

STATISTICA DESCRITTIVA - SCHEDA N. 3

VARIABILI QUANTITATIVE

Indici di centralità, dispersione e forma

Esistono altri indici che forniscono informazioni sulla distribuzione dei dati osservati, oltre a quelli basati sui quantili, visti nella scheda n. 2.

In seguito indicheremo con n il numero dei dati osservati e con x_i , l' i -esimo dato osservato (non necessariamente ordinato).

1. Indici di centralità (o posizione)

Forniscono indicazioni sulla posizione dei dati, ovvero indicano intorno a quali valori numerici si distribuisce la variabile osservata X .

media empirica (mean)	\bar{x}	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
media spuntata (trimmed mean)		Media dei dati calcolata considerando solo il 90% dei dati centrali, cioè compresi fra il 5% e il 95% dei dati ordinati
moda/e		Valore/i con frequenza massima
mediana (median)	Q_2	il minimo valore osservato tale che almeno il 50% dei dati è minore o uguale a questo

Sofferamoci sulla media.

Scriviamo la formula della media utilizzando la distribuzione della variabile X .

Ricordiamo che la distribuzione della variabile è l'insieme delle coppie (x_k, f_k) , per k da 1 a m , avendo indicato con x_k gli m differenti valori assunti dalla variabile e con f_k le corrispondenti frequenze relative:

$$\bar{x} = \sum_{k=1}^m f_k x_k$$

Osserviamo che se i valori assunti dalla variabile sono tutti diversi, la frequenza di ciascun dato è $1/n$ e si ritrova la formula precedente.

La media gode delle seguenti proprietà:

1. la somma degli errori che si commettono sostituendo il valore della media a tutte le osservazioni

(scarto) è nullo, ovvero $\sum_{i=1}^n (x_i - \bar{x}) = 0$;

2. la media rende minima la somma dei quadrati degli scarti, cioè, se scegliamo qualunque altro numero a e consideriamo i quadrati degli scarti dei dati da a , abbiamo la seguente disuguaglianza

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2$$

La media viene anche detta **baricentro dei dati**. Infatti se interpretiamo i diversi valori assunti dalla variabile come pesi "attaccati" all'asse reale, la media è il punto di equilibrio dei dati. Proprio in quanto baricentro dei dati, la media risente molto della posizione dei valori estremi, la media troncata ovvia in parte questo problema. La mediana non è influenzata dai valori estremi.

ESEMPIO 1. Consideriamo il peso e l'altezza di 92 studenti.

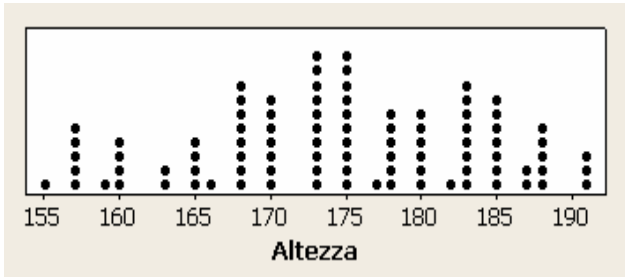


Figura 1. Dotplot Altezza

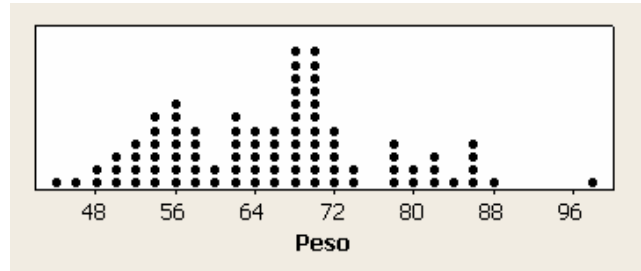


Figura 1. Dotplot Peso

Variable	Mean	TrMean	Median
Altezza	174.54	174.71	175.26
Peso	65.84	65.56	65.77

NB: per determinare la media spuntata, si cancellano i primi 5 (5% di 92) e gli ultimi 5 dei valori ordinati e si calcola la media dei rimanenti.

Una proprietà simile a quella sopra considerata per la media che riguarda la **mediana** è la seguente. La mediana rende minima la somma degli scarti assoluti, cioè, se scegliamo qualunque altro numero a e consideriamo gli scarti assoluti dei dati da a , abbiamo la seguente *disuguaglianza*

$$\sum_{i=1}^n |x_i - Q2| \leq \sum_{i=1}^n |x_i - a|$$

La media è preferibile in molte circostanze come indice di posizione perché ha buone proprietà che permettono di costruire modelli statistici previsionali a partire dai dati osservati. D'altra parte la mediana è un indice di posizione che è meno influenzato dai valori estremi e quindi può risultare più stabile, come possiamo vedere nel seguente esempio.

ESEMPIO 1 (continua)

Supponiamo nei 92 dati dell'esercizio precedente per un errore di battitura, sia stata digitata un'altezza 1.905 anziché 190.5. Il tal caso la mediana resta invariata (175.26), mentre la media diminuisce (172.49).

Supponiamo che i dati delle Altezze dell'esercizio precedente siano già stati suddivisi in 9 classi. Si può ancora determinare un valor medio e una mediana, approssimando ogni classe con il suo valore centrale.

intervalli	conteggi	Valore centrale
150-155	1	152.5
155-160	6	157.5
160-165	6	162.5
165-170	13	167.5
170-175	17	172.5
175-180	17	177.5
180-185	15	182.5
185-190	14	187.5
190-195	3	192.5

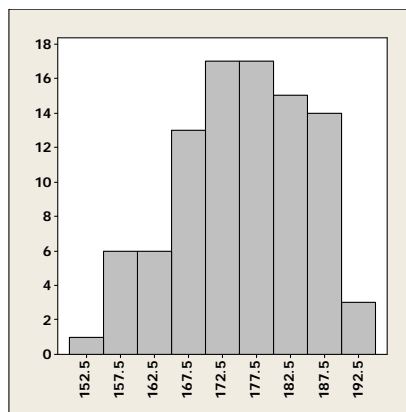


Figura 3. Istogramma Altezze

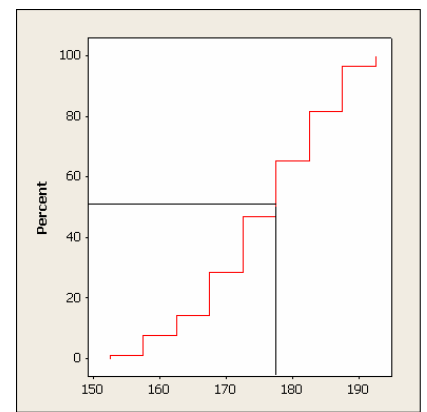


Figura 4. Cumulata Altezze

Il valore medio dei dati raggruppati in classi è dato da: $\frac{152.5 \times 1 + \dots + 192.5 \times 3}{92} = 175.435$

La mediana dei dati raggruppati in classi si ottiene dalla funzione di distribuzione cumulata calcolata a partire dai valori centrali delle classi.

2. Indici di dispersione

Esaminiamo i principali.

varianza	σ^2	$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
scarto quadratico medio	σ	radice quadrata positiva della varianza
inter quartile range	IQR	Q3-Q1
intervallo di variazione	R	valore massimo – valore minimo

In alcuni contesti si usa una definizione leggermente differente di varianza, dove al denominatore si sostituisce n con (n-1). Osserviamo che nel caso in cui i dati osservati siano numerosi, le due definizioni coincidono.

Anche la varianza può essere espressa utilizzando la distribuzione della variabile X:

$$\sigma^2 = \sum_{k=1}^m f_k (x_k - \bar{x})^2$$

La varianza può essere scritta in modo più semplice per i calcoli, svolgendo il quadrato:

$$\sigma^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 \quad \text{oppure} \quad \sigma^2 = \left(\sum_{k=1}^m f_k x_k^2 \right) - \bar{x}^2$$

ovvero come la differenza fra la media dei dati al quadrato e la media al quadrato.

La varianza gode di alcune proprietà importanti

1. la varianza è sempre positiva;
2. σ^2 vale 0 se e solo se la variabile quantitativa osservata è costante, cioè i dati osservati coincidono con un unico valore (la media).

Queste proprietà spiegano perché la varianza misura quanto la variabile osservata si discosta dal suo valor medio. In particolare tanto è maggiore la varianza tanto più la variabile è dispersa, mentre valori bassi della varianza corrispondono a variabili con valori maggiormente concentrati attorno alla media.

La radice quadrata positiva della varianza è detta **scarto quadratico medio** (o **deviazione standard**). Osserviamo che se i dati sono espressi in kg, la varianza sarà in kg², mentre lo scarto sarà ancora in kg. Quindi, entrambi gli indici risentiranno della scelta delle unità di misura. Spesso si usa lo scarto quadratico medio come indice di dispersione invece della varianza proprio perché ha lo stessa unità di misura dei dati.

Vale al seguente relazione fra lo scarto quadratico medio e l'intervallo di variazione:

$$\sigma \leq \frac{R}{2}$$

NB: Esistono altri modi (poco utilizzati e meno efficaci da un punto di vista matematico) per misurare la "distanza" fra la variabile e la sua media. Ad esempio, si può definire lo scarto assoluto medio

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|. \quad \text{Come abbiamo già osservato, lo scarto medio } \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} \text{ vale sempre 0.}$$

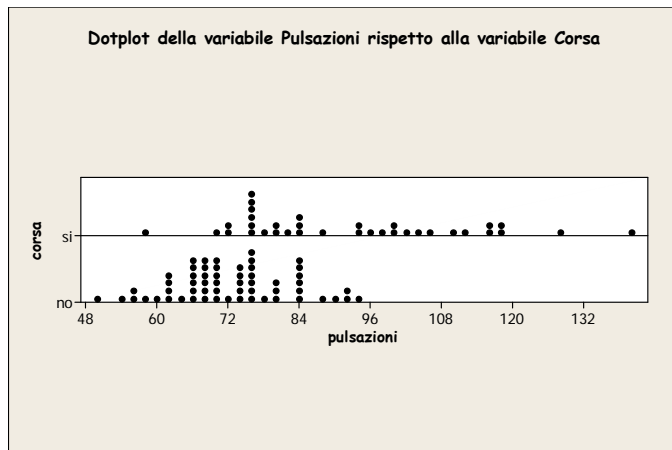
ESEMPIO 1 (continua) Con i dati dell'esempio precedente, si ha

Variable	StDev	Variance	Range	IQR
Altezza	9.29	86.39	35.56	15.24
Peso	10.77	115.95	54.43	14.29

Graficamente più la varianza è grande più il dotplot risulta "schiacciato" o disperso, più la varianza è piccola più il dotplot risulta "concentrato" su pochi valori, come si può osservare comparando i dotplot delle variabili Pesi e Altezze.

ESEMPIO 2.

Riprendiamo l'Esempio 2 della Scheda 2 riguardante le rilevazioni delle pulsazioni cardiache,



Variable	Corsa	Mean	StDev	Variance
Pulsazioni	si	92.51	18.94	358.85
	no	72.32	9.95	98.97

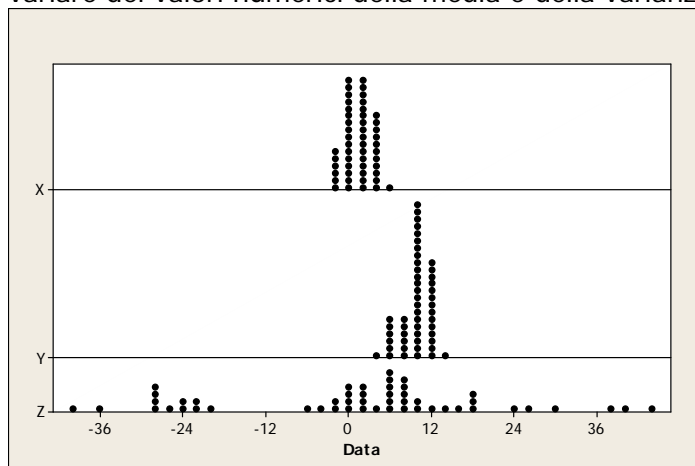
Come abbiamo già commentato in precedenza, coloro che hanno effettuato la corsa hanno una distribuzione molto più dispersa rispetto a chi non ha corso; lo scarto quadratico medio delle pulsazioni per chi ha corso è circa il doppio di quello per chi non ha corso.

Figura 5. Dotplot delle pulsazioni suddiviso per sottogruppi

In Appendice è riportato il calcolo della media e della varianza dell'intera popolazione a partire dalle medie e varianze nei sottogruppi.

ESEMPIO 3.

Nell'esempio seguente mostriamo come variano la posizione e la forma della distribuzione dei dati al variare dei valori numerici della media e della varianza.



Variable	Mean	Variance
X	1.324	4.162
Y	9.814	4.443
Z	1.29	375.25

Osserviamo che le variabili X e Z hanno circa la stessa media ma varianze molto diverse; invece le variabili X e Y hanno circa la stessa varianza ma medie molto diverse.

Figura 5. Dotplot di tre variabili

3) Indici di forma

Forniscono informazioni sulla forma della distribuzione e quindi del dotplot.

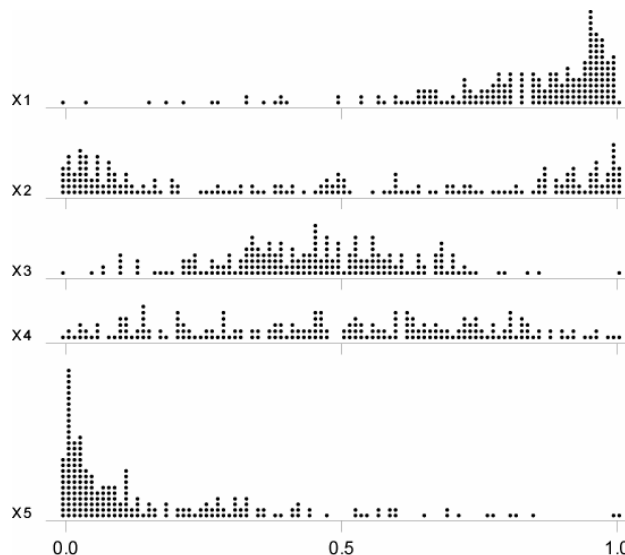
Alcune informazioni possono essere tratte dal confronto fra la posizione della media, della media spuntata e della mediana. Ad esempio, se la distribuzione è simmetrica i tre indici coincidono, invece (completare con minore e maggiore).

- se i valori sono più concentrati a sinistra, la mediana è _____ della media
- se i valori sono più concentrati a destra, la mediana è _____ della media

Indicate approssimativamente la media e la mediana delle distribuzioni riportate a fianco.

	\bar{x}	Q2
X1		
X2		
X3		
X4		
X5		

Indicate per quali distribuzioni la media troncata assume valori molto diversi dalla media.



Esistono, inoltre, alcuni indici che danno informazione sulla simmetria della distribuzione, anche se a volte sono di difficile interpretazione. Riportiamo i due più semplici.

confronto media, mediana e scarto quadratico	$\frac{\bar{x} - Q2}{\sigma}$
confronto fra quartili	$\frac{(Q3 - Q2) - (Q2 - Q1)}{Q3 - Q1}$

Entrambi gli indici assumono valori compresi fra -1 e 1 ; risultano positivi se i dati sono più concentrati a sinistra e valori negativi se i dati sono più concentrati a destra.

NB: Gli indici introdotti sono sintesi dell'informazione contenuta nella totalità dei dati e se esaminati singolarmente possono far perdere informazioni essenziali dei dati.

APPENDICE

Un approfondimento: media e varianza nei sottogruppi e nel gruppo totale.

Talvolta la popolazione studiata è suddivisibile in sottogruppi (ad esempio M e F), in questo caso è interessante studiare i legami fra la media e la varianza dei sottogruppi e i corrispondenti indici nella popolazione complessiva. Supponiamo per semplicità di avere due sottogruppi A e B di numerosità n_A e n_B , frequenza relativa f_A e f_B , medie \bar{x}_A e \bar{x}_B , varianze σ_A^2 e σ_B^2 . Avremo che

$$\bar{x}_{tot} = f_A \bar{x}_A + f_B \bar{x}_B$$

cioè la media totale è la media pesata delle medie dei sottogruppi (analogo al caso dei profili visto nella scheda n. 1)

$$\sigma_{tot}^2 = (f_A \sigma_A^2 + f_B \sigma_B^2) + (f_A (\bar{x}_A - \bar{x}_{tot})^2 + f_B (\bar{x}_B - \bar{x}_{tot})^2)$$

cioè la varianza totale è la media (pesata) delle varianze dei sottogruppi sommata alla varianza (pesata) delle medie dei sottogruppi.

Naturalmente i risultati precedenti si estendono anche al caso di un numero maggiore di sottogruppi.

ESERCIZI

1) Sotto sono riportati i pesi di bambini di 2 anni espressi in kg:

8.9 8.8 8.7 8.9 8.6 8.6 8.8 9.0 9.1 9.0 9.2 9.2 9.3 9.3
 9.4 9.5 9.3 9.4 9.5 9.3 8.5 9.0 9.4 9.5 9.6 9.3 9.7 9.8

a) Costruire una tabella con le frequenze assolute, la distribuzione dei pesi, la cumulata assoluta e la funzione di distribuzione cumulata.

b) Determinare i valori dei quartili Q1, Q2 e Q3 e la lunghezza dell'intervallo interquartile.

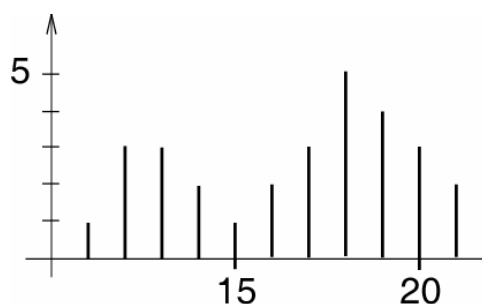
c) Calcolare la media e lo scarto quadratico medio del peso.

2) Nell'istogramma qui sotto è riportato il numero di vasi che una ditta artigianale ha venduto quotidianamente.

a) Qual è il numero minimo di vasi che sono stati prodotti in un giorn? e il numero massimo?

b) Per quanti giorni sono stati prodotti 18 vasi?

c) Quanti giorni sono stati considerati?



d) Completare la tabella a fianco.

e) Determinare i valori dei quartili Q1, Q2 e Q3 e la lunghezza dell'intervallo interquartile.

f) Calcolare la media e lo scarto quadratico medio del numero di vasi venduti quotidianamente da quella ditta, nel periodo di tempo esaminato.

	x_k	n_k	$n_k x_k$	x_k^2	$n_k x_k^2$
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
TOTALI					

3) Si effettuano 50 misure di una quantità e si registra la somma dei valori e la somma dei quadrati dei valori:

$$\sum_{i=1}^{50} x_i = 2480.82$$

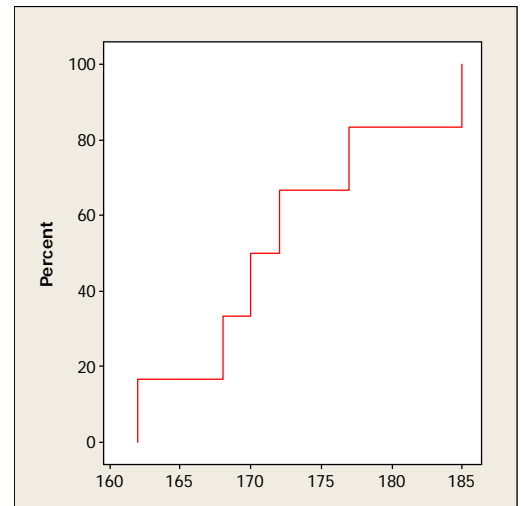
$$\sum_{i=1}^{50} x_i^2 = 125147$$

Calcolare media e scarto quadratico medio dei dati.

4) Qui sotto è riportata la funzione di distribuzione cumulata per le altezze di alcuni studenti maschi e a fianco si trova una sua rappresentazione grafica.

x	$F(x)$
162	0.10000
168	0.26667
170	0.60000
172	0.80000
177	0.93333
185	1.00000

Nota: sono riportati solo i valori di F calcolata nei valori delle altezze osservate.



a) Calcolare la media delle altezze degli studenti.

b) Calcolare la mediana delle altezze degli studenti.

5) Sotto sono riportati i pesi di bambini di 2 anni espressi in kg:

8.9 8.8 8.7 8.9 8.6 8.6 8.8 9.0 9.1 9.0 9.2 9.2 9.3 9.3
 9.4 9.5 9.3 9.4 9.5 9.3 8.5 9.0 9.4 9.5 9.6 9.3 9.7 9.8

a) Costruire una tabella con le frequenze assolute, la distribuzione dei pesi, la funzione cumulata assoluta e la funzione di distribuzione cumulata.

b) Determinare i valori dei quartili Q1, Q2 e Q3 e l'ampiezza dell'intervallo interquartile.

c) Calcolare la media e lo scarto quadratico medio del peso.