Esercitazione 2: Analisi di due variabili: correlazione, indipendenza e amici

Angela Andreella

20/10/2020

Variabili Quantitative

Esercizio 1

Si rilevano il peso e l'altezza di 9 giocatori di rugby

Peso	Altezza
93	184
79	168
86	180
94	184
84	185
83	188
80	180
70	177
75	178

1. Calcolare la media della variabile Peso e della variabile Altezza;
2. Calcolare la varianza delle due variabili, la covarianza, e verificare quale delle due variabili risulta più variabile;
3. Creare un grafico appropriato;
4. Misurare la relazione esistente tra peso e altezza

Soluzione

1. Indicando con X il peso, e Y l'altezza. Le medie risultano pari a:

$$\bar{X} = \frac{93 + 79 + 86 + 94 + 84 + 83 + 80 + 70 + 75}{9} = 82.667$$

$$\bar{Y} = \frac{184 + 168 + 180 + 184 + 185 + 188 + 180 + 177 + 178}{9} = 180.444$$

In R:

```
Peso <- c(93,79,86,94,84,83,80,70,75)
Altezza <- c(184,168,180,184,185,188,180,177,178)
```

```r
mean(Peso)
```

```
## [1] 82.66667
```

```r
mean(Altezza)
```

```
## [1] 180.4444
```

2. Per calcolare la varianza costruiamo la seguente tabella:

	X	Y	X - mean(X)	Y - mean(Y)	(X - mean(X))^2	Y - mean(Y))^2
1	93	184	10.3333333	3.5555556	106.7777778	12.6419753
2	79	168	-3.6666667	-12.4444444	13.4444444	154.8641975
3	86	180	3.3333333	-0.4444444	11.1111111	0.1975309
4	94	184	11.3333333	3.5555556	128.4444444	12.6419753
5	84	185	1.3333333	4.5555556	1.7777778	20.7530864
6	83	188	0.3333333	7.5555556	0.1111111	57.0864198
7	80	180	-2.6666667	-0.4444444	7.1111111	0.1975309
8	70	177	-12.6666667	-3.4444444	160.4444444	11.8641975
9	75	178	-7.6666667	-2.4444444	58.7777778	5.9753086
Totale	744	1624	0.0000000	0.0000000	488.0000000	276.2222222

dove la terza colonna equivale a $X - \bar{X}$, la quarta $Y - \bar{Y}$, la quinta $(X - \bar{X})^2$ e la sesta $(Y - \bar{Y})^2$ ricordando che:

$$S_X^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n}$$

$$S_Y^2 = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{n}$$

Le varianze sono dunque pari a:

$$S_X^2 = \frac{488}{9} = 54.22222$$

$$S_Y^2 = \frac{276.2222222}{9} = 30.69136$$

In R:

```r
var(Peso) * (9-1)/9
```

```
## [1] 54.22222
```

```r
var(Altezza) * (9-1)/9
```

```
## [1] 30.69136
```

moltiplichiamo per $(n-1)/n$ con $n = 9$ numero totale di osservazioni perchè R calcola la formula della varianza ponendo al denominatore $n-1$ invece di n.

Per confrontare Altezza e Peso in termini di variabilità, calcoliamo i coefficienti di variazione:

$$CV_X = \frac{S_X}{\bar{X}} = \frac{\sqrt{54.22222}}{82.66667} = 0.08907549$$

$$CV_Y = \frac{S_Y}{\bar{Y}} = \frac{\sqrt{30.69136}}{180.4444} = 0.03070185$$

la variabile X, ovvero il Peso risulta più variabile.

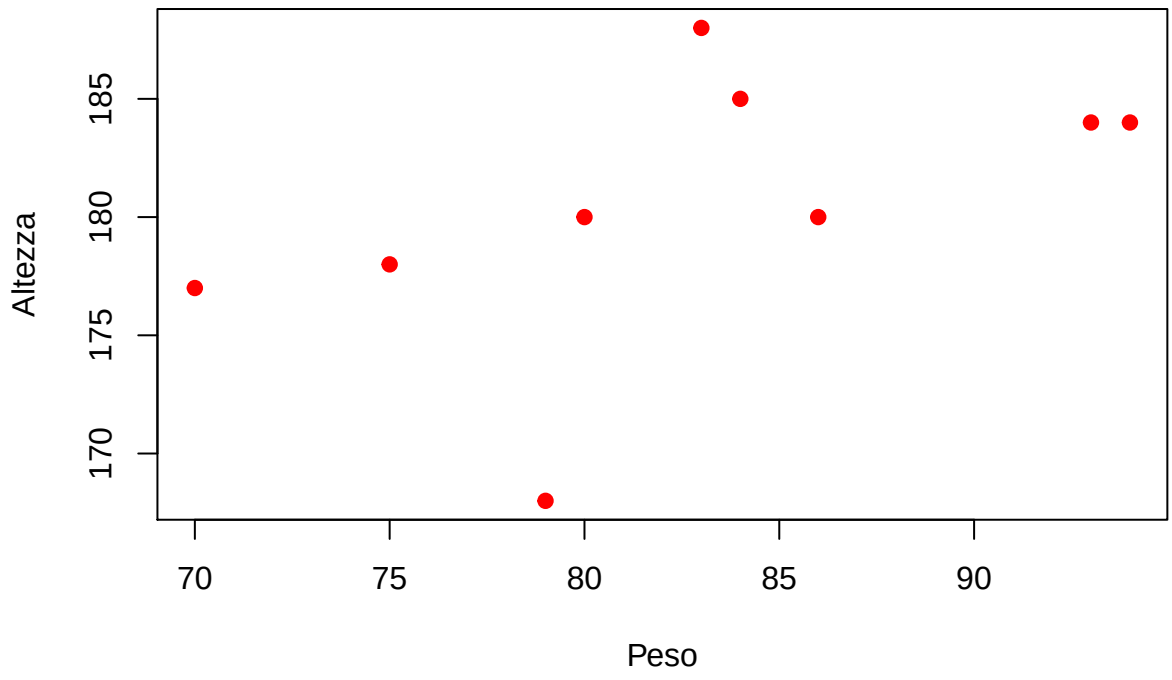
In R:

```
CV_x <- sqrt(var(Peso) * (9-1)/9)/mean(Peso)
```

```
CV_x <- sqrt(var(Altezza) * (9-1)/9)/mean(Altezza)
```

3. Un grafico appropriato è lo scatterplot:

```
plot(Peso, Altezza, col = "red", pch = 19)
```



4. Per capire la relazione tra Altezza e Peso, calcoliamo il coefficiente di correlazione:

$$R_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

Possiamo continuare la tabella del punto 2:

	X	Y	X - mean(X)	Y - mean(Y)	(X - mean(X))^2	Y - mean(Y))^2	(X - mean(X))(Y - mean(Y))
1	93	184	10.3333	3.5556	106.7771	12.6423	36.7411
2	79	168	-3.6667	-12.4444	13.4447	154.8631	45.6299
3	86	180	3.3333	-0.4444	11.1109	0.1975	-1.4813
4	94	184	11.3333	3.5556	128.4437	12.6423	40.2967
5	84	185	1.3333	4.5556	1.7777	20.7535	6.0740
6	83	188	0.3333	7.5556	0.1111	57.0871	2.5183
7	80	180	-2.6667	-0.4444	7.1113	0.1975	1.1851
8	70	177	-12.6667	-3.4444	160.4453	11.8639	43.6292
9	75	178	-7.6667	-2.4444	58.7783	5.9751	18.7405
Totale	744	1624	-0.0003	0.0004	488.0001	276.2223	193.3335

dunque:

$$R_{XY} = \frac{193.3335/9}{7.363574 \cdot 5.539978} = 0.5265838$$

In R:

```
cor(Peso, Altezza)
```

```
## [1] 0.5265838
```

oppure

```
cov(X,Y)/(sd(X)*sd(Y))
```

```
## [1] 0.5265838
```

Variabili Qualitative

Si rilevano il sesso e l'essere favorevoli ad una certa proposta di legge di 303 studenti del dipartimento di Psicologia

	Maschi	Femmine	Totale
Favorevoli	38	74	112
Non favorevoli	45	146	191
Totale	83	220	303

1. Calcolare le tabelle di frequenza assoluta e relativa per la variabile Sesso
2. Calcolare la tabella di frequenza della variabile Sesso avendo risposte favorevoli
3. Calcolare l'indice χ^2 e commentarlo
4. Quante femmine sono favorevoli alla proposta di legge?

Soluzioni:

1. La tabelle di frequenza assoluta per la variabile Sesso è pari alle frequenze marginali che trovate nell'ultima riga *Totale*, ovvero:

	f_i
Maschi	83
Femmine	220
Totale	303

per le frequenze relative dobbiamo dividere f_i per il totale delle osservazioni ovvero:

	f_i	p_i
Maschi	83	0.2739274
Femmine	220	0.7260726
Totale	303	1.0000000

2. Per calcolare la frequenza relativa condizionata della variabile Sesso avendo risposte favorevoli dobbiamo concentrarci solamente sulle osservazioni che hanno modalità favorevole, e dividere per il totale delle persone favorevoli:

	p_i
Maschi	0.3392857
Femmine	0.6607143
Totale	1.0000000

3. Per calcolare l'indice χ^2 per prima cosa dobbiamo calcolare le frequenze attese definite come

$$f_{ij}^A = \frac{f_{\cdot j} \cdot f_{i\cdot}}{n}$$

dove: - $f_{i\cdot}$: corrisponde alle frequenze marginali riga: $f_{1\cdot} = 112$, $f_{2\cdot} = 191$ - $f_{\cdot j}$: corrisponde alle frequenze marginali colonna: $f_{\cdot 1} = 83$, $f_{\cdot 2} = 220$

Dunque abbiamo:

- $f_{11}^A = \frac{112 \cdot 83}{303} = 30.67987$;
- $f_{12}^A = \frac{112 \cdot 220}{303} = 81.32013$;
- $f_{21}^A = \frac{191 \cdot 83}{303} = 52.32013$;
- $f_{22}^A = \frac{191 \cdot 220}{303} = 138.6799$;

Infine dobbiamo utilizzare la formula del χ^2:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(f_{ij} - f_{ij}^A)^2}{f_{ij}^A}$$

Perchè è costruito in questo modo il χ^2? Al numeratore abbiamo la somma delle differenze tra frequenze osservate f_{ij} e quelle attese (attese nel caso di indipendenza! ovvero quelle che ci aspettiamo se fossimo in

caso di indipendeza) f_{ij}^A. Le differenze sono poste al quadrato per rendere le differenze tutte positive, alla fine ci interessa quanto è grande la differenza non se è positiva o negativa (visto che questo indice ci indica solo se vi è dipendenza o meno, e non la direzione (come il coefficiente di correlazione) della relazione tra le due variabili). Tutte queste differenze vengono rapportate con quelle attese corrispondenti e sommate con quelle due sommatorie.

Nel nostro caso risulta pari a:

$$\chi^2 = \frac{(38-30.67987)^2}{30.67987} + \frac{(74-81.32013)^2}{81.32013} + \frac{(45-52.32013)^2}{52.32013} + \frac{(146-138.6799)^2}{138.6799} = 3.8159$$

Si può notare che le frequenze osservate e attese sono abbastanza diverse, per esempio 146 e 138.6799, ma comunque calcolando il valore soglia del χ^2 pari a $3 \cdot (\text{numero righe} - 1) \cdot (\text{numero colonne} - 1) = 3$, possiamo dire che la connessione delle due variabili è di debole intensità.

In R è molto semplice, creaiamo la tabella iniziale:

```
db <- as.table(rbind(c(38,45), c( 74, 146)))
dimnames(db) <- list(Sesso = c("F", "M"),
                     Risposta = c("Favorevoli","Non favorevoli"))
db
```

```
##      Risposta
## Sesso Favorevoli Non favorevoli
##     F         38             45
##     M         74            146
```

e usiamo il seguente comando:

```
(Xsq <- chisq.test(db, correct = F))
```

```
##
##  Pearson's Chi-squared test
##
## data:  db
## X-squared = 3.816, df = 1, p-value = 0.05076
```

Potete vedere il risultato del test nella parte dove c'è scritto 3.816 e con i seguenti comandi potete vedere le frequenze osservate:

```
Xsq$observed
```

```
##      Risposta
## Sesso Favorevoli Non favorevoli
##     F         38             45
##     M         74            146
```

e attese:

```
Xsq$expected
```

```
##      Risposta
## Sesso Favorevoli Non favorevoli
##     F   30.67987       52.32013
##     M   81.32013      138.67987
```

4. Per rispondere a questa domanda basta guardare la frequenza congiunta rispetto alle modalità Femmina e Favorevole, ovvero 74.