Correzione dell' Esame di Statistica Descrittiva (Mod. B)
1° Appello - 28 Marzo 2007
Facoltà di Astronomia

ESERCIZIO 1

La seguente tabella riporta la distribuzione congiunta della situazione lavorativa e dello stato civile di 1618 uomini in età 40-60:

	Sposato	**Celibe**	**Vedovo**	**Divorziato**	**Totale**
occupato	679	217	110	309	**1315**
disoccupato	63	30	25	21	**139**
pensionato	42	43	39	40	**164**
Totale	**784**	**290**	**174**	**370**	**1618**

1.1 Qual è l'unità statistica, quali sono le variabili rilevate e di che tipo.
1.2 Si ottenga la distribuzione (congiunta) delle frequenze relative.
1.3 Si può affermare che i pensionati sono equi – distribuiti rispetto allo stato civile? Rispondere a questa affermazione motivandola con un indice opportuno.
1.4 Si fornisca un indice di posizione per lo stato civile.

1.1
L'unità statistica è l'uomo in età 40-60 anni. Le variabili rilevate sono di tipo qualitativo sconnesso (categoriali non ordinali). Le variabili rilevate sono lo "Stato Civile" e la "Situazione Lavorativa".

1.2
La distribuzione congiunta delle frequenze relative è per definizione la distribuzione congiunta delle frequenze assolute divise per la numerosità del campione. Quindi:

fi	**sposato**	**celibe**	**vedovo**	**divorziato**	**totale**
occupato	679	217	110	309	**1315**
disoccupato	63	30	25	21	**139**
pensionato	42	43	39	40	**164**
totale	**784**	**290**	**174**	**370**	**1618**

Distribuzione congiunta di "Stato Civile" e "Situazione Lavorativa", frequenze assolute.

pi	**sposato**	**celibe**	**vedovo**	**divorziato**	**totale**
occupato	0.42	0.13	0.07	0.19	**0.81**
disoccupato	0.04	0.02	0.02	0.01	**0.09**
pensionato	0.03	0.03	0.02	0.02	**0.10**
totale	**0.48**	**0.18**	**0.11**	**0.23**	**1.00**

Distribuzione congiunta di "Stato Civile" e "Situazione Lavorativa", frequenze relative.

1.3
Per rispondere al quesito bisogna innanzitutto ottenere la distribuzione della variabile "Stato civile" condizionata alla modalità "pensionato" della variabile "Situazione Lavorativa". Quindi, se X = "Stato civile" e Y ="Situazione Lavorativa", si deve ottenere **p(X|Y="pensionato")**:

	sposato	**celibe**	**vedovo**	**divorziato**	**totale**
p(X|Y="pensionato")	0.256	0.262	0.238	0.244	1.000

Un indice opportuno per valutare l'equi-distribuzione dello "Stato Civile" è l'indice di Gini, o in alternativa, l'indice di Shannon:

$$G = 1 - \sum_i p(X_i|Y="pensionato") = 0.750$$

$$H = - \sum_i p(X_i|Y="pensionato)\ ln[p(X_i|Y="pensionato)] = 0.602$$

Gli indici così calcolati devono essere normalizzati in modo da assumere valori compresi tra 0 *e* 1. *Gli indici così ottenuti valgono* 0 *in caso di minima mutabilità e* 1 *in caso di massima mutabilità* (*o equidistribuzione*).

Nel nostro caso, detto K *il numero di modalità assunte dalla variabile "Stato Civile"*:

$$G_{norm} = G \times K/(K-1) = 1$$

$$H_{norm} = H / ln(K) = 0.99$$

Entrambi gli indici normalizzati assumono valori molto vicini a 1, *quindi si può concludere che i pensionati sono equi-distribuiti rispetto allo stato civile.*

ESERCIZIO 2

Per 200 laureati, di cui 100 hanno svolto un tirocinio presso una struttura privata e 100 presso una struttura pubblica, è stato rilevato il tempo di attesa dalla laurea al primo lavoro. I risultati sono stati raccolti nella seguente tabella:

	mesi			
tirocinio	**<1**	**[1-6)**	**>6**	
privato	53	45	2	**100**
pubblico	43	51	6	**100**
	96	**96**	**8**	**200**

2.1 Si ottengano le distribuzioni del tempo dalla laurea al primo lavoro condizionate alle modalità della variabile tirocinio.

2.2 Si rappresentino opportunamente le distribuzioni calcolate al punto precedente.

2.3 Si calcoli un indice opportuno per valutare se le variabili "tempo di attesa" e "tirocinio" sono stocasticamente indipendenti.

2.1

	Tempo d'attesa (mesi)			
Tirocinio	**<1**	**[1-6)**	**>6**	
privato	53	45	2	**100**
pubblico	43	51	6	**100**
	96	**96**	**8**	**200**

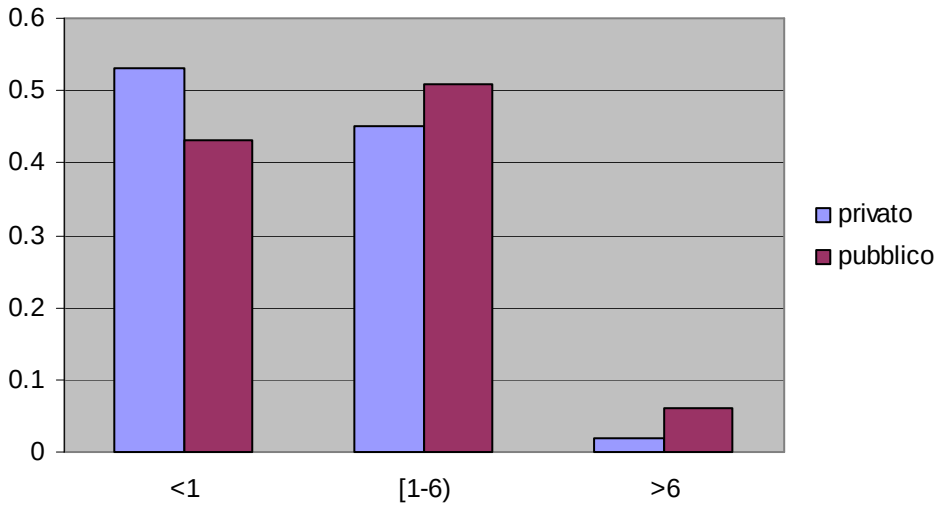
Distribuzione congiunta di X = *"Tempo d'attesa" e* Y = *"Tirocinio"*.

	Tempo d'attesa (mesi)			
Tirocinio	**<1**	**[1-6)**	**>6**	
p(X|Y="privato")	0.53	0.45	0.02	1
p(X|Y="pubblico")	0.43	0.51	0.06	1

Distribuzione condizionata di X = *"Tempo d'attesa" a* Y = *"Tirocinio"*.

2.2

La rappresentazione più opportuna è il diagramma a barre: si tratta la varibile "Tempo d'attesa" come una variabile categoriale (*ordinale*) *e si rappresentano le due distribuzioni condizionate appena ottenute:*



In alternativa, si poteva considerare la variabile "Tempo d'attesa" per quello che è, cioè una variabile quantitativa continua e rappresentare le distribuzioni condizionate tramite la funzione di ripartizione empirica (*bisogna però stabilire un punto "centrale" per la modalità "*>6*"*).

2.3

L'indice richiesto è il χ^2 *di Pearson:*

Tirocinio	**<1**	**[1-6)**	**>6**	
privato	48	48	4	**100**
pubblico	48	48	4	**100**
	96	**96**	**8**	**200**

Frequenze attese sotto l'ipotesi di Indipendenza Stocastica tra X *e* Y.

L'indice richiesto vale:

$$\chi^2 = 3.416$$

Per rispondere al quesito, tuttavia, è opportuno normalizzare l'indice in modo che esso assuma valori nell'intervallo [0,1]. *Anche in questo caso,* $\chi^2 = 0$ *implica l'indipendenza e* $\chi^2 = 1$ *implica la dipendenza funzionale tra le variabili. In alternativa, si può calcolare l'indice V di Cramér*:

$$\chi^2_{norm} = \chi^2 / [n \times min\,(C\text{-}1, R\text{-}1)] = 0.0085$$

$$V = [\chi^2_{norm}]^{1/2} = 0.092$$

In entrambi i casi l'indice è molto basso per cui si conclude che tra le due variabili considerate c'è indipendenza stocastica.

ESERCIZIO 3

La seguente tabella riporta il numero di incidenti stradali osservato nel periodo 1991-2000 nella regione Veneto:

X = Anno	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Y = Incidenti	5413	6122	6705	6824	7790	7698	8571	8688	9422	9904

Sono inoltre note le seguenti quantità:

$\sum_i x_i$	19955
$\sum_i y_i$	77137
$\sum_i x_i^2$	39820285
$\sum_i y_i^2$	614053443
$\sum_i x_i y_i$	10777283

3.1 Si rappresentino i dati in modo opportuno.

3.2 Si ottenga l'equazione della retta di regressione lineare:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

3.3 Si può affermare che gli incidenti stradali nel Veneto sono aumentati nel periodo 1991-2000? Spiegare perché.

3.4 Si ottenga una previsione del numero di incidenti stradali relativi all'anno in corso.

3.5 Si giudichi, tramite un indice opportuno, se il modello proposto si adatta bene ai dati.

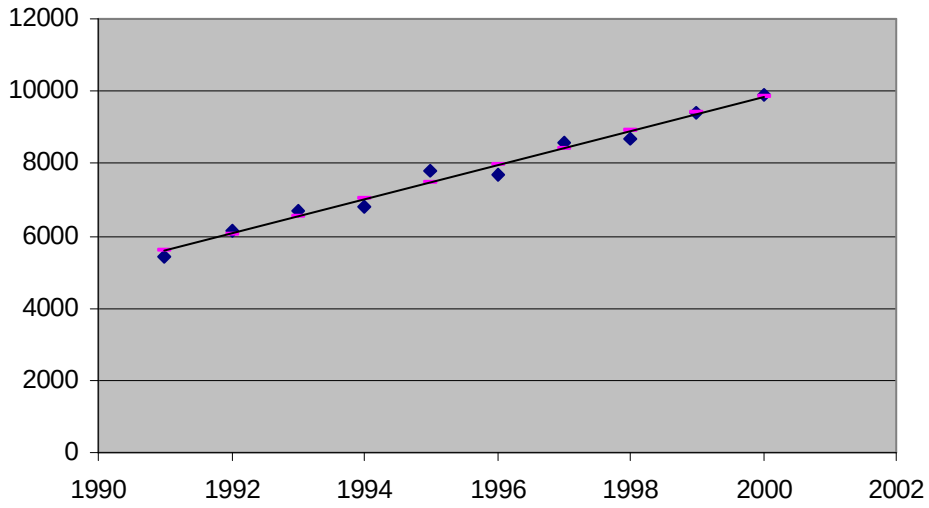
3.6 Al fine di calcolare l'ammontare dei risarcimenti (**Z**), una compagnia assicuratrice monopolista applica il seguente modello lineare basato sui costi fissi e sul numero di incidenti:

$$\hat{z}_i = 100{,}000 + 1.6\,\hat{y}_i$$

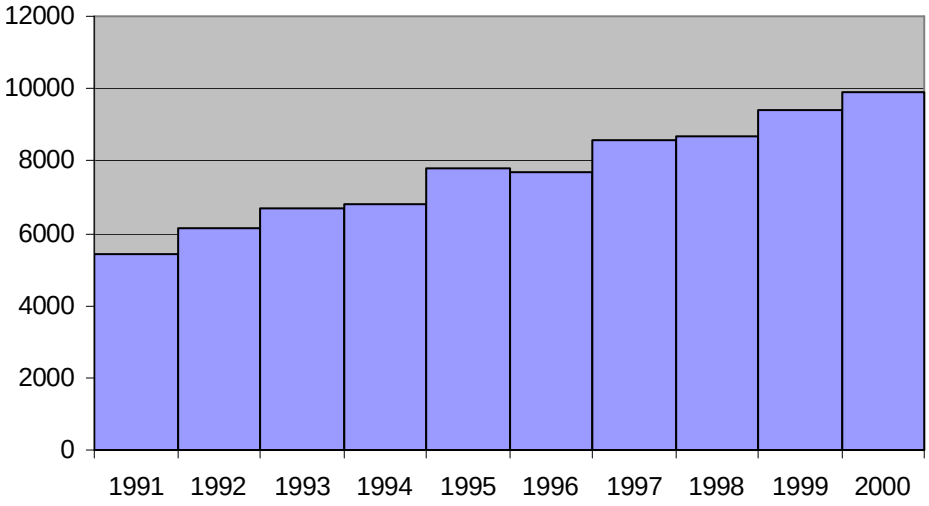
- si ottenga la spesa media sostenuta dall'assicurazione monopolista nel periodo 1991-2000.
- si ottenga la varianza di spesa per risarcimenti sostenuta dall'assicurazione monopolista nel periodo 1991-2000.

3.1

La rappresentazione più opportuna, visto che si tratta di un esercizio sulla regressione lineare, è il diagramma di dispersione:



In alternativa, anche l'istogramma poteva andare bene. Si noti però che in questo caso la variabile "Numero di incidenti" è intesa come "frequenza assoluta del numero di incidenti" e non come variabile quantitativa discreta. Poiché gli incidenti avvengono durante tutto l'anno (con continuità), la rappresentazione adeguata se si sceglie l'istogramma è la seguente (*barre attaccate*).



3.2

Per ottenere l'equazione della retta di regressione lineare bisogna innanzitutto ottenere le medie, le varianze e la covarianza delle variabili X = *"Anno" e* Y = *"Numero di Incidenti"*:

$$\bar{x} = \sum_i x_i / n = 1995.5$$

$$\bar{y} = \sum_i y_i / n = 7713.7$$

$$\sigma_X^2 = \sum_i x_i^2/n - \bar{x}^2 = 8.25$$

$$\sigma_Y^2 = \sum_i y_i^2/n - \bar{y}^2 = 1904176.61$$

$$\sigma_{XY} = \sum_i x_i\, y_i/n - \bar{x}\,\bar{y} = 3929.15$$

A questo punto:

$$\hat{\beta} = \frac{\sigma_{XY}}{\sigma_X^2} = 476.26$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = -942664.3$$

3.3

Si, si può affermare che gli incidenti stradali sono aumentati perché il coefficiente angolare della retta di regressione è positivo.

3.4

Per ottenere una previsione del numero di incidenti per l'anno in corso si ricorre al modello lineare appena determinato:

$$\hat{y}_{2007} = \hat{\alpha} + \hat{\beta} \times 2007 = 13191 \text{ (arrotondato all'intero più prossimo)}$$

3.5

Al fine di valutare la bontà del modello si ricorre al coefficiente di determinazione R^2, *che è equivalente al coefficiente di correlazione tra* X *e* Y *al quadrato*:

$$R^2 = \rho(X,Y)^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = 0.982$$

Questo coefficiente dà la proporzione di varianza spiegata sulla varianza totale di Y. *Essendo molto vicino al proprio massimo* (1), *si conclude che il modello si adatta quasi perfettamente ai dati* (*vedere anche il diagramma di dipersione*).

3.6

Questo esercizio si risolveva facilmente senza bisogno di fare chissà quali calcoli: *basta ricordare che la somma dei valori previsti* ($\hat{Y}$) *è uguale alla somma dei valori osservati* (y_i), *quindi anche le relative medie sono uguali. Il resto è la conseguenza delle proprietà della trasformazione lineare sulla media*:

$$\bar{z} = 100{,}000 + 1.6\bar{y} = 112341.9$$

Per quanto riguarda la varianza invece, si ha:

$$\sigma_Z^2 = 1.6^2\,\sigma(\hat{Y})^2 = 2.56\frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2 = 2.56 \times SSE(Y)$$

dove SSE(Y) *è la varianza spiegata dal modello. In conclusione*:

$$\sigma_Z^2 = 2.56 \times SSE(Y) = 2.56 \times \hat{\beta}^2\,\sigma_X^2 = 4709526.4$$