M5_T01

February 26, 2023

1 Sprint 5

1.1 Tasca M5 T01

1.2 Exercici 1

Descarrega el dataset adjunt de dades oficials de la UEFA i selecciona un atribut del conjunt de dades. Calcula el p-valor i digues si rebutja la hipòtesi nul · la agafant un alfa de 5%.

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df=pd.read_csv('Lligues europees UEFA.csv', sep=';',encoding='unicode-escape')
df.head(5)
```

[1]:

	Rk	Squad	Country	LgRk	MP	W	D	L	GF	GA	…	Pts	Pts/G	\
0	1	Manchester City	ENG	1	37	28	6	3	96	24	…	90	2.43	
1	2	Liverpool	ENG	2	36	26	8	2	89	24	…	86	2.39	
2	3	Real Madrid	ESP	1	37	26	7	4	80	31	…	85	2.30	
3	4	Bayern Munich	GER	1	34	24	5	5	97	37	…	77	2.26	
4	5	Paris S-G	FRA	1	37	25	8	4	85	36	…	83	2.24	

	xG	xGA	xGD	xGD/90	Last 5	Attendance	Top Team Scorer	\
0	86.1	26.8	59.3	1.60	W W W W D	52739	Kevin De Bruyne - 15	
1	84.6	33.1	51.4	1.43	W W W D W	53367	Mohamed Salah - 22	
2	73.0	45.8	27.2	0.73	W W L W D	40624	Karim Benzema - 27	
3	88.1	37.1	51.0	1.50	W W L D D	33176	Robert Lewandowski - 35	
4	71.6	38.1	33.4	0.90	W D D D W	41188	Kylian Mbappé - 25	

	Goalkeeper
0	Ederson
1	Alisson
2	Thibaut Courtois
3	Manuel Neuer
4	Keylor Navas

[5 rows x 21 columns]

1.2.1 Hypothesis Testing

- H0: La distribució de Punts dels 5 millors equips de les lligues corresponen a una distribució Gaussiana
- H1: La distribució de Punts dels 5 millors equips de les lligues no corresponen a una distribució Gaussiana

[2]:
```
from scipy.stats import normaltest
alpha=0.05
col='Pts'
data = df[df['LgRk']<=5]
stat, p = normaltest(data[col])

print('stat=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('\033[1m'+'Probably Gaussian'+'\033[0m')
    print('\n\033[1m'+'Cannot reject null hypothesis''\033[0m')
else:
    print('\033[1m'+'Probably not Gaussian'+'\033[0m')
    print('\n\033[1m'+'Reject null hypothesis'+'\033[0m')
```

```
stat=2.021, p=0.364
Probably Gaussian

Cannot reject null hypothesis
```

- Veiem que no es pot rebutjar la hipòtesis nul · la i per tant la *distribució de punts* correspon a una Gaussiana.

1.2.2 Hypothesis Testing

- H0: La distribució d'Assistents dels equips espanyols i anglesos és la mateixa
- H1: La distribució d'Assistents dels equips espanyols i anglesos no és la mateixa

[3]:
```
from scipy.stats import ttest_ind

data1 = df[df['Country']=='ESP']
data2 = df[df['Country']=='ENG']
col='Attendance'
stat, p = ttest_ind(data1[col], data2[col])

print('stat=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('\033[1m'+'Probably the same distribution'+'\033[0m')
    print('\n\033[1m'+'Cannot reject null hypothesis''\033[0m')
else:
    print('\033[1m'+'Probably different distributions'+'\033[0m')
    print('\n\033[1m'+'Reject null hypothesis'+'\033[0m')
```

```
stat=-3.535, p=0.001
Probably different distributions

Reject null hypothesis
```

- Veiem que es rebutja la hipòtesis nul · la i per tant la *distribució del nombre d'assistents* de la lliga espanyola és la mateixa que la anglesa.

1.2.3 Hypothesis Testing

- H0: La distribució d'Assistents dels equips espanyols i alemanys és la mateixa
- H1: La distribució d'Assistents dels equips espanyols i alemanys no és la mateixa

[4]:
```
data1 = df[df['Country']=='ESP']
data2 = df[df['Country']=='GER']
col='Attendance'
stat, p = ttest_ind(data1[col], data2[col])

print('stat=%.3f, p=%.3f' % (stat, p))

if p > alpha:
    print('\033[1m'+'Probably the same distribution'+'\033[0m')
    print('\n\033[1m'+'Cannot reject null hypothesis''\033[0m')
else:
    print('\033[1m'+'Probably different distributions'+'\033[0m')
    print('\n\033[1m'+'Reject null hypothesis'+'\033[0m')
```

```
stat=0.498, p=0.622
Probably the same distribution

Cannot reject null hypothesis
```

- Veiem que no es pot rebutjar la hipòtesis nul · la i per tant la *distribució del nombre d'assistents* de la lliga espanyola no és la mateixa que la alemana.

1.3 Exercici 2

Amb el mateix dataset selecciona dos altres atributs del conjunt de dades. Calcula els p-valors i digues si rebutgen la hipòtesi nul · la agafant un alfa de 5%.

1.3.1 Hypothesis Testing

- H0: No hi ha correlació entre els gols esperats (xG) i els punts (Pts)
- H1: Hi ha dependència entre els gols esperats (xG) i els punts (Pts)

[5]:
```
from scipy.stats import pearsonr

# calculate the correlation coefficient and p-value
corr, p_value = pearsonr(df['xG'], df['Pts'])
```

```
# print the results
print('Pearson correlation coefficient:', corr)
print('p-value:', p_value)

# check if null hypothesis is rejected at alpha=0.05
if p_value < alpha:
    print('\n\033[1m'+'Reject null hypothesis'+'\033[0m')
else:
    print('\n\033[1m'+'Cannot reject null hypothesis''\033[0m')
```

```
Pearson correlation coefficient: 0.8481665702817157
p-value: 3.0809793792070877e-28
```

Reject null hypothesis

- Veiem que es rebutja la hipòtesis nul · la i per tant hi ha una *relació* entre gols esperats i punts.

1.3.2 Hypothesis Testing

- H0: No hi ha una notable diferència entre gols marcats pels màxims golejadors de les lligues anglesa i espanyola
- H1: Hi ha una notable diferència entre gols marcats pels màxims golejadors de les lligues anglesa i espanyola

[8]:
```
def goals(ds):
    goalst=[]

    for ii in ds:
        goalst.append(ii.split(' - ')[1])

    goalst=pd.Series(goalst,dtype=int)
    return goalst

espg = df.loc[df['Country'] == 'ESP', 'Top Team Scorer']
engg = df.loc[df['Country'] == 'ENG', 'Top Team Scorer']
esp_goals= goals(espg)
eng_goals= goals(engg)

# perform the t-test
t, p_value = ttest_ind(esp_goals, eng_goals, equal_var=False)

# print the results
print('t-value:', t)
print('p-value:', p_value)

# check if null hypothesis is rejected at alpha=0.05
```

```
if p_value < alpha:
    print('\n\033[1m'+'Reject null hypothesis'+'\033[0m')
else:
    print('\n\033[1m'+'Cannot reject null hypothesis'+'\033[0m')
```

```
t-value: -0.10072416522709038
p-value: 0.9203085943952437
```

Cannot reject null hypothesis

- Veiem que no es pot rebutjar la hipòtesis nul · la i per tant els gols marcats pels màxims golejadors de la lliga espanyola i la anglesa són similars.

1.4 Exercici 3

Continua amb el conjunt de dades adjunt i selecciona tres atributs del conjunt de dades. Calcula el p-valor i digues si rebutja la hipòtesi nul · la agafant un alfa de 5%.

1.4.1 Hypothesis Testing

- H0: No hi ha una diferència significant en la mitja de gols a favor (GF) de les lligues anglesa, espanyola i alemana
- H1: Hi ha una diferència significant en la mitja de gols a favor (GF) de les lligues anglesa, espanyola i alemana

[7]:
```
from scipy.stats import f_oneway

# select the data for the three leagues
eng_data = df[df['Country'] == 'ENG']['GF']
esp_data = df[df['Country'] == 'ESP']['GF']
ger_data = df[df['Country'] == 'GER']['GF']

# perform the one-way ANOVA test
f_stat, p_value = f_oneway(eng_data, esp_data, ger_data)

# print the results
print("F-statistic: {:.2f}".format(f_stat))
print("p-value: {:.4f}".format(p_value))

# check if null hypothesis is rejected at alpha=0.05
if p_value < alpha:
    print('\n\033[1m'+'Reject null hypothesis'+'\033[0m')
else:
    print('\n\033[1m'+'Cannot reject null hypothesis'+'\033[0m')
```

```
F-statistic: 0.70
p-value: 0.4988
```

`Cannot reject null hypothesis`

- Veiem que no es pot rebutjar la hipòtesis nul · la i per tant la mitja de gols a favor de les lligues anglesa, espanyola i alemana es similar, no hi ha una notable diferència.