Università degli Studi di Padova
Facoltà di Scienze Statistiche, CL. in Astronomia

Esame di Statistica Descrittiva mod. B, Appello del 4/7/2007
CORREZIONE

1) Un'azienda produce bulloni di diametro pari a 1 cm. (diametro nominale). Per controllare la qualità dei pezzi prodotti viene rilevato il diametro di 20 bulloni prodotti dall'impianto A e altri 20 prodotti dall'impianto B. La seguente tabella riporta le differenze in millimetri tra diametri rilevati e quello nominale:

Impianto A	0.1	0.2	-0.3	0.1	0.1	0.2	0.0	-0.1	0.0	0.1
	0.0	0.1	0.1	0.0	0.0	-0.2	-0.1	0.3	0.1	0.0
Impianto B	0.0	-0.1	-0.1	0.1	-0.4	0.0	0.2	0.1	0.0	-0.1
	-0.1	-0.2	-0.2	-0.1	0.0	0.0	0.1	0.0	-0.2	0.0

a) Calcolare le distribuzioni di frequenze assolute, cumulate, relative e relative cumulate per ciascun impianto:

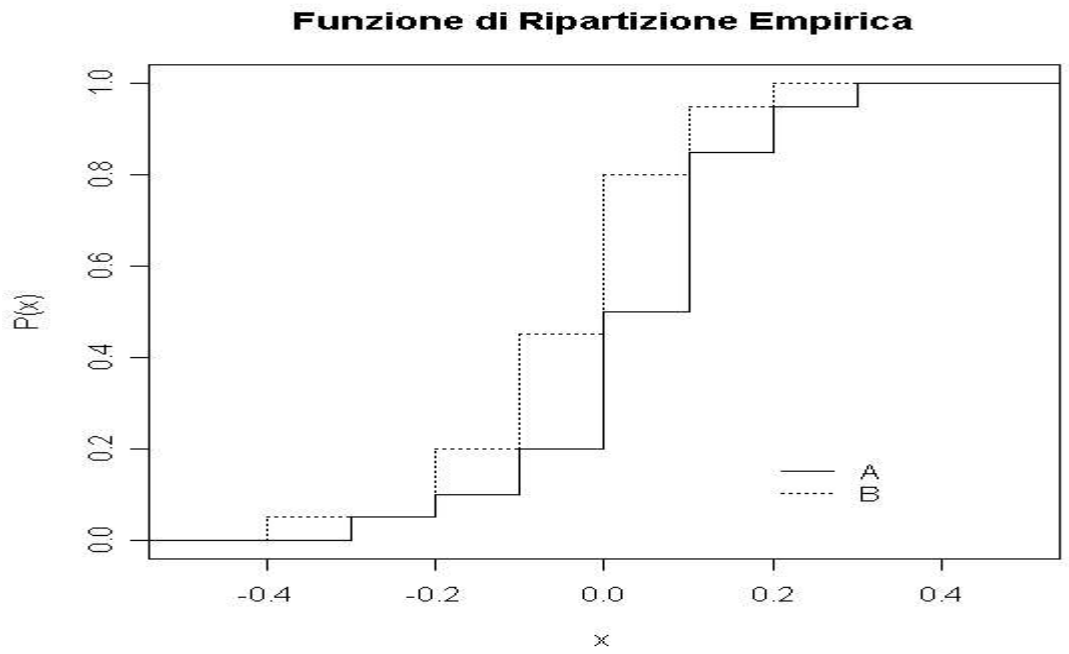
Impianto A				
Xi	**fi**	**Fi**	**pi**	**Pi**
-0.4	0	0	0	0
-0.3	1	1	0.05	0.05
-0.2	1	2	0.05	0.10
-0.1	2	4	0.10	0.20
0	6	10	0.30	0.50
0.1	7	17	0.35	0.85
0.2	2	19	0.10	0.95
0.3	1	20	0.05	1.00
Totale	**20**		**1**	

Impianto B				
Xi	**fi**	**Fi**	**pi**	**Pi**
-0.4	1	1	0.05	0.05
-0.3	0	1	0.00	0.05
-0.2	3	4	0.15	0.20
-0.1	5	9	0.25	0.45
0	7	16	0.35	0.80
0.1	3	19	0.15	0.95
0.2	1	20	0.05	1.00
0.3	0	20	0.00	1.00
Totale	**20**		**1**	

Note: - per trovare le distribuzioni richieste bisogna ordinare i dati in senso crescente!

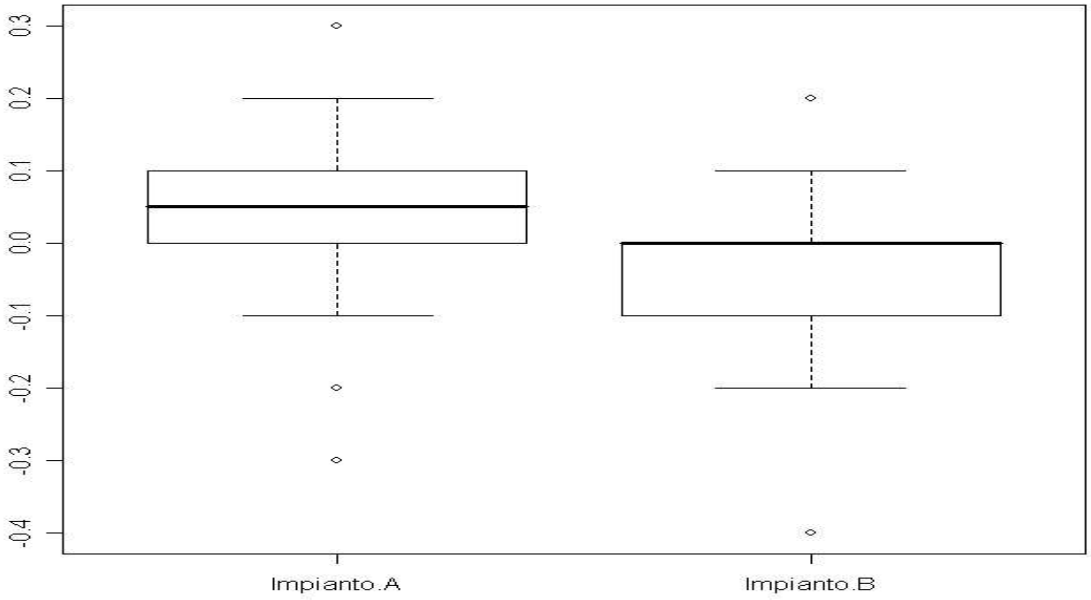
- Per costruire la funzione di ripartizione empirica bisogna utilizzare la prima colonna della tabella qui sopra come ascissa e l'ultima colonna come ordinata e poi unire i punti con una funzione "a gradini".

b) Si traccino due grafici delle funzioni di ripartizione empirica e in base a questi si confrontino le due distribuzioni:



Dal grafico si nota che l'impianto B tende a produrre bulloni con scarti dal valore nominale che sono più piccoli di quelli dell'impianto A. In altre parole la distribuzione della variabile X = "scarti tra valore effettivo e valore nominale del diametro" è più spostata a sinistra nell'impianto B.

c) si traccino due Boxplot e in base a questi si confrontino le due distribuzioni:



Dal boxplot si evincono le stesse considerazioni fatte al punto precedente. Inoltre, la distribuzione dei dati nell'impianto A presenta una simmetria che non si riscontra nella distribuzione dei dati relativa all'impianto B.
Se consideriamo come indice di posizione la mediana, si nota che il centro della distribuzione relativa all'impianto A è maggiore di quello relativo all'impianto B.
La variabilità delle due distribuzioni sembra essere la stessa, se consideriamo come indice di variabilità lo scarto interquartile $Q_3 - Q_1$. Entrambe le distribuzioni presentano valori anomali (outliers).

Note: alcuni di voi hanno calolcato i boxplot non a partire dai dati originali, ma dalla loro distribuzione di frequenze assolute. E' evidente che ciò non ha senso e questo è un errore molto grave. Meglio studiare un po' di più….

d) Si calolino media e varianza delle due distribuzioni e si argomenti in merito alla qualità dei pezzi prodotti dai due impianti (usare almeno due posizioni decimali):

Impianto:		
	A	**B**
media	0.035	-0.05
varianza	0.018275	0.0175

Sembra migliore l'impianto A, perché in media i pezzi prodotti hanno uno scarto dal valore nominale che è minore in valore assoluto di quelli prodotti dall'impianto B (0.035 contro 0.05). Tuttavia, l'impianto A sembra essere meno preciso, perchè ha variabilità superiore a quella dell'impianto B.

Note: Nessuno di voi ha fatto questo commento.

e) Supponendo che ai due impianti vengano apportate delle modifiche tecniche tali che le differenze tra diametro dei pezzi prodotti e diametro nominale si dimezzino per l'impanto A e diminuiscano di 0.05 per l'impianto B, rispondere nuovamente alla domanda (d):

Le modifiche tecniche cambiano la qualità della produzione nei due impianti. Si tratta di due diverse trasformazioni lineari dei dati. Sia X_i^A il generico dato relativo all'impianto A prima delle modifiche, e Y_i^A il generico dato ottenuto dopo le modifiche tecniche. Con la stessa notazione per l'impianto B, abbiamo che i nuovi dati seguono le trasformazioni lineari:

$$Y_i^A = 0.5 \times X_i^A$$
$$Y_i^B = (1\text{-}0.05) \times X_i^B = 0.95 \times X_i^B$$

Allora, **senza bisogno di ricalcolare tutti i dati (come si è visto a lezione)**, si ha che per una trasformazione lineare del tipo:

$$Y = a + b \times X$$

$M(Y) = a + b \times M(X)$
$V(Y) = b^2 V(X)$

Dove M(X) e V(X) sono rispettivamente la media e la varianza di X. Per cui:
$M(Y^A) = 0.5 \times 0.035 = 0.0175$
$V(Y^A) = 0.5^2 \times 0.018275 = 0.25 \times 0.018275 = 0.00456875$

$M(Y^B) = 0.95 \times (-0.05) = -0.0475$
$V(Y^B) = 0.95^2 \times 0.0175 = 0.9025 \times 0.0175 = 0.01579375$

Note: anche questo argomento è stato trattato più volte a lezione.

2) La seguente tabella riporta la distribuzione di frequenze assolute del livello di soddisfazione espresso da 50 cittadini riguardo un certo servizio pubblico distintamente per maschi e femmine.

	sesso		
Soddisfazione	**Maschi**	**Femmine**	**Tot.**
scarsa	6	1	**7**
insuff	10	9	**19**

buona	7	14	**21**
ottima	2	1	**3**
Totale	**25**	**25**	**50**

a) Calcolare le frequenze relative della distribuzione doppia, delle distribuzioni della soddisfazione condizionate al sesso e della distribuzione marginale della soddisfazione:

Distribuzione delle frequenze relative doppia: $p_{ij} = f_{ij}/n$

p_{ij}	**Y = sesso**		
X = Soddisfazione	Maschi	Femmine	
scarsa	0.12	0.02	**0.14**
insuff	0.20	0.18	**0.38**
buona	0.14	0.28	**0.42**
ottima	0.04	0.02	**0.06**
	0.5	**0.5**	**1**

Distribuzione della soddisfazione condizionata al sesso: $P(X|Y) = p_{i|j} = f_{ij} / f_{\cdot j}$

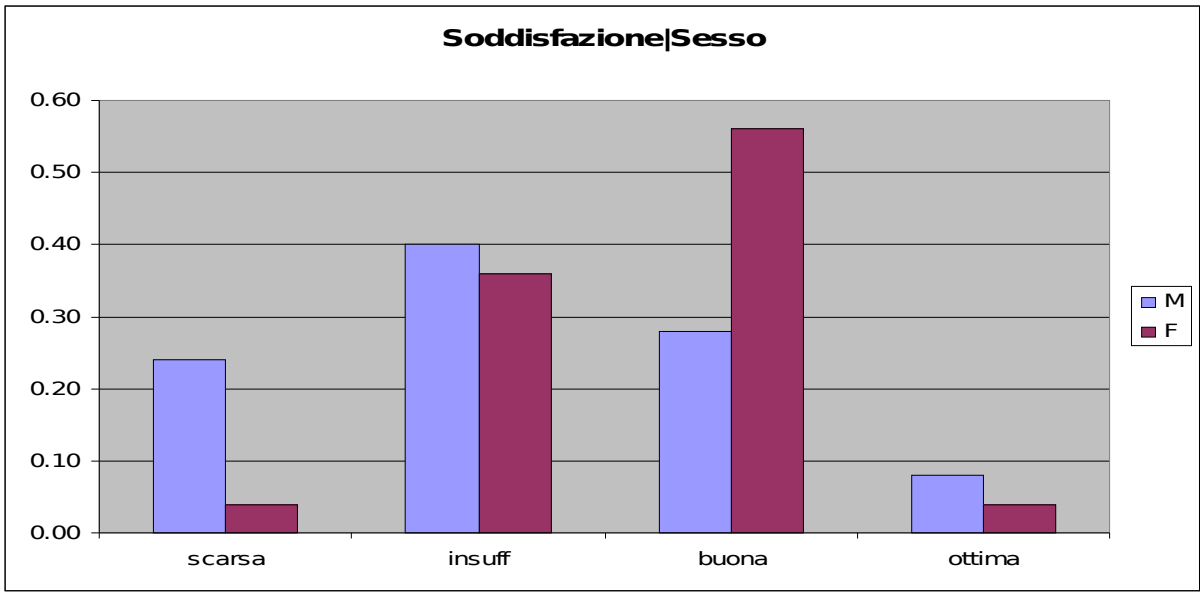
$p_{i|j}$		
Soddisfazione	**X| Y = M**	**X| Y = F**
scarsa	0.24	0.04
insuff	0.40	0.36
buona	0.28	0.56
ottima	0.08	0.04
	1	**1**

Distribuzione marginale della soddisfazione: $p_{i\cdot} = f_{i\cdot}/n$

Soddisfazione	$p_{i\cdot}$
scarsa	0.14
insuff	0.38
buona	0.42
ottima	0.06
	1

Note: alcuni di voi hanno calcolato la distribuzione del sesso condizionata alla soddisfazione, ma questo non era richiesto dall'esercizio, oltre ad avere scarsa utilità pratica come analisi.

b) Confrontare le due distribuzioni della soddisfazione condizionate al sesso graficamente e commentare:



Dal grafico si evidenzia che le femmine sono generalmente più soddisfatte del servizio.

c) Confrontare le due distribuzioni della soddisfazione condizionate al sesso in base ad un indice di posizione e commentare:

Come indice di posizione andava bene la moda ("Insufficiente" per i maschi e "buona" per le femmine), ma dal momento che **X** è una variabile ordinale, era anche possibile calcolare la mediana, che si può ottenere a partire dalle due distribuzioni condizionate. In questo modo si ricava che la mediana per i maschi è ancora la modalità "Insufficiente" e per le femmine la modalità "buona".

d) Confrontare le due distribuzioni della soddisfazione condizionate al sesso in base ad un indice di mutabilità normalizzato e commentare:

Gli indici di mutabilità che abbiamo visto a lezione sono due: l'indice di Gini e l'indice di Shannon. Andavano calcolati a partire dalle distribuzioni condizionate calcolate al punto (2a).

$$G = \Sigma_i \; p_{i|j} \times (1- p_{i|j}) = 1 - \Sigma_i \; p_{i|j}^2$$

$$H = -\Sigma_i \; p_{i|j} \log(p_{i|j})$$

Note: da notare che l'indice G per costruzione assume valori nell'intervallo [0,(K-1)/K] (dove K è il numero delle modalità della variabile oggetto di studio), mentre l'indice H assume valori in $R^+ \cup \{0\}$. Ad alcuni di voi è risultato un indice G non normalizzato maggiore di 1 (errore molto grave) e

a qualcun altro, peggio ancora, un indice negativo. Gli indici normalizzati per definizione assumono valori nell'intervallo [0,1]. Anche qui qualcuno di voi è riuscito ad ottenere indici maggiori di 1.

Gli indice suddetti assumono il valore minimo (zero) in corrispondenza di minima mutabilità, cioè quando tutta la distribuzione di frequenza è concentrata in una sola modalità. Assumono il valore massimo nella situazione di massima mutabilità, cioè quando la distribuzione di frequenza è distribuita in modo uniforme (o equi-distribuita), cioè quando $p_{i|j} = 1/K$. Quindi:

$$G_{max} = 1 - \Sigma_i \ (1/K)^2 = 1 - K/K^2 = 1-1/K = (K-1)/K$$

$$H_{max} = -\Sigma_i \ (1/K) \log(1/K) = -K/K \log(1/K) = -\log(1/K)$$

Gli indici non normalizzati e normalizzati risultavano:

$G^M = 0.6976$	$G^M/G_{max}^M = 0.9301$
$G^F = 0.5536$	$G^F/G_{max}^F = 0.7381$
$H^M = 0.5505$	$H^M/H_{max}^M = 0.9143$
$H^F = 0.4126$	$H^F/H_{max}^F = 0.6853$

Dagli indici calolati si evince che la mutabilità è maggiore nella distribuzione della soddisfazione condizionata di **X|Y=M**.

e) Calcolare le fequenze attese in caso di indipendenza tra soddisfazione e sesso e determinare il valore dell'indice χ^2 di Pearson:

f_{ij}*	**sesso**		
Soddisfazione	Maschi	Femmine	
scarsa	3.5	3.5	**7**
insuff	9.5	9.5	**19**
buona	10.5	10.5	**21**
ottima	1.5	1.5	**3**
	25	**25**	**50**

Poiché i totali di riga sono uguali, anche le frequenze attese sono uguali in ciascuna riga.

$$\chi^2 = 6.29$$

Se si volesse ottenere un indice normalizzato (anche se non era richiesto), si può ricorrere all'indice V di Cràmer:

$$V = \sqrt{\frac{X^2}{n \min(R-1, C-1)}} = 0.355$$

L'indice normalizzato indica una discreta dipendenza tra sesso e soddisfazione. Ovviamente V assume valori in [0,1].

3) La seguente tabella riporta per ciascuno di 10 maratoneti il numero di Km mediamente percorsi in allenamento in una settimana (X) e il tempo impiegato in una data gara (Y):

X	80	90	100	110	120	130	140	150	160	170
Y	150	160	145	142	157	150	163	155	140	137

a) Calcolare il coefficiente di correlazione lineare tra le due variabili. Commentare il risultato:

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} = \frac{COV(X,Y)}{\sqrt{V(X)V(Y)}}$$

M(X) = 125
M(Y) = 149.9
V(X) = 916.66
V(Y) = 77.87
COV(X,Y) = -75.5

ρ = **-0.3139**

Commento: tra le variabili considerate c'è una correlazione negativa, sembra cioè che all'aumentare dei Km percorsi in allenamento diminuisca il tempo impiegato nella gara. Il risultato non dovrebbe sorprendere, perché più uno si allena per la gara meglio riesce a farla, cioè nel in un tempo minore.

Note: le varianze negative non le voglio più vedere…. E il coefficiente di correlazione assume valori nell'intervallo [-1, 1]….

b) Calcolare i parametri della retta di regressione lineare che spiega il tempo di gara in funzione dei Silometri percorsi in allenamento. Commentare i risultati.

$\hat{\alpha}$ = 160.20
$\hat{\beta}$ = -0.08

Commento: ad ogni Km percorso in più in allenamento, il tempo impiegato in gara diminuisce di 0.08 unità, cioè di 4.8 secondi. L'intercetta rappresenta il tempo medio "di base", indipendentemente dall'allenamento, che risulta essere di 2 ore e 40 minuti.

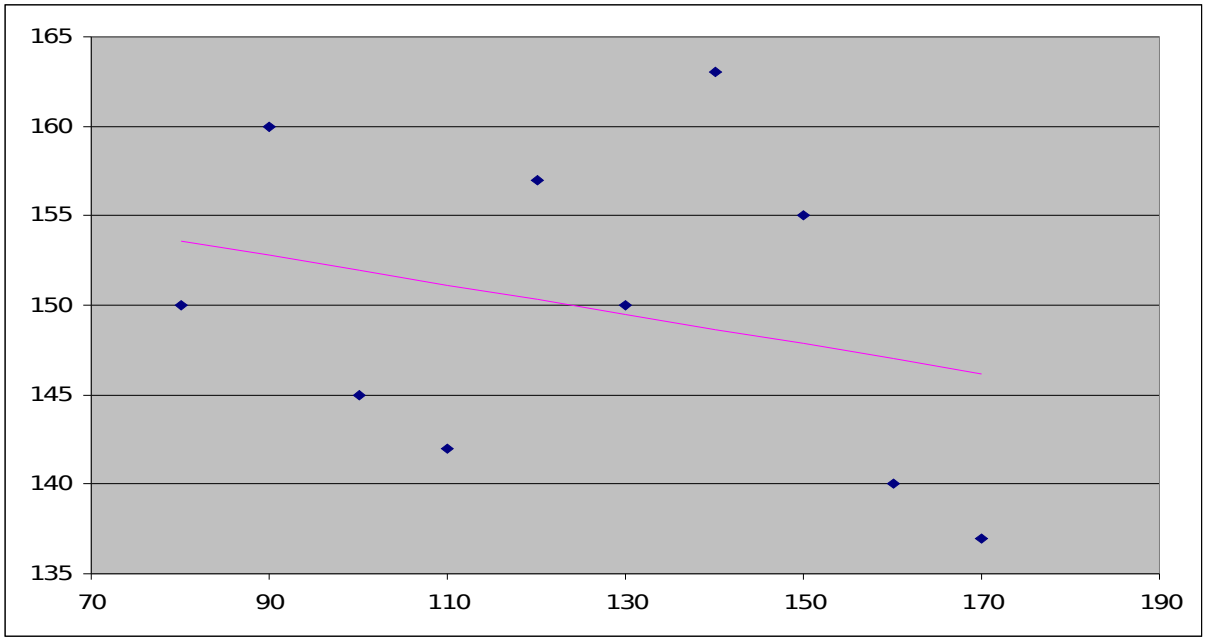
c) Calcolare il coefficiente di determinazione. Commentare il risultato.

$R^2 = \rho^2 = 0.0986$

Commento: il modello lineare ottenuto non è un buon modello, in quanto solo l'1% circa della variabilità della risposta (Y) è spiegato dalla variabile esplicativa (X).

Note: non voglio più vedere anche i coefficienti di determinazione negativi, perché questo è matematicamente assurdo. L'indice rivela nel simbolo che si tratta di un quadrato, positivo per definizione.

d) Tracciare un diagramma di dispersione su cui riportare la retta di regressione:



Note: quando si disegna la retta di regressione bisogna stare attenti alla scala utilizzata nella rappresentazione. Se la scala delle ascisse va da zero a 180 (ad esempio), allora la retta parte dal punto di coordinate [0, $\hat{\alpha}$] = [0, 160.20]. Se, come nel grafico sopra, la scala delle ascisse va da 70 a 190 allora il primo punto della retta ha coordinate [80, 153.6], dove 153.6 è il valore previsto dal modello in corrispondenza di X = 80. Ricordiamo inoltre che per disegnare una retta sono sufficienti due punti…..

e) Quanti Km bisognerebbe percorrere per ottenere un tempo pari a 2 ore e 20 minuti?

Innanzitutto bisogna trasformare il tempo in minuti, quindi il tempo di gara ambito è pari a 140 minuti. Poi si tratta di risolvere un'equazione lineare in cui è nota la Y. L'equazione è data dalla retta di regressione ottenuta al punto (3b):

$$Y = 160.20 – 0.08X$$

Quindi:

$$X = (Y - 160.20) / 0.08 = \mathbf{252.5\ Km}$$

La commisione