Esercizi di Statistica per Biotecnologie

Francesca Pizzorni Ferrarese

Esercitazione IV– Statistica descrittiva II

Es. 1
Dopo il disastro, una commissione d'inchiesta del British Board of Trade ha compilato una lista di tutti i 1316 passeggeri con alcune informazioni aggiuntive riguardanti: l'esito (salvato, non salvato), la classe (I, II, III) in cui viaggiavano, il sesso, l'età,. . . .
In questo esercizio ci limitiamo nella prima parte a considerare le nformazioni sull'esito e la classe.

	Classe			
Esito	I	II	III	Totale
Salvato	203	118	178	499
Non Salvato	122	167	528	817

a) Se possibile, calcolare la moda, la mediana e la media della serie
b) Determinare la percentuale di passeggeri di prima classe non salvati
c) Quale percentuale dei passeggeri salvati era di prima classe?

Es. 2
I seguenti dati sono i punteggi che 10 studenti hanno conseguito in due esami di Analisi
Matematica (punteggio massimo = 100).

Analisi I	*Analisi* II
51	74
68	70
97	93
55	67
95	99
74	73
20	33
91	91
74	80
80	86

a) Rappresentare i dati mediante scatter plot
b) Se possibile, calcolare la moda, la mediana e la media della serie
b) Calcolare la covarianza e la correlazione.

Es. 3

Alla fine di una giornata di lavoro un intervistatore si accorge di aver perso i dati raccolti su un certo numero di famiglie relativamente al numero X di televisori

	Numero componenti Y			
Numero televisori X	1	2	3	Totale
0	0	3	1	?
1	3	?	7	16
2	1	?	?	?
Totale	?	13	11	?

posseduti e al numero Y di componenti della famiglia. Ricostruendo a memoria le interviste fatte, arriva alla seguente tabella

a) Si completi la tabella e si dica se i due caratteri sono indipendenti.
b) Si calcolino le medie e le varianze del numero di televisori condizionatamente al numero di componenti della famiglia.

Es. 4
Nel settore delle cucitrici a macchina in una fabbrica di capi di abbigliamento viene assegnato un punteggio a ogni capo finito sulla base della sua qualità. La paga di ciascuna cucitrice dipende in parte dal numero di capi completati. In una particolare

Cucitrice	Numero capi (X)	Punteggio (Y)
1	14	7,2
2	13	7,3
3	17	6,9
4	16	7,3
5	17	7,5
6	18	7,6
7	19	6,8
8	32	3,7
9	18	6,5
10	15	7,9
11	15	6,8
12	19	7,1

giornata si sono registrati i seguenti punteggi medi per pezzo:
a) Calcolare il coefficiente di correlazione tra X e Y e commentare.
b) Tracciare il diagramma di dispersione (scatter plot) dei dati
c) Si cotruisca la retta di regressione e sovrapporre il grafico della retta di regressione al diagramma di dispersione di X e Y.
d) Calcolare l'indice della bontà di adattamento della retta ai dati.
e) In base al modello di regressione si preveda il punteggio corrispondente ad un numero di capi pari a 20.
e) Ricalcolare la retta di regressione dopo aver eliminato l'osservazione sull'ottava cucitrice e tracciarne il grafico.

Es. 5
Si considerino i seguenti dati relativi alle importazioni e alle esportazioni in migliaia di milioni di euro di alcune regioni italiane nel 2010:

	Piemonte	Lombardia	Veneto	Toscana	Lazio
Importazioni	33,6	118,5	35,9	21,1	22,5
Esportazioni	51,0	115,8	54,0	32,3	14,1

a) Si calcoli la retta di regressione delle esportazioni rispetto alle importazioni.
b) Si tracci la retta di regressione sul grafico di dispersione delle due variabili.
c) Si calcoli la bontà di adattamento.
d) Si prevedano le esportazioni della regione Emilia-Romagna sapendo che le importazioni sono state pari a 22,58.

Es. 6
Nella seguente tabella si riportano le misure dell'ossigeno consumato da una persona che cammina, in corrispondenza a varie velocità della persona.

velocità (Km/h)	*ossigeno (litri/h)*
0	19
1	20
2	20.5
3	21.5
4	22
5	23
6	23
7	23.5
8	24

Verificare che esiste dipendenza lineare del consumo di ossigeno dalla velocità, determinare la retta di regressione e verificare la bontà del modello.

ESERCITAZIONE 4

① Serie bi-variata: serie statistica in cui ad ogni unità statistica si rilevano 2 caratteri $O_i = (x_i ; y_i)$ $O = \{O_i\}$

Frequenza congiunta: valore associato ad ogni modalità i, j

Tabella a doppia entrata: $n_{i,j}$: frequenze assolute

SALVATO S	CLASSE I	CLASSE II	CLASSE III	TOTALE
SI	203	118	178	499
NO	122	167	528	817
TOTALE	325	285	706	1316
	$n_{*,j}$: frequenze assolute			totale osservaz.

a) moda: modalità con frequenza maggiore $mo = (NO ; III)$

media: osservazione che bipartisce la popolazione → solo dati quantitativi

→ la media non si applica alle serie bivariate → NO

media: media delle osservazioni → solo dati quantitativi → NO

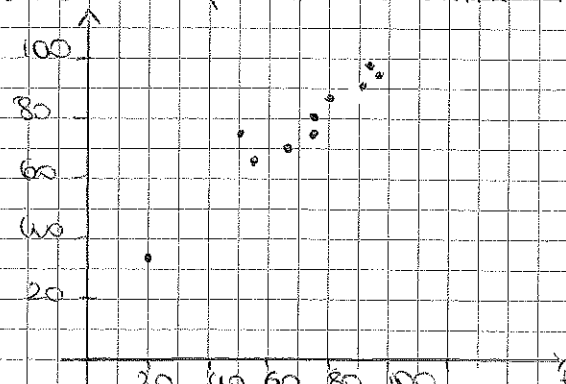
b) Distribuzione condizionata

$$P(S = NO \mid C = I) = \frac{122}{325} \cdot 100 = 37,54\,\%$$

c) $$P(C = I \mid S = SI) = \frac{203}{499} \cdot 100 = 40,68\,\%$$

② Scatter plot: rappresentazione grafica usata per serie quantitative, in cui ogni osservazione viene riportata in un piano cartesiano come se fosse un punto

a)



b) $mo = (97 ; 99)$

mediana: NO

media: somma dei valori e divido il numero per n°

$$[(81;74) + (68;70) + (97;93) + (85;67) + (95;99) + (74;73) + (20;33) + (91;91) + (74;80) + (80;86)] / 10 =$$

$= (51+68+97+55+95+76+20+91+74+80 \;;\; 74+70+93+67+99+$

$+73+33+91+80+86)/10 =$

$= (705/10 \;;\; 766/10) = (70,5 \;;\; 76,6)$ ← la media della media è data dalla media dei campioni

c) Non è possibile ottenere il gruppo di valori di correlazione → essendo solo la varianza

Come misuro il "grado di legame" fra 2 caratteri?

Dati n' osservazioni $(x_i ; y_i)$ valuto il prodotto dei 2 scarti

$(x_i - \bar{x})(y_i - \bar{y}) \longrightarrow \begin{cases} > 0 & \text{scarti concordi} \Rightarrow \text{↗ direzione} \\ < 0 & \text{↘ inversa} \end{cases}$ ↑ proporzionalità

⇒ Dati una serie numerica delle N osservazioni $(x_i ; y_i)$ definisco covarianza

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Varianza $x = (51^2 + 68^2 + \ldots + 80^2)/10 - 70,5^2 =$

$= 5171,7 - 4970,25 = 501,45$

Varianza $y = (74^2 + 70^2 + \ldots + 86^2)/10 - 76,6^2 =$

$= 6181 - 5776 = 405$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
51	74	-19,5	-2,6	50,7
68	70	-2,5	-6,6	16,5
97	93	26,5	16,4	434,6
55	67	-15,5	-9,6	148,8
95	99	24,5	22,4	548,8
74	73	3,5	-3,6	-12,6
20	33	-50,5	-43,6	2201,8
91	91	20,5	14,4	295,2
74	80	3,5	3,4	11,9
80	86	9,5	9,4	89,3

$\Rightarrow \sigma_{xy} = \frac{3785}{10} = 378,5$ → matrice var/cov $= \Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 501,45 & 378,5 \\ 378,5 & 405 \end{bmatrix}$

Correlazione di Pearson: data una sua bivariata (x_i, y_i) diciamo correlazione il valore $R = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

$\rightarrow R = \frac{378,5}{\sqrt{501,45}\sqrt{405}} = 0,84$ R è meno influenzato dalla variabilità

(3) a)

NUMERO TELEFONI X	NUMERO COMPONENTI Y = 1	2	3	TOTALE
0	0	3	1	4
1	3	6	7	16
2	1	4	3	8
TOTALE	4	13	11	28

X	$f(x)$	$f(x \mid y_1)$	$f(x \mid y_2)$	$f(x \mid y_3)$
0	4/28	0/4	3/13	1/11
1	16/28	3/4	6/13	7/11
2	8/28	1/4	4/13	3/11

non sono indipendenti.

b) Data una sua bivariata (x_i, y_i) avente $M = M_x M_y$ modalità con frequenza relativa f_{ij} diciamo covarianza

$$\sigma_{xy} = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} (x_i - \bar{x})(y_j - \bar{y}) f_{ij}$$

$$\bar{x} = \frac{4 \cdot 0 + 16 \cdot 1 + 8 \cdot 2}{28} = 1,28$$

$$\bar{y} = \frac{4 \cdot 1 + 13 \cdot 2 + 11 \cdot 3}{28} = 2,25$$

$$\sigma_{xy} = \sum_{i=1}^{3} \sum_{j=1}^{3} (x_i - 1,28)(y_j - 2,25) f_{ij} =$$

$$= (0-1,28)(1-2,25) \cdot 0 + (0-1,28)(2-2,28) \cdot 3 +$$
$$+ (0-1,28)(3-2,25) \cdot 1 + (1-1,28)(1-2,25) \cdot 3 +$$
$$+ (1-1,28)(2-2,25) \cdot 6 + (1-1,28)(3-2,25) \cdot 7 +$$
$$+ (2-1,28)(1-2,25) \cdot 1 + (2-1,28)(2-2,25) \cdot 4 +$$
$$+ (2-1,28)(3-2,25) \cdot 3 = 0 + 0,96 - 0,96 + 1,05 + 0,42 +$$
$$- 1,47 - 0,9 - 0,72 + 1,62 = 0$$

(4) a) $R = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

$$me(x) = \frac{14+13+17+\ldots+15+19}{12} = 47,75$$

$$me(y) = \frac{7,2+7,3+6,9+\ldots+6,8+7,1}{12} = 6,88$$

$$Var(x) = \frac{14^2+13^2+17^2+\ldots+15^2+19^2}{12} - 47,75^2 =$$

$$= 336,92 - 315,0625 = 21,8575$$

$$Var(y) = \frac{7,2^2+7,3^2+6,9^2+\ldots 6,8^2+7,1^2}{12} - 6,88^2 =$$

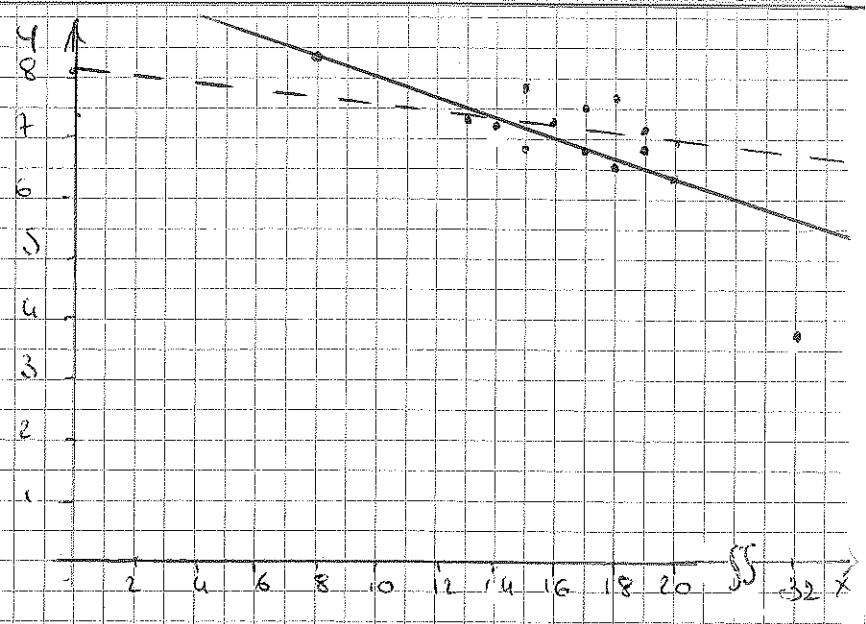
$$= 48,44 - 47,3344 = 1,1056$$

		16,45 ↓		7,17 ↓		
x_i	y_i	$x_i - \bar{x}$		$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	
14	7,2	-3,75	-2,45	0,32	0,03	-1,2	-0,0735
13	7,3	-4,75	-3,45	0,42	0,13	-2,995	-0,4485
17	6,9	-0,75	0,55	0,02	-0,27	-0,015	-0,1485
16	7,3	-1,75	-0,45	0,42	0,13	-0,735	-0,0585
17	7,5	-0,75	0,55	0,62	0,33	-0,665	0,1815
18	7,6	0,75	1,55	0,72	0,43	0,54	0,6665
19	6,8	1,75	2,55	-0,08	-0,37	-0,14	-0,9435
32	3,7	19,25		-3,18		-45,315	
18	6,5	0,75	1,55	-0,38	-0,67	-0,285	-1,0385
15	7,9	-2,75	-1,45	1,02	0,73	-2,805	-1,0585
15	6,8	-2,75	-1,45	-0,08	-0,37	-0,22	0,5365
19	7,1	1,75	2,55	0,22	-0,07	0,385	-0,1785
						-52,25	-2,5635

$$\sigma_{xy} = \frac{-52,25}{12} = -4,35$$

$$\Rightarrow R = \frac{-4,35}{\sqrt{21,8575}\sqrt{1,1056}} = -\frac{4,35}{4,92} = -0,88$$

b)



c) Punteggio y = variabile spiegata

nº ore x = variabile esplicativa

Retta di regressione

- modello di tipo affine $m_i \cdot \hat{y}_i = a x_i + b$
- parametri stimati secondo minimo

$$SSR = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{N} (a x_i + b - y_i)^2$$

$$\arg\min_{a,b} \sum_{i=1}^{N} (a x_i + b - y_i)^2$$

- si dimostra che $a = \frac{S_{xy}}{\sigma_x^2}$ $\quad b = \bar{y} - a\bar{x}$

$$\Rightarrow a = -\frac{4,35}{21,86} = -0,199$$

$$b = 6,88 + 0,199 \cdot 17,75 = 10,41$$

$\Rightarrow$ retta di regressione stimata è $\hat{y} = -0,199x + 10,41$

d) Valuto il coefficiente di Pearson

$|R| < 0,3$ scarsa prob. di legame

$0,3 < |R| < 0,7$ moderata prob. di legame lineare

$0,7 < |R|$ alta prob. " " $\leftarrow |R| = 0,88$

e) $\hat{y} = ax + b = -0,199 \cdot 20 + 10,41 = 6,43$

↑

n° copie

f) $me(x) = \frac{47,75 \cdot 12 - 32}{11} = 46,45$

$me(y) = \frac{6,88 \cdot 12 - 3,7}{11} = 7,17$

$\sigma_{xy} = \frac{-2,5635}{11} = -0,2330$

$\sigma_a(x) = \frac{3019}{11} - 46,45^2 = 274,45 - 270,6025$

$= 3,8520$

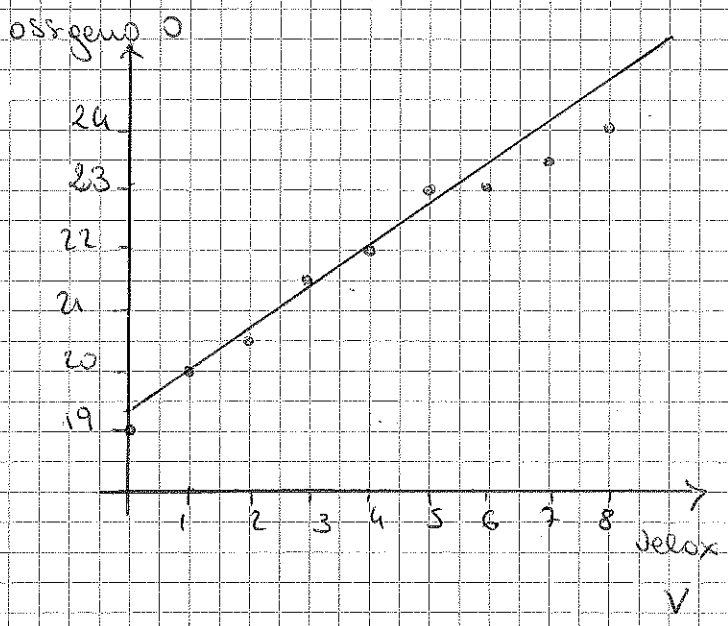
$\sigma_a(y) = \frac{567,89}{11} - 7,17^2 = 51,6 - 51,4089 = 0,12$

$a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-0,2330}{3,8520} = -0,0605$

$b = 7,17 + 0,0605 \cdot 46,45 = 8,1652$

$\hat{y}_1 = -0,0605x + 8,1652$

⑥ Il diagramma di dispersione mostra che il consumo di O_2 cresce e all'incirca una funzione lineare della velocità dell'aereo



$\bar{x} = \frac{0 + 1 + \ldots + 7 + 8}{9} = \frac{36}{9} = 4$

$\bar{y} = \frac{19 + 20 + \ldots + 23,5 + 24}{9} = \frac{196,5}{9} = 21,83$

$\sigma_x^2 = \frac{0^2 + 1^2 + \ldots + 7^2 + 8^2}{9} - 4^2 = 22,67 - 16 = 6,67$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	19	-4	-2,83	11,32
1	20	-3	-1,83	5,49
2	20,5	-2	-1,33	2,66
3	21,5	-1	-0,33	0,33
4	22	0	0,17	0
5	23	1	1,17	1,17
6	23	2	1,17	2,34
7	23,5	3	1,67	5,01
8	24	4	2,17	8,68
				37

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$$

$$= \frac{37}{9} = 4,11$$

Reta de regressão $\quad a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{4,11}{6,67} = 0,62$

$$b = \bar{y} - a\bar{x} = 21,83 - 0,62 \cdot 4 = 19,35$$

$\Rightarrow \hat{y}_i = a x_i + b$ modelo de regressão

$$= 0,62\, x_i + 19,35$$

$$R = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

$$\sigma_y^2 = \frac{19^2 + 20^2 + \ldots + 23,5^2 + 24^2}{9} - 21,83^2 = \frac{4313,75}{9} - 476,55 =$$

$$= 479,30 - 476,55 = 2,75$$

$$\Rightarrow R = \frac{4,11}{\sqrt{6,67 \cdot 2,75}} = 0,96 > 0,7$$ ajuste linear

(3) cont. matrice varianza / covarianza

$$\Sigma = \begin{bmatrix} \sigma^2_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2_y \end{bmatrix} = \begin{bmatrix} 0,07 & 0 \\ 0 & 0,47 \end{bmatrix}$$

$$\sigma^2_x = \frac{(0^2)\cdot 4 + (1^2)\cdot 16 + (2^2)\cdot 8}{28} - 1,28^2 = 1,71 - 1,64 = 0,07$$

$$\sigma^2_y = \frac{(1^2)\cdot 4 + (2^2)\cdot 13 + (3^2)\cdot 11}{28} - 2,25^2 = 5,53 - 5,06 = 0,47$$

N.B. Proprietà R

a) $-1 \leq R \leq 1$

b) se per ogni osservazione $y_i = a x_i + b$ allora $|R| = 1$

c) se $(x; y)$ sono ottenute da realizzazioni di v.v. c.c. X ed Y marginali ⟹

R è una v.c.

$E(R) = 0$