

|  |
| --- |
| Report statistico Esame di Statistica MatematicaDocente: Prof. Marco GadaletaA.A 2019/2020 |
| Corso di laurea diInformatica e Comunicazione digitaleSede di Taranto |
| **24 febbraio 2020** Matteo Luceri |

|  |
| --- |
| Traccia 1Introduzione La traccia richiede la verifica di alcuni valori ricavati da una ricerca che studia l’efficacia dell’uso del touch screen.  Il campione di riferimento sarà così formato da 15 persone, con capacità fra loro omogenee.  I dati raccolti, rappresentanti le velocità d’uso, saranno quindi composti da una prima e da una seconda rilevazione, sempre di 15 valori.  E’ richiesta la verifica e il riscontro, con un test di livello α = 0.05, di un cambiamento di velocità media di utilizzo dei dispositivi fra le due rilevazioni. Svolgimento Avremo quindi due campioni, **X** e **Y**, rappresentanti le due rilevazioni di **15** valori ciascuno.  Immagine che contiene screenshot  Descrizione generata automaticamente  Da una prima analisi è evidente la correlazione fra i due campioni, che **non** risulteranno quindi indipendenti. Si decide di effettuare un test di confronto delle medie.  Si effettua un’analisi del campione **X.** Si ottiene:  Come possiamo notare dai primi dati, la **varianza** di X è grande, quindi i valori di discosteranno notevolmente dalla media.  Immagine che contiene screenshot  Descrizione generata automaticamenteIl **boxplot** di X risulterà il seguente:  Da una prima analisi del grafico è evidente la presenza di un valore *outlier,* causato da un possibile errore di misurazione.  Verrà quindi rimosso il valore outlier in X e il corrispettivo in Y, e riproposti i nuovi dati e il nuovo boxplot per la variabile X:    Immagine che contiene screenshot  Descrizione generata automaticamente  La varianza di X si è dimezzata rimuovendo il valore outlier, ma rimane comunque un valore moderatamente grande. Analizzando la simmetria, in particolare il valore della **mediana**, si può affermare che il campione X presenta una distribuzione **leggermente** **asimmetrica** a sinistra, in quanto il primo quartile è più distante rispetto al terzo.  Si presenta adesso **l’istogramma** della variabile X: |
| Osservando l’istogramma della variabile X, e la **curva di densità** tracciata su di esso, è possibile notare il carattere **unimodale** della densità, oltre alla presenza di un solo massimo.  Si effettua adesso un’analisi del campione **Y**.Si ottiene:  Come per X, la **varianza** di Y è grande, quindi i valori di discosteranno notevolmente dalla media.  Il **boxplot** di Y risulterà il seguente:    Si può notare (anche grazie l’assenza di valori *outliers*) la lunga distribuzione dei dati, evidenziata dalla distanza da e ai rispettivi primo e terzo quartile. Anche in questo caso si evidenzia, grazie alla **mediana**, una più netta asimmetria a sinistra in Y, evidenziata dala vicinanza della mediana al terzo quartile.  Si presenta adesso **l’istogramma** della variabile Y:  Osservando l’istogramma della variabile Y, e la **curva di densità** tracciata su di esso, è possibile notare il carattere **unimodale** della densità. L’istogramma infatti presenta un solo massimo, risultando sbilanciato a sinistra, con la coda sinistra leggermente più lunga.  Si effettua adesso il test del *chi quadro* per stabilire se le variabili **X** e **Y** seguono una distribuzione normale. Si ottiene **per X**:  Confrontando X con un campione normale generato random risulta che X **non** segue legge Normale, in quanto la statistica ottenuta è pari a “27,18” con un *p-value* basso (0.01173). Infatti confrontando la statistica ottenuta con i valori della tavola numerica del chi quadro, rifiuteremo l’ipotesi nulla (“X segue distribuzione normale”) qualora la statistica ottenuta (con 13 gradi di libertà) sia maggiore di 22.36203. Possiamo dunque affermare che X **non** segue legge Normale. Per avvalorare l’ipotesi si proprone un ulteriore calcolo del test. Infatti per campioni estremamente piccoli il calcolo della *p-value* può risultare scorretto. Una soluzione può costituirsi nel raggruppare le classi per aumentare così il numero di casi possibili. Non riuscendo in questo intento si è scelto di adottare la c.d. correzione di Yates, che propone di apporre una leggera correzione (0.5) per ovviare al problema. Ma anche questa misurazione darà valore di *p-value* molto piccoli e di statistiche molto grandi, consolidando l’ipotesi che X non segue la legge Normale.  e **per Y**:  Anche in questo caso possiamo affermare, che Y non segue legge Normale in quanto la statistica ottenuta è pari a “71.34” e la tavola numerica ci porta ad accettare l’ipotesi alternativa (“Y **non** segue legge normale”) poiché la statistica ottenuta (con 13 gradi di libertà) è maggiore di 22.36203. Anche in questo caso il *p-value* ottenuto è decisamente basso (4.53940033037... \* (10 ^ (-10)).  Analagomente per X si sono effettuati calcoli di correzione consolidanto anche in questo caso che Y **non** segue la legge Normale.  Nel caso in cui X e Y avessero seguito la legge Normale, avremmo eseguito il test di Fisher, supponendo che le variabili abbiano la stessa varianza.  Si è deciso di eseguire e proporre ugulmente il test di Fisher per X e Y.  Si ottiene così per X:  e per Y:  I risultati del test di Fisher affermano che essi, con un margine di errore del 5%, hanno la stessa varianza. Per sicurezza si è svolto il test invetendo il numeratore con il dominatore e anche in questo caso il test afferma di non rifiutare l’ipotesi nulla. Ma tali ipotesi non hanno valenza a fini dello studio.  Si procede infine, effetuando il *test di Student* per verificare se le rilevazioni in questione hanno risentito di un cambiamento di velocità media. Il test T effettuato con 14(n-1) gradi di libertà ci suggerisce rifutare la nostra ipotesi (“Le due rilevazioni hanno la stessa media”) poiché paragonando la nostra statistica con il quantile ,risulterà maggiore, per cui al livello il campione è in regione critica, e quindi possiamo dire che nel tempo **la velocità media di uso del dispositivo è aumentata**. Traccia 2Introduzione La traccia richiede di effettuare, date le qualità del vino Bordeaux (Y) e la somma delle temperature gionaliere del mese di aprile(X), delle previsioni sulla bontà del vino (Y) in previsione di valori medi di temperatura (X) pari a 2961 e 3363.  Si richiede di effettuare le previsioni, sui dati raccolti, tramite modello di regressione lineare, determinando il migliore. Svolgimento Si sintetizzano i dati raccolti attraverso la seguente tabella:  Il boxplot relativo alla variabile X è il seguente:  Immagine che contiene screenshot  Descrizione generata automaticamente  Da una prima analisi del grafico è evidente il valore della **mediana**.  Si può affermare che il campione X presenta una distribuzione **asimmetrica** a destra, in quanto il primo quantile è più vicino rispetto al terzo.  Mentre il boxplot relativo a Y è il seguente: |
| Analizzando il grafico è evidente il valore della mediana.  Si può affermare che il campione Y presenta una distribuzione asimmetrica a sinistra, in quanto il primo quantile è più lontano rispetto al terzo.  Si presenta adesso l’istogramma delle variabili X e Y:  L’ asimmetria è ancor più visibile osservando i due istogrammi. Nonostante ciò è evidenziabile un comportamento bimodale di entrambe le distribuzioni, con picchi più alti ai lati e più bassi al centro.  In particolare, in Y l’asimmetria è fortemente smorzata.  Rappresentando sul piano cartesiano i punti per si ha il seguente diagramma di dispersione:  Immagine che contiene testo  Descrizione generata automaticamente |

Dal diagramma raffigurato pare ci sia una effettiva relazione tra X e Y in quanto all’aumentare del valore di X aumenta anche Y.

Costruiamo quindi un modello di regressione lineare per studiare in maniera più approfondita la relazione che intercorre tra X e Y.

Si è scelto di aggiungere una costante +3 a ogni valore di Y in modo tale da rendere migliore il calcolo del modello. Infatti il valore dell’intercetta non andrà ad influire sui risultati.

Immagine che contiene screenshot

Descrizione generata automaticamente

Dal primo modello di regressione lineare possiamo notare che siamo riusciti a rappresentare circa il 75,4% dei dati (R-squared), il che significa che il nostro modello è abbastanza rappresentativo della realtà.

Da R-squared possiamo ricavarci il valore assoluto del coefficiente di correlazione lineare,e ricavando un valore pari a 0.868383 possiamo dire quindi che tra le due variabili c’è una correlazione forte.

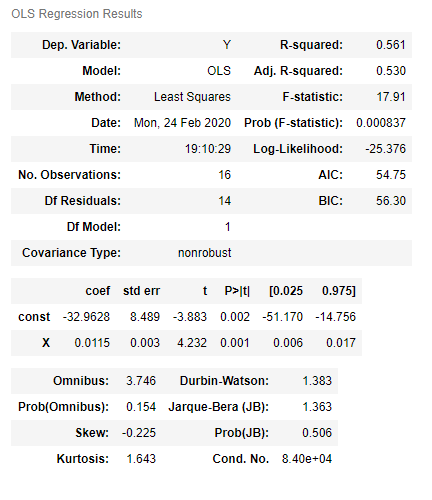
La pendenza è pari a 0.001 ciò vuol dire che approssimativamente la nostra variabile Y si comporta secondo la retta

La significatività del test effettuato sulla stima del parametro è molto alta come riportato dal valore di P >|t|.

L’intervallo di confidenza del coefficiente è ristretto, cioè la stima effettuata è abbastanza precisa.

Possiamo inoltre analizzare l’indice di curtosi che essendo maggiore di 0 la curva si definisce leptocurtica, cioè più “appuntita” di una normale.

Proviamo ora a predire un nuovo modello di regressione lineare aggiungendo un'altra intercetta.



In questo caso la variabile Y si comporta secondo la retta Y = 0.0115\*X -32.96 .

Come possiamo vedere, il modello peggiora. In questo caso siamo riusciti a rappresentare solo il 56.1% dei dati rispetto all’ 75.4% del modello precedente. Ma questo non è il solo indice a peggiorare.

Difatti, anche la significatività del test peggiora nella stima del parametro dell’intercetta,con un intervallo di confidenza eccessivamente grande quindi molto impreciso.

Alla luce di quanto messo in risalto dai due modelli proposti, si decide di definire il primo modello come quello più rappresentativo. Per tanto la variabile Y si comporta approssimativamente secondo la retta:

Pertanto per effettuare delle previsioni sulla bontà avremo:

e