ISTITUZIONI DI STATISTICA – A. A. 2007/2008
Marco Minozzo e Annamaria Guolo
Laurea in Economia del Commercio Internazionale
Laurea in Economia e Amministrazione delle Imprese
Università degli Studi di Verona (sede di Vicenza)

Statistica Descrittiva
Soluzioni 6. Indici di variabilità, asimmetria e curtosi

Introduzione

Dato un carattere X, la sua **variabilità** rappresenta l'attitudine ad assumere modalità diverse. Prenderemo in considerazione solamente i caratteri quantitativi. Per i caratteri qualitativi, la variabilità si traduce in mutabilità.
Esistono diversi tipi di indici di variabilità. In ogni caso, un indice di variabilità deve a) annullarsi nel caso in cui tutti i termini siano uguali e b) aumentare di valore al crescere della diversità tra modalità. Distinguiamo tre gruppi di indici di variabilità.

- **Intervalli di variazione**. Sono indici basati sul confronto di alcuni termini della successione ordinata di termini della distribuzione.
  - Campo di variazione (o range):
    $$R = x_{max} - x_{min},$$
    dove x_{max} e x_{min} indicano rispettivamente il massimo ed il minimo valore osservato di X.
  - Scarto interquartilico:
    $$D = Q_3 - Q_1,$$
    ottenuto come differenza tra il terzo ed il primo quartile della distribuzione di X.
- **Scostamenti da un valore medio**. Sono indici che sintetizzano gli scarti tra i termini della distribuzione di X ed un valore centrale.
  - Scostamento semplice medio dalla media aritmetica:
    $$_mS_1 = \frac{\sum_i |x_i - m| \cdot f_i}{\sum_i f_i}.$$
  - Scostamento semplice medio dalla mediana:
    $$_{me}S_1 = \frac{\sum_i |x_i - me| \cdot f_i}{\sum_i f_i}.$$
    Si noti che $_{me}S_1$ è sempre minore di $_mS_1$.
- **Scarto quadratico medio.** È la misura di variabilità più diffusa.
  $$_mS_2 = \sqrt{\frac{\sum_i (x_i - m)^2 \cdot f_i}{\sum_i f_i}}.$$
  Tale misura si indica solitamente con σ.

- **Varianza**. È data dal quadrato dello scarto quadratico medio:

$$V(X) = \sigma^2 = \frac{\sum_i (x_i - m)^2 \cdot f_i}{\sum_i f_i}.$$

  - La varianza si può calcolare sia applicando la formula sopra indicata (metodo diretto) sia attraverso al seguente formula (metodo indiretto):

$$V(X) = \frac{\sum_i x_i^2 \cdot f_i}{\sum_i f_i} - m^2,$$

    vale a dire come media dei quadrati dei valori di X meno la media aritmetica di X elevata al quadrato.
  - Proprietà pitagorica della varianza:

$$\frac{\sum_i (x_i - m)^2 \cdot f_i}{\sum_i f_i} = \frac{\sum_i (x_i - A)^2 \cdot f_i}{\sum_i f_i} + (A - m)^2 = \sigma^2 + (A - m)^2,$$

    dove A è un valore arbitrario.
  - Varianza di una trasformazione lineare:

$$V(a + b \cdot X) = b^2 V(X),$$

    con a e b valori reali arbitrari.
  - Standardizzazione. Dato un carattere X di media m e varianza σ^2, il carattere $U = (X - m)/\sigma$ ha media nulla e varianza pari a 1.

- **Differenze medie.** Sono indici basati sul confronto tra tutti i termini della distribuzione.

  - Differenze medie assolute semplici

$$\Delta = \frac{\sum_{j=1}^N \sum_{i=1, i \neq j}^N |x_i - x_j|}{N(N-1)}$$

    e con ripetizione

$$\Delta_R = \frac{\sum_{j=1}^N \sum_{i=1}^N |x_i - x_j|}{N^2}.$$

  - Differenze medie quadratiche semplici

$${}_2\Delta = \sqrt{\frac{\sum_{j=1}^N \sum_{i=1, i \neq j}^N (x_i - x_j)^2}{N(N-1)}}$$

    e con ripetizione

$${}_2\Delta_R = \sqrt{\frac{\sum_{j=1}^N \sum_{i=1}^N (x_i - x_j)^2}{N^2}}.$$

- **Indici relativi di variabilità**. Gli indici di variabilità esaminati sono assoluti, nel senso che mantengono la dipendenza dalle unità di misura in cui sono osservati i valori di X. Si possono modificare al fine di eliminare tale dipendenza e permettere quindi il confronto tra distribuzioni di fenomeni diversi. Si ottengono allora gli indici relativi seguenti.

  - Range relativo:

$$R_m = \frac{x_{max} - x_{min}}{m}.$$

  - Differenza interquartilica relativa:
$$D_m = \frac{Q_3 - Q_1}{m}.$$
  - Scostamento semplice medio relativo
$$\frac{{}_mS_1}{m}.$$
  - Coefficiente di variazione
$$Cv = \frac{\sigma}{m}.$$

- **Asimmetria**. Dato un carattere X, si valuta il grado di lontananza della distribuzione del carattere da una situazione di simmetria tramite i seguenti indici:
  - coefficiente di skewness di Pearson, per distribuzioni unimodali,
$$s_k = \frac{m - m_0}{\sigma};$$
  - indice β_1 di Pearson
$$\beta_1 = \frac{({}_m\mu_3)^2}{(\sigma^2)^3} = \frac{\{\sum_i (x_i - m)^3 f_i / \sum_i f_i\}^2}{(\sigma^2)^3};$$
  - indice γ_1 di Fisher
$$\gamma_1 = \frac{{}_m\mu_3}{\sigma^3} = \frac{\sum_i (x_i - m)^3 f_i / \sum_i f_i}{\sigma^3}$$

- **Curtosi**. Dato un carattere X, si valuta il grado di aderenza della sua distribuzione a quello di una distribuzione normale, con particolare attenzione alle code della distribuzione stessa, tramite i seguenti indici:
  - indice β_2 di Pearson
$$\beta_2 = \frac{{}_m\mu_4}{\sigma^4} = \frac{\sum_i (x_i - m)^4 f_i / \sum_i f_i}{\sigma^4};$$
  - indice γ_2 di Fisher
$$\gamma_2 = \frac{{}_m\mu_4}{\sigma^4} - 3 = \frac{\sum_i (x_i - m)^4 f_i / \sum_i f_i}{\sigma^4} - 3.$$

Esercizio A.
Ponendo le nove osservazioni in ordine crescente, si ricava che la mediana ed il primo e terzo quartile sono pari, rispettivamente,a

$$me = 173,\ Q_1 = 167,\ Q_3 = 177.$$

Inoltre, la media aritmetica delle osservazioni è pari a $m = 170,89$. Si ricava quindi quanto richiesto, come segue.

a) Range $R = x_{max} - x_{min} = 180 - 156 = 24$.
Scarto interquartilico $D = Q_3 - Q_1 = 177 - 167 = 10$.

b) Scostamento medo semplice dalla media:

$$ {}_mS_1 = \frac{\sum_{i=1}^{9} |x_i - m|}{9} = \frac{|156 - 170,89| + \cdots + |180 - 170,89|}{9} = 6,123. $$

Scostamento medio semplice dalla mediana:

$$ {}_{me}S_1 = \frac{\sum_{i=1}^{9} |x_i - me|}{9} = \frac{|156 - 173| + \cdots + |180 - 173|}{9} = 5,889. $$

Come noto dalla teoria, lo scostamento medio semplice dalla mediana è minore della scostamento medio semplice dalla media.

c) Varianza:

$$ V(X) = \sigma^2 = \frac{\sum_{i=1}^{9} (x_i - m)^2}{9} = \frac{(156 - 170,89)^2 + \cdots + (180 - 170,89)^2}{9} = 51,654. $$

d) La varianza di $1 + 0,9 \cdot X$ è pari a $0,9^2 \cdot V(X) = 41,840$.

Esercizio B.
a) La seguente tabella contiene le informazioni necessarie al calcolo dello scostamento semplice medio dalla mediana ${}_{me}S_1$ e dello scarto quadratico medio dalla media artimetica σ. si indica con x_i il valore centrale della calsse di età, con d_i l'ampiezza della classe, con p_i la frequenza relativa e con P_i la frequenza relativa cumulata.

			Senza titolo		Scuola media inf.	
Classe di età	x_i	d_i	p_i	P_i	p_i	P_i
$[15 - 20)$	17,5	5	0,01	0,01	0,04	0,04
$[20 - 25)$	22,5	5	0,02	0,03	0,09	0,13
$[25 - 30)$	27,5	5	0,03	0,06	0,15	0,28
$[30 - 40)$	35	10	0,10	0,16	0,33	0,62
$[40 - 50)$	45	10	0,29	0,44	0,24	0,86
$[50 - 60)$	55	10	0,38	0,83	0,11	0,97
$[60 - 65)$	62,5	5	0,12	0,95	0,02	0,99
$[65 - 70)$	67,5	5	0,05	1	0,01	1

Utilizzando le informazioni contenute nella tabella, risulta che:

Titolo di Studio	m	σ	m_e	${}_{me}S_1$
Senza titolo	49,88	10,69	51,435	8,623
Scuola media inf.	37,43	10,99	36,440	8,790

b) Dai risultati in tabella si nota che la seconda distribuzione presenta una maggiore variabilità, sia che essa venga misurata tramite lo scostamento medio semplice dalla mediana che tramite lo scarto quadratico medio.

Esercizio C.
a) La media delle 5 osservazioni risulta essere pari a $m = 3574,2$. La varianza calcolata secondo il metodo diretto è pari a

$$ \sigma^2 = \frac{\sum_{i=1}^{5} (x_i - 3574,2)^2}{5} = 5077683. $$

In base al metodo indiretto si ha

$$V(X) = \frac{\sum_{i=1}^{5} x_i^2}{5} - m^2.$$

Essendo la media dei valori di X elevati al quadrato pari a $(7214^2 + \cdots + 3185^2)/5 = 17852588$, si ha

$$V(X) = 17852588 - 3574,2^2 = 5077683,$$

che, come ci si attende, coincide col valore ottenuto tramite il calcolo diretto.

Esercizio D.

a) Per trovare la differenza semplice media del numero di abitanti calcoliamo le differenze:

$\lvert x_i - x_j \rvert$	$x_1 = 831714$	$x_2 = 3527303$	$x_3 = 5108067$	$x_4 = 1661429$
$x_1 = 831714$	0	2695589	4276353	829715
$x_2 = 3527303$	2695589	0	1580764	1865874
$x_3 = 5108067$	4276353	1580764	0	3446638
$x_4 = 1661429$	829715	1865874	3446638	0

da cui si ricava, ponendo $N = 4$, che

$$\triangle = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| = 2449155,50.$$

b) I calcoli per trovare la differenza quadratica del numero di morti sono i seguenti:

$(x_i - x_j)^2$	$x_1 = 9133$	$x_2 = 41286$	$x_3 = 47086$	$x_4 = 13707$
$x_1 = 9133$	0	1033815409	1440430209	20921476
$x_2 = 41286$	1033815409	0	33640000	760601241
$x_3 = 47086$	1440430209	33640000	0	1114157641
$x_4 = 13707$	20921476	760601241	1114157641	0

da cui si ricava che

$${}_2\triangle = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2} = 27091,10.$$

Esercizio E.

a) Si consideri la regione Veneto. Fissando l'altezza minima e massima rispettivamente pari a 150 cm e 199 cm, si ha

Statura	[149,5–159,5)	[159,5–164,5)	[164,5–169,5)	[169,5–179,5)	[179,5–184,5)	[184,5–189,5)	[189,5–199,5)
p_i	0,008	0,036	0,127	0,550	0,187	0,071	0,021
x_i	154,5	162	167	174,5	182	187	194,5
$x_i^2 p_i$	190,962	944,784	3541,903	16747,64	6194,188	2482,799	794,4353

Quindi, essendo $m = 175,65$, lo scarto quadratico medio dalla media aritmetica si può ottenere tramite il metodo indiretto, come segue:

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 p_i - m^2 = 30.896,71 - (175,65)^2 = 43,7875,$$

e si ha $\sigma = 6,6172$. Per la regione Sicilia, avendo $m = 171,81$, si ottiene $\sigma = 6,6003$.

b) Per la regione Veneto, non conoscendo le stature di ogni singolo individuo, con una leggera forzatura il campo di variazione si può calcolare come $c_n - c_0 = 199,5 - 149,5 = 50$, mentre l'intervallo interquartile è pari a $Q_3 - Q_1 = 180,2754 - 170,9364 = 9,339$.

Esercizio F.

a) Una ponderazione naturale è data dalla popolazione residente (in milioni) dei singoli Paesi. Indicando con w_i la popolazione dei singoli Paesi, con w la loro somma e con $m = (1/w)\sum_{i=1}^{4} x_i w_i$ la media ponderata del quoziente di natalità, la varianza del quoziente di natalità dei quattro Paesi è data da

$$\sigma^2 = \frac{1}{w}\sum_{i=1}^{4} x_i^2 w_i - m^2 = 155,1836 - (12,4376)^2 = 0,4897$$

e quindi $\sigma = 0,6998$.

Esercizio G.

a) Assegnando il valore di 99 anni compiuti all'estremo superiore dell'ultima classe si ha

x_i	f_i	p_i	P_i	$x_i p_i$	$(x_i - m)^2 p_i$	$(x_i - m)^3 p_i$	$(x_i - m)^4 p_i$
0,5	11.941	0,0082	0,0082	0,0041	16,032	-708,911	31.346,185
3,0	47.077	0,0325	0,0407	0,0975	56,561	-2.359,573	98.435,151
7,5	62.862	0,0433	0,0840	0,3248	59,976	-2.232,154	83.074,872
12,5	65.171	0,0449	0,1289	0,5613	46,604	-1.501,467	48.373,287
20,0	169.866	0,1171	0,2460	2,3420	71,542	-1.768,327	43.708,361
35,0	421.981	0,2908	0,5368	10,1780	27,459	-266,832	2.592,900
55,0	369.483	0,2547	0,7915	14,0085	26,930	276,913	2.847,404
82,5	302.498	0,2085	1,0000	17,2013	297,640	11.245,617	424.889,227
	1.450.879	1,0000		44,7174	602,745	2.685,265	735.267,390

da cui si ricava che

$$\gamma_1 = {}_m\mu_3/(\sigma)^3 = 2.685,265/\left(\sqrt{602,745}\right)^3 = 0,1815.$$

Inoltre $Q_1 = 25,2751$, $m_e = 42,4691$ e $Q_3 = 61,7413$ da cui

$$\frac{(Q_3 - m_e) - (m_e - Q_1)}{(Q_3 - m_e) + (m_e - Q_1)} = 0,0387;$$

gli indici segnalano un livello di asimmetria molto basso.
b) L'indice di curtosi è pari a $\gamma_2 =_m \mu_4/(\sigma)^4 - 3 = 735.267,390/(\sqrt{602,745})^4 - 3 = -0,9762$.

Esercizio H.
a) L'indice di asimmetria basato sui quartili fornisce il valore 0,494. Per il calcolo dell'indice γ_1, essendo disponibile la distribuzione di quantità, si possono utilizzare le medie parziali al posto dei valori centrali di classe ottenendo $\gamma_1 = {}_m\mu_3/(\sigma)^3 = 494,6/(58,897)^3 = 2,42$.