Gennaio 2020

In [31]:

```
import pandas as pd
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt
import math
```

Esercizio 0

Dato un evento C in uno spazio campionario Ω, sia P(C) la probabilità che C si verifichi. Data una partizione $A_1, A_2, \ldots, A_n$ di Ω, supponiamo che siano note, per ogni i = 1, ... , n:

a) Le probabilità marginali dei singoli eventi A_i e

b) le probabilità condizionate che accada C dato che è accaduto A_i

Esprimete, in funzione delle opportune probabilità marginali e condizionate, la probabilità che accada l'evento C.

$$P(C) = \sum_{i=1}^{n} P(C|A_i)P(A_i)$$

Esercizio 1

Sia X una variabile casuale di valore atteso μ e varianza σ², indichiamo con X* la corrispondente variabile standardizzata.

1 Esprimete X* in funzione di X, μ e σ.

$$X^* = \frac{X - \mu}{\sigma}$$

2 Controllate che il valore atteso di X* è uguale a 0.

$$E[X^*] = E[\frac{X - \mu}{\sigma}]$$

Per proprietà di linearità del valore atteso

$$E[X^*] = \frac{E[X] - \mu}{\sigma}$$

$$E[X^*] = \frac{\mu - \mu}{\sigma}$$

$$E[X^*] = 0$$

3 Controllate che la varianza di X* è uguale a 1.

$$Var(X^*) = \frac{1}{\sigma^2} Var(X - \mu) = \frac{Var(X)}{\sigma^2} = 1$$

4 Supponiamo solo in questo punto che X segua una variabile uniforme discreta con punti di massa nell'insieme {1,2}.

- **4.1** Tracciate il grafico di massa di probabilità di X.

- **4.2** Quali valori può assumere X*? Poichè ho $D_X = \{0, 2\}$ quindi n = 2

$$E(X^*) = \frac{n+1}{2} = \frac{3}{2}$$

$$Var(X^*) = \frac{n^2+1}{12} = \frac{1}{4}$$

$$\sigma = \frac{1}{2}$$

Per x = 0

$$x^* = \frac{X - \mu}{\sigma} = \frac{1 - \frac{3}{2}}{\frac{1}{2}} = -1$$

Per x = 2

$$x^* = 1$$

- **4.3** Tracciate il grafico di massa di probabilità di X*.

Esercizio 2

Sia X una variabile casuale normale di parametri μ e σ². Sia k > 0 un valore fissato.

1 Esprimete, in funzione della funzione di ripartizione Fx, la probabilità che X assuma volori compresi tra μ - k e μ + k.

$$P(\mu - k <= X <= \mu + k)$$

Applico la definizione di Funzione di Ripartizione e ottengo che

$$P(\mu - k \leq X \leq \mu + k) = F_X(k + \mu) - F_X(\mu - k)$$

2 Controllate che P(μ - k <= X <= μ + k) = P(|X*|<= k/σ).

$$X^* = \frac{X - \mu}{\sigma}$$

$$P(-k \leq X - \mu \leq k) = P(\frac{-k}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{k}{\sigma}) = P(|X^*| \leq \frac{k}{\sigma})$$

3 Indichiamo con ϕ la funzione di ripartizione di una variabile normale standard. Controllate che P(|X*| <= k/σ) = 2ϕ(k/σ) - 1

$$\phi(\frac{k}{\sigma}) - \phi(-\frac{k}{\sigma}) = 2\phi(\frac{k}{\sigma}) - 1$$

Per la simmetria intorno all'origine della normale (ho standardizzato ma rimane normale)

$$\phi(\frac{k}{\sigma}) - (1 - \phi(\frac{k}{\sigma})) = 2\phi(\frac{k}{\sigma}) - 1$$

$$2\phi(\frac{k}{\sigma}) - 1 = 2\phi(\frac{k}{\sigma}) - 1$$

Esercizio 3

Sia $\bar{X}n$ la media campionaria di un campione casuale X1, ... , Xn estratto da una popolazione normale X di valore atteso μ e di cui è disponibile una stima della varianza pari a $\sigma^2 = 3$.

1 Fissati i valori epsilon = 0.25 e alpha = 0.9, determinate una condizione sufficiente per n affinchè sia maggiore di alpha la probabilità $P(|\bar{X}_n - \mu| <= \text{epsilon})$.

NON SERVE APPLICARE IL TEOREMA DEL LIMITE CENTRALE O CHEBYSHEV PERCHE' HO GIA' UNA NORMALE STANDARD

$$P(|\overline{X_n} - \mu| \leq \epsilon) \geq \alpha$$

Standardizzo:

$$P(|\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}| \leq \frac{\epsilon\sqrt{n}}{\sigma}) \geq \alpha$$

Per la riproducibilità rimane normale

$$P(|Z| \leq \frac{\epsilon\sqrt{n}}{\sigma}) = 2\Phi(\frac{\epsilon\sqrt{n}}{\sigma}) - 1$$

$$\sqrt{n} \geq \Phi^{-1}(\frac{\alpha+1}{2}) \cdot \frac{\sigma}{\epsilon}$$

In [2]:

```
Z = st.norm()
sigma = math.sqrt(3)
alpha = 0.9
epsilon = 0.25
(Z.ppf((alpha+1)/2)*sigma/epsilon)**2
# n >= 130
```

Out[2]:

```
129.8660857965798
```

2 Proponete uno stimatore, chiamiamolo Tn di μ

$$T_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

3 Lo sitmatore che avete proposto al punto precedente è non distorto? Giustificate la risposta.

Lo stimatore Tn è la media campionaria ed essa è sempre uno stimatore non distorto del VALORE ATTESO della popolazione

Esercizio 4

Scaricare mtcars.txt Quessto file contiene, tra le altre, le seguenti informazioni riguardo al design e alle prestazion di diversi modelli di automobili:

- modello: identificatore univoco:
- consumo: espresso in km/l;
- cilindrata: cilindrata (espressa in cavalli vapore);
- peso: peso, espresso in tonnellate;
- test400metri: tempo (espresso in secondi) impiegato per percorrere 400 metri partendo da fermo;
- trasmissione: tipo di trasmissione 0 se si tratta di trasmissione automatica, 1 se si tratta di trasmissione manuale;
- marce: numero di marce, senza contare la retromarcia.

In questo file il carattere di tabulazione ("\t") separa le colonne ed i numeri reali sono stati registrati usando il carattere "," come separatore dei decimali.

In [3]:

```
df = pd.read_csv("mtcars.txt",sep="\t",decimal=",")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 7 columns):
modello         32 non-null object
consumo         32 non-null float64
cilindrata      32 non-null int64
peso            32 non-null float64
test400metri    32 non-null float64
trasmissione    32 non-null int64
marce           32 non-null int64
dtypes: float64(3), int64(3), object(1)
memory usage: 1.8+ KB
```

1 Quanti e quali sono i caratteri scalari presenti nel dataset?

3 caratteri: consumo, peso, test400metri

2 Quante osservazioni contiene il dataset?

In [4]:

```
len(df)
```

Out[4]:

```
32
```

3 Qual'è la moda del carattere peso?

In [5]:

```
df.peso.mode()
```

Out[5]:

```
0    3.44
dtype: float64
```

4 Qual'è il modello di auto più pesante?

In [6]:

```
df[df.peso == df.peso.max()].modello
```

Out[6]:

```
15    Lincoln Continental
Name: modello, dtype: object
```

5 Quali sono i possibili valori per il carattere marce?

In [7]:

```
df.marce.unique()
```

Out[7]:

```
array([4, 3, 5], dtype=int64)
```

6 Dei caratteri marce e trasmissione visualizzate:

- **6.1** La tabella delle frequenze congiunte assolute;

In [26]:

```
pd.crosstab(index=df.marce, columns=df.trasmissione,margins=True)
```

Out[26]:

trasmissione	0	1	All
marce			
3	15	0	15
4	4	8	12
5	0	5	5
All	19	13	32

- **6.2** La tabella delle frequenze congiunte relative.

In [27]:

```
pd.crosstab(index=df.marce, columns=df.trasmissione, normalize=True,margins=True)
```

Out[27]:

trasmissione	0	1	All
marce			
3	0.46875	0.00000	0.46875
4	0.12500	0.25000	0.37500
5	0.00000	0.15625	0.15625
All	0.59375	0.40625	1.00000

7 Utilizzate le informazioni contenute nelle tabelle prodotte al punto precedente per rispondere alle seguenti domande:

- **7.1** Quanti sono i modelli di auto a 5 marce con trasmissione automatica?

0%

- **7.2** Quanti sono i modelli di auto a 5 marce con trasmissione manuale?

5%

- **7.3** Qual'è la percentuale di modelli che hanno 5 marce e trasmissione automatica?

0

- **7.4** Qual'è la percentuale di modelli che hanno 5 marce e trasmissione manuale?

15%

- **7.5** Qual'è la percentuale di modelli che hanno 4 marce?

37%

- **7.6** Tra i modelli che hanno trasmissione manuale, quale percentuale ha 4 marce?

8/13 = 0.61

61%

Esercizio 5

Caratterizziamo ora i vari modelli di automobile rispetto al numero di marce, cioè dividiamo l'insieme delle nostre osservazioni in tre sottoinsiemi, in cui rispettivamente le auto hanno 3, 4 oppure 5 marce.

1 Compilate la Tabella 1 delle frequenze relative del carattere marce.

----------**modello a** --- **modello a** --- **modello a**

----------**3 marce** ------ **4 marce** ------ **5 marce** > > > **Frequenza Relativa**

In [10]:

```
df.marce.value_counts(normalize=True)
```

Out[10]:

```
3    0.46875
4    0.37500
5    0.15625
Name: marce, dtype: float64
```

2 Qual'è la probabilità che, sorteggiando un modello di auto dal dataset, quel modello sia a 4 marce?

0.375

3 In Tabella 2, per ogni categoria di auto (a 3, 4 oppure 5 marce) è mostrata la probabilità che l'auto abbia consumi alti (queste probabilità sono state stimate tramite un'analisi precedente). Fornite una stima della probabilità che, sorteggiando un modello di auto dal dataset, quel modello abbia consumi alti.

----------**modello a** --- **modello a** --- **modello a**

----------**3 marce** ------ **4 marce** ------ **5 marce**

probabilità di alti consumi 0.8 0.17 0.6

C = "alti consumi", M_i = "modelli di macchina"

$$P(C) = \sum_{i=1}^{3} P(C|M_i)P(M_i)$$

In [11]:

```
0.8*0.46875 + 0.17*0.37500 + 0.6*0.15625
```

Out[11]:

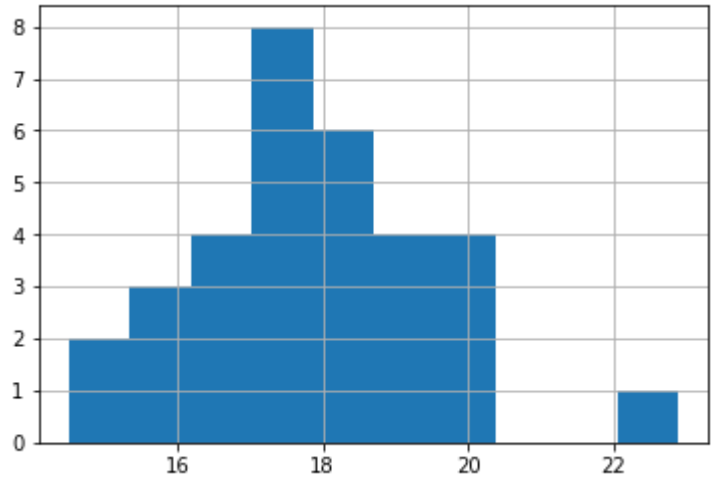
0.5325

Esercizio 6

1 Tracciate l'istogramma del tempo impiegato a percorrere 400 metri.

In [28]:

```
import matplotlib.pyplot as plt
df.test400metri.hist()
plt.show()
```



2 Per il tempo impiegato a percorrere 400 metri calcolate:

- **2.1** due indici di posizione centrale,

MEDIA, MEDIANA

In [13]:

```
print(df.test400metri.mean(), df.test400metri.median())
```

17.848750000000003 17.71

- **2.2** due indici di dispersione.

VARIANZA, RANGE INTERQUARTILE

In [29]:

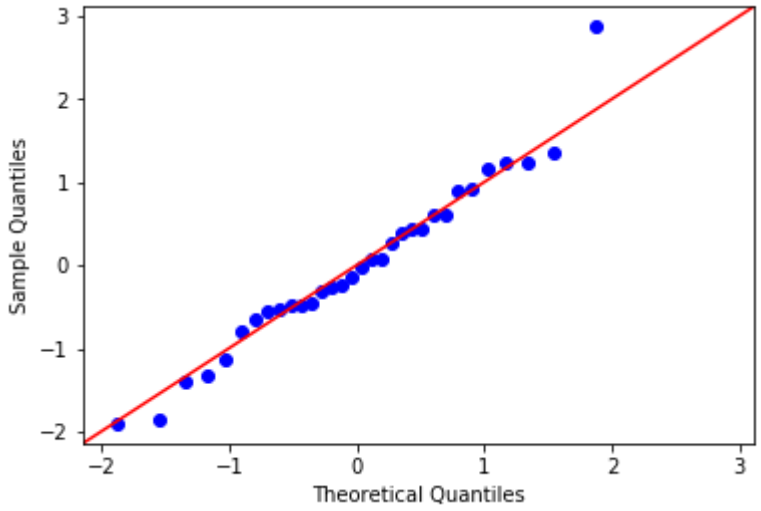
```
print(df.test400metri.var(), df.test400metri.quantile(0.75) - df.test400metri.quantile(
0.25))
```

3.193166129032258 2.0075000000000003

3 Tracciate un grafico opportuno per controllare se è plausibile l'ipotesi che il carattere test400metri segua una legge normale (In caso di errori, controllate che i dati coinvolti non contengano valori mancanti).

In [15]:

```
import statsmodels.api as sm
sm.qqplot(df.test400metri.dropna(), fit=True, line='45')
plt.show()
```



4 Ritenete plausibile l'ipotesi che il carattere test400metri segua una legge normale? Giustificate la risposta avvalendovi dei risultati trovati nei due punti precedenti.

Il grafico qq mostra che il carattere segue una legge normale appoggiandosi bene sulla bisettrice. Inoltre media e mediana simile fanno intendere la stessa ipotesi

Esercizio 7

1 Selezionate in una variabile chiamata bolidi i modelli di cilindrata superiore a 180. In questo esercizio ci occuperemo soltato dei bolidi.

In [16]:

```
bolidi = df[df.cilindrata > 180]
```

2 Quanti sono i casi selezionati?

In [17]:

```
len(bolidi)
```

Out[17]:

7

3 Calcolate il primo ed il terzo quantile dei consumi di modelli selezionati.

In [18]:

```
(bolidi.consumo.quantile(0.25), bolidi.consumo.quantile(0.75))
```

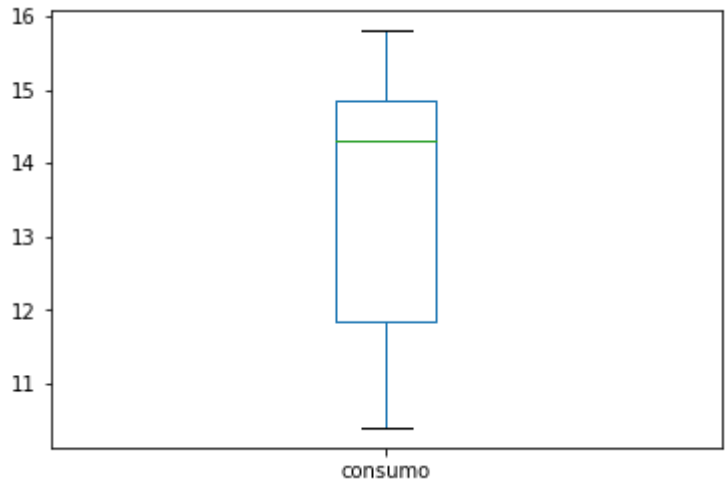
Out[18]:

```
(11.850000000000001, 14.85)
```

4 Tracciate il grafico che vi sembra più adatto a comunicare l'informazione che questi modelli hanno consumi alti. Commentate la vostra risposta.

In [32]:

```
bolidi.consumo.plot.box()
plt.show()
```

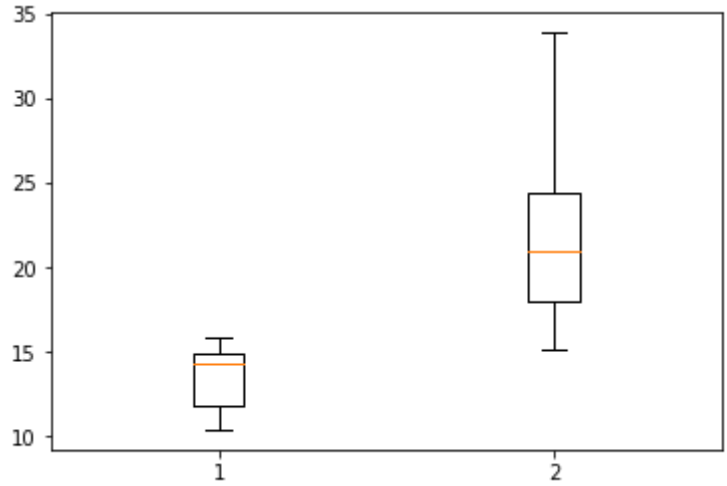


In [33]:

```
nobolidi  = df[df.cilindrata <= 180]
data = [bolidi.consumo, nobolidi.consumo]

# Multiple box plots on one Axes
fig, ax = plt.subplots()
ax.boxplot(data)

plt.show()
```



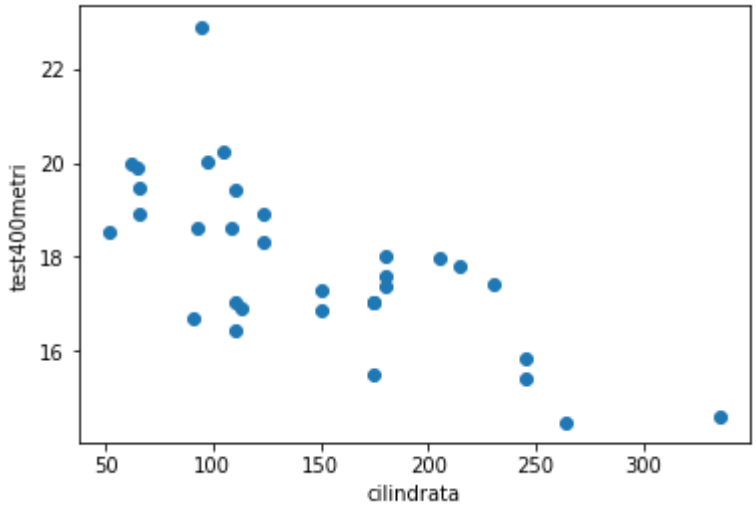
Esercizio 8

Studiamo ora la relazione tra cilindrata, prestazioni e consumi.

1 Tracciate un grafico opportuno per controllare se esiste una relazione tra i caratteri cilindrata e test400metri.

In [20]:

```
plt.scatter(df.cilindrata, df.test400metri)
plt.xlabel('cilindrata')
plt.ylabel('test400metri')
plt.show()
```



2 Come varia il tempo necessario per percorrere 400 metri all'aumentare della cilindrata?

All'aumentare della cilindrata il tempo sui 400 metri diminuisce

3 Utilizzate un'opportuno indice numerico per supportare la risposta che avete dato al punto precedente (In caso di errori, controllate valori mancanti).

In [21]:

```
df.cilindrata.corr(df.test400metri)
```

Out[21]:

```
-0.7082233888619532
```

Esercizio 9

Possiamo considerare i valori del carattere test400metri come la realizzazione campionaria di un campione casuale estratto dalla popolazione X = "tempo necessario per percorrere 400 metri.

1 Fornite una stima del valore atteso del tempo necessario per percorrere 400 metri.

In [22]:

```
df.test400metri.mean()
```

Out[22]:

```
17.848750000000003
```

2 Qual'è la taglia del campione che avete utilizzato per la stima?

IMPORTANTE: attenzione ai valori mancanti

In [23]:

```
len(df.test400metri.dropna())
```

Out[23]:

```
32
```

3 Lo stimatore che avete utilizzato è non distorto? Giustificate la risposta.

Lo stimatore è non distorto in quanto la media campionaria è sempre uno stimatore non distorto per il VALORE ATTESO

4 Fornite una stima della deviazione standard di X.

In [24]:

```
df.test400metri.std()
```

Out[24]:

```
1.7869432360968431
```

5 Nell'ipotesi che il tempo necessario per percorrere 400 metri segua una legge normale, calcolate la probabilità che, nela stima del valore atteso di X, si commetta un errore al più uguale a 0.25, in eccesso o in difetto.

$$P(|\bar{X}_n - \mu| <= 0.25)$$

$$P(-0.25 <= \bar{X}_n - \mu <= 0.25)$$

$$P(-\frac{0.25\sqrt{n}}{\sigma} <= \frac{(\bar{X}_n - \mu)\sqrt{n}}{\sigma} <= \frac{0.25\sqrt{n}}{\sigma})$$

$$P(-\frac{0.25\sqrt{n}}{\sigma} <= Z <= \frac{0.25\sqrt{n}}{\sigma})$$

$$2\phi(\frac{0.25\sqrt{n}}{\sigma}) - 1$$

In [34]:

```
sigma = df.test400metri.std()
n = len(df.test400metri.dropna())
Z = st.norm()
2*Z.cdf(0.25*(n**0.5)/sigma) - 1
```

Out[34]:

```
0.5712981015960152
```