9. GLI INDICI DI DISPERSIONE

Ritorniamo alla congettura del professor Curiosi (pagina 354) riguardo alle sue nuove classi I A e I B.

Il professore aveva avuto l'impressione, da una iniziale sommaria conoscenza,
che in una di esse gli studenti fossero "meno omogenei nella preparazione":
che ci fosse, insomma, un gruppo abbastanza nutrito di allievi molto bravi e un altro gruppo sostanzioso di scarsi.
Nell'altra classe la situazione gli era sembrata diversa, più equilibrata.

Dopodiché il professore aveva somministrato alle due classi il medesimo test di ingresso,
che aveva fatto registrare i punteggi seguenti (M = media):

I A 51 62 42 58 60 68 61 68 64 70 71 60 51 62 41 51 36 47 58 73 37 54 63 65 ($M \approx 57{,}2$)

I B 45 48 51 63 51 60 29 52 47 41 52 50 56 62 57 70 55 64 59 55 67 ($M = 54$)

Ci domandiamo ora:
esisterà un indicatore statistico adeguato a valutare se il test effettuato conferma l'impressione iniziale?

Un primo indicatore di "dispersione" (= di "sparpagliamento" dei dati) potrebbe essere
la *differenza fra il dato massimo e il dato minimo* in ciascuna delle due classi.
Vediamo che

- per la I A questa differenza, detta in statistica "**campo di variabilità**", vale $73 - 36 = 37$
- mentre in I B vale $70 - 29 = 41$.

$$\textit{campo di variabilità} = \textit{dato massimo} - \textit{dato minimo} = x_{\text{MAX}} - x_{\text{min}}$$

A giudicare dal "campo di variabilità", sembrerebbero quindi più disomogenee le prestazioni della I B …

… tuttavia, va osservato che il "campo di variabilità" tiene conto di DUE SOLI valori (quelli estremi)
mentre non risente per nulla di tutti i valori intermedi … la presenza, nella classe, anche di *un singolo*
caso isolato di alunno molto bravo o molto poco preparato potrebbe allora condizionarlo pesantemente.
Le prestazioni della "massa" degli allievi non influiscono in alcun modo sul calcolo di questo indicatore!

Riflettiamo. Quello che veramente ci interessa è di investigare
in quale delle due classi i valori "sono mediamente più lontani dalla media aritmetica".

Potremmo allora pensare, per ciascuna classe, di elencare tutti gli "scarti dalla media".

I A		51	62	42	58	…
$M \approx 57{,}2$	Scarti	–6,2	+4,8	–15,2	+0,8	…
I B		45	48	51	63	…
$M = 54$	Scarti	–9	–6	–3	+9	…

Questo sarebbe un buon inizio, ma poi?
Se ora andassimo a calcolare la *media aritmetica di questi scarti*, per entrambe le classi *otterremmo 0*!
E certo! Come sappiamo, infatti, la somma algebrica degli scarti dalla media aritmetica è sempre 0.

Sorge allora l'idea di calcolare la media aritmetica … *non* degli scarti, ma del VALORE ASSOLUTO di questi.
Tale media si dice "**scarto medio**" o (più correttamente) "**scarto assoluto medio**".

$$\textit{scarto medio} = \textit{media aritmetica dei valori assoluti degli scarti dalla media aritmetica} = \frac{|x_1 - M| + |x_2 - M| + \ldots + |x_n - M|}{n} \quad (\textit{è più corretto dire: "scarto assoluto medio"})$$

Così facendo, otteniamo (verificalo con un foglio elettronico!) $\textit{scarto}(\text{IA}) \approx 8{,}74$; $\textit{scarto}(\text{IB}) \approx 7{,}05$.

Vediamo di trarre qualche conclusione.
Per la I A, abbiamo ottenuto $\textit{campo di variabilità}\ (\text{IA}) = 37$; $\textit{scarto assoluto medio}\ (\text{IA}) \approx 8{,}74$
e per la I B $\textit{campo di variabilità}\ (\text{IB}) = 41$; $\textit{scarto assoluto medio}\ (\text{IB}) \approx 7{,}05$

♪ **La I A ha uno scarto assoluto medio maggiore**:
i punteggi sono mediamente più lontani, in questa classe, dalla media aritmetica della classe,
segno della presenza "importante" di fasce di allievi che si allontanano alquanto dalla media

♫ D'altra parte, il **campo di variabilità è maggiore per la I B**:
di ciò è responsabile il povero alunno che, purtroppo, ha conseguito un punteggio bassissimo (29 punti).

Anziché fare la media dei valori assoluti degli scarti,
avremmo potuto anche *elevare ciascuno scarto al quadrato*, ottenendo così un valore certamente positivo,
per poi fare la media aritmetica dei QUADRATI degli scarti (detta "varianza").

$$\textit{varianza} = \textit{media aritmetica dei quadrati degli scarti dalla media aritmetica} = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}$$

In questo modo avremmo avuto $\textit{varianza}\ (\text{IA}) \approx 109{,}8$; $\textit{varianza}\ (\text{IB}) \approx 83{,}2$

Varianza *maggiore* comporta *maggiore dispersione dei dati rispetto alla media della popolazione*:
la varianza, in accordo con lo scarto assoluto medio, indica dunque nella I A la classe più disomogenea.

Son pronto a scommettere che la "varianza" ti appare d'istinto più "antipatica" rispetto allo "scarto assoluto medio", che a prima vista sembra assai più semplice e più "spontaneo" da usare, come indice di dispersione.
Tuttavia, ti segnalo che nella pratica si preferisce invece utilizzare la "varianza", e ancora di più
la sua radice quadrata che è chiamata "scarto quadratico medio", anziché lo "scarto assoluto medio".

I motivi per cui la "varianza" ha un rilievo speciale in statistica sono parecchi.
Qui ci limitiamo a citarne soltanto due.

1) La **varianza** $\dfrac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}$

   è legata alla media aritmetica in modo assai peculiare.
   Infatti si può dimostrare che essa è sempre inferiore a qualsivoglia analoga quantità
   $\dfrac{(x_1 - a)^2 + (x_2 - a)^2 + \ldots + (x_n - a)^2}{n}$ nella quale gli scarti vengano calcolati,
   invece che rispetto alla media aritmetica M, rispetto ad un altro qualsiasi valore a.

   Lo "**scarto assoluto medio**" dal canto suo **si ricollega piuttosto ad *un altro* indice** di posizione centrale:
   la mediana. In effetti la quantità $\dfrac{|x_1 - a| + |x_2 - a| + \ldots + |x_n - a|}{n}$ è minima (come si potrebbe dimostrare)
   quando il valore a è la *mediana*, NON la media aritmetica dei dati $x_1, x_2, \ldots, x_n$.

2) **La varianza è il quadrato dello "scarto quadratico medio"**, di cui andiamo a parlare qui di seguito,
   e lo "scarto quadratico medio" ha un'importanza colossale in svariate questioni,
   come la teoria degli errori di misura.

Lo "**scarto quadratico medio**" o "**deviazione standard**" è la radice quadrata della varianza:

$$\textit{scarto quadratico medio o deviazione standard} = \textit{radice quadrata della varianza} = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}} = \textit{media quadratica degli scarti}$$

Lo scarto quadratico medio viene generalmente indicato con σ ("sigma"), e la varianza con σ^2.

Nell'esempio precedentemente considerato dei punteggi delle due classi I A e I B, si ha:

$$\sigma^2(\text{I A}) = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n} \approx \frac{(51 - 57{,}2)^2 + (62 - 57{,}2)^2 + \ldots + (65 - 57{,}2)^2}{24} \approx 109{,}8$$

$$\sigma^2(\text{I B}) = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n} = \frac{(45 - 54)^2 + (48 - 54)^2 + \ldots + (67 - 54)^2}{21} \approx 83{,}2$$

da cui $\sigma(\text{I A}) = \sqrt{\sigma^2(\text{I A})} \approx 10{,}5$; $\sigma(\text{I B}) = \sqrt{\sigma^2(\text{I B})} \approx 9{,}1$

Se i dati provengono da una tabella con le frequenze, evidentemente sarà, dette f_i le frequenze (assolute):

$$\sigma = \sqrt{\frac{(x_1 - M)^2 f_1 + (x_2 - M)^2 f_2 + \ldots + (x_p - M)^2 f_p}{f_1 + f_2 + \ldots + f_p}}$$

Le ragioni per cui lo scarto quadratico medio è preferito alla varianza sono sostanzialmente due.

1) La prima è che, se i dati sono, ad esempio, dei metri, la "varianza" sarebbe espressa in "metri quadrati",
   e lo scarto quadratico medio invece ancora in metri. Insomma,
   lo scarto quadratico medio ha il pregio di avere la stessa unità di misura dei dati dei quali proviene.

2) La seconda ragione è il **ruolo cruciale dello scarto quadratico medio nella cosiddetta "gaussiana"**,
   alla quale accenneremo parlando, più avanti, di "**errori di misura**".

Per il calcolo dello scarto quadratico medio, anziché la formula $\sigma = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}}$,

si può anche utilizzare una formula equivalente, più comoda, che è $\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - M^2}$.

Per *confrontare* due distribuzioni in quanto alla loro "variabilità", alla loro "dispersione",
si utilizza un indice che è detto "coefficiente di variazione" (di solito espresso come percentuale,
non calcolabile se la media dei dati è 0, e comunque poco significativo quando la media dei dati è vicina a 0):

$$\textit{coefficiente di variazione} = \frac{\textit{scarto quadratico medio}}{|\textit{media aritmetica}|} = \frac{\sigma}{|M|}$$

NOTA - **Il coefficiente di variazione**, essendo il rapporto fra due quantità, σ e $|M|$, che sono espresse nella stessa unità di misura, **è un numero puro, senza unità di misura** (si dice che è "adimensionale").

Ad esempio, se si vanno a misurare i pesi dei bambini nati in un certo periodo in un grande ospedale, e simultaneamente i pesi delle loro mamme, si osserverà certamente una deviazione standard molto inferiore nell'insieme dei bambini … Per forza! Infatti i bambini appena nati pesano soltanto due-tre o quattro chili … quindi anche gli scarti dalla media dei loro pesi saranno piccolini!!! Volendo confrontare le due "variabilità" (quella dei pesi dei neonati con quella dei pesi delle mamme) si farà ricorso allora al coeff. di variazione.

RIASSUNTO SCHEMATICO (INDICI DI DISPERSIONE)

Indicatori di "DISPERSIONE" o di "VARIABILITÀ": ci dicono
QUANTO, GLOBALMENTE, I DATI SONO LONTANI DALLA LORO MEDIA ARITMETICA *M*.

Ogni indicatore di dispersione ha la proprietà di essere maggiore quando i dati si allontanano maggiormente, nel loro complesso, dalla centralità.

CAMPO DI VARIABILITA' = $= \textit{dato massimo} - \textit{dato minimo} = x_{MAX} - x_{min}$	E' un indicatore piuttosto "grezzo", perché dipende esclusivamente dai due valori estremi ignorando quelli intermedi	EXCEL, OPENOFFICE: **MAX()–MIN()**
SCARTO MEDIO = SCARTO ASSOLUTO MEDIO = $= \textit{m. aritm. dei val. ass. degli scarti dalla m. aritm.} = \frac{|x_1 - M| + |x_2 - M| + \ldots + |x_n - M|}{n}$	Sarebbe minimo qualora al posto della media *M* ci fosse, nella formula, la mediana	EXCEL, OPENOFFICE: **MEDIA.DEV()**
VARIANZA = $= \sigma^2 = \textit{m. aritm. dei quadr. degli scarti dalla m. aritm.} = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}$	Ha il difetto di non essere espressa nella stessa unità di misura dei dati	EXCEL, OPENOFFICE: **VAR.POP()** (NOTA)
SCARTO QUADR. MEDIO o DEVIAZ. STANDARD = $= \sigma = \textit{radice quadrata della varianza} = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}} = \textit{media quadratica degli scarti}$	E' l'indicatore di dispersione più utilizzato in statistica; è espresso nella stessa unità di misura dei dati, e ha un'importanza decisiva nella teoria degli errori di misura, e, in generale, nelle distribuzioni che tendono a identificarsi con la cosiddetta "gaussiana"	EXCEL, OPENOFFICE: **DEV.ST.POP()** (NOTA)
Comodissima formula alternativa: $\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - M^2}$		
COEFF. DI VARIAZ. = $\frac{\textit{scarto quadratico medio}}{|\textit{media aritmetica}|} = \frac{\sigma}{|M|}$	E' un numero puro, senza unità di misura, ottimo per confrontare fra loro distribuzioni differenti.	

NOTA su alcune funzioni statistiche nel foglio elettronico

$$\text{VAR.POP} = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}, \quad \text{VAR} = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n-1}$$

VAR è dunque, per il foglio elettronico, la cosiddetta "**varianza corretta**", ossia un indicatore statistico che, calcolato su di un campione, permette di stimare meglio la varianza incognita dell'intera popolazione. La "varianza corretta" e l'analoga "deviazione standard corretta" si utilizzano quindi in statistica inferenziale … questo tuttavia è un discorso che, se affrontato seriamente, presenta grande interesse ma anche una certa difficoltà.

TM	TF	BM	BF
96,3	96,4	58	57
96,7	96,7	63	57
96,9	96,8	64	59
97	97,2	64	61
97,1	97,2	64	61
97,1	97,4	65	62
97,1	97,6	66	62
97,2	97,7	66	64
97,3	97,7	67	64
97,4	97,8	67	64
97,4	97,8	68	65
97,4	97,8	68	65
97,4	97,9	68	66
97,5	97,9	69	66
97,5	97,9	69	68
97,6	98	70	68
97,6	98	70	68
97,6	98	70	69
97,7	98	70	69
97,8	98	70	69
97,8	98,1	70	69
97,8	98,2	71	70
97,8	98,2	71	71
97,9	98,2	71	72
97,9	98,2	71	73
98	98,2	71	73
98	98,2	72	73
98	98,3	72	73
98	98,3	72	73
98	98,3	72	74
98	98,4	73	74
98,1	98,4	73	75
98,1	98,4	73	76
98,2	98,4	73	76
98,2	98,4	73	77
98,2	98,5	74	77
98,2	98,6	74	77
98,3	98,6	74	77
98,3	98,6	74	77
98,4	98,6	75	78
98,4	98,7	75	78
98,4	98,7	75	78
98,4	98,7	75	79
98,5	98,7	76	79
98,5	98,7	77	79
98,6	98,7	77	79
98,6	98,8	78	79
98,6	98,8	78	79
98,6	98,8	78	80
98,6	98,8	78	80
98,6	98,8	78	81
98,7	98,8	78	81
98,7	98,8	78	81
98,8	98,9	79	82
98,8	99	80	82
98,8	99	80	83
98,9	99,1	81	83
99	99,1	81	84
99	99,2	82	84
99	99,2	82	84
99,1	99,3	82	85
99,2	99,4	83	86
99,3	99,9	83	87
99,4	100	84	89
99,5	100,8	86	89

I dati qui a sinistra sono tratti dal
Journal of the American Medical Association, vol. 268.

Di 130 soggetti, 65 uomini e 65 donne,
rappresentanti un campione casuale della popolazione locale,
sono stati misurati

- la temperatura corporea, in gradi Fahrenheit,
- e il numero di battiti cardiaci al minuto.

Utilizza un foglio elettronico per calcolare, di ciascuna colonna,

- la media
- lo scarto quadratico medio o deviazione standard
- lo scarto quadratico medio "corretto"
- il coefficiente di variazione
  (prendi lo sc. q. m. "non corretto" per determinarlo)

Le risposte sono qui in fondo alla pagina, capovolte,
ma tu guardale solo alla fine!

Per trovare altri gruppi di dati reali "grezzi" su cui lavorare,
puoi ad esempio consultare le pagine web

www.amstat.org/publications/jse/jse_data_archive.htm
e
http://www2.stetson.edu/~jrasp/data.htm

LA STATISTICA di *Trilussa*

Sai ched'è la statistica? È 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che spósa.
Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perchè c'è un antro che ne magna due.

In tutti i casi seguenti
c'è chi mangia 0 polli e chi ne mangia di più:
secondo te, quali situazioni sono
più equilibrate, meno ingiuste?

Prova a calcolare media,
scarto quadratico medio,
coefficiente di variazione …

2 persone: 0 polli, 2 polli
3 persone: 0 1 2
5 persone: 0 1 1 1 2
6 persone: 0 0 1 1 2 2
4 persone: 0 0 0 4
6 persone: 0 0 0 1 1 4
3 persone: 0 2 4
5 persone: 0 3 3 3 6

media	74,15384615	73,36923077	98,39384615	98,10461538
devstandard	8,042637855	5,829815225	0,737746449	0,693359884
coeffvariaz	0,108458809	0,079458585	0,007497892	0,007067556
devstandardcorr	8,105227421	5,875184122	0,743487753	0,698755762