DIMA

CORSO LAUREA MAGISTRALE IN **DIGITAL MARKETING**

STATISTICA DESCRITTIVA

Prof. ssa ANNA LINA SARRA

MISURE di VARIABILITÀ

Variabilità

Per variabilità si intende l'attitudine dei fenomeni, naturali e sociali, a manifestarsi in modi differenti.

- **La variabilità** è l'attitudine di un carattere a presentare modalità differenti nel collettivo in esame.
- La distribuzione di un carattere presenta **variabilità nulla** se su tutte le unità statistiche si rileva la stessa modalità. In tal caso tutti gli indici di variabilità assumono valore zero.

Variabilità

Le misure variabilità sono indici che

1. Segnalano quanto sono tra loro diversi i valori della variabile
2. Evidenziano il grado di dispersione di ciascun valore rispetto ad un punto di riferimento
3. Misurano la diversità tra due termini della distribuzione o tra due quantili

Variabilità

MISURE DI VARIABILITA'

- ✓ assumono sempre il valore zero se i valori della variabile sono fra loro uguali
- ✓ assumono valori crescenti positivi per livelli progressivamente crescenti di variabilità: quanto più i termini della distribuzione sono fra loro diversi, tanto più l'indice assume valori elevati
- ✓ sono espressi nella stessa unità di misura della variabile

Indici di dispersione:
Scarto quadratico medio o deviazione standard

La deviazione standard rappresenta la distanza media fra tutte le osservazioni e la media.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Prese le distanze fra ogni osservazione e la media ("scarti"), se ne fa una media non aritmetica - *quadratica*

Es. X = peso paziente, std = 4.5kg: è la "distanza rilevante" fra due pazienti

Scarto quadratico medio

distribuzione disaggregata: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Numero di prodotti acquistati nell'ultimo mese

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

μ=7.14

0, 5, 6, 8, 9, 10, 12

❑ ***Scostamento quadratico medio:***

$$\sigma = \sqrt{\frac{(0-7.14)^2 + (5-7.14)^2 + \cdots + (12-7.14)^2}{7}}$$

$$= \sqrt{\frac{92.86}{7}} = 3.64$$

Numero di prodotti

σ per una distribuzione di frequenze a modalità singole: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu)^2 n_i}{N}}$$

Numero di prodotti acquistati nell'ultimo mese da 19 persone

x_i	n_i	$x_i * n_i$	$(xi - \mu)^2$	$(x_i - \mu)^2 \cdot n_i$
1	1	1	19.54	19.55
2	2	4	11.70	23.41
3	3	9	5.86	17.58
4	3	12	2.02	6.06
5	1	5	0.18	0.18
6	2	12	0.34	0.67
7	1	7	2.50	2.49
8	2	16	6.66	13.30
9	3	27	12.82	38.43
10	1	10	20.98	20.97
totale	19	103		142.63

- **Media aritmetica**

$$\mu = \frac{103}{19} = 5.42$$

- ***Scostamento quadratico medio***

$$\sigma = \sqrt{\frac{(1-5.42)^2 \cdot 1 + \cdots + (1-5.42)^2 \cdot 10}{19}} =$$

$$= \sqrt{\frac{142.63}{19}} = 7.51$$

Varianza

Il quadrato della deviazione standard prende il nome **di VARIANZA**

$$\sigma^2 = \frac{\sum_i^n (xi - \mu)^2}{N}$$

- Il numeratore è noto come «*somma dei quadrati degli scarti dalla media*»
- Si misura in unità quadratiche (Es. se x è l'altezza in *cm* la varianza sarà espressa in *cm* elevati al quadrato)
- Per questo conviene avere una misura espressa nelle unità originarie di x, e ciò si realizza calcolando la radice quadrata della varianza.

Diversità tra due termini della distribuzione

La distanza tra il valore più piccolo e il valore più grande è indicata come **Campo di variazione (*Range*) .**

Se utilizziamo la mediana come indice del centro della distribuzione, dividendo la distribuzione in due parti, possiamo usare la stessa idea per misurare la dispersione .

La distanza tra Q1 e Q3 è una misura di variabilità detta **INTERVALLO INTERQUARTILE**

Intervallo interquartile

N° ore attività fisica	Valori assoluti	Valori %	Cumulata
0	51	8.8	8.8
1	8	1.4	10.2
2	86	14.9	25.1
3	42	7.3	32.4
4	89	15.4	47.8
5	51	8.8	56.7
6	59	10.2	66.9
7	28	4.9	71.8
8	55	9.5	81.3
9	13	2.3	83.5
10	16	2.8	86.3
11	6	1.0	87.3
12	44	7.6	95.0
14	29	5.0	100.0

Intervallo interquartile $\Delta_q = Q_3 - Q_1$
individuato dal terzo e dal primo quartile : intervallo in cui è compreso il 50% delle osservazioni

Q3 =8 ore
Q1 =2 ore

Può essere calcolato per
VARIABILI QUANTITATIVE
E VARIABILI QUALITATIVE ORDINALI

Indici di variabilità relativa

Il problema si pone nel confrontare gli indici di variabilità di 2 o più distribuzioni *diverse* .

Indice di variabilità relativa: numero puro a-dimensionale che elimina l'influenza dell'unità di misura e dell'intensità media

Coefficiente di variazione

- Si chiama **indice di variabilità percentuale** il rapporto, moltiplicato per 100, tra un indice di variabilità assoluto e la media aritmetica.
- Particolare rilievo per le applicazioni ha **coefficiente di variazione**

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Il CV è una misura relativa di variabilità: esprime la variabilità in proporzione
alla dimensione media del carattere; inoltre, è un numero senza unità di misura, è quindi una misura adatta a confrontare la variabilità fra popolazioni diverse, e anche fra caratteri diversi.

Confrontare la variabilità di due distribuzioni

Peso neonato: media = 3.2 kg, std = 0.5 kg
Altezza neonato: media = 51 cm, std = 3.5 cm
Peso Madre: media = 64 kg, std = 4.5 kg

→ I neonati sono più variabili rispetto al peso o all'altezza?

→ Il peso è più variabile nei neonati o nelle madri?

Peso: CV = (0.5 kg / 3.2 kg)·100 = 15.6
Altezza: CV = (3.5 cm / 51 cm) = 6.9
Peso Madre: CV = (4.5 kg / 64 kg) = 7.0

→ I neonati sono più variabili rispetto al peso che all'altezza (circa il doppio) e in termini di peso sono variabili del doppio anche rispetto alle madri

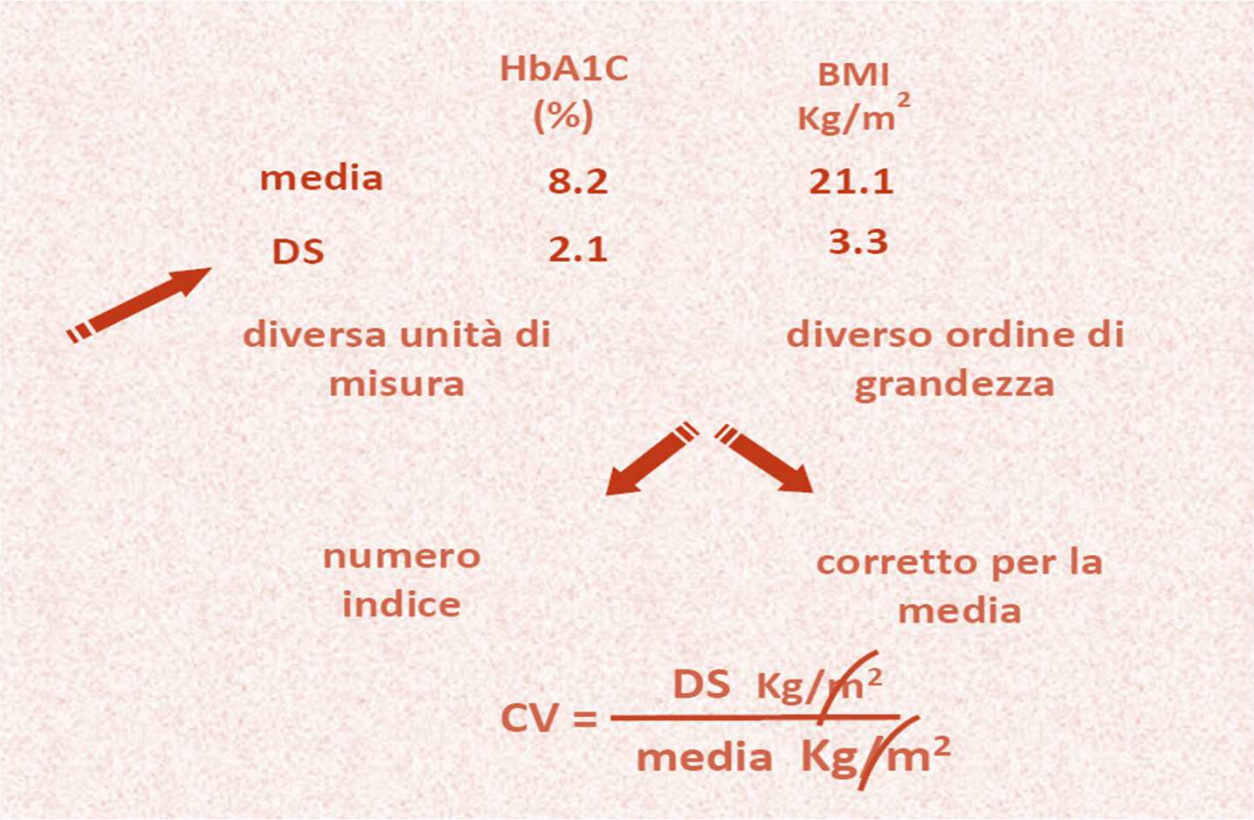
Confrontare la variabilità di due distribuzioni

Soggetti	HbA1C (%)	BMI Kg/m^2
1	7.5	20.17
2	6.3	19.48
3	8.4	19.26
4	8.6	22.21
...	...	...
577	8.5	16.41
578	6.4	17.31
577	7.3	22.46
media	8.2	21.1
DS	2.1	3.3

Quale delle due distribuzioni risulta essere caratterizzata da maggiore variabilità?

- diversa unità di misura
- diverso ordine di grandezza

Coefficiente di variazione



Coefficiente di variazione

	A1C (%)	BMI Kg/m^2	
media	8.2	21.1	
DS	2.1	3.3	
$CV = \frac{DS}{media}$	0.26	0.16	
$CV = \frac{DS}{media} \times 100$	26%	16%	può essere espresso in %

Il BOX-PLOT

- Una descrizione sintetica e abbastanza completa di una distribuzione di frequenze secondo un carattere quantitativo è data dal **box-plot**; questo è un riassunto a cinque numeri.
- I numeri sono i seguenti:

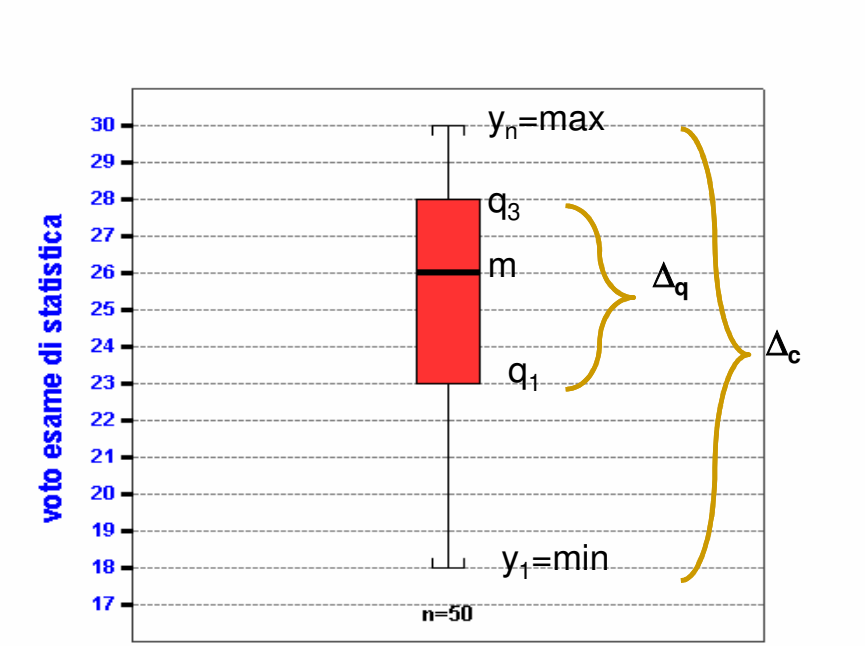
  - - la mediana (che dà informazioni sulla tendenza centrale)
  - - il primo e terzo quartile (la cui differenza dà informazioni sulla variabilità)
  - - i due estremi (la modalità più grande e la modalità più piccola)

- Questi numeri forniscono una descrizione sintetica di un insieme di dati anche quando il numero di unità osservate è elevato.

Box plot o Diagramma a scatola

Il box plot di una distribuzione è un grafico caratterizzato da tre elementi principali:

- una linea che indica la posizione della **mediana** della distribuzione
- un rettangolo (box) i cui estremi sono determinati in base ai **quartili Q1 e Q3** della distribuzione e la cui altezza indica la variabilità dei valori prossimi alla mediana
- due segmenti che partono dal rettangolo i cui estremi sono determinati in base ai **valori minimo e massimo** della distribuzione



	y1	q1	m	q3	yn
voto esame di statistica	18	23	26	28	30

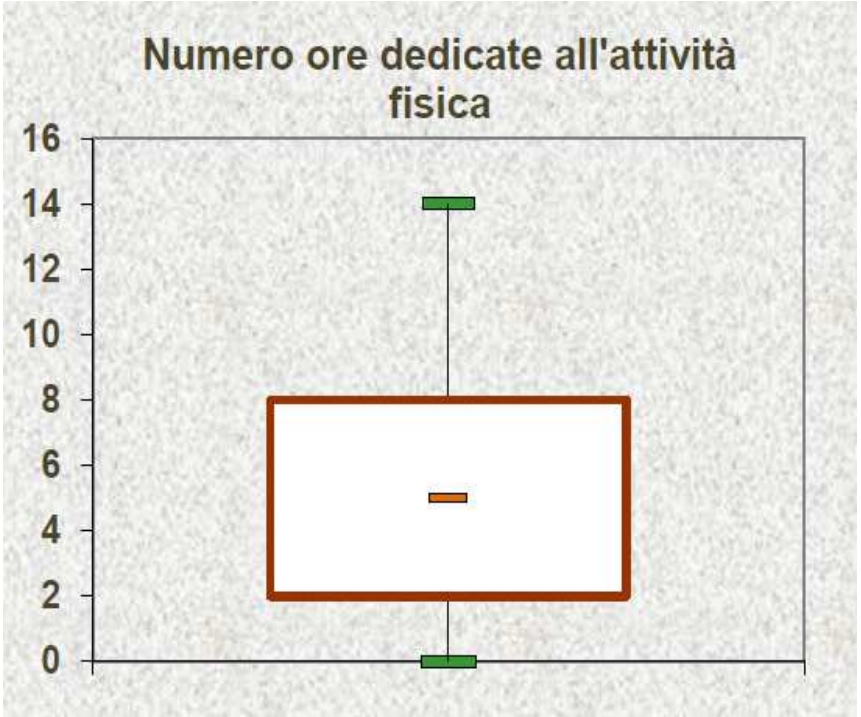
INTERPRETAZIONE DEL BOXPLOT

Il box-plot è utile perché riassume mediante pochi numeri molte informazioni su una distribuzione di frequenze.

- La mediana riassume la tendenza centrale della distribuzione.
- I quartili danno un'indicazione sulla variabilità, perché con essi si calcola lo scarto interquartile (misura più robusta del campo di variazione).
- La posizione della mediana rispetto ai quartili fornisce altre utili informazioni (in particolare sulla asimmetria della distribuzione).
- Gli estremi forniscono indicazioni non solo sul valore massimo e valore minimo ma soprattutto sull'eventuale presenza di dati con caratteristiche anomale .

Il BOX-PLOT

N° ore attività fisica	Valori assoluti	Valori %	Valori % Cumulati
0	51	8.8	8.8
1	8	1.4	10.2
2	86	14.9	25.1
3	42	7.3	32.4
4	89	15.4	47.8
5	51	8.8	56.7
6	59	10.2	66.9
7	28	4.9	71.8
8	55	9.5	81.3
9	13	2.3	83.5
10	16	2.8	86.3
11	6	1.0	87.3
12	44	7.6	95.0
14	29	5.0	100.0



Valori anomali ed estremi

Un dato è **anomalo** se:

- è maggiore del valore $Q3+1.5\Delta_q$
- è minore del valore $Q1-1.5\Delta_q$

Un dato è **estremo** (estremamente anomalo) se

- è maggiore del valore $Q3+3\Delta_q$
- è minore del valore $Q1-3\Delta_q$