Modalità d'esame e osservazioni

Denominazione: tesina con discussione e valutazione, da consegnarsi via mail al docente entro la data dell'appello scelto.
Lo studente, previa iscrizione all'appello (chi non risultasse iscritto non sosterrà la prova), svolgerà una relazione, sotto forma di report scientifico, dopo aver scelto un data set, sul quale applicherà una o più tecniche fra quelle riferite al programma (regressione, componenti principali, analisi dei cluster, analisi discriminante, analisi delle corrispondenze), utilizzando il software SPSS, che può essere scaricato gratuitamente mediante le credenziali in avviso in piattaforma e-learning nello spazio dedicato al modulo. Il report, di 15-50 pagine più eventuali centinaia in apposite appendici, va scritto presupponendo la conoscenza della teoria statistica da parte del lettore, secondo le seguenti linee guida.

- Deve essere presupposta la conoscenza da parte del lettore della teoria statistica matematica. Qualora occorresse riferirsi a formule, ciò è possibile attraverso note a piè pagina, oppure richiamando le formule numerate in apposita appendice

- Dopo una breve presentazione del data-set, lo studente inserirà nel corpo del testo l'output ritenuto rilevante e lo commenterà in modo scientifico, al fine di estrarre l'informazione rilevante e presentare le conclusioni scientifiche che si possono ottenere

Lo studente sappia che:

- si è optato per tale modalità **per precisa e personale scelta di efficacia didattica**, ampiamente verificata su altri insegnamenti, al

fine di garantire la obiettiva valutazione della preparazione del candidato, che avrà un tempo ampiamente necessario per redigere il lavoro in autonomia, tenuto conto che:

- Il software SPSS è un super-foglio xcell, pertanto con interfaccia molto amichevole

- In altri corsi da me tenuti, la presente modalità è molto apprezzata dagli studenti, che, in 20 anni di adozione, hanno potuto lavorare in autonomia senza che si riscontrasse alcun problema

- Lo studente potrebbe ragionevolmente preparare la prova in 5-10 pomeriggi; tenendo conto che 1 CFU corrisponde idealmente a 25 ore di lavoro a casa, pertanto 25*5 CFU=125 ore complessive di preparazione: ad es. 5 pomeriggi 5*5=25 ore per preparazione report più 100 ore di studio, 10 pomeriggi 10*5=50 ore per preparazione report più 75 ore di studio.

Si chiede cortesemente agli studenti **di evitare di scrivere al docente**: *tutte le informazioni dettagliate, aggiornate quotidianamente, riguardanti ogni aspetto della didattica sono presenti nello spazio dedicato all'insegnamento in piattaforma e-learning*.

Linee guida per prova d'esame di metodi statistici per il management, per argomenti

E' consigliabile che lo studente scelga un data set di interesse o, quantomeno, che riguardi un problema che desti interesse e curiosità nello studente medesimo, in modo che egli sia motivato

ad estrarre in modo scientifico informazioni e conclusioni rilevanti dal data set selezionato.

Regressione

Si tratta di prevedere una variabile quantitativa in funzione lineare di altre (*k*-1), anzi, problema di "model selection", di sceglierne s≤(k-1), le quali spieghino al meglio la variabile dipendente.
Questi i passi principali:

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc

- Selezione delle variabili esplicative, due approcci (lo studente scelga uno dei due):

a) Top down

1. Si considera il modello con tutte le k-1 variabili

2. Si escludono quelle il cui coefficiente di regressione risulta non significativamente diverso da zero (test *t*)

3. Con altre considerazioni, che dipendono anche dalla sensibilità del ricercatore, si perviene ad un modello con un numero s<k-1 variabili esplicative, che si ritiene candidato a prevedere la variabile dipendente

4. Si effettua il test *F* che verifica l'ipotesi nulla che valga il modello con *s* variabili, contro l'alternativa che valga il modello con tutte le k-1 variabili. Si accetta il modello associato all'ipotesi ritenuta vera.

b) Bottom up

1. Si considera il modello che prevede la variabile dipendente in funzione lineare di quella che ha coefficiente di correlazione più elevato con essa

2. E via via si aggiungono ad una ad una le variabili con più alta correlazione (sarebbe meglio considerare le correlazioni parziali, meno bene utilizzare le correlazioni grezze) sino a pervenire ad un modello tale che aggiunta di altre variabili parrebbe non migliorare la previsione. Oppure, si aggiunge una variabile fra le restanti escluse, includendo quella che opera la maggior riduzione della varianza resisidua o, che è lo stesso, il maggior incremento di varianza spiegata.
3. Si effettua il test F che verifica l'ipotesi nulla che valga il modello di cui al punto 2., contro l'alternativa che valga il modello con tutte le k-1 variabili. Si accetta il modello associato all'ipotesi ritenuta vera.

- Si deve verificare che le ipotesi del modello sui residui siano soddisfatte, in particolare:

1. Normalità: si costruisce il grafico delle frequenze relative specifiche dei residui, e si vede se imita l'andamento di una normale standard N(0,1), quasi simmetrico unimodale campanulare. Esistono anche dei test di verifica dell'ipotesi, Kolmogorov-Smirnov e altri (si veda SPSS)

2. Incorrelazione: test di Durbin-Watson e altri test

3. Omoschedasticità: si vede se in intervalli di valori differenti i residui presentino variabilità differente, se ciò non avviene si può concludere per varianza costante, in linea con le ipotesi del modello.

- Previsione

- Conclusioni scientifiche finali

- Bibliografia, sitografia, appendici.

Warning: se la numerosità del campione è almeno 100, le verifiche d'ipotesi del modello sui residui possono essere omesse, ma comunque si controlli la più o meno sussistenza delle ipotesi.

Cluster analysis

Si consideri una matrice *X* di *n* righe e *p* colonne, composta da *p* variabili quantitative rilevate su *n* unità statistiche.

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc

- Eventuale riduzione delle variabili:

a) Se *p* ed *n* sono ragionevolmente piccoli, si procede con l'analisi dei cluster

b) Se *p* e/o *n* sono grandi, al fine di far girare gli algoritmi di cluster in tempi ragionevoli sarà necessario ridurre le variabili attraverso l'analisi delle componenti principali, scegliendo le componenti principali ottenute dalla matrice di correlazione, associate ad autovalori maggiori di uno, valutando altresì la percentuale cumulata di varianza spiegata. L'analisi dei cluster sarà effettuata sulle nuove variabili, le componenti principali (le quali vanno interpretate nel loro legame con le variabili iniziali)

- Determinazione del numero ottimale dei cluster: si applica un metodo di clustering gerarchico, pervenendo ad un dendrogramma, che verrà tagliato in corrispondenza della soluzione "a gomito" (si vedano tecniche opportune), il numero di

cluster potrebbe essere anche aumentato o diminuito di 1, e fare confronti di performance e d'interpretazione.

- Identificazione dei cluster e loro interpretazione: si considerano le sintesi delle distribuzioni delle variabili sulle unità statistiche di tutta la popolazione e i loro principali quantili. Sintesi: min, max, media, scarto quadratico, varianza, asimmetria, curtosi. Quantili principali di ordine 0,05 0,10 0,25 0,5 (mediana) 0,75 0,90 0,95. Queste sintesi e quantili di popolazione vanno confrontati con i corrispondenti delle distribuzioni delle variabili per le unità statistiche di ogni cluster, al fine di avere una descrizione delle caratteristiche delle unità statistiche di cui è composto ogni cluster rispetto alle variabili.

[WARNING: se la clusterizzazione è avvenuta sulle componenti principali, l'interpretazione dei cluster avverrà sulle sintesi delle distribuzioni delle variabili di partenza (più informative) e NON sulle sintesi sulle componenti principali (meno informative)

- Applicazione di una tecnica di clustering non gerarchico:

1. si considera un numero di cluster determinato attraverso il dendrogramma ottenuto precedentemente dall'applicazione di una tecnica di clustering gerarchico.
2. Interpretazione dei cluster con le sintesi

- Confronto fra i cluster formati con il metodo gerarchico con quelli formati col metodo non gerarchico

- Conclusioni scientifiche finali

- Bibliografia, sitografia, appendici.

Analisi discriminante

k gruppi, l'*i*-esimo di numerosità n_i, su ognuna delle *n* unità statistiche si rilevano p variabili quantitative.
- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, variabili quasi collineari, etc

- Diagonalizzazione della matrice $S^{-1} \cdot S_{(b)}$, determinazione degli autovalori (componenti varianza) e autovettori (funzioni discriminanti).

- Scelta del numero di funzioni discriminati, in base alla percentuale di varianza cumulata

- Interpretazione logico-statistica delle funzioni discriminanti, in base al loro legame con le *p* variabili

- Tasso di corretta classificazione

- Se vi sono delle osservazioni aggiuntive, imputazione di ogni unità statistica a un gruppo con confronto con i centroidi di gruppo

- Conclusioni scientifiche finali

- Bibliografia, sitografia, appendici.

Analisi delle corrispondenze

- Descrizione e presentazione del dataset, e sua preparazione – eliminazione dati mancanti, etc

- Due casi

1. Bivariata

Si trovano le soluzioni per i fattori e gli assi fattoriali, e se ne sceglie il numero in base alla percentuale di varianza cumulata. Si svolge l'analisi quantitativa dei contributi puntuali e globali di fattori alle modalità e viceversa, al fine di interpretare in senso logico-statistico i fattori in relazioni alle variabili. Se s'intendono sufficienti due fattori, si può procedere mediante l'interpretazione di punti riga e colonna e del b-plot, mediante commenti appropriati.

2. Multivariata.

Si considera un approccio, matrice disgiuntiva, soluzione di Burt, e si procede all'interpretazione secondo quando prescritto dalla teoria (si vedano teoria, applicazioni, e manuale SPSS)

- Qualora esistano osservazioni supplementari, si può pensare di imputare ogni singola unità statistica ad un fattore, similmente a come si procede nell'analisi discriminate

- Conclusioni scientifiche finali

- Bibliografia, sitografia, appendici.

Si consiglia di esaminare con attenzione **anche** articoli, report e applicazioni proposti nelle sezioni presenti nello spazio dedicato al modulo di **analisi quantitative di mercato** del mio insegnamento "statistica per il marketing", ECOMARK 2019-20, in modo da trarre ispirazione ed impratichirsi, al fine di raggiungere il migliore risultato possibile nella stesura della relazione.