探索性資料分析(EDA)_探討變數之間的關係

簡報閱讀 範例與作業 問題討論 學習心得(完成)

重要知識點

掌握不同變數類型間的關係，以數值的方式描述下列三種資料型態的關係

- 連續 vs 連續

透過以下範例，了解變數之間的關係

20 筆資料，收集到 5 種資料，包含

- sex：性別
- insomnia：失眠
- age：年齡
- height：身高
- weight：體重

	sex	insomnia	age	height	weight
0	Male	Y	23	180	100
1	Male	N	40	170	68
2	Male	N	5	100	20
3	Male	N	30	176	70
4	Male	N	1	70	10
5	Female	N	40	160	45
6	Female	Y	16	170	50
7	Female	Y	27	166	58
8	Female	Y	43	155	58
9	Female	N	8	35	17
10	Male	Y	23	170	101
11	Male	N	39	168	65
12	Male	N	5	101	22
13	Male	N	29	175	79
14	Male	N	1	72	12
15	Female	N	42	163	40
16	Female	Y	13	169	53
17	Female	Y	29	163	52
18	Female	Y	41	151	56
19	Female	N	10	40	14

挖掘變數之間的關係

- 圖像型就是把圖畫出來，透過資料在圖形上走勢，判斷變數間的相關性。
- 數值型就是運用一些數值的特性，判斷變數間是否存在某些相關性。

今天的課程介紹如何透過數值來挖掘變數之間的特性。

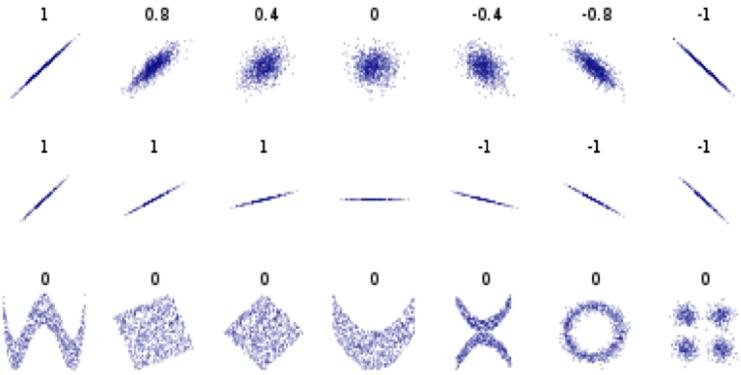
每一個類型皆介紹一種常見的用法

- 連續 vs 連續
  # Pearson 相關係數
- 離散 vs 離散
  # Cramer's V 係數
- 離散 vs 連續
  # Point biserial's correlation
  # **Cohen's ds**
  # eta-squared

連續 vs 連續：Pearson 相關係數

用於量測兩個連續型變數之間，線性相依的程度
在今天的課程範例中，height：身高；weight：體重為連續型。
常用 r 作為代表符號

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=0}^{n}(X_i-\bar{X})^2\sum_{i=0}^{n}(Y_i-\bar{Y})^2}}$$



Pearson 相關係數只能呈現線性關係，第三列的圖形，圖形呈現對稱關係，但Pearson 相關係數為0。

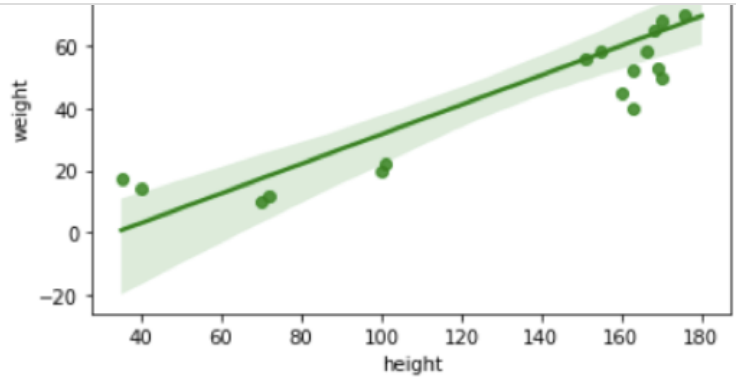
Pearson 判斷相關的準則

相關係數範圍	變數之間關聯程度
1	完全線性相關
0.7-0.99	高度線性相關
0.4-0.69	中度線性相關
0.1-0.39	低度線性相關
< 0.1	無線性相關

stats.pearsonr 計算相關性

```
# 由於 pearsonr 有兩個回傳結果，我們只需取第一個回傳值為相關係數
corr, _=stats.pearsonr(data['height'], data['weight'])
print(corr)
#代表身高和體重有高度線性相關
```

```
0.8380879580762451
```



圖形和數值皆呈現高度的線性相關性。

離散 vs 離散：Cramer’s V 係數

Cramer's V 運用卡方檢定的結果來運算出一 個可以估算離散型變數的相關性的指標。

定義如下：

$$V = \sqrt{\frac{\chi^2}{n \times (MIN(r,c)-1)}}$$

r：交叉列連表(contingency table) 的行數
C：交叉列連表(contingency table) 的列數
n：資料總筆數

計算 Cramer’s V 係數的步驟

目標：我們想要探討性別與失眠的關係

- Step1：用交叉列連表(contingency table) 整理資料

$$df^* = MIN(r, c) - 1$$

- Step3：運用 researchpy 套件，計算出 Cramer's V 係數

Cramer's V 判斷相關的準則

不同的自由度，有對應判斷兩個離散型變數強度的準則。

df*	negligible	small	medium	large
1	0 ~ .10	.10 ~ .30	.30 ~ .50	.50 or more
2	0 ~ .07	.07 ~ .21	.21 ~ .35	.35 or more
3	0 ~ .06	.06 ~ .17	.17 ~ .29	.29 or more
4	0 ~ .05	.05 ~ .15	.15 ~ .25	.25 or more
5	0 ~ .05	.05 ~ .13	.13 ~ .22	.22 or more

researchpy.crosstab 計算相關性

Step 1：用交叉列連表(contingency table) 整理資料

```
contTable = pd.crosstab(data['sex'], data['insomnia'])
contTable
```

insomnia	N	Y
sex		
Female	4	6
Male	8	2

Step 2：計算資料自由度 df

$df = min(r, c) - 1$

```
df = min(contTable.shape[0], contTable.shape[1]) - 1
df
```

```
1
```

Step 3：運用 researchpy 套件，計算出 Cramer's V 係數

- 失眠狀態和性別呈現中度相關

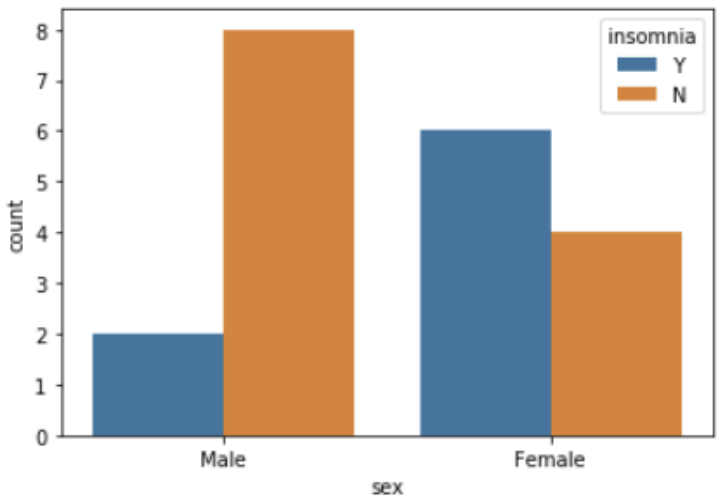
```
crosstab, res = researchpy.crosstab(data['sex'], data['insomnia'], test='chi-square')
#print(res)
print("Cramer's value is",res.loc[2,'results'])

#這邊用卡方檢定獨立性，所以採用的 test 參數為卡方 "test =" argument.
# 採用的變數在這個模組中，會自己根據資料集來判斷，Cramer's Phi if it a 2x2 table, or Cramer's V is larger than 2x2.
```

```
Cramer's value is 0.4082
```

```
valiate_CramerV(df,res.loc[2,'results'])
```

```
'medium'
```



離散 vs 連續：Eta Squared (η2)

Eta Squared (η^2)

- 描述一個離散型變數和連續型變數的相關性

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}}$$

$$SS_{total} = SS_{within} + SS_{between}$$

以連續型變數的數值計算變異數

總變異數：

$$SS_{total} = \sum_{all}(X_i - \bar{X}_{all})^2$$

組內變異數：

$$SS_{within} = \sum_{g\in 各組} \sum_{i\in 各組資料} (X_{gi} - \bar{X}_g)^2$$

組間變異數：

$$SS_{between} = \sum_{g\in 各組} n_g(\bar{X}_g - \bar{X}$$

Eta Squared (η2) 判斷相關的準則

η^2	Interpretation
0.00 < 0.01	Negligible
0.01 < 0.06	Small
0.06 < 0.14	Medium
0.14 <= 1.00	Large

離散 vs 連續：Eta Squared (η2)

目標：失眠和體重的相關性

- Step1：取出失眠和體重資料
- Step2：運用 pg.anova 計算三種變異數
- Step3：變異數換算得到 Eta Squared (η^2)

pg.anova 計算變異數

```
#!pip install pingouin
import pingouin as pg
```

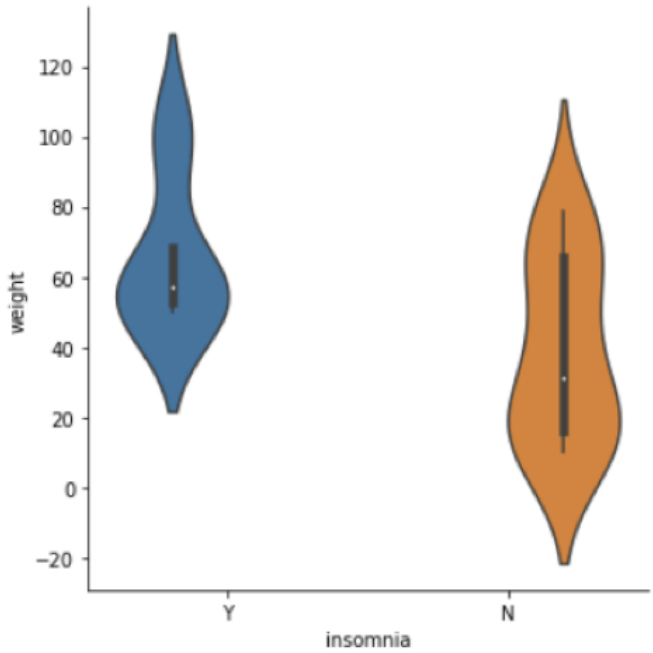
```
aov = pg.anova(dv='weight', between='insomnia', data=data, detailed=True)
aov
```

		Source	SS	DF	MS	F	p-unc	np2
組間變異數 →	0	insomnia	3630.0	1	3630.000000	6.123137	0.023521	0.253828
組內變異數 →	1	Within	10671.0	18	592.833333	NaN	NaN	NaN

計算相關性 Eta Squared (η2)

```
def valiate_etaSq(etaSq):
    if etaSq < .01:
        qual = 'Negligible'
    elif etaSq < .06:
        qual = 'Small'
    elif etaSq < .14:
        qual = 'Medium'
    else:
        qual = 'Large'
    return(qual)
valiate_etaSq(etaSq)
```

```
'Large'
```



失眠狀態和體重呈現高度相關

總結

今天的課程中，透過數值的方法，偵測三種不同資料類型的相關性。

方法	Pearson 皮爾森	Cramer's V 克雷莫	
變數特性	兩個成對**連續**變數	兩個成對**離散**變數	成對的- 一個離

回到今天的程式範例

- pip install pingouin
- pip install researchpy

延伸閱讀

何謂卡方檢定？

在Cramer's 我們用到卡方檢定，驗證從兩個變數抽出的配對觀察值組是否互相獨立，如果不獨立則代表變數間可能相關。

	男	女	總計
右	43	44	87
左	9	4	13
總計	52	48	100

	男	女	總計
右	43 (45.24)	44 (41.76)	87
左	9 (6.76)	4 (6.24)	13
總計	52	48	

圖表資料來源：皮爾森卡方檢驗

下一步：閱讀範例與完成作業