

Indici di variabilità

- Prendendo in considerazione il numero di addetti per classi e le aziende:

Addetti	Aziende
0 - 5	195
5 - 10	225
10 - 25	360
25 - 40	180
40 - 100	40

- determinare la varianza e lo scarto quadratico medio.
- se il numero di addetti venisse espresso in decine piuttosto che in unità, quali valori otterremmo per media aritmetica, varianza e scarto quadratico medio?
- supponiamo che per un altro gruppo di 1000 aziende il numero medio di addetti sia pari a 45.54 e la devianza sia 382744. In quale gruppo si riscontra maggiore variabilità?

Soluzione

X	n_i	x_i	$x_i n_i$	$x_i^2 n_i$
0 - 5	195	2.5	487.5	1218.75
5 - 10	225	7.5	1687.5	12656.25
10 - 25	360	17.5	6300	110250
25 - 40	180	32.5	5850	190125
40 - 100	40	70	2800	196000
	1000		17125	510250

Media	$M = \frac{1}{n} \sum_{i=1}^K x_i n_i$	17.125
Varianza	$\sigma^2 = \frac{\sum_{i=1}^K x_i^2 n_i}{n} - M^2$	$510.25 - 293.2656 = 216.9844$
Sq	$\sigma = \sqrt{\sigma^2}$	14.73039
$M\left(\frac{X}{10}\right) = \frac{M(X)}{10}$	$\frac{17.125}{10}$	1.7125
$\text{Var}\left(\frac{X}{10}\right) = \frac{\text{Var}(X)}{100}$	$\frac{216.9844}{100}$	2.169844
$\sigma\left(\frac{X}{10}\right) = \frac{\sigma(X)}{10}$	$\frac{14.73039}{10}$	1.473039
CV_2	$\frac{\sqrt{\text{dev}(X)/n}}{ M }$	$\frac{\sqrt{382744/1000}}{45.54} = 0.43$
CV_1	$\frac{\sigma}{ M }$	$\frac{14.73039}{17.125} = 0.86$

2. Conoscendo il coefficiente di variazione $CV=0.36$ e la media $M= 6.95$ per 15 unità, come varia il CV con l'introduzione di una ulteriore unità $X_{16}=8.65$?

Soluzione

CV_{15}	M_{15}	n_i	X_{16}	n_{i+1}
0.36	6.95	15	8.95	16

Otteniamo il coefficiente di variazione per le 15 osservazioni poiché abbiamo a disposizione gli elementi per il suo calcolo. Tramite formula inversa otteniamo la deviazione standard delle 15 osservazioni.

Per il calcolo della deviazione standard delle 16 osservazioni abbiamo tuttavia bisogno della somma del quadrato delle 15 osservazioni, poiché se facessimo il quadrato della somma delle osservazioni i.e $(6.95 * 15)^2$ otterremmo un risultato erraneo potendo affermare che la somma del quadrato dei termini non è pari al quadrato della somma degli stessi ($2^2 + 3^2 \neq 5^2$), non conoscendo i 15 termini possiamo solo ottenerne il loro quadrato dalla formula inversa in cui sono stati impiegati.

Si definisce poi la deviazione standard per le 16 unità e infine il CV.

$CV_{15} = \frac{\sigma}{ M } ; \quad \sigma_{15} = CV * M_{15}$	$(6.95*0,36) =$	2.502
$\sum_{i=1}^{15} x_i^2 = N(\sigma_{15}^2 + M_{15}^2)$	$15[(2.502)^2 + (6.95)^2] =$	818.438
$M_{16} = \frac{(NM_{15}) + X_{16}}{n_{i+1}}$	$\frac{(6.95 * 15) + 8.95}{16} =$	7.075
$\sigma_{16}^2 = \frac{(x_{15}^2 + x_{16}^2)}{N} - M_{16}^2$	$\left(\frac{818.438 + 8.95^2}{16} \right) - 7.075^2 =$	6.10
$CV_{16} = \frac{\sigma_{16}}{ M_{16} }$	$\frac{\sqrt{6.10}}{7.075} =$	0.35

3. Il data set contiene 21 osservazioni relative al peso prima del trattamento (Peso pre) e dopo il trattamento (Peso post) di pazienti con anoressia. I pazienti sono stati assegnati a caso a tre gruppi di trattamento: CBT (terapia cognitivo comportamentale), FT (terapia familiare) e Cont (controllo). In questo caso l'informazione rilevante sta nella variazione prima e dopo il trattamento. Costruire i box plot e commentare i risultati.

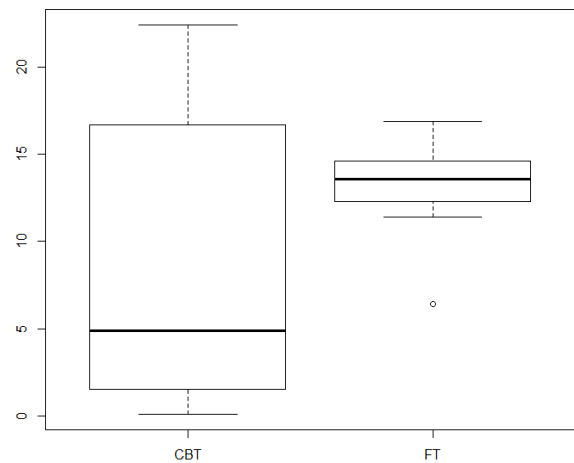
Paziente	Trattamento	Diff peso
1	FT	13.6
2	FT	13.21
3	FT	6.39
4	FT	11.39

5	FT	15.68
6	FT	16.88
7	FT	13.6
8	CBT	14.35
9	CBT	22.41
10	CBT	2.28
11	CBT	4.89
12	CBT	0.11
13	CBT	19.06
14	CBT	0.8

Soluzione

I box plot sono stati costruiti prendendo come riferimento le due distribuzioni. Gli elementi necessari sono di seguito elencati:

	FT	CBT
1q	11.39	0.8
Min	6.39	0.11
Mediana	13.6	4.89
Max	16.88	22.41
3q	13.62	14.35
Linf	8.075	-19.525
Lsup	16.915	34.675



La prima operazione da svolgere è quella di ordinare in senso non decrescente le osservazioni, per ciascuna delle due distribuzioni.

Paziente	Trattamento	Diff peso	
1	FT	6.39	
2	FT	11.39	Q1
3	FT	13.21	
4	FT	13.6	Med
5	FT	13.62	Q3
6	FT	15.68	
7	FT	16.88	
8	CBT	0.11	
9	CBT	0.8	Q1
10	CBT	2.28	
11	CBT	4.89	Med
12	CBT	14.35	Q3
13	CBT	19.06	
14	CBT	22.41	

- Primo quartile (Q1) è il minimo valore osservato tale che almeno il 25% ($=1/4$) dei dati è minore o uguale a questo valore $7/4=1.75$, si sceglie il valore che si posiziona al 1.75esimo posto, di solito nel caso del primo quartile si arrotonda per eccesso. Il primo quartile sarà pari a 11.39 (FT) e 0.8 (CBT).
 - Il Terzo quartile (Q3) è il valore osservato che lascia il 75% ($=3/4$) a sinistra della distribuzione, il valore sarà $7*(3/4)=5.25$, i valori posizionati al 5.25esimo posto sono 13.62 (FT) e 14.35 (CBT).
 - La mediana suddivide la distribuzione in due parti $50\%=(1/2)$ lasciando un numero uguale di termini a destra e sinistra, si sceglie il valore che si trova al $(7+1)/2=4$ poiché il numero di unità è dispari, la mediana coincide con 13.6 (FT) e 4.89 (CBT).
 - Il Limite inferiore viene calcolato $Q1-1.5(Q3-Q1)=11.39-1.5(13.6-11.39) = 8.075$ (FT); $0.8-1.5*(14.35-0.8)=-19.525$
 - Il limite superiore viene calcolato $Q3+1.5(Q3-Q1)=13.6+1.5(13.6-11.39) = 16.915$ (FT); $14.35+1.5*(14.35-0.8)=34.675$
 - Nel grafico verranno disegnati gli estremi o baffi seguendo la regola che se $x_{min} < L_{inf}$ si disegnerà il limite inferiore se $x_{max} > L_{sup}$ si disegnerà L_{sup} . Nel caso di FT sono i due limiti ad essere disegnati, nel caso di CBT sarà disegnato X_{min} e il Limite superiore.
4. Date le seguenti distribuzioni di frequenze relative per studenti appartenenti a gruppi di corsi di laurea differenti, in quale gruppo c'è maggiore eterogeneità?

Corsi	M	F
Lingue	0.265	0.735
Letteratura	0.452	0.548
Economia	0.459	0.541
Statistica	0.522	0.478
Scienze Politiche	0.502	0.498

Soluzione

Corsi	M	F	$f_M \ln f_M$	$f_F \ln f_F$	H	H/Hmax
Lingue	0.265	0.735	-0.35193	-0.2263	0.58	0.83
Letteratura	0.352	0.648	-0.36753	-0.28114	0.65	0.93
Economia	0.402	0.541	-0.36634	-0.33236	0.69	1.00
Statistica	0.422	0.578	-0.36408	-0.31685	0.68	0.98
Scienze Politiche	0.645	0.355	-0.28284	-0.36765	0.65	0.94

E' stato usato l'indice di entropia H (p.171 del manuale)

5. Prendendo in considerazione la distribuzione del PIL pro capite per abitante in 10 Paesi Europei nel 2013 se ne misuri il grado di concentrazione e lo si rappresenti graficamente. Infine si valuti la simmetria della distribuzione e la curtosi.

PIL (migliaia euro)	31300	44400	9900	15800	25600	29800	34500	83400	14200	43800
---------------------	-------	-------	------	-------	-------	-------	-------	-------	-------	-------

Soluzione

La concentrazione di un collettivo può essere misurata tramite il rapporto di concentrazione di Gini:

$$R = \sum_{i=1}^{N-1} \frac{(p_i - q_i)}{p_i} = 1 - \sum_{i=1}^{N-1} \frac{q_i}{p_i}$$

Con max equidistribuzione $0 \leq R \leq 1$ max concentrazione.

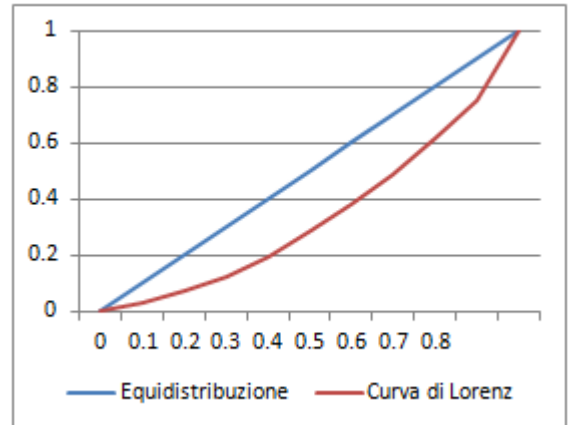
Ordinare le unità in senso non decrescente (carattere quantitativo trasferibile)

Definire le intensità cumulate

Calcolare le frequenze relative cumulate

Calcolare il rapporto tra intensità cumulate

i	X _i	X _i	S _i	p_i	q_i	$(p_i - q_i)$
1	31300	9900	9900	0.1	0.029757	0.070243
2	44400	14200	24100	0.2	0.072438	0.127562
3	9900	15800	39900	0.3	0.119928	0.180072
4	15800	25600	65500	0.4	0.196874	0.203126
5	25600	29800	95300	0.5	0.286444	0.213556
6	29800	31300	126600	0.6	0.380523	0.219477
7	34500	34500	161100	0.7	0.48422	0.21578
8	83400	43800	204900	0.8	0.61587	0.18413
9	14200	44400	249300	0.9	0.749324	0.150676
10	43800	83400		1	1	
		332700	976600	4.5	2.935377	1.564623



$$R = 1.564623/4.5 = 1 - (1.564623/4.5) = 0.3476$$

Una ulteriore formula è

$$R = 1 - \frac{2 \cdot \sum_{i=1}^{(n-1)} S_i}{(N-1)X_n} = 1 - \frac{2 \cdot 976600}{9 \cdot 332700} = 1 - 0.652306 = 0.3476$$

Per l'asimmetria della distribuzione calcoliamo l'indice di Fisher:

$$\beta = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - M}{\sigma} \right)^3$$

i	X _i	x _i - M	(x _i - M)/σ	[(x _i - M)/σ] ³
1	9900	-23370	-1.10316	-1.34252
2	14200	-19070	-0.90018	-0.72945
3	15800	-17470	-0.82466	-0.56082
4	25600	-7670	-0.36206	-0.04746
5	29800	-3470	-0.1638	-0.00439
6	31300	-1970	-0.09299	-0.0008
7	34500	1230	0.058061	0.000196
8	43800	10530	0.497061	0.122808
9	44400	11130	0.525383	0.14502
10	83400	50130	2.366349	13.25062

10.8332

L'asimmetria sarà pari a $10.8332/10=1.083$ (c'è una asimmetria positiva)
 La curtosi sarà pari a 3.41

6. Sono stati estratti a campione i risultati del test di accesso all'Università di 200 studenti suddivisi in due turni. Quale turno ha riscontrato uguale preparazione degli studenti?

	Primo turno	Secondo turno
Insufficiente	10	16
Sufficiente	27	43
Discreto	115	92
Buono	42	44
Ottimo	6	5
	200	200

Soluzione:

	Primo turno	Secondo turno	f_i	f_i^2	f_i	f_i^2
Insufficiente	10	16	0.05	0.0025	0.08	0.0064
Sufficiente	27	43	0.135	0.018225	0.215	0.046225
Discreto	115	92	0.575	0.330625	0.46	0.2116
Buono	42	44	0.21	0.0441	0.22	0.0484
Ottimo	6	5	0.03	0.0009	0.025	0.000625
TOT	200	200		0.39635		0.31325
G_primo	0.60365					
G_secondo	0.68675					

Il primo turno risulta più omogeneo, è stato utilizzato l'indice di eterogeneità del Gini (p.169 del manuale).

7. Data la seguente distribuzione effettuare la scomposizione della devianza:

Età	Fumatori	Non fumatori
18- 20	12	8
20- 25	9	21
25- 35	15	11
35- 55	5	9

Soluzione

Età media	Fumatori	Non fumatori	TOT
-----------	----------	--------------	-----

19	22	8	30
22.5	9	21	30
30	15	11	26
45	5	9	24
TOT	51	49	110

$$\bar{X} = \frac{(19 * 30) + (22.5 * 30) + (30 * 26) + (45 * 24)}{110} = 28.23$$

$$\bar{x}_1 = \frac{(19 * 22) + (22.5 * 9) + (30 * 15) + (45 * 5)}{51} = 25.40$$

$$\bar{x}_2 = \frac{(19 * 8) + (22.5 * 21) + (30 * 11) + (45 * 9)}{49} = 31.62$$

$$Dev(X) = (19 - 28.23)^2 * 30 + (22.5 - 28.23)^2 * 30 + (30 - 28.23)^2 * 26 + (45 - 28.23)^2 * 24 = 10371.81$$

$$Dev(B) = (25.40 - 28.23)^2 * 51 + (31.62 - 28.23)^2 * 49 = 972.38$$

$$Dev(W) = 10371.81 - 972.38 = 9398.62$$

8. Calcolare la mediana per le seguenti distribuzioni

Tempo	Auto tedesche	Auto italiane
6- 8	2	1
8- 10	6	2
10- 13	8	7
13- 15	2	3
15- 17	1	5
17- 20	1	2
	20	20

Soluzione:

Tempo	Auto tedesche	Auto italiane	C_i	f_i	F_i	C_i	f_i	F_i
6- 8	2	1	2	0.1	0.1	1	0.05	0.05
8- 10	6	2	8	0.3	0.4	3	0.1	0.15
10- 13	8	7	16	0.4	0.8	10	0.35	0.5
13- 15	2	3	18	0.1	0.9	13	0.15	0.65
15- 17	1	5	19	0.05	0.95	18	0.25	0.9
17- 20	1	2	20	0.05	1	20	0.1	1
	20	20		1			1	

$$Med = x_{i-1} * \frac{(x_i - x_{i-1}) \left(\frac{n}{2} - C_{i-1} \right)}{n_i} = 10 + \frac{(13 - 10)(10 - 8)}{8} = 10.75$$

$$Med = x_{i-1} * \frac{(x_i - x_{i-1})(0.5 - F_{i-1})}{F_i - F_{i-1}} * \Delta = 10 + \frac{(0.5 - 0.4)}{(0.8 - 0.4)} * 3 = 10.75$$

Auto italiane=10.85

9. Calcolare lo sqm relativo alla spesa (in centinaia di euro) per libri effettuata da 30 famiglie. Trovare lo sqm della spesa nell'ipotesi che ogni famiglia debba aggiungere una ulteriore spesa di 0.5.

x_i	n_i
1	2
1.5	3
1.8	5
2	12
2.5	8

Soluzione

x_i	n_i	$(x_i)^2$	n
1	2	1	2
1.5	3	2.25	6.75
1.8	5	3.24	16.2
2	12	4	48
2.5	8	6.25	50
			122.95

- sqm: $122.95/30 = \text{sqrt}(4.1) = 2.02$

- lo sqm è invariante per traslazione, aggiungendo una costante ad ogni termine della distribuzione lo sqm non cambia

10. Data la seguente distribuzione delle entrate provenienti dai redditi (in migliaia di dollari) soggetti ad imposta di una certa categoria di lavoratori

entrate	0- 2	2- 4	4- 6	6- 8	8- 10	10- 15	15- 20	20- 50	50- 100	100- 200	TOT
frequenze	250	1589	1768	1473	1172	1298	306	200	21	3	8080

Calcolare:

- I quartili
- Il rapporto di concentrazione

Soluzione

- a. Nella distribuzione data, il primo quartile è l'elemento il cui posto si trova approssimativamente $\frac{1}{4}=0,25$ dunque $8080/4=2020$, il terzo $8080/(3/4)=6060$. Si fa riferimento alle distribuzioni cumulate, il quartile si riferisce alla classe.

classi	n_i	C_i	f_i	F_i
0- 2	250	250	0.030941	0.030941
2- 4	1589	1839	0.196658	0.227599
4- 6	1768	3607	0.218812	0.446411
6- 8	1473	5080	0.182302	0.628713
8- 10	1172	6252	0.14505	0.773762
10- 15	1298	7550	0.160644	0.934406
15- 20	306	7856	0.037871	0.972277
20- 50	200	8056	0.024752	0.99703
50- 100	21	8077	0.002599	0.999629
100- 200	3	8080	0.000371	1
8080				

$$Q_1 = x_{i-1} * \frac{(x_i - x_{i-1}) \left(\frac{n}{4} - C_{i-1} \right)}{n_i} = 4 + \frac{(6 - 4)(2020 - 1839)}{1768} = 4.20$$

$$Q_1 = x_{i-1} * \frac{(x_i - x_{i-1})(0.25 - F_{i-1})}{F_i - F_{i-1}} * \Delta = 4 + \frac{(0.25 - 0.22)}{(0.44 - 0.22)} * 2 = 4.20$$

$$Q_3 = x_{i-1} * \frac{(x_i - x_{i-1}) \left(\frac{3}{4}n - C_{i-1} \right)}{n_i} = 8 + \frac{(10 - 8)(6060 - 5080)}{1172} = 9.67$$

$$Q_3 = x_{i-1} * \frac{(x_i - x_{i-1})(0.75 - F_{i-1})}{F_i - F_{i-1}} * \Delta = 8 + \frac{(0.75 - 0.628)}{(0.77 - 0.628)} * 2 = 9.67$$

b. Rapporto di concentrazione:

calcolare il valore centrale, la numerosità, le frequenze cumulate, p_i=frequenze cumulate/numerosità, le intensità cumulate, le frequenze delle intensità cumulate.

n_i	val centr	x_in_i	C_i	p_i	S_i	q_i=S_i/S_n	q_i+q_i-1	p_i-p_(i-1)	
250	1	250	250	0.030941	250	0.003827	0.003827	0.030941	0.000118
1589	3	4767	1839	0.227599	5017	0.076805	0.080633	0.196658	0.015857
1768	5	8840	3607	0.446411	13857	0.212137	0.288942	0.218812	0.063224
1473	7	10311	5080	0.628713	24168	0.369988	0.582125	0.182302	0.106123
1172	9	10548	6252	0.773762	34716	0.531468	0.901456	0.14505	0.130756
1298	12.5	16225	7550	0.934406	50941	0.779856	1.311324	0.160644	0.210656
306	17.5	5355	7856	0.972277	56296	0.861836	1.641693	0.037871	0.062173
200	35	7000	8056	0.99703	63296	0.968999	1.830835	0.024752	0.045318
21	75	1575	8077	0.999629	64871	0.993111	1.96211	0.002599	0.0051
3	150	450	8080	1	65321	1			0.639324

$$R=1-0.64=0.36$$