Metodi Statistici per la Biologia

Paolo Dai Pra
Università di Padova

A.A. 2006/07

1 Introduzione

Qualche informazione di carattere organizzativo

Docenti: Paolo Dai Pra e Francesco Caravenna

Orari di ricevimento (Dai Pra): Lunedì ore 14:30 e per appuntamento

email: daipra@math.unipd.it

Pagina web: www.math.unipd.it/∼daipra/didattica/bio07 o vedi e-learning

Testo: Sheldon M.Ross, PROBABILITA E STATISTICA per l'ingegneria e le scienze, Apogeo.

Modalità d'esame: Prova scritta (l'eventuale prova orale è facoltativa). E obbligatoria l'iscrizione alla lista elettronica.

Un po' di chiacchiere

La statistica nasce nel '700 in Inghilterra come strumento per organizzare e presentare dati relativi ai membri di una popolazione (reddito, sesso, età .....)

Per molto tempo la statistica venne usata con sola valenza *descrittiva*.

Solo verso la fine dell' '800 ci si rese conto dell'interesse (politico, organizzativo, finanziario...) di avere dati relativi a determinate variabili (opinioni, abitudini di vita, ...) per cui la rilevazione individuale su tutti i membri della popolazione in esame sia irrealizzabile.

Sorse in modo naturale la seguente domanda: se i dati a disposizione sono relativi a solo una parte della popolazione, è ragionevole trarne conclusioni che coinvolgano l'intera popolazione?

Supponiamo, ad esempio, di essere interessati a quanti italiani adulti (diciamo 40 milioni di individui) utilizzano abitualmente il web.

Da un'intervista su 1000 individui risulta che 450 usano il web abitualmente. Cosa possiamo dedurre?

1. Almeno 450 italiani usano il web: corretto ma del tutto inutile.
2. Più di un quarto degli italiani usano il web: non è conseguenza logica dei dati, ma è plausibile.
3. Almeno 39 milioni di italiani usano il web: non contraddice i dati, ma non è plausibile.

La Statistica definisce in modo non ambiguo e quantitativo tale plausibilità.

- La plausibilità di un'inferenza dipende dalle modalità con cui è stato selezionato il campione di 1000 individui della popolazione.
- Se scegliessi tutti i 1000 tra i laureati in informatica, o tra coloro che risiedono oltre i 2000 metri di quota, otterrei ovviamente un campione non significativo della popolazione.

- La metodologia corretta è quella della selezione casuale: è come estrarre 1000 palline da una gigantesca urna che ne contiene 40 milioni. La presenza di casualità nel campionamento rende indispensabile in Statistica il Calcolo delle Probabilità (la matematica (sigh..) che studia i fenomeni casuali o aleatori).

In gran parte dei casi un'analisi statistica ha, schematicamente, le caratteristiche di un processo penale.

1. Un'ipotesi investigativa: un crimine e un'imputato.
2. Un'accurata indagine da parte della polizia.
3. Un pubblico ministero che presenti i risultati delle indagini in modo chiaro, accurato e non tendenzioso.
4. Una giuria equa.

Tradotto nel linguaggio della statistica:

1. Un'ipotesi statistica: ad esempio "più di un quarto degli italiani usano il web".
2. Un'indagine sperimentale: una procedura di campionamento per raccogliere dati sulle variabili di interesse per l'ipotesi statistica.
3. Delle procedure per presentare quanto raccolto nell'indagine sperimentale. Nell'esempio in esame è sufficiente dire: su 1000 individui campionati casualmente 450 usano il web. Vedremo esempi in cui la descrizione dei dati sperimentali è più complessa: questa è materia della statistica descrittiva.
4. Dei metodi ben fondati per accettare o rigettare l'ipotesi statistica, sulla base di una ben definita nozione di plausibilità: questa è materia della statistica inferenziale.

La statistica nelle scienze della vita

L'impatto e la necessità della statistica nelle scienze della vita (Biologia, Medicina .....) sono cresciuti in modo enorme negli ultimi decenni, tanto da rendere (in apparenza) la Biostatistica una disciplina a sè stante.

Lo sviluppo straordinario di Biologia e Biomedicina ha reso possibile esperimenti prima impensabili, che forniscono numerosi dati che devono essere analizzati e interpretati. Ecco alcuni esempi.

- Analisi dei fattori di rischio in medicina.
- Test clinici (efficacia di farmaci e terapie).
- Modelli e studio statistico di sequenze nel DNA

2 Statistica descrittiva

Obiettivo: presentare i dati di una, ricerca, indagine..... in modo sintetico attraverso grafici, tabelle e altre forme di sintesi.

I dati si possono riferire a una o più quantità misurate (*variabili*). I dati relativi ad una variabile si presentano come una sequenza $x_1, x_2, \ldots, x_n$ di valori della variabile (*campione*), ottenuti in n "misurazioni".

Le variabili vengono divise in due grandi categorie: le variabili *numeriche*, che assumono cioè valori numerici, e le variabili *categoriche* (tutte le altre)

2.1 Statistica descrittiva per variabili categoriche

I valori possibili di una variabile categorica vengono chiamati categorie. I dati vengono rappresentati tramite tabelle e grafici.

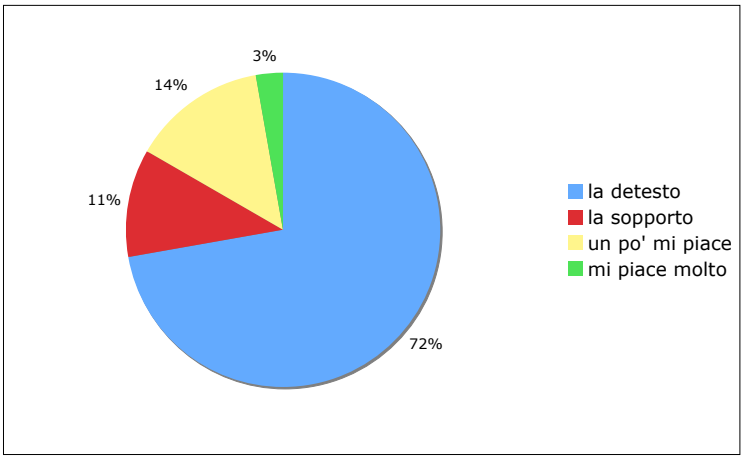
Lo illustro con un esempio, tratto da un'indagine in cui si è chiesto a studenti di area Biologica: quanto vi piace la Statistica?

	frequenza assoluta	**frequenza relativa**	**frequenza percentuale**
la detesto	78	0,7573	75,73%
la sopporto	12	0,1165	11,65%
un po' mi piace	15	0,1456	14,56%
mi piace molto	3	0,0291	2,91%

Frequenza assoluta = numero di occorrenze di un dato valore

$$\text{Frequenza relativa} = \frac{\text{frequenza assoluta}}{\text{numero di osservazioni}}$$

Frequenza percentuale = Frequenza relativa $\times$ 100 %



Più variabili categoriche relative ad una medesima popolazione possono venire usate per studi di *correlazione*. Ad esempio gli studenti di cui sopra possono essere suddivisi a seconda dell'appartenenza ad uno dei tre corsi di laurea: Biologia, Biologia Molecolare, Biotecnologia.

	biologia	**biologia molecolare**	**biotecnologia**	
la detesto	37	21	20	78
la sopporto	3	6	3	12
un po' mi piace	5	6	4	15
mi piace molto	1	2	0	3
	46	35	27	108

Da tabelle di questo tipo, dette *tabelle di contingenza*, si vogliono rilevare eventuali *correlazioni* tra le due variabili: in questo caso se la disposizione verso la statistica sia sostanzialmente la stessa nei tre corsi di laurea, o vi siano differenze significative. Questo tipo di analisi verrà affrontata verso la fine di questo corso.

2.2 Statistica descrittiva per variabili numeriche

Le variabili numeriche possono assumere valori *discreti* (il punteggio di un dado, il numero di studenti iscritti ad un corso di laurea, il numero di sigarette fumate in un giorno....) oppure *continui* (la pressione del sangue, la densità di polveri sottili nell'aria, il tempo che trascorre tra due successive eruzioni di un vulcano...).

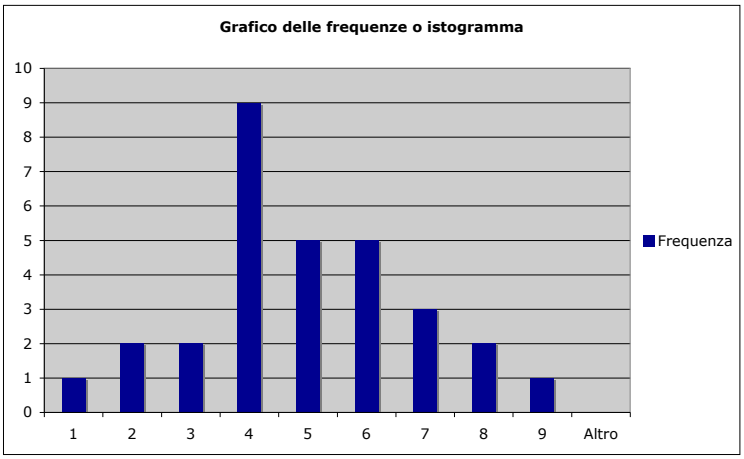
I dati relativi a variabili con valori numerici discreti possono essere rappresentati con tabelle in modo simile a variabili categoriche. Tuttavia, l'ordine

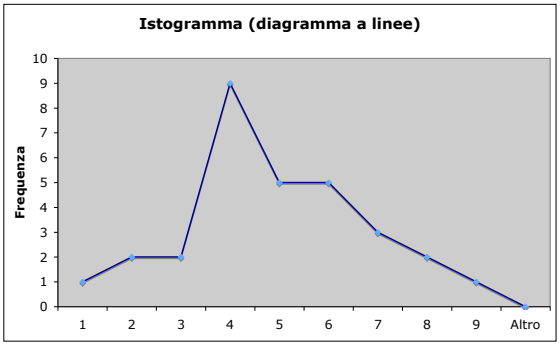
naturale tra i numeri rende rappresentazioni grafiche tipo *istogramma* più espressive di quelle tipo "a torta" che abbiamo visto per variabili categoriche.

Consideriamo, a titolo di esempio, i seguenti dati, che danno il numero di casi classificato "gravi" giunti dal 1 al 30 settembre 2006 al pronto soccorso dell'Ospedale di Borgo Trento a Verona.

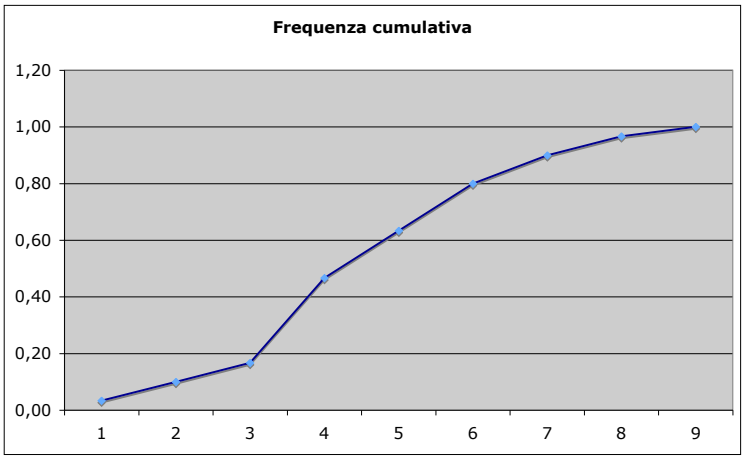
giorno	n. di casi gravi
1	4
2	7
3	5
4	2
5	1
6	5
7	7
8	3
9	4
10	9
11	6
12	4
13	5
14	5
15	5
16	6
17	8
18	8
19	2
20	4
21	4
22	7
23	4
24	6
25	4
26	4
27	3
28	6
29	6
30	4

N. casi gravi	*Frequenza*	*Frequenza relativa*	*Frequenza cumulativa*
1	1	0,0333	0,0333
2	2	0,0667	0,1000
3	2	0,0667	0,1667
4	9	0,3000	0,4667
5	5	0,1667	0,6333
6	5	0,1667	0,8000
7	3	0,1000	0,9000
8	2	0,0667	0,9667
9	1	0,0333	1,0000
Altro	0		





Distribuzioni di dati con istogrammi di questo tipo sono chiamate unimodali: sono caratterizzate da un unico "picco", con frequenze crescenti alla sinistra del picco e decrescenti a destra.

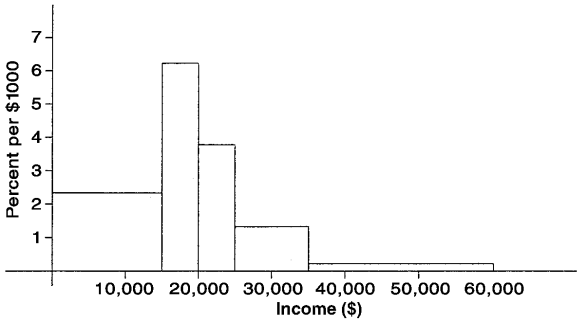


Per variabili a valori continui, per fornire rappresentazione tabulari e grafiche è prima opportuno suddividere l'insieme dei valori possibili in *intervalli disgiunti* (classi).

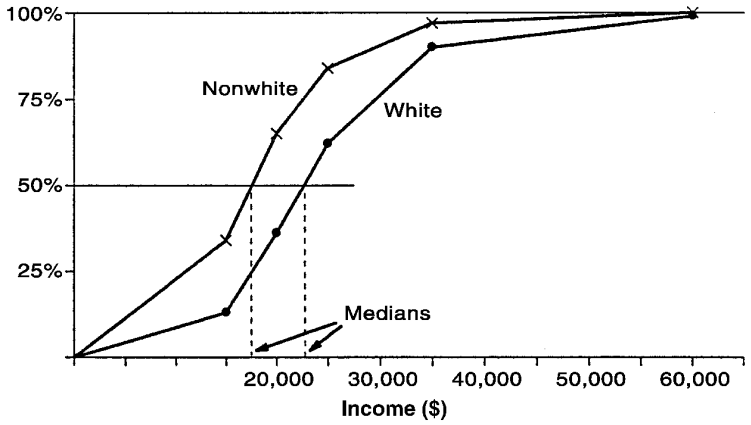
Consideriamo i seguenti dati, relativi al reddito delle famiglie americane nel 1983. I dati riguardano due variabili: reddito delle famiglie di etnia bianca; reddito delle famiglie di diversa etnia.

	Percent of Families	
Income ($)	White	Nonwhite
0–14,999	13	34
15,000–19,999	24	31
20,000–24,999	26	19
25,000–34,999	28	13
35,000–59,999	9	3
60,000 and over	1	Negligible
Total	100	100

Il seguente istogramma riguarda i dato relativi alle famiglie "non bianche".



Dato che gli intervalli delle classi non sono tutti della stessa ampiezza, l'altezza di ogni barra è uguale alla percentuale di appartenenti alla classe divisa per l'ampiezza dell'intervallo della classe. In tal modo la percentuale della classe è proporzionale all'*area* della barra. Interessanti sono i digrammi a linee delle frequenze percentuali cumulative per le due variabili. Si vede chiaramente la differenza tra le distribuzioni delle due variabili



Indici numerici

A fianco delle rappresentazioni grafiche, importanti indicazioni sintetiche dei dati sono fornite da *indici numerici*.

Siano $x_1, x_2, \ldots, x_n$ i dati per una variabile numerica x ottenuti da n osservazioni *campione di dati*).

Una funzione $f(x_1, x_2, \ldots, x_n)$ di tali dati viene chiamata *statistica campionaria*.

Tra le statistiche campionarie che introdurremo distinguiamo gli *indici di posizione* e gli *indici di dispersione*.

Indici di posizione

I primi due indici che introduciamo rappresentano due diversi modi di identificare il "centro" di una distribuzione di dati.

MEDIA CAMPIONARIA

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

Esempio. La media campionaria dei dati $2, 1, 5, 3, 2$ è

$$\frac{2+1+5+3+2}{5} = \frac{13}{5} = 2.6$$

Il valore $(x_i - \overline{x})^2$ viene chiamato lo *scarto quadratico* del dato x_i. Più in generale, dato un numero reale c, $(x_i - c)^2$ viene chiamato *scarto quadratico da* c del dato x_i. Si può dimostrare che la somma degli scarti quadratici da c

$$\sum_{i=1}^{n}(x_i - c)^2$$

raggiunge il suo valore minimo per $c = \overline{x}$.

MEDIANA

Per definire la mediana, per prima cosa si riordinino i dati $x_1, x_2, \ldots, x_n$ in ordine crescente:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

La mediana m_x è definita da

$$m_x := \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ è dispari} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{se } n \text{ è pari} \end{cases}$$

In altre parole, per un numero dispari di dati, la mediana è il dato "centrale" nella disposizione ordinata. Per un numero pari di dati è la media aritmetica dei "due dati centrali".

Ad esempio, per calcolare la mediana dei dati $2, 1, 5, 3, 2$ si procede come segue:

1. Si ordina i dati secondo un ordine crescente: $1, 2, 2, 3, 5$.
2. Essendo $n = 5$, m_x è il dato di posto $\frac{n+1}{2} = 3$ nella disposizione ordinata, cioè $m_x = 2$

Similmente, per calcolare la mediana dei dati $2, 1, 5, 3, 2, 7$, una volta ordinati $(1, 2, 2, 3, 5, 7)$, la mediana è la media aritmetica tra il dato di posto $n/2 = 3$ e il successivo, cioè

$$m_x = \frac{2+3}{2} = 2.5$$

La mediana ha, rispetto alla media campionaria, la caratteristica di essere poco sensibile a dati "molto atipici": questo può essere un vantaggio se non si è certi dell'affidabilità di alcune misurazioni. Tuttavia la mediana è raramente usata in statistica inferenziale, contrariamente alla media campionaria.

Si può dimostrare che la mediana m_x *minimizza* in c lo *scarto totale*

$$\sum_{i=1}^{n} |x_i - c|$$

Nel caso di variabili numeriche discrete viene talvolta usato un altro indice di posizione, detto *moda*: si tratta del valore a cui corrisponde la frequenza massima. La definizione ambigua (la frequenza massima potrebbe essere raggiunta da più di un valore) e difficilmente estendibile a variabili continue, rendono la *moda* una statistica campionaria scarsamente usata.

PERCENTILI CAMPIONARI

La mediana può essere grossolanamente descritta come quel valore m_x tale che "esattamente" la metà (o il 50%) dei dati hanno valore non maggiore di m_x. Generalizzando, consideriamo $p \in [0, 1]$ e chiamiamo $100p$-esimo percentile quel valore c per cui "esattamente" il $100p\%$ sono minori di c. Più precisamente:

- Consideriamo, come sopra, l'ordinamento crescente $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ dei dati.
- Se np non è un numero intero, sia k l'intero immediatamente successivo a np. Il $100p$-esimo percentile è, per definizione, $x_{(k)}$.
- Se np è un numero intero, allora si definisce come $100p$-esimo percentile la media aritmetica di $x_{(np)}$ e $x_{(np+1)}$ (eccetto il caso $p = 0$ e $p = 1$), cioè $\frac{x_{(np)}+x_{(np+1)}}{2}$ (per $p = 0$ lo 0-esimo percentile è $x_{(1)}$, cioè il dato minimo; per $p = 1$ il 100^o percentile è $x_{(n)}$, cioè il dato massimo).

Il 50^o percentile (cioè $p = 1/2$) non è altro che la mediana.
Per esempio, consideriamo i dati $-2, 4, 3, 1, 2, 4, 3, 6$. Ordinandoli si ottiene

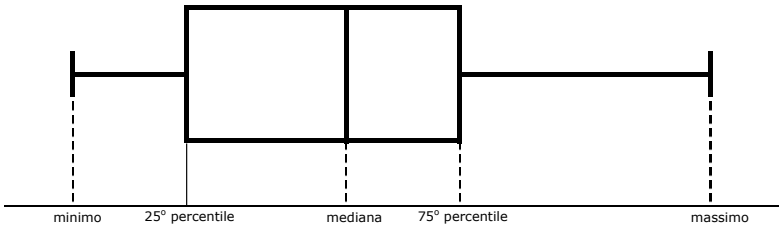
$$-2 \quad 1 \quad 2 \quad 3 \quad 3 \quad 4 \quad 4 \quad 6$$

Calcoliamo il 25^o percentile: $n = 8$, $p = 0.25$, $np = 2$. Il 25^o percentile è 1.5.

Calcoliamo il 40^o percentile: $np = 8 \times 0.4 = 3.2$. Il 40^o percentile è 3.

I percentili più usati, a parte gli estremi ($p = 0$ e $p = 1$) sono il 25^o, detto anche *primo quartile* e denotato con Q_1, la mediana e il 75^o, detto anche *terzo quartile* e denotato con Q_3.

L'informazione contenuta in questi percentili viene rappresentata graficamente tramite i diagrammi *boxplot*.



Indici di dispersione

Tali indici intendono misurare quanto i dati siano concentrati/dispersi.

L'indice di dispersione di gran lunga più usato è la **VARIANZA CAMPIONARIA** definita da

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

Notare che $s_x^2 \geq 0$, e $s_x^2 = 0$ *se e solo se i dati sono tutti uguali*.

Notare anche che la varianza campionaria non ha la stessa unità di misura dei dati, bensì il *quadrato* di quell'unità di misura. La *deviazione standard campionaria*, definita da

$$s_x := \sqrt{s_x^2}$$

ha invece la stessa unità di misura dei dati.

Si noti che un naturale indice di dispersione è dato dallo *scarto quadratico medio*:

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

che differisce dalla varianza campionaria per un fattore $\frac{n}{n-1}$. La ragione di questo fattore è di natura probabilistica, e verrà chiarita più avanti nel corso.

Un altro indice di dispersione largamente usato in statistica descrittiva è la *differenza interquartile*

$$Q_3 - Q_1$$

che fornisce l'ampiezza di un intervallo "centrale" che contiene (circa) la metà dei dati.

Trasformazione di media e varianza campionarie per cambi di unità di misura

I dati $x_1, x_2, \ldots, x_n$ di un campione sono usualmente espressi in una determinata unità di misura (metri, Kg, gradi centigradi....). Qualora si esegua

un cambiamento di unità di misura (ad esempio da gradi centigradi a gradi Farenheit), i singoli dati si trasformano secondo una *trasformazione affine*:

$$y_i = ax_i + b$$

dove le y_i sono i dati espressi nella nuova unità di misura, a e b sono costanti che dipendono dalle due unità di misura. Ad esempio, nella conversione da gradi centigradi a gradi Farenheit, $a = \frac{9}{5}$ e $b = 32$.

Non è difficile verificare le seguenti *formule di trasformazione*:

$$\overline{y} = a\overline{x} + b \qquad\qquad s_y^2 = a^2 s_x^2$$

Correlazione tra due variabili

Consideriamo due variabili numeriche x e y misurate sugli *stessi* individui di una popolazione. In altre parole abbiamo *due* campioni di dati

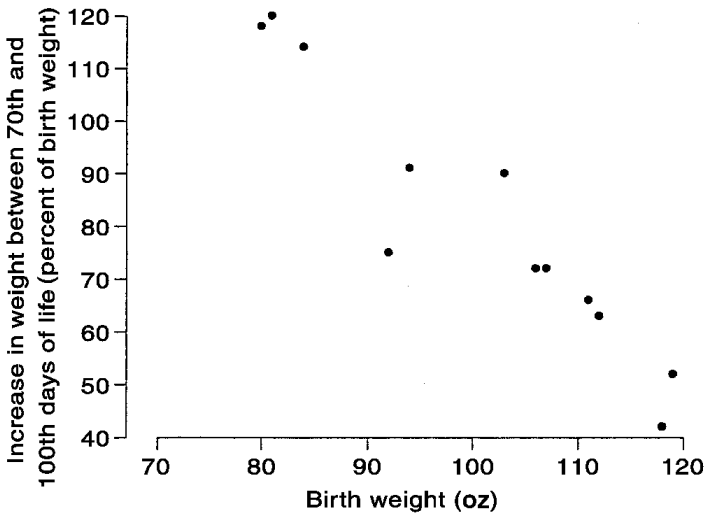
$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{array}$$

dove x_i e y_i sono i valori delle due variabili misurate sullo stesso individuo.

Nell'esempio che segue è stato considerato un gruppo di 12 neonati per i quali è stato misurato: il peso alla nascita (x) e l'aumento percentuale di peso tra il 70^o e il 100^o giorno di vita.

x (oz)	y (%)	x (oz)	y (%)
112	63	81	120
111	66	84	114
107	72	118	42
119	52	106	72
92	75	103	90
80	118	94	91

Per avere un'idea della correlazione tra le due variabili, è opportuno rappresentare le coppie di dati (x_i, y_i) come punti di un *piano cartesiano*, ottenendo il relativo *diagramma di dispersione*



Tale diagramma indica una *correlazione negativa* tra le due variabili: ad un minor peso alla nascita corrisponde normalmente un maggior aumento di peso successivo.

Per quantificare tale correlazione viene introdotto il *coefficiente di correlazione*:

$$r := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

Si può dimostrare (non è semplicissimo!) che

- $-1 \leq r \leq 1$
- $r = 1$ (risp. $r = -1$) se e soltanto se i punti del diagramma di dispersione sono *esattamente* allineati lungo una retta con coefficiente angolare positivo (risp. negativo).

Più in generale, se $|r|$ è "vicino" a 1 i punti del diagramma di dispersione sono disposti "vicino" ad una retta non orizzontale: in questo caso si dice che le due variabili sono *fortemente correlate*.

Se invece $|r| \simeq 0$ i punti del diagramma di dispersione non sono "approssimabili" con una retta, e ciò si esprime dicendo che le variabili sono sostanzialmente *scorrelate*.

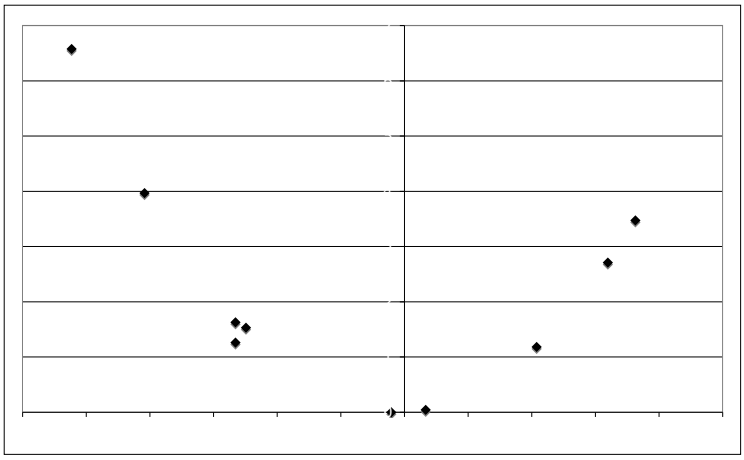
Nell'esempio precedente $r = -0.946$, che indica effettivamente una forte correlazione negativa.

È importante avere presente un limite fondamentale del coefficiente di correlazione: esso è in grado di rilevare la *correlazione lineare* tra due variabili, ma non necessariamente altri tipi di correlazioni.

Consideriamo, ad esempio, i seguenti dati:

x	y
-1,245	1,551
1,597	2,551
-2,042	4,171
1,815	3,294
-1,327	1,761
0,166	0,028
-0,104	0,011
1,039	1,079
-1,328	1,764
-2,615	6,837

Essi producono il diagramma di dispersione



Si vede che c'è una forte "dipendenza" tra le due variabili, ma di tipo *non lineare*.

Il coefficiente di correlazione (-0.438) è sostanzialmente privo di significato.

3 Probabilità e modelli probabilistici

3.1 La Probabilità e le sue proprietà fondamentali

Useremo il termine fenomeno aleatorio per indicare per indicare qualsiasi vicenda reale o ideale di cui si possano descrivere gli *esiti possibili*, e per la quale l'esito sia almeno parzialmente imprevedibile.

Per formulare un *modello matematico* di un fenomeno aleatorio è necessario, per prima cosa, identificare un insieme, che indicheremo con S, che contenga tutti gli esiti possibili.

Esempi

- S = insieme dei punteggi ottenibili nel lancio di un dado = $\{1, 2, 3, 4, 5, 6\}$
- S = il patrimonio genetico del futuro figlio di una coppia
- S = il tempo che impiegherà il mio computer a scaricare (illegalmente!) l'ultimo film di Spielberg = $[0, +\infty)$

Affermazioni del tipo

- Il punteggio ottenuto lanciando il dado sarà almeno 3
- La coppia avrà un figlio maschio con i capelli neri e gli occhi verdi
- Riuscirò a scaricare il film in meno di 20 ore

individuano, in ognuno degli esempi precedenti, sottoinsiemi dell'insieme S.

Una Probabilità è una regola per assegnare in modo "coerente" un *grado di fiducia* (che chiameremo, appunto, probabilità) ai sottoinsiemi di un insieme S

Introduciamo nozioni e notazioni in modo più preciso.

L'insieme S degli esiti verrà (talvolta) chiamato *spazio campionario*. I sottoinsiemi di S verranno chiamati *eventi*

L'insieme di tutti gli eventi verrè indicato con $\mathcal{P}(S)$.

Definizione. Una *probabilità* P è una funzione

$$P : \mathcal{P}(S) \ \rightarrow \ [0, 1]$$

dall'insieme degli eventi all'intervallo dei numeri reali compresi fra 0 e 1, per cui siano soddisfatte le seguenti proprietà:

(P1). $P(S) = 1$ (l'evento "certo" ha probabilità 1)

(P2). Se $A, B \in \mathcal{P}(S)$ e $A \cap B = \emptyset$, allora

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

(la probabilità che si verifichi uno di due eventi incompatibili è la somma delle probabilità dei due eventi)

Prima di vedere degli esempi di probabilità, è utile vedere alcune semplici conseguenze delle proprietà $(P1)$ e $(P2)$. Nel seguito useremo le seguenti notazioni insiemistiche:

$$A \setminus B := \{s \in A \text{ tali che } s \notin B\}$$

$$A^c = S \setminus A \quad (\text{complementare di } A)$$

Le seguenti conseguenze di $(P1)$ e $(P2)$ si dimostrano facilmente

1. $P(\emptyset) = 0$

   Infatti: $\emptyset \cap S = \emptyset$ e $\emptyset \cup S = S$. Quindi per $(P1)$ e $(P2)$

   $$1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset) = 1 + P(\emptyset)$$

   da cui segue $P(\emptyset) = 0$.

2. Se $A \subseteq B$ allora $P(B \setminus A) = P(B) - P(A)$. In particolare $P(A^c) = 1 - P(A)$

   Infatti: $B = A \cup (B \setminus A)$ e $A \cap (B \setminus A) = \emptyset$. Perciò, per $(P2)$

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

   da cui si conclude facilmente.

3. Se $A_1, A_2, \ldots, A_n$ sono n eventi a due a due disgiunti, cioè $A_i \cap A_j = \emptyset$ per $i \neq j$, allora

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

   Tale proprietà si ottiene applicando ripetutamente $(P2)$.

4. Se A e B sono due eventi generici $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

   Per dimostrarlo, scriviamo

   $$A \cup B = [A \setminus (A \cap B)] \cup [B \setminus (A \cap B)] \cup (A \cap B)$$

   I tre eventi di quest'ultima unione sono a due a due disgiunti. Perciò, per la proprietà 3. appena dimostrata, e quindi usando la proprietà 2.

   $$\begin{aligned} P(A \cup B) &= P[A \setminus (A \cap B)] + P[B \setminus (A \cap B)] + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

ESEMPI

Ex. 1

Sia S un insieme con un numero finito di elementi. Per $A \subseteq S$ denotiamo con $|A|$ il numero di elementi di A. Definiamo

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

È facile vedere che le proprietà (P1) e (P2) sono soddisfatte. Questa probabilità viene talvolta chiamata *uniforme*.
Questa probabilità può essere caratterizzata anche in un'altro modo: se $s \in S$, $P(\{s\})$ è la stessa per tutti gli s.

Gli esempi qui di seguito si sviluppano sulla base di questo esempio fondamentale.

Ex. 2

Si consideri il fenomeno aleatorio consistente nell'estrarre "a caso" 5 carte (distinte) da un mazzo di 52 carte da poker. Dunque S è l'insieme di tutte le possibili scelte di 5 carte e, essendo la scelta del tutto casuale, la probabilità "naturale" su S è quella uniforme.

Per calcolare la probabilità di un evento in S, devo prima calcolare il numero di elementi di S. Notare che un elemento di S corrisponde ad un sottoinsieme di 5 carte dell'insieme di tutte le 52 carte del mazzo.

In generale, per un insieme di n elementi, e $k = 0, 1, \ldots, n$, il numero di sottoinsiemi con k elementi si denota con $\binom{n}{k}$ e vale

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

dove $0! = 1$ e, per $k > 0$,

$$k! = k \cdot (k-1) \cdot (k-2) \cdots 2 \cdot 1$$

Perciò

$$|S| = \binom{52}{5} = \frac{52!}{5!47!} = \frac{52 \cdot 51 \cdots 48}{5!} = 2598960$$

Consideriamo ora l'evento costituito dalle scelte di 5 carte che contengono almeno un asso. Denotiamo con A questo evento, e calcoliamone la probabilità. Poichè $P(A) = \frac{|A|}{|S|}$, si tratta di determinare il numero di elementi di A.
Come spesso accade quando un evento è definito dal quantificatore logico "almeno uno", è più facile determinare la probabilità del suo complementare (qualche testo chiama questa evidenza il "principio dell'almeno uno"). Infatti gli elementi di A^c sono tutte le scelte di 5 carte che *non contengono alcun asso*; contare queste scelte equivale a contare le scelte di 5 carte da un mazzo di 48 carte a cui sono stati tolti i 4 assi.

Perciò

$$|A^c| = \binom{48}{5}$$

Allora

$$P(A^c) = \frac{|A^c|}{|S|} = \frac{\binom{48}{5}}{\binom{52}{5}} = \frac{\frac{48!}{5!43!}}{\frac{52!}{5!47!}} = \frac{48 \cdot 47 \cdots 44}{52 \cdot 51 \cdots 48} = 0,658841998$$

$$P(A) = 1 - P(A^c) = 0,341158002$$

Ex. 3

Consideriamo un semplice modello di trasmissione dei caratteri. Ad un dato carattere corrispondono i tre genotipi aa, aA e AA, ognuno di essi identificato da una coppia di geni.

Consideriamo una coppia di individui entrambi di genotipo aA. Quale sarà il genotipo del prossimo figlio? O meglio: qual'e' la probabilità corretta da assegnare alle diverse possibilità?

$S = \{aa, aA, AA\}$ è l'insieme dei possibili genotipi del figlio. Secondo le regole della trasmissione dei caratteri, il figlio "sceglierà a caso" un gene dal padre e uno dalla madre. Indicando con il colore rosso il gene scelto dalla madre, sono possibili 4 configurazioni:

$$aa, aA, aA, AA$$

ognuna delle quali ha probabilità $1/4$. Le due centrali danno tuttavia luogo allo stesso elemento di S, cioè aA.

Pertanto

$$P(\{aa\}) = \frac{1}{4} \qquad P(\{aA\}) = \frac{1}{2} \qquad P(\{AA\}) = \frac{1}{4}$$

Ad esempio, la probabilità che il figlio *manifesti* il carattere dominante A è

$$P(\{aA, AA\}) = P(\{aA\}) + P(\{AA\}) = \frac{3}{4}$$

3.2 Probabilità condizionata

Sia S uno spazio campionario, e P una fissata probabilità sui suoi eventi. Dati due eventi A e B con $P(B) > 0$, la *probabilità di A condizionata a B* è definita da

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

Intuitivamente $P(A|B)$ rappresenta la probabilità con cui riteniamo si verifichi l'evento A *se sappiamo* l'evento B si verificherà sicuramente.

Non è difficile mostrare che, fissato B, la funzione $A \mapsto P(A|B)$ soddisfa alle proprietà (P1) e (P2) che definiscono una probabilità. Da ciò segue, ad esempio, che se A_1, A_2 sono disgiunti, allora $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B)$, e che $P(A^c|B) = 1 - P(A|B)$.

Esempi

Ex. 1

Supponiamo di giocare un gioco che consiste nel lanciare due dadi, uno alla volta: vinciamo se il punteggio totale ottenuto è almeno 9. Chiamiamo A l'evento che corrisponde alla vittoria. Supponiamo che il punteggio ottenuto nel lancio del primo dado sia 5. Indichiamo con B questo secondo evento. Quale valore assegneremo alla probabilità di vincere dopo aver visto il punteggio del primo dado?

In altre parole, quanto vale $P(A|B)$?

Per rispondere a questa domanda, indichiamo con il colore rosso il punteggio del primo dado. Ogni esito possibile del lancio dei due dadi è dunque rappresentabile nella forma (i, j), dove i e j possono assumere i valori $1, 2, 3, 4, 5, 6$. Si vede facilmente che l'insieme S di tutti tali esiti ha 36 elementi che, nel caso di dadi equilibrati, sono equiprobabili, cioè P è la probabilità uniforme su Ss.

Inoltre

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

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

Da ciò si ricava

$$A \cap B = \{(5,4), (5,5), (5,6)\}$$

Pertanto $P(A \cap B) = 3/36 = 1/12$, $P(B) = 6/36 = 1/6$, e quindi

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/12}{1/6} = \frac{1}{2}$$

Ex. 2

Nella pratica biomedica una grande importanza è rivestita dagli *screening test*. Esistono delle patologie la cui precoce e sicura individuazione è molto costosa o invasiva. I pazienti vengono pertanto prima sottoposti ad un test clinico più semplice (screening test), quindi solo coloro che risultano positivo al test vengono sottoposti ad un'analisi più accurata.

Consideriamo un individuo che venga sottoposto allo screening test. Consideriamo gli eventi

$$A := \text{“risulta positivo allo screening test”}$$

$$B := \text{“è affetto dalla patologia”}$$

Le probabilità $P(A|B)$ e $P(A|B^c)$ riguardano la *sensibilità* del test. Esse vengono stimate su base sperimentale. Il test viene applicato a due gruppi di individui: i membri del primo gruppo sono *sicuramente* affetti dalla patologia, i membri del secondi sono *sicuramente* sani. La frazione dei membri del primo

gruppo che risultano positivi è una *stima* per $P(A|B)$, mentre la frazione dei membri del secondo gruppo che risultano positivi è una *stima* per $P(A|B^c)$.

Un'altra probabilità stimabile sperimentalmente è $P(B)$, ossia la probabilità che un'individuo scelto da una fissata popolazione sia affetto dalla patologia. In altre parole, l'*incidenza* della patologia nella popolazione, o la *frazione* di popolazione affetta. Questo dato è spesso stimabile sulla base di dati "storici".

Le probabilità condizionate $P(B|A)$ e $P(B|A^c)$ riguardano invece la *predittività* del test. Le modalità per il loro calcolo saranno illustrate nel seguito.

Restiamo ancora un attimo nel contesto dell'esempio precedente. Supponiamo siano note da dati sperimentali la sensibilità del test $P(A|B)$ e $P(A|B^c)$ e l'incidenza della patologia $P(B)$.

Qual è la probabilità che un individuo scelto a caso risulti positivo allo screening test? In altre parole: qual è la percentuale di popolazione che risulterebbe positiva al test?

Si tratta cioè di calcolare $P(A)$ sulla base dei dati.

A questo scopo, in tutta generalità, usiamo la seguente formula nota come *formula delle probabilità totali* o *formula di disintegrazione* o *formula di fattorizzazione*:

- Dati due eventi A e B, con $0 < P(B) < 1$, si ha

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Per dimostrare la formula di disintegrazione, osserviamo che, per la proprietà (P2)

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Inoltre, per definizione di probabilità condizionata,

$$P(A \cap B) = P(A|B)P(B) \qquad P(A \cap B^c) = P(A|B^c)P(B^c)$$

Sostituendo nella precedente la formula è dimostrata.

Tornando all'esempio dello screening test, se la sensibilità del test è $P(A|B) = 0.995$, $P(A|B^c) = 0.12$, e l'incidenza della patologia è $P(B) = 0.03$, allora la probabilità che un test risulti positivo è

$$\begin{aligned} P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) \\ = 0.995 \cdot 0.03 + 0.12 \cdot 0.97 = 0.14625 \end{aligned}$$

La formula delle probabilità totali ammette la seguente generalizzazione, talvolta utile.

- Siano $B_1, B_2, \ldots, B_n$ n eventi a due a due disgiunti la cui unione sia tutto S (*partizione* di S), e assumiamo $P(B_i) > 0$ per $i = 1, 2, \ldots, n$. Allora, per ogni altro evento A

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

Si noti che per $n = 2$ si ricade nella formula di prima.

Esempio

In un Dipartimento Universitario ci sono 100 computers indistinguibili. Di questi 40 sono stati assemblati dalla ditta AA, 35 dalla ditta BB e gli altri dalla ditta CC. I computer della ditta AA hanno una probabilità 0.87 di funzionare almeno tre anni senza guasti; tale probabilità è 0.98 per i computer della ditta BB e 0.70 per quelli della ditta CC. Io non ho idea da quale ditta venga il mio computer. Qual è la probabilità che funzioni senza guasti almeno tre anni?

Siano:

$B_1 =$ "il mio computer viene dalla ditta AA"
$B_2 =$ "il mio computer viene dalla ditta BB"
$B_3 =$ "il mio computer viene dalla ditta CC"

e $A =$"il mio computer funzionerà senza guasti almeno tre anni".

I dati sono: $P(A|B_1) = 0.87$, $P(A|B_2) = 0.98$, $P(A|B_3) = 0.70$, $P(B_1) = 0.4$, $P(B_2) = 0.35$, $P(B_3) = 0.25$

Perciò

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) = 0,866$$

Tornando un'attimo al problema degli screening test, le quantità più rilevanti dal punto di vista delle applicazioni sono le probabilità condizionate $P(B|A)$ e $P(B|A^c)$, che caratterizzano la *predittività* del test.

Il calcolo di tali probabilità fa uso di una formula generale, chiamata *Formula di Bayes*

- Siano A e B due eventi tali che $P(A) > 0$ e $P(B) > 0$. Allora

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

La dimostrazione di questa formula si ottiene semplicemente osservando che

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

Con la Formula di Bayes possiamo calcolare la predittività del test. Ricordiamo che con i dati sulla sensibilità del test $P(A|B) = 0.995$, $P(A|B^c) = 0.12$, e sull'incidenza della patologia $P(B) = 0.03$, avevamo calcolato

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = 0.14625$$

Ora calcoliamo

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.995 \cdot 0.03}{0.14625} = 0.204102564$$

e

$$P(B|A^c) = \frac{P(A^c|B)P(B)}{P(A^c)} = \frac{[1 - P(A|B)]P(B)}{1 - P(A)} = 0,000188466$$

Possiamo concludere che la probabilità di ottenere un "falso negativo" ($P(B|A^c)$) è molto piccola, ma la probabilità di ottenere un "falso positivo" ($P(B^c|A) = 1 - P(B|A)$) è piuttosto elevata: circa l'80% di coloro che risultano positivi al test sono in realtà sani!

3.3 Indipendenza

Sia S uno spazio campionario e P una probabilità sui suoi eventi.

- Due eventi A e B si dicono *indipendenti* se

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

Si noti che, se $P(B) > 0$, gli eventi A e B sono indipendenti se e solo se

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

N.B.: non confondere l'indipendenza con l'incompatibilità

Esempio

Si consideri l'esempio già visto in precedenza del lancio di due dadi, in cui

$$S = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$$

e i due eventi

$A =$ "il punteggio del primo dado è 5"
$B =$ "il punteggio del secondo dado è almeno 4"

Verifichiamo che questi due eventi sono indipendenti.

Possiamo riscrivere

$$A = \{(5, j) : j = 1, 2, 3, 4, 5, 6\}$$
$$B = \{(i, j) : i = 1, 2, 3, 4, 5, 6;\ j = 4, 5, 6\}$$

Notare, in particolare, che

$$P(A) = \frac{6}{36} = \frac{1}{6} \qquad P(B) = \frac{18}{36} = \frac{1}{2}$$

Inoltre

$$A \cap B = \{(5, j) : j = 4, 5, 6\}$$

Quindi

$$P(A \cap B) = \frac{3}{36} = \frac{1}{12} = P(A)P(B)$$

I due eventi sono perciò indipendenti.

Un'utile proprietà dell'indipendenza è la seguente.

- Siano A e B due eventi indipendenti. Allora anche A^c e B sono eventi indipendenti, così come lo sono A e B^c, e A^c e B^c

In due parole: l'indipendenza è "insensibile" alla complementazione.

Non è difficile dimostrare la proprietà precedente. Mostriamo, ad esempio, che A^c e B sono indipendenti:

$$\begin{aligned} P(A^c \cap B) = P(B \setminus (A \cap B)) &= P(B) - P(A \cap B) \\ &= P(B) - P(A)P(B) = (1 - P(A))P(B) = P(A^c)P(B) \end{aligned}$$

Fin qui abbiamo visto la nozione di indipendenza di due eventi. Essa può essere generalizzata ad un numero n arbitrario di eventi.

- n eventi $A_1, A_2, \ldots, A_n$ si dicono indipendenti se, per ogni $k = 2, 3, \ldots, n$ e per ogni sottogruppo $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ degli eventi $A_1, A_2, \ldots, A_n$ si ha

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

Ad esempio, per mostrare che A, B e C sono indipendenti, non è sufficiente mostrare che

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

ma bisogna anche mostrare che

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

Esempio: lo schema delle prove ripetute indipendenti

Consideriamo una "prova" il cui esito possa essere successo o insucesso. e che inoltre la probabilità che l'esito della prova sia un successo sia $p \in [0, 1]$. Supponiamo di eseguire n ripetizioni della prova, in modo tale che gli esiti di prove distinte siano indipendenti.

Più precisamente, per $i = 1, 2, \ldots, n$, si consideri l'evento

$$A_i = \text{"la prova } i\text{-esima è un successo"}$$

Assumiamo che gli eventi $A_1, A_2, \ldots, A_n$ siano indipendenti, e $P(A_i) = p$.

- Qual è la probabilità di ottenere *almeno* un successo?
- Qual è la probabilità che il primo successo si realizzi all'n-esimo tentativo?
- Qual è la probabilità di ottenere k successi e $n - k$ insuccessi, con $k = 0, 1, \ldots, n$?

Sia B l'evento "si ottiene almeno un successo". Si noti che

$$B^c = A_1^c \cap A_2^c \cap \cdots \cap A_n^c$$

Poiché gli eventi $A_1^c, A_2^c, \ldots, A_n^c$ sono indipendenti (si ricordi che l'indipendenza è "insensibile" alla complementazione),

$$P(B^c) = P(A_1^c)P(A_2^c)\cdots P(A_n^c) = (1-p)^n$$

e quindi

$$P(B) = 1 - (1-p)^n$$

Sia C l'evento "il primo successo è realizzato all'n-esimo tentativo".

$$C = A_1^c \cap A_2^c \cap \cdots \cap A_{n-1}^c \cap A_n$$

Usando di nuovo l'ipotesi di indipendenza

$$P(C) = (1-p)^{n-1}p$$

Sia D_k l'evento "si realizzano k successi e $n-k$ insuccessi".

Sia I una sequenza di k elementi distinti di $\{1, 2, \ldots, n\}$, e sia $\mathcal{I}_k$ l'insieme di tutte tali sequenze. Per un dato $I \in \mathcal{I}_k$ definiamo

$$E_I = \text{"la prova } i\text{-esima è un successo se e solo se } i \in I\text{"}$$

Si osservi che

$$E_I = \left(\bigcap_{i \in I} A_i\right) \cap \left(\bigcap_{j \notin I} A_j^c\right)$$

Segue allora dall'indipendenza che

$$P(E_I) = p^k(1-p)^{n-k}$$

Le osservazioni "chiave" sono:

$$E_I \cap E_J = \emptyset \qquad \text{per } I, J \in \mathcal{I}_k \text{ con } I \neq J$$

e

$$D_k = \bigcup_{I \in \mathcal{I}_k} E_I$$

Seque dalla proprietà (P2) della probabilità che

$$P(D_k) = \sum_{I \in \mathcal{I}_k} P(E_I) = |\mathcal{I}_k|\, p^k(1-p)^{n-k}.$$

Notando che $\mathcal{I}_k$ è l'insieme dei sottoinsiemi di k elementi di $\{1, 2, \ldots, n\}$, segue che $|\mathcal{I}_k| = \binom{n}{k}$, e perciò

$$P(D_k) = \binom{n}{k} p^k(1-p)^{n-k}$$

4 Variabili casuali

4.1 Variabili casuali discrete e continue, e loro distribuzioni

Nel Capitolo di Statistica Descrittiva abbiamo chiamato *variabile* una quantità numerica che vegna rilevata o misurata. Nella gran parte delle situazioni, le variabili che vengono effettivamente misurate e analizzate corrispondono a particolari caratteristiche di un *fenomeno più complesso*. Ad esempio

- In un'atleta che partecipa ai Giochi Olimpici, per la prova antidoping, tra tutte le caratteristiche ematiche vengono misurati il livello di ematocrito e la concentrazione di ormone della crescita.
- Per rilevare il tasso di inquinamento dell'aria nella Zona Industriale di Padova, della composizione dell'aria vengono misurati il livello di Diossina e la concentrazione di polveri sottili.
- Per verificare la preparazione acquisita durante la Laurea Triennale, alla fine del triennio uno o più studenti scelti a caso nelle Lauree Triennali di Biologia e Biologia Molecolare vengono sottoposti ad un *ulteriore esame di Statistica*, con relativa valutazione.

Con riferimento ad uno degli esempi precedenti, sia X una delle "quantità misurate". Tale quantità è variabile all'interno della popolazione in esame (per il primo e terzo esempio), nello spazio e nel tempo (nel secondo esempio).

Il massimo di informazione che è possibile fornire sta pertanto nella risposta a domande del tipo

- qual è la probabilità che in una futura misurazione il valore di X sia compreso tra a e b, dove $a < b$ sono due numeri reali assegnati?

Tali probabilità identificano la *distribuzione* della variabile casuale.

Lo scopo delle definizioni che seguono è di tradurre in termini rigorosi e operativi le precedenti considerazioni intuitive.

Sia S uno spazio campionario, con una fissata probabilità P. Indicheremo con $\mathbb{R}$ l'insieme dei numeri reali.

- Una funzione $X : S \to \mathbb{R}$ si dice *variabile casuale* o *variabile aleatoria*.

Ad una variabile casuale sono associati in modo "naturale" alcuni eventi. Se $A \subseteq \mathbb{R}$, possiamo considerare l'evento

$$\{s \in S : X(s) \in A\}$$

cioè l'insieme degli esiti s in corrispondenza dei quali il valore della variabile X è un elemento di A. Tale evento lo indicheremo in modo sintetico con

$$\{X \in A\}$$

e la sua probabilità con

$$P(X \in A)$$

In particolare, se $A = (a, b] = \{x : a < x \leq b\}$, con $a < b$, scriveremo

$$\{a < X \leq b\} \quad \text{e} \quad P(a < X \leq b)$$

e se $A = \{x\}$ scriveremo

$$\{X = x\} \quad \text{e} \quad P(X = x)$$

Consideriamo l'insieme

$$X(S) := \{X(s) : s \in S\}$$

di tutti i valori che assume la variabile casuale X. Se tale insieme è un insieme discreto

$$X(S) = \{x_1, x_2, \ldots, x_n, \ldots\}$$

diciamo che X è una variabile casuale discreta.

La funzione $p_X : \mathbb{R} \in [0, 1]$ definita da

$$p_X(x) = P(X = x)$$

si dice *funzione di massa* o *densità discreta* di X. Notare che $p_X(x) \neq 0$ solo se $x \in X(S)$.

Si noti che con la funzione di massa si può ricostruire l'intera distribuzione di X. Infatti, sia

$$X(S) = \{x_1, x_2, \ldots, x_n, \ldots\}$$

e sia $A \subseteq \mathbb{R}$. Essendo

$$\{X \in A\} = \bigcup_{i:x_i \in A} \{X = x_i\}$$

per l'additività delle probabilità abbiamo

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

In particolare

$$\sum_i p_X(x_i) = P(X \in \mathbb{R}) = 1.$$

Una variabile casuale può assumere valori in un insieme "continuo", ad esempio in *un intervallo di* $\mathbb{R}$.

Diremo che una variabile casuale X è una *variabile casuale continua* se esiste una funzione $f_X : \mathbb{R} \to [0, +\infty)$ tale che per ogni a, b con $-\infty \leq a \leq b \leq +\infty$ si ha

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

La funzione f_X viene chiamata *densità* della variabile casuale X.

Essa gode della proprietà

$$\int_{-\infty}^{+\infty} f_X(x)dx = P(X \in \mathbb{R}) = 1$$

Si noti che, se $x \in \mathbb{R}$

$$P(X = x) = P(x \leq X \leq x) = \int_x^x f_X(z)dz = 0$$

Ciò implica che

$$P(a < X \leq b) = P(a \leq X \leq b) - P(X = a) = P(a \leq X \leq b)$$

Lo stesso si può fare mettendo una disuguaglianza stretta nell'estremo b.

Sintetizzando: per una variabile continua ogni singolo valore ha probabilità 0 di essere ottenuto. Inoltre, se $\mathcal{I}$ è un intervallo di $\mathbb{R}$, allora $P(X \in \mathcal{I})$ non dipende dal fatto che $\mathcal{I}$ contenga o meno i suoi estremi.

Sia per variabili discrete che per variabili continue (soprattutto per queste ultime) è talvolta utile usare la seguente nozione.

- Sia X una variabile casuale. La funzione $F_X : \mathbb{R} \to [0, 1]$ definita da

  $$F_X(x) := P(X \leq x)$$

  è chiamata *funzione di ripartizione* o *distribuzione cumulativa* di X

Si noti che nel caso di variabili casuali continue

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz$$

e quindi, per il *Teorema fondamentale del Calcolo Integrale*,

$$F'_X(x) = f_X(x)$$

4.2 Valore atteso

- Sia X una variabile casuale discreta, con $X(S) = \{x_1, x_2, \ldots, x_n, \ldots\}$ e funzione di massa p_X. Il *valore atteso* (o *valor medio* o *media* di X) è denotato da $E(X)$ e definito da

  $$E(X) := \sum_i x_i p_X(x_i)$$

  (nel caso di somma infinita $E(X)$ si definisce solo nel caso tale somma abbia significato).

- Sia X una variabile casuale continua, con densità f_X. Il suo *valor atteso* è definito da
$$E(X) := \int_{-\infty}^{+\infty} x f_X(x) dx$$
(ammesso che tale integrale abbia senso).

Esempio

Un giocatore di roulette adotta la seguente strategia: punta sempre sul rosso, iniziando a puntare un euro, quindi raddoppiando la puntata fino al primo successo. Dopo il primo successo smette di giocare. In suo capitale iniziale è tale che dopo N sconfitte deve comunque smettere di giocare, per mancanza di capitale ulteriore. Sia X la differenza tra il capitale alla fine del gioco e il capitale iniziale. Calcoliamo la distribuzione e la media di X.

Supponiamo il giocatore perda nei primi $n-1$ giri di roulette, e abbia fatta l'n-esima puntata. Il totale delle puntate fatte è
$$1 + 2 + 2^2 + \cdots + 2^{n-2} + 2^{n-1} = \frac{2^n - 1}{2 - 1} = 2^n - 1$$
("ricordare" la relazione $1 + x + \cdots x^{n-1} = \frac{x^n - 1}{x - 1}$, valida per ogni $x \neq 1$).

Se il giocatore vince all'n-esima puntata, ottiene il doppio di quanto puntato, cioè $2 \cdot 2^{n-1} = 2^n$. Quindi rimane in attivo di
$$2^n - (2^n - 1) = 1.$$

Quindi, se il giocatore vince almeno una volta nei primi N tentativi, X assume sempre il valore 1.

L'unico caso in cui X assume un valore diverso, è nella sfortunata eventualità di N sconfitte consecutive. In questo caso il giocatore perde tutto quanto puntato, cioè
$$1 + 2 + 2^2 + \cdots + 2^{N-1} = 2^N - 1$$
In questo caso X assume il valore (negativo!) $1 - 2^N$.

Qual è la probabilità con cui ciò avviene, cioè quanto vale $P(X = 1 - 2^N)$? Sia $p = \frac{18}{37}$ la probabilità di successo in un giro di ruota. La probabilità di N insuccessi su N prove è
$$(1 - p)^N$$

Considerando il fatto che X assume solo due valori, abbiamo completamente determinato la sua distribuzione: $X(S) = \{1, 1 - 2^N\}$,
$$p_X(1 - 2^N) = (1 - p)^N \qquad p_X(1) = 1 - p_X(1 - 2^N) = 1 - (1 - p)^N$$

Pertanto:
$$\begin{aligned} E(X) &= 1 \cdot p_X(1) + (1 - 2^N) p_X(1 - 2^N) \\ &= [1 - (1 - p)^N] + (1 - 2^N)(1 - p)^N \\ &= 1 - [2(1 - p)]^N = 1 - \left(\frac{38}{37}\right)^N \end{aligned}$$

Notare che per ogni valore di N tale valor medio è negativo, e decresce al crescere di N!

Proprietà del valor medio

- (Linearità) Siano X e Y due variabili casuali definite nello stesso spazio campionario, e siano $a, b, c \in \mathbb{R}$. Si definisce in modo naturale la variabile casuale $aX + bY + c$:

  $$(aX + bY + c)(s) = aX(s) + bY(s) + c$$

  Allora

  $$E(aX + bY + c) = aE(X) + bE(Y) + c$$

- Sia $X : S \to \mathbb{R}$ una variabile casuale, discreta o continua. Sia inoltre $g : \mathbb{R} \to \mathbb{R}$. È allora possibile definire la *composizione* di X e g

  $$g \circ X(s) := g(X(s))$$

  Dunque $g \circ X$, che noi denoteremo sempre con $g(X)$, è una funzione $S \to \mathbb{R}$, dunque è una variabile casuale.

  Conoscendo la funzione di massa di X (nel caso in cui X sia discreta) o la densità di X (nel caso in cui X sia continua), è possibile *calcolare* il valor medio di $g(X)$:

  - Se X è discreta

    $$E[g(X)] = \sum_i g(x_i) p_X(x_i)$$

  - Se X è continua

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

4.3 Varianza e deviazione standard

Sia X una variabile casuale, e sia $\mu := E(X)$. La *Varianza* di X si denota con $Var(X)$ ed è definita da

$$Var(X) := E\left[(X - \mu)^2\right]$$

Scrivendo $(X - \mu)^2 = X^2 - 2\mu X + \mu^2$ e usando la linearità vista prima, si vede che

$$Var(X) = E\left(X^2\right) - \mu^2$$

- Se X è discreta

  $$Var(X) = \sum_i (x_i - \mu)^2 p_X(x_i) = \sum_i x_i^2 p_X(x_i) - \mu^2$$

- Se X è continua

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx = \int_{-\infty}^{+\infty} x^2 f_X(x)dx - \mu^2$$

Notare che $Var(X) \geq 0$ e $Var(X) = 0$ se e solo se X assume un unico valore.

La varianza di X è un indice di quanto la distribuzione di X sia *dispersa* attorno alla media $E(X)$. In particolare, posto $E(X) = \mu$, si può dimostrare che, per ogni $c > 0$

$$P(|X - \mu| \geq c) \leq \frac{Var(X)}{c^2}$$

(disuguaglianza di Chebischev).

In altre parole, per variabili casuali con varianza piccola, è piccola la probabilità di assumere valori "distanti" dalla media.

Notare che che $a, b \in \mathbb{R}$

$$Var(aX + b) = a^2 Var(X)$$

Se $Var(X) = \sigma^2$, allora $\sigma = \sqrt{Var(X)}$ è chiamata *deviazione standard* della variabile casuale X.

Notare che se X è una variabile casuale con media μ e varianza σ^2, allora

$$Y := \frac{X - \mu}{\sigma}$$

ha media 0 e varianza 1. Una variabile Y con queste proprietà viene detta *standardizzata*

4.4 Indipendenza di variabili casuali

- Siano $X_1, X_2, \ldots, X_n$ variabili casuali definite nello stesso spazio campionario. Diciamo che esse sono indipendenti se per ogni scelta di $A_1, A_2, \ldots, A_n \subseteq \mathbb{R}$ gli eventi

$$\{X_1 \in A_1\},\ \{X_2 \in A_2\}, \ldots,\ \{X_n \in A_n\}$$

sono eventi indipendenti.

Una proprietà, rilevante in Statistica Inferenziale, delle variabili casuali indipendenti è la seguente

- Siano X e Y due variabili casuali indipendenti. Allora

$$Var(X + Y) = Var(X) + Var(Y)$$

Attenzione: tale "additività" della varianza NON è vera in generale.

4.5 Classi rilevanti di variabili casuali discrete

Le variabili binomiali

Consideriamo uno schema di n prove ripetute e indipendenti, con probabilità di successo p. Sia X il numero di successi. Chiaramente

$$X(S) = \{0, 1, \ldots, n\}$$

Inoltre già sappiamo che, se $k = 0, 1, \ldots, n$

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

Una variabile casuale con questa distribuzione viene detta Binomiale di parametri n e p, e scriveremo

$$X \sim B(n, p)$$

Vedrete nelle esercitazioni che, se $X \sim B(n, p)$

$$E(X) = np \qquad Var(X) = np(1-p)$$

Questo è in realtà una facile conseguenza del fatto che, posto

$$X_i = \begin{cases} 1 & \text{se l'}i\text{-esima prova è un successo} \\ 0 & \text{altrimenti} \end{cases}$$

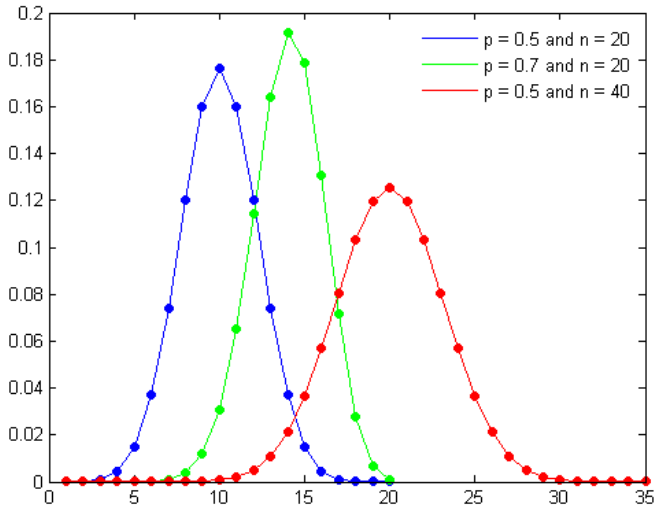
le variabili $X_1, X_2, \ldots, X_n \sim B(1, p)$, sono indipendenti e

$$X = X_1 + X_2 + \cdots + X_n$$

Un'ulteriore conseguenza di ciò è che se $X \sim B(n, p)$ e $Y \sim B(m, p)$ sono indipendenti, allora

$$X + Y \sim B(n + m, p)$$

Funzione di massa di una variabile binomiale



Le variabili geometriche

Si consideri uno schema di prove ripetute e indipendenti, che assumiamo di poter ripetere un numero arbitrario di volte. Sia X il numero di prove necessarie ad ottenere il primo successo. In questo caso

$$X(S) = \mathbb{N} \setminus \{0\} = \text{ insieme dei numeri naturali} > 0 = \{1, 2, 3, \ldots\}$$

e l'evento $\{X = n\}$ = "le prime $n-1$ prove sono insuccessi mentre l'n-esima è un successo". Pertanto

$$p_X(n) = P(X = n) = p(1-p)^{n-1}$$

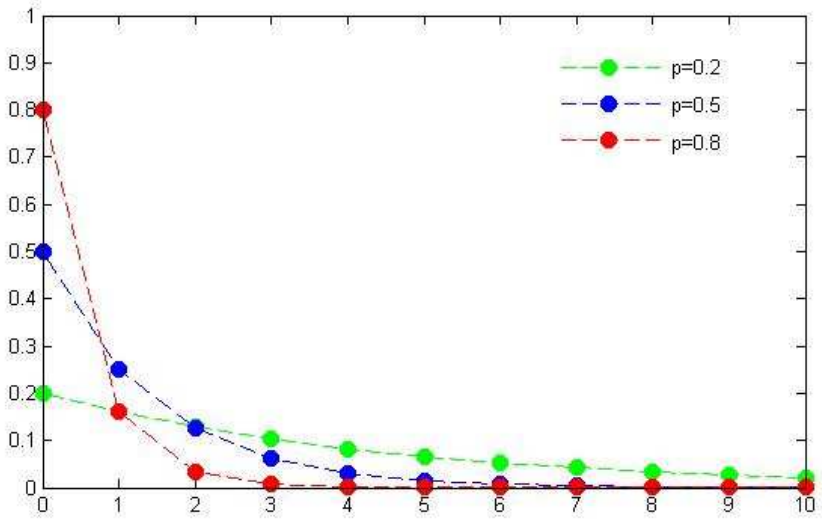
Una variabile con questa distribuzione viene chiamata Geometrica di parametro p, e scriveremo

$$X \sim Ge(p)$$

Si può dimostrare che se $X \sim Ge(p)$ allora

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{1-p}{p^2}$$

Funzione di massa di una variabile geometrica



Le variabili di Poisson

Le variabili di Poisson, che assumono valori in $\mathbb{N}$, vengono usate per modellare il numero di volte in cui verifica un determinato accadimento "aleatorio" in una determinata regione di tempo o di spazio: il numero di ricoveri in un anno per una determinata malattia, il numero di guasti in un'automobile nei suoi primi 100000 km, il numero di eruzioni in un secolo di un certo vulcano, il numero di fiori in un ramo di pesco,.....

Le variabili di Poisson sono caratterizzate da una funzione di massa della forma

$$P(X = n) = p_X(n) = e^{-\lambda}\frac{\lambda^n}{n!}$$

dove $\lambda > 0$ è un parametro. Una variabile con questa distribuzione viene chiamata di Posson di parametro λ, e scriveremo

$$X \sim Po(\lambda)$$

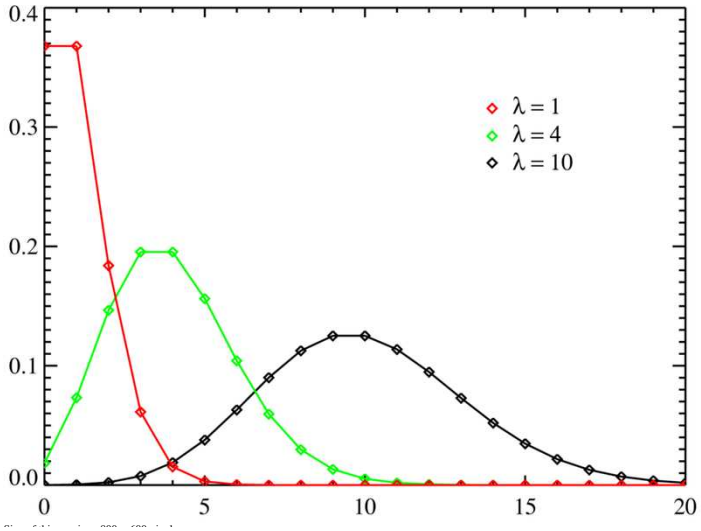
Si può dimostrare che se $X \sim Po(\lambda)$ allora

$$E(X) = \lambda \qquad Var(X) = \lambda$$

Una ulteriore proprietà fondamentale delle variabili di Poisson è la seguente: se X e Y sono due variabili casuali *indipendenti*, $X \sim Po(\lambda)$ e $Y \sim Po(\mu)$, allora

$$X + Y \sim Po(\lambda + \mu)$$

Funzione di massa di una variabile di Poisson



4.6 Classi rilevanti di variabili casuali continue

Classi rilevanti di variabili casuali continue

Le variabili normali o Gaussiane

Le variabili normali sono di gran lunga le più usate nella modellistica dei fenomeni aleatori, per ragioni che verranno illustrare nelle prossime lezioni. Si tratta di variabili casuali continue, con densità

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

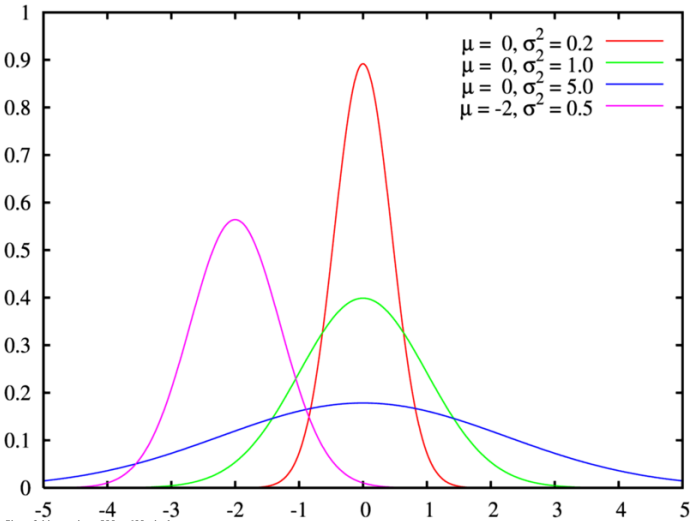
dove $\mu \in \mathbb{R}$ e $\sigma > 0$ sono due parametri. Si può mostrare che se una variabile X ha tale densità allora

$$E(X) = \mu \qquad Var(X) = \sigma^2$$

Diremo allora che X ha distribuzione normale di media μ e varianza σ^2, e scriveremo

$$X \sim N(\mu, \sigma^2)$$

Densità di una variabile Normale



Principali proprietà delle variabili Normali

- La trasformazione affine di una variabile Normale è ancora una variabile Normale: se $X \sim N(\mu, \sigma^2)$, allora

  $$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

  In particolare

  $$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

  Un variabile con distribuzione $N(0, 1)$ viene detta *normale standard*.

- Se $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ sono indipendenti, allora

  $$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

  cioè *la somma di Normali indipendenti è ancora Normale*.

Le variabili esponenziali

Queste variabili vengono usate per modellare *tempi di attesa* di eventi "imprevedibili": una catastrofe naturale, lo svilupparsi di una mutazione in una specie, il decadimento di un atomo radioattivo.....

Si tratta di variabili continue con densità

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{per } x \geq 0 \\ 0 & \text{per } x < 0 \end{cases}$$

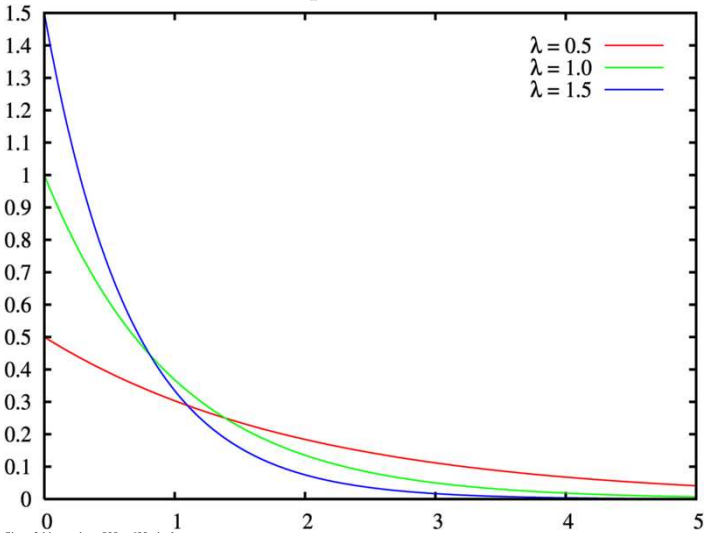
dove $\lambda > 0$ è un parametro. Una variabile con questa densità è detta *Esponenziale di parametro* λ, e scriveremo

$$X \sim Exp(\lambda)$$

Una semplice integrazione che vedrete nelle esercitazioni mostra che

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

Densità di una variabile esponenziale



5 Distribuzioni delle statistiche campionarie

5.1 Statistiche campionarie

Gran parte della statistica inferenziale riguarda l'analisi di dati relativi a misure ripetute. Abbiamo visto nel Capitolo precedente come la nozione di variabile casuale e di sua distribuzione costituiscano un modello matematico per descrivere la "variabilità" di una misura. Trattando misure ripetute è pertanto naturale considerare una sequenza $X_1, X_2, \ldots, X_n$ di n variabili casuali indipendenti e tutte con la stessa distribuzione (parleremo di variabili casuali *indipendenti e identicamente distribuite*, in breve *i.i.d.*).

La sequenza $X_1, X_2, \ldots, X_n$ viene anche chiamata *campione aleatorio* di *taglia* n.

- Una variabile casuale che sia funzione del campione aleatorio, cioè del tipo $f(X_1, X_2, \ldots, X_n)$ si dice *statistica campionaria*.

Le statistiche campionarie più usate in statistica inferenziale sono le seguenti.

- La *media campionaria* $\overline{X}$ definita da

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

- La *varianza campionaria* S^2 definita da

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Supponiamo che il campione aleatorio $X_1, X_2, \ldots, X_n$ sia tale che

$$E(X_i) = \mu \quad \text{e} \quad Var(X_i) = \sigma^2$$

Il seguente enunciato illustra alcune semplici ma fondamendati proprietà della media e varianza campionarie.

- $$E(\overline{X}) = \mu$$
$$Var(\overline{X}) = \frac{\sigma^2}{n}$$
$$E(S^2) = \sigma^2$$

Nota. La validità dell'ultima delle precedenti uguaglianze giustifica il fattore $\frac{1}{n-1}$ nella definizione di varianza campionaria.

5.2 La legge dei grandi numeri

Le due identità appena enunciate

$$E(\overline{X}) = \mu \quad \text{e} \quad Var(\overline{X}) = \frac{\sigma^2}{n}$$

esprimono il fatto che, per n grande, la distribuzione di $\overline{X}$ è "concentrata" attorno alla media μ.

Per dare una versione "quantitativa" di tale affermazione, ricordiamo la *disuguaglianza di Chebischev*: se X è una variabile casuale con $E(X) = \mu$, allora

$$P(|X - \mu| \geq c) \leq \frac{Var(X)}{c^2}$$

Applicando tale disuguaglianza a $\overline{X}$ otteniamo

$$P(|\overline{X} - \mu| \geq c) \leq \frac{\sigma^2}{c^2 n}$$

Abbiamo pertanto ottenuto:

- *La legge dei grandi numeri*: considerato un campione aleatorio $X_1, X_2, \ldots, X_n$ con distribuzione di media μ, per ogni $c > 0$ la probabilità

  $$P(|\overline{X} - \mu| \geq c)$$

  tende a 0 per $n \to +\infty$.

La legge dei grandi numeri, oltre a fornire un legame profondo tra media campionaria e media probabilistica, collega in modo analogo le distribuzioni di frequenza alle distribuzioni in senso probabilistico.

- Sia $X_1, X_2, \ldots, X_n$ un campione aleatorio con distribuzione discreta avente funzione di massa p. Sia k uno dei valori per cui $p(k) > 0$. Denotiamo con $F_k^{(n)}$ la frequenza relativa del valore k nel campione, cioè

  $$F_k^{(n)} := \frac{|\{i = 1, 2, \ldots, n \text{ tali che } X_i = k\}|}{n}$$

  Allora per ogni $c > 0$

  $$P(|F_k^{(n)} - p(k)| \geq c) \to 0$$

  per $n \to +\infty$.

Il risultato precedente segue dalla seguente osservazione: definito

$$Y_i := \begin{cases} 1 & \text{se } X_i = k \\ 0 & \text{altrimenti} \end{cases}$$

si ha che $Y_i \sim B(1, p(k))$, e

$$F_k^{(n)} = \overline{Y}$$

Basta perciò applicare la legge dei grandi numeri al campione $Y_1, Y_2, \ldots, Y_n$.

5.3 Teorema del limite centrale e approssimazione normale

Buona parte della statistica inferenziale è basata sulla conoscenza delle distribuzioni delle statistiche campionarie.

Determinare le distribuzioni della media e della varianza campionarie per un campione con distribuzione assegnata è di solito assai difficile. Uno dei pochi casi facili, è quello in cui si cerca la distribuzione della media campionaria per un *campione normale*, cioè un campione avente distribuzione normale.

Sia $X_1, X_2, \ldots, X_n$ un campione con distribuzione $N(\mu, \sigma^2)$. Poichè:

- la somma di normali indipendenti è normale
- la trasformazione affine di una normale è normale

è facile vedere che

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

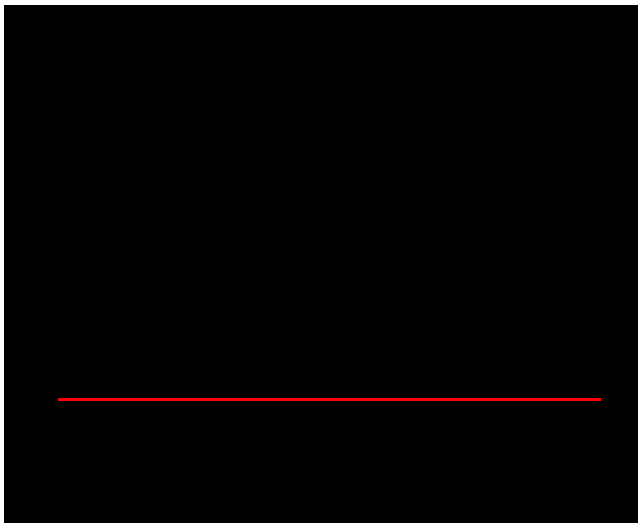
COSA POSSIAMO DIRE SE LA DISTRIBUZIONE DEL CAMPIONE NON È NORMALE?

I grafici che seguono si riferiscono alla densità della media campionaria di un campione, per diversi valori della taglia, avente distribuzione *Uniforme*: si tratta di una distribuzione continua con densità

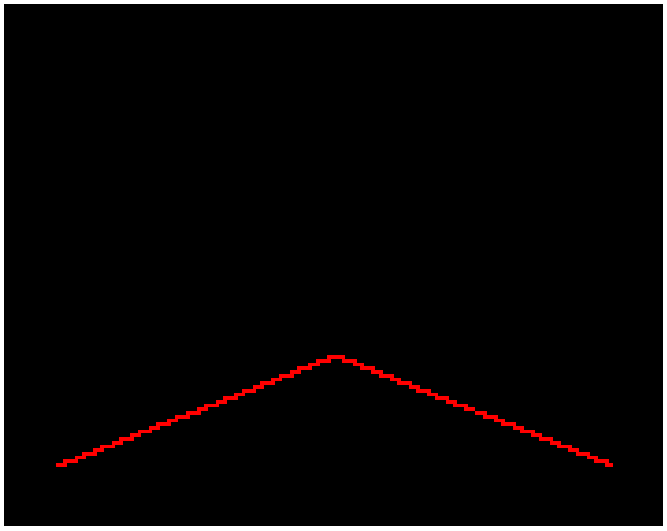
$$f(x) = \begin{cases} 1 & \text{se } x \in [0,1] \\ 0 & \text{altrimenti} \end{cases}$$

Si tenga presente che uan variabile casuale con questa distribuzione ha media $1/2$ e varianza $1/12$.

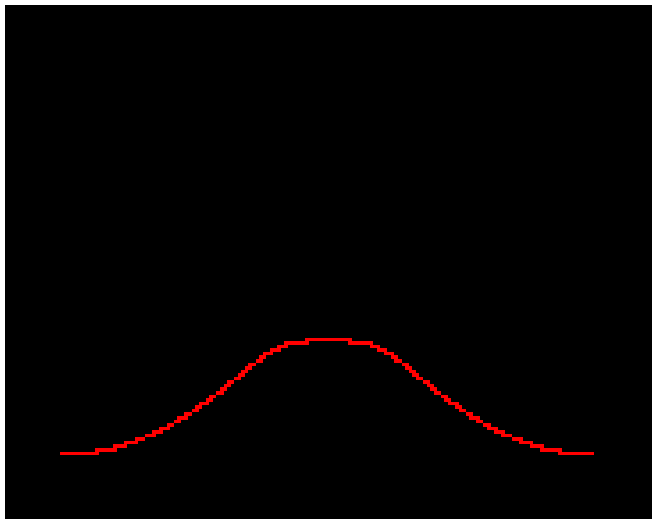
$$n = 1$$



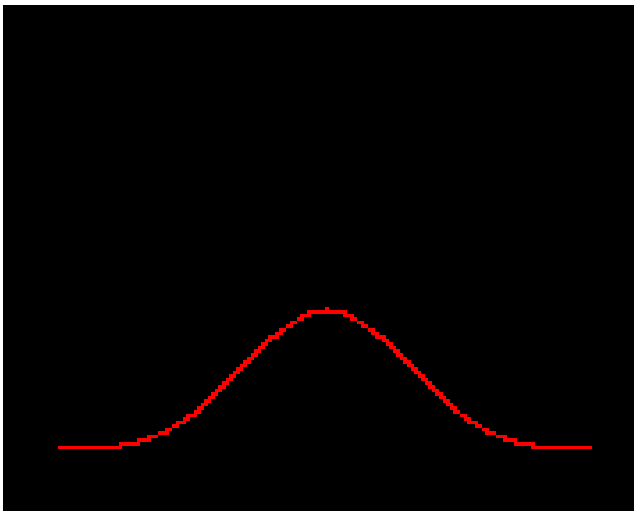
$n = 2$



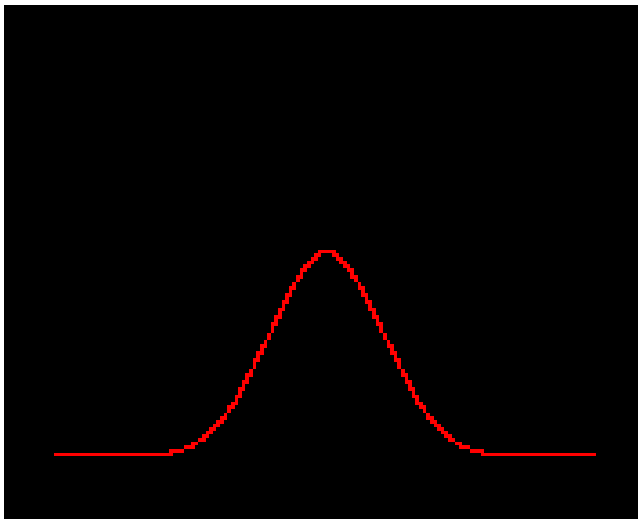
$n = 3$



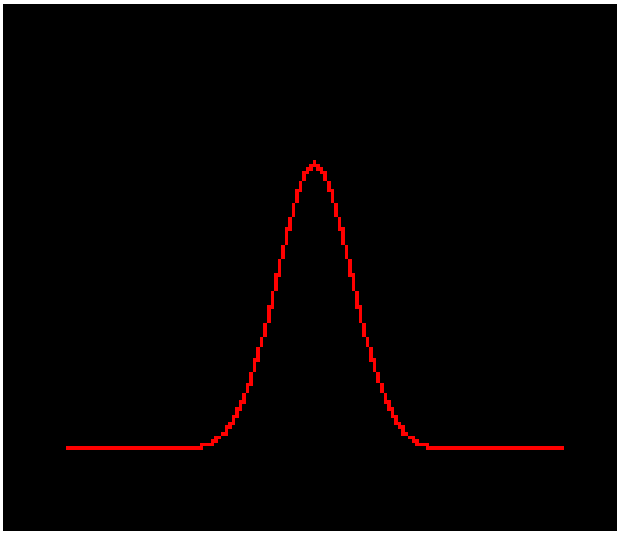
$n = 4$



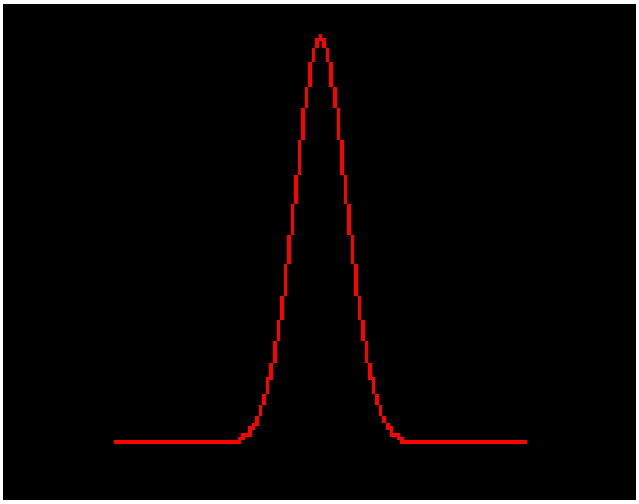
$n = 8$



$n = 16$



$n = 32$



Si può notare che la distribuzione, oltre a diventare sempre più "concentrata" attorno alla media, tende ad assumere una forma *"normale"* : se la densità di $\overline{X}$ con $n = 32$ venisse confrontata con quella di una normale avente *la stessa media e la stessa varianza*, i due grafici risulterebbero pressochè indistinguibili.

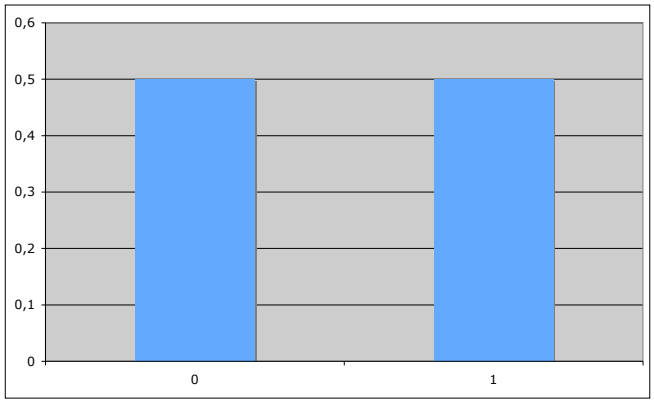
Questo fenomeno non ha nulla a che vedere con la distribuzione uniforme, o con il fatto che la distribuzione del campione sia continua.

Nei grafici che seguono viene rappresentata la funzione di massa di

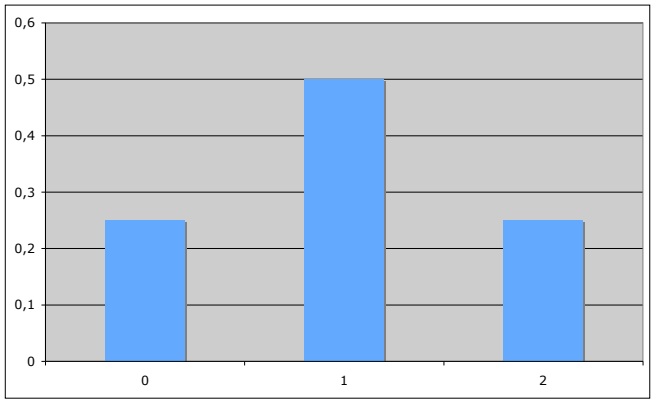
$$X_1 + X_2 + \cdots + X_n$$

dove il campione $X_1, X_2, \ldots, X_n$ ha distribuzione $B(1, 0.5)$. Si noti che la funzione di massa di $\overline{X}$ si ottiene da un semplice cambio di "unità di misura".

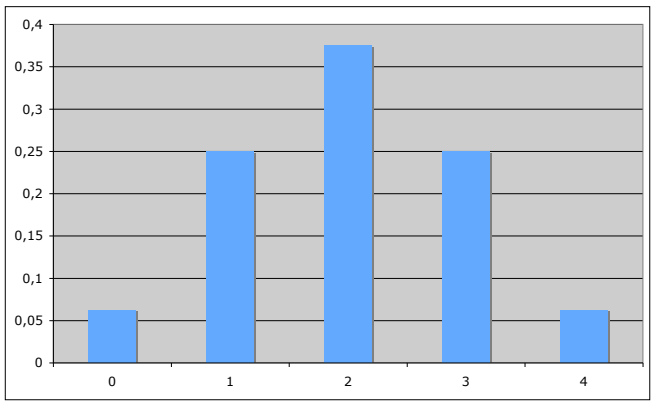
$n = 1$



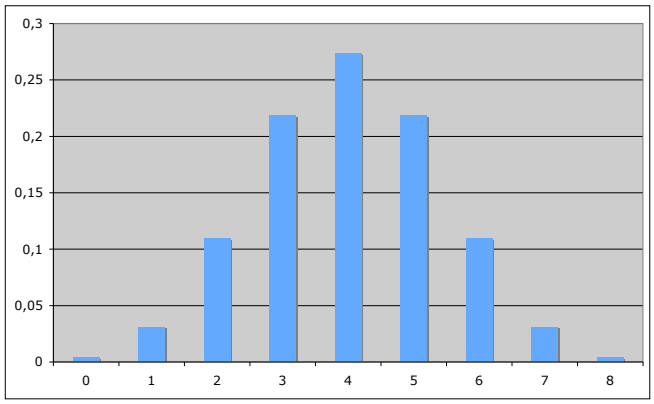
$n = 2$



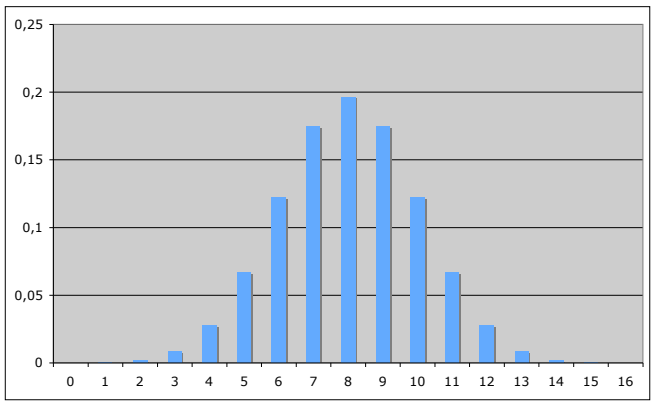
$n = 4$


0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
0
1
2
3
4

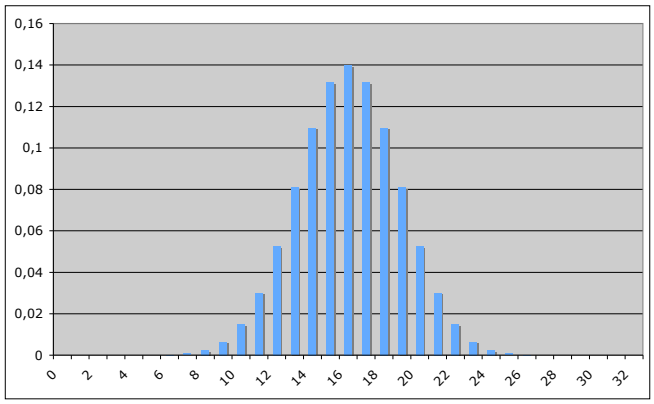
$n = 8$


0,3
0,25
0,2
0,15
0,1
0,05
0
0
1
2
3
4
5
6
7
8

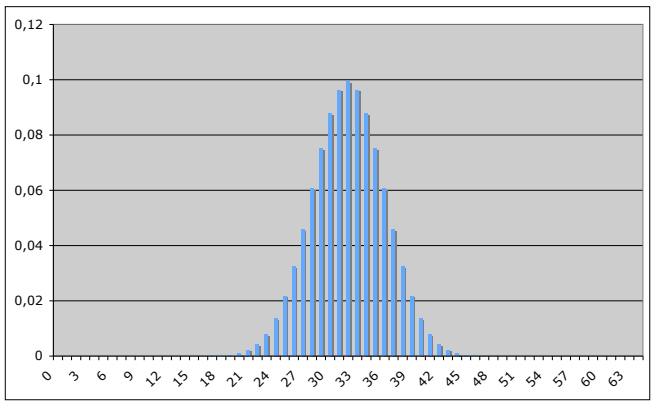
$n = 16$


0,25
0,2
0,15
0,1
0,05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

$n = 32$


0,16
0,14
0,12
0,1
0,08
0,06
0,04
0,02
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32

$n = 64$



Anche in questo caso, la distribuzione ottenuta per $n = 64$ viene "interpolata" in modo pressochè perfetto dalla densità di una normale *con la stessa media e la stessa varianza*

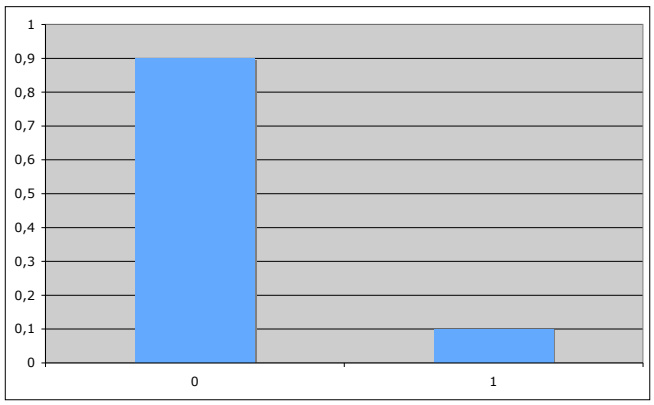
Nei due esempi precedenti, il campione ha una distribuzione "simmetrica rispetto alla media".

Per vedere cosa succede in un caso in cui tale simmetria è assente, consideriamo la funzione di massa di

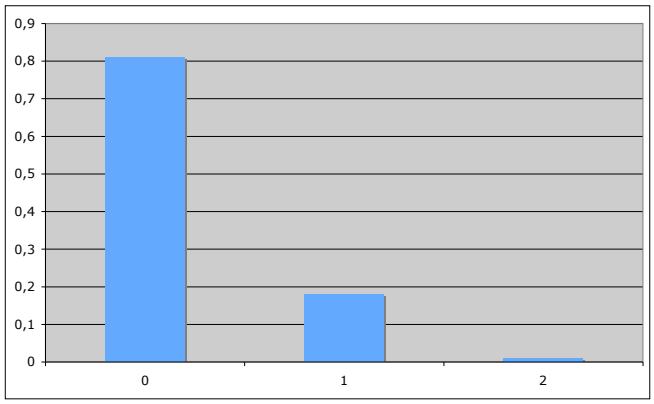
$$X_1 + X_2 + \cdots + X_n$$

dove il campione $X_1, X_2, \ldots, X_n$ ha distribuzione $B(1, 0.1)$.

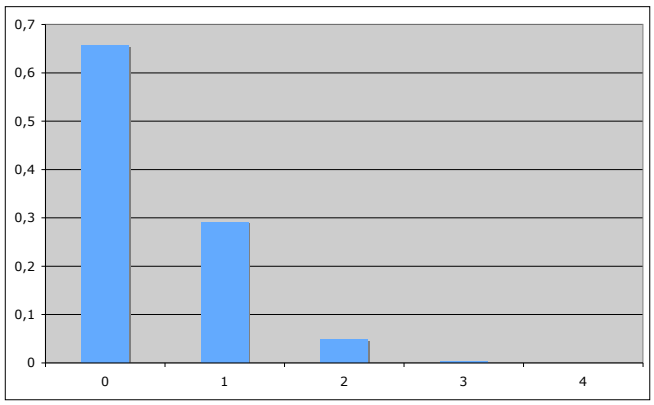
$$n = 1$$



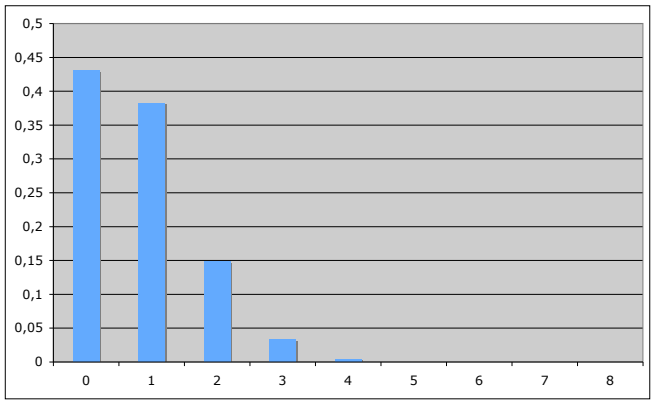
$n = 2$



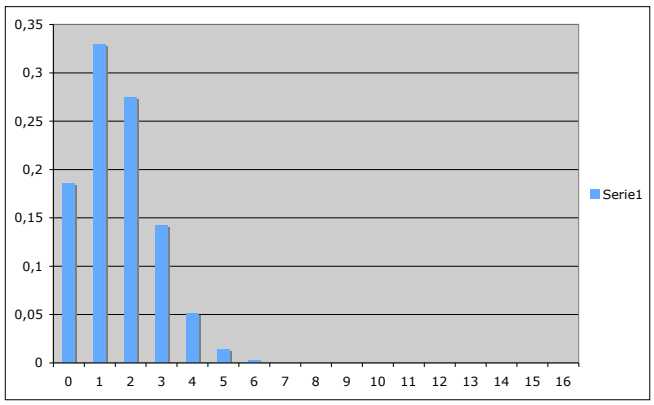
$n = 4$


0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
1
2
3
4

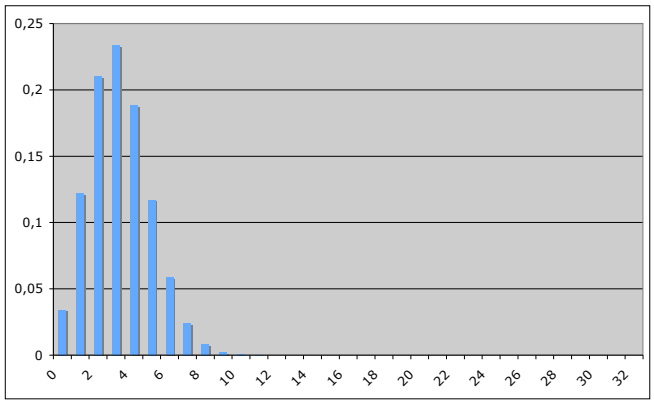
$n = 8$


0,5
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
0
1
2
3
4
5
6
7
8

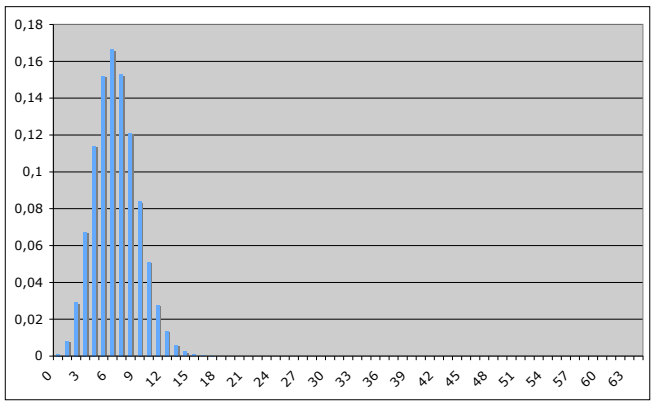
$n = 16$


0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Serie1

$n = 32$


0,25
0,2
0,15
0,1
0,05
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32

$n = 64$



Quello che si può notare è che la convergenza verso una distribuzione normale avviene anche nel caso asimmetrico, ma in modo *più lento*

Per avere un confronto accurato tra la distribuzione di una variabile casuale e quella di una normale con le stesse media e varianza, è opportuno *standardizzare* le variabili.

Sia $X_1, X_2, \ldots, X_n$ un campione aleatorio con distribuzione arbitraria, con media μ e varianza σ^2. Consideriamo la variabile casuale Z ottenuta dalla standardizzazione della media campionaria:

$$Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Il seguente risultato traduce in modo rigoroso le osservazioni sulla "approssimata normalità" fatte in precedenza.

- *TEOREMA DEL LIMITE CENTRALE*. Qualunque sia la distribuzione del campione $X_1, X_2, \ldots, X_n$, per ogni $x \in \mathbb{R}$

  $$P(Z \leq x) \to \Phi(x) \text{ per } n \to +\infty$$

  dove Φ è la funzione di ripartizione di una normale standard.

Per dirla in breve, il Teorema del limite centrale afferma che la statistica campionaria $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ha, per n grande, una distribuzione approssimativamente $N(0,1)$. Scriveremo:

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

Usando le tavole per Φ, questo risultato può essere usato per il calcolo approssimativo di probabilità di eventi esprimibili in termini della media campionaria.

Esempio

Si lanci 100 volte una moneta equlibrata. Qual è la probabilità che il numero di teste sia compreso tra 40 e 70 (inclusi gli estremi)?

Sia

$$X_i := \begin{cases} 1 & \text{se l'}i\text{-mo lancio dà come esito una testa} \\ 0 & \text{altrimenti} \end{cases}$$

L'evento in questione si può scrivere nella forma ($n = 100$)

$$\{40 \leq X_1 + \cdots + X_n \leq 70\}$$

Notando che $X_i \sim B(1, 1/2)$, perciò $\mu = 1/2$ e $\sigma^2 = 1/4$, e che Z si può anche scrivere come

$$Z = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}},$$

otteniamo

$$\begin{aligned} P(40 \leq X_1 + \cdots + X_n \leq 70) &= P\left(\frac{40-50}{5} \leq Z \leq \frac{70-50}{5}\right) \\ &= P(-2 \leq Z \leq 4) \simeq \Phi(4) - \Phi(-2) \end{aligned}$$

dove nell'ultimo passaggio abbiamo usato il Teorema del Limite Centrale. Perciò

$$\begin{aligned} P(40 \leq X_1 + \cdots + X_n \leq 70) &\simeq \Phi(4) - \Phi(-2) \\ &= \Phi(4) - [1 - \Phi(2)] = 0,977218197 \end{aligned}$$

Questo procedimento viene chiamato *approssimazione normale*. In questo caso, la distribuzione di $X_1 + \cdots + X_n$ è nota, essendo $B(n, 1/2)$. Con l'ausilio di una calcolatore, la probabilità cercata si può calcolare "quasi" esattamente:

$$P(40 \leq X_1 + \cdots + X_n \leq 70) = 0,98238382$$

Gran parte dell'errore è dovuto al fatto che $X_1 + \cdots + X_n$ ha distribuzione discreta che, con l'approssimazione normale, approssimiamo con una distribuzione continua.

Per migliorare l'approssimazione, osserviamo anzitutto che

$$P(40 \leq X_1 + \cdots + X_n \leq 70) = P(39 < X_1 + \cdots + X_n < 71)$$

Adottiamo la cosidetta *correzione di continuità*, che consiste nello scegliere un "via di mezzo" tra le precedenti alternative:

$$P(39,5 \leq X_1 + \cdots + X_n \leq 70,5)$$

Usando l'approssimazione normale per quest'ultima probabilità, si ottiene come risultato

$$0,982114922$$

che rappresenta un'eccellente approssimazione del valore "vero".

Per poter usare con fiducia il metodo dell'approssimazione normale, bisognerebbe sapere a priori: quanto grande dev'essere n affinchè l'approssimazione sia buona?

In generale è difficile dare una risposta "precisa" a questa domanda. Dipende da vari fattori, in particolare dalla *asimmetria* della distribuzione del campione.

Nella gran parte dei casi concreti $n \geq 30$ è una condizione sufficiente per avere una buona approssimazione.
Qualcosa di più preciso si può dire per campioni con distribuzione $B(1,p)$: l'approssimazione normale, corretta con la correzione di continuità, fornisce risultati buoni non appena

$$np \geq 5 \quad \text{e} \quad n(1-p) \geq 5$$

Si vede come in casi molto asimmetrici (p vicino a 0 o a 1), è necessario n più grande.

Un'interpretazione del Teorema del Limite Centrale

Un modo per esprimere il contenuto del Teorema del Limite Centrale è il seguente: se $X_1, X_2, \ldots, X_n$ sono variabili casuali i.i.d., allora

$$X_1 + X_2 + \ldots + X_n$$

ha distribuzione "approssimativamente" normale.

Questo risultato può essere dimostrato anche indebolendo l'ipotesi che le variabili siano i.i.d.: a parte casi "patologici", è sufficiente che le X_i siano indipendenti, ma non è necessario che abbiano le stessa distribuzione. In altre parole, *la somma di numerose variabili casuali indipendenti ha distribuzione approssimativamente normale*.

In molte situazioni concrete l'aleatorietà di una misura è la risultante di una somma di molti fattori indipendenti. PER QUESTO MOTIVO MOLTE VARIABILI MISURATE RISULTANO AVERE DISTRIBUZIONE GAUSSIANA, O APPROSSIMATIVAMENTE GAUSSIANA

5.4 Distribuzione delle statistiche campionarie di un campione normale

Sia $X_1, X_2, \ldots, X_n$ un campione normale di media μ e varianza σ^2. Abbiamo già osservato che

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

In alcuni problemi di statistica inferenziale che vedremo in seguito, è utile conoscere la distribuzione della varianza campionaria S^2. Cominciamo con la seguente definizione.

- Siano $Z_1, Z_2, \ldots, Z_n \sim N(0,1)$ indipendenti. La distribuzione della variabile casuale
$$Q := Z_1^2 + Z_2^2 + \cdots + Z_n^2$$
è denotata con χ^2_n, e chiamata *chi-quadro a n gradi di libertà*.

La variabile casuale Q è continua, ed è esplicitamente nota la sua densità. Tale densità non verrà da noi usata. Useremo invece degli altri valori "caratteristici" della distribuzione di Q.

Fissiamo $0 < \alpha < 1$. Non è difficile mostrare che esiste un unico numero reale z, necessariamente positivo, tale che

$$P(Q > z) = \alpha$$

Tale valore z viene indicato con $\chi^2_{n,\alpha}$ e chiamato *α-quantile della distribuzione χ^2_n*. Tali valori possono essere ricavati da tavole numeriche.

Il risultato fondamentale sulla distribuzione della varianza campionaria di un campione normale è:

- La variabile casuale
$$\frac{(n-1)S^2}{\sigma^2}$$
ha distribuzione χ^2_{n-1}.

Consideriamo due variabili casuali indipendenti $Z \sim N(0,1)$ e $Q \sim \chi^2_n$. La variabile casuale continua

$$T := \frac{Z}{\sqrt{Q/n}}$$

ha distribuzione nota, chiamata *t-di-Student a n gradi di libertà*, indicata con t_n.

Anche di tale distribuzione saranno utili i quantili: per $0 < \alpha < 1$, chiameremo *α-quantile della distribuzione t-di-Student a n gradi di libertà* quell'unico numero reale z tale che

$$P(T > z) = \alpha$$

Tale quantile viene indicato con $t_{n,\alpha}$, e può essere ricavato da tavole numeriche.

La distribuzione t-di-Student appare quando si considera la seguente statistica campionaria

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}-\mu}{\sigma\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}}$$

Si vede che tale statistica campionaria, se il campione è normale, è il rapporti tra una $N(0,1)$ e una χ^2_{n-1} divisa per $n-1$, che si può dimostrare sono indipendenti. Dunque

$$T \sim t_{n-1}$$

Per completare le notazioni che ci serviranno in seguito, indichiamo con z_α l'α-quantile della distribuzione normale standard, caratterizzato da

$$P(Z > z_\alpha) = \alpha$$

dove $Z \sim N(0,1)$, o equivalentemente

$$\Phi(z_\alpha) = 1 - \alpha$$

Si può mostrare che, per ogni fissato α, $t_{n,\alpha} \geq z_\alpha$, e

$$t_{n,\alpha} \to z_\alpha \quad \text{per} \quad n \to +\infty$$

In pratica $t_{n,\alpha}$ e z_α sono indistinguibili per $n \geq 200$.

6 Stima di media e varianza, e intervalli di confidenza

Lo scopo essenziale della Statistica inferenziale è quello di *ricavare da un campione di dati sperimentali, informazioni sulla distribuzione di una variabile*.

Ad esempio, nel caso della *Statistica parametrica*, si suppone che la distribuzione di una variabile sia nota *a meno di uno o più parametri incogniti*, ad esempio la media e/o la varianza. In questo caso l'obiettivo è di *stimare* questi parametri sulla base di dati sperimentali.

Stima della media per campioni normali: il caso di varianza nota

Sia $X_1, X_2, \ldots, X_n$ un campione con distribuzione $N(\mu, \sigma^2)$. Assumiamo che la varianza σ^2 sia nota, e che la media μ sia incognita. Lo scopo è quello di stimare μ sulla base di un campione di dati $x_1, x_2, \ldots, x_n$.

La legge dei grandi numeri stabilisce che la media campionaria $\overline{X}$ assume, con probbilità elevata se n è sufficientemente grande, un valore vicino a μ. È perciò naturale considerare

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

come *stima di* μ. La questione rilevante è tuttavia stabilire l'*affidabilità* di tale stima. A questo scopo osserviamo che

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

da cui segue che, fissato $0 < \alpha < 1$ "piccolo"

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) \\ &= \Phi(z_{\alpha/2}) - [1 - \Phi(z_{\alpha/2})] = 2\Phi(z_{\alpha/2}) - 1 = 1 - \alpha \end{aligned}$$

Perciò

$$\begin{aligned} 1 - \alpha = P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right)) \\ = P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) \end{aligned}$$

In altre parole, con probabilità $1-\alpha$, l'intervallo aleatorio di estremi

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

contiene μ con probabilità $1-\alpha$.

Sulla base di questo, diciamo che

- l'intervallo
$$\left[\overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]$$
è un *intervallo di confidenza per μ con livello di confidenza $1-\alpha$*.

L'intervallo di confidenza appena ottenuto è della forma $\overline{x} \pm e$, cioè è simmetrico rispetto a $\overline{x}$. Un tale intervallo viene detto *intervallo di confidenza bilatero*. La quantità $e = \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$ è detta *semiampiezza* dell'intervallo.

In alcuni casi può essere sufficiente fornire un intervallo di confidenza che abbia solo un limite superiore o un limite inferiore. A questo scopo osserviamo che, posto ancora

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

si ha

$$1 - \alpha = P(Z \leq z_\alpha) = P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_\alpha \leq \mu\right)$$

Pertanto l'intervallo

$$\left[\overline{x} - \frac{\sigma}{\sqrt{n}} z_\alpha, +\infty\right)$$

è detto *intervallo di confidenza unilatero destro per μ di livello di confidenza $1-\alpha$*.

In modo del tutto analogo, si mostra che

$$\left(-\infty, \overline{x} + \frac{\sigma}{\sqrt{n}} z_\alpha\right]$$

è un *intervallo di confidenza unilatero sinistro per μ di livello di confidenza $1-\alpha$*.

Esempio.
In un esperimento in un centro ricerche per l'agricoltura, si vuole verificare gli effetti dell'aggiunta di nitrato inorganico nella dieta di un gruppo di bovini, sulla quantità e qualità del latte prodotto. In particolare, si vuole stimare la produzione annuale di burro per individuo. Si può ritenere che la distribuzione di questa quantità sia normale; esperimenti precedenti hanno portato ad una valutazione di $\sigma^2 = 6400\ (lb/anno)^2$, che si assume non sia modificata in modo significativo dalla modifica della dieta.

I dati su 25 esemplari della popolazione di bovini in esame danno una media campionaria $\overline{x} = 465\ lb/anno$. Determiniamo un intervallo di confidenza bilatero per la media μ della distribuzione, al 95%, cioè $\alpha = 0.05$. Si ha $z_{\alpha/2} = 1.96$. Pertanto l'intervallo di confidenza richiesto è

$$465 - 1.96\left(\frac{80}{5}\right) \leq \mu \leq 465 + 1.96\left(\frac{80}{5}\right)$$

Stima della media per campioni normali: il caso di varianza ignota

In questo caso la statistica $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ non può essere usata, essendo σ incognito. Usiamo invece la statistica

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

L'argomento usato prima per la statistica Z, essendo basato solo sulla simmetria della distribuzione di Z (usata nella forma $\Phi(-x) = 1 - \Phi(x)$), può essere ripetuto, con la sola modifica di rimpiazzare i quantili della distribuzione normale standard con i quantili della distribuzione t-di-Student.

Pertanto si trovano gli intervalli di confidenza di livello di confidenza $1 - \alpha$:

- **Bilatero**
$$\left[\overline{x} - \frac{s}{\sqrt{n}}t_{n-1,\alpha/2}, \overline{x} + \frac{s}{\sqrt{n}}t_{n-1,\alpha/2}\right]$$
- **Unilatero destro**
$$\left[\overline{x} - \frac{s}{\sqrt{n}}t_{n-1,\alpha}, +\infty\right)$$
- **Unilatero sinistro**
$$\left(-\infty, \overline{x} + \frac{s}{\sqrt{n}}t_{n-1,\alpha}\right]$$

Esempio

In un esperimento rivolto a valutare alcuni benefici fisici della pratica della corsa, viene misurato il *massimo volume di assorbimento di ossigeno* (VO_2), la cui distribuzione si può assumere normale. In un gruppo di 25 "runners" sono stati raccolti i seguenti dati

$$\begin{aligned} \overline{x} &= 47.5\ ml/Kg \\ s &= 4.8\ ml/Kg \end{aligned}$$

Si ottiene un intervallo di confidenza al 95%

$$\overline{x} \pm \frac{s}{\sqrt{25}}t_{24,0.025} = 47.5 \pm \frac{4.8}{\sqrt{25}}2.064 = [45.5, 49.5]$$

Per confronto, nella stessa ricerca è stato misurato il VO_2 in 26 individui che non praticano la corsa, ottenendo

$$\begin{aligned} \overline{x} &= 37.5\ ml/Kg \\ s &= 5.1\ ml/Kg \end{aligned}$$

ottenendo un intervallo di confidenza al 95%

$$\overline{x} \pm \frac{s}{\sqrt{25}} t_{25,0.025} = [35.4, 39.6]$$

OSSERVAZIONE IMPORTANTE: le formule appena viste per gli intervalli di confidenza per la media con varianza incognita, vengono usate anche por campioni *non normali*, purchè la numerosità del campione sia sufficientemente elevata ($n \geq 30$). Esse forniscono un intervallo di livello di confidenza *approssimativamente* uguale ad $1-\alpha$. Questo è basato sul fatto che, per campioni anche non normali ma sufficientemente numerosi, la distribuzione della statistica T si discosta "poco" da quella di una t-di-Student.

Stima della varianza per campioni normali

Sia $X_1, X_2, \ldots, X_n$ un campione con distribuzione $N(\mu, \sigma^2)$, di cui assumiamo ignota tanto la media quanto la varianza. Vediamo come sia possibile determinare un intervallo di confidenza per la varianza σ^2 della distribuzione.

Il procedimento è basato sul fatto che

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Rispetto alla stima della media, dobbiamo qui tenere conto che

- la distribuzione χ^2 è "concentrata" sui reali positivi;
- la distribuzione χ^2 non è simmetrica.

Fissato $\alpha \in (0,1)$

$$P\left(\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\right) = 1-\alpha$$

Risolvendo per σ^2

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) = 1-\alpha$$

cioè

- l'intervallo
$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$
è un intervallo di confidenza bilatero di livello $1-\alpha$ per σ^2.

In modo del tutto analogo si determinano gli intervalli di confidenza unilateri.

- $$\left[0, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}}\right]$$
- $$\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha}}, +\infty\right)$$

Esempio

In una città è di grande rilevanza avere informazioni sulla distribuzione del consumo di energia elettrica per unità abitativa. Nel caso di unità abitative di metratura confrontabile, la varianza indica la variabilità nei livelli di efficienza energetica, un dato di interesse tanto per l'impresa erogatrice quanto per l'amministrazione locale.

In un campione di 101 unità abitative "omogenee" si è osservata una varianza campionaria

$$s^2 = 1.21 \text{ migliaia di } kWh^2$$

In questo caso può essere ragionevole essere interessati solo ad un "limite superiore" per la varianza, e quindi considerare l' intervallo unilatero

$$\left[0, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}}\right]$$

Usando i dati e le tavole, scelto $\alpha = 0.05$

$$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}} = \frac{100(1.21)}{77.929} = 1.5527$$

Possiamo perciò affermare, con una *confidenza* del 95%, che la varianza della distribuzione è inferiore a 1.5527.

Stima di una proporzione, cioè il parametro di una distribuzione $B(1,p)$

Consideriamo un campione aleatorio $X_1, X_2, \ldots, X_n$ con distribuzione $B(1,p)$, e consideriamo il problema di stimare il parametro p, sulla base di un campione di dati $x_1, x_2, \ldots, x_n$. Essendo p la media della distribuzione $B(1,p)$, anche in questo caso $\overline{x}$ *è la stima per* p.

Per ottenere un intervallo di confidenza, possiamo considerare la statistica

$$\tilde{Z} := \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

che, per il Teorema del limite centrale, ha distribuzione *approssimativamente* $N(0,1)$ se $np \geq 5$ e $n(1-p) \geq 5$.

Dall'uguaglianza

$$P(-z_{\alpha/2} \leq \tilde{Z} \leq z_{\alpha/2}) \simeq 1 - \alpha$$

si può "evidenziare" p, e ricavare un *intervallo di confidenza approssimato*.

Tuttavia

- Non conoscendo p, le condizioni $np \geq 5$ e $n(1-p) \geq 5$ non possono essere verificate.
- Isolare p da $-z_{\alpha/2} \leq \tilde{Z} \leq z_{\alpha/2}$ conduce a disequazioni di secondo grado, con formule risolutive un po' complicate

Per risolvere queste complicazioni conviene modificare la statistica $\tilde{Z}$, sostituendola con

$$Z := \frac{\overline{X} - p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}}$$

che si può dimostrare avere distribuzione approssimativamente $N(0,1)$ per n abbastanza grande. Questa approssimazione è da considerarsi buona se

$$n\overline{X} \geq 5 \quad \text{e} \quad n(1-\overline{X}) \geq 5$$

Pertanto

$$1 - \alpha \simeq P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

da cui si ricava, sotto l'ipotesi $n\overline{x} \geq 5$ e $n(1-\overline{x}) \geq 5$,

- Intervallo di confidenza bilatero per p di livello di confidenza *approssimativamente* $1-\alpha$:

$$\overline{x} \pm \sqrt{\frac{\overline{x}(1-\overline{x})}{n}} z_{\alpha/2}$$

Notare che $\overline{x}(1-\overline{x}) \leq \frac{1}{4}$, da cui si ricava

$$e := \sqrt{\frac{\overline{x}(1-\overline{x})}{n}} z_{\alpha/2} \leq \frac{z_{\alpha/2}}{2\sqrt{n}}$$

È dunque possibile determinare a priori il numero n di osservazioni sufficienti ad avere la semiampiezza dell'intervallo di confidenza al di sotto di una soglia prefissata.

Come al solito si possono determinare gli intervalli di confidenza unilateri di livello di confidenza approssimativamente $1-\alpha$:

- $$\left(-\infty, \overline{x} + \sqrt{\frac{\overline{x}(1-\overline{x})}{n}} z_{\alpha}\right]$$

- $$\left[\overline{x} - \sqrt{\frac{\overline{x}(1-\overline{x})}{n}} z_\alpha, +\infty\right)$$

Esempio

Una ricerca vuole verificare l'incidenza del melanoma in donne di età compresa tra i 45 e i 54 anni. In un gruppo, selezionato casualmente, di 5000 donne, 28 hanno la malattia. Determiniamo un intervallo di confidenza per la percentuale di donne che hanno la malattia.

Il campione da considerare è

$$X_i = \begin{cases} 1 & \text{se l'}i\text{-ma donna ha la malattia} \\ 0 & \text{altrimenti} \end{cases}$$

I dati forniscono

$$\overline{x} = \frac{28}{5000} = 0.0056$$

Notare che $n\overline{x} = 28 > 5$ (così come ovviamente, $n(1-\overline{x}) > 5$), pertanto è lecito considerare l'intervallo di confidenza approssimato al 95%

$$\overline{x} \pm \sqrt{\frac{\overline{x}(1-\overline{x})}{5000}} z_{0.025} = 0.0056 \pm (0.0011)(1.96) = (0.0034, 0.0078)$$

7 Verifica di ipotesi

7.1 Nozioni generali

Nella gran parte delle applicazioni, la statistica inferenziale viene usata per verificare *ipotesi statistiche*, cioè delle *affermazioni sulla distribuzione della variabile in esame*. Nell'ambito della statistica parametrica, queste affermazioni si riferiscono ai parametri incogniti della distribuzione, ad esempio media e varianza. esempi di ipotesi statistiche sono

- La media della distribuzione è uguale a 2 (*ipotesi bilatera*)
- La media della distribuzione è minore di 10 (*ipotesi unilatera*)

Un'ipotesi statistica verrà denotata con H_0 e chiamata *ipotesi nulla*, mentre chiameremo *ipotesi alternativa* la sua negazione, denotata con H_1 (o H_a).

Lo scopo di una *verifica di ipotesi* è quello di determinare una regola che consenta, sulla base un campione di dati $x_1, x_2, \ldots, x_n$, di propendere per l'ipotesi nulla o quella alternativa.

Un *test di verifica di ipotesi* consiste nel determinare una regione C di valori del campione $x_1, x_2, \ldots, x_n$, detta *regione critica*, tale che

$$\text{se } (x_1, x_2, \ldots, x_n) \in C \quad \text{si rifiuta } H_0, \text{ e quindi si accetta } H_1$$

$$\text{se } (x_1, x_2, \ldots, x_n) \notin C \quad \text{si accetta } H_0$$

Due tipi di errori sono possibili.

- **Errore di prima specie**: rifiutare H_0 quando H_0 è vera.
- **Errore di seconda specie**: accettare H_0 quando H_0 è falsa.

Una regione critica "ideale" dovrebbe rendere "piccole" tanto la probabilità di commettere un errore di prima specie, quanto la probabilità di commettere un errore di seconda specie.

Questo spesso non è possibile: restringendo la regione critica la probabilità di commettere un errore di prima specie diminuisce, ma può aumentare quella di commettere un errore di seconda specie. Il contrario accade allargando la regione critica.

La scelta usuale nella teoria della verifica di ipotesi è di tenere "sotto controllo" la probabilità di errore di prima specie, a scapito, eventualmente, della probabilità di errore di seconda specie.

- Diciamo che un test per la verifica dell'ipotesi H_0 con regione critica C ha livello di significatività α se *per ogni distribuzione del campione* $X_1, X_2, \ldots, X_n$ *che soddisfi* H_0 si ha

$$P((X_1, X_2, \ldots, X_n) \in C) \leq \alpha$$

La scelta di privilegiare il controllo dell'errore di prima specie rende *asimmetrici* i ruoli dell'ipotesi nulla e dell'ipotesi alternativa. Consideriamo un test per la verifica dell'ipotesi H_0 con regione critica C e livello di significatifità $\alpha \ll 1$

- Se $(x_1, x_2, \ldots, x_n) \in C$, cioè si rifiuta H_0 (o equivalentemente, si accetta H_1), allora possiamo concludere che i dati sperimentali sono in *contraddizione significativa* con l'ipotesi H_0.
- Se $(x_1, x_2, \ldots, x_n) \notin C$, cioè si accetta H_0, possiamo soltanto concludere che i dati sperimentali *non sono in contraddizione significativa* con l'ipotesi H_0: questo *non significa affatto che essi siano in contraddizione con* H_1, *ma soltanto che essi non escludono in modo significativo che* H_0 *sia vera*

Questa asimmetria ha una rilevante implicazione: uno sperimentatore che desideri "dimostrare" con dati sperimentali una certa ipotesi sulla distribuzione di una variabile, adotterà l'ipotesi da dimostrare come *ipotesi alternativa*

7.2 Test per un campione normale

z-test su una media di un campione normale con varianza nota

Sia $X_1, X_2, \ldots, X_n$ un campione con distribuzione $N(\mu, \sigma^2)$, dove σ^2 è nota. Consideriamo il problema di verificare l'ipotesi *bilatera*

$$H_0 : \mu = \mu_0$$

dove μ_0 è un valore assegnato.

Questo test è basato su una particolare statistica campionaria, detta *statistica-test*,

$$Z := \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

L'osservazione chiave è la seguente: *se H_0 è vera, allora $Z \sim N(0,1)$*. Pertanto, se H_0 è vera

$$P(|Z| > z_{\alpha/2}) = \alpha$$

In altre parole, la scelta della regione critica

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2} \right\}$$

individua un test per la verifica di H_0 con livello di significatività α.

OSSERVAZIONE: notare che

$$(x_1, x_2, \ldots, x_n) \in C \text{ se e solo se } \mu_0 \notin \left(\overline{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

cioè *l'ipotesi $\mu = \mu_0$ viene rifiutata a livello di significatività α se e solo se μ_0 non appartiene all'intervallo di confidenza per μ di livello di confidenza $1 - \alpha$*

Un altro fatto importante, è che l'appartenenza di $(x_1, x_2, \ldots, x_n)$ alla regione critica, cioè

$$\left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

dipende dalla scelta del livello di significatività α.

Per continuità e monotonìa dei quantili della normale standard, esiste un unico $\overline{\alpha} \in (0,1)$ tale che

$$\left| \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \right| = z_{\overline{\alpha}/2}$$

e quindi

- se $\alpha > \overline{\alpha}$ allora H_0 viene rifiutata;
- se $\alpha \leq \overline{\alpha}$ allora H_0 viene accettata.

$\overline{\alpha}$ viene detto *p-value* (o *p-dei-dati*) del test.

Tanto più il p-value di un test è vicino a 0, tanto più i dati sono in contraddizione con l'ipotesi H_0.

Nel caso del test appena visto, il p-value è caratterizzato dall'uguaglianza

$$\left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right| = z_{\overline{\alpha}/2}$$

che è equivalente a

$$1-\frac{\overline{\alpha}}{2} = \Phi\left(\left|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right|\right)$$

che permette di calcolare esplicitamente $\overline{\alpha}$.

Per verificare ipotesi unilatere della forma $H_0 : \mu \leq \mu_0$ oppure $H_0 : \mu \geq \mu_0$, si ragiona in modo analogo, solo leggermente più complicato. Ci limitiamo a riportare i risultati.

- Per verificare $\mu \leq \mu_0$ a livello di significatività α si usa come regione critica

  $$C := \left\{(x_1, x_2, \ldots, x_n) \text{ tali che } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha\right\}$$

  e il p-value è dato da

  $$1-\overline{\alpha} = \Phi\left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

- Per verificare $\mu \geq \mu_0$ a livello di significatività α si usa come regione critica

  $$C := \left\{(x_1, x_2, \ldots, x_n) \text{ tali che } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha\right\}$$

  e il p-value è dato da

  $$\overline{\alpha} = \Phi\left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

Esempio

Per una variabile con distribuzione normale con media incognita e deviazione standard $\sigma = 2$, si raccoglie un campione di 10 dati, che forniscono $\overline{x} = 18.58$. Si verifichi l'ipotesi $H_0 : \mu = 20$ al 5%, e si calcoli quindi il p-value del test.

La regione critica è data da

$$z := \frac{\sqrt{10}}{2}|\overline{x}-20| > z_{0.025} = 1.96.$$

Essendo $z = 2.2452$, il campione cade nella regione critica, e quindi H_0 viene rifiutata.

Il p-value si ottiene da

$$1-\frac{\overline{\alpha}}{2} = \Phi(2.2452)$$

da cui si ottiene $\overline{\alpha} = 0.0248$.

t-test su una media di un campione normale con varianza ignota

Nel caso di campioni normali con media e varianza ignota, i precedenti argomenti vengono modificati in completa analogia con quanto fatto per gli intervalli di confidenza: in altre parole, alla statistica Z si sostituisce la statistica

$$T := \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

e ai quantili della normale standard i quantili della t_{n-1}.

Si ottengono pertanto le seguenti regioni critiche, a livello di significatività α

- $H_0 : \mu = \mu_0$:

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right| > t_{n-1,\alpha/2} \right\}$$

- $H_0 : \mu \leq \mu_0$:

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \frac{\overline{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha} \right\}$$

- $H_0 : \mu \geq \mu_0$:

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \frac{\overline{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha} \right\}$$

Esempio

I ragazzi di una determinata età hanno un peso medio di 42.5 Kg. In un sobborgo in cui si teme ci possano essere ragazzi malnutriti, viene misurato il peso di 25 ragazzi, ottenendo una media campionaria $\overline{x} = 40.47$ Kg, e una deviazione standard campionaria $s = 5.8$ Kg. Si assume la normalità della distribuzione della variabile in esame. Quale conclusione si può trarre?

Sia $H_0 : \mu \geq 42.5$, cioè $H_1 : \mu < 42.5$. H_0 viene rifiutata al 5% se

$$t := \frac{\overline{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,0.05}$$

In questo caso: $t = -1.75$, $t_{24,0.05} = 1.71$. Pertanto H_0 viene rifiutata: a questo livello di significatività si può concludere che i ragazzi del sobborgo siano, in media, malnutriti rispetto alla popolazione complessiva.

7.3 Test su una proporzione

Consideriamo un campione aleatorio $X_1, X_2, \ldots, X_n$ con distribuzione $B(1,p)$, dove p è incognito. Sulla base di un campione di dati $x_1, x_2, \ldots, x_n$, sottoponiamo a verifica l'ipotesi

$$H_0 : p = p_0$$

dove $p_0 \in (0,1)$ è un valore assegnato.

Consideriamo la *statistica test*

$$Z := \frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Se H_0 è vera, Z è *approssimativamente distribuita* come un $N(0,1)$, se $np_0 \geq 5$ e $n(1-p_0) \geq 5$. Ne segue che la regione

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \left| \frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| > z_{\alpha/2} \right\}$$

è una regione critica di un test per la verifica dell'ipotesi H_0 a livello di significatività α.

Inoltre il p-value $\overline{\alpha}$ del test è dato dall'identità

$$1 - \frac{\overline{\alpha}}{2} = \Phi\left(\left| \frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| \right)$$

In modo analogo si determinano le regioni critiche dei test per le ipotesi unilateri:

- $H_0 : p \leq p_0$.

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_\alpha \right\}$$

$$1 - \overline{\alpha} = \Phi\left(\frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right)$$

- $H_0 : p \geq p_0$.

$$C := \left\{ (x_1, x_2, \ldots, x_n) \text{ tali che } \frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -z_\alpha \right\}$$

$$\overline{\alpha} = \Phi\left(\frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right)$$

Esempio

Ognuno di noi ha un occhio dominante, che punta direttamente all'oggetto di interesse, mentre il secondo occhio si adatta al primo allo scopo di mettere a fuoco l'oggetto. Un gruppo di ricercatori sospetta che nei bambini con problemi di lettura l'occhio dominante prevalente sia quello destro. Allo scopo di verificare questa ipotesi viene esaminato un gruppo di 225 bambini con problemi di lettura, e di questi 144 hanno l'occhio destro dominante.

Sia p la percentuale di bambini con problemi di lettura che hanno l'occhio destro come occhio dominante. Sottoponendo a verifica l'ipotesi

$$H_0 : p \leq 1/2$$

si ottiene un valore della statistica test

$$z = \frac{\overline{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{144}{225} - 1/2}{\sqrt{\frac{1}{4 \cdot 225}}} \simeq 4.2$$

Pertanto il p-value è

$$\overline{\alpha} = 1 - \Phi(4.2) \simeq 0$$

I dati sono quindi nettamente a favore dell'ipotesi dei ricercatori.

7.4 Test su due campioni normali

Molti problemi si statistica applicata possono essere formulati in termini di *confronto* tra le distribuzioni di *due o più variabili*.

In questo corso tratteremo alcuni aspetti del caso di due variabili: i problemi con più di due variabili richiedono tecniche più avanzate.

Confronto di medie per dati appaiati

Considereremo ora un metodo largamente usato in statistica medica ed epidemiologia. Consideriamo una variabile misurata sugli individui di una popolazione. Il problema è verificare gli effetti si un "*trattamento*" su questa variabile. Sia $X_1, X_2, \ldots, X_n$ una campione aleatorio della variabile misurata *prima del trattamento* su n individui della popolazione, e sia $Y_1, Y_2, \ldots, Y_n$ il campione aleatorio della variabile misurata *dopo il trattamento* sugli stessi individui soggetti alla prima misurazione. Si assume che entrambi i campioni abbiano distribuzione normale, con medie rispettivamente μ_x e μ_y, entrambe incognite, come incognite sono anche le varianze .

Per verificare gli effetti del trattamento, possiamo sottoporre a verifica una delle seguenti ipotesi

- $H_0 : \mu_x = \mu_y$
- $H_0 : \mu_x \leq \mu_y$

- $H_0 : \mu_x \geq \mu_y$

A tale scopo consideriamo il campione aleatorio $D_1, D_2, \ldots, D_n$ relativo alla variabile che forisce la differenza tra i valori *prima del trattamento* e quelli *dopo il trattamento*:

$$D_i := X_i - Y_i$$

Assumiamo che anche quest'ultimo campione abbia distribuzione normale. La sua media, μ_d, è ovviamente data da

$$\mu_d = \mu_x - \mu_y$$

Si vede allora che l' ipotesi statistica $H_0 : \mu_x = \mu_y$ (risp. $H_0 : \mu_x \leq \mu_y$ o $H_0 : \mu_x \geq \mu_y$) si può riscrivere nella forma

$$H_0 : \mu_d = 0 \quad (\text{risp. } H_0 : \mu_d \leq 0 \text{ o } H_0 : \mu_d \geq 0)$$

Pertanto il problema si riduce a verificare un'ipotesi sulla media di *un* campione normale $D_1, D_2, \ldots, D_n$ con varianza incognita.

Esempio

Un'elevata concentrazione di zinco nell'acqua da bere, oltre ad alterarne il sapore, può provocare problemi di salute. Per migliorare le modalità di raccolta dell'acqua, si vuole stabilire se vi sia una differenza significativa nella concentrazione di zinco tra la superficie ed il fondo di un certo corso d'acqua. In 6 punti distinti viene misurata la concentrazione di zinco (in mg/L) al fondo e alla superficie, ottenendo i seguenti risultati:

fondo	superficie
0.430	0.415
0.266	0.238
0.567	0.390
0.531	0.410
0.707	0.605
0.716	0.609

Quali conclusioni si possono trarre?

Denotiamo con x_i i dati relativi al fondo, y_i quelli relativi alla superficie, e $d_i = x_i - y_i$. Si trova: $\overline{d} \simeq 0.0917$, $s_d \simeq 0.0607$. Essendo $\overline{d} \geq 0$, i dati suggeriscono che, in media, la concentrazione di zinco sul fondo sia più alta che in superficie. Per verificare se i dati a disposizione sia sufficienti a trarre questa conclusione, sottoponiamo a verifica l'ipotesi $H_0 : \mu_d \leq 0$.

La statistica test è

$$T := \frac{\overline{d}}{s_d/\sqrt{6}} \simeq 3.7$$

H_0 viene rifiutata al 5% se $T > t_{0.05,5} \simeq 2.015$. Pertanto H_0 viene rifiutata: la concentrazione di zinco in superficie è significativamente più bassa che sul fondo.

Confronto di media per campioni indipendenti

Nel problema precedente, le variabili casuali X_i e Y_i, riferendosi a misure relative al medesimo individuo, non possono essere considerate indipendenti.

Il problema del confronto delle medie si pone tuttavia anche nel caso in cui i due campioni aleatori siano *indipendenti*. Questo accade, ad esempio, quando i due campioni di dati sono ottenuti da due diversi gruppi di individui.

Consideriamo due campioni $X_1, X_2, \ldots, X_{n_x}$ e $Y_1, Y_2, \ldots, Y_{n_y}$, che *possono avere numerosità diversa*, entrambi con distribuzione normale, rispettivamente $N(\mu_x, \sigma_x^2)$ e $N(\mu_y, \sigma_y^2)$. Per quanto segue, assumeremo ignote sia le medie che le varianze, ma supporremo che esse *siano uguali*, cioè

$$\sigma_x = \sigma_y = \sigma$$

Tale assunzione, per certi aspetti innaturale, è essenziale per la determinazione di una statistica test, e verrà ulteriormente discussa tra poco. Supponiamo di voler sottoporre a verifica l'ipotesi

$$H_0 : \mu_x = \mu_y$$

Definiamo la *varianza campionaria combinata*:

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$$

Si può dimostrare che, *se H_0 è vera*, la statistica campionaria

$$T := \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

ha distribuzione $t_{n_x+n_y-2}$.

Pertanto

$$C := \left\{ (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_x}) : \left| \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \right| > t_{n_x+n_y-2,\alpha/2} \right\}$$

è una regione critica di un test per la verifica di H_0 a livello di significatività α.

Le regioni critiche, sempre a livello di significatività α, per ipotesi unilatere si ottengono in modo analogo

- $H_0 : \mu_x \leq \mu_y$:

$$C := \left\{ (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_x}) : \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} > t_{n_x+n_y-2,\alpha} \right\}$$

- $H_0 : \mu_x \geq \mu_y$:

$$C := \left\{ (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_x}) : \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} < -t_{n_x+n_y-2,\alpha} \right\}$$

Osservazione importante

Come sopra osservato, questo test è basato sull'assunzione che $\sigma_x = \sigma_y$, cosa che tipicamente non è nota a priori. In generale i risultati di questo test si possono ritenere affidabili se le corrispondenti *varianze campionarie* assumono valori "non troppo diversi".

Una condizione comunemente usata è

$$\frac{s_x^2}{s_y^2} \in \left(\frac{1}{2}, 2\right)$$

Esempio

Un problema molto studiato negli ultimi anni da equipes mediche è quello degli effetti del fumo passivo nei bambini. In quest'indagine è stato misurato il contenuto di *cotinina* (un metabolite della nicotina) nell'urine di due gruppi di bambini: il primo gruppo è costituito da bambini che vivono in una famiglia di non fumatori, mentre i bambini del secondo gruppo sono esposti in famiglia al fumo passivo.

non esposti ($n_x = 7$)	42	13	37	54	23	32	45		
esposti ($n_y = 9$)	67	70	53	42	37	74	53	57	64

Sottoponiamo a verifica l'ipotesi $H_0 : \mu_x = \mu_y$, cioè che non vi sia alcuna influenza del fumo passivo nella variabile misurata.

Anzitutto si calcola

$$s_x^2 \simeq 191.81 \quad s_y^2 \simeq 157.78$$

Perciò il rapporto s_x^2/s_y^2 è nell'intervallo di valori per cui il t-test sopra descritto può essere usato.

Un po' di calcoli portano a valutare la statistica test

$$t := \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = -3,370732381$$

L'ipotesi H_0 viene respinta a livello α se $|t| > t_{14,\alpha/2}$. Ad esempio $t_{14,0.005} = 2,976842734$, e pertanto H_0 viene respinta all'1%: in altre parole il p-value del test è inferiore all'1%. Possiamo quindi concludere che i dati mostrano una significativa influenza del fumo passivo nel livello di cotinina nell'urine.

8 Regressione lineare

In quanto appena visto abbiamo studiato alcuni aspetti del problema di *confrontare* le distribuzioni di due variabili. Un altro problema rilevante in statistica è quello della *dipendenza* tra due variabili. Ecco alcuni esempi di questioni che ricadono in questa classe di problemi:

- qual è la dipendenza tra il dosaggio di un farmaco antiipertensivo e la pressione arteriosa?
- qual è la dipendenza tra la temperatura media giornaliera e il consumo di energia per mq di unità abitativa?

In entrambi gli esempi vi sono *due* variabili in questione: x (dosaggio di un farmaco, temperatura media giornaliera) e y (pressione arteriosa, consumo di energia).

Mentre nel secondo esempio entrambe le variabili sono soggette ad "aleatorietà", nel primo esempio la variabile x è del tutto *deterministica*, cioè il suo valore è scelto da chi esegue l'esperimento.

Il metodo che descriveremo è del tutto insensibile al fatto che la variabile x sia aleatoria o deterministica: infatti l'oggetto dell'analisi statistica sarà la *distribuzione di y* per ogni *fissato valore della variabile x* .

La variabile x verrà chiamata *ingresso* o *predittore* o *input*, la variabile y *uscita* o *output*.

Nell'applicare il metodo della regressione lineare, si *assume* che il legame tra ingresso e uscita sia della forma

$$y = \alpha + \beta x + e$$

dove e è una variabile di "errore" con distribuzione normale di media 0.

In altre parole, se x_i è il valore della variabile di ingresso in una certa osservazione, la corrispondente variabile di uscita Y_i ha distribuzione

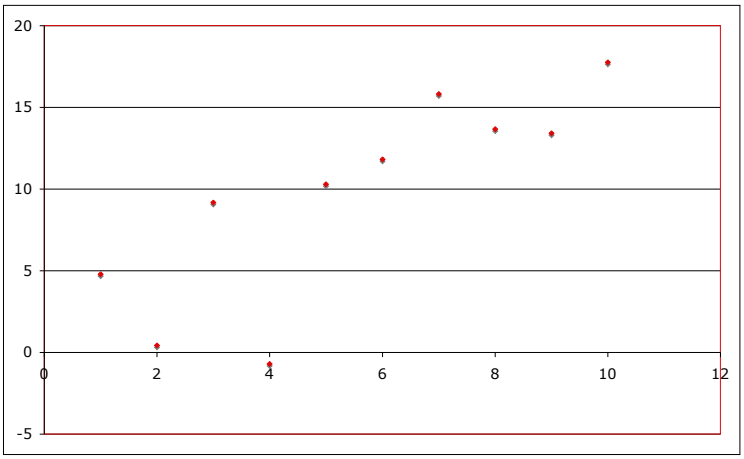
$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

e si assume inoltre che le variabili $Y_1, Y_2, \ldots, Y_n$, corrispondenti a n distinte osservazioni con valori $x_1, x_2, \ldots, x_n$ dell'ingresso, siano indipendenti.

Il problema presenta *tre parametri incogniti*: α, β e σ^2: sulla base dei valori $x_1, x_2, \ldots, x_n$ dell'ingresso e dei corrispondenti valori *osservati* $y_1, y_2, \ldots, y_n$ dell'uscita si vogliono ottenere informazioni sui parametri incogniti.

8.1 Inferenza statistica sui parametri della regressione

La via più intuitiva per giungere alla stima dei parametri α e β è quella geometrica. I valori $x_1, x_2, \ldots, x_n$ dell'ingresso e dei corrispondenti valori *osservati* $y_1, y_2, \ldots, y_n$ dell'uscita possono essere pensati come n coppie (x_i, y_i), rappresentabili su un piano cartesiano.



L'idea è ora quella di determinare la retta di equazione $y = a+bx$ che *meglio approssima* i punti del diagramma di dispersione. Per chiarire il significato di questa approssimazione, fissata una retta "candidata" $y = a + bx$, chiamiamo i valori $a + bx_i$ i *valori previsti* per l'uscita, e consideriamo gli *scarti quadratici* $(y_i - a - bx_i)^2$ tra il valore effettivo dell'uscita e quello previsto.

L'obiettivo è di determinare i coefficienti a e b della retta che rende *più piccolo possibile* lo *scarto quadratico totale*

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

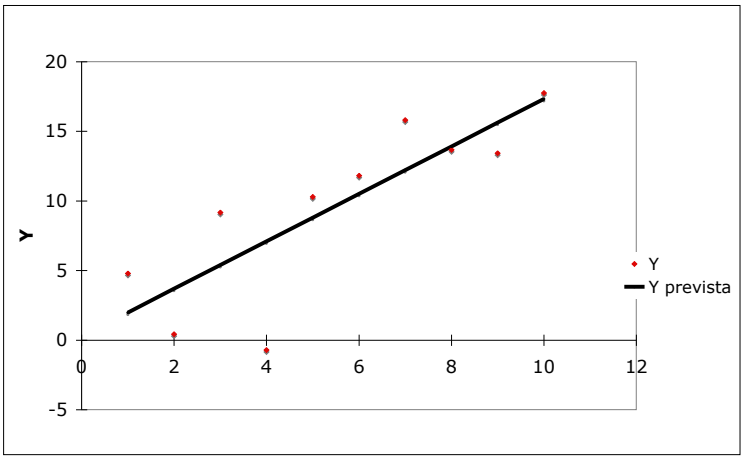
Si tratta di un problema di ottimizzazione abbastanza semplice, le cui soluzioni sono date da

$$b = \frac{\sum_i x_i y_i - n\overline{x}\,\overline{y}}{\sum_i x_i^2 - n\overline{x}^2}$$
$$a = \overline{y} - b\overline{x}$$

La retta di equazione

$$y = a + bx$$

è chiamata *retta di regressione di y su x*.



L'idea è ora quella di usare le statistiche campionarie a e b come *stime* dei parametri incogniti α e β.

Siano dunque $Y_1, Y_2, \ldots, Y_n$ variabili casuali indipendenti, con

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

e consideriamo le statistiche campionarie

$$B := \frac{\sum_i x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_i x_i^2 - n\overline{x}^2}$$

$$A = \overline{Y} - B\overline{x}$$

Posto

$$S_{xx} := \sum_i x_i^2 - n\overline{x}^2 = (n-1)s_x^2,$$

(dove s_x^2 è la varianza campionaria di $x_1, \ldots, x_n$), si dimostra abbastanza facilmente che

$$B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

$$A \sim N\left(\alpha, \frac{\sigma^2 \sum_i x_i}{nS_{xx}}\right)$$

Resta da determinare una stima per σ^2. A questo scopo consideriamo gli scarti

$$R_i := Y_i - A - Bx_i$$

che sono chiamati *residui*, e definiamo

$$SS_R := \sum_{i=1}^{n} R_i^2$$

Si può dimostrare che

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

Da cui segue che

$$E\left(\frac{SS_R}{n-2}\right) = \sigma^2$$

Il valore sui dati della statistica $SS_R/(n-2)$ viene pertanto usata come *stima* per σ^2.

Avendo determinato le distribuzioni di A, B e SS_R, non è difficile ottenere intervalli di confidenza per i parametri incogniti α, β e σ^2.

Notare che essendo $B \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$, si ha che

$$\frac{B-\beta}{\sigma/\sqrt{S_{xx}}} \sim N(0,1).$$

Rimpiazzando σ^2 con la statistica campionaria $SS_R/(n-2)$ che ne fornisce la stima, si può mostrare che

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2}$$

Questo permette di determinare facilmente un intervallo di confidenza bilatero per β di livello di confidenza $1-\alpha$:

$$B \pm \sqrt{\frac{SS_R}{(n-2)S_{xx}}} t_{n-2,\alpha/2}$$

Il fatto che

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2}$$

permette, con i metodi oramai usuali, di formulare dei test per la verifica di ipotesi sul parametro β. Per semplicità, non tratteremo ipotesi generali, ma solo quella più usata nel contesto della verifica di ipotesi:

$$H_0 : \beta = 0$$

Tale ipotesi esprime il fatto che la distribuzione dell'uscita Y *non* dipenda dall'ingresso x. Pertanto, se i dati sperimentali conducono al rifiuto di H_0, è

possibile concludere che la dipendenza tra ingresso ed uscita sia significativa. Lo scopo di molte applicazioni della regressione lineare è proprio di rilevare una tale dipendenza.

Da quanto appena visto,

$$C := \left\{ (y_1, y_2, \ldots, y_n) : \left| \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(b - \beta) \right| > t_{n-2,\alpha/2} \right\}$$

è una regione critica di un test per l'ipotesi $H_0 : \beta = 0$ con livello di significatività α.

Considerazioni analoghe permettono di determinare intervalli di confidenza e test per la verifica di ipotesi anche per i parametri α e σ^2. Per brevità, non le includeremo in queste lezioni.

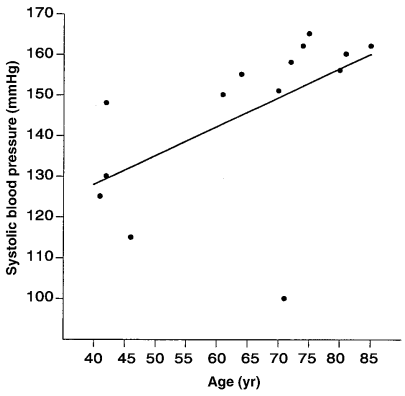
Esempio

La seguente tabella mostra i dati del peso (x, in anni) e della pressione arteriosa sistolica (y, in mmHg) per 15 donne.

x	y
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

Si vuole studiare la correlazione tra le due variabili.

Il diagramma di dispersione



mostra che una relazione di dipendenza lineare è ragionevole.

La stima dei parametri della regressione fornisce

$$b = 0.705 \qquad\qquad a = 99.96$$

cioè la retta di regressione è

$$y = 99.96 + 0.705x$$

L'intervallo di confidenza per β al 95% risulta

$$(0,0869, 1,3229)$$

mentre quello al 99% è

$$(-0,1568, 1,5666)$$

Ricordando il fatto che *l'ipotesi $H_0 : \beta = 0$ viene rifiutata a livello α se e solo se 0 non appartiene all'intervallo di confidenza per β di livello di confidenza $1 - \alpha$*, ne deduciamo che l'ipotesi H_0 viene rifiutata al 5% ma accettata all'1%.

Nella parte del corso dedicata alla statistica descrittiva, abbiamo introdotto la nozione di *coefficiente di correlazione campionario* per i dati relativi a due variabili:

$$r := \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}} = \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

Vogliamo ora dare un'interpretazione di r in termini del modello di regressione lineare. La quantità

$$SS_R = \sum_{i=}^{n} (y_i - a - bx_i)^2$$

è un indice della *variabilità intrinsica* delle uscite, cioè la loro tendenza a scostarsi dalla retta di regressione a causa del termine e di errore nel modello di regressione lineare

$$y = \alpha + \beta x + e$$

La quantità

$$S_{yy} := \sum_{i=1}^{n} (y_i - \overline{y})^2 = (n-1)s_y^2$$

è invece un indice della *variabilità totale* delle uscite.

Non è difficile dimostrare che $SS_R \leq S_{yy}$. L'indice

$$R^2 := 1 - \frac{SS_R}{S_{yy}}$$

rappresenta la *frazione di variabilità dovuta alla dipendenza tra le due variabili*. Non è difficile dimostrare che

- $r^2 = R^2$
- Il segno di r è uguale al segno di b

Nell'esempio precedente, che riguardava la correlazione tra età e pressione arteriosa, si trova

$$R^2 = 0.32$$

In altre parole, solo il 32% della variabilità di y può essere spiegata in termini della sua dipendenza da x, mentre il restante 68% è dovuta ad *aleatorietà intrinseca*.

9 Test non parametrici

Le ultime lezioni di questo corso saranno dedicate a due test che non ricadono, almeno nella loro formulazione più generale, nell'ambito della statistica parametrica. Si tratta di test che, previo un opportuno adattamento, possono essere applicate a variabili sia discrete che continue, ma anche a variabili categoriche, ossia a valori non numerici.

9.1 Test χ^2 di buon adattamento

I test di buon adattamento, in generale, hanno lo scopo di verificare se una variabile in esame abbia o meno un certa *distribuzione ipotizzata* sulla base, come al solito, di dati sperimentali. Ecco alcuni esempi in cui il problema del buon adattamento sorge in modo naturale.

- A volte la distribuzione di una variabile viene suggerita da considerazioni di tipo "fondamentale", di natura fisica, chimica o biologica. Ad esempio in *genetica* vengono usati vari modelli probabilistici per descrivere la

trasmissione dei caratteri, la selezione naturale dei caratteri favorevoli, lo sviluppo di mutazioni ..... Tutti questi modelli sono basati su ipotesi che in qualche modo semplificano la realtà. L'aderenza di questi modelli ai dati reali è un problema primario, e si pone proprio come problema di buon adattamento.

- Una buona parte dei metodi di inferenza statistica che abbiamo visto sono basati sull'assunzione che la distribuzione delle variabili sia normale. È possibile verificare la "ragionevolezza" di quest'assunzione sulla base dei dati? In altre parole, possiamo *verificare* l'assunzione di normalità come fosse un'*ipotesi statistica*?

Il caso di una variabile che assume un numeri finito di valori possibili

Consideriamo una variabile che può assumere solo m valori possibili: $1, 2, \ldots, m$. Questi valori non vanno visti come "veri" valori numerici, ma piuttosto come *etichette*. Non useremo *mai* le usuali statistiche campionarie per variabili numeriche (media e varianza campionarie) ma soltanto le *frequenze* che abbiamo introdotto in statistica descrittiva.

Sia $X_1, X_2, \ldots, X_n$ una campione aleatorio per tale variabile e, per $k = 1, 2, \ldots, m$ consideriamo le frequenze (assolute)

$$N_k := |\{i : X_i = k\}|$$

Sia $\pi = (\pi(1), \pi(2), \ldots, \pi(m))$ una funzione di massa, che consideriamo come *candidata* ad essere la distribuzione della variabile in esame.

Sulla base di dati sperimentali $x_1, x_2, \ldots, x_n$ vogliamo verificare la seguente ipotesi statistica

$$H_0 : \text{il campione ha distribuzione } \pi$$

Come sempre accade nei test di verifica di ipotesi, il punto chiave è determinare una *statistica test* che fornisca una "misura" della deviazione di dati rispetto all'ipotesi nulla, e la cui distribuzione sia nota, almeno in qualche senza approssimato.

Una soluzione, assolutamente non banale, di questo problema fu proposta da *K. Pearson* in un celebre articolo pubblicato nel 1900. Egli considerò le quantità

$$f_k := n\pi(k)$$

che chiameremo *frequenze teoriche* o *frequenze attese*. Si tratta infatti del *valor medio di N_k se la distribuzione delle X_i è π*.

Pearson introdusse la statistica campionaria

$$\mathcal{P} := \sum_{k=1}^{m} \frac{(N_k - f_k)^2}{f_k}$$

e dimostrò che, se H_0 è vera, la distribuzione di questa statistica è, per n sufficientemente grande, approssimativamente χ^2_{m-1}.

Ne segue immediatamente che una regione critica per un test di livello di significatività *approssimativamente* α per H_0 è data da

$$C := \left\{ (x_1, x_2, \ldots, x_n) : \sum_{k=1}^{m} \frac{(n_k - f_k)^2}{f_k} > \chi^2_{m-1,\alpha} \right\}$$

dove le n_k sono le frequenze calcolate sui dati $x_1, x_2, \ldots, x_n$.

Esempio

Sono stati ottenuti dall'anagrafe i dati relativi ai primi tre figli di 1054 famiglie con almeno tre figli. Ne risulta che 157 famiglie hanno tre figli maschi, 371 due maschi e una femmina, 362 un maschio e due femmine, 164 tre femmine. Questi dati sono compatibili con l'ipotesi che ogni figlio nasca maschio (o femmina) con probabilità 1/2 indipendentemente dagli altri?

Se quest'ultima ipotesi è verificata, il numero di figlie femmine ha distribuzione $B(3, 1/2)$. Pertanto poniamo, per $k = 0, 1, 2, 3$,

$$\pi(k) = \binom{3}{k} \left(\frac{1}{2}\right)^3$$

Si calcolano allora le frequenze attese:

$$f_k = 1054\pi(k)$$

Si calcola quindi il valore della statistica di Pearson

$$\mathcal{P} = 17.02$$

Supponendo che $n = 1054$ sia in questo esempio sufficientemente grande da poter applicare il test χ^2 di Pearson, calcoliamo il quantile

$$\chi^2_{3,0.01} = 11.34$$

Essendo $\mathcal{P} > \chi^2_{3,0.05}$, l'ipotesi

$$H_0 : \text{la distribuzione del campione è } \pi$$

viene rifiutata all' 1%.

Come facciamo a sapere se n è sufficientemente grande da poter applicare il test χ^2 di Pearson?

Una regola empirica accettabile è la seguente: *tutte le f_k sono maggiori o uguali a 1, e almeno l'80% di esse sono maggiori o uguali a 5*.

Si vede facilmente che tale condizione è verificata nell'esempio precedente.

Se questa condizione non è verificata *si raggruppano alcuni valori della variabile in una singola classe*.

Supponiamo di dover verificare l'adattamento di certi dati ad una distribuzione $B(3, 0.1)$. Abbiamo $n = 1000$ dati a disposizione, che forniscono $n_0 = 735$, $n_1 = 240$, $n_2 = 23$, $n_3 = 2$.

Le corrispondenti frequenze attese sono: $f_0 = 729$, $f_1 = 243$, $f_2 = 27$, $f_3 = 1$.

In questo caso una frequenza teorica su 4 è minore di 5. Non è perciò verificata la condizione di applicabilità del test χ^2 di Pearson. Una semplice soluzione consiste nel raggruppare i dati di valore 2 e 3 in una singola classe. In questo modo otteniamo le frequenze osservate $n_0 = 735$, $n_1 = 240$, $n_2 = 25$, e le frequenze attese $f_0 = 729$, $f_1 = 243$, $f_2 = 28$, che soddisfano la condizione di applicabilità del test χ^2 di Pearson.

Si trova

$$\mathcal{P} = \frac{(735 - 729)^2}{729} + \frac{(240 - 243)^2}{243} + \frac{(25 - 28)^2}{28} = 0.408$$

Questo valore va confrontato con il quantile $\chi^2_{2,\alpha}$. Essendo $\chi^2_{2,0.05} = 5.99$, l'ipotesi di adattamento alla distribuzione $B(3, 0.1)$ viene accettata: i dati mostrano un ottimo adattamento a questa distribuzione.

Il caso di una variabile discreta che assume infiniti valori

Questo caso, rispetto al precedente, non presenta alcuna difficoltà concettuale. Si tratta di *raggruppare i valori della variabile in classi*, in modo che la condizione di applicabilità del test di Pearson risulti verificata.

Esempio

Si vuole verificare l'adattamento dei dati della seguente tabella, che si riferisce al numero settimanale (in $n = 130$ settimane successive) di operazioni di manutenzione straordinaria in un centro di calcolo, alla distribuzione $Po(3)$:

numero di casi	frequenza
0	8
1	20
2	27
3	26
4	24
5	12
6	6
7 o più	7

I dati sono quindi raggruppati in 7 classi. La distribuzione "teorica" π è

$$\pi(k) = e^{-3}\frac{3^k}{k!} \qquad \text{per } k = 0, 1, 2, 3, 4, 5, 6$$

e

$$\pi(7) = 1 - \sum_{k=0}^{6} \pi(k)$$

che, moltiplicati per $n = 130$ danno le frequenze attese: 6.47 19.42 29.13 29.13 21.84 13.11 6.55 4.36 . Soltanto una su otto di queste frequenze è minore di 5, ed in ogni caso maggiore di 1. È quindi lecito applicare il test di Pearson.

Si trova:

$$\mathcal{P} = 2.83$$

Essendo $\chi^2_{7,0.05} = 14.067$, ne segue che l'ipotesi di adattamento è accettata al 5%: i dati sono in ottimo accordo con la distribuzione $Po(3)$.

Il caso di variabili continue

Nel caso in cui la variabile in esame sia continua, è necessario suddividere l'insieme dei valori possibili in *classi*, come descritto nel Capitolo di Statistica Descrittiva. Il test di buon adattamento viene quindi applicato alle frequenze di tali classi.

Nel caso più frequente, la *distribuzione ipotizzata* della variabile viene assegnata in termini della sua *distribuzione cumulativa* F.

Se $C_1, C_2, \ldots, C_m$ sono le *classi*, ognuna di esse è della forma $C_k = [a_k, b_k]$. La probabilità *ipotizzata* della classe C_k è

$$\pi(k) = F(b_k) - F(a_k)$$

da cui si calcola la frequenza attesa $f_k = n\pi(k)$.

Distribuzione assegnata a meno di parametri incogniti

In buona parte delle situazioni applicative, la distribuzione ipotizzata π contiene parametri incogniti. In questo caso il test χ^2 di Pearson viene applicato seguendo la seguente procedura:

- Sia k il numero di parametri incogniti (di solito $k = 1$ o 2). Anzitutto questi parametri vanno stimati con i dati: la media con la media campionaria e la varianza con la varianza campionaria.
- Si calcola quindi da statistica $\mathcal{P}$ di Pearson dove, nelle frequenze attese, i parametri incogniti vengono sostituiti dai loro valori *stimati*.
- La regione critica per un test a livello α, è determinata dalla disuguaglianza

  $$\mathcal{P} > \chi^2_{m-1-k,\alpha}$$

  Quindi ad ogni parametro stimato corrisponde la perdita di un grado di libertà.

Esempio 1

La seguente tabella riporta il numero di errori di diagnosi appurati nell'anno 1986 da parte di 44 medici che lavorano nel pronto soccorso di un grande ospedale:

Numero di errori	Frequenza

0	1
1	12
2	12
3	5
4	1
5	4
6	2
7	2
8	2
9	1
10	1
11	1

Questi dati sono compatibili con l'ipotesi che il numero di errori di diagnosi per anno di un medico abbia distribuzione di Poisson?

Il parametro incognito di un distribuzione di Poisson è uguale alla media, che pertanto *stimiamo con la media campionaria*

$$\overline{x} = 3.34$$

A questo punto dobbiamo verificare l'adattamento dei dati della tabella precedente alla distribuzione $Po(3.34)$. I valori della funzione di massa di tale distribuzione, e le relative frequenze attese sono riportati nella seguente tabella.

k	$e^{-3.34}\frac{3.34^k}{k!}$	$44e^{-3.34}\frac{3.34^k}{k!}$
0	$0,0354$	$1,5592$
1	$0,1184$	$5,2078$
2	$0,1977$	$8,6971$
3	$0,2201$	$9,6827$
4	$0,1838$	$8,0851$
5	$0,1227$	$5,4008$
6	$0,0683$	$3,0065$
7	$0,0326$	$1,4345$
8	$0,0136$	$0,5989$
9	$0,0051$	$0,2223$
10	$0,0017$	$0,0742$
≥ 11	$0,0007$	$0,0309$

Ci sono troppe frequenze attese "piccole" per poter applicare il test χ^2 di

Pearson. È perciò necessario congiungere alcuni valori in classi.

classe	π	frequenza attesa
$\{0,1\}$	$0,1538$	$6,7670$
2	$0,1977$	$8,6971$
3	$0,2201$	$9,6827$
4	$0,1838$	$8,0851$
5	$0,1227$	$5,4008$
≥ 6	$0,1220$	$5,3673$

Confrontiamo ora questi valori con le frequenze osservate

classe	frequenza osservata	frequenza attesa
$\{0,1\}$	13	$6,7670$
2	12	$8,6971$
3	5	$9,6827$
4	1	$8,0851$
5	4	$5,4008$
≥ 6	9	$5,3673$

Da questa tabella si calcola immediatamente la statistica di Pearson:

$$\mathcal{P} = \frac{(13-6,7670)^2}{6,7670} + \frac{(12-8,6971)^2}{8,6971} + \frac{(5-9,6827)^2}{9,6827} + \frac{(1-8,0851)^2}{8,0851} + \frac{(4-5,4008)^2}{5,4008} + \frac{(9-5,3673)^2}{5,3673} = 18,2908$$

Il problema è stato ricondotto a 6 classi, e un parametro è stato stimato. Pertanto in numero di gradi di libertà della χ^2 di Pearson è $6-1-1=4$. Pertanto il valore $\mathcal{P} = 18,2908$ va confrontato con il quantile $\chi^2_{4,\alpha}$. Per $\alpha = 0.01$ si trova

$$\chi^2_{4,0.01} = 13,2767 < \mathcal{P}$$

L'ipotesi di adattamento ad una distribuzione di Poisson viene rifiutata all'1%: i dati sono fortemente in contrasto con tale ipotesi.

9.2 Tabelle di contingenza e test χ^2 di indipendenza

Ricordate la seguente tabella, vista all'inizio del corso?

	biologia	**biologia molecolare**	**biotecnologia**	
la detesto	37	21	20	78
la sopporto	3	6	3	12
un po' mi piace	5	6	4	15
mi piace molto	1	2	0	3
	46	35	27	108

La tabella riporta i valori di due variabili (y = corso di studi di appartenenza e x = gradimento della statistica), "misurate" sugli stessi 108 individui. Lo scopo della seguente analisi è di determinare eventuali correlazioni tra le due variabili. Le tabelle come la precedente sono chiamate *tabelle di contingenza*.

In generale, sia $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ un campione aleatorio per due variabili x e y. Assumiamo che la variabile x possa assumere i valori $\{1, 2, \ldots, m\}$, e la variabile y i valori $\{1, 2, \ldots, r\}$.

Una tabella di contingenza riporta, nell'incrocio tra la colonna k e la riga h, la *frequenza congiunta*

$$N_{h,k} = |\{i : X_i = h \text{ e } Y_i = k\}|$$

L'ultima riga riporta le *frequenze marginali*

$$N_k^y = |\{i : Y_i = k\}|$$

mentre nell'ultima colonna compaiono le *frequenze marginali*

$$N_h^x = |\{i : X_i = h\}|$$

Con queste frequenze è possibile verificare l'ipotesi statistica

$$H_0 : \text{ le due variabili sono indipendenti}$$

Per la legge dei grandi numeri, se n è grande,

$$\frac{N_{h,k}}{n} \simeq P(X_i = h, Y_i = k)$$

$$\frac{N_k^y}{n} \simeq P(Y_i = k)$$

$$\frac{N_h^x}{n} \simeq P(X_i = h)$$

Se H_0 è vera,

$$P(X_i = h, Y_i = k) = P(Y_i = k)P(X_i = h)$$

Quindi, se n è grande, è molto probabile che

$$\frac{N_{h,k}}{n} \simeq \frac{N_k^y}{n}\frac{N_h^x}{n}$$

La statistica campionaria seguente è basata proprio sul confronto tra le frequenze relative $\frac{N_{h,k}}{n}$ e i prodotti delle frequenze relative marginali $N_k^y n \frac{N_h^x}{n}$.

$$\mathcal{P} = \sum_{h=1}^{m}\sum_{k=1}^{r} \frac{\left(N_{h,k} - \frac{N_h^x N_k^y}{n}\right)^2}{\frac{N_h^x N_k^y}{n}}$$

Si può mostrare che, per n sufficientemente grande, se H_0 è vera $\mathcal{P}$ è approssimativamente distribuita come una $\chi^2_{(m-1)(r-1)}$.

L'approssimazione è da ritenersi buona se *i denominatori* $\frac{N_h^x N_k^y}{n}$ *sono tutti maggiori o uguali a* 1 *e almeno l'*80% *di essi sono maggiori o uguali a* 5.

Ne segue che la regione critica di un test per H_0 a livello α è

$$C := \left\{ (x_1, y_1, \ldots, x_n, y_n) : \sum_{h=1}^{m} \sum_{k=1}^{r} \frac{\left(n_{h,k} - \frac{n_h^x n_k^y}{n} \right)^2}{\frac{n_h^x n_k^y}{n}} \geq \chi^2_{(m-1)(r-1),\alpha} \right\}$$

Esempio

Consideriamo la seguente tabella di contingenza, che riporta i risultati ad un appello di esame degli studenti di tre corsi di studio

	Biologia Molecolare	**Biologia**	**Biotecnologia**	
esame superato	30	15	50	95
esame non superato	40	8	37	85
	70	23	85	180

Calcoliamo la statistica $\mathcal{P}$:

$$\mathcal{P} = \frac{(30 - \frac{95\cdot 70}{180})^2}{\frac{95\cdot 70}{180}} + \frac{(15 - \frac{95\cdot 23}{180})^2}{\frac{95\cdot 23}{180}} + \frac{(50 - \frac{95\cdot 85}{180})^2}{\frac{95\cdot 85}{180}}$$
$$+ \frac{(40 - \frac{85\cdot 70}{180})^2}{\frac{85\cdot 70}{180}} + \frac{(8 - \frac{85\cdot 23}{180})^2}{\frac{85\cdot 23}{180}} + \frac{(37 - \frac{85\cdot 85}{180})^2}{\frac{85\cdot 85}{180}} = 4.9613$$

Questo valore va confrontato con il quantile $\chi^2_{2,\alpha}$ $(2 = (2-1)(3-1))$. Per $\alpha = 0.05$

$$\chi^2_{2,0.05} = 5.991$$

che conduce all'accettazione dell'ipotesi H_0 di indipendenza. *Non* si può perciò dedurre da questi dati che vi sia dipendenza tra corso di studio e profitto in quel particolare esame.