UNIVERSITA' DEGLI STUDI DI PERUGIA
DIPARTIMENTO DI FILOSOFIA SCIENZE SOCIALI UMANE E DELLA FORMAZIONE
Corso di Laurea in Scienze per l'Investigazione e la Sicurezza

5. GLI INDICI DI VARIABILITA'

Prof. Maurizio Pertichetti

Statistica sociale

5. GLI INDICI DI VARIABILITA'

Un valore centrale, utile a precisare la "posizione" di una distribuzione di frequenze , vale a dire la "tendenza centripeta" dei dati, da solo non è però sufficiente a caratterizzarla, perché non consente di percepire quali siano le effettive grandezze dei valori da cui esso è ricavato. Si sa che due distribuzioni possono avere uguale media, ma essere composte da valori diversi.

Consideriamo il seguente esempio di tre studenti che hanno superato ciascuno tre esami:

A	24	24	24
B	23	24	25
C	18	24	30

è facile vedere che il voto medio e quello mediano per ciascun studente è pari a 24.

Sintetizzando dunque un insieme di dati con un unico valore centrale, si deve pertanto prendere atto che si vengono a perdere informazioni essenziali sulla natura dei dati stessi.

Ne consegue allora che è necessario disporre di un ulteriore parametro che, in modo altrettanto sintetico, dia conto anche del modo con il quale i valori si distribuiscono tra le diverse modalità del carattere, ovvero misuri la loro **variabilità**, la quale altro non è se non la caratteristica di un fenomeno ad essere predisposto ad assumere misure diverse.

Gli **indici di variabilità** rispondono quindi all'esigenza di rappresentare in modo sintetico quanto possono essere dissimili tra loro i valori di una distribuzione.

I valori assunti dagli indici risultano diversi se si misura la variabilità:

- delle singole osservazioni rispetto ad un valore centrale. Ovvero, si calcola determinando gli scostamenti o scarti medi, tra ciascuna modalità del carattere e un valore medio, a loro volta sintetizzati come media. I questo caso la variabilità viene intesa come **dispersione**;
- tra tutte le modalità prese a due a due. Ovvero, si calcola determinando le differenze medie, cioè le differenze in valore assoluto, tra le modalità del carattere prese appunto a coppia, e sintetizzate come media. I questo caso la disuguaglianza viene intesa come **disuguaglianza**.

I conseguenti indici che si ottengono, così come già detto per tutti quelli in generale utilizzati dalla statistica, si distinguono in:

- **indici assoluti**, che sono espressi nella stessa unità di misura del fenomeno in esame e non si adattano a consentire comparazioni;
- **indici relativi**, che prescindono dall'unità di misura del fenomeno esaminato e sono dunque adatti per effettuare confronti tra distribuzioni diverse. Come già detto, si ottengono rapportando tra loro due misure assolute o un indice assoluto al suo massimo. In tale ultimo caso si ottengono gli **indici normalizzati**, la cui caratteristica è quella di assumere valori che variano tra 0 ed 1.

Tutti gli indici di variabilità devono comunque rispondere alle condizione di:

- assumere solo valori positivi (non ha senso parlare di dispersione o variabilità negative) ed essere nulli, quando tutti i termini della distribuzione sono uguali fra loro $X_1 = X_2 = \ldots = X_k$;
- registrare valori crescenti all'aumentare della variabilità, ciò in quanto una misura di questa deve essere tanto più grande quanto maggiore è la diversità fra i termini.

Questo corso nei propri approfondimenti sulle misure della variabilità non tratta delle differenze medie.

Il campo di variazione

Un indice assoluto della diversità di dati di una successione, di immediata percezione e assai semplice da calcolarsi, è dato dal **campo di variazione** (o **range**), che si ottiene come differenza tra il valore massimo e il valore minimo della successione. Di fatto costituisce l'ampiezza dell'intervallo dei dati. In simboli:

$$R = x_{max} - x_{min}$$

L'indice in questione è poco utilizzato in quanto prende in considerazione solo la dispersione esistente tra i valori estremi della distribuzione, per cui, oltre a non tener conto di tutte le informazioni su una variabile statistica, risente di eventuali valori anomali nei dati.

Data la seguente serie: 1 2 3 6 9 10 15
- Il valore più alto è 15, il più basso 1
- Il range è dato dalla differenza tra i due valori R = 15-1 = 14

Data la seguente serie: -11 -2 3 9 10 18
- Il valore più alto è 18, il più basso -11
- Il range è dato dalla differenza tra i due valori R = 18 - (-11) = 18 + 11 = 29

Il campo di variazione, che è espresso nella stessa unità di misura dei dati, tanto più è piccolo tanto più i dati sono concentrati, viceversa tanto più è grande tanto più i dati sono dispersi.

Lo scostamento semplice medio dalla media aritmetica

Lo **scostamento semplice medio dalla media aritmetica** è un indice di variabilità dato dalla media aritmetica dei valori assoluti degli scarti dalla media aritmetica, ovvero a seconda che si abbia una successione di dati tra loro diversi o una distribuzione di frequenza:

$$S_\mu = \frac{\sum_{i=1}^{k} |x_i - \mu|}{k} \qquad S_\mu = \frac{\sum_{i=1}^{k} |x_i - \mu| n_i}{n}$$

Questi indici sono calcolati considerando i valori assoluti degli scarti, in quanto nel caso della media aritmetica, come sappiamo, la media degli scarti, presi con il loro segno, è uguale a zero. Va detto che l'uso dei dei valori assoluti rende questa misura poco utilizzata.

Per conoscenza va detto che analogo allo scostamento semplice medio dalla media aritmetica è lo scostamento semplice medio dalla mediana, che si ottiene utilizzando le stesse formule ma ponendo la mediana al posto della media aritmetica.

Esempio di calcolo dello scostamento medio dalla media aritmetica

Data la seguente successione: 4 5 15 23 28

$$\mu = \frac{4 + 5 + 15 + 23 + 28}{5} = \frac{75}{5} = 15{,}00$$

$$S_\mu = \frac{|4-15| + |5-15| + |15-15| + |23-15| + |28-15|}{5} = \frac{42}{5} = 8{,}4$$

Il valore così ricavato indica che i dati della distribuzione si discostano, dalla loro media aritmetica, mediamente di 8,4 unità in più o in meno.

La varianza, lo scostamento quadratico medio e la devianza

La **varianza** di una qualunque successione di dati statistici è una misura di dispersione che si ottiene come **media dei quadrati degli scarti dalla media aritmetica**. In simboli, ovvero a seconda che si abbia una successione di dati o una distribuzione di frequenza:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2}{k} \qquad \sigma^2 = \frac{\sum_{i=1}^{k} (xi - \mu)^2 n_i}{n}$$

E' anche possibile usare una formula semplificata ottenuta sviluppando la precedente (sviluppo che qui si omette):

$$\sigma^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - \mu^2$$

La varianza essendo espressa come quadrato dell'unità di misura delle osservazioni fa emergere tuttavia una complicazione che si manifesta nel concreto nella impossibilità di rappresentare su uno stesso diagramma i valori della varianza stessa e quelli della distribuzione delle osservazioni.

Per ovviare alla complicazione anzidetta, si ricorre allora all'uso della radice quadrata della varianza, che consente di ottenere un importante indice di variabilità, assai utilizzato, denominato **scostamento quadratico medio** o **deviazione standard**, riportato alla stessa unità di misura delle osservazioni. In simboli, ovvero a seconda che si abbia una successione di dati o una distribuzione di frequenza:

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

Lo scostamento quadratico medio o deviazione standard è un indice assai rappresentativo del livello di addensamento dei dati intorno al loro valore medio ed è preferibile allo scostamento semplice medio perché da evidenza anche alle variazioni più piccole delle distribuzioni, ovvero rappresenta una misura della variabilità più sensibile.

Da ultimo va infine va dato un rilievo al numeratore della varianza che si presenta anch'esso come un'ulteriore misura di dispersione denominata **devianza**. Per un carattere X la sua espressione analitica, a seconda che si abbia una successione di dati o una distribuzione di frequenza::

$$D(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \qquad D(X) = \sum_{i=1}^{k} (xi - \mu)^2 n_i$$

Esempi di calcolo dello scostamento quadratico medio dalla media aritmetica

Data la seguente successione:

4 5 15 23 28

$$\mu = \frac{4 + 5 + 15 + 23 + 28}{5} = \frac{75}{5} = 15{,}00$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
4	-11,0	121
5	-10,0	100
15	0,0	0
23	8,0	64
28	13,0	169
totale		454

$$\sigma^2 = \frac{D(X)}{n} = \frac{454}{5} = 90{,}80$$

$$\sigma = \sqrt{\sigma^2} = 9{,}53$$

Data la seguente successione:

3,9 8,9 4,8 5,0 9,9

$$\mu = \frac{3{,}9 + 8{,}9 + 4{,}8 + 5{,}0 + 9{,}9}{5} = \frac{32{,}5}{5} = 6{,}50$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$
3,9	-2,6	6,76
8,9	2,4	5,76
4,8	-1,7	2,89
5	-1,5	2,25
9,9	3,4	11,56
totale		29,22

$$\sigma^2 = \frac{D(X)}{n} = \frac{29{,}22}{5} = 5{,}84$$

$$\sigma = \sqrt{\sigma^2} = 2{,}42$$

Data la seguente distribuzione di frequenze:

x_i	n_i
173	14
178	18
183	28
188	33
193	17
198	15
tot	125

$\mu = 186$

$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
-12,64	159,77	2.236,77
-7,64	58,37	1.050,65
-2,64	6,97	195,15
2,36	5,57	183,80
7,36	54,17	920,88
12,36	152,77	2.291,54
totale		6.878,80

$x_i^2 n_i$
419.006
570.312
937.692
1.166.352
633.233
588.060
4.314.655

$$\sigma^2 = \frac{D(X)}{n} = \frac{6.878{,}80}{125} = 55{,}03$$

$$\sigma^2 = \frac{\Sigma x_i^2 n_i}{n} - \mu^2 = \frac{4.314.655}{125} - 34.462 = 55{,}03 \qquad \sigma = \sqrt{\sigma^2} = 7{,}418$$

Data la seguente distribuzione di frequenze:

x_i	n_i
1	60
2	80
3	30
4	25
5	5
tot	200

$x_i n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
60	-1,18	1,38	82,84
160	-0,18	0,03	2,45
90	0,83	0,68	20,42
100	1,83	3,33	83,27
25	2,83	7,98	39,90
435	totale		228,88

$x_i^2 n_i$
60
320
270
400
125
1.175

$\mu = 435 / 200 = 2{,}18$

$$\sigma^2 = \frac{D(X)}{n} = \frac{228{,}88}{200} = 1{,}14 \qquad \sigma^2 = \frac{\Sigma x_i^2 n_i}{n} - \mu^2 = \frac{1.175}{200} - 4{,}73 = 1{,}14$$

$$\sigma = \sqrt{\sigma^2} = 1{,}070$$

Il coefficiente di variazione

In realtà operare confronti sulla deviazione standard non è di grande aiuto, perché essa dipende fortemente dalla media dei dati su cui è stata calcolata. In questo senso un indice relativo cui si ricorre molto spesso, se i valori della distribuzione sono positivi e comunque la media risulta maggiore di zero, è il **coefficiente di variazione**, un indice relativo che si calcola come rapporto tra scostamento quadratico medio o deviazione standard e media aritmetica. E' in sostanza un numero puro, non espresso in alcuna unità di misura, che consente di effettuare confronti fra distribuzioni diverse per fenomeni omogenei. La sua espressione analitica è:

$$Cv = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{\Sigma (x_i - \mu)^2 n_i}{n}}}{\mu}$$

Questo coefficiente di variazione, espresso in genere in termini percentuali moltiplicando Cv per 100, è indipendente dall'unità di misura, ovvero è un numero puro utilizzato, sia per misurare la variazione media del fenomeno in rapporto alla sua media aritmetica, sia per confrontare la variabilità relativa di un fenomeno in circostanze differenti (ad esempio, la variabilità della distribuzione per età tra le varie regioni, la distribuzione dei redditi per paesi e per anno, la variabilità del peso rispetto al sesso, ...).

Il coefficiente di variazione, inoltre, è necessario come già detto tutte le volte che si intende confrontare la variabilità di due fenomeni espressi in unità di misure diverse (ad esempio, la variabilità del peso rispetto a quella dell'altezza, ecc.).

Esempi di calcolo del coefficiente di variazione

Date le seguenti distribuzioni:

x_i	n_i	x_i*n_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(xi-\mu)^2 n_i$
1	2	2	-2	4	8,00
2	2	4	-1	1	2,00
3	2	6	0	0	0,00
4	2	8	1	1	2,00
5	2	10	2	4	8,00
	10	30			20,00

$\mu = 30/10 =$ 3

SQM = $\sigma = \sqrt{\Sigma(xi-\mu)^2 n_i / n} =$ 1,414

$Cv = \sigma/\mu =$ 0,471 (*100 = 47,1 %)

x_i	n_i	x_i*n_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(xi-\mu)^2 n_i$
1	0	0	-2	4	0,00
2	0	0	-1	1	0,00
3	10	30	0	0	0,00
4	0	0	1	1	0,00
5	0	0	2	4	0,00
	10	30			0,00

$\mu = 30/10 =$ 3

SQM = $\sigma = \sqrt{\Sigma(xi-\mu)^2 n_i / n} =$ 0,000

$Cv = \sigma/\mu =$ 0,000 (*100 = 0 %)

Date le seguenti distribuzioni :

Un fenomeno riferito ai maschi

x_i	n_i	$x_i n_i$	$x_i-\mu$	$(x_i-\mu)^2$	$(x_i-\mu)^2 n_i$
34,5	81	2.794,5	-11,4	129,2	10.464,3
44,5	31	1.379,5	-1,4	1,9	57,9
54,5	36	1.962,0	8,6	74,5	2.683,6
64,5	35	2.257,5	18,6	347,2	12.152,8
	183	8.393,5			25.358,5

$$\mu = 45{,}866 \qquad \sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2 n_i}{n}} = \frac{25.358{,}5}{183} = 11{,}772$$

$$Cv = \frac{\sigma}{\mu} = \frac{11{,}772}{45{,}866} = 0{,}257 \qquad (*100 = 25{,}7\ \%)$$

Stesso fenomeno riferito alle femmine

x_i	n_i	$x_i n_i$	$x_i-\mu$	$(x_i-\mu)^2$	$(x_i-\mu)^2 n_i$
34,5	30	1.035,0	-18,5	343,2	10.294,8
44,5	42	1.869,0	-8,5	72,7	3.052,1
54,5	36	1.962,0	1,5	2,2	78,4
64,5	75	4.837,5	11,5	131,7	9.876,4
	183	9.703,5			23.301,6

$$\mu = 53{,}025 \qquad \sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2 n_i}{n}} = \frac{23.301{,}6}{183} = 11{,}284$$

$$Cv = \frac{\sigma}{\mu} = \frac{11{,}284}{53{,}025} = 0{,}213 \qquad (*100 = 21{,}3\ \%)$$

Per l'esempio fatto i maschi presentano una variazione media del fenomeno intorno alla media aritmetica pari al 25% contro il 21% delle femmine.

Alcuni valori particolari del CV che possono essere utili nello studio di una distribuzione di dati:

- $CV = 0$ in questo caso la deviazione standard è pari a 0. Tutti i dati sono uguali tra loro e la media può essere considerata come un indice perfetto per rappresentarli.
- $CV \geq 0.5$ in questo caso la deviazione standard è più della metà della media. La media, in questo caso, non può essere considerata un buon indice per rappresentare i dati.
- $CV \leq 0.5$ in questo caso la deviazione standard è meno della metà della media. La media, in questo caso, può essere considerata un buon indice per rappresentare i dati.

L'Indice di eterogeneità

Per disporre di indici di variabilità utilizzabili con qualsiasi tipo di carattere (anche qualitativo non ordinabile) occorre che la definizione dell'indice coinvolga solo le frequenze delle diverse modalità, senza richiedere relazioni di ordine fra le modalità stesse. Un esempio è fornito dall'**indice di eterogeneità di Gini**. In formula

$$G = 1 - \sum_{i=1}^{k} f_i^2$$

E' un indice assoluto di eterogeneità in quanto è massimo ossia pari a 1-(1/k) quando le modalità hanno tutte la medesima frequenza o, se si vuole, quando le frequenze sono equidistribuite tra tutte le modalita, mentre è minimo (ossia nullo e quindi c'è massima omogeneità) quando tutte le frequenze si addensano in una sola modalità .

Esempio di calcolo dell'indice assoluto

condizione di massima eterogeneità

x_i	n_i	f	f^2
1	2	0,20	0,04
2	2	0,20	0,04
3	2	0,20	0,04
4	2	0,20	0,04
5	2	0,20	0,04
	10	1,00	0,20

condizione di eterogeneità nulla

x_i	n_i	f	f^2
1	0	0,00	0,00
2	0	0,00	0,00
3	10	1,00	1,00
4	0	0,00	0,00
5	0	0,00	0,00
	10	1,00	1,00

$G = 1 - \Sigma f^2$ $\quad G = 1 - 0,20 = 0,80$ $\quad G = 1 - 1,00 = 0,00$

$k = 5$ $\quad max = 1 - \frac{1}{k} = 0,80$ $\quad min = 0,00$

Per rendere pero confrontabili fra loro due indici calcolati su due diversi caratteri con frequenze n_i diverse delle modalità occorre utilizzare un indice relativo, che si ottiene dividendo l'indice assoluto per il massimo valore che esso può assumere.

Nel caso dell'indice di Gini, essendo $G_{max} = 1-(1/k)$, l'**indice normalizzato G_N** si ottiene:

$$\mathbf{G_N} = \frac{G}{G_{max}} = \frac{G}{1-(1/k)} = \mathbf{G\ \frac{k}{(k-1)}}$$

L'indice così ottenuto è un indice relativo che varia tra 0 che è il minimo e 1 che è il massimo.

$$0 \leq G_N \leq 1$$

Riprendendo l'esempio precedente :

- l'indice assoluto in corrispondenza della massima eterogeneità era $G = 0,8$ per cui:
  $G_N = G * k/(k-1) = 0,80 * 5/4 = 0,80 * 1,25 =$ **1,0**

- l'indice assoluto in corrispondenza della minima eterogeneità era $G = 0$ per cui:
  $G_N = G * k/(k-1) = 0 * 5/4 =$ **0**