

Appunti di probabilità e statistica

Una approccio da studenti per studenti

Rafael Mosca

Scritto con **LATEX**

Questo libro è stato realizzato in base agli argomenti del corso di Probabilità e Statistica impartito da Egidio Battistini nel corso di studi (2016-2017).

First release, September 2017

Quest'opera è stata rilasciata con licenza Creative Commons Attribuzione:
- Non commerciale
- Non opere derivate 4.0 Internazionale.

This work is licensed under a Creative Commons “Attribution-NonCommercial-NoDerivatives 4.0 International” license.



Per leggere una copia della licenza visita il sito web <https://creativecommons.org/licenses/by-nc-nd/4.0/> o spedisci una lettera a Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.



Indice

1	Calcolo Combinatorio	9
1.1	Permutazioni	9
1.1.1	Permutazioni semplici	9
1.1.2	Permutazioni con ripetizioni	10
1.2	Disposizioni e combinazioni	10
1.2.1	Disposizioni semplici	10
1.2.2	Combinazioni semplici	11
1.2.3	Disposizioni con ripetizioni	11
1.2.4	Combinazioni con ripetizioni	12
1.2.5	Uso nel calcolo delle probabilità	12
2	Statistica	13
2.1	Statistica descrittiva	13
2.1.1	Indici Riassuntivi	14
2.1.2	Istogramma	15
3	Probabilità	17
3.1	Definizioni	18
3.1.1	Definizione classica	18
3.1.2	Definizione frequentista	18
3.1.3	Definizione soggettivista	18
3.1.4	Definizione assiomatica	18
3.2	Teoremi fondamentali della probabilità	19

3.3	Probabilità condizionata	20
3.3.1	Teorema di Bayes	21
3.3.2	Problema di Monty-Hall	23
3.3.3	Filtro bayesiano	24
4	Variabili aleatorie	25
4.1	Indipendenza di variabili aleatorie	26
4.2	Funzioni di variabili aleatorie	27
4.2.1	Formula di Convoluzione	28
4.3	Indici caratteristici di una variabile aleatoria	31
4.3.1	Proprietà degli indici caratteristici	32
4.4	Disuguaglianza di Cebicev	33
5	Modelli probabilistici discreti	35
5.1	Bernoulliana	35
5.2	Binomiale	37
5.2.1	Valore atteso	37
5.2.2	Varianza	37
5.2.3	Indice di asimmetria	38
5.3	Geometrica	40
5.3.1	Valore atteso	41
5.3.2	Varianza	41
5.4	Binomiale negativa	43
5.4.1	Valore Atteso	43
5.4.2	Varianza	44
5.5	Proprietà del modello bernoulliano	44
5.6	Uniforme discreta	45
5.6.1	Origini	45
5.6.2	Valore atteso	45
5.6.3	Varianza	46
5.7	Campionamento e modelli statistici	47
5.7.1	Somma e media campionaria	47
5.7.2	Legge debole dei grandi numeri / convergenza in probabilità	48
6	Modelli probabilistici continui	49
6.0.1	Aprossimazione di Stirling	50
6.1	Campo Poissoniano	50
6.2	Assiomi di Poisson	50
6.3	Poissoniana	51
6.3.1	Valore atteso	51

6.3.2	Varianza	52
6.4	Variabili aleatorie continue	52
6.4.1	PDF	52
6.4.2	Correzione di continuità	53
6.4.3	CDF	53
6.5	Distribuzioni particolari	55
6.5.1	Legge di trasformazione	55
6.6	Variabili assolutamente continue e continue singolari	56
6.7	Esponenziale	56
6.7.1	Valore atteso	57
6.7.2	Varianza	57
6.7.3	Proprietà di assenza di memoria	58
6.8	Laplace	59
6.9	Weibull	59
6.10	Rayleigh	60
6.11	Maxwell	60
6.12	Gamma	61
6.12.1	Valore atteso	61
6.12.2	Varianza	61
6.12.3	Proprietà di chiusura	62
6.13	Standardizzata	63
6.14	Teorema centrale del limite	63
6.14.1	Convergenza in legge	64
6.15	Normale	65
6.15.1	Proprietà della normale	66
6.15.2	Curtosi	67
6.15.3	Momenti	67
6.15.4	Chiusura rispetto alla somma	67
6.15.5	Somma di due distribuzioni normali	67
6.16	Chi-Quadro	68
6.17	Student - t	69
6.18	Uniforme continua	70
6.18.1	Origini	70
6.18.2	Valore atteso	71
6.18.3	Varianza	71
6.18.4	Proprietà di riscrittura	71
6.18.5	Somma di uniformi	72

6.19	Generare uniformi	72
6.20	Da uniformi a una variabile qualunque	72
6.21	Massimo	73
6.21.1	"Contatore" (CDF) e funzione densità	73
6.22	Minimo	74
6.22.1	"Contatore" (CDF) e funzione densità (PDF)	74
6.23	R-esima statistica d'ordine	76
6.23.1	Statistica "a metà" di uniformi	76
6.24	Range	77
7	Statistica inferenziale	79
7.1	Statistiche e proprietà degli stimatori	79
7.1.1	Condizione sufficiente per la non correttezza	80
7.1.2	Condizione sufficiente per la consistenza	81
7.2	Stima puntuale dei parametri	82
7.2.1	Efficienza degli stimatori	84
7.3	Dettagli su gli stimatori della varianza	84
7.4	Intervalli di confidenza	85
7.4.1	Intervalli esatti per il valore atteso	86
7.4.2	Intervalli asintotici per il valore atteso	86
7.4.3	Intervalli di confidenza per un parametro	89
7.4.4	Intervalli di confidenza per la varianza	90
7.5	Tavole	91
7.6	Test di verifica d'ipotesi	95
7.6.1	Complementi	96
7.7	Test parametrici per il valore atteso	96
7.7.1	Test Z bilatero	97
7.7.2	Test T bilatero	98
7.7.3	Test Z e Test T unilateri	99
7.8	Test parametrici per la varianza	100
7.8.1	Test bilatero	100
7.8.2	Test unilatero	100
7.9	Ulteriori tipi di test parametrici	101
7.9.1	Test con entrambe ipotesi composte	101
7.9.2	Test con entrambe ipotesi semplici	101
7.10	Test Non-Parametrici	102
7.10.1	Test χ^2 di buon adattamento	102
7.10.2	Test di Kolmogorov-Smirnov	105
7.10.3	Test di normalità	107
7.10.4	Test di poissonianità	109

7.11	Test di casualità	110
7.12	Significatività statistica: p-value	111
8	Vettori aleatori	113
8.1	Funzione generatrice dei momenti	113
8.1.1	MGF Discrete	114
8.1.2	MGF continue	115
8.1.3	Non esistenza MGF	116
8.1.4	Dimostrazioni con MGF	117
8.2	Funzione caratteristica	118
8.3	Vettori aleatori	118
8.3.1	PDF e CDF	118
8.3.2	Momenti	121
8.3.3	Varianza e Covarianza	121
8.3.4	Coefficiente di correlazione	123
8.4	Normale bivariata e multivariata	124
8.5	Normale bivariata marginale e condizionata	127
8.5.1	Normale bivariata marginale	127
8.5.2	Normale bivariata condizionata	128
8.5.3	Chiusura della Normale bivariata	129
8.6	Funzione generatrice dei momenti per vettori	129
8.7	Teorema di Cochran	130
9	Regressione	133
9.1	Ipotesi del modello lineare	135
9.2	Metodo dei minimi quadrati	135
9.3	Verifica delle proprietà degli stimatori	137
9.4	Intervalli di confidenza sui parametri	139
9.5	Decomposizione della devianza	140
9.6	Coefficiente di determinazione	141
9.7	Rappresentazione vettoriale della regressione	142
9.8	Trasformazioni del modello lineare	146
9.8.1	Lineare-Logaritmica	146
9.8.2	Potenza	146
9.8.3	Esponenziale	146
9.8.4	Iperbolica	146
9.9	Test globali sui predittori	147
9.9.1	Test F	147
9.9.2	Test del modello ridotto	148

9.10	Test su singoli parametri	149
9.11	Intervalli di previsione	150
9.12	Distribuzione di Fisher	151
9.13	Analisi della varianza	154
9.13.1	Analisi della varianza a un fattore	154

1. Calcolo Combinatorio

Il calcolo combinatorio è quella branca della matematica che si occupa di capire in quanti e quali modi si possono disporre gli oggetti di un insieme.

1.1 Permutazioni

Una *permutazione* è un modo di presentare ordinatamente un insieme di n oggetti distinti, come quando si realizza l'anagramma di una parola.

1.1.1 Permutazioni semplici

Si parla di *permutazioni semplici* quando gli oggetti vengono presentati una ed una sola volta, ragione per cui prendono anche il nome di permutazioni senza ripetizioni.

Per calcolare quante siano le permutazioni possibili di n oggetti, basta osservare che il primo elemento di una possibile permutazione può essere scelto in n modi diversi, il secondo in $(n - 1)$ modi, e così via fino ad arrivare all'ultimo oggetto. Il numero delle permutazioni di n oggetti si indica con P_n ed è pari al fattoriale di n .

$$P_n = n \cdot (n - 1) \cdot (n - 2) \cdots 1 = n! \quad (1.1)$$

ESEMPIO. Se si prende, ad esempio tre palline di colori diversi e tre caselle; il numero di permutazioni semplici ottenibili dalle tre palline sono $3! = 3 \cdot 2 \cdot 1 = 6$.



1.1.2 Permutazioni con ripetizioni

Molte volte un insieme può contenere elementi che si ripetono, per cui, di tutte le permutazioni ottenibili ci saranno alcune permutazioni che risulteranno uguali. Per calcolare quante siano le permutazioni diverse di n oggetti con l'elemento n_i che si ripete k_i volte, basta osservare che ad esempio, se un elemento si ripete due volte, allo scambio degli elementi, si ha sempre la stessa configurazione.

Il numero di permutazioni, è dato da:

$$P_n^* = \frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_i!} \quad (1.2)$$

e prende il nome di *coefficiente multinomiale* (poiché estensione del coefficiente binomiale).

$$\binom{n}{\mathbf{k}} = \binom{n}{k_1, \dots, k_i} = \frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_i!} \quad (1.3)$$

ESEMPIO. Se supponiamo di avere tre palline, di cui due sono dello stesso colore; il numero di permutazioni uniche ottenibili dalle 3 palline sono $3!/2! = 3$, com'è evidenziato dalla figura sottostante.



1.2 Disposizioni e combinazioni

Una *disposizione* è un sequenza ordinata di k elementi presa da un insieme di n oggetti, con $k \leq n$.

1.2.1 Disposizioni semplici

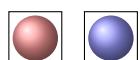
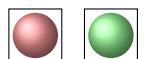
Quando si realizza una disposizione su un insieme in cui tutti gli oggetti sono distinti tra loro si parla di *disposizione semplice* di n elementi di classe k oppure di n elementi presi k alla volta.

Il numero di disposizioni semplici si trova come:

$$D_{n,k} = \frac{n!}{(n-k)!} = n \cdot (n-1) \cdot \dots \cdot (n-k+1) \quad (1.4)$$

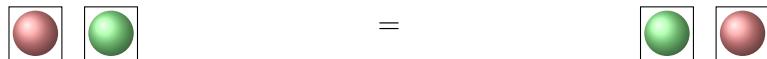
Inoltre, si può notare facilmente che quando $k = n$, il numero di disposizioni semplici coincide con il numero di permutazioni semplici.

ESEMPIO. Supponiamo di avere sempre tre palline di colori diversi ($n = 3$), e due caselle ove collocarle ($k = 2$); il numero di disposizioni semplici è $3!/(3-2)! = 6$



1.2.2 Combinazioni semplici

Notiamo che se nell'esempio precedente l'ordine non interessa, allora:



Definiamo *combinazioni semplici* come disposizioni semplici in cui l'ordine non ha alcuna importanza. Quindi si parla di combinazioni semplici quando si vuole raggruppare k elementi di un insieme di cardinalità n in modo che il raggruppamento contenga tutti elementi distinti fra loro e indipendente dall'ordine.

Il numero di combinazioni semplici ottenibili è dato da:

$$C_{n,k} = \frac{D_{n,k}}{P_k} = \binom{n}{k} = \frac{n!}{(n-k)!k!} \quad (1.5)$$

Quindi le combinazioni semplici ottenibili nel esempio precedente sono $3!/2!(3-2)! = 3$.

Se consideriamo lo stesso esempio ma con una sola casella, cioè $k = 1$, le combinazioni semplici ottenibili sono $3!/1!(3-1)! = 3$, da questo esempio possiamo vedere che le combinazioni semplici godono di un'interessante proprietà giacché vale:

$$\binom{n}{k} = \binom{n}{k-1} \quad (1.6)$$

Cioè il numero delle combinazioni semplici ottenibili con $k = 1$ e $k = 2$ rimane invariato. Ma perché? Sembrebbe illogico pensare che scegliere due oggetti su tre sia la stessa cosa di sceglierne uno su tre, ma in realtà ha molto senso, infatti sceglierne due su tre è come sceglierne uno dei tre da scartare.

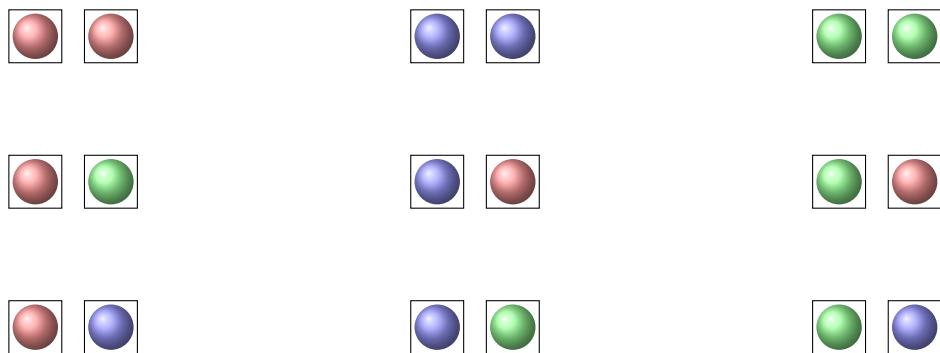
1.2.3 Disposizioni con ripetizioni

Le *disposizioni con ripetizioni* sono disposizioni in cui ogni elemento può essere scelto più di una volta, come se dopo aver fatto l'estrazione di una pallina da una sacca la rimettesse di nuovo dentro la sacca, con la possibilità che venga ripescata alla prossima estrazione.

Il numero di disposizioni con ripetizioni si ottiene come:

$$D_{n,k}^* = n^k \quad (1.7)$$

ESEMPIO. Considerando tre palline di colori diversi e due caselle, assumendo che si possano ripetere, il numero di disposizioni sono $3^2 = 9$, che sono quelle mostrate in figura:

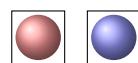
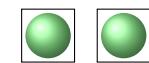
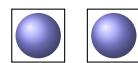
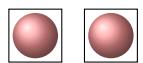


1.2.4 Combinazioni con ripetizioni

Le *combinazioni con ripetizioni* sono disposizioni in cui non conta l'ordine e ogni elemento può essere scelto più di una volta. Il numero di combinazioni con ripetizioni è dato da:

$$C_{n,k}^* = C_{n+k-1,k} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!k!} \quad (1.8)$$

ESEMPIO. Considerando le solite tre palline di colori distinti si hanno $4!/2!2! = 6$ combinazioni che sono mostrate in figura.



1.2.5 Uso nel calcolo delle probabilità

ESEMPIO. Prendiamo in considerazione una famiglia che ha 3 figli, vogliamo calcolare qual è la probabilità che la famiglia abbia 3 figli maschi.

$$P(3 m) = \frac{C_{3,3}}{D_{2,3}^*} = \frac{1}{2^3} = \frac{1}{8} = 12.5\% \quad (1.9)$$

Come possiamo vedere dalla figura i casi possibili sono otto e i casi favorevoli, soltanto uno.



Si noti che se si calcolasse la probabilità di avere due maschi, i casi favorevoli sarebbero tre ($C_{3,2} = 3$).

2. Statistica

Le statistiche sono come i bikini. Ciò che rivelano è suggestivo, ma ciò che nascondono è più importante.
(AARON LEVENSTEIN)

La statistica è strettamente legata alla probabilità nello studio dei fenomeni aleatori, ma ciascuna di queste discipline svolge in esso un ruolo diverso; mentre la probabilità si occupa principalmente di fornire modelli teorici che definiscono come è ripartita la probabilità nel caso di studio, la statistica parte dal caso di studio e descrive le sue proprietà statistiche oppure prova a risalire o inferire il modello probabilistico sotteso ad esso.

La statistica ha dunque due grandi categorie, la statistica descrittiva e la statistica inferenziale.

2.1 Statistica descrittiva

La statistica descrittiva è la branca della statistica che prova soltanto a riassumere e rappresentare i dati di un caso di studio mediante diversi strumenti descrittivi.

Prima di iniziare a conoscere i diversi strumenti della statistica descrittiva è importante definire per prima, diversi termini ricorrenti:

- *Dati campionari.* dati osservati sui campioni.
- *Dati grezzi.* i dati appena rilevati, senza alcuna sintesi o compressione che comporti perdita di informazione.
- *Serie storica.* dati grezzi ordinati temporalmente.
- *Frequenza assoluta.* quante volte una certa osservazione è apparsa nella rilevazione dei dati.
- *Frequenza relativa.* quante volte *sul totale* una certa osservazione è stata rilevata.

- *Dati aggregati.* dati raggruppati in classi rispetto a un certo criterio.

2.1.1 Indici Riassuntivi

Molte volte quando si rilevano dei dati è conveniente aggregarli rispetto un certo criterio con la finalità di rendere più agevole la sua analisi o presentazione, ma come si potrebbe supporre facilmente, quest'aggregazione comporta una perdita sostanziale di informazione, per questo e altri motivi, ci sono dei metodi per comprimere l'informazione, che chiamiamo indici riassuntivi, che anche se comportano delle perdite, risaltano una particolare proprietà dei dati, facilitando l'analisi complessiva a l'analisi di un insieme di indici.

Supponiamo di aver campionato i valori x_1, x_2, \dots, x_n allora possiamo utilizzare questi strumenti per riassumere i dati:

- ① *Media aritmetica campionaria.* Rappresenta il *baricentro* dei dati ove ogni campione ha lo stesso peso.

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.1)$$

- ② *Varianza campionaria corretta.* Rappresenta quanto i dati si discostino quadraticamente dalla media aritmetica.

$$S_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (2.2)$$

- ③ *Deviazione standard corretta.* Rappresenta una misura della volatilità dei dati.

$$\sqrt{S_{n-1}^2} = \bar{S} \quad (2.3)$$

- ④ *Momento campionario di indice k.*

$$m_k = \frac{1}{n} \sum_{i=1}^n x_i^k \quad (2.4)$$

- ⑤ *Momento centrale campionario di indice k*

$$m'_k = \frac{\sum_{i=1}^n (x_i - \bar{x})^k}{n} \quad (2.5)$$



Il momento campionario di indice 1 coincide con la media campionaria mentre il momento centrale campionario di indice 2 e la varianza campionaria corretta si legano come:

$$S_{n-1}^2 = m'_2 \frac{n}{n-1}$$

Ulteriori indici riassuntivi possono essere formulati se si effettua un riordine dei dati, per indicare che è stato effettuato un riordine si mette tra parentesi l'indice dell'osservazione, per fare ciò ci basta prendere i campioni x_1, x_2, \dots, x_n e ordinarli in forma crescente $x_{(1)}, x_{(2)}, \dots, x_{(k)}, \dots, x_{(n)}$ modo che:

$$x_{(1)} = \min\{x_i\} \quad x_{(n)} = \max\{x_i\} \quad \forall j, k \quad j < k \quad x_{(j)} \leq x_{(k)} \quad (2.6)$$

Con un riordine crescente possiamo usare i seguenti indici riassuntivi:

- ① *Range*. anche chiamato intervallo campionario; è l'insieme di valori che possono assumere i dati.

$$x_{(n)} - x_{(1)} \quad (2.7)$$

- ② *Midrange*.

$$\frac{x_{(1)} + x_{(n)}}{2} \quad (2.8)$$

- ③ *Mediana campionaria*. Rappresenta il valore o insiemi di valori che si trovano al centro della distribuzione.

$$x_{\left(\frac{n}{2}\right)} = \begin{cases} \text{se } n = 2r \text{ (n pari)} & \frac{x_{(r)} + x_{(r+1)}}{2} \\ \text{se } n = 2r + 1 \text{ (n dispari)} & x_{(r+1)} \end{cases} \quad (2.9)$$

- ④ *Quantili di ordine* $\alpha \in [0, 1]$. Sono valori che dividono i dati in $n = \frac{1}{\alpha}$ parti uguali.

$$x_{\alpha n} = \begin{cases} \text{se } (n-1) \cdot \alpha \text{ è intero } r & x_{(r+1)} \\ \text{se } (n-1) \cdot \alpha \in (r-1, r) \text{ si interpola fra } x_{(r)} \text{ e } x_{(r+1)} & \end{cases} \quad (2.10)$$

La mediana è il quantile di ordine $1/2$, i quartili sono quantili di ordini $1/4, 2/4$ e $3/4$, i *percentili* dividono i dati in cento parti uguali.

- ⑤ *Differenza interquartile*. L'ampiezza della range di valori che contiene la metà "centrale" dei valori osservati

$$IQR = x_{\left(\frac{3}{4}n\right)} - x_{\left(\frac{1}{4}n\right)} \quad (2.11)$$

2.1.2 Istogramma

Nel ambito della statistica sono di rilevante importanza le distribuzioni di frequenza, oltre ai dati numerici, queste ammettono anche rappresentazioni grafiche come diagrammi a barre o a torta se il carattere è qualitativo o con istogrammi se il carattere è quantitativo.

NB

Un grafico a barre verticale è diverso da un istogramma, l'istogramma è un grafico costituito da barre non distanziate, mentre sono distanziate nel caso di grafici a barre.

Numero di classi

Quando il numero di dati è molto alto, per poter dare una rappresentazione grafica dei dati, si deve suddividere l'intervallo in cui variano i dati in gruppi di dati detti *classi* e assegnare ogni osservazione rilevata alla classe corrispondente. La scelta del numero di classi non è banale: troppo poche appiattiscono il grafico fino a renderlo insignificante; troppe classi distruggono la semplicità dell'istogramma (che ha impatti sia nella lettura che nell'analisi) e possono esaltare il "rumore" legata alla variabilità dei dati.

Classi di ampiezza uguale

Fissata un'ampiezza h della classe e x i valori assunti nei dati si può calcolare il numero di classi con la formula:

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil$$

Regola della radice

Se non si vuole fissare un'ampiezza delle classi, il metodo più semplice per calcolare il numero k di classi ottimali è dato dalla regola della radice dove n rappresenta la numerosità dei dati/campione.

$$k = \sqrt{n}$$

Non perché questa regola sia molto intuitiva e facile da ricordare significa che non sia utilizzata nella pratica, in realtà è la regola utilizzata da Excel® .

Regola di Sturges

Sempre senza fissare un'ampiezza delle classi, un altro metodo per calcolare il numero k di classi ottimali è dato dalla regola di Sturges, questa però assume in modo implicito una distribuzione approssimativamente normale. (Più avanti si capirà cosa si intende)

$$k = \lceil \log_2 n \rceil + 1$$

Regola di Rice

In alternativa a Sturges si può usare

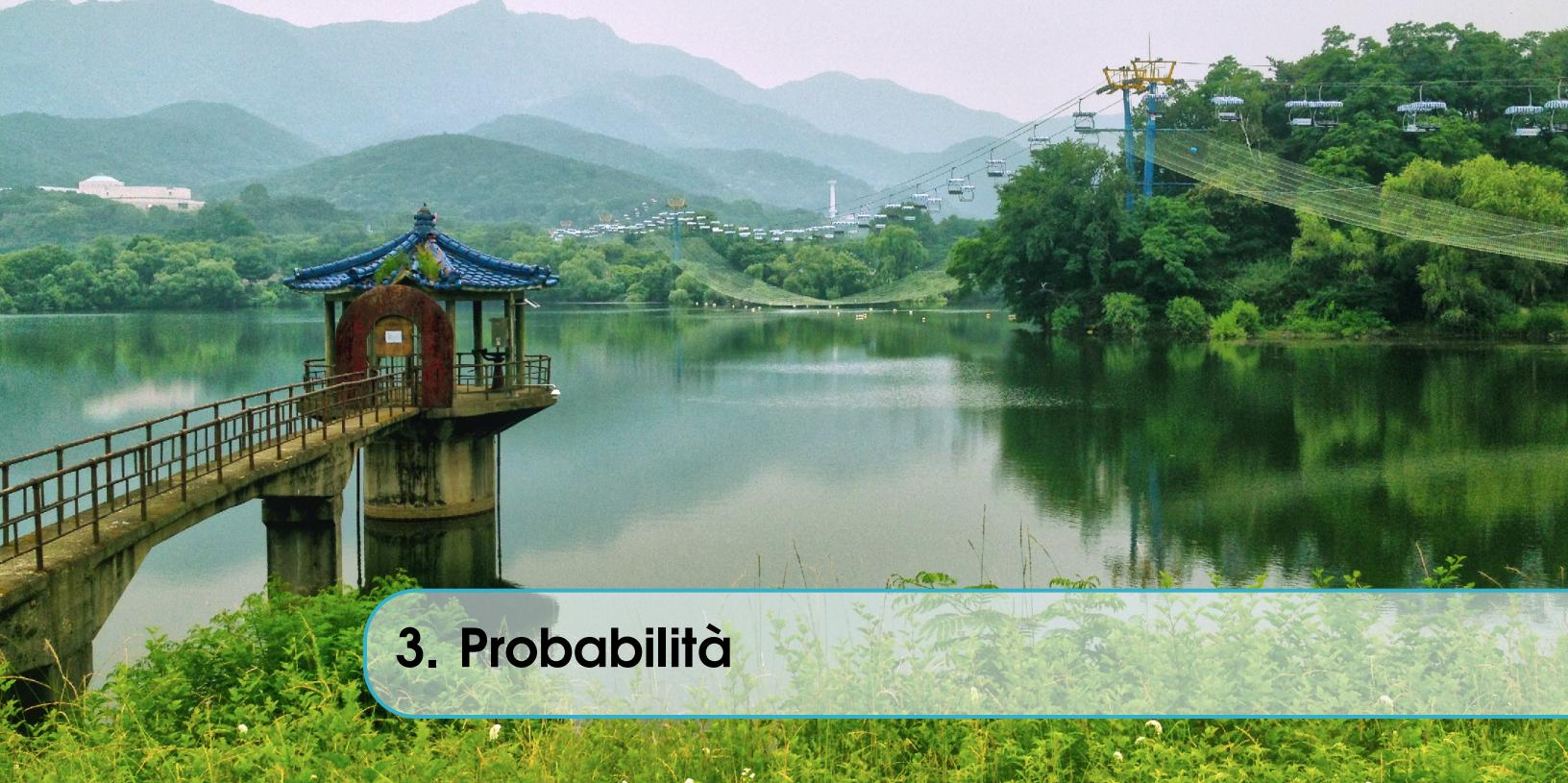
$$k = \lceil 2n^{1/3} \rceil$$

Regola di Scott

Ulteriore alternativa a Sturges e Rice

$$k = \frac{3.5S}{n^{1/3}},$$

dove S è la deviazione standard.



3. Probabilità

Il concetto di probabilità è il più importante della scienza moderna, soprattutto perché nessuno ha la più pallida idea del suo significato.

(BERTRAND RUSSELL, 1872-1970)

La probabilità , utilizzata sin dal XVI secolo da Cardano, può essere definita come la teoria matematica dell'incertezza e così come ogni modello matematico, consente la trattazione di un problema di interesse in modo logico e rigoroso. Per parlare di probabilità è bene chiarire i termini che verranno usati con molta frequenza:

- *Esiti elementari.* I risultati di un esperimento casuale; di solito indicati come $\omega_1, \omega_2, \dots$
- *Spazio campionario.* L'insieme che contiene come elementi tutti i possibili esiti dell'esperimento sotto considerazione; viene indicato con il simbolo Ω
- *Esempio casuale.* è un esperimento che dà luogo a differenti risultati se ripetuto più volte sotto le stesse condizioni.
- *Evento.* Si intende qualsiasi insieme di esiti elementari, di solito indicati con A, B, C, \dots

$$A \subset \Omega$$

- *Evento casuale o aleatorio.* Un evento che può verificarsi oppure non verificarsi, dato che le cause che lo producono non si possono oggettivamente controllare o governare.
- *Eventi indipendenti.* Due eventi si dicono indipendenti se il verificarsi dell'uno non influisce sul verificarsi dell'altro.
- *Eventi incompatibili.* Un evento si dice incompatibile o disgiunto se non ha eventi elementari in comune con un altro evento, cioè: $A \cap B = \emptyset$.

- Lo spazio campionario associato ad un esperimento si dice *discreto* se è uno spazio finito o infinito numerabile oppure *continuo* se è uno spazio infinito non numerabile.

3.1 Definizioni

3.1.1 Definizione classica

Nel caso della definizione classica, attribuita a *Laplace*, la probabilità di un evento (A) è definita come il rapporto fra due numeri, il numero di casi favorevoli all'evento e il numero di casi possibili, purché questi ultimi siano tutti equiprobabili.

$$\text{Probabilità} = \frac{\# \text{ casi favorevoli}}{\# \text{ casi possibili}} \quad \text{ossia} \quad P(A) = \frac{n_A}{n} \quad (3.1)$$

3.1.2 Definizione frequentista

La definizione classica presenta un evidente problema, non definisce la probabilità in casi di eventi non equiprobabili, per questo motivo il matematico austriaco *von Mises* propose di definire la probabilità di un evento come il limite cui tende la frequenza relativa dell'evento al crescere del numero degli esperimenti.

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n} \quad (3.2)$$

Cioè la probabilità di A è il limite per il numero di prove che va all'infinito del rapporto fra il numero delle prove nelle quali l'evento A si è verificato e il numero n delle prove effettuate. Questa definizione è fondamentata nella legge empirica del caso o *legge dei grandi numeri*.

NB

È importante notare che nella definizione classica la probabilità è nota *a priori*, ossia prima di guardare i dati; mentre per la definizione frequentista la probabilità è definita *a posteriori*, cioè dall'esame dei dati.

3.1.3 Definizione soggettivista

La definizione frequentista ha un problema nella sua formulazione, infatti non è detto che si possano fare esperimenti più volte sotto le stesse condizioni, per questo motivo *De Finetti* introdusse una nuova definizione di probabilità , la probabilità di un evento è il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica, 0 se l'evento non si verifica. Questa definizione è fondata, tuttavia, sull'opinione delle persone, quindi non è rigorosa e varia da persona a persona giacché ognuno presenta diverse propensioni al rischio.

3.1.4 Definizione assiomatica

La definizione assiomatica della probabilità , proposta da *Kolmogorov*, come soluzione a uno dei 23 problemi posti da Hilbert al congresso di Parigi nel 1900, non è una definizione operativa ne diretta, ma piuttosto è un'impostazione che accetta qualunque approccio, purché questo rispetti determinati assiomi.

A partire da questi assiomi, sono stati formulati diversi teoremi e leggi, che costituiscono la base della moderna teoria della probabilità . Vale la pena risaltare che in questo approccio assiomatico la probabilità viene vista come una *funzione* che associa a ciascun sottoinsieme di Ω un numero reale non negativo tale che la somma delle probabilità di tutti gli eventi è pari a 1.

Assiomi della probabilità

- 1** *Positività*. Ad ogni evento è assegnato un numero reale non negativo detto probabilità di A.

$$P(A) \geq 0$$

- 2** *Certezza*. La probabilità dell'evento certo e quindi dello spazio campionario Ω è sempre 1.

$$P(\Omega) = 1$$

- 3** *Unione*. Se A e B sono due eventi incompatibili, allora la probabilità della loro unione è la somma delle singole probabilità di A e B.

$$P(A \cup B) = P(A) + P(B)$$

Quest'ultimo assioma può essere esteso all'unione di un infinità numerabile di eventi, tale postulato viene detto *postulato della sommabilità completa*.

3.2 Teoremi fondamentali della probabilità

- 1** *Evento impossibile*. La probabilità dell'evento impossibile e quindi dell'insieme nullo \emptyset è sempre 0.

$$P(\emptyset) = 0$$

- 2** *Evento negazione*. La probabilità dell'evento negazione di A è pari al complemento a 1 della probabilità di A.

$$P(\bar{A}) = 1 - P(A)$$

- 3** *Monotonia*. La probabilità di un evento A incluso in un altro B è sempre minore uguale alla probabilità dell'evento in cui è incluso.

$$A \subseteq B \Rightarrow P(A) \leq P(B)$$

- 4** *Probabilità unione*. Se A e B sono due eventi compatibili, allora la probabilità della loro unione è la somma delle singole probabilità di A e B meno la loro intersezione.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

che può essere generalizzato come segue:

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \dots \\ &\quad + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \dots \end{aligned} \tag{3.3}$$

Ciò si tolgo le intersezioni che coinvolgono un numero pari di insiemi e si sommano quelle che coinvolgono un numero dispari di insiemi.

3.3 Probabilità condizionata

Si parla di probabilità condizionata o *condizionamento* quando si vuole calcolare la probabilità di un evento A sapendo che l'evento B si è verificato. Vale la formula

$$P(A|B) = \sum_{\omega \in A \cap B} P(\omega|B) = \frac{P(A \cap B)}{P(B)} \quad (3.4)$$

Questa definizione ci permette di definire la probabilità "semplice" come un caso particolare di probabilità condizionata in cui la condizione data è lo spazio campionario.

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

Finora sappiamo pochissime cose su $P(A \cap B)$, grazie agli assiomi e teoremi fondamentali, sappiamo che:

$$\max\{P(A) + P(B) - 1, 0\} \leq P(A \cap B) \leq \min\{P(A), P(B)\}$$

intuitivamente questa formula ha molto senso, basta pensare che se la probabilità di tutti gli eventi è maggiore di 1, vuol dire che c'è almeno una intersezione para all'eccidente di uno; questa intersezione può essere più grande, ma al massimo vale la probabilità del evento con la probabilità più bassa (per il fatto di essere l'intersezione fra insiemi).

Gli eventi si dicono *stocasticamente indipendenti* se il fatto che si sia verificato l'evento B non cambia la probabilità dell'evento A .

$$P(A|B) = P(A) \quad P(B|A) = P(B) \quad (3.5)$$

Nel caso gli eventi siano stocasticamente indipendenti possiamo calcolare la probabilità dell'intersezione sfruttando la definizione di probabilità condizionata, ottenendo facilmente:

$$P(A \cap B) = P(A)P(B) \quad \text{se } A \text{ e } B \text{ sono stocasticamente indipendenti.} \quad (3.6)$$

È importante sottolineare che la indipendenza stocastica non è una proprietà intrinseca all'evento ma piuttosto una proprietà della distribuzione della probabilità, più in generale diciamo che gli eventi sono indipendenti se la probabilità di un gruppo di eventi (prodotto logico) si può scrivere come prodotto delle probabilità dei singoli eventi. L'indipendenza è più difficile che si verifichi quanto più grande è il numero di condizioni da verificarsi.

$$\text{infatti deve soddisfare } 2^n - n - 1 \text{ condizioni}$$

Come conseguenza della probabilità condizionata possiamo dedurre i seguenti teoremi:

Teorema della probabilità composta

Come conseguenza della probabilità condizionata possiamo scrivere:

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

Questo teorema generalizzato prende il nome di **regola della catena**:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|A_1 \cap \dots \cap A_{n-1}) \quad (3.7)$$

Teorema delle probabilità totali

Consideriamo una partizione dello spazio campionario, $\Omega = A_1 \cup A_2 \cup \dots \cup A_n$ essendo $A_i \cap A_j \neq \emptyset$ con $i \neq j$. L'evento B si può scrivere come:

$$B \cap \Omega = B(A_1 \cup A_2 \dots \cup A_n) = B \cap A_1 \cup B \cap A_2 \cup \dots \cup B \cap A_n$$

evidenziando il fatto che l'evento B si può scomporre come la somma di tanti diversi pezzettini. Giacché abbiamo supposto una partizione, possiamo sfruttare la proprietà che tutti i pezzettini sono incompatibili fra loro per scrivere:

$$\begin{aligned} P(B) &= P(B \cap A_1 \cup B \cap A_2 \cup \dots \cup B \cap A_n) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + \dots + P(B|A_n)P(A_n) \\ &= \sum_{j=1}^n P(B|A_j)P(A_j) \end{aligned} \tag{3.8}$$

Teorema della probabilità marginale

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \tag{3.9}$$

NB

Bisogna avere ben chiaro che

$$\left\{ \begin{array}{l} P(A) + P(\bar{A}) = 1 \\ P(A|B) + P(\bar{A}|B) = 1 \\ P(A|B) + P(A|\bar{B}) \begin{cases} > 1 \\ < 1 \\ = 1 \end{cases} \end{array} \right.$$

3.3.1 Teorema di Bayes

È uno dei teoremi più importanti deducibili dalla probabilità condizionata. Prende il nome del matematico britannico Thomas Bayes (1702-1761) che fu il primo ad usare questa formula come un potente metodo per correggere una probabilità alla luce di nuove informazioni.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \tag{3.10}$$

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

La prima delle formule scrive $P(B)$ con la formula della probabilità marginale e la seconda scrive $P(B)$ con la formula delle probabilità totali.

ESEMPIO

Una classe del POLIMI è composta da 65% ingegneri gestionali e 35% di ingegneri informatici, si sa che solo il 30% dei gestionali si laurea in 3 anni mentre solo il 15% degli informatici ci riesce. Dopo tanti anni, presa una classe già laureata, qual è la probabilità che preso uno studente di cui si sa che si è laureato in 3 anni, sia ingegnere informatico?

$$P(I|L) = \frac{P(L \cap I)}{P(L)} = \frac{P(L|I)P(I)}{P(L|I)P(I) + P(L|G)P(G)} = \frac{0,15 \cdot 0,35}{(0,15 \cdot 0,35) + (0,3 \cdot 0,65)} = 21,21\%$$

Dal fatto che lo studente può essere solo gestionale o informatico si ha che:

$$P(G|L) = 1 - P(I|L) = 78,78\%$$

Ma qual è il ragionamento dietro la formula di Bayes?

Cominciamo calcolando alcune probabilità , ad esempio la probabilità che lo studente scelto si sia laureato in 3 anni e sia ingegnere informatico:

$$P(L \cap I) = P(L|I)P(I) = 5,25\%$$

e possiamo anche ottenere la probabilità che lo studente scelto si sia laureato in 3 anni e sia ingegnere gestionale:

$$P(L \cap G) = P(L|G)P(G) = 19,5\%$$

Grazie alla probabilità marginale possiamo ottenere la probabilità che uno studente della classe si sia laureato in 3 anni:

$$P(L) = P(L|I)P(I) + P(L|G)P(G) = 24,75\%$$

Infatti ora che abbiamo queste probabilità si tratta di sfruttare l'informazione che abbiamo in più, cioè il fatto di sapere che si sia laureato in 3 anni, infatti nel calcolo delle probabilità precedenti, consideravamo la possibilità che non si fosse laureato in 3 anni, quindi ora si tratta di "spalmare" proporzionalmente la probabilità del caso già verificato ("si è laureato in 3 anni") nei sottocasi ("è informatico" oppure "è gestionale"). Infatti:

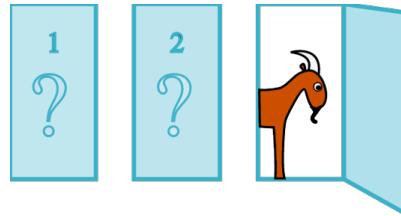
$$P(I|L) = \frac{5,25\%}{24,75\%} = 21,21$$

$$P(G|L) = \frac{19,5\%}{24,75\%} = 78,78$$

Che ci conferma il risultato ottenuto precedentemente

3.3.2 Problema di Monty-Hall

Supponiamo che un certo Vittorio stia partecipando ad un gioco a premi alla televisione; lui può scegliere fra tre porte; dietro una di esse c'è un'automobile e dietro le altre due, capre. Supponiamo che scelga la porta numero 1, e il conduttore televisivo, che sa cosa c'è dietro ciascuna porta, apre la numero 3, rivelando (ovviamente) una capra. A questo punto li domanda: "Vorresti cambiare la tua scelta per la numero 2?" È conveniente che cambi?



Questo problema è uno dei problemi di probabilità più famosi al mondo (insieme al problema delle tre carte che ha una costruzione molto simile). La maggior parte delle persone pensa che avendo due porte chiuse, si abbia una probabilità di $1/2$ di trovare l'automobile dietro ognuna di queste, e che quindi non ci sia motivo per cambiare porta. Ma non è affatto vero!

Chiamiamo l'evento che l'automobile si trovi dietro una certa porta rispettivamente A_1, A_2 e A_3 e C_3 la probabilità che il conduttore apra la porta 3. All'inizio:

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$$

Ma $P(C_3) = \frac{1}{2}$ poiché:

- Nel caso in cui la macchina sia dietro la porta 1, il conduttore televisivo può scegliere a caso fra la porta 2 o 3.

$$P(C_3|A_1) = \frac{1}{2}$$

- Nel caso in cui la macchina sia dietro la porta 2 sarà obbligato ad aprire la porta 3.

$$P(C_3|A_2) = 1$$

- Nel caso in cui la macchina sia dietro la porta 3 sarà obbligato ad aprire la porta 2. Pertanto

$$P(C_3|A_3) = 0$$

Dunque:

$$P(C_3) = \sum_{i=1}^3 P(C_3|A_i)P(A_i) = \frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \frac{1}{2}$$

È sempre conveniente che Vittorio cambi la sua porta 1 per la 2 perché:

$$P(A_2|C_3) = 1 - P(A_1|C_3) = 1 - \frac{P(C_3|A_1)P(A_1)}{P(C_3)} = 1 - \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Il "trucco" della soluzione sta nel notare che il conduttore televisivo **deve sempre** offrire la possibilità di cambiare e deve per forza aprire una porta che non contenga il premio.

3.3.3 Filtro bayesiano

Molti dei sistemi anti-spam di posta elettronica usati oggigiorno sono i cosiddetti *filtri bayesiani*; questo tipo di filtri eseguono un filtraggio basandosi sull'analisi sul contenuto delle email e appunto, si chiamano bayesiani, perché utilizzano il teorema di Bayes per fare tale filtraggio. In modo molto semplificato un filtro bayesiano calcola la probabilità che una email sia spam oppure no con la formula:

$$P(\text{Spam} | \text{Frase}) = \frac{P(\text{Frase} | \text{Spam}) P(\text{Spam})}{P(\text{Frase})}$$

Ad esempio un mail che contiene la frase “*This is not a scam*” o “*You’re a Winner!*” ha una probabilità del 98% di essere spam. Il termine $P(\text{Spam})$ è la probabilità che un qualunque messaggio inviato nel mondo sia spam, ovviamente questa probabilità viene stimata con la frequenza relativa; la cosa più sorprendente è che le statistiche stimano questa probabilità come 0.8, cioè, per ogni 10 email che si inviano nel mondo, 8 sono spam!



4. Variabili aleatorie

“La probabilità non è in fondo che il buon senso ridotto a calcolo” (LAPLACE)

In questo capitolo analizzeremo le *variabili aleatorie*, per il momento ci limiteremo a le variabili discrete finite. Una variabile aleatoria anche se è detta variabile, in realtà è una *funzione* che ha come dominio lo spazio campionario e assume valori diversi a seconda del risultato di un esperimento.

$$X : \Omega \rightarrow \mathbb{R} \quad \text{si scrive} \quad x = X(\omega) \quad (4.1)$$

In un esperimento, ciò che è veramente importante sono le osservazioni, dette anche *determinazioni*. Per convenzione, le variabili aleatorie si indicheranno con lettere maiuscole (X) e le determinazioni con lettere minuscole x ; per cui in realtà, d'ora in poi tenderemo a dimenticare tutte le variabilità esperimentalì che sono contenute in Ω e ω .

La funzione che associa ad ogni valore della variabile casuale una probabilità si chiama “distribuzione di probabilità” oppure “*densità discreta*” (nel caso discreto) oppure PMF (Probability Mass Function) e si scrive come:

$$P_X(A) \stackrel{\text{def}}{=} P(x \in A) \quad P_X(x_i) \stackrel{\text{def}}{=} P(X = x_i) \quad (4.2)$$

date due variabili aleatorie X e Y , si definisce la loro **distribuzione congiunta** come la distribuzione di probabilità associata al vettore aleatorio (X, Y)

$$P(X = x, Y = y) = P(Y = y | X = x) \cdot P(X = x) = P(X = x | Y = y) \cdot P(Y = y) \quad (4.3)$$

Si definisce come **distribuzione marginale** la distribuzione di una singola variabile aleatoria del insieme. Vale la pena notare che è sempre possibile ricavare la distribuzione marginale dalla distribuzione congiunta, ma il viceversa è solo possibile nel caso di indipendenza, concetto che verrà chiarito successivamente.

NB

La proprietà fondamentale delle variabili aleatorie è che la somma delle probabilità di tutte le determinazioni (valori assunti x_i) è sempre 1.

ESEMPIO

Tabella 4.1: Distribuzione di probabilità congiunta di X e Y

$Y X$	x_1	x_2	x_3	x_4
y_1	4/100	12/100	16/100	8/100
y_2	3/100	9/100	12/100	16/100
y_3	2/100	6/100	8/100	4/100
y_4	1/100	3/100	4/100	2/100

Tabella 4.2: Distribuzione di probabilità congiunta e marginale di X e Y

$Y X$	x_1	x_2	x_3	x_4	$P_Y(y)$
y_1	4/100	12/100	16/100	8/100	4/10
y_2	3/100	9/100	12/100	6/100	3/10
y_3	2/100	6/100	8/100	4/100	2/10
y_4	1/100	3/100	4/100	2/100	2/10
$P_X(x)$	1/10	3/10	4/10	2/10	1

Dalle tabelle si può facilmente dedurre che la probabilità marginale si ottiene sommando lungo le righe oppure lungo le colonne e trascrivendo il risultato appunto *a margine* rispettivamente della riga o colonna sommata, cioè $P_X(x_i) = \sum_{y_j} P_{X,Y}(x_i, y_j)$

4.1 Indipendenza di variabili aleatorie

La formulazione di indipendenza di variabili aleatorie è molto simile alla formulazione di indipendenza di eventi, ma non si devono confondere, giacché sono cose ben diverse. Due variabili aleatorie si dicono indipendenti se:

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j) \quad \forall i, j \quad (4.4)$$

Un’ulteriore differenza è che l’indipendenza tra variabili aleatorie si può ridurre in un’unica condizione (anche se ne riassume tante) mentre per eventi sono necessarie $2^n - n - 1$ condizioni.

Indipendenza Eventi A, B, C

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

$$P(A \cap B) = P(A)P(B)$$

$$P(A \cap C) = P(A)P(C)$$

$$P(B \cap C) = P(B)P(C)$$

Indipendenza variabili aleatorie X, Y, Z

$$P(X = x_i, Y = y_j, Z = z_k)$$

$$= P(X = x_i) P(Y = y_j) P(Z = z_k) \quad \forall i, j, k$$

ESEMPIO

Data una tabella che ci dice com'è distribuita la probabilità , come facciamo a sapere se le variabili sono indipendenti? Supponiamo di avere la tabella seguente:

Tabella 4.3: Distribuzione di probabilità

$Y X$	x_1	x_2	x_3	x_4
y_1	$10/81$	$5/81$	$15/81$	$15/81$
y_2	$2/81$	$1/81$	$3/81$	$3/81$
y_3	$4/81$	$2/81$	$6/81$	$6/81$
y_4	$2/81$	$1/81$	$3/81$	$3/81$

Per prima cosa occorre scrivere le leggi marginali:

Tabella 4.4: Scriviamo le leggi marginali di X e Y

$Y X$	x_1	x_2	x_3	x_4	$P_Y(y)$
y_1	$10/81$	$5/81$	$15/81$	$15/81$	$5/9$
y_2	$2/81$	$1/81$	$3/81$	$3/81$	$1/9$
y_3	$4/81$	$2/81$	$6/81$	$6/81$	$2/9$
y_4	$2/81$	$1/81$	$3/81$	$3/81$	$1/9$
$P_X(x)$	$2/9$	$1/9$	$3/9$	$3/9$	1

Ora occorre verificare che *ogni* cella della probabilità congiunta può essere scritta come prodotto della probabilità delle due leggi marginali, in questo caso, ciò è verificato, per cui X e Y sono indipendenti.

4.2 Funzioni di variabili aleatorie

Se le variabili aleatorie in realtà sono funzioni, perché diamine si chiamano variabili? Perché così come per una qualunque variabile, è possibile realizzare funzioni di variabili aleatorie, queste funzioni non solo cambiano gli eventuali valori che può assumere la determinazione, ma comportano anche una ridistribuzione della probabilità . Sia $Y = g(X)$

$$P_Y(y) := \sum_{x:g(x)=y} P_X(x) \quad (4.5)$$

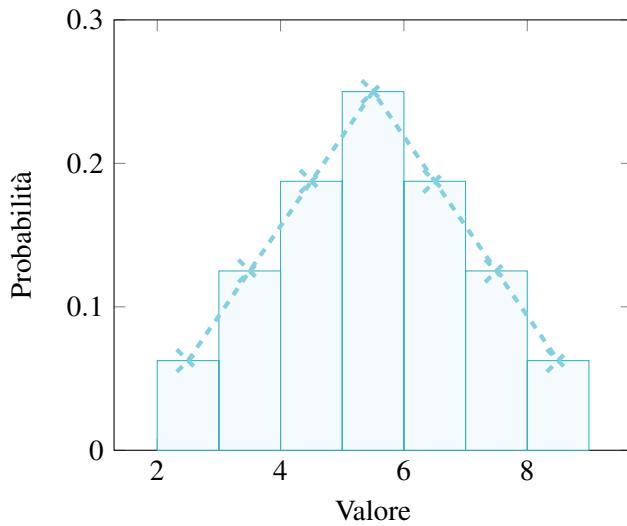
ESEMPIO

Consideriamo la funzione somma $Y = X_1 + X_2$ ove sia X_1 che X_2 sono variabili aleatorie che possono assumere i valori $\{1, 2, 3, 4\}$ con la stessa probabilità ; la distribuzione di probabilità congiunta e marginale è:

$X_1 X_2$	1	2	3	4	$P_{X_2}(x)$
1	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
2	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
4	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$P_{X_1}(x)$	$1/4$	$1/4$	$1/4$	$1/4$	1

La funzione somma, non solo crea nuovi valori, ma cambia anche la probabilità, infatti, il numero 5 ha più probabilità di comparire giacché può essere ottenuto come: $5 = 1 + 4 = 2 + 3 = 3 + 2 = 4 + 1$. Su questa funzione in particolare, basta sommare le probabilità lungo le diagonali secondarie per ottenere le probabilità dei nuovi valori possibili (com'è evidenziato dai colori). Facendo il poligono di frequenza della legge di distribuzione notiamo che assume una forma particolare, a forma di triangolo, per questo motivo si dice che la variabile aleatoria Y ha una *distribuzione triangolare*.

$X_1 + X_2$	2	3	4	5	6	7	8
$P_Y(y)$	1/16	2/16	3/16	4/16	3/16	2/16	1/16



4.2.1 Formula di Convoluzione

Introduciamo la **formula di convoluzione** che ci dice che $\forall X_1, X_2$, se si vuole conoscere la distribuzione di $Y = X_1 + X_2$, vale la formula:

$$P_Y(y) = \sum_{(x_1, x_2): x_1 + x_2 = y} P_{X_1}(x_1) P_{X_2}(x_2) = \sum_{x_1} P_{X_1}(x_1) P_{X_2}(y - x_1) \quad (4.6)$$

Per ricavare la somma di due dadi cubici, anziché scrivere la probabilità marginale e da lì ricavare la probabilità congiunta, poi sommare adeguatamente, ecc. (come si è appena fatto), è molto più facile realizzare questa procedura con la formula di convoluzione, cioè, sapendo che la PMF di un dado cubico è $P_{X_i}(x) = \frac{1}{6} I_{\{1,2,\dots,6\}}(x)$ possiamo applicare la formula di convoluzione come segue:

$$\sum_x P_{X_1}(x) P_{X_2}(y - x) = \sum_x \frac{1}{6} I_{\{1,2,\dots,6\}}(x) \cdot \frac{1}{6} I_{\{1,2,\dots,6\}}(y - x)$$

Questa sommatoria è complicata, poiché il valore da sommare cambia a seconda del valore che assume la x , infatti le funzioni I chiamate indicatrici, servono principalmente a ricordarci dove varia la variabile, perciò cominciamo per abbordare il problema vedendo le indicatrici.

Le indicatrici ci impongono:

$$1 \leq x \leq 6 \quad 1 \leq y - x \leq 6$$

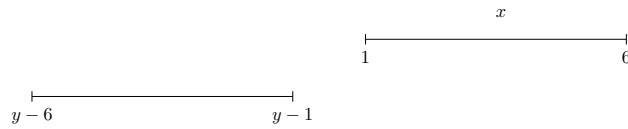
Il primo passo da fare è esprimere tutte le disequazioni rispetto alla variabile che si sta sommando, con lo scopo di ottenere gli intervalli dove si deve sommare, in questo caso lo facciamo rispetto a x e si ottiene:

$$1 \leq x \leq 6 \quad y - 6 \leq x \leq y - 1$$

Il problema ora è che negli intervalli compare un'altra variabile (y) di cui non sappiamo il suo valore. Per questo motivo dobbiamo considerare tutti i casi possibili, in questo caso i casi sono 4:

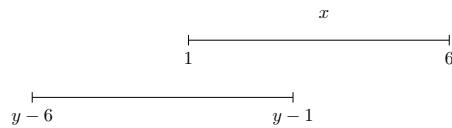
Caso 1

I valori che assume la y sono *tutti* valori più piccoli dei valori assunti dalla x , per cui non esiste alcun intervallo dove si abbia valori non nulli per entrambe le PMF (si ricorda che l'indicatrice impone il valore 0 per tutti i valori diversi dal dominio stabilito).



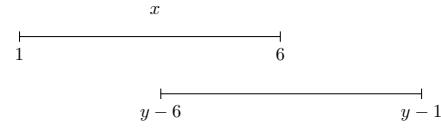
Caso 2

Alcuni valori che assume la y sono più piccoli dei valori assunti dalla x , per cui si ha un intervallo con valori non nulli per entrambe le PMF. Dalla figura si apprezza che l'intervallo è da 1 a $y - 1$



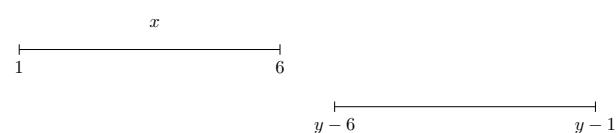
Caso 3

Alcuni valori che assume la y sono più grandi dei valori assunti dalla x , per cui si ha un intervallo con valori non nulli per entrambe le PMF. Dalla figura si apprezza che l'intervallo è $y - 6$ a 6.



Caso 4

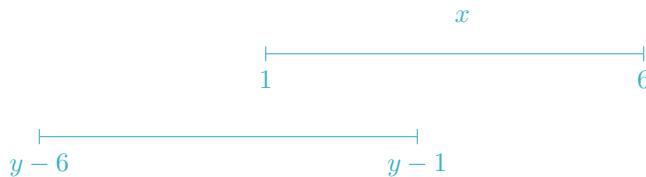
I valori che assume la y sono *tutti* valori più grandi dei valori assunti dalla x , per cui non esiste alcun intervallo dove si abbia valori non nulli per entrambe le PMF (si ricorda che l'indicatrice impone il valore 0 per tutti i valori diversi dal dominio stabilito).



Ora che abbiamo trovato tutti gli intervalli non nulli da sommare, realizziamo la somma per ognuno di questi:

Caso 2

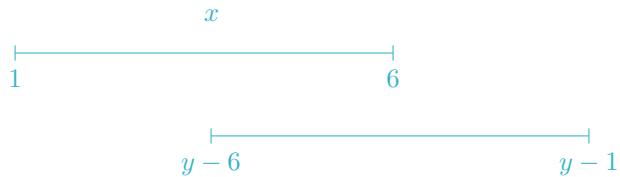
$$\sum_1^{y-1} \frac{1}{36} = \frac{1}{36} \sum_1^{y-1} 1 = \frac{(y-1)}{36}$$



Dalla figura si può apprezzare che $y-1$ si trova tra 1 e 6 per cui questo risultato è valido per $2 \leq y \leq 7$

Caso 3

$$\sum_{y=6}^6 \frac{1}{36} = \frac{1}{36} \sum_{y=6+7-y}^{6+7-y} 1 = \frac{(13-y)}{36}$$



Dalla figura si può apprezzare che $y-6$ si trova tra 1 e 6 per cui questo risultato è valido per $7 \leq y \leq 12$.

L'unica cosa che ci resta da fare è definire uno dei risultati per $y = 7$ che come si può verificare facilmente è definita per entrambe le espressioni con lo stesso valore come risultato. arbitrariamente, sceglieremo il risultato con i valori di y più grandi, cioè per $y = 7$ risulterà definita $\frac{(13-y)}{6}$ e non $\frac{(y-1)}{6}$.

A questo punto non ci resta altro che scrivere l'espressione finale:

$$P_Y(y) = \frac{(y-1)}{36} I_{\{2,3,\dots,6\}}(y) + \frac{(13-y)}{36} I_{\{7,8,\dots,12\}}(y)$$

NB

Si noti che per $y = 7$ si è dovuto scegliere uno dei due risultati, perché altrimenti entrambe le espressioni sarebbero state definite per quel valore, dandoci come risultato $2/6$ anziché $1/6$

4.3 Indici caratteristici di una variabile aleatoria

Molto spesso è conveniente trovare un modo di descrivere una variabile aleatoria nel modo più sintetico possibile, in questo caso gli indici caratteristici, ci permettono di cogliere le caratteristiche essenziali come la forma, posizione e variabilità della distribuzione tramite un numero ridotto di valori.

- 1** *Valore atteso.* Viene chiamato anche *speranza matematica della distribuzione* e rappresenta il baricentro della distribuzione della probabilità (quindi si può vedere come una media pesata); è indicato con $E[X]$ (dall'inglese *expectation* e dal francese *espérance*).

$$\left\{ \begin{array}{l} \text{numero finito} = E[X] = \sum_{i=1}^n x_i P_X(x_i) \\ \text{infinità numerabile} = E[X] = \sum_{i=1}^{\infty} x_i P_X(x_i) \end{array} \right. \quad (4.7)$$

Se X è una variabile aleatoria discreta (assume un numero finito di valori) e tutti i valori assunti sono equiprobabili allora il valore atteso è la media aritmetica dei valori.

$$E[X] = \frac{1}{n} \sum_{i=1}^n x_i \quad (4.8)$$



Una variabile aleatoria con valore atteso negativo non necessariamente indica che la maggior parte delle osservazioni sono negative



- 2** *Varianza.* Quantifica la misura della dispersione dei valori di X

$$\sigma^2 = Var(X) = \sum_{i=1}^n P_X(x_i) (x_i - E[X])^2 \quad (4.9)$$

- 3** *Deviazione standard.*

$$\sigma = \sqrt{Var(X)} = \sqrt{\sum_{i=1}^n P_X(x_i) (x_i - E[X])^2} \quad (4.10)$$

questi indici, molto simili ma diversi da quelli della statistica, verranno approfonditi ulteriormente.

4.3.1 Proprietà degli indici caratteristici

1 VALORE ATTESO

a Operatore lineare/Funzionale

$$E[\alpha X + \beta Y + \gamma] = \alpha E[X] + \beta E[Y] + \gamma$$

Si noti che il valore atteso di una costante è la costante stessa $E[\gamma] = \gamma$

b Omogeneità

Se $X \geq Y$ allora:

$$E[X] \geq E[Y]$$

2 VARIANZA

a Non negatività

$$Var(X) \geq 0$$

b Omogeneità di secondo grado

$$Var(\alpha X) = \alpha^2 X$$

c Invariante per traslazione

$$Var(\alpha X + b) = Var(\alpha X) = \alpha^2 X$$

d Riscrittura

$$Var(X) = E[(X - E[X])^2]$$

$$\begin{aligned} Var(X) &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] - E[2XE[X]] + E[E[X]^2] \quad (\text{linearità del valore atteso}) \quad (4.11) \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \quad (\text{valore atteso di una costante}) \\ &= E[X^2] - E[X]^2 \end{aligned}$$

e Somma di due variabili aleatorie

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Questa si ottiene come conseguenza della proprietà di riscrittura, infatti:

$$Var(X + Y) = E[(X + Y)^2] - E[X + Y]^2 = E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 -$$

Che per le proprietà del valore atteso si ottiene:

$$Var(X + Y) = E[X^2] + 2E[X]E[Y] + E[Y^2] - (E[X])^2 - (E[Y])^2 - 2E[X]E[Y]$$

e raggruppando opportunamente:

$$Var(X + Y) = Var(X) + Var(Y) + 2\underbrace{(E[XY] - E[X]E[Y])}_{\text{Covarianza}=\text{Cov}(X,Y)}$$

3 COVARIANZA

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] \quad (4.12)$$

La covarianza appena introdotta è un numero che fornisce una misura di quanto due variabili aleatorie varino assieme, perciò prende il nome di covarianza.

a *Riscrittura*

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

b *Simmetria*

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

c *Covarianza per variabili indipendenti*

Se due variabili aleatorie sono tra loro indipendenti vale:

$$E[XY] = E[X]E[Y]$$

che comporta:

$$\text{Cov}(X, Y) = 0$$

in questo caso si dice che le variabili aleatorie sono *incorelate*. Come conseguenza, vale la proprietà di additività della varianza per variabili aleatorie indipendenti.

$$\text{Se } X, Y \text{ sono indipendenti} \Rightarrow \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

4.4 Disuguaglianza di Cebicev

La disuguaglianza di Cebyshev / Cebicev è molto importante nell'ambito della probabilità e ci dice che qualunque sia la variabile aleatoria X , purché questa abbia media e varianza finita, soddisfa la disuguaglianza:

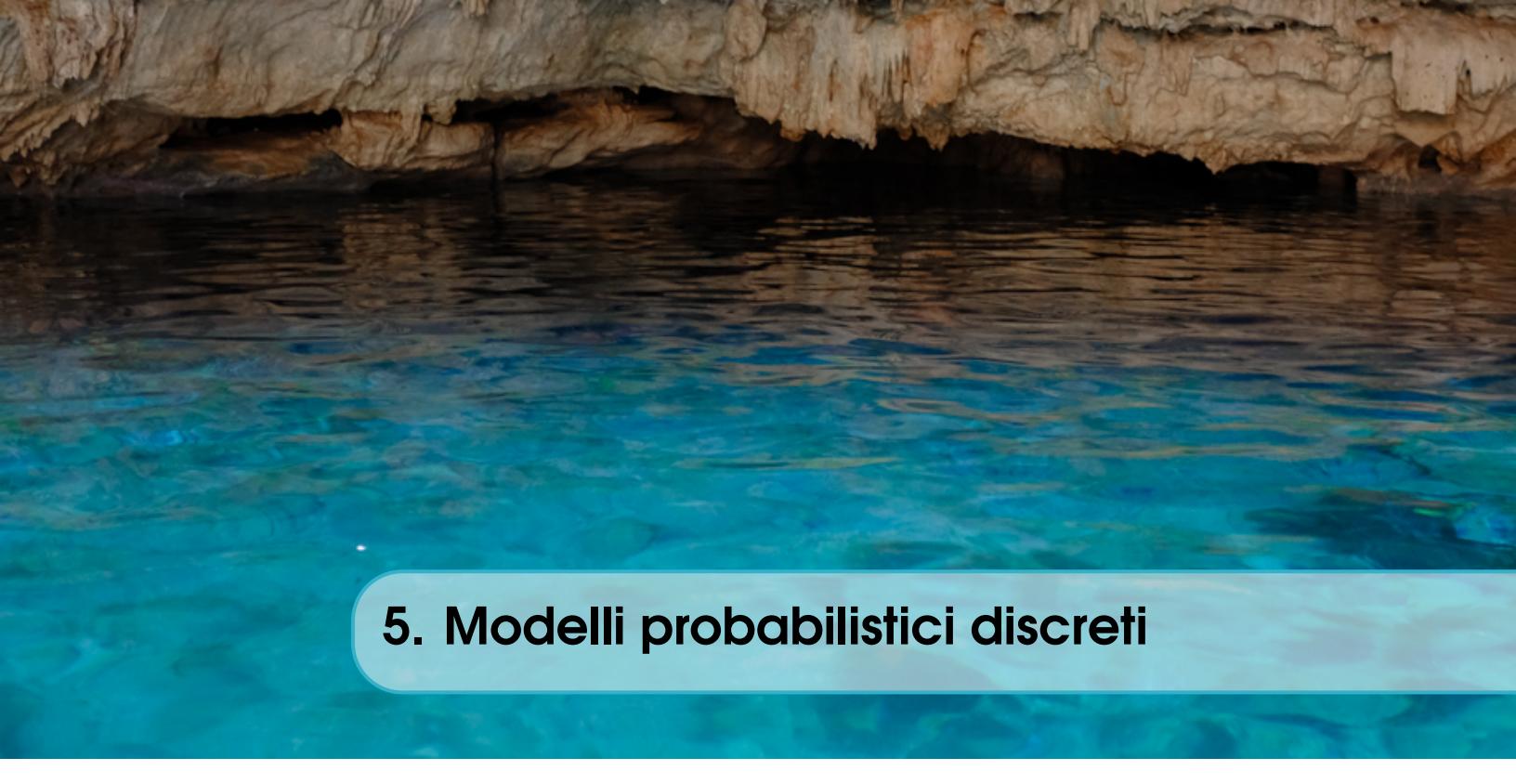
$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}. \quad (4.13)$$

$$P(|X - \mu| < h\sigma) \geq 1 - \frac{1}{h^2}.$$

In sintesi, quello che ci dice questa disuguaglianza è che prendendo un intero positivo h , siamo in grado di trovare qual è la probabilità che la variabile aleatoria si distribuisca nell'intervallo $\mu - h\sigma$ e $\mu + h\sigma$, questa stima è però pessimista a causa della sua generalità, quindi ci dice *almeno* quanta probabilità si distribuisce nel intervallo, ne può essere di più.

Da questa disuguaglianza si deduce:

- almeno il 75% dei valori si trova tra $\mu - 2\sigma$ e $\mu + 2\sigma$
- almeno l'88,8% dei valori si trova tra $\mu - 3\sigma$ e $\mu + 3\sigma$



5. Modelli probabilistici discreti

My mother said to me, 'If you are a soldier, you will become a general. If you are a monk, you will become the Pope.' Instead, I was a painter, and became Picasso.

(PABLO PICASSO, 1881-1973)

Il concetto di variabile aleatoria ci permette di formulare modelli che ci permettono di studiare molti fenomeni aleatori, uno di questi modelli è il processo di Bernoulli.

5.1 Bernoulliana

Una *prova di Bernoulli* è un esperimento aleatorio che ha due esiti possibili: il successo con probabilità θ , o il fallimento con probabilità $(1 - \theta)$. Il numero θ viene chiamato *parametro* della prova di Bernoulli.

All'esperimento viene associato una variabile aleatoria X che assume il valore 1 in caso di successo e assume 0 in caso di fallimento, questa variabile aleatoria si chiama **bernoulliana di parametro θ** , e si scrive $X \sim B(\theta)$.

$$\begin{aligned} \text{successo} &\rightarrow P(X_i = 1) = \theta \\ \text{fallimento} &\rightarrow P(X_i = 0) = (1 - \theta) \end{aligned} \tag{5.1}$$

Possiamo scrivere la bernoulliana in funzione della **funzione indicatrice** che come già accennato, è una funzione che ci indica in un certo senso il dominio di una espressione, questa è definita come:

$$I_A(x) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases} \tag{5.2}$$

Dunque:

$$X \sim B(\theta) \rightarrow P_X(x) = \theta^x (1 - \theta)^{1-x} I_{\{0,1\}}(x)$$

Valore atteso

Il valore atteso di una variabile bernoulliana è data dalla formula già introdotta:

$$E[X] = \sum_{i=1}^n \underbrace{x_i}_{val.\text{assunto}} \underbrace{P_X(x_i)}_{prob.\text{val.\text{assunto}}}$$

solo che in questo caso sappiamo che la $P_X(x_i)$ è data dalla 5.1, quindi si ha:

$$E[X] = \sum_{i=1}^n x_i \theta^x (1 - \theta)^{1-x} I_{\{0,1\}}(x)$$

Ora scomponiamo questa formula indicando le determinazioni (che per definizione di variabile bernoulliana sono solo 1 in caso di successo o 0 in caso di insuccesso), quindi:

$$E[X] = \underbrace{1}_{succ.} \cdot \underbrace{\theta}_{prob.\text{succ.}} + \underbrace{0}_{insucc.} \cdot \underbrace{(1 - \theta)}_{prob.\text{insucc.}} = \theta$$

Varianza

La varianza si calcola con la formula già introdotta:

$$Var(X) = E[X^2] - (E[X])^2$$

dato che è molto più semplice ottenere il risultato che con $Var(X) = \sum_{i=1}^n P_X(x_i) (x_i - E[X])^2$. A questo punto conosciamo già il valore atteso della variabile aleatoria bernoulliana, per poter applicare la prima formula ci basta calcolare $E[X^2]$ che ha come formula:

$$E[X^2] = \sum_{i=1}^n x_i^2 P_X(x_i)$$

Scomponendo la formula otteniamo:

$$E[X^2] = 1^2 \cdot \theta + 0^2 \cdot (1 - \theta) = \theta$$

Ora sapendo che $E[X^2] = \theta$ e $E[X] = \theta$ possiamo applicare la formula:

$$Var(X) = E[X^2] - (E[X])^2 = \theta - \theta^2 = \theta(1 - \theta)$$

In sintesi:

$$\underline{\text{Variabile Bernoulliana} \quad X \sim B(\theta)}$$

$$E[X] \qquad \qquad \qquad \theta$$

$$Var(X) \qquad \qquad \qquad \theta(1 - \theta)$$

5.2 Binomiale

Chiamiamo *processo di Bernoulli* un'iterazione finita o infinita di esperimenti di Bernoulli di uguale parametro θ , tra loro indipendenti.

In un processo Bernoulliano, la probabilità di ottenere, in n prove, una particolare sequenza di k successi e $(n - k)$ insuccessi è pari a :

$$\theta^k(1 - \theta)^{n-k} \quad (5.3)$$

mentre la probabilità di avere almeno un successo in n prove è pari a:

$$1 - (1 - \theta)^n \quad (5.4)$$

È logico pensare di introdurre una variabile aleatoria Y_n che conta il numero totale di successi ottenuti nelle n prove; questa variabile aleatoria prende il nome di **binomiale di parametro θ** , e si scrive $Y_n \sim bin(n, \theta)$ o alcune volte $Y_n \sim bin(x; n, \theta)$ specificando la *determinazione*. Poiché questa variabile conta il numero di successi, si dice che è una variabile di conteggio. Questa variabile aleatoria può essere vista come non altro che la somma dei successi delle singole prove (questo è dovuto al fatto che consideriamo le prove indipendenti ed identicamente distribuite); ossia la binomiale è la somma di n variabili bernoulliane (dunque per $n = 1$ la binomiale coincide con la bernoulliana) .

Basandoci sulla 5.3 e la definizione di combinazioni semplici possiamo facilmente concludere che la probabilità (densità discreta) di avere una qualunque sequenza di x successi è pari a:

$$Y_n \sim bin(n, \theta) \rightarrow P_{Y_n}(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} I_{\{1, 2, \dots, n\}}(x) \quad (5.5)$$

La distribuzione di probabilità (discreta) che descrive il numero di successi in un processo di Bernoulli viene chiamata **distribuzione binomiale**.

5.2.1 Valore atteso

Il valore atteso di una variabile binomiale si trova così:

$$E[Y_n] \underset{def. bin}{=} E \left[\sum_{i=1}^n X_i \right] \underset{ind.}{=} \sum_{i=1}^n E[X_i] \underset{id. distrib.}{=} \sum_{i=1}^n \theta = n\theta$$

5.2.2 Varianza

La varianza si calcola in modo simile:

$$Var[Y_n] \underset{def. bin}{=} Var \left[\sum_{i=1}^n X_i \right] \underset{ind.}{=} \sum_{i=1}^n Var[X_i] \underset{id. distrib.}{=} \sum_{i=1}^n \theta(1 - \theta) = n\theta(1 - \theta)$$

In sintesi:

Variabile Binomiale $Y_n \sim bin(n; \theta)$

$$E[Y_n] = n\theta$$

$$Var(Y_n) = n\theta(1 - \theta)$$

ESEMPIO

Consideriamo il lancio di una moneta di due facce di cui vogliamo sapere la densità discreta (cioè come si distribuisce la probabilità), sapendo che una moneta normale ha come parametro $\theta = 0,5$ (che esca l'evento testa) mentre una moneta truccata ha parametro $\theta = 0,6$.

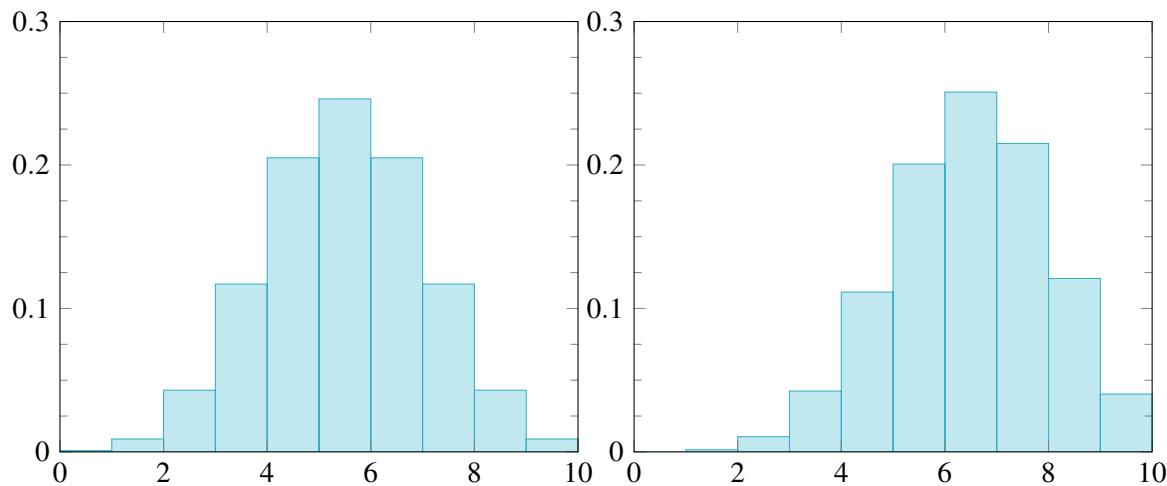


Figura 5.1: $\theta = 0,5$

Figura 5.2: $\theta = 0,6$

Nel caso in cui $\theta = 0,5$ la densità è simmetrica. Nel caso $\theta = 0,6$ si può percepire che la distribuzione è asimmetrica; è dunque opportuno introdurre un nuovo indice, l'indice di asimmetria.

5.2.3 Indice di asimmetria

L'indice di asimmetria anche chiamato *skewness* ci da una misura di quanto sia asimmetrica la distribuzione. Affinché una distribuzione sia simmetrica, è una condizione necessaria, ma non sufficiente, che l'indice di asimmetria sia nullo.

Occorre introdurre il concetto di asimmetria sia per i dati che per le variabili aleatorie (che anche se sono cose estremamente legate sono formalmente diverse) quindi ricordando la definizione di momento centrato possiamo definirlo come:

DATI	VARIABILI ALEATORIE
$a = \frac{m'_3}{m''_2}$	$A = \frac{\mu'_3}{\mu''_2}$

DATI	VARIABILI ALEATORIE
Momento campionario di ordine k	Momento di ordine k
$m_k = \frac{1}{n} \sum_i^n (x_i)^k$	$\mu_k = E[(X)^k]$
Momento campionario centrato di ordine k	Momento centrato di ordine k
$m'_k = \frac{1}{n} \sum_i^n (x_i - \bar{x})^k$	$\mu'_k = E[(X - E[X])^k]$

In particolare si avrà asimmetria negativa se i dati tendono a destra ($\theta > 0,5$) e asimmetria positiva se i dati tendono a sinistra ($\theta < 0,5$).

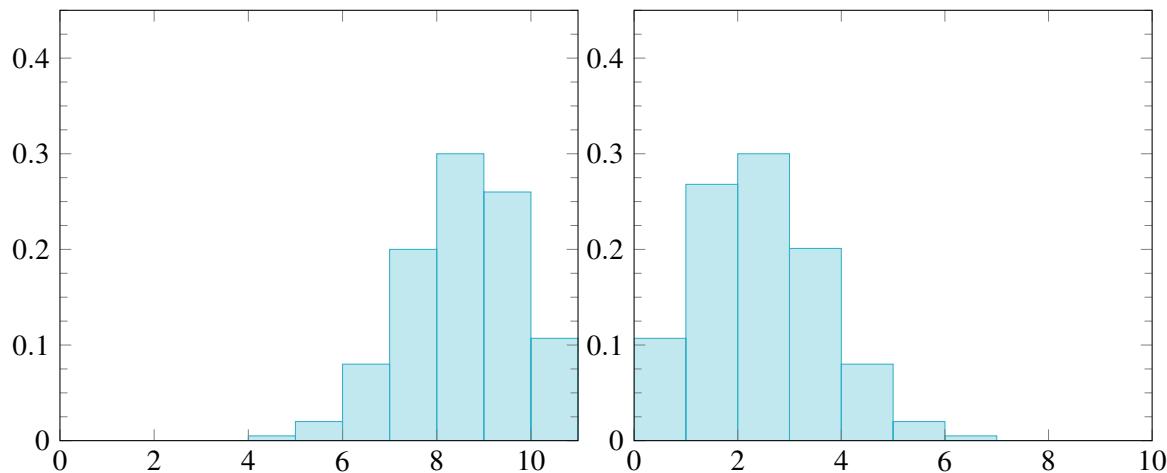


Figura 5.3: $\theta = 0,8$ $skewness = -0.47$

Figura 5.4: $\theta = 0,2$ $skewness = 0.47$

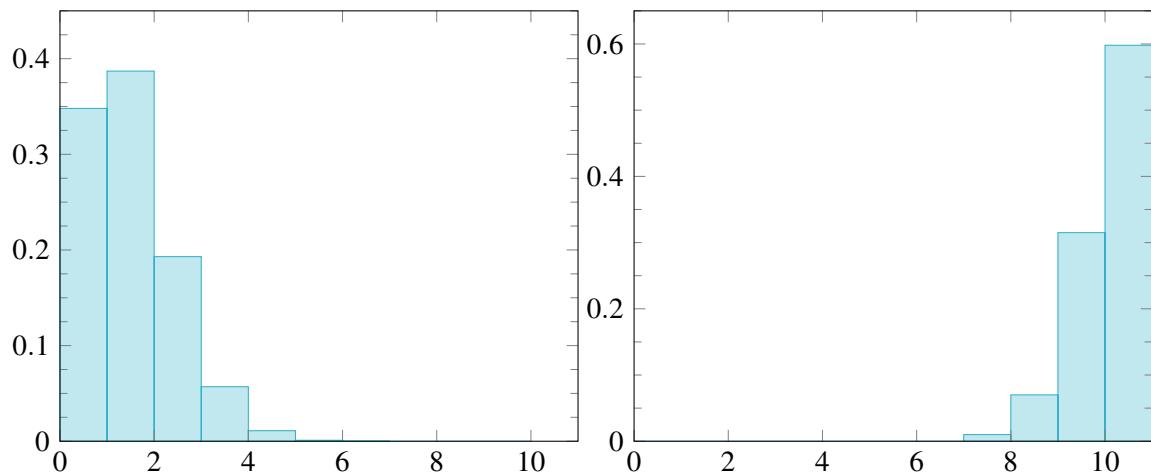


Figura 5.5: $\theta = 0,1$ $skewness = 0.84$

Figura 5.6: $\theta = 0,95$ $skewness = -1.3$

5.3 Geometrica

Supponiamo di poter fare un numero illimitato ma numerabile di prove in un processo di Bernoulli. Per questo processo illimitato introduciamo una nuova variabile aleatoria chiamata **geometrica di parametro θ** e indicata come $T \sim geom(\theta)$; questa variabile conta il numero di prove necessarie prima che si presenti il primo successo (quindi si dice che è una variabile d'attesa). La sua legge di distribuzione chiamata ovviamente **distribuzione geometrica** è data da:

$$T \sim geom(\theta) \rightarrow P_T(x) = \theta(1 - \theta)^{x-1} I_{\{1,2,\dots\}}(x) \quad (5.6)$$

Ma perché è fatta così? Se chiamiamo x il numero della prova in cui si è verificato il primo successo significa che si sono verificati $x - 1$ insuccessi e finalmente nella prova x si è avuto il primo successo, quindi la probabilità di ogni evento ha come esponente la frequenza dell'evento stesso. Notiamo che l'indicatrice comincia da 1 poiché, indicandoci il numero d'ordine del primo successo, questo non può essere 0.

$$P_T(x) = \underbrace{\theta}_{prob.\ succ.} \underbrace{(1 - \theta)^{x-1}}_{prob.\ insucc.} I_{\{1,2,\dots\}}(x)$$

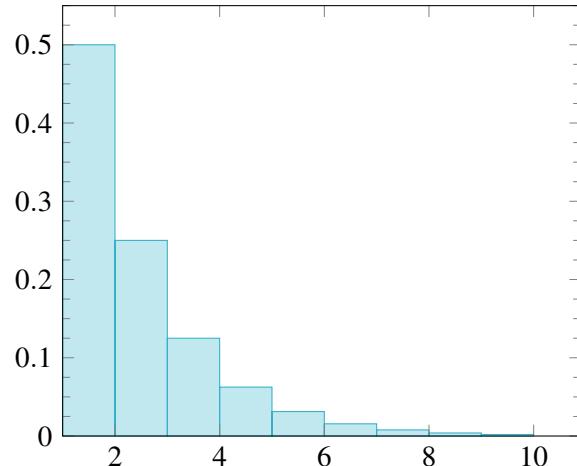
Al non limitare il processo, la variabile aleatoria può assumere infiniti valori, infatti si può prendere un qualunque $x \geq 1$, $x \in \mathbb{N}$.

È opportuno notare che la legge conta anche il numero di prove necessarie per ottenere il prossimo successo, a partire di una prova qualsiasi, questo come conseguenza dell'indipendenza delle prove.

ESEMPIO

Un esempio chiaro di distribuzione geometrica è un processo che è distribuito come la serie geometrica di ragione pari a $1/2$ e indice di partenza pari a 1 (per ottenerla basta considerare la distribuzione con $\theta = \frac{1}{2}$). Infatti, questa serie converge a 1 e dunque soddisfa le condizioni necessarie per poter essere considerata come una legge di distribuzione.

$$\sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k = 1$$



In questo caso la distribuzione può essere considerata come quella di una moneta non truccata e come si evidenzia dal grafico, la probabilità di ottenere una sequenza di tutte teste o tutte croci decresce all'aumentare dei lanci.

CURIOSITÀ (La probabilità di morire oggi è la stessa di lanciare 16 volte testa [1] e la probabilità di vincere il SuperEnalotto è la stessa di lanciare 30 volte testa [2], cioè la probabilità di morire oggi è 16384 volte quella di vincere il SuperEnalotto).

Se anziché contare il numero di successi si vuole contare il numero x di fallimenti prima del primo successo, si introduce una variabile aleatoria molto simile alla geometrica ma con i suoi valori "traslati" di una unità : $T' = T - 1$. Se per esempio il primo successo si è verificato alla x -esima + 1 prova vuol dire che ho avuto x insuccessi. Dato al suo evidente legame con la geometrica la legge si chiamerà sempre geometrica ma diremo che è "traslata" e si indicherà $T' \sim geom(\theta)$ o $T' \sim G'(\theta)$, la distribuzione è data da:

$$T \sim geom'(\theta) \rightarrow P_T(x) = \theta(1-\theta)^x I_{\{0,1,\dots\}} \quad (5.7)$$

Notiamo che in questo caso l'indicatrice parte da 0 e non da 1 poiché se il successo l'ottengo "al primo colpo" allora il numero di fallimenti è nullo. Quindi:

$$P_T(x) = \underbrace{\theta}_{prob.\ succ.} \underbrace{(1-\theta)^x}_{prob.\ insucc.} I_{\{0,1,\dots\}}(x)$$

5.3.1 Valore atteso

Il valore atteso di una geometrica e geometrica traslata si ottengono come:

$$E[T] = \sum_{x=1}^n x P_T(x) = \sum_x x \theta (1-\theta)^{x-1}$$

Se portiamo fuori θ dalla sommatoria notiamo che dentro la sommatoria c'è la derivata di $(1-\theta)^x$ quindi riscriviamo il tutto come:

$$E[T] = \theta \sum_x \frac{-d}{d\theta} [(1-\theta)^x]$$

In questo caso (non sempre si può fare) scambiamo la sommatoria delle derivate come la derivata della sommatoria:

$$E[T] = \theta \cdot \frac{-d}{d\theta} \sum_x [(1-\theta)^x]$$

Ora la sommatoria è diventata una sommatoria geometrica con somma $\frac{1}{1-q}$ ove q è la ragione della serie, quindi:

$$E[T] = \theta \cdot \frac{-d}{d\theta} \left[\frac{1}{1-(1-\theta)} \right] = \theta \cdot \frac{1}{\theta^2} = \frac{1}{\theta}$$

Per la geometrica traslata ci basta sfruttare la definizione.

$$E[T'] \underset{per.\ def.}{=} E[T-1] \underset{prop.\ val.\ att.}{=} E[T] - E[1] = \frac{1}{\theta} - 1$$

5.3.2 Varianza

La varianza si calcola con la solita formula:

$$Var(X) = E[X^2] - (E[X])^2$$

Calcoliamo per prima $E[X^2]$:

$$E[T^2] = \sum_{i=1}^n x_i^2 P_X(x_i) = \sum_x x^2 \theta (1-\theta)^{x-1}$$

Sfruttando l'artificio $x^2 = x(x-1) + x$ riscriviamo il tutto come due sommatorie

$$E[T^2] = \sum_x x(x-1)\theta(1-\theta)^{x-1} + \underbrace{\sum_x x\theta(1-\theta)^{x-1}}_{E[X]}$$

quindi ci concentriamo sulla prima sommatoria, applicando lo stesso ragionamento di prima possiamo fare in modo che dentro la sommatoria compaia la derivata in questo caso seconda di $(1-\theta)^x$:

$$(1-\theta)\theta \sum_x x(x-1)(1-\theta)^{x-2} = (1-\theta)\theta \sum_x \frac{d^2}{d\theta^2} [(1-\theta)^x]$$

Scambiando derivazione e sommatoria:

$$(1-\theta)\theta \frac{d^2}{d\theta^2} \sum_x [(1-\theta)^x] = (1-\theta)\theta \frac{d^2}{d\theta^2} \left[\frac{1}{1-(1-\theta)} \right]$$

Derivando e considerando l'altra sommatoria si ottiene:

$$E[T^2] = (1-\theta)\theta \cdot \frac{2}{\theta^3} + E[X] = \frac{2(1-\theta)}{\theta^2} + \frac{1}{\theta}$$

Quindi:

$$Var(T) = E[T^2] - (E[T])^2 = \frac{2(1-\theta)}{\theta^2} + \frac{1}{\theta} - \left(\frac{1}{\theta}\right)^2 = \frac{(1-\theta)}{\theta^2}$$

Per la geometrica traslata si sfrutta sempre la definizione e si ottiene grazie all'invarianza per traslazione della varianza:

$$Var(T') = Var(T-1) \underset{inv.\,trasl.}{=} Var(T) = \frac{(1-\theta)}{\theta^2}$$

In sintesi:

Geometrica	$T \sim geom(\theta)$	Geometrica Traslata	$T' \sim geom'(\theta)$
$E[T]$	$\frac{1}{\theta}$	$E[T']$	$\frac{1}{\theta} - 1$
$Var(T)$	$\frac{1-\theta}{\theta^2}$	$Var(T')$	$\frac{1-\theta}{\theta^2}$

5.4 Binomiale negativa

Introduciamo una nuova variabile aleatoria che chiamiamo **binomiale negativa** e indichiamo con $T_r \sim \text{bineg}(r, \theta)$; questa variabile aleatoria conta il numero T_r della prova in cui si ottengono r successi (non solo il primo). Evidentemente la binomiale negativa è somma di r variabili geometriche indipendenti. Nel caso di geometriche la distribuzione è data da:

$$T_n \sim \text{bineg}(r, \theta) \rightarrow P_{T_r}(k) = \binom{k-1}{r-1} \theta^r (1-\theta)^{k-r} I_{\{r,r+1,\dots\}}(k) \quad (5.8)$$

Ma da dove viene la formula? Se otteniamo alla prova k l' r -esimo successo, vuol dire che nelle $k-1$ prove precedenti si sono ottenuti $r-1$ successi, ma come non sappiamo in che ordine si sono verificati i successi, dobbiamo considerare tutte le possibili combinazioni, questo lo si fa con il coefficiente binomiale.

Poiché consideriamo r successi abbiamo θ^r e poiché si hanno $k-1-(r-1)=k-r$ insuccessi si ha $(1-\theta)^{k-r}$.

Alcuni autori scrivono la formula in modo diverso ma il significato resta invariato, di solito si tratta solo del coefficiente binomiale

$$\binom{k-1}{r-1} = \binom{k-1}{k-r} = \frac{(k-1)!}{(r-1)!(k-r)!}$$

Nel caso di geometriche traslate la binomiale negativa conta il numero di insuccessi ottenuti prima di ottenere r successi; la distribuzione è data da:

$$T'_r \sim \text{bineg}'(r, \theta) \rightarrow P_{T'_r}(k) = \binom{r+k-1}{k} \theta^r (1-\theta)^k I_{\{0,1,\dots\}}(k) \quad (5.9)$$

Per ottenere questa formula basta notare che il numero di prove necessarie per ottenere r successi è uguale al numero di insuccessi più il numero di successi cioè $T_r = T'_r + r \rightarrow T'_r = T_r - r$. Anche in questo caso il coefficiente binomiale ci esprime la stessa cosa che nel caso non traslato (le combinazioni dei successi su tutte le prove precedenti), solo che i successi ora sono r e le prove sono

$$\underbrace{k}_{\text{insucc.}} + \underbrace{r}_{\text{succ.}} - 1.$$

Si noti come se scriviamo la 5.8 nel modo alternativo, si ha in entrambi i casi:

$$T \sim \binom{\diamond + * - 1}{*} \theta^\diamond (1-\theta)^* I_{\{*=0,\dots\}}$$

5.4.1 Valore Atteso

$$E[T_r] = \underbrace{E}_{per.def.} \left[\sum_i^r T_i \right] = \underbrace{\sum_i^r E[T_i]}_{ind.} = \underbrace{\sum_i^r \frac{1}{\theta}}_{id.dist.} = \frac{r}{\theta}$$

Si dimostra in modo analogo il valore atteso della traslata.

5.4.2 Varianza

$$\text{Var}[T_r] = \underbrace{\text{Var}}_{\text{per.def.}} \left[\sum_i^r T_i \right] \underbrace{=}_{\text{ind.}} \sum_i^r \text{Var}[T_i] \underbrace{=}_{\text{id.dist.}} \sum_i^r \frac{(1-\theta)}{\theta^2} = \frac{r(1-\theta)}{\theta^2}$$

Si dimostra in modo analogo la varianza della traslata.

In sintesi:

Geometrica	$T \sim \text{geom}(\theta)$	Binomiale Negativa	$T_r \sim \text{bineg}(r, \theta)$
$E[T]$	$\frac{1}{\theta}$	$E[T_r]$	$\frac{r}{\theta}$
$\text{Var}(T)$	$\frac{1-\theta}{\theta^2}$	$\text{Var}(T_r)$	$\frac{r(1-\theta)}{\theta^2}$
Geometrica Traslata	$T' \sim \text{geom}(\theta)$	Binomiale Negativa	$T'_r \sim \text{bineg}(r, \theta)$
$E[T']$	$\frac{1}{\theta} - 1$	$E[T'_r]$	$r \left(\frac{1}{\theta} - 1 \right)$
$\text{Var}(T')$	$\frac{1-\theta}{\theta^2}$	$\text{Var}(T'_r)$	$\frac{r(1-\theta)}{\theta^2}$

5.5 Proprietà del modello bernoulliano

Le proprietà più importanti del modello bernoulliano sono:

PROPRIETÀ DI CHIUSURA per alcune variabili aleatorie, cioè:

- Sia $Y_1 \sim \text{bin}(n_1, \theta)$ e $Y_2 \sim \text{bin}(n_2, \theta)$ allora:

$$Y_1 + Y_2 \sim \text{bin}(n_1 + n_2, \theta) \quad (5.10)$$

- Sia $Y_1 \sim \text{bineg}(r_1, \theta)$ e $Y_2 \sim \text{bineg}(r_2, \theta)$ allora:

$$Y_1 + Y_2 \sim \text{bineg}(r_1 + r_2, \theta) \quad (5.11)$$

Per la variabile aleatoria geometrica (e anche per tutte le variabili di attesa) vale un'altra proprietà che viene chiamata *assenza di memoria* (memoryless property); e ci dice che il fatto che un successo non ancora successo non cambia la probabilità attendere tanto o poco tempo affinché avvenga, e come se ad ogni evento il processo “cancellassi” la sua memoria sul passato. In formule:

$$P(T_r \geq k+h | T_r \geq k) = P(T_r \geq h) \quad (5.12)$$

oppure:

$$P(T_r \geq k+h) = P(T_r \geq k) \cdot P(T_r \geq h) \quad (5.13)$$

5.6 Uniforme discreta

La variabile aleatoria uniforme discreta è una variabile ove le determinazioni assumono un numero finito di valori con stessa probabilità . La variabile uniforme discreta si indica $T \# \sim \text{unif}\{x_1 \dots x_k\}$

$$\begin{aligned} \text{Se } T \# \sim \text{unif}\{1 \dots n\} \quad P_{T \#}(x) &= \frac{1}{n} I_{\{1, \dots, n\}}(x) \\ (5.14) \end{aligned}$$

$$\text{Se } T \# \sim \text{unif}\{0 \dots n\} \quad P_{T \#}(x) = \frac{1}{n+1} I_{\{0, \dots, n\}}(x)$$

ESEMPIO

Un dado cubico si distribuisce in modo uniforme; ognuna delle 6 facce ha la stessa probabilità . Una moneta non truccata può essere considerata sia come una bernoulliana di parametro θ che come una uniforme, infatti entrambe le facce hanno la stessa probabilità .

5.6.1 Origini

Consideriamo la variabile $T \#$ come variabile di attesa condizionata al fatto che è avvenuto un successo in una delle n prove, per convenzione chiamiamo Y_n la rispettiva come variabile di conteggio e T la variabile di attesa.

$$P_{T \#}(x) = P(T \# = x) = P(T = x | Y_n = 1)$$

applicando il teorema di Bayes si ha:

$$P_{T \#}(x) = \frac{P(Y_n = 1 | T = x) P(T = x)}{P(Y_n = 1)}$$

A numeratore, l'espressione a sinistra vale 1 poiché sappiamo già che si è verificato un successo, mentre quella a destra si calcola come una geometrica. A denominatore si ha ovviamente una binomiale, poiché è la probabilità che la variabile di conteggio assuma un numero in particolare. Quindi:

$$P_{T \#}(x) = \frac{(1 - \theta)^{n-1} \theta}{\binom{n}{1} (1 - \theta)^{n-1} \theta} = \frac{1}{n} I_{\{1, \dots, n\}}(x)$$

5.6.2 Valore atteso

CASO $T \# \sim \text{unif}\{1 \dots n\}$

$$E[T \#] = \sum_x x P_{T \#}(x) = \sum_{x=1}^n x \frac{1}{n} = \frac{1}{n} \sum_{x=1}^n x = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

CASO $T \# \sim \text{unif}\{0 \dots n\}$

$$E[T \#] = \sum_x x P_{T \#}(x) = \sum_{x=0}^n x \frac{1}{n+1} = \frac{1}{n+1} \sum_{x=0}^n x = \frac{1}{n+1} \cdot \frac{n(n+1)}{2} = \frac{n}{2}$$

5.6.3 Varianza

Prima di calcolare la varianza, si può prevedere facilmente che la varianza aumenterà nel caso $T\# \sim \text{unif}\{0 \dots n\}$ poiché si considerano più elementi.

CASO $T\# \sim \text{unif}\{1 \dots n\}$

Ricordando la formula della somma dei primi n quadrati si ha:

$$E[T\#^2] = \sum_x x^2 P_{T\#}(x) = \sum_{x=1}^n x^2 \frac{1}{n} = \frac{1}{n} \sum_{x=1}^n x^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6}$$

$$\text{Var}(T\#) = E[T\#^2] - (E[T\#])^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \dots = \frac{n^2-1}{12}$$

CASO $T\# \sim \text{unif}\{0 \dots n\}$

$$E[T\#^2] = \sum_x x^2 P_{T\#}(x) = \sum_{x=1}^n x^2 \frac{1}{n+1} = \frac{1}{n+1} \sum_{x=0}^n x^2 = \frac{n(n+1)(2n+1)}{6(n+1)}$$

$$\text{Var}(T\#) = E[T\#^2] - (E[T\#])^2 = \frac{n(2n+1)}{6} - \frac{n^2}{4} = \dots = \frac{(n+1)^2-1}{12}$$

5.7 Campionamento e modelli statistici

Quello che cerca di fare la statistica è studiare una *popolazione*, cioè, un insieme non necessariamente di persone, ma un insieme di elementi che hanno qualche caratteristica in comune. Per studiare una popolazione ci serve un **campione casuale** (*sample*) che per noi sarà una n -upla di variabili indipendenti e identicamente distribuite, $\underline{X} = (X_1, X_2, \dots, X_n)$ da estrarre dalla popolazione, successivamente indicheremo con (x_1, x_2, \dots, x_n) i valori effettivamente osservati sul campione casuale. Quindi prima del *campionamento* si avranno variabili aleatorie indicate con maiuscole, dopo il *campionamento* si hanno valori numerici e la variabile che assumono questi numeri sono indicati con minuscole.

Come vedremo nel capitolo 7, a partire da questi campioni, è possibile congetturare sulle caratteristiche di tutta la popolazione e vedremo che più campioni si hanno o più numeroso è il campione, diventa sempre più probabile che le congetture fatte corrispondano alle vere caratteristiche della popolazione.

5.7.1 Somma e media campionaria

Consideriamo una nuova variabile, la variabile aleatoria "somma campionaria":

$$Y_n = \sum_{i=1}^n X_i \quad (5.15)$$

e la variabile aleatoria "media campionaria":

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (5.16)$$

E' facile vedere che la media campionaria non è altro che la somma campionaria diviso n . La variabile media campionaria ha delle proprietà particolari infatti:

$$E[\bar{X}_n] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} n\mu = \mu$$

Quindi il valore atteso della variabile aleatoria \bar{X}_n è proprio μ .

Questa variabile ha anche un'altra proprietà :

$$Var(\bar{X}_n) = \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i \right) =_{per\ ind.} \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Con questo risultato possiamo vedere che la varianza diminuisce al crescere di n , quindi quanto più grande sia il campione, minore sarà la dispersione dei dati e migliore sarà per congetturare sulla popolazione.

5.7.2 Legge debole dei grandi numeri / convergenza in probabilità

Consideriamo un campione casuale (X_1, X_2, \dots, X_n) dotato di valore atteso e varianza finita, grazie alla diseguaglianza di Cebyshev e ponendo $\eta = \sigma^2/t$ possiamo dire che:

$$\forall \varepsilon > 0 \quad \forall \eta > 0 \quad \exists \bar{n} : \forall n > \bar{n} \quad P(|\bar{X}_n - \mu| \leq \varepsilon) \geq 1 - \eta \quad (5.17)$$

In particolare per $n \rightarrow \infty$ si ha che:

$$\lim_{n \rightarrow +\infty} P(|\bar{X}_n - \mu| \leq \varepsilon) = 1 \quad (5.18)$$

In questo caso si dice che la successione di variabili aleatorie tende in probabilità a μ per $n \rightarrow \infty$. E si scrive:

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{P} \mu$$

In pratica la legge dei grandi numeri ci dice che è possibile che i dati ricavati non riflettano la distribuzione della probabilità, (si pensi ad esempio a 5 lanci di una moneta equiprobabile in cui nei cinque lanci è uscito "croce"), ma la probabilità che questo succeda diminuisce all'incrementare della grandezza del campione casuale.

ANALOGIA

Supponiamo che un amico abbia una moneta truccata, di cui ignoriamo le probabilità di ciascuna faccia, questo amico non ci dirà mai quanto vale la probabilità della sua moneta ma è disposto a lanciare la moneta e dirci se è uscito "testa" o "croce" un numero arbitrario di volte. Quindi noi possiamo scoprire la probabilità di ciascuna faccia con questi dati; infatti la 5.17 ci dice che fissato un errore di calcolo $\varepsilon > 0$ e noti μ, σ^2 possiamo trovare il numero minimo di lanci necessari \bar{n} che ci permette di ottenere la probabilità dentro il margine d'errore desiderato.

La 5.18 ci dice una cosa sorprendente, che se lanciamo la moneta un numero infinito di volte, avremo la probabilità esatta di entrambe le facce.

Si noti bene che per la legge dei grandi numeri si suppongono noti μ, σ^2 ma nella maggior parte dei casi, entrambi non sono noti. Come facciamo? La risposta è: *stimarli*, questo argomento verrà trattato nel capitolo 7.

6. Modelli probabilistici continui

Il concetto di probabilità è il più importante della scienza moderna, soprattutto perché nessuno ha la più pallida idea del suo significato.

(BERTRAND RUSSELL, 1872-1970)

Per introdurre il campo poissoniano possiamo farlo in due modi, noi lo facciamo per quello che lo lega al Campo Bernoulliano. Prima di ciò introduciamo un'ultima variabile aleatoria discreta.

Supponiamo di avere un esperimento bernoulliano con un parametro θ molto piccolo e n molto grande, la distribuzione binomiale è data dunque da:

$$P_{Y_n} = \binom{n}{k} \theta^k (1-\theta)^{n-k} I_{\{0,1,\dots,n\}}(k)$$

A questo punto poniamo $n\theta = \lambda$, riscrivendo la legge di distribuzione abbiamo:

$$P_{Y_n} = \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} I_{\{0,1,\dots,n\}}(k)$$

Ora facciamo tendere $n \rightarrow \infty$ supponendo che n sia molto grande; a questo punto sfruttiamo qualche asintotico e limite notevole e troviamo:

$$\lim_{n \rightarrow \infty} P_{Y_n} = \underbrace{\frac{n!}{(n-k)!n^k}}_{\approx 1} \underbrace{\frac{\lambda^k}{k!}}_{e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\approx 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\approx 1} I_{\{0,1,\dots,n\}}(k)$$

Ottenendo:

$$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda} I_{\{\mathbb{N}\}}(k) \quad (6.1)$$

Questa distribuzione prende il nome di **distribuzione di Poisson** in onore a Siméon-Denis Poisson, ed esprime in qualche modo la probabilità che occorra un numero determinato di eventi in un dato intervallo di tempo sapendo che in media si verifica un numero pari a: $\lambda = n\theta$; poiché viene usata in casi ove la probabilità degli eventi è molto piccola è anche nota come **legge degli eventi rari**.

La distribuzione di Poisson usa il valore $k!$ che non è facilmente calcolabile, infatti una calcolatrice standard riesce a calcolare solo fino $69!$; poiché l'approssimazione di Poisson coinvolge dati molto grandi, nasce dunque la necessità di calcolare il fattoriale anche quando k è molto grande.

6.0.1 Aproximazione di Stirling

L'aproximazione di Stirling ci permette di ottenere una stima accurata del fattoriale di un numero a patto che il numero sia molto grande.

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \quad (6.2)$$

6.1 Campo Poissoniano

La variabile poissoniana non è soltanto utile per approssimare binomiali ma può anche modellare fenomeni di "eventi/arrivi/successi nell'unità di tempo/lunghezza/superficie/volume", guardandolo in questo modo, ci basta porre $\lambda = v\tau$ ove τ è la misura di ampiezza del intervallo di "conteggio" o di tempo, e v l'intensità di quello che chiamiamo *campo poissoniano*, ossia il numero medio di successi/arrivi per unità di tempo. In questo contesto scriveremo che la probabilità che nell'intervallo di ampiezza τ si verifichino x successi si distribuisce con la legge:

$$P_N(x) = \frac{v\tau^x}{x!} e^{-v\tau} I_{\{0,\dots\}}(x) \quad (6.3)$$

Una delle particolarità più importanti del campo poissoniano è che non esiste più il concetto di prove come nel campo bernoulliano e tutte le altre proprietà sono descritte da quelli che chiamiamo assiomi del campo poissoniano o assiomi di Poisson.

6.2 Assiomi di Poisson

① Assioma di rarità 1.

$$P(N(0, \tau) = 1) = v\tau + o(\tau) \quad \text{per } \tau \rightarrow 0 \quad (6.4)$$

Questo assioma ci dice che la probabilità di effettuare un successo nell'intervallo $(0, \tau)$ è proporzionale all'ampiezza dell'intervallo.

② Assioma di rarità 2.

$$P(N(0, \tau) > 1) = o(\tau) \quad \text{per } \tau \rightarrow 0 \quad (6.5)$$

Questo assioma ci dice che la probabilità di contare più di un successo nell'intervallo $(0, \tau)$ è infima, infatti quando supponevamo $v\tau = \lambda = n\theta$; θ era molto piccolo.

(3) Assioma di indipendenza.

$$P[N(t_1, t_2) = k, N(t_3, t_4) = h] = P[N(t_1, t_2) = k] \cdot P[N(t_3, t_4) = h] \Leftrightarrow (t_1, t_2) \cap (t_3, t_4) = \emptyset \quad (6.6)$$

Questo assioma ci dice che la probabilità che ci siano k successi in un intervallo è h in un altro, è uguale al prodotto delle probabilità se e solo se gli intervalli non ricoprono alcun valore più di una volta.

(4) Assioma di omogeneità .

$$P(N(t, t + \tau) = k) = P(N(0, \tau) = k) \quad \forall t \in \mathbb{R} \quad (6.7)$$

Questo assioma ci dice che la probabilità che ci siano k successi è indipendente dall'intervallo scelto.

6.3 Poissoniana

A questo punto introduciamo una nuova variabile aleatoria N che ha distribuzione poissoniana e soddisfa gli assiomi di Poisson; questa variabile prende nome di *poissoniana*; diremo anche che N ha legge di Poisson e scriveremo $N \sim \text{poiss}(\lambda)$ o $N \sim P(\lambda)$ oppure $N \sim \text{poiss}(v\tau)$ poiché si ricordi $\lambda = v\tau$. Si noti che la variabile poissoniana è una variabile aleatoria discreta.

6.3.1 Valore atteso

Il valore atteso per la poissoniana si calcola con la formula per una infinità numerabile, cioè:

$$E[N] = \sum_{x=1}^{\infty} x P_N(x) = \sum_{x=1}^{\infty} x \frac{v\tau^x}{x!} e^{-v\tau}$$

Poiché il termine $e^{-v\tau}$ non dipende da x , può essere portato fuori dalla serie, e $x!$ può essere scritto come $x \cdot (x - 1)!$

$$E[N] = \sum_{x=1}^{\infty} x \frac{v\tau^x}{x!} e^{-v\tau} = e^{-v\tau} \sum_{x=1}^{\infty} x \frac{(v\tau)^x}{x \cdot (x - 1)!}$$

Ora scriviamo $(v\tau)^x$ come $(v\tau)^{x-1} \cdot (v\tau)$ e portiamo $(v\tau)$ fuori dalla serie.

$$E[N] = e^{-v\tau} (v\tau) \sum_{x=1}^{\infty} \frac{(v\tau)^{x-1}}{(x - 1)!}$$

A questo punto basta una cambiare l'indice della serie per ricondursi alla serie esponenziale.

$$E[N] = e^{-v\tau} (v\tau) \sum_{s=0}^{\infty} \frac{(v\tau)^s}{s!} = e^{-v\tau} (v\tau) e^{v\tau} = v\tau$$

6.3.2 Varianza

La varianza si calcola con la formula $Var(N) = E[N^2] - (E[N])^2$ di cui si deve calcolare ancora $E[N^2]$ quindi:

$$E[N^2] = \sum_{x=1}^{\infty} x^2 \frac{v\tau^x}{x!} e^{-v\tau}$$

A questo punto sfruttiamo il solito artificio $x^2 = x(x-1) + x$.

$$E[N^2] = \sum_{x=1}^{\infty} x(x-1) \frac{v\tau^x}{x!} e^{-v\tau} + \underbrace{\sum_{x=1}^{\infty} x \frac{v\tau^x}{x!} e^{-v\tau}}_{E[N]}$$

Alla prima parte della serie applichiamo lo stesso ragionamento che nel valore atteso

$$e^{-v\tau} \sum_{x=1}^{\infty} x(x-1) \frac{v\tau^x}{x \cdot (x-1) \cdot (x-2)!} = e^{-v\tau} (v\tau)^2 \sum_{x=2}^{\infty} \frac{(v\tau)^{x-2}}{(x-2)!}$$

Cambiando indice della serie $s = x - 2$

$$e^{-v\tau} (v\tau)^2 \sum_{s=0}^{\infty} \frac{(v\tau)^s}{(s)!} = e^{-v\tau} (v\tau)^2 e^{v\tau} = (v\tau)^2$$

Ottenendo finalmente che $E[N^2] = (v\tau)^2 + (v\tau)$; ora applicando la formula della varianza si ottiene
In sintesi

$$Var(N) = E[N^2] - (E[N])^2 = (v\tau)^2 + (v\tau) - (v\tau)^2 = v\tau$$

Variabile Poissoniana $N \sim poiss(v\tau)$

$$E[N] \quad \lambda = v\tau$$

$$Var(N) \quad \lambda = v\tau$$

Proprietà di chiusura

Sia $N_1 \sim poiss(\lambda_1)$ e $N_2 \sim poiss(\lambda_2)$ allora:

$$N_1 + N_2 \sim poiss(\lambda_1 + \lambda_2)$$

6.4 Variabili aleatorie continue

Le variabili aleatorie discrete sono molto utili per modellare tanti fenomeni ma ci sono alcune situazioni che non possono essere descritte da variabili discrete, come ad esempio il tempo di vita di una persona, il suo peso, la sua altezza, ecc. Ogniqualvolta si ha a che fare con una grandezza continua (lunghezze, volumi, velocità, tempo) il fenomeno sarà sempre rappresentato da una variabile continua.

6.4.1 PDF

Una variabile aleatoria continua T è definita da una funzione detta densità continua o anche PDF (Probability Density Function) definita come $f_T : \mathbb{R} \rightarrow \mathbb{R}$ che gode delle seguenti proprietà :

$$\int_{\mathbb{R}} f_T(t) dt = 1 \quad f_T(t) \geq 0 \quad \forall t \in \mathbb{R} \tag{6.8}$$

Analogamente al caso discreto si definisce:

1 Valore atteso

$$E[T] \stackrel{\text{def}}{=} \int_{\mathbb{R}} t f_T(t) dt$$

2 Varianza

$$\text{Var}(T) \stackrel{\text{def}}{=} \int_{\mathbb{R}} (t - E[T])^2 f_T(t) dt$$

3 Momento di indice k

$$E[T^k] \stackrel{\text{def}}{=} \int_{\mathbb{R}} t^k f_T(t) dt$$

Così come nel caso discreto la densità discreta ci dava la legge di ripartizione della probabilità della variabile aleatoria discreta, nel caso continuo è la densità continua chi ci dà la legge di ripartizione della probabilità , e si definisce come:

$$P(T \in I) := \int_I f_T(t) dt \quad \forall t \in I \tag{6.9}$$

Come conseguenza si ha che :

$$P(T = t) \stackrel{\text{def}}{=} 0 \quad \forall t \in \mathbb{R} \tag{6.10}$$

Cioè, la probabilità che una variabile aleatoria continua assuma un valore specifico è sempre nulla! Infatti se applichiamo la definizione 6.9, ciò equivale a calcolare l'integrale di un unico punto e per definizione d'integrale, questo è sempre zero. Quindi ha senso unicamente calcolare la probabilità che una variabile aleatoria T assuma valori in un intervallo di ampiezza positiva, non in punti prefissati!

Per questo motivo calcolare $P(a \leq T \leq b)$ è la stessa cosa che calcolare $P(a < T < b)$.

6.4.2 Correzione di continuità

Poiché l'integrale in un determinato punto è sempre nulla, ciò non vuol dire come nel caso discreto, l'evento risulti impossibile, ma impone che se si vuole calcolare la probabilità di un valore specifico, si deve creare un intervallo piccolo, di solito ampio 1, cioè ampio 0,5 a sinistra e 0,5 destra del valore in considerazione in modo da eliminare la nullità della probabilità .

$$P(T = x) \quad \rightarrow \quad P(x - 0,5 < T < x + 0,5)$$

6.4.3 CDF

Chiamiamo *funzione di distribuzione cumulativa* o anche CDF (Cumulative Distribution Function) o colloquialmente "contatore", la funzione che conta la probabilità fino a t :

$$F_T(t) \stackrel{\text{def}}{=} P(T \leq t) = P(T \in (-\infty, t]) \quad \forall t \in \mathbb{R} \tag{6.11}$$

V.A. DISCRETE

$$F_X(t) \stackrel{\text{def}}{=} \sum_{x_k \leq t} P_X(x_k)$$

V.A. CONTINUE

$$F_T(t) \stackrel{\text{def}}{=} \int_{-\infty}^t f_T(y) dy$$

Questa funzione ha alcune proprietà :

- 1** $F_T(t)$ è una funzione monotona crescente (non strettamente)

Questa è più che ovvia, infatti se stiamo accumulando probabilità , il “contatore” aumenta sempre o al limite rimane invariato, ma sicuramente non decremente mai, altrimenti ciò significherebbe che esiste la probabilità negativa (assurdo).

- 2** $F_T(t) \rightarrow 1$ per $t \rightarrow \infty$

- 3** $F_T(t) \rightarrow 0$ per $t \rightarrow -\infty$

- 4** $F_T(t)$ è la funzione integrale della rispettiva funzione densità , ed è dunque sempre continua e, se la densità è continua (solo nei casi v.a. continua), risulta anche derivabile e vale il secondo teorema fondamentale del calcolo integrale:

$$F'_T(t) = f_T(t) \quad (6.12)$$

- 5** $P(a < T \leq b) = F_T(b) - F_T(a)$

Queste proprietà hanno conseguenze molto utili, infatti dall’ultima si ha che:

$$\begin{aligned} P(t < T \leq t + \Delta\tau) &= F_T(t + \Delta\tau) - F_T(t) \\ &= \int_{-\infty}^{t+\Delta\tau} f_T(y) dy - \int_{-\infty}^t f_T(y) dy \\ &= \int_t^{t+\Delta\tau} f_T(y) dy \end{aligned} \quad (6.13)$$

Che ci dice che la probabilità è l’area sottesa alla funzione densità .

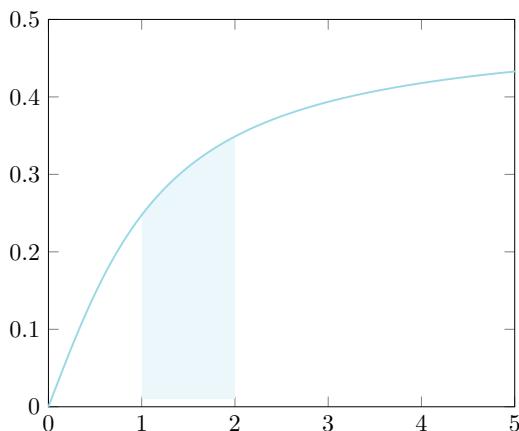
ESEMPIO

Calcolo della probabilità di una variabile aleatoria con **distribuzione di Cauchy**

$$f_T(y) = \frac{dy}{\pi(1+y^2)}$$

$$P(1 < T < 2) = \int_1^2 \frac{dy}{\pi(1+y^2)}$$

$$P(1 < T < 2) = \frac{1}{\pi} \arctan(2) - \arctan(1) = 0.321$$



6.5 Distribuzioni particolari

Per ottenere la densità della somma di due variabili indipendenti si usa l'integrale di convoluzione.

$$U = X + Y \quad \Rightarrow \quad f_U(u) = \int_{-\infty}^{\infty} f_X(x) f_Y(u-x) dx \quad (6.14)$$

Per ottenere la densità del prodotto di due variabili indipendenti si usa l'integrale:

$$U = XY \quad \Rightarrow \quad f_U(u) = \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y(u/x) dx \quad (6.15)$$

Per ottenere la densità del quoziente di due variabili indipendenti si usa l'integrale:

$$U = Y/X \quad \Rightarrow \quad f_U(u) = \int_{-\infty}^{\infty} |x| f_X(x) f_Y(ux) dx \quad (6.16)$$

6.5.1 Legge di trasformazione

Se $Y = g(X)$, possiamo scrivere la densità di questa variabile aleatoria come:

$$f_Y(y) = |\det \mathbf{J}| f_X(g^{-1}) \quad (6.17)$$

6.6 Variabili assolutamente continue e continue singolari

La funzione distribuzione cumulativa ci permette di suddividere le variabili aleatorie continue in due sottoclassi, le variabili aleatorie assolutamente continue e le continue singolari. Si chiameranno *variabili aleatorie assolutamente continue* quelle in cui vale sempre che:

$$\int_{-\infty}^t \frac{d}{dT} F_T(x) dx = F_T(t)$$

queste variabili aleatorie rivestono la maggior parte delle variabili aleatorie usate per descrivere casi di studio d'interesse; ahimè, non tutte le variabili aleatorie continue godono di questa proprietà, quelle che non la godono fanno parte di un'altra sottoclasse chiamata *variabili continue singolari*, queste variabili sono di solito esempi strani come la funzione di *Vitali-Cantor*:

La funzione di Vitali-Cantor è un esempio di funzione continua e crescente che assomiglia ad una scala con infiniti gradini, la particolarità risiede nel fatto che nonostante cresca in altezza progressivamente la sua pendenza è nulla (derivata zero).

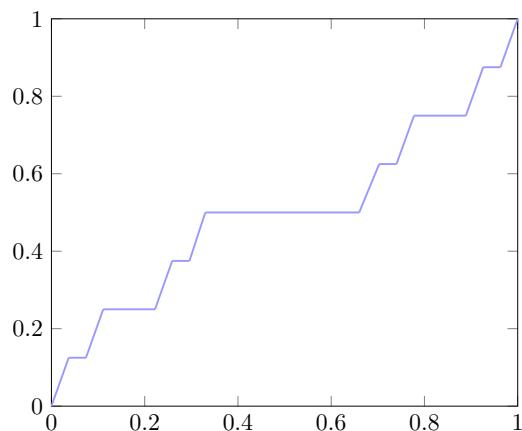


Figura 6.1: Funzione di Vitali-Cantor, (ovviamente non una scala al cielo)

6.7 Esponenziale

Così come nel caso discreto c'erano degli strumenti che ci permettevano di definire l'attesa e conteggio nel campo bernoulliano, possiamo ora definire degli strumenti analoghi per le variabili aleatorie continue. Chiamiamo **variabile aleatoria esponenziale** di parametro ν , $T \sim \exp(\nu)$ la variabile che misura il tempo del primo successo o "arrivo" in un processo di Poisson, cioè $T \leq \tau$ modella "il successo è avvenuto prima dell'istante τ ".

La variabile esponenziale modella l'attesa di un successo, non più in un campo bernoulliano, ma in un campo poissoniano, cioè, laddove si ha a che vedere con cose continue come dimensioni, tempo, peso, ecc. e valgono gli assiomi di Poisson.

Così come in tutte le altre variabili, si vuole capire come si distribuisce la probabilità di questa variabile aleatoria.

Consideriamo una binomiale che può essere approssimata mediante Poisson, ricordando che la binomiale conta il numero di successi, "il successo è avvenuto prima dell'istante τ " con $P(T \leq \tau)$ è l'evento complementare di "il successo non è avvenuto prima dell'istante τ " con $P(T \leq \tau) = 1 - P(T > \tau)$ che a sua volta è equivalente a considerare una binomiale con $k = 0$, con $P_{N_\tau}(0)$ (non ci sono stati

successi fino all'istante τ), dalla approssimazione si ha:

$$P(N_\tau = 0) = 1 - \frac{(\nu\tau)^0}{0!} e^{-\nu\tau}$$

Quindi :

$$T \sim \exp(\nu) \rightarrow F_T(\tau) = (1 - e^{-\nu\tau}) I_{[0,+\infty)}(\tau) \quad (6.18)$$

essendo continua la funzione, la densità continua è la derivata di $F_T(t)$:

$$T \sim \exp(\nu) \rightarrow f_T(\tau) = \nu e^{-\nu\tau} I_{[0,+\infty)}(\tau) \quad (6.19)$$

6.7.1 Valore atteso

Il valore atteso si calcola con la formula continua mettiamo t anziché τ (il significato non cambia).

$$E[T] = \int_{\mathbb{R}} t f_T(t) dt = \int_0^{\infty} t \nu e^{-\nu t} dt$$

Ora procediamo con l'integrazione per parti $\int u dv = uv - \int v du$

$$E[T] = \left[-te^{-\nu t} - \int -e^{-\nu t} dt \right]_0^{\infty} = \left[-te^{-\nu t} + \frac{e^{-\nu t}}{\nu} \right]_0^{\infty} = \frac{1}{\nu}$$

6.7.2 Varianza

La varianza si calcola con la solita formula $Var(T) = E[T^2] - (E[T])^2$, calcoliamo $E[T^2]$ sempre per parti.

$$E[T^2] = \int_{\mathbb{R}} t^2 f_T(t) dt = \int_0^{\infty} t^2 \nu e^{-\nu t} dt = \left[-t^2 e^{-\nu t} + \frac{2}{\nu} \left[\int t e^{-\nu t} dt \right] \right]_0^{\infty}$$

$$E[T^2] = \int_{\mathbb{R}} t^2 f_T(t) dt = \int_0^{\infty} t^2 \nu e^{-\nu t} dt = \left[-t^2 e^{-\nu t} + \frac{2}{\nu} \left[-te^{-\nu t} + \frac{e^{-\nu t}}{\nu} \right] dt \right]_0^{\infty} = \dots = \frac{2}{\nu^2}$$

$$Var(T) = E[T^2] - (E[T])^2 = \frac{2}{\nu^2} - \left(\frac{1}{\nu} \right)^2 = \frac{1}{\nu^2}$$

La variabile aleatoria esponenziale gode della proprietà di chiusura rispetto al prodotto per una costante positiva.

Come si riesce ad apprezzare dalla figura, l'esponenziale inizia dal reciproco della costante e tende sempre di più a zero al diminuire di α .

Se $T \sim \exp(\nu)$ allora:

$$\alpha T \sim \exp\left(\frac{\nu}{\alpha}\right)$$

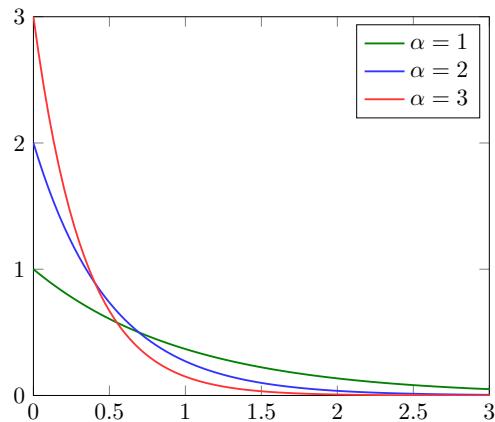


Figura 6.2:
Esponenziali di $\nu = 1$ e α variabile

6.7.3 Proprietà di assenza di memoria

Come previamente accennato, la variabile esponenziale gode della proprietà di assenza di memoria così come la geometrica, dato che è una variabile aleatoria che modella l'attesa. Cioè:

$$P(T > t + s | T > s) = P(T > t)$$

Questa proprietà è facilmente dimostrabile. Applicando il teorema di Bayes e notando che dato che il successo deve avvenire dopo dell'istante s e $t + s$ per soddisfare entrambe le disequazioni basta che il successo avvenga dopo l'istante $t + s$ (e ciò è modellato da $P(T > t + s)$).

$$P(T > t + s | T > s) = \frac{P(T > t + s, T > s)}{P(T > s)} = \frac{P(T > t + s)}{P_z(T > s)}$$

passando all'evento complemento per facilitare i conti si ha una esponenziale a destra del numeratore così come a parte destra del denominatore) :

$$P(T > t + s | T > s) = \frac{1 - P(T \leq t + s)}{1 - P(T \leq s)} = \frac{1 - (1 - e^{-\nu(t+s)})}{1 - (1 - e^{-\nu s})} = \frac{e^{-\nu(t+s)}}{e^{-\nu s}} = e^{-\nu t} = P(T > t)$$

6.8 Laplace

La distribuzione di Laplace, anche chiamata distribuzione esponenziale doppia è la distribuzione della variabile Y :

$$Y = T_1 - T_2 \quad (6.20)$$

con T_1, T_2 esponenziali dello stesso parametro v . Questa distribuzione ha:

$$E[Y] = 0 \quad \text{Var}(Y) = \frac{2}{v^2}$$

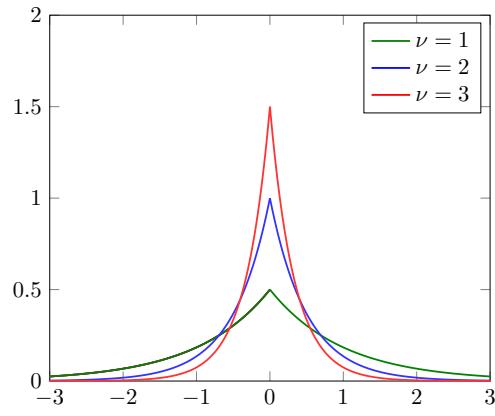


Figura 6.3: Distribuzione di Laplace al variare di ν

La sua densità è ricavabile dall'integrale di convoluzione

$$\begin{aligned} f_Y(y) &= \int_R f_{T_1}(x) f_{T_2}(x-y) dx \\ f_Y(y) &= \int_{\mathbb{R}} v e^{-vx} I_{[0,+\infty]}(x) v e^{-v(y-x)} I_{[0,+\infty)}(x-y) dx \end{aligned}$$

Infine si ottiene:

$$f_Y(y) = \frac{v}{2} e^{vy} I_{(-\infty,0)}(y) + \frac{v}{2} e^{-vy} I_{[0,+\infty)}(y) \quad (6.21)$$

6.9 Weibull

La distribuzione di Weibull è una distribuzione che ci dice il "tempo di vita" di un fenomeno che gode la proprietà di assenza di memoria.

Se $T \sim \exp(v)$ allora la variabile:

$$W = \left(\frac{T}{v} \right)^k \quad (6.22)$$

ha distribuzione di Weibull.

6.10 Rayleigh

Un campo poissoniano non necessariamente dev'essere unidimensionale, è anche possibile avere un campo bidimensionale e tridimensionale; in questi casi v rappresenterà l'intensità di un successo per unità di misura e τ rappresentera l'area o il volume di una porzione di campo.

La variabile aleatoria continua di tipo esponenziale in un campo bidimensionale prende il nome di *variabile aleatoria di Rayleigh* e il suo significato è di distanza del primo successo dall'origine. Per ottenere la sua distribuzione ci basta notare che, essendo in un campo bidimensionale, il successo si presenta ai bordi della circonferenza di area πt^2 , quindi:

$$R \sim \text{Rayleigh}(v) \rightarrow F_R(t) = 1 - e^{-v\pi t^2} I_{\mathbb{R}^+}(t) \quad (6.23)$$

Derivando la CDF si ottiene:

$$R \sim \text{Rayleigh}(v) \rightarrow f_R(t) = 2\pi v t e^{-v\pi t^2} I_{\mathbb{R}^+}(t) \quad (6.24)$$

Se si integra per parti si trova il valore atteso e la varianza:

Variabile Rayleigh $R \sim \text{rayleigh}(v)$

$$E[R] = \frac{1}{2\sqrt{v}}$$

$$Var(TR) = \frac{1}{v} \left(\frac{1}{\pi} - \frac{1}{4} \right)$$

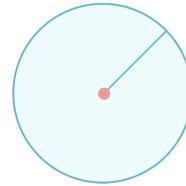


Figura 6.4: Successo più vicino in un campo bidimensionale

6.11 Maxwell

Esponenziale generalizzata ad uno spazio tridimensionale.

$$M \sim \text{Maxwell}(v) \rightarrow F_R(t) = 1 - e^{-v^{\frac{4}{3}}\pi t^3} I_{\mathbb{R}^+}(t) \quad (6.25)$$

Derivando la CDF si ottiene:

$$M \sim \text{Maxwell}(v) \rightarrow f_R(t) = 4\pi v t^2 e^{-v^{\frac{4}{3}}\pi t^3} I_{\mathbb{R}^+}(t) \quad (6.26)$$

6.12 Gamma

Chiamiamo **variabile aleatoria gamma** di parametro r e v la variabile aleatoria continua che misura il tempo in cui avviene l' r -esimo successo. Quindi è la somma di r esponenziali indipendenti, scriveremo $T_r \sim \Gamma(r, v)$.

$$T_r \sim \Gamma(r, v) \rightarrow F_T(\tau) = 1 - \sum_{k=0}^{r-1} e^{-v\tau} \frac{(v\tau)^k}{k!} I_{[0,+\infty)}(\tau) \quad (6.27)$$

e con densità continua:

$$T_r \sim \Gamma(r, v) \rightarrow f_T(\tau) = v e^{-v\tau} \frac{(v\tau)^{r-1}}{(r-1)!} I_{[0,+\infty)}(\tau) \quad (6.28)$$

Poiché la gamma è la somma di r esponenziali indipendenti, per $r = 1$ la gamma coincide con l'esponenziale.

In figura sono rappresentate diverse distribuzioni al variare di r , si può facilmente notare che le distribuzioni si "appiattiscono" (questo per conservare la proprietà che l'area sottesa, cioè la probabilità dei tutti i casi, rimanga pari a 1), e che all'aumentare di r le distribuzioni si spostano via a via verso destra, assumendo come massimo il valore $r - 1$.

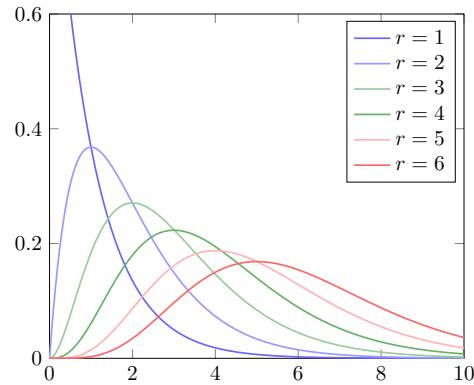


Figura 6.5: Gamma al variare di r

Per r sufficientemente grande, che dipende dal parametro v , la distribuzione assume una forma molto simile alla "campana di Gauss". (questo argomento verrà trattato nella sezione 6.15.1). Per $r \rightarrow \infty$ la probabilità si ripartisce in modo praticamente uniforme (cfr. 6.18) (infatti si appiattisce così tanto che praticamente diventa una linea orizzontale).

6.12.1 Valore atteso

$$E[T_r] \underset{\text{def.gamm.}}{=} E \left[\sum_{i=1}^r T_i \right] \underset{\text{ind.}}{=} \sum_{i=1}^r E[T_i] \underset{\text{id.distrib.}}{=} \sum_{i=1}^r \frac{1}{v} = \frac{r}{v}$$

6.12.2 Varianza

$$\text{Var}(T_r) \underset{\text{def.gamm}}{=} \text{Var} \left[\sum_{i=1}^r T_i \right] \underset{\text{ind.}}{=} \sum_{i=1}^r \text{Var}[T_i] \underset{\text{id.distrib.}}{=} \sum_{i=1}^r \frac{1}{v^2} = \frac{r}{v^2}$$

6.12.3 Proprietà di chiusura

Lipsum

In sintesi:

Variabile Esponenziale $T \sim \exp(\nu)$

$$E[T] = \frac{1}{\nu}$$

$$Var(T) = \frac{1}{\nu^2}$$

Variabile Gamma $T_r \sim \Gamma(r, \nu)$

$$E[T_r] = \frac{r}{\nu}$$

$$Var(T_r) = \frac{r}{\nu^2}$$

6.13 Standardizzata

Data una qualunque variabile X possiamo definire la variabile "centrata" come:

$$X' = X - E[X] \quad (6.29)$$

Così come la *standardizzata*

$$X^* = \frac{X - E[X]}{\sqrt{Var(X)}} \quad (6.30)$$

La standardizzata della somma campionaria di un campione casuale è molto particolare:

$$Y_n^* = \frac{Y_n - E[Y_n]}{\sqrt{Var(Y_n)}} = \frac{Y_n - n\mu}{\sqrt{n}\sigma}$$

se dividiamo sia il numeratore che il denominatore per n si ha:

$$Y_n^* = \frac{\frac{Y_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Ma qual è la particolarità di questo risultato? Che $\frac{Y_n}{n}$ è la media campionaria, eppure μ e $\frac{\sigma}{\sqrt{n}}$ che compaiono nell'espressione sono rispettivamente $E[\bar{X}_n]$ e $\sqrt{Var(\bar{X}_n)}$. Quindi la standardizzata del campione casuale è identica alla standardizzata della media campionaria:

$$Y_n^* \equiv (\bar{X}_n)^* \quad (6.31)$$

6.14 Teorema centrale del limite

È arrivato il momento di enunciare uno dei teoremi più importanti nell'ambito della teoria della probabilità.

Sia $Y_n = X_1 + \dots + X_n$ un campione casuale (ed ivi indipendenti e identicamente distribuiti) di una variabile aleatoria X e sia Y_n^* la rispettiva standardizzata; allora se $\exists E[X] = \mu$ eppure $\exists Var(X) = \sigma^2$ entrambi finiti, allora:

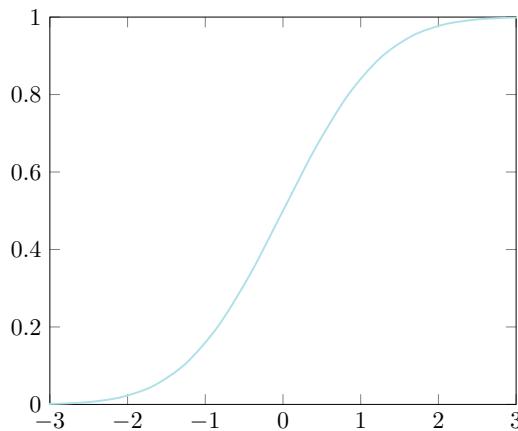
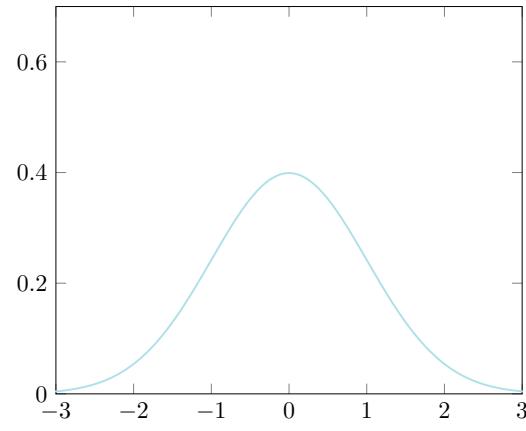
$$F_{Y_n^*}(x) \xrightarrow{n \rightarrow \infty} \Phi(x) \quad (6.32)$$

ove

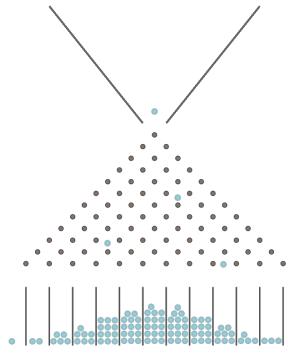
$$\begin{aligned} \Phi(x) &\stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy \\ \varphi(x) &\stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \end{aligned} \quad (6.33)$$

NB

La scelta di enunciare il teorema rispetto alla CDF e non rispetto alla funzione densità è che definito così, il teorema vale anche per le variabili aleatorie discrete.

Figura 6.6: Funzione contatore $\Phi(x)$.Figura 6.7: Funzione densità $\varphi(x)$; comunemente nota come **campana gaussiana**.

Questo teorema fu dimostrato sperimentalmente da Francis Galton nel caso in cui la distribuzione normale approssima quella di una binomiale al aumentare la numerosità del campione.

Figura 6.8: Macchina costruita da Galton chiamata *scatola di Galton* o *quinconce*.

La dimostrazione sperimentale consiste in una tavola verticale sulla quale sono posizionati dei chiodi configurati in modo analogo al triangolo di Tartaglia, in modo tale che, al far cadere dalla fessura delle palline, ogni pallina abbia 0,5 come probabilità di andare a destra e 0,5 di probabilità di andare a sinistra. Cioè si tratta di una binomiale $\theta = 0,5$.

6.14.1 Convergenza in legge

Il teorema può essere enunciato in un altro modo; sia Z una variabile aleatoria che ha come densità φ , diciamo che la variabile ha **distribuzione normale** e la si scrive $Z \sim N(0, 1)$ (0 è il valore atteso e 1 la varianza). Allora sotto le stesse ipotesi di prima si ha:

$$(Y_n)^* \xrightarrow[n \rightarrow \infty]{L} Z \quad (6.34)$$

la L significa che la *convergenza è in legge* è ciò non significa altro che hanno la stessa CDF al divergere di n , cioè siano A e B variabili aleatorie, la convergenza in legge significa:

$$A \xrightarrow[n \rightarrow \infty]{L} B \quad \stackrel{\text{def}}{=} \quad F_A(x) \xrightarrow[n \rightarrow \infty]{} F_B(x) \quad (6.35)$$

NB Poiché si è visto nella 6.13 che $Y_n^* \equiv (\bar{X}_n)^*$ si possono riformulare le due versioni in termini della standardizzata della media campionaria.

6.15 Normale

Sapendo che $Y_n^* \xrightarrow[n \rightarrow \infty]{L} Z$ non è possibile riformulare il tutto in modo di non avere a che fare con la standardizzata? La risposta ovviamente è sì.

Ricordando la definizione di standardizza lo possiamo anche scrivere come

$$Y_n \simeq \sigma_{Y_n} Y_n^* + E[Y_n]$$

Per il teorema centrale del limite:

$$Y_n = \sigma_{Y_n} Z + E[Y_n]$$

A questo punto possiamo generalizzare a una variabile W qualunque, tanto non cambia niente per ipotesi del teorema, riscriviamo la funzione densità :

$$f_{\sigma Z + \mu}(x) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2} \quad (6.36)$$

Una qualunque variabile W ,

$$W \rightsquigarrow N(\mu_W, \sigma_W^2)$$

ove il simbolo \rightsquigarrow si legge: "è asintoticamente distribuito come".

Quanto detto significa che qualunque sia la nostra variabile, se $n \rightarrow \infty$ la distribuzione della variabile sarà praticamente identica a una campana gaussiana, centrata in μ ed "appiattata" a seconda della varianza. Quando n è sufficientemente grande, si può supporre la distribuzione della variabile uguale a quella di $N(\mu_W, \sigma_W^2)$, questa approssimazione è nota sotto il nome di *approssimazione normale*.

Ovviamente possiamo applicare questo risultato anche alla somma campionaria Y_n che si distribuirà asintoticamente come:

$$Y_n \rightsquigarrow N(n\mu_X, n\sigma_X^2)$$

e alla media campionaria che si distribuirà come:

$$\bar{X}_n \rightsquigarrow N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

ove certamente μ_X e σ_X^2 variano a seconda della distribuzione delle variabili in considerazione.

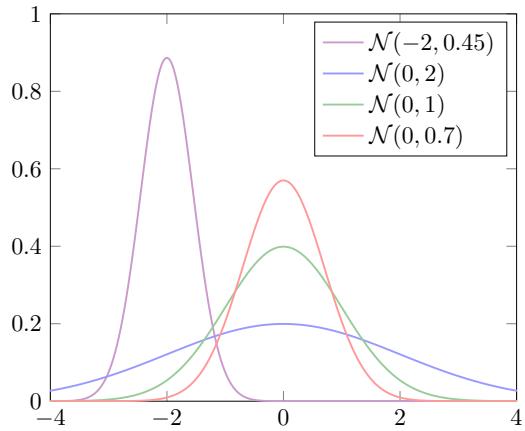


Figura 6.9: Campane Gaussiane

Possiamo confermare facilmente quanto detto con una distribuzione poissoniana, infatti quanto più λ è grande, quanto più la densità poissoniana assume un aspetto simmetrico rispetto al valore medio e quanto più assomiglia alla campana gaussiana della distribuzione normale, ai fini pratici per $\lambda >> 15$ possiamo considerare $poiss(\lambda) \rightsquigarrow N(\lambda, \lambda)$.

6.15.1 Proprietà della normale

Analizziamo alcune proprietà della variabile normale $N(\mu, \sigma^2)$.

La probabilità che la variabile assuma valori nell'intervallo (a, b) è pari a:

$$P(a < T < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \Phi(b) - \Phi(a) \quad (6.37)$$

oppure dalla diseguaglianza di Cebicev si ha:

$$P(|X - \mu| \leq \sigma t) = P\left(\frac{|X - E[X]|}{\sigma} \leq t\right)$$

che per definizione è il contatore della variabile: $F_{\frac{|X-\mu|}{\sigma}}(t)$ che per la normale è $\Phi(t)$. Notando che:

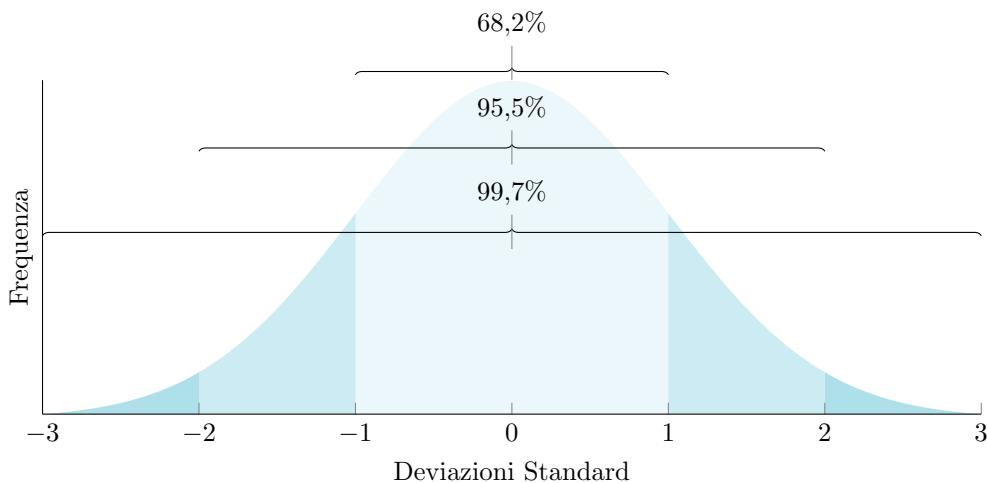
$$\Phi(-t) = 1 - \Phi(t)$$

Si può concludere che:

$$P\left(\frac{|X - E[X]|}{\sigma} \leq t\right) = 2\Phi(t) - 1$$

Con questa formula possiamo verificare che per $N(\mu, \sigma^2)$ si verifica che:

- Il 68,2% della probabilità si trova nell'intervallo $\mu - \sigma \leq x \leq \mu + \sigma$.
- Il 95,5% della probabilità si trova nell'intervallo $\mu - 2\sigma \leq x \leq \mu + 2\sigma$.
- Il 99,7% della probabilità si trova nell'intervallo $\mu - 3\sigma \leq x \leq \mu + 3\sigma$.



NB

Nella distribuzione normale la media, moda e mediana coincidono.

6.15.2 Curtosi

La curtosi, così come l'indice di assimetria incontrato in precedenza (5.2.3) è un indice statistico relativo alla forma assunta dalla distribuzione.

La curtosi misura quanto una distribuzione si allontana dalla distribuzione normale; cioè, quantifica quanto la distribuzione di discosta dalla forma a campana gaussiana. L'indice più noto per la sua misurazione è *l'indice di Pearson*.

$$\text{Curtosi}[X] = \frac{\mu'_4}{(\mu'_4)^2} \quad (6.38)$$

- Se $\text{Curtosi}[X] > 3$ la distribuzione è più "appuntita" di una normale e si dice **lepticurtica**.
- Se $\text{Curtosi}[X] < 3$ la distribuzione è più "piatta" di una normale e si dice **platicurtica**
- Se $\text{Curtosi}[X] = 3$ la distribuzione è "piatta" come una normale e si dice **mesocurtica**.

6.15.3 Momenti

lipsum

6.15.4 Chiusura rispetto alla somma

Siano X e Y variabili aleatorie con distribuzione $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$:

$$\text{se } Z = X + Y \quad \text{allora} \quad Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \quad (6.39)$$

6.15.5 Somma di due distribuzioni normali

La somma di due distribuzioni normali non deve confondersi in alcun modo con la somma di due variabili aleatorie con distribuzioni normale, nel caso in cui due distribuzioni normali si sommino si ha come risultato una **mistura di distribuzioni** come quella in figura.

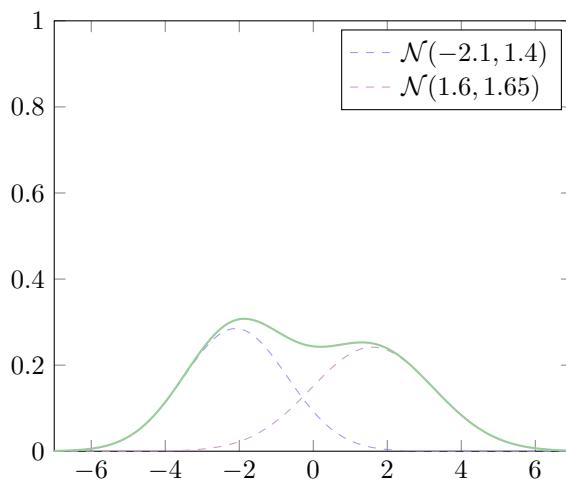


Figura 6.10: Non è un cappello né un boa che digerisce un elefante, è la somma di due normali! (cfr. Il Piccolo Principe)

6.16 Chi-Quadro

Siano Z_i variabili normali $N(0, 1)$ e sia $W = Z_1^2 + Z_2^2 + \dots + Z_n^2$, allora la variabile si distribuisce come una chi-quadro a n gradi di libertà e si scriverà $W \sim \chi_{(n)}^2$.

NB

Il nome gradi di libertà è solo un nome complicato per il parametro della distribuzione.

La variabile $\chi_{(1)}^2$ coincide con la distribuzione di $\Gamma(\frac{1}{2}, \frac{1}{2})$, infatti l'espressione:

$$\chi_{(1)}^2 \sim \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y} y^{\frac{1}{2}-1} I_{\mathbb{R}^+}(y)$$

è una gamma con parametri $r = \frac{1}{2}$ e $v = \frac{1}{2}$

$$ve^{-vy} \frac{(vy)^{r-1}}{(r-1)!} I_{[0,+\infty)}(y) = \frac{v^r}{(r-1)!} e^{-vy} y^{r-1} I_{\mathbb{R}^+}(y)$$

NB

L'ultima espressione richiede il calcolo del fattoriale di un numero non intero, il suo risultato dev'essere calcolato con una funzione speciale chiamata *gamma di Eulero* (che però richiede conoscenze avanzate di matematica, per cui il suo utilizzo esulerà gli scopi di questo libro).

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt \quad (6.40)$$

In realtà tutta questa complicazione è che legando la chi-quadro a la gamma (godendo la gamma della proprietà di chiusura per variabili indipendenti) possiamo ricondurre tutto questo concetto complicato come segue:

$$\chi_{(n)}^2 = \sum_i^n \chi_{(1)}^2 = \sum_i^n \Gamma\left(\frac{1}{2}, \frac{1}{2}\right) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

Quindi Chi-Quadro non è altro che un caso particolare della gamma! Sempre grazie alla gamma scopriamo che:

$$E(W) = n \quad Var(W) = 2n$$

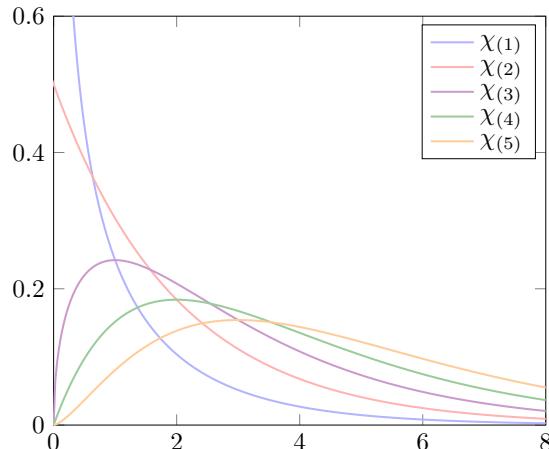


Figura 6.11: Densità della Chi-Quadro

NB

Si noti che per $n = 1$ la densità risulta infinita, ma ciò non implica che ci sia "massa concentrata".

$$\begin{array}{ccc} \text{massa concentrata} & \Rightarrow & \text{densità infinita} \\ \text{densità infinita} & \not\Rightarrow & \text{massa concentrata} \end{array} \quad (6.41)$$

6.17 Student - t

La variabile T di Student è un'altra variabile legata alla normale, ed è stata scoperta da William Gosset. Gosset era uno statistico alla Guinness e faceva spesso delle pubblicazioni scientifiche. La birreria irlandese, che temeva che la concorrenza venisse a sapere che le statistiche delle pubblicazioni di Gosset erano quelle della fabbrica costrinse Gosset a pubblicare, anziché con il suo vero nome, sotto uno pseudonimo. Gosset scelse il pseudonimo di Student. Da qui il nome della variabile.

Se la variabile:

$$T = \frac{Z}{\sqrt{\frac{W}{n}}} \Rightarrow T \sim t_{(n)}$$

ove $W \sim \chi^2_{(n)}$ e $Z \sim N(0, 1)$; si dice che la distribuzione ha n gradi di libertà .

La distribuzione $t_{(n)}$ ha delle particolari proprietà :

- ① E' simmetrica come la normale
- ② Ha le code più "grasse" della normale (Heavier tailed function)
- ③ Per $n = 1$ la distribuzione è la distribuzione di Cauchy

$$f_T(y) = \frac{1}{\pi(1+y^2)}$$

- ④ Per $n = k$ la distribuzione non ammette momenti a partire dal k -esimo quindi:
per $n = 1 \Rightarrow \exists E[T]$
per $n = 2 \Rightarrow \exists Var(T)$
- ⑤ Per $n \geq 2$ si ha che $E[T] = 0$

La proprietà più importante è che la distribuzione tende alla normale all'aumentare di n quindi:

$$T \xrightarrow[n \rightarrow \infty]{L} Z \quad (6.42)$$

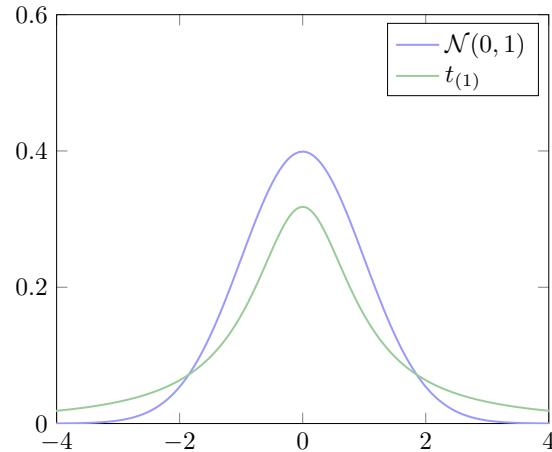


Figura 6.12: Confronto della densità normale e di student

6.18 Uniforme continua

La variabile uniforme è una variabile aleatoria (in questo caso continua) che si distribuisce uniformemente, questa variabile aleatoria è il legame tra il concetto di conteggio e il concetto d'attesa, come vedremo, una variabile che ha distribuzione uniforme si può vedere come un'attesa condizionata da un conteggio.

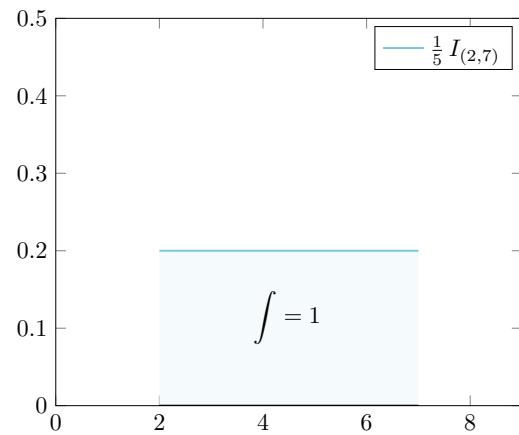
Ma cosa significa distribuzione uniforme? Il concetto di distribuzione uniforme è molto intuitivo; consideriamo il caso in cui in un campo poissoniano si sia verificato un unico successo nell'intervallo $[0, t]$, quello che ci dice la distribuzione uniforme è che ogni istante di tempo nell'intervallo considerato ha la stessa probabilità di "ospitare" il successo, cioè, non esiste un motivo per il quale un istante di tempo risulti "privilegiato" e si verifichi il successo il quel preciso istante.

ESEMPIO

A destra è riportata la distribuzione di una variabile di tipo uniforme nell'intervallo $(2, 7)$

$$f_{T\#}(t) = \frac{1}{5} I_{(2,7)}(t)$$

La particolarità di tutte le variabili uniformi è che il grafico della funzione densità è sempre una linea orizzontale e quello del contatore ha sempre la stessa pendenza.



Indicheremo con $T\# \sim \text{unif}(0, \tau)$ una variabile che si distribuisce uniformemente nell'intervallo $(0, \tau)$. Di solito si indica la densità uniforme su l' intervallo $(0, \tau)$ ma può essere indicato su un generico intervallo (a, b) . La funzione di densità è :

$$f_{T\#}(t) = \frac{1}{\tau} I_{(0,\tau)}(t) \quad \text{oppure} \quad f_{T\#}(t) = \frac{1}{b-a} I_{(a,b)}(t) \quad (6.43)$$

È anche possibile calcolare la probabilità che la variabile aleatoria assuma valori in un sotto-intervallo J , questa probabilità risulta essere data da quanto sia lungo J rispetto all'intervallo (a, b) .

$$P(T \in J) = \int_J \frac{1}{b-a} I_{(a,b)}(t) dt = \frac{1}{b-a} \cdot |(a, b) \cap J| \quad (6.44)$$

6.18.1 Origini

Consideriamo la variabile $T\#$ come variabile di attesa condizionata al fatto che è avvenuto un successo nell'intervallo $(0, \tau)$, per convenzione chiamiamo N la rispettiva come variabile di conteggio e T la variabile di attesa.

$$F_{T\#}(x) = P(T\# \leq x) = P(T \leq x | N_\tau = 1)$$

applicando il teorema di Bayes si ha:

$$F_{T\#}(x) = \frac{P(T \leq x, N_\tau = 1)}{P(N_\tau = 1)} = \frac{P(N_x = 1, N(x, \tau) = 0)}{P(N_\tau = 1)}$$

Applicando l'assioma di indipendenza (terzo assioma di Poisson) possiamo scrivere:

$$F_{T^\#}(x) = \frac{P(N_x = 1) P(N(\tau-x) = 0)}{P(N_\tau = 1)}$$

A numeratore, l'espressione a sinistra è una poissoniana di conteggio pari a 1; quella a destra è sempre una poissoniana di conteggio nullo nell'intervallo $(\tau-x, \tau)$ quindi l'ampiezza dell'intervallo è $\tau - x$. A denominatore è evidente che si ha ancora una volta poissoniana di conteggio pari a 1 con ampiezza dell'intervallo pari a x . Mettendo tutto insieme si ha:

$$F_{T^\#}(x) = \frac{\frac{(vx)^1}{1!} e^{-vx} \cdot \frac{(v(\tau-x))^0}{0!} e^{-v(\tau-x)}}{\frac{(v\tau)^1}{1!} e^{-v\tau}} = \frac{x}{\tau} I_{[0,\tau]}(x) + I_{(\tau,+\infty)}(x)$$

per ottenere la densità non ci basta altro che derivare la CDF:

$$f_{T^\#}(x) = \frac{1}{\tau} I_{(0,\tau)}(x)$$

6.18.2 Valore atteso

Per calcolare il valore atteso della uniforme continua si deve soltanto applicare la definizione

$$E[T^\#] = \int_I t f_{T^\#}(t) dt = \int_a^b \frac{t}{b-a} = \left[\frac{t^2}{2(b-a)} \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2} \quad (6.45)$$

Per una variabile $W \sim \text{unif}[0, 1]$ si ha $E[W] = \frac{1}{2}$

6.18.3 Varianza

La varianza si calcola nel solito modo $\text{Var}[T^\#] = E[T^{\#2}] - (E[T^\#])^2$, quindi calcoliamo $E[T^{\#2}]$

$$E[T^{\#2}] = \int_I t^2 f_{T^\#}(t) dt = \int_a^b \frac{t^2}{b-a} = \left[\frac{t^3}{3(b-a)} \right]_a^b = \dots = \frac{1}{3} (a^2 + ab + b^2)$$

Mettendo assieme i risultati precedenti si trova:

$$\text{Var}[T^\#] = \frac{1}{3} (a^2 + ab + b^2) - \left(\frac{a+b}{2} \right)^2 = \frac{(a-b)^2}{12} = \frac{(b-a)^2}{12} \quad (6.46)$$

Per una variabile $W \sim \text{unif}[0, 1]$ si ha $\text{Var}[W] = \frac{1}{12}$

6.18.4 Proprietà di riscrittura

Una qualunque variabile uniforme si può riscrivere come combinazione lineare di una variabile uniforme $X \sim \text{unif}[0, 1]$. Se $Y \sim \text{unif}[a, b]$ allora:

$$Y = (b-a)X + a$$

Ove il coefficiente moltiplicativo realizza una dilatazione/contrazione e il coefficiente a "sposta" l'inizio dell'indicatrice.

6.18.5 Somma di uniformi

La somma di n uniformi non è più una uniforme, per $n = 2$ è triangolare e per $n \geq 3$ la distribuzione assume una forma sempre più simile a una campana gaussiana.



In realtà è possibile generare una distribuzione praticamente identica alla gaussiana con la somma di soltanto $12 X$ con $X \sim \text{unif}[-\frac{1}{2}, \frac{1}{2}]$

6.19 Generare uniformi

Sia X una variabile aleatoria continua con una qualsiasi distribuzione. Se $F_X(\cdot)$ è la funzione di ripartizione della variabile (CDF), allora la nuova variabile aleatoria L definita come $L = F_X(X)$ si distribuisce come una uniforme.

$$\begin{aligned} F_L(x) &= P(L \leq x) = P(F_X(X) \leq x) = P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(x)) = P(X \leq F_X^{-1}(x)) \\ &= F_X(F_X^{-1}(x)) = x \end{aligned}$$

Dato che un contatore (CDF) deve valere come minimo 0 e al massimo 1, si ha che:

$$F_L(x) = xI_{[0,1]}(x) + I_{(1,\infty)}(x)$$

che non è altro che la CDF della variabile uniforme continua.

6.20 Da uniformi a una variabile qualunque

Se V è una variabile aleatoria continua con una qualsiasi distribuzione e $F_V(\cdot)$ è la funzione di ripartizione della variabile (CDF), la nuova variabile aleatoria $S = F_V^{-1}(X)$ con $X \sim \text{unif}[0, 1]$ si distribuisce come la variabile V .

$$\begin{aligned} F_S(x) &= P(S \leq x) = P(F_V^{-1}(X) \leq x) = P(F_V(F_V^{-1}(X)) \leq F_V(x)) = P(X \leq F_V(x)) \\ &= F_X(F_V(x)) = F_V(x) \end{aligned}$$

Questo risultato ha conseguenze molto importanti. Ogniqualvolta si è davanti a un campione casuale \underline{X} ottenuto da una popolazione uniforme, è possibile generare da questo campione un altro che abbia una qualunque distribuzione desiderata purché si trasformino i dati rilevati con la funzione inversa della CDF.

ESEMPIO

Per generare un campione di distribuzione esponenziale basta fare la trasformazione:

$$-\frac{1}{\nu} \ln(1 - X) \quad X \sim \text{unif}[0, 1]$$

ove $-\frac{1}{\nu} \ln(1 - x)$ è la funzione inversa di $y = 1 - e^{-\nu x}$

6.21 Massimo

Così come nel caso discreto chiamiamo massimo di funzioni aleatorie continue la variabile:

$$X_{(n)} = \max_i \{X_i\} \quad (6.47)$$

ove le X_i sono variabili aleatorie ottenute da un campione casuale di qualunque distribuzione.

6.21.1 "Contatore" (CDF) e funzione densità

Proviamo a calcolare qual è la funzione di distribuzione cumulativa della variabile aleatoria massimo, comunemente nota come CDF o "contatore".

Per definizione di questa è pari a:

$$F_{X_{(n)}}(x) = P(X_{(n)} \leq x) = P(\max\{X_1 \dots X_n\} \leq x)$$

Se ci concentriamo sull'ultima espressione, è facile vedere che dentro le parentesi l'espressione non ci dice altro che il massimo è minore o uguale di x , logicamente questo vuol dire che tutte le variabili X_1, \dots, X_n sono anche minori di x per cui possiamo riscrivere l'espressione come:

$$F_{X_{(n)}}(x) = P(X_1 \leq x, \dots, X_n \leq x) \underset{\text{ind.(c.c)}}{=} P(X_1 \leq x) \cdot \dots \cdot P(X_n \leq x)$$

Ora possiamo riapplicare la definizione di contatore a ogni singola variabile, trovando:

$$F_{X_{(n)}}(x) = F_{X_1}(x) \cdot \dots \cdot F_{X_n}(x)$$

Sapendo però che per definizione di campione casuale le variabili aleatorie sono identicamente distribuite, il contatore della variabile X_1 è lo stesso di quello di $X_2 \dots X_n$, quindi:

$$X_{(n)} = \max_i \{X_i\} \quad \rightarrow \quad F_{X_{(n)}}(x) = [F_X(x)]^n \quad (6.48)$$

Per trovare la funzione densità non ci basta altro che applicare la definizione, cioè derivare la CDF.

$$X_{(n)} = \max_i \{X_i\} \quad \rightarrow \quad f_{X_{(n)}}(x) = n [F_X(x)]^{n-1} f_X(x) \quad (6.49)$$

Massimo per uniformi

Per quanto detto questa definizione è sempre valida, qualunque sia la distribuzione della variabile X . Consideriamo il caso in cui $X \sim \text{unif}[0, \theta]$.

Per quanto detto nella sezione 6.18 la variabile uniforme assume come funzione densità :

$$f_X(x) = \frac{1}{\theta} I_{[0, \theta]}(x)$$

e funzione "contatore":

$$F_X(x) = \int_0^x \frac{1}{\theta} dt = \frac{x}{\theta} I_{[0, \theta]}(x) + I_{(\theta, +\infty)}(x)$$

Quindi la funzione contatore del massimo è:

$$F_{X_{(n)}}(x) = [F_X(x)]^n = \frac{x^n}{\theta^n} I_{[0,\theta]}(x) + I_{(\theta,+\infty)}(x)$$

Dalla figura si apprezza che all'aumentare di n il contatore tende alla funzione scalino o funzione di Heaviside traslata di θ . Questo vuol dire che se \underline{X} è campione casuale da $X \sim \text{unif}[0, \theta]$

$$X_{(n)} \xrightarrow[n \rightarrow \infty]{L} \theta$$

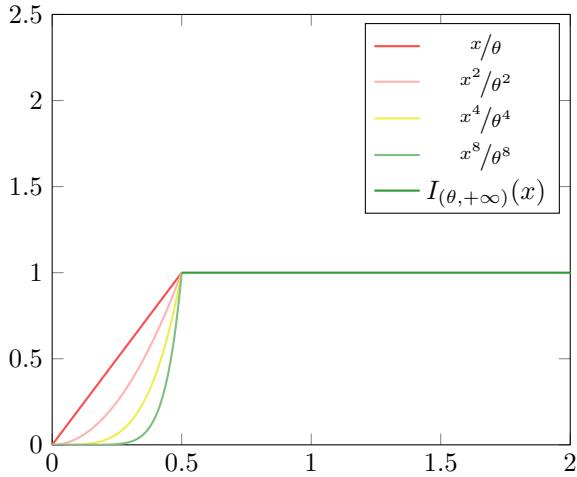


Figura 6.13: Caso con $\theta = 0.5$

La figura ci mostra che il grafico è sempre meno "sibilanciato" al crescere di n per cui:

$$E[X_{(n)}] \xrightarrow[n \rightarrow \infty]{} \theta$$

La funzione densità non è altro che la derivata del contatore già trovato:

$$f_{X_{(n)}}(x) = n \left(\frac{x}{\theta} \right)^{n-1} \frac{1}{\theta} I_{[0,\theta]}(x)$$

6.22 Minimo

Anche in questo caso è possibile definire il minimo di funzioni aleatorie continue :

$$X_{(1)} = \min_i \{X_i\} \quad (6.50)$$

ove le X_i sono variabili aleatorie ottenute da un campione casuale di qualunque distribuzione.

6.22.1 "Contatore" (CDF) e funzione densità (PDF)

Proviamo a calcolare la funzione di distribuzione cumulativa del caso duale a quello del massimo, che come vedremo è strettamente legata a quella del massimo. Per definizione di questa è pari a:

$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = P(\min\{X_1 \dots X_n\} \leq x)$$

Passiamo all'evento complementare poiché in questo modo si può ragionare meglio sull'espressione.

$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - P(\min\{X_1 \dots X_n\} > x)$$

A destra dell'espressione si ha che il minimo è maggiore di x , questo significa che se il minimo è maggiore di x lo sono per definizione anche tutti gli altri, quindi possiamo riscrivere l'espressione come:

$$F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - P(X_1 > x, \dots, X_n > x) \underset{\text{ind.}}{=} 1 - [P(X_1 > x) \cdot \dots \cdot P(X_n > x)]$$

A questo punto passiamo una seconda volta all'evento complementare, questa volta con l'intenzione di ricondurci a la definizione di contatore.

$$F_{X_{(1)}}(x) = 1 - [(1 - P(X_1 \leq x)) \cdot \dots \cdot (1 - P(X_n \leq x))]$$

Sapendo che $X_1 \dots X_n$ sono identicamente distribuite, tutte hanno lo stesso contatore, quindi:

$$X_{(1)} = \min_i \{X_i\} \quad \rightarrow \quad F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n \quad (6.51)$$

Per trovare la funzione densità non ci basta altro che applicare la definizione, cioè derivare la CDF.

$$X_{(1)} = \min_i \{X_i\} \quad \rightarrow \quad f_{X_{(1)}}(x) = n[1 - F_X(x)]^{n-1} f_X(x) \quad (6.52)$$

Minimo per uniformi

Dato che la definizione è sempre valida, qualunque sia la distribuzione della variabile X . Consideriamo il caso in cui $X \sim \text{unif}[0, \theta]$.

Per quanto detto nella sezione 6.18 la variabile uniforme assume come funzione densità :

$$f_X(x) = \frac{1}{\theta} I_{[0, \theta]}(x)$$

e funzione "contatore":

$$F_X(x) = \int_0^x \frac{1}{\theta} dt = \frac{x}{\theta} I_{[0, \theta]}(x) + I_{(\theta, +\infty)}(x)$$

Quindi la funzione contatore del minimo è:

$$F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n = 1 - \left[1 - \frac{x}{\theta}\right]^n$$

Dalla figura si apprezza che all'aumentare di n , il contatore si tende a comportarsi come se tutta la massa di probabilità fosse concentrato in 0. Questo vuol dire che se \underline{X} è campione casuale da $X \sim \text{unif}[0, \theta]$ allora:

$$X_{(1)} \xrightarrow[n \rightarrow \infty]{L} 0$$

Cioè $F_{X_{(1)}}(x)$ tende a 1. Questo contatore è il contatore di un'altra variabile aleatoria, chiamata **variabile aleatoria degenere** che si caratterizza per avere una PDF (densità) pari a un delta di Dirac in 0 e CDF costante a 1.

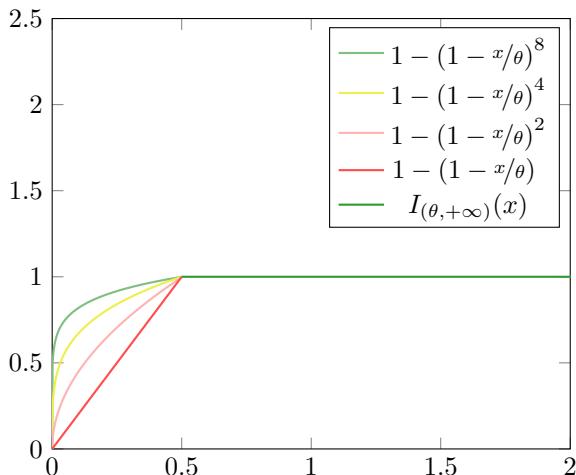


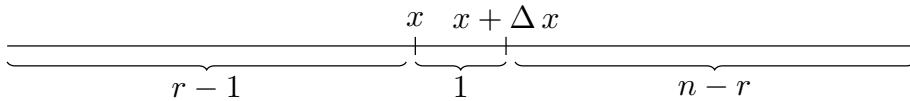
Figura 6.14: Caso con $\theta = 0.5$

6.23 R-esima statistica d'ordine

Vogliamo capire in generale qual è la distribuzione della r -esima statistica d'ordine ($X_{(r)}$), di cui il massimo ($X_{(n)}$) e il minimo ($X_{(1)}$) costituiscono dei casi particolari. Per ottenere la sua distribuzione sfruttiamo il fatto che:

$$f_{X_{(r)}}(x) \Delta x \simeq P(x < X_{(r)} \leq x + \Delta x)$$

A questo punto ci basta interpretare la formula appena scritta, se volgiamo trovare l' r -esima statistica d'ordine, significa che ci devono essere $r - 1$ osservazioni prima di x , 1 osservazione nell'intervallo $[x, x + \Delta x]$ e $n - r$ osservazioni dopo $x + \Delta x$. Inoltre dobbiamo considerare con degli opportuni coefficienti binomiali tutte le possibili scelte per gli $r - 1$ e gli $n - r$ elementi.



Dalla figura si può vedere facilmente che:

$$f_{X_{(r)}}(x) \Delta x \simeq P(X_1 \leq x, \dots, X_{r-1} \leq x, \dots, x \leq X_r \leq x + \Delta x, X_{r+1} > x + \Delta x, \dots, X_n > x + \Delta x)$$

$$f_{X_{(r)}}(x) = \binom{n}{r-1} \binom{n-r+1}{1} [F_X(x)]^{r-1} [1 - F_X(x)]^{n-r} f_X(x) \quad (6.53)$$

NB

Si verifica facilmente che per $r = 1$ si ha la densità del minimo e per $r = n$ si ha la densità del massimo.

6.23.1 Statistica "a metà" di uniformi

Analizziamo qual è la distribuzione della statistica d'ordine situata a metà delle osservazioni. Supponiamo per semplicità che si abbia un numero dispari di dati ($2k + 1$), l'osservazione "a metà" è dunque la statistica d'ordine $X_{(k+1)}$.

Con le approssimazioni di Stirling si ottiene:

$$f_{X_{(k+1)}}(x) = \frac{(2k+1)!}{(k!)^2} [x(1-x)]^k I_{[0,1]}(x)$$

NB

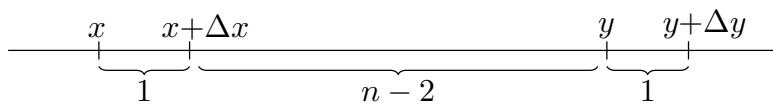
Si può dimostrare facilmente che $X_{(k+1)}$ è uno "stimatore" (argomento trattato nel capitolo 7) per la mediana del modello.

6.24 Range

Vogliamo capire come si distribuisce quello che chiamiamo *range*, cioè, vogliamo capire qual è la densità della variabile aleatoria $R = X_{(n)} - X_{(1)}$. Per ottenere la sua distribuzione usiamo lo stesso scamotage usato nella r-esima statistica d'ordine:

$$f_{X_{(1)}, X_{(n)}}(x, y) \Delta x \Delta y = P(x \leq X_{(1)} \leq x + \Delta x, y \leq X_{(n)} \leq y + \Delta y)$$

A questo punto si deve interpretare la formula appena scritta, se volgiamo trovare il range significa che tutte le osservazioni devono essere maggiori del minimo, che presenta un'osservazione tra x e $x + \Delta x$. La stessa cosa succede per il massimo che presenta un'osservazione tra y e $y + \Delta y$. le restanti $n - 2$ osservazioni di trovano tra $x + \Delta x$ e y per definizione. Anche in questo caso saranno necessari dei coefficienti binomiali che ci considerano le possibili scelte per gli elementi. Per il minimo dobbiamo scegliere 1 osservazione su n , e di conseguenza per il massimo dev'essere scelta un'osservazione fra le $n - 1$ rimanenti.



Dalla figura si evince più facilmente che:

$$f_{X_{(1)}, X_{(n)}}(x, y) \Delta x \Delta y \simeq P(x \leq X_{(1)} \leq x + \Delta x, x + \Delta x \leq X_{(2)} \leq y, \dots, y \leq X_{(n)} \leq y + \Delta y)$$

$$f_{X_{(1)}, X_{(n)}}(x, y) = \binom{n}{1} \binom{n-1}{1} [F_Y(y) - F_X(x)]^{n-2} f_X(x) f_Y(y) I_{\{x < y\}}(x, y) \quad (6.54)$$

Chiusura rispetto al minimo

Le due variabili d'attesa, la geometrica in campo bernoulliano e l'esponenziale in campo poissoniano, godono della proprietà di chiusura rispetto al minimo, cioè il minimo di queste variabili ha la stessa distribuzione delle singole variabili anche se con parametri un po' diversi. Verifichiamolo tenendo ben presente che la formula generale per il minimo è:

$$F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n$$

ma vale solo se i contatori sono identici per tutte le n variabili, altrimenti si ha

$$F_{X_{(1)}}(x) = 1 - [1 - F_{X_1}(x)] \cdots [1 - F_{X_n}(x)]$$

Esponenziale

$$T_{(1)} = \text{Min}\{T_1, \dots, T_n\} \quad \text{con} \quad T_i \sim \exp(v_i)$$

$$T_{(1)} = 1 - [1 - F_{T_1}(x)] \cdots [1 - F_{T_n}(x)] = 1 - [1 - (1 - e^{v_1 x})] \cdots [1 - (1 - e^{v_n x})] = 1 - [e^{v_1 x}] \cdots [e^{v_n x}]$$

Che per le proprietà degli esponenti:

$$1 - [e^{v_1 x}] \cdots [e^{v_n x}] = 1 - \left[e^{(v_1 + \dots + v_n)x} \right]$$

Cioè abbiamo appena trovato che sotto l'ipotesi che i campi poissoniani delle singole variabili esponenziali si sovrappongano nell'intervallo x , allora il minimo si distribuisce come un'esponenziale di un campo poissoniano con la somma delle intensità sei singoli campi, in formule

$$T_{(1)} \sim \exp(v_1 + \dots + v_n)$$

Geometrica

$$T_{(1)} = \text{Min}\{T_1, \dots, T_n\} = 1 - [1 - F_{T_1}(x)] \cdots [1 - F_{T_n}(x)]$$



Si ricordi che $F_{T_n}(x) \stackrel{\text{def}}{=} P(T_n \leq x)$ passando all'evento complementare si ha $1 - P(T_n > x)$ che significa che ho avuto x insuccessi, cioè:

$$1 - P(T_n > x) = 1 - (1 - \theta)^x$$

$$\begin{aligned} T_{(1)} &= 1 - [1 - (1 - (1 - \theta)^x)] \cdots [1 - (1 - (1 - \theta)^x)] = 1 - [(1 - \theta_1)^x] \cdots [(1 - \theta_n)^x] \\ &\quad 1 - [(1 - \theta_1)^x] \cdots [(1 - \theta_n)^x] = 1 - [(1 - \theta_1) \cdot (1 - \theta_n)]^x \end{aligned}$$

alla luce di quanto detto si ottiene:

$$T_{(1)} \sim \text{geom}(1 - [(1 - \theta_1) \cdots (1 - \theta_n)])$$



7. Statistica inferenziale

La risposta a tutto è 57 o 23. E non per caso ($57=1*3*19$) e ($23=1+3+19$).

(STUDENTE CHE HA SUPERATO CDP)

7.1 Statistiche e proprietà degli stimatori

Supponiamo di avere una *popolazione* qualsiasi. Come abbiamo già accennato prima, lo studio di una popolazione richiede uno o più **campioni casuali** che come detto in precedenza non è altro che una n -upla di variabili indipendenti e identicamente distribuite, $\underline{X} = (X_1, X_2, \dots, X_n)$. La statistica inferenziale è una branca della statistica che prova a risalire o a *stimare* a partire dai campioni casuali il valore del o dei parametri delle ipotetica distribuzione che assume la popolazione, in altri casi (piuttosto rari) prova a risalire direttamente alla funzione densità dato che non si conosce precisamente la distribuzione. Vedremo come la statistica inferenziale attraversa le stesse tappe di uno studente universitario, intraprende un processo di apprendimento (raccolta informazioni) e più informazioni si raccolgono, maggiore sarà l'esperienza e la capacità di affrontare problemi (in questo caso una migliore stima dei parametri). Detto questo addentriamoci nel mondo della statistica inferenziale definendo alcuni concetti molto importanti:

- **Statistica.** Una variabile aleatoria che è in funzione del campione casuale e non contiene quantità non note.

$$T = g(\underline{X})$$

- **Stimatore.** Una statistica che si usa per stimare un parametro ad esempio ϑ ; la stima del parametro si indica di solito con $\widehat{\vartheta}$.

- **Stimatore corretto.** Uno stimatore che ha come valore atteso il parametro che cerca stimare.

$$E[g(\underline{X})] = \vartheta$$

- *Stimatore distorto.* Uno stimatore che NON ha come valore atteso il parametro che cerca stimare.
- *Stimatore asintoticamente corretto.* Uno stimatore che non ha valore atteso *uguale* a ϑ , ma *tende* a ϑ all'aumentare della dimensione del campione

$$E[g(\underline{X})] \xrightarrow{n \rightarrow \infty} \theta$$

- *Stimatore consistente.* Uno stimatore che converge in probabilità al parametro che si cerca di stimare.

$$g(\underline{X}) \xrightarrow[n \rightarrow \infty]{P} \vartheta$$

Le proprietà di correttezza e consistenza ci quantificano in alcun modo “l'affidabilità” e “bontà” dello stimatore.

7.1.1 Condizione sufficiente per la non correttezza

Se lo stimatore (funzione del campione casuale per stimare ϑ) non è in grado di stimare il parametro attraverso una trasformazione lineare del campione casuale, una eventuale trasformazione non lineare da applicare implicherebbe la perdita della proprietà di correttezza, poiché in generale non si verifica:

$$E[g(X)] \stackrel{?}{=} g(E[X])$$

L'espressione soprastante è vera solo nel caso di trasformazioni lineari a causa della linearità del valore atteso, per questo motivo la proprietà di correttezza viene conservata per queste trasformazioni. Per trasformazioni non lineari, possiamo approssimare la funzione $g(x)$ applicando Taylor in $x_0 = \mu$:

$$g(X) \simeq g(\mu) + g'(\mu)(X - \mu)$$

Ottenendo:

$$g(X) \underset{\sim}{\approx} N(g(\mu), (g'(\mu))^2 var(X))$$

La trasformazione non lineare più comune da dover fare è il passaggio al reciproco, questa trasformazione implica la perdita della correttezza ma conserva la correttezza asintotica. Proviamo a verificare questo fatto applicando questa trasformazione alla variabile aleatoria X :

$$g(X) = \frac{1}{X}$$

$$g(X) \simeq \frac{1}{\mu} - \frac{1}{\mu^2}(X - \mu)$$

Calcoliamo per primo $E[g(X)]$

$$E[g(X)] = E\left[\frac{1}{\mu} - \frac{1}{\mu^2}(X - \mu)\right] = E\left[\frac{1}{\mu}\right] - E\left[\frac{1}{\mu^2}(X - \mu)\right] = E\left[\frac{1}{\mu}\right] - \frac{1}{\mu^2}E[(X - \mu)]$$

$$E\left[\frac{1}{\mu}\right] - \frac{1}{\mu^2}E[X] + \frac{1}{\mu^2}E[\mu] = \frac{1}{\mu} - \frac{1}{\mu^2}E[X] + \frac{\mu}{\mu^2} = \frac{1}{\mu} - \frac{\mu}{\mu^2} + \frac{\mu}{\mu^2} = \frac{1}{\mu}$$

Procediamo ora al calcolo della varianza tenendo ben in mente le sue proprietà di invarianza per traslazione e omogeneità di secondo grado .

$$\begin{aligned} Var(g(X)) &= Var\left(\frac{1}{\mu} - \frac{1}{\mu^2}(X - \mu)\right) \underset{\text{inv tras.}}{=} Var\left(-\frac{1}{\mu^2}(X - \mu)\right) \\ &\underset{\text{om 2.}}{=} \left(\frac{1}{\mu^2}\right)^2 Var(X - \mu) \underset{\text{inv tras.}}{=} \left(\frac{1}{\mu^2}\right)^2 Var(X) \end{aligned}$$

Se X consente l'approssimazione normale

$$g(X) \approx N\left(\frac{1}{\mu}, \left(\frac{1}{\mu^2}\right)^2 var(X)\right)$$

Infatti questa approssimazione ricavata è un caso particolare della formula

$$g(X) \approx N(g(\mu), (g'(\mu))^2 var(X)) \quad (7.1)$$



Questa formula rimane valida anche per trasformazioni lineari. (la dimostrazione si ricava in modo analogo a quanto fatto in precedenza)

ESEMPIO

Se $X \sim geom(\theta)$ allora

$$\frac{1}{X} \approx N\left(\theta, (\theta^2)^2 \frac{(1-\theta)}{\theta^2}\right) = N\left(\theta, \theta^2(1-\theta)\right)$$

Per questo motivo possiamo concludere che lo stimatore $\frac{1}{\bar{X}_n}$ per θ tratto da una popolazione geometrica si distribuisce come:

$$\frac{1}{\bar{X}_n} \approx N\left(\mu_X, \frac{\sigma_X^2}{n}\right) = N\left(\theta, \frac{\theta^2(1-\theta)}{n}\right)$$

7.1.2 Condizione sufficiente per la consistenza

Una proprietà di facile verifica che assicura la convergenza in legge e dunque la consistenza dello stimatore è:

- ① Lo stimatore $T = g(\underline{X})$ è almeno asintoticamente corretto (la correttezza implica la correttezza asintotica ma non viceversa).

$$E[T] \xrightarrow[n \rightarrow \infty]{L} \vartheta$$

- ② La varianza dello stimatore tende a zero all'aumentare della numerosità del campione.

$$Var(T) \xrightarrow[n \rightarrow \infty]{L} 0$$

Dato un campione casuale di una popolazione con distribuzione nota, il processo di stabilire come si possono stimare i parametri della distribuzione prende il nome di *stima puntuale dei parametri*.

7.2 Stima puntuale dei parametri

Variabile aleatoria	Stimatore	Proprietà
$X \sim \mathbf{b}(\theta)$	$\bar{X}_n \longleftrightarrow \theta$	<ul style="list-style-type: none"> ● Correttezza ● Consistenza
$Y_m \sim \mathbf{bin}(m, \theta)$	$\frac{1}{m} \bar{X}_n \longleftrightarrow \theta$	<ul style="list-style-type: none"> ● Correttezza ● Consistenza
$T \sim \mathbf{geom}(\theta)$	$\frac{1}{\bar{X}_n} \longleftrightarrow \theta$	<ul style="list-style-type: none"> ● Correttezza asintotica ● Consistenza
$T_r \sim \mathbf{bineg}(\theta)$	$\frac{r}{\bar{X}_n} \longleftrightarrow \theta$	<ul style="list-style-type: none"> ● Correttezza asintotica ● Consistenza
$N \sim \mathbf{poiss}(\lambda)$	$\bar{X}_n \longleftrightarrow \lambda$	<ul style="list-style-type: none"> ● Correttezza ● Consistenza
$T \sim \mathbf{exp}(\nu)$	$\frac{1}{\bar{X}_n} \longleftrightarrow \nu$	<ul style="list-style-type: none"> ● Correttezza asintotica ● Consistenza
$T_r \sim \mathbf{\Gamma}(r, \nu)$	$\frac{r}{\bar{X}_n} \longleftrightarrow \nu$	<ul style="list-style-type: none"> ● Correttezza asintotica ● Consistenza
$T\# \sim \mathbf{unif}[0, \theta]$	$2\bar{X}_n \longleftrightarrow \theta$	<ul style="list-style-type: none"> ● Correttezza ● Consistenza

Figura 7.1: Si noti che n è la dimensione del campione casuale, per questo motivo non tutte le variabili assumono i loro indici “standard”.

Per la normale bisogna fare qualche considerazione in più poiché il parametro da stimare non è soltanto uno ma sono ben due parametri da stimare: μ e σ^2 . Per stimare μ c'è un solo modo:

Variabile aleatoria	Stimatore	Proprietà
$Z \sim N(\mu, \sigma^2)$	$\bar{X}_n \longleftrightarrow \mu$	<ul style="list-style-type: none"> • Correttezza • Consistenza

Per stimare σ^2 ci sono 3 casi, che dipendono se μ era già noto e delle proprietà associate ad ogni stimatore.

Situazione	Stimatore	Proprietà
μ noto	$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \leftrightarrow \sigma^2$	<ul style="list-style-type: none"> • Correttezza • Consistenza
μ incognito	$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \leftrightarrow \sigma^2$	<ul style="list-style-type: none"> • Correttezza asintotica • Consistenza
μ incognito	$S_C^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \leftrightarrow \sigma^2$ $= \frac{n}{n-1} \left(\frac{\sum_{i=1}^n X_i^2}{n} - (\bar{X}_n)^2 \right) \leftrightarrow \sigma^2$	<ul style="list-style-type: none"> • Correttezza • Consistenza

E' anche possibile stimare il parametro θ di una uniforme con la sola conoscenza del valore più alto del campione, cioè il massimo $X_{(n)}$.

Situazione	Stimatore	Proprietà
$T \# \sim \text{unif}[0, \theta]$	$X_{(n)} \longleftrightarrow \theta$	<ul style="list-style-type: none"> • Correttezza asintotica • Consistenza
$T \# \sim \text{unif}[0, \theta]$	$\frac{n+1}{n} X_{(n)} \longleftrightarrow \theta$	<ul style="list-style-type: none"> • Correttezza • Consistenza

7.2.1 Efficienza degli stimatori

A questo punto potrebbe aver sorto la domanda, se ho più modi di calcolare un parametro, come nel caso del parametro θ quando la variabile si distribuisce come una uniforme, quale stimatore si deve scegliere? Ovviamente quello migliore! Quello migliore è quello che risulta più efficiente, quindi formalizziamo quella che è la risposta intuitiva a questo problema.

Se T_1 è uno stimatore per ϑ e anche T_2 è uno stimatore sempre per ϑ . Diremo che T_1 è più efficiente di T_2 , e dunque da preferirsi, se:

$$\text{Var}(T_1) \leq \text{Var}(T_2)$$

NB

Questa scala di efficienza si dice relativa, si può definire anche una scala di efficienza assoluta, argomento che però non verrà trattato.

ESEMPIO

Se consideriamo il caso in cui la variabile è uniforme, sono 2 gli stimatori corretti e consistenti, $\frac{n+1}{n}X_{(n)}$ e $2\bar{X}_n$; poiché si verifica che:

$$\text{Var}\left(\frac{n+1}{n}X_{(n)}\right) = \frac{k}{n^2}$$

$$\text{Var}(2\bar{X}_n) = \frac{k}{n}$$

(k indica una costante della varianza che non è rilevante ai fini dell'analisi della efficienza, infatti ciò che conta è l'esponente di n) quindi si ha che lo stimatore $\frac{n+1}{n}X_{(n)}$ è da preferirsi poiché la varianza tende a zero più rapidamente di quanto lo faccia $2\bar{X}_n$.

SUGGERIMENTO

Ora che si conoscono tutti gli stimatori per le variabili viste, come facciamo a ricordarli tutti? basta ricordare che $\bar{X} \longleftrightarrow E[X]$ qualunque sia la distribuzione della variabile X ; quindi ciò che resta da fare è manipolare l'espressione finché non si isola il parametro da stimare.

7.3 Dettagli su gli stimatori della varianza

Prima di proseguire è importante conoscere la distribuzione asintotica così come gli indici caratteristici degli stimatori della varianza dato che verranno usati con molta frequenza nel argomento successivo.

Consideriamo per prima lo stimatore S_C^2

$$\frac{(n-1) S_C^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}_n}{\sigma} \right)^2 \xrightarrow{n \text{ grande}} \sim \chi_{(n-1)}^2 \equiv \Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right)$$

Quindi $S_C^2 \sim \frac{\sigma^2}{n-1} W$ con $W = \chi_{(n-1)}^2 = \Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right)$

Ora che sappiamo come si distribuisce è immediato calcolare il valore atteso e la varianza:

$$E[S_C^2] = E\left[\frac{\sigma^2}{n-1} W\right] = \frac{\sigma^2}{n-1} E[W] = \frac{\sigma^2}{n-1} \cdot \frac{\frac{n-1}{2}}{\frac{1}{2}} = \sigma^2$$

$$Var(S_C^2) = Var\left(\frac{\sigma^2}{n-1}W\right) = \left(\frac{\sigma^2}{n-1}\right)^2 Var(W) = \left(\frac{\sigma^2}{n-1}\right)^2 \frac{\frac{n-1}{2}}{\left(\frac{1}{2}\right)^2} = \frac{\sigma^4}{(n-1)^2} 2(n-1)$$

Trovando in definitiva

$$Var(S_C^2) = \frac{2\sigma^4}{n-1}$$

Quindi la sua distribuzione asintotica è:

$$S_C^2 \rightsquigarrow N\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$



Si noti che nella distribuzione asintotica si è scritto a denominatore della varianza n anziché $n-1$ poiché per n grande non c'è gran differenza.

Consideriamo ora S_0^2

$$\frac{n S_0^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x-\mu}{\sigma}\right)^2 \xrightarrow{n \text{ grande}} \sim \chi_{(n)}^2 \equiv \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

Dunque allo stesso modo di prima si ottiene:

$$E[S_0^2] = E\left[\frac{\sigma^2}{n}W\right] = \frac{\sigma^2}{n}E[W] = \frac{\sigma^2}{n} \frac{\frac{n}{2}}{\frac{1}{2}} = \sigma^2$$

$$Var(S_0^2) = Var\left(\frac{\sigma^2}{n}W\right) = \left(\frac{\sigma^2}{n}\right)^2 Var(W) = \left(\frac{\sigma^2}{n}\right)^2 \frac{\frac{n}{2}}{\left(\frac{1}{2}\right)^2} = \frac{\sigma^4}{(n)^2} 2(n) = \frac{2\sigma^4}{n}$$

Quindi la sua distribuzione asintotica è:

$$S_0^2 \rightsquigarrow N\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

7.4 Intervalli di confidenza

Gli stimatori non sono l'unico strumento che ci permettono di conoscere il valore di un parametro; esiste un altro strumento chiamato *intervalli di confidenza* che anziché fornirci un valore preciso del parametro come lo fanno gli stimatori, ci forniscono, appunto, un intervallo di valori che può assumere il parametro ϑ , con la probabilità γ che questo accada.

Cioè, un intervallo di confidenza è un intervallo funzione del campione casuale \underline{X} tale che:

$$P(\vartheta \in I(\underline{X})) \geq \gamma \tag{7.2}$$

In pratica, un generico intervallo di confidenza con probabilità gamma verrà indicato

$$I_\gamma(\underline{X})$$

7.4.1 Intervalli esatti per il valore atteso

Gli intervalli di confidenza più semplici sono quelli che sfruttano la disuguaglianza di Cebicev. Supponiamo di dover stimare il parametro μ , conoscendo σ ; dalla disuguaglianza di Cebicev sappiamo che:

$$P\left(|\bar{X}_n - \mu| \leq t \frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{t^2}$$

Cioè:

$$P\left(-t \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq t \frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{t^2}$$

A questo punto ci è ancora ignoto t , ma come vogliamo che l'intervallo abbia probabilità γ , risolviamo:

$$1 - \frac{1}{t^2} = \gamma \quad \rightarrow \quad t = \frac{1}{\sqrt{1-\gamma}}$$

Per ottenere l'intervallo di confidenza, basta riscrivere l'intervallo in modo tale che sia centrato sul parametro e non abbia quantità non note a destra e a sinistra.

$$-t \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq t \frac{\sigma}{\sqrt{n}} \quad = \quad \bar{X}_n - \frac{1}{\sqrt{1-\gamma}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{1}{\sqrt{1-\gamma}} \frac{\sigma}{\sqrt{n}}$$

Quindi possiamo concludere che l'intervallo di confidenza per μ :

$$I_\gamma(\underline{X}) = \left[\bar{X}_n \pm \frac{\sigma}{\sqrt{1-\gamma}\sqrt{n}} \right]$$

7.4.2 Intervalli asintotici per il valore atteso

L'intervallo di confidenza ottenuto con Cebicev non è molto utile dato che ci fornisce un intervallo molto ampio di valori assumibili dal parametro incognito ϑ . Per ottenere una migliore approssimazione si usa la conoscenza della distribuzione asintotica della variabile.

Normale

Se è noto che la variabile $X \sim N(\mu, \sigma^2)$ allora sempre assumendo che σ sia nota, possiamo ottenere dalla disuguaglianza di Cebicev:

$$P\left(|\bar{X}_n - \mu| \leq k_\gamma \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

Questa volta la costante incognita dipendente da γ , indicata come k_γ , può essere calcolata più accuratamente sapendo che la variabile si distribuisce asintoticamente come una normale. Facendo queste considerazioni possiamo scrivere la disuguaglianza come:

$$P\left(\frac{|\bar{X}_n - \mu|}{\frac{\sigma}{\sqrt{n}}} \leq k_\gamma\right) = \gamma$$

sapendo dal teorema centrale del limite che $\frac{|\bar{X}_n - \mu|}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ per definizione di contatore si ha:

$$P\left(-k_\gamma \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k_\gamma\right) = \Phi(k_\gamma) - \Phi(-k_\gamma) = 2\Phi(k_\gamma) - 1$$

A questo punto non ci basta altro che trovare k_γ :

$$2\Phi(k_\gamma) - 1 = \gamma \quad \rightarrow \quad k_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right)$$

Come sappiamo dall'analisi, Φ è una funzione che non può essere calcolata con i metodi standard di integrazione, per questo motivo, esistono delle tabelle che riportano i valori più comuni di $\Phi(t)$, per $\Phi^{-1}(t)$ basta leggere la stessa tabella "al contrario". Le tabelle della $\Phi(t)$ e di altre distribuzione sono riportate nella sezione 7.5. Detto questo, è possibile scrivere l'intervallo di confidenza come :

$$I_\gamma(\underline{X}) = \left[\bar{X}_n \pm \Phi^{-1}\left(\frac{1+\gamma}{2}\right) \frac{\sigma}{\sqrt{n}} \right]$$

Student

Finora abbiamo supposto che σ fosse noto, ma se non lo fosse? Dobbiamo stimare σ^2 con S_C^2 dato che μ non ci è noto. facendo un po' gli stessi passaggi di prima si ha :

$$P\left(\frac{|\bar{X}_n - \mu|}{\frac{\sqrt{S_C^2}}{\sqrt{n}}} \leq k_\gamma\right) = \gamma \quad \equiv \quad P\left(-k_\gamma \leq \frac{\bar{X}_n - \mu}{\sqrt{\frac{S_C^2}{n}}} \leq k_\gamma\right)$$

il termine $\frac{\bar{X}_n - \mu}{\sqrt{\frac{S_C^2}{n}}}$ non ha più distribuzione normale, ma ha distribuzione di student a $n-1$ gradi di libertà . Con questa conoscenza otteniamo γ come:

$$2t_{(n-1)}(k_\gamma) - 1 = \gamma \quad \rightarrow \quad k_\gamma = t_{(n-1)}^{-1}\left(\frac{1+\gamma}{2}\right)$$

Ottenendo in definitiva:

$$I_\gamma(\underline{X}) = \left[\bar{X}_n \pm t_{(n-1)}^{-1}\left(\frac{1+\gamma}{2}\right) \sqrt{\frac{S_C^2}{n}} \right]$$

ESEMPIO

Intervalli di confidenza con $\gamma = 0,95$

$$\text{Cebicev :} \quad I_{0,95}(\underline{X}) = \left[\bar{X}_n \pm 4,46 \frac{\sigma}{\sqrt{n}} \right]$$

Il risultato ci fornisce un intervallo molto ampio.

$$\text{Normale :} \quad I_{0,95}(\underline{X}) = \left[\bar{X}_n \pm 1,96 \frac{\sigma}{\sqrt{n}} \right]$$

Se confrontiamo con il risultato ottenuto precedentemente, vediamo che l'intervallo fornito utilizzando l'approssimazione normale è molto più efficiente che quello ottenuto con Cebicev.

$$\text{Student :} \quad I_{0,95}(\underline{X}) = \left[\bar{X}_n \pm 2,06 \sqrt{\frac{S_C^2}{n}} \right]$$

Vediamo che l'intervallo fornito è più efficiente che quello ottenuto con Cebicev ma meno efficiente di quello ottenuto con la normale, questo calo nelle prestazioni è dovuto al fatto che si è stimato il parametro σ^2 che era incognito.

Supponiamo ora di voler stabilire un intervallo di confidenza di $\gamma = 0,95$ per il valore atteso $\frac{1}{v}$ di una variabile esponenziale. Scegliamo di farlo sfruttando l'approssimazione normale per un campione casuale molto numeroso:

$$\frac{1}{v} \in I_{0,95}(\underline{X}) = \left[\bar{X}_n \pm 1,96 \frac{\sigma}{\sqrt{n}} \right]$$

Finora abbiamo sempre lasciato la radice della varianza indicata, per cui non si ha ancora un vero proprio intervallo di confidenza ma un intervallo di livello γ . Per renderlo un intervallo di confidenza dobbiamo far sì che l'intervallo contenga agli estremi tutte quantità note. Proviamo a sostituire il suo valore: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{v^2}}$

$$\frac{1}{v} \in I_{0,95}(\underline{X}) = \left[\bar{X}_n \pm 1,96 \frac{1}{v} \frac{1}{\sqrt{n}} \right]$$

Ma ha senso quello che si è scritto? Certamente no! stiamo cercando di trovare un intervallo per $\frac{1}{v}$ che richiede la conoscenza di $\frac{1}{v}$!

A questo punto ci sono due strade:

- **Asintotico "esatto"** (Cercare di raccogliere $\frac{1}{v}$)

Riscriviamo l'intervallo per esteso:

$$\left[\bar{X}_n - 1,96 \frac{1}{v} \frac{1}{\sqrt{n}} \leq \frac{1}{v} \leq \bar{X}_n + 1,96 \frac{1}{v} \frac{1}{\sqrt{n}} \right] = \left[\frac{1}{v} \left(1 - 1,96 \frac{1}{\sqrt{n}} \right) \leq \bar{X}_n \leq \frac{1}{v} \left(1 + 1,96 \frac{1}{\sqrt{n}} \right) \right]$$

Ora riscrivendo l'intervallo per $\frac{1}{v}$ si ha:

$$I_{0,95}(\underline{X}) = \left[\frac{\bar{X}_n}{1 + 1,96 \frac{1}{\sqrt{n}}} \leq \frac{1}{v} \leq \frac{\bar{X}_n}{1 - 1,96 \frac{1}{\sqrt{n}}} \right]$$

- **Asintotico approssimato** (Fare una approssimazione della varianza)

Dalla stima puntuale dei parametri sappiamo che uno stimatore corretto e consistente per $\frac{1}{v}$ è \bar{X}_n ; essendo la varianza $\frac{1}{v^2}$ il quadrato del valore atteso, possiamo approssimarla con \bar{X}_n^2 , dunque σ può essere approssimato con \bar{X}_n

$$\left[\bar{X}_n - 1,96 \frac{1}{v} \frac{1}{\sqrt{n}} \leq \frac{1}{v} \leq \bar{X}_n + 1,96 \frac{1}{v} \frac{1}{\sqrt{n}} \right] = \left[\bar{X}_n - 1,96 \bar{X}_n \frac{1}{\sqrt{n}} \leq \frac{1}{v} \leq \bar{X}_n + 1,96 \bar{X}_n \frac{1}{\sqrt{n}} \right]$$

Quindi l'intervallo di confidenza è:

$$\left[\bar{X}_n \left(1 - 1,96 \frac{1}{\sqrt{n}} \right) \leq \frac{1}{v} \leq \bar{X}_n \left(1 + 1,96 \frac{1}{\sqrt{n}} \right) \right]$$

NB

Ovviamente la strada da preferire è quella di raccogliere la quantità da stimare, giacché ovviamente l'introdurre approssimazioni comporta una "riduzione nelle prestazioni".

7.4.3 Intervalli di confidenza per un parametro

Supponiamo di voler calcolare un intervallo di confidenza per un parametro qualunque che non sia necessariamente il valore atteso, come lo si realizza?

Supponiamo di voler calcolare l'intervallo di confidenza per θ da una popolazione di $X \sim \text{unif}[0, \theta]$, l'intervallo avrà una forma identica a quella del valore atteso solo che in questo caso lo stimatore non è più \bar{X}_n ma $2\bar{X}_n$.

Se T_n è uno stimatore, la forma generale di un intervallo di confidenza è:

$$E[T_n] \in I_\gamma(\underline{X}) = \left[T_n \pm k_\gamma \sqrt{\text{Var}(T_n)} \right] \quad (7.3)$$

ove k_γ si ricava sfruttando Cebicev, la distribuzione normale o di student.

Poiché $E[2\bar{X}_n] = \theta$ e si può ricavare dalla (7.1) che la varianza è pari a $(g'(x))^2 \text{Var}(\bar{X}_n)$ cioè:

$$\text{Var}(2\bar{X}_n) = 2^2 \text{Var}(\bar{X}_n) = 4 \frac{\text{Var}(X)}{n} = 4 \frac{1}{n} \frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

Quindi l'intervallo di confidenza per θ sarà

$$I_\gamma(\underline{X}) \left[2\bar{X}_n \pm k_\gamma \sqrt{\frac{\theta^2}{3n}} \right]$$

Poiché l'intervallo richiede il parametro da stimare, si procede con i metodi già visti in precedenza e si trova:

$$I_\gamma(\underline{X}) \left[\frac{\bar{X}_n}{\left(1 \pm k_\gamma \sqrt{\frac{1}{3n}} \right)} \right] \quad \text{oppure} \quad I_\gamma(\underline{X}) \left[2\bar{X}_n \left(1 \pm k_\gamma \sqrt{\frac{1}{3n}} \right) \right]$$

In alternativa si poteva ricavare un intervallo di confidenza per $\frac{\theta}{2}$ (che è il valore atteso) e poi moltiplicare tutto per 2:

$$I_\gamma(\underline{X}) \left[\bar{X}_n \pm k_\gamma \sqrt{\frac{\theta^2}{12n}} \right] = \bar{X}_n - k_\gamma \sqrt{\frac{\theta^2}{12n}} \leq \frac{\theta}{2} \leq \bar{X}_n + k_\gamma \sqrt{\frac{\theta^2}{12n}}$$

$$I_\gamma(\underline{X}) \left[2\bar{X}_n \pm 2k_\gamma \sqrt{\frac{\theta^2}{12n}} \right] = 2\bar{X}_n - 2k_\gamma \sqrt{\frac{\theta^2}{12n}} \leq \theta \leq 2\bar{X}_n + 2k_\gamma \sqrt{\frac{\theta^2}{12n}}$$

portando il 2 dentro la radice

$$2\bar{X}_n - k_\gamma \sqrt{\frac{4\theta^2}{12n}} \leq \theta \leq 2\bar{X}_n + k_\gamma \sqrt{\frac{4\theta^2}{12n}} = I_\gamma(\underline{X}) \left[2\bar{X}_n \pm k_\gamma \sqrt{\frac{\theta^2}{3n}} \right]$$

che è identico a quello ricavato prima. ATTENZIONE! Risultano uguali perché la trasformazione è lineare, altrimenti risulterebbero diversi!

7.4.4 Intervalli di confidenza per la varianza

Per gli intervalli di confidenza per la varianza è conveniente adottare una forma generale più semplice

$$I_\gamma(\underline{X}) = [c_1 S_i^2, c_2 S_i^2]$$

ove S_i^2 può essere o S_C^2 oppure S_0^2 e i coefficienti c_1, c_2 adottano la convenzione di *code di ugual peso*

Code di ugual peso

Supponiamo di conoscere μ , l'intervallo di confidenza corrispondente è :

$$P(c_1 S_0^2 \leq \sigma^2 \leq c_2 S_0^2) = \gamma$$

manipolando l'intervallo possiamo ottenere:

$$P\left(\frac{n}{c_2} \leq \frac{nS_0^2}{\sigma^2} \leq \frac{n}{c_1}\right) = \gamma$$

Sapendo che $\frac{nS_0^2}{\sigma^2} \sim \chi_{(n)}^2$ si possono ottenere i coefficienti come:

$$F_{\chi_{(n)}^2}\left(\frac{n}{c_1}\right) - F_{\chi_{(n)}^2}\left(\frac{n}{c_2}\right) = \gamma \quad F_{\chi_{(n)}^2}\left(\frac{n}{c_2}\right) - F_{\chi_{(n)}^2}\left(\frac{n}{c_1}\right) = 1 - \gamma$$

Sfortunatamente ci sono infiniti valori di c_1, c_2 che soddisfano l'equazione. Per questo motivo si introduce la convenzione di code di ugual peso, cioè per convenzione si stabilisce che la probabilità a sinistra di c_1 dev'essere la stessa probabilità a destra di c_2 .

Cioè:

$$P\left(\chi_{(n)}^2 \leq \frac{n}{c_2}\right) = \frac{1-\gamma}{2}$$

$$P\left(\chi_{(n)}^2 > \frac{n}{c_1}\right) = \frac{1-\gamma}{2}$$

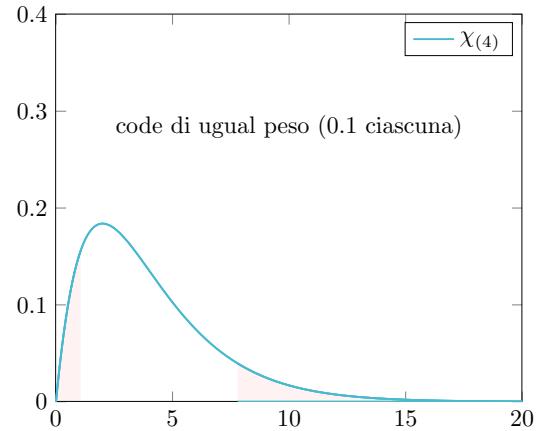


Figura 7.2: Code di ugual peso

Ottenendo:

$$c_1 = \frac{n}{F_{\chi_{(n)}^2}^{-1}\left(\frac{1+\gamma}{2}\right)} \quad c_2 = \frac{n}{F_{\chi_{(n)}^2}^{-1}\left(\frac{1-\gamma}{2}\right)}$$

NB

Se μ è incognito si fa lo stesso procedimento ma si sfrutta che $\frac{(n-1)S_C^2}{\sigma^2} \sim \chi_{(n-1)}^2$, quindi i coefficienti c_1 e c_2 varieranno solo nella funzione contatore inversa che anziché essere $F_{\chi_{(n)}^2}^{-1}$ sarà $F_{\chi_{(n-1)}^2}^{-1}$

7.5 Tavole

$$\Phi(t) / \Phi^{-1}(t)$$

n	$Student \quad F_{t_n} / F_{t_n^{-1}}$											
	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%	
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31	
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327	
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215	
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173	
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893	
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208	
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785	
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501	
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297	
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144	
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025	
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930	
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852	
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787	
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733	
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686	
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646	
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610	
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579	
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552	
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527	
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505	
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485	
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467	
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450	
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435	
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421	
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408	
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396	
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385	
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340	
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307	
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281	
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261	
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245	
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232	
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090	

n	$Chi - quadro \quad F_{\chi_n^2} / F_{\chi_n^2}^{-1}$											
	0.1%	0.5%	1.0%	2.5%	5.0%	10.0%	12.5%	20.0%	25.0%	33.3%	50.0%	
1	0.000	0.000	0.000	0.001	0.004	0.016	0.025	0.064	0.102	0.186	0.455	
2	0.002	0.010	0.020	0.051	0.103	0.211	0.267	0.446	0.575	0.811	1.386	
3	0.024	0.072	0.115	0.216	0.352	0.584	0.692	1.005	1.213	1.568	2.366	
4	0.091	0.207	0.297	0.484	0.711	1.064	1.219	1.649	1.923	2.378	3.357	
5	0.210	0.412	0.554	0.831	1.145	1.610	1.808	2.343	2.675	3.216	4.351	
6	0.381	0.676	0.872	1.237	1.635	2.204	2.441	3.070	3.455	4.074	5.348	
7	0.598	0.989	1.239	1.690	2.167	2.833	3.106	3.822	4.255	4.945	6.346	
8	0.857	1.344	1.646	2.180	2.733	3.490	3.797	4.594	5.071	5.826	7.344	
9	1.152	1.735	2.088	2.700	3.325	4.168	4.507	5.380	5.899	6.716	8.343	
10	1.479	2.156	2.558	3.247	3.940	4.865	5.234	6.179	6.737	7.612	9.342	
11	1.834	2.603	3.053	3.816	4.575	5.578	5.975	6.989	7.584	8.514	10.341	
12	2.214	3.074	3.571	4.404	5.226	6.304	6.729	7.807	8.438	9.420	11.340	
13	2.617	3.565	4.107	5.009	5.892	7.042	7.493	8.634	9.299	10.331	12.340	
14	3.041	4.075	4.660	5.629	6.571	7.790	8.266	9.467	10.165	11.245	13.339	
15	3.483	4.601	5.229	6.262	7.261	8.547	9.048	10.307	11.037	12.163	14.339	
16	3.942	5.142	5.812	6.908	7.962	9.312	9.837	11.152	11.912	13.083	15.338	
17	4.416	5.697	6.408	7.564	8.672	10.085	10.633	12.002	12.792	14.006	16.338	
18	4.905	6.265	7.015	8.231	9.390	10.865	11.435	12.857	13.675	14.931	17.338	
19	5.407	6.844	7.633	8.907	10.117	11.651	12.242	13.716	14.562	15.859	18.338	
20	5.921	7.434	8.260	9.591	10.851	12.443	13.055	14.578	15.452	16.788	19.337	
21	6.447	8.034	8.897	10.283	11.591	13.240	13.873	15.445	16.344	17.720	20.337	
22	6.983	8.643	9.542	10.982	12.338	14.041	14.695	16.314	17.240	18.653	21.337	
23	7.529	9.260	10.196	11.689	13.091	14.848	15.521	17.187	18.137	19.587	22.337	
24	8.085	9.886	10.856	12.401	13.848	15.659	16.351	18.062	19.037	20.523	23.337	
25	8.649	10.520	11.524	13.120	14.611	16.473	17.184	18.940	19.939	21.461	24.337	
26	9.222	11.160	12.198	13.844	15.379	17.292	18.021	19.820	20.843	22.399	25.336	
27	9.803	11.808	12.879	14.573	16.151	18.114	18.861	20.703	21.749	23.339	26.336	
28	10.391	12.461	13.565	15.308	16.928	18.939	19.704	21.588	22.657	24.280	27.336	
29	10.986	13.121	14.256	16.047	17.708	19.768	20.550	22.475	23.567	25.222	28.336	
30	11.588	13.787	14.953	16.791	18.493	20.599	21.399	23.364	24.478	26.165	29.336	
35	14.688	17.192	18.509	20.569	22.465	24.797	25.678	27.836	29.054	30.894	34.336	
40	17.916	20.707	22.164	24.433	26.509	29.051	30.008	32.345	33.660	35.643	39.335	
45	21.251	24.311	25.901	28.366	30.612	33.350	34.379	36.884	38.291	40.407	44.335	
50	24.674	27.991	29.707	32.357	34.764	37.689	38.785	41.449	42.942	45.184	49.335	
55	28.173	31.735	33.570	36.398	38.958	42.060	43.220	46.036	47.610	49.972	54.335	
60	31.738	35.534	37.485	40.482	43.188	46.459	47.680	50.641	52.294	54.770	59.335	

n	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
55	57.016	58.945	61.665	63.577	67.211	68.796	73.311	77.380	82.292	85.749	93.168
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607

7.6 Test di verifica d'ipotesi

Il test di verifica d'ipotesi è un test che verifica la "bontà" di un'ipotesi, nell'ambito statistico quest'ipotesi è una congettura sul valore di un certo parametro (es. θ); quindi ciò che cercano di fare i test di verifica d'ipotesi è confermare o smentire un'ipotesi, di solito espressa come una uguaglianza o disegualità, attraverso dei dati ricavati sperimentalmente (campioni casuali). In molti casi pratici, questo test consiste nel rifiutare una ipotesi H_0 e accettare l'ipotesi alternativa H_1 o viceversa.

Terminologia

- L'ipotesi H_0 si dice *ipotesi nulla* (dall'inglese *null hypothesis*)
- L'ipotesi H_1 si dice *ipotesi alternativa*.
- L'ipotesi si dice *semplice* se considera un solo valore fissato

$$H_0 : \theta = \theta_0$$

- L'ipotesi si dice *composta* se non è semplice, cioè se considera più valori o intervalli.
- Un'ipotesi composta si dice *unilaterale* o *a una coda* se è del tipo:

$$H_0 : \theta > \theta_0 \quad \vee \quad H_0 : \theta < \theta_0$$

- Un'ipotesi composta si dice *bilaterale* o *a due code* se è del tipo:

$$H_0 : \theta \neq \theta_0 \quad \text{cioè} \quad H_0 : \theta > \theta_0 \quad \wedge \quad H_0 : \theta < \theta_0$$

I test di verifica d'ipotesi si suddividono principalmente in due categorie:

1 Test Parametrici

Sono test che vengono applicati a campioni di cui è nota la distribuzione e hanno ipotesi riguardanti il valore di un parametro quale il valore atteso e la varianza.

2 Test Non-Parametrici

Sono test che vengono applicati quando non si conoscono i parametri relativi alla distribuzione dei campioni. Le ipotesi non riguardano il valore di un parametro ma riguardano la distribuzione del campione.

Prima di addentrarci su i diversi test, chiariamo per primo a cosa ci servono, e che considerazioni si devono fare.

Supponiamo di voler dire se il parametro θ (esce testa) di una moneta è $\theta = 0,5$ (non è truccata) oppure no (è truccata), cioè vogliamo stabilire in base ad un campione casuale se:

$$\begin{aligned} H_0 &: \theta = 0,5 \\ H_1 &: \theta \neq 0,5 \end{aligned} \tag{7.4}$$

Supponiamo che il campione casuale sia composta da 20 osservazioni di cui solo 5 sono risultate testa. Il test parametrico ci aiuta a determinare se accettare H_0 o rifiutarla.

Il fatto che solo 5 osservazioni siano risultate testa, non esclude la possibilità che il parametro sia effettivamente 0,5 ma riduce la probabilità che l'ipotesi sia vera, in altre parole, poiché sappiamo che la probabilità che 5 su 20 monete vengano testa (avendo come parametro 0,5) è $\approx 0,014$, è possibile che il parametro sia effettivamente 0,5 e al rifiutare H_0 si commetterebbe un errore, ma la probabilità di commettere questo errore è molto piccola.

Queste considerazioni ci portano a definire i possibili errori che si possono commettere:

- **Errore di primo tipo** (*falso positivo*) si rifiuta l'ipotesi H_0 quando in realtà risulta vera. A questo errore viene associata una probabilità indicata con α e viene chiamata *significance level* o semplicemente livello.

$$P_{H_0}(\text{rifiuto } H_0) = \alpha \quad (7.5)$$

- **Errore di secondo tipo** (*falso negativo*) si accetta l'ipotesi H_0 quando in realtà risulta falsa, o equivalentemente si rifiuta H_1 quando in realtà risulta vera. A questo errore viene associata una probabilità indicata con β . La probabilità $(1 - \beta)$ viene chiamata *statistical power*.

$$P_{H_1}(\text{rifiuto } H_1) = \beta \quad (7.6)$$



Si ricordi che: rifiutare $H_1 \Rightarrow$ accettare H_0 e rifiutare $H_0 \Rightarrow$ accettare H_1 .

Il pedice nella probabilità indica che questa è calcolata nel caso in cui si verifica l'ipotesi a pedice.

I test di verifica d'ipotesi ci aiutano a rifiutare o ad accettare H_0 minimizzando gli errori. Ovviamente quanto più grande è il campione, quanto meno sono gli errori e quanto più difficile è sbagliare.

Ipotesi	Accettata	Rifiutata
Valida	Buona decisione (<i>vero positivo</i>) $1 - \beta$	Errore di primo tipo (<i>falso positivo</i>) α
Errata	Errore di secondo tipo (<i>falso negativo</i>) β	Buona decisione (<i>vero negativo</i>) $1 - \alpha$

Figura 7.3: In figura vengono rappresentati i casi possibili di un test di ipotesi.

7.6.1 Complementi

In medicina, la **specificità** di un test misura la frequenza di veri negativi, mentre la **predittività** di un test misura invece la frequenza di veri positivi. Questi parametri aiutano a misurare quanto sia preciso o meno un test medico.

7.7 Test parametrici per il valore atteso

Analizziamo i diversi tipi di test in cui è nota la distribuzione della popolazione, ovvero i test parametrici.

7.7.1 Test Z bilatero

Il test Z permette di valutare quando $X \sim N(\mu, \sigma^2)$ e σ^2 nota, le ipotesi:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned} \tag{7.7}$$

Cioè, questo test ha come scopo il verificare se il valore atteso della distribuzione si discosta molto da un valore fissato ($\mu \neq \mu_0$) o se si trova vicino ($\mu = \mu_0$).

Sorge quindi la domanda, come si quantifica se è vicino o lontano?

Fissato un intervallo $[\mu_0 - c, \mu_0 + c]$ con un c in modo tale da considerare tutti i valori di quel intervallo "vicini", rifiutare significa che la stima del parametro ottenuta dal campione casuale (data da \bar{X}_n) si discosta molto da μ_0 cioè, non si trova nel intervallo che io considero "vicino":

$$\bar{X}_n \notin [\mu_0 - c, \mu_0 + c] \tag{7.8}$$

Quindi l'errore di primo tipo, cioè che dal campione casuale si conclude che μ non è vicino a μ quando in realtà lo è; è uguale a:

$$P_{\mu=\mu_0} (\bar{X}_n \notin [\mu_0 - c, \mu_0 + c]) = \alpha \tag{7.9}$$

Scrivendo per bene l'intervallo e passando all'evento complementare si ha:

$$1 - P(\mu_0 - c \leq \bar{X}_n \leq \mu_0 + c) = \alpha$$

Normalizzando si trova:

$$1 - P\left(\frac{-c}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq \frac{c}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$$

Cioè:

$$P\left(\frac{-c}{\frac{\sigma}{\sqrt{n}}} \leq "Z \sim N(0, 1)" \leq \frac{c}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \alpha$$

con diversi passaggi algebrici si trova:

$$c = \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left(1 - \frac{\alpha}{2}\right)$$

Grazie a queste considerazioni, possiamo concludere che fissato un valore di α che quantifichi la tolleranza agli errori, è possibile fornire una regola di decisione che mi permetta di dire se accettare o rifiutare H_0 :

Rifiutiamo H_0 se $\bar{X}_n \notin \left[\mu_0 \pm \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left(1 - \frac{\alpha}{2}\right)\right]$

$$\bar{X}_n \notin \left[\mu_0 \pm \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left(1 - \frac{\alpha}{2}\right)\right] \tag{7.10}$$

NB

Il test si chiama test Z poiché ci si è ricondotto ad una variabile $Z \sim N(0, 1)$

7.7.2 Test T bilatero

Il test T permette di valutare quando $X \sim N(\mu, \sigma^2)$ e σ^2 incognita, le ipotesi:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned} \tag{7.11}$$

Così come nel caso precedente, l'errore di primo tipo è definito come :

$$1 - P(\mu_0 - c \leq \bar{X}_n \leq \mu_0 + c) = \alpha$$

Questa volta poiché σ^2 non è nota non si può normalizzare... ma si può fare qualcosa di simile!

$$1 - P\left(\frac{-c}{\sqrt{\frac{s_c^2}{n}}} \leq \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{s_c^2}{n}}} \leq \frac{c}{\sqrt{\frac{s_c^2}{n}}}\right) = \alpha$$

Cioè:

$$P\left(\frac{-c}{\sqrt{\frac{s_c^2}{n}}} \leq "T \sim t_{(n-1)}" \leq \frac{c}{\sqrt{\frac{s_c^2}{n}}}\right) = 1 - \alpha$$

con diversi passaggi algebrici si trova:

$$c = \sqrt{\frac{s_c^2}{n}} t_{(n-1)}^{-1} \left(1 - \frac{\alpha}{2}\right)$$

Quindi la regola di decisione è:

$$\text{Rifiutiamo } H_0 \text{ se } \bar{X}_n \notin \left[\mu_0 \pm \sqrt{\frac{s_c^2}{n}} t_{(n-1)}^{-1} \left(1 - \frac{\alpha}{2}\right) \right] \tag{7.12}$$

NB

Il test si chiama test T poiché ci si è ricondotto ad una variabile $T \sim t_{(n-1)}$

Riformulazione

Le regole di decisione possono essere riformulate in funzione di intervalli di confidenza di parametro $\gamma = 1 - \alpha$ cioè:

$$\begin{aligned} \text{Test Z} \quad \text{Rifiutiamo } H_0 \text{ se } \mu_0 \notin I_{1-\alpha}(\underline{X}) &= \left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left(1 - \frac{\alpha}{2}\right) \right] \\ \text{Test T} \quad \text{Rifiutiamo } H_0 \text{ se } \mu_0 \notin I_{1-\alpha}(\underline{X}) &= \left[\bar{X}_n \pm \sqrt{\frac{s_c^2}{n}} t_{(n-1)}^{-1} \left(1 - \frac{\alpha}{2}\right) \right] \end{aligned} \tag{7.13}$$

È possibile anche formulare dei test che non coinvolgono il valore atteso, ad esempio, si vuole trovare una regola di decisione per le seguenti ipotesi assumendo che il campione casuale sia stato campionato da una popolazione esponenziale:

$$\begin{aligned} H_0 : v &= b \\ H_1 : v &\neq b \quad (b \text{ è un valore fissato}). \end{aligned} \tag{7.14}$$

Anche se questo intervallo non coinvolge il valore atteso (μ) è possibile riformulare il test in modo equivalente ad un altro test con il valore atteso. Ricordando che per le esponenziali $\mu = \frac{1}{v}$:

$$H_0 : \frac{1}{v} = \frac{1}{b} \Rightarrow H_0 : \mu = \frac{1}{b} \tag{7.15}$$

$$H_1 : \frac{1}{v} \neq \frac{1}{b} \Rightarrow H_1 : \mu \neq \frac{1}{b}$$

Poiché per le esponenziali si verifica $Var(T) = (E[T])^2$ allora $\sigma^2 = (\frac{1}{b})^2$ Quindi la regola di decisione è:

$$\text{Rifiutiamo } H_0 \text{ se } \bar{X}_n \notin \left[\mu_0 \pm \frac{1}{b\sqrt{n}} \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right] \tag{7.16}$$

7.7.3 Test Z e Test T unilateri

È possibile anche formulare dei test riguardanti il valore atteso con ipotesi unilaterale:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned} \tag{7.17} \quad \text{oppure} \quad \begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &< \mu_0 \end{aligned} \tag{7.18}$$

Per questi test il concetto di "vicino" cambia radicalmente, ad esempio nel caso con $\mu > \mu_0$, essere "vicino" a μ vuol dire che esiste un k tale per cui \bar{X}_n è minore di k , e quindi la probabilità di rifiutare H_0 , cioè di dire che μ è lontano quando in realtà non lo è, si ottiene da:

$$P_{\mu=\mu_0}(\bar{X}_n \geq k) = \alpha \tag{7.19}$$

Normalizzando e trovando k fissando un valore di α , si trova la regola di decisione (caso Test Z, per il test T le differenze sono ovvie):

$$\begin{aligned} H_1 : \mu > \mu_0 &\quad \text{Accettiamo } H_0 \text{ se } \bar{X}_n \leq \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1-\alpha) \\ H_1 : \mu < \mu_0 &\quad \text{Accettiamo } H_0 \text{ se } \bar{X}_n \geq \mu_0 - \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1-\alpha) \end{aligned} \tag{7.20}$$

7.8 Test parametrici per la varianza

7.8.1 Test bilatero

Il test permette di valutare quando $X \sim N(\mu, \sigma^2)$ le ipotesi:

$$\begin{aligned} H_0 : \sigma^2 &= a \\ H_1 : \sigma^2 &\neq a \end{aligned} \tag{7.21}$$

Formuliamo la regola di decisione considerando un intervallo che quantifica l'essere "vicino".

$$\alpha = 1 - P(ac_1 \leq S_0^2 \leq ac_2) \quad \mu \text{ noto} \tag{7.22}$$

Oppure

$$\alpha = 1 - P(ac_1 \leq S_C^2 \leq ac_2) \quad \mu \text{ incognito} \tag{7.23}$$

Anche i coefficienti c_i cambiano a seconda che μ sia noto oppure incognito. Moltiplicando tutta l'espressione per $\frac{n}{a}$ se μ noto oppure per $\frac{n-1}{a}$ se μ incognito si trova la regola di decisione:

$$\begin{aligned} \mu \text{ noto} \quad \text{Rifiutiamo } H_0 \text{ se} \quad S_0^2 &\notin \left[\frac{a}{n} F_{\chi_{(n)}^2}^{-1} \left(\frac{\alpha}{2} \right), \frac{a}{n} F_{\chi_{(n)}^2}^{-1} \left(1 - \frac{\alpha}{2} \right) \right] \\ \mu \text{ incognito} \quad \text{Rifiutiamo } H_0 \text{ se} \quad S_C^2 &\notin \left[\frac{a}{n-1} F_{\chi_{(n-1)}^2}^{-1} \left(\frac{\alpha}{2} \right), \frac{a}{n-1} F_{\chi_{(n-1)}^2}^{-1} \left(1 - \frac{\alpha}{2} \right) \right] \end{aligned} \tag{7.24}$$

7.8.2 Test unilatero

Facendo le stesse considerazioni fatte per i test unilateri del valore atteso, si trova che per un test unilatero della varianza, la regola di decisione è:

$$\begin{aligned} \mu \text{ noto} \quad \text{Rifiutiamo } H_0 \text{ se} \quad S_0^2 &\geq \frac{a}{n} F_{\chi_{(n)}^2}^{-1} (1 - \alpha) \\ \mu \text{ incognito} \quad \text{Rifiutiamo } H_0 \text{ se} \quad S_C^2 &\geq \frac{a}{n-1} F_{\chi_{(n-1)}^2}^{-1} (1 - \alpha) \end{aligned} \tag{7.25}$$

NB

In questi casi ci si è ricondotto al fatto che $\frac{nS_0^2}{\sigma^2} \sim \chi_{(n)}^2$ e che $\frac{(n-1)S_C^2}{\sigma^2} \sim \chi_{(n-1)}^2$. Per questo motivo sia nel test bilatero che nel test unilatero compare il contatore di una chi-quadro. Si noti che nel test bilatero si è utilizzata l'ipotesi di code di ugual peso dato che la distribuzione della chi-quadro non è simmetrica e non si può concludere come per la distribuzione normale o di student che: $F_{\chi_{(n-1)}^2}^{-1} \left(\frac{\alpha}{2} \right) = -F_{\chi_{(n-1)}^2}^{-1} \left(1 - \frac{\alpha}{2} \right)$

7.9 Ulteriori tipi di test parametrici

7.9.1 Test con entrambe ipotesi composte

Cosa succede se entrambe le ipotesi sono ipotesi composte? Ad esempio:

$$\begin{aligned} H_0 : \mu &\leq \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned} \tag{7.26}$$

Notiamo che l'errore di primo tipo, cioè di rifiutare H_0 quando in realtà è vera, può assumere diversi valori a seconda di dove si trovi μ_0 . Più \bar{X}_n si trova a sinistra di μ_0 , minore è la probabilità di sbagliare e commettere un errore di primo tipo. Ovviamente l'errore più grande lo si ha quando $\bar{X}_n = \mu_0$, questo caso è detto di *massimo rischio* (poiché l'errore di primo tipo assume il suo valore massimo), cioè:

$$P_{H_0}(\text{rifiuto } H_0) = P_{\mu \leq \mu_0}(\bar{X}_n \geq k) = \alpha$$

Normalizzando:

$$\begin{aligned} \alpha &= P_{\mu \leq \mu_0} \left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{k - \mu}{\frac{\sigma}{\sqrt{n}}} \right) = 1 - P \left("Z \sim N(0, 1)" \leq \frac{k - \mu}{\frac{\sigma}{\sqrt{n}}} \right) \\ \alpha &= 1 - \Phi \left(\frac{k - \mu}{\frac{\sigma}{\sqrt{n}}} \right) = g(\mu) \end{aligned}$$

Cioè l'errore di primo tipo è una funzione del parametro $g(\mu)$. Nel caso considerato, questa funzione è una funzione crescente, cioè più $\bar{X}_n \leftrightarrow \mu$ si sposta a destra, più l'errore aumenta; e come già detto:

$$\max_{\mu \leq \mu_0} \{g(\mu)\} = g(\mu_0)$$

Cioè il caso di *massimo rischio* lo si ha quando $\bar{X}_n = \mu_0$. Avendo fatto queste considerazioni possiamo concludere che una volta fissata la tolleranza di un errore di primo tipo (α), il test deve essere fatto considerando il caso di massimo rischio; detto questo, il test si trasforma nel test già trattato:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned} \tag{7.27}$$

7.9.2 Test con entrambe ipotesi semplici

Cosa succede se entrambe le ipotesi sono ipotesi semplici?, Ad esempio:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &= \mu_1 \end{aligned} \tag{7.28}$$

Fissato un k tra μ_0 e μ_1 , supponendo che la situazione sia :



Si rifiuta μ_0 se $\bar{X}_n \geq k$, cioè se lo stimatore di μ mi da un valore "lontano" a μ_0 e "vicino" a μ_1 o equivalentemente si rifiuta μ_1 se $\bar{X}_n < k$. L'errore di primo tipo è dato da:

$$P_{H_0}(\bar{X}_n \geq k) = \alpha$$

e quindi la regola di decisione per rifiutare μ_0 è la regola già vista:

$$\text{Rifiutiamo } H_0 \text{ se } \bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha)$$

In questo caso è facile calcolare anche l'errore di secondo tipo:

$$P_{H_1}(\bar{X}_n < k) = \beta$$

Sapendo però dal errore di primo tipo che $k = \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha)$ la probabilità diventa:

$$P_{H_1}\left(\bar{X}_n < \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha)\right) = \beta$$

Normalizzando trova:

$$\beta = \Phi\left(\frac{\mu_0 - \mu_1}{\frac{\sigma}{\sqrt{n}}} + \Phi^{-1}(1 - \alpha)\right) \quad (7.29)$$

Il **Teorema di Neyman-Pearson** ci dice che questa regola di decisione è quella che minimizza al minimo l'errore di secondo tipo.

7.10 Test Non-Parametrici

Dentro la categoria dei test non-parametrici possiamo distinguere due sotto-categorie:

1 Test di buon adattamento

I test di buon adattamento ci permettono di dire se una variabile ha o meno un certa distribuzione ipotizzata.

- *Tipo Generale*: hanno validità per tutte o una grande classe di funzioni come lo sono le variabili aleatorie continue e le variabili aleatorie discrete.
- *Tipo Speciale*: hanno validità solo per particolari tipi di distribuzione ipotizzata.

2 Test di causalità

I test di causalità provano a determinare se c'è o meno una dipendenza tra le determinazioni, cioè i valori osservati.

7.10.1 Test χ^2 di buon adattamento

Il test chi-quadro di Pearson detto anche test χ^2 del buon adattamento è un test che ha come obiettivo stabilire se un campione casuale molto grande è stato estratto da una popolazione con una predeterminata distribuzione oppure no. Quindi il test consiste nel rifiutare o accettare l'ipotesi:

$$\begin{aligned} H_0 : \underline{X} &\sim F_X(x) \\ H_1 : \underline{X} &\not\sim F_X(x) \end{aligned} \quad (7.30)$$

NB

Si faccia attenzione al fatto che la conoscenza della distribuzione è stata espressa come il conoscere il contatore della variabile. La scelta del contatore anziché la densità assicura la validità di questo test sia per variabili aleatorie continue che per variabili aleatorie discrete.

Ma come facciamo a determinare se una variabile aleatoria segue la distribuzione ipotizzata? Cominciamo per suddividere il range di osservazioni del campione casuale $\underline{X} = (X_1, X_2, \dots, X_n)$ in r classi "1", "2", ..., "r" (non necessariamente tutte della stessa dimensione) e siano Y_1, Y_2, \dots, Y_r le frequenze assolute delle rispettive classi.

Ad esempio una classe potrebbe essere: "1" : $x \in [0, 2]$ e di conseguenza Y_1 è il numero di volte che il campione casuale riporta un elemento della classe "1".

Per determinare se una variabile aleatoria segue una determinata distribuzione, si confronta la probabilità che dovrebbe avere ogni classe secondo il modello ipotizzato con la frequenza relativa delle classi. Cioè ci interessa capire:

$$\text{Quanto si discosta } P(X_j \in "i") \text{ da } fr(i) = \frac{Y_i}{n}$$

Valutiamo questo discostamento; poiché non ci interessa se la differenza è per eccesso o per difetto, ne prendiamo il quadrato. E poiché non ci interessa la differenza assoluta ma quella relativa (cioè la differenza assoluta tra 100 e 101 così come per 0,1 e 0,5 è bassa in entrambi casi, ma la differenza relativa è diversissima, in un caso la differenza è $\approx 1\%$ e nell'altro $\approx 400\%$) per questo motivo è conveniente introdurre un fattore che "pesi" i diversi discostamenti. Avendo fatto queste considerazioni chiamiamo χ_{oss}^2 la somma degli scarti quadratici pesati ricavate dalle osservazioni:

$$\chi_{oss}^2 = n \sum_{i=1}^r \frac{(P_i - fr_i)^2}{P_i} = \sum_{i=1}^r \frac{(Y_i - nP_i)^2}{nP_i} \quad \text{che per } n \text{ grande } \rightsquigarrow \chi_{r-1}^2 \quad (7.31)$$

La nostra regola di decisione si baserà ovviamente su quanto sia grande il discostamento del modello ipotizzato e quello osservato. Per fare ciò ci serve una soglia c oltre alla quale considereremo il discostamento troppo grande ($\chi_{oss}^2 \geq c$) e rifiuteremo l'ipotesi H_0 . Il valore di c può essere calcolato prendendo in considerazione l'errore di primo tipo α (massima tolleranza all'errore).

$$\alpha = P_{H_0}(\text{rifiuto } H_0) = P_{H_0}(\chi_{oss}^2 \geq c)$$

Facendo i conti si trova che la regola di decisione è:

$\text{Si rifiuta } H_0 \text{ se } \chi_{oss}^2 \geq F_{\chi_{(r-1)}^2}^{-1}(1 - \alpha) \quad (7.32)$

Suddivisione in classi ottimale

Abbiamo visto che la prima cosa da fare per effettuare questo test è la suddivisione in classi. La cosa più "furba" da fare, è suddividere tutte le osservazioni in r classi, tutte contenti la stessa area di probabilità. In questo modo la probabilità P_i non dipende più dalla classe scelta ma diventa un valore costante pari a $1/r$.

ESEMPIO

Vogliamo ricavare r classi da $X \sim \exp(\nu)$ allora quello che facciamo è trovare i quantili di ordine m/r con $m = 1, 2, \dots, r-1$, cioè i separatori tra classi. Per fare ciò impostiamo:

$$\begin{aligned} P(0 \leq x \leq S_1) &= 1/r & F_X(S_1) - F_X(0) &= 1/r \\ P(S_1 \leq x \leq S_2) &= 1/r & F_X(S_2) - F_X(S_1) &= 1/r \\ &\vdots & & \\ P(S_{r-2} \leq x \leq S_{r-1}) &= 1/r & F_X(S_{r-1}) - F_X(S_{r-2}) &= 1/r \\ P(S_{r-1} \leq x \leq +\infty) &= 1/r & F_X(+\infty) - F_X(S_{r-1}) &= 1/r \end{aligned} \tag{7.33}$$

Poiché il separatore S_2 non si può calcolare a meno che S_1 sia già stato calcolato possiamo dare la seguente definizione ricorsiva per i $r-1$ separatori:

$$S_1 = F_X^{-1}\left(\frac{1}{r}\right) \quad S_{i+1} = F_X^{-1}\left(\frac{1}{r} + S_i\right) \quad S_{r-1} = F_X^{-1}\left(1 - \frac{1}{r}\right) \tag{7.34}$$

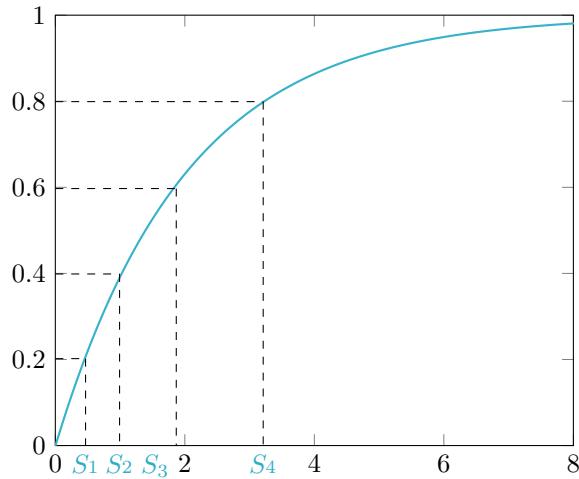


Figura 7.4: In figura si mostrano gli $r-1$ separatori per $r=5$ classi di una distribuzione esponenziale.

Impatto della stima di un parametro

Precedentemente abbiamo detto che χ_{oss}^2 si distribuisce asintoticamente per n grande come $\chi_{(r-1)}^2$ questo in realtà richiede qualche precisazione, infatti questa affermazione vale solo sotto l'ipotesi che i parametri della distribuzione ipotizzata siano tutti i noti. Se uno o più parametri non sono noti, allora la stima di questi parametri introduce un cambio nella distribuzione asintotica di χ_{oss}^2 , in generale vale la regola:

$$\chi_{oss}^2 \rightsquigarrow \chi_{(r-1-p)}^2 \quad p \text{ numero di parametri stimati} \tag{7.35}$$

7.10.2 Test di Kolmogorov-Smirnov

Il test di Kolmogorov è un test di buon adattamento valido solo per le variabili aleatorie continue, questo test permette di giudicare se una variabile ha o meno una distribuzione ipotizzata confrontando la CDF del modello teorico ipotizzato e la CDF empirica. Il confronto avviene facendo per prima un riordino delle osservazioni e poi una valutazione sulla discrepanza tra le due funzioni, che è data dal termine:

$$D_n = \sup_x |F_X(x) - F_X^*(x)| \quad (7.36)$$

Ove $F_X(x)$ è la CDF del modello teorico ipotizzato e $F_X^*(x)$ è la CDF empirica definita come:
 $F_X^*(x) = fr(X \leq x) = \frac{\text{num. di oss.} \leq x}{n}$

NB Poiché la $F_X^*(x)$ è ricavata da un numero finito di osservazioni, questa non è una funzione continua ma piuttosto una "scalinata". Il test quindi consisterà nel confrontare quanto la "scalinata" assomigli ad uno "scivolo" $F_X(x)$.

Il termine D_n che rappresenta la massima differenza tra la funzione continua $F_X(x)$ e la funzione costante a tratti $F_X^*(x)$, rappresenta sempre la differenza in corrispondenza di qualche punto di salto (discontinuità di prima specie). Possiamo dunque riscrivere D_n in un modo più chiaro:

$$D_n = \max_i \{d(i)\} \quad d(i) = \max \{d(i)^-, d(i)^+\} = \max \left\{ \left| F_X(x) - \frac{i}{n} \right|, \left| F_X(x) - \frac{i-1}{n} \right| \right\} \quad (7.37)$$

Ove $d(i)^-$ è la differenza al limite sinistro del salto i -esimo e $d(i)^+$ la differenza al limite destro del salto.

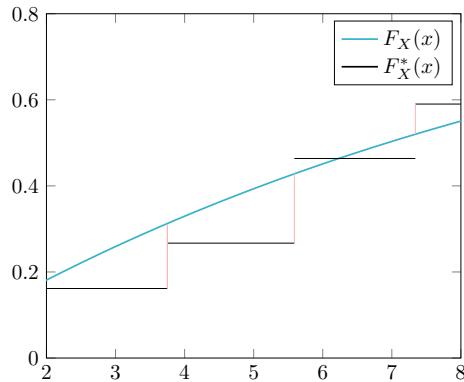


Figura 7.5: In rosso sono indicati i diversi $d(i)$

Per quanto detto, questo test intende valutare l'ipotesi:

$$H_0 : F_X^*(x) = F_X(x) \quad (7.38)$$

e rifiutarla se la differenza D_n è maggiore di una qualche soglia prefissata dipendente dalla tolleranza al errore di primo tipo, cioè se $D_n \geq c(\alpha)$.

Il termine c si ottiene da:

$$P_{H_0}(D_n \geq c) = \alpha$$

Il problema è che per ottenere c ci serve conoscere la distribuzione di D_n , che al momento non è nota. Non entriamo nei dettagli ma si è dimostrato che:

$$\sqrt{n} D_n \rightsquigarrow 1 - \left(2 \sum_{j=1}^{+\infty} (-1)^{j+1} e^{2j^2 x^2} \right) I_{\mathbb{R}^+}(x)$$

NB

Questa CDF oltre ad essere una funzione non elementare, è una serie non convergente uniformemente, cioè fissato un errore ε , non è possibile trovare un indice \bar{n} valido per *tutti* i valori di x tale che:

$$\left| \left[1 - \left(2 \sum_{j=1}^{+\infty} (-1)^{j+1} e^{-2j^2 x^2} \right) \right] - \left[1 - \left(2 \sum_{j=1}^{\bar{n}} (-1)^{j+1} e^{-2j^2 x^2} \right) \right] \right| \leq \varepsilon$$

Con queste considerazioni è possibile trovare il valore di c e dunque la regola di decisione:

Si rifiuta H_0 se $\sqrt{n} D_n \geq F_{\sqrt{n} D_n}^{-1}(1 - \alpha)$ (7.39)

NB

Qualora non fosse disponibile un calcolatore per fare i calcoli precisi, possiamo approssimare il termine $F_{\sqrt{n} D_n}^{-1}(1 - \alpha)$ considerando solo il primo addendo con $\sqrt{-\frac{1}{2} \ln(\frac{\alpha}{2})}$

Il **teorema di Glivenko-Cantelli** ci assicura che la CDF empirica di una variabile aleatoria ($F_X^*(x)$) converge **uniformemente** verso l'effettiva CDF ($F_X(x)$), detto in un altro modo si dice che la convergenza è *quasi certa* cioè:

$$P\left(\lim_{n \rightarrow \infty} D_n = 0\right) = 1 \quad (7.40)$$

Convergenza quasi certa

La nozione di convergenza quasi certa è quella più forte di tutte perché ci dice che all'aumentare della numerosità del campione, la probabilità che i campioni casuali X_n non tendano a coincidere con le determinazioni di X è zero.

$$X_n \xrightarrow{q.c} X \quad \text{se} \quad P\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1 \quad (7.41)$$

Essendo la convergenza quasi certa quella più forte di quelle viste finora si ricava che:

$$\text{Se } X_n \xrightarrow{q.c} X \quad \text{allora} \quad X_n \xrightarrow{P} X \quad \text{e} \quad X_n \xrightarrow{L} X$$

NB

Si noti che la formulazione della convergenza in probabilità pur essendo molto simile è completamente diversa concettualmente

$$X_n \xrightarrow{P} X \quad \text{se} \quad \lim_{n \rightarrow \infty} P(X_n = X) = 1 \quad (7.42)$$

Grazie alla convergenza quasi certa è possibile riformulare la legge dei grandi numeri però nella sua versione forte e non debole come quella presentata con convergenza in probabilità.

7.10.3 Test di normalità

Il test di normalità è un test di adattamento di tipo speciale, che è pensato per una particolare distribuzione, la distribuzione normale. Essendo un test per una particolare distribuzione, il test risulta più preciso dei test generali riportati precedentemente. Il test di normalità consiste nella valutazione congiunta o singola di due sotto-test, il *test di asimmetria*, che per brevità verrà chiamato Test A, e il *test di curtosi* che per brevità verrà chiamato Test C. Questi test provano a confermare o smentire l'ipotesi:

$$H_0 : X \sim "N(\mu, \sigma^2)" \quad (7.43)$$

Ovviamente la regola di decisione risulta:

Valutazione congiunta	Si accetta H_0 se	Test A accetta $H_0 \wedge$ Test C accetta H_0
Valutazione singola	Si accetta H_0 se	Test A accetta $H_0 \vee$ Test C accetta H_0

(7.44)

NB

La valutazione singola del solo test di asimmetria è molto scadente e pressoché inutile poiché il fatto di avere una distribuzione simmetrica è condizione necessaria ma non sufficiente affinché la distribuzione sia normale, infatti ci sono distribuzioni simmetriche come la distribuzione dell'uniforme continua che pur essendo simmetrica non è certamente normale. Questo test in realtà è più utile per smentire l'ipotesi, cioè, se non si supera questo test, si può concludere immediatamente che la distribuzione non è normale.

Se si decide di fare una valutazione congiunta dei due test, il livello della valutazione non è più α , si potrebbe pensare che è pari a $1 - (1 - \alpha)^2$ (che si otterrebbe valutando la probabilità congiunta di eventi indipendenti); il problema è che i test non sono indipendenti tra loro (vengono dallo stesso campione e usano lo stesso momento (M'_2), quindi la formula relativa a eventi indipendenti non può essere applicata). Dalle considerazioni fatte nella sezione (3.3) sulla probabilità $P(A \cap C)$ si ha che:

$$\max\{0, (1 - \alpha) + (1 - \alpha) - 1\} \leq P(A \cap C) \leq \min\{(1 - \alpha), (1 - \alpha)\}$$

Passando al evento complementare si trova che:

$$\alpha \leq \text{livello intervallo} \leq \min\{1, 2\alpha\}$$

(nel caso i livelli dei test fossero diversi):

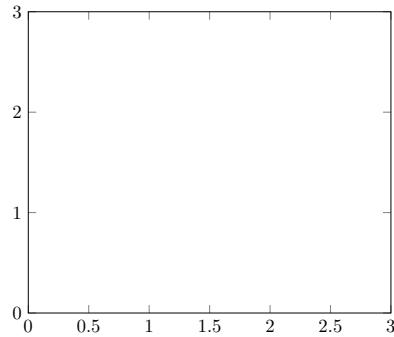
$$\max\{\alpha_A, \alpha_C\} \leq \text{livello intervallo} \leq \min\{1, \alpha_A + \alpha_C\}$$

Ulteriori test di normalità

Un altro modo per verificare rapidamente se una popolazione ha distribuzione normale o meno consiste nel verificare le proprietà di distribuzione della normale, cioè, che approssimativamente il 68,2% dei dati si trovi in $\mu \pm \sigma$, il 95,5% dei dati si trovi in $\mu \pm 2\sigma$ e il 99,7% dei dati si trovi in $\mu \pm 3\sigma$.

Q-Q plot

Un modo alternativo per verificare la normalità , è attraverso il confronto grafico fra la rappresentazione grafica dei quantili della distribuzione sotto analisi.

**Test di asimmetria o *skewness***

Il test di asimmetria è un test che ci dice se una distribuzione ha un asse di simmetria o meno basandosi sull'omonimo indice statistico. Questo indice statistico può essere stimato con la variabile

$$A = \frac{M'_3}{(M'_2)^{3/2}}$$

Il test di asimmetria sfrutta il fatto che per una normale i momenti di ordini dispari sono nulli.

$$E[(Z \sim N(0, 1))^3] = 0 \quad E[(X^*)^3] = \frac{E[(X - \mu)^3]}{E[\sigma^3]} = \frac{E[(X - \mu)^3]}{(\sigma^2)^{\frac{3}{2}}}$$

Che scrivendo per esteso i momenti centrati di ordine k e considerando che la dimensione del campione casuale è molto grande, si ha:

$$A \underset{n \rightarrow \infty}{\approx} N\left(0, \frac{6}{n}\right) \quad (7.45)$$

Nel Test A, ovviamente la regola di decisione è rifiutare se lo stimatore A assume un valore sufficientemente grande. Quantifichiamo l'essere "grande" con un intervallo soglia determinato dall'errore di primo tipo :

$$P\left(-c < N\left(0, \frac{6}{n}\right) < c\right) = 1 - \alpha \quad (7.46)$$



Si è passato al evento complementare poiché si ricordi che l'indice di asimmetria può essere sia positivo che negativo, quindi l'essere "grande" equivale a stare fuori dal intervallo soprastante.

La regola di decisione è la seguente:

$$\text{Si accetta } H_0 \text{ se } |A| < \sqrt{\frac{6}{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (7.47)$$

Test di curtosì

Il test di curtosì è un test basato sull'omonima proprietà del modello. Lo stimatore per questa proprietà è:

$$C = \frac{M'_4}{(M'_2)^2} \longleftrightarrow \frac{\mu'_4}{(\mu'_2)^2}$$

Se consideriamo un campione casuale molto grande, si scopre che:

$$A \rightsquigarrow N\left(3, \frac{24}{n}\right) \quad (7.48)$$

Allo stesso modo che nel Test A, nel Test C la regola di decisione consiste nel rifiutare se lo stimatore C assume un valore "grande". Quantificando l'essere "grande" con un intervallo soglia determinato dall'errore di primo tipo si ottiene:

$$P\left(-c < "N\left(3, \frac{24}{n}\right)" < c\right) = 1 - \alpha \quad (7.49)$$

La regola di decisione è la seguente:

$$\text{Si accetta } H_0 \text{ se } |C - 3| < \sqrt{\frac{24}{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (7.50)$$

7.10.4 Test di poissonianità

Il test di poissonianità è un test di buon adattamento speciale che prova a determinare se una variabile ha distribuzione poissoniana sfruttando il fatto che per questa distribuzione valore atteso e varianza coincidono. Cioè per poissoniane si verifica che:

$$\frac{E[X]}{\sigma^2} = 1 \quad \frac{E[X]}{\sigma^2} - 1 = 0$$

Il test rifiuterà l'ipotesi:

$$H_0 : \underline{X} \text{ proviene da } poiss(\lambda) \quad (7.51)$$

se la variabile: $Y = \left| \frac{\bar{X}_n}{S^2} - 1 \right|$ si discosta molto da zero.

Fissando una soglia c in base ad α , a partire dalla quale consideriamo "grande" il discostamento, si trova la regola di decisione:

$$\text{Si accetta } H_0 \text{ se } F_Y^{-1}\left(\frac{\alpha}{2}\right) < \frac{\bar{X}_n}{S^2} - 1 < F_Y^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (7.52)$$

NB

La CDF della variabile Y non è un caso notevole o che si possa approssimare mediante CDF notevoli. Per il suo calcolo ci si deve ricondurre a delle "tabelle" analoghe a quelle delle distribuzioni notevoli, solo che in questo caso l'individuazione dei valori viene fatta con una esaustiva simulazione al computer con campioni virtuali di dimensioni molto grandi.

7.11 Test di casualità

I test di casualità servono, come il nome ce lo indica, a verificare la casualità o meno da $\underline{X} = X_1, X_2, X_3, \dots, X_n$. Una condizione necessaria ma non sufficiente per la casualità è l'uniformità delle X_i . Il problema è che pur essendo uniformi (PMF orizzontale), possono non essere casuali. Si pensi ai seguenti esempi :

$$E_1 : 1010101010$$

$$E_2 : 0000011111$$

Per rifiutare questi casi che pur essendo uniformi non sono casuali, si può verificare l'uniformità in più dimensioni (Per 2 dimensioni si considerano le X_i a coppie, cioè: $(X_1, X_2), (X_3, X_4), \dots, (X_{n-1}, X_n)$). Se non è uniforme in una dimensione, si può subito concludere che non è una variabile casuale, ma bisogna fare attenzione poiché anche se sono uniformi in tutte le dimensioni ciò non implica la casualità , cioè sono condizioni necessarie ma non sufficienti.

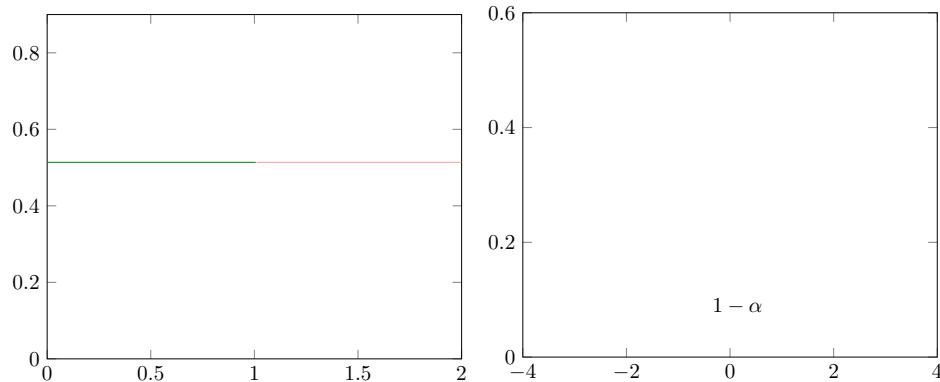


Figura 7.6: Anche se E_1 risulta uniforme in una dimensione, in due dimensione si può apprezzare la sua non-uniformità .

Per verificare l'effettiva distribuzione casuale si deve dimostrare l'indipendenza, questa verifica si riconduce a svolgere sostanzialmente 5 passi:

- ① Trasformazione delle osservazioni in osservazioni di una uniforme $\text{unif}[0, 1]$.
- ② Suddivisione delle osservazioni così trasformate in m campioni di cardinalità k ($n = m \cdot k$).
- ③ Si calcola il massimo di ogni riga, $W_m = \max_{1 \leq i \leq k} \{X_{mi}\}$.
Si ricordi che $F_{W_m}(x) = (F_X(x))^k$ se le X_i sono indipendenti; poiché la casualità implica l'indipendenza il contatore teorico dovrebbe essere (essendo uniforme) :

$$F_{W_m}(x) = x^k I_{[0,1]}(x) + I_{[1,+\infty)}(x)$$

- ④ Calcolo di $Y_m = (W_m)^k$.

$$F_Y(y) = P(Y \leq y) = P(W^k \leq y) = P(W \leq y^{\frac{1}{k}}) = F_W(y^{\frac{1}{k}})$$

Si trova che teoricamente $Y \sim \text{unif}[0, 1]$

- ⑤ Si verifica se esiste un'effettiva uniformità delle Y_m con i test di buon adattamento.

7.12 Significatività statistica: p-value

Il p-value (α^*), detto anche livello di significatività osservato, è il valore **massimo** di α per il quale si accetta l'ipotesi H_0 . Per chiarire il concetto, utilizziamo un esempio molto usato nell'ambito di test d'ipotesi. Supponiamo che un giudice deva valutare le ipotesi:

$$\begin{aligned} H_0 &: \text{l'accusato è innocente} \\ H_1 &: \text{l'accusato è colpevole} \end{aligned} \quad (7.53)$$

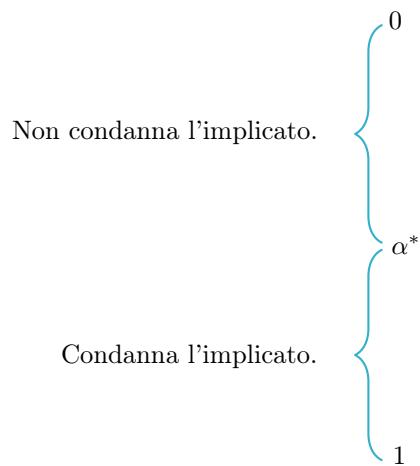
in base a delle prove raccolte che lo aiuteranno a prendere una decisione (X).

Il giudice vuole prendere una decisione di livello α , cioè la sua tolleranza al errore di primo tipo P_{H_0} (rifiuto H_0) è proprio α . Ma cosa vuol dire tutto ciò? Ciò vuol dire che il giudice non vuole condannare persone che in realtà sono innocenti ammettendo che nella sua decisione ci possa essere un errore pari ad α .

Uno si potrebbe chiedere, cosa succederebbe se il giudice non volesse *mai* condannare degli innocenti? La soluzione per non condannare mai degli innocenti ($\alpha = 0$) sarebbe non condannare nessuno, e ciò è inaccettabile.

Se invece non gli interessa proprio condannare degli innocenti ($\alpha = 1$), il giudice dovrebbe condannare tutti gli indagati.

In questa situazione il p-value rappresenta il valore massimo di α per il quale si ritiene innocente l'indagato.



Da questo caso di studio notiamo che il p-value ci misura quanto o meno sia “robusta” una decisione. Inoltre possiamo generalizzare i risultati trovati affermando che le regole di decisione dei test d'ipotesi possono essere viste in funzione del p-value (α^*):

Si accetta H_0	$\forall \alpha \leq \alpha^*$	(7.54)
Si rifiuta H_0	$\forall \alpha > \alpha^*$	

Dato che più α è grande, più severo è il test; più grande è il p-value maggiore è la robustezza della decisione.

Ma in pratica, come si fa a trovare il p-value? Per ottenere il p-value basta applicare la definizione, cioè, ci si mette nel caso limite della regola di decisione e si trova il valore di α . Il valore di α trovato è il p-value (α^*)

ESEMPIO

Si consideri il Test Z unilatero:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned} \tag{7.55}$$

La cui regola di decisione è

$$\text{Rifiutiamo } H_0 \text{ se } \bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha)$$

Ci chiediamo ora quale è il p-value, cioè quale è il valore di α più grande per il quale si accetta H_0 . Quindi ciò che bisogna trovare è il valore di α^* a partire dal quale si comincia a rifiutare, cioè :

$$\mu_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \alpha^*) = \bar{X}_n$$

Svolgendo i calcoli si trova che:

$$\alpha^* = 1 - \Phi\left(\frac{\bar{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \tag{7.56}$$

Per chiarezza riportiamo un'ulteriore forma di vedere il concetto di p-value:

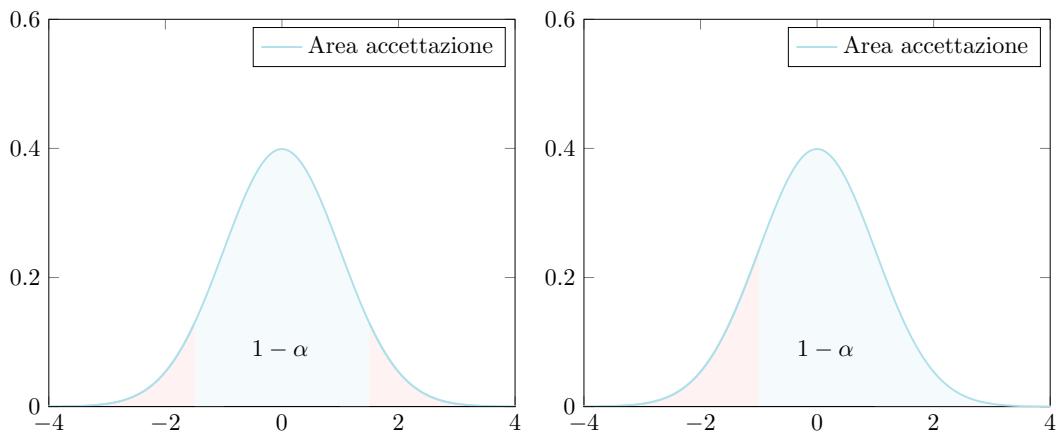


Figura 7.7: A sinistra: Test Bilatero, a destra: Test Unilatero

All'aumentare di α si riduce l'ampiezza dell'intervallo di accettazione e al diminuire di α si aumenta l'ampiezza dell'intervallo di accettazione. Il p-value è il valore di α limite per il quale la statistica legata alla regola di decisione cade nella zona di accettazione.



8. Vettori aleatori

$$\begin{bmatrix} 8 & 11 & 37 & 1 \\ 36 & 2 & 7 & 12 \\ 3 & 39 & 9 & 6 \\ 10 & 5 & 4 & 38 \end{bmatrix}$$

(La matrice preferita del Prof. Battistini)

8.1 Funzione generatrice dei momenti

Prima dell'introduzione dei vettori aleatori, è conveniente introdurre un concetto molto importante, la **funzione generatrice dei momenti** o MGF (dall'inglese Moment-Generating Function). La funzione generatrice dei momenti è una funzione di una variabile aleatoria X definita come:

$$m_X(t) \stackrel{\text{def}}{=} E[e^{tX}] \quad (8.1)$$

Variabili aleatorie discrete

$$m_X(t) = \sum_x e^{tx} P_X(x)$$

Variabili aleatorie continue

$$m_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x) dx$$

Questa funzione gode di varie proprietà :

① *Prodotto per uno scalare*

$$m_{\alpha X}(t) = E[e^{t\alpha X}] = E[e^{\alpha tX}] = m_X(\alpha t)$$

② *Traslazione*

$$m_{X_1+\beta}(t) = E[e^{t(X_1+\beta)}] = E[e^{tX_1} \cdot e^{t\beta}] = e^{t\beta} m_X(t)$$

③ *Biunivocità* tra MGF e PMF/PDF , cioè per ogni densità continua o densità discreta esiste una e una sola funzione generatrice dei momenti e viceversa, ogni funzione generatrice dei momenti è associata ad una e una sola densità .

- ④ Tutte le MGF sono *funzioni analitiche* in un intorno dell'origine (sviluppabili in serie di MacLaurin).

$$m_X(t) = m_X(0) + m'_X(0)t + m''_X(0)\frac{t^2}{2} + \dots + o(t^2) = \sum_{k=0}^{+\infty} m_X^{(k)}(0) \frac{t^k}{k!}$$

- ⑤ Per tutte le MGF si verifica che $m_X(0) = 1$
- ⑥ Se X_1 e X_2 sono variabili aleatorie indipendenti si ha

$$m_{X_1+X_2}(t) = E[e^{t(X_1+X_2)}] = E[e^{tX_1} e^{tX_2}] \stackrel{\text{ind}}{=} E[e^{tX_1}] E[e^{tX_2}] = m_{X_1}(t) \cdot m_{X_2}(t)$$

- ⑦ *Bicontinuità*. la biunivocità è conservata al passaggio al limite.

$$\begin{array}{ccc} \overline{X}_n & \xrightarrow[n \rightarrow \infty]{L} & Y \\ \uparrow & & \uparrow \\ m_{\overline{X}_n}(t) & \xrightarrow[n \rightarrow \infty]{} & m_Y(t) \end{array} \quad (8.2)$$

NB Le proprietà 1,2 e 6 della MGF sono molto simili a quelle della trasformata di Laplace.

Ma la proprietà più importante in assoluto, di cui prende il nome MGF, è la proprietà di generare tutti i momenti della variabile aleatoria.

$$\left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0} = E(X^k) \quad (8.3)$$

A questo punto ricaviamo a mo' d'esempio le diverse MGF delle variabili aleatorie studiate finora.

8.1.1 MGF Discrete

MGF degenero

Se la variabile aleatoria è degenero, cioè ha la probabilità concentrata in un unico punto x_0 (la sua PMF è : $P_X(x) = I_{x_0}(x)$) allora la sua MGF è:

$$m_X(t) = \sum_x e^{tx} P_X(x) = e^{tx_0}$$

MGF bernoulliana

Se $X \sim b(\theta)$ si ha:

$$m_X(t) = \sum_x e^{tx} P_X(x) = e^{t \cdot 0} \theta^0 (1-\theta)^{1-0} + e^{t \cdot 1} \theta^1 (1-\theta)^{1-1} = (1-\theta) + e^t \theta$$

MGF binomiale

Se $Y_n \sim bin(n, \theta)$ si ha:

$$m_{Y_n}(t) = \sum_y e^{ty} P_{Y_n}(y) \stackrel{\text{prop. 6}}{=} ((1-\theta) + e^t \theta)^n$$

Alternativamente, non sfruttando la proprietà 6 si ha:

$$m_{Y_n}(t) = \sum_y e^{ty} P_{Y_n}(y) = \sum_y e^{ty} \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Ricordando che $(a+b)^n = \sum_{x=0}^n \binom{n}{x} a^x b^{n-x}$ e che in questo caso $a = e^t \theta$ e $b = (1-\theta)$ si trova la stessa espressione di prima.

MGF geometrica

Se $X \sim geom(\theta)$ si ha:

$$\begin{aligned} m_X(t) &= \sum_x e^{tx} P_X(x) = \sum_{x=1}^{+\infty} e^{tx} \theta (1-\theta)^{x-1} \xrightarrow{s=x-1} \sum_{s=0}^{+\infty} e^{t(s+1)} \theta (1-\theta)^s \\ &= \theta e^t \sum_{s=0}^{+\infty} e^{ts} (1-\theta)^s = e^t \sum_{s=0}^{+\infty} (e^t (1-\theta))^s = \text{serie geometrica di } q = e^t (1-\theta) \end{aligned}$$

Ricordando che la serie geometrica converge a $\frac{1}{1-q}$ per $|q| < 1$ si ha:

$$m_X(t) = \frac{e^t \theta}{1 - e^t (1-\theta)} I_{\{-\infty, -\ln(1-\theta)\}}(t)$$

NB

$-\ln(1-\theta)$ si è ricavato imponendo $|q| < 1$ cioè: $e^t (1-\theta) < 1 \rightarrow t < \ln(\frac{1}{1-\theta})$

MGF binomiale negativa

Se $X_r \sim bineg(r, \theta)$ si ha

$$m_{\{X_1+X_2+\dots+X_r\}}(t) = \sum_x e^{t(x_1+x_2+\dots+x_r)} P_{X_r}(x) \stackrel{\text{prop. 6}}{=} \left(\frac{e^t \theta}{1 - e^t (1-\theta)} \right)^r I_{\{-\infty, -\ln(1-\theta)\}}(t)$$

MGF poissoniana

Se $X \sim poisson(\lambda)$ si ha:

$$m_X(t) = \sum_{x=0}^{+\infty} e^{tx} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{+\infty} \frac{(e^t \lambda)^x}{x!} = e^{-\lambda} e^{e^t \lambda} = e^{\lambda(e^t - 1)}$$

8.1.2 MGF continue**MGF esponenziale**

Se $X \sim exp(v)$

$$m_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x) dx = \int_0^{+\infty} e^{tx} v e^{-vx} dx = -\frac{v}{v-t} \left[e^{-x(v-t)} \right]_0^{+\infty} = \frac{v}{v-t} I_{\{-\infty, v\}}(t)$$

NB

Per $v \leq t$ si ha una funzione esponenziale divergente, è per questo motivo che la MGF vale per $v > t$

MGF gamma

Se $X_r \sim \Gamma(r, v)$ si ha

$$m_{\{X_1+X_2+\dots+X_r\}}(t) = \int_{\mathbb{R}} e^{t(x_1+x_2+\dots+x_r)} f_{X_r}(x) \stackrel{\text{prop. 6}}{=} \left(\frac{v}{v-t}\right)^r I_{\{-\infty, v\}}(t)$$

MGF chi-quadro

Ricordando che $\chi^2 \equiv \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ si ottiene immediatamente dalla MGF della gamma:

$$m_{\chi_{(n)}^2}(t) = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n}{2}} I_{\{-\infty, \frac{1}{2}\}}(t)$$

MGF uniforme continua

Se $X \sim \text{unif}[a, b]$

$$m_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x) dx = \int_a^b e^{tx} \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{e^{tx}}{t} \right]_a^b = \frac{1}{b-a} \left(\frac{e^{tb} - e^{ta}}{t} \right) = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

NB

Si noti che per una uniforme in $t = 0$ si ha una discontinuità ; per fortuna questa discontinuità è di tipo eliminabile. In modo rigoroso la MGF ad esempio per una uniforme in $[0, 1]$ è:

$$m_X(t) \begin{cases} 1 & t = 0 \\ \frac{e^t - 1}{t} & t \neq 0 \end{cases} \quad \text{in questi casi si calcola} \quad \lim_{\epsilon \rightarrow 0} \frac{d^k m_X(t)}{dt^k} \Big|_{t=\epsilon} = E(X^k)$$

MGF normale

Se $Z \sim N(0, 1)$ allora:

$$m_Z(t) = \int_{\mathbb{R}^+} e^{tx} f_X(x) dx = \int_{-\infty}^{+\infty} e^{tx} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \int_{-\infty}^{+\infty} \frac{e^{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)}}{\sqrt{2\pi}} dx = e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{+\infty} \frac{e^{-\frac{1}{2}(x-t)^2}}{\sqrt{2\pi}} dx}_{\sim N(t, 1)}$$

poiché l'integrale è per definizione uguale a 1: $m_Z(t) = e^{\frac{t^2}{2}}$.

Per una normale e $W \sim N(\mu, \sigma^2)$ che può essere scritta come $W = \sigma Z + \mu$ si deduce facilmente dalla proprietà 1 e 2 che :

$$m_Z(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

8.1.3 Non esistenza MGF**MGF student-t**

La variabile aleatoria $T \sim t_{(n)}$ non ammette MGF poiché ammetterne una, vorrebbe dire che ogni momento può essere calcolato derivando opportunamente la funzione; sfortunatamente è noto che una variabile T di student a n gradi di libertà non ammette momenti di grado maggiore a n , rendendo impossibile il fatto che ci sia una funzione generatrice dei momenti per la variabile aleatoria.

MGF log-normale

Sia $Y = e^Z$ con $Z \sim N(0, 1)$ la variabile Y si distribuisce come: $F_Y(y) = P(Y \leq y) = P(e^Z \leq y) = P(Z \leq \ln(y)) = \Phi(\ln(y))$ Per cui derivando si ottiene:

$$f_X(y) = \frac{\phi(\ln(y))}{y} = \frac{1}{\sqrt{2\pi}y} e^{-\frac{(\ln(y))^2}{2}}$$

Questa variabile prende il nome di **log-normale**



La variabile log-normale, pur avendo momenti di ogni ordine, non ammette una MGF. [Cosa simile a quello che accade con alcune funzioni C^∞ che non sono però analitiche] Il problema è che non soddisfa la proprietà che $m_X(0) = 1$. Infatti si verifica che $\forall k \geq 0$

$$\int_0^{+\infty} y^{k-1} \frac{e^{-\frac{(\ln(y))^2}{2}}}{\sqrt{2\pi}} dy = 0$$

8.1.4 Dimostrazioni con MGF

Con la funzione generatrice dei momenti possono essere dimostrati i teoremi più importanti della probabilità anche se in una loro formulazione più debole (poiché si richiede l'esistenza di una MGF più l'analiticità della stessa, e ciò non è sempre garantito).

Legge debole dei grandi numeri

Dimostriamo la convergenza della media campionaria \bar{X}_n al valore atteso per $n \rightarrow \infty$.

$$m_{\bar{X}_n}(t) = m_{\frac{1}{n} \sum_{i=1}^n X_i}(t) \stackrel{\text{prop.1}}{=} m_{\sum_{i=1}^n X_i} \left(\frac{t}{n} \right) \stackrel{\text{prop.2}}{=} \left(m_X \left(\frac{t}{n} \right) \right)^n$$

Se facciamo tendere $n \rightarrow \infty$ la MGF $m_{\bar{X}_n}(t)$ tenendo ben presente il limite notevole $e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$ si ha:

$$\lim_{n \rightarrow \infty} \left(m_X \left(\frac{t}{n} \right) \right)^n \stackrel{\text{prop. 4}}{=} \lim_{n \rightarrow \infty} \left(1 + \frac{\mu t}{n} + o\left(\frac{1}{n}\right) \right)^n = e^{\mu t}$$

Il termine trovato non è altro che la MGF di una variabile aleatoria degenere in μ . Per la proprietà di bicontinuità si ha:

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{L} \mu$$

Teorema Centrale del Limite

Dimostriamo la convergenza della media campionaria standardizzata $(\bar{X}_n)^*$ alla CDF di una $Z \sim N(0, 1)$.

$$m_{(\bar{X}_n)^*}(t) = m_{\frac{\sum_{i=1}^n X_i^*}{\sqrt{n}}}(t) \stackrel{\text{prop.1}}{=} m_{\sum_{i=1}^n X_i^*} \left(\frac{t}{\sqrt{n}} \right) \stackrel{\text{prop.2}}{=} \left(m_{X^*} \left(\frac{t}{\sqrt{n}} \right) \right)^n$$

Facendo tendere $n \rightarrow \infty$ e ricordando che la MGF di una normale/standardizzata è $e^{\frac{t^2}{2}}$ e il limite notevole di e^x si ha:

$$\lim_{n \rightarrow \infty} \left(m_{X^*} \left(\frac{t}{\sqrt{n}} \right) \right)^n \stackrel{e^t = 1 + t + \dots}{=} \lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right)^n = e^{\frac{t^2}{2}}$$

Per la proprietà di bicontinuità è immediato vedere che :

$$(\bar{X}_n)^* \xrightarrow[n \rightarrow \infty]{L} Z \sim N(0, 1)$$

8.2 Funzione caratteristica

Abbiamo visto che data una variabile aleatoria, non è detto che esista la sua MGF. La funzione caratteristica o CF (*Characteristic Function*) è una funzione complessa di variabile aleatoria che esiste qualunque sia la variabile aleatoria, e estende il concetto introdotto dalla MGF. Questa funzione non verrà trattata poiché esula dagli scopi di questo libro, ad ogni modo diamo la definizione e un paio di considerazioni.

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \int_{\mathbb{R}} e^{itx} dF_X(x) = \int_{-\infty}^{+\infty} f_X(x) e^{itx} dx. \quad (8.4)$$



Così come la MGF era strettamente legata alla trasformazione di Laplace, la CF è strettamente legata alla trasformata di Fourier.

8.3 Vettori aleatori

Un vettore aleatorio è una variabile aleatoria a valori in \mathbb{R}^n , in altre parole, un vettore aleatorio non è altro che un vettore colonna dove ogni componente è una variabile aleatoria continua.

$$\underline{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad (8.5)$$

Si definisce allo stesso modo che per le variabili unidimensionali la probabilità e la funzione di distribuzione cumulativa.

$$P(\underline{X} = \underline{x}) \stackrel{\text{def}}{=} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \quad (8.6)$$

$$F_{\underline{X}}(\underline{x}) \stackrel{\text{def}}{=} P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \quad (8.7)$$

D'ora in poi ci concentreremo principalmente di vettori aleatori sia continui che discreti in \mathbb{R}^2 .

8.3.1 PDF e CDF

Sussiste un legame fra PDF e CDF analogo a quello unidimensionale. La densità per un vettore aleatorio in \mathbb{R}^2 è definita come:

$$f_{X_1, X_2}(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} F_{X_1, X_2}(x_1, x_2) \quad (8.8)$$

La densità marginale per un vettore aleatorio in \mathbb{R}^2 è definita come l'integrale sulla variabile "da eliminare":

$$f_{X_1}(x_1) = \int_{\mathbb{R}} f_{X_1, X_2}(x_1, x_2) dx_2 \quad (8.9)$$

Il contatore è definito come l'integrale del quarto di piano sinistro inferiore al punto:

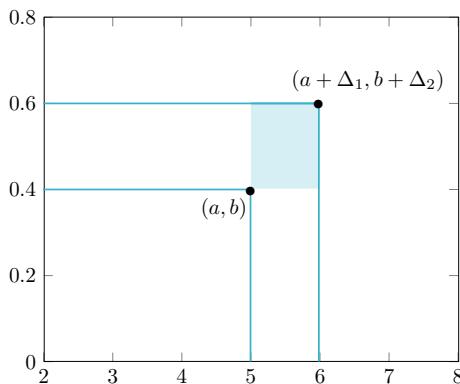
$$F_{X_1, X_2}(x_1, x_2) = \iint_{\substack{t_1 \leq x_1 \\ t_2 \leq x_2}} f_{X_1, X_2}(t_1, t_2) dt_1 dt_2 \quad (8.10)$$

È evidente l'estensione al caso \mathbb{R}^n .

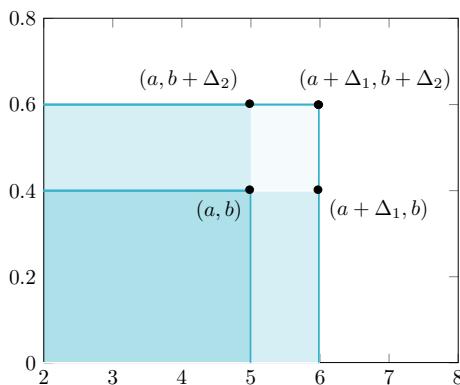
NB

Il teorema di Schwarz ci dice che sotto opportune condizioni l'ordine di derivazione parziale è indifferente. Il teorema di Fubini ci dice che sotto opportune condizioni l'ordine d'integrazione è indifferente.

Per dimostrare questo legame consideriamo la situazione in figura:



Si vuole stabilire qual è la probabilità puntuale, cioè la densità continua. Per stabilire la probabilità puntuale prendiamo una sezione di piano, sommiamo la probabilità in essa contenuta e dividiamo per l'area della sezione considerata. In questo modo, facendo tendere Δ_1 e Δ_2 a zero avremo la probabilità puntuale cercata.



Per ottenere l'area desiderata sottraiamo da $F(a + \Delta_1, b + \Delta_2)$ sia $F(a, b + \Delta_2)$ che $F(a + \Delta_1, b)$ che togliono i rettangoli $(0, b) \times (a, b + \Delta_2)$ e $(a, 0) \times (a + \Delta_1, b)$.

Il problema è che togliendo i rettangoli sopracitati si toglie due volte il rettangolo $F(a, b)$ quindi dobbiamo risommarlo.

Con queste considerazioni si trova:

$$\frac{F(a + \Delta_1, b + \Delta_2) - F(a, b + \Delta_2) - F(a + \Delta_1, b) + F(a, b)}{\Delta_1 \Delta_2}$$

Questa la possiamo riscrivere come:

$$\begin{aligned} \frac{\frac{F(a + \Delta_1, b + \Delta_2) - F(a, b + \Delta_2)}{\Delta_1} - \frac{F(a + \Delta_1, b) - F(a, b)}{\Delta_1}}{\Delta_2} &= \begin{array}{ll} \Delta_1 \rightarrow 0 & \frac{\partial}{\partial \Delta_1} F(a, b + \Delta_2) - \frac{\partial}{\partial \Delta_1} F(a, b) \\ \Delta_2 \rightarrow 0 & \Delta_2 \end{array} \\ \frac{\partial^2}{\partial \Delta_1 \partial \Delta_2} F(a, b) &= f_{X_1, X_2}(x_1, x_2) \end{aligned}$$

Proprietà della CDF

La CDF o “contatore”, gode di proprietà molto simili a quelle del caso unidimensionale:

- ① Se *una* componente x_i tende a $-\infty$ allora:

$$\lim_{x_i \rightarrow -\infty} F_{\underline{X}}(\underline{x}) = 0$$

- ② Se *tutte* le componenti (x_1, x_2 per un vettore bidimensionale) tendono a $+\infty$ allora :

$$\lim_{\substack{x_1 \rightarrow +\infty \\ x_2 \rightarrow +\infty}} F_{\underline{X}}(\underline{x}) = 1$$

- ③ Se *una* componente tende a $+\infty$ allora si ottiene un *contatore marginale*, cioè un contatore senza la variabile che si è fatta tendere all’infinito. Nel caso di due dimensioni, si passa da un quarto di piano a un semipiano:

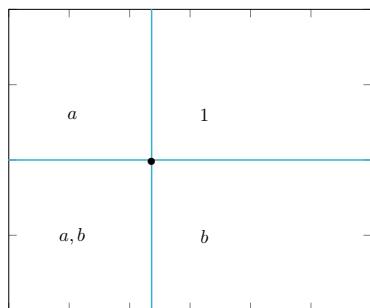
$$\lim_{x_1 \rightarrow +\infty} F_{\underline{X}}(\underline{x}) = F_{X_2}(x_2)$$

- ④ $F_{\underline{X}}(\underline{x})$ è continua da destra per ogni variabile:

$$F_{\underline{X}}(\hat{x}, x_2) = \lim_{x_1 \rightarrow \hat{x}^+} F_{\underline{X}}(x_1, x_2) \quad \wedge \quad F_{\underline{X}}(x_1, \hat{x}) = \lim_{x_2 \rightarrow \hat{x}^+} F_{\underline{X}}(x_1, x_2)$$

- ⑤ È una funzione monotona crescente in ogni componente se le altre rimangono fissate. Oppure se $\underline{a} < \underline{b}$ allora si ha:

$$F_{\underline{X}}(\underline{a}) < F_{\underline{X}}(\underline{b}) \quad \Rightarrow \quad 0 < F_{\underline{X}}(\underline{b}) - F_{\underline{X}}(\underline{a}) \quad \Rightarrow \quad P(\underline{a} < \underline{X} \leq \underline{b}) > 0$$



Per scrivere l’espressione di un contatore bisogna individuare i diversi casi ove possono variare le variabili e calcolare il contatore di ciascuna zona. Prendiamo come esempio il vettore uniforme.

$$f_{X_1, X_2}(x_1, x_2) = I_{[0,1] \times [0,1]}(\underline{x})$$

$$F_{X_1, X_2}(x_1, x_2) = (a \cdot b) I_{[0,1] \times [0,1]}(\underline{x}) + I_{(1,+\infty) \times (1,+\infty)}(\underline{x}) + a I_{[0,1] \times (1,+\infty)}(\underline{x}) + b I_{(1,+\infty) \times [0,1]}(\underline{x})$$

In due variabili, se il supporto di un vettore aleatorio non ha la struttura di prodotti cartesiani (rettangoli o un rettangolo fatto di rettangoli disgiunti), si può concludere immediatamente che non c’è indipendenza fra le variabili che compongono il vettore. La indipendenza di variabile vuol dire che la CDF può essere ottenuta come il prodotto dei singoli contatori marginali.

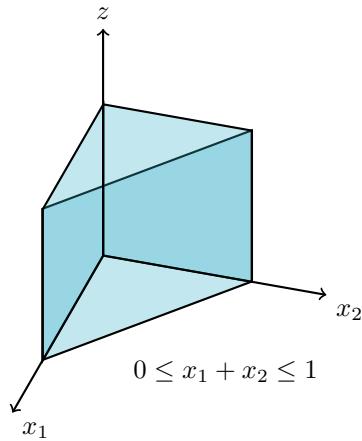


Figura 8.1:

ESEMPIO

$$f_{X_1, X_2} = 2 I(x_1, x_2) \text{ t.c. } 0 \leq x_1 + x_2 \leq 1$$

Dalle considerazioni precedenti, le variabili non sono indipendenti. I contatori marginali sono definiti come:

$$F_{X_1}(x_1) = (2 - 2x_1)I_{[0,1]}(x_1) + I_{(1,+\infty)}(x_1)$$

$$F_{X_2}(x_2) = (2 - 2x_2)I_{[0,1]}(x_2) + I_{(1,+\infty)}(x_2)$$

In questo caso la funzione \$x_1 + x_2 \leq 1\$ è invariante rispetto alla permutazione delle variabili, in questi casi si dice che le variabili sono **variabili scambiabili**. In generale, se il grafico è simmetrico rispetto alla bisettrice, le variabili sono scambiabili.

8.3.2 Momenti

Il valore atteso (momento di ordine 1) di un vettore aleatorio è definito come il valore atteso delle singole componenti.

$$E[\underline{X}] = \begin{bmatrix} E[X_1] \\ E[X_2] \end{bmatrix} = \begin{bmatrix} \iint_{\mathbb{R}^2} x_1 f_{\underline{X}}(\underline{x}) \\ \iint_{\mathbb{R}^2} x_2 f_{\underline{X}}(\underline{x}) \end{bmatrix} \quad (8.11)$$

Per i momenti successivi, il momento \$k\$-esimo di un vettore aleatorio sarà composto da una matrice simmetrica con tutti i termini di grado \$k\$-esimo, ad esempio per il momento di grado 2 di un vettore a due dimensioni, la matrice sarà fatta così:

$$E(\underline{X}) = \begin{bmatrix} E[X_1^2] & E[X_1, X_2] \\ E[X_1, X_2] & E[X_2^2] \end{bmatrix} \quad (8.12)$$

8.3.3 Varianza e Covarianza

La covarianza fra due vettori aleatori \$X, Y\$ è definita attraverso una matrice detta matrice di covarianza. Questa matrice ha come elementi \$\text{Cov}(X_i, Y_j)\$:

$$\text{Cov}(\underline{X}, \underline{Y}) = \begin{bmatrix} \text{Cov}(X_1, Y_1) & \text{Cov}(X_1, Y_2) & \dots & \text{Cov}(X_1, Y_n) \\ \text{Cov}(X_2, Y_1) & \text{Cov}(X_2, Y_2) & \dots & \text{Cov}(X_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, Y_1) & \text{Cov}(X_n, Y_2) & \dots & \text{Cov}(X_n, Y_n) \end{bmatrix} \quad (8.13)$$

Poiché la varianza può essere vista come un caso particolare di covarianza (ciò è evidenziabile facilmente dalle proprietà di riscrittura, basta porre \$Y = X\$ per ottenere \$\text{Var}(X) = \text{Cov}(X, X)\$):

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] \quad \text{se } X = Y \Rightarrow \quad E[(X - E[X])(X - E[X])] = \text{Var}(X)$$

Con queste considerazioni, possiamo scrivere la varianza come la covarianza fra X e se stesso:

$$Var(\underline{X}) = \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \dots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \dots & Cov(X_n, X_n) \end{bmatrix} \quad (8.14)$$

Che per un vettore aleatorio di due dimensioni diventa:

$$Var(\underline{X}) = \Sigma_{\underline{X}} \stackrel{\text{def}}{=} \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{bmatrix} \quad (8.15)$$

Per quanto evidenziato, la matrice $\Sigma_{\underline{X}}$ è a volte chiamata *matrice di varianza-covarianza*, noi la chiameremo semplicemente matrice di covarianza di un vettore aleatorio.

Valgono le proprietà per il valore atteso e la matrice di varianza-covarianza:

1 *Linearità affine.*

$$E[A\underline{X} + B] = AE[\underline{X}] + B$$

2 *Riscrittura.*

$$\Sigma_{\underline{X}} = E[(\underline{X} - E[\underline{X}])(\underline{X} - E[\underline{X}])^T]$$

3 Omogeneità di secondo grado

$$Var(A\underline{X}) = A \cdot \Sigma_{\underline{X}} \cdot A^T$$

4 $\Sigma_{\underline{X}}$ è definita positiva, cioè tutti i suoi autovalori sono positivi, di conseguenza:

$$\text{tr}(\Sigma_{\underline{X}}) > 0 \quad \det(\Sigma_{\underline{X}}) > 0$$

5 $\Sigma_{\underline{X}}$ è una matrice simmetrica poiché:

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_1)$$

6 *Riscrittura.*

$$\text{Cov}(\underline{X}, \underline{Y}) = E[(\underline{X} - E[\underline{X}])(\underline{Y} - E[\underline{Y}])^T] = E[\underline{XY}^T] - E[\underline{X}]E[\underline{Y}]^T$$

7 *Quasi simmetrico.*

$$\text{Cov}(\underline{X}, \underline{Y}) = [\text{Cov}(\underline{Y}, \underline{X})]^T$$

8 *Bilinearità.*

$$\text{Cov}(\underline{X} + \underline{Z}, \underline{Y}) = \text{Cov}(\underline{X}, \underline{Y}) + \text{Cov}(\underline{Z}, \underline{Y})$$

$$\text{Cov}(\underline{X}, \underline{Y} + \underline{Z}) = \text{Cov}(\underline{X}, \underline{Y}) + \text{Cov}(\underline{X}, \underline{Z})$$

8.3.4 Coefficiente di correlazione

Il *coefficiente di correlazione* indicato con ρ è definito come :

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sigma_1 \sigma_2} \quad (8.16)$$

In realtà il coefficiente non è altro che la covarianza standardizzata:

$$\frac{E[(X_1 - E[X_1])(X_2 - E[X_2])^T]}{\sigma_1 \sigma_2} = E\left[\left(\frac{X_1 - E[X_1]}{\sigma_1}\right)\left(\frac{X_2 - E[X_2]}{\sigma_2}\right)^T\right] = \text{Cov}(X_1^*, X_2^*)$$

Avendo definito il coefficiente di correlazione possiamo scrivere $\text{Cov}(X_1, X_2) = \rho \sigma_1 \sigma_2$ e dunque riscrivere la matrice di covarianza come:

$$\Sigma_{\underline{X}} = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}$$

Si noti che $-1 \leq \rho \leq 1$, a seconda dei diversi valori di ρ si possono dunque distinguere vari casi:

- Se $\rho > 0$, le variabili X e Y si dicono **direttamente correlate**.

Se $0 \leq \rho \leq 0,3$ la correlazione si dice **debole**.

Se $0,3 \leq \rho < 0,7$ la correlazione si dice **moderata**.

Se $0,7 \leq \rho < 1$ la correlazione si dice **forte**.

- Se $\rho = 0$, le variabili X e Y si dicono **incorrelate**.
- Se $\rho < 0$, le variabili X e Y si dicono **inversamente correlate**.
- $|\rho| = 1$ solo se X è ottenibile come combinazione lineare di Y o viceversa.



Si noti che l'indipendenza implica l'incorrelazione:

$$X \text{ e } Y \text{ indipendenti} \rightarrow \text{Cov}(X, Y) = 0 \quad \text{Cov}(X, Y) = 0 \leftrightarrow \rho = 0$$

Ma l'incorrelazione non sempre implica l'indipendenza, cioè, è possibile trovare che ρ sia nullo senza che le variabili siano indipendenti. Come vedremo, solo in caso di normalità congiunta, l'incorrelazione implica l'indipendenza.

8.4 Normale bivariata e multivariata

Sia \underline{Z} un vettore aleatorio, se il vettore aleatorio ha come componenti variabili aleatorie $Z_i \sim N(0, 1)$ indipendenti fra loro, allora il suo valore atteso sarà la matrice nulla e la sua matrice di covarianza è la matrice identità. In questo caso si dice che il vettore si distribuisce come una normale bivariata (nel caso che la dimensione del vettore sia 2) o multivariata (nel caso di dimensione superiore a 2) e si scrive:

$$\underline{Z} \sim N(\underline{0}, I_n)$$

Sempre sotto l'ipotesi di indipendenza, la densità della normale bivariata, cioè della normale standardizzata in due dimensioni è:

$$f_{\underline{Z}}(\underline{x}) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} \quad (8.17)$$

Più in generale possiamo dire che la densità della normale bivariata standardizzata con matrice di covarianza:

$$\Sigma_{\underline{X}} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

cioè indipendentemente se le variabili sono dipendenti o indipendenti è:

$$f_{\underline{Z}}(\underline{x}) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)} \quad (8.18)$$

NB Si noti che la formula 8.17 è un caso particolare della 8.18 con $\rho = 0$ (incorelate per ipotesi di indipendenza).

NB Si noti che è possibile passare da

$$\Sigma_{\underline{X}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \xrightarrow{AX} \quad \Sigma_{\underline{X}} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

moltiplicando il vettore $\underline{X} \sim N(\underline{0}, I_n)$ per una matrice A triangolare superiore:

$$A = \begin{bmatrix} \sqrt{1-\rho^2} & \rho \\ 0 & 1 \end{bmatrix}$$

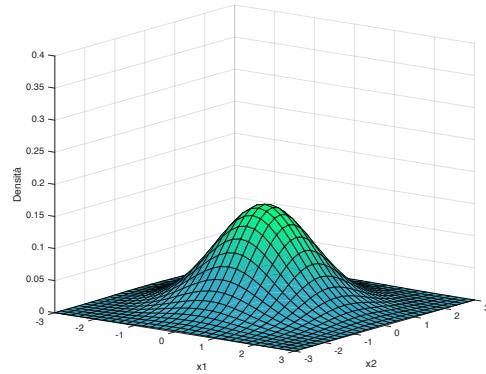


Figura 8.2: Grafico della normale bivariata standardizzata di variabili indipendenti $\rho = 0$

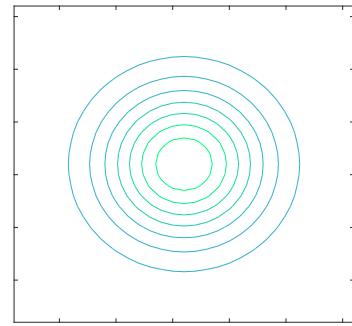


Figura 8.3: Le curve di livello della normale bivariata standardizzata di variabili indipendenti sono delle circonferenze

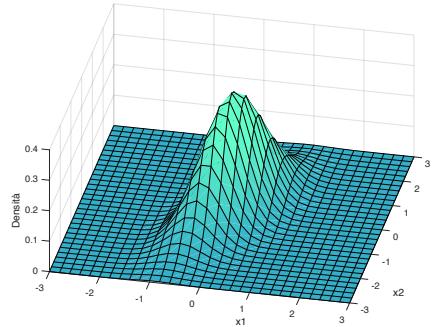


Figura 8.4: Grafico della normale bivariata standardizzata di variabili correlate moderatamente $\rho = 2/3$

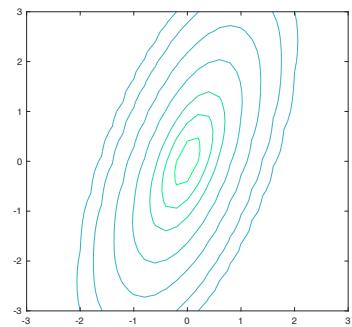


Figura 8.5: Le curve di livello della normale bivariata sono delle ellissi centrate in 0 e orientate lungo la retta $x_2 = \frac{3}{2} x_1$

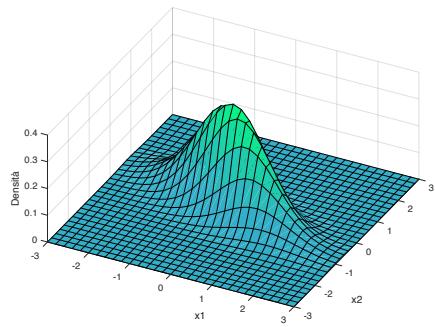


Figura 8.6: Grafico della normale bivariata standardizzata di variabili di variabili correlate moderatamente $\rho = -2/3$

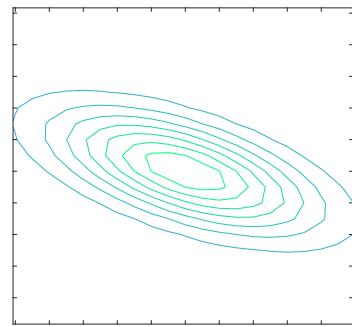
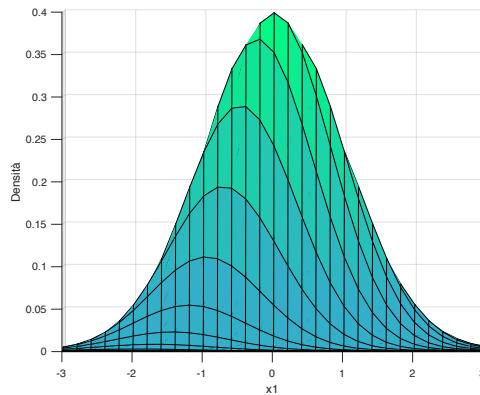


Figura 8.7: Le curve di livello della normale bivariata sono delle ellissi centrate in 0 e orientate lungo la retta $x_2 = -\frac{2}{3} x_1$

Come è evidenziato dai grafici vale:

- Se $\rho = 1$ il grafico degenera nella bisettrice del primo e terzo quadrante.
- Se $0 < \rho < 1$ il grafico risulta simmetrico rispetto la retta $x_2 = \frac{1}{\rho} x_1$ quindi si approssima sempre di più alla bisettrice $I - III$ al aumentare di ρ .
- Se $\rho = 0$, il grafico ha infiniti assi di simmetria (simmetria assiale).
- Se $-1 < \rho < 0$ il grafico risulta simmetrico rispetto la retta $x_2 = \rho x_1$ quindi si approssima sempre di più alla bisettrice $II - IV$ al diminuire di ρ .
- Se $\rho = -1$ il grafico degenera nella bisettrice del secondo e quarto quadrante.
- Le sezioni parallele e ortogonali lungo l'asse di simmetria sono delle campane gaussiane.



- Le ellissi delle curve di livello sono più schiacciate lungo la variabile con varianza minore, quindi:

Se $\sigma_1 > \sigma_2$ sono più schiacciate lungo x_2

Se $\sigma_2 > \sigma_1$ sono più schiacciate lungo x_1

- Se $\sigma_1 = \sigma_2 = 0$ il grafico degenera in un punto.
- Applicare una rotazione di un qualunque angolo θ alla normale bivariata di parametro $\rho = 0$, non cambia la distribuzione grazie alla simmetria assiale.
- Se si applica una rotazione di $\theta = \pi/2$ a una normale bivariata di $\rho = |1|$, la distribuzione ha la stessa distribuzione normale bivariata ma di parametro $-\rho$

Qualora si volesse generalizzare a più dimensioni, la formula per la normale standardizzata multivariata di variabili **indipendenti** è:

$$f_{\underline{Z}}(\underline{x}) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}(\underline{x}^T \underline{x})} \quad (8.19)$$

Così come nel caso unidimensionale, vogliamo vedere come si distribuisce un vettore aleatorio che sia combinazione lineare di una normale bivariata, cioè vogliamo capire qual è la densità di $\underline{W} = A\underline{Z} + \mu$.

Si trova che nel caso multidimensionale la densità di una normale multivariata è:

$$f_{\underline{W}}(\underline{v}) = \frac{e^{-\frac{1}{2}(\underline{v}-\mu)^T \Sigma_W^{-1} (\underline{v}-\mu)}}{(\sqrt{2\pi})^n |\det \Sigma_W|^{1/2}} \quad (8.20)$$

Nel caso di due dimensioni si trova che la densità della normale bivariata di media μ e matrice di covarianza $\Sigma_W = AA^T$ è:

$$f_{\underline{W}}(\underline{v}) = \frac{e^{-\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right]}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \quad (8.21)$$

NB

Si noti che la variabile si distribuisce come $\underline{W} \sim N(\mu, AA^T)$ e che $\Sigma_W = AA^T$ poiché:

$$\Sigma_W = \text{Var}(\underline{W}) = \text{Var}(A\underline{Z} + \mu) = \text{Var}(A\underline{Z}) = A \text{Var}(\underline{Z}) A^T = A \Sigma_Z A^T$$

Ricordando che per un vettore normale $\Sigma_Z = I_n$ (identità) si ha per definizione di identità

$$A \Sigma_Z A^T = A I_n A^T = A A^T$$

8.5 Normale bivariata marginale e condizionata

8.5.1 Normale bivariata marginale

Integrando lungo la variabile che si vuole eliminare, si trova che la distribuzione marginale della normale bivariata di media μ e $\sigma_1 = \sigma_2 = 1$ è ancora una normale, **qualunque sia il valore di ρ** .

$$f_{X_i}(x_i) = \frac{e^{-\frac{1}{2}x_i^2}}{\sqrt{2\pi}} \quad i = 1, 2 \quad (8.22)$$

NB

Analogamente a quello che succedeva nel caso monodimensionale, se le variabili aleatorie sono indipendenti, la densità congiunta può essere espressa come il prodotto delle densità marginali.

$$\text{Indipendenza} \Rightarrow f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$$

8.5.2 Normale bivariata condizionata

Se si vuole applicare un condizionamento ad un vettore aleatorio, si definisce la densità condizionata come:

$$f_{X_1|X_2}(x_1|x_2) \stackrel{\text{def}}{=} \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} \quad (8.23)$$

Se si svolgono i conti si trova che:

$$f_{X_1|X_2}(x_1|x_2) = \frac{e^{-\frac{1}{2(1-\rho^2)}[x_1^2 - 2\rho x_1 x_2 + \rho^2 x_2^2]}}{\sqrt{2\pi} \sqrt{1-\rho^2}} \quad (8.24)$$

La densità condizionata è ancora una normale, ma monodimensionale di media ρx_2 (se si condiziona x_2) o ρx_1 (se si condiziona x_1) e di varianza $1 - \rho^2$.

Applicare un condizionamento, significa sezionare o "affettare" verticalmente la densità , con un piano parallelo al piano xy , posizionato nello spazio a seconda del condizionamento.

Argomento strettamente legato al condizionamento sono i massimi e minimi vincolati. Come sappiamo dall'analisi, gli eventuali massimi e minimi vincolati sono tangentie alle curve di livello quindi un modo veloce di trovare i massimi vincolati e vedere i punti di tangenza o più facilmente, trovare l'intersezione dell'asse di simmetria della distribuzione bivariata e la retta (proiettata) che impone il piano condizionante.

È anche utile definire il valore atteso condizionato:

$$E[X|Y=y] \stackrel{\text{def}}{=} \int_{\mathbb{R}} x f_{X|Y}(x|y) \quad E[Y|X=x] \stackrel{\text{def}}{=} \int_{\mathbb{R}} y f_{Y|X}(y|x) \quad (8.25)$$

Si nota che il valore atteso condizionato non è altro che il baricentro della curva normale creata dal sezionamento, cioè, ad esempio $E[Y|X=x]$ è la migliore previsione che uno può fare sul valore di Y sapendo che $X=x$

Teorema

Se un vettore aleatorio ha distribuzione normale multivariata/bivariata e tutte le componenti sono incorrelate fra loro, allora le componenti sono tra loro indipendenti. Questo vuol dire che dimostrare l'indipendenza delle componenti di un vettore aleatorio con distribuzione normale multivariata si riduce al semplice calcolo della covarianza delle componenti.

NB

Si faccia attenzione al fatto che per essere valido il teorema, la distribuzione normale dev'essere congiunta (cioè deve avere distribuzione normale multivariata), non basta che le componenti siano marginalmente normali!

8.5.3 Chiusura della Normale bivariata

Se si ha normalità congiunta e $X_i \sim N(\mu_i, \sigma_i)$ allora $\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, Var(\sum_{i=1}^n X_i))$
 Quello che richiede un'attenzione è la varianza poiché si ricordi che:

$$Var(X_1 + X_2 + X_3) = Var(X_1) + Var(X_2) + Var(X_3) + 2Cov(X_1, X_2) + 2Cov(X_2, X_3) + 2Cov(X_1, X_3)$$

Quindi in generale:

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) + 2 \sum_{1 \leq i < j \leq n} Cov(X_i, X_j)$$

NB

È importante notare che sotto ipotesi di normalità congiunta, la somma di variabili aleatorie sia in caso di indipendenza che in caso di dipendenza, si ha in ogni caso una legge normale. Senza l'ipotesi di normalità congiunta, la somma di variabili aleatorie indipendenti è ancora normale mentre la somma di variabili aleatorie dipendenti non necessariamente è normale.

Più in generale e da un punto di vista alternativo diciamo che un vettore aleatorio ha distribuzione normale multivariata (normalità congiunta) se e solo se tutte le combinazioni lineari delle componenti hanno distribuzione normale.

8.6 Funzione generatrice dei momenti per vettori

Per vettori di definisce la funzione generatrice dei momenti (MGF) come:

$$m_X(t) \stackrel{\text{def}}{=} E[e^{t^T X}] \quad (8.26)$$

Variabili aleatorie continue

$$m_{\underline{X}}(t) = \int \cdots \int_{\mathbb{R}^n} e^{t_1 x_1 + \cdots + t_n x_n} f_{\underline{X}}(x_1, \dots, x_n) dx_1, \dots, dx_n$$

Variabili aleatorie discrete

$$m_{\underline{X}}(\underline{t}) = \sum_{x_1, \dots, x_n} e^{t_1 x_1 + \cdots + t_n x_n} P_{\underline{X}}(x_1, \dots, x_n)$$

Questa funzione gode di varie proprietà analoghe a quelle del caso monodimensionale. La proprietà fondamentale è riscritta come:

$$\left. \frac{\partial^k m_{\underline{X}}(\underline{t})}{\partial t_1^k \partial t_2^k \cdots \partial t_n^k} \right|_{\underline{t}=0} = E[X_i^{k_1} \cdots X_n^{k_n}] \quad (8.27)$$

svolgendo i calcoli si trova:

$$\text{Vettore Normale } \underline{Z} \sim N(\underline{0}, \Sigma_Z)$$

$$\text{Vettore Normale } \underline{W} \sim N(\underline{\mu}, \Sigma_W)$$

$$\begin{aligned} m_{\underline{Z}}(t) &= m_{\underline{Z}}(t_1, \dots, t_n) = m_{Z_1}(t_1) \cdots m_{Z_n}(t_n) & m_{\underline{W}}(t) &= m_{\underline{W}}(t_1, \dots, t_n) = m_{W_1}(t_1) \cdots m_{W_n}(t_n) \\ &= e^{\frac{1}{2}t_1^2} \cdots e^{\frac{1}{2}t_n^2} = e^{\frac{1}{2}(t_1^2 + \cdots + t_n^2)} = e^{\frac{1}{2}(t^T t)} & &= e^{\frac{1}{2}(\underline{t}^T \Sigma_W \underline{t}) + \underline{t}^T \underline{\mu}} \end{aligned} \quad (8.28) \quad (8.29)$$

8.7 Teorema di Cochran

Dato un fenomeno con distribuzione multivariata di dimensione maggiore di 2, come ad esempio in \mathbb{R}^3 risulta impossibile visualizzare la densità della normale multivariata, per fortuna, per visualizzare la densità, ciò che possiamo fare sono le proiezioni. Dalla geometria si sa che un vettore di una qualsiasi dimensione può essere scomposto in due spazi vettoriali ortogonali fra loro, V_1 e V_1^\perp , cioè:

$$\underline{X} = \underline{V} + \underline{W} \quad \underline{x} \in \mathbb{R}^n \quad \underline{v} \in V_1 \quad \underline{w} \in V_1^\perp \quad (8.30)$$

CURIOSITÀ. Una cosa analoga succede per il tesseratto (l'ipercubo quadrimensionale) che anche non essendo fisicamente visualizzabile, può essere visualizzato tramite le sue proiezioni.

Il teorema di Cochran ci dice che se \underline{X} è un vettore aleatorio distribuito come $N(\mu, \sigma^2 I_n)$ e $\mathbb{R}^n = V_1 \oplus V_2 \oplus \dots \oplus V_k$ e P_{V_i} sono le rispettive matrici di proiezione (cioè matrici che se applicate su un vettore producono la proiezione sullo spazio vettoriale considerato) allora:

$$P_{V_i} \underline{X} \perp P_{V_j} \underline{X} \quad (j \neq i) \Rightarrow P_{V_i} \underline{X} \text{ indipendente da } P_{V_j} \underline{X}$$

Cioè se si può proiettare un fenomeno in diversi spazi vettoriali, allora se le proiezioni del vettore sui sottospazi sono ortogonali, si ha l'indipendenza fra le proiezioni.

Infatti se le proiezioni del vettore sui sottospazi sono ortogonali la covarianza è:

$$\text{Cov}(P_{V_1} \underline{X}, P_{V_2} \underline{X}) \stackrel{\text{bilin.}}{=} P_{V_1} \text{Cov}(\underline{X}, \underline{X}) P_{V_2}^T = P_{V_1} \sigma^2 I_n P_{V_2}^T \stackrel{\perp}{=} \sigma^2 \underline{0} = \underline{0}$$

Poiché sappiamo che sotto ipotesi di normalità congiunta, l'incorrelazione (covarianza nulla) implica l'indipendenza, rimane così dimostrato il teorema.

$$\text{Se inoltre si verifica che } E[P_{V_i} \underline{X}] = \underline{0} \text{ allora: } \frac{\|P_{V_i} \mu_{\underline{X}}\|^2}{\sigma^2} \sim \chi^2_{(\dim V_i)} \quad (8.31)$$

Uno si potrebbe chiedere a cosa potrebbe mai servire questo teorema formulato in un modo così strano, in realtà questo teorema ci serve a dimostrare cose importantissime come il fatto che sotto ipotesi di normalità, la media e varianza campionaria sono tra loro indipendenti!

Supponiamo di voler proiettare un insieme di osservazioni di un vettore casuale in \mathbb{R}^n su due spazi vettoriali $\mathbb{1}_n$ che è lo spazio generato da $[1, 1, 1, \dots, 1]^T$ e $\mathbb{1}_n^\perp$ che è il corrispondente spazio ortogonale.

Dalla geometria sappiamo che la proiezione di un vettore \underline{X} su un vettore \underline{W} è definita come:

$$P_{\underline{W}}(\underline{X}) = \frac{\underline{X} \cdot \underline{W}}{\|\underline{W}\|^2} \underline{W}$$

In questo caso in particolare sappiamo che la proiezione di un vettore $\underline{X} = [X_1, X_2, \dots, X_n]$ su $\mathbb{1}_n$ è:

$$\frac{X_1 + X_2 + \dots + X_n}{(\sqrt{1+1+\dots+1})^2} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \overline{X_n} \\ \overline{X_n} \\ \vdots \\ \overline{X_n} \end{bmatrix} \quad \text{cioè la matrice di proiezione è} = \begin{bmatrix} 1/n & 1/n & \dots & 1/n \\ 1/n & 1/n & \dots & 1/n \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/n & \dots & 1/n \end{bmatrix}$$

NB

A questo risultato si può arrivare diversamente notando che il vettore delle \bar{X}_n è l'unico vettore che gode della proprietà di essere a distanza minima di \underline{X} , cioè, è il vettore che minimizza la quantità $\|\underline{X} - P_{\mathbb{1}_n^\perp}(\underline{X})\|$

Per calcolare la matrice di proiezione sullo spazio $\mathbb{1}_n^\perp$ possiamo sfruttare la 8.30:

$$\underline{X} = \underline{V} + \underline{W} \quad \underline{x} \in \mathbb{R}^n \quad \underline{V} \in \mathbb{1}_n^\perp \quad \underline{W} \in \mathbb{1}_n^\perp \quad \Rightarrow \quad \underline{X} - \underline{V} = \underline{W}$$

Poiché abbiamo visto che la proiezione su $\mathbb{1}_n^\perp$ è il vettore $[\bar{X}_n, \bar{X}_n, \dots, \bar{X}_n]$ si ottiene:

$$\underline{X} - \underline{V} = \underline{W} \quad \equiv \quad \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} - \begin{bmatrix} \bar{X}_n \\ \bar{X}_n \\ \vdots \\ \bar{X}_n \end{bmatrix} = \begin{bmatrix} X_1 - \bar{X}_n \\ X_2 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{bmatrix} \quad \text{quindi} \quad P_{\mathbb{1}_n^\perp} \underline{X} = \begin{bmatrix} X_1 - \bar{X}_n \\ X_2 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{bmatrix}$$

Cioè la matrice di proiezione è:

$$\begin{bmatrix} 1 - 1/n & 1/n & \dots & 1/n \\ 1/n & 1 - 1/n & \dots & 1/n \\ \vdots & \ddots & \ddots & \vdots \\ 1/n & 1/n & \dots & 1 - 1/n \end{bmatrix}$$

Inoltre si verifica l'ipotesi aggiuntiva del teorema, cioè:

$$E \begin{bmatrix} X_1 - \bar{X}_n \\ X_2 - \bar{X}_n \\ \vdots \\ X_n - \bar{X}_n \end{bmatrix} = \underline{0}$$

Per cui si ha che:

$$\frac{\|P_{\mathbb{1}_n^\perp} \bar{X}\|^2}{\sigma^2} \sim \chi^2_{(\dim \mathbb{1}_n^\perp)} \quad \equiv \quad \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Se ci facciamo caso l'espressione trovata è molto simile a quella della varianza campionaria, basta moltiplicare S_C^2 per $(n-1)$ per ottener la stessa espressione. Con questo si dimostra:

$$\frac{(n-1)S_C^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

e inoltre per il fatto che le proiezioni sono ortogonali segue che grazie al teorema di Cochran e ricordando che $\chi^2_{(n)} \equiv \Gamma(\frac{n-1}{2}, \frac{n-1}{2\sigma^2})$ sappiamo ora che:

$$\bar{X}_n \quad \text{indipendente da} \quad S_C^2 \quad \text{e} \quad S_C^2 \sim \Gamma\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2}\right) \quad (8.32)$$

9. Regressione

There are no two words more harmful than “good job”... I was there to push people beyond what’s expected of them. I believe that’s an absolute necessity. (Whiplash)

La regressione è una tecnica che prova a costruire un modello attraverso il cui si possano prevedere i possibili valori di una variabile, con la conoscenza di altre variabili. Le variabili sono in questo caso variabili indipendenti e variabili dipendenti, ma quest’indipendenza non ha nulla a che vedere con l’indipendenza statistica, ma riguarda il fatto se la variabile è impostata dallo sperimentatore o meno; per evitare confusioni, le variabili indipendenti sono dette anche **variabile esplicative o predittori** (*explanatory variables*) e le variabili dipendenti prendono il nome di **variabili di risposta** (*response variable*). Se si studia il legame fra una o una sola variabile esplicativa con una sola variabile di risposta la regressione si dice *semplice*; se si studia il legame fra una variabile di risposta con più di un predittore, la regressione si dice *multipla*.

Noi adotteremo utilizzeremo principalmente la regressione lineare semplice, cioè il modello che ci aiuterà a prevedere i valori della variabile di risposta sarà una linea retta. In realtà molte volte la linea “retta” del modello non è veramente una linea retta nella sua scala, ma applicando una opportuna trasformazione sugli assi è possibile ottenere una linea retta.

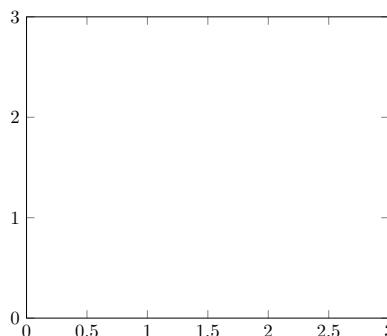


Figura 9.1: Prima della trasformazione

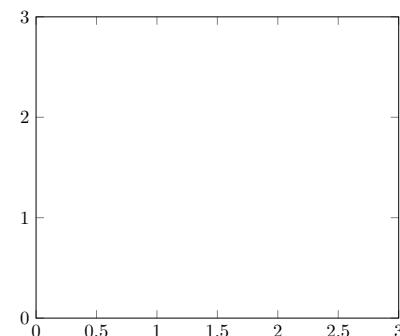


Figura 9.2: Dopo la trasformazione

Supponiamo di voler rispondere alla domanda: esiste un legame fra la variabile X e la variabile Y ? cioè ad esempio, esiste un legame tra le ore di allenamento settimanali e il numero di chilometri che una persona può correre senza fermarsi? Idealmente, se facciamo il grafico dei dati, è evidente che sussiste una relazione tra le due variabili, a questo punto quello che ci interessa sapere è: esiste una funzione, nel nostro caso lineare, una retta, che meglio approssima il legame fra le variabili, in modo che, fissato un condizionamento su una variabile si sia capaci di dare la migliore stima sul valore dell'altra variabile? Cioè, esiste una funzione che mi dica ad esempio quanti chilometri ci si può aspettare di correre se si hanno a disposizione soltanto 2 ore a settimana per allenarsi? In tal caso, qual è tale funzione? Per scoprirlo la regressione usa come strumento di analisi dei campioni casuali, che in questo caso non saranno singoli valori, ma le osservazioni saranno composte da due valori, cioè le ore dedicate allo sport da parte della persona i -esima e il corrispettivo numero chilometri corsi.

Per quanto detto, la regressione lineare rappresenta un metodo di stima del valore atteso condizionato, cioè ci dice cosa ci possiamo aspettare in Y (risposta) dati i diversi valori dei predittori (fissati/determinati dallo sperimentatore).

$$E[Y|X_1, X_2, \dots]$$

Quando assumiamo una *regressione lineare semplice* fra x e la variabile Y , assumiamo che la funzione cercata sia una retta di pendenza data dal coefficiente β_1 e intercetta pari a β_0 , cioè:

$$E[Y|x] = \beta_0 + \beta_1 x \quad (9.1)$$

La retta che meglio approssima è dunque:

$$Y = \beta_0 + \beta_1 x + \underbrace{E}_{\text{variabilità}} \quad (9.2)$$

I valori effettivamente osservati non cadranno nella retta ipotizzata ma si discosteranno di una lunghezza d_i che varia di caso in caso a causa della variabilità .

$$d_i = \underbrace{Y_i}_{\text{osservato}} - \underbrace{\hat{Y}_i}_{\text{previsto}} \quad (9.3)$$

Si noti che l'errore dovuto alla variabilità è considerato dalla variabile aleatoria E (per questo motivo Y è anch'essa una variabile aleatoria mentre x assume soltanto valori delle osservazioni).

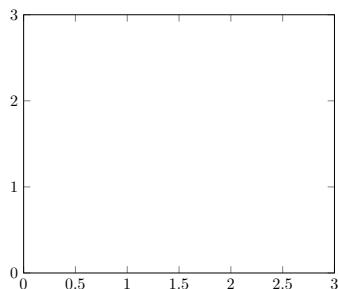


Figura 9.3: Retta di regressione

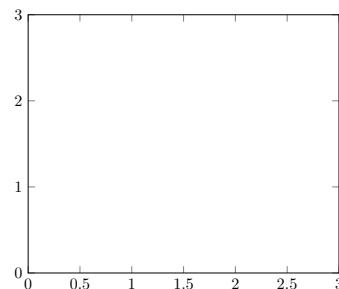


Figura 9.4: Scarti legati alla variabilità

9.1 Ipotesi del modello lineare

Quando assumiamo una regressione lineare, ci mettiamo sotto certe ipotesi, che non è sempre detto che si verifichino.

- 1** Si assume che gli errori dovuti alla variabilità sono indipendenti e identicamente distribuiti e di media nulla.
- 2** IMPORTANTE Si assume che l'errore è *omoschedastico* (di ugual varianza), cioè che la varianza del errore vale sempre σ^2 , qualunque sia il valore della variabile esplicativa (cosa non sempre vera).

Mettendo assieme con la prima ipotesi si ottiene che le E_i sono dunque distribuite come

$$E \sim \sigma Z_i \quad \text{con} \quad Z_i \sim N(0, 1)$$

Queste ipotesi implicano che $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

- 3** La linearità del modello è intesa nei parametri e non nei predittori.

A questo punto, da dove si prendono i parametri β_0 , β_1 e σ ? Così come si fa per tutti i parametri non noti, gli stimiamo. Chiamiamo dunque *retta di regressione stimata* la retta:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\sigma} Z \quad (9.4)$$

9.2 Metodo dei minimi quadrati

Il metodo dei minimi quadrati è un metodo che ci aiuta a stimare i parametri incogniti β_0 e β_1 . Questo metodo sceglie la miglior retta approssimante, assumendo che la miglior retta sia quella che ha la somma del quadrato degli errori più bassa di tutti.

Questo metodo si basa in realtà nello stesso principio con cui si definisce il valore atteso, il valore atteso è la quantità che rende minima la funzione $g(a) = E[(x - a)^2]$, in questo caso però la funzione che vogliamo rendere minima è $h(a) = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ cioè si vuole minimizzare

$$g'(\beta_0, \beta_1) = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

e il risultato non si tratterà più di un punto come nel caso del valore atteso, bensì di una retta. Il calcolo degli stimatori si riduce alla ricerca del minimo di funzione a due variabili o $n + 1$ variabili nel caso di n predittori.



Si noti inoltre che per costruzione, il metodo dei minimi quadrati posiziona la retta in modo che :

$$\sum_i d_i \equiv \sum_{i=1}^n (Y_i - \bar{Y}) = 0$$

Si è dimostrato che questo metodo è il miglior metodo possibile se si verifica effettivamente che gli errori si distribuiscono normalmente (Teorem di Gauss-Markov o *BLUE*).

Come risultato del metodo dei minimi quadrati si trova che:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad (9.5)$$

Dove:

$$\begin{aligned} S_{XX} &= \sum_{i=1}^n (x_i - \bar{x})^2 \\ S_{YY} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ S_{XY} &= \sum_{i=1}^n Y_i(x_i - \bar{x}_n) \quad \text{o} \quad \sum_{i=1}^n (Y_i - \bar{Y})x_i \quad \text{o} \quad \sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}_n) \end{aligned} \quad (9.6)$$

Ma com'è possibile che tre formule per S_{XY} diano lo stesso risultato?

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}_n) &= \sum_{i=1}^n (Y_i - \bar{Y})x_i - \sum_{i=1}^n (Y_i - \bar{Y})\bar{x}_n \\ &= \sum_{i=1}^n (Y_i - \bar{Y})x_i - \bar{x}_n \underbrace{\sum_{i=1}^n (Y_i - \bar{Y})}_{0 \text{ per def media}} \\ &= \sum_{i=1}^n (Y_i - \bar{Y})x_i \end{aligned} \quad (9.7)$$

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}_n) &= \sum_{i=1}^n Y_i(x_i - \bar{x}_n) - \sum_{i=1}^n \bar{Y}(x_i - \bar{x}_n) \\ &= \sum_{i=1}^n Y_i(x_i - \bar{x}_n) - \bar{Y} \underbrace{\sum_{i=1}^n (x_i - \bar{x}_n)}_{0 \text{ per def media}} \\ &= \sum_{i=1}^n Y_i(x_i - \bar{x}_n) \end{aligned} \quad (9.8)$$

Per stimare σ^2 viene usato:

$$\sigma^2 = \frac{\sum_i^n d_i^2}{n-2} = \frac{\sum_i^n (Y_i - \hat{Y}_i)^2}{n-2} = \frac{RSS}{n-2} \quad (9.9)$$



Si noti che si divide per $n - 2$ poiché al non conoscere β_0 e β_1 si perdono con la stima due gradi di libertà.

9.3 Verifica delle proprietà degli stimatori

Stimatore $\hat{\beta}_1$

$$\begin{aligned}
 E[\hat{\beta}_1] &= E\left[\frac{S_{XY}}{s_{XX}}\right] = \frac{E[S_{XY}]}{s_{XX}} = \frac{E[\sum Y_i(x_i - \bar{x})]}{s_{XX}} = \frac{\sum E[Y_i(x_i - \bar{x})]}{s_{XX}} \\
 &= \frac{\sum E[Y_i](x_i - \bar{x})}{s_{XX}} = \frac{\sum(\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{s_{XX}} \\
 &= \cancel{\frac{\sum \beta_0(x_i - \bar{x})}{s_{XX}}} + \frac{\sum \beta_1 x_i(x_i - \bar{x})}{s_{XX}} = \frac{\beta_1 \sum x_i(x_i - \bar{x})}{s_{XX}} \\
 &= \frac{\beta_1 \sum (x_i - \bar{x})(x_i - \bar{x})}{s_{XX}} = \frac{\beta_1 \sum (x_i - \bar{x})^2}{s_{XX}} = \frac{\beta_1 s_{XX}}{s_{XX}} = \beta_1
 \end{aligned} \tag{9.10}$$

NB Si noti che si è aggiunto il termine $\sum \bar{x}(x_i - \bar{x})$ in quanto è un termine nullo, portando fuori dalla sommatoria \bar{x} che non dipende da alcun indice si ha: $\bar{x} \sum (x_i - \bar{x})$, ove il termine $\sum (x_i - \bar{x})$ che è la somma degli scarti è nullo per costruzione.

$$\begin{aligned}
 Var(\hat{\beta}_1) &= Var\left(\frac{S_{XY}}{s_{XX}}\right) = \frac{1}{(s_{XX})^2} Var(S_{XY}) = \frac{Var(\sum Y_i(x_i - \bar{x}))}{(s_{XX})^2} \\
 &= \frac{\sum (x_i - \bar{x})^2 Var(Y_i)}{(s_{XX})^2} = \frac{\sum (x_i - \bar{x})^2 \sigma^2}{(s_{XX})^2} = \frac{\sigma^2 \sum (x_i - \bar{x})^2}{(s_{XX})^2} \\
 &= \frac{\sigma^2 s_{XX}}{(s_{XX})^2} = \frac{\sigma^2}{(s_{XX})} \text{ che è consistente se } s_{XX} \rightarrow \infty \text{ per } n \rightarrow \infty
 \end{aligned} \tag{9.11}$$

NB Non è detto che $\frac{\sigma^2}{(s_{XX})}$ tenda a zero per $n \rightarrow \infty$. Affinché sia consistente è necessario che il termine a denominatore diverga ciò succede se e solo se la serie $\sum (x_i - \bar{x})$ diverge. Anche se si ha un numero molto grande o addirittura infinito di osservazioni è possibile che la serie considerata converga, questo è dovuto al fatto che le osservazioni sono molto vicini tra di sé. È come se avessimo prese tutte le osservazioni in un solo punto del modello.

Se prendiamo infinite osservazioni nei punti:

$$x_i : \{-1, 1, -1/2, 1/2, -1/3, 1/3, -1/4, 1/4, \dots\}$$

la serie converge.

Basta prendere le osservazioni un poco più spaziate affinché la serie diverga:

$$x_i : \{-1, 1, -1/\sqrt{2}, 1/\sqrt{2}, -1/\sqrt{3}, 1/\sqrt{3}, -1/\sqrt{4}, 1/\sqrt{4}, \dots\}$$

Quindi lo stimatore $\hat{\beta}_1$ è *corretto e condizionatamente consistente*.

Stimatore $\hat{\beta}_0$

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E[\bar{Y}] - E[\hat{\beta}_1 \bar{x}] = \beta_0 + \beta_1 \bar{x} - E[\hat{\beta}_1] \bar{x} \\ &\stackrel{\text{per correttezza di } \hat{\beta}_1}{=} \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned} \quad (9.12)$$

$$\begin{aligned} \text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{x}) = \text{Var}(\bar{Y}) + \text{Var}(\hat{\beta}_1 \bar{x}) - 2 \text{Cov}(\bar{Y}, \hat{\beta}_1 \bar{x}) \\ &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) + (\bar{x})^2 \text{Var}(\hat{\beta}_1) - 2\bar{x} \text{Cov}(\bar{Y}, \hat{\beta}_1) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n Y_i\right) + (\bar{x})^2 \text{Var}(\hat{\beta}_1) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) + (\bar{x})^2 \frac{\sigma^2}{s_{XX}} = \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{s_{XX}} \right) \end{aligned} \quad (9.13)$$

Si noti che $2\bar{x} \text{Cov}(\bar{Y}, \hat{\beta}_1)$ è nullo in quanto:

$$\begin{aligned} \text{Cov}(\bar{Y}, \hat{\beta}_1) &= \text{Cov}\left(\bar{Y}, \frac{S_{XY}}{s_{XX}}\right) = \text{Cov}\left(\bar{Y}, \frac{\sum_{i=1}^n Y_i(x_i - \bar{x})}{s_{XX}}\right) = \frac{\text{Cov}(\bar{Y}, \sum_{i=1}^n Y_i(x_i - \bar{x}))}{s_{XX}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \text{Cov}(\bar{Y}, Y_i)}{s_{XX}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \text{Cov}(\sum_{j=1}^n Y_j, Y_i)}{s_{XX}} = \frac{\sum (x_i - \bar{x}) \sum_{i=1}^n \text{Cov}(Y_i, Y_i)}{s_{XX}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \sigma^2}{s_{XX}} = \frac{\sigma^2 \sum_{i=1}^n (x_i - \bar{x})}{s_{XX}} = \frac{\sigma^2 \cdot 0}{s_{XX}} = 0 \end{aligned}$$

NB

Anche in questo caso la consistenza è condizionale, in quanto non è detto che $\frac{\sigma^2}{(s_{XX})}$ tenda a zero per $n \rightarrow \infty$. Basta prendere le osservazioni un poco più spaziate affinché la serie $\sum(x_i - \bar{x})$ diverga. Quindi lo stimatore $\hat{\beta}_0$ è anch'esso *corretto e condizionatamente consistente*.

Stimatore $\hat{\sigma}^2$

$$\begin{aligned} E[\hat{\sigma}^2] &= E\left[\frac{\|Y - \hat{Y}\|^2}{n-2}\right] \stackrel{\text{per Cochran}}{=} E\left[\frac{\sigma^2 \chi_{n-2}^2}{n-2}\right] \\ &= \frac{\sigma^2 E[\chi_{n-2}^2]}{n-2} = \frac{\sigma^2(n-2)}{n-2} = \sigma^2 \end{aligned} \quad (9.14)$$

$$\begin{aligned} \text{Var}(\hat{\sigma}^2) &= \text{Var}\left(\frac{\|Y - \hat{Y}\|^2}{n-2}\right) \stackrel{\text{per Cochran}}{=} \text{Var}\left(\frac{\sigma^2 \chi_{n-2}^2}{n-2}\right) \\ &= \frac{\sigma^4}{(n-2)^2} \text{Var}(\chi_{n-2}^2) = \frac{\sigma^4}{(n-2)^2} 2(n-2) = \frac{2\sigma^4}{n-2} \end{aligned} \quad (9.15)$$

Lo stimatore $\hat{\sigma}^2$ è *corretto e consistente* e non dipende da come sono state prese le osservazioni, infatti tende a zero per $n \rightarrow \infty$.

9.4 Intervalli di confidenza sui parametri

Gli intervalli di confidenza per i parametri del modello lineare hanno la forma già nota, ma in questi casi la varianza è sempre stimata e il coefficiente k_γ è sempre calcolato con la funzione t di student a $n-p$ gradi di libertà.

$$\beta_i \in I_\gamma(\underline{X}) = \left[\hat{\beta}_i \pm \sqrt{\widehat{\text{Var}}(\hat{\beta}_i)} t_{(n-p)}^{-1} \left(\frac{1+\gamma}{2} \right) \right] \quad (9.16)$$

Sotto l'ipotesi di regressione lineare semplice gli intervalli di confidenza sono:

Per β_1 l'intervallo è:

$$\left[\hat{\beta}_1 \pm \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} t_{(n-2)}^{-1} \left(\frac{1+\gamma}{2} \right) \right] \rightarrow \left[\frac{S_{XY}}{s_{xx}} \pm \sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} t_{(n-2)}^{-1} \left(\frac{1+\gamma}{2} \right) \right]$$

Per β_0 l'intervallo è:

$$\left[\hat{\beta}_0 \pm \sqrt{\widehat{\text{Var}}(\hat{\beta}_0)} t_{(n-2)}^{-1} \left(\frac{1+\gamma}{2} \right) \right] \rightarrow \left[\bar{y} - \hat{\beta}_1 \bar{x} \pm \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)} t_{(n-2)}^{-1} \left(\frac{1+\gamma}{2} \right) \right]$$

Per σ^2 è l'intervallo :

$$\left[\frac{n-2}{F_{\chi_{(n-2)}^2}^{-1} \left(\frac{1+\gamma}{2} \right)} \hat{\sigma}^2 \leq \sigma^2 \leq \frac{n-2}{F_{\chi_{(n-2)}^2}^{-1} \left(\frac{1-\gamma}{2} \right)} \hat{\sigma}^2 \right]$$

9.5 Decomposizione della devianza

Per procedere lo studio della regressione è necessario introdurre tre termini di fondamentale importanza:

- **RSS (Residual Sum of Squares)**

Questo è un termine molto facile da capire, se ricordiamo che la retta di regressione stimata \hat{Y} non è altro che la retta effettiva Y "sporcata" dalla variabilità E , ovvero sostituendo $\hat{Y} = Y + E$ si ottiene che il termine è semplicemente la somma di ogni singolo errore elevato al quadrato.

$$RSS = \sum_i^n (Y_i - \hat{Y}_i)^2 = \sum_i^n (E_i)^2$$

- **SS_{REG} (Sum of Squares due to Regression)** chiamato anche **ESS (Explained Sum of Squares)** è la somma delle differenze tra i dei valori predetti e la media elevate al quadrato.

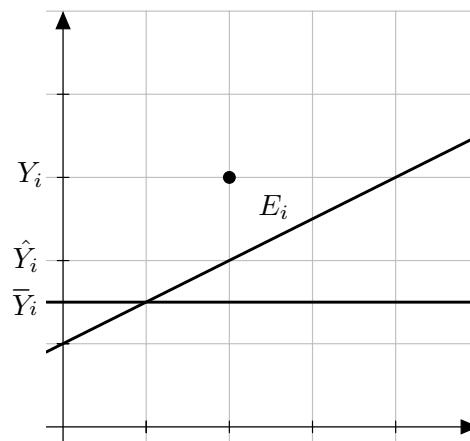
$$SS_{REG} = \sum_i^n (\hat{Y}_i - \bar{Y})^2$$

- **SYY** , talvolta chiamato **TSS (Total Sum of Squares)** è la somma delle differenze tra i dei valori reali e la media elevate al quadrato.

$$SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Svolgendo opportunamente i quadrati si può notare la seguente relazione:

$$SYY = SS_{REG} + RSS \quad (9.17)$$



9.6 Coefficiente di determinazione

La relazione 9.17 appena introdotta serve per introdurre un indice statistico che misura la bontà di adattamento della retta di regressione ai dati osservati; questo indice è chiamato *coefficiente di determinazione* ed è rappresentato dal simbolo R^2 .

Il coefficiente di determinazione assume valori compresi nell'intervallo $[0, 1]$, assumendo il valore 0 nei casi in cui non c'è alcuna relazione statistica tra le osservazioni e il modello o il valore 1 in cui esiste una dipendenza lineare (perfetta) tra il modello e le osservazioni empiriche. È importante notare che nei valori intermedi, ovvero quelli nell'intervallo $(0, 1)$, la differenza tra la linea di regressione e i punti che rappresentano l'osservazione diminuisce al crescere di R^2 , quindi all'aumentare di R^2 la bontà del modello di regressione aumenta.

Il termine RSS può essere visto come un termine che quantifica l'incertezza poiché è nullo solo quando tutti i valori previsti e i valori osservati coincidono sempre. La bontà della regressione può essere visto come una percentuale che può essere ottenuta come $100\% - Incertezza\%$ ovvero:

$$R^2 = 1 - \frac{RSS}{SYY} \quad (9.18)$$

Utilizzando la relazione precedente possiamo definire il coefficiente di determinazione come:

$$R^2 = 1 - \frac{RSS}{SS_{REG} + RSS} = \frac{SS_{REG}}{RSS} \quad (9.19)$$

Uno si potrebbe chiedere: come mai il simbolo del coefficiente di correlazione è R^2 e non semplicemente R ? questo perché si può dimostrare che R^2 corrisponde al quadrato del *coefficiente di correlazione* ρ , ovvero :

$$R^2 = \rho_{XY}^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} \quad (9.20)$$

9.7 Rappresentazione vettoriale della regressione

- *Matrice \mathbf{X}*

E una matrice formata dall'accostamento del vettore $\mathbb{1}_n$ e i predittori x

$$\mathbf{X} = [\mathbb{1}_n \ x_i]$$

- *Vettore dei coefficienti di regressione*

$$\underline{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- *Vettore della variabile di risposta stimata*

$$\hat{\mathbf{Y}} = \underline{\beta} \mathbf{X}$$

- *Vettori delle medie*

$$\bar{\mathbf{X}}_n = \mathbf{P}_{\mathbb{1}_n} \mathbf{X}$$

$$\bar{\mathbf{Y}}_n = \mathbf{P}_{\mathbb{1}_n} \mathbf{Y}$$

- *SYY*

$$SYY = \| \mathbf{Y} - \mathbf{P}_{\mathbb{1}_n} \mathbf{Y} \|^2 = \| \mathbf{Y} - \bar{\mathbf{Y}}_n \|^2$$

- *RSS*

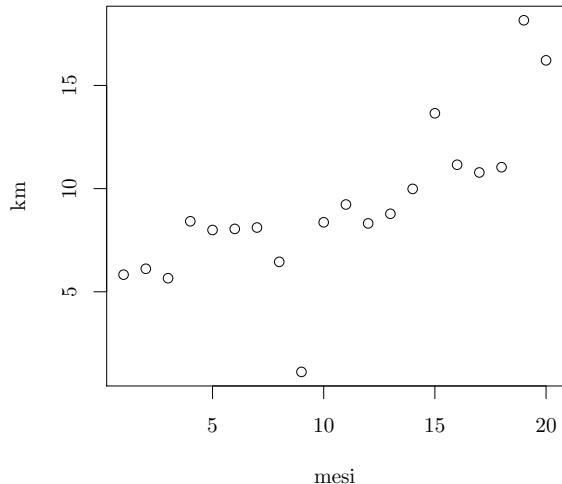
$$RSS = \| \mathbf{Y} - \underline{\beta} \mathbf{X} \|^2 = \| \mathbf{Y} - \hat{\mathbf{Y}} \|^2$$

- *SS_{REG}*

$$SS_{REG} = \| \mathbf{P}_{\mathbb{1}_n^\perp} \mathbf{Y} \|^2 = \| \hat{\mathbf{Y}} - \mathbf{P}_{\mathbb{1}_n} \mathbf{Y} \|^2 = \| \hat{\mathbf{Y}} - \bar{\mathbf{Y}}_n \|^2$$

ESEMPIO

Un aspirante maratonista vuole capire quanto tempo manca affinché sia capace di correre i 42,195 km che prevede la competizione. A questo scopo, registra in un diario la media dei chilometri percorsi ogni mese per 20 mesi ricavando così i dati: (1, 5.83)(2, 6.11)(3, 5.65) ... riportati nel grafico a destra.



Come primo tentativo proviamo a impostare il modello di regressione lineare semplice.

$$X = \begin{bmatrix} \mathbb{1}_{20} & x_i \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \end{bmatrix} \quad (X^T X) = \begin{bmatrix} 20 & 210 \\ 210 & 2870 \end{bmatrix}$$

NB Si noti che $(X^T X)$ è sempre una matrice simmetrica che la numerosità del campione in esame nel primo elemento della diagonale principale.

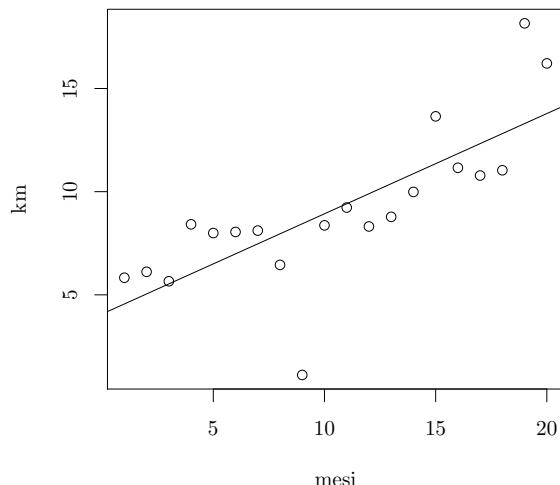
Procediamo a invertire la matrice con il metodo dei cofattori (chi non dovesse ricordarlo può rinfrescare la memoria su [GAL MOSCA sezione 6.3.2])

$$(X^T X)^{-1} = \frac{1}{\det(X^T X)} \begin{bmatrix} 2870 & -210 \\ -210 & 20 \end{bmatrix} = \begin{bmatrix} 0.215789474 & -0.015789474 \\ -0.015789474 & 0.001503759 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} \mathbb{1}_{20} \\ x_i \end{bmatrix} Y = \begin{bmatrix} 183.429952 \\ 2248.972909 \end{bmatrix}$$

A questo punto possiamo calcolare la matrice dei predittori stimati, $\hat{\beta}$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 0.215789474 & -0.015789474 \\ -0.015789474 & 0.001503759 \end{bmatrix} \begin{bmatrix} 183.429952 \\ 2248.972909 \end{bmatrix} = \begin{bmatrix} 4.07215424 \\ 0.485651749 \end{bmatrix}$$



La nostra retta di regressione è dunque

$$Y = 4.07215 + 0.48565X$$

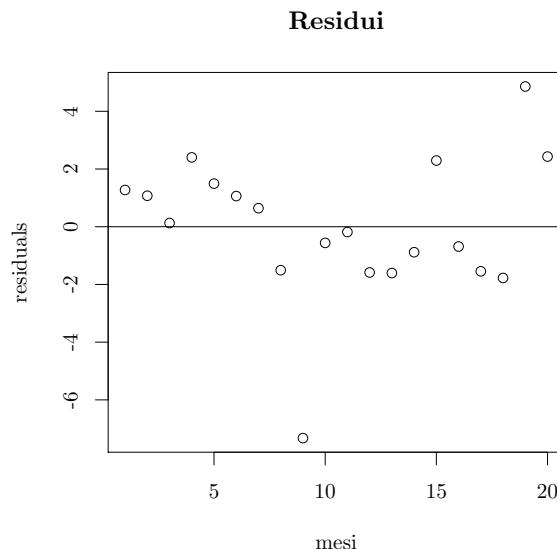
Usiamo i valori previsti \hat{Y}_i dalla nostra retta di regressione e i valori effettivamente osservati per calcolare l'RSS e ulteriormente $\hat{\sigma}^2$

$$RSS = \sum_i^n (Y_i - \hat{Y}_i)^2 = (5.83 - 4.55)^2 + (6.11 - 5.04)^2 + (5.65 - 5.52)^2 + \dots = 115.2950106$$

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{115.2950106}{18} = 6.4052$$

Prima di rispondere alla domanda del problema, dobbiamo verificare che il nostro modello lineare abbia senso e soddisfi le diverse ipotesi assunte.

Se facciamo un grafico dei residui possiamo apprezzare che i residui effettivamente si comportano come una normale di media zero e varianza $\hat{\sigma}^2$. Si può verificare con una occhiata veloce al grafico dei residui che la maggior parte dei residui si trova entro 1 deviazione standard e tutti i residui si trovano entro 3 deviazioni standard .



Poiché le ipotesi sono verificate, rispondiamo alla domanda iniziale; l'aspirante maratonista può aspettarsi di correre i 42.195 km al mese $x = 78.4986$, cioè fra circa 58 mesi e mezzo/ circa 5 anni!

Calcoliamo ora il coefficiente di determinazione R^2

$$R^2 = 1 - \frac{RSS}{SYY} = 1 - \frac{115.2950106}{272.1403} = 0.5763398$$

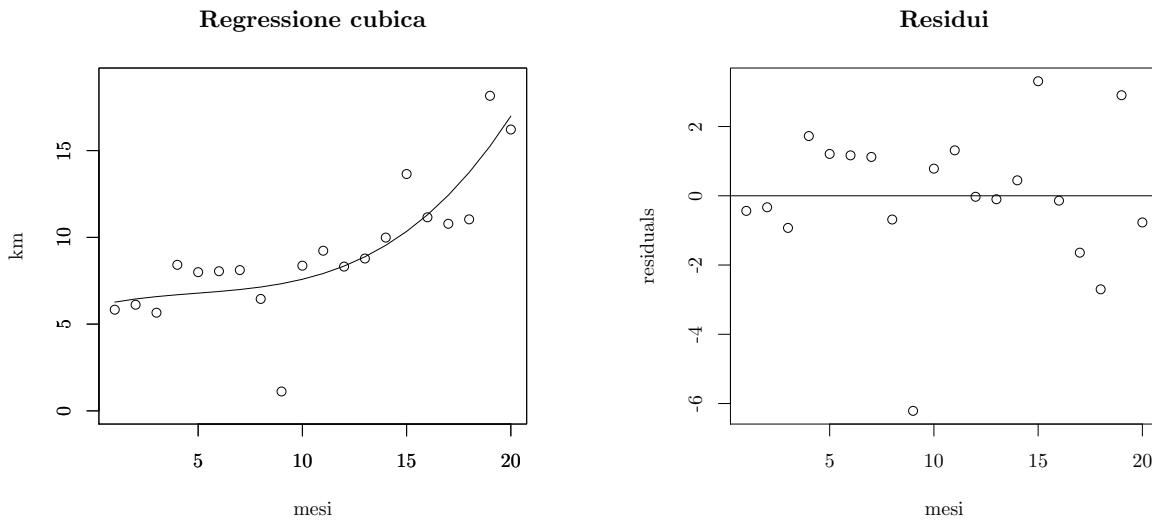
Ci rendiamo conto che il coefficiente di determinazione è basso, si può fare di meglio? Proviamo ad adottare un modello di regressione anche in questo caso lineare nei parametri ma non lineare nelle variabili di predizione, in questo caso scegliamo il modello cubico dove ci aspettiamo un modello di forma $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$. Cominciamo per scrivere la matrice dei predittori

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad (X^T X) = \begin{bmatrix} 20 & 210 & 2870 & 44100 \\ 210 & 2870 & 44100 & 722666 \\ 2870 & 44100 & 722666 & 12333300 \\ 44100 & 722666 & 12333300 & 216455810 \end{bmatrix}$$

Facendo gli stessi passaggi di prima troviamo:

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 6.025760655 \\ 0.276670545 \\ -0.037750401 \\ 0.002565931 \end{bmatrix} \quad \hat{\sigma}^2 = \frac{RSS}{n-4} = \frac{79.8130}{16} = 4.9883$$

In questo caso il coefficiente di determinazione R^2 aumenta a 0.7067



Comunque il nostro valore di R^2 non è altissimo, se il modello di regressione spiegassi molto bene il modello, il coefficiente di determinazione dovrebbe essere vicino a 1. È dunque opportuno fare un test d'ipotesi riguardo al modello di regressione. DA COMPLETARE....

9.8 Trasformazioni del modello lineare

9.8.1 Lineare-Logaritmica

Forma Ipotizzata	Trasformazione	Forma linearizzata
------------------	----------------	--------------------

$$T = \alpha + \beta \ln t \quad x = \ln t \quad Y = \alpha + \beta x$$

9.8.2 Potenza

Forma Ipotizzata	Trasformazione	Forma linearizzata
------------------	----------------	--------------------

$$T = \alpha t^\beta \quad Y = \ln T \quad x = \ln t \quad Y = \ln \alpha + \beta x$$

9.8.3 Esponenziale

Forma Ipotizzata	Trasformazione	Forma linearizzata
------------------	----------------	--------------------

$$T = \alpha e^{\beta x} \quad Y = \ln T \quad Y = \ln \alpha + \beta x$$

9.8.4 Iperbolica

Forma Ipotizzata	Trasformazione	Forma linearizzata
------------------	----------------	--------------------

$$T = \frac{t}{\alpha t + \beta} \quad Y = \frac{1}{T} \quad x = \frac{1}{t} \quad Y = \alpha + \beta x$$

9.9 Test globali sui predittori

9.9.1 Test F

Una volta determinato un possibile modello di regressione, è possibile realizzare diversi test d'ipotesi che verifichino la fattibilità di questo modello. Test molto comuni sono quelli che verificano se tutti i coefficienti β_i sono nulli. Questi tipi di test hanno come scopo di verificare se è inutile il modello di regressione dato che in realtà da quello si riesce ad osservare potrebbe essere spiegato da un modello di variabile aleatoria e la variabilità presente non è altro che rumore.

$$\begin{aligned} H_0 : \beta_1 &= \beta_2 = \dots = \beta_p = 0 \\ H_1 : \exists 1 \leq i \leq p : \beta_i &\neq 0 \end{aligned} \quad (9.21)$$

Ma qual è la regola di decisione in questo caso? In caso che valga H_0 , staremo dicendo che $Y \sim N(\beta_0 I_n, \sigma^2 I_n)$ e che non c'è regressione. Quindi al non esserci regressione la matrice $\underline{\beta}$ sarà composta dal solo coefficiente β_0 e per il resto, zeri. Per cui:

$$E(\hat{Y}) = \underline{X}\underline{\beta} = \begin{bmatrix} \mathbb{1}_n | & x_1 | & \dots | & x_p \end{bmatrix} \begin{bmatrix} \beta_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbb{1}_n \beta_0$$

Ma al non esserci regressione

$$E(P_{\mathbb{1}_n^\perp} Y) \equiv E(\bar{Y}_n \mathbb{1}_n) = \frac{1}{n} \sum E(Y_i \mathbb{1}_n) = \frac{1}{n} \sum (\beta_0 + \beta_1 x_i + \dots + \beta_p x_p) \mathbb{1}_n \stackrel{H_0}{=} \begin{bmatrix} \beta_0 + 0 + \dots \\ \beta_0 + 0 + \dots \\ \vdots \\ \beta_0 + 0 + \dots \end{bmatrix} = \mathbb{1}_n \beta_0$$

Quindi se vale H_0 , per quanto appena mostrato, si verifica che:

$$E(P_{(\mathbb{1}_n^\perp)_X} Y) = E(\hat{Y} - P_{\mathbb{1}_n} Y) = 0$$

Dato che avevamo scomposto lo spazio $\mathbb{R}_n = \mathbb{1}_n \oplus (\mathbb{1}_n^\perp)_X \oplus X^\perp$ e si è appena mostrato che $E(P_{(\mathbb{1}_n^\perp)_X} Y) = 0$, siamo nell'ipotesi aggiuntiva di Cochran e possiamo dire che:

$$\frac{\overbrace{\|\hat{Y} - P_{\mathbb{1}_n} Y\|^2}^{SS_{REG}}}{\sigma^2} \sim \chi^2_{(p)}$$

La statistica con cui prenderemo una decisione è il rapporto fra le chi-quadro appena trovate divise per il rispettivo grado di libertà

$$F = \frac{SS_{REG}/(p' - 1)}{RSS/(n - p')} \sim F_{(p, n-p')} \quad (9.22)$$

Questa statistica (variabile aleatoria che non contiene quantità non note) assume valori quanto più grandi quanto migliore è la relazione di regressione, e assume valori piccoli quando il fenomeno può essere spiegato da una variabile aleatoria. Questo comportamento è dovuto al fatto che a numeratore si trova la quantità SS_{REG} che come abbiamo visto dal grafico rappresenta l'incremento di spiegazione del modello di variabile aleatoria rispetto al modello di regressione.

Il test lo si formula in modo analogo ai test visti finora, rifiuteremo H_0 , cioè diremo che non sussiste un legame di variabile aleatoria ma piuttosto uno di regressione se il valore della statistica è maggiore di una certa soglia determinata dalla tolleranza all'errore di primo tipo α .

$$P_{H_0}(\text{rifiuto } H_0) = \alpha \quad P_{H_0}(F \geq c) = \alpha$$

$$P_{H_0}(F \leq c) = 1 - \alpha \quad c = F_{F_{(p'-1,n-p')}}^{-1}(1 - \alpha)$$

Di conseguenza la regola di decisione risulta:

$$\text{Si accetta } H_0 \text{ se} \quad F \leq F_{F_{(p'-1,n-p')}}^{-1}(1 - \alpha) \quad (9.23)$$

9.9.2 Test del modello ridotto

Oltre al test che verifica che è effettivamente conveniente scegliere un modello di regressione anziché di variabile aleatoria, è possibile formulare altri test che ci dicono se è conveniente scegliere il modello così com'è stato formulato o piuttosto è più opportuno sceglierne una versione ridotta visto che non tutte le variabili di predizione risultano in realtà significative.

$$\begin{aligned} H_0 & : (\text{Ridotto}) \quad Y_i = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r + \sigma Z_i \\ H_1 & : (\text{Esteso}) \quad Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \sigma Z_i \end{aligned} \quad (9.24)$$

In modo analogo a quanto fatto prima, possiamo determinare una statistica con cui prenderemo una decisione:

$$F = \frac{\|\mathbf{P}_{(X_{rid}^\perp)} \underline{Y}\|^2 / (p' - r)}{\|\mathbf{P}_{X^\perp} \underline{Y}\|^2 / (n - p')} \underset{H_0}{\sim} F_{(p'-r, n-p')}$$

$$\text{Si accetta } H_0 \text{ se} \quad F \leq F_{F_{(p'-r,n-p')}}^{-1}(1 - \alpha) \quad (9.26)$$

NB

Si faccia molta attenzione alle omonimie, la prima F è una statistica, $F_{(p'-r,n-p')}$ è una variabile aleatoria con distribuzione di Fisher e F^{-1} è il contatore inverso della variabile aleatoria in pedice.

9.10 Test su singoli parametri

Molte volte si vuole determinare se risulta inutile o meno un certo predittore considerato. Per determinarlo si formula un test bilatero con le ipotesi:

$$\begin{aligned} H_0 : \beta_i &= 0 \\ H_1 : \beta_i &\neq 0 \end{aligned} \tag{9.27}$$

Rifiuteremo l'ipotesi H_0 se il valore previsto del predittore non è prossimo a zero, cioè $|\hat{\beta}_i| \geq c$. Per aiutarci a formulare il test facciamo una studentizzazione, cioè l'analogo alla standardizzazione quando la varianza non è nota:

$$\left| \frac{\hat{\beta}_i - \beta_i}{\sqrt{Var(\hat{\beta}_i)}} \right| \geq c$$

Avendo fatto questo possiamo sfruttare il fatto che :

$$\left| \frac{\hat{\beta}_i - \beta_i}{\sqrt{Var(\hat{\beta}_i)}} \right| \sim t_{(n-p')}$$

A questo punto facendo i soliti passaggi si trova la regola di decisione

$$\text{Si accetta } H_0 \text{ se } \left| \beta_i - \hat{\beta}_i \right| \leq \sqrt{Var(\hat{\beta}_i)} t_{(n-p')}^{-1} \left(1 - \frac{\alpha}{2} \right) \tag{9.28}$$

NB

Si noti che $Var(\hat{\beta}_i)$ può essere espressa come $\sigma^2 v_{ii}$ dove v_{ii} è l' i -esimo elemento diagonale della matrice $(X^T X)^{-1}$

9.11 Intervalli di previsione

Gli intervalli di previsione si assomigliano sia in nome che in forma agli intervalli di confidenza, ma concettualmente sono due cose ben diverse. Gli intervalli di confidenza ci danno un intervallo riguardo ad un parametro, cioè ci dicono tra che due numeri può cadere il vero parametro del modello con una probabilità γ . Gli intervalli di previsione ci danno un intervallo, ma non più riguardo ad un parametro, ma riguardo alle future osservazioni, cioè, dato un certo numero di osservazioni, in che intervallo ci aspettiamo che cadano le osservazioni future? (si ricordi che l'sperimentatore fissa le variabili di predizione x_i quindi l'intervallo calcolato è sui valori della variabile di risposta y).

Intervallo di confidenza

$$I_\gamma(\underline{X}) = \left[\tilde{x}_0^T \hat{\beta} \pm \sqrt{\hat{\sigma}^2 (\tilde{x}^T (\underline{x}^T \underline{x})^{-1} \tilde{x}_0)} t_{(n-p')}^{-1} \left(\frac{1+\gamma}{2} \right) \right] \quad (9.29)$$

Intervallo di previsione

$$I_\gamma(\underline{X}) = \left[\tilde{x}_0^T \hat{\beta} \pm \sqrt{\hat{\sigma}^2 (1 + \tilde{x}^T (\underline{x}^T \underline{x})^{-1} \tilde{x}_0)} t_{(n-p')}^{-1} \left(\frac{1+\gamma}{2} \right) \right] \quad (9.30)$$

La differenza sostanziale è la seguente:

$$\text{CONFIDENZA } \frac{\tilde{x}_0^T \hat{\beta} - \tilde{x}_0^T \beta}{\sqrt{Var(\tilde{x}_0^T \hat{\beta} - \tilde{x}_0^T \beta)}} \quad \text{PREVISIONE } \frac{\tilde{x}_0^T \hat{\beta} - Y_0}{\sqrt{Var(\tilde{x}_0^T \hat{\beta} - Y_0)}} \quad (9.31)$$

NB

Si noti che l'intervallo di previsione è più grande di quello di confidenza, questo è dovuto al fatto che l'intervallo di previsione fa una stima, su un modello, che a sua volta è già stimato; in un certo senso negli intervalli di previsioni c'è una doppia "stima".

9.12 Distribuzione di Fisher

La distribuzione di Fisher scritta come $F \sim F(n, m)$ è la densità continua di una variabile aleatoria F con n e m gradi di libertà :

$$F = \frac{Q_1/n}{Q_2/m} \quad (9.32)$$

dove $Q_1 \sim \chi^2_{(n)}$ e $Q_2 \sim \chi^2_{(m)}$. La variabile F è dunque la proporzione fra due χ^2 . Ma a cosa ci potrebbe mai servire una variabile aleatoria fatta in questo modo? In realtà la distribuzione di Fisher sorge molto di frequente quando si lavora con rapporti di stimatori per le varianze e la regressione.

Le cose più importanti da considerare sono:

- La densità di $F_1 = \frac{Q_1/n}{Q_2/m}$ è diversa da quella di $F = \frac{Q_2/m}{Q_1/n}$
- La densità ha asimmetria negativa (densità più a sinistra).
- $E[F] = \frac{m}{m-2}$ per $m > 2$
- Se $n = m$ la mediana è pari a 1

$\frac{n}{m}$	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	0.500	1.50	1.71	1.82	1.89	1.94	1.98	2.00	2.04	2.07	2.09	2.12	2.15	2.17
	0.600	2.63	2.93	3.09	3.20	3.27	3.32	3.36	3.41	3.45	3.48	3.52	3.56	3.59
	0.667	4.00	4.42	4.64	4.78	4.88	4.95	5.00	5.08	5.13	5.18	5.24	5.29	5.33
	0.750	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.32	9.41	9.50	9.58	9.67	9.74
	0.800	12.0	13.1	13.6	14.0	14.3	14.4	14.6	14.8	14.9	15.0	15.2	15.3	15.6
2	0.500	1.00	1.13	1.21	1.25	1.28	1.30	1.32	1.35	1.36	1.38	1.39	1.41	1.42
	0.600	1.50	1.64	1.72	1.76	1.80	1.82	1.84	1.86	1.88	1.89	1.91	1.92	1.94
	0.667	2.00	2.15	2.22	2.27	2.30	2.33	2.34	2.37	2.38	2.40	2.42	2.43	2.45
	0.750	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.38	3.39	3.41	3.43	3.44	3.46
	0.800	4.00	4.16	4.24	4.28	4.32	4.34	4.36	4.38	4.40	4.42	4.43	4.45	4.47
3	0.500	0.88	1.00	1.06	1.10	1.13	1.15	1.16	1.18	1.20	1.21	1.23	1.24	1.25
	0.600	1.26	1.37	1.43	1.47	1.49	1.51	1.52	1.54	1.55	1.56	1.57	1.58	1.59
	0.667	1.62	1.72	1.77	1.80	1.82	1.83	1.84	1.86	1.87	1.88	1.89	1.90	1.91
	0.750	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.45	2.46	2.46	2.47	2.47
	0.800	2.89	2.94	2.96	2.97	2.97	2.97	2.98	2.98	2.98	2.98	2.98	2.98	2.98
4	0.500	0.83	0.94	1.00	1.04	1.06	1.08	1.09	1.11	1.13	1.14	1.15	1.16	1.18
	0.600	1.16	1.26	1.31	1.34	1.36	1.37	1.38	1.40	1.41	1.42	1.43	1.43	1.45
	0.667	1.46	1.55	1.58	1.61	1.62	1.63	1.64	1.65	1.65	1.66	1.67	1.67	1.68
	0.750	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	0.800	2.47	2.48	2.48	2.48	2.47	2.47	2.47	2.46	2.46	2.45	2.44	2.44	2.43
5	0.500	0.80	0.91	0.96	1.00	1.02	1.04	1.05	1.07	1.09	1.10	1.11	1.12	1.13
	0.600	1.11	1.20	1.24	1.27	1.29	1.30	1.31	1.32	1.33	1.34	1.34	1.35	1.36
	0.667	1.38	1.45	1.48	1.50	1.51	1.52	1.53	1.53	1.54	1.54	1.54	1.55	1.55
	0.750	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.87
	0.800	2.26	2.25	2.24	2.23	2.22	2.21	2.20	2.19	2.18	2.18	2.17	2.16	2.15
6	0.500	0.78	0.89	0.94	0.98	1.00	1.02	1.03	1.05	1.06	1.07	1.08	1.10	1.11
	0.600	1.07	1.16	1.20	1.22	1.24	1.25	1.26	1.27	1.28	1.29	1.29	1.30	1.31
	0.667	1.33	1.39	1.42	1.44	1.44	1.45	1.45	1.46	1.46	1.47	1.47	1.47	1.47
	0.750	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.76	1.76	1.75	1.74
	0.800	2.13	2.11	2.09	2.08	2.06	2.05	2.04	2.03	2.02	2.01	2.00	1.98	1.97
7	0.500	0.77	0.87	0.93	0.96	0.98	1.00	1.01	1.03	1.04	1.05	1.07	1.08	1.09
	0.600	1.05	1.13	1.17	1.19	1.21	1.22	1.23	1.24	1.24	1.25	1.26	1.26	1.27
	0.667	1.29	1.35	1.38	1.39	1.40	1.40	1.41	1.41	1.41	1.41	1.41	1.42	1.42
	0.750	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.68	1.68	1.67	1.66	1.65
	0.800	2.04	2.02	1.99	1.97	1.96	1.94	1.93	1.92	1.91	1.89	1.88	1.86	1.85
8	0.500	0.76	0.86	0.91	0.95	0.97	0.99	1.00	1.02	1.03	1.04	1.05	1.07	1.09
	0.600	1.03	1.11	1.15	1.17	1.19	1.20	1.20	1.21	1.22	1.22	1.23	1.24	1.25
	0.667	1.26	1.32	1.35	1.36	1.36	1.37	1.37	1.37	1.37	1.38	1.38	1.38	1.37
	0.750	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.62	1.62	1.61	1.60	1.59
	0.800	1.98	1.95	1.92	1.90	1.88	1.87	1.86	1.84	1.83	1.81	1.80	1.78	1.76

$\frac{n}{m}$	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	0.900	49.5	53.6	55.8	57.2	58.2	59.1	59.7	60.5	61.0	61.5	62.0	62.6	63.0
	0.950	199.	216.	225.	230.	234.	237.	239.	242.	244.	246.	248.	250.	252.
	0.975	800.	864.	900.	922.	937.	948.	957.	969.	977.	985.	993.		254.
	0.990													
	0.999													
2	0.900	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39	9.41	9.43	9.44	9.46	9.47
	0.950	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5
	0.975	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5
	0.990	99.0	99.2	99.2	99.3	99.3	99.4	100.	100.	100.	100.	100.	100.	99.5
	0.999	999.	999.											
3	0.900	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.23	5.22	5.20	5.18	5.17	5.15
	0.950	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58
	0.975	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.4	14.3	14.3	14.2	14.1	14.0
	0.990	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.2	27.1	26.9	26.7	26.5	26.4
	0.999	149.	141.	137.	135.	133.	132.	131.	129.	128.	127.	126.	125.	123.
4	0.900	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.92	3.90	3.87	3.84	3.82	3.79
	0.950	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70
	0.975	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.84	8.75	8.66	8.56	8.46	8.38
	0.990	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.5	14.4	14.2	14.0	13.8	13.5
	0.999	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.0	47.4	46.8	46.1	45.4	44.9
5	0.900	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.30	3.27	3.24	3.21	3.17	3.15
	0.950	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44
	0.975	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.62	6.52	6.43	6.33	6.23	6.14
	0.990	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.1	9.89	9.72	9.55	9.38	9.24
	0.999	37.1	33.2	31.1	29.8	28.8	28.2	27.6	26.9	26.4	25.9	25.4	24.9	24.4
6	0.900	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.94	2.90	2.87	2.84	2.80	2.77
	0.950	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.94	3.87	3.81	3.75
	0.975	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.46	5.37	5.27	5.17	5.07	4.98
	0.990	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.87	7.72	7.56	7.40	7.23	7.09
	0.999	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.4	18.0	17.6	17.1	16.7	15.7
7	0.900	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.70	2.67	2.63	2.59	2.56	2.52
	0.950	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.51	3.44	3.38	3.32
	0.975	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.76	4.67	4.57	4.47	4.36	4.28
	0.990	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.62	6.47	6.31	6.16	5.99	5.86
	0.999	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.1	13.7	13.3	12.9	12.5	11.7
8	0.900	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.54	2.50	2.46	2.42	2.38	2.35
	0.950	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.22	3.15	3.08	3.02
	0.975	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.29	4.20	4.10	4.00	3.89	3.81
	0.990	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.81	5.67	5.52	5.36	5.20	5.07
	0.999	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.5	11.2	10.8	10.5	10.1	9.80

9.13 Analisi della varianza

L'analisi della varianza chiamato anche ANOVA (*Analysis of Variance*) è un insieme di tecniche statistiche che si usano per analizzare e fare dei test d'ipotesi sulle differenze tra le medie di tra gruppi e.

I test d'ipotesi valutano nella maggior parte l'ipotesi H_0 : tutti i gruppi considerati hanno la stessa distribuzione, e le differenze osservate non sono altro che rumore dovuto alla variabilità intrinseca del fenomeno. Le tecniche ANOVA vengono suddivise a seconda se il modello prevede una o più cause o addirittura una interazione fra le cause. Noi ci concentreremo nella sua formulazione più semplice, cioè quella che prevede una sola causa.

9.13.1 Analisi della varianza a un fattore

L'analisi della varianza a un fattore oppure *one-way ANOVA* è il caso più semplice di analisi della varianza, che prevede una solo causa. Questo metodo di analisi è usato per fare dei test con l'ipotesi:

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 = \mu_3 = \dots \\ H_1 : \exists r \neq s \quad \mu_r &\neq \mu_s \end{aligned} \tag{9.33}$$

Questo test quindi pretende analizzare se è vero che 3 o più gruppi di campioni sono stati campionati da una stessa popolazione (ipotizzare la stessa media grazie all'ipotesi di omoschedasticità (stessa varianza) in realtà vuol dire stessa popolazione).

NB

Si noti che questo test considera 3 o più gruppi, questo è dovuto al fatto che nel caso di 2 gruppi esistono dei test più semplici che forniscono gli stessi risultati (già trattato nella sezione 7.9.2).

Si noti che di conseguenza si possono ottenere gli stessi risultati per più gruppi facendo questo test più volte. Per $\mu_1 = \mu_2 = \mu_3$ basterebbe realizzare $\mu_1 = \mu_2$, $\mu_2 = \mu_3$ e $\mu_1 = \mu_3$.

La cosa favolosa di questo problema è che può essere ricondotto al modello lineare. Consideriamo di avere G gruppi di osservazioni, ognuno di numerosità n_i .

$$F_{OSS} = \frac{\|\hat{Y} - P_{\mathbb{1}_n} \underline{Y}\|^2 / (\dim G - 1)}{\|\underline{Y} - \hat{Y}\|^2 / (n - \dim G)} \sim F_{(\dim G - 1, n - \dim G)} \tag{9.34}$$

Questo può essere riscritto come il rapporto fra la somma delle varianze fra i diversi gruppi e la somma delle varianze "interne" ai gruppi.

$$F_{OSS} = \frac{\sum_{i=1}^G n_i (\mu_i - \bar{Y}_n)^2 / (\dim G - 1)}{\sum_{i=1}^G \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{n_i})^2 / (n - \dim G)} \tag{9.35}$$

Con base a questa statistica la regola di decisione diventa:

Si accetta H_0 se $F_{OSS} \leq F_{F_{(\dim G - 1, n - \dim G)}}^{-1} (1 - \alpha)$

$$(9.36)$$