Introduzione alla statistica

La statistica descrittiva

Dicembre, 2016

Attilio Peloni, Istituto Alberti, Bormio

Overview

Premessa

Statistica: strumento fondamentale per il supporto decisionale.
Consiste nella:

- raccolta dati
- elaborazione e analisi dati con strumenti adatti
  (per esempio tabelle e grafici)
- interpretazione dati e valutazione con opportuni metodi statistici

Svolgimento indagine statistica

Si seguono i seguenti passi:

1. definire un tema (Es. calcolare il valore di π utilizzando un metodo definito)
2. definire le variabili che interessa misurare
3. fissare i metodi di presa dati (Es. definire il campione che desidera utilizzzare)
4. misura dei dati (Es. misura lunghezza stuzzicadenti, distanza fra linee, conto del numero casi favorevoli,...)
5. elaborazione dei dati
6. interpretazione dei dati

L'istogramma

I dati statistici e le loro frequenze si possono rappresentare graficamente in modi diversi.
Rappresentiamo i dati raccolti nell'esperimento di Buffon.
Nella tabella sono riportati 10 esperimenti (10 lanci), per ognuno e' calcolato il valore di π, utilizzando la formula assegnata

L'istogramma

Numero esperimento	π calcolato		
1	3,15		
2	3,50		
3	3,50		
4	2,90		
5	3,15		
6	3.50		
7	3.50		
8	4,00		
9	3.50		
10	3.40		

L'istogramma

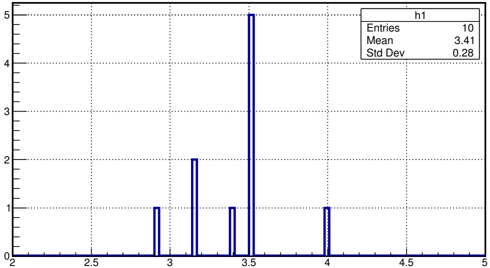


Figure: Istogramma delle frequenze di π calcolato

- Asse orizzontale: valori degli estremi delle classi, ottenendo lunghezze che rappresentano le ampiezze degli intervalli (basi dei rettangoli)
- Asse verticale: valori della frequenza con cui e' stato osservato un evento.

L'istogramma

Osservando l'istogramma e la serie di dati si nota che alcuni valori si ripetono. E' naturale introdurre una nuova quantita' per descrivere il fenomeno: la frequenza

Frequenza assoluta

di un evento (modalita') e' il numero di volte in cui essa si presenta

Frequenza relativa p

di un evento (modalita') e' il rapporto fra la frequenza assoluta f e la totalita' T delle osservazioni:

$$p = \frac{f}{T}$$

L'istogramma

	π calcolato	f	p	p(%)
	2,90	1	1/10	10%
	3,15	2	2/10	20%
	3,40	1	1/10	10%
	3,50	5	5/10	50%
	4,00	1	1/10	10%
Totale		**10**	**1**	**100%**

L'istogramma

Dall'esempio si verificano le seguenti proprieta':

$$\sum_{i=1}^{n} p_i = 1$$

$$\sum_{i=1}^{n} p_i(\%) = 100\%$$

Indici di posizione

Esistono dei valori che riassumono e rappresentano un insieme di dati. Permettono di:

1. dedurre le caratteristiche di una situazione statistica;
2. di confrontare diverse situazioni;
3. studiare la variabilita' dei dati

Media aritmetica

Media aritmetica semplice

fra n numeri $x_1, x_2, \ldots, x_n$, e' il rapporto fra la loro somma ed n:

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

$$M = \frac{3.15 + 3.50 + 3.50 + 2.90 + 3.15 + 3.50 + 3.50 + 4.00 + 3.50 + 3.40}{10}$$

$$= 3.41$$

Media ponderata

Media ponderata

quando ogni dato ha un suo peso (rappresentato dalla sua frequenza). Se $f_1, f_2, \ldots, f_n$ sono le frequenze degli eventi (modalita') di $x_1, x_2, \ldots, x_n$, la media aritmetica e' data dalla formula:

$$M = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

$$M = \frac{3.15x2 + 3.50x5 + 2.90x1 + 3.40x1 + 4.00x1}{10} = 3.41$$

Scarto della media

Scarto della media

e' la differenza tra il valore osservato e la media stessa. Dati n valori $x_1, x_2, ..., x_n$, gli scarti dalla loro media M sono i valori:

$$(x_1 - M), (x_2 - M), ..., (x_n - M)$$

per cui vale:

$$(x_1 - M) + (x_2 - M) + ... + (x_n - M) = x_1 + x_2 + ... + x_n - nM = nM - nM = 0$$

da cui:

Scarto della media

$$\sum_{i=1}^{n} (x_n - M) = 0$$

Campo di variabilita'

Esistono dei valori che riassumono e rappresentano un insieme di dati. Essi permettono di dedurre le caratteristiche di una situazione statistica e di confrontare diverse situazioni. Studiamo come si disperdono i dati rispetto a uno dei valori indice scelto come riferimento.

Campo di variabilita'

di un insieme di n dati numerici $x_1 + x_2 + ... + x_n$ e' la differenza tra il valore massimo ed il valore minimo degli x_i:

$$x_{massimo} - x_{minimo}$$

Il campo di variabilita' dell'esperimento e' di 1.10 unita':
$\pi = 4.00 - 2.90 = 1.10$

Scarto quadratico medio o deviazione standard

Il campo di variabilita':

1. e' un indice grossolano. Dipende fortemente dai valori estremi considerati
2. non misura la variabilita' dei dati in modo significativo

La media, da sola, ignora la dispersione dei dati.
E' necessario introdurre un indice di variabilita' che indica di quanto un dato si discosta dal valor medio.

Scarto quadratico medio o deviazione standard

Deviazione standard

E' la media quadratica degli scarti dalla media aritmetica.

$$\sigma = \sqrt{\frac{\sum_i (x_i - M)^2}{n}}$$

oppure

$$\sigma = \sqrt{\frac{\sum_i (x_i - M)^2 f_i}{\sum_i f_i}}$$

Scarto quadratico medio o deviazione standard

Numero esperimento	x_i (π calcolato)	$x_i - M$	$(x_i - M)^2$
1	3,15	-0,26	0,0067
2	3,50	0.09	0,0081
3	3,50	0,09	0,0081
4	2,90	-0,51	0,2601
5	3,15	-0,26	0,0676
6	3.50	0,09	0,0081
7	3.50	0,09	0,0081
8	4,00	0,59	0,3491
9	3.50	0,09	0,0081
10	3.40	-0,01	0,0001

Scarto quadratico medio o deviazione standard

Dalla formula data e dai dati in tabella il risultato e':

$$M = 3,41 \qquad \sigma = 0,28$$

Questo significa che dal valore medio di $M = 3,41$ ci si puo' aspettare uno spostamento dei dati tra un valore massimo ($M = 3,41 + 0,28 = 3,69$) e uno minimo ($M = 3,41 - 0,28 = 3,13$)
Nel linguaggio della fisica questo risultato si scrive:

$$\pi = 3,41 \pm 0,28$$

Questo risultato e' molto buono poiche' l'errore relativo commesso in questa misura e':

$$\epsilon = \frac{\sigma}{M} = 8\%$$

in altri termini, il valore "vero" di π greco cade all'interno dell'intervallo $(M - \sigma, M + \sigma)$.

Scarto quadratico medio o deviazione standard

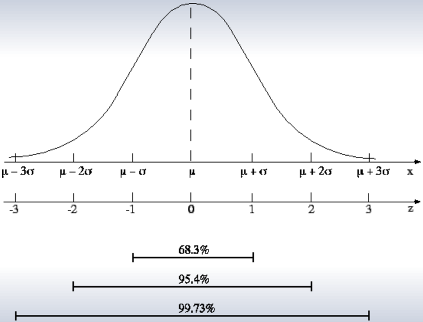
Immaginiamo di confrontare i dati ottenuti da tre gruppi di studenti differenti. Tutti i gruppi hanno ottenuto la stessa media di π.
Quale e' il gruppo che ha ottenuto il risultato piu' attendibile?

Gruppo	valore medio di π calcolato	σ	
1	3,41	0,28	
2	3,41	0.09	
3	3,41	0,80	

Scarto quadratico medio o deviazione standard

Risposta: il gruppo Nr. 2 poiche' rappresenta la variabilita' minore (variazione attorno al valor medio).
Significa che il risutato medio di π e' 3.41 ma ci si deve preparare a superare tale valore "sopra" e "sotto" pari a 0.09 ovvero $4,31$ o $3,32$.
Il gruppo Nr. 3 mostra una maggiore variabilita' con un valore minimo $3,41 - 0.80 = 2,61$ e un valore massimo $3,41 + 0.80 = 4,21$

La distribuzione gaussiana



La distribuzione gaussiana

Se aumetiamo le misure l'istogramma Nr. 1 del valore di π calcolati tenderebbe a diventare una curva a campana detta curva gaussiana.
La σ ha particolare importanza nelle distribuzioni gaussiane, perché è collegato al modo in cui le frequenze si distribuiscono attorno al valore medio M.
Da un'analisi del grafico si possono fare alcune osservazioni:

- la simmetria della curva rispetto alla retta $x = M$. Intorno al valore medio tutti gli altri si distribuiscono con la stessa frequenza per valori equidistanti da M
- σ grande significa grande dispersione. Diminuendo σ diminuisce la dispersione e la curva diventa sempre piu' "piccata" (allungata e stretta) attorno al valor medio.

Si puo' dimostrare che:

- il 68,4% dei casi osservati e' compreso tra $M - \sigma$ e $M + \sigma$
- il 95,4% dei casi osservati e' compreso tra $M - 2\sigma$ e $M + 2\sigma$
- il 99,7% dei casi osservati e' compreso tra $M - 3\sigma$ e $M + 3\sigma$

Ad esempio possiamo immaginare, dopo un gran numero di lanci, di avere una percentuale di valore di π calcolati pari al $0,3\%$ maggiori o minori del valore medio calcolato di $\pm 3\sigma$.

Bibliografia

Libro di testo