Università degli studi di Ferrara Dipartimento di Matematica

A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

SIMULAZIONE ESAME 2020

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: lunedì su appuntamento previa mail

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Nome e cognome:_____ Numero di matricola: _____

Corso di Laurea Magistrale in Matematica

STATISTICA MULTIVARIATA

Crediti formativi 6

A.A. 2018/2019

_____

Appello XXXXXXXXXX

Aula A1

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

INDICAZIONI PER GLI STUDENTI

L'esame è interamente scritto e individuale. Non è ammesso parlare durante l'esame. Non è ammesso l'utilizzo di apparecchi elettronici, telefoni e simili, calcolatrici programmabili. Non è consentito l'utilizzo di note, libri, testi e parti di testo o fogli diversi da quelli consegnati dal docente. Non è consentito copiare né confrontare le proprie risposte durante la sessione d'esame.

Ogni studente avrà sul banco penna e tessera universitaria accompagnata da documento di riconoscimento. Borse, zaini, cappotti e simili vengono depositati ai lati dell'aula.

La copiatura è sanzionata con il ritiro dell'esame, il conseguente annullamento e la segnalazione del caso alla Commissione Didattica.

Il tempo a disposizione per completare l'esame è 90 minuti. Chi termina prima del tempo totale a disposizione può consegnare l'intera documentazione d'esame e uscire dall'aula senza turbare la tranquillità dei colleghi. Chi si ritira può uscire dall'aula dopo aver consegnato tutta la documentazione d'esame.

L'esame è composto da tre parti:

1) Parte teorica: domande a risposta multipla (15 domande: 11 punti)
2) Parte pratica: domande a risposta multipla ed esercizi a completamento (15 domande: 11 punti)
3) Parte applicata utilizzando R: eseguire l'analisi richiesta utilizzando R. Salvare lo script eseguito, i commenti inseriti e la conclusione analitica che date. Tutte le componenti verranno valutate (8 punti)

La risposta deve essere chiaramente indicata con una croce apposta sulla lettera scelta. Non sono ammesse correzioni, risposte multiple, cambi di risposta; per questo si chiede massima attenzione nell'indicazione della risposta scelta. Ogni risposta sbagliata o non data, ogni risposta multipla o non chiara vale zero punti. La completezza e correttezza dell'esame in tutte le sue parti porta ad una valutazione di 30/30. La lode indica uno sforzo aggiuntivo sia in termini applicati che interpretativi.

La lettura attenta delle domande è un prerequisito fondamentale per la comprensione e la conseguente buona riuscita dell'esame.

Lo script, i commenti e la conclusione deve essere salvato sul Desktop del pc in uso: verrà trasformato in pdf non trasformabile e salvato su pendrive in vostra presenza. Una stampa dell'analisi verrà allegata al vostro esame.

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

PARTE TEORICA

Q1. Il prodotto tra un vettore riga e un vettore colonna è

a) una matrice
b) uno scalare
c) una matrice diagonale

Q2. Nell'analisi della regressione lineare semplice, l'approccio più comune per stimare i coefficienti di regressione è il metodo dei minimi quadrati, attraverso il quale:

a) si minimizza la somma dei quadrati degli scarti tra i valori osservati e quelli stimati
b) si minimizza la differenza dei quadrati degli scarti tra i valori osservati e quelli stimati
c) si massimizza la differenza dei quadrati degli scarti tra i valori osservati e quelli stimati

Q3. Quali delle seguenti caratteristiche non fa parte delle assunzioni base del modello di regressione lineare?

a) Normalità dei residui
b) Costanza dei residui
c) Linearità dei residui

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q4. All'interno della'Analisi di regressione lineare semplice, per testare la presenza di una relazione lineare nella popolazione si imposta il seguente sistema di ipotesi:

a) H0: β_1=0 ; H1: $\beta_1 \neq 0$
b) H0: $\beta_1 \neq$0 ; H1: $\beta_1 \neq 0$
c) H0: α=0 ; H1: $\alpha \neq 0$

Q5. Nell'analisi di regressione lineare multipla, una misura della multicollinearità per la *j-esima* variabile è fornita da:

a) Lindice CIF_j, pari a $\frac{1}{1+r_j}$
b) L'indice VIF_j, pari a $\frac{1}{1-R^2{}_j}$
c) L'indice r, pari a $\frac{1}{R_j^2}$

Q6. Nell'analisi di regressione lineare multipla, la significatività generale del modello è indicata da:

a) T test
b) R^2
c) F test

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q.7 Nell'analisi per componenti principali (anche nota come Principal Component Analysis), la prima componente principale (CP) è:

a) la combinazione lineare delle n variabili di partenza avente massima varianza

b) la combinazione lineare delle n variabili di partenza avente minima varianza

c) la variabile di partenza con minima varianza

Q.8 Nell'analisi per componenti principali, il criterio di scelta del numero di componenti non si basa su:

a) soglia minima di variabilità totale spiegata (ca. 70%)

b) auto valori maggiori di uno

c) analisi del termine di errore

Q.9 Nell'analisi fattoriale l'analisi delle correlazioni è fondamentale. Uno dei test applicati è il test della sfericità di Bartlett, il quale tuttavia:

a) dipende dal numero delle variabili e dalla numerosità del campione

b) dipende dalla tipologia di variabili

c) dipende dalla tipologia dei soggetti intervistati

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q. 10 L'analisi fattoriale esplorativa (AFE) e l'analisi fattoriale confermativa (AFC) si basano su due procedure diverse, infatti:

a) l'AFE si fonda su una procedura induttiva e l'AFC su una procedura deduttiva

b) l'AFE si basa sulla verifica delle ipotesi di ricerca e l'AFC si basa sulla osservazione del metodo

c) l'AFE si fonda su una procedura deduttiva e l'AFC su una procedura induttiva

Q11. Considerando l'analisi fattoriale e l'analisi per componenti principali, quali delle seguenti affermazioni è falsa?

a) Nella analisi fattoriale si distingue tra fattori comuni e fattori di unicità, mentre nell'analisi per componenti principali si hanno solo fattori comuni

b) Nell'analisi fattoriale la comunalità è sconosciuta e deve essere stimata, mentre nell'analisi per componenti principali la comunalità è pari a 1.

c) Nell'analisi fattoriale il numero di fattori comuni è pari al numero delle variabili osservate, mentre per l'analisi per componenti principali il numero di componenti è inferiore al numero di variabili osservate.

Q12. Considerando i metodi di rotazione applicabili ad un'analisi fattoriale, il metodo "Varimax" è:

a) Un metodo di rotazione ortogonale

b) Un metodo di rotazione obliquo

c) Un metodo di rotazione misto

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q13. Una delle fasi centrali nell'analisi per gruppi è la misurazione di similarità o diversità fra unità statistiche attraverso la distanza. Considerando la metrica di Minkoski $[d_m(i,i')]$, indicare tra le seguenti l'affermazione corretta:

a) Quando m=1, vi è una corrispondenza con la distanza euclidea
b) Quando m=2, vi è una corrispondenza con la distanza euclidea
c) Quando m=3, vi è una corrispondenza con la distanza euclidea

Q14. Nell'analisi per gruppi (o Cluster Analysis), nel procedimento di definizione dei gruppi attraverso strategia gerarchica, il metodo del legame singolo si basa su:

a) Un criterio di distanza massima tra unità
b) Un criterio di distanza minima tra unità
c) Un criterio di distanza media tra unità

Q15. In una analisi per gruppi (Cluster Analysis), si consideri WD quale devianza nei gruppi, TD devianza totale e g il numero di gruppi. Quale tra le seguenti affermazioni riguardanti l'indice R^2 è vera?

a) R^2 cresce all'aumentare di g
b) R^2 cresce al diminuire di g
c) R^2 non è una funzione monotona di g

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

PARTE PRATICA

Q1. Data la matrice quadrata $A = \begin{bmatrix} 2 & -2 & 3 \\ 1 & 1 & 1 \\ 1 & 3 & -1 \end{bmatrix}$ e uno scalare λ, l'autovalore con massimo valore assoluto (detto dominante) è pari a:

a)3
b)4
c)2

Q2. Data la matrice $A = \begin{pmatrix} 2 & 5 & 1 \\ 7 & 5 & 9 \\ 9 & 7 & 5 \end{pmatrix}$, la traccia di A = tr(A) è pari a:

a)12
b)2*5*5
c)15

Q3. Data la matrice $A = \begin{bmatrix} 3 & 4 & 1 \\ 4 & 6 & 2 \\ 5 & 7 & 4 \end{bmatrix}$, il determinante di a *det(A)* è pari a

a) - 4
b) 4
c) 16

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q4. Si consideri l'output sotto riportato:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indicare il modello di regressione semplice stimato:

a) Y = 0.5262+X_i*0.1569

b) Y = 1.833 + X_i*10.641

c) Y = 0.9645+ X_i*1.6699

Q5. Si consideri l'output sotto riportato:

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x           1 105.748 105.748  113.23 1.823e-07 ***
Residuals 12  11.207   0.934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Indicare il valore di R^2 relativo all'analisi di regressione lineare effettuata risulta essere:

a) 0.90871

b) 0.90417

c) 0.87904

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q6. Si consideri l'output sotto riportato relativo all'analisi di regressione lineare semplice:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.87406 -0.74834  0.08121  0.86255  1.15032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9645     0.5262   1.833   0.0917 .
x             1.6699     0.1569  10.641 1.82e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9664 on 12 degrees of freedom
```

Indicare il valore del *t-statistico (*grandezza utile per testare l'esistenza di relazione lineare tra le variabili x e y nella realtà):

a)10.641
b)1.6699
c) 1.833

Q7. Si osservi l'output di seguito riportato, relativo ad un'analisi per componenti principali:

```
> summary(pcal)
Importance of components%s:
                            PC1    PC2    PC3
Standard deviation     1.2629 0.9946 0.6448
Proportion of Variance 0.5316 0.3297 0.1386
Cumulative Proportion  0.5316 0.8614 1.0000
```

seguendo il metodo della varianza comune cumulata, quante componenti si estraggono?

a) 1
b) 2
c) 3

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

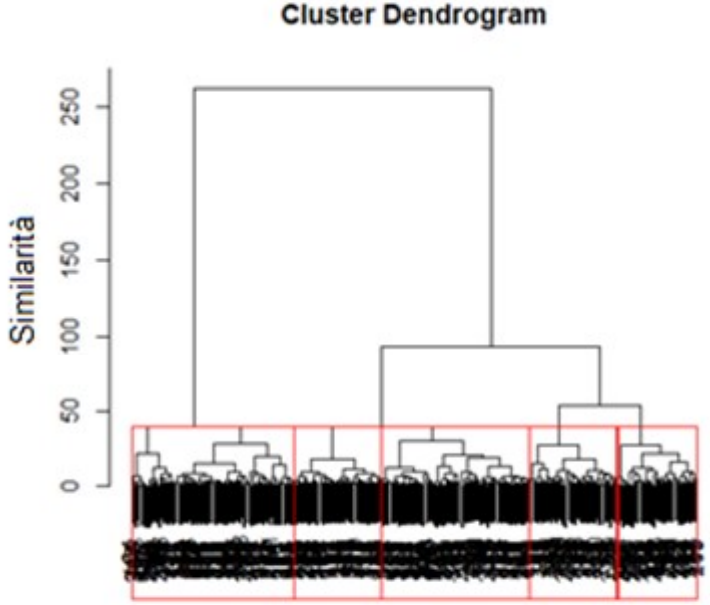
Q8. L'immagine di seguito riportata si riferisce ad uno scree-plot relativo al grafico decrescente degli auto valori di un'analisi fattoriale esplorativa.

In base al grafico, quanti fattori non si dovrebbero estrarre?

a) 1 (teoria di Citrouille)
b) 2 (teoria di Harman)
c) 3 (teoria di Cattell)

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q9. Si osservi l'immagine sotto riportata relativa ad una analisi per gruppi (Cluster Analysisi) condotta su un campione di 386 rispondenti ad un questionario sulle preferenze relative alla tipologia di vino consumato.



Ad un livello di similarità pari a 75, quanti gruppi possono essere individuati?

a) 2
b) 3
c) 5

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q10. Da un'analisi di regressione lineare multipla si ottiene il seguente risultato:

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.1751 -0.4982  0.1616  0.6278  2.3758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12803    0.17372   0.737 0.000331 ***
LIKE_AROMA   0.42853    0.05601   7.652 1.63e-13 ***
LIKE_SWEET   0.19714    0.05441   3.623 0.000331 ***
LIKE_TASTE   0.24836    0.05409   4.591 5.99e-06 ***
---
:

Residual standard error: 1.042 on 382 degrees of freedom
Multiple R-squared:  0.5919,    Adjusted R-squared:  0.5887
F-statistic: 184.7 on 3 and 382 DF,  p-value: < 2.2e-16
```

Utilizzando i valori ottenuti, il modello studiato è:

a) Y=0.128+0.428*x1+0.197*x2+0.248*x3
b) Y=0.128+0.056*x1+0.0544*x2+0.0540*x3
c) Y=0.1737+0.056*x1+0.0544*x2+0.0540*x3

Q11. Osservando l'output riportato in Q.10, quale percentuale di variabilità di Y (LIKE_PAS) è spiegata dal modello?

a) circa il 10%
b) circa il 59%
c) circa il 62%

Q12. Se nell'analisi che ha prodotto l'output riportato in Q.10, il VIF è maggiore di 5 allora possiamo dire che:

a) vi è presenza di eteroschedasticità
b) vi presenza di multicollinearità
c) nessuna delle precedenti

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

Q13. Osservando l'output Q.10, ricavare il coefficiente di correlazione:

a) 0.769
b) 0.592
c) 184.7

Q14. Il valore espresso da "R aggiustato" (si veda Adjusted R-squared nell'output Q.10) è utile per:

a) Definire la bontà dell'intercetta
b) Definire la bontà dei coefficienti
c) Confrontare la bontà tra diversi modelli

Q15. Osservando l'output riportato in Q.10 possiamo fare inferenza statistica riferendoci ad un livello di...

a) significatività del 99%
b) confidenza del 99%
c) confidenza del 1-0.99

SIMULAZIONE ESAME STATISTICA MULTIVARIATA

APPLICAZIONE IN R

Si dispone di un database creato dai dati ufficiali della FAO, relativo ai consumi alimentari procapite suddivisi per tipologia di cibo. Sono stati considerati 126 paesi con una popolazione superiore ai 3 milioni di abitanti.

Le variabili includono il nome del paese, l'ID paese, la popolazione e diverse tipologie di alimenti.

L'Organizzazione delle Nazioni Unite per l'Alimentazione e l'Agricoltura sta pensando di promuovere un programma internazionale per la riduzione del consumo di Bevande Alcoliche. A tal fine vi chiede di indagare la relazione di causalità tra il consumo di bevande alcoliche (Alcoholic Beverages) e il consumo di carne (Meat), pesce (Fish) e zucchero (Sugar).

Effettuare l'analisi richiesta utilizzando l'ambiente R, seguendo l'impostazione analitica vista a lezione.

Commentare, dove si ritiene necessario, la scelta dei comandi e le fasi dell'analisi.

Spiegare le caratteristiche dei risultati ottenuti.

Interpretare il risultato alla luce della domanda di ricerca.

Salvare tutto in uno script (formato txt) sul desktop nel formato "Esame SM_Nome Cognome.txt"