CORSO DI STATISTICA DI BASE (Prof. GIORGIO ALLEVA)

Anno Accademico 2019-2020

Prova scritta dell'8 gennaio 2020

PROVA A

ESERCIZI (9 PUNTI ciascuno)

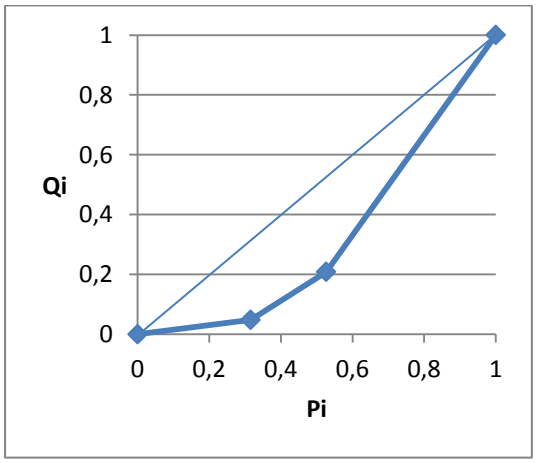
Esercizio 1. Si disponga della seguente distribuzione di frequenza delle variabili X e Y, osservate su un campione di 38 unità.

X\Y	*1*	*3*	*8*	*Tot*
1	2	4	6	12
5	-	2	6	8
11	10	-	8	18
Tot	12	6	20	38

a) Si misuri l'eterogeneità di X e si valuti se questa sia o meno elevata;
Si poteva utilizzare uno dei due seguenti indici di eterogeneità relativi (compresi tra o e 1).
$S_{rel} = [1- \Sigma f_u^2] / (k-1/k) = [1 – (12/38)^2- (8/38)^2- (18/38)^2] / (2/3) = 0,6316/0,6666 = 0,9474$
$H_{rel} = -(\Sigma f_u \log f_u)/\log k = -[(12/38)\log(12/38)+(8/38)\log(8/38)+(18/38)\log(18/38)]/(\log 3) = 0,443/0,477=0,9521$
(Essendo k=3, il numero di diverse modalità di X).

b) si determini l'intervallo che contiene l'80% dei dati campionari centrali di X;
Si tratta di determinare la differenza tra il nono e il primo decile $X_{0,9} – X_{0,1}$
$X_{0,1} = 1$ *(rango = 4)*
$X_{0,0} = 11$ *(rango = 35)*
Intervallo che contiene l'80% dei dati centrali: 1 ; 11 (ampiezza pari a 10)

c) si rappresenti la concentrazione di X attraverso il diagramma di Lorenz;
P_i: *frequenze cumulate relative:* 12/38 20/38 38/38
Q_i: *ammontari cumulati relativi:* 12/250 52/250 190/250
Coordinate P_i; Q_i = (0; 0) (0,3158; 0,048) (0,5263; 0,208) (1; 1)



d) si confronti la variabilità di X e di Z=2X attraverso: d1) la varianza; d2) il coefficiente di variazione.
Var(X)= $[(1-6,579)^2 12 + (5-6,579)^2 8 + (11-6,579)^2 18] / 38 = 19,612$ (oppure $\mu(X^2)–(\mu_X)^2= 62,895-6,579^2=19,612$)
Var(Z) =Var(2X) = 4 Var(X) = 78,449
CV(X)= σ_X / μ_X = rad(19,612)/6,579 = 0,673 = CV(Z) *in quanto* $\sigma_Z= 2\sigma_X$ *e* $\mu_Z= 2\mu_X$

Esercizio 2. Sulla base dei dati campionari contenuti nella precedente tabella, una volta specificate le assunzioni necessarie:

a) si determini l'intervallo di confidenza della media di X per α = 0,02 e si esponga il suo significato;

Essendo ignota la varianza σ^2 della popolazione: $p\left\{\overline{X} - t_{(n-1;\alpha/2)}\frac{\overline{S}}{\sqrt{n}} \le \mu \le \overline{X} + t_{(n-1;\alpha/2)}\frac{\overline{S}}{\sqrt{n}}\right\} = 1-\alpha$

Media campionaria (già calcolata in precedenza) = 6,579

Varianza calcolata in precedenza (S^2) = 19,612; varianza campionaria corretta = 19,612x38/37 = 20,142

S corretto = rad(20,142) =4,488 rad(n=38)=6,164

$t_{(37;\ 0,01)}$ = 2,431 $t_{(37;\ 0,01)}$ = - 2,431

p(6,579 - 2,431 x 4,488/6,164 ≤ μ ≤ 6,579 + 2,431 x 4,488/6,164) = p(4,809 ≤ μ ≤ 8,349)= 0,98

E' 0,98 la probabilità che l'intervallo 4,809 ; 8,349 contenga la media μ nella popolazione.

b) si indichi come varierebbe l'ampiezza di tale intervallo qualora il campione fosse di 61 unità (a parità delle altre caratteristiche);

A parità di media e varianza campionaria Rad(n=61)=7,81 $t_{(60;\ 0,01)}$ = 2,39 $t_{(60;\ 0,01)}$ = - 2,39

p(6,579 - 2,39 x 4,488/7,81 ≤ μ ≤ 6,579 + 2,39 x 4,488/7,81) = p(5,206 ≤ μ ≤ 7,952)= 0,98

c) si indichi sotto quali condizioni lo stimatore della media sia consistente;

Uno stimatore si dice consistente se converge in probabilità al parametro della popolazione, cioè se:

lim n→∞ p(|T-Θ|≤ ε) = 1. La condizione che assicura la consistenza è che l' MSE dello stimatore tenda a zero al crescere di n.

Essendo la media campionaria uno stimatore corretto di μ la condizione è che tenda a zero la sua varianza, che è σ^2/n. La media campionaria è quindi consistente perché la sua varianza tende a zero al crescere di n.

d) si verifichi attraverso un test se la varianza di X possa essere considerata nella popolazione uguale a 15 per α= 0,02;

H_0 : σ^2 = 15

H_1: $\sigma^2 \neq 15$

Essendo ignota μ rifiutiamo H_0 se $\frac{\sum_{1=1}^{n}(X_i - \overline{X})^2}{\sigma^2_0} \le \chi^2_{1-\alpha/2;n-1}$ *oppure*

$\frac{\sum_{1=1}^{n}(X_i - \overline{X})^2}{\sigma^2_0} \ge \chi^2_{\alpha/2;n-1}$

La devianza campionaria è uguale a 38 x la varianza campionaria già calcolata in precedenza = 38x19,612=745,2632

$\chi^2_{(37;\ 0,01)}$ = 19,96 $\chi^2_{(37;\ 0,99)}$ = 59,89 Dev Camp/σ^2_0 = 49,684

Quindi rifiutiamo se 49,684 ≤ 19,96 o se 49,64 ≥ 59,89

Poiché non è vero non rifiutiamo H_0, ossia non possiamo affermare che la varianza sia diversa da 15 con a=0,99

e) si indichi cosa si intenda per livello di significatività α e per potenza del test.

α è l'errore di prima specie, la probabilità di rifiutare H_0 quando è vera: α = p(S∈R|H_0).

1- β è la potenza del test, la probabilità di rifiutare H_0 quando questa è falsa: :1- β = p(S∈R|H_1), essendo β la probabilità di commettere l'errore di seconda specie.

(dove S è la statistica campionaria, R la regione di rifiuto, A la regione di accettazione).

Esercizio 3. Da una rilevazione di una popolazione si conoscano le informazioni seguenti:

$\mu(X) = 12$ $\mu(Y) = 15$ $\mu(X^2) = 225$ $\mu(Y^2) = 369$ $\mu(XY) = 85$

a) Determinare l'equazione della retta di regressione di Y su X e indicare il significato dei parametri;

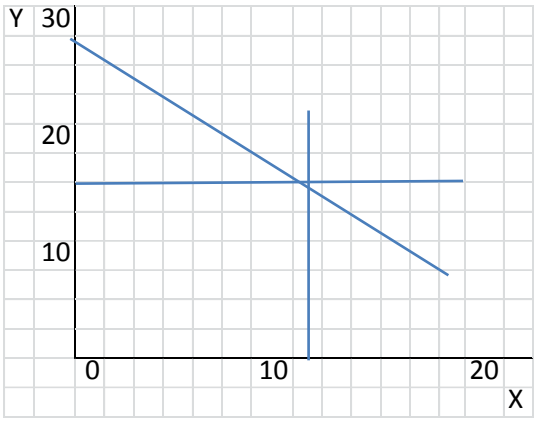
B_1=Cov(X,Y)/Var(X) =(85-12x15)/(225-12^2)=-95/81= - 1,173 $B_0 = \mu(Y) - \mu(X)$ *B_1 = 15 – (-1,173)12=29,074*

$\hat{Y}$ = 29,074 – 1,173 X

B_0 è l'intercetta della retta con l'asse Y, e stima il valore di Y quando X=0

B_1 è il coefficiente angolare della retta e stima quanto varia Y per un incremento unitario di X

rappresentare graficamente la retta e stimare Y per X = 18;



$\hat{Y}$ |X=18 = 29,074 – 1,173x18 = 7,963

b) misurare la bontà dell'adattamento della retta di regressione e indicare in che intervallo di valori sia compresa in generale tale misura;

Bontà di adattamento: r^2 = 0,74 *$0 \le r^2 \le 1$*

c) come varia il precedente intervallo qualora di conosca che $\eta^2_{Y|X}$= 0,85 e $\eta^2_{X|Y}$= 0,9 ?

$0 \le r^2 \le min(\eta^2_{Y|X}, \eta^2_{X|Y}) = 0,85$

d) aggiungendo nel modello una seconda variabile esplicativa Z si indichi l'equazione del piano di regressione e l'espressione della devianza residua di Y;$\hat{Y}$

$\hat{Y} = B_0 + B_1 X + B_2 Z$ *Devianza residua di Y = $\Sigma (Y - B_0 - B_1 X - B_2 Z)^2$* *anche = Dev(Y) $(1-R^2_{Y(X,Z)})$*

e) in quale caso la scelta della variabile Z comporta che il piano di regressione sia indeterminato?

Nel caso di perfetta collinearità ossia quando $r_{xz} = \pm 1$ In questo caso sia i parametri sia R^2 sono forme indeterminate 0/0.

QUESITI (barrare la risposta ritenuta esatta)

(PUNTI 2 per risposta corretta, PUNTI -1 per risposta sbagliata, PUNTI 0 per assenza di risposta)

Uno stimatore è corretto se:

- ❒ $E(\Theta) = T$
- ☒ $E(T) = \Theta$
- ❒ $MSE(T) = 0$

La curva di regressione di Y su X (o interpolatrice ottima):

- ❒ passa per le medie condizionate di X (*NO, passa per le medie condizionate di Y)*
- ❒ la residua di Y è la devianza esterna *(NO, la devianza residua è la devianza interna)*
- ☒ nessuna delle precedenti

L’indipendenza assoluta implica:

- ❒ l’indipendenza in media
- ❒ l’indipendenza lineare
- ☒ entrambe

L’indice dei prezzi di Laspeyres al tempo 1 in base 0 è una media aritmetica degli indici elementari dei prezzi con pesi:

- ☒ $p_0 q_0$
- ❒ $p_0 q_1$
- ❒ nessuno dei precedenti

La v.c. di Poisson con parametro λ:

- ❒ ha media e varianza entrambe uguali a λ
- ❒ è una v.c. discreta con $0 \leq X \leq \infty$
- ☒ entrambe

CORSO DI STATISTICA DI BASE (Prof. GIORGIO ALLEVA)

Anno Accademico 2019-2020

Prova scritta dell'8 gennaio 2020

PROVA B

ESERCIZI (9 PUNTI ciascuno)

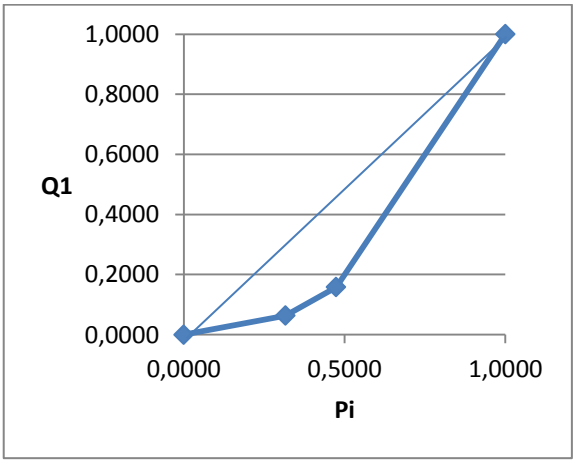
Esercizio 1. Si disponga della seguente distribuzione di frequenza delle due variabili X e Y, osservate su un campione di 38 unità.

X\Y	*1*	*3*	*8*	*Tot*
1	2	4	6	12
5	-	2	6	8
11	10	-	8	18
Tot	12	6	20	38

a) Si determinino i quartili di Y e si misurino, sulla base di questi, la variabilità e l'asimmetria della variabile Y;
Q1 = 1 (rango 10) *Q2 = 8 (rango 19 e 20)* *Q3 = 8 (rango 30)*
Differenza interquartile = Q3-Q1 = 7 *Asimmetria = (Q3-Q2) – (Q2-Q1) = -7*

b) si calcoli il coefficiente di variazione di Y:
media campionaria = (1x12 + 3x6 + 8x20)/38 = 5 *media dei quadrati = (1x12 + 9x6 + 64x20)/38 = 35,421*
varianza campionaria = 35,421-25= 10,421 *varianza campionaria corretta = 10,421 x 38/37 = 10,703*
sqm campionario= 3,228 *sqm campionario corretto = 3,271*
CV campionario = 0,646 *CV campionario corretto = 0,654*

c) si rappresenti la concentrazione di Y attraverso il diagramma di Lorenz Y.
P_i: frequenze cumulate relative: 12/38 18/38 38/38
Q_i: ammontari cumulati relativi 12/190 30/190 190/190
Coordinate P_i; Q_i = (0; 0) (0,3158; 0,0632) (0,4737; 0,1579) (1; 1)



Esercizio 2. Sulla base dei dati campionari contenuti nella precedente tabella, una volta specificate le assunzioni necessarie:

Si assume che il campione provenga da n v.c. estrazione NIID (normali, indipendenti, identicamente distribuite)

a) si determini l'intervallo di confidenza della varianza di Y per α = 0,02 e si esponga il suo significato;

Essendo ignota la media μ nella popolazione: $$p\left\{\frac{\sum(X_i-\bar{X})^2}{\chi^2_{(n-1;\alpha/2)}} \le \sigma^2 \le \frac{\sum(X_i-\bar{X})^2}{\chi^2_{(n-1;1-\alpha/2)}}\right\} = 1-\alpha$$

Media campionaria (già calcolata in precedenza) = 5

Devianza campionaria = n S^2 (già calcolato in precedenza) = 38x10,421=396

$\chi^2_{(37;\,0,01)}$ = 19,96 $\chi^2_{(37;\,0,99)}$ = 59,89

p(396/59,89 ≤ σ^2 ≤ 396/19,96) p(6,6121 ≤ σ^2 ≤ 19,84)= 0,98

E' 0,98 la probabilità che l'intervallo 6,6121 ; 19,84 contenga la varianza σ^2 nella popolazione.

b) si verifichi attraverso un test se la media di Y possa essere considerata maggiore di 4 con un livello di significatività dell'1%;

H_0 : $\mu = 4$

H_1: $\mu > 4$

Rifiutiamo H_0 se $\overline{X} \geq \mu_0 + t_{\alpha;n-1}\frac{\overline{S}}{\sqrt{n}}$

Cioè se 5 ≥ 4 + 2,431 (3,271/0,531) = 5,29 poiché non è vero non rifiutiamo H_0, ossia non possiamo affermare che la media sia maggiore di 4 con a=0,99

(essendo $t_{370,99}$=2,431; sqm corretto =3,271 e la radice di 38 = 0,531)

c) si mostri se la decisione sarebbe la stessa qualora si conoscesse che la varianza della popolazione fosse pari a 5;

Rifiutiamo se $\overline{X} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

Cioè se 5 ≥ 4 + 2,326 (2,236/0,531) = 4,844 poiché è vero rifiutiamo H_0, ossia possiamo affermare che la media sia maggiore di 4 con a=0,99

(essendo z_{99}=2,326; sqm nella popolazione =2,236 e la radice di 38 = 0,531)

d) indicare cosa si intenda per livello di significatività α e per potenza del test;

α è l'errore di prima specie, la probabilità di rifiutare H_0 quando è vera: $\alpha = p(S \in R \mid H_0)$.

1- β è la potenza del test, la probabilità di rifiutare H_0 quando questa è falsa: :1- $\beta = p(S \in R \mid H_1)$, essendo β la probabilità di commettere l'errore di seconda specie.

(dove S è la statistica campionaria, R la regione di rifiuto, A la regione di accettazione).

e) Indicare sotto quale condizioni lo stimatore della media sia corretto e pienamente efficiente.

Uno stimatore è corretto se il suo valore atteso è uguale al parametro $E(T)=\Theta$.

Uno stimatore è pienamente efficiente se la sua varianza è minima.

Se si stima la media con il metodo dei minimi quadrati lo stimatore è corretto e pienamente efficiente se valgono le condizioni del teorema di Gauss Markov (residui con media nulla, omoschedastici e incorrelati tra loro).

Esercizio 3. Da una rilevazione condotta su una popolazione di 120 unità si conoscano le informazioni seguenti:

$\mu(X) = 8$ $\mu(Y) = 10$ $\mu(X^2) = 80$ $\mu(Y^2) = 200$ $\mu(XY) = 44$

a) Determinare l'equazione della retta di regressione di Y su X e indicare il significato dei parametri;

B_1=Cov(X,Y)/Var(X) =(44-8x10)/(80-64^2)=-36/16= -2,25 *$B_0 = \mu(Y) - \mu(X)$ B_1 = 15 –8 (-2,25) = 28*

$\hat{Y}$ = 28 - 2,25 X

B_0 è l'intercetta della retta con l'asse Y, e stima il valore di Y quando X=0

B_1 è il coefficiente angolare della retta e stima quanto varia Y per un incremento unitario di X

b) rappresentare graficamente la retta e stimare Y per X = 12;

$\hat{Y}$ |X=12 = 28 - 2,25x12 = 1

c) misurare la devianza spiegata e la devianza residua di Y dalla retta di regressione, valutando la bontà di adattamento della retta di regressione;

Bontà di adattamento: r^2 = 0,81

Devianza spiegata = Dev(Y) r^2 =12.000x0,81 = 9.720

Devianza residua = Dev(Y) (1-r^2) = 12.000x0,19 =2.280

(Essendo Dev(Y) = n Var(Y) = 120 (200-10^2) = 120x100 =12.000)

d) aggiungendo nel modello una seconda variabile esplicativa Z si indichi l'equazione del piano di regressione e l'espressione della devianza residua di Y;

$\hat{Y} = B_0 + B_1 X + B_2 Z$ *Devianza residua di Y = $\Sigma (Y - B_0 - B_1 X - B_2 Z)^2$* *anche = Dev(Y) (1-$R^2_{Y(X,Z)}$)*

e) in quale caso la scelta della variabile Z comporta che il piano di regressione sia indeterminato?

Nel caso di perfetta collinearità ossia quando $r_{xz} = \pm 1$. In questo caso sia i parametri sia R^2 sono forme indeterminate 0/0.

QUESITI (barrare la risposta ritenuta esatta)

(PUNTI 2 per risposta corretta, PUNTI -1 per risposta sbagliata, PUNTI 0 per assenza di risposta)

Uno stimatore T_1 è più efficiente di uno stimatore T_2 se:

- ❒ $\sigma^2(T_1)$ tende a zero al crescere della dimensione del campione
- ❒ $\sigma^2(T_1) < \sigma^2(T_2)$ *(NO, solo se lo stimatore è corretto)*
- ☒ $MSE(T_1) < MSE(T_2)$

La curva di regressione di Y su (o interpolatrice ottima):

- ❒ passa per le medie condizionate di X (*NO, passa per le medie condizionate di Y)*

- ❒ rende minima la devianza esterna *(NO, è minima la devianza residua, che è la devianza interna)*
- ☒ nessuna delle precedenti

$\chi^2 = 0$ implica:

- ❒ $\eta^2_{X|Y} = 0$
- ❒ $r^2 = 0$
- ☒ $\eta^2_{X|Y} = r^2 = 0$

L'indice dei prezzi di Paasche al tempo 1 in base 0 è una media aritmetica degli indici elementari *dei prezzi* con pesi:

- ❒ $p_0\, q_0$
- ❒ $p_1\, q_1$
- ☒ nessuno dei precedenti *(i pesi sono $p_0\, q_1$)*

La v.c. binomiale con parametro p:

- ❒ ha media pari a np e varianza pari a npq
- ❒ è una v.c. discreta con $0 \le X \le n$
- ☒ entrambe