- **Le misure della variabilità**

Gli indici di variabilità

- In tutti gli esempi visti nell'ultima lezione, abbiamo visto che le grandezze considerate - pur nelle diverse SCALE DI MISURA - non erano mai UNIFORMI, ma presentavano sempre una – più o meno forte – VARIABILITÀ
- Gli indici della tendenza centrale visti nella precedente lezione servivano proprio a fissare un punto di sintesi della distribuzione, una distribuzione che era sempre VARIABILE
- Il passo in avanti che facciamo oggi consiste nel MISURARE il "grado di variabilità" di una distribuzione

Come si misura la variabilità?

- Analogamente a quanto visto nell'ultima lezione a proposito degli indici della tendenza centrale, anche le misure della variabilità saranno DIVERSE a seconda della SCALA DI MISURA nella quale si trovano i nostri dati

Riprendiamo lo schema visto nella lezione precedente

Scala di misura	**Nominale**	**Ordinale**	**Ad intervalli**	**A rapporti**
Indici di tendenza centrale ammissibili	**Moda** (la modalità con la frequenza più alta)	Idem + **Mediana** (la modalità che divide in due parti uguali la distribuzione dei soggetti, ordinati sulla base della proprietà in questione)	Idem + **Media aritmetica** (la somma delle modalità della variabile in questione per i tutti i soggetti diviso il numero dei soggetti)	Idem + **Media geometrica**

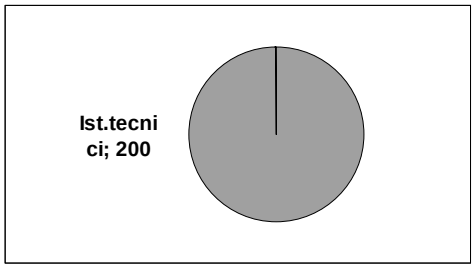
La scala nominale: la scuola frequentata

- Riprendiamo l'esempio della lezione precedente

Scuola	*Freq.ass.*	*Freq.rel.*	*%*
Licei	50	0,25	25%
Ist.tecnici	70	0,35	35%
Ist.professionali	80	0,4	40%
TOTALE	200	1	100%

C'è variabilità?

- La risposta è sì, perché altrimenti avremmo avuto TUTTI GLI STUDENTI concentrati in un'unica scuola; ad esempio, tutte le 200 matricole provenienti dagli istituti tecnici

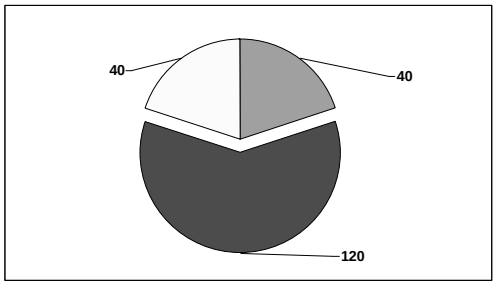


Consideriamo la distribuzione dell'anno precedente

Scuola	*Freq.ass.*	*Freq.rel.*	*%*
Licei	40	0,2	20%
Ist.tecnici	120	0,6	60%
Ist.professionali	40	0,2	20%
TOTALE	200	1	100%

È più o meno variabile rispetto a quella di quest'anno?

- A prima vista, sembra meno variabile, perché più studenti (120) sono concentrati nella modalità "Istituti tecnici"



Esiste un modo per misurare in modo rigoroso la variabilità?

- La risposta è sì. Esistono molte misure della variabilità, anche per i dati su scala nominale
- Noi ne consideriamo solo due, le più semplici: l'indice di entropia e l'indice di eterogeneità

L'indice di entropia

- Fu messo a punto da Claude Shannon, nell'ambito della teoria dell'informazione
- Indicando con $\{p_i;\ i=1,2,\ldots, k\}$ la serie delle frequenze relative delle k classi della distribuzione considerata (nel nostro caso 3), la formula per il calcolo dell'indice di entropia è:

L'indice di entropia

$$E = -\sum_{i=1}^{k} p_i \ln p_i$$

Data una distribuzione, E è la somma
(con il segno cambiato)
delle frequenze relative, ciascuna moltiplicata per il proprio logaritmo naturale (ln)

Nei dati del primo esempio

Scuola	*Freq.ass.*	*Freq.rel.*	*%*
Licei	50	0,25	25%
Ist.tecnici	70	0,35	35%
Ist.professionali	80	0,4	40%
TOTALE	200	1	100%

E(1)= -(0,25*ln0,25+0,35*ln0,35+0,4*ln0,4)=
= **1,0805**

Nei dati del secondo esempio

Scuola	*Freq.ass.*	*Freq.rel.*	*%*
Licei	40	0,2	20%
Ist.tecnici	120	0,60	60%
Ist.professionali	40	0,2	20%
TOTALE	200	1	100%

E(2)= -(0,2*ln0,2+0,6*ln0,6+0,2*ln0,2)=
= **0,9503**

In conclusione

- E(1) > E(2)
- Dunque:
- La distribuzione delle matricole dell'anno scorso è sensibilmente **meno variabile** rispetto alla distribuzione delle matricole di quest'anno

Un altro indice di variabilità: l'indice di eterogeneità

$$ET = 1 - \sum_{i=1}^{k} p_i^2$$

ET(1)= 1-(0,25^2+0,35^2+0,4^2)=
= **0,655**

ET(2)= 1-(0,2^2+0,6^2+0,2^2)=
= **0,560**

ET(1)>ET(2)

La scala ordinale

- Un dirigente scolastico di una scuola media vuole valutare i rendimenti in italiano, a fine anno, di due classi: la Prima A e la Prima B
- Il dirigente scolastico vuole vedere quale delle due classi è più variabile
- Il d.s. dispone delle distribuzioni dei giudizi nelle due classi:

La Prima A (n=16)

Giudizio	Frequenza assoluta	Frequenza cumulata	%	Percentuale cumulata
Insufficiente	2	2	12,5	**12,5**
Scarso	2	4	12,5	**25,0**
Sufficiente	6	10	37,5	**62,5**
Discreto	2	12	12,5	**75,0**
Buono	4	16	25,0	**100,0**
	16		100,0	

La Prima B (n=24)

Giudizio	Frequenza assoluta	Frequenza cumulata	%	Percentuale cumulata
Insufficiente	2	2	8,3	**8,3**
Scarso	4	6	16,7	**25,0**
Sufficiente	8	14	33,3	**58,3**
Discreto	4	18	16,7	**75,0**
Buono	4	22	16,7	**91,7**
Ottimo	2	24	8,3	**100,0**
	24		100,0	

Gli indici di variabilità

- Per dati su scala ordinale, gli indici di variabilità – analogamente a quanto accadeva per la mediana – sono indici di posizione, cioè si basano sull'ordine della variabile nella distribuzione
- Sono sostanzialmente due: l'intervallo di variazione (o RANGE) e la DIFFERENZA INTERQUARTILE

Gli indici di variabilità

- RANGE = differenza tra il massimo e il minimo della distribuzione
- DIFFERENZA INTERQUARTILE = differenza tra il terzo e il primo quartile

La Prima A (n=16)

Giudizio	Frequenza assoluta	Frequenza cumulata	%	Percentuale cumulata
Insufficiente	2	2	12,5	**12,5**
Scarso	2	4	12,5	**25,0**
Sufficiente	6	10	37,5	**62,5**
Discreto	2	12	12,5	**75,0**
Buono	4	16	25,0	**100,0**
	16		100,0	

RANGE = Tra "insufficiente" e "buono"

DIFF.INTERQUARTILE = Tra "scarso" e "discreto"

La Prima B (n=24)

Giudizio	Frequenza assoluta	Frequenza cumulata	%	Percentuale cumulata
Insufficiente	2	2	8,3	**8,3**
Scarso	4	6	16,7	**25,0**
Sufficiente	8	14	33,3	**58,3**
Discreto	4	18	16,7	**75,0**
Buono	4	22	16,7	**91,7**
Ottimo	2	24	8,3	**100,0**
	24		100,0	

RANGE = Tra "insufficiente" e "ottimo"

DIFF.INTERQUARTILE = Tra "scarso" e "discreto"

Conclusione: la Prima B è più variabile della Prima A nelle fasce "estreme" (è maggiore il RANGE), ma è ugualmente variabile nelle fasce intermedie (la DIFF.INTERQUARTILE è la stessa)

Dati (almeno) su scala ad intervalli: l'età in anni compiuti

Matricole: quest'anno

Età anni	Freq.ass.	%
19	**18**	**9**
20	**160**	**80**
21	**18**	**9**
22	**4**	**2**
Totale	**200**	**100**

Media = 20,04

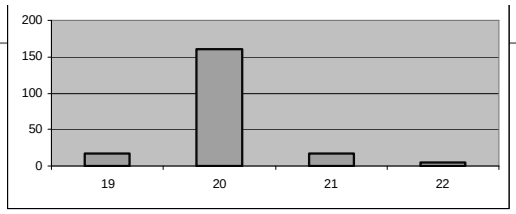
Matricole: l'anno precedente

Età anni	Freq.ass.	%
19	**30**	**15**
20	**170**	**85**
21	**0**	**0**
22	**0**	**0**
Totale	**200**	**100**

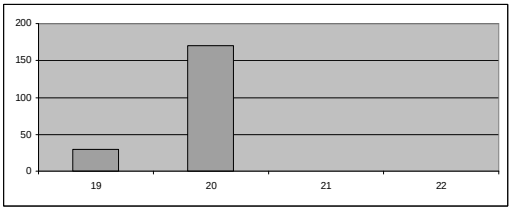
Media = 19,85

Oltre ad avere medie diverse, quale delle due distribuzioni è più variabile?

(1)



(2)



A prima vista, sembrerebbe più variabile la distribuzione (1), cioè quella di quest'anno; mentre la seconda sembra più concentrata sul valore "20 anni"

Come si misura la variabilità (o dispersione)?

- Se siamo (almeno) su scala ad intervalli, l'indice che misura la variabilità (attorno alla media) si chiama SCARTO QUADRATICO MEDIO (o DEVIAZIONE STANDARD);
- Lo S.Q.M. è la media degli scarti rispetto alla media aritmetica, e si calcola con la seguente formula:

Lo scarto quadratico medio (σ):

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{k} \left(x_i - \bar{x}\right)^2 n_i}{n}}$$

Con una distribuzione degli *n* elementi in *k* classi

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - \bar{x}^2}$$

La formula è equivalente, ma più agevole dal punto di vista calcolatorio

Quale delle due distribuzioni è più variabile?

- Deviazione standard (1) = **0,508** anni (media degli scarti)
- Deviazione standard (2) = **0,357** anni (media degli scarti)
- La distribuzione (1) è più variabile della distribuzione (2), come avevamo congetturato

Le misure <u>relative</u> della variabilità

- La deviazione standard della prima distribuzione è maggiore della d.s. della seconda distribuzione
- Dobbiamo però tenere presente che le medie, quindi gli ordini di grandezza, sono diversi (le medie erano 20,04 e 19,85)
- La deviazione standard è una misura assoluta, che non tiene conto del fatto che le medie possono essere diverse
- Sarebbe opportuno disporre anche di una **misura relativa** della variabilità

Il coefficiente di variazione (CV)

- Se siamo su scala a rapporti (è questo il nostro caso, perché la variabile "età" è una differenza tra due "anni di nascita", e dunque è su scala a rapporti), si può calcolare il CV;
- Il CV è il rapporto tra la Deviazione standard di una distribuzione e la rispettiva media:

Il coefficiente di variazione (CV)

$$CV(X) = \frac{\sigma(X)}{\bar{x}}$$

CV (1) = **0,508/20,04 = 0,025**

CV (2) = **0,357/19,85 = 0,018**

Pertanto, possiamo concludere che la distribuzione (1) è PIÙ VARIABILE della distribuzione (2) NON SOLO IN SENSO ASSOLUTO, MA ANCHE IN SENSO RELATIVO.

Riassumendo:

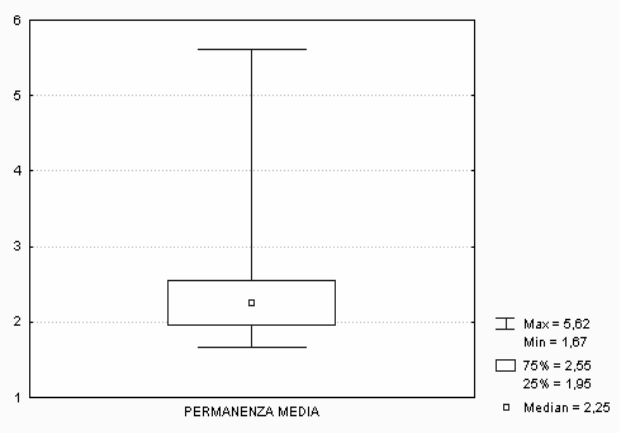
Scala di misura	**Nominale**	**Ordinale**	**Ad intervalli**	**A rapporti**
Indici di variabilità ammissibili	**Indice di entropia** **Indice di eterogeneità**	Idem + **Range**, **Differenza interquartile**	Idem + **Scarto quadratico medio (deviazione standard)**	Idem + **Coefficiente di variazione (CV)**

LA DIFFERENZA INTERQUARTILE E IL "BOX-PLOT"

- Come abbiamo visto, la DIFFERENZA INTERQUARTILE è la differenza tra il PRIMO ed il TERZO quartile della distribuzione.
- La tabella che segue mostra la "permanenza media" dei turisti in alcune grandi città europee:

Città	Permanenza media (giorni)
Amsterdam	2,19
Antwerp	4,15
Barcelona	2,25
Berlin	2,52
Bern	2,05
Budapest	3,23
Düsseldorf	1,83
Firenze	2,31
Frankfurt	1,85
Genere	2,53
Hamburg	1,89
Helsinki	1,82
London	5,62
Milano	2,61
Oslo	1,67
Paris	2,24
Roma	4,48
Salzburg	1,95
Venezia	2,26
Wien	2,55
Zürich	2,09

BOX-PLOT



- Media = 2,58
- Mediana = 2,25

- Come si può osservare nel *box-plot*, si tratta di una distribuzione FORTEMENTE ASIMMETRICA (il *RANGE*, cioè la differenza tra il minimo e il massimo, è molto grande), ma al tempo stesso poco variabile.

- Il valore della *DIFFERENZA INTERQUARTILE* (rappresentata nel grafico dalla "lunghezza" della scatola) è infatti pari ad appena **0,6**.