Var. num. **continua**	Var. **cat.**

I dati presentati in una tabella così come sono raccolti sono detti **dati grezzi**.
Sono difficili da analizzare soprattutto se molto numerosi.

Un primo modo di analizzare i dati è quello di produrre dei **dati raggruppati** in **classi**.

Esempio 2.1.1. Consideriamo i dati relativi all'età degli individui appartenenti al campione della tabella che supponiamo essere composto da 200 persone e raggruppiamoli in **classi** di età:

Cl.	Freq. Ass.	Freq. Cum.	Freq. Rel.	Freq. Rel. Cum.	Freq. Perc.	Freq. Perc. Cum.
10-14	3.	3.	0.015	0.015	1.5	1.5
15-19	7.	10.	0.035	0.05	3.5	5.
20-24	17.	27.	0.085	0.135	8.5	13.5
25-29	19.	46.	0.095	0.23	9.5	23.
30-34	28.	74.	0.14	0.37	14.	37.
35-39	19.	93.	0.095	0.465	9.5	46.5
40-44	22.	115.	0.11	0.575	11.	57.5
45-49	21.	136.	0.105	0.68	10.5	68.
50-54	20.	156.	0.1	0.78	10.	78.
55-59	16.	172.	0.08	0.86	8.	86.
60-64	8.	180.	0.04	0.9	4.	90.
65-69	11.	191.	0.055	0.955	5.5	95.5
70-74	2.	193.	0.01	0.965	1.	96.5
75-79	1.	194.	0.005	0.97	0.5	97.
80-84	6.	200.	0.03	1.	3.	100.
85-90	0.	200.	0.	1.	0.	100.

Si considerino i dati grezzi $\{x_i\}_{i=1}^n$ (il campione ha quindi ampiezza n). Si considerino degli insiemi disgiunti $I_1, \dots, I_{N_c}$ tali che per ogni $i = 1, \dots, n$ si abbia $x_i \in \cup_{j=1}^{N_c} I_j$; definiamo la classe j-esima come l'insieme degli indici $C_j := \{i \in \{1, \dots, n\} : x_i \in I_j\}$.

Un caso particolare si ha quando ogni insieme I_j è un **singoletto** cioè quando $\#I_j = 1$ per ogni $j = 1, \dots, N_c$ (dove $\#$ denota la **cardinalità**).

Definizione 2.1.2. La **frequenza assoluta** $f_a(k)$ relativa alla k-esima classe è il numero di osservazioni che ricadono in quella classe.

$$f_a(k) = \#C_k \quad (k = 1, \dots, N_c)$$

Proprietà: $\sum_{k=1}^{N_c} f_a(k) = n$ essendo n il numero totale delle osservazioni (200 nell'esempio 2.1.1).

Definizione 2.1.3. La **frequenza relativa** $f_r(k)$ della k-esima classe è il rapporto $f_a(k)/n$
Proprietà: $\sum_{k=1}^{N_c} f_r(k) = 1$.

Definizione 2.1.4. La **frequenza percentuale** $f_p(k)$ è la quantità $f_p(k) = f_r(k) \cdot 100$.
Proprietà: $\sum_{k=1}^{N_c} f_p(k) = 100$.

Definizione 2.1.5. La **frequenza assoluta cumulativa** $F_a(k)$ della k-esima classe è il numero totale delle osservazioni che ricadono nelle classi fino a k-esima compresa:

$$F_a(k) = \sum_{j=1}^{k} f_a(j).$$

Proprietà: F_a è una funzione non decrescente e $F_a(N_c) = n$.

Definizione 2.1.6. La **frequenza relativa cumulativa** è il rapporto $F_r(k) = F_a(k)/n \equiv \sum_{j=1}^{k} f_r(j)$, ed è sempre compresa fra 0 ed 1.
Proprietà: F_r è una funzione non decrescente e $F_r(N_c) = 1$.

Definizione 2.1.7. La **frequenza percentuale cumulativa** $F_p(k)$ è la quantità $F_p(k) = F_r(k) \cdot 100$. Proprietà: F_p è una funzione non decrescente e $F_p(N_c) = 100$.

Il raggruppamento in classi costituite da intervalli contigui vale sia per variabili numeriche *discrete* che per variabili numeriche *continue*. Nel nostro esempio possiamo definire tanti intervalli di altezze in metri (es. [1.50-1.55], [1.55-1.60], [1.60-1.65], ...). Le frequenze sono definite nello stesso modo di prima.

Per le variabili categoriche le classi sono costituite in maniera naturale dalle categorie.

$$f_a(\text{cat. } k) = \#\{i \in \{1, \dots, n\} : x_i = \text{cat. } k\} \quad k = 1, \dots, N_c$$

e definizioni analoghe per le altre quantità f_r e f_p. Non ha senso invece definire la frequenza cumulativa a meno che non vi siano, in virtù di qualche specifica ragione, indicazioni per introdurre una relazione d'ordine totale sull'insieme delle categorie (ad esempio se le categorie fossero determinate da colori, uno potrebbe introdurre l'ordine indotto dalle lunghezze d'onda o dalle frequenze).

Nel nostro esempio:

$$f_a(\text{Rock}) = \#\{i \in \{1, \dots, n\} : x_i = \text{Rock}\} = 3.$$

- Arbitrarietà nella scelta delle suddivisioni.

- Dalle frequenze non si può risalire alle osservazioni in quanto i dati sono stati raggruppati (*perdita di informazione*).

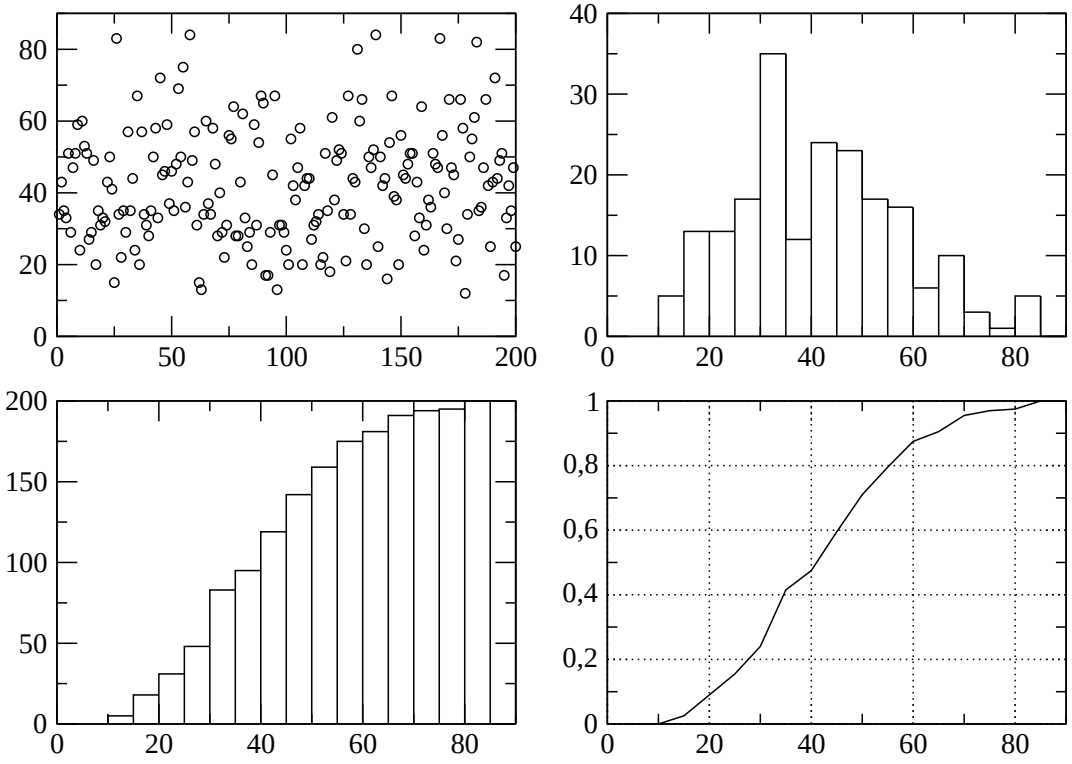
- Si ha quindi una facilità di comprensione e maggior chiarezza nell'esposizione dei dati.

2.2 Metodi grafici

- **Istogramma**: grafico della distribuzione di frequenze per dati *numerici*. Le basi dei rettangoli adiacenti sono gli intervalli che definiscono le classi. Comandi Matlab *hist* e *histc*. Le altezze possono essere scelte in due modi: nel primo l'altezza $h_r(j) = f_r(j)$ (risp. $h_a(j) = f_a(j)$), nel secondo $h_r(j) = f_r(j)/(b_j - a_j)$ (risp. $h_a(j) = f_a(j)/(b_j - a_j)$) dove $I_j = (a_j, b_j)$ è l'intervallo relativo alla classe j-esima (in questo caso è necessario che le basi siano proporzionali all'ampiezza della classe $b_j - a_j$). Nel primo modo l'altezza è proporzionale alla frequenza, nel secondo sarà l'area ad essere proporzionale alla frequenza.

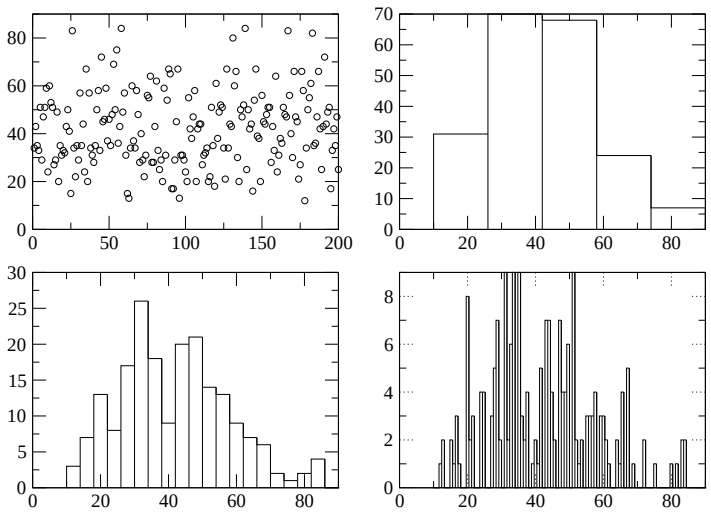
- **Diagramma a barre (o di Pareto)**: ad ogni classe corrisponde una barra la cui base non ha significato. Le barre non si disegnano adiacenti. Utili per rappresentare variabili di tipo *categorico*. Comandi Matlab *bar, pareto*. L'altezza di ogni barra è proporzionale alla frequenza.

- **Grafico di frequenza cumulativa**: si usa per dati *numerici*. in ascissa si riportano i valori osservati, oppure nella suddivisione in classi gli estremi degli intervalli di variabilità. In ordinata le frequenze cumulative corrispondenti. Comando Matlab *plot*. Nel caso del diagramma a barre cumulativo, l'altezza è proporzionale alla frequenza cumulativa.



Distribuzione delle età del campione di 200 persone: grafico dei dati grezzi, Istogramma della distribuzione delle frequenze assolute per le classi di età (comando Matlab *histc*), istogramma

della distribuzione delle frequenze cumulative assolute, grafico della distribuzione delle frequenze cumulative relative (comando Matlab *plot*).



Istogrammi della distribuzione delle frequenze assolute per le classi di età. Sono state usate diverse scelte dei numeri di classi.

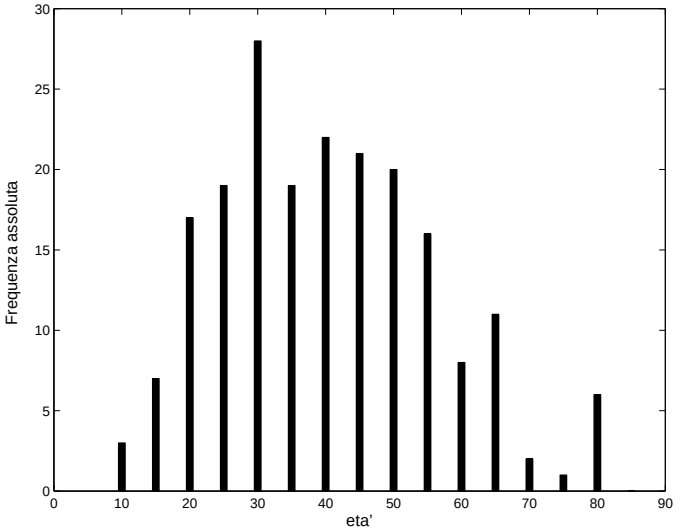


Diagramma a barre (o di Pareto) della distribuzione delle frequenze assolute per le classi di età.

2.3 Indici di Posizione, Dispersione e Forma

Si definiscono degli indici numerici che forniscono un'idea di massima di dove (indici di posizione) e come (indici di dispersione e di forma) i dati sono distribuiti.

2.3.1 Indici di Posizione

Gli indici di posizione più usati sono la **media**, la **mediana** e la **moda** associata al grafico della frequenza.

- **media** o **media campionaria** di n dati numerici $\{x_i\}_{i=1}^n$ (comando di Matlab *mean*):

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

Esempio 2.3.1. Supponiamo di aver misurato i seguenti 10 valori di una variabile discreta x:

$$x = [18 \quad 6 \quad 31 \quad 71 \quad 84 \quad 17 \quad 23 \quad 1 \quad 9 \quad 43]$$

allora la media è:

$$\overline{x} = (18 + 6 + 31 + 71 + 84 + 17 + 23 + 1 + 9 + 43)/10 = 30.3$$

Proprietà:

– La media fornisce sempre un valore compreso fra il minimo ed il massimo valore dell'insieme dei dati (strettamente compreso ogni qualvolta esistano almeno due dati differenti). Infatti supponiamo di avere ordinato i dati in ordine crescente: $x_1 \leq x_2 \cdots \leq x_n$. Allora:

$$x_n - \overline{x} = x_n - \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_n - \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(x_n - x_i) \geq 0$$

Analogamente

$$\overline{x} - x_1 = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_1) \geq 0$$

e quindi in definitiva $x_1 \leq \overline{x} \leq x_n$. Inoltre essendo $\sum_{i=1}^{n}(x_n - x_i) = 0$ se e solo se $x_1 = x_2 = \cdots = x_n$, allora $\overline{x} = x_n \equiv \max(x_1, \ldots, x_n)$ se e solo se $x_1 = x_2 = \cdots = x_n$ o, equivalentemente, se e solo se $\overline{x} = x_1 = \min(x_1, \ldots, x_n)$. Pertanto $\max(x_1, \ldots, x_n) > \overline{x} > \min(x_1, \ldots, x_n)$ se e solo se esistono i, j tali che $x_i \neq x_j$.

– Media calcolata a partire dai dati raggruppati in classi.
Dividiamo i dati in N_c classi indicando con x_{kl} il dato l-esimo della k-esima classe e con $f_a(k)$ la frequenza assoluta della k-esima classe. Possiamo riorganizzare il calcolo della media nel seguente modo

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{N_c}\sum_{l=1}^{f_a(k)} x_{kl}$$

ma, per definizione,

$$\sum_{l=1}^{f_a(k)} x_{kl} = f_a(k)\overline{x}_k$$

se $\overline{x}_k$ è la media dei dati della classe k-esima.
Sostituendo si ha:

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{N_c} f_a(k)\overline{x}_k = \sum_{k=1}^{N_c} f_r(k)\overline{x}_k$$

La media si ottiene dalle frequenze assolute o relative delle classi dei dati raggruppati se sono noti i valori medi dei dati in ciascuna classe. Poiché di solito questi ultimi non sono noti, si sostituisce a ciascun $\overline{x}_k$ il valore centrale dell'intervallo associato alla classe k (questo è un esempio di perdita di informazioni). In tal modo si ottiene un valore approssimato della media.

– Trasformazione affine di dati.
Abbiamo delle osservazioni $\{x_1, x_2, \ldots, x_n\}$ di cui abbiamo calcolato il valor medio $\overline{x}$. Ci interessa conoscere la media dei *dati trasformati in maniera affine* $y_i = ax_i + b$. Risulta

$$\overline{y} = a\overline{x} + b$$

Infatti

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = b + a\frac{1}{n}\sum_{i=1}^{n} x_i = a\overline{x} + b$$

Esempio 2.3.2. Siano $\{x_1, x_2, \ldots, x_n\}$ misure di temperatura in gradi Fahrenheit con valore medio $\overline{x}_F = 50$. Quale è la media in gradi centigradi?

$$\overline{x}_C = \frac{100}{180}(\overline{x}_F - 32) = 10^o C$$

– Aggregazione di dati.
Siano due campioni di osservazioni $\{x_1, x_2, \ldots, x_l\}$ e $\{y_1, y_2, \ldots, y_m\}$, di medie campionarie rispettive $\overline{x}$ e $\overline{y}$. Consideriamo quindi un nuovo campione costituito dai dati aggregati $\{z_1, z_2, \ldots, z_n\} = \{x_1, x_2, \ldots, x_l, y_1, y_2, \ldots, y_m\}$, $n = l + m$. La media $\overline{z}$ di questo campione è:

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{l}{n}\overline{x} + \frac{m}{n}\overline{y}.$$

- **mediana** di n dati numerici $\{x_i,\ i = 1 \ldots n\}$ (comando Matlab *median*):
si dispongono i dati in ordine crescente (ad esempio con il comando Matlab *sort*). La mediana è il dato nella posizione centrale se n è dispari, oppure la media aritmetica dei due dati in posizione centrale, se n è pari.

  Nell'esempio 2.3.1

$$x = [1 \quad 6 \quad 9 \quad 17 \quad 18 \quad 23 \quad 31 \quad 43 \quad 71 \quad 84]$$

  $n = 10$ è pari e quindi la mediana è $\frac{(18+23)}{2} = 20.5$.

  Proprietà:

  – **media** e **mediana** non coincidono necessariamente; sono tanto più vicine quanto più i dati sono disposti regolarmente. Entrambi gli indici forniscono un valore più o meno centrato rispetto ai dati. La media è più facile da calcolare. La mediana è meno sensibile alla presenza di valori aberranti nei dati.

  – Mediana di dati raggruppati: si può definire come quel valore che divide l'insieme dei dati raggruppati in due gruppi ugualmente numerosi.
  Infatti per definizione di mediana avremo che almeno metà dei dati sarà minore (o uguale) e almeno metà maggiore (o uguale) di essa.
  Per stimare la mediana basterà allora determinare il valore in corrispondenza del quale la frequenza cumulativa relativa o percentuale prende il valore 0.5 o 50, rispettivamente. a tal proposito segnaliamo due metodi differenti (in presenza di N_c classi).

    1. Si determina in quale classe cade: si dirà che cade nella classe k se e solo se $F_r(k-1) < 0.5 \leq F_r(k)$ (dove $F_r(0) := 0$); tale valore k evidentemente esiste ed è unico.

2. Si procede come nel punto precedente e quindi si stima per interpolazione lineare

$$med := \frac{0.5 - F_r(k-1)}{F_r(k) - F_r(k-1)}(M_k - m_k) + m_k$$

dove M_k ed m_k sono, rispettivamente, i valori massimo e minimo della classe k-esima (cioè, ad esempio, $I_k = (m_k, M_k]$).

- **moda** di n dati numerici raggruppati $\{x_i,\ i = 1 \dots n\}$: punto di massimo assoluto nella distribuzione di frequenza. La moda è dunque il valore o, più in generale, la classe in corrispondenza del quale si ha la popolazione più numerosa. Se il valore massimo è raggiunto in più punti, allora la distribuzione delle frequenze si dice plurimodale, altrimenti è detta unimodale.

Esempio 2.3.3. Campione di dati $\{1, 2, 2, 2, 4, 5, 6, 6, 8\}$.
Media campionaria: 4; mediana: 4; moda: 2.
Campione di dati $\{1, 2, 2, 2, 3, 4, 5, 6, 6, 6, 53\}$.
Media campionaria: 9; mediana: 4; mode: 2, 6.

2.3.2 Indici di dispersione

Si vuole valutare come si disperdono i dati intorno alla media.

Osservazione 2.3.4. Definiamo lo scarto del dato i-esimo come $s_i = x_i - \overline{x}$. La somma degli scarti **non** rappresenta un indice di dispersione poiché è identicamente nullo:

$$\sum_{i=1}^{n} s_i = \sum_{i=1}^{n} (x_i - \overline{x}) = 0.$$

- **range** di un insieme di dati $\{x_i, i = 1, \dots\}$ non necessariamente finito:

$$r = \sup\{x_i : i = 1, \dots\} - \inf\{x_i : i = 1, \dots\}$$

in particolare se l'insieme di dati è finito si ha $r = \max\{x_i : i = 1, \dots\} - \min\{x_i : i = 1, \dots\} \in \mathbb{R}$. dove $x_{\max}$ e $x_{\min}$ sono il valore massimo e minimo dell'insieme di dati.
Il range fornisce un'informazione piuttosto grossolana, poiché non tiene conto della distribuzione dei dati all'interno dell'intervallo che li comprende.

- **Varianza campionaria** (comando di Matlab *var*):

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

Deviazione standard o **scarto quadratico medio** (comando di Matlab *std*). è la radice quadrata della varianza:

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

Si osservi che $s^2 \geq 0$ e vale l'uguaglianza se e solo se $x_i = \overline{x}$ per ogni i; pertanto la varianza campionaria è positiva se e solo se esistono i e j tali che $x_i \neq x_j$.

Esempio 2.3.5. i tempi per il taglio di una lastra in sei parti sono (espressi in minuti): $\{0.6, 1.2, 0.9, 1.0, 0.6, 0.8\}$. Calcoliamo s.

$$\overline{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 0.85 \text{ (minuti)}$$

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
0.6	-0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	-0.25	0.0625
0.8	-0.05	0.0025

$$s = \sqrt{\frac{0.0625 + 0.1225 + 0.0025 + 0.0225 + 0.0625 + 0.0025}{5}}$$
$$= \sqrt{\frac{0.2750}{5}} \approx 0.23(\text{minuti})$$

Proprietà della varianza:

– In alcuni testi si trova la definizione seguente di varianza:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n}s^2$$

e analogamente per la deviazione standard. La definizione data in precedenza è da preferirsi a questa, per ragioni che verranno esposte nella Sezione 8.2.2, o più in generale nel Capitolo 8.1 quando parleremo di stimatori corretti. La maggior parte dei pacchetti software di analisi statistica usa la prima definizione. Per n grande la differenza è trascurabile.

– Modo alternativo di calcolare la varianza:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - \frac{n}{n-1}\overline{x}^2$$

Con l'altra definizione di varianza si ottiene:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

In questo caso la varianza è pari alla differenza fra la media dei quadrati e il quadrato della media.

Varianza calcolata in base ai dati raggruppati:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{k=1}^{N_c}\sum_{l=1}^{f_a(k)}(x_{lk} - \overline{x})^2$$

sostituiamo x_{lk} con $\overline{x}_k$

$$\begin{aligned} s^2 &\approx \frac{1}{n-1}\sum_{k=1}^{N_c}\sum_{l=1}^{f_a(k)}(\overline{x}_k - \overline{x})^2 \\ &= \frac{1}{n-1}\sum_{k=1}^{N_c}f_a(k)(\overline{x}_k - \overline{x})^2 = \frac{1}{n-1}\sum_{k=1}^{N_c}f_a(k)\overline{x}_k^2 - \frac{n}{n-1}\overline{x}^2 \\ &= \frac{n}{n-1}\sum_{k=1}^{N_c}f_r(k)(\overline{x}_k^2 - \overline{x}^2) = \frac{n}{n-1}\left(\sum_{k=1}^{N_c}f_r(k)\overline{x}_k^2 - \overline{x}^2\right). \end{aligned}$$

Si dimostra facilmente che l'approssimazione sarebbe esatta se sostituissimo $\overline{x}_k$ con $\sqrt{\sum_{i=1}^{f_a(k)}x_{ik}^2/f_a(k)}$ (cosa non possibile se non si è più in possesso dei dati grezzi).

– Trasformazione affine di dati.
Abbiamo delle osservazioni $\{x_1, x_2, \ldots, x_n\}$ di cui abbiamo calcolato la varianza s_x^2. Ci interessa conoscere la varianza dei *dati trasformati in maniera affine* $y_i = ax_i + b$. Risulta

$$s_y^2 = a^2 s_x^2$$

Infatti

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\overline{x})^2$$

$$= a^2 \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2 s_x^2$$

– Dato il campione $\{x_1, \ldots, x_n\}$, definiamo la **variabile standardizzata** operando la seguente trasformazione

$$y = \frac{x - \overline{x}}{s_x}$$

Il **campione standardizzato** corrispondente $y_i = (x_i - \overline{x})/s_x$ ha media nulla e varianza 1. Infatti

$$\overline{y} = \frac{1}{s_x}(\overline{x} - \overline{x}) = 0$$

$$s_y^2 = \frac{1}{s_x^2} s_x^2 = 1.$$

- **Quantili, percentili & C.**
Sia $\{x_i\}_{i=1}^n$ un campione di dati numerici *ordinato* in maniera *non decrescente* (i.e. $x_i \le x_j$ per ogni $i \le j$). Con la seguente definizione generalizziamo il concetto di mediana cercando un valore q_p con la proprietà che almeno una frazione p dei dati sia non superiore a q_p ed almeno una frazione $1-p$ sia non inferiore a q_p (qui, inferiore e superiore sono da intendersi *strettamente*).

Definizione 2.3.6. Definizione *non interpolata* di p**-esimo quantile** q_p $(0 < p < 1)$:
se np non è intero, sia k l'intero tale che $k < np < k+1$: $q_p = x_{k+1}$.
se $np = k$ con k intero, allora $q_p = (x_k + x_{k+1})/2$.
Definizione *interpolata* di p**-esimo quantile** q_p $(0 < p < 1)$:
alcuni programmi (ad esempio Matlab) restituiscono un valore interpolato calcolato come segue:

$$q_p := \begin{cases} x_1 & np < 1/2 \\ (np - k + 1/2)(x_{k+1} - x_k) + x_k & k - 1/2 \le np < k + 1/2 \\ x_n & n - 1/2 \le np. \end{cases}$$

Si noti che le due definizioni date sopra coincidono se e solo se $p = (k - 1/2)/n$ per qualche $k = 1, 2, \ldots, n$. Verifichiamo immediatamente che le definizioni date soddisfano le richieste fatte, ci limitiamo a tal proposito solo alla prima definizione che è l'unica che utilizzeremo (per la seconda la verifica è analoga). Calcoliamo quindi la frazione di dati non superiore (risp. non inferiore) a q_p:

$$\frac{\#\{i : x_i \le q_p\}}{n} \ge \begin{cases} \frac{k+1}{n} \ge p & k < np < k+1 \\ \frac{k}{n} = p & k = np \end{cases}$$

$$\frac{\#\{i : x_i \ge q_p\}}{n} \ge \frac{n-k}{n} = 1 - \frac{k}{n} \ge 1 - p,$$

e la seconda disuguaglianza nella prima riga (risp. la disuguaglianza nell'ultima riga) diventa un'uguaglianza se e solo se np è un valore intero. La prima disuguaglianza invece dipende dal fatto che $x_{k+1} = x_{k+2}$ (risp. $x_{k+1} = x_k$) o meno.

- Il *p-esimo quantile* viene anche detto $100p$**-esimo percentile**.
- Il p-esimo quantile o $100p$-esimo percentile forniscono un valore che risulta maggiore o uguale del 100p% dei dati del campione.
- Il 25^o, 50^o e 75^o percentile vengono detti anche primo, secondo e terzo **quartile**, e indicati con Q_1, Q_2, Q_3. Q_1, Q_2, Q_3 sono tre numeri che dividono l'insieme di osservazioni in 4 gruppi contenenti ciascuno circa un quarto dei dati.
- Il secondo quartile Q_2 coincide con la mediana.
- Anche se è piu corretto annoverare i quantili tra gli indici di posizione, è possibile ricavare un indice di dispersione, chiamato **differenza interquartile**, o **IQR** (dall'inglese *interquartile range*) definito come la distanza fra il primo ed il terzo quartile:

$$IQR = Q_3 - Q_1$$

Esempio 2.3.7. Altezze di 20 persone di sesso maschile. Campione ordinato in modo crescente espresso in metri:
{1.58, 1.60, 1.66, 1.68, 1.70, 1.74, 1.74, 1.75, 1.75, 1.76, 1.78, 1.78, 1.78, 1.79, 1.80, 1.81, 1.82, 1.84, 1.88, 1.91}.

$$\overline{x} = 1.7575m, \quad s^2 = 0.0069m^2, \quad s = 8.3cm$$

Calcolo di $q_{0.5}$: $np = 20 \cdot 0.5 = 10$ è intero. Pertanto:

$$q_{0.50} = \frac{x_{10} + x_{11}}{2} = \frac{1.76 + 1.78}{2} = 1.77m$$

Analogamente:

$$q_{0.25} = \frac{x_5 + x_6}{2} = \frac{1.70 + 1.74}{2} = 1.72m$$

$$q_{0.75} = \frac{x_{15} + x_{16}}{2} = \frac{1.80 + 1.81}{2} = 1.805m$$

$IQR = q_{0.75} - q_{0.25} = 1.805 - 1.72 = 8.5cm$. range$=x_{20} - x_1 = 1.91 - 1.58 = 33cm$.

In presenza di dati raggruppati, analogamente al caso della mediana, si può calcolare il quantile q_α individuando la classe cui appartiene e successivamente stimandone il valore.

1. La classe in cui cade sarà quella in corrispondenza alla quale si supera il valore α nel calcolo della frequenza relativa cumulativa, cioè

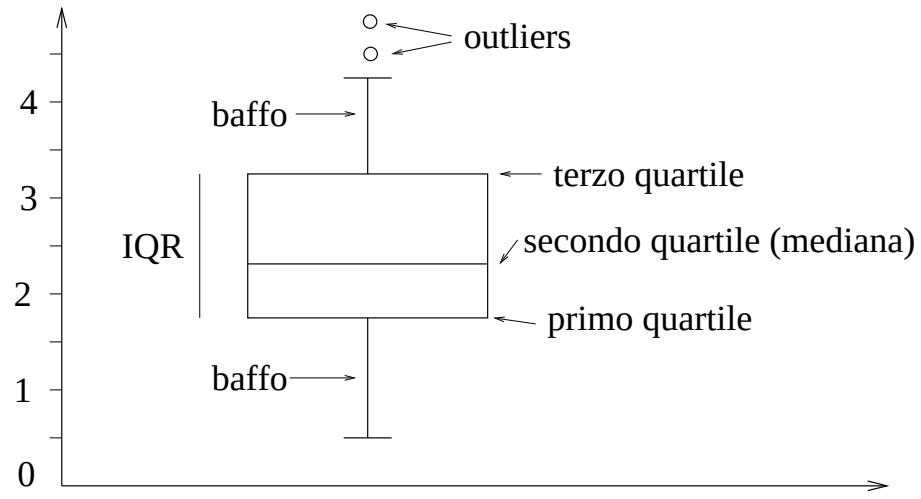
$$q_\alpha \in I_k \text{ dove } k := \min\{i = 1, \ldots, N_c : F_r(i) \geq \alpha\}.$$

2. Determinata la classe è possibile procedere alla stima interpolata del quantile tramite la relazione

$$q_\alpha = \frac{\alpha - F_r(k-1)}{F_r(k) - F_r(k-1)}(M_k - m_k) + m_k$$

dove M_k e m_k sono, rispettivamente, il massimo ed il minimo valore della classe k e k è definito come al punto precedente.

- Alcune informazioni contenute nella distribuzione di frequenza (e in particolare nei quartili) possono essere visualizzate graficamente con un **boxplot**.



Gli *outliers* sono dati che *giacciono fuori dai limiti*, la cui correttezza andrebbe accertata. Possono essere definiti in vari modi.
Ad esempio come quei dati che stanno sotto il 5^o percentile o sopra il 95^o percentile.
Altra definizione (usata da Matlab): si calcola un limite superiore $U = Q_3 + 1.5IQR$; il baffo superiore viene prolungato fino all'ultima osservazione che risulta minore o uguale di U. Gli outliers sono i dati che eccedono questo valore. Si segue analoga procedura per gli outliers inferiori.

2.3.3 Indici di forma

- La **skewness**

$$\gamma_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^3$$

È una grandezza *adimensionale*. Può assumere valori sia positivi che negativi.
Se è negativa denota una *coda* verso sinistra.
Se è positiva denota una *coda* verso destra.
se la distribuzione è simmetrica, allora la skewness è nulla, ma l'inverso non è vero.
Per trasformazioni lineari $y_i = ax_i + b$ la skewness non cambia: $\gamma_3^y = \gamma_3^x$.

- La **curtosi**

$$\gamma_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^4$$

È una grandezza adimensionale e non negativa. Misura (in un certo senso) l'appiattimento della distribuzione delle frequenze, poiché assegna un peso elevato agli scarti grandi: valori elevati della curtosi segnalano distribuzioni significativamente diverse da $\overline{x}$ per grandi scarti, piccoli valori distribuzioni *appuntite* in corrispondenza di $\overline{x}$.
Per trasformazioni lineari $y_i = ax_i + b$ la curtosi non cambia: $\gamma_4^y = \gamma_4^x$.

- Momento centrato di ordine k:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$$

Se k è pari, allora $m_k \geq 0$: indice di dispersione ($m_2 = \sigma^2$) e di forma. Si mostra che esiste $k \in \mathbb{N}^*$ pari (dove $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$) tale che $m_k = 0$ se e solo se per ogni $k \in \mathbb{N}^*$ pari si ha $m_k = 0$ cioè se e solo se tutti i dati coincidono.
Se k è dispari, allora $m_k \in \mathbb{R}$: indice di simmetria.

2.4 Analisi comparative, correlazione tra variabili

Si effettuano osservazioni simultanee di più variabili su una medesima popolazione (ad esempio peso e altezza in un campione di persone). I dati in questo caso si dicono *multivariati*. Ci si domanda se esistono dei legami (associazione, dipendenza, correlazione) tra le variabili considerate.

2.4.1 Frequenze congiunte per dati multivariati raggruppati in classi

Come nel caso dei dati univariati è spesso utile raggruppare i dati in classi. Si considerino n osservazioni vettoriali $\{x_i := (x_i(1), \dots, x_i(k))\}_{i=1}^n$ e si considerino N_j classi in cui si suddividono i valori relativi alla j-esima variabile (o j-esima componente del vettore) $C_1(j), \dots, C_{N_j}(j)$ $(j = 1, \dots, k)$; le classi risultino definite, come nel caso unidimensionale da una scelta di insiemi $\{I_i(j)\}_{i=1}^{N_j}$ per ogni $j = 1, \dots, k$ come

$$C_i(j) := \{w \in \{1, \dots, n\} : x_w(j) \in I_i(j)\}.$$

Si ottengono così delle classi intersezione

$$C_{i_1,\dots,i_k} := \bigcap_{j=1}^{k} C_{i_j}(j).$$

Definiamo di seguito le frequenze congiunte per dati bivariati.

Definizione 2.4.1. la **frequenza assoluta congiunta** $f_a(i_1, \dots, i_k)$ relativa alla i_j-esima classe della j-esima variabile (per $j = 1, 2, \dots, k$) è il numero delle osservazioni che ricadono in quelle classi.

$$\begin{aligned} f_a(i_1, \dots, i_k) &:= \#\{h \in \{1, 2, \dots, n\} : x_h(j) \in I_{i_j}(j), \forall j = 1, 2, \dots, k\} \\ &\equiv \#\{h \in \{1, 2, \dots, n\} : h(j) \in C_{i_j}(j), \forall j = 1, 2, \dots, k\} \\ &= \#C_{i_1,\dots,i_k}. \end{aligned}$$

essendo n il numero totale delle osservazioni ed $1 \le i_j \le N_j$ per ogni $j = 1, \dots, k$.

Esempio 2.4.2. Il numero di persone per le quali l'altezza è compresa tra 1.65 e 1.70 metri, e il peso è compreso tra 75 e 80 Kg.

Proprietà:

$$\sum_{\substack{i_1=1,\dots,N_1 \\ \vdots \\ i_k=1,\dots,N_k}} f_r(i_1, \dots, i_k) = n.$$

Definizione 2.4.3. La **frequenza relativa congiunta** $f_r(i_1, \dots, i_k)$ è definita come il rapporto $f_a(i_1, \dots, i_k)/n$.
Proprietà:

$$\sum_{\substack{i_1=1,\dots,N_1 \\ \vdots \\ i_k=1,\dots,N_k}} f_r(i_1, \dots, i_k) = 1.$$

Definizione 2.4.4. La **frequenza cumulativa congiunta assoluta** $F_a(i_1,\dots,i_k)$ è il numero totale delle osservazioni che ricadono in una delle classi fino alla i_1 compresa per la prima variabile, fino alla i_2 compresa per la seconda variabile e così dicendo fino alla i_k compresa per la k-esima variabile:

$$F_a(i_1,\dots,i_k) = \sum_{\substack{j_1=1,\dots,i_1 \\ \vdots\ \vdots\ \vdots \\ j_k=1,\dots,i_k}} f_a(j_1,\dots,j_k)$$

dove le classi sono state ordinate per valori crescenti. Analogamente si definisce la **frequenza cumulativa congiunta relativa**

Definizione 2.4.5. Si supponga di dividere le k variabili in due insiemi disgiunti $\{x(y_1),\dots,x(y_h)\}$ e $\{x(r_1),\dots,x(r_d)\}$ cosicché $d+h=k$; in tal caso chiameremo **frequenza marginale assoluta** relativa al primo gruppo di variabili la quantità

$$\begin{aligned} f_{a,\,x(y_1)\cdots x(y_h)}(i_{y_1},\dots,i_{y_h}) &= \sum_{\substack{i_{r_1}=1,\dots,N_{r_1} \\ \vdots\ \vdots\ \vdots \\ i_{r_d}=1,\dots,N_{r_d}}} f_a(i_1,\dots,i_k) \\ &\equiv \#\{w\in\{1,\dots,n\} : x_w(y_1)\in I_{i_{y_1}}(y_1),\dots,x_w(y_h)\in I_{i_{y_h}}(y_h)\} \\ &\equiv \#\bigcap_{i=1}^{h} C_{i_{y_h}}(y_h). \end{aligned}$$

Analogamente si definisce la **frequenza marginale relativa**:

$$f_{r,\,x(y_1)\cdots x(y_h)}(i_{y_1},\dots,i_{y_h}) = f_{a,\,x(y_1)\cdots x(y_h)}(i_{y_1},\dots,i_{y_h})/n.$$

Dalle frequenze marginali si può ricavare la frequenza congiunta solo in casi molto specifici, ossia se le due variabili sono *indipendenti*, come vedremo più avanti nel corso. In generale nemmeno dalla conoscenza di tutte le marginali si può risalire alla congiunta.

La media e la varianza per ciascuna variabile si definiscono nel modo naturale a partire dalle marginali.

Osservazione 2.4.6. Nel caso bidimensionale, si considerano coppie $\{(x_i,y_i)\}_{i=1}^n$, suddivise in N_x e N_y classi rispetto alla prima e seconda coordinata, siano esse $\{C_x(i)\}_{i=1}^{N_x}$ e $\{C_y(j)\}_{j=1}^{N_y}$. Allora, ad esempio, le frequenze assolute congiunta e marginali prendono la forma

$$f_a(i,j) := \#C_x(i)\cap C_y(j),\ \ f_a^x(i) = \sum_{j=1}^{N_y} f_a(i,j)\ \ f_a^y(j) = \sum_{i=1}^{N_x} f_a(i,j)$$

e analogamente per la frequenza relativa.

2.4.2 Covarianza, coefficiente di correlazione

In questo paragrafo supporremo che k sia pari a 2 avremo quindi due set di variabili. Almeno inizialmente supporremo di avere i dati grezzi $\{(x_i,y_i)\}_{i=1}^n$.

Definizione 2.4.7. La **covarianza campionaria** delle variabili x e y è il numero

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \frac{1}{n-1}\sum_{i=1}^{n}x_iy_i - \frac{n}{n-1}\overline{x}\ \overline{y}$$

- Come per la varianza, anche nel caso della covarianza si trova in alcuni testi la definizione con n al denominatore al posto di $n-1$.

- Vale la proprietà: $s_{xy} = s_{yx}$.

- Vale la proprietà:
$$s^2_{x+y} = s^2_x + s^2_y + 2s_{xy}$$

Dimostrazione. Mostriamo innanzitutto che $\overline{x+y} = \overline{x} + \overline{y}$:
$$\overline{x+y} = \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{x} + \overline{y}$$

Dunque:
$$s^2_{x+y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i + y_i - \overline{x} - \overline{y})^2 =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 + (y_i - \overline{y})^2 + 2(x_i - \overline{x})(y_i - \overline{y}) = s^2_x + s^2_y + 2s_{xy}$$

□

- Vale la proprietà:
$$s_{\alpha x, \beta y} = \alpha\beta s_{xy}$$

Infatti:
$$s_{\alpha x, \beta y} = \frac{1}{n-1}\sum_{i=1}^{n}(\alpha x_i - \alpha\overline{x})(\beta y_i - \beta\overline{y}) =$$

$$= \alpha\beta\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \alpha\beta s_{xy}$$

- Covarianza calcolata in base ai dati raggruppati:
$$s_{xy} \approx \frac{1}{n-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} f_a(i,j)(\overline{x}_i - \overline{x})(\overline{y}_j - \overline{y})$$

Definizione 2.4.8. Il **coefficiente di correlazione campionario** di x e y è il numero
$$\rho_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{k=1}^{n}(y_k - \overline{y})^2}}$$

- Il coefficiente di correlazione ha lo stesso segno della covarianza.

- Le variabili x e y si dicono
direttamente correlate se $s_{xy} > 0$ (e dunque se $\rho_{xy} > 0$),
inversamente correlate se $s_{xy} < 0$,
non correlate se $s_{xy} = 0$.

- Vale la proprietà:
$$-1 \leq \rho_{xy} \leq 1$$
Per mostrarlo si noti che
$$0 \leq s^2_{\frac{x}{s_x}+\frac{y}{s_y}} = s^2_{\frac{x}{s_x}} + s^2_{\frac{y}{s_y}} + 2s_{\frac{x}{s_x},\frac{y}{s_y}} = 2 + 2\frac{s_{xy}}{s_x s_y}$$
da cui $\rho_{xy} \geq -1$. Ragionando in modo analogo su $s^2_{\frac{x}{s_x}-\frac{y}{s_y}}$ deduciamo che $\rho_{xy} \leq 1$.
- $\rho_{xy} = \pm 1$ se e solo se esistono a e b tale che $y_i = ax_i + b$. ρ_{xy} ha lo stesso segno di a.
- ρ_{xy} è invariante per trasformazione affine: se $x'_i = ax_i + b$, $y'_i = cy_i + d$, si ha $\rho_{x'y'} = \rho_{xy}$

Osservazione 2.4.9. Trovare una correlazione tra due variabili x e y non significa aver trovato un legame di causa-effetto tra loro, ma semplicemente un'indicazione qualitativa sulla monotonia "congiunta" delle due variabili (questa indiazione diventerà meno qualitativa e più quantitativa nella regressione lineare). Una prima ragione risiede nel fatto che la correlazione è simmetrica nello scambio tra x e y quindi non è ben chiaro se il fenomeno rilevato con x debba essere causato dal fenomento rilevato con y o viceversa. Una seconda ragione è che entrambi gli effetti potrebbero essere causati da un elemento che non abbiamo preso in considerazione.

Un esempio è il seguente. È noto che il succo di arancia ha effetti benefici sui postumi di un consumo elevato di alcool. Per cui conduciamo il seguente esperimento: facciamo consumare una grande quantità di alcool ad un gruppo di persone e poi facciamo fare a loro un test cognitivo. Infine facciamo bere una grande quantità di succo di arancia e successivamente facciamo ripetere un test analogo. Se una persona all'oscuro dell'esperimento (e della sua sequenza temporale) confrontiasse i dati "quantità di succo di arancia assunta" e "errori nel test" vedrebbe probabilmente una correlazione positiva. Potrebbe quindi erroneamente concludere che (1) fare molti errori nei test aumenta il consumo di succo di arancia oppure (2) consumare succo di arancia fa aumentare il numero di errori nel test. Se poi si decidesse di confrontare i dati "quantità di succo di arancia assunta" e "differenza tra errori nel secondo test ed errori nel primo test" vedrebbe probabilmente una correlazione negativa che porterebbe a concludere che assumere succo di arancia migliora le capacità cognitive. Se infine decidesse di confrontare i dati "quantità di succo di alcool" e "differenza tra errori nel secondo test ed errori nel primo test" vedrebbe probabilmente ancora una correlazione negativa che porterebbe a concludere che assumere alcool migliora le capacità cognitive.

Approfondimento

Ricordiamo che dato uno spazio vettoriale normato $(X, \|\cdot\|)$ completo (si pensi ad esempio a $\mathbb{R}^k$), gli elementi di una collezione al più numerabile (i.e. finita o numerabile) $\{v_i\}_{i\in I}$ si dicono *linearmente indipendenti* se e solo se ogni volta che $\sum_{i\in I} a_i v_i = 0$ si ha $a_i = 0$ per ogni $i \in I$. Si dimostra immediatamente che se la norma deriva da un prodotto scalare $\langle\cdot,\cdot\rangle$ allora $\{v_i\}_{i\in I}$ sono linearmente dipendenti se e solo se esiste $\{a_i\}_{i\in I}$ successione di scalari non tutti nulli tale che $\sum_{i\in I} a_i v_i$ converge e $\langle v_j, \sum_{i\in I} a_i v_i\rangle = 0$ per ogni $j \in I$. Come corollario si ha che $\{v_1, \dots, v_n\}$ sono linearmente dipendenti se e solo se la matrice $n \times n$
$$C := (\langle v_i, v_j\rangle)_{i,j=1}^n$$
non è invertibile i.e. $\det(C) = 0$.
Un altro corollario è che tra $\{v_1, \dots, v_n\}$ (questa volta si supponga che lo spazio sia $\mathbb{R}^k$) esiste una relazione affine non banale, i.e. esistono degli scalari $\{a_0, a_1, \dots, a_n\}$ non tutti nulli tali che
$$a_0\mathbf{1} + a_1 v_1 + \cdots + a_n v_n = 0 \tag{2.1}$$
se e solo se $\det(Cov(v_1, \dots, v_n)) = 0$ dove $\mathbf{1}$ è il vettore con tutte le coordinate pari ad 1, mentre
$$Cov(v_1, \dots, v_n) := (\mathrm{Cov}(v_i, v_j))_{i,j=1}^n.$$

Si osservi, infatti, che se vale l'equazione 2.1 si ha necessariamente $a_0 = -\sum a_i \overline{v}_i$ dove $\overline{v}_i$ è la media delle coordinate del vettore i-esimo

$$\overline{v}_i := \frac{1}{k} \sum_{j=1}^{k} v_i(j).$$

Quindi $a_0, \ldots, a_n$ non sono tutti nulli e se definiamo $\widehat{v}_i := v_i - \overline{v}_i \mathbf{1}$ allore l'equazione 2.1 è equivalente a

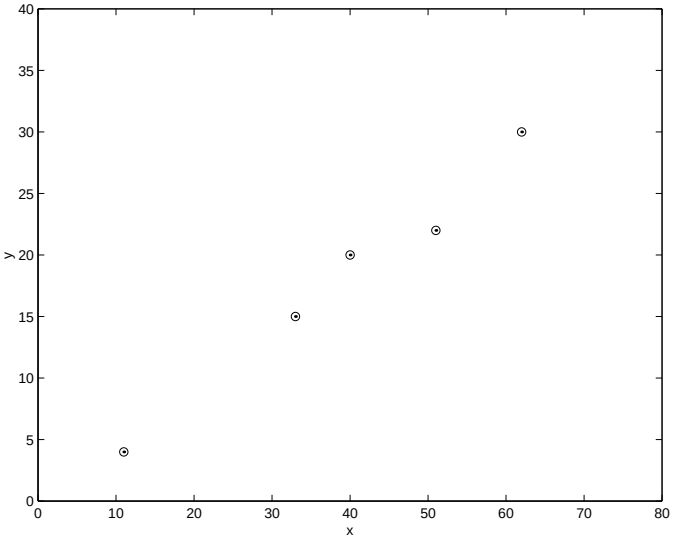
$$a_1 \widehat{v}_1 + \cdots + a_n \widehat{v}_n = 0$$

applico quindi il precedente risultato.

2.4.3 Scatterplot, o diagramma di dispersione

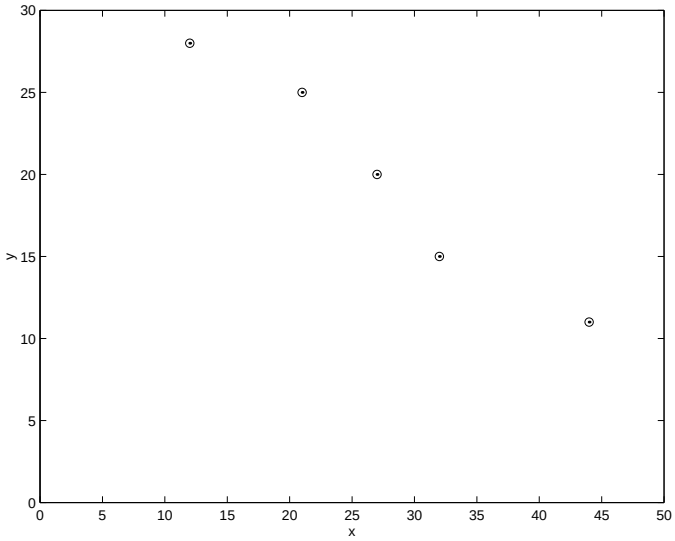
Lo **scatterplot** (comando di Matlab *plot*) è un metodo grafico utile per stimare se esistono delle correlazioni tra due variabili. Si mette in ascissa una variabile, in ordinata un'altra, e si rappresentano le singole osservazioni con dei punti.
Se punti con ascissa piccola hanno ordinata piccola, e punti con ascissa grande hanno ordinata grande, allora esiste una correlazione diretta tra le due variabili ($\rho_{xy} > 0$).
Viceversa quando al crescere dell'una l'altra decresce si ha correlazione inversa ($\rho_{xy} < 0$).
Se i punti formano una nuvola indistinta i dati sono pressoché scorrelati.

Esempio 2.4.10. $(x_i, y_i) = (11, 4), (51, 22), (40, 20), (62, 30), (33, 15)$.



I dati sono fortemente correlati. Infatti $\rho_{xy} = 0.9913$.

Esempio 2.4.11. $(x_i, y_i) = (21, 25), (12, 28), (32, 15), (44, 11), (27, 20)$.



I dati sono inversamente correlati. Infatti $\rho_{xy} = -0.9796$.

2.4.4 Regressione lineare semplice

Ricerca di una relazione affine tra le variabili x e y. Stiamo supponendo di avere

$$y_i = ax_i + b + r_i \qquad (1)$$

dove r_i è un residuo che vogliamo quanto più piccolo possibile in qualche senso che dovremo specificare.

- Chiameremo x_i *predittore* e y_i *responso*.
- La retta che cerchiamo si chiama **retta di regressione semplice** (si dice semplice perché coinvolge un solo predittore), o anche **retta dei minimi quadrati**.
- Alla forma (1) si può essere arrivati dopo eventuali trasformazioni dei dati.

Per stimare al meglio i coefficienti a e b utilizziamo il **principio dei minimi quadrati**: minimizziamo la quantità

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

Dal calcolo differenziale sappiamo che dobbiamo imporre:

$$\frac{\partial f(a,b)}{\partial a} = -\sum_{i=1}^{n} 2x_i \, [y_i - (ax_i + b)] = 0$$

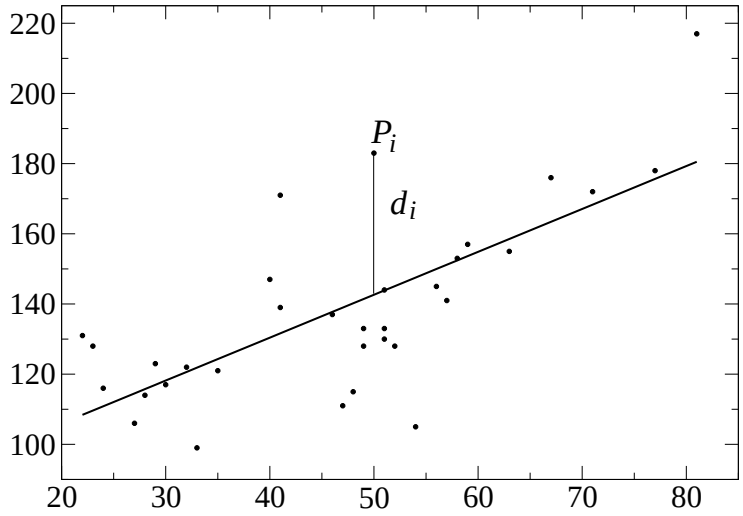
$$\frac{\partial f(a,b)}{\partial b} = -\sum_{i=1}^{n} 2 \, [y_i - (ax_i + b)] = 0$$

.

Otteniamo quindi:

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$b = \overline{y} - a\overline{x} = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x}$$



La retta di regressione è quella che rende minima la somma dei quadrati delle lunghezze d_i dei segmenti verticali congiungenti i punti osservati con la retta stessa.

. La matrice hessiana risulta

$$Hf(a,b) = 2\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{pmatrix}$$

quindi, se $n \geq 2$ e per almeno due coppie (x_i, y_i) e (x_j, y_j) si ha $x_i \neq x_j$, si ottiene immediatamente

$$\sum_{i=1}^n x_i^2 > 0, \qquad \det(Hf(a,b)) = n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 > 0$$

che implica che ogni punto stazionario è un minimo. Nota: $\sum_{i=1}^n r_i = 0$.

Per stimare la qualità di una regressione possiamo utilizzare i seguenti criteri:

- Il coefficiente di correlazione ρ_{xy} deve essere vicino a 1 o a -1 (ρ_{xy}^2 vicino ad 1).
- L'esame visivo dello scatterplot delle due variabili: i dati devono essere vicini alla retta di regressione.
- L'esame del grafico dei residui: in ascissa i valori previsti, in ordinata i valori dei residui. La nuvola dei punti deve avere un aspetto omogeneo, senza segni di curvatura, allargamenti o restringimenti.
  - Un grafico dei residui che presenti curvatura è un indizio che una dipendenza lineare non spiega bene i dati. Si può tentare di correggere questo difetto con trasformazioni di x e/o y, oppure si può provare a passare a una regressione multipla (che definiremo più avanti).
  - Un allargarsi/restringersi della nuvola di punti è un indizio che gli errori non sono tutti dello stesso tipo al variare di i. Si scelga quella combinazione di trasformazioni che danno la nuvola dei residui più omogenea possibile.

2.4.5 Regressione lineare multipla

Il responso y è *spiegato* da più predittori $\underline{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})$. Ipotizziamo il modello teorico

$$y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)} + r_i \qquad (1)$$

I coefficienti $a_0, a_1, \ldots, a_d$ sono stimati usando il principio dei minimi quadrati: si rende minima la quantità

$$f(a_0, a_1, \ldots, a_d) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left[y_i - (a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)}) \right]^2$$

Dobbiamo imporre:

$$\frac{\partial f(a_0, a_1, \ldots, a_d)}{\partial a_0} = -\sum_{i=1}^{n} \left[y_i - (a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)}) \right] = 0$$

$$\frac{\partial f}{\partial a_k} = -\sum_{i=1}^{n} 2x_i^{(k)} \left[y_i - (a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)}) \right] = 0, \quad k = 1, \ldots, d$$

Questo sistema lineare di $d+1$ equazioni in $d+1$ incognite ammette una soluzione unica (supponendo che il determinante sia non nullo); tale soluzione, essendo $\lim_{\|\underline{a}\| \to +\infty} f(\underline{a}) = +\infty$, è sicuramente un minimo.

È comodo riscrivere il sistema di equazioni in forma matriciale: posto

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(d)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(d)} \end{bmatrix} \quad r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} \quad a = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} \qquad (2.2)$$

Il sistema da risolvere

$$X^T X a = X^T y$$

ammette soluzione è

$$a = (X^T X)^{-1} X^T y.$$

L'equazione $y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)}$ è l'equazione di un iperpiano. Esso rappresenta quell'iperpiano che rende minima la somma dei quadrati delle lunghezze d_i dei segmenti congiungenti i punti osservati all'iperpiano stesso

- Come per la regressione lineare semplice possiamo essere arrivati al modello lineare (1) dopo aver fatto trasformazioni sul responso e/o sui predittori.
- Tra i predittori possiamo inserire anche potenze e prodotti dei predittori fondamentali.
- Se i predittori sono tutti potenze di un unico predittore fondamentale, si parla di **regressione polinomiale**.
- Il grafico dei residui, ossia lo scatterplot dei punti $(a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)}, r_i)$, è anche in questo caso uno strumento di analisi grafica per controllare la bontà della regressione. Valgono le considerazioni già fatte nel caso della regressione semplice.
- Definiamo
  la *devianza totale* $DT = \sum_{i=1}^{n} (y_i - \overline{y})^2$,
  la *devianza spiegata* $DS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ (dove $\hat{y}_i$ sono i valori previsti $\hat{y}_i = a_0 + \sum_{k=1}^{d} a_k x_i^{(k)}$),
  la *devianza dei residui* $DR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Vale la proprietà $DT = DS + DR$. Infatti:

$$\begin{aligned} DT - DS - DR &= \sum_{i=1}^{n}[(y_i - \overline{y})^2 - (\hat{y}_i - \overline{y})^2 - (y_i - \hat{y}_i)^2] \\ &= \sum_{i=1}^{n}[(y_i - 2\overline{y} + \hat{y}_i)(y_i - \hat{y}_i) - (y_i - \hat{y}_i)^2] = \sum_{i=1}^{n} 2(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) \\ &= -\sum_{i=1}^{n}\sum_{k=1}^{d} a_k \frac{\partial f}{\partial a_k} + 2(a_0 - \overline{y})\frac{\partial f}{\partial a_0} \\ &= 0. \end{aligned}$$

Il coefficiente di *determinazione multipla* R^2 definito da

$$R^2 = \frac{DS}{DT} = 1 - \frac{DR}{DT}$$

è sempre compreso tra 0 e 1 ed è un indice della frazione della variabilità di y spiegata dal modello. R^2 vicino a 1 è un buon indizio. Si noti che nel caso della regressione semplice $R^2 \equiv \rho_{xy}^2$.

Modelli con retta di regressione per l'origine: Si ipotizza che il responso deve essere nullo quando i predittori sono nulli. In altre parole il coefficiente a_0 viene posto uguale a 0: $y_i = a_1 x_i^{(1)} + a_2 x_i^{(2)} + \cdots + a_d x_i^{(d)} + r_i$. Si procede come prima col principio dei minimi quadrati, ma si ottengono d equazioni nelle d incognite $a_1, \ldots, a_d$. La soluzione cambia.
Nota: non è più vero che $DT = DS + DR$.

Approfondimento

Metodo dei minimi quadrati. Questo metodo si realizza in $\mathbb{R}^n$ (dove n è l'ampiezza del campione). Si cerca di minimizzare la distanza euclidea

$$\|\underline{y} - a_0 - a_1\underline{x}^{(1)} - \cdots - a_d\underline{x}^{(d)}\|$$

dove

$$\begin{aligned} \underline{y} &= (y_1, \ldots, y_n) \\ \underline{x}^{(1)} &= (x_1^{(1)}, \ldots, x_n^{(1)}) \\ &\cdots\cdots\cdots\cdots \\ \underline{x}^{(d)} &= (x_1^{(d)}, \ldots, x_n^{(d)}). \end{aligned}$$

Dalla teoria degli *spazi di Hilbert* (o dallo studio di $\mathbb{R}^n$) si sa che il minimo della distanza tra un punto ed un sottospazio lineare chiuso si realizza nella proiezione del punto sul sottospazio, in questo caso il sottospazio è quello generato dai vettori $\mathbb{1}, \underline{x}^{(1)}, \ldots, \underline{x}^{(d)}$. Inoltre si sa che se S è un sottospazio lineare chiuso di uno spazio di Hilbert $\mathcal{H}$ (si pensi pure al caso $\mathcal{H} = \mathbb{R}^n$ con l'usuale prodotto scalare) allora per ogni $y \in \mathcal{H}$ e $v \in S$ si ha

$$\|y - v\|^2 = \|y - P_S y\|^2 + \|P_S y - v\|^2,$$

dove P_S rappresenta l'operatore proiezione su S. Nel caso specifico la precedente equazione risulta

$$\|\underline{y} - \overline{y}\mathbb{1}\|^2 = \|\underline{y} - \widehat{\underline{y}}\|^2 + \|\widehat{\underline{y}} - \overline{y}\mathbb{1}\|^2.$$

Approfondimento

Alla forma matriciale del sistema di equazioni date dal sistema dei minimi quadrati si arriva in maniera semplice nel seguente modo. Se f e g sono due

funzioni a valori in uno spazio di Hilbert reale $\mathcal{H}$ (si pensi al solito caso $\mathcal{H} = \mathbb{R}^n$ con l'usuale prodotto scalare) entrambe F-differenziabili in x_0 (nel caso $\mathcal{H} = \mathbb{R}^n$ significa "differenziabili" in x_0) e se denotiamo con $J_F f(x_0, \cdot)$ e $J_F g(x_0, \cdot)$ i due differenziali allora $\langle f, g\rangle$ è F-differenziabile in x_0 e vale

$$J_F\langle f, g\rangle(x_0, h) = \langle J_F f(x_0, h), g(x_0)\rangle + \langle f(x_0), J_F g(x_0, h)\rangle$$

per ogni valore del vettore incremento h. Nel caso

$$f(\underline{a}) = g(\underline{a}) := \underline{y} - X\underline{a}$$

si ha pertanto $J_F f(a, \underline{h}) = -X\underline{h}$ e quindi il minimo è soluzione dell'equazione

$$0 = J_F\langle f, g\rangle(x_0, \underline{h}) =\equiv \langle X\underline{h}, \underline{y} - X\underline{a}\rangle, \qquad \forall \underline{h} \in \mathbb{R}^{n+1},$$

ma essendo $\langle X\underline{h}, \underline{y} - X\underline{a}\rangle = \langle \underline{h}, X^T\underline{y} - X^T X\underline{a}\rangle$ ed essendo $\langle \underline{h}, \underline{x}\rangle = 0$ per ogni $\underline{h}$ se e solo se $\underline{x} = \underline{0}$, si ha, equivalentemente,

$$X^T\underline{y} - X^T X\underline{a} = \underline{0}.$$

Cap. 3. Calcolo delle probabilità

Scopo: si vogliono ricavare dei modelli matematici per esperimenti aleatori.

3.1 Definizione assiomatica: spazi misurabili, misure, misure di probabilità

Incominciamo ad introdurre alcuni concetti astratti che ci saranno utili in seguito. In tutto questo paragrafo supporremo di avere un'insieme Ω che chiameremo **spazio campionario** oppure **spazio degli eventi elementari**. Utilizzeremo la solita notazione insiemistica: in particolare $A \setminus B := \{x \in A : x \notin B\}$ e, avendo lo spazio Ω in mente, per ogni $A \subseteq \Omega$ chiameremo $A^{\complement} := \Omega \setminus A$ il complementare di A.

Definizione 3.1.1. Chiamiamo **σ-algebra** su Ω una collezione $\mathcal{F}$ di sottoinsiemi di Ω soddisfacente le seguenti proprietà:

(i) $\emptyset \in \mathcal{F}$;

(ii) se $A \in \mathcal{F}$ allora $A^{\complement} \in \mathcal{F}$

(iii) se $\{A_i\}_{i\in\mathbb{N}}$ è una collezione di insiemi di $\mathcal{F}$ allora $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

Valgono le proprietà (provare per esercizio):

1. $\Omega \in \mathcal{F}$;
2. se $\{A_i\}_{i\in\mathbb{N}}$ è una collezione di insiemi di $\mathcal{F}$ allora $\bigcap_{i\in\mathbb{N}} A_i \in \mathcal{F}$;
3. se $\{A_i\}_{i=1}^{n}$ è una collezione di insiemi di $\mathcal{F}$ allora $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$;
4. se $\{A_i\}_{i=1}^{n}$ è una collezione di insiemi di $\mathcal{F}$ allora $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$;
5. se $A, B \in \mathcal{F}$ allora $A \setminus B \in \mathcal{F}$.

Definizione 3.1.2. Una coppia $(\Omega, \mathcal{F})$ dove $\mathcal{F}$ è una σ-algebra su Ω, prende il nome di **spazio misurabile** e gli elementi della σ-algebra prendono il nome di **insiemi misurabili**.

In particolare quando $\mathcal{A}$ è una collezione di sottoinsiemi di Ω esiste ed è unica la **σ-algebra generata** $\sigma(\mathcal{A})$ così definita:

$$\sigma(\mathcal{A}) := \bigcap_{\mathcal{B}\,\sigma\text{-algebre su }\Omega:\mathcal{B}\supseteq\mathcal{A}} \mathcal{B}.$$

Nel caso in cui lo spazio misurabile sia su $\mathbb{R}^n$ la σ-algebra sarà sempre quella generata dai rettangoli (nel caso $n = 1$ sono gli intervalli) detta **σ-algebra di Borel** e simboleggiata da $\mathcal{R}^n$.

Esempio 3.1.3. Dato un insieme Ω vi sono due σ-algebre immediatamente a disposizione:

- la σ-algebra banale $\{\emptyset, \Omega\}$,
- la σ-algebra totale costituita $\mathcal{P}(\Omega)$ da tutti i sottoinsiemi di Ω.

Definizione 3.1.4. Dati due spazi misurabili $(\Omega, \mathcal{F})$ ed $(\Omega_1, \mathcal{F}_1)$, una funzione $f : \Omega \to \Omega_1$ prende il nome di **funzione $\mathcal{F} - \mathcal{F}_1$–misurabile** (o più semplicemente *funzione misurabile*) se e solo se

$$f^{-1}(A) := \{\omega \in \Omega : f(\omega) \in A\} \in \mathcal{F}, \qquad \forall A \in \mathcal{F}_1.$$

Osservazione 3.1.5. Spesso nel seguito utilizzeremo le seguenti abbreviazioni:

$$\begin{array}{lll} \{f = a\} & \text{indica} & \{\omega \in \Omega : f(\omega) = a\} \\ \{a < f \leq b\} & \text{indica} & \{\omega \in \Omega : a < f(\omega) \leq b\} \\ \{f \in I\} & \text{indica} & \{\omega \in \Omega : f(\omega) \in I\} \end{array}$$

Si mostra immediatamente che se $\mathcal{F}_1 = \sigma(\mathcal{A})$ allora una funzione f è misurabile se e solo se $f^{-1}(A) \in \mathcal{F}$ per ogni $A \in \mathcal{A}$; pertanto per tutte le funzioni $f : \Omega \to \mathbb{R}$ la misurabilità (rispetto a $\mathcal{R}$) equivale a $f^{-1}(I) \in \mathcal{F}$ per ogni $I \subseteq \mathbb{R}$ intervallo (per semplicità potrete assumere questa come definizione, almeno nel caso reale); in particolare ogni funzione $f : \mathbb{R}^n \to \mathbb{R}^m$ continua risulta $\mathcal{R}^n - \mathcal{R}^m$–misurabile.

Una proprietà che si potrebbe dimostrare senza troppo sforzo è che data $f : \Omega \to \mathbb{R}^n$ di componenti $(f_1, \ldots, f_n)$ allora f è misurabile se e solo se lo sono $f_1, \ldots, f_n$.

Approfondimento

In realtà data una collezione di funzioni $\{f_\alpha\}_{\alpha \in I}$ definite su $(\Omega, \mathcal{F})$ a valori rispettivamente in $(\Omega_\alpha, \mathcal{F}_\alpha)$ allora esiste una σ-algebra, detta **σ-algebra prodotto** $\otimes_{\alpha \in I} \mathcal{F}_\alpha$ su $\prod_{\alpha \in I} \Omega_\alpha$ tale che la funzione da Ω a valori in $\prod_{\alpha \in I} \Omega_\alpha$

$$\omega \mapsto \{f_\alpha(\omega)\}_{\alpha \in I}$$

è misurabile se e solo se lo è f_α per ogni $\alpha \in I$. La σ-algebra di Borel $\mathcal{R}^n$ è la σ-algebra prodotto di n copie di $\mathcal{R}$.

Esercizio 3.1.6. Mostrare che data una funzione $f : \Omega \to \Omega_1$ e dei sottoinsiemi di Ω_1, $A, B, \{A_\alpha\}_{\alpha \in I}$, allora:

1. $$f\left(\bigcup_{\alpha \in I} A_\alpha\right) = \bigcup_{\alpha \in I} f\left(A_\alpha\right)$$
2. $$f^{-1}\left(\bigcup_{\alpha \in I} A_\alpha\right) = \bigcup_{\alpha \in I} f^{-1}\left(A_\alpha\right)$$
3. $$f^{-1}\left(A \setminus B\right) = f^{-1}\left(A\right) \setminus f^{-1}\left(B\right)$$
4. $$f^{-1}\left(\bigcap_{\alpha \in I} A_\alpha\right) = \bigcap_{\alpha \in I} f^{-1}\left(A_\alpha\right)$$
5. $$f^{-1}\left(A^{\complement}\right) = f^{-1}\left(A\right)^{\complement}$$
6. $$f\left(\bigcap_{\alpha \in I} A_\alpha\right) \subseteq \bigcap_{\alpha \in I} f\left(A_\alpha\right).$$

Data una funzione $f : \Omega \to \Omega_1$ e dove $(\Omega_1, \mathcal{F}_1)$ è uno spazio misurabile allora la più piccola σ-algebra su Ω che rende misurabile f è

$$f^{-1}(\mathcal{F}_1) := \{B \subseteq \Omega : \exists A \in \mathcal{F}_1 : B = f^{-1}(A)\};$$

tale σ-algebra acquisterà un particolare significato nel prossimo paragrafo. A tal proposito si può mostrare che data una collezione $\mathcal{A}$ di sottoinsiemi di Ω_1 la sigma algebra generata dalla collezione $f^{-1}(\mathcal{A})$ coincide con $f^{-1}(\sigma(\mathcal{A}))$ cioè $\sigma(f^{-1}(\mathcal{A})) = f^{-1}(\sigma(\mathcal{A}))$.

Introduciamo infine la cosiddetta **misura di probabilità**.

Definizione 3.1.7. Sia $(\Omega, \mathcal{F})$ uno spazio misurabile; una funzione $\mathbb{P} : \mathcal{F} \to [0, +\infty]$ si dice **misura di probabilità** se e solo se soddisfa le seguenti proprietà:

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

(ii) se $\{A_i\}_{i=1}^{\infty}$ è una collezione di insiemi misurabili tali che $i \neq j$ implica $A_i \cap A_j = \emptyset$ allora

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Valgono le seguenti proprietà:

- $\mathbb{P}(\emptyset) = 0$;
- se $\{A_i\}_{i=1}^{n}$ è una collezione di insiemi misurabili a due a due disgiunti (i.e. $i \neq j$ implica $A_i \cap A_j = \emptyset$) allora
$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i);$$

Approfondimento

Questa proprietà di *additività* non implica la (ii) della definizione precedente (detta σ-*additività*). L'esempio si può costruire in $\Omega = \mathbb{N}$ ispirandosi al Teorema 3.2.6 e alla teoria degli spazi di Banach (troppo complessa per essere trattata in queste note). Infatti si potrebbe mostrare che $(l^{\infty})^*$, il duale di l^{∞}, può essere identificato con lo spazio delle misure finite additive su $\mathbb{N}$, mentre l^1 può essere identificato con il sottoinsieme (proprio) di $(l^{\infty})^*$ contenente tutte e sole le misure σ-additive. Questi risultati sono legati alla teoria dei limiti di Banach.

- se $A, B \in \mathcal{F}$ sono tali che $A \supseteq B$ allora $\mathbb{P}(A) \geq \mathbb{P}(B)$;
- $\mathbb{P}(A) \in [0, 1]$ per ogni $A \in \mathcal{F}$;
- se $A, B \in \mathcal{F}$ sono tali che $A \supseteq B$ allora $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(B)$;
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
- $\{A_i\}_{i=1}^{n}$ una collezione finita di insiemi misurabili allora
$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{j=1}^{n} (-1)^{j+1} \sum_{1 \leq i_1 < \cdots < i_j \leq n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_j})$$
(**formula di Poincarè** o **principio di inclusione-esclusione**, per la dimostrazione si veda il Paragrafo 3.6);

- se $\{A_i\}_{i\in I}$ è una collezione al più numerabile di insiemi misurabili tali che $i \neq j$ implica $\mathbb{P}(A_i \cap A_j) = 0$) allora
$$\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i);$$
- Se $\{A_i\}_{i\in I}$ è una $\mathbb{P}$-partizione al più numerabile (i.e. (i) $A_i \in \mathcal{F}$ per ogni $i \in I$, (ii) $\mathbb{P}(\cup_{i\in I} A_i) = 1$ e (iii) se $i \neq j$ allora $\mathbb{P}(A_i \cap A_j) = 0$) si ha
$$\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(B \cap A_i)$$
dove $B \in \mathcal{F}$ (**formula delle probabilità totali**).
- Se gli eventi soddisfano $A_{i+1} \supseteq A_i$ (risp. $A_{i+1} \subseteq A_i$) per ogni $i \in \mathbb{N}$, allora
$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \uparrow \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \qquad \left(\text{risp. } \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \downarrow \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right)\right)$$
se $n \to +\infty$ (continuità della misura).
- Per qualsiasi famiglia di eventi $\{A_i\}_{i\in I}$ con I al più numerabile si ha
$$\mathbb{P}\Big(\bigcup_{i\in I} A_i\Big) \leq \sum_{i\in I} \mathbb{P}(A_i)$$
(si dimostra per induzione sulla cardinalità di I per I finito e quindi passando al limite ed utilizzando la continuità della misura). In particolare si ha che se $\mathbb{P}(A_i) = 0$ (risp. $\mathbb{P}(A_i) = 1$) per ogni $i \in I$ allora
$$\mathbb{P}\Big(\bigcup_{i\in I} A_i\Big) = 0, \qquad \mathbb{P}\Big(\bigcap_{i\in I} A_i\Big) = 1.$$

Approfondimento

Una misura di probabilità è un particolare caso di **misura positiva** secondo la seguente definizione.

Definizione 3.1.8. Sia $(\Omega, \mathcal{F})$ uno spazio misurabile; una funzione $\mathbb{P} : \mathcal{F} \to [0, +\infty]$ si dice **misura positiva** se e solo se soddisfa le seguenti proprietà:

(i) $\mathbb{P}(\emptyset) = 0$

(ii) se $\{A_i\}_{i=1}^{\infty}$ è una collezione di insiemi misurabili tali che $i \neq j$ implica $A_i \cap A_j = \emptyset$ allora
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Per mostrare che questa definizione è più generale (per esercizio) si dimostri innanzitutto che dalla proprietà (ii) si ha che $\mathbb{P}(\emptyset) \in \{0, +\infty\}$ e che se $\mathbb{P}(\Omega) = 1$ allora necessariamente $\mathbb{P}(\emptyset) = 0$; quindi se $\mathbb{P}$ soddisfa le richieste della Definizione 3.1.7 allora deve necessariamente soddisfare quelle della Definizione 3.1.8.
Si osservi inoltre che data una serie a termini non negativi $\sum_{i=1}^{\infty} \alpha_i$ per ogni funzione biettiva $\phi : \mathbb{N} \to \mathbb{N}$ si ha
$$\sum_{i=1}^{\infty} \alpha_i = \sum_{i=1}^{\infty} \alpha_{\phi(i)} \in \mathbb{N} \cup \{+\infty\}$$
convergenza e divergenza incondizionata.

Data una funzione misurabile $f : \Omega \to \Omega_1$, allora la funzione

$$\mathbb{P}_f(A) := \mathbb{P}(f^{-1}(A)), \qquad \forall A \in \mathcal{F}_1$$

risulta essere una misura di probabilità su $(\Omega_1, \mathcal{F}_1)$ (provare per esercizio) che prende il nome di **legge di** f. Se la funzione misurabile è a valori in $\mathbb{R}^n$, sia $f := (f_1, \dots, f_n)$ allora la sua legge prende il nome di **legge congiunta** delle funzioni $f_1, \dots, f_n$. Se $g : \Omega_1 \to \Omega_2$ (dove $(\Omega_2, \mathcal{F}_2)$ è un nuovo spazio misurabile) è un altra funzione misurabile allora $g \circ f$ risulta misurabile e la legge $\mathbb{P}_{g\circ f} = (\mathbb{P}_f)_g$. La legge congiunta di $f_1, \dots, f_n$ è univocamente determinata dai suoi valori

$$\mathbb{P}_{(f_1,\dots,f_n)}(I_1 \times \cdots \times I_n) \equiv \mathbb{P}\left(\bigcap_{i=1}^{n} f_i^{-1}(I_i)\right), \qquad \forall I_1, \dots, I_n \subseteq R \text{ intervalli.}$$

Definizione 3.1.9. Uno spazio misurabile $(\Omega, \mathcal{F})$ è detto **discreto** se e solo se Ω è al più numerabile. In tal caso, se non esplicitamente detto, considereremo $\mathcal{F} := \mathcal{P}(\Omega)$.
Una misura di probabilità $\mathbb{P}$ su uno spazio misurabile $(\Omega, \mathcal{F})$ si dice **misura discreta** se e solo se esiste un sottoinsieme misurabile $S \in \mathcal{F}$ al più numerabile tale che $\mathbb{P}(S) = 1$. Se $\mathbb{P}$ è una misura di probabilità su $(\Omega, \mathcal{F})$ e $(\Omega_1, \mathcal{F}_1)$ è un altro spazio misurabile, una funzione misurabile $f : \Omega \to \Omega_1$ si dice **discreta** se esiste un sottoinsieme $S \subseteq \Omega_1$ discreto tale che $\{f \in S\} \in \mathcal{F}$ e $\mathbb{P}(f \in S) = 1$ (o equivalentemente $\mathbb{P}_f(S) = 1$).

In particolare una funzione misurabile f è discreta se e solo se lo è la sua legge $\mathbb{P}_f$; inoltre se $(\Omega, \mathcal{F})$ è discreto, ogni misura di probabilità su di esso è discreta (ma esistono misure discrete definite su spazi non discreti), mentre se $(\Omega_1, \mathcal{F}_1)$ è discreto allora ogni funzione misurabile a valori in esso è discreta (ma esistono funzioni discrete definite su spazi non discreti).

Osservazione 3.1.10. Sia $\{f_i\}_{i\in I}$ una famiglia di funzioni misurabili tali che $f : \Omega_i \to \Omega$ dove $(\Omega, \mathcal{F})$ è uno spazio misurabile e gli elementi della collezione $\{(\Omega_i, \mathcal{F}_i, \mathbb{P}^{(i)})\}$ sono spazi di probabilità; si dice che le funzioni $\{f_i\}_{i\in I}$ sono **identicamente distribuite** se e solo se per ogni $i, j \in I$ le leggi soddisfano $\mathbb{P}^{(i)}_{f_i} = \mathbb{P}^{(j)}_{f_j}$.
Questo non significa che le variabili coincidano, infatti non è detto nemmeno che siano definite sullo stesso spazio.

Osservazione 3.1.11. Tutte le definizioni date, a proposito di insiemi misurabili e funzioni misurabili, nel presente paragrafo si estendono ai paragrafi successivi dove adotteremo un nuovo linguaggio e parleremo, rispettivamente, di eventi e variabili aleatorie.

3.2 Definizione assiomatica: eventi e variabili aleatorie

Un **esperimento aleatorio** è un esperimento che a priori può avere diversi esiti possibili, e il cui esito effettivo dipende dal caso.

Esempio 3.2.1.

1. (a) Si estraggono sei palline da un campione di 90 palline numerate progressivamente, e si guardano i numeri estratti.
   (b) Si entra in una classe di studenti e si conta il numero di assenti.
2. (a) Si lancia ripetutamente una moneta finché non esce testa; si conta il numero di lanci.
   (b) Si telefona ogni minuto a un determinato numero finché non lo si trova libero. Si conta il numero di tentativi.
3. (a) Si accende una lampadina e si misura il suo tempo di vita.
   (b) Si misura l'altezza di un individuo scelto a caso in un gruppo di persone.

Lo spazio su cui ambientiamo i possibili stati che determinano gli esiti di un esperimento aleatorio è uno **spazio di probabilità**, cioè uno spazio misurabile con una misura di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. L'insieme dei possibili stati è Ω i cui elementi a volte prendono il nome di **eventi elementari**; lo spazio Ω prende il nome di **spazio campionario** o **spazio degli eventi elementari**.

Gli insiemi della σ-algebra $\mathcal{F}$ prendono il nome di **eventi**. Si osservi che in generale un evento elementare non è un evento (in quanto non è un sottoinsieme di Ω ma un elemento di Ω); in generale, dato un evento elementare $\omega \in \Omega$, non è detto nemmeno che il singoletto $\{\omega\} \in \mathcal{F}$ e quindi che $\{\omega\}$ sia un evento. Quando i singoletti appartengono tutti ad $\mathcal{F}$ (caso molto frequente), li chiameremo ancora eventi elementari abusando un po' della nomenclatura.

La nostra interpretazione è: il caso "pesca" un evento elementare $\omega \in \Omega$; noi diremo che un evento $A \in \mathcal{F}$ accade in corrispondenza alla scelta di ω se e solo se $\omega \in A$. La probabilità dell'evento A sarà $\mathbb{P}(A)$.

Esempio 3.2.2. Sia $\Omega := \{(0,0), (0,1), (1,0), (1,1)\}$ dove (a,b) lo interpretiamo come il risultato di due lanci di una moneta (0 è testa, 1 è croce). Scegliamo $\mathcal{F} := \mathcal{P}(\Omega)$ e come misura di probabilità $\mathbb{P}(A) := \#A/4$. Allora l'evento "al primo lancio esce croce" sarà $A := \{(1,0), (1,1)\}$ e la sua probabilità è $1/2$. L'evento $\{(0,0)\}$ si interpreta come "in entrambi i lanci è uscita testa" ed ha probabilità $1/4$, mentre l'evento $\{(1,0), (0,1), (1,1)\}$ è l'evento "in almeno un lancio è uscita croce" ed ha probabilità $3/4$.

Lo spazio campionario è quindi detto **discreto**, in accordo con la Definizione 3.1.9, se i suoi elementi sono in numero finito oppure un'infinità numerabile (es. 1 e 2). È detto **continuo** se è più numeroso, ad esempio $\mathbb{R}$ o un suo intervallo (es. 3).

Negli esempi 1a e 1b gli eventi elementari sono in numero finito.
Negli esempi 2a e 2b sono un'infinità numerabile ($\Omega = \mathbb{N}$).
Negli esempi 3a e 3b sono un'infinità non numerabile ($\Omega = \mathbb{R}$ o un intervallo di $\mathbb{R}$).

Esempi di eventi.
1a: le palline estratte hanno numeri progressivi contigui.
1b: non vi sono più di 3 assenti.
2a: si ottiene testa dopo non meno di 10 lanci e non più di 20.
2b: non si aspetta più di 10 minuti.
3a: la lampadina dura almeno 300 ore.
3b: la persona misura meno di 1.80 metri.

Rappresentazione insiemistica degli eventi.

Linguaggio degli insiemi	*Linguaggio degli eventi*
Ω	evento certo
$\emptyset$	evento impossibile
insieme A	si verifica l'evento A
insieme $A^{\complement}$	non si verifica A
$A \cup B$	si verificano A o B (qui "o" ha il significato latino di *vel*, cioè almeno uno dei eventi due si verifica)
$\bigcup_{\alpha \in I} A_\alpha$	almeno uno degli eventi della collezione $\{A_\alpha\}_{\alpha \in I}$ si verifica
$A \cap B$	si verificano sia A che B
$\bigcap_{\alpha \in I} A_\alpha$	tutti gli eventi della collezione $\{A_\alpha\}_{\alpha \in I}$ si verificano
$A \setminus B$	si verifica A e non si verifica B
$A \cap B = \emptyset$	A e B sono incompatibili
$B \subseteq A$	B implica A
$A \triangle B := (A \setminus B) \cup (B \setminus A)$	uno ed uno solo dei due eventi si verifica (è il latino *aut*

Alcune proprietà degli insiemi; A, B e C sono sottoinsiemi qualsiasi di Ω:

$A \cup A = A$	idempotenza dell'unione
$A \cap A = A$	idempotenza dell'intersezione
$A \cup \emptyset = A$	
$A \cap \emptyset = \emptyset$	
$A \cup \Omega = \Omega$	
$A \cap \Omega = A$	
$A \cup A^{\complement} = \Omega$	
$A \cap A^{\complement} = \emptyset$	
$A \cup B = B \cup A$	commutatività dell'unione
$A \cap B = B \cap A$	commutatività dell'intersezione
$A \cup (B \cup C) = (A \cup B) \cup C$	associatività dell'unione
$A \cap (B \cap C) = (A \cap B) \cap C$	associatività dell'intersezione
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	distributività dell'unione risp. intersez.
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	distributività dell'intersez. risp. unione
$(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}$	legge di De Morgan
$(A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}$	legge di De Morgan
$(A^{\complement})^{\complement} = \mathrm{A}$	

Quindi in relazione a un evento siamo interessati a calcolarne la probabilità. Nei nostri esempi:
1a: prob. che le palline estratte abbiano numeri progressivi contigui.
1b: prob. che non vi siano più di 3 assenti.
2a: prob. che si ottenga testa dopo non meno di 10 lanci e non più di 20.
2b: prob. di aspettare non più di 10 minuti.
3a: prob. che la lampadina duri almeno 300 ore.
3b: prob. che la persona misuri meno di 1.80 metri.

Esercizio 3.2.3. Supponiamo che i pezzi prodotti da una certa macchina possano presentare due tipi di difetti, che chiameremo a e b. È stato stabilito che la probabilità che un pezzo presenti il difetto a è 0.1, la probabilità che non presenti il difetto b è 0.8, la probabilità che presenti entrambi i difetti è 0.01.
Qual è la probabilità che un pezzo non presenti alcun difetto?

Soluzione.
Indichiamo con A l'evento *il pezzo presenta il difetto a* e con B l'evento *il pezzo presenta il difetto b*. Le informazioni si traducono in: $\mathbb{P}(A) = 0.1$, $\mathbb{P}(B^{\complement}) = 0.8$, $\mathbb{P}(A \cap B) = 0.01$.
L'evento richiesto è l'evento $A^{\complement} \cap B^{\complement} = (A \cup B)^{\complement}$. Pertanto, dalla Definizione 3.1.7 e proprietà seguenti, si ha

$$\mathbb{P}(A^{\complement} \cap B^{\complement}) = \mathbb{P}((A \cup B)^{\complement}) = 1 - \mathbb{P}(A \cup B) = 1 - [\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)] =$$

$$= \mathbb{P}(B^{\complement}) + \mathbb{P}(A \cap B) - \mathbb{P}(A) = 0.8 + 0.01 - 0.1 = 0.71$$

Definizione 3.2.4. Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ ed uno spazio misurabile $(\Omega_1, \mathcal{F}_1)$ (generalmente $(\mathbb{R}^n, \mathcal{R}^n)$), conveniamo di chiamare ogni funzione $\mathcal{F} - \mathcal{F}_1$–misurabile una **variabile aleatoria** (o più semplicemente **variabile**) a valori in Ω_1. Tutte le definizioni inerenti alle funzioni misurabili si estendono quindi alle variabili aleatorie.

Osservazione 3.2.5. L'evento $\emptyset$ prende il nome di **evento impossibile**, mentre l'evento Ω prende il nome di **evento certo**. Sarà chiaro nel Capitolo 5 che possono esistere eventi $A \in \mathcal{F}$ non impossibili (i.e. $A \neq \emptyset$) tali che $\mathbb{P}(A) = 0$ (di conseguenza esisteranno anche eventi non certi (i.e. $A \neq \Omega$) tali che $\mathbb{P}(A) = 1$ detti eventi **quasi certi**). Matematicamente questi eventi non creano problemi, la loro interpretazione sarà quella di eventi più improbabili di qualsiasi evento di probabilità strettamente positiva, ma comunque possibili.

Operativamente la scelta di includere certi eventi nella σ-algebra dipenderà dalla interpretazione nell'ambito dell'esperimento che stiamo modellizzando: se rappresentano un evento che risulta teoricamente (per ragioni fisiche o di altra natura) impossibile allora possono essere eliminati.

Approfondimento

Il problema della scelta dello spazio campionario. Supponiamo di avere degli spazi misurabili $\{(\Omega_\alpha, \mathcal{F}_\alpha)\}_{\alpha\in I}$, una famiglia di leggi $\{\mathbb{P}_S\}_{S\subseteq I: S \text{ finito}}$, tali che $\left(\prod_{\alpha\in S}\Omega_\alpha, \otimes_{\alpha\in S}\mathcal{F}_\alpha, \mathbb{P}_S\right)$ è uno spazio di probabilità; qui con

$$\prod_{\alpha\in I}\Omega_\alpha := \{f : I \to \bigcup_{\alpha\in I}\Omega_\alpha : f(\alpha)\in\Omega_\alpha,\ \forall\alpha\in I\}$$

si intende il prodotto cartesiano degli spazi $\{\Omega_\alpha\}_{\alpha\in I}$ e $\otimes_{\alpha\in I}\mathcal{F}_\alpha$ è una particolare σ-algebra su $\prod_{\alpha\in I}\Omega_\alpha$ detta *σ-algebra prodotto* di $\{\mathcal{F}_\alpha\}_{\alpha\in I}$. Sotto alcune ipotesi di consistenza (date dal **Teorema di Kolmogorov** che qui non affrontiamo), l'interpretazione è che esse siano le leggi congiunte di sottofamiglie finite di variabili aleatorie. Quello che il teorema dimostra è che esiste un'unica misura di probabilità $\mathbb{P}$ sullo spazio $\left(\prod_{\alpha\in I}\Omega_\alpha, \otimes_{\alpha\in I}\mathcal{F}_\alpha\right)$ che "estende" la famiglia di leggi $\{\mathbb{P}_S\}_{S\subseteq I: S \text{ finito}}$; inoltre è possibile costruire su questo spazio la famiglia di variabili aleatorie $\{Z_\alpha\}_{\alpha\in i}$ di cui queste sono le leggi congiunte. Se ora $(\Omega, \mathcal{F}, \overline{\mathbb{P}})$ è un altro spazio in cui vive una famiglia di variabili $\{X_\alpha\}_{\alpha\in I}$; tutte gli eventi di cui possiamo calcolare la probabilità relativi a questa famiglia, sono nelle σ-algebre, $\otimes_{\alpha\in i}\mathcal{F}_\alpha$ e $\mathcal{F}_1 \subseteq \mathcal{F}$, generate da eventi del tipo

$$\{Z_{\alpha_1}\in E_1, \dots, Z_{\alpha_n}\in E_n\} \qquad n\in\mathbb{N},\ \alpha_i\in I\ \forall i=1,\dots,n,\ E_i\in\mathcal{F}_{\alpha_i}\ \forall i=1,\dots,n$$
$$\{X_{\alpha_1}\in E_1, \dots, X_{\alpha_n}\in E_n\} \qquad n\in\mathbb{N},\ \alpha_i\in I\ \forall i=1,\dots,n,\ E_i\in\mathcal{F}_{\alpha_i}\ \forall i=1,\dots,n$$

per cui vale

$$\mathbb{P}(\{Z_{\alpha_1}\in E_1, \dots, Z_{\alpha_n}\in E_n\}) = \mathbb{P}_S(E_1\times\cdots\times E_n) = \overline{\mathbb{P}}(\{X_{\alpha_1}\in E_1, \dots, X_{\alpha_n}\in E_n\})$$

dove $S := \{\alpha_1, \dots, \alpha_n\}$. Si dimostra ancora che esiste una funzione misurabile $J : \Omega \to \prod_{\alpha\in I}\Omega_\alpha$ tale che

$$J^{-1}(\{Z_{\alpha_1}\in E_1, \dots, Z_{\alpha_n}\in E_n\}) = \{X_{\alpha_1}\in E_1, \dots, X_{\alpha_n}\in E_n\}.$$

Si conclude quindi che $\mathcal{F}_1 = J^{-1}(\otimes_{\alpha\in I}\mathcal{F}_\alpha)$ ed $\mathbb{P} = \overline{\mathbb{P}}_J$, pertanto non importa la realizzazione specifica dello spazio di probabilità che si sceglie ma solo la legge. Nel seguito spesso non faremo riferimento ad alcuno spazio di probabilità, ma solo alla legge.

Interpretazione. A priori il risultato di un esperimento si può modellizzare con una variabile aleatoria X definita sullo spazio degli eventi elementari Ω che rappresenta l'insieme degli stati che contengono tutte le informazioni necessarie al mio esperimento; nei casi più semplici il risultato dell'esperimento sarà modellizzabile più semplicemente con un evento (esperimento a due risultati). Se lo stato del sistema è $\omega\in\Omega$ allora, a posteriori, il risultato dell'esperimento sarà $X(\omega)$. Viceversa, se sappiamo che l'esperimento ha dato un risultato nell'intervallo I allora lo stato del sistema ω appartiene a $X^{-1}(I)$. In questo modo appare chiaro che lo stato ω può essere osservato, in generale, solo attraverso la conoscenza dei risultati di uno o più esperimenti.

3.2.1 Come si assegnano le probabilità

Ora sappiamo cos'è uno spazio di probabilità; ma come si sceglie lo spazio giusto, o meglio, la legge giusta per rappresentare il fenomeno che si sta studiando? Questo problema, di natura strettamente applicativa, è simile a quello che si incontra continuamente nello studio delle scienze fisiche o naturali.

Negli spazi discreti (sotto l'ipotesi che la σ-algebra sia quella totale, i.e. ogni sottoinsieme è un evento) c'è una procedura che permette di costruire tutte le misure possibili a partire dalla determinazione del valore della misura su alcuni eventi speciali.

Sia quindi Ω uno spazio discreto e $\{\omega_i\}_{i\in I}$ (dove $I \subseteq \mathbb{N}$), $i = 1, \dots,$ gli eventi elementari. Ogni evento A può essere visto come unione finita o infinita (numerabile) di eventi elementari (e perciò disgiunti). Allora, dalla Definizione 3.1.7,

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i:\omega_i\in A} \{\omega_i\}\right) = \sum_{i:\omega_i\in A} \mathbb{P}(\{\omega_i\})$$

Quindi se conosciamo le probabilità $p_i = \mathbb{P}(\{\omega_i\})$ degli eventi elementari, risulta completamente definita la funzione di probabilità su Ω.

Vale infatti il seguente Teorema.

Teorema 3.2.6. *Se $\{p_i\}_{i\in I}$ è una successione di numeri positivi allora*

$$\mathbb{P}(A) := \sum_{i:\omega_i\in A} p_i$$

definisce una misura di probabilità su $(\Omega, \mathcal{P}(\Omega))$. Tale misura è l'unica con la proprietà $\mathbb{P}(\{\omega_i\}) = p_i$ per ogni $i \in I$.

Esempio 3.2.7.

$$\Omega = \mathbb{N}, \quad p_i = \frac{1}{2^i}, \quad i = 1, 2, \dots$$

Verifichiamo che p_i definisce una probabilità:

$$0 \le p_i \le 1, \quad \mathbb{P}(\Omega) = \sum_{i=1}^{\infty} p_i = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = \frac{1}{1-1/2} - 1 = 1$$

Calcoliamo ad esempio la probabilità dell'evento A *numero pari*:

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(\{2i\}) = \sum_{i=1}^{\infty} p_{2i} = \sum_{i=1}^{\infty} \left(\frac{1}{4}\right)^i = \frac{1}{1-1/4} - 1 = \frac{1}{3}.$$

Vedremo in seguito che la misura appena introdotta descrive l'esperimento "quanti lanci di una moneta non truccata devo fare prima che esca testa per la prima volta?"

Esaminiamo di seguito due approcci possibili alla costruzione di particolari misure.

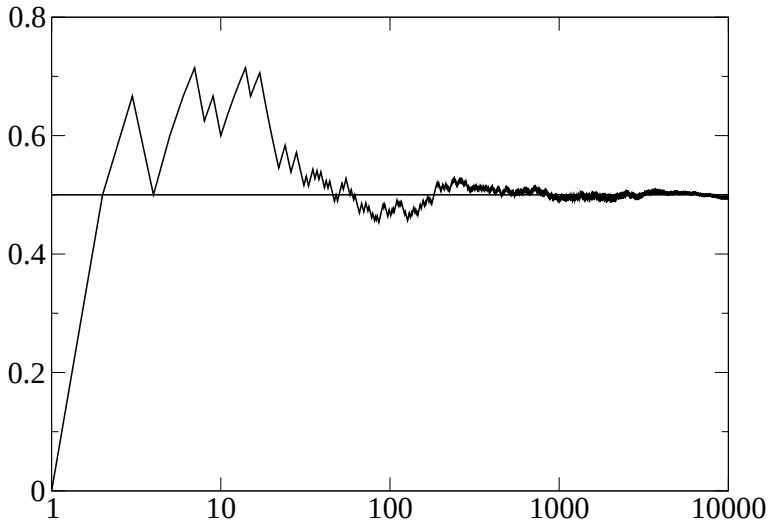
La probabilità classica. Consideriamo il caso in cui lo spazio campionario è *finito*. Facciamo l'ulteriore ipotesi che gli eventi elementari siano *equiprobabili*:

$$\Omega = \{\omega_1, \dots, \omega_n\}, \quad p_k = \frac{1}{n} \text{ per ogni } k = 1, \dots, n.$$

Seguendo lo schema generale per spazi discreti, la probabilità dell'evento A è

$$\mathbb{P}(A) = \sum_{\omega_i\in A} \frac{1}{n} = \frac{\#A}{n} = \frac{\#A}{\#\Omega}$$

dove $\#A$ rappresenta il numero degli eventi elementari che costituiscono l'evento A (detti *casi favorevoli* all'evento A).
Dunque la probabilità *classica* di un evento è *il rapporto tra il numero dei casi favorevoli e il numero dei casi possibili*.



Frequenza relativa dell'evento testa in una successione di lanci

Esempio 3.2.8. Estraiamo due palline da un'urna che contiene 60 palline bianche e 40 palline nere. In questo caso $n = C_{100,2} = 100 \cdot 99/2 = 4950$
($C_{n,k}$ è il *coefficiente binomiale* e rappresenta il numero di **combinazioni di** k **oggetti tra** n):

$$C_{n,k} = \binom{n}{k} = \frac{n(n-1)(n-2)\dots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

I modi possibili di estrarre 2 palline nere è $\#A = C_{40,2} = 40 \cdot 39/2 = 780$.
La probabilità che le due palline estratte siano nere è: $p = \#A/n = 0.158$.

L'idea frequentista di probabilità. La probabilità dell'evento A è il limite della frequenza relativa con cui A si verifica in una lunga serie di prove ripetute sotto condizioni simili.
Da questo punto di vista la probabilità è dunque una frequenza relativa. La **legge forte dei grandi numeri** a cui faremo qualche cenno nel Paragrafo 7.2.3, in qualche modo giustifica questo approccio.

Esempio 3.2.9. Si lancia una moneta n volte e si considera la frequenza relativa dell'evento *Testa* (numero di volte in cui si presenta T diviso per n). All'aumentare di n tale frequenza relativa tende a stabilizzarsi intorno al valore limite 0.5, che è la probabilità di T.

3.3 Probabilità condizionata

Ci chiediamo quale sia la probabilità di un evento A nell'ipotesi che l'evento B si verifichi.

Definizione 3.3.1. Sia B un evento con $\mathbb{P}(B) > 0$. Si chiama **probabilità di** A **condizionata a** B il numero

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

- Nel caso della probabilità classica (ossia di n eventi elementari equiprobabili con $p_i = 1/n$), dato $B \neq \emptyset$ si ha:
$$\mathbb{P}(A|B) = \frac{\frac{\#(A \cap B)}{\#\Omega}}{\frac{\#B}{\#\Omega}} = \frac{\#(A \cap B)}{\#B}$$
Si considera B come nuovo spazio campionario e si fa riferimento solo agli eventi elementari che appartengono sia ad A che a B.

- La mappa $A \mapsto \mathbb{P}(A|B)$, fissato B, è effettivamente una probabilità. Infatti $\mathbb{P}(A|B) \geq 0$; inoltre $\mathbb{P}(\Omega|B) = 1$ poiché $\Omega \cap B = B$; infine data una successione di eventi incompatibili A_i,
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i|B\right) = \frac{\mathbb{P}((\cup_{i=1}^{\infty} A_i) \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}((A_1 \cap B) \cup (A_2 \cap B) \cup \dots)}{\mathbb{P}(B)}$$
Essendo $\{A_i \cap B\}_{i=1}^{\infty}$ incompatibili, $\mathbb{P}((A_1 \cap B) \cup (A_2 \cap B) \dots) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)$. Pertanto
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i|B\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

- La mappa $B \mapsto \mathbb{P}(A|B)$, fissato A, non è una probabilità.
Ad esempio, mentre è vero che $\mathbb{P}(\Omega|B) = 1$, in generale $\mathbb{P}(A|\Omega) = \mathbb{P}(A)$ che può essere strettamente minore di 1.

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ per ogni coppia di eventi A, B tali che $\mathbb{P}(B) > 0$ (altrimenti, ricordiamo, la probabilità condizionata $\mathbb{P}(A|B)$ non è definita).

- In generale, se $\mathbb{P}(B) \in (0, 1)$,
$$\frac{\mathbb{P}(A)}{\max(\mathbb{P}(B), \mathbb{P}(B^{\complement}))} \leq \mathbb{P}(A|B) + \mathbb{P}(A|B^{\complement}) \leq \frac{\mathbb{P}(A)}{\min(\mathbb{P}(B), \mathbb{P}(B^{\complement}))}$$

- Se lo spazio campionario è finito (e gli eventi elementari sono equiprobabili), la definizione data sopra trova una piena giustificazione: poiché l'evento B si è verificato, si tratta di determinare la probabilità dell'evento $A \cap B$ prendendo come nuovo spazio campionario l'insieme B. Agli eventi elementari di A si attribuiscono nuove probabilità $\pi_i = p_i/p(B)$. Si ha dunque:
$$\mathbb{P}(A|B) = \sum_{A \cap B} \pi_i = \frac{\sum_{A \cap B} p_i}{\sum_B p_i} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- Se $\{A_i\}_{i \in I}$ (con $I \subseteq \mathbb{N}$) è una $\mathbb{P}$-partizione di Ω tale che $\mathbb{P}(A_i) > 0$ per ogni $i \in I$ e B è un evento qualsiasi allora
$$\mathbb{P}(B) = \sum_{j \in I} \mathbb{P}(B|A_j)\mathbb{P}(A_j)$$
(si veda la formula delle probabilità totali).

- Se $\{A_i\}_{i \in I}$ (con $I \subseteq \mathbb{N}$) è una $\mathbb{P}$-partizione di Ω tale che $\mathbb{P}(A_i) > 0$ per ogni $i \in I$ e B è un evento con probabilità $\mathbb{P}(B) > 0$ allora vale, per ogni $i \in I$ fissato,
$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j \in I} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$
(**formula di Bayes**).

- *Esempio* 3.3.2. Una confezione contiene 25 transistors di buona qualità, 10 difettosi (cioè che si rompono dopo qualche ora), e 5 guasti. Un transistor viene scelto a caso e messo in funzione. Sapendo che non è guasto, qual è la probabilità che sia di buona qualità?
Evento A: il transistor scelto a caso è di buona qualità.
Evento B: il transistor scelto a caso è difettoso.
Evento C: il transistor scelto a caso è guasto.
$\mathbb{P}(A) = 25/40$, $\mathbb{P}(B) = 10/40$, $\mathbb{P}(C) = 5/40$.

$$\mathbb{P}(A|C^{\complement}) = \frac{\mathbb{P}(A \cap C^{\complement})}{\mathbb{P}(C^{\complement})} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(C)} = \frac{25/40}{35/40} = \frac{5}{7}$$

- *Esercizio* 3.3.3. Problema delle tre carte: supponiamo di avere tre carte da gioco, una con faccia rossa e l'altra nera, una con entrambe le facce rosse e una con entrambe le facce nere. Si estrae una carta a caso e la si mette sul tavolo. Se la faccia visibile è rossa, qual è la probabilità che la faccia coperta sia rossa?

  Soluzione.
  Sia A l'evento *la faccia coperta è rossa*.
  Sia B l'evento *la faccia visibile è rossa*.
  Dobbiamo calcolare $\mathbb{P}(A|B)$.
  L'evento $A \cap B$ è l'evento *abbiamo scelto la carta con entrambe le facce rosse* la cui probabilità è pari a $1/3$.
  L'evento B ha probabilità $1/2$ poichè vi sono in totale tante facce rosse quante facce nere. Quindi

$$\mathbb{P}(B|A) = \frac{1/3}{1/2} = \frac{2}{3}$$

- *Esercizio* 3.3.4. Quando piove Adalfonso esce di casa con probabilità $1/3$, quando il tempo è sereno esce con probabilità $2/3$. Sapendo che piove con probabilità $1/4$, calcolare:

  1. la probabilità che in un giorno qualsiasi Adalfonso esca di casa;
  2. la probabilità che oggi piova dato che è uscito di casa.

  Soluzione.
  Sia $P :=$"oggi piove" e $U :=$"Adalfonso oggi esce di casa", allora $\mathbb{P}(P) = 1/4$ e $\mathbb{P}(U|P) = 1/3$ mentre $\mathbb{P}(U|P^{\complement}) = 2/3$. Dalla formula delle probabilità totali

$$\mathbb{P}(U) = \mathbb{P}(U|P)\mathbb{P}(P) + \mathbb{P}(U|P^{\complement})\mathbb{P}(P^{\complement}) = \frac{1}{3}\frac{1}{4} + \frac{2}{3}\frac{3}{4} = \frac{7}{12},$$

  da cui direttamente (oppure utilizzando la formula di Bayes),

$$\mathbb{P}(P|U) = \frac{\mathbb{P}(U|P)\mathbb{P}(P)}{\mathbb{P}(U)} = \frac{1/12}{7/12} = \frac{1}{7}.$$

- *Esercizio* 3.3.5. In un gioco televisivo viene messo in palio un 1 milione di euro. Per vincerlo il concorrente dovrà indovinare fra tre buste qual è quella che contiene l'assegno. Il concorrente sceglie a caso una busta; a questo punto il conduttore mostra una delle due buste che sa essere vuota, offrendo al concorrente di cambiare la propria busta con quella rimanente.

  Qual è la probabilità di vincere il premio conservando la prima busta scelta?

Qual è la probabilità di vincere cambiando la busta?

Qual è la probabilità di vincere se gioca a testa e croce fra le due strategie?

Soluzione.
Cominciamo con un ragionamento intuitivo. Il presentatore può sempre aprire una busta vuota, quindi l'apertura della busta vuota non cambia il contenuto di quella scelta dal concorrente; pertanto visto che la probabilità a priori di scegliere la busta contenente la promessa di pagamento è 1/3, se il concorrente decide di conservare la prima busta scelta, la probabilità di vincere è 1/3. Con la seconda strategia, consistente nel cambiare la busta che si ha in mano con la busta rimanente dopo che il conduttore ne ha mostrata una vuota, il concorrente vince se e solo se inizialmente ha scelto una delle due buste vuote. Pertanto, con la strategia del cambio della busta, la probabilità di vincere è pari a 2/3.

Rendiamo più rigoroso il ragionamento. Sia $W = \{$Il concorrente sceglie la busta vincente$\}$ e $P = \{$il presentatore apre una busta vuota$\}$. Chiaramente $\mathbb{P}(W) = 1/3$ e $\mathbb{P}(P) = 1$, pertanto V è indipendente da ogni altro evento (incluso W); infatti $\mathbb{P}(W)\mathbb{P}(P) = \mathbb{P}(W) = \mathbb{P}(W \cap P) + \mathbb{P}(W \cap P^{\complement}) = \mathbb{P}(W \cap P)$, poiché $0 \leq \mathbb{P}(W \cap P^{\complement}) \leq \mathbb{P}(P^{\complement}) = 0$. Quindi $\mathbb{P}(W|P) = \mathbb{P}(W \cap P)/\mathbb{P}(P) = \mathbb{P}(W) = 1/3$. Quindi se non cambia la busta vince con probabilità 1/3, il che implica che se cambia la busta vince con probabilità 2/3.

Per l'ultimo punto, poniamo $T = \{$Esce testa$\}$, $V = \{$Il concorrente vince$\}$ e supponiamo che, se esce testa, il concorrente sceglie la prima strategia, ovvero non cambia la busta. Se gioca a testa o croce fra le due strategie abbiamo, per la formula delle probabilità totali, che:

$$\mathbb{P}(V) = \mathbb{P}(V|T)\mathbb{P}(T) + \mathbb{P}(V|T^{\complement})\mathbb{P}(T^{\complement}) = 1/3 * 1/2 + 2/3 * 1/2 = 1/2.$$

•

Esercizio 3.3.6. Un signore ha due figli e supponiamo che il sesso di ciascuno dei sue sia indipendente da quello dell'altro e che la probabilità che nasca un maschio sia pari ad 1/2.

1. Ci dice di avere almenu un maschio, qual è la probabilità di avere due maschi?
2. Lo incontriamo in giro con uno dei suoi figli e vediamo che è un maschio, qual è la probabilità che entrambi siano maschi?

Soluzione.

1. Sia $\Omega := \{(M,M),(M,F),(F,M),(F,F)\}$ con la probabilità uniforme. L'evento "almeno uno dei due è un maschio è $\{(M,M),(M,F),(F,M)\} =: A$ e $\mathbb{P}(A) = 3/4$, l'evento entrambi sono maschi è $\{(M,M)\} =: B$ da cui
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{1}{3}.$$
2. Sia $\Omega_1 := \{1,2\} \times \{(M,M),(M,F),(F,M),(F,F)\}$ con la probabilità uniforme e sia Y così definita:
$$Y(i, w_1, w_2) := w_i$$
che rappresenta il sesso del figlio che incontro. Quindi l'evento "incontro un figlio maschio" è $\{(1,M,M),(2,M,M),(1,M,F),(2,F,M)\} =: A$ $(\mathbb{P}(A) = 1/2))$, mentre l'evento "entrambi i figli sono maschi è $\{(1,M,M),(2,M,M)\}$ $(\mathbb{P}(B) = 1/4)$. Pertanto
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{1}{2}.$$

Una seconda soluzione è la seguente: siano X e Y due variabili aleatorie equidistribuite con range rispettivamente in $\{0,1\}$ e $\{\omega_1 := (0,0), \omega_2 := (0,1), \omega_3 := (1,0), \omega_4 := (1,1)\}$ (con la convenzione 0 maschio, 1 femmina) e con probabilità condizionate

$$\mathbb{P}(X=0|Y=\omega_i) := \begin{cases} 1 & i=1 \\ 1/2 & i \in \{2,3\} \\ 0 & i=4. \end{cases}$$

La variabile X ci dice il sesso del figlio che vediamo e Y ci dice la coppia dei sessi dei due figli (nell'ordine). L'ambientazione esiste per il Teorema di Kolmogorov e la formula di Bayes ci da

$$\mathbb{P}(Y=\omega_1|X=0) = \frac{\mathbb{P}(X=0|Y=\omega_1)\mathbb{P}(Y=\omega_1)}{\sum_{i=1}^4 \mathbb{P}(X=0|Y=\omega_i)\mathbb{P}(Y=\omega_i)} = \frac{1/4}{1/4+2\cdot 1/4\cdot 1/2} = \frac{1}{2}.$$

Approfondimento

Cerchiamo di capire meglio e, al tempo stesso, generalizzare l'interpretazione della probabilità condizionata. Supponiamo di avere una $\mathbb{P}$-partizione $\{A_i\}_{i\in I}$. A priori si ha $\mathbb{P}(\cdot) = \sum_{i\in I} \alpha_i \mathbb{P}(\cdot|A_i)$ con $\alpha_i = \mathbb{P}(A_i)$ (formula delle probabilità totali); in particolare, a priori, la probabilità di un evento B è quindi $\mathbb{P}(B) = \sum_{i\in I} \mathbb{P}(A_i)\mathbb{P}(B|A_i)$. L'informazione che A_{i_0} avviene, ci porta a modificare i pesi $\{\alpha_i\}_{i\in I}$ (precisamente $\alpha_{i_0} = 1$ ed $\alpha_i = 0$ per ogni $i \neq i_0$) ottenendo una misura di probabilità a posteriori $\overline{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot|A_{i_0})$.

Immagniamo ora di avere un'informazione differente, cioè che le probabilità degli eventi $\{A_i\}_{i\in I}$ sono date da una successione $\{\bar{\alpha}_i\}_{i\in I}$ (dove $\bar{\alpha}_i \geq 0$ per ogni $i \in I$ e $\sum_{i\in I} \bar{\alpha}_i = 1$). In questo caso utilizziamo questa nuova informazione ottenendo una misura a posteriori $\overline{\mathbb{P}}(\cdot) = \sum_{i\in I} \bar{\alpha}_i \mathbb{P}(\cdot|A_i)$. Questo generalizza il caso precedente.

Esercizio 3.3.7. Durante un'indagine si sa che un sospetto è colpevole con probabilità pari a 0.6. Il sospetto ha una certa caratteristica fisica che è condivisa dal 20% della popolazione. Ulteriori indagini mostrano che il colpevole ha la stessa caratteristica. Qual è la nuova probabilità di colpevolezza del sospetto? Se invece le nuove indagini mostrassero che la probabilità che il colpevole abbia la caratteristica è 0.9, e quella che il sospetto abbia la caratteristica 0.7, qual è la nuova probabilità di colpevolezza del sospetto in questo caso?

Soluzione.
Siano $G =$"il sospetto è colpevole" e $C =$"il sospetto ha la caratteristica". I dati del problema si traducono in $\mathbb{P}(G) = 0.6$, $\mathbb{P}(C|G) = 1$ e $\mathbb{P}(C|G^{\complement}) = 0.2$. Utilizzando la formula di Bayes si ha

$$\mathbb{P}(G|C) = \frac{1\cdot 0.6}{1\cdot 0.6 + 0.2\cdot 0.4} = \frac{15}{16} \approx 0.882$$
$$\mathbb{P}(G|C^{\complement}) = 0;$$

in particolare, a priori, $\mathbb{P}(C) = 1\cdot 0.6 + 0.2\cdot 0.4 = 0.68$. Applichiamo quando visto in precedenza dove la $\mathbb{P}$-partizione è $A_1 = C$ e $A_2 = C^{\complement}$.

Nel primo caso l'informazione mi porta ad utilizzare una nuova misura $\overline{\mathbb{P}}(\cdot) = \alpha_1\mathbb{P}(\cdot|C) + (1-\alpha)\mathbb{P}(\cdot|C^{\complement})$ dove $\alpha = 1$, pertanto la risposta è $\overline{\mathbb{P}}(G) = \mathbb{P}(G|C) = 15/17 \approx 0.882$.

Nel secondo caso invece $\alpha = 0.8$ pertanto $\overline{\mathbb{P}}(G) = 0.7\,\mathbb{P}(G|C) + 0.3\,\mathbb{P}(G|C^{\complement})$. I dati del problema si traducono in $\mathbb{P}(G) = 0.6$, $\mathbb{P}(C|G) = 0.9$ e $\mathbb{P}(C|G^{\complement}) = 0.2$.

Utilizzando la formula di Bayes si ha

$$\mathbb{P}(G|C) = \frac{0.9 \cdot 0.6}{0.9 \cdot 0.6 + 0.2 \cdot 0.4} = \frac{27}{31} \approx 0.871$$
$$\mathbb{P}(G|C^{\complement}) = \frac{0.1 \cdot 0.6}{0.1 \cdot 0.6 + 0.8 \cdot 0.4} = \frac{3}{19} \approx 0.158;$$

in particolare, a priori, $\mathbb{P}(C) = 0.9 \cdot 0.6 + 0.2 \cdot 0.4 = 0.62$. Pertanto $\overline{\mathbb{P}}(G) = 0.7 \cdot 27/31 + 0.3 \cdot 3/19 \approx 0.657$.

3.4 Indipendenza di eventi

Intuitivamente, due eventi A e B si dicono indipendenti se il verificarsi di uno dei due non modifica la probabilità che l'altro accada.

Definizione 3.4.1. Due eventi A e B si dicono **indipendenti** se e solo se

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

o equivalentemente (nel caso in cui $\mathbb{P}(A) > 0$)

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

e analogamente, sotto l'ipotesi $\mathbb{P}(B) > 0$,

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

Esempio 3.4.2. Un'urna contiene 6 palline rosse e 4 palline bianche. Ne estraggo una, ne guardo il colore, la reintroduco e ne estraggo una seconda. Qual è la probabilità che entrambe siano bianche?

Gli eventi B_i = *la i-esima pallina estratta è bianca* si considerano indipendenti. Pertanto $\mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1)\mathbb{P}(B_2) = \frac{4}{10} \times \frac{4}{10} = 0.16$.

Si noti che se l'estrazione fosse avvenuta senza reimmissione, i due eventi B_1 e B_2 non sarebbero stati più indipendenti.

Osservazione 3.4.3. Quando la misura di probabilità è fornita, allora due eventi sono indipendenti o meno in accordo alla Definizione 3.4.1. Quando invece si cerca di modellizzare un problema o un'esperimento, in generale la misura non è fornita, ma va costruita in base ai dati e ad alcune considerazioni: una di queste potrebbe essere, ad esempio, la richiesta di indipendenza di alcune coppie (o alcuni insiemi) di eventi.

Definizione 3.4.4. Gli eventi $\mathcal{A} := \{A_i\}_{i \in I}$ si dicono indipendenti se per ogni sottofamiglia finita $A_{i_1}, \dots A_{i_k}$ di $\mathcal{A}$ (dove $i_1, \dots, i_k \in I$) vale

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

Osservazione 3.4.5. Eventi indipendenti sono sicuramente indipendenti a due a due. *Non è vero però il viceversa: controesempio di Bernstein.* Consideriamo un tetraedro con le facce di questi colori: 1 blu, 1 rossa, 1 gialla, 1 rossa blu e gialla. Lanciamo il tetraedro e osserviamo se un certo colore compare sulla faccia appoggiata. Consideriamo i tre eventi:
B = esce il colore blu.
R = esce il colore rosso.
G = esce il colore giallo.
Chiaramente $\mathbb{P}(B) = \mathbb{P}(R) = \mathbb{P}(G) = 1/2$.
$\mathbb{P}(B \cap R) = \mathbb{P}(R \cap G) = \mathbb{P}(B \cap G) = 1/4 = \mathbb{P}(B)\mathbb{P}(R) = \mathbb{P}(R)\mathbb{P}(G) = \mathbb{P}(B)\mathbb{P}(G)$: gli eventi B, R e G sono a due a due indipendenti.
Però $\mathbb{P}(B \cap R \cap G) = 1/4 \neq \mathbb{P}(B)\mathbb{P}(R)\mathbb{P}(G) = 1/8$: B, R e G non sono indipendenti.

Valgono le seguenti proprietà:

- Se A, B sono due eventi soddisfacenti $A \supseteq B$ allora A e B sono indipendenti se e solo se $(1 - \mathbb{P}(A))\mathbb{P}(B) = 0$.
  Due eventi incompatibili ciascuno di probabilità strettamente positiva non sono mai indipendenti!

- Sia $\mathcal{A} := \{A_i\}_{i\in I}$ una famiglia di eventi indipendenti, e sia $\mathcal{B}$ una collezione ottenuta da $\mathcal{A}$ prendendo per ogni evento A, l'evento stesso oppure il suo complementare (ma mai entrambi) (i.e. scelta una funzione $f : \mathcal{A} \to \cup_{A\in\mathcal{A}}\{A, A^{\complement}\}$ con la proprietà che $f(a) \in \{A, A^{\complement}\}$, allora $\mathcal{B} = \mathcal{B}_f := \cup_{A\in\mathcal{A}}\{f(A)\}$); la collezione $\mathcal{B}$ così ottenuta è una famiglia di eventi indipendenti. La dimostrazione si conduce per induzione.

Esempio 3.4.6. Siano $\{A_1, \dots, A_n\}$ eventi indipendenti, allora per definizione sappiamo che $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$, d'altro canto è possibile calcolare immediatamente anche l'unione

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} A_1^{\complement}\right) = 1 - \prod_{i=1}^{n} \mathbb{P}(A_i^{\complement}) = 1 - \prod_{i=1}^{n}(1 - \mathbb{P}(A_i)).$$

Teorema 3.4.7. *Siano $A_1, \dots, A_n$ eventi indipendenti. Allora:*

1. *se $\mathbb{P}(A_i) < 1$ per ogni i allora si ha che $\mathbb{P}(\cup_{i=1}^n A_i) < 1$ e quindi $\cup_{i=1}^n A_i \neq \Omega$;*
2. *se $\mathbb{P}(A_i) > 0$ per ogni i allora si ha che $\mathbb{P}(\cap_{i=1}^n A_i) > 0$ e quindi $\cap_{i=1}^n A_i \neq \emptyset$.*

Dimostrazione. Le due affermazioni sono equivalenti (basta passare ai complementari). Per mostrare (2) si noti che

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i) > 0.$$

□

Quindi se gli eventi $A_1, \dots, A_n$ sono tali che $\mathbb{P}(A_i) < 1$ per ogni i e $\mathbb{P}(\cup_{i=1}^n A_i) = 1$ allora gli eventi non possono essere indipendenti.

Esercizio 3.4.8. Siano $\{A_i\}_{i\in E}$ e $\{B_j\}_{j\in H}$ due famiglie ciascuna composta da eventi a due a due disgiunti tali che, per ogni $i \in E$, $j \in H$, A_i e B_j sono indipendenti. Mostrare che, se E ed H sono al più numerabili, allora anche $\bigcup_{i\in E} A_i$ e $\bigcup_{j\in H} B_j$ sono eventi indipendenti.

Soluzione.
Essendo $A_i \cap A_j = \emptyset$ per ogni $i, j \in E$ tali che $i \neq j$ e $B_i \cap B_j = \emptyset$ per ogni $i, j \in H$ tali che $i \neq j$ allora

$$\begin{aligned}\mathbb{P}\Big(\Big(\bigcup_{i\in E} A_i\Big) \cap \Big(\bigcup_{j\in H} B_j\Big)\Big) &= \mathbb{P}\Big(\bigcup_{i\in E, j\in H} A_i \cap B_j\Big) = \sum_{i\in E, j\in H} \mathbb{P}(A_i \cap B_j) \\ &= \sum_{i\in E, j\in H} \mathbb{P}(A_i)\mathbb{P}(B_j) = \sum_{i\in E} \mathbb{P}(A_i) \sum_{j\in H} \mathbb{P}(B_j) = \mathbb{P}\Big(\bigcup_{i\in E} A_i\Big)\mathbb{P}\Big(\bigcup_{j\in H} B_j\Big)\end{aligned}$$

dove nella seconda uguaglianza si è utilizzato il fatto che $\{A_i \cap B_j\}_{i\in E, j\in H}$ è una famiglia di eventi a due a due disgiunti.

La definizione di indipendenza si estende anche alle variabili aleatorie nel seguente modo.

Definizione 3.4.9. Una famiglia di variabili aleatorie $\mathcal{X} := \{X_\alpha\}_{\alpha\in I}$ ($X_\alpha : \Omega \to \Omega_\alpha$, dove $\{(\Omega_\alpha, \mathcal{F}_\alpha)\}_{\alpha\in I}$ è una collezione di spazi misurabili) si dice composta da variabili indipendenti se e solo se per ogni sottoinsieme finito $S \subseteq I$ e per ogni collezione $\{E_\alpha\}_{\alpha\in S}$ tale che $E_\alpha \in \mathcal{F}_\alpha$ si ha che $\{X_\alpha^{-1}(E_\alpha)\}_{\alpha\in S}$ è una famiglia di eventi indipendenti.

Una proprietà immediata è che $X_1, \dots, X_n$ sono variabili aleatorie indipendenti a valori in $\mathbb{R}$ se e solo se per ogni scelta di $E_1, \dots, E_n \in \mathcal{R}$

$$\mathbb{P}_{(X_1,\dots,X_n)}(E_1 \times \cdots \times E_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} X_i^{-1}(E_i)\right) = \prod_{i=1}^{n} \mathbb{P}(X_i^{-1}(E_i)) = \prod_{i=1}^{n} \mathbb{P}_{X_i}(E_i).$$

Nel seguito, spesso modificheremo la notazione e scriveremo $\{X_1 \in E_1, \dots, X_n \in E_n\}$ al posto di $\{X_1 \in E_1\} \cap \cdots \cap \{X_n \in E_n\}$.

Definizione 3.4.10. Una famiglia di variabili $\{X_i\}_{i\in I}$ si dice composta da variabili **i.i.d** se sono **indipendenti ed identicamente distribuite** (cioè ammettono la stessa legge).

Osservazione 3.4.11. La definizione di indipendenza di n variabili aleatorie è equivalente a

$$\mathbb{P}_{(X_1,\dots,X_n)}(E_1 \times \cdots \times E_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} X_i^{-1}(E_i)\right) = \prod_{i=1}^{n} \mathbb{P}(X_i^{-1}(E_i)) = \prod_{i=1}^{n} \mathbb{P}_{X_i}(E_i).$$

per ogni scelta di $E_1, \dots, E_n \in \mathcal{R}$. Tale proprietà richiede una sola condizione, mentre quella di indipendenza di n eventi richiedeva che $\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \dots \mathbb{P}(A_{i_k})$ per ogni $k \leq n$ e per ogni scelta degli indici $i_1, \dots, i_k$ tutti distinti. Le due definizioni sembrerebbero diverse, tuttavia, per l'arbitrarietà della scelta degli intervalli I_i, esse in realtà sono analoghe. Infatti se scegliamo $I_i = \mathbb{R}$ per $i > k$, allora $\mathbb{P}(X_i) = 1$ per $i > k$, e

$$\mathbb{P}(X_1 \in I_1, X_2 \in I_2, \dots, X_k \in I_k) = \mathbb{P}(X_1 \in I_1)\mathbb{P}(X_2 \in I_2) \dots \mathbb{P}(X_k \in I_k)$$

Infine notiamo che se $\{X_{i,j}\}_{i\in I, j=1,\dots,n_i}$ è una famiglia di variabili indipendenti e $f_1, \dots, f_k$ sono funzioni misurabili tali che $f_i : \mathbb{R}^{n_i} \to C_i$ (dove (C_i, Σ_i) sono spazi misurabili generici, si pensi ad esempio al caso $(C_i, \Sigma_i) = (\mathbb{R}, \mathcal{R})$ per ogni i) allora le variabili Z_i definite come

$$Z_i := f_i(X_{i,1}, \dots, X_{i,n_i})$$

sono indipendenti. In altre parole, funzioni che agiscono su insiemi disgiunti di variabili indipendenti sono, a loro volta, indipendenti.

Se invece sapessimo solo che X_1 e X_3 sono indipendenti e che X_2 e X_3 sono indipendenti allora in generale non è vero che $f(X_1, X_2)$ e X_3 sono indipendenti. Lo mostriamo con un cotroesempio che coinvolge tre eventi (prendendo le tre funzioni indicatrici si ha l'esempio per variabili aleatorie). Supponiamo che $\Omega = \{1, 2, \dots 8\}$ e $\mathbb{P}$ sia equidistribuita. Siano $A := \{1, 2, 3, 4\}$, $B := \{3, 4, 5, 6\}$ e $C := \{3, 4, 7, 8\}$. Allora $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$ e $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \mathbb{P}(A \cap B \cap C) = 1/4$. Quindi sono addirittura a due a due indipendenti (ma non indipendenti), ma, ad esempio, $A \cap B$ non è indipendente da C.

3.5 Funzione di ripartizione e funzione dei quantili

3.5.1 Funzione di ripartizione

Di grande importanza è lo studio della funzione di ripartizione.

Definizione 3.5.1. Data una misura di probabilità ν su $(\mathbb{R}^n, \mathcal{R}^n)$ si dice **funzione di ripartizione** di ν di $F_\nu(t_1, \dots, t_n) := \nu((-\infty, t_1] \times \cdots \times (-\infty, t_n])$. La funzione di ripartizione di una variabile aleatoria X è definita come la funzione di ripartizione della sua legge $\mathbb{P}_X$; cioè, nel caso unidimensionale, $F_X(t) = \mathbb{P}_X((-\infty, t]) = \mathbb{P}((-\infty, t])$.

Per una funzione di ripartizione F valgono le seguenti proprietà.

- $F : \mathbb{R}^n \to [0, 1]$

- Si dimostra che F è una funzione di ripartizione se e solo se valgono le proprietà seguenti:
  1. dati $s_1 \le t_1, \dots, s_n \le t_n$ allora $F(s_1, \dots, s_n) \le F(t_1, \dots, t_n)$
  2. $$\lim_{\substack{(s_1,\dots,s_n)\to(t_1,\dots,t_n)\\ s_1\ge t_1,\dots,s_n\ge t_n}} F(s_1, \dots, s_n) = F(t_1, \dots, t_n)$$
  3. $$\lim_{\substack{t_1\to+\infty\\ \dots\\ t_n\to+\infty}} F(t_1, \dots, t_n) = 1$$
     $$\lim_{\substack{t_1\to-\infty\\ \dots\\ t_n\to-\infty}} F(t_1, \dots, t_n) = 0$$
- Se μ e ν sono due leggi le cui funzioni di ripartizione soddisfano $F_\mu = F_\nu$ allora $\mu = \nu$.
- Una famiglia $X_1, \dots, X_n$ di variabili aleatorie è composta da elementi indipendenti se e solo se
  $$F_{(X_1,\dots,X_n)}(t_1, \dots, t_n) = \prod_{i=1}^{n} F_{X_i}(t_i).$$
- Nel caso monodimensionale $n = 1$ la funzione di ripartizione è evidentemente non decrescente, continua da destra, avente limite da sinistra, $\lim_{x\to-\infty} F(x) = 0$ e $\lim_{x\to+\infty} F(x) = 1$; inoltre l'insieme dei punti di discontinuità
  $$\{x \in \mathbb{R} : \lim_{t\to x^-} F(t) \neq \lim_{t\to x^+} F(t)\} \equiv \{x \in \mathbb{R} : \lim_{t\to x^-} F(t) \neq F(x)\}$$
  è al più numerabile. Inoltre se X è la variabile aleatoria di cui F è la funzione di ripartizione allora per ogni $x \in \mathbb{R}$ si ha $F(x) - \lim_{t\to x^-} F(t) = \mathbb{P}_X(x)$ e $\lim_{x\to x^-} F(x) = \mathbb{P}(X < x)$.
- Date le variabili $X_1, \dots, X_n$, la funzione di ripartizione $F_{(X_1,\dots,X_n)}$ prende il nome di **funzione di ripartizione congiunta**. Si dimostra facilmente che, supponendo di aver diviso le n variabili in due insiemi $\{X_{y_1}, \dots, X_{y_h}\}$ e $\{X_{r_1}, \dots, X_{r_d}\}$ cosicché $d + h = n$, per la **funzione di ripartizione marginale** $F_{(X_{y_1}\cdots X_{y_h})}(t_{y_1}, \dots, t_{y_h})$ relativa al primo gruppo di variabili vale:
  $$F_{(X_{y_1}\cdots X_{y_h})}(t_{y_1}, \dots, t_{y_h}) = \mathbb{P}(X_{y_1} \le t_{y_1}, \dots, X_{y_h} \le t_{y_h}) \equiv \lim_{\substack{t_{r_1}\to+\infty\\ \dots\\ t_{r_n}\to+\infty}} F_{(X_1,\dots,X_n)}(t_1, \dots, t_n)$$
- Sia X una variabile aleatoria ed F_X la sua funzione di ripartizione. Se $a, b \in \mathbb{R}$ sono tali che $F_X(a), F_X(b) \in (0, 1)$, allora scelti $A \in \{\{X \ge a\}, \{X \le a\}\}$ e $B \in \{\{X \ge b\}, \{X \le b\}\}$, si ha che A e B non possono essere indipendenti (è conseguenza del Teorema 3.4.7).

3.5.2 Funzione dei quantili

Concentriamo ora la nostra attenzione sulle funzioni di ripartizione a valori reali (anche se analoghe definizioni e proprietà potrebbero essere estese anche al caso multidimensionale).

Definizione 3.5.2. Per ogni funzione di ripartizione F (cioè per ogni funzione reale non decrescente, continua da destra e tale che $\lim_{x\to+\infty} F(x) = 1$ e $\lim_{x\to-\infty} F(x) = 0$) si chiama **funzione quantile** Q_F o **pseudoinversa** di F la funzione così definita
$$Q_F(x) := \inf\{t : F(t) \ge x\} \equiv \min\{t : F(t) \ge x\}, \qquad x \in (0, 1).$$
Se $F = F_X$ allora si scrive Q_X al posto di Q_{F_X}.

Si osservi che l'uguaglianza tra il min e l'inf segue dalla continuità da destra della funzione di ripartizione F.

Valgono a tal proposito le seguenti proprietà:

- la funzione $Q_F : (0,1) \to \mathbb{R}$ è non decrescente (e quindi con limiti da destra e sinistra, essendo monotona), continua da sinistra.
- $$F \circ Q_F(x) \geq x, \quad \forall x \in (0,1)$$
  in particolare il segno di uguaglianza vale se e solo se $x \in \mathrm{Rg}(F)$, dove quest'ultimo è l'insieme delle immagini di F (cioè $\mathrm{Rg}(F) := \{y \in \mathbb{R} : \exists x \in \mathbb{R} : F(x) = y\}$).
- $$Q_F \circ F(t) \leq t, \quad \forall t \in \mathbb{R}$$
  in particolare il segno di uguaglianza vale se e solo se per ogni $s < t$ si ha $F(s) < F(t)$.
- Se esiste $(a,b) \subseteq \mathbb{R}$ tale che $F|_{(a,b)} : (a,b) \to (0,1)$ è una funzione biettiva (cioè iniettiva e suriettiva) allora Q_F è la funzione inversa di $F|_{(a,b)}$.
- Se $Y = aX + b$ con $a > 0$ allora $Q_Y = aQ_X + b$ infatti
  $$F_Y(t) \geq \alpha \iff \mathbb{P}(aX + b \leq t) \geq \alpha \iff \mathbb{P}\left(X \leq \frac{t-b}{a}\right) \geq \alpha \iff F_X\left(\frac{t-b}{a}\right) \geq \alpha,$$
  da cui passando all'estremo inferiore su t e notando che $t \to (t-b)/a$ è continua e crescente si ha
  $$Q_X(\alpha) = \frac{Q_Y(\alpha) - b}{a}.$$

La conoscenza della funzione quantile permette di risolvere i problemi del tipo: data una variabile aleatoria X reale ed un valore $\alpha \in (0,1)$ calcolare il minimo valore t tale che $\mathbb{P}(X \leq t) \geq \alpha$; tale valore è $t = Q_X(\alpha)$.

Osservazione 3.5.3. Ovviamente
$$Q_F \circ F(t) = \inf\{s : F(s) \geq F(t)\} \leq t$$
inoltre, per la continuità da destra di F,
$$F \circ Q_F(x) = F(\inf\{s : F(s) \geq x\}) \geq t.$$

3.6 Principio di Inclusione-Esclusione

Approfondimento

Riprendiamo il principio di inclusione-esclusione che abbiamo visto nel paragrafo 3.1. Ne esiste una versione più generale che è la seguente. In questo paragrafo consideriamo un insieme X generico; sia $\mathbb{R}^{S_X}$ l'insieme delle funzioni f definite su S_X e chiamiamo $S_X := \{A \subseteq X : A \text{ finito}\}$.

Teorema 3.6.1. *Definiamo $I, J : \mathbb{R}^{S_X} \to \mathbb{R}^{S_X}$ come*
$$(If)(A) := \sum_{B:B\subseteq A} i(A,B)f(B), \qquad \forall A \in S_X$$
$$(Jf)(A) := \sum_{B:B\subseteq A} j(A,B)f(B), \qquad \forall A \in S_X$$

dove $i, j : \{(A, B) \in S_X \times S_X : A \supseteq B\}$. *Se per ogni* $A, A_1 \in S_X$ *tale che* $A_1 \subseteq A$

$$\sum_{B:A_1\subseteq B\subseteq A} i(A,B)j(B,A_1) = \delta(A,A_1) := \begin{cases} 0 & A \neq A_1 \\ 1 & A = A_1 \end{cases}$$

allora $I(Jf) = f$

Dimostrazione. Per ogni $A \in S_X$

$$\begin{aligned}(I(Jf))(A) &= \sum_{B:B\subseteq A} i(A,B)(Jf)(B) = \sum_{B:B\subseteq A} i(A,B) \sum_{A_1:A_1\subseteq B} j(B,A_1)f(A_1) \\ &= \sum_{A_1:A_1\subseteq A} \left(\sum_{B:A_1\subseteq B\subseteq A} i(A,B)j(B,A_1) \right) f(A_1) \\ &= \sum_{A_1:A_1\subseteq B} \delta(A,A_1)f(A_1) = f(A)\end{aligned}$$

□

Il seguente corollario è più chiaro dal punto di vista intuitivo e di più immediata applicazione.

Corollario 3.6.2. *Date* f, g *allora le seguenti affermazioni sono equivalenti*

1. $g(A) = \sum_{S:S\subseteq A} f(S)$, *per ogni* $A \in S_X$,
2. $f(A) = \sum_{S:S\subseteq A} (-1)^{|A|-|S|} g(S)$, *per ogni* $A \in S_X$,

dove $|\cdot|$ *denota la cardinalità. Similmente, se* X *è finito allora le seguenti affermazioni sono equivalenti*

1. $g(A) = \sum_{S:X\supseteq S\supseteq A} f(S)$, *per ogni* $A \in S_X$,
2. $f(A) = \sum_{S:X\supseteq S\supseteq A} (-1)^{|A|-|S|} g(S)$, *per ogni* $A \in S_X$,

Dimostrazione. Per dimostrare l'enunciato è sufficiente mostrare che $i(A,S) = 1$ e $j(A,S) = (-1)^{|A|-|S|}$ soddisfano $\sum_{B:A_1\subseteq B\subseteq A} i(A,B)j(B,A_1) = \sum_{B:A_1\subseteq B\subseteq A} j(A,B)i(B,A_1) = \delta(A,A_1)$. Infatti

$$\begin{aligned}\sum_{B:A_1\subseteq B\subseteq A} i(A,B)j(B,A_1) &= \sum_{B:A_1\subseteq B\subseteq A} (-1)^{|B|-|A_1|} \\ &= \sum_{j=0}^{|A|-|A_1|} \sum_{B:A_1\subseteq B\subseteq A, |B|=|A_1|+j} (-1)^j \\ &= \sum_{j=0}^{|A|-|A_1|} |\{B : A_1 \subseteq B \subseteq A, |B| = |A_1| + j\}|(-1)^j \\ &= \sum_{j=0}^{|A|-|A_1|} \binom{|A| - |A_1|}{j} (-1)^j \\ &= (1-1)^{|A|-|A_1|} = \begin{cases} 1 & |A| = |A_1| \\ 0 & |A| \neq |A_1|. \end{cases}\end{aligned}$$

Inoltre

$$\begin{aligned}\sum_{B:A_1\subseteq B\subseteq A} j(A,B)i(B,A_1) &= \sum_{B:A_1\subseteq B\subseteq A}(-1)^{|A|-|B|}\\ &= (-1)^{|A|-|A_1|}\sum_{B:A_1\subseteq B\subseteq A}(-1)^{|B|-|A_1|}\\ &= \begin{cases}1 & |A|=|A_1|\\ 0 & |A|\neq|A_1|.\end{cases}\end{aligned}$$

Per mostrare la seconda parte si applichi la prima alle funzioni f_1 e g_1 definite da $f_1(A) := f(X \setminus A)$ e $g_1(A) := g(X \setminus A)$. □

Vediamo ora alcune forme alternative del principio di inclusione-esclusione che sono utili nelle applicazioni.

Proposizione 3.6.3. *Siano $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità e $\{A_i\}_{i=1}^n$ una collezione finita di insiemi misurabili allora*

$$\mathbb{P}\left(\bigcup_{i=1}^k A_i\right) = \sum_{j=1}^n (-1)^{j+1} \sum_{1\le i_1<\cdots<i_j\le n} \mathbb{P}(A_{i_1}\cap\cdots\cap A_{i_j}).$$

La stessa formula vale nel caso $(\Omega, \mathcal{F}, \mathbb{P})$ sia un generico spazio misurabile se, in più, $\mathbb{P}\left(\bigcup_{i=1}^k A_i\right) < +\infty$.

Dimostrazione. Sia $X := \{1, \dots, n\}$ e si definiscano

$$f(A) := \mathbb{P}\left(\left(\bigcap_{i\in X\setminus A} A_i\right)\cap\left(\bigcap_{i\in A}(\Omega\setminus A_i)\right)\right) \qquad g(A) := \mathbb{P}\left(\bigcap_{i\in X\setminus A} A_i\right).$$

Chiaramente $g(A) = \sum_{S:S\subseteq A} f(S)$, per ogni $A \in S_X$, pertanto dal Corollario 3.6.2 si ha $f(A) = \sum_{S:S\subseteq A}(-1)^{|A|-|S|}g(S)$, per ogni $A \in S_X$. Poiché $f(X) = \mathbb{P}(\bigcap_{i\in X}(\Omega\setminus A_i)$ e $g(X) = \mathbb{P}(\Omega)$ si ha

$$\mathbb{P}\left(\bigcap_{i=1}^k(\Omega\setminus A_i)\right) = \sum_{B:B\subseteq X}(-1)^{|X\setminus B|}\mathbb{P}\left(\bigcap_{i\in X\setminus B} A_i\right)$$

(dove $\bigcap_{i\in\emptyset} A_i = \Omega$) da cui

$$\mathbb{P}\left(\bigcup_{i=1}^k(A_i)\right) = \mathbb{P}(\Omega) - \mathbb{P}\left(\bigcap_{i=1}^k(\Omega\setminus A_i)\right) = \sum_{A:A\subseteq X, A\neq\emptyset}(-1)^{|A|+1}\mathbb{P}\left(\bigcap_{i\in A} A_i\right).$$

Per il caso generale si può sempre scegliere $\Omega := \bigcup_{i\in X} A_i$ che ha misura totale finita per ipotesi. □

Un'altra importante applicazione del principio di inclusione-esclusione è la seguente.

Si supponga che $X = \{P_i\}$ un insieme finito di proprietà che possono essere godute o meno dagli elementi di un generico insieme finito Ω. Definiamo, per ogni

$S \subseteq P$

$$N_{\geq}(S) = |\{\omega \in \Omega : \omega \text{ soddisfa tutte le } P_i \in S\}|$$
$$N_{=}(S) = |\{\omega \in \Omega : \omega \text{ soddisfa tutte e sole le } P_i \in S\}|$$

dove $N_{\geq}(\emptyset) = |\Omega|$. Chiaramente $f(S) := N_{\geq}(S)$ e $g(S) := N_{=}(S)$ soddisfano $g(A) = \sum_{S:X \supset S \supseteq A} f(S)$, per ogni $A \in S_X$; dal Corollario 3.6.2 si ha il seguente risultato.

Proposizione 3.6.4. *Per ogni $S \subseteq X$ si ha che*

$$N_{=}(S) = \sum_{J:S \subseteq J \subseteq X} (-1)^{|J|-|S|} N_{\geq}(J).$$

3.7 Affidabilità

Supponiamo di avere un sistema costituito da vari sottositsemi in serie o in parallelo. Si consideri una caratteristica Z che ciascun sottosistema può avere o non avere (ad esempio, X potrebbe essere "il sistema ha durata superiore ad un tempo T"). La probabilità che il sistema (o il sottosistema) abbia la caratteristica Z si chiama **affidabilità**.

Supponiamo che vi siano n sottosistemi e che l'i-esimo sottosistema abbia quindi affidabilità a_i. Supponiamo inoltre che gli eventi $\{$"l'elemento i-esimo ha la caratteristica Z"$\}_{i=1}^n$ siano indipendenti. Si considerino i due sistemi P ed S ottenuti mettendo gli n sottosistemi in parallelo e in serie rispettivamente.

Quando gli elementi sono messi in serie si suppone che il sistema S abbia la caratteristica Z se e solo se tutti i sottosistemi ce l'hanno. Pertanto l'affidabilità a_S di S si calcola come

$$a_S = \prod_{i=1}^{n} a_i.$$

Viceversa quando gli elementi sono messi in parallelo si suppone che il sistema P abbia la caratteristica Z se e solo se almeno un sottosistema ce l'ha. Pertanto l'affidabilità a_P di P si calcola passando attraverso gli eventi complementari $\{$"l'elemento i-esimo non ha la caratteristica Z"$\}_{i=1}^n$, ottenendo

$$a_P = 1 - \prod_{i=1}^{n} (1 - a_i).$$

Iterando opportunamente queste due formule si possono costruire le affidabilità di sistemi complessi.

Esercizio 3.7.1. Abbiamo a disposizione n tipi di componenti di affidabilità $a_1, \ldots, a_n$; di ciascun tipo ne vogliamo utilizzare k e sono indipendenti gli uni dagli altri. Cosa ci conviene fare, in termini di affidabilità del sistema finale, mettere in serie n sistemi ciascuno formato da k componenti di ugual affidabilità in parallelo oppure mettere in parallelo k sistemi ciascuno formato da n elementi (uno per ciascun tipo) in serie?

Cosa si può dire nel caso in cui i componenti siano generici (cioè non necessariamente indipendenti) e di affidabilità qualsiasi?

Soluzione.
Mostriamo che mettere in serie n elementi ciascuno formato da k elementi dello stesso tipo in parallelo è sempre più conveniente piuttosto che che mettere in parallelo k elementi ciascuno formato da n componenti (uno per tipo) in serie. In realtà i due sistemi hanno ugual affidabilità se (e solo se) $n = 1$ oppure $k = 1$ oppure $a_i = 0$ per qualche i oppure $a_i = 1$ per ogni i.

Le due affidabilità si calcolano facilmente

$$A_1 = \prod_{i=1}^{n}(1-(1-a_i)^k)$$

$$A_2 = 1-(1-\prod_{i=1}^{n} a_i)^k.$$

Mostriamo che se $a_i \in [0,1]$ per ogni $i = 1,\dots,n$ allora

$$\prod_{i=1}^{n}(1-(1-a_i)^k) \geq 1-(1-\prod_{i=1}^{n} a_i)^k$$

e l'uguaglianza si verifica se e solo se $k = 1$ oppure $n = 1$ oppure $a_i = 0$ per qualche i oppure $a_i = 1$ per ogni i.

Che ciascuna di queste condizioni implichi l'uguaglianza è ovvio, supponiamo quindi che $a_i \in (0,1)$ per ogni i, $n,k > 1$. Definiamo le seguenti funzioni:

$$G_n(a_1,\dots,a_n) := \prod_{i=1}^{n}(1-(1-a_i)^k) - (1-(1-\prod_{i=1}^{n} a_i)^k)$$
$$F(a,b) := (1-(1-a)^k)(1-(1-b)^k) - (1-(1-ab)^k);$$

si vede facilmente che

$$G_n(a_1,\dots,a_n) = (1-(1-a_n)^k)G_{n-1}(a_1,\dots,a_{n-1}) + F(\prod_{i=1}^{n-1} a_i, a_n).$$

Mostriamo quindi che $F(a,b) \geq 0$ e che la disuguaglianza è stretta se $a,b \in (0,1)$ e $k > 1$. Infatti

$$\begin{aligned} F(a,b) &= (1-a)^k((1-b)^k-1) + b(1-a)\sum_{j=0}^{k-1}(1-ab)^j(1-b)^{k-1-j} \\ &= b(1-a)\sum_{j=0}^{k-1}(1-ab)^j(1-b)^{k-1-j} - (1-a)^k b\sum_{j=0}^{k-1}(1-b)^j \\ &= b(1-a)\left(\sum_{j=0}^{k-1}(1-b)^j\left((1-ab)^{k-j-1} - (1-a)^{k-1}\right)\right) \geq 0 \end{aligned}$$

dal momento che

$$(1-ab)^{k-1-j} \geq (1-a)^{k-1-j} \geq (1-a)^{k-1}$$

e la prima disuguaglianza è stretta se $b \neq 1$.

Terminiamo mostrando l'asserto per induzione su n; per $n = 1$ $g_1(a_1) = 0$ pertanto la proposizione è verificata. Se vale per $n-1$ allora

$$G_n(a_1,\dots,a_n) = (1-(1-a_n)^k)G_{n-1}(a_1,\dots,a_{n-1}) + F(\prod_{i=1}^{n-1} a_i, a_n) > 0$$

poiché $\prod_{i=1}^{n-1} a_i, a_n \in (0,1)$ e quindi $F(\prod_{i=1}^{n-1} a_i, a_n) > 0$.

Nel caso generico basta osservare che, detta A_{ij} la probabilità che il componente nella riga i e colonna j funzioni, allora evidentemente si ha che la probabilità che il primo sistema funzioni è $\mathbb{P}\left(\bigcap_i \bigcup_j A_{ij}\right)$, mentre quella del secondo sistema è $\mathbb{P}\left(\bigcup_j \bigcap_i A_{ij}\right)$. Facilmente $\bigcap_i \bigcup_j A_{ij} \supseteq \bigcup_j \bigcap_i A_{ij}$ da cui si ha che l'affidabilità del primo sistema non è inferiore a quella del secondo.

Cap. 4. Variabili aleatorie discrete

Da questo capitolo in poi, se non altrimenti specificato, considereremo solo variabili aleatorie a valori in $\mathbb{R}$ oppure, più in generale, in $\mathbb{R}^n$.

Abbiamo già introdotto le variabili aleatorie discrete, ma vediamo un altro esempio.

Esempio 4.0.2. Sia Ω lo spazio campionario generato dal lancio di due dadi: la scelta minimale per questo spazio è $\Omega = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\} = \{(1, 1), (1, 2), \ldots, (6, 6)\}$. Definiamo X la somma dei numeri che si verificano:

$$(i, j) \longrightarrow i + j$$

In genere quello che ci interessa di una variabile aleatoria è di *calcolare la probabilità che essa assuma determinati valori*.

Nell'esempio precedente del lancio di due dadi ci può interessare di conoscere la probabilità che la somma dei numeri sia pari a 5, oppure che sia inferiore a 7, ecc..

Abbiamo visto che una legge discreta è univocamente determinata a partire da un insieme S al più numerabile e dai valori che assume sugli eventi elementari.

Più precisamente data una qualsiasi misura $\mathbb{P}$ di probabilità $S := \{\omega : \mathbb{P}(\{\omega\}) > 0\}$ è al più numerabile e $\mathbb{P}(S) = \sum_{\omega \in S} \mathbb{P}(\{\omega\}) \leq 1$ dove l'uguaglianza è verificata se e solo se la misura è discreta.

Pertanto, per una v.a. discreta a valori in un insieme di numeri reali $S := \{x_i : i \in I\}$ dove $I \subseteq \mathbb{N}$, la legge univocamente determinata dai valori p_i che rappresentano, al variare di $i \in I$, la probabilità che la variabile aleatoria assuma il valore x_i.

La funzione reale

$$x_i \longrightarrow p_X(x_i) = \mathbb{P}(X = x_i)$$

viene chiamata **densità (discreta) di X, o funzione di probabilità**.
Precisamente, la legge è data da:

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

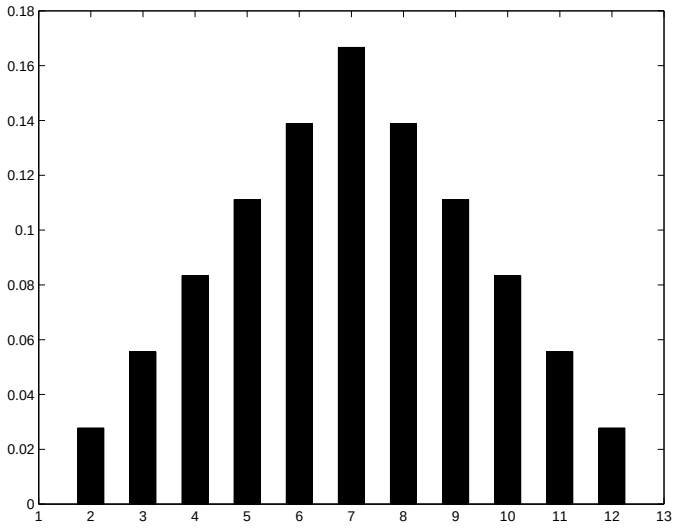
Alcuni testi indicano con $x \to \mathbb{P}(X = x)$, $x \in \mathbb{R}$ con il nome di funzione densità (discreta).

Nell'esempio precedente del lancio di due dadi, la v.a. assume valori interi compresi tra 2 e 12. La densità di X è data dalla seguente tabella:

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

la funzione di ripartizione della v.a. X discreta si calcola come:

$$F_X(x) = \sum_{i : x_i \leq x} p_X(x_i)$$



Densità discreta della v.a. *somma dei punti di due dadi*

Osservazione 4.0.3. Se esiste un riordinamento $\varphi : \mathbb{N} \to \mathbb{N}$ tale che $i \mapsto x_{\varphi(i)}$ sia crescente, allora la funzione di ripartizione della una v.a. discreta è costante a tratti: nell'intervallo $[x_i, x_{i+1})$ è costante, mentre in x_{i+1} cresce della quantità $p_X(x_{i+1})$.

Nel nostro esempio del lancio di due dadi la funzione di ripartizione è data dalla seguente tabella:

X	$x < 2$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,7)$
F_X	0	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$

X	$[7,8)$	$[8,9)$	$[9,10)$	$[10,11)$	$[11,12)$	$x \geq 12$
F_X	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

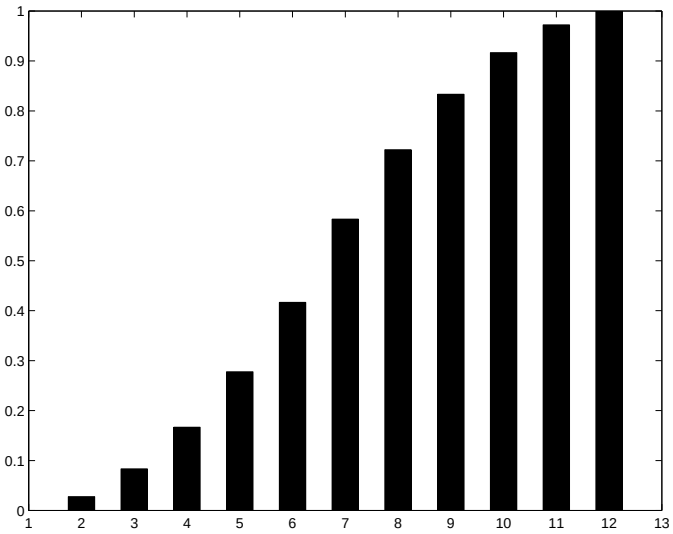
Quella che segue è una serie di condizioni equivalenti per una variabile aleatoria per essere discreta.

Teorema 4.0.4. *Sia X una variabile aleatoria a valori in Y, allora le seguenti condizioni sono equivalenti.*

1. *X è una v.a. discreta;*
2. *esiste una successione $\{(y_i, p_i)\}_{i\in J}$ con $J \subset \mathbb{N}$ tale che $(y_i, p_i) \in Y \times (0, 1]$ per ogni $i \in J$, $\sum_{i\in J} p_i = 1$ e $\mathbb{P}(X = y_i) = p_i$:*
3. *la legge $\mathbb{P}_X$ è una misura di probabilità discreta;*
4. *$\sum_{y\in Y:\mathbb{P}(X=y)>0} \mathbb{P}(X = y) = 1$.*

In tal caso l'insieme dei valori possibili è $V = \{y \in Y : \mathbb{P}(X = y) > 0\} = \{y_i : i \in J\}$ e per ogni $A \subset Y$ si ha

$$\mathbb{P}_X(A) := \mathbb{P}(X \in A) = \sum_{y\in A:\mathbb{P}(X=y)>0} \mathbb{P}(X = y) \equiv \sum_{y\in A\cap V} p_X(y).$$



Funzione di ripartizione della v.a. *somma dei punti di due dadi*

Definizione 4.0.5. Date n variabili aleatorie $X_1, \dots, X_n$, la funzione $p_{X_1,\dots,X_n}$ (o $\mathbb{P}_{(X_1,\dots,X_n)}$) definita da

$$p_{X_1,\dots,X_n}(y_1, \dots, y_n) = \mathbb{P}(X_1 = y_1, \dots, X_n = y_n) \equiv \mathbb{P}_{(X_1,\dots,X_n)}(\{y_1\} \times \cdots \times \{y_n\})$$

è detta **densità congiunta** di $X_1, \dots, X_n$.

Osservazione 4.0.6. Si osservi che la legge congiunta $\mathbb{P}_{(X_1,\dots,X_n)}$ è discreta se e solo se ciascuna delle leggi $\{\mathbb{P}_{X_i}\}_{i=1}^n$ è discreta (per mostrarlo si utilizzi il fatto che disuguaglianza $\sum_{\omega\in S} \mathbb{P}(\{\omega\}) \leq 1$ diviene un uguaglianza se e solo se la legge è discreta).

Come già detto nei paragrafi precedenti si definisce la funzione di ripartizione congiunta delle v.a. $X_1, \cdots, X_n$

$$F_{X_1,\dots,X_n}(t_1, \dots, t_n) = \mathbb{P}(X_1 \leq t_1, \dots, X_n \leq t_n) = \sum_{\substack{u_1 \leq t_1 \\ \cdots \\ \cdots \\ u_n \leq t_n}} p_{X_1,\dots,X_n}(u_1, \dots, u_n).$$

Definizione 4.0.7. Supponiamo di dividere le n variabili $X_1, \dots, X_n$ in due insiemi $\{X_{y_1}, \dots, X_{y_h}\}$ e $\{X_{r_1}, \dots, X_{r_d}\}$ cosicché $d + h = n$, allora la **densità marginale** $p_{X_{y_1},\cdots,X_{y_h}}(u_{y_1}, \dots, u_{y_h})$, relativa al primo gruppo di variabili, è:

$$p_{X_{y_1},\cdots,X_{y_h}}(u_{y_1}, \dots, u_{y_h}) = \mathbb{P}(X_{y_1} = u_{y_1}, \cdots, X_{y_h} = u_{y_h}) = \sum_{u_{r_1},\dots,u_{r_d}} p_{X_1,\dots,X_n}(u_1, \dots, u_k).$$

L'indipendenza di n variabili aleatorie si traduce nella seguente:

$$p_{X_1,\dots,X_n}(y_1, \dots, y_n) = p_{X_1}(y_1) \times \cdots \times p_{X_n}(y_n)$$

dove p_{X_i} indica la densità di probabilità della v.a. X_i.

Esempio 4.0.8. Un'urna contiene 5 palline numerate da 1 a 5. Estraiamo due palline, con reimmissione della prima pallina. Siano X_1 e X_2 i risultati della prima e della seconda estrazione rispettivamente. Se si suppone che ogni coppia di risultati abbia la stessa probabilità, allora

$$\mathbb{P}(X_1 = i, X_2 = j) = \frac{1}{25} = \mathbb{P}(X_1 = i)\mathbb{P}(X_2 = j)$$

pertanto X_1 e X_2 sono eventi indipendenti.

Esempio 4.0.9. Un'urna contiene 5 palline numerate da 1 a 5. Estraiamo due palline senza reimmissione. Siano Y_1 e Y_2 i risultati della prima e della seconda estrazione rispettivamente. Se si suppone che ognuna delle 20 coppie di risultati abbia la stessa probabilità, allora per $i \neq j$ $\mathbb{P}(Y_1 = i, Y_2 = j) = \frac{1}{20}$ mentre $\mathbb{P}(Y_1 = i)\mathbb{P}(Y_2 = j) = \frac{1}{25}$ pertanto Y_1 e Y_2 non sono eventi indipendenti.

4.1 Valore atteso per variabili aleatorie discrete

Nei prossimi due paragrafi introdurremo il calcolo di media e varianza per variabili aleatorie discrete; per un cenno al calcolo di questi indici nel caso di una variabile aleatoria generica si veda il capitolo 6.

Sia X una v.a. discreta che assume i valori $x_1, \ldots, x_n$, e sia p_X la sua densità di probabilità. Si chiama **valore atteso, o media, o speranza matematica** di X, e la si denota con $\mathbb{E}(X)$, la quantità

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_X(x_i).$$

Se i valori della v.a. sono un'infinità numerabile, la somma diventa una serie; in tal caso si dice che X ammette valor medio se e solo se $\sum_{i=1}^{\infty} |x_i| p_X(x_i) < \infty$ (**convergenza assoluta* della serie**). In questo caso il valor medio è definito da

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_X(x_i);$$

il valore atteso $\mathbb{E}(X)$ quindi è definito a condizione che la serie converga assolutamente. In tal caso il valore medio è ben definito, nel senso che, preso un qualsiasi riordinamento $\phi : \mathbb{N} \in \mathbb{N}$ della serie, di ha

$$\sum_{i=1}^{\infty} x_{\phi(i)} \mathbb{P}_X(x_{\phi(i)}) = \sum_{i=1}^{\infty} x_i \mathbb{P}_X(x_i)$$

(**convergenza incondizionata**).

Esempio 4.1.1. Consideriamo la variabile aleatoria associata al lancio di due dadi. Il valore atteso è

$$\mathbb{E}(X) = 2\frac{1}{36} + 3\frac{2}{36} + 4\frac{3}{36} + 5\frac{4}{36} + 6\frac{5}{36} + 7\frac{6}{36} +$$
$$+8\frac{5}{36} + 9\frac{4}{36} + 10\frac{3}{36} + 11\frac{2}{36} + 12\frac{1}{36} = 7$$

Proprietà del valore atteso (per dimostrare alcune di esse sono necessarie le definizioni generali del Capitolo 6 le cui nozioni sono però facoltative).

1. Per una trasformazione affine della v.a. il valore atteso si trasforma in maniera affine, cioè:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$$

2. Valore atteso di una funzione di v.a.: sia X una v.a. e f una funzione continua su $\mathbb{R}$. Allora il valore atteso di $f(X)$ esiste se e solo se

$$\sum_k |f(x_k)| p_X(x_k) < +\infty$$

e vale

$$\mathbb{E}(f(X)) = \sum_k f(x_k) p_X(x_k).$$

3. Valore atteso di una funzione di più v.a.: si supponga di avere una collezione finita di v.a. (discrete) $X_1, \dots, X_n$ di cui si conosca la legge congiunta $\mathbb{P}_{(X_1,\dots,X_n)}$ e sia $f : \mathbb{R}^n \to \mathbb{R}$ una funzione misurabile; allora la variabile $f(X_1, \dots, X_n)$ ammette media se e solo se

$$\sum_{u_1,\dots,u_n} p_{X_1,\dots,X_n}(u_1, \dots, u_n)|f(u_1, \dots, u_n)| < +\infty$$

e vale

$$\mathbb{E}(f(X_1, \dots, X_n)) = \sum_{u_1,\dots,u_n} p_{X_1,\dots,X_n}(u_1, \dots, u_n) f(u_1, \dots, u_n).$$

4. Se $X_1, \dots, X_n$ sono v.a. indipendenti allora $X_1 X_2 \cdots X_n$ ammette valore atteso se e solo se ciascuna variabile X_i ammette valore atteso; inoltre

$$\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \cdots \mathbb{E}(X_n).$$

Si noti che la precedente relazione non implica l'indipendenza, ma ne è solo implicata!

Esercizio 4.1.2. Nella trasmissione di un'immagine il colore di ogni pixel è descritto da un vettore a 8 bits $(a_1, \dots, a_8)$, dove gli a_i possono valere 0 oppure 1. Durante la trasmissione di ogni bit si può avere un errore con probabilità $p_b = 2 \times 10^{-4}$, indipendentemente da un bit all'altro.

1. Qual è la probabilità che un singolo pixel venga trasmesso correttamente?

2. Per un'immagine composta da $512 \times 256 = 131072$ pixels quale sarà il numero medio di pixels distorti?

Soluzione.

1. Consideriamo l'evento A_i = l'i-esimo bit non è stato distorto ($i = 1, \dots, 8$). La probabilità p_p che un singolo pixel venga trasmesso correttamente è

$$p_p = \mathbb{P}(A_1 \cap \cdots \cap A_8) = \mathbb{P}(A_1) \dots \mathbb{P}(A_8) = (1 - p_b)^8 \approx 0.9984$$

2. Definiamo la v.a. X_i che vale 1 se l'i-esimo pixel è stato distorto, 0 altrimenti. Si ha $\mathbb{P}(X_i = 1) = 1 - 0.9984 = 1.6 \times 10^{-3}$. Chiamiamo $H_n = \sum_{i=1}^n X_i$. Il valor medio che cerchiamo è $\mathbb{E}(H_n)$ con $n = 131072$:

$$\mathbb{E}(H_n) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = n\mathbb{E}(X_1) \approx 131072 \times 1.6 \times 10^{-3} = 209.7$$

4.2 Varianza per variabili aleatorie discrete

Definiamo ora gli indici caratteristici di *dispersione* (la varianza, la deviazione standard) di una variabile aleatoria discreta.

Definizione 4.2.1. Sia X una v.a. discreta avente valore atteso finito. Si definisce **varianza** di X la quantità

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

purché questo valore sia finito. In caso contrario X non ha varianza finita.

La varianza di una v.a. dunque rappresenta una misura della sua dispersione rispetto al valore atteso $\mathbb{E}(X)$.

Proprietà della varianza (per dimostrare alcune di esse sono necessarie le definizioni generali del Capitolo 6 le cui nozioni sono però facoltative); si supponga, a tal proposito, che la variabile assuma valori $\{x_i\}_{i \in J}$ dove $J \subseteq \mathbb{N}$.

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

2. Per calcolare la varianza possiamo ricorrere alla formula:
$$\mathrm{Var}(X) = \sum_{i\in J}[x_i - \mathbb{E}(X)]^2 p_X(x_i) \equiv \sum_{i\in J} x_i^2 p_X(x_i) - \mathbb{E}(X)^2.$$
che si ottiene sfruttando le proprietà del valore atteso di una funzione di variabili aleatorie.
3. $\mathrm{Var}(X) \geq 0$ e vale $\mathrm{Var}(X) = 0$ se e solo se esiste $x \in \mathbb{R}$ tale che $\mathbb{P}(X = x) = 1$.
4. $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$ per qualsiasi valore di $a, b \in \mathbb{R}$
5. Se $X_1, X_2, \ldots, X_n$ sono v.a. *indipendenti*, allora
$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)$$
Se le v.a. non sono indipendenti questa proprietà non vale. Ad esempio, scelta una qualsiasi variabile aleatoria X con varianza non nulla, allora $\mathrm{Var}(X + X) = \mathrm{Var}(2X) = 4\mathrm{Var}(X) \neq \mathrm{Var}(X) + \mathrm{Var}(X)$.

Definizione 4.2.2. La **deviazione standard, o scarto quadratico medio**, è definito come
$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Definizione 4.2.3. Sia X una v.a. con valore atteso μ_X e varianza σ_X^2 finite. Si dice **standardizzata** di X la v.a.
$$Z = \frac{X - \mu_X}{\sigma_X}$$
Z ha valore atteso nullo e varianza pari a 1. Infatti
$$\mathbb{E}\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X}(\mathbb{E}(X) - \mu_X) = \frac{1}{\sigma_X}(\mu_X - \mu_X) = 0$$
$$\mathrm{Var}\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X}\mathrm{Var}(X) = \frac{\sigma_X}{\sigma_X} = 1.$$

4.3 Analisi comparative tra variabili aleatorie discrete

Definizione 4.3.1. Siano X e Y due v.a. aventi varianza finita; si definisce la **covarianza** di X e Y come:
$$\mathrm{cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right)$$
La covarianza viene anche denotata con il simbolo σ_{XY}.
Proprietà della covarianza:

1. $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$
2. $\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
3. $\mathrm{cov}(X, X) = \mathrm{Var}(X)$
4. $\mathrm{cov}(aX + b, cY + d) = a \cdot c \cdot \mathrm{cov}(X, Y)$
5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{cov}(X, Y)$
Da questa proprietà discende che condizione necessaria (ma non suffiente!) affinché due v.a. X e Y siano indipendenti è che $\mathrm{cov}(X, Y) = 0$. Infatti, se sono indipendenti allora $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Perciò $\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

Esempio 4.3.2. Si consideri la seguente legge congiunta (per $p \in (0, 1/3)$)

$p_{X,Y}$	$X=-1$	$X=0$	$X=1$	Marginale di Y
$Y=0$	$p/4$	$(1-p)/2$	$p/4$	$1/2$
$Y=1$	$3p/4$	$(1-3p)/2$	$3p/4$	$1/2$
Marginale di X	p	$1-2p$	p	

allora $\mathbb{E}(X)=0, \mathbb{E}(Y)=1/2, \mathbb{E}(XY)=0$ da cui $\mathrm{cov}(X,Y)=0$ ma, ad esempio, $\mathbb{P}_{X,Y}(1,0) = p/4 \neq p/2 = p_X(1)p_Y(0)$ pertanto X e Y non sono indipendenti.

Definizione 4.3.3. Siano X e Y due v.a. aventi entrambe varianza finita e strettamente positiva; si definisce **coefficiente di correlazione** di X, Y, la quantità

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Il coefficiente di correlazione è sempre compreso tra -1 e 1. Due variabili aleatorie si dicono *incorrelate* o *scorrelate* se $\mathrm{cov}(X,Y)=0$ e dunque se $\rho_{XY}=0$. Se due v.a. X, Y sono indipendenti, allora sono incorrelate (ma non vale il viceversa!).

Nel caso in cui $\rho_{XY} = \pm 1$, si dimostra che le v.a. sono linearmente dipendenti: $Y = aX + b$. La costante a ha lo stesso segno di ρ_{XY}.

Esercizio 4.3.4. Un'urna contiene k palline nere e $n-k$ palline bianche. Se ne estraggono due senza rimpiazzo. Definiamo le v.a. X_1 e X_2:
$X_1 = 1$ se la 1^a pallina estratta è *nera*,
$X_1 = 0$ se la 1^a pallina estratta è *bianca*,
$X_2 = 1$ se la 2^a pallina estratta è *nera*,
$X_2 = 0$ se la 2^a pallina estratta è *bianca*.

Calcolare il coefficiente di correlazione tra X_1 e X_2.

Soluzione.

$$\mathbb{P}(X_1=1) = \mathbb{P}(X_2=1) = \frac{k}{n}; \quad \mathbb{P}(X_1=0) = \mathbb{P}(X_2=0) = \frac{n-k}{n}$$

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = \frac{k}{n}1 + \frac{n-k}{n}0 = \frac{k}{n}; \quad \mathbb{E}(X_1^2) = \mathbb{E}(X_2^2) = \frac{k}{n}1 + \frac{n-k}{n}0 = \frac{k}{n}$$

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = \mathbb{E}(X_1^2) - [\mathbb{E}(X_1)]^2 = \frac{k(n-k)}{n^2}$$

$$\mathrm{cov}(X_1,X_2) = \left(1-\frac{k}{n}\right)\left(1-\frac{k}{n}\right)\mathbb{P}(X_1=1,X_2=1) +$$

$$+\left(1-\frac{k}{n}\right)\left(-\frac{k}{n}\right)\mathbb{P}(X_1=1,X_2=0) + \left(-\frac{k}{n}\right)\left(1-\frac{k}{n}\right)\mathbb{P}(X_1=0,X_2=1) +$$

$$+\left(-\frac{k}{n}\right)\left(-\frac{k}{n}\right)\mathbb{P}(X_1=0,X_2=0)$$

$$= \left(1-\frac{k}{n}\right)\left(1-\frac{k}{n}\right)\frac{k}{n}\frac{(k-1)}{(n-1)} + 2\left(-\frac{k}{n}\right)\left(1-\frac{k}{n}\right)\frac{k}{n}\frac{(n-k)}{(n-1)} +$$

$$+\left(-\frac{k}{n}\right)\left(-\frac{k}{n}\right)\frac{n-k}{n}\frac{(n-k-1)}{(n-1)}$$

$$= \frac{-k(n-k)}{n^2(n-1)}$$

$$\rho_{X_1X_2} = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}} = -\frac{1}{n-1}.$$

Analogie tra variabile aleatoria e insiemi di dati numerici:

dati numerici	*variabile aleatoria*
n-upla di numeri $(x_1, \ldots, x_n)$	v.a. X
$f_r(k) = \#\{x_i \mid x_i \in \text{classe}(k)\}/n \ (i = 1, \ldots, n)$	$x_i \longrightarrow p_X(x_i) = \mathbb{P}(X = x_i)$
$\overline{x} = \sum_{i=1}^{N_c} \overline{x}_i f_r(i)$	$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_X(x_i)$
$\sigma^2 = \sum_{i=1}^{N_c} f_r(i)(\overline{x}_i - \overline{x})^2$	$\mathrm{Var}(X) = \sum_{i=1}^{n}(x_i - \mathbb{E}(X))^2 p_X(x_i)$
$\sigma_{xy} = \sum_{i=1}^{N_c} f_r(i)(\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y})$	$\mathrm{cov}(X, Y) = \sum_{i=1}^{n}(x_i - \mathbb{E}(X))\times$ $\times(y_i - \mathbb{E}(Y))p_{XY}(x_i, y_i)$

4.4 Modelli discreti di variabili aleatorie

4.4.1 Variabili di Bernoulli e Binomiali, processo di Bernoulli

Definizione 4.4.1. Si chiama **variabile di Bernoulli** una variabile aleatoria discreta X a valori in $\{0, 1\}$. con probabilità, rispettivamente, $1 - p$ e p (dove p è un numero reale in $[0, 1]$). p viene chiamato *parametro* della variabile di Bernoulli. Convenzionalmente l'evento $\{X = 1\}$ con probabilità p viene chiamato *successo* mentre quello con probabilità $1-p$ viene chiamato *insuccesso*.

A volte si generalizza la definizione chiamando **variabile di Bernoulli generalizzata** (o più semplicemente ancora **variabile di Bernoulli**) ogni variabile aleatoria che ammette solo due valori.

Esempio 4.4.2. Il lancio di una moneta è un esperimento di Bernoulli. Se la moneta non è truccata il parametro p vale $1/2$.

Esempio 4.4.3. Lancio due dadi e considero *successo* l'evento *la somma dei punti dei due dadi è* 7, e insuccesso l'evento complementare. Il parametro p vale $1/6$.

La variabile aleatoria di Bernoulli si indica con $X \sim B(p)$ e prende il valore 1 in caso di successo e 0 in caso di insuccesso:

$$p_X(1) = p \quad p_X(0) = 1 - p.$$

In modo compatto:

$$p_X(a) = p^a(1-p)^{1-a}, \qquad a = 0, 1.$$

L'esperimento associato ad una variabile di Bernoulli prende il nome di **esperimento di Bernoulli** o **prova di Bernoulli**.

Definizione 4.4.4. Si dice **processo di Bernoulli** una successione al più numerabile $\{X_i\}_{i \in I}$ (dove $I \subseteq \mathbb{N}$) di esperimenti di Bernoulli di uguale parametro p, tra loro indipendenti.

Nel caso I numerabile si parla di processo di Bernoulli *illimitato*.

Definizione 4.4.5. Consideriamo un processo di Bernoulli di parametro p, di n prove. Si definisce **binomiale** di parametri n e p, e la si scrive $X \sim B(n, p)$, la v.a. che conta il numero complessivo di successi ottenuti nelle n prove. Dunque $B(n, p)$ è la somma di n v.a. Bernoulliane di parametro p, indipendenti tra loro:

$$X = \sum_{i=1}^{n} X_i, \quad X \sim B(n, p), \quad X_i \sim B(p).$$

Esempio 4.4.6. Si controllano 100 pezzi prodotti e si registra il numero di pezzi difettosi.

Teorema 4.4.7. *La v.a. binomiale di parametri n e p può assumere valori interi compresi tra 0 e n. La sua densità discreta è:*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots, n,$$

dove

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

Dimostrazione. La probabilità $\mathbb{P}_{X_1 \dots X_n}(a_1, \dots, a_n)$, per l'indipendenza degli eventi, vale:

$$\mathbb{P}_{X_1 \dots X_n}(a_1, \dots, a_n) = p_{X_1}(a_1) \dots p_{X_n}(a_n) =$$
$$= p^{a_1}(1-p)^{1-a_1} \dots p^{a_n}(1-p)^{1-a_n} = p^{\sum_{i=1}^n a_i}(1-p)^{n-\sum_{i=1}^n a_i}$$

L'evento *k successi nelle n prove* si esprime come

$$\{(a_1, \dots, a_n) : a_i \in \{0,1\}, \forall i = 1, \dots, n, \sum_{i=1}^n a_i = k\} \equiv \bigcup_{(a_1,\dots,a_n) \in \{0,1\}^n : \sum_{i=1}^n a_i = k} \{(a_1, \dots, a_n)\},$$

dove l'unione è disgiunta e ciascun addendo ha probabilità

$$\mathbb{P}_{X_1 \dots X_n}(a_1, \dots, a_n) = p^k (1-p)^{n-k}.$$

Tale evento può quindi essere ottenuto come unione disgiunta delle $\binom{n}{k}$ scelte diverse delle n-uple $(a_1, \dots, a_n)$, pertanto

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

□

Osservazione 4.4.8. Ricordiamo che n v.a. $X_1, \dots, X_n$ sono identicamente distribuite se e solo se hanno la stessa legge; ad esempio nel caso discreto questo è equivalente ad

$$p_{X_1}(x) = p_{X_2}(x) = \dots = p_{X_n}(x) \quad \forall x.$$

Questo non vuol dire che le variabili siano identiche!

L'indipendenza può infatti cambiare drasticamente un modello. Ad esempio, sia $H_n = X_1 + \dots + X_n$ con X_i Bernoulliane indipendenti tra di loro: H_n è una binomiale $H_n \sim B(n,p)$. Invece, se $X_1 = X_2 = \dots = X_n$, le v.a. X_i non sono indipendenti, e $H_n = X_1 + \dots + X_n = nX_1$. H_n assume soltanto i due valori 0 e n, e la densità di probabilità associata è $p(n) = p$, $p(0) = 1-p$.

Esercizio 4.4.9. Una compagnia aerea sa che in media il 10% dei passeggeri che hanno prenotato non si presenta alla partenza. In base a questa considerazione accetta 32 prenotazioni su 30 posti liberi. Supponendo che i comportamenti dei passeggeri siano indipendenti, qual è la probabilità che almeno uno rimanga a terra?

Soluzione.
Sia X la v.a. che vale 0 se il passeggero con prenotazione non si presenta, 1 se si presenta. $X \sim B(0.9)$. Gli eventi sono 32, e cerchiamo la probabilità che la binomiale di parametri $n = 32$ e $p = 0.9$ abbia valore maggiore di 30:

$$p(H_{32} > 30) = \binom{32}{31} 0.9^{31} 0.1^1 + \binom{32}{32} 0.9^{32} 0.1^0 \approx 0.122 + 0.034 = 0.156$$

Sia $X \sim B(P)$. Allora

$$\mathbb{E}(X) = 0 \cdot p_X(0) + 1 \cdot p_X(1) = p$$
$$\mathrm{Var}(X) = (0-p)^2 p_X(0) + (1-p)^2 p_X(1) = p^2(1-p) + (1-p)^2 p = p(1-p).$$

Sia ora $X \sim B(n,p)$. Siccome X è somma di n v.a. Bernoulliane *indipendenti*, si avrà:

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}\left(X_i\right) = np$$
$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) = np(1-p)$$

dove nella seconda catena di uguaglianze abbiamo utilizzato l' indipendenza delle variabili di Bernoulli $\{X_i\}_{i=1}^n$.

4.4.2 Variabili Geometriche

Definizione 4.4.10. Consideriamo una successione $\{X_i\}_{i=1}^{\infty}$ di prove Bernoulliane indipendenti di parametro p. Si chiama **geometrica** la v.a. che rappresenta il numero di prove necessario affinché si presenti per la prima volta l'evento *successo*. La indicheremo $X \sim \mathrm{Geo}(p)$ ed è definita come

$$X := \min\{i : X_i = 1\}$$

dove $\min \emptyset := +\infty$.

La funzione di probabilità associata è:

$$p_X(k) = p(1-p)^{k-1}, \quad k = 1, 2, 3, \ldots$$

Infatti questa è la probabilità che in k prove Bernoulliane le prime $k-1$ siano insuccessi e l'ultimo sia un successo. Si osservi che $\mathbb{P}(X > k) = (1-p)^k$ per ogni $k \in \mathbb{N}$.

La variabile X è a priori discreta a valori in $\mathbb{N}^* \bigcup \{+\infty\}$. Se $p = 0$ evidentemente $p_X(k) = 0$ per ogni $k \in \mathbb{N}^*$ da cui $p_X(+\infty) = 1$. Nel caso in cui $p \in (0,1]$ allora si possono seguire due strade equivalenti:

- $\mathbb{P}(X = +\infty) = \lim_{k \to +\infty} \mathbb{P}(X > k) = 0$
- $$\mathbb{P}(X = +\infty) = 1 - \mathbb{P}(X \in \mathbb{N}^*) = 1 - \sum_{i=1}^{+\infty} p_X(i)$$
$$= 1 - \sum_{i=1}^{+\infty} p(1-p)^{i-1} = 1 - p\frac{1}{1-(1-p)} = 0$$

dove si è utilizzato il seguente risultato (provare per induzione)

$$\sum_{i=0}^{n} a^i = \begin{cases} n+1 & \text{se } a = 1 \\ \frac{1-a^{n+1}}{1-a} & \text{se } a \neq 1. \end{cases}$$

In qualsiasi dei due modi lo si affronti, si ha che $\mathbb{P}(X \in \mathbb{N}^*) = 1$ pertanto la variabile può essere considerata discreta a valori in $\mathbb{N}^*$. D'ora in poi considereremo $X \sim Geom(p)$ con $p \in (0,1]$.

Assenza di memoria della legge geometrica. Sia $X \sim \mathrm{Geo}(p)$ Si considerino gli eventi $A_k := \{X > k\}$ (per ogni $k = 1, 2, \ldots$). Supponiamo che si sappia che $X > k$ e ci si chieda quale sia la probabilità che $X > k+n$ con $n \in \mathbb{N}$. Quello che vogliamo calcolare è la probabilità condizionata (si osservi che $A_{k+n} \subseteq A_k$)

$$\mathbb{P}(X > k+n | X > k) = \mathbb{P}(A_{k+n}|A_k) := \frac{\mathbb{P}(A_{k+n} \cap A_k)}{\mathbb{P}(A_k)} = \frac{\mathbb{P}(A_{k+n})}{\mathbb{P}(A_k)}$$
$$= \frac{(1-p)^{k+n}}{(1-p)^k} = (1-p)^n = \mathbb{P}(X > n).$$

Questa proprietà, detta **assenza di memoria**, significa che se sappiamo che il primo successo di un processo di Bernoulli non è ancora avvenuto dopo k prove, allora la probabilità che si debbano attendere almeno altre n prove è la stessa per ogni $k \in \mathbb{N}$. Si potrebbe mostrare che l'unica legge discreta sull'insieme $\mathbb{N}$ avente la proprietà di assenza di memoria, è proprio la legge geometrica.

Se $X \sim \text{Geo}(p)$ allora si può mostrare che

$$\mathbb{E}(X) = \begin{cases} \sum_{i=1}^{\infty} ip(1-p)^{i-1} = \frac{1}{p} & p \in (0,1] \\ +\infty & p = 0. \end{cases}$$

Nel caso in cui $p \in (0,1]$ calcoliamo la varianza come

$$\text{Var}(X) = \sum_{i=1}^{\infty} i^2 p(1-p)^{i-1} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p}.$$

Approfondimento

Esempio 4.4.11. Consideriamo il seguente esperimento. Siano $\{X_i\}_{i\in\mathbb{N}}$ una sequenza di variabili discrete iid a valori in Y e con legge $\{p_y\}_{y\in Y}$ tale che $p_y > 0$ per ogni $y \in Y$. Si consideri una sequenza finita $w = (y_1, \dots, y_l) \in Y^l$ e sia $N := \min\{k : X_{k-l+j} = y_j, \forall j = 1, \dots, l\}$ l'istante in cui per la prima volta esce la stringa w nella sequenza $\{X_i\}_{i\in\mathbb{N}}$. Più avanti daremo un cenno al calcolo di $\mathbb{E}(N)$ che però necessita di tecniche più raffinate (teoria delle **Martingale**).

Per ora occupiamoci del caso più semplice in cui $Y = \{0,1\}$, $p_0 = p_1 = 1/2$ e $w_1 = (0,0)$, $w_2 = (0,1)$. Osserviamo che $\mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(X_1 = 0, X_2 = 1) = 1/4$. Inoltre se cerchiamo la probabilità dell'evento A_1 :="esce prima $(0,0)$ di $(0,1)$" vediamo immediatamente che se T è la variabile $\mathcal{G}(1/2)$ che governa la prima uscita di 0 allora la probabilità di A_1 si può scivere, usando l'indipendenza delle $\{X_i\}_{i\in\mathbb{N}}$, come

$$\begin{aligned} \mathbb{P}(X_{T+1} = 0) &= \sum_{i=0}^{\infty} \mathbb{P}(X_{T+1} = 0 | T = i)\mathbb{P}(T = i) = \sum_{i=0}^{\infty} \mathbb{P}(X_{i+1} = 0 | T = i)/2^i \\ &= \frac{1}{2} \sum_{i=0}^{\infty} \frac{1}{2^i} = 1/2. \end{aligned}$$

Quindi la probabilità che esca prima $(0,0)$ oppure $(0,1)$ è identica.

Sia ora N_i l'istante di prima uscita di w_i (per $i = 1, 2$). Sia, come sopra T l'istante di prima uscita di 0. Affinché esca $(0,1)$ bisogna che prima esca uno 0 (al tempo T) e poi si attenda la prima uscita di un 1, chiamiamo questo ulteriore lasso di tempo $\bar{T}$. Quindi $T_2 = T + \bar{T}$. Per l'assenza di memoria (o, equivalentemente, per l'indipendenza) $T, \bar{T} \sim \mathcal{G}(1/2)$ da cui $\mathbb{E}(T_2) = \mathbb{E}(T) + \mathbb{E}(\bar{T}) = 4$. Viceversa, per l'uscita di $(0,0)$ dobbiamo attendere che esca il primo 0, se il carattere seguente è 0 abbiamo finito altrimenti se è 1 dobbiamo attendere il primo 0 successivo e così via. In maniera rigorosa, sia la sequenza $\{S_i\}_{i\in\mathbb{N}}$ definita induttivamente da $S_{i+1} := \min\{n \geq 1 : X_{S_i+1+n} = 0\}$. Chiaramente $\{S_i\}_{i\in\mathbb{N}}$ sono iid distribuite come $\mathcal{G}(1/2)$ e S_{i+1} rappresenta il tempo di attesa dopo l'istante $S_i + 1$ per vedere comparire di nuovo O per la prima volta. Notiamo che se definiamo $Z_n := X_{\sum_{i=1}^{n}(S_i+1)}$ si ha che $\{Z_n\}_{n\in\mathbb{N}}$ sono iid distribuite come $\mathcal{B}(1/2)$. Sia quindi, Q la variabile $\mathcal{G}(1/2)$ che conta la prima uscita di 0 nella sequenza $\{Z_n\}_{n\in\mathbb{N}}$. Allora $\mathbb{E}(N_1) = \mathbb{E}(\sum_{i=1}^{Q}(S_i + 1)) = \mathbb{E}(Q) + \mathbb{E}(\sum_{i=1}^{Q} S_i)$. Si osservi che, essendo $\{Q, S_1, S_2, \dots\}$ una famiglia di variabili indipendenti, allora, utilizzando il Teorema di convergenza

monotona (o il Teorema di Beppo Levi) per commutare il valore atteso e la serie,

$$\begin{aligned}\mathbb{E}(N_1) &= \mathbb{E}(Q) + \mathbb{E}\Big(\sum_{i=1}^{Q} S_i\Big) = 2 + \mathbb{E}\Big(\sum_{k=0}^{\infty} \mathbb{1}_{\{Q=k\}} \sum_{i=1}^{Q} S_i\Big) \\ &= 2 + \sum_{k=0}^{\infty} \mathbb{E}\Big(\mathbb{1}_{\{Q=k\}} \sum_{i=1}^{k} S_i\Big) = 2 + \sum_{k=0}^{\infty} \mathbb{E}\Big(\mathbb{1}_{\{Q=k\}}\Big)\mathbb{E}\Big(\sum_{i=1}^{k} S_i\Big) \\ &= 2 + \sum_{k=0}^{\infty} \mathbb{P}(Q=k) k \mathbb{E}(S_1) = 2 + \mathbb{E}(Q)\cdot\mathbb{E}(S_1) = 6\end{aligned}$$

(lo stesso risultato si sarebbe ottenuto con l'attesa condizionata $\mathbb{E}(\cdot|Q=k)$; si veda ad esempio l'Esercizio 10.5.20). Naturalmente si ha anche, per simmetria nello scambio tra 0 e 1, che i valori attesi per le prime uscite di $(1,1)$ e $(1,0)$ sono 6 e 4 rispetivamente.

Nel caso generale,si può mostrare che, definito

$$I := \{i = 1, \dots, l : y_j = y_{l-i+j}, \forall j = 1, \dots, i\}$$

(ovviamente $l \in I$), allora

$$\mathbb{E}(N) = \sum_{i\in I} \prod_{j=1}^{i} p_{y_j}^{-1}.$$

Nel caso di w_1 si ha $I_1 = \{1,2\}$ mentre per w_2 si ha $I_2 = \{2\}$ da cui $\mathbb{E}(N_1) = 1/(p_0 + p_0 p_1) = 6$, mentre $\mathbb{E}(N_2) = 1/(p_0 p_1) = 4$.

4.4.3 Variabili di Poisson, processo di Poisson

La legge di Poisson si costruisce come "limite" (in maniera precisa si dice "limite in legge" o "limite in distribuzione) di una successione di variabili binomiali $B(n, \lambda/n)$ ($n \geq [1/\lambda]$ dove $\lambda > 0$).

Approfondimento

Vediamo in che senso una variabile di Poisson è limite di una successione di Binomiali. Sia $X_k \sim B(n_k, p_k)$ dove $n_k \to \infty$ e $n_k p_k \to \lambda \in \mathbb{R}$ se $k \to \infty$. Allora

$$\mathbb{P}(X_k = i) = \binom{n_k}{i} p_k^i (1-p_k)^{n_k - i}$$

Si osservi che l'esistenza di $p_i := \lim_{k\to+\infty} \mathbb{P}(X_k = i)$ implica in generale $\sum_{i=0}^{\infty} p_i \leq 1$. In questo caso, considerando che

$$\lim_{k\to+\infty} \left(1 - \frac{\alpha_k}{n_k}\right)^{n_k} = e^{-\alpha}$$

se $\alpha_k \to \alpha$ (quando $k \to \infty$) allora si ha

$$\begin{aligned}&\lim_{n\to+\infty} (1-p_k)^{n_k - i} = \lim_{k\to+\infty} \left(1 - \frac{p_k(n_k - i)}{n_k - i}\right)^{n_k - i} = e^{-\lambda}, \\ &\lim_{n\to+\infty} \frac{n_k!}{(n_k - i)!} p_k^i = \lambda^i,\end{aligned}$$

pertanto

$$\mathbb{P}(X_k = i) = \frac{1}{i!} \frac{n_k! p_k^i}{(n_k - i)!} (1-p_k)^{n_k - i} \sim \frac{\lambda^i}{i!} e^{-\lambda}.$$

Definizione 4.4.12. Una variabile X si dice di Poisson di parametro $\lambda > 0$ se

$$\mathbb{P}(X = i) = \begin{cases} \frac{\lambda^i}{i!} e^{-\lambda} & i \in \mathbb{N} \\ 0 & \text{altrimenti.} \end{cases}$$

Si scrive pertanto $X \sim P(\lambda)$.

Per quanto faremo in seguito ci interessa generalizzare la precedente definizione al caso $\lambda = 0$ e $\lambda = +\infty$: diremo che una variabile X a valori in $\overline{\mathbb{N}} := \mathbb{N} \cup \{+\infty\}$ ha legge $P(0)$ (risp. $P(+\infty)$) se e solo se $\mathbb{P}(X = 0) = 1$ (risp. $\mathbb{P}(X = +\infty) = 1$).

Osservazione 4.4.13. Sia X il numero di utenti che chiamano un centralino telefonico in un giorno. Si vuole conoscere la distribuzione di probabilità di X, sapendo che il numero n delle persone che *potrebbero* chiamare il centralino è molto grande , che le azioni di questi utenti sono indipendenti, e che in media si verificano λ chiamate al giorno.

Allora X è una variabile di Bernoulli $X \sim B(n, \lambda/n)$, e la sua densità di probabilità è

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Sapendo che, se $X \sim B(n, p)$, allora $\mathbb{E}(X) = np = \lambda$ e supponendo ignoto (ma grande) il numero n allora ha senso modellizzare con una sorta di limite per $n \to +\infty$:

$$\begin{aligned} p(k) &= \lim_{n\to+\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots \\ &= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k e^{-\lambda}}{k!}. \end{aligned}$$

Esempio 4.4.14. Un filo di rame di lunghezza L metri possiede delle imperfezioni. Sappiamo che in media si verificano λ imperfezioni ogni L metri, e che le posizioni delle imperfezioni sul filo sono variabili casuali indipendenti. Vogliamo sapere la funzione di probabilità della v.a. *numero di imperfezioni* del filo di lunghezza L.

Altri esempi che danno luogo alla stessa distribuzione di probabilità sono:

1. il numero di automobili che passa per un determinato incrocio in un determinato intervallo di tempo;
2. il numero di persone che si reca in un negozio in un giorno feriale;
3. il numero di guasti che si verificano in un impianto in un giorno lavorativo;
4. il numero di pixels difettosi di uno schermo a cristalli liquidi;

Approfondimento

Più delicata è la definizione del processo di Poisson: premettiamo alla definizione vera e propria un esempio.

Esempio 4.4.15. Riprendiamo l'esempio 1 e chiediamoci come varia la variabile aleatoria che governa il numero delle imperfezioni in L metri al variare della lunghezza L? Quello che ci si aspetta è che sezioni disgiunte del filo si comportino tutte come Poisson indipendenti tra loro e che a lunghezza uguale corrisponda parametro della legge di Poisson uguale (uniformità del filo). Queste condizioni sono compatibili e danno origine ad un processo (la cui esistenza è garantita da un teorema).

Definizione 4.4.16. Sia (X, Σ, μ) uno spazio con misura μ. Una famiglia di variabili aleatorie $\{X_\Delta\}_{\Delta\in\Sigma}$ si dice **processo di Poisson di intensità** μ se e solo se

1. Se $\{\Delta_i\}_{i=1}^n$ sono disgiunti allora $X_{\Delta_1}, \dots, X_{\Delta_n}$ sono indipendenti;
2. $X_\Delta \sim P(\mu(\Delta))$;
3. $\Delta \to X_\Delta(\omega)$ è una misura di conteggio per ogni $\omega \in \Omega$ (i.e. esiste $A = A_\omega$ tale che $\mu(B) = \#A \cap B$).

Esempio 4.4.17. Supponiamo che il numero di telefonate medio che arrivano ad un centralino in un giorno sia 10. Ci chiediamo quale sia la probabilità che in 3.5 giorni ne arrivino 25.
Lo spazio (X, Σ, μ) in questo caso è $(\mathbb{R}, \mathcal{R}, \mu)$ dove μ è univocamente determinata dai valori sugli intervalli $(a, b]$ $(b > a)$

$$\mu((a, b]) = |b - a|\epsilon \qquad (\textbf{Processo stazionario}).$$

Sappiamo che esiste un processo di Poisson di cui μ sia l'intensità: interpretiamo le variabili aleatorie $X_{(a,b]}$ come il numero di telefonate che arrivano al centralino nell'intervallo di tempo $(a, b]$. In questo caso $X_{(a,b]} \sim P(\mu((a, b])) = P(|b-a|\epsilon)$, pertanto il numero di telefonate che arrivano al centralino dipende solo dalla lunghezza dell'intervallo. La costante ϵ si calcola sapendo che $10 = \mathbb{E}(X_{(0,1]}) = \epsilon$. Pertanto $X_{(0,3.5]} \sim P(35)$ quindi

$$\mathbb{P}(X_{(0,3.5]} = 25) = e^{-35}\frac{35^{25}}{25!} \approx 1.62 \cdot 10^{-2}.$$

Osservazione 4.4.18. In generale, stabilita l'intensità del processo μ, la legge di X_Δ è univocamente determinata dal parametro $\mu(\Delta)$ che ne rappresenta anche il valore medio.

Perché si sceglie un processo di Poisson per modellizzare fenomeni come il numero di telefonate ad un centralino in un dato intervallo di tempo o il numero di difetti in una data area di materiale e così via?
La risposta è contenuta nel seguente teorema.

Teorema 4.4.19. *Sia $\{N_t\}_{n\geq 0}$ una collezione di variabili aleatorie a valori in $\mathbb{N} \cup \{+\infty\}$. Allora, dato $\lambda > 0$ le seguenti condizioni*

(a1) per ogni $t \geq 0$ si ha $N_t \sim P(\lambda t)$,
(a2) se $\{(a_i, b_i]\}_{i=1}^n$ sono intervalli disgiunti allora $\{N_{b_i} - N_{a_i}\}_{i=1}^n$ sono variabili indipendenti,

sono equivalenti alle seguenti

(b1) $N_0 = 0$ q. c.
(b2) se $\{(a_i, b_i]\}_{i=1}^n$ sono intervalli disgiunti allora $\{N_{b_i} - N_{a_i}\}_{i=1}^n$ sono variabili indipendenti
(b3) $\{N_{h+t} - N_h\}_{h\geq 0}$ hanno la stessa distribuzione

(b4) $\dfrac{\mathbb{P}(N_t \geq 2)}{t} \xrightarrow{t\to 0} 0$

(b5) $\dfrac{\mathbb{P}(N_t \geq 1) - \lambda t}{t} \xrightarrow{t\to 0} 0.$

Proprietà.

- Calcoliamo il valore atteso della legge di Poisson $P(\lambda)$:

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k p_X(k) = \sum_{k=0}^{\infty} \frac{k e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda$$

- Calcoliamo la varianza:

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{k=0}^{\infty} k^2 p_X(k) - \mathbb{E}(X)^2 = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} + \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} - \lambda^2 \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \mathbb{E}(X) - \lambda^2 = \lambda^2 e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} + \lambda - \lambda^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda.\end{aligned}$$

- Siano $X_1, \ldots, X_n$ v.a. indipendenti con legge di Poisson $X_i \sim P(\lambda_i)$ (λ_i possono essere diversi tra loro). Allora

$$X_1 + \cdots + X_n \sim P(\lambda_1 + \cdots + \lambda_n).$$

Cap. 5. Variabili aleatorie assolutamente continue

Per semplicità in questo capitolo ci limiteremo al caso reale unidimensionale, cioè al caso di variabili aleatorie a valori in $\mathbb{R}$.

Assumiamo che lo spazio campionario degli eventi sia uno spazio non discreto, e che i valori che può assumere la v.a. X formino un insieme continuo su $\mathbb{R}$.

La **legge** della v.a., come nel caso della v.a. discreta, è l'applicazione

$$A \mapsto \mathbb{P}(X \in A) =: \mathbb{P}_X(A) \quad \text{per ogni } A \in \mathcal{R};$$

osserviamo che la legge è univocamente determinata dai suoi valori

$$I \mapsto \mathbb{P}(X \in I) =: \mathbb{P}_X(I) \quad \text{per ogni intervallo } I \subseteq \mathbb{R}.$$

Definizione 5.0.20. Una variabile aleatoria (o la sua legge) si dice **assolutamente continua** (o più semplicemente **continua**) se e solo se esiste una funzione (misurabile ed integrabile) $f : \mathbb{R} \to \mathbb{R}$ tale che

$$\mathbb{P}_X(A) = \int_A f_X(t)dt \quad \forall A \in \mathcal{R}$$

(o equivalentemente, per ogni A intervallo reale), tale funzione prende il nome di **densità (continua)** della v.a. X.

Osservazione 5.0.21. Condizione necessaria e sufficiente affinché f_X sia una densità continua è

$$\begin{cases} f_X(t) \geq 0 \quad \forall t \in \mathbb{R} \\ \int_{\mathbb{R}} f_X(t)dt = 1 \end{cases}$$

- Attenzione: $f_X(x)$ non è una funzione di probabilità nel senso del precedente capitolo. In particolare $f_X(x) \neq \mathbb{P}(X = x)$, e non necessariamente $f_X(x) \leq 1$.

- $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b)$ per ogni $a, b \in \mathbb{R}$.

- La legge della v.a. continua è non nulla solo su intervalli non degeneri, mentre è nulla in singoli punti o insiemi numerabili di punti. Pertanto $\mathbb{P}(X = a) = 0$ per ogni $a \in \mathbb{R}$.

Nel caso di variabili aleatorie continue il calcolo della funzione di ripartizione è piuttosto semplice:

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(x)dx.$$

Nei punti in cui la densità $f_X(t)$ è continua, $F_X(t)$ è derivabile, e $F'_X(t) = f_X(t)$.

Approfondimento

Una condizione necessaria e sufficiente affinché X sia una v.a. continua è che la sua funzione di ripartizione F_X soddisfi la seguente proprietà: per ogni $\epsilon > 0$ esiste $\delta = \delta_\epsilon > 0$ tale che per ogni scelta di n intervalli reali disgiunti $\{[a_i, b_i]\}_{i=1}^n$ (con $n \in \mathbb{N}^*$) soddisfacenti $\sum_{i=1}^n |b_i - a_i| \leq \delta$ si abbia $\sum_{i=1}^n |F_X(b_i) - F_X(a_i)| \leq \epsilon$ (**assoluta continuità**).
Si dimostra inoltre che se una variabile aleatoria è continua allora esiste una densità che coincide con la derivata della funzione di ripartizione F_X dove quest'ultima è definita.
Diamo di seguito due condizioni sufficienti affinché una funzione di ripartizione F ammetta densità:

- (Per chi conosce l'integrale di Lebesgue) Una funzione di ripartizione F ammette densità se è derivabile ovunque, tranne al più un numero finito di punti dove però è continua; una densità è $f := F'$ (definita 0 per esempio dove F' non è definita).
- (Per chi conosce l'integrale di Riemann) Una funzione di ripartizione F ammette densità se è derivabile ovunque, tranne al più un numero finito di punti dove però è continua, e la derivata è Riemann integrabile; una densità è $f := F'$ (definita 0 per esempio dove F' non è definita).

5.1 Valore atteso per variabili aleatorie continue

Diamo nel seguito definizioni e proprietà per le v.a. continue di quantità analoghe a quelle già definite per le v.a. discrete.

Definizione 5.1.1. Una variabile aleatoria assolutamente continua X con densità f_X si dice che **ammette media** se e solo se

$$\int_{\mathbb{R}} |t| f_X(t) dt < +\infty;$$

in tal caso si chiama **valore atteso, o media, o speranza matematica** della v.a. X, il numero

$$\mathbb{E}(X) = \int_{\mathbb{R}} t f_X(t) dt$$

a condizione che l'integrale sia finito.

Proprietà

- Sia X una variabile continua, $a \neq 0$ e $b \in \mathbb{R}$ allora $Y = aX + b$ è una variabile continua con densità
$$f_Y(t) = f_X\left(\frac{t-b}{a}\right)\frac{1}{|a|}.$$
- $\mathbb{E}(aX_1 + b) = a\mathbb{E}(X_1) + b$
- $\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$
- $X_1, \ldots, X_n$ indipendenti: $\mathbb{E}(X_1 X_2 \ldots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2)\ldots\mathbb{E}(X_n)$.
- valore atteso di una funzione di v.a.: data una funzione $g : \mathbb{R} \to \mathbb{R}$ misurabile (ad esempio g continua) allora $g \circ X$ ammette media se e solo se
$$\int_{\mathbb{R}} |g(t)| f_X(t) dt < +\infty$$

(ad esempio g misurabile e limitata); in tal caso il valore medio è

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(t) f_X(t) dt.$$

5.2 Varianza e covarianza per variabili aleatorie continue

Definizione 5.2.1. Si definisce **varianza** di una v.a. continua X il numero

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int_{\mathbb{R}} [t - \mathbb{E}(X)]^2 f_X(t) dt \equiv \int_{\mathbb{R}} t^2 f_X(t) dt - \mathbb{E}(X)^2.$$

a condizione che l'integrale esista finito.

Proprietà

- $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \int_{\mathbb{R}} t^2 f_X(t) dt - \left(\int_{\mathbb{R}} t f_X(t) dt\right)^2$
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$
- $X_1, \ldots, X_n$ indipendenti: $\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$
- Per ogni v.a. continua X si ha $\mathrm{Var}(X) > 0$; la variabile

$$Y := \frac{X - \mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}}$$

  si dice **standardizzata** di X.

Come per le v.a. discrete (si veda la Definizione 3.4.1) se X e Y sono due v.a. aventi varianza finita allora la covarianza di X e Y risulta:

$$\mathrm{cov}(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right)$$

Ricordiamo che la covarianza viene anche denotata con il simbolo σ_{XY} e che ha le seguenti proprietà:

1. $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$
2. $\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
3. $\mathrm{cov}(X, X) = \mathrm{Var}(X)$
4. $\mathrm{cov}(aX + b, cY + d) = a \cdot c \cdot \mathrm{cov}(X, Y)$
5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{cov}(X, Y)$

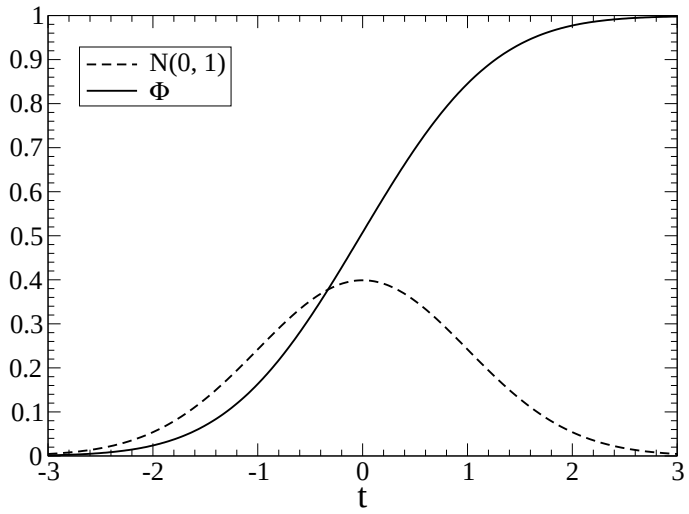
Anche nel caso delle v.a. continue una condizione necessaria (ma non suffiente!) affinché due v.a. X e Y siano indipendenti è che $\mathrm{cov}(X, Y) = 0$.

5.3 Modelli continui di variabili aleatorie

5.3.1 Densità uniforme

Diciamo che la v.a. X ha densità uniforme sull'intervallo $[a, b]$ ($X \sim U(a, b)$) se e solo se

$$f_X(t) = \frac{1}{b - a} \mathbb{1}_{[a,b]}(t)$$



Densità normale standard $\mathcal{N}(0,1)$ e sua funzione di ripartizione Φ

dove $\mathbb{1}_{[a,b]}$ è la *funzione indicatrice* dell'intervallo $[a,b]$ che vale 1 all'interno dell'intervallo e 0 fuori.

$$\mathbb{P}(t_1 < X < t_2) = \int_{t_1}^{t_2} \frac{1}{b-a} \mathbb{1}_{[a,b]}(t)dt = \frac{|[a,b] \cap [t_1,t_2]|}{b-a}$$

Nota: la funzione $f_X(t)$ in questo caso è discontinua.

Calcoliamo il valore atteso e la varianza di una variabile uniforme $X \sim U([a,b])$.

$$\mathbb{E}(X) = \int_a^b \frac{t}{b-a} \mathrm{d}t = \frac{b+a}{2};$$

mentre

$$\mathrm{Var}(X) = \int_a^b \frac{t^2}{b-a} \mathrm{d}t - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

5.3.2 Densità gaussiana (o normale)

- Densità gaussiana standard
  Una v.a. X si dice **normale standard** (e si scrive $X \sim \mathcal{N}(0,1)$) se la sua densità f_X è

  $$f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

  Si dimostra che questa funzione è integrabile su $\mathbb{R}$ con integrale pari a 1. Ovviamente si ha

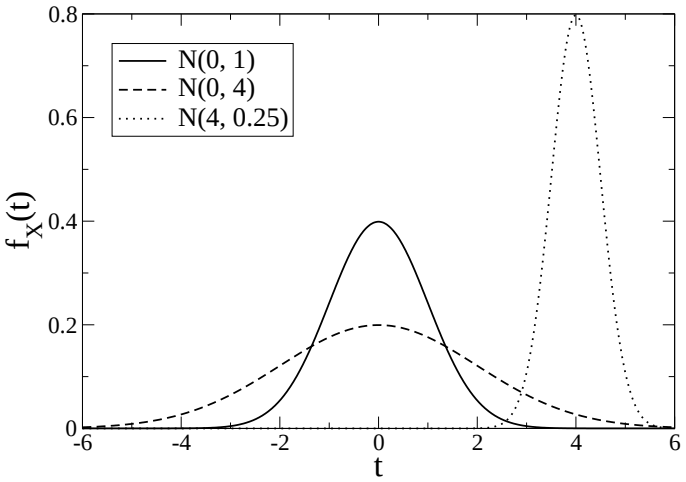
  $$\mathbb{P}(a < X < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

  e la funzione di ripartizione vale

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

  Calcoliamo valore atteso e varianza della normale standard.

  $$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} t e^{-t^2/2} dt = 0$$



Alcuni esempi di densità gaussiana

poiché l'integranda è una funzione dispari.

$$\mathbb{E}(X^2) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} t^2 e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} t \left[t e^{-t^2/2} \right] dt =$$
$$= \frac{-t e^{-t^2/2}}{\sqrt{2\pi}} \Bigg|_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-t^2/2} dt = 1$$
$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = 1$$

- Densità gaussiana.
  Una v.a. Y si dice **gaussiana** (o **normale**) (e si scrive $Y \sim \mathcal{N}(\mu, \sigma^2)$ con $\mu \in \mathbb{R}$ e $\sigma > 0$) quando è possibile scriverla come:

$$Y = \sigma X + \mu \quad \text{con } X \sim \mathcal{N}(0,1)$$

Si ha pertanto

$$\mathbb{P}(a < Y < b) = \mathbb{P}(a < \sigma X + \mu < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < X < \frac{b-\mu}{\sigma} \right) =$$
$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-t^2/2} dt = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

Quindi una v.a. continua è gaussiana se e solo se la sua densità è

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

per qualche $\mu \in \mathbb{R}$ e $\sigma > 0$.

La funzione di ripartizione vale

$$F_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \Phi\left(\frac{x-\mu}{\sigma} \right)$$

Calcoliamo valore atteso e varianza della gaussiana $Y \sim \mathcal{N}(\mu, \sigma^2)$. Sia X la v.a. normale standard: $X \sim \mathcal{N}(0,1)$.

$$\mathbb{E}(Y) = \mathbb{E}(\sigma X + \mu) = \sigma \mathbb{E}(X) + \mu = \mu$$
$$\mathrm{Var}(Y) = \mathrm{Var}(\sigma X + \mu) = \sigma^2 \mathrm{Var}(X) = \sigma^2$$

Si dimostra che se $X_1, \dots, X_n$ sono n v.a. *indipendenti* con distribuzione gaussiana $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, e $\{a_i\}_{i=1}^n$, $\{b_i\}_{i=1}^n$ sono due successioni reali, allora

$$\sum_{i=1}^{n}(a_i X_i + b_i) \sim \mathcal{N}\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n} a_i^2\sigma_i^2\right). \tag{5.1}$$

Osservazione 5.3.1. In generale se due variabili aleatorie normali non sono indipendenti, non si può concludere che la loro somma sia ancora una variabile normale. Inoltre, anche assumendo che due variabili aleatorie X, Y siano normali, non è detto che $\mathrm{cov}(X, Y) = 0$ implichi che le variabili siano indipendenti. Mostriamo tutto ciò con un esempio.

Sia X una variabile aleatoria con distribuzione $\mathcal{N}(0, 1)$ e si definisca la variabile Y_α (al variare di $\alpha \in \mathbb{R}$) come segue

$$Y_\alpha(\omega) := \begin{cases} X(\omega) & \text{if } |X(\omega)| \geq \alpha \\ -X(\omega) & \text{if } |X(\omega)| < \alpha. \end{cases}$$

Si osservi che

- Y_α ha distribuzione $\mathcal{N}(0, 1)$;
- X ed Y_α non sono indipendenti: infatti

$$\mathbb{P}(X \geq \alpha, Y \leq -\alpha) = 0 \neq \mathbb{P}(X \geq \alpha)\mathbb{P}(Y \leq -\alpha);$$

- $$X(\omega) + Y_\alpha(\omega) = \begin{cases} 2X(\omega) & \text{if } |X(\omega)| \geq \alpha \\ 0 & \text{if } |X(\omega)| < \alpha, \end{cases}$$

  quindi $X + Y_\alpha$ non è normale s $\alpha > 0$;

- $$\begin{aligned} \mathrm{cov}(X, Y_\alpha) =& E[XY_\alpha] - E[X]E[Y_\alpha] = E[X]E[Y_\alpha] \\ =& E[X^2\mathbb{I}_{|X|\geq\alpha}] - E[X^2\mathbb{I}_{|X|<\alpha}] = 2E[X^2\mathbb{I}_{|X|\geq\alpha}] - 1, \end{aligned}$$

  ora essendo $\alpha \mapsto E[X^2\mathbb{I}_{|X|\geq\alpha}]$ continua e poiché

$$\lim_{\alpha\to+\infty} E[X^2\mathbb{I}_{|X|\geq\alpha}] = 0, \qquad \lim_{\alpha\to 0} E[X^2\mathbb{I}_{|X|\geq\alpha}] = 1,$$

  allora esiste un valore $\alpha_0 > 0$ (precisamente $\alpha_0 = q_{3/4}$, dove $q \equiv \phi^{-1}$ è la funzione quantile della normale standard, si vedano i Paragrafi 3.5 e 5.4) tale che $\mathrm{cov}(X, Y_{\alpha_0}) = 0$.

Come costruzione concreta per X si prenda lo spazio di probabilità $((0, 1), \mathbb{B}_{(0,1)}, m_l)$ e come variabile aleatoria la funzione quantile ϕ^{-1}.

5.3.3 La legge esponenziale

Sia $X = \{X_t\}_{t\geq 0}$ un processo di Poisson di intensità scalare ν (ossia $X_t \sim P(\nu t)$). Si dice *v.a. esponenziale* di parametro ν, e si denota con $Y \sim \mathrm{Esp}(\nu)$ la v.a. Y che misura l'istante del primo successo del processo di Poisson. Y dunque è una v.a. continua.

Ad esempio se il processo di Poisson è quello che descrive il numero di guasti nel tempo in un apparecchio meccanico, la v.a. che descrive il tempo di attesa del suo primo guasto è una v.a. esponenziale.

Troviamo l'espressione della densità esponenziale. Dalla definizione:

$$\mathbb{P}(Y > t) = \mathbb{P}(X_t = 0) = e^{-\nu t} \quad \text{per } t > 0, \qquad \mathbb{P}(Y > t) = 1 \quad \text{per } t \leq 0$$

Quindi la funzione di ripartizione è

$$F_Y(t) = \mathbb{P}(Y \leq t) = 1 - \mathbb{P}(Y > t) = 1 - e^{-\nu t} \quad \text{per } t > 0, \quad F_Y(t) = 0 \quad \text{per } t \leq 0$$

La densità di Y è la derivata della f.d.r.:

$$f_Y(t) = \nu e^{-\nu t} \quad \text{per } t > 0, \qquad f_Y(t) = 0 \quad \text{per } t \leq 0$$

Calcoliamo valore atteso e varianza della legge esponenziale:

$$\mathbb{E}(Y) = \int_{\mathbb{R}} t f_Y(t) dt = \int_0^{+\infty} t\nu e^{-\nu t} dt = -te^{-\nu t}\big|_0^{+\infty} + \int_0^{+\infty} e^{-\nu t} dt = \frac{1}{\nu}$$

$$\mathbb{E}(Y^2) = \int_{\mathbb{R}} t^2 f_Y(t) dt = -t^2 e^{-\nu t}\big|_0^{+\infty} + \int_0^{+\infty} 2te^{-\nu t} dt = \frac{2}{\nu^2}$$

perciò

$$\mathrm{Var}(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \frac{1}{\nu^2}$$

La legge esponenziale descrive anche il tempo di attesa tra due successi consecutivi in un processo di Poisson.

Nell'esempio dei guasti di un apparecchio meccanico essa fornisce la probabilità del tempo di attesa tra due guasti successivi. Il valor medio $1/\nu$ viene anche detto *tempo medio tra due guasti*.

Una proprietà importante della legge esponenziale è la sua **assenza di memoria**: se aspettiamo un successo nel processo di Poisson, e dopo un tempo T non si è verificato, la probabilità di dover aspettare ancora per un tempo t è uguale a quella che avevamo in partenza. Formalmente:

$$\mathbb{P}(Y > T + t | Y > T) = \mathbb{P}(Y > t)$$

Infatti

$$\begin{aligned}\mathbb{P}(Y > T + t | Y > T) &= \frac{\mathbb{P}(Y > T + t, Y > T)}{\mathbb{P}(Y > T)} = \frac{\mathbb{P}(Y > T + t)}{\mathbb{P}(Y > T)} \\ &= \frac{e^{-\nu(T+t)}}{e^{-\nu T}} = e^{-\nu t} = \mathbb{P}(Y > t).\end{aligned}$$

Osservazione 5.3.2. Si può dimostrare che se una variabile assolutamente continua gode della proprietà di assenza di memoria, allora ha legge esponenziale.

5.3.4 La legge gamma

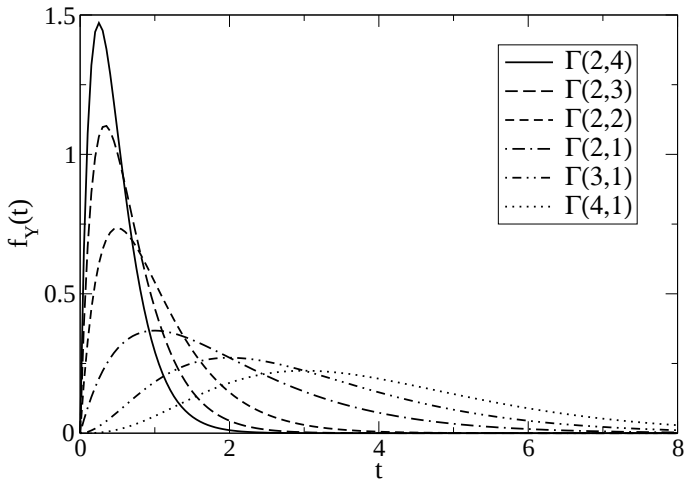
Sia X un processo di Poisson di intensità ν. Si dice **gamma** la v.a. che misura l'istante dell'n-esimo successo del processo di Poisson. Si scrive $Y \sim \Gamma(n, \nu)$.

Poiché il tempo di attesa tra due successi consecutivi è una v.a. di legge $\mathrm{Esp}(\nu)$, e poiché gli eventi sono indipendenti, la legge gamma è somma di n v.a. i.i.d. di legge $\mathrm{Esp}(\nu)$. La funzione di ripartizione $F_Y(t) = 1 - \mathbb{P}(Y > t)$ si ottiene notando che l'evento $Y > t$ significa *l'n-esimo successo avviene dopo l'istante t* e coincide con l'evento *il numero di successi fino all'istante t è* $\leq n - 1$. Perciò

$$F_Y(t) = 1 - \mathbb{P}(Y > t) = 1 - \mathbb{P}(X \leq n - 1) = 1 - \sum_{k=0}^{n-1} e^{-\nu t} \frac{(\nu t)^k}{k!}$$

La densità continua è la derivata della f.d.r.:

$$f_Y(t) = \frac{d}{dt} F_Y(t) = \nu \sum_{k=0}^{n-1} e^{-\nu t} \frac{(\nu t)^k}{k!} - \nu \sum_{k=1}^{n-1} e^{-\nu t} \frac{(\nu t)^{k-1}}{(k-1)!} = \nu e^{-\nu t} \frac{(\nu t)^{n-1}}{(n-1)!}$$



Densità delle leggi $\Gamma(n, \nu)$

Definizione 5.3.3. Una variabile assolutamente continua X si dice **gamma** di parametri $\alpha \geq 0$ e $\nu > 0$ (si scrive $X \sim \Gamma(\alpha, \nu)$) se e solo se la sua densità è

$$\rho_X(t) := \mathbb{1}_{(0,+\infty)}(t)\nu^\alpha e^{-\nu t}\frac{t^{\alpha-1}}{\Gamma(\alpha)}$$

dove $\Gamma(\alpha) := \int_0^{+\infty} e^{-t}t^{\alpha-1}\mathrm{d}t$.

- Nel caso particolare $n = 1$ la legge gamma coincide con la legge esponenziale: $\Gamma(1, \nu) = \mathrm{Esp}(\nu)$, e $f_Y(t) = \nu e^{-\nu t}$
- Il valore atteso e la varianza della legge gamma sono semplici da ottenere poiché $Y \sim \Gamma(n, \nu)$ è somma di n v.a. i.i.d. $X \sim \mathrm{Esp}(\nu)$:

  $$\mathbb{E}(Y) = n\mathbb{E}(X) = \frac{n}{\nu}, \qquad \mathrm{Var}(Y) = n\mathrm{Var}(X) = \frac{n}{\nu^2}.$$

  Più in generale si può mostrare che se $Y \sim \Gamma(\alpha, \nu)$ si ha $\mathbb{E}(Y) = \alpha/\nu$ e $\mathrm{Var}(Y) = \alpha/\nu^2$ per ogni $\alpha \geq 0$, $\nu > 0$.
- Se $X_1, \ldots, X_n$ sono variabili aleatorie indipendenti tali che $X_i \sim \Gamma(\alpha_i, \nu)$ allora

  $$\sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n \alpha_i, \nu\right).$$

5.4 Quantili per una variabile aleatoria assolutamente continua

In molti problemi statistici occorre ragionare in direzione inversa, ossia assegnato $\alpha \in (0, 1)$, determinare x tale che $\mathbb{P}(X \leq x) = \alpha$. È evidente che esiste almeno una soluzione a questa equazione in x se e solo se $\alpha \in \mathrm{Rg}(F_X)$ (i.e. $\#F_X^{-1}(\alpha) \geq 1$), mentre esiste al più una soluzione se e solo se $\#F_X^{-1}(\alpha) \leq 1$.

Per una variabile aleatoria assolutamente continua, essendo la funzione di ripartizione F una funzione continua, allora per ogni $\alpha \in (0, 1)$ esiste una soluzione all'equazione $F_X(x) = \alpha$. Pertanto $F_X(Q_{F_X}(\alpha)) = \alpha$. Tale soluzione è unica se e solo se F_X è strettamente crescente; questo avviene,

per esempio, nel caso in cui la densità ρ_X soddisfa $\rho_X(t) > 0$ per ogni $t \in \mathbb{R}$ tranne un insieme al più numerabile di punti. Infine notiamo che $\mathbb{P}(X \leq x) = \mathbb{P}(X < x)$ poiché, per ogni variabile assolutamente continua, vale $\mathbb{P}(X = x) = 0$.

Nota: nel resto di questo paragrafo supporremo sempre che la funzione di ripartizione F_X sia strettamente crescente.

Sotto queste ipotesi quindi la funzione dei quantili (si veda il Paragrafo 3.5) è l'inversa della funzione di ripartizione:

$$\mathbb{P}(X \leq q_\alpha) \equiv F(q_\alpha) = \alpha, \quad q_\alpha = F^{-1}(\alpha) \equiv Q_{F_X}(\alpha).$$

Definizione 5.4.1. Si definisce **quantile di ordine** α (o quantile α-esimo) per una variabile aleatoria continua con funzione di ripartizione strettamente crescente X, il valore della funzione dei quantili in corrispondenza al valore α, i.e. $Q_{F_X}(\alpha)$ (spesso indicato semplicemente come q_α).

Ovviamente vale la proprietà:

$$\mathbb{P}(X > q_{1-\alpha}) \equiv \mathbb{P}(X \geq q_{1-\alpha}) = \alpha.$$

Definizione 5.4.2. Una legge $\mathbb{P}_X$ su $\mathbb{R}$, si dice **simmetrica** se e solo se per ogni $A \in \mathcal{R}$

$$\mathbb{P}_X(A) = \mathbb{P}_X(-A)$$

dove $-A := \{-t : t \in A\}$.

Teorema 5.4.3. *Una variabile assolutamente continua X ha legge simmetrica se e solo se esiste una versione della densità f tale che $f(x) = f(-x)$. Se X soddisfa la precedente condizione allora $F_X(t) + F_X(-t) = 1$ per ogni $t \in \mathbb{R}$ e $q_\alpha + q_{1-\alpha} = 0$. Inoltre*

$$\mathbb{P}(|X| \leq q_{(1+\alpha)/2}) = \alpha$$

Dimostrazione. Dimostriamo solo la seconda parte. Dalla simmetria si ha

$$F_X(-t) = \mathbb{P}(X \leq -t) = \mathbb{P}(X \geq t) = \mathbb{P}(X > t) = 1 - \mathbb{P}(X \leq t) = 1 - F_X(t).$$

Poichè $F_X(q_\alpha) = \alpha$ e $q_{F_X(t)} = t$ si ha

$$q_{1-\alpha} = q_{1-F_X(q_\alpha)} = q_{F_X(-q_\alpha)} = -q_\alpha.$$

Infine dalla simmetria e dall'assoluta continuità, si ha $\mathbb{P}(X \leq -x) = \mathbb{P}(X \geq x) = 1 - \mathbb{P}(X \leq x)$ da cui

$$\mathbb{P}(|X| < x) = 1 - \mathbb{P}(X \leq x) - \mathbb{P}(X > x) = 2\mathbb{P}(X \leq x) - 1$$

pertanto

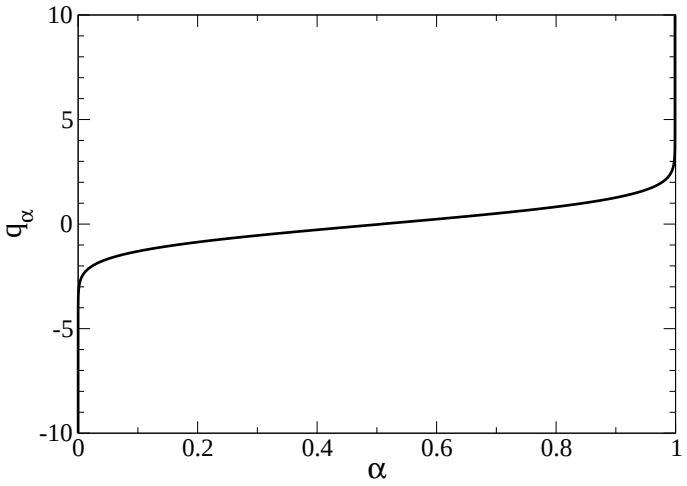
$$\mathbb{P}(|X| \leq x) = \alpha \iff \mathbb{P}(X \leq x) = \frac{1+\alpha}{2} \iff x = q_{(1+\alpha)/2}.$$

□

Un caso particolare estremamente utile riguarda i quantili della normale standard. Consideriamo la v.a. normale standard $X \sim \mathcal{N}(0,1)$ di densità $f_X(t) = \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ e f.d.r. $F_X(t) = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du$

Il quantile q_α è l'unico numero che soddisfa

$$\Phi(q_\alpha) = \mathbb{P}(X < q_\alpha) = \alpha$$



Quantili della normale standard

Per la simmetria della legge gaussiana rispetto a $x = 0$ si ha:

$$\Phi(-x) = 1 - \Phi(x),\ \forall x \in \mathbb{R}, \qquad q_{1-\alpha} = -q_\alpha,\ \forall \alpha \in (0,1).$$

$$\Phi(-q_\alpha) = 1 - \Phi(q_\alpha) = 1 - \alpha$$

Sempre per la proprietà di simmetria si ha anche:

$$\mathbb{P}(|X| < q_{\frac{1+\alpha}{2}}) = \alpha \qquad \mathbb{P}(|X| > q_{1-\frac{\alpha}{2}}) = \alpha.$$

Infine è possibile calcolare i quantili di una normale $\mathcal{N}(\mu, \sigma^2)$ da quelli di una normale standard qualora questi ultimi fossero noti. Infatti se $Y = \sigma X + \mu$ dove $X \sim \mathcal{N}(\mu, \sigma^2)$ allora $Q_Y(\alpha$ soddisfa

$$\alpha = \mathbb{P}(Y \leq Q_Y(\alpha)) = \mathbb{P}\left(X \leq \frac{Q_Y(\alpha) - \mu}{\sigma}\right)$$

da cui, per definizione, $Q_Y(\alpha) = \sigma\Phi(\alpha) + \mu$ (come abbiamo già mostrato nel Paragrafo 3.5).

Alcuni quantili della normale standard sono riassunti nella seguente tabella:

α	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
q_α	1.2816	1.6449	1.96	2.3263	2.7578	3.0902	3.2905

5.5 Utilizzo delle tavole e approssimazione della normale standard

Si possono utilizzare le tavole per calcolare i valori assunti dalla funzione di ripartizione e della sua inversa (cioè per calcolare i quantili).

Nelle tavole "standard" si possono trovare valori della funzione di ripartizione $\Phi(t)$ in corrispondenza a tutti i valori $t \in [0, 3.3]$ con scarto di 0.01. Nella tabella al posto (i, j) (riga i e colonna j) si trova il valore di $\Phi((i-1)\cdot 0.1 + (j-1)\cdot 0.01)$.
Supponiamo, ad esempio, di voler calcolare il valore $\Phi(2.33)$; si guarda nella tabella il valore sulla 24-esima riga (contrassegnata da 2.3) e 4-a colonna (contrassegnata da 0.03) cioè 0.9901. Nel caso di valori t negativi si può utilizzare la formula $\Phi(t) = 1 - \Phi(-t)$. Se invece il valore t non fosse in tabella possiamo distinguere due casi:

(i) se $t > 3.3$ allora essendo la funzione di ripartizione crescente si ha che $\Phi(t) > \Phi(3.3) \approx 0.997$, in genere si accetta l'approssimazione ad 1;

(ii) se $t_1 < t < t_2$, dove t_1, t_2 sono valori tabulati, si può applicare la cosiddetta approssimazione lineare:

$$\Phi(t) \approx \frac{t - t_1}{t_2 - t_1}(\Phi(t_2) - \Phi(t_1)) + \Phi(t_1).$$

Ad esempio se si volesse calcolare il valore $\Phi(2.334)$, procediamo scegliendo $t_1 := 2.33$, $t_2 := 2.34$ da cui

$$\Phi(2.334) \approx \frac{2.334 - 2.33}{2.34 - 2.33}(0.9904 - 0.9901) + 0.9901 \approx 0.9902$$

contro il valore approssimato calcolato da Matlab **normcdf**$(2.334) \approx 0.9902$.

Per il calcolo del quantile si procede in maniera inversa, essendo quest'ultimo il valore Φ^{-1}. Supponiamo di voler calcolare il quantile $q_{0.7324}$, si cerca nella tabella (zona centrale) il valore che in questo caso è al posto $(7, 3)$ corrispondente alla riga identificata da 0.6 e la colonna identificata da 0.02, pertanto $q_{0.7324} \approx 0.62$. Se il valore di α fosse in $(0, 1/2)$ si utilizza l'uguaglianza $q_\alpha = -q_{1-\alpha}$.

Se il valore non è compreso in tabella, allora si cercano due valori α_1, α_2 in tabella tali che $\alpha_1 < \alpha < \alpha_2$; supponendo che $\Phi(t_1) = \alpha_1$ e $\Phi(t_2) = \alpha_2$ si può utilizzare l'approssimazione lineare

$$q_\alpha \approx \frac{\alpha - \alpha_1}{\alpha_2 - \alpha_1}(t_2 - t_1) + t_1.$$

Ad esempio si voglia calcolare $q_{0.23}$; utilizzando la formula $q_\alpha = -q_{1-\alpha}$ cerchiamo il valore $q_{0.77}$. Dalla tabella vediamo che

$$\Phi(0.73) = 0.7673 < 0.77 < 0.7704 = \Phi(0.74)$$

da cui

$$q_{0.77} \approx \frac{0.77 - 0.7673}{0.7704 - 0.7673}(0.74 - 0.73) + 0.73 \approx 0.7387$$

pertanto $q_{0.23} \approx -0.7387$ contro il valore approssimato calcolato con Matlab **norminv**$(0.23) \approx -0.7388$.

Cap. 6. Alcuni cenni al calcolo degli indici per variabili aleatorie generiche

6.1 Integrazione rispetto ad una misura positiva

Approfondimento

Si supponga di avere uno spazio $(\Omega, \mathcal{F}, \mathbb{P})$ dove $\mathbb{P}$ è una misura positiva qualsiasi (si pensi pure ad una misura di probabilità). Una funzione $f : \Omega \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$ si dice **semplice** se e solo se $\#\mathrm{Rg}(f) < +\infty$ i.e. se e solo se esistono $a_1, \ldots, a_n \in \mathbb{R}$ e $M_1, \ldots, M_n \in \mathcal{F}$ tali che

$$f(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{M_i}(\omega), \qquad \forall \omega \in \Omega.$$

Si estendono a $\overline{\mathbb{R}}$ le operazioni nella maniera seguente:

$$\begin{aligned} a + +\infty &:= +\infty, \quad a + -\infty := -\infty, \\ a - +\infty &:= -\infty, \quad a - -\infty := +\infty, \end{aligned}$$

$$a \cdot \pm\infty := \begin{cases} \pm\infty & a > 0 \\ \mp\infty & a < 0 \\ 0 & a = 0. \end{cases}$$

Si definisce l'integrale di una funzione semplice a valori in $[0, +\infty]$

$$\int_\Omega f \mathrm{d}\mathbb{P} := \sum_{i=1}^{n} a_i \mu(M_i),$$

lo si estende alle funzioni misurabili non negative f a valori in $[0, +\infty]$ come

$$\int_\Omega f \mathrm{d}\mathbb{P} := \sup \left\{ \int_\Omega g \mathrm{d}\mu : g \text{ funzione semplice}, f \geq g \right\}.$$

Una funzione misurabile f si dice **integrabile** se e solo se $\int_\Omega |f| \mathrm{d}\mathbb{P} < +\infty$; questo equivale a dire che le due funzioni misurabili non negative $f^+(\omega) := \max(0, f(\omega))$ e $f^-(\omega) := -\min(0, f(\omega))$ sono entrambe integrabili e per definizione

$$\int_\Omega f \mathrm{d}\mathbb{P} := \int_\Omega f^+ \mathrm{d}\mathbb{P} - \int_\Omega f^- \mathrm{d}\mathbb{P}.$$

A volte si preferisce dare una definizione più debole richiedendo che almeno una delle due funzioni f^+ ed f^- sia integrabile, la definizione è poi la stessa con le convenzioni adottate sopra (**debole integrabilità**). Infine si definisce, per ogni $M \in \mathcal{F}$, e per ogni funzione integrabile f

$$\int_M f \mathrm{d}\mathbb{P} := \int_\Omega \mathbb{1}_F \cdot f \mathrm{d}\mathbb{P}.$$

Definizione 6.1.1. Una proprietà $\mathcal{P}$, dipendente da $\omega \in \Omega$, si dice che vale **quasi ovunque** (o in gergo probabilistico **quasi certamente**) se e solo se esiste un insieme $M \in \mathcal{F}$ tale che $\mathbb{P}(M^\complement) = 0$ e per ogni $\omega \in M$ la proprietà $\mathcal{P}$ vale.

Valgono le seguenti proprietà:

- Se f e g sono due funzioni integrabili e $a, b \in \mathbb{R}$ allora $af + bg$ è integrabile e vale
$$\int_\Omega (af + bg)\mathrm{d}\mathbb{P} = a \int_\Omega f \mathrm{d}\mathbb{P} + b \int_\Omega g \mathrm{d}\mathbb{P}.$$
- Se f e g sono due funzioni integrabili e $f \geq g$ quasi certamente allora
$$\int_\Omega f \mathrm{d}\mathbb{P} \geq \int_\Omega g \mathrm{d}\mathbb{P}$$
e vale l'uguaglianza se e solo se $f = g$ quasi certamente.
- Se f è integrabile allora f è quasi certamente finita.
- Sia $f : \Omega \to \mathbb{R}$ integrabile e $g : \mathbb{R} \to \mathbb{R}$ misurabile; allora $g \circ f$ (è sicuramente misurabile) è integrabile rispetto alla misura $\mathbb{P}$ se e solo se g è integrabile rispetto alla legge $\mathbb{P}_f$ e vale
$$\int_\Omega g \circ f \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g \mathrm{d}\mathbb{P}_f.$$
In particolare se definiamo $\int_A f \mathrm{d}\mathbb{P} := \int_\Omega \mathbb{1}_A \cdot f \mathrm{d}\mathbb{P}$ (dove A è un qualsiasi sottoinsieme misurabile di Ω) si ha
$$\int_{g^{-1}(A)} g \circ f \mathrm{d}\mathbb{P} = \int_A g \mathrm{d}\mathbb{P}_f$$
dove A è un sottoinsieme misurabile di $\mathbb{R}$.
- Se $\mathbb{P}_f(A) = \int_A \rho_f \mathrm{d}x$ si ha che
$$\int_{g^{-1}(A)} g \circ f \mathrm{d}\mathbb{P} = \int_A g \mathrm{d}\mathbb{P}_f = \int_A g \cdot \rho_f \mathrm{d}x.$$

6.2 Media e varianza

Approfondimento

Da qui in poi supporremo che $\mathbb{P}$ sia una misura di probabilità e che una variabile sia a valori in $\mathbb{R}$ (e non in $\overline{\mathbb{R}}$).

Definizione 6.2.1. Si definisce **valore medio** di una variabile integrabile X il valore

$$\mathbb{E}(X) := \int_\Omega X \mathrm{d}\mathbb{P}.$$

Valgono le seguenti proprietà

- Sia X è una variabile, è integrabile se e solo se

$$\int_{\mathbb{R}} |x| \mathrm{d}\mathbb{P}_X < +\infty$$

  e vale

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \mathrm{d}\mathbb{P}_X.$$

- Se $F(t) := \mathbb{P}(X \leq t)$ allora X è integrabile se e solo se

$$\int_{(0,+\infty)} (1 - F(t)) \mathrm{d}t + \int_{(-\infty,0)} F(f) \mathrm{d}t < +\infty$$

  e vale

$$\mathbb{E}(X) = \int_{(0,+\infty)} (1 - F(t)) \mathrm{d}t - \int_{(-\infty,0)} F(f) \mathrm{d}t.$$

- Più in particolare

$$\int_\Omega X^+ \mathrm{d}\mathbb{P} = \int_{(0,+\infty)} (1 - F(t)) \mathrm{d}t$$
$$\int_\Omega X^- \mathrm{d}\mathbb{P} = \int_{(-\infty,0)} F(f) \mathrm{d}t.$$

Definizione 6.2.2. Si definisce **varianza** di una variabile X il valore

$$\mathrm{Var}(X) := \mathbb{E}((X - \mathbb{E}(X))^2)$$

qualora $\mathbb{E}(X^2) < +\infty$ (la variabile risulta automaticamente integrabile).

Valgono le seguenti proprietà.

- $\mathrm{Var}(X) \geq 0$ e vale $\mathrm{Var}(X) = 0$ se e solo se X è quasi certamente costante.

- Per il calcolo della varianza vale:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \int_\Omega (X - \mathbb{E}(X))^2 \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} (x - \mathbb{E}(X))^2 \mathrm{d}\mathbb{P}_X$$
$$= \int_\Omega X^2 \mathrm{d}\mathbb{P} - \mathbb{E}(X)^2 = \int_{\mathbb{R}} x^2 \mathrm{d}\mathbb{P}_X - \mathbb{E}(X)^2.$$

- Il valore atteso minimizza la funzione $v(y) := \mathbb{E}((X-y)^2)$. Infatti $v(y) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 + (\mathbb{E}(X) - y)^2$ che ha un minimo in $y = \mathbb{E}(X)$.

È possibile dimostrare che le definizioni appena date si riducono a quelle viste nei precedenti capitoli nel caso di variabili discrete ed assolutamente continue.

Approfondimento

Analogamente a quanto visto in uno degli approfondimenti del Capitolo 2, si ha che tra le relazioni seguenti, valide per variabili aleatorie $X_1, \ldots, X_n$ tali che $\mathbb{E}(|X_i|^2) < +\infty$ per ogni i,

1. esistono $\{a_i\}_{i=1}^n$ scalari non tutti nulli tali che
$$\sum_{i=1}^n a_i X_i = 0, \qquad q.c. \quad \text{lineare dipendenza;}$$
2. esistono $\{a_i\}_{i=0}^n$ scalari non tutti nulli tali che
$$\sum_{i=1}^n a_i X_i + a_0 = 0, \qquad q.c. \quad \text{dipendenza affine;}$$
3. $\det(\{\mathbb{E}(X_i X_j)\}_{i,j}) = 0$;
4. $\det(\{\mathrm{cov}(X_i, X_j)\}_{i,j}) = 0$;
5. le variabili non sono a due a due scorrelate;
6. le variabili non sono indipendenti (in senso probabilistico);

sussistono le seguenti implicazioni

$$\begin{array}{ccc} (1) & \Longleftrightarrow & (3) \\ \Downarrow & & \\ (2) & \Longleftrightarrow & (4) \\ \Downarrow & & \\ (5) & & \\ \Downarrow & & \\ (6) & & \end{array}$$

Quindi l'indipendenza probabilistica implica quella lineare (e quella affine), ma non vale il viceversa. Come controesempio basta considerare una coppia X e Y di variabili gaussiane a media nulla, scorrelate ma non indipendenti. Con queste prescrizioni

$$\det \begin{pmatrix} \mathrm{Var}(X) & 0 \\ 0 & \mathrm{Var}(Y) \end{pmatrix} \neq 0$$

che implica sia l'indipendenza affine che quella lineare.

6.3 Supporto di una misura e valori assunti da una funzione misurabile

Approfondimento

Supponiamo ora che lo spazio di misura $(\Omega, \mathcal{F}, \mathbb{P})$ sia dotato della σ-algebra di borel $\mathcal{F} = \sigma(\tau)$ rispetto ad una topologia τ. L'obbiettivo è quello di trovare un insieme $M \in \mathcal{F}$ ragionevolmente piccolo tale che $\mathbb{P}(M^{\complement}) = 0$ e quindi $\mathbb{P}(A) = \mathbb{P}(A \cap M)$ per ogni $A \in \mathcal{F}$.

Definiamo **range essenziale** di una funzione misurabile $f : X \to Y$ dove $(Y, \sigma(\tau_Y))$ è uno spazio misurabile con σ-algebra di borel, l'insieme $\text{essrg}(f) := \{y \in Y : \mathbb{P}(f^{-1}(V)) > 0, \forall V \in \tau_Y \text{ t.c. } x \in V\}$ (per questa definizione non è importante che $\mathcal{F}$ sia una σ-algebra di borel).

Per risolvere il nostro problema, si potrebbe utilizzare la cosiddetta **misura esterna** ed individuare un insieme M in generale non misurabile, oppure attenersi alla seguente strategia. Ricordiamo che τ si dice **topologia II-numerabile** se e solo se ammette una base numerabile di aperti. Esempi di spazi topologici II-numerabile sono gli spazi $\mathbb{R}^n$ con l'usuale topologia euclidea.

Teorema 6.3.1. *Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di misura dotato di una σ-algebra di borel generata da una topologia τ. Siano*

$$M := \{x \in X : \mathbb{P}(V) > 0, \forall V \in \tau \text{ t.c. } x \in V\}, \qquad M_1 := \bigcup_{V \in \tau : \mathbb{P}(V) = 0} V.$$

M è chiuso, $M_1 = M^{\complement}$ è aperto, quindi sono entrambi misurabili. Inoltre, se τ è II-numerabile allora $\mathbb{P}(M_1) = 0$.
In particolare se $\mathbb{P} = \mathbb{P}_Z$ con Z funzione misurabile a valori in X, allora $M = \text{essrg}(Z)$. Se $\mathbb{P}_Z$ è discreta allora $M := \overline{\{x \in X : \mathbb{P}_Z(\{x\}) > 0\}}$. Se $X = \mathbb{R}^n$ e $\mathbb{P}_Z$ è assolutamente continua rispetto alla misura di Lebesgue allora $M = \text{ess supp}(\rho)$; in particolare se $\text{supp}(\rho) := \{x : \rho(X) > 0\}$ è un intervallo ristretta al quale ρ è continua, allora $M = \overline{\text{supp}(\rho)}$.

In generale M, generato da $\mathbb{P}_Z$, si interpreta come l'insieme dei valori possibili per la funzione misurabile Z.

Cap. 7. Teorema Centrale del Limite e Legge dei grandi numeri

7.1 Teorema Centrale del Limite ed approssimazioni gaussiane

7.1.1 Teorema Centrale del Limite

Teorema 7.1.1. *Sia $\{X_i\}_{i\geq 1}$ una successione di v.a. (a valori in $\mathbb{R}$) indipendenti identicamente distribuite con valore atteso $\mathbb{E}(X_i) = \mu$ e varianza finita $\mathrm{Var}(X_i) = \sigma^2 > 0$. Definiamo quindi la v.a.* **media campionaria** *$\overline{X}_n$ nel modo seguente*

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

evidentemente la variabile aleatoria $\overline{X}_n$ restituisce, per ogni $\omega \in \Omega$ fissato, la media campionaria dei valori numerici $X_1(\omega), \ldots, X_n(\omega)$. Allora definita la **media campionaria standardizzata** *H_N^* come*

$$H_n^* = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{\mathrm{Var}(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

vale, per ogni t fissato e per $n \to \infty$:

$$H_n^* \overset{n\to+\infty}{\longrightarrow} Y \sim \mathcal{N}(0,1), \text{ in legge i.e. } \mathbb{P}(H_n^* \leq t) \overset{n\to+\infty}{\longrightarrow} \Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx \tag{7.1}$$

dove il secondo limite è uniforme rispetto a t.

Di conseguenza possiamo ritenere accettabili le seguenti approssimazioni:

$$\begin{aligned} \overline{X}_n \approx Y \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \mathbb{P}(\overline{X}_n \leq t) \approx \Phi\left(\frac{\sqrt{n}(t-\mu)}{\sigma}\right) \\ H_n = X_1 + \cdots + X_n \approx Z \sim \mathcal{N}(n\mu, n\sigma^2), \quad \mathbb{P}(H_n \leq t) \approx \Phi\left(\frac{t-n\mu}{\sqrt{n}\sigma}\right). \end{aligned} \tag{7.2}$$

Nota. Le precedenti approssimazioni sono in realtà delle uguaglianze nel caso in cui la legge delle variabili $\{X_i\}_{i\geq 1}$ sia una normale $\mathcal{N}(\mu, \sigma^2)$ (per l'equazione (5.1)).

L'interpretazione statistica di questo teorema è la seguente: sia $X_1, \ldots, X_n$ un campione casuale di ampiezza n estratto da una popolazione di distribuzione qualsiasi, avente valore atteso μ e varianza σ^2. Allora, al crescere di n, la media campionaria standardizzata H_n^* tende a distribuirsi con legge normale standard.

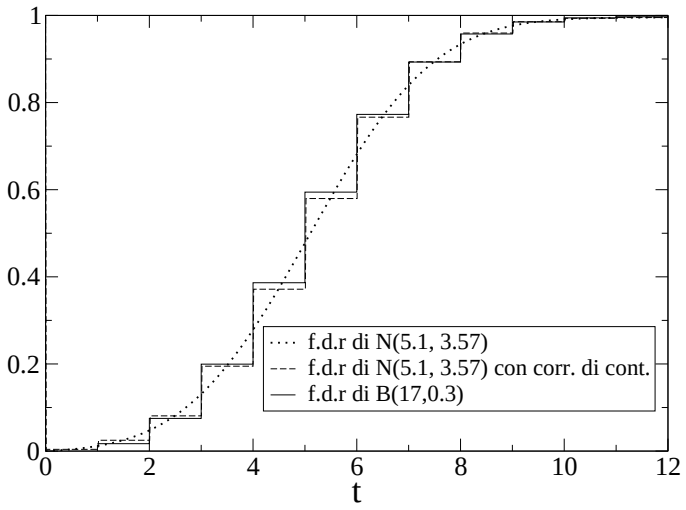


Grafico della f.d.r. della legge $B(17, 0.3)$, della sua approssimazione normale e della sua approssimazione normale con correzione di continuità.

Quanto debba essere grande n affinché l'approssimazione della media campionaria con la legge normale standard sia accettabile dipende dalla legge di partenza. Una regola empirica è di richiedere che $n \geq 30$. Questo valore va aumentato se la legge di partenza è fortemente asimmetrica, e può essere diminuito se essa è fortemente simmetrica: ad es. per la legge uniforme l'approssimazione è già buona per $n = 10$.

Correzione di continuità. Se la variabile aleatoria che vogliamo approssimare con la legge normale è discreta, conviene correggere la legge normale in modo da tenere conto del fatto che la funzione di probabilità è costante a tratti.

Precisamente, supponiamo che di avere una v.a. X che assume valori interi, come nel caso della binomiale e che per essa valga un'approssimazione del tipo $X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2)$ (ad esempio la variabile H_n dove X_i sono variabili che assumono valori interi). Allora essendo $\mathbb{P}(X \leq t)$ costante per $t \in [k, k+1)$ con k intero, (ricordiamo infatti che $\mathbb{P}(X \leq k) = \mathbb{P}(X < k+1)$) si pone il problema di quale valore utilizzare nell'approssimazione. Infatti si ha,

$$\begin{array}{ccc} \mathbb{P}(X \leq k) & = & \mathbb{P}(X < k+1) \\ \approx & & \approx \\ \Phi\left(\frac{\sqrt{n}(k-\mu)}{\sigma}\right) & \neq & \Phi\left(\frac{\sqrt{n}(k+1-\mu)}{\sigma}\right). \end{array}$$

Convieniamo allora di usare l'approssimazione

$$\mathbb{P}(X \leq k) = \mathbb{P}(X < k + 0.5) \simeq \Phi\left(\frac{\sqrt{n}(k + 0.5 - \mu)}{\sigma}\right);$$

questa procedura prende il nome di **correzione di continuità**.

Approfondimento

Definizione 7.1.2. Date $f, f_1, f_2, \ldots : X \to Y$ dove (Y, d) è uno spazio metrico (si può pensare a $(\mathbb{R}^n, \|\cdot\|_n)$), si dice che la convergenza $f_n \to f$ è uniforme rispetto a x se e solo se per ogni $\epsilon > 0$ esiste $n = n_\epsilon$ tale che per ogni $m \geq n$ e per ogni $x \in X$ si ha che $\epsilon \geq d(f(x), f_n(x))$.

Si può dimostrare che $f_n \to f$ uniformemente rispetto ad x se e solo se per ogni successione $\{x_n\}$ si ha

$$\lim_{n \to +\infty} d(f_n(x_n), f(x_n)) = 0.$$

In particolare se $F_{\widetilde{X}_n}(t) \to F_X(t)$ per ogni t ed F_X è continua allora la convergenza è uniforme rispetto a t. Questo è il caso descritto dal Teorema Centrale del Limite.

7.1.2 Approssimazioni gaussiane

Il Teorema Centrale del Limite viene spesso utilizzato per approssimare con una legge normale una legge più difficile da trattare o addirittura incognita; non affrontiamo qui il problema dell'accuratezza dell'approssimazione. Tali approssimazioni prendono il nome di **approssimazioni gaussiane** o **normali** e si basano sulle equazioni (7.2).

Osservazione 7.1.3. Supponiamo che $X_1, \dots, X_n$ sia una successione di variabili i.i.d. provenienti tutte da una legge che ammette media μ e varianza σ, allora, detto $\overline{X}_n := H_n/n$ dove $H_n := \sum_{i=1}^n X_i$, si ha

$$\begin{aligned} \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \equiv \frac{H_n - n\mu}{\sigma\sqrt{n}} &\approx Y \sim \mathcal{N}(0,1) \\ H_n &\approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \\ \overline{X}_n &\approx Y \sim \mathcal{N}(\mu, \sigma^2/n) \end{aligned}$$

dove, come detto in precedenza, le approssimazioni sono in realtà uguaglianze nel caso in cui la legge di partenza sia una legge normale $\mathcal{N}(\mu, \sigma^2)$ in virtù dell'equazione (5.1).

Le approssimazioni più utilizzate sono le seguenti.

- *Approssimazione normale di una Binomiale.* Siano $X_i \sim B(p)$ n v.a. Bernoulliane i.i.d., e sia $H_n = X_1 + \cdots + X_n \sim B(n,p)$. La media di una variabile di Bernoulli è $\mu = p$ e la sua varianza $\sigma^2 = p(1-p)$.

  Dal Teorema Centrale del Limite possiamo affermare che, per n grande,

  $$\begin{aligned} H_n &\approx Y \sim \mathcal{N}(n\mu, n\sigma^2) = \mathcal{N}(np, np(1-p)), \\ \overline{X}_n &\approx Z \sim \mathcal{N}(\mu, \sigma^2/n) = \mathcal{N}(p, p(1-p)/n), \end{aligned}$$

  $$\mathbb{P}(H_n \le t) \approx \Phi\left(\frac{t-np}{\sqrt{np(1-p)}}\right) \qquad \text{senza correzione di continuità}$$

  $$.\mathbb{P}(H_n \le k) \approx \Phi\left(\frac{k+0.5-np}{\sqrt{np(1-p)}}\right) \qquad \text{con correzione di continuità, } k \in \mathbb{N}$$

  Una buona norma è quella di applicare l'approssimazione normale della binomiale quando sono verificate le condizioni $np > 5$, $n(1-p) > 5$. Ricordiamo che se n è grande e p è piccolo, la binomiale può essere approssimata dalla legge di Poisson $P(np)$ (si veda il Paragrafo 4.4.3).

- *Approssimazione normale di una Poisson.* Poichè se $X_1, \dots, X_n$ i.i.d. distribuite come $P(1)$ implica $H_n = \sum_{i=1}^n X_i \sim P(n)$ si mostra in generale che una buona approssimazione per una variabile di Poisson $X \sim P(\lambda)$ con $\lambda \ge 5$ (λ non necessariamente intero) è rappresentata da $Y \sim \mathcal{N}(\lambda, \lambda)$. Quidni

  $$\mathbb{P}(X \le t) \approx \Phi\left(\frac{t-\lambda}{\sqrt{\lambda}}\right) \qquad \text{senza correzione di continuità}$$

  $$.\mathbb{P}(X \le k) \approx \Phi\left(\frac{k+0.5-\lambda}{\sqrt{\lambda}}\right) \qquad \text{con correzione di continuità, } k \in \mathbb{N}.$$

- *Approssimazione normale di una variabile di legge* Γ. Sappiamo che se $X_1, \dots, X_n$ sono i.i.d. di legge $\Gamma(1,\nu)$ allora $H_n : \sum_{i=1}^n X_i \sim \Gamma(n,\nu)$ si mostra in generale che una buona approssimazione per una variabile $X \sim \Gamma(\alpha,\nu)$ è $Y \sim \mathcal{N}(\alpha/\mu, \alpha/\mu^2)$, da cui

  $$\mathbb{P}(X \le t) \approx \Phi\left(\frac{t-\alpha/\mu}{\sqrt{\alpha}/\mu}\right).$$

Esercizio 7.1.4. Si lanci una moneta non truccata per 400 volte e si calcoli la probabilità che il numero di teste che escono sia compreso tra 180 e 210 (compresi gli estremi).

Soluzione.
Poichè i lanci sono indipendenti, la variabile N che conta il numero di teste uscite ha legge $B(400, 1/2)$ da cui, utilizzando l'approssimazione normale suggerita dal Teorema Centrale del Limite, si ha $N \approx Y \sim \mathcal{N}(200, 100)$. Pertanto

$$\begin{aligned}\mathbb{P}(180 \leq N \leq 210) &= \mathbb{P}(N \leq 210) - \mathbb{P}(N < 180) = \mathbb{P}(N \leq 210.5) - \mathbb{P}(N \leq 179.5) \\ &\approx \Phi\left(\frac{210.5 - 200}{10}\right) - \Phi\left(\frac{179.5 - 200}{10}\right) = \Phi(1.05) - \Phi(-2.05) \\ &= \Phi(1.05) + \Phi(2.05) - 1 \approx 0.8531 + 0.9798 - 1 = 0.8329.\end{aligned}$$

7.2 Legge dei Grandi Numeri

Date n v.a. X_i $(i = 1, \ldots, n)$, richiamiamo la definizione di v.a. **media campionaria** $\overline{X}_n$:

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

introdotta nei precedenti paragrafi.

Il valore atteso della v.a. media campionaria è

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)$$

Siano ora $X_1, X_2, \ldots$ v.a. indipendenti identicamente distribuite con media finita μ e varianza finita σ^2. Il valore atteso della v.a. media campionaria è

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$$

Mentre la varianza vale:

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

Ossia la varianza diminuisce al crescere del campione di v.a..

In definitiva, la media campionaria, a priori, non è detto che coincida con il valore atteso μ della legge del campione (cioè la legge comune a tutte le variabili $\{X_i\}_{i\geq 1}$), tuttavia il valore atteso della media campionaria è il valore atteso della legge del campione. In più la varianza tende a 0 in maniera monotona se n tende all'infinito, questo induce a pensare che la probabilità che $\overline{X}_n$ sia "vicino" a μ aumenti all'aumentare di n.

7.2.1 Disuguaglianza di Chebychev

Proposizione 7.2.1. *Sia X una v.a. con valore atteso μ e varianza σ^2. Sia δ un numero reale positivo prefissato. Vale la seguente disuguaglianza:*

$$\mathbb{P}(|X - \mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}$$

Dimostrazione. Diamo la dimostrazione per una v.a. discreta X con legge determinata, al solito, dai pesi $\{p_i\}_i$, ma notiamo che la disuguaglianza vale per tutte le variabili aleatorie (e la dimostrazione è sostanzialmente identica nel caso delle variabili continue).

Sia A l'intervallo dei valori di X compresi tra $\mu-\delta$ e $\mu+\delta$, e sia $A^{\complement}$, al solito, il complementare di A rispetto ad $\mathbb{R}$:

$$A = \{x : |x-\mu| < \delta\}, \qquad A^{\complement} = \{x : |x-\mu| \geq \delta\}$$

Scriviamo la varianza come:

$$\sigma^2 = \sum_i (x_i-\mu)^2 p_X(x_i) = \sum_{i:x_i\in A} (x_i-\mu)^2 p_X(x_i) + \sum_{j:x_j\in A^{\complement}} (x_j-\mu)^2 p_X(x_j)$$

Considerato che $\sum_{i:x_i\in A}(x_i-\mu)^2 p_X(x_i)$ è una quantità certamente non negativa, e che per ogni $x \in A^{\complement}$ vale $(x-\mu)^2 \geq \delta^2$, alla fine si ha:

$$\sigma^2 \geq \sum_{j:x_j\in A^{\complement}} (x_j-\mu)^2 p_X(x_j) \geq \delta^2 \sum_{j:x_j\in A^{\complement}} p_X(x_j).$$

Ovviamente $\sum_{j:x_j\in A^{\complement}} p_X(x_j) = \mathbb{P}(|X-\mu| \geq \delta)$, pertanto, dalla precedente disuguaglianza, discende direttamente la disuguaglianza di Chebychev. □

La disuguaglianza di Chebychev può essere scritta equivalentemente nelle seguenti utili forme alternative:

$$\mathbb{P}(|X-\mu| \geq \delta\sigma) \leq \frac{1}{\delta^2}$$

$$\mathbb{P}(|X-\mu| < \delta\sigma) \geq 1 - \frac{1}{\delta^2}$$

7.2.2 Legge debole dei grandi numeri

Teorema 7.2.2. *Siano $X_1, X_2, \ldots$ v.a. indipendenti identicamente distribuite con media finita μ. Allora per ogni $\epsilon > 0$ vale*

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \to 0 \quad \textit{per} \quad n \to \infty$$

Questo fatto si esprime dicendo che la successione di v.a. $\{\overline{X}_n\}$ tende in probabilità a μ per $n \to \infty$.

Dimostrazione. Dimostriamo la legge nell'ipotesi che esista finita la varianza $\mathrm{Var}(X_i) = \sigma^2$ (la legge però è vera anche se non esiste la varianza).
Applichiamo la disuguaglianza di Chebychev alla v.a. $\overline{X}_n$:

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq \delta\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{\delta^2} \quad \text{per ogni } \delta > 0$$

Scegliamo $\delta = \epsilon\sqrt{n}/\sigma$:

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2 n} \quad \text{per ogni } \epsilon > 0$$

Per $n \to \infty$ si ottiene la tesi. □

Osservazione 7.2.3. Nel caso esista finita la varianza σ, la legge debole dei grandi numeri può anche essere provata a partire dal Teorema Centrale del Limite. Se $\sigma = 0$ è banale. Se $\sigma \neq 0$, essendo la convergenza nell'equazione (7.1) uniforme, si ha, equivalentemente che, per ogni successione $\{t_n\}$,

$$\lim_{n\to+\infty} |\mathbb{P}(H_n^*(t_n) - \Phi(t_n)| = 0.$$

Pertanto,

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = \lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu|\sqrt{n}/\sigma > \epsilon\sqrt{n}/\sigma) = \lim_{n\to\infty} \left(2\phi(\epsilon\sqrt{n}/\sigma) - 1\right) = 1.$$

7.2.3 Legge forte dei grandi numeri

Teorema 7.2.4. *Siano $X_1, X_2, \dots$ v.a. indipendenti identicamente distribuite con media finita μ. Allora*

$$\mathbb{P}\left(\{\omega : \lim_{n\to\infty} \overline{X}_n = \mu\}\right) = 1.$$

La dimostrazione è più complessa che nel caso della legge debole e pertanto la omettiamo. Si dimostra inoltre che la legge forte dei grandi numeri implica la legge debole.

Il significato del precedente teorema è che con probabilità 1 la media campionaria di n estrazioni indipendenti di valori dalla stessa legge tende, per n che tende all'infinito al valore atteso della legge. Questo oltre a "giustificare" in qualche modo l'approccio frequentista, suggerisce, unitamente alle proprietà del valore atteso e della varianza della v.a. $\overline{X}_n$ viste in precedenza, di stimare il valore μ con il valore *a posteriori* $\overline{X}_n(\omega) =: \overline{x}_n$; tutto questo sarà oggetto dello studio del prossimo capitolo.

Applichiamo il precedente teorema al seguente caso: sia $\{X_i\}_i$ una successione i.i.d. di variabili aleatorie e Q la loro comune funzione dei quantili. Sia $\alpha \in (0,1)$ e sia $N_n := \#\{i = 1, 2, \cdots, n : X_i \leq Q(\alpha)\}$. La legge dei grandi numeri ci dice che, con probabilità 1, si ha $\lim_{n\to\infty} N_n/n = \alpha$, quindi Q_α diviene un buon modello a priori del quantile di ordine α (calcolato a posteriori sul campione).

Cap. 8. Statistica inferenziale: stime

La distribuzione di probabilità di una v.a., dipendente da uno o più parametri θ, permette di assegnare una probabilità a qualsiasi campione. Scopo della statistica inferenziale è di procedere all'inverso, ossia a partire dai dati di un campione di una popolazione, si vuole determinare il parametro incognito θ.

Esempio 8.0.5. Un sondaggio eseguito su un campione di n votanti mette in luce che una frazione p di essi ha votato per un certo partito. Il modello che potremmo utilizzare è il seguente: la persona i-esima sarà rappresentata da una variabile di Bernoulli di parametro incognito q e supporremo le variabili indipendenti. La legge è determinata univocamente dal parametro $q \in [0,1]$. Eseguito il sondaggio (quindi *a posteriori*) qual è la miglior stima che possiamo dare per q? Quanto affidabile è la nostra stima?
La prima domanda è un problema di *stima puntuale*, mentre la seconda è un problema di *stima per intervalli*.

Esempio 8.0.6. Una macchina produce componenti meccanici di dimensioni specificate con un livello di tolleranza dato. Al di fuori dei limiti di tolleranza il pezzo viene giudicato difettoso. Il produttore vuole garantire che la percentuale dei pezzi difettosi non superi il 5%. Il modello di produzione è ben rappresentato mediante un processo di Bernoulli di v.a. X_i che vale 1 se l'i-esimo pezzo è difettoso e 0 altrimenti: $X_i \sim B(p)$. Il parametro incognito è p, e si vuole stimarlo sulla base di osservazioni a campione.

Nei prossimi paragrafi utilizzeremo spesso la notazione vettoriale (considereremo quindi stimatori per parametri vettoriali reali); pertanto, ad esempio, considereremo valori attesi di v.a. vettoriali: quest'ultime ammettono valore atteso se e solo se ogni componente ammette valore atteso ed il vettore dei valori attesi ha come componenti i valori attesi di ciascuna componente. Utilizzeremo la notazione $\|\cdot\|$ e $\langle\cdot,\cdot\rangle$ rispettivamente per indicare la norma ed il prodotto scalare.

8.1 Modello statistico parametrico

Definizione 8.1.1. Un **modello statistico parametrico** è una famiglia di leggi di v.a., dipendenti da uno o più parametri θ: $\{\mathcal{L}_\theta\}_{\theta\in\Theta}$. La legge è nota a meno dei parametri $\theta \in \Theta$ dove Θ è un opportuno insieme (in genere $\Theta \subseteq \mathbb{R}$ oppure $\Theta \subseteq \mathbb{R}^n$).

Definizione 8.1.2. Un **campione casuale** di dimensione n estratto da una popolazione di legge $\mathcal{L}_\theta$ è una n-upla di v.a. $X_1, \dots, X_n$ i.i.d. ciascuna con legge $\mathcal{L}_\theta$.

Osservazione 8.1.3. Si osservi che lo spazio di probabilità su cui sono definite $X_1, \dots, X_n$, fissato il modello statistico parametrico, in generale ha una misura di probabilità $\mathbb{P}_\theta$ dipendente da $\theta \in \Theta$.

Definizione 8.1.4. Consideriamo una v.a. X avente densità di probabilità $\mathcal{L}_\theta$, dove $\theta \in \Theta \subseteq \mathbb{R}^m$. Si dice **statistica** una v.a. $T = t_n(X_1, X_2, \dots, X_n)$ funzione del campione casuale $(X_1, X_2, \dots, X_n)$. La funzione $t_n : \mathbb{R}^n \to \mathbb{R}^m$ deve essere $\mathcal{R}^n - \mathcal{R}^m$–misurabile.
N.B.: T **NON** deve dipendere in modo esplicito da θ.

Ad esempio, la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ di n v.a. è una statistica.

Definizione 8.1.5. Si dice **stimatore** del parametro θ una statistica usata per stimare θ o, più in generale, una sua funzione $g(\theta)$. Assegnata la statistica $T = t_n(X_1, X_2, \ldots, X_n)$, una volta estratto un particolare campione $(x_1, x_2, \ldots, x_n)$, il valore $\tau = t_n(x_1, x_2, \ldots, x_n)$ si dice **stima** di $g(\theta)$.

Osservazione 8.1.6. lo stimatore è una variabile aleatoria, mentre la stima è un numero reale. Ricordiamo che il nostro modello di campionamento, a priori, è rappresentato da una successione finita $X_1, \ldots, X_n$ di variabili aleatorie i.i.d. ciascuna delle quali rappresenta il risultato dell'i-esimo elemento del campione; se lo stato è $\omega \in \Omega$ allora in corrispondenza ad esso (cioè a posteriori) osserveremo il campione $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$. Allo stesso modo se osserviamo il campione $x_1, \ldots, x_n$ sappiamo che si è verificato l'evento $\{X_i = x_i,\ \forall i = 1, \ldots, n\}$. Se decidiamo quindi di calcolare a posteriori $t(x_1, \ldots, x_n)$ per stimare $g(\theta)$, il modello di questa stima, a priori, è la variabile aleatoria $t(X_1, \ldots, X_n)$.

Nota. In generale per stimare $g(\theta)$ si utilizzerà una procedura valida per ogni ampiezza n del campione, questo significa che si determinerà una successione $\{T_n\}_{n\in\mathbb{N}^*}$ dove ciascuna statistica T_n si costruisce da una funzione misurabile t_n definita su $\mathbb{R}^n$.

Proprietà degli stimatori.

- *Correttezza.* Uno stimatore T di $g(\theta)$ si dice **corretto**, o **non distorto**, se $\mathbb{E}_\theta(T) = g(\theta)$ per ogni θ (ricordiamo che la legge, e quindi anche il valor medio, di T dipende da θ; nel seguito talvolta, per semplicità di notazione, sottintenderemo i pedici θ).

  Uno stimatore non corretto si dice **distorto** e la quantità $\mathbb{E}_\theta(T) - g(\theta)$ si dice **distorsione** dello stimatore.

- *Correttezza Asintotica.* Una famiglia di stimatori $T_n = t_n(X_1, X_2, \ldots, X_n)$, $n = 1, 2, \ldots$, si dice **asintoticamente corretta** se la distorsione si annulla al crescere dell'ampiezza del campione:
$$\lim_{n\to\infty} \mathbb{E}_\theta(T_n) - g(\theta) = 0, \quad \forall\theta \in \Theta.$$

- *Consistenza.* Una famiglia di stimatori si dice **semplicemente consistente** o **debolmente consistente** se
$$\lim_{n\to\infty} \mathbb{P}_\theta(|T_n - g(\theta)| \le \epsilon) = 1 \quad \forall\epsilon > 0, \quad \forall\theta \in \Theta.$$

  Si può mostrare che vi sono famiglie di stimatori corretti non consistenti e, viceversa, famiglie di stimatori consistenti che non sono nemmeno asintoticamente corretti.

  Esempio 8.1.7. Siano $\{X_n\}$ i.i.d. con media μ e tali che $\mathbb{P}(X_i = \mu) < 1$. Si consideri la funzione reale $t_n(x_1, \ldots, x_n) := x_1$; allora la famiglia di stimatori corrispondente $\{T_n\}$, dove $T_n = X_1$, è una famiglia di stimatori corretti ma non è consistente.
  L'esempio di una famiglia di stimatori non asintoticamente corretta ma consistente è più complesso e non lo vedremo.

- **Errore quadratico medio (EQM)** di uno stimatore T di $g(\theta)$ è definito come
$$\mathrm{EQM}(T) \equiv \mathrm{EQM}_\theta(T) := \mathbb{E}_\theta(\|T - g(\theta)\|^2).$$

  Si osservi che,
$$\mathbb{E}_\theta(\|T - g(\theta)\|^2) = \mathbb{E}_\theta(\|T\|^2) - 2\langle\mathbb{E}_\theta(T), g(\theta)\rangle + \|g(\theta\|^2) = \mathrm{Var}_\theta(T) + \|\mathbb{E}_\theta(T) - g(\theta)\|^2.$$

  Una famiglia di stimatori $\{T_n\}$ di $g(\theta)$ si dice **consistente in media quadratica** se e solo se
$$\lim_{n\to+\infty} \mathrm{EQM}_\theta(T_n) = 0, \qquad \forall\theta \in \Theta.$$

Teorema 8.1.8. *Per una famiglia di stimatori $\{T_n\}$ di $g(\theta)$ le seguenti affermazioni sono equivalenti.*

(i) la famiglia è consistente in media quadratica;

(ii) la famiglia è semplicemente consistente e $\lim_{n\to+\infty} \mathrm{Var}_\theta(T_n) = 0$*;*

(iii) la famiglia è asintoticamente corretta e $\lim_{n\to+\infty} \mathrm{Var}_\theta(T_n) = 0$.

Dimostrazione.

Approfondimento

(i) $\iff$ (iii). Dalla definizione di EQM si ha che

$$\lim_{n\to+\infty} \mathrm{EQM}_\theta(T_n) = 0 \iff \begin{cases} \lim_{n\to+\infty} \mathrm{Var}_\theta(T_n) = 0 \\ \lim_{n\to+\infty} (\mathbb{E}_\theta(T_n) - g(\theta)) = 0. \end{cases}$$

(ii) $\implies$ (iii). Si osservi che da $\mathrm{Var}_\theta(T_n) \to 0$ se $n \to +\infty$, utilizzando la disuguaglianza di Chebychev, per ogni $\theta \in \Theta$,

$$\mathbb{P}_\theta(\|T_n - \mathbb{E}_\theta(T_n)\| > \epsilon) \le \frac{\mathrm{Var}_\theta(T_n)}{\epsilon^2} \to 0, \qquad n \to +\infty,$$

pertanto

$$\lim_{n\to+\infty} \mathbb{P}_\theta(\|T_n - \mathbb{E}_\theta(T_n)\| \le \epsilon) = 1, \qquad \forall \theta \in \Theta.$$

Utilizzando la disuguaglianza triangolare

$$\|g(\theta) - \mathbb{E}_\theta(T_n)\| \le \|g(\theta) - T_n\| + \|T_n - \mathbb{E}_\theta(T_n)\| \qquad (\text{per ogni } \omega \in \Omega)$$

se, per assurdo, $\|g(\theta) - \mathbb{E}_\theta(T_n)\| > \epsilon > 0$ per infiniti valori di n allora $\{\|g(\theta) - T_n\| \le \epsilon/2\}$ e $\{\|T_n - \mathbb{E}_\theta(T_n)\| \le \epsilon/2\}$ sono disgiunti pertanto

$$1 \ge \mathbb{P}_\theta(\|T_n - \mathbb{E}_\theta(T_n)\| \le \epsilon/2) + \mathbb{P}_\theta(\|g(\theta) - T_n\| \le \epsilon/2)$$

e quindi $\mathbb{P}_\theta(\|g(\theta) - T_n\| \le \epsilon/2) \not\to 1$ se $n \to +\infty$ da cui l'assurdo.
(iii) $\implies$ (ii). Ancora dalla disuguaglianza triangolare

$$\|g(\theta) - T_n\| \le \|T_n - \mathbb{E}_\theta(T_n)\| + \|g(\theta) - \mathbb{E}_\theta(T_n)\| \qquad (\text{per ogni } \omega \in \Omega)$$

e da (iii) esiste n_0 tale che per ogni $n \ge n_0$ si ha $\|g(\theta) - \mathbb{E}_\theta(T_n)\| < \epsilon/2$ quindi, per ogni $n \ge n_0$,

$$\begin{aligned} \mathbb{P}(\|g(\theta) - T_n\| \le \epsilon) &\ge \mathbb{P}(\|T_n - \mathbb{E}_\theta(T_n)\| + \|g(\theta) - \mathbb{E}_\theta(T_n)\| \le \epsilon) \\ &\ge \mathbb{P}(\|T_n - \mathbb{E}_\theta(T_n)\| + \epsilon/2 \le \epsilon) \\ &= \mathbb{P}(\|T_n - \mathbb{E}_\theta(T_n)\| \le \epsilon/2) \to 1 \qquad \text{se } n \to +\infty, \end{aligned}$$

da cui si ha la semplice consistenza.

□

It teorema precedente garantisce che, sotto l'ipotesi $\mathrm{Var}(T_n) \to 0$ per $n \to +\infty$ allora le semplice consistenza e l'asintotica correttezza sono equivalenti.

Osservazione 8.1.9. Poichè in generale $\mathbb{E}(h \circ T) \neq h(\mathbb{E}(T))$, si ha che se T è uno stimatore non distorto di $g(\theta)$ allora $h \circ T$ è uno stimatore, in generale distorto, di $h(g(\theta))$ (la non distorsione si ha nel caso $h(x) := ax + b$).
Un esempio è la ricerca di uno stimatore non distorto per il parametro λ di una distribuzione

esponenziale. Sappiamo che il valore atteso di una variabile $X \sim \text{Esp}(\lambda)$ è $\mathbb{E}(X) = 1/\lambda$ pertanto saremmo tentati di stimare λ utilizzando la famiglia di stimatori $1/\overline{X}_n = n/\sum_{i=1}^n X_i$. Questo è uno stimatore soltanto asintoticamente corretto, mentre calcoli semplici (si utilizzi la densità gamma introdotta nel Paragrafo 5.3.4) mostrano che uno stimatore corretto è

$$T_n := \frac{n-1}{\sum_{i=1}^n X_i} \equiv \frac{n-1}{n\overline{X}_n}.$$

8.2 Stima puntuale

Scopo della *stima puntuale* è di utilizzare opportune statistiche per stimare i valori dei parametri incogniti della distribuzione di partenza. Vedremo nei prossimi paragrafi esempi di statistiche per stimare il valore atteso e la varianza di una distribuzione. Nella sezione successiva ci proporremo di fornire degli intervalli ai quali riteniamo plausibile che tali parametri appartengano. Questa parte della statistica inferenziale viene chiamata *stima per intervalli*. Noi cercheremo, se possibile, famiglie consistenti di stimatori corretti.

8.2.1 Stima puntuale della media

La media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ di n v.a. i.i.d. è una statistica; si può affermare che è uno stimatore corretto e debolmente consistente (in virtù della legge forte dei grandi numeri) del valore atteso comune $g(\theta) := \mathbb{E}_\theta(X_i)$. Nel caso in cui $\text{Var}(X_i) = \sigma^2 < +\infty$, allora il teorema precedente garantisce la consistenza in media quadratica poiché $\text{Var}(\overline{X}_n) = \sigma^2/n \to 0$ se $n \to +\infty$.

Esempi. Nell'esempio visto in precedenza della produzione di componenti meccanici, il modello statistico parametrico è la famiglia di leggi di Bernoulli $B(p)$, p è il parametro da determinare; La media campionaria $T_n = \overline{X}_n$ è uno stimatore non distorto e consistente di p.

Per una v.a. normale $\mathcal{N}(\mu, \sigma^2)$ i parametri sono $\theta = (\mu, \sigma^2)$. La media campionaria è uno stimatore non distorto e consistente di $g(\theta) := \mu$.

Lo stesso discorso si applica alle altre v.a. notevoli:
Per la binomiale $X \sim B(n, p)$, $\theta = (n, p)$, la media campionaria è uno stimatore di $\mathbb{E}_\theta(X) = np =: g(\theta)$.
Per l'esponenziale $X \sim Exp(\lambda)$, $\theta = \lambda$, la media campionaria è uno stimatore di $\mathbb{E}_\theta(X) = 1/\lambda$.
Per il modello Gamma $X \sim \Gamma(n, \lambda)$, $\theta = (n, \lambda)$, la media campionaria è uno stimatore di $\mathbb{E}_\theta(X) = n/\lambda$.

Date n v.a. i.i.d. X_i, e scelta una serie di sequenze di numeri reali $\{\lambda_i^{(n)}\}_{i=1}^n$ $(n = 1, 2, \ldots)$ la v.a. $T_n^\lambda = \sum_{i=1}^n \lambda_i^{(n)} X_i$ con $\sum_{i=1}^n \lambda_i^{(n)} = 1$ è uno stimatore corretto del valore atteso $\theta = \mathbb{E}_\theta(X_i)$. È facile altresì vedere che se $\sum_{i=1}^n \lambda_i^{(n)} = 1$, e la legge di X_i ammette varianza σ^2 allora $\text{EQM}(T_n^\lambda) = \text{Var}(T_n^\lambda) = \sum_{i=1}^n (\lambda_i^{(n)})^2$; (il cui valore minimo si ha per $\lambda_1 = \cdots = \lambda_n = 1/n$) ed è anche debolmente consistente se e solo se $\sum_{i=1}^n (\lambda_i^{(n)})^2 \to 0$ per $n \to \infty$.

8.2.2 Stima puntuale della varianza

Siano n variabili aleatorie i.i.d. X_i aventi valore atteso comune $\mu = \mathbb{E}_\theta(X_i)$ e varianza $\sigma^2 = \text{Var}_\theta(X_i)$.

Consideriamo la statistica, dipendente da una successione di numeri reali positivi $\{a_n\}_{n=1}^{+\infty}$

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

Calcoliamo il valor medio di $\hat{S}_n^2$:

$$\mathbb{E}(\hat{S}_n^2) = \frac{1}{a_n}\sum_{i=1}^n \mathbb{E}[(X_i - \overline{X}_n)^2]$$

Per completare il calcolo riscriviamo $\hat{S}_n^2$ come

$$\begin{aligned}\hat{S}_n^2 &= \frac{1}{a_n}\sum_{i=1}^{n}[(X_i-\mu)-(\overline{X}_n-\mu)]^2\\ &= \frac{1}{a_n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{n}{a_n}(\overline{X}_n-\mu)^2-2(\overline{X}_n-\mu)\frac{1}{a_n}\sum_{i=1}^{n}(X_i-\mu)\\ &= \frac{1}{a_n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{n}{a_n}(\overline{X}_n-\mu)^2\end{aligned}$$

da cui

$$\begin{aligned}\mathbb{E}(\hat{S}_n^2) &= \mathbb{E}\left(\frac{1}{a_n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{n}{a_n}(\overline{X}_n-\mu)^2\right)\\ &= \frac{1}{a_n}\left(\sum_{i=1}^{n}\mathbb{E}[(X_i-\mu)^2]-n\mathbb{E}[(\overline{X}_n-\mu)^2]\right)\\ &= \frac{1}{a_n}n\sigma^2-\frac{1}{a_n}\sigma^2=\frac{n-1}{a_n}\sigma^2.\end{aligned}$$

La statistica $\hat{S}_n^2$ è dunque uno stimatore asintoticamente corretto della varianza $\sigma^2=\mathrm{Var}_\theta(X_1)=:g(\theta)$ se e solo se $a_n/n\to 1$ se $n\to+\infty$; la distorsione è $\mathbb{E}(\hat{S}_n^2)-\sigma^2=\sigma^2((n-1)/a_n-1)$ e quindi $\hat{S}_n^2$ è uno stimatore corretto se e solo se $a_n=n-1$. Questo giustifica la scelta fatta nel capitolo sulla statistica descrittiva per stimare la varianza.

In maniera euristica si può spiegare questo fatto notando che la presenza dell'addendo $\overline{X}_n$ al posto della media vera μ introduce un errore nel calcolo messo in mostra dall'equazione

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

Definiamo quindi uno stimatore corretto di σ^2 come:

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

per questo nuovo stimatore vale quindi $\mathbb{E}(S_n^2)=\sigma^2$. La v.a. S_n^2 prende il nome di **varianza campionaria**. Si può dimostrare con calcoli facili ma noiosi che S_n^2 è uno stimatore consistente in media quadratica:

$$\lim_{n\to\infty}\mathrm{Var}(S_n^2)=0$$

purché $\mathbb{E}((X_i-\mu)^4)<\infty$ o, equivalentemente, $\mathbb{E}(X_i^4)<+\infty$

Approfondimento

Si mostri prima per esercizio che

$$S_n^2=\frac{1}{2n(n-1)}\sum_{j=1}^{n}\sum_{i=1}^{n}(X_i-X_j)^2=\frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{i>j}(X_i-X_j)^2.$$

Ovviamente, senza perdita di generalità si può supporre che $\mathbb{E}(X_i)=0$ per ogni i, altrimenti si utilizzino $\widetilde{X}_i:=X_i-\mathbb{E}(X_i)$ osservando che $\mathrm{Var}(\widetilde{X}_i)=\sigma^2$ e che lo stimatore della varianza campionaria $\widetilde{S}_n^2$ relativo alle nuove variabili coincide con S_n^2.

Definiamo $\mu^{(4)} := \mathbb{E}(X^4)$; con facili calcoli si ottiene

$$\begin{aligned}
\mathrm{Var}(S_n^2) &= \frac{1}{4n^2(n-1)^2} \sum_{i,j,h,k=1}^{n} \mathbb{E}((X_i - X_j)^2(X_h - X_k)^2) - \sigma^4 \\
&= \frac{1}{4n^2(n-1)^2} \sum_{i,j,h,k=1}^{n} \mathbb{E}\Big(X_i^2 X_h^2 - 2X_i^2 X_h X_k + X_i^2 X_k^2 - 2X_i X_j X_h^2 \\
&+ 4X_i X_j X_h X_k - 2X_i X_j X_k^2 + X_j^2 X_h^2 - 2X_j^2 X_h X_k + X_j^2 X_k^2 \Big) - \sigma^4 \\
&= \frac{1}{4n^2(n-1)^2} \Big(4n^2 \sum_{i,h=1}^{n} \mathbb{E}(X_i^2 X_h^2) - 8n \sum_{i,j,k=1}^{n} \mathbb{E}(X_i X_j X_k^2) \\
&+ 4 \sum_{i,j,h,k=1}^{n} \mathbb{E}(X_i X_j X_h X_k) \Big) - \sigma^4
\end{aligned}$$

da cui, ricordando che $\mathbb{E}(X_i) = 0$ per ogni i ed utilizzando l'indipendenza, si ha

$$\begin{aligned}
\mathrm{Var}(S_n^2) &= \frac{1}{4n^2(n-1)^2} \Big((4n^2 - 8n + 12)n(n-1)\sigma^4 + (4n^2 - 8n + 4)n\mu^{(4)} \Big) - \sigma^4 \\
&= \left(\frac{n^2 - 2n + 3}{n(n-1)} - 1 \right) \sigma^4 + \frac{\mu^{(4)}}{n} \\
&= \frac{3 - n}{n(n-1)} \sigma^4 + \frac{\mu^{(4)}}{n} \to 0 \quad \text{se } n \to +\infty.
\end{aligned}$$

In pratica, una volta estratto un particolare campione $(x_1, x_2, \ldots, x_n)$, si ottiene il valore corrispondente di s_n^2:

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

ossia s_n^2 è la varianza campionaria dei dati $x_1, \ldots, x_n$.

Osservazione 8.2.1. Se è noto il valore atteso $\mathbb{E}(X_i) = \mu$ della v.a. X_i, allora per stimare la varianza si può usare la statistica seguente:

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

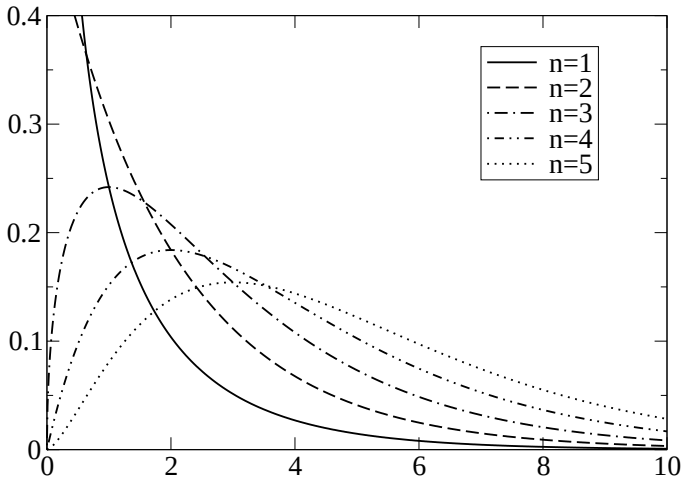
Attenzione: T_n è una statistica solo se il valore atteso $\mathbb{E}(X_i) = \mu$ è noto; altrimenti μ è un parametro incognito e T_n non è più una statistica.

Dimostriamo che T_n è uno stimatore corretto di σ^2:

$$\mathbb{E}(T_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(X_i - \mu)^2] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i^2 - 2\mu X_i + \mu^2) =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \big(\mathbb{E}(X_i^2) - 2\mu \mathbb{E}(X_i) + \mu^2 \big) = \frac{1}{n} \sum_{i=1}^{n} \big(\mathbb{E}(X_i^2) - \mu^2 \big) = \sigma^2$$

Esempio. Vogliamo stimare i parametri r e λ di una popolazione con distribuzione $\Gamma(r, \lambda)$. Effettuiamo un campionamento e consideriamo gli stimatori $\overline{X}_n$ e S_n^2. I loro valori attesi sono rispettivamente $\mathbb{E}(\overline{X}_n) = r/\lambda$ e $\mathbb{E}(S_n^2) = r/\lambda^2$. Dunque $\overline{X}_n$ e S_n^2 sono stimatori non distorti dei parametri r/λ e r/λ^2. Nella pratica, a campionamento effettuato otteniamo i valori $\overline{x}_n$ e s_n^2. Risolvendo per r e λ otteniamo le stime $\hat{\lambda} = \overline{x}_n / s_n^2$ e $\hat{r} = \overline{x}_n^2 / s_n^2$.
Si può anche dire che $\overline{X}_n / S_n^2$ e $\overline{X}_n^2 / S_n^2$ sono stimatori (distorti!) rispettivamente di λ ed r.



Funzione densità della v.a. chi-quadrato $\chi^2(n)$ per alcuni valori di n.

8.3 Stima per intervalli: leggi notevoli

Uno stimatore T, come ad esempio $\overline{X}_n$, fornisce, *a campionamento eseguito*, una stima del valore di θ del quale è però ignota l'accuratezza. Descriviamo questa proprietà degli stimatori dicendo che forniscono una stima puntuale del/dei parametro(i) incogniti. Se lo stimatore è asintoticamente corretto e consistente $\mathbb{E}(T)$ darà una stima sempre più accurata al crescere dell'ampiezza del campione (in virtù della disuguaglianza di Chebychev); tuttavia non sempre è possibile aumentare n. È necessario quindi un metodo per ottenere dal campione stesso anche una stima dell'accuratezza della stima puntuale. Questo metodo consiste nella costruzione di un intervallo, detto **intervallo di confidenza o intervallo fiduciario**, che verosimilmente contenga il valore vero del parametro incognito. In tale ottica, parliamo di stima per *intervalli* di θ.

Per stimare i parametri di una distribuzione normale, è utile definire alcune distribuzioni continue.

8.3.1 Legge chi-quadrato

Definizione 8.3.1. Si dice **legge chi-quadrato con n gradi di libertà**, la legge di una variabile aleatoria

$$Y = \sum_{i=1}^{n} X_i^2,$$

dove X_i sono n v.a. indipendenti, ciascuna di legge $\mathcal{N}(0,1)$. Si scrive $Y \sim \chi^2(n)$ ed è univocamente determinata.

Come vedremo, la legge chi-quadrato è utile per stimare la varianza di una popolazione normale.

Proprietà. Si dimostra che la legge $\chi^2(n)$ coincide con la legge gamma di parametri $n/2$, $1/2$: $\chi^2(n) = \Gamma(n/2, 1/2)$. Da questa proprietà si possono ricavare molte informazioni sulla legge chi-quadrato:

- La funzione densità è:

$$f_Y(t) = c_n t^{n/2-1} e^{-t/2} \quad \text{per } t > 0, \quad f_Y(t) = 0 \quad \text{per } t \leq 0$$

  Il valore di c_n viene ottenuto imponendo la relazione $\int_0^{+\infty} c_n t^{n/2-1} e^{-t/2} = 1$.

- il valore atteso: ricordando la proprietà che il valore atteso della $\Gamma(r, \nu)$ è pari a r/ν,

$$\mathbb{E}(Y) = \mathbb{E}\left(\sum_{i=1}^{n} X_i^2\right) = \frac{n/2}{1/2} = n$$

- la varianza: ricordando la proprietà che la varianza della $\Gamma(r,\nu)$ è pari a r/ν^2,vale:

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i^2\right) = \frac{n/2}{1/4} = 2n$$

- Se Y_1 e Y_2 sono v.a. indipendenti con leggi rispettive $Y_1 \sim \chi^2(n)$, $Y_2 \sim \chi^2(m)$, allora

$$Y_1 + Y_2 \sim \Gamma\left(\frac{n}{2} + \frac{m}{2}, \frac{1}{2}\right) = \chi^2(n+m)$$

- Dal Teorema Centrale del Limite se $Y_n \sim \chi^2(n)$ si ha che Y_n è approssimabile con una variabile di legge $\mathcal{N}(n, 2n)$ nel senso che

$$\lim_{n\to+\infty}\left(\mathbb{P}(Y \le t) - \Phi\left(\frac{t-n}{\sqrt{2n}}\right)\right) = 0.$$

- Indichiamo con $\chi^2_\alpha(n)$ i quantili della legge chi-quadrato:

$$\mathbb{P}(Y \le \chi^2_\alpha(n)) = \alpha$$

  I valori dei quantili sono tabulati per i primi valori di n. Per n grande ($n > 30$) si possono determinare i quantili da quelli della normale, sfruttando l'approssimazione normale:

$$\Phi\left(\frac{t-n}{\sqrt{2n}}\right) = \alpha, \quad q_\alpha \simeq \frac{\chi^2_\alpha(n) - n}{\sqrt{2n}}, \quad \chi^2_\alpha(n) \simeq q_\alpha\sqrt{2n} + n$$

  Un'approssimazione leggermente migliore di questa, valida sempre per $n > 30$, è

$$\chi^2_\alpha(n) \simeq \frac{1}{2}\left(q_\alpha + \sqrt{2n-1}\right)^2.$$

L'importanza della legge chi-quadrato è dovuta alle seguenti proprietà:
Siano $X_1, X_2, \ldots, X_n$, n v.a. normali i.i.d. di legge $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Allora

- La somma delle standardizzate al quadrato vale

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

  Questa proprietà discende direttamente dalla definizione della legge chi-quadrato come somma di quadrati di v.a. normali standard indipendenti.

- Se $\overline{X}_n$ è la media campionaria,

$$\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 \sim \chi^2(n-1)$$

  la media μ viene sostituita con la media campionaria $\overline{X}_n$, e la v.a. trovata ha legge chi-quadrato con un grado di libertà in meno.
  Non dimostreremo questa proprietà. Intuitivamente si può capire che le n v.a. $X_i - \overline{X}_n$ non sono più indipendenti, poiché la loro somma è nulla. Questa relazione sottrae un grado di libertà alla somma dei loro quadrati.

  In termini della varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$, la formula precedente si può riscrivere come

$$\frac{(n-1)}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

- Si dimostra infine che, se il campione proviene da una famiglia normale, la varianza campionaria S_n^2 e la media campionaria $\overline{X}_n$ sono v.a. tra loro indipendenti. Questa proprietà non è semplice da dimostrare e non vale in generale per una legge qualsiasi.

Esercizio 8.3.2. Una ditta produce bulloni del diametro medio di $2cm$. Dall'esperienza passata è noto che la deviazione standard del loro diametro è di $0.1cm$. Si può supporre inoltre che il diametro effettivo di un bullone abbia una distribuzione normale. Una seconda ditta intende comprare una partita di bulloni ma non crede ai parametri forniti dalla prima ditta sul valor medio e sulla varianza, e pone come requisito che la varianza campionaria di 20 pezzi scelti a caso non superi $(0.12cm)^2$. Qual è la probabilità che la partita venga scartata?

Soluzione.
Applichiamo la formula $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$ con $n = 20$, $\sigma = 0.1cm$. Poniamo $Y \sim \chi^2(19)$.

$$\mathbb{P}\left(S_n^2 > (0.12cm)^2\right) = \mathbb{P}\left(Y > \frac{19 \cdot 0.12^2}{0.1^2}\right) = \mathbb{P}(Y > 27.36) \simeq 0.1$$

Il valore di $\mathbb{P}(Y > 27.36)$ è stato ricavato dalle tavole.

8.3.2 Legge t di Student

La legge t di Student è utile per stimare il valor medio di una popolazione normale quando non sia nota la varianza.

Definizione 8.3.3. Si dice **Legge t di Student con n gradi di libertà**, la legge di una v.a.

$$T = \frac{Z}{\sqrt{Y/n}}, \quad \text{dove } Z \sim \mathcal{N}(0,1),\ Y \sim \chi^2(n)$$

e si richiede che Z e Y siano indipendenti. Si usa scrivere $T \sim t(n)$.

Si può calcolare esplicitamente la densità della $t(n)$:

$$f_T(t) = c_n \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

il coefficiente di normalizzazione c_n si ricava imponendo che $\int_{-\infty}^{+\infty} f_T(t)dt = 1$; si potrebbe mostrare che

$$c_n = \frac{\Gamma((n+1)/2)}{n^{1/2}\Gamma(n/2)}.$$

.

Proprietà.

- Per $n \to \infty$, la legge $t(n)$ tende alla normale standard $\mathcal{N}(0,1)$. Infatti è facile stabilire che

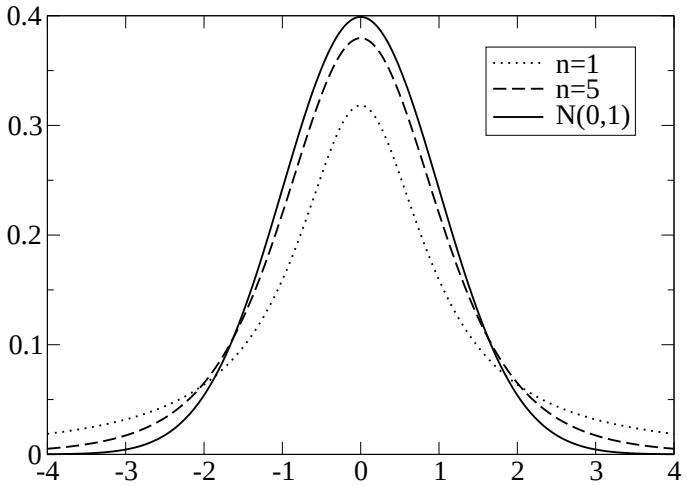
  $$\lim_{n\to\infty} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} = e^{-t^2/2};$$

  la giustificazione precisa viene da un teorema della teoria della misura che prende il nome di Teorema della Convergenza Dominata osservando che

  $$c_n \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \leq \frac{1}{\sqrt{2\pi}(1+t^2)}, \qquad \forall t \in \mathbb{R}, , \forall n \in \mathbb{N}.$$

- La densità della $t(n)$ è una funzione simmetrica pari, perciò il valore atteso è nullo.

  La varianza vale $n/(n-2)$ per $n > 2$ (mentre è infinita se $n = 1, 2$); pertanto la varianza è sempre maggiore di uno, e tende a 1 per $n \to \infty$.



Funzione densità della v.a. t di Student $t(n)$ e confronto con la normale standard $\mathcal{N}(0,1)$

- Indichiamo con $t_\alpha(n)$ i quantili della legge $t(n)$:

$$\mathbb{P}(T \leq t_\alpha(n)) = \alpha$$

Per la simmetria della funzione densità, valgono le seguenti proprietà, del tutto simili a quelle relative ai quantili della normale:

$$\mathbb{P}(T \geq t_{1-\alpha}(n)) = \alpha, \quad \mathbb{P}(|T| \geq t_{1-\frac{\alpha}{2}}(n)) = \alpha, \quad \mathbb{P}(|T| \leq t_{\frac{1+\alpha}{2}}(n)) = \alpha.$$

Per valori di n maggiori di 120 si possono approssimare i quantili della $t(n)$ con quelli della normale standard. Per n minore di 120 i valori si ricavano dalle tavole.

L'importanza della legge t di Student è dovuta alla seguente proprietà: siano $X_1, X_2, \ldots, X_n$, n v.a. normali i.i.d. di legge $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Allora

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

Infatti, sappiamo che $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, e dunque

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

D'altra parte, abbiamo visto in precedenza che

$$\frac{(n-1)}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

Essendo S_n^2 e $\overline{X}_n$ indipendenti, otteniamo

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{S_n^2/\sigma^2}} \sim t(n-1)$$

Esercizio 8.3.4. La ditta che vuole decidere se comprare la partita di bulloni dell'esempio precedente, procede a una misurazione a campione di 50 bulloni, e trova che il diametro medio del campione è di $2.04cm$ con una deviazione standard campionaria di $0.15cm$. Supponendo ancora

che il diametro dei bulloni segua una legge normale, calcolare la probabilità che il valore medio differisca di meno di $0.1cm$ dal valore dichiarato di $2cm$.

Soluzione.
Si considera la v.a.

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

dove S_n è la deviazione standard campionaria. Per quanto visto prima T_n ha distribuzione t di Student con $n-1$ gradi di libertà. Dunque

$$\mathbb{P}\left(t_\alpha < \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} < t_\beta\right) = \beta - \alpha$$

$$\mathbb{P}(\overline{X}_n - t_\beta S_n/\sqrt{n} < \mu < \overline{X}_n - t_\alpha S_n/\sqrt{n}) = \beta - \alpha$$

Imponiamo

$$\overline{X}_n - t_\beta S_n/\sqrt{n} = 1.99, \quad \overline{X}_n - t_\alpha S_n/\sqrt{n} = 2.01$$

$$t_\beta = (\overline{X}_n - 1.99)\sqrt{n}/S_n \simeq 2.357, \quad t_\alpha = (\overline{X}_n - 2.01)\sqrt{n}/S_n \simeq 1.414$$

Dalle tavole risulta $\beta \simeq 0.9888$, $\alpha \simeq 0.9182$. Pertanto la probabilità cercata è $\beta - \alpha \simeq 0.07$.

8.4 Stima per intervalli: intervalli di confidenza

Sia $\{X_1, \ldots, X_n\}$ un campione aleatorio estratto da una popolazione di legge $\mathcal{L}_\theta$. Siano $T_1 = t_1(X_1, \ldots, X_n)$ e $T_2 = t_2(X_1, \ldots, X_n)$ due statistiche, e sia $g(\theta)$ una funzione del (dei) parametro(i) θ. Fissato $\alpha \in [0,1]$, l'intervallo aleatorio (T_1, T_2) si dice **intervallo di confidenza per $g(\theta)$, al livello del** $100\alpha\%$ se

$$\mathbb{P}_\theta(T_1 < g(\theta) < T_2) = \alpha$$

per ogni $\theta \in \Theta$. A campionamento eseguito, l'intervallo ottenuto $[t_1(x_1, \ldots, x_n), t_2(x_1, \ldots, x_n)]$ si chiama intervallo di confidenza per $g(\theta)$, al livello del $100\alpha\%$, calcolato dal campione. Questo intervallo perde il significato di probabilità: **non** è vero che la probabilità che $g(\theta)$ sia compresa tra $t_1(x_1, \ldots, x_n)$ e $t_2(x_1, \ldots, x_n)$ è pari ad α. Per questo motivo si parla di confidenza e non di probabilità. A priori l'affermazione è vera, ma a posteriori non c'è più nulla di aleatorio.
È vero invece che se effettuassimo numerosi campionamenti e calcolassimo per ciascuno di questi l'intervallo di confidenza allo stesso livello, ci aspettiamo, in virtù della Legge dei Grandi Numeri Forte, che una proporzione del $100\alpha\%$ degli intervalli contenga il valore di $g(\theta)$.

Approfondimento

In generale dati due insiemi X e Y ed una funzione $f : X \to \mathcal{P}(Y)$ (dove $\mathcal{P}(Y)$ è l'insieme delle parti di Y), si definisce $f^* : Y \to \mathcal{P}(X)$, la **funzione d'insieme coniugata**, come

$$x \in f^*(y) \iff y \in f(x);$$

ovviamente $(f^*)^* = f$. Sia ora $(X, \mathcal{F})$ uno spazio misurabile, una funzione $f : X \to \mathcal{P}(Y)$ si dice misurabile se e solo se per ogni $y \in Y$ si ha $f^*(y) = \{x \in X : y \in f(x)\} \in \mathcal{F}$ (cioè misurabile rispetto alla σ-algebra $\mathcal{F}_1$ su $\mathcal{P}(Y)$ generata dagli insiemi del tipo

$$A_y := \{A \subseteq Y : y \in A\}.$$

Se ora $(X, \mathcal{F}, \mathbb{P}) = (\Omega, \mathcal{F}, \mathbb{P})$ è uno spazio di probabilità e $g : \Theta \to Y$ dove Θ è lo spazio dei parametri, allora una famiglia di funzioni misurabili, parametrizzate da $\theta \in \Theta$ $f_\theta : \Omega \to \mathcal{P}(Y)$ prende il nome di **intervallo di confidenza** per $g(\theta)$ **a livello** $\alpha \in [0,1]$ se e solo se

$$\mathbb{P}_\theta(g(\theta) \in f_\theta) \geq \alpha.$$

Quindi in generale esistono infiniti intervalli di confidenza di livello fissato; vedremo in seguito quali ragionevoli ipotesi aggiungere, caso per caso, per determinare univocamente la "forma" dell'intervallo.

8.4.1 Intervalli di confidenza per la media

Utilizzando i risultati descritti sopra ci proponiamo ora di costruire gli *intervalli fiduciari per la media* nei due casi in cui rispettivamente la varianza sia nota e la varianza sia incognita. Ricordiamo che in generale $\mathbb{P} = \mathbb{P}_\theta$ e di conseguenza $\mathbb{E} = \mathbb{E}_\theta$ e $\mathrm{Var} = \mathrm{Var}_\theta$; nel seguito sottintenderemo quasi sempre la dipendenza da θ così da rendere più compatta la notazione.

- *Intervallo fiduciario per la media di una popolazione con varianza nota.*

  Consideriamo un campione casuale $(X_1, X_2, \ldots, X_n)$ di ampiezza n estratto da una popolazione avente valor medio μ incognito e varianza σ^2 nota. Lo stimatore per μ è la media campionaria $\overline{X}_n$ per la quale supporremo che

  $$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

  Questa relazione è esatta se la legge della popolazione è normale, mentre vale solo asintoticamente per $n \to \infty$ altrimenti (in virtù del Teorema Centrale del Limite).

  La standardizzata di $\overline{X}_n$

  $$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

  è distribuita secondo la normale standard

  $$Z_n \sim \mathcal{N}(0,1).$$

  Fissato il livello di confidenza α, possiamo affermare che

  $$\mathbb{P}\left(\frac{|\overline{X}_n - \mu|}{\sigma/\sqrt{n}} \leq q_{\frac{1+\alpha}{2}}\right) = \alpha$$

  ovvero

  $$\mathbb{P}(|\overline{X}_n - \mu| \leq q_{\frac{1+\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = \alpha$$

  Gli estremi dell'intervallo sono

  $$\mu_\pm = \overline{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{\frac{1+\alpha}{2}}$$

  Abbiamo dunque costruito un intervallo *casuale*, centrato sul valore (casuale) $\overline{X}_n$, avente ampiezza fissata nota $2q_{(1+\alpha)/2}\sigma/\sqrt{n}$. Tale intervallo casuale ha probabilità α di contenere il valore vero μ.
  Una volta eseguito il campionamento ed ottenuta la stima del valor medio $\overline{x}_n$ si ottiene l'intervallo di confidenza

  $$\left[\overline{x}_n - q_{\frac{1+\alpha}{2}}\sigma/\sqrt{n}, \overline{x}_n + q_{\frac{1+\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right].$$

  N.B. A campionamento eseguito, non si può più tuttavia affermare che esso contiene μ con probabilità α.

  Osservazione 8.4.1. Si noti che l'ampiezza dell'intervallo fiduciario, fissati σ ed n, è tanto più grande quanto maggiore è il livello di fiducia poiché $q_{(1+\alpha)/2} \to \infty$ per $\alpha \to 1$. Pertanto innalzare il livello fiduciario aumenta il margine di errore su μ. Se si vuole mantenere un margine di errore prefissato, e nel contempo un livello fiduciario elevato, è necessario aumentare n; si noti come l'ampiezza $2q_{(1+\alpha)/2}\sigma/\sqrt{n}$ decresca come $1/\sqrt{n}$; quindi per diminuire l'errore di un ordine di grandezza è necessario aumentare l'ampiezza del campione di due ordini di grandezza.

Se la variabile aleatoria di partenza è una Bernoulliana, per un campione sufficientemente grande si ottiene l'intervallo di confidenza inserendo al posto di σ il valore $\sqrt{\overline{x}_n(1-\overline{x}_n)}$ (si veda il paragrafo 8.4.3 per maggiori dettagli).

Approfondimento

In generale, come abbiamo già detto, l'equazione $\mathbb{P}(g(\theta) \in I_\alpha) = \alpha$ (o più in generale la disequazione $\mathbb{P}(g(\theta) \in I_\alpha) \geq \alpha$) può avere infinite soluzioni rispetto all'intervallo di confidenza I_α (che chiameremo **intervallo di confidenza a livello** α). Quello che abbiamo fatto in precedenza suggerisce un metodo piuttosto generale per ottenere intervalli di confidenza. Supponiamo di avere la famiglia di stimatori T_n per la grandezza $g(\theta)$ e di costruire una famiglia di variabili aleatorie parametriche $Q(T_n, \theta)$ la cui legge non dipenda da θ; $Q(T_n, \theta)$ prende il nome di **quantità pivotale**. A questo punto se g è una funzione a valori in $\mathbb{R}$ e la legge della quantità pivotale ha funzione di ripartizione F si ha $\mathbb{P}(Q(T_n, \theta) \in (a, b]) = F(b) - F(a)$; estendiamo per comodità il dominio di F nel seguente modo

$$F(+\infty) := 1, \qquad F(-\infty) := 0.$$

Si osservi che una buona richiesta potrebbe essere $Q(\theta, \theta) \in (a, b]$ per α sufficientemente grande e per ogni $\theta \in \Theta$.
A questo punto supponiamo di poter risolvere rispetto a θ nel seguente senso

$$Q(T_n, \theta) \in (a, b] \iff g(\theta) \in I_{a,b}$$

dove l'intervallo $I_{a,b}$ è casuale e dipende anche da n. Se siamo in grado di fare tutto questo, quello che ci rimane da fare è operare una scelta opportuna per la coppia (a, b) in maniera che $F(b) - F(a) = \alpha \in (0, 1)$ (ove questo sia possibile). Per fare questo dobbiamo mettere altre condizioni (non molto restrittive) sulla legge F.

Supponiamo che la legge F sia strettamente crescente e continua e sia q_α la sua funzione quantile estesa nel seguente modo

$$q_0 := -\infty, \qquad q_1 := +\infty.$$

Quindi $F : [-\infty, +\infty] \to [0, 1]$ è una funzione biettiva continua con inversa q anch'essa continua. Sia $\alpha \in (0, 1)$; si vede facilmente che per ogni scelta di $a \in [-\infty, +\infty]$ tale che $F(a) \in [0, 1 - \alpha]$ si ha che esiste un unico valore $b \in [q_\alpha, +\infty]$ tale che

$$F(b) - F(a) = \alpha$$

(quindi $(a, b) = (q_\beta, q_{\alpha+\beta})$ per qualche $\beta \in [0, 1 - \alpha]$).

Definizione 8.4.2. La scelta $\beta = 0$ porta ad un **intervallo unilatero sinistro**, $\beta = 1 - \alpha$ porta ad un **intervallo unilatero destro**, mentre la scelta $\beta = (1 - \alpha)/2$ porta ad un **intervallo bilatero**.

Nel caso della media abbiamo scelto un intervallo bilatero di lunghezza minima.

- *Intervallo fiduciario per la media di una popolazione con varianza incognita.*

  Consideriamo come nel caso precedente un campione casuale $(X_1, X_2, \ldots, X_n)$ di ampiezza n estratto da una popolazione con legge normale $\mathcal{N}(\mu, \sigma^2)$ (o, alternativamente, si supponga

che la popolazione abbia legge con media μ e varianza σ^2 e l'ampiezza del campione sia molto elevata). Costruiamo la v.a.

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

Fissato il livello fiduciario α abbiamo:

$$\mathbb{P}(|T_n| \leq t_{\frac{1+\alpha}{2}}(n-1)) = \alpha$$

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

$$\mathbb{P}\left(|\overline{X}_n - \mu| \leq t_{\frac{1+\alpha}{2}}(n-1)\frac{S_n}{\sqrt{n}}\right) = \alpha$$

I due valori estremi di μ sono

$$\mu_\pm = \overline{X}_n \pm \frac{S_n}{\sqrt{n}} t_{\frac{1+\alpha}{2}}(n-1)$$

Osservazione 8.4.3. – Anche in questo caso l'intervallo fiduciario è centrato su $\overline{X}_n$, tuttavia la sua ampiezza $2t_{\frac{1+\alpha}{2}}(n-1)S_n/\sqrt{n}$ non è più nota a priori, ma è a sua volta una v.a..

– Fissato un livello fiduciario α, l'ampiezza dell'intervallo e quindi l'errore nella stima di μ tende a zero (quasi certamente) per $n \to \infty$ (in generale la convergenza non è monotona), poiché S_n^2 è uno stimatore consistente di σ^2.

Esercizio 8.4.4. Un laboratorio di analisi controlla il quantitativo medio di catrame contenuto in una certa marca di sigarette. In un campione di 30 sigarette si trovano i seguenti valori per la media campionaria $\overline{x}_n$ e la deviazione standard campionaria s_n:

$$\overline{x}_n = 10.92mg, \qquad s_n = 0.51mg$$

Si determini l'intervallo fiduciario per il quantitativo medio di catrame al livello del 99%

Soluzione.

$$\alpha = 0.99 \qquad \frac{1+\alpha}{2} = 0.995 \qquad t_{0.995}(29) \simeq 2.756$$

Gli estremi dell'intervallo sono

$$10.92 - 2.756\frac{0.51}{\sqrt{30}}, \qquad 10.92 + 2.756\frac{0.51}{\sqrt{30}}$$

$$10.92 - 0.25, \qquad 10.92 + 0.25$$

Si noti che se avessimo considerato la deviazione standard campionaria come il valore vero ed avessimo considerato il quantile $q_{0.995} \simeq 2.33$ avremmo trovato un intervallo fiduciario leggermente più stretto:

$$10.92 \pm \frac{0.51}{\sqrt{30}} 2.33 \simeq 10.92 \pm 0.22$$

8.4.2 Intervalli di confidenza per la varianza

Ci proponiamo di costruire gli *intervalli fiduciari per la varianza* nei due casi in cui rispettivamente il valor medio sia noto e il valor medio sia incognito.

- *Intervallo fiduciario per la varianza di una popolazione con media nota.*

  Partiamo come prima da un campione casuale $(X_1, X_2, \ldots, X_n)$ di ampiezza n estratto da una popolazione con legge normale $\mathcal{N}(\mu, \sigma^2)$ (o, alternativamente, si supponga che la popolazione abbia legge con media μ e varianza σ^2 e l'ampiezza del campione sia molto elevata).

  Essendo μ nota, la v.a.

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

  è uno stimatore corretto per la varianza σ^2. Si ha poi, in virtù di quanto visto nel paragrafo 8.3.1,

  $$\frac{nT_n^2}{\sigma^2} \sim \chi^2(n).$$

  Nei casi visti in precedenza del calcolo di intervalli di confidenza per il valor medio si procedeva a costruire un intervallo centrato sull'origine e tale che la probabilità che la v.a. standardizzata appartenga a detto intervallo sia α. Qui il procedimento va modificato perché la densità della legge chi-quadrato non è simmetrica rispetto all'origine.

  Si possono adottare più punti di vista: quello di maggiorare la varianza (è il caso ad esempio in cui si voglia avere una stima di errori sperimentali), quello in cui si voglia costringere la varianza all'interno di un intervallo (è il caso in cui si vuole determinare il valore esatto della varianza) e, più raramente, quello di minorare la varianza.

  Nel primo caso otteniamo una maggiorazione sul valore della varianza imponendo che

  $$\mathbb{P}\left(\frac{nT_n^2}{\sigma^2} \leq \chi^2_\alpha(n)\right) = \alpha \quad \Rightarrow \quad \mathbb{P}\left(\frac{nT_n^2}{\sigma^2} \geq \chi^2_{1-\alpha}(n)\right) = \alpha$$

  $$\mathbb{P}\left(\sigma^2 \leq \frac{nT_n^2}{\chi^2_{1-\alpha}(n)}\right) = \alpha.$$

  Otteniamo l'intervallo fiduciario:

  $$\sigma^2 \in \left[0, \frac{nT_n^2}{\chi^2_{1-\alpha}(n)}\right].$$

  Nel secondo caso, poniamo $Y \sim \chi^2(n)$. È ragionevole considerare un intervallo $[a, b]$, $0 < a < b$, tale che $\mathbb{P}(a < Y < b) = \alpha$, e inoltre che $\mathbb{P}(Y < a) = \mathbb{P}(Y > b)$ (ossia le code hanno uguale probabilità):

  $$\mathbb{P}(Y < a) = \mathbb{P}(Y > b) = \frac{1-\alpha}{2}$$

  Si ricava:

  $$a = \chi^2_{\frac{1-\alpha}{2}}(n), \qquad b = \chi^2_{\frac{1+\alpha}{2}}(n).$$

  Con questa scelta dell'intervallo possiamo scrivere

  $$\mathbb{P}\left(\chi^2_{\frac{1-\alpha}{2}}(n) \leq \frac{nT_n^2}{\sigma^2} \leq \chi^2_{\frac{1+\alpha}{2}}(n)\right) = \alpha$$

ovvero

$$\mathbb{P}\left(\frac{nT_n^2}{\chi^2_{\frac{1+\alpha}{2}}(n)} \leq \sigma^2 \leq \frac{nT_n^2}{\chi^2_{\frac{1-\alpha}{2}}(n)}\right) = \alpha$$

Il valore esatto σ^2 della varianza ha probabilità α di essere contenuto nell'intervallo aleatorio

$$\left[\frac{nT_n^2}{\chi^2_{\frac{1+\alpha}{2}}(n)}, \frac{nT_n^2}{\chi^2_{\frac{1-\alpha}{2}}(n)}\right].$$

Nell'ultimo caso si ha, analogamente al primo caso

$$\mathbb{P}\left(\frac{nT_n^2}{\sigma^2} \leq \chi^2_\alpha(n)\right) = \alpha \quad \Rightarrow \quad \mathbb{P}\left(\sigma^2 \geq \frac{nT_n^2}{\chi^2_\alpha(n)}\right) = \alpha$$

da cui otteniamo l'intervallo fiduciario:

$$\sigma^2 \in \left[\frac{nT_n^2}{\chi^2_\alpha(n)}, +\infty\right).$$

- *Intervallo fiduciario per la varianza di una popolazione con media incognita.*

  Partiamo sempre da un campione casuale $(X_1, X_2, \ldots, X_n)$ di ampiezza n estratto da una popolazione con legge normale $\mathcal{N}(\mu, \sigma^2)$ (o, alternativamente, si supponga che la popolazione abbia legge con media μ e varianza σ^2 e l'ampiezza del campione sia molto elevata).

  La v.a. T_n^2 che abbiamo usato in precedenza, ora *non* è più una statistica poiché è funzione della media che non è nota.
  La v.a. varianza campionaria invece è uno stimatore corretto:

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

  dove, per quanto visto nel paragrafo 8.3.1,

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1).$$

  Come nel caso precedente, distinguiamo i tre casi in cui vogliamo rispettivamente maggiorare la varianza, costringerla in un intervallo opportuno oppure minorarla.

  Nel primo caso imponiamo che

$$\mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_\alpha(n-1)\right) = \alpha \quad \Rightarrow \quad \mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma^2} \geq \chi^2_{1-\alpha}(n-1)\right) = \alpha$$

$$\mathbb{P}\left(\sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{1-\alpha}(n-1)}\right) = \alpha.$$

  L'intervallo di confidenza della varianza con livello α è dunque

$$\left[0, \frac{(n-1)S_n^2}{\chi^2_{1-\alpha}(n-1)}\right].$$

  Nel secondo caso poniamo $Y \sim \chi^2(n-1)$ e consideriamo un intervallo di confidenza $[a, b]$ tale che

$$\mathbb{P}(Y < a) = \mathbb{P}(Y > b) = \frac{1-\alpha}{2}$$

$$\mathbb{P}\left(\chi^2_{\frac{1-\alpha}{2}}(n-1) \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{\frac{1+\alpha}{2}}(n-1)\right) = \alpha$$

ovvero

$$\mathbb{P}\left(\frac{(n-1)S_n^2}{\chi^2_{\frac{1+\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi^2_{\frac{1-\alpha}{2}}(n-1)}\right) = \alpha$$

Il valore esatto σ^2 della varianza ha probabilità α di essere contenuto nell'intervallo aleatorio

$$\left[\frac{(n-1)S_n^2}{\chi^2_{\frac{1+\alpha}{2}}(n-1)}, \frac{(n-1)S_n^2}{\chi^2_{\frac{1-\alpha}{2}}(n-1)}\right].$$

Nel terzo ed ultimo caso si ha,

$$\mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_\alpha(n-1)\right) = \alpha \quad \Rightarrow \quad \mathbb{P}\left(\sigma^2 \geq \frac{(n-1)S_n^2}{\chi^2_\alpha(n-1)}\right) = \alpha$$

da cui otteniamo l'intervallo fiduciario:

$$\sigma^2 \in \left[\frac{(n-1)S_n^2}{\chi^2_\alpha(n-1)}, +\infty\right).$$

8.4.3 Intervalli di confidenza per una popolazione

Supponiamo di voler campionare una legge di Bernoulli di parametro p incognito. Siano quindi $X_1, \ldots, X_n$ i.i.d. con legge $B(p)$.

Premettiamo il seguente lemma.

Lemma. Siano $x_1, \ldots, x_n$ numeri reali in $\{0, 1\}$. Allora se $\overline{x}_n := (1/n)\sum_{i=1}^n x_i$

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

Dim. Osserviamo che $x_i \in \{0, 1\}$ per ogni i se e solo se $x_i = x_i^2$, pertanto

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

Si osservi che dal precedente lemma si ottiene immediatamente che la varianza campionaria S_n^2 di un campione proveniente da una legge bernoulliana è noto se si conosce la media campionaria $\overline{x}_n$ (ma in generale non vale il viceversa, essendo S_n^2 lo stesso se prendiamo $\overline{y}_n := 1 - \overline{x}_n$ al posto di $\overline{x}_n$). Infatti si ottiene

$$S_n^2 = \frac{n}{n-1}\overline{x}_n(1 - \overline{x}_n).$$

Approfondimento

Lemma. Sia $f : X \to \mathbb{R}$ una funzione misurabile rispetto a $(X, \mathcal{F}, \mu)$. Se $f(x)(1 - f(x)) = 0$ per quasi ogni x e f è integrabile, allora

$$\int_X \left(f - \int_X f\mathrm{d}\mu\right)^2 \mathrm{d}\mu = \int_X f\mathrm{d}\mu \left(1 - \int_X f\mathrm{d}\mu\right).$$

Essendo p corrispondente al valore medio, utilizziamo come stimatore non distorto la media campionaria

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

Dal Teorema Centrale del Limite abbiamo che

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \approx Y \sim \mathcal{N}(0,1)$$

da cui

$$\mathbb{P}\left(\left|\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}}\right| \leq q_{\frac{1+\alpha}{2}}\right) \approx \alpha$$

pertanto

$$\left|\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}}\right| \leq q_{\frac{1+\alpha}{2}}$$

$$\Updownarrow$$

$$(\overline{X}_n - p)^2 \leq q^2_{\frac{1+\alpha}{2}} \frac{p(1-p)}{n}$$

$$\Updownarrow$$

$$p^2\left(1 + \frac{q^2_{\frac{1+\alpha}{2}}}{n}\right) - p\left(2\overline{X}_n + \frac{q^2_{\frac{1+\alpha}{2}}}{n}\right) + \overline{X}_n^2 \leq 0$$

$$\Updownarrow$$

$$\overline{X}_n - \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\, q_{\frac{1+\alpha}{2}} + \frac{q_{\frac{1+\alpha}{2}}}{2n} \leq p \leq \overline{X}_n + \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\, q_{\frac{1+\alpha}{2}} + \frac{q_{\frac{1+\alpha}{2}}}{2n}$$

che è una prima possibile scelta per l'intervallo di confidenza bilatero per il parametro p.

Una seconda possibilità è data dalla seguente approssimazione

$$\frac{\overline{X}_n - p}{S_n/\sqrt{n}} \approx Y_n \sim t(n-1)$$

da cui, essendo

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

si ha

$$\mathbb{P}\left(\left|\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)/(n-1)}}\right| \leq t_{\frac{1+\alpha}{2}}(n-1)\right) \approx \alpha$$

cioè

$$\left|\frac{\overline{X}_n - p}{\sqrt{\overline{X}_n(1-\overline{X}_n)/(n-1)}}\right| \leq t_{\frac{1+\alpha}{2}}(n-1)$$

$$\Updownarrow$$

$$\overline{X}_n - t_{\frac{1+\alpha}{2}}(n-1)\sqrt{\overline{X}_n(1-\overline{X}_n)/(n-1)} \leq p \leq \overline{X}_n + t_{\frac{1+\alpha}{2}}(n-1)\sqrt{\overline{X}_n(1-\overline{X}_n)/(n-1)}.$$

che è una seconda possibile scelta per l'intervallo di confidenza bilatero.

Per comodità di calcolo, giustificata anche dal fatto che generalmente n è molto grande, prendiamo come intervallo di confidenza bilatero per il parametro p di una popolazione il seguente

$$p \in \left[\overline{X}_n - q_{\frac{1+\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},\ \overline{X}_n + q_{\frac{1+\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

Osserviamo che, per n grande, i due intervalli calcolati in precedenza non sono dissimili da quest'ultimo.

Esempio 8.4.5. Affrontiamo un problema reale: ci viene commissionato un exit-poll per un certo partito.

Soprassediamo sulla problematica della scelta del campione se non per osservare che l'ampiezza n deve essere sufficientemente grande per poter applicare il Teorema Centrale de Limite ma non troppo relativamente al numero degli elettori, per poter assumere che ogni elettore del campione, rispetto alla domanda "ha votato per il partito A?" sia ben rappresentato da una prova di Bernoulli (ciascun elettore del campione viene assunto indipendente dagli altri). Se il campione fosse troppo ampio bisognerebbe ricorrere all'uso della cosiddetta "legge ipergeometrica". Un campione sufficiente per il numero di elettori italiani (circa 40000000) ha un ampiezza tra i 5000 ed i 10000. Il problema è che l'intervistato potrebbe avere dei problemi a rivelare la sua vera scelta, per cui si ricorre al seguente stratagemma: si prende una moneta truccata che abbia probabilità $q \neq 1/2$ di successo (diciamo "testa"), a tale scopo un dado è una buona scelta. Ciascun intervistato lancia la moneta (senza che nessuno lo veda) e se esce testa dirà la verità altrimenti dirà il falso (perchè questo "mascheramento" funzioni, è necessario che $p \in (0,1)$).

Come stimiamo a livello di confidenza di α la percentuale di elettori del partito A?

Modellizziamo il problema come segue: si prendono due processi di Bernoulli indipendenti $\{X_i\}_{i=1}^n$ ed $\{Y_i\}_{i=1}^n$ di parametri, rispettivamente, p e q; il primo processo, che vale 1 se l'i-esimo intervistato ha votato per A e 0 altrimenti, ha parametro p da stimare, mentre il secondo, che simula il lancio della moneta e vale 1 se esce testa e 0 altrimenti, ha parametro noto e differente da $1/2$. La risposta dell'i-esimo intervistato alla domanda "ha votato per il partito A" è determinata nella maniera seguente: sarà "si" se $X_i = 1$ ed $Y_i = 1$ o se $X_i = 0$ ed $Y_i = 0$, sarà 0 altrimenti. In definitiva è una variabile Z_i che vale

$$Z_i := \begin{cases} 1 & \text{se } X_i = Y_i \\ 0 & \text{se } X_i \neq Y_i \end{cases}$$

cioè è un processo di Bernoulli di parametro

$$\mathbb{P}(X_i = Y_i) = pq + (1-p)(1-q) = 2pq + 1 - p - q =: h.$$

Si noti che la funzione $p \mapsto 2pq + 1 - p - q$ è iniettiva (da $[0,1]$ su $[\min(q, 1-q), \max(q, 1-q)]$) se e solo se $q \neq 1/2$ (altrimenti vale $1/2$ per ogni valore di p); l'inversa vale $p = (h + q - 1)/(2q - 1)$. Pertanto è possibile risalire al valore di p se e solo se $q \neq 1/2$. Il problema della stima di p si riduce così a quello della stima di h. L'intervallo di confidenza a livello α sappiamo essere

$$h \in \left[\overline{z}_n - q_{\frac{1+\alpha}{2}} \sqrt{\frac{\overline{z}_n(1-\overline{z}_n)}{n}},\ \overline{z}_n + q_{\frac{1+\alpha}{2}} \sqrt{\frac{\overline{z}_n(1-\overline{z}_n)}{n}} \right]$$

da cui, nel caso $q \in (1/2, 1]$,

$$p \in \left[\frac{\overline{z}_n - q_{\frac{1+\alpha}{2}} \sqrt{\frac{\overline{z}_n(1-\overline{z}_n)}{n}}}{2q-1} + \frac{q-1}{2q-1},\ \frac{\overline{z}_n + q_{\frac{1+\alpha}{2}} \sqrt{\frac{\overline{z}_n(1-\overline{z}_n)}{n}}}{2q-1} + \frac{q-1}{2q-1} \right]$$

mentre, nel caso $q \in [0, 1/2)$,

$$p \in \left[\frac{1-q}{1-2q} - \frac{\overline{z}_n + q_{\frac{1+\alpha}{2}} \sqrt{\frac{\overline{z}_n(1-\overline{z}_n)}{n}}}{2q-1},\ \frac{1-q}{1-2q} + \frac{\overline{z}_n - q_{\frac{1+\alpha}{2}} \sqrt{\frac{\overline{z}_n(1-\overline{z}_n)}{n}}}{2q-1} \right].$$

Notiamo che, per rendere piccola la probabilità che la stima di h porti ad un valore di p negativo o superiore ad 1 è meglio prendere q vicino ad $1/2$.

Cap. 9. Statistica inferenziale: test d'ipotesi

9.1 Definizioni

9.1.1 Ipotesi statistica

Sia $\{\mathcal{L}_\theta\}_{\theta\in\Theta}$ una famiglia di leggi di probabilità di una certa popolazione statistica, θ uno o più parametri tutti o in parte incogniti (nella maggiorparte dei casi la legge $\mathcal{L}_\theta$ sarà completamente specificata, per ogni $\theta \in \Theta$ da una densità $f_X(\cdot, \theta)$).
Si dice **ipotesi statistica** un'asserzione sul valore vero dei parametri θ incogniti.
Un'ipotesi statistica si dice **semplice** se specifica completamente la legge $f_X(x, \theta)$, altrimenti si dice **composta**.

Esempi. Supponiamo che una certa grandezza sia distribuita normalmente secondo la legge $\mathcal{N}(\mu, 4)$, allora

- L'ipotesi $\mu = 5$ *è semplice* perché specifica completamente la legge normale.
- L'ipotesi $\mu \leq 3$ *è composta* perché non specifica completamente la legge normale.

In generale, come nel capitolo precedente, il valore di θ varia all'interno di un insieme Θ detto **spazio dei parametri** e un'ipotesi su θ ha la forma

$$\theta \in \Theta_0$$

dove Θ_0 è un sottoinsieme di Θ. L'ipotesi è semplice se e solo se Θ_0 contiene un solo punto.

L'ipotesi che intendiamo sottoporre a verifica si dice **ipotesi nulla**, si indica con H_0 e viene ritenuta *vera fino a prova contraria*. Nella costruzione di un test, si sceglie come H_0 l'ipotesi alla quale si è disposti a rinunciare solo in caso di forte evidenza del contrario.

Esempi.

- $H_0 : \mu = 4$
- $H_0 : \mu > 5$
- $H_0 : 0.1 \leq p \leq 0.7$

L'**ipotesi alternativa** H_1 sarà del tipo $\theta \in \Theta_1$ dove $\Theta_1 \cap \Theta_0 = \emptyset$. Nella maggiorparte dei casi sceglieremo la cossiddetta **ipotesi complementare** ad H_0, cioè $\theta \in \Theta_0^\complement \equiv \Theta \setminus \Theta_0$.

Esempi.

- $H_0 : \mu = 4$, $H_1 : \mu \neq 4$
- $H_0 : \mu > 5$, $H_1 : \mu \leq 5$
- $H_0 : 0.1 \leq p \leq 0.7, \quad H_1 : p < 0.1, p > 0.7$

9.1.2 Verifica d'ipotesi

Si dice **verifica d'ipotesi**, o **test d'ipotesi** il procedimento con cui si decide, sulla base di una stima ottenuta dai dati campionari, se accettare o meno l'ipotesi.

Ad esempio, se l'ipotesi nulla fosse $H_0 : \mu = 4$ per una distribuzione normale $\mathcal{N}(\mu, 1)$, si potrebbe pensare di usare come stimatore la media campionaria di un campione. Non sarebbe però ragionevole richiedere che il valore della media campionaria ottenuto sia esattamente uguale a 4, perché entrano in gioco le fluttuazioni statistiche. È più sensato richiedere che il valore medio si situi in un intorno opportunamente piccolo del valore 4.

Nella esecuzione di un test si possono avere i seguenti esiti

- il test *accetta* H_0 quando (dato che) H_0 *è vera*. La decisione è corretta.
- il test *rifiuta* H_0 quando (dato che) H_0 *è vera*. In questo caso si commette un errore di **I tipo** o **Ia specie**.
- il test *accetta* H_0 quando (dato che) H_0 *è falsa*. In questo caso si commette un errore di **II tipo** o **IIa specie**.
- il test *rifiuta* H_0 quando (dato che) H_0 *è falsa*. La decisione è corretta.

Utilizzeremo il verbo "accettare" come sinonimo di "non rifiutare"; tuttavia dal punto di vista matematico si potrebbe vedere che è più probabile commettere un errore quando si accetta H_0 rispetto a quando la si rifiuta, per questo motivo a volte si sostituisce "accettare H_0" con "non posso rifiutare H_0" o "non ho elementi sufficienti per rifiutare H_0". "Non rifiutare H_0" prende il nome di **conclusione debole**, mentre "rifiutare H_0" si dice **conclusione forte**.

Riassumendo:

	H_0 è vera	H_0 è falsa
Rifiutiamo H_0	*Errore di Ia specie*	Decisione corretta
Accettiamo H_0	Decisione corretta	*Errore di IIa specie*

L'errore del Ia specie è considerato più grave di quello di IIa specie. In altre parole noi cercheremo di impostare il test (cioè sceglieremo H_0 ed H_1) affinchè l'errore di Ia specie sia quello che vorremmo evitare di commettere.

Esempio 9.1.1. Consideriamo il processo ad un imputato. Formuliamo prima l'ipotesi H_0: *l'imputato è colpevole*. Otteniamo la tabella seguente:

	H_0: L'imputato è colpevole	H_1: L'imputato è innocente
Viene assolto	*Errore di Ia specie*	Decisione corretta
Viene condannato	Decisione corretta	*Errore di IIa specie*

Mentre se assumessimo come ipotesi H_0: *l'imputato è innocente* otterremmo:

	H_0: L'imputato è innocente	H_1: L'imputato è colpevole
Viene condannato	*Errore di Ia specie*	Decisione corretta
Viene assolto	Decisione corretta	*Errore di IIa specie*

Ritenendo che sia più grave condannare un innocente rispetto a lasciare un colpevole in libertà, dobbiamo scegliere l'ipotesi nulla H_0 come nel secondo caso: *l'imputato è innocente*.

Esempio 9.1.2. Due persone giocano con un dado. Una delle due persone ha il sospetto che il dado sia truccato. Decide di effettuare un gran numero di lanci e di registrare il numero di volte in cui esce il 6. L'ipotesi nulla è $H_0 : p(6) = 1/6$ (ipotesi *innocentista*). Il test d'ipotesi sarà del tipo: *rifiuto* H_0 *se* $|\overline{X}_n - 1/6| > k$ (per un valore opportuno di k), dove $\overline{X}_n$ è la media campionaria delle v.a. Bernoulliane $X_i \sim B(1/6)$ che valgono 1 se all'i-esimo lancio del dado è venuto il 6.

Esempio 9.1.3. Una ditta produce bicchieri con spessore medio alla base dichiarato di $4mm$. Prima di decidere se mettere in vendita il prodotto vuole effettuare delle misurazioni su un campione. Sapendo che gli acquirenti riterranno importante che lo spessore abbia una valore minimo garantito (per ragioni di robustezza dei bicchieri), ma soprattutto essendo importante far partire le vendite del prodotto al più presto formula l'ipotesi H_0 : *lo spessore medio della base è almeno pari a quello dichiarato*. Il test d'ipotesi sarà: *rifiuto* H_0 *se* $\overline{X}_n < k$ per un opportuno valore di k.

La scelta di H_0 ed H_1 dipende quindi dai punti di vista, cioè dipende da quale errore si cerca di non commettere.

Esempio 9.1.4. Un'epidemia di meningite è scoppiata e si dispone della percentuale $\overline{p}$ di ammalati nella prima settimana (stima della probabilità di ammalarsi di quest'anno p), nonché della stessa percentuale $\overline{p}_0$ relativa all'anno precedente (stima della probabilità di ammalarsi dell'anno precedente p_0). L'istituto di sanità vuole capire se l'epidemia dell'anno in corso sia più pericolosa di quella dell'anno precedente.
Punto di vista cautelativo: non si vuole correre il rischio di sottovalutare l'epidemia, quindi non si vuole correre il rischio di pensare che $p < p_0$ se non è vero. Quest'ultimo deve diventare quindi l'errore di Ia specie e pertanto la scelta di H_0 (risp. H_1) sarà $p \geq p_0$ (risp. $p < p_0$).
Punto di vista non allarmistico: non si vuole correre il rischio di sopravvalutare l'epidemia e diffondere il panico tra la popolazione, quindi non si vuole correre il rischio di pensare che $p > p_0$ se non è vero. Questa volta la scelta di H_0 (risp. H_1) sarà $p \leq p_0$ (risp. $p > p_0$).

9.1.3 Regione critica e funzione potenza

Fissate le ipotesi

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

il test d'ipotesi consiste nello scegliere una statistica appropriata $T = t(X_1, \ldots, X_n)$ (in generale a valori in $\mathbb{R}^m$), e nello stabilire una regola di decisione per accettare o rifiutare l'ipotesi. In generale T sarà uno stimatore di θ.
Precisamente, adottiamo la seguente *regola di decisione*: *si rifiuti* H_0 *se (dopo il campionamento)* $t(x_1, \ldots, x_n) \in I$, dove $I \subseteq \mathbb{R}^m$. Allora l'insieme $\mathcal{RC}$ delle realizzazioni campionarie che portano a rifiutare H_0, cioè

$$\mathcal{RC} = \{(x_1, \ldots, x_n) : t(x_1, \ldots, x_n) \in I\}$$

è detta *regione critica*, o *di rifiuto* del test.

Esempio 9.1.5. H_0 è l'ipotesi dell'esempio 9.1.3: *lo spessore medio della base del bicchiere è almeno pari a quello dichiarato*. La statistica è la media campionaria $\overline{X}_n$; fissiamo k e stabiliamo la regola di decisione si rifiuti H_0 se $\overline{X}_n < k$; l'insieme $\mathcal{R}$ è l'insieme dei possibili risultati campionari che forniscono una media campionaria nella regione critica:

$$\mathcal{RC} = \{(x_1, \ldots, x_n) : \overline{x}_n < k\}$$

Una volta definita la regione critica, si può pensare di calcolare, per ogni valore possibile del parametro incognito, la probabilità che l'ipotesi venga rifiutata.

Definizione 9.1.6. Fissato il test, lo stimatore T e l'insieme I che determina la regione di rifiuto dell'ipotesi nulla, si definisce la **funzione potenza** $\mathrm{Pot} : \Theta \to [0, 1]$ nel seguente modo:

$$\mathrm{Pot}(\theta) = \mathrm{Pot}_I(\theta) := \mathbb{P}_\theta(T \in I).$$

Nel caso in cui $\theta \in \Theta_0$ allora $\mathrm{Pot}(\theta)$ rappresenta la probabilità di un errore di Ia specie, mentre se $\theta \in \Theta_1$ allora $1 - \mathrm{Pot}(\theta)$ è la probabilità di commettere un errore di IIa specie.

Nota. Qualche volta scriveremo, abusando un po' della notazione, $\mathrm{Pot}(y) = \mathbb{P}(T \in I | \theta = y)$; in realtà, nel nostro approccio (detto *non Bayesiano*), θ non è una variabile aleatoria e quindi $\{\theta \in A\}$ non è un evento significativo (i.e. è Ω oppure $\emptyset$).

Nell'esempio precedente, supponiamo che lo spessore della base dei bicchieri segua una legge normale con varianza σ^2 nota; il parametro incognito è μ, e

$$\mathrm{Pot}(\mu) = \mathbb{P}_\mu(\overline{X}_n < k)$$

Si può calcolare esplicitamente $\mathrm{Pot}(\mu)$, ricordando che $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$:

$$\mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < q_{\alpha(\mu)}\right) = \alpha(\mu)$$

$$\mathbb{P}\left(\overline{X}_n < \mu + q_{\alpha(\mu)}\frac{\sigma}{\sqrt{n}}\right) = \alpha(\mu)$$

Poniamo $\mu + q_{\alpha(\mu)}\frac{\sigma}{\sqrt{n}} = k$:

$$q_{\alpha(\mu)} = \frac{(k-\mu)\sqrt{n}}{\sigma}, \qquad \mathrm{Pot}(\mu) = \alpha(\mu) = \Phi\left(\frac{(k-\mu)\sqrt{n}}{\sigma}\right)$$

9.1.4 Livello di significatività

Definiamo ora il **livello di significatività**, anche detto **ampiezza del test**. Consideriamo il problema di verifica dell'ipotesi

$$H_0 : \theta \in \Theta_0$$

contro

$$H_1 : \theta \in \Theta_1$$

(dove, ricordiamo, $\Theta_0 \cap \Theta_1 = \emptyset$; solitamente $\Theta_1 = \Theta_0^\complement$). L'*ampiezza*, o *livello di significatività* α del test basato su un campione di dimensione n con regione critica

$$\mathcal{RC} = \{(x_1, \ldots, x_n) : t(x_1, \ldots, x_n) \in I\}$$

è definito come

$$\alpha = \sup_{\theta \in \Theta_0} \mathrm{Pot}(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}(t(X_1, \ldots, X_n) \in I);$$

quindi α rappresenta la massima probabilità di rifiutare l'ipotesi nulla, quando questa è vera; è cioè la massima probabilità di fare un errore di Ia specie. Più α è piccolo, più siamo confidenti in una decisione corretta, quando la regola di decisione ci porta a rifiutare l'ipotesi nulla. Contemporaneamente si può calcolare la massima probabilità di errore di IIa specie

$$\beta := \sup_{\theta \in \Theta_1} (1 - \mathrm{Pot}(\theta)) = 1 - \inf_{\theta \in \Theta_1} \mathrm{Pot}(\theta).$$

Nella pratica il valore di α viene stabilito a priori, prima di eseguire il campionamento, e l'insieme $I = I_\alpha$ viene ottenuto di conseguenza.

Quindi quello che cercheremo di fare è, per ciascun test, operare una scelta di insiemi $\{I_\alpha\}_{\alpha \in (0,1)}$ con alcune proprietà:

(i) $\alpha \geq \beta$ implica $I_\alpha \supseteq I_\beta$

(ii) $\bigcap_{\alpha \in (0,1)} I_\alpha = \emptyset$

(iii) $\bigcup_{\alpha \in (0,1)} I_\alpha = \mathbb{R}^m$

(iv) $\sup_{\theta\in\Theta_0} \mathrm{Pot}_{I_\alpha}(\theta) = \alpha$, per ogni $\alpha \in (0,1)$.

Per alcuni aspetti che saranno chiari nel Paragrafo 9.1.5 può essere utile estendere la definizione della regione critica anche ai casi $\alpha \in \{0,1\}$ nel seguente modo:

(v) si consideri $\{I_\alpha\}_{\alpha\in[0,1]}$ con la convenzione $I_0 := \emptyset$ e $I_1 := \mathbb{R}^m$.

Quindi si possono prendere in considerazione le proprietà (i), (ii), (iii) e (iv) oppure (i), (iv) e (v); quest'ultima implica evidentemente

$$\bigcap_{\alpha\in[0,1]} I_\alpha = \emptyset \qquad \bigcup_{\alpha\in[0,1]} I_\alpha = \mathbb{R}^m$$

analoghe, rispettivamente, di (ii) e (iii).

Valori tipici per il livello di significatività α sono 0.1, 0.05, 0.01.
Nell'esempio precedente scegliamo un livello di significatività α e cerchiamo il valore di k corrispondente.

$$\sup_{\mu\geq 4} \mathbb{P}(\overline{X}_n < k) = \sup_{\mu\geq 4} \Phi\left(\frac{k-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{k-4}{\sigma/\sqrt{n}}\right)$$

Scegliendo

$$k = 4 + \frac{q_\alpha \sigma}{\sqrt{n}} = 4 - \frac{q_{1-\alpha}\sigma}{\sqrt{n}}$$

otteniamo proprio

$$\sup_{\mu\geq 4} \mathbb{P}(\overline{X}_n < k) = \Phi(q_\alpha) = \alpha$$

Dunque al livello di significatività α, la regola di decisione dell'ipotesi nulla *lo spessore alla base è almeno pari a quello dichiarato* è "si rifiuti H_0 se $\overline{x}_n < 4 - \frac{q_{1-\alpha}\sigma}{\sqrt{n}}$".

Riassumiamo i passi di un test statistico:

1. Si scelgono l'ipotesi nulla H_0 e la sua alternativa H_1. Nella scelta va condotto un giudizio su quale delle due ipotesi sia la più importante o, più precisamente quale ipotesi non vorremmo rifiutare nel caso fosse vera (quest'ultima diventerà H_0).

2. Si sceglie una statistica per stimare il parametro su cui effettuare il test, e si stabilisce la forma della regione critica (ad esempio: si rifiuti H_0 se $\overline{X}_n < k$; k è ancora indeterminato).

3. Si sceglie il livello di significatività α a cui si vuole eseguire il test. Più α è piccolo e più difficilmente rifiuteremo l'ipotesi nulla, e più certi saremo di non sbagliare quando la rifiutiamo.

4. Si determina la regione del rifiuto in funzione del valore α scelto (ad esempio si rifiuti H_0 se $\overline{X}_n < 4 - \frac{q_{1-\alpha}\sigma}{\sqrt{n}}$).

5. Si esegue il campionamento, si calcola la statistica definita nel punto 2 e si vede se il risultato appartiene o meno alla regione di rifiuto: in caso positivo si rifiuta l'ipotesi nulla, in caso negativo la si accetta.

9.1.5 P-value

Supponiamo che le proprietà (i), (ii), (iii) e (iv) (oppure (i), (iv) e (v)) definite nel paragrafo 9.1.4 siano soddisfatte. Poiché k dipende dal livello di significatività impostato, una stessa ipotesi che è stata rifiutata diciamo al livello dell' 1% può essere invece accettata ad un livello inferiore.
Esiste un livello di significatività limite, detto **P-value**, pari al più basso livello di significatività a cui i dati campionari consentono di rifiutare l'ipotesi nulla.

Definizione 9.1.7. Dato un test,

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

uno stimatore $T : \mathbb{R}^n \to \mathbb{R}^m$ e l'insieme delle regioni critiche $\mathcal{RC}_\alpha := \{T \in I_\alpha\}$ con le proprietà specificate sopra, allora, dopo aver campionato ed ottenuto la stima $t = t(x_1, \ldots, x_n)$, si chiama **P-value** il valore $\overline{\alpha}$ definito da

$$\overline{\alpha} := \inf\{\alpha \in (0,1) : t \in I_\alpha\} \equiv \sup\{\alpha \in (0,1) : t \notin I_\alpha\}.$$

Approfondimento

La seconda uguaglianza si dimostra ricordando che in uno spazio totalmente ordinato $(X, \geq)$ (ad esempio $\mathbb{R}$ con il suo ordinamento naturale) per ogni $A \subset X$ sono equivalenti:

1. $x \in A, y \geq x \Longrightarrow y \in A$;
2. $x \in A^{\complement}, y \leq x \Longrightarrow y \in A^{\complement}$;

Inoltre se per $(X, \geq)$ vale la proprietà che per ogni x, y tali che $x < y$ esiste z tale che $x < y < z$ (per esempio $\mathbb{R}$ con il suo ordinamento naturale), allora ciascuna delle due precedenti equivale a ciascuna delle due seguenti:

1. Esiste $\sup(A^{\complement})$ se e solo se esiste $\inf(A)$ ed in tal caso $\sup(A) = \inf(A^{\complement})$.
2. Esiste $\sup(A^{\complement})$ se e solo se esiste $\inf(A)$ ed in tal caso $\sup(A^{\complement}) \geq \inf(A)$.

L'esistenza di un tale valore $\overline{\alpha}$ è garantita dalle proprietà (i), (ii) e (iii) della famiglia $\{I_\alpha\}_{\alpha \in (0,1)}$ (un'identica colnclusione si otterrebbe utilizzando l'altro set (i), (iv) e (v)).

La procedura decisionale diviene quindi "Rifiuto H_0" a livello α se e solo se $\alpha > \overline{\alpha}$ (in virtù della proprietà (i)); il caso $\alpha = \overline{\alpha}$ va esaminato separatamente, ad esempio se I_α sono intervalli aperti (e sotto opportune ipotesi di continuità) $t \notin I_{\overline{\alpha}}$. In generale valori di α vicini al P-value non danno indicazioni forti né in un senso né nell'altro.

Sotto opportune ipotesi di continuità che non è qui il caso di specificare (saranno tutte soddisfatte nei casi che considereremo), $\overline{\alpha}$ soddisfa l'equazione $t \in \partial I_{\overline{\alpha}}$ (dove ∂A è l'usuale frontiera di $A \subseteq \mathbb{R}^m$).
In tutti i casi che considereremo, lo si otterrà ponendo l'uguaglianza al posto della disuguaglianza che definisce la regione critica del test.

Nell'esempio precedente, il *P-value* è soluzione dell'equazione

$$\overline{x}_n = 4 - \frac{q_{1-\overline{\alpha}}\sigma}{\sqrt{n}}$$

ossia

$$\overline{\alpha} = \Phi\left(\frac{\overline{x}_n - 4}{\sigma/\sqrt{n}}\right)$$

- Un *P-value* molto piccolo significa che H_0 può venire rifiutata con tranquillità.
- Un *P-value* basso ma non piccolissimo, dell'ordine dei consueti livelli di significatività (cioè 0.01, 0.05, ...) vuol dire che la decisione di rifiutare H_0 dipende fortemente dal livello di significatività impostato.
- Un *P-value* alto vuol dire che H_0 può essere plausibilmente accettata.

Ritornando alle proprietà (i)–(iv) del Paragrafo 9.1.4 e al loro legame con il P-value, osserviamo che

- se cadesse la proprietà (ii) allora potrebbe accadere che $t_n(x_1, \dots, x_n) \in I_\alpha$ per ogni $\alpha \in (0,1)$, in tal caso definiamo $\overline{\alpha} := 0$ (si veda ad esempio il Paragrafo 9.2 nel caso $H_0 : \mu \neq \mu_0$ e $\overline{x}_n = \mu_0$);

- se cadesse la proprietà (iii) allora potrebbe accadere che $t_n(x_1, \dots, x_n) \notin I_\alpha$ per ogni $\alpha \in (0,1)$, in tal caso definiamo $\overline{\alpha} := 1$ (si veda ad esempio il Paragrafo 9.2 nel caso $H_0 : \mu = \mu_0$ e $\overline{x}_n = \mu_0$).

Ovviamente a questo punto potremmo equivalentemente scegliere di soddisfare il set di proprietà (i), (iv) e (iv).

9.1.6 Confronto tra errore di Ia specie ed errore di IIa specie

Approfondimento

Consideriamo il test

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1.$$

Supponiamo di scegliere lo stimatore T e le regioni critiche

$$\{T \in I_\alpha\}$$

al variare di $\alpha \in (0,1)$ soddisfacenti le proprietà del paragrafo 9.1.4. Sia inoltre la funzione potenza (che dipende anche dal parametro $\alpha \in (0,1)$)

$$\mathrm{Pot}(\theta, \alpha) = \mathbb{P}_\theta(T \in I_\alpha).$$

Definizione 9.1.8. Se vale

$$\sup_{\theta \in \Theta_0} \mathrm{Pot}(\theta, \alpha) \leq \inf_{\theta \in \Theta_1} \mathrm{Pot}(\theta, \alpha)$$

per ogni $\alpha \in (0,1)$ allora il test prende il nome di **test non distorto**.

Se un test è non distorto e vale $\alpha = \sup_{\theta \in \theta_0} \mathrm{Pot}(\theta, \alpha)$ allora, definito il valore massimo dell'errore di IIa specie $\beta_\alpha := \sup_{\theta \in \theta_1}(1 - \mathrm{Pot}(\theta, \alpha))$, si ha $\alpha + \beta_\alpha \leq 1$. Tuttavia nel caso in cui valga la proprietà molto frequente

$$\sup_{\theta \in \Theta_0} \mathrm{Pot}(\theta, \alpha) = \inf_{\theta \in \Theta_1} \mathrm{Pot}(\theta) \quad (\equiv \alpha)$$

allora $\alpha + \beta_\alpha = 1$ e quindi se diminuiamo il rischio di errore di Ia specie aumentiamo automaticamente il rischio di commettere un errore di II specie.
Quindi scegliere valori di α troppo bassi porta ad accettare H_0 ma la probabilità di commettere un errore (IIa specie) può essere alta; viceversa scegliere valori di α troppo alti porta a rifiutare H_0 ma la probabilità di commettere un errore (Ia specie) può essere alta. Pertanto valori tipici per il livello di significatività del test vanno da 0.01 a 0.1.

Esempio 9.1.9. Nell'esempio 9.1.3 si ha

$$\mathrm{Pot}(\mu, \alpha) = \phi\left(q_\alpha + \frac{4-\mu}{\sigma/\sqrt{n}}\right) \Longrightarrow \inf_{\mu<4} \phi\left(q_\alpha + \frac{4-\mu}{\sigma/\sqrt{n}}\right) = \alpha.$$

9.1.7 Scambio delle ipotesi

Approfondimento

Consideriamo i due test

$$\mathcal{T}_1 \begin{cases} H_0 & : \theta \in \Theta_0 \\ H_1 & : \theta \in \Theta_1 \end{cases} \qquad \mathcal{T}_2 \begin{cases} H_0 & : \theta \in \Theta_1 \\ H_1 & : \theta \in \Theta_0, \end{cases}$$

dove, al solito, Θ_0, Θ_1 sono due sottoinsiemi disgiunti dello spazio Θ. Supponiamo di scegliere lo stimatore T e le regioni critiche, rispettivamente,

$$\{T \in I_\alpha\} \qquad \{T \in J_\alpha\},$$

con $\alpha \in (0,1)$ soddisfacenti le usuali proprietà. Definiamo, a campionamento avvenuto, i due P-value

$$\overline{\alpha} := \inf\{\alpha \in (0,1) : t \in I_\alpha\}, \qquad \overline{\beta} := \inf\{\alpha \in (0,1) : t \in J_\alpha\}$$

e le rispettive funzioni potenza (che dipendono anche dal parametro $\alpha \in (0,1)$)

$$\mathrm{Pot}(\theta, \alpha) = \mathbb{P}_\theta(T \in I_\alpha), \qquad \mathrm{Pot}_1(\theta, \alpha) = \mathbb{P}_\theta(T \in J_\alpha).$$

Ricordiamo quindi che

$$\sup_{\theta \in \Theta_0} \mathrm{Pot}(\theta, \alpha) = \alpha, \qquad \sup_{\theta \in \Theta_1} \mathrm{Pot}_1(\theta, \alpha) = \alpha,$$

per ogni $\alpha \in (0,1)$. Distinguiamo tre casi:

1. È possibile scegliere le regioni di rifiuto in maniera che $I_\alpha \subseteq J^{\complement}_{1-\alpha}$ (per ogni $\alpha \in (0,1)$) se e solo se

   $$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T \in I_\alpha) \geq \inf_{\theta \in \Theta_1} \mathbb{P}_\theta(T \in I_\alpha),$$

   ed in tal caso $\overline{\alpha} + \overline{\beta} \geq 1$.
   Per sincerarsene basta osservare che $T \in I_\alpha$ se e solo se $\alpha \geq \overline{\alpha}$, se $J^{\complement}_{1-\alpha} \supseteq I_\alpha$ allora si ha che

   $$\alpha \geq \overline{\alpha} \Longrightarrow T \in I_\alpha \Longrightarrow T \notin J_{1-\alpha} \Longrightarrow \alpha \leq 1 - \overline{\beta}$$

   e quindi passando al $\sup_{\alpha \geq \overline{\alpha}}$ si ha $\overline{\alpha} + \overline{\beta} \geq 1$. Per il viceversa si procede in maniera analoga (sotto opportune ipotesi di continuità).

2. È possibile scegliere le regioni di rifiuto in maniera che $I_\alpha \supseteq J^{\complement}_{1-\alpha}$ (per ogni $\alpha \in (0,1)$) se e solo se

   $$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T \in I_\alpha) \leq \inf_{\theta \in \Theta_1} \mathbb{P}_\theta(T \in I_\alpha),$$

   (cioè se e solo se il test $\mathcal{T}_0$ è non distorto) ed in tal caso $\overline{\alpha} + \overline{\beta} \leq 1$.

3. È possibile scegliere le regioni di rifiuto in maniera che $I_\alpha = J^{\complement}_{1-\alpha}$ (per ogni $\alpha \in (0,1)$) se e solo se

   $$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T \in I_\alpha) = \inf_{\theta \in \Theta_1} \mathbb{P}_\theta(T \in I_\alpha)$$

   ed in tal caso $\overline{\alpha} + \overline{\beta} = 1$. Per provarlo basta utilizzare i due punti precedenti.

Il caso di gran lunga più frequente nel seguito sarà l'ultimo. In questo caso, supponendo $\overline{\alpha} \leq 1/2$ (senza perdità di generalità), allora $\overline{\beta} = 1 - \overline{\alpha} > 1/2$.

a) I due test sono concordi a livello α se e solo se $\alpha \in (\overline{\alpha}, 1 - \overline{\alpha})$.

b) I due test sono discordi a livello α se e solo se $\min(\alpha, 1 - \alpha) < \overline{\alpha}$.

9.1.8 Confronto tra regioni di rifiuto e intervalli di confidenza

Approfondimento

Osserviamo che c'è una relazione molto stretta tra intervalli di confidenza e **regioni di accettazione** (cioè il complementare delle regioni di rifiuto). Lo schema è il seguente, dato un campione $X_1, \ldots, X_n$ proveniente da una legge $\mathcal{L}_\theta$ e uno stimatore T di θ, quello che si fa è cercare una quantità pivotale $Q(T, \theta)$ di legge $\mathbb{P}_X$ e degli insiemi H_α per $\alpha \in (0, 1)$ tali che $\mathbb{P}_X(H_\alpha) = \mathbb{P}(Q(T, \theta) \in H_\alpha) = \alpha$. Ad esempio una buona richiesta potrebbe essere che per ogni $\theta \in \Theta$ esista $\alpha_\theta \in [0, 1)$ tale che per ogni $\alpha > \alpha)\theta$ si abbia $Q(\theta, \theta) \in H_\alpha$. A questo punto si costruiscono due famiglie di insiemi $S_{T,\alpha} \subseteq \Theta$ e $J_{\theta,\alpha} \subset \mathbb{R}^m$ tali che

$$\theta \in S_{T,\alpha} \iff Q(T, \theta) \in H_\alpha \iff T \in J_{\theta,\alpha};$$

osserviamo che $S_{T,\alpha}$ è un insieme di confidenza a livello α, mentre l'insieme $\{T \in J^{\complement}_{\theta,\alpha}\}$ è una regione di rifiuto a livello $1 - \alpha$, cioè $\{T \in J_{\theta,\alpha}\}$ è una regione di accettazione. Se quindi definiamo, con $\alpha \in (0, 1)$ fissato, $f : \mathbb{R}^m \to \mathcal{P}(\Theta)$

$$f(t) := S_{t,\alpha}$$

allora la funzione di insieme coniugata è, per definizione,

$$f^*(\theta) = J_{\theta,\alpha}.$$

Esempio. Si consideri il test $\mu = \mu_0$ (Paragrafo 9.2).

9.2 Verifica di ipotesi sulla media (varianza nota)

Affrontiamo in modo sistematico il problema del test di ipotesi sulla media. Iniziamo dal caso in cui la varianza σ^2 della popolazione sia nota.

- Supponiamo per cominciare che l'ipotesi nulla sia

  $$H_0 : \mu \geq \mu_0$$

  mentre

  $$H_1 : \mu < \mu_0$$

  è l'ipotesi alternativa.

  – Il primo passo nella costruzione del test è la scelta di una statistica, *detta statistica test*, mediante la quale si stima il parametro incognito a partire dai dati campionari. Nel caso della media la statistica è naturalmente la media campionaria

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

per la quale vale (esattamente o asintoticamente)

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Osserviamo che stiamo verificando l'ipotesi $\mu \geq \mu_0$ che rigetteremo solo nel caso che la stima ottenuta dal campione sia "nettamente" al di sotto di μ_0.
  Fissiamo allora un valore $k < \mu_0$ e decidiamo di accettare H_0 se risulterà $\overline{x}_n \geq k$ e di rifiutarla in caso contario.
- La regione critica del test è l'insieme dei valori campionari

$$\mathcal{RC} = \{(x_1, \ldots, x_n) : \overline{x}_n < k\}$$

- *Come scegliere k?*
  Fissando il livello di significatività α del test, viene fissato di conseguenza il valore di k. Per legare k ad α partiamo dalla seguente espressione

$$\mathbb{P}_\mu\left(\overline{X}_n < k\right) = \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < \frac{k-\mu}{\sigma/\sqrt{n}}\right) = \alpha(\mu) \equiv \mathrm{Pot}(\mu).$$

  Poiché $X = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ (vale approssimativamente per un campione numeroso in virtù del Teorema Centrale del Limite, o per un campione qualsiasi estratto da una popolazione gaussiana), otteniamo

$$\mathbb{P}_\mu\left(X < \frac{k-\mu}{\sigma/\sqrt{n}}\right) = \alpha(\mu).$$

  Il valore massimo di questa probabilità, al variare di μ in H_0 (cioè per $\mu \geq \mu_0$), viene assunto in $\mu = \mu_0$. Pertanto

$$\frac{k-\mu_0}{\sigma/\sqrt{n}} = q_\alpha \equiv -q_{1-\alpha}$$

$$k = \mu_0 - \frac{\sigma q_{1-\alpha}}{\sqrt{n}}$$

  L'ipotesi $H_0 : \mu \geq \mu_0$ sarà accettata (al livello α) se la media stimata $\overline{x}_n$ risulta *maggiore o uguale* di k, altrimenti sarà rigettata:
  rifiuto H_0 se

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -q_{1-\alpha}$$

  Poiché k dipende dal livello di significatività impostato, una stessa ipotesi che è stata rifiutata ad un certo livello può essere invece accettata ad un livello inferiore.

- Il *P-value* si ottiene ponendo l'uguaglianza al posto della disuguaglianza precedente:

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} = -q_{1-\overline{\alpha}}$$

$$\overline{\alpha} = \Phi\left(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right)$$

Osservazione 9.2.1. Notiamo che il *P-value* è in generale una funzione complicata di n perché dipende non solo dal valore numerico n ma anche dal valore numerico della media campionaria $\overline{x}_n$. Inoltre, anche supponendo che la varianza campionaria sia costante in n il *P-value* può avere andamenti monotoni opposti. Prendiamo ad esempio il test appena visto,

il cui *P-value* è $\overline{\alpha} = \Phi\left(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\right)$. Si vede immediatamente che, supponendo $\overline{x}_n$ costante in n, se per qualche n_0 si ha $\overline{\alpha} > 1/2$ (i.e $\overline{x}_{n_0} - \mu_0 > 0$) allora $\overline{\alpha} \uparrow 1$ quando $n \to \infty$; viceversa se $\overline{\alpha} < 1/2$ (i.e $\overline{x}_{n_0} - \mu_0 < 0$) allora $\overline{\alpha} \downarrow 0$ quando $n \to \infty$.

Il tipico esempio è il seguente. In un test sulla media a varianza nota dove $H_0 : \mu \geq \mu_0$ e $H_1 : \mu < \mu_0$ si rifiuta (risp. non si rifiuta) H_0 ad un livello $\alpha < 1/2$ (risp. $\alpha > 1/2$) in corrispondenza ad un campione di ampiezza n. Se ora prendiamo un campione di ampiezza $m > n$ per il quale vale $\overline{x}_m = \overline{x}_n$ cosa succede al *P-value*?
R. Il *P-value* diminuisce (risp. aumenta). Ciò è dovuto al fatto che rifiutare (risp. non rifiutare) H_0 ad un livello $\alpha < 1/2$ (risp. $\alpha > 1/2$) implica che $\overline{\alpha} < 1/2$ (risp. $\overline{\alpha} < 1/2$).

- Il caso in cui l'ipotesi nulla e quella alternativa sono

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0$$

si tratta in modo del tutto analogo al precedente o, in alternativa, ci si riconduce al caso precedente ponendo $Y_i := -X_i$, $\mathbb{E}(Y_i) = -\mu =: \widetilde{\mu}$ e $\widetilde{\mu}_0 := -\mu_0$ e testando l'ipotesi $H_0 : \widetilde{\mu} \geq \widetilde{\mu}_0$ contro $H_1 : \widetilde{\mu} < \widetilde{\mu}_0$.

Si rifiuta H_0 se

$$\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > q_{1-\alpha}$$

Il *P-value* è

$$\overline{\alpha} = \Phi\left(\frac{\mu_0 - \overline{x}_n}{\sigma/\sqrt{n}}\right)$$

- Consideriamo il caso in cui l'ipotesi nulla e quella alternativa sono

$$H_0 : \mu = \mu_0, \qquad H_1 : \mu \neq \mu_0$$

In questo caso il test deve essere costruito in modo da rifiutare uno scostamento, di qualunque segno, maggiore di un certo k da determinarsi a partire da α.
Per determinare la relazione tra k ed α scriviamo

$$\mathbb{P}_{\mu_0}\left(|\overline{X}_n - \mu_0| > k\right) = \mathbb{P}_{\mu_0}\left(\frac{|\overline{X}_n - \mu_0|}{\sigma/\sqrt{n}} > \frac{k}{\sigma/\sqrt{n}}\right) = \alpha$$

da cui, ricordando che per una v.a. $X \sim \mathcal{N}(0,1)$ si ha

$$\mathbb{P}(|X| > q_{1-\frac{\alpha}{2}}) = \alpha$$

otteniamo

$$\frac{k}{\sigma/\sqrt{n}} = q_{1-\frac{\alpha}{2}}$$
$$k = q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Fissato α si rifiuterà H_0 se

$$|\overline{x}_n - \mu_0| > q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Il P-value viene ottenuto ponendo l'uguaglianza

$$|\overline{x}_n - \mu_0| = q_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

da cui si ricava

$$\overline{\alpha} = 2 - 2\Phi\left(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right).$$

- Consideriamo infine il caso in cui l'ipotesi nulla e quella alternativa siano

$$H_0 : \mu \neq \mu_0, \qquad H_1 : \mu = \mu_0$$

Analogamente al caso precedente rifiuteremo l'ipotesi nulla se la media campionaria è "ragionevolmente vicina al valore μ_0". Per determinare la relazione tra k ed α scriviamo

$$\mathbb{P}_\mu\left(|\overline{X}_n - \mu| < k\right) = \mathbb{P}_\mu\left(\frac{|\overline{X}_n - \mu|}{\sigma/\sqrt{n}} < \frac{k}{\sigma/\sqrt{n}}\right) = \alpha(\mu).$$

Osserviamo che l'estremo superiore della funzione $\alpha(\mu)$ su $\mu \neq \mu_0$, per continuità si ha in $\mu = \mu_0$; inoltre, ricordando ancora che per una v.a. $X \sim \mathcal{N}(0,1)$ si ha

$$\mathbb{P}(|X| < q_{\frac{1+\alpha}{2}}) = \alpha$$

otteniamo

$$\frac{k}{\sigma/\sqrt{n}} = q_{\frac{1+\alpha}{2}}$$

$$k = q_{\frac{1+\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Fissato α si rifiuterà H_0 se

$$|\overline{x}_n - \mu_0| < q_{\frac{1+\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Il P-value viene ottenuto, al solito, ponendo l'uguaglianza

$$|\overline{x}_n - \mu_0| = q_{\frac{1+\overline{\alpha}}{2}}\frac{\sigma}{\sqrt{n}}$$

da cui si ricava

$$\overline{\alpha} = 2\Phi\left(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right) - 1.$$

Esempio 9.2.2. Da una popolazione normale di media incognita e deviazione standard $\sigma = 3$ si estrae un campione di ampiezza 20, e si sottopone a test l'ipotesi nulla $H_0 : \mu = 100$.

a) Troviamo la regione critica ai livelli dell'1%, del 5% e del 10%.
Per quanto visto sopra la regione critica del test è data da quei valori di $\overline{x}_n$ per cui si ha

$$\frac{|\overline{x}_n - 100|}{\sigma/\sqrt{n}} > q_{1-\frac{\alpha}{2}}$$

con

α	0.01	0.05	0.1
$q_{1-\frac{\alpha}{2}}$	2.578	1.96	1.6449

b) Supponendo di avere estratto un campione per cui $\overline{x}_n = 98.5$, si tragga una conclusione, per ciascuno dei tre livelli di significatività.
Sostituiamo nella formula precedente $\overline{x}_n$ con 98.5:

$$\frac{|\overline{x}_n - 100|}{\sigma/\sqrt{n}} = \frac{|98.5 - 100|}{3/\sqrt{20}} \approx 2.2361$$

al livello dell'1% l'ipotesi nulla viene accettata, mentre ai livelli del 5% e del 10% viene rifiutata.

c) Calcoliamo infine il P-value:

$$\overline{\alpha} = 2 - 2\Phi\left(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\right) = 2 - 2\Phi\left(\frac{|98.5 - 100|}{3/\sqrt{20}}\right) \approx 0.0253$$

Tutte le ipotesi nulle relative a test con livello di significatività inferiore al 2.53% sono accettate, mentre quelle con livello maggiore sono rifiutate.

Riassumiamo i risultati ottenuti in questa sezione nella tabella seguente:

H_0	H_1	Rifiutare H_0 se	P-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z\rvert > q_{1-\alpha/2}$	$2 - 2\Phi(\lvert z\rvert)$
$\mu \leq \mu_0$	$\mu > \mu_0$	$z > q_{1-\alpha}$	$\Phi(-z)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$z < -q_{1-\alpha}$	$\Phi(z)$
$\mu \neq \mu_0$	$\mu = \mu_0$	$\lvert z\rvert < q_{(1+\alpha)/2}$	$2\Phi(\lvert z\rvert) - 1$

$$\text{dove } \; z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

9.3 Test su una frequenza (grandi campioni)

Vogliamo sottoporre a verifica d'ipotesi un campione tratto da una popolazione Bernoulliana $X_i \sim B(p)$.

Consideriamo le ipotesi nulle

$$H_0 : p = p_0; \quad H_0 : p \leq p_0; \quad H_0 : p \geq p_0; \quad H_0 : p \neq p_0$$

e le loro rispettive alternative

$$H_1 : p \neq p_0; \quad H_1 : p > p_0; \quad H_1 : p < p_0; \quad p = p_0.$$

Utilizziamo la proprietà che, se il campione è sufficientemente numeroso, la media campionaria tende a una v.a. normale:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Perciò

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0,1).$$

Possiamo ragionare come nella sezione precedente. L'unica differenza è che la deviazione standard σ viene sostituita con $\sqrt{p_0(1-p_0)}$; tale risultato si raggiunge con calcoli analoghi a quelli fatti in precedenza per la media, tenendo conto del fatto che la funzione $p \mapsto (a-p)/\sqrt{p(1-p)/n}$ è una funzione decrescente da $(0,1)$ in $\mathbb{R}$, per ogni $a \in [0,1]$ ed $n \in \mathbb{N}^*$; vale infatti

$$\frac{a-p}{\sqrt{p(1-p)/n}} = \sqrt{n}\left(a\sqrt{\frac{1-p}{p}} - (1-a)\sqrt{\frac{p}{1-p}}.\right)$$

Pertanto, come nel caso precedente, il valore massimo della potenza (su Θ_0) viene assunto in $p = p_0$; otteniamo in definitiva la tabella seguente

H_0	H_1	Rifiutare H_0 se	P-value
$p = p_0$	$p \neq p_0$	$\lvert z\rvert > q_{1-\alpha/2}$	$2 - 2\Phi(\lvert z\rvert)$
$p \leq p_0$	$p > p_0$	$z > q_{1-\alpha}$	$\Phi(-z)$
$p \geq p_0$	$p < p_0$	$z < -q_{1-\alpha}$	$\Phi(z)$
$p \neq p_0$	$p = p_0$	$\lvert z\rvert < q_{(1+\alpha)/2}$	$2\Phi(\lvert z\rvert) - 1$

$$\text{dove } \; z = \frac{\overline{x}_n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Esempio 9.3.1. Un partito politico ha ricevuto nelle ultime elezioni il 35% dei voti. Quattro anni dopo, da un sondaggio d'opinione basato su 300 interviste si è trovato che il 32% degli intervistati ha dichiarato di essere disposto a votare per quel partito. Ci si chiede se, rispetto al risultato elettorale, la situazione del partito sia peggiorata.

Si tratta di un test d'ipotesi sul parametro p di una popolazione Bernoulliana $B(p)$. Se prendiamo il punto di vista *propagandistico* (il partito non vuole ammettere facilmente di aver perso consensi) allora l'ipotesi da verificare (ipotesi nulla) è

$$H_0 : p \geq 0.35$$

mentre l'ipotesi alternativa è

$$H_1 : p < 0.35$$

La standardizzata vale

$$z = \frac{\overline{x}_n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.32 - 0.35}{\sqrt{0.35 \cdot 0.65/300}} \approx -1.0894$$

Il P-value corrispondente al dato campionario è

$$\overline{\alpha} = \Phi(z) \approx \Phi(-1.0894) \approx 0.1380$$

L'ipotesi H_0 viene accettata da ogni test il cui livello di significatività sia inferiore al P-value, cioè al 13.8%.

9.4 Verifica di ipotesi sulla media (varianza incognita)

Consideriamo ora il caso in cui la varianza σ^2 della popolazione sia incognita.

Riprendiamo per cominciare la verifica dell'ipotesi

$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

Come nel caso a varianza nota, fissiamo un valore $k < \mu_0$ e decidiamo di rifiutare H_0 se $\overline{x}_n$ dovesse risultare inferiore a k. La probabilità che $\overline{X}_n < k$, *dato* μ è

$$\mathbb{P}_\mu\left(\overline{X}_n < k\right) = \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} < \frac{k-\mu}{S_n/\sqrt{n}}\right) =$$

$$= \mathbb{P}_\mu(T_n < t_{\alpha(\mu)}) = \alpha(\mu).$$

Nella espressione precedente si è sostituito σ con il suo stimatore S_n e siamo passati dalla standardizzata $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ alla v.a.

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

Seguendo lo stesso ragionamento del caso a varianza nota possiamo dire che

$$\alpha = \sup_{\mu \geq \mu_0} \alpha(\mu) = \alpha(\mu_0)$$

e quindi fissato il livello di significatività α abbiamo

$$\mathbb{P}(T_n < -t_{1-\alpha}) = \alpha$$

ovvero

$$\frac{k-\mu_0}{S_n/\sqrt{n}} = -t_{1-\alpha}$$

Una volta effettuato il campionamento ed ottenute le stime $\overline{x}_n$ ed s_n rifiuteremo H_0 se

$$\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}} < -t_{1-\alpha}$$

Osservazione 9.4.1. Il procedimento seguito è esatto se la popolazione da cui si estrae il campione è normale; è ancora approssimativamente valido, per popolazioni non normali, se il campione è sufficientemente numeroso.

Il P-value viene ottenuto ponendo

$$t_{1-\overline{\alpha}} = -\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}$$

Ragionando in modo del tutto analogo si ottengono facilmente le regioni critiche e i P-value nei casi $H_0 : \mu \leq \mu_0$, $H_0 : mu = \mu_0$ e $H_0 : \mu \neq \mu_0$.

I risultati sono riassunti nella tabella seguente.

H_0	H_1	Rifiutare H_0 se	P-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert > t_{1-\alpha/2}(n-1)$	$t_{1-\overline{\alpha}/2}(n-1) = \lvert t\rvert$
$\mu \leq \mu_0$	$\mu > \mu_0$	$t > t_{1-\alpha}(n-1)$	$t_{1-\overline{\alpha}}(n-1) = t$
$\mu \geq \mu_0$	$\mu < \mu_0$	$t < -t_{1-\alpha}(n-1)$	$t_{1-\overline{\alpha}}(n-1) = -t$
$\mu \neq \mu_0$	$\mu = \mu_0$	$\lvert t\rvert < t_{(1+\alpha)/2}(n-1)$	$t_{(1+\overline{\alpha})/2}(n-1) = \lvert t\rvert$

$$\text{dove } \; t = \frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}$$

9.5 Verifica d'ipotesi sulla varianza

Sia $X_1, \ldots, X_n$ un campione aleatorio estratto da una popolazione normale $\mathcal{N}(\mu, \sigma^2)$. Ci proponiamo di sottoporre a verifica l'ipotesi H_0 riguardante la varianza σ^2.

Esempio 9.5.1. In un processo di produzione di wafer al silicio si richiede che la varianza dello spessore del singolo wafer sia al più di 0.5 micron. Avendo riscontrato una varianza campionaria di 0.64 micron su un campione di 50 wafer, si vuole sottoporre a verifica con livello di significatività $\alpha = 0.05$ l'ipotesi H_0: *la deviazione standard dello spessore dei wafer è minore o uguale a 0.5 micron*.

L'analisi è diversa a seconda che il valore medio μ sia noto o incognito.

- Trattiamo innanzitutto il caso in cui μ sia ignoto. Una statistica test appropriata per la varianza della popolazione è la varianza campionaria

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

Sottoponiamo a verifica l'ipotesi

$$H_0 : \sigma^2 \leq \sigma_0^2$$

Il test d'ipotesi sarà del tipo:

$$\textit{rifiuto } H_0 \textit{ se } S_n^2 > k$$

per un valore opportuno di k da stabilire in funzione del livello di significatività α scelto.

Precisamente k viene ottenuto imponendo che

$$\sup_{\sigma\in(0,\sigma_0]} \mathbb{P}(S_n^2 > k) = \alpha.$$

La v.a. $(n-1)S_n^2/\sigma^2$ ha una legge chi-quadrato con $n-1$ gradi di libertà.

Il sup della probabilità viene assunto per $\sigma = \sigma_0$, infatti:

$$\mathbb{P}_\sigma(S_n^2 > k) = \mathbb{P}_\sigma\left(\frac{(n-1)S_n^2}{\sigma^2} > \frac{k(n-1)}{\sigma^2}\right) = \mathbb{P}\left(Y_n > \frac{k(n-1)}{\sigma^2}\right)$$

dove $Y_n \sim \chi^2(n-1)$. Il il secondo membro della disuguaglianza è decrescente in σ^2, e pertanto la probabilità è crescente. Dunque l'estremo superiore viene assunto in corrispondenza al valore massimo per σ, cioè in σ_0.

Otteniamo dunque:

$$\mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n-1)\right) = \alpha$$

dove $\chi^2_{1-\alpha}(n-1)$ è il quantile $1-\alpha$ della legge $\chi^2(n-1)$.

$$\mathbb{P}\left(S_n^2 > \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha}(n-1)\right) = \alpha$$

il valore di k è perciò

$$k = \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha}(n-1)$$

e la regola di decisione del test è in definitiva:

rifiuto H_0 *se* $(n-1)S_n^2/\sigma_0^2 > \chi^2_{1-\alpha}(n-1)$.

Il P-value $\overline{\alpha}$ viene ottenuto ponendo l'uguaglianza

$$\chi^2_{1-\overline{\alpha}} = \frac{(n-1)S_n^2}{\sigma_0^2}.$$

Nell'esempio 9.5.1 dei wafer di silicio, abbiamo

$$\frac{(n-1)S_n^2}{\sigma_0^2} = \frac{49 \cdot 0.64}{0.5} = 62.72, \quad \chi^2_{0.95}(49) \simeq 66.34,$$

pertanto l'ipotesi H_0 viene accettata.

Il P-value è dato da:

$$\chi^2_{1-\overline{\alpha}} = 62.72 \quad \Rightarrow \quad \overline{\alpha} \simeq 0.09$$

Per ogni livello di significatività inferiore al 9% l'ipotesi nulla viene accettata, mentre per ogni livello di significatività superiore al 9% viene rifiutata.

Ragionando in modo analogo per le ipotesi nulle di altro tipo (H_0: $\sigma^2 = \sigma_0^2$, H_0: $\sigma^2 \neq \sigma_0^2$ e H_0: $\sigma^2 \geq \sigma_0^2$), si ottiene la seguente tabella:

H_0	H_1	Rifiutare H_0 se	P-value
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{1-\frac{\alpha}{2}}(n-1)$ o $\chi^2 < \chi^2_{\frac{\alpha}{2}}(n-1)$	$\overline{\alpha} = \min(\alpha_0, 1-\alpha_0)$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{1-\alpha}(n-1)$	$\chi^2_{1-\overline{\alpha}}(n-1) = \chi^2$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{\alpha}(n-1)$	$\chi^2_{\overline{\alpha}}(n-1) = \chi^2$
$\sigma^2 \neq \sigma_0^2$	$\sigma^2 = \sigma_0^2$	$\chi^2_{(1-\alpha)/2}(n-1) < \chi^2 < \chi^2_{(1+\alpha)/2}(n-1)$	$\overline{\alpha} = \lvert 2\alpha_0 - 1\rvert$

$$\text{dove } \chi^2 = \frac{(n-1)s_n^2}{\sigma_0^2}, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \mu)^2, \quad \chi^2_{\alpha_0}(n-1) = \chi^2.$$

- Nel caso in cui il valore medio μ sia noto, la statistica test appropriata è

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

Si ha poi

$$\frac{nT_n^2}{\sigma^2} \sim \chi^2(n)$$

Le formule della tabella precedente rimangono invariate, a patto di sostituire χ^2 con la quantità nt_n^2/σ_0^2, e di sostituire $n-1$ con n nel numero dei gradi di libertà della legge chi-quadrato:

H_0	H_1	Rifiutare H_0 se	P-value
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi^2_{1-\frac{\alpha}{2}}(n)$ o $\chi^2 < \chi^2_{\frac{\alpha}{2}}(n)$	$\overline{\alpha} = \min(\alpha_0, 1-\alpha_0)$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{1-\alpha}(n)$	$\chi^2_{1-\overline{\alpha}}(n) = \chi^2$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{\alpha}(n)$	$\chi^2_{\overline{\alpha}}(n) = \chi^2$
$\sigma^2 \neq \sigma_0^2$	$\sigma^2 = \sigma_0^2$	$\chi^2_{(1-\alpha)/2}(n) < \chi^2 < \chi^2_{(1+\alpha)/2}(n)$	$\overline{\alpha} = \lvert 2\alpha_0 - 1\rvert$

$$\text{dove } \chi^2 = \frac{nt_n^2}{\sigma_0^2}, \quad t_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \quad \chi^2_{\alpha_0}(n) = \chi^2.$$

9.6 Test chi-quadrato di buon adattamento

Supponiamo di avere n osservazioni di una variabile X raggruppate in N_c classi. Le classi possono rappresentare

- valori assunti da una variabile discreta: ogni classe raggruppa le osservazioni che assumono un determinato valore o un gruppo di valori.
- Intervalli di valori assunti da una variabile continua.
- caratteristiche qualitative assunte da una variabile categorica (colori, pariti votati, ecc.).

Sia $f_r(k)$ la frequenza relativa della k-esima classe.

Supponiamo di possedere una stima teorica dei valori che dovrebbe assumere la frequenza relativa. Ci poniamo il problema di valutare la bontà di adattamento delle frequenze osservate alle frequenze teoriche ipotizzate.

Ci chiediamo se la legge esponenziale sia adeguata a descrivere il fenomeno osservato.

Per risolvere questo tipo di problema si considera la seguente statistica test:

$$\begin{aligned} Q &:= \sum_{i=1}^{N_c}\frac{(np_i - f_a(i))^2}{np_i} = n\sum_{i=1}^{N_c}\frac{(p_i - f_r(i))^2}{p_i} \\ &= \sum_{i=1}^{N_c}\frac{f_a(i)^2}{np_i} - n = n\left(\sum_{i=1}^{N_c}\frac{f_r(i)^2}{p_i} - 1\right) \end{aligned} \tag{9.1}$$

dove $f_a(i)$ sono le frequenze assolute introdotte nel Capitolo 2 ($f_a(i) = nf_r(i)$), cioè il numero di osservazioni del campione appartenente alla i-esima classe. Si noti che np_i sono le frequenze assolute ipotizzate dalla teoria. In dipendenza dai dati a disposizione (frequenze assolute o relative si può utilizzare indifferentemente una qualsiasi delle forme equivalenti della statistica Q definita dall'equazione 9.1.

La statistica Q viene detta **chi-quadrato calcolato dal campione**. Q è tanto più piccola quanto migliore è l'adattamento delle frequnze osservate a quelle ipotizzate. Si può allora pensare di utilizzare Q per fare un **test di adattamento**, nel modo seguente:

L'ipotesi nulla è H_0: *le osservazioni provengono da una popolazione distribuita secondo le frequenze relative attese* $p_1, p_2, \dots, p_{N_c}$.

La procedura decisionale del test è del tipo:

si rifiuti H_0 *se* $Q > k$, per un valore di k opportuno.

Il teorema seguente permette di determinare la costante k una volta fissato il livello di significatività α:

Teorema 9.6.1. *Estraiamo un campione casuale di ampiezza n da una popolazione ripartita in N_c classi di frequenze relative teoriche $p_1, p_2, \dots, p_{N_c}$ (dove $p_i > 0$ e $\sum_{i=1}^{N_c} p_i = 1$). Sia $f_a(i)$ la frequenza assoluta osservata (estratta) relativa alla classe i-esima. Allora la statistica*

$$Q = \sum_{i=1}^{N_c} \frac{(np_i - f_a(i))^2}{np_i}$$

è una v.a. la cui legge tende (in legge) alla legge chi-quadrato $\chi^2(N_c - 1)$ per $n \to \infty$.
Se le frequenze relative attese p_i, invece di essere assegnate a priori, sono calcolate dopo aver stimato r parametri incogniti dai dati del campione*, allora la legge limite di Q è $\chi^2(N_c - 1 - r)$.*

Il teorema ci permette di calcolare una regione di rifiuto a livello α; infatti sotto l'ipotesi H_0 (cioè se H_0 è vera) il teorema garantisce che la legge di Q è (approssimativamente) una $\chi^2(N_c - 1 - r)$ (con r eventualmente pari a 0), pertanto la probabilità di entrare in una regione critica del tipo $Q > k$ (che sotto l'ipotesi H_0 è la probabilità di commettere un errore di Ia specie) è calcolabile esplicitamente. Più precisamente, il suo valore è α in corrispondenza di un preciso valore di k che si ricava come segue

$$\alpha = \mathbb{P}(Q > k) \iff 1 - \alpha = \mathbb{P}(Q \leq k) \iff k = \chi^2_{1-\alpha}(N_c - 1 - r).$$

Quindi la procedura decisionale del test d'ipotesi è:

si rifiuti H_0 *se* $Q > \chi^2_{1-\alpha}(N_c - 1)$ (oppure $Q > \chi^2_{1-\alpha}(N_c - 1 - r)$ se r è il numero di parametri stimati dai dati).

Alternativamente si calcola il P-value $\overline{\alpha}$ risolvendo l'equazione $Q = \chi^2_{1-\overline{\alpha}}(N_c - 1)$ (oppure $Q = \chi^2_{1-\overline{\alpha}}(N_c - 1 - r)$).

Il teorema è applicabile a condizione che le frequenze assolute attese verifichino $np_i \geq 5$ per ogni i, altrimenti la legge per Q non sarebbe ben approssimabile con la legge chi-quadrato. Se risultasse che $np_i < 5$ per qualche valore di i, allora bisognerebbe accorpare opportunamente alcune classi contigue, finché la condizione non è verificata.

Esempio 9.6.2. La legge ipotizzata per il tempo di vita in mesi di una lampadina è una legge esponenziale $X \sim Exp(0.33)$. Su un campione di 100 lampadine sono state riscontrate le seguenti durate:

	freq. oss. f_r	freq. ipotizzata p
$X \leq 1$	0.39	$1 - e^{-0.33} \simeq 0.281$
$1 < X \leq 2$	0.24	$e^{-0.33} - e^{-0.66} \simeq 0.202$
$2 < X \leq 3$	0.12	$e^{-0.66} - e^{-0.99} \simeq 0.145$
$3 < X \leq 5$	0.16	$e^{-0.99} - e^{-1.65} \simeq 0.180$
$5 < X \leq 10$ $(\equiv 5 < X)$	0.09	$e^{-1.65} \simeq 0.192$

Effettuiamo il test di adattamento per un livello di significatività del 10% (dove l'ultima classe diviene $\{5 < X\}$):

$$Q = 100 \left[\frac{(0.39 - 0.281)^2}{0.281} + \frac{(0.24 - 0.202)^2}{0.202} + \frac{(0.12 - 0.145)^2}{0.145} + \right.$$

$$\left. + \frac{(0.16 - 0.180)^2}{0.180} + \frac{(0.09 - 0.192)^2}{0.192} \right] \simeq 8.3219$$

mentre

$$\chi^2_{1-\alpha}(N_c - 1) = \chi^2_{0.9}(4) \simeq 7.7794$$

Risulta $Q > 7.7794$: l'ipotesi nulla viene rifiutata.
Il valore del P-value per il test è dato da:

$$\chi^2_{1-\overline{\alpha}}(N_c - 1) = Q \quad \Rightarrow \quad \overline{\alpha} \simeq 8.1\%$$

Per livelli di significatività minori dell'8.1% l'ipotesi nulla viene accettata.

Supponiamo adesso che la durata di vita delle lampadine segua una legge esponenziale di parametro incognito, e ricaviamo il valore di ν dal campione: $\nu = 99/(100 \cdot \overline{x}) \simeq 0.46$ (per la stima utilizziamo le classi originali del problema in modo da ottenere un risultato più preciso). Il nuovo valore di Q è pari a 0.6834. Lo dobbiamo confrontare con $\chi^2_{1-\alpha}(N_c - 1 - 1) = \chi^2_{0.9}(3) \simeq 6.2514$. Questa volta l'ipotesi nulla viene accettata. Si noti che la legge chi-quadrato possiede solo tre gradi di libertà, in quanto ν è stato stimato a partire dal campione e ciò toglie un grado di libertà. Il P-value è pari a 0.877, valore estremamente elevato: l'approssimazione esponenziale risulta ottima.

Osservazione 9.6.3. Il test di buon adattamento in realtà controlla solo che le frequenze sperimentali $\{N_i/n\}_{i=1}^{N_c}$ siano in buon accordo (o meno) con quelle teoriche $\{p_i\}$ (relative alle classi $\{X \in I_i\}$), pertanto qualsiasi altra variabile aleatoria Y per la quale $\mathbb{P}(Y \in I_i) = p_i \equiv \mathbb{P}(X \in I_i)$ otterrebbe lo stesso risultato tramite il test. Anche in questo caso la conclusione "forte" è "rifiutare H_0".

9.7 Test chi-quadrato di indipendenza

Questo test viene applicato al seguente problema: date n osservazioni congiunte di due variabili, ci si chiede se le due variabili sono indipendenti tra loro.
Il problema era stato affrontato per variabli numeriche nel primo capitolo mediante il calcolo del coefficiente di correlazione. Il metodo che esponiamo ora è alternativo e può essere applicato anche a variabili di tipo categorico.

Consideriamo il caso di due variabili X e Y associate alla medesima popolazione; effettuiamo un campionamento e raggruppiamo i dati in classi. Se le due variabili sono indipendenti, allora

$$\mathbb{P}(X \in A_i, Y \in B_j) = \mathbb{P}(X \in A_i)\mathbb{P}(Y \in B_j)$$

dove A_i sono le classi relative alla variabile X, e B_j quelle relative a Y.

La probabilità $\mathbb{P}(X \in A_i)$ può essere stimata con la frequenza marginale relativa $f_{rX}(i)$, e analogamente $\mathbb{P}(Y \in B_j) \simeq f_{rY}(j)$. L'ipotesi di indipendenza si traduce nella

$$f_r^{teor}(i, j) = f_{rX}(i) f_{rY}(j)$$

Mentre per le frequenze assolute:

$$f_a^{teor}(i, j) = f_{aX}(i) f_{aY}(j)/n$$

Costruiamo la statistica chi-quadrato calcolata dai dati:

$$Q = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \frac{(f_a(i,j) - f_{aX}(i)f_{aY}(j)/n)^2}{f_{aX}(i)f_{aY}(j)/n}$$

Per quanto abbiamo visto prima, per n sufficientemente grande e per una suddivisione in classi con almeno 5 elementi in ogni classe, la statistica Q è approssimabile con una legge chi-quadrato.

I gradi di libertà della legge chi-quadrato si calcolano in questo modo: Il numero di classi è in totale $N_c = N_1N_2$. I valori di $f_{rX}(i)$ per $i = 1, \dots, N_1 - 1$ sono stimati dal campione (l'ultimo valore, per $i = N_1$ viene ricavato dal fatto che deve essere $\sum_{i=1}^{N_1} f_{rX}(i) = 1$). Analogamente anche i valori di $f_{rY}(j)$ per $j = 1, \dots, N_2 - 1$ sono stimati dal campione. In totale i parametri stimati dal campione sono in numero $N_1 - 1 + N_2 - 1$. Dunque il numero dei gradi di libertà della legge chi-quadrato è

$$N_c - 1 - (N_1 - 1 + N_2 - 1) = N_1N_2 - N_1 - N_2 + 1 = (N_1 - 1)(N_2 - 1)$$

L'ipotesi nulla è

$$H_0\text{: } \textit{le variabili } X \textit{ e } Y \textit{ sono indipendenti tra loro.}$$

Il test sull'ipotesi di indipendenza è:

$$\textit{si rifiuti } H_0 \textit{ se } Q > \chi^2_{1-\alpha}((N_1 - 1)(N_2 - 1)).$$

Esercizio 9.7.1. A un campione di 150 persone è stato chiesto il colore e l'animale preferiti. I risultati sono presentati nella seguente tabella:

	rosso	blu	verde	giallo	totale
gatto	7	17	16	13	53
cane	8	28	22	9	67
cavallo	5	10	9	6	30
totale	20	55	47	28	150

Ci chiediamo se il colore preferito è indipendente dall'animale preferito, per un livello di significatività del 10%.

Soluzione.
Applichiamo la formula trovata sopra: si rifiuta l'ipotesi d'indipendenza se $Q > \chi^2_{1-\alpha}((N_1 - 1)(N_2 - 1))$.

$Q \simeq 3.2983$, mentre $\chi^2_{1-\alpha}((N_1-1)(N_2-1)) = \chi^2_{0.9}(6) \simeq 10.6446$: l'ipotesi nulla viene accettata, ossia si può concludere che le due variabili colore e animale sono indipendenti. Il P-value è dato da $\chi^2_{1-\overline{\alpha}}(6) = 3.2983$: $\overline{\alpha} \simeq 77.1\%$.

9.8 Verifica d'ipotesi sulla differenza tra due medie

Vogliamo confrontare le medie di due popolazioni diverse, estraendo un campione casuale da ciascuna.

Si consideri il caso di due popolazioni normali indipendenti: $X \sim \mathcal{N}(\mu_X, \sigma^2_X)$, $Y \sim \mathcal{N}(\mu_Y, \sigma^2_Y)$. Estraiamo dalla prima un campione casuale di ampiezza n_X, e dalla seconda una campione di ampiezza n_Y; n_X e n_Y non sono necessariamente uguali.

Vogliamo confrontare le medie delle due popolazioni; formuliamo a tale fine una delle seguenti ipotesi nulle:

$$H_0 : \mu_x - \mu_Y \geq \delta, \quad H_0 : \mu_X - \mu_Y = \delta, \quad H_0 : \mu_X - \mu_Y \neq \delta, \quad H_0 : \mu_x - \mu_Y \leq \delta$$

dove δ è un numero reale fissato.

I test di verifica delle ipotesi cambiano a seconda che le varianze siano note oppure incognite. Tratteremo due casi: quello in cui σ^2_X e σ^2_Y sono entrambe note, quello in cui σ^2_X e σ^2_Y sono

incognite ma uguali e quello in cui sono incognite ma non necessariamente uguali. Nella pratica, il caso in cui σ_X^2 e σ_Y^2 sono entrambe incognite si può ricondurre a quello in cui esse siano note, purché i campioni siano sufficientemente grandi ($n_X, n_Y > 30$), usando le varianze campionarie come se fossero i valori esatti delle varianze.

- Le varianze σ_X^2 e σ_Y^2 sono entrambe note. Si costruisce la statistica test

$$\overline{X}_{n_X} - \overline{Y}_{n_Y} - \delta \sim \mathcal{N}\left(\mu_X - \mu_Y - \delta, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

ossia

$$\frac{\overline{X}_{n_X} - \overline{Y}_{n_Y} - \delta - (\mu_X - \mu_Y - \delta)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0,1)$$

Le regole di decisione dei test si ricavano allo stesso modo di quelle trovate nel precedente capitolo sulla media di una popolazione gaussiana a varianza nota.

Riassumiamo i risultati nella tabella seguente:

H_0	H_1	Rifiutare H_0 se	P-value
$\mu_X - \mu_Y = \delta$	$\mu_X - \mu_Y \neq \delta$	$\|z\| > q_{1-\alpha/2}$	$2 - 2\Phi(\|z\|)$
$\mu_X - \mu_Y \leq \delta$	$\mu_X - \mu_Y > \delta$	$z > q_{1-\alpha}$	$\Phi(-z)$
$\mu_X - \mu_Y \geq \delta$	$\mu_X - \mu_Y < \delta$	$z < -q_{1-\alpha}$	$\Phi(z)$
$\mu_X - \mu_Y \neq \delta$	$\mu_X - \mu_Y = \delta$	$\|z\| < q_{(1+\alpha)/2}$	$2\Phi(\|z\|) - 1$

$$\text{dove } \; z = \frac{\overline{x}_{n_X} - \overline{y}_{n_Y} - \delta}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

- Il secondo caso è quello in cui le varianze sono entrambe incognite ma uguali: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Si considera la seguente statistica:

$$T = \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}$$

Dove S_X^2 e S_Y^2 sono le varianze campionarie dei due campioni. T ha una legge t di Student con $n_X + n_Y - 2$ gradi di libertà. Infatti:

$$\frac{(n_X - 1)S_X^2}{\sigma^2} \sim \chi^2(n_X - 1), \quad \frac{(n_Y - 1)S_Y^2}{\sigma^2} \sim \chi^2(n_Y - 1)$$

per la proprietà della legge χ^2 la loro somma è anch'essa una legge chi-quadrato, con $n_X + n_Y - 2$ gradi di libertà:

$$S^2 = \frac{(n_X - 1)S_X^2}{\sigma^2} + \frac{(n_Y - 1)S_Y^2}{\sigma^2} \sim \chi^2(n_X + n_Y - 2)$$

La variabile aleatoria

$$\frac{\overline{X}_{n_X} - \overline{Y}_{n_Y} - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)}}$$

ha legge $\mathcal{N}(0,1)$. Pertanto per definizione della legge t di Student:

$$\frac{\overline{X}_{n_X} - \overline{Y}_{n_Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} = \frac{\overline{X}_{n_X} - \overline{Y}_{n_Y} - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{\sigma^2}{n_X}+\frac{\sigma^2}{n_Y}\right)}} \Big/ \sqrt{S^2/(n_X+n_Y-2)} \sim t(n_X+n_Y-2)$$

Le regole di decisione dei test si ricavano allo stesso modo di quelle trovate nel precedente capitolo sulla media di una popolazione gaussiana a varianza incognita.

H_0	H_1	Rifiutare H_0 se	P-value
$\mu_X - \mu_Y = \delta$	$\mu_X - \mu_Y \neq \delta$	$\lvert t\rvert > t_{1-\alpha/2}(n)$	$t_{1-\overline{\alpha}/2}(n) = \lvert t\rvert$
$\mu_X - \mu_Y \leq \delta$	$\mu_X - \mu_Y > \delta$	$t > t_{1-\alpha}(n)$	$t_{1-\overline{\alpha}}(n) = t$
$\mu_X - \mu_Y \geq \delta$	$\mu_X - \mu_Y < \delta$	$t < -t_{1-\alpha}(n)$	$t_{1-\overline{\alpha}}(n) = -t$
$\mu_X - \mu_Y \neq \delta$	$\mu_X - \mu_Y = \delta$	$\lvert t\rvert < t(n)_{(1+\alpha)/2}$	$t_{(1+\overline{\alpha})/2}(n) = \lvert t\rvert$

$$\text{dove } \; t = \frac{\overline{x}_{n_X} - \overline{y}_{n_Y} - \delta}{\sqrt{\frac{(n_X-1)s_X^2 + (n_Y-1)s_Y^2}{n_X+n_Y-2}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}, \quad n = n_X + n_Y - 2$$

- L'ultimo caso che contempliamo è quello in cui entrambe le varianze sono incognite e non necessariamente uguali; in tal caso si potrebbe mostrare che lo stimatore da utilizzare è

$$\frac{\overline{X}_{n_X} - \overline{Y}_{n_Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n_X + S_Y^2/n_Y}} \sim t(\nu)$$

dove il grado di libertà dipende dal campionamento e vale

$$\nu = \frac{(s_X^2/n_X + s_Y^2/n_Y)^2}{(s_X^2/n_X)^2/(n_X-1) + (s_Y^2/n_Y)^2/(n_Y-1)}.$$

Pertanto

H_0	H_1	Rifiutare H_0 se	P-value
$\mu_X - \mu_Y = \delta$	$\mu_X - \mu_Y \neq \delta$	$\lvert t\rvert > t_{1-\alpha/2}(\nu)$	$t_{1-\overline{\alpha}/2}(\nu) = \lvert t\rvert$
$\mu_X - \mu_Y \leq \delta$	$\mu_X - \mu_Y > \delta$	$t > t_{1-\alpha}(\nu)$	$t_{1-\overline{\alpha}}(\nu) = t$
$\mu_X - \mu_Y \geq \delta$	$\mu_X - \mu_Y < \delta$	$t < -t_{1-\alpha}(\nu)$	$t_{1-\overline{\alpha}}(\nu) = -t$
$\mu_X - \mu_Y \neq \delta$	$\mu_X - \mu_Y = \delta$	$\lvert t\rvert < t(\nu)_{(1+\alpha)/2}$	$t_{(1+\overline{\alpha})/2}(\nu) = \lvert t\rvert$

$$\text{dove } \; t = \frac{\overline{x}_{n_X} - \overline{y}_{n_Y} - \delta}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}}$$

Esercizio 9.8.1. L'osservazione dei tempi di cui hanno bisogno i clienti di un ufficio postale per effettuare le loro operazioni ha dato i seguenti risultati: su 150 persone il tempo medio ad operazione allo sportello A è risultato pari a 85 secondi, con una deviazione standard campionaria di 15 secondi, mentre allo sportello B su 200 persone la media è stata di 81 secondi e deviazione standard 20 secondi.

Al livello di confidenza del 5% ci domandiamo se è plausibile che i clienti passino più tempo al primo sportello che al secondo.

Soluzione.
L'ipotesi nulla è

$$H_0\colon\ \mu_A \leq \mu_B$$

dove μ_A e μ_B sono i tempi medi passati rispettivamente agli sportelli A e B. Siamo nel caso di varianze incognite, non necessariamente uguali. Approssimiamo i valori delle varianze vere con quelli delle varianze campionarie essendo $n_A, n_B > 30$ e quindi utilizziamo il test per due medie a varianze note: $\sigma_A^2 = 15^2$, $\sigma_B^2 = 20^2$.

La regola di decisione del test è: *si rifiuti* H_0 *se* $z > q_{0.95}$, con

$$z = \frac{\overline{x}_A - \overline{y}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{85 - 81}{\sqrt{\frac{225}{150} + \frac{400}{200}}} \simeq 2.1381$$

Il quantile vale $q_{0.95} = 1.6449$. Pertanto l'ipotesi nulla viene rifiutata.
Calcoliamo infine il P-value: $\overline{\alpha} = \Phi(-2.1381) \simeq 1.6\%$.

Esercizio 9.8.2. In Lombardia negli ultimi 4 inverni sono stati registrati, durante i mesi di Novembre, Dicembre e Gennaio, i seguenti casi di meningite

99/00	13
00/01	15
01/02	20
02/03	18

Supponendo che il numero di abitanti sia rimasto sostanzialmente invariato in questi anni (e pari a 8.940.000) e tenendo come livello di riferimento la media dei primi 3 anni di monitoraggio, si vuole decidere se l'epidemia di quest'anno sia più preoccupante.

1. Si adotti il punto di vista *precauzionale* (non si vuole correre il rischio di sottovalutare l'epidemia): si formuli l'ipotesi nulla adeguata e si discuta la validità dell'ipotesi con un livello di significatività del 5%.

2. Si adotti il punto di vista *non allarmistico* (non si vuole sopravvalutare l'epidemia): si formuli l'ipotesi nulla adeguata, si calcoli il P-value e si discuta la validità dell'ipotesi.

3. Si ripeta il punto (1) utilizzando il confronto tra le medie del 2002 e del 2003 per decidere se quest'anno l' epidemia sia più virulenta.

4. Si studi la possibilità che la media dei primi 3 anni sia uguale a quella dell'ultimo anno supponendo che nell'ultimo anno non siano stati registrati 10 casi (e che quindi in totale siano 28).

Soluzione.

1. La media campionaria dei primi tre anni (assunta come vera) sul campione di ampiezza $n = 8940000$ è $p_0 := (13 + 15 + 20)/(3n) \approx 1.7897 \cdot 10^{-6}$, mentre $\overline{p} = 18/n \approx 2.0134 \cdot 10^{-6}$. Il test

$$H_0 : p \geq p_0$$
$$H_1 : p < p_0$$

ha come regione critica (o regione di rifiuto) e P-value

$$\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < q_\alpha$$
$$\overline{\alpha} = \phi\left(\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right).$$

Essendo $\overline{p} > p_0$ l'ipotesi nulla non può essere rifiutata a livelli inferiori a 0.5. In ogni caso si ha

$$\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx 0.5$$
$$q_{0.05} \approx -1.6449$$
$$\overline{\alpha} = \phi\left(\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right) \approx 0.6915.$$

2. Questa volta si tratta di studiare il test

$$H_0 : p \leq p_0$$
$$H_1 : p > p_0$$

ha come regione critica e P-value

$$\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > q_{1-\alpha}$$
$$\overline{\alpha} = 1 - \phi\left(\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right).$$

Eseguendo i calcoli si ha

$$q_{1-0.05} \approx 1.6449$$
$$\overline{\alpha} = \phi\left(\frac{\overline{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right) \approx 0.3185$$

pertanto ancora non si può rifiutare l'ipotesi nulla (equivalentemente utilizzando la regione critica o il P-value).

3. Si considerano $\overline{p}_1 = 18/n$ e $\overline{p}_2 = 20/n$ ed il test

$$H_0 : p_1 \geq p_2$$
$$H_1 : p_1 < p_2$$

ha come regione critica (o regione di rifiuto) e P-value

$$\frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n + \overline{p}_2(1-\overline{p}_2)/n}} < q_\alpha$$
$$\overline{\alpha} = \phi\left(\frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n + \overline{p}_2(1-\overline{p}_2)/n}}\right).$$

Eseguendo i calcoli si ha

$$\frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n + \overline{p}_2(1-\overline{p}_2)/n}} \approx \frac{-2}{\sqrt{18+20}} \approx -0.3244$$
$$\overline{\alpha} = \phi\left(\frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n + \overline{p}_2(1-\overline{p}_2)/n}}\right) \approx 0.3728$$

quindi non possiamo rifiutare l'ipotesi nulla al livello di significatività del 5%.

4. Si considerano $n_1 := n$, $n_2 := 3n$, $\overline{p}_1 = 28/n_1 \approx 3.132 \cdot 10^{-6}$ e $\overline{p}_2 = (20+13+15)/n_2 \approx 1.7897 \cdot 10^{-6}$ ed il test

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

ha come regione di accettazione e P-value

$$q_{\alpha/2} < \frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n_1 + \overline{p}_2(1-\overline{p}_2)/n_2}} < q_{1-\alpha/2}$$
$$\overline{\alpha} = 2\left(1 - \phi\left(\frac{|\overline{p}_1 - \overline{p}_2|}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n_1 + \overline{p}_2(1-\overline{p}_2)/n_2}}\right)\right).$$

Eseguendo i calcoli si ha

$$\frac{\overline{p}_1 - \overline{p}_2}{\sqrt{\overline{p}_1(1-\overline{p}_1)/n_1 + \overline{p}_2(1-\overline{p}_2)/n_2}} \approx 2.0785$$
$$q_{0.975} = -q_{0.025} \approx 1.9600$$
$$\overline{\alpha} \approx 0.0377$$

pertanto rifiutiamo l'ipotesi nulla al livello di significatività del 5%.

9.9 Verifica d'ipotesi per due variabili accoppiate

Sipponiamo di avere un campione di $X_1, \ldots, X_n$ variabili che rappresentano la misurazione di una certa grandezza su soggetti differenti (non è richiesta indipendenza o identica distribuzione). Supponiamo ora di operare sui soggetti una procedura che introduca in maniera additiva un disturbo W_i di legge (μ, σ^2) (o di legge qualsiasi se n è sufficientemente grande). Se ora rimisuriamo per gli stessi soggetti la grandezza in questione ci aspettiamo di osservare una variabile $Y_i = X_i + W_i$; nell'ipotesi di indipendenza (ed identica distribuzione per $W_1, \ldots, W_n$ potremo testare μ.

L'esempio classico potrebbe essere l'efficacia di un farmaco per abbassare la pressione arteriosa dei pazienti; per accertare se il farmaco è efficace si può rilevare la pressione di n pazienti prima e dopo la somministrazione e vedere se "mediamente" la pressione arteriosa è scesa.

Analogamente potremmo pensare di dover accertare l'efficacia di un nuovo tipo di marmitta per auto; la procedura potrebbe quindi consistere nel misurare le emissioni di alcune sostanze prima e dopo l'utilizzo della nuova marmitta.

Definiamo quindi $\overline{W}_n := \overline{X}_n - \overline{Y}_n$ e supponiamo di voler testare una delle seguenti ipotesi nulle

$$H_0 : \mu_x - \mu_Y \geq \delta, \quad H_0 : \mu_X - \mu_Y = \delta, \quad H_0 : \mu_X - \mu_Y \neq \delta, \quad H_0 : \mu_x - \mu_Y \leq \delta$$

contro le rispettive alternative; questo equivale a voler testare rispettivamente

$$H_0 : \mu_W \geq \delta, \quad H_0 : \mu_W = \delta, \quad H_0 : \mu_W \neq \delta, \quad H_0 : \mu_W \leq \delta$$

a varianza incognita. Utilizziamo quindi il seguente stimatore

$$T := \frac{\overline{X}_n - \overline{Y}_n - \delta}{S_n/\sqrt{n}} \equiv \frac{\overline{W}_n - \delta}{S_n/\sqrt{n}}$$

dove

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

I risultati sono riassunti nella tabella seguente.

H_0	H_1	Rifiutare H_0 se	P-value
$\mu_X - \mu_Y = \delta$	$\mu_X - \mu_Y \neq \delta$	$\lvert t\rvert > t_{1-\alpha/2}(n-1)$	$t_{1-\overline{\alpha}/2}(n-1) = \lvert t\rvert$
$\mu_X - \mu_Y \leq \delta$	$\mu_X - \mu_Y > \delta$	$t > t_{1-\alpha}(n-1)$	$t_{1-\overline{\alpha}}(n-1) = t$
$\mu_X - \mu_Y \geq \delta$	$\mu_X - \mu_Y < \delta$	$t < -t_{1-\alpha}(n-1)$	$t_{1-\overline{\alpha}}(n-1) = -t$
$\mu_X - \mu_Y \neq \delta$	$\mu_X - \mu_Y = \delta$	$\lvert t\rvert < t(n-1)_{(1+\alpha)/2}$	$t_{(1+\overline{\alpha})/2}(n-1) = \lvert t\rvert$

$$\text{dove } \; t = \frac{\overline{x}_n - \overline{y}_n - \delta}{s_n/\sqrt{n}} \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - y_i - (\overline{x}_n - \overline{y}_n))^2.$$

Esercizio 9.9.1. Un'industria ha messo a punto un programma per ridurre le ore perse per incidenti sul lavoro. Si sono calcolate, per ciascuno dei 10 stabilimenti simili le ore perse (siano X_i) per gli incidenti nel mese antecedente all'introduzione della procedura e nel mese successivo (siano Y_i). Sapendo che $\sum_{i=1}^{10}(Y_i - X_i) = -21.5$ mentre $\sum_{i=1}^{10}(Y_i - X_i)^2 = 127.25$ si può ritenere davvero efficace il programma al 5%? Si può ritenere che la media delle ore perse si sia ridotta di 2 ore al 5%?

Soluzione.

Per stabilire l'efficacia del programma tramite una conclusione forte (cioè per avere una forte evidenza della sua efficacia) scegliamo $H_0 : \mu_Y \geq \mu_X$ e $H_1 : \mu_Y < \mu_X$.

Se $Z_i := Y_i - X_i$ allora $\overline{Z}_{10} = -2.15$ mentre $S_{10}^2 = (127.25 - 10 \cdot (-2.15)^2)/9 \approx 9.003$. Utilizziamo la seguente regione di rifiuto a livello α

$$T = \frac{\overline{Z}_{10} - \delta}{S_{10}/\sqrt{n}} < t_\alpha(9)$$

dove la stima di T è $t = -2.266$ (essendo $\delta = 0$). Dalle tavole $t_{0.05}(9) = -t_{0.95}(9) = -1.833114$ quindi rifiuto H_0 al livello 0.05.

Alternativamente, stimiamo il P-value come segue: $-2.266 = t_{\overline{\alpha}}(9) \equiv -t_{1-\overline{\alpha}}(9)$ da cui $2.266 = t_{1-\overline{\alpha}}(9)$. Essendo

$$t_{0.975}(9) = 2.262159 < 2.266 < 2.821434 = t_{0.99}(9)$$

si ha $t_{0.975}(9) < t_{1-\overline{\alpha}}(9) < t_{0.99}(9)$ da cui $0.025 < \overline{\alpha} < 0.01$ (anche se di fatto $t_{0.975}(9) \approx t_{1-\overline{\alpha}}(9)$ da cui $0.025 \approx \overline{\alpha}$).

Per la seconda parte si utilizza la seguente regione di rifiuto a livello α

$$T = \frac{\overline{Z}_{10} - \delta}{S_{10}/\sqrt{n}} < t_\alpha(9)$$

dove stavolta la stima di T è $t = -0.1581$ (essendo $\delta = 2$) che non ci consente di rifiutare H_0 a livello 0.05. Similmente al caso precedente si ottiene una stima per il P-value $\overline{\alpha} > 0.2$.

9.10 Test sulla regressione lineare

In questo paragrafo studieremo l'affidabilità dei parametri di una regressione lineare (si vedano i Paragrafi 2.4.4 e 2.4.5 per i dettagli) e calcoleremo gli intervalli di confidenza per i coefficienti della regressione stessa e per le previsioni del nostro modello lineare.

Assumiamo la seguente relazione tra i regressori $X_1, \ldots, X_k$ e la variabile Y

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

dove ϵ è una variabile $\mathcal{N}(0, \sigma^2)$; pertanto se consideriamo un campione di ampiezza n avremo

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_k x_i^{(k)} + \epsilon_i, \qquad \forall i = 1, 2, \ldots, n$$

oppure, in forma vettoriale, $y = X\beta$ dove la matrice X sarà definita nel prossimo paragrafo.

Nel Paragrafo 2.4.5 abbiamo visto come stimare il coefficienti della regressione utilizzando degli stimatori $\widehat{\beta}_j$ (dove $j = 1, \ldots, k$), qui approfondiremo lo studio stimando la varianza σ^2 comune a tutti gli errori ϵ_i (che supporremo indipendenti).

Successivamente introdurremo dei test in grado di giudicare l'affidabilit'a della regressione e dei singoli coefficienti. In questo modo saremo in grado di capire se esiste almeno un regressore X_i significativo (con β_i significativamente diverso da 0) e quali siano invece i regressori trascurabili.

Ci occuperemo infine delle previsioni, cioè di dare un intervallo di confidenza per la **risposta del sistema** (il valore di Y) in corrispondenza ad una k-upla $(x_1, \ldots, x_k)$; tale risposta dovrà tenere conto dell'errore sulla stima dei coefficienti e del suo errore intrinseco ϵ di misurazione. Per questo motivo introduciamo anche il concetto di **risposta media del sistema**, cioè il valore atteso di Y condizionato alla k-upla $(x_1, \ldots, x_k)$; questo valore atteso coinvolge solo l'errore ϵ, pertanto l'intervallo di confidenza per questa grandezza dipende solo dall'incertezza con cui conosciamo i coefficienti della regressione.

Ricordiamo alcune grandezze introdotte nei Paragrafi 2.4.4 e 2.4.5. Abbiamo definito
la *devianza totale* $DT = \sum_{i=1}^n (y_i - \overline{y})^2$,
la *devianza spiegata* $DS = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$ dove $\hat{y}_i$ sono i valori previsti (o più precisamente i valori attesi delle previsioni) $\hat{y}_i = \beta_0 + \sum_{j=1}^k \beta_j x_i^{(j)}$),
ed infine la *devianza dei residui* $DR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$;
vale inoltre la relazione $DT = DS + DR$. In alcuni testi si utilizzano le abbreviazioni SS_y o SST, SSR e SSE rispettivamente per DT, DS e DR.

Definiamo i tre gradi di libertà delle devianze $f_{DT} := n - 1$, $f_{DS} := k$ e $f_{DR} := f_{DT} - f_{DS} = n - (k + 1)$; ovviamente ha senso fare la regressione solo in presenza di un numero sufficiente di dati ($n \geq k + 1$).

9.10.1 Analisi della varianza

Ricordiamo le definizioni date in precedenza

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

$$s_y^2 := \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^n y_i^2 - \frac{n}{n-1}\bar{y}^2 = \frac{1}{n-1}(\sum_{i=1}^n y_i^2 - \frac{1}{n}(\sum_{i=1}^n y_i)^2)$$

$$s_{xy} := \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^n x_i y_i - \frac{n}{n-1}\bar{x}\bar{y} = \frac{1}{n-1}(\sum_{i=1}^n x_i y_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)).$$

Supponendo che gli errori ϵ_i siano indipendenti con media 0 e varianza σ^2, ricordando che lo stimatore (vettoriale) dei coefficienti della regressione è $\widehat{\beta} := (X^T X)^{-1} X^T y$ dove

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(k)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(k)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(k)} \end{bmatrix}$$

si ottiene immediatamente che

$$\mathbb{E}(\widehat{\beta}) = \mathbb{E}((X^T X)^{-1} X^T y) = \mathbb{E}((X^T X)^{-1} X^T (X\beta + \epsilon)) = \beta$$

dove ϵ è il vettore aleatorio di componenti ϵ_i. Quindi lo stimatore è corretto. Similmente si calcola la varianza dei coefficienti in termini della matrice simmetrica

$$C := (X^T X)^{-1}$$

dove, per convenzione, chiameremo $C_{i,j}$ l'elemento della riga $i+1$ e colonna $j+1$ (con $i,j = 0,\ldots,k$); precisamente la matrice di covarianza di $\widehat{\beta}$ è $\text{cov}(\widehat{\beta}) = \sigma^2 C$ cioè $\text{cov}(\widehat{\beta}_i, \widehat{\beta}_j) = \sigma^2 C_{i,j}$, pertanto in particolare $\text{Var}(\widehat{\beta}_i) = \sigma^2 C_{i,i}$. Si può dimostrare che uno stimatore corretto per σ^2 è

$$\widehat{\sigma}^2 = \frac{DR}{n-k-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k-1}.$$

Attraverso l'uso di questo stimatore definiamo la stima dell'errore standard $\text{se}(\widehat{\beta}_i)$ del coefficiente $\widehat{\beta}_i$ come

$$\text{se}(\widehat{\beta}_i) := \sqrt{\widehat{\sigma}^2 C_{i,i}}.$$

Nel caso di regressione semplice le formule si semplificano nel seguente modo

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{(n-1)s_x^2} & -\frac{\bar{x}}{(n-1)s_x^2} \\ -\frac{\bar{x}}{(n-1)s_x^2} & \frac{\sum_{i=1}^n x_i^2}{n(n-1)s_x^2} \end{bmatrix}$$

da cui

$$\text{se}(\widehat{\beta}_1) = \sqrt{\frac{\widehat{\sigma}^2}{(n-1)s_x^2}} = \sqrt{\frac{DR}{(n-2)(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}}$$
$$\text{se}(\widehat{\beta}_0) = \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)} = \sqrt{\frac{DR}{n(n-2)} \frac{\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)}}.$$

9.10.2 Intervalli di confidenza per i coefficienti della regressione

Sotto le ipotesi che gli errori ϵ_i siano indipendenti ed identicamente distribuiti con distribuzione $\mathcal{N}(0,\sigma^2)$ si ha che le osservazioni $Y_1,\ldots,Y_n$ sono anch'esse variabili normali, i.i.d. con media $\beta_0 + \sum_{j=1}^k x_i^{(j)} \beta_j$ e varianza σ^2. Da un noto teorema di teoria delle probabilità (chi fosse interessato veda i vettori aleatori ed in particolare i vettori gaussiani) si ha che le variabili $\widehat{\beta}_i$ sono normali con media β_i e varianza $\sigma^2 C_{i,i}$. Si mostra che le statistiche

$$\frac{\widehat{\beta}_i - \beta_i}{\sqrt{\widehat{\sigma}^2 C_{i,i}}}, \qquad i = 0,1,\ldots,k$$

sono distribuite come una t-student con $n-k-1$ gradi di libertà. Pertanto l'intervallo di confidenza bilatero a livello α per il coefficiente β_i è

$$\widehat{\beta}_i - t_{((1+\alpha)/2,n-k-1} \sqrt{\widehat{\sigma}^2 C_{i,i}} \le \beta_i \le \widehat{\beta}_i + t_{(1+\alpha)/2,n-k-1} \sqrt{\widehat{\sigma}^2 C_{i,i}}.$$

Nel caso di regressione semplice gli intervalli sono

$$\widehat{\beta}_1 - t_{(1+\alpha)/2,n-k-1} \sqrt{\frac{\widehat{\sigma}^2}{(n-1)s_x^2}} \le \beta_1 \le \widehat{\beta}_1 + t_{(1+\alpha)/2,n-k-1} \sqrt{\frac{\widehat{\sigma}^2}{(n-1)s_x^2}}$$
$$\widehat{\beta}_0 - t_{(1+\alpha)/2,n-k-1} \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)} \le \beta_0 \le \widehat{\beta}_0 + t_{(1+\alpha)/2,n-k-1} \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)}.$$

9.10.3 Test sui coefficienti della regressione

Il primo test che affrontiamo è quello di **significatività della regressione** le cui ipotesi sono

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \exists i \in \{1,2,\ldots,k\} \text{ tale che } \beta_i \neq 0.$$

In caso H_0 non possa essere rifiutata dovremo rigettare la regressione come non significativa.

La statistica corretta con cui eseguire il test si dimostra essere

$$F := \frac{DS/f_{DS}}{DR/f_{DR}} = \frac{DS/k}{DR/(n-k-1)} = \frac{MS_S}{MS_R}$$

dove $MS_S = DS/k$ e $MS_R = DR/(n-k-1)$ sono rispettivamente l'errore medio spiegato e l'errore medio residuo.

Il criterio di rifiuto o regione critica del test al livello di significatività α è

$$F > F_{1-\alpha,k,n-k-1}$$

dove $F_{\alpha,i,j}$ è il quantile della **distribuzione F** con i e j gradi di libertà. Al solito il P-value $\overline{\alpha}$ soddisfa l'equazione $F = 1 - F_{1-\overline{\alpha},k,n-k-1}$.

Il secondo test che analizziamo serve a comprendere se un regressore sia significativo (i.e. ha coefficiente differente da 0) oppure no. Pertanto formalizziamo il seguente test:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

cioè supponiamo il regressore non significativo fino a prova contraria. La statistica test appropriata è

$$T := \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 C_{j,j}}} = \frac{\widehat{\beta}_j}{\mathrm{se}(\widehat{\beta}_j)}$$

con regione di rifiuto a livello α

$$t > t_{1-\alpha/2,n-k-1} \text{ o } t < -t_{1-\alpha/2,n-k-1}$$

Se h_0 non può essere rifiutata allora il regressore X_j può essere cancellato dal modello. Si noti che, a rigore, questo test controlla la significatività del regressore X_j quando tutti gli altri regressori sono presenti; infatti il test dipende da tutti i coefficienti. Quindi quello che stiamo confrontanto è un modello con tutti i regressori compreso X_j contro un modello con tutti i regressori tranne X_j.

9.10.4 Intervalli di confidenza per una previsione

Supponiamo di voler stimare la risposta del sistema in corrispondenza al valore

$$\begin{pmatrix} x_0^{(0)} \\ x_0^{(1)} \\ \vdots \\ x_0^{(k)} \end{pmatrix}$$

sia pertanto

$$x_0 := \begin{pmatrix} 1 \\ x_0^{(0)} \\ x_0^{(1)} \\ \vdots \\ x_0^{(k)} \end{pmatrix}$$

Sappiamo che la risposta del sistema sarà $y(x_0) := x_0^T\beta + \epsilon$ mentre la risposta media sarà $y_m(x_0) := x_0^T\beta$ (il termine "media" si riferisce solo all'errore ϵ che è l'unico variabile in gioco in questo momento).

Tuttavia noi non conosciamo i veri valori dei coefficienti del vettore β, ma solo le stime

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}$$

pertanto è naturale attendersi di poter calcolare un intervallo di confidenza per le due risposte; l'intervallo per la risposta, in particolare, dovrà tener conto della presenza dell'errore di misurazione ϵ.

Cominciamo dall'intervallo per la risposta media $y_m(x_0)$; lo stimatore naturale è $\widehat{y}_m(x_0) := x_0^T\widehat{\beta}$ che ha media $x_0^T\beta$ e varianza $\sigma^2 x_0^T(X^TX)x_0$. Si potrebbe mostrare che l'intervallo di confidenza a livello α è

$$\widehat{y}_m(x_0) - t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2 x_0^T(X^TX)x_0} \leq y_m(x_0) \leq \widehat{y}_m(x_0) + t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2 x_0^T(X^TX)x_0}.$$

Nel caso di regressione semplice l'intervallo si riduce a

$$\widehat{y}_m(x_0) - t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0^{(1)} - \overline{x})^2}{(n-1)s_x^2}\right)} \leq y_m(x_0) \leq \widehat{y}_m(x_0) + t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0^{(1)} - \overline{x})^2}{(n-1)s_x^2}\right)},$$

dove $\widehat{y}_m(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0^{(1)}$.

La stima puntuale per la risposta $y(x_0)$ è $\widehat{y}(x_0) := x_0^T\widehat{\beta}$, mentre si può mostrare che l'intervallo a livello α è

$$\widehat{y}(x_0) - t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2(1 + x_0^T(X^TX)x_0)} \leq y(x_0) \leq \widehat{y}(x_0) + t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2(1 + x_0^T(X^TX)x_0)}.$$

Nel caso di regressione semplice la formula precedente si riduce a

$$\widehat{y}(x_0) - t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0^{(1)} - \overline{x})^2}{(n-1)s_x^2}\right)} \leq y(x_0) \leq \widehat{y}(x_0) + t_{(1+\alpha)/2,n-k-1}\sqrt{\widehat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0^{(1)} - \overline{x})^2}{(n-1)s_x^2}\right)},$$

dove $\widehat{y}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0^{(1)}$.