Report statistico

Esame di Statistica Matematica
Docente: Prof. Marco Gadaleta
A.A 2019/2020

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO
DI INFORMATICA

Corso di laurea di
Informatica e Comunicazione digitale
Sede di Taranto

24 febbraio 2020

Matteo Luceri

Traccia 1
Introduzione

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 $\alpha = 0.05$, 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.

```
X=np.array([ 67, 64, 69, 88, 72, 80, 85, 116, 77, 78, 81, 66, 91, 68, 73])
```

```
Y=np.array([ 57, 53, 71, 61, 73, 50, 53, 80, 63, 41, 78, 68, 86, 70, 74 ])
```

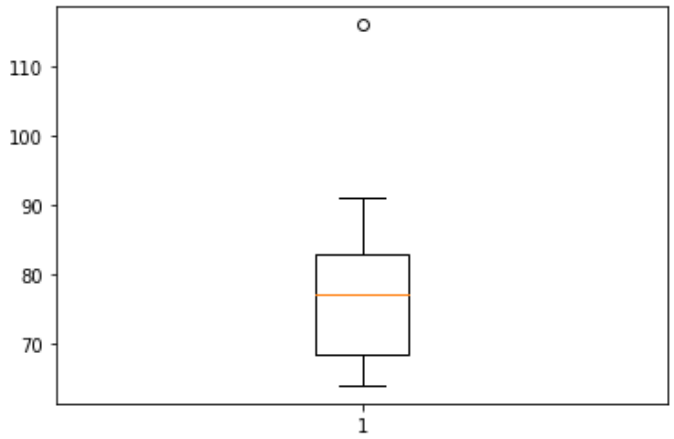
Da una prima analisi è evidente la correlazione fra i due campioni, che **non** risulteranno quindi indipendenti.
Si decide di effettuare un test di confronto delle medie.

Si effettua un'analisi del campione **X.**
Si ottiene:

```
Massimo di X: 116
Minimo di X: 64
Media di X: 78.33333333333333
Varianza di X: 165.15555555555557
```

Come possiamo notare dai primi dati, la **varianza** di X è grande, quindi i valori di discosteranno notevolmente dalla media.

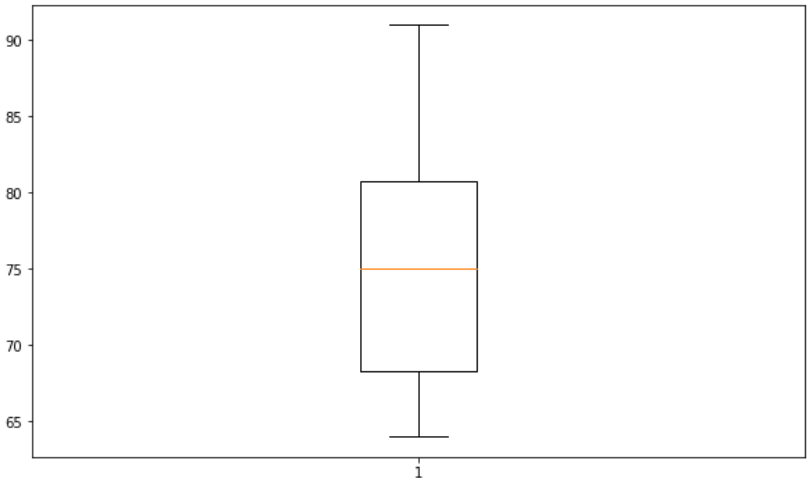
Il **boxplot** di X risulterà il seguente:



Da una prima analisi del grafico è evidente la presenza di un valore *outlier,* causato da un possibile errore di misurazione.

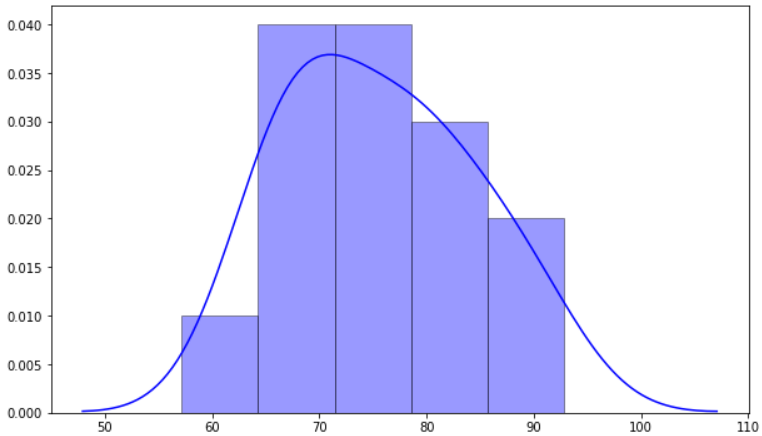
Verrà quindi rimosso il valore outlier in X e il corrispettivo in Y, e riproposti i nuovi dati e il nuovo boxplot per la variabile X:

```
Massimo di X: 91
Minimo di X: 64
Media di X: 75.64285714285714
Varianza di X: 68.37244897959184
```



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:



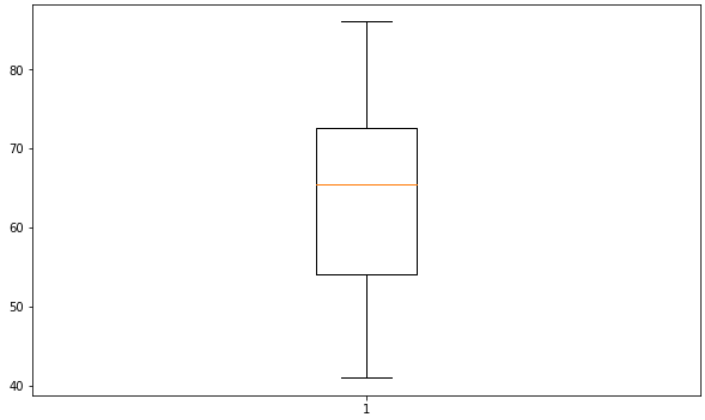
Osservando l'istogramma della variabile X, e la **curva di densità** tracciata su di esso, è possibile notare il carattere **unimodale** della densità, oltre alla presenza di un solo massimo.

Si effettua adesso un'analisi del campione **Y**.
Si ottiene:

```
Massimo di Y: 86
Minimo di Y: 41
Media di Y: 64.14285714285714
Varianza di Y: 141.9795918367347
```

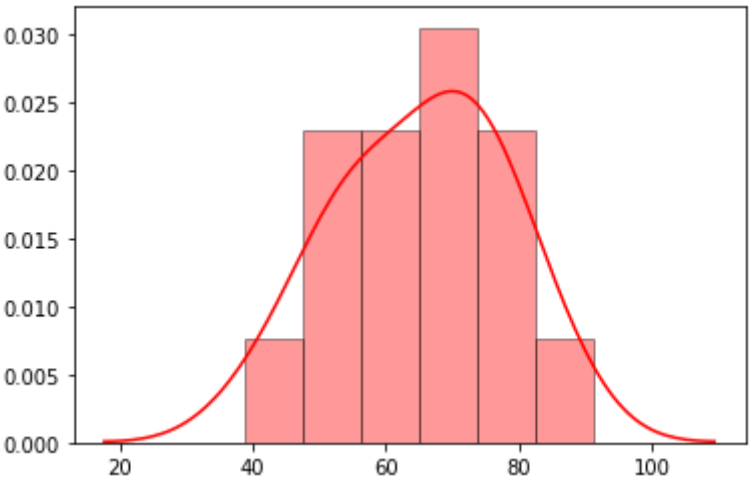
Come per X, la **varianza** di Y è grande, quindi i valori di discosteranno notevolmente dalla media.

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 y_{min} e y_{max} 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**:

```
Power_divergenceResult(statistic=27.18639957303996, pvalue=0.011729718373811345)
```

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**:

```
Power_divergenceResult(statistic=71.34248186126645, pvalue=4.539400330371238e-10)
```

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:

```
Cant Reject H0 0.1005252260265182
```

e per Y:

```
Cant Reject H0 0.8994747739734817
```

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 H_0 ("Le due rilevazioni hanno la stessa media") poiché paragonando la nostra statistica $T_0 = 2.8588814819356774$ con il quantile $t_{0.95}(13) = 1.77093$,risulterà maggiore, per cui al livello $\alpha = 0.05$ il campione è in regione critica, e quindi possiamo dire che nel tempo **la velocità media di uso del dispositivo è aumentata**.

Traccia 2

Introduzione

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.

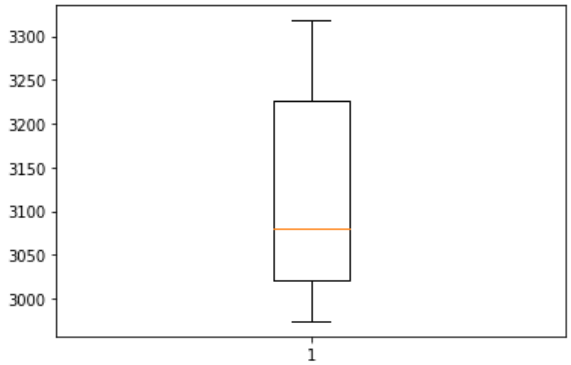
Si richiede di effettuare le previsioni, sui dati raccolti, tramite modello di regressione lineare, determinando il migliore.

Svolgimento

Si sintetizzano i dati raccolti attraverso la seguente tabella:

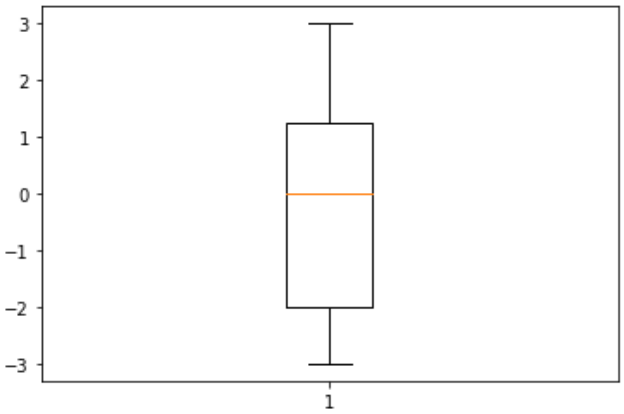
	Y	X
count	16.000000	16.000000
mean	-0.062500	3123.125000
std	1.842779	120.101554
min	-3.000000	2974.000000
25%	-2.000000	3021.250000
50%	0.000000	3080.000000
75%	1.250000	3227.000000
max	3.000000	3318.000000

Il boxplot relativo alla variabile X è il seguente:



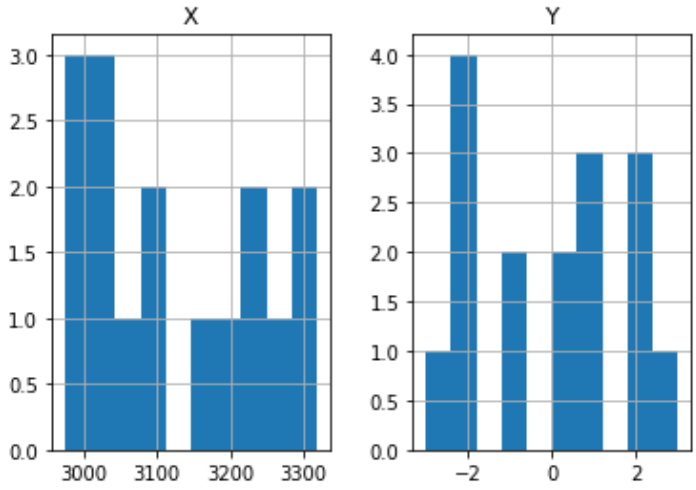
Da una prima analisi del grafico è evidente il valore della **mediana**.
Si può affermare che il campione X presenta una distribuzione **asimmetrica** a destra, in quanto il primo quantile è più vicino rispetto al terzo.

Mentre il boxplot relativo a Y è il seguente:



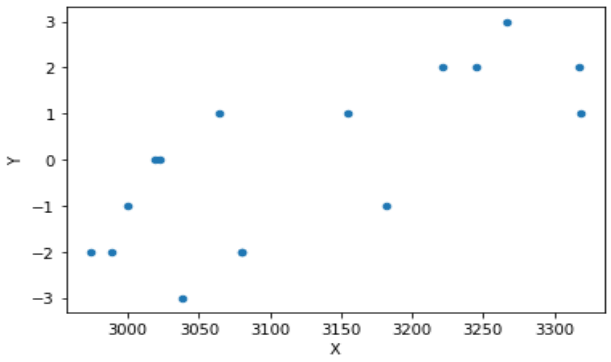
Analizzando il grafico è evidente il valore della **mediana**.
Si può affermare che il campione Y presenta una distribuzione **asimmetrica a sinistra,** in quanto il primo quantile è più lontano rispetto al terzo.

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 (x_i, y_i) per $i = 1,2,\dots,N$ si ha il seguente diagramma di dispersione:



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.

OLS Regression Results

Dep. Variable:	Y	**R-squared (uncentered):**	0.754
Model:	OLS	**Adj. R-squared (uncentered):**	0.738
Method:	Least Squares	**F-statistic:**	46.08
Date:	Mon, 24 Feb 2020	**Prob (F-statistic):**	6.12e-06
Time:	16:00:57	**Log-Likelihood:**	-31.224
No. Observations:	16	**AIC:**	64.45
Df Residuals:	15	**BIC:**	65.22
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
X	0.0010	0.000	6.788	0.000	0.001	0.001

Omnibus:	2.545	**Durbin-Watson:**	2.095
Prob(Omnibus):	0.280	**Jarque-Bera (JB):**	1.068
Skew:	-0.004	**Prob(JB):**	0.586
Kurtosis:	1.734	**Cond. No.**	1.00

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

$$Y = 0.0010 * X - 3.$$

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.

OLS Regression Results

Dep. Variable:	Y	**R-squared:**	0.561
Model:	OLS	**Adj. R-squared:**	0.530
Method:	Least Squares	**F-statistic:**	17.91
Date:	Mon, 24 Feb 2020	**Prob (F-statistic):**	0.000837
Time:	19:10:29	**Log-Likelihood:**	-25.376
No. Observations:	16	**AIC:**	54.75
Df Residuals:	14	**BIC:**	56.30
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-32.9628	8.489	-3.883	0.002	-51.170	-14.756
X	0.0115	0.003	4.232	0.001	0.006	0.017

Omnibus:	3.746	**Durbin-Watson:**	1.383
Prob(Omnibus):	0.154	**Jarque-Bera (JB):**	1.363
Skew:	-0.225	**Prob(JB):**	0.506
Kurtosis:	1.643	**Cond. No.**	8.40e+04

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:

$$Y = 0.001 * X - 3$$

Pertanto per effettuare delle previsioni sulla bontà avremo:

$$Y = 0.001 * 2961 - 3 = -0.039$$

e

$$Y = 0.001 * 3363 - 3 = 0.363$$