PROVA SCRITTA DI STATISTICA

cod. 4038 CLEA-CLAPI-CLEFIN-CLELI
cod. 5047 CLEA-CLAPI-CLEFIN-CLEMIT

5 Novembre 2003

SOLUZIONI
MOD. A

In 8 facoltà di un ateneo italiano vengono rilevati i seguenti dati campionari sui laureati:

F : facoltà
$\overline{P}$: *media* dei premi di laurea P = (voto finale – media esami in 110decimi)
S_P : scarto quadratico medio campionario di P (stima corretta)
n : numerosità del gruppo

I dati sono contenuti nella seguente tabella:

F	$\overline{P}$	S_p	n
Agraria	8.11	2.27	100
Farmacia	8.64	2.80	90
Giursisprudenza	4.86	3.01	336
Lettere e filosofia	4.91	2.91	304
Medicina Veterinaria	4.86	2.14	46
Medicina e Chirurgia	7.57	3.47	280
Scienze Matematiche, Fisiche e Naturali	7.97	2.47	458
Scienze Politiche	5.33	2.44	260

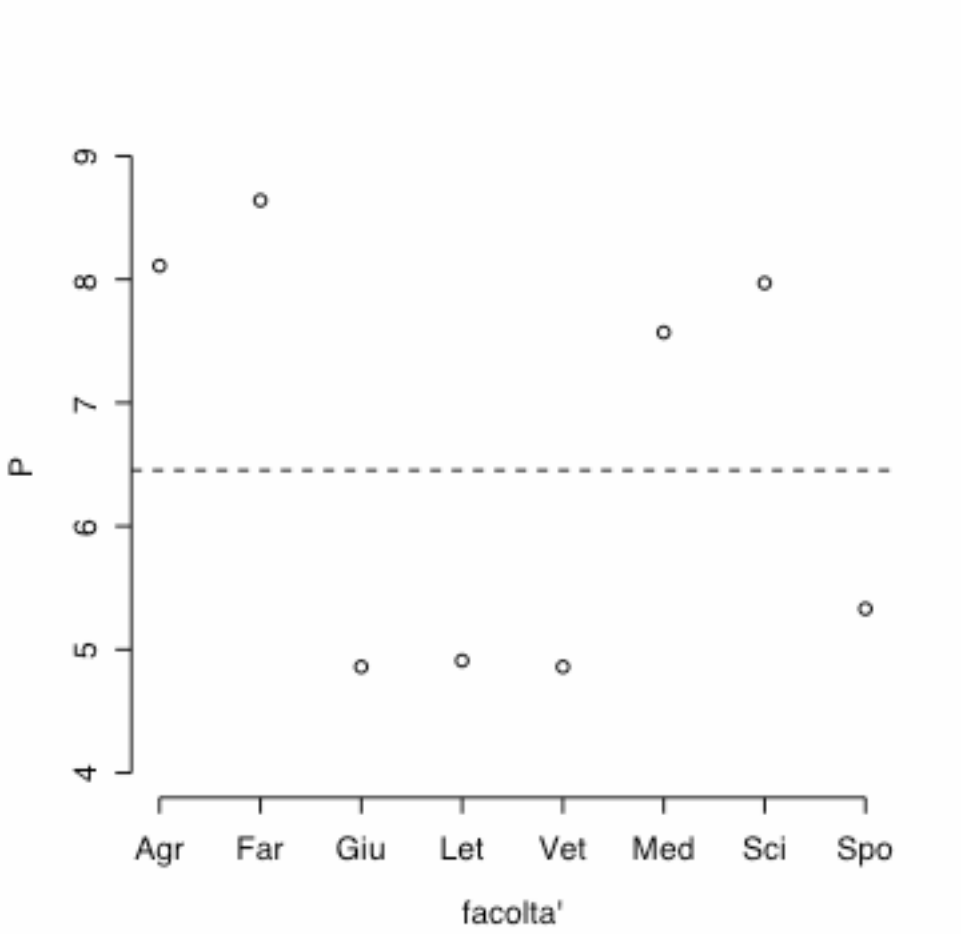
1. (3 punti) Si rappresenti la funzione di regressione di $\overline{P}$ al variare di F e si dica, argomentando in modo opportuno, se il premio di laurea è indipendente dalla facoltà in cui viene conseguito il titolo.

SOLUZIONE:

Media generale del premio di laurea :
*6.4533 = (8.11*100+8.64*90+4.86*336+4.91*304+4.86*46+7.57*280+7.97*458+5.33*260)/1874*

Se vi fosse indipendenza tra il premio di laurea e la facoltà di appartenenza dovremmo osservare “piccoli” scostamenti delle medie di P per facoltà e di tali medie dalla media generale. Si noti che per alcune facoltà (Giur., Lett., Vet. e Sc. Pol) le medie si trovano tutte al di sotto della media generale (linea tratteggiata nel

grafico). Qui "piccolo" deve intendersi in termini di scarto quadratico medio del fenomeno. In molti casi la distanza tra le medie di due facoltà è superiore allo scarto quadratico medio (si consideri ad esempio, la coppia Giurisprudenza/Agraria). Si conclude che non ci sono elementi per ritenere che vi sia indipendenza tra il premio di laurea medio e la facoltà in cui viene conseguito il titolo.



2. (2 punti) Si fornisca la definizione di indipendenza stocastica per due variabili casuali discrete X ed Y.

SOLUZIONE:
Sia X la v.c. che assume i k valori x_i con probabilità $P(X=x_i), i=1,\ldots,k$ e Y la v.c. che assume gli h valori y_j con probabilità $P(Y=y_j), j=1,\ldots,h$. Le due variabili casuali sono indipendenti se e solo se si verifica, per ogni coppia di indici i e j

$$P(X=x_i , Y=y_j) = P(X=x_i) * P(Y=y_j)$$

3. (7 punti) Con riferimento alla variabile $\overline{P}$ della tabella sopra riportata, assumendo l'ipotesi di indipendenza tra le commissioni di laurea delle diverse facoltà, si costruisca un test (di livello α=0.01) per verificare se vi è difformità di comportamento tra le facoltà di *Giurisprudenza* e *Agraria*.

a) (3 punti) Scrivere la statistica test (in simboli) e calcolarne il valore campionario.
b) (2 punti) Scrivere la regione di rifiuto del test (in simboli).
c) (2 punti) Si rifiuta oppure no l'ipotesi nulla? (Ovvero, c'è uniformità di comportamento?)
SOLUZIONE:
L'ipotesi nulla da sottoporre a test è $H_0 : \mu_1 = \mu_2$ contro l'alternativa che le medie siano diverse. Per poter eseguire il test si devono assumere le seguenti ipotesi: 1) il voto di laurea è una variabile casuale normale di

media m_1 per la prima facoltà e m_2 per la seconda facoltà; 2) le varianze nelle due popolazioni sono uguali, ovvero $s_1^2 = s_2^2 = s^2$.

Una stima corretta di s^2 è data dalla quantità

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(336-1)*3.01^2 + (100-1)*2.27^2}{336+100-2} = 8.1688$$

a) La statistica test risulta quindi essere

$$T = \frac{\overline{X_1} - \overline{X_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

dove $S_p = \sqrt{S_p^2}$ e $\overline{X}_1, \overline{X}_2$ sono i premi medi di laurea rispettivamente di Giurisprudenza e Agraria. Il valore campionario di *t* è pari a

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{4.86 - 8.11}{\sqrt{8.1688 * \left(\frac{1}{336} + \frac{1}{100}\right)}} = -9.9823$$

b) essendo *t* distribuita come una t di Student con (336+100-2) gradi di libertà, la regione di rifiuto del test è della seguente forma

$$R := \left\{ x_1, x_2 : | t | = \frac{| \bar{x}_1 - \bar{x}_2 |}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{1-\frac{a}{2}}^{(n_1+n_2-2)} \right\}$$

c) nel nostro caso |*t*| = 9.9823 è maggiore di $t_{1-\frac{a}{2}}^{(n_1+n_2-2)} = z_{1-\frac{a}{2}} = z_{0.995} = 2.57$, quindi si rifiuta l'ipotesi nulla di uguaglianza tra i comportamenti delle commissioni di laurea nelle due facoltà.

4. (6 punti) Per andare in una scuola di Milano, un bambino di città deve utilizzare la metropolitana M, mentre un bambino di campagna l'autobus A e la metropolitana M. Si supponga che i tempi di percorrenza della metropolitana M siano ben approssimati da una variabile casuale normale di media 15 e varianza 1, mentre quelli dell'autobus A da una variabile casuale normale di media 20 e varianza 4.

- **a) (2 punti)** Con quale probabilità un bambino di città impiega meno 17 minuti per arrivare a scuola?
- **b) (2 punti)** Si consideri una classe di 5 alunni. Con quale probabilità almeno 3 bambini di città arrivano a scuola nei primi 17 minuti? (Si utilizzino le opportune ipotesi necessarie ai calcoli esplicitandole nel testo della risposta).
- **c) (1 punto)** Quale variabile casuale descrive il tempo totale di viaggio di un bambino di campagna?
- **d) (1 punto)** Quanto tempo impiega in media un bambino di campagna per arrivare a scuola?

SOLUZIONE:

$$M \sim N(15,1), A \sim N(20,4)$$

a) $P(M<17)=P\left(Z<\frac{17-15}{1}\right)=\Phi(2)=0.9772$

b) Si assume l'indipendenza nel comportamento dei 5 bambini. Si utilizza la variabile casuale *X* Binomiale di parametri n=5 e $q=0.9772$. Dobbiamo quindi calcolare

$$P(X\geq 3)=P(X=3)+P(X=4)+P(X=5)$$

$$P(X=3)=\binom{5}{3}0.9772^3(1-0.9772)^2=0.0048$$

$$P(X=4)=\binom{5}{4}0.9772^4(1-0.9772)^1=0.1039$$

$$P(X=5)=\binom{5}{5}0.9772^5=0.8911$$

$$P(X\geq 3)=0.0048+0.1039+0.8911=0.9998$$

c) La variabile casuale $T = A+M$ descrive il tempo totale di viaggio, assumendo l'indipendenza tra *A* ed *M*, allora *T* è una varaibile casuale normale di media 20+15 =35 e varianza 4+1=5

d) La risposta è *E(T)=35*

5. (2 punti) Si consideri un fenomeno statistico X tale per cui la distanza tra il primo quartile e la mediana è molto più piccola di quella tra la mediana e il terzo quartile. Si dica se è credibile che la media si trovi a sinistra (cioè sia minore) della mediana circostanziando la risposta.

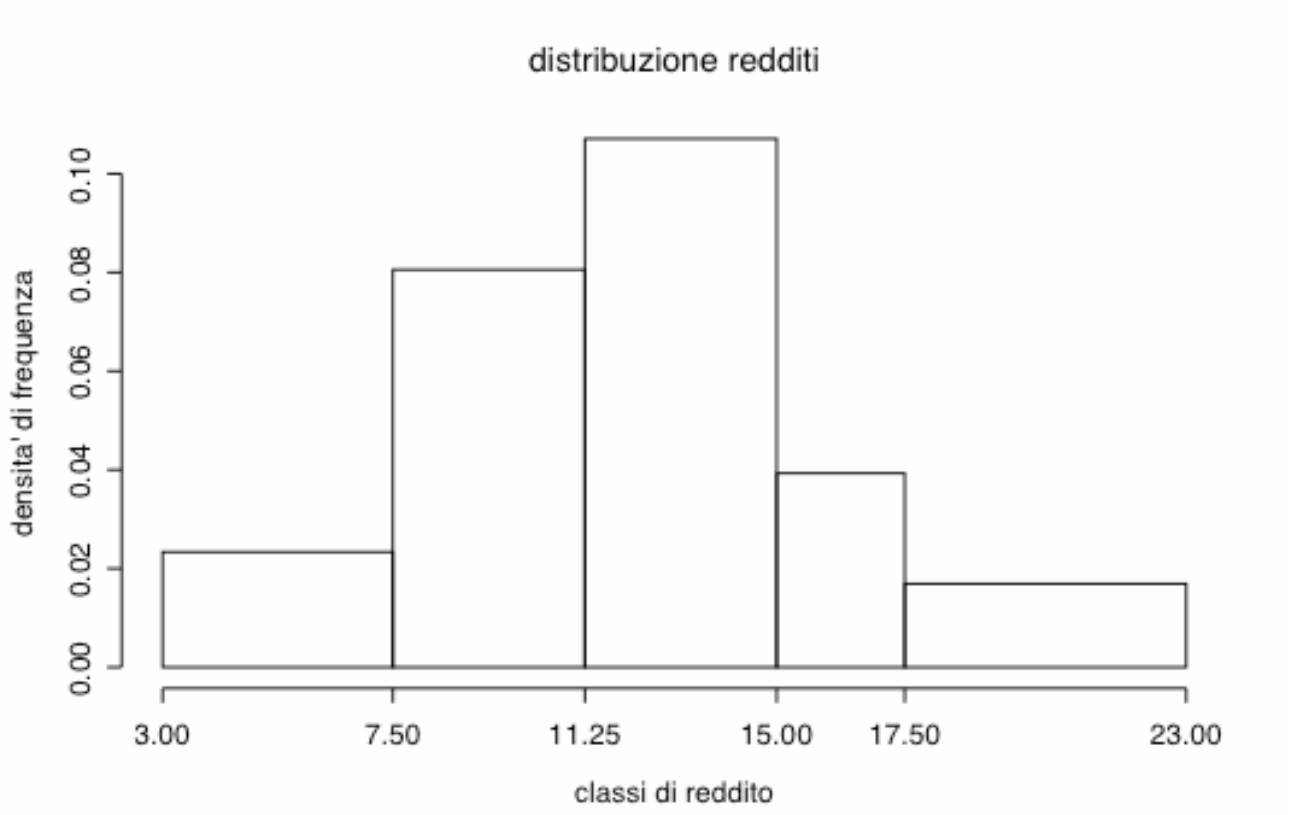
SOLUZIONE:
Un fenomeno come quello descritto presenta una distribuzione con una coda destra molto allungata. Poiché la media artimetica risente dei valori estremi, ci può aspettare che la media sia più grande della mediana, ovvero si trovi a destra della mediana. La risposta è quindi NO.

6. (7 punti) Si consideri la seguente distribuzione dei redditi relativi al primo impiego di neolaureati di un ateneo italiano.

Classe di reddito (in centinaia di euro)	redditieri	Reddito totale della classe (centinaia di euro)
[3.00 - 7.50]	183	732
(7.50 - 11.25]	526	5154
(11.25 - 15.00]	699	9087
(15.00 - 17.50]	171	2907
(17.50 - 23.00]	162	3078
Totale	1741	20958

a) **(3 punto)** Si rappresenti graficamente tale distribuzione
b) **(1 punto)** Per quale motivo in presenza di fenomeni quantitativi continui raccolti in classi, si fa preferibilmente riferimento alle densità di frequenza?
c) **(3 punti)** Si calcolino moda, media e mediana dei redditi dei neolaureati

Classe di reddito (in centinaia di euro)	redditieri	Reddito totale della classe (centinaia di euro)	Ampiezza della classe a_i	Frequenze relative p_i	Frequenze relative cumulate F_i	Densità di frequenza $c_i=p_i/a_i$
[3.00 - 7.50]	183	732	4.5	0.1051	0.1051	0.0234
(7.50 - 11.25]	526	5154	3.75	0.3021	0.4072	0.0806
(11.25 - 15.00]	699	9087	3.75	0.4015	0.8087	0.1071
(15.00 - 17.50]	171	2907	2.5	0.0982	0.9069	0.0393
(17.50 - 23.00]	162	3078	5.5	0.0930	1.0000	0.0169
Totale	1741	20958				



b) vedere Picarreta-Molteni

c) Come classe modale si sceglie la classe (11.25 – 15.00] poiché presenta densità di frequenza più elevata. Come moda si può assumere il valore centrale della classe: 13.125. Per quanto riguarda la media aritmetica si ricorre ai totali di classe (intensità) senza utilizzare i valori centrali delle classi, quindi: 20958/1741= 12.0379. Per il calcolo della mediana si applica la formula:

$$Me = x_i + (0.5 - F_{i-1})/c_i = 11.25 + (0.5 - 0.4072)/0.1071 = 12.1165$$

dove con i ci si riferisce alla classe (la terza) che contiene la mediana, cioè tale per cui le frequenze cumulate raggiungono almeno il valore di 0.5