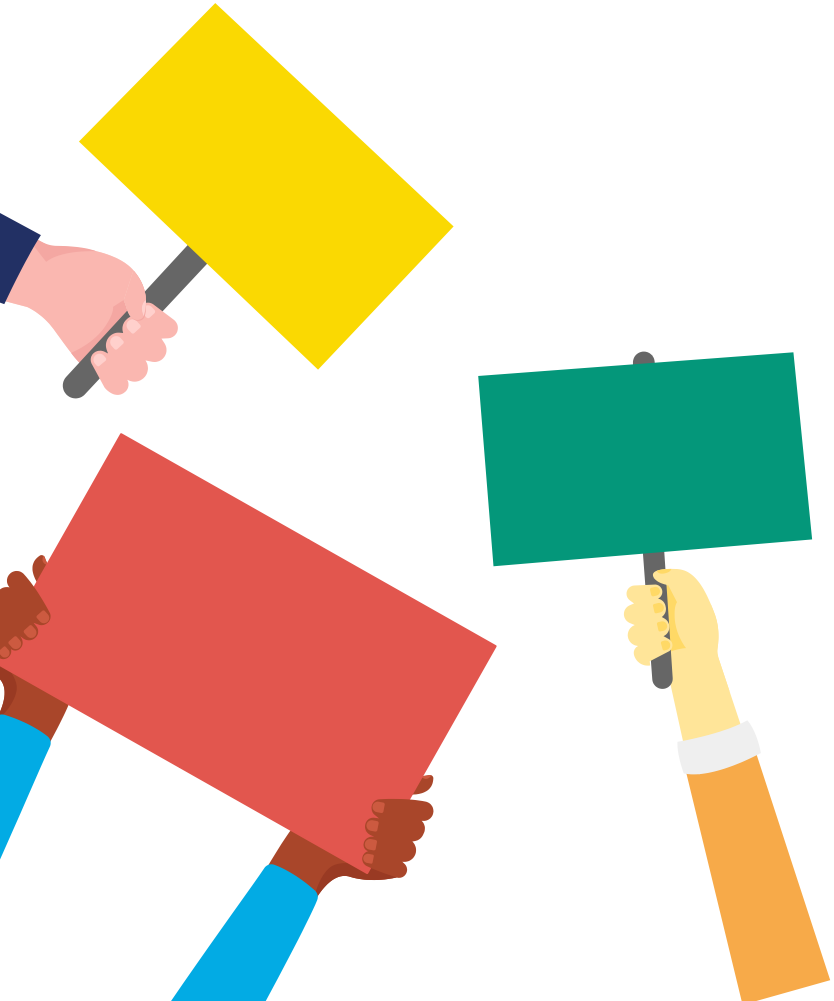
HW1

利用統計方法檢測學生總成績

111426025 盧盈穎

2022/09/19

目錄



01

研究動機

01 研究動機
Research motivation

- 「龍生龍鳳生鳳，老鼠生的孩子會打洞」—俗諺

- 學生學習成績是否受學生有沒有先修上課或父母的教育程度等因素影響

02

資料集介紹

使用kaggle所提供Student Performance In Exams資料集

https://www.kaggle.com/code/spscientist/student-performance-in-exams/data

02. 資料集介紹 / 資料前處理 / 資料觀察

Intro. to Datasets

gender	race/ ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
female	group D	some college	free/ reduced	none	77	86	86

總共:1000筆資料

資料欄位統計敘述

主欄位	類別	總數
test preparation course	none	2
	completed	
parental level of education	some college	6
	associate's degree	
	high school	
	some high school	
	bachelor's degree	
	master's degree	

以上為本次研究會使用到的資料欄位。

1. 檢查資料型態

```
df.dtypes
```

```
gender                         object
race/ethnicity                 object
parental level of education    object
lunch                          object
test preparation course        object
math score                      int64
reading score                   int64
writing score                   int64
dtype: object
```

2. 檢查資料是否有空值

```
df.isnull().sum()
```

```
gender                         0
race/ethnicity                 0
parental level of education    0
lunch                          0
test preparation course        0
math score                     0
reading score                  0
writing score                  0
dtype: int64
```

3. 將數學、閱讀、寫作三科分數進行加總及平均，得到學生總成績平均。

```
df['Total'] = (df['math score']+df['reading score']+df['writing score'])/3
```

```
for i in range(len(df['Total'])):
    df['Total'][i] = round(df['Total'][i],2)
```

4. 刪除多餘的欄位。

```
df = df.drop('math score',axis=1)
df = df.drop('reading score',axis=1)
df = df.drop('writing score',axis=1)
```

03

研究方法

03. 研究方法
Methods

目的：

學生有沒有先修上課是否會影響學生總成績？

統計方法：

T-Test 檢測學生有沒有上先修課程是否會影響學生總成績。

03. 研究方法
Methods

將1000筆學生成績資料進行分類、去除不必要欄位整理後資料

實際程式中跑的資料

	Total	Test
0	72.67	none
1	92.67	none
2	49.33	none
3	76.33	none
. **.** **.**		
999	74.33	completed

test preparation course

```
df_no  =  df[df['test  preparation  course']=="none"]
df_co  =  df[df['test  preparation  course']=="completed"]
```

```
df_no.count()
```

```
gender                       642
race/ethnicity               642
parental level of education  642
lunch                        642
test preparation course      642
Total                        642
dtype: int64
```

```
df_co.count()
```

```
gender                       358
race/ethnicity               358
parental level of education  358
lunch                        358
test preparation course      358
Total                        358
dtype: int64
```

兩種類別數量皆大於30

03. 研究方法
Methods

1. 檢查資料集是否符合常態分配

根據中央極限定理，樣本數夠大，樣本和減去平均數再除以標準差，將會趨近平均數為0。

```
a = df_no['Total'].sum()
```

```
import statistics
d = st_dev = statistics.pstdev(df_no['Total'])
d
```

```
14.175514823953018
```

```
a = a/642
```

```
a = a-65.038801
```

```
c = 642**0.5
```

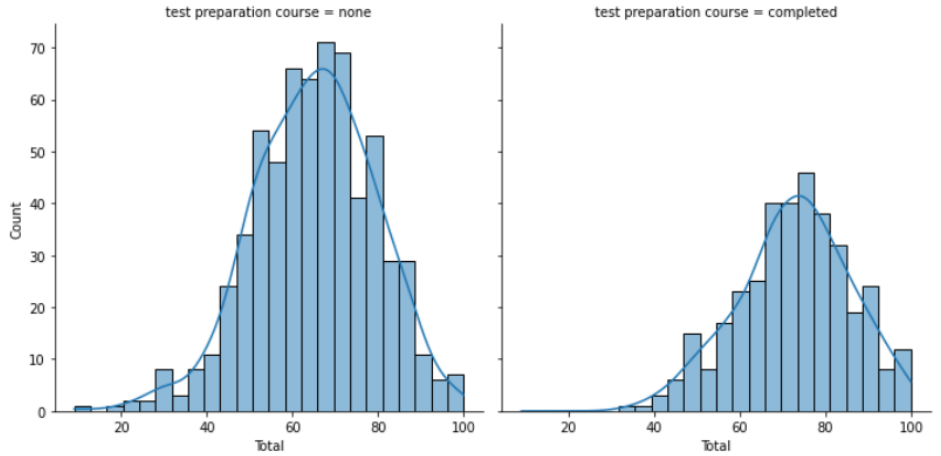
```
d = d/c
```

```
a/d
```

```
-6.737658792570465e-07
```

```
import seaborn as sns
sns.displot(data=df_mix2, x='Total', col='test preparation course', kde='true')
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb819561190>
```



03. 研究方法
Methods

2. 檢查資料集是否為同性質

```
print('None:',np.var(df_shs_RR),'Completed:'  ,np.var(df_hs_RR))
```

```
None: 200.945220524111176 Completed: 169.49196703676537
```

```
223.26228499734717/181.60957426853395
```

```
1.229353055292251
```

此資料集方差的比率為1.2293，小於4，表示可以假設總體方差相等。

3. 虛無假設

H_0: $\mu_{A1} = \mu_{A2}$

H_a:學生有沒有上先修課程考試成績平均值不完全相同

α = 0.05

03. 研究方法
Methods

結論：

```
mean1 = np.mean(df_shs_RR)
mean2 = np.mean(df_hs_RR)
```

```
std1 = np.std(df_shs_RR)
std2 = np.std(df_hs_RR)
```

```
nobs1 = len(df_shs_RR)
nobs2 = len(df_hs_RR)
```

```
modified_std1 = np.sqrt(np.float32(nobs1)/np.float32(nobs1-1)) * std1
modified_std2 = np.sqrt(np.float32(nobs2)/np.float32(nobs2-1)) * std2
```

```
statistic, pvalue = stats.ttest_ind_from_stats( mean1=mean1, std1=modified_std1, nobs1=nobs1,
                                                mean2=mean2, std2=modified_std2, nobs2=nobs2 )
```

```
statistic
```

```
-8.391129735727805
```

```
pvalue
```

```
1.631376440841688e-16
```

p<0.05，拒絕H_0，學生有沒有上先修課程考試成績平均值不完全相同。

因此，學生有上先修課程和沒有上先修課程有顯著差異。

04

延伸問題

04. 延伸問題
Extension problem

目的：

如果加入父母的教育程度、學生有沒有上先修課程是否會影響學生總成績？

統計方法：

Two way ANOVA 檢測學生有沒有上先修課程和父母的教育程度是否會影響學生總成績。

04. 延伸問題
Extension problem

1. 取得資料

	Total	Parent	Test
0	72.67	bachelor's degree	none
1	92.67	master's degree	none
2	49.33	associate's degree	none
3	76.33	some college	none
. **.** **.**	**.** **.** **.**	**.** **.** **.**	**.** **.** **.**
999	74.33	some college	completed

04. 延伸問題
Extension problem

2. 檢查Parent欄位中各類別數量

```
a = N_df[N_df['Parent']=="some high school"]
b = N_df[N_df['Parent']=="high school"]
c = N_df[N_df['Parent']=="associate's degree"]
d = N_df[N_df['Parent']=="some college"]
e = N_df[N_df['Parent']=="bachelor's degree"]
f = N_df[N_df['Parent']=="master's degree"]
```

bachelor‘s degree

```
e.count()
```

```
Total     118
Test      118
Parent    118
dtype: int64
```

mater‘s degree

```
f.count()
```

```
Total     59
Test      59
Parent    59
dtype: int64
```

some high school

```
a.count()
```

```
Total     179
Test      179
Parent    179
dtype: int64
```

high school

```
b.count()
```

```
Total     196
Test      196
Parent    196
dtype: int64
```

associate’s degree

```
c.count()
```

```
Total     222
Test      222
Parent    222
dtype: int64
```

some college

```
d.count()
```

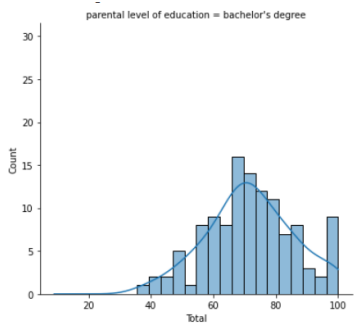
```
Total     226
Test      226
Parent    226
dtype: int64
```

以上各類別數量皆大於30，固可根據中央極限定理視為常態分佈

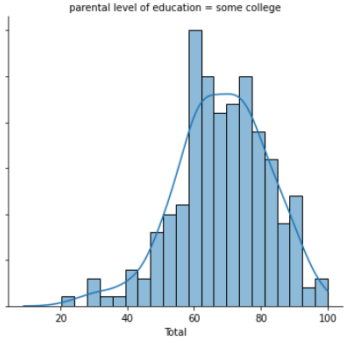
04. 延伸問題
Extension problem

3. 檢查資料集是否符合常態分配

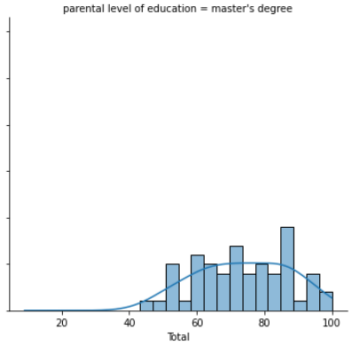
根據中央極限定理，樣本數夠大，可視為常態分佈。以下為Parent各類別的分布圖:



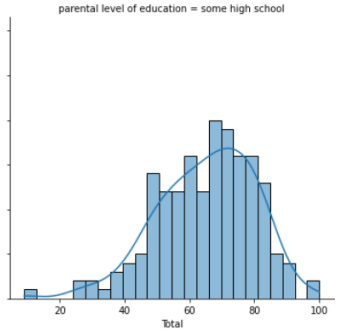
bachelor‘s degree



some college

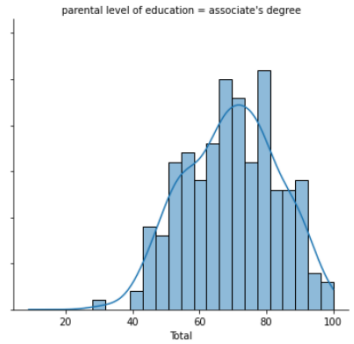


mater‘s degree

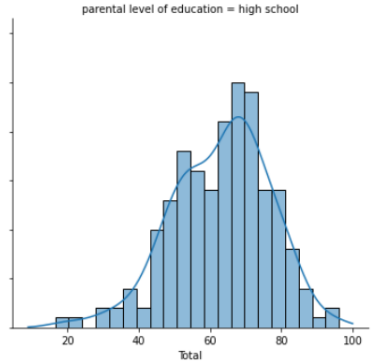


some high school

associate’s degree



high school



04. 延伸問題
Extension problem

4. 檢查資料集是否有相同變異數

Test 因子

```
stats.levene(aaa1,bbb1)
```

```
LeveneResult(statistic=2.8836464775685, pvalue=0.08979489565937057)
```

Parent因子

```
stats.levene(aaa,bbb,ccc,ddd,eee,fff)
```

```
LeveneResult(statistic=0.5080914858918754, pvalue=0.7703017393081184)
```

兩因子的P值皆大於0.05，故他們的變異數相同

04. 延伸問題
Extension problem

5. 虛無假設

Test 因子

μ_{A1} : 學生有上先修課程

μ_{A2} : 學生沒有上先修課程

H_0: $\mu_{A1} = \mu_{A2}$

H_a: 學生有沒有上先修課程成績平均值不完全相同

Parent因子

μ_{B1} : some college

μ_{B2} : associate's degree

μ_{B3} : high school

μ_{B4} : some high school

μ_{B5} : bachelor's degree

μ_{B6} : master's degree

H_0: $\mu_{B1} = \mu_{B2} = \mu_{B3} = \mu_{B4} = \mu_{B5} = \mu_{B6}$

H_a: 至少有一組父母的教育程度成績平均值不完全相同

04. 延伸問題
Extension problem

結論：

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
mod = ols('Total ~ Test+Parent+Test:Parent', data = N_df).fit()
sm.stats.anova_lm(mod, typ = 2)
```

	sum_sq	df	F	PR(>F)
Test	12862.375092	1.0	70.906113	1.306095e-16
Parent	9899.843741	5.0	10.914927	3.058591e-10
Test:Parent	561.825220	5.0	0.619432	6.850331e-01
Residual	179223.286814	988.0	NaN	NaN

學生有沒有上先修課程，p<0.05，拒絕H_0 ，學生有上先修課程和沒有上先修課程有顯著差異。

父母的教育程度，p<0.05 ，拒絕H_0 ，所以父母是碩士還是高中畢業對於學生總成績會產生顯著差異。

04. 延伸問題
Extension problem

Tukey HSD -父母的教育程度

```
import statsmodels.api as sm
mc = sm.stats.multicomp.MultiComparison(N_df["Total"],N_df["Parent"])
mc_result = mc.tukeyhsd()
print(mc_result)
```

```
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
=============================================================================
      group1              group2      meandiff p-adj   lower    upper  reject
-----------------------------------------------------------------------------
associate's degree bachelor's degree     2.355 0.6515  -2.1735  6.8835  False
associate's degree       high school  -6.4718  0.001 -10.3678 -2.5758   True
associate's degree   master's degree    4.0297 0.3574  -1.7925  9.8518  False
associate's degree      some college   -1.0929    0.9   -4.849  2.6633  False
associate's degree  some high school   -4.4613 0.0183  -8.4543 -0.4683   True
 bachelor's degree       high school   -8.8268  0.001 -13.4584 -4.1952   True
 bachelor's degree   master's degree    1.6747    0.9  -4.6634  8.0127  False
 bachelor's degree      some college   -3.4479 0.2476  -7.9625  1.0667  False
 bachelor's degree  some high school   -6.8163  0.001 -11.5298 -2.1028   True
       high school   master's degree   10.5015  0.001   4.5988 16.4041   True
       high school      some college    5.3789 0.0011   1.4991  9.2587   True
       high school  some high school    2.0105 0.7022  -2.0991    6.12  False
   master's degree      some college   -5.1225 0.1204 -10.9339  0.6888  False
   master's degree  some high school   -8.491  0.001 -14.4582 -2.5238   True
      some college  some high school   -3.3685 0.1508  -7.3457  0.6088  False
-----------------------------------------------------------------------------
```

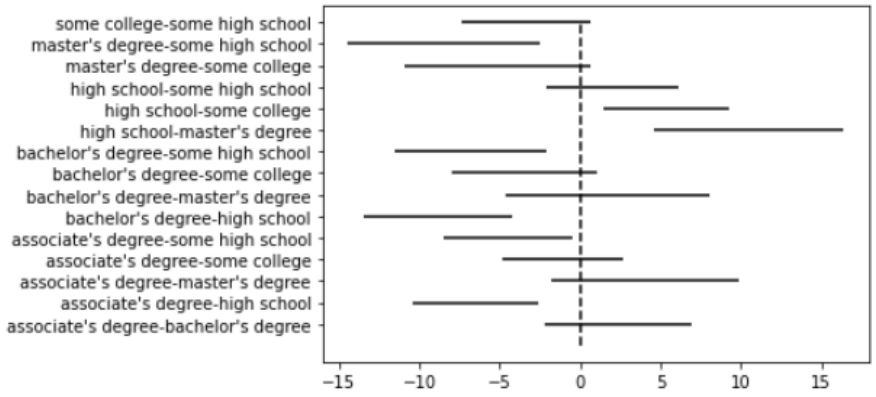
在畫黃色底線的類別，由於他們的p值小於0.05，他們之間存在差異。

以下會以區間圖及箱形圖顯示他們之間的交互關係是否存在差異

04. 延伸問題
Extension problem

Tukey HSD - 父母的教育程度

```
import matplotlib.pyplot as plt
rows = mc_result.summary().data[1:]
plt.hlines( range(len(rows)), [row[4] for row in rows], [row[5] for row in rows] )
plt.vlines( 0, -1, len( rows )-1, linestyles='dashed' )
plt.gca().set_yticks( range( len( rows ) ) )
plt.gca().set_yticklabels( [ f'{x[0]}-{x[1]}' for x in rows ] )
plt.show()
```



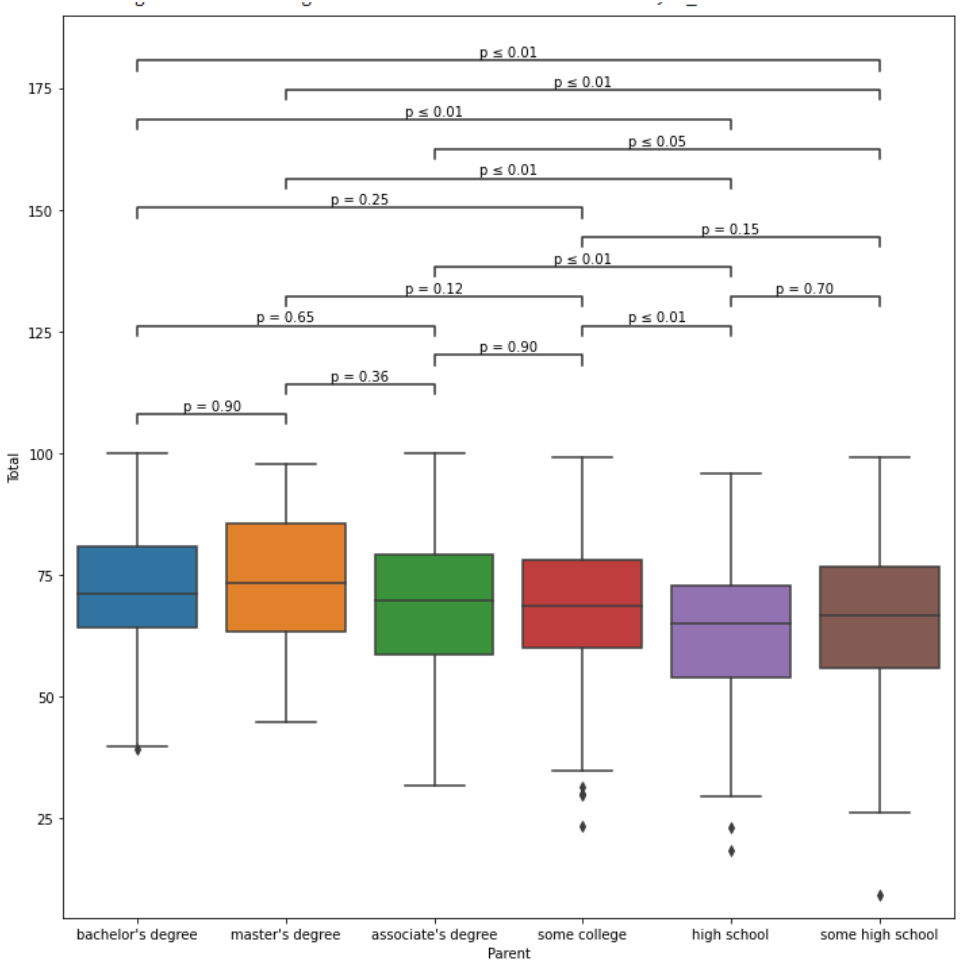
04. 延伸問題
Extension problem

Tukey HSD - 父母的教育程度

```
from statannotations.Annotator import Annotator
from matplotlib import pyplot as plt
plt.figure(figsize=(10,10))
ax = sns.boxplot(data=N_df, x="Parent", y="Total")

pairs = [(i[1]["index"], i[1]["variable"]) for i in molten_df.iterrows()]
p_values = [i[1]["value"] for i in molten_df.iterrows()]

annotator = Annotator(
        ax, pairs, data=N_df, x="Parent", y="Total")
annotator.configure(text_format="simple", loc="inside")
annotator.set_pvalues_and_annotate(p_values)
plt.tight_layout()
```



謝謝大家

111426025 盧盈穎