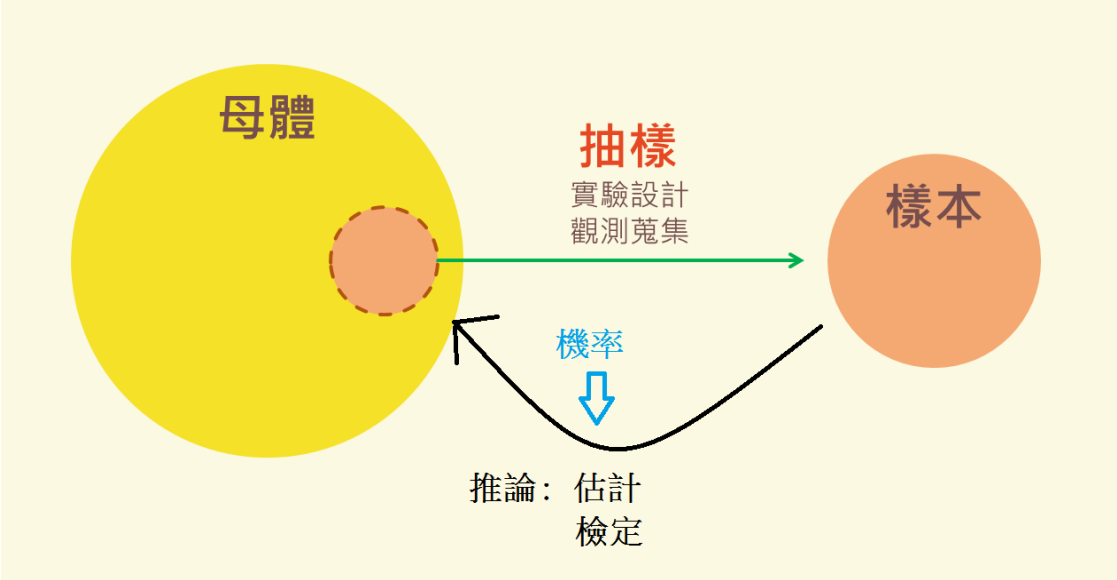
應用統計方法

Lecture 1 2/24/2021

課程簡介

- 請同學在紙上寫下專業背景，修課動機，是否修過相關課程,特別想聽的主題，在下課前給老師 (不一定要留姓名)
- 要先在 new E3 下載講義帶來上課
- 旁聽同學請給我姓名與學號，可以在 new E3 獲得上課資訊

統計學概述



*** 上學期課程與本學期比較**

	上學期	本學期
變數	單一：X	數個 $(Y, X_1, ..., X_p)$，$(X_1, ..., X_p)$
機率假設	單維度 X	單維度 $Y \mid X_1, ..., X_p$；高維度 $(X_1, ..., X_p)$
Interest:	X 分配的參數	建立模型或變數間的相關性分析
樣本:	均質	不一定均質 (模型會解釋變數間的關係)
推論	估計 & 檢定	估計檢定+額外的議題(選模與配適度問題)
計算	計算機 + 查表	統計軟體 (R, Excel, SPSS, SAS)
		本課程會學習 R
內容	基礎統計	各種方法
	資料分析	雙變量相關分析，迴歸分析
	機率	實驗設計，變異數分析
	信賴區間	類別資料分析, Logistic regression
	假設檢定	進階方法(多變量分析,存活分析,時間序列)

大數據時代：母體 = 樣本；雜亂缺乏結構與精確

主題一：關聯性分析

— 尋找現象間的'關聯性'是科學的基本問題

— 統計方法為探討科學問題的有用工具

Bivariate Data: (X_i, Y_i) $i = 1,...,n$

Examples:

a. (math score, physics score) for the ith student → 數理智能

b. (父親身高，兒子身高) for 第 i 對父子 → 遺傳顯現在智能

c. (父親 IQ，兒子 IQ) for 第 i 對父子 → 遺傳顯現在身高

d. (血壓，是否有心臟病) for the ith person → 心血管疾病風險因子

e. (是否有某個基因，是否有某個疾病) for the ith person → 疾病與遺傳

f. (治療方法,病人存活時間) for the ith patient → 哪個治療方法好

思考:"關聯性" 在以上例子的科學意義

Remark:

- 統計分析方法需要依據變數型態與變數角色來設計

  * Numerical (continuous) variable: 可做算數運算 $(+, -, \times, \div, \exp(.), \log(.), \sqrt{.})$

  * Ordinal: 很喜歡/有點喜歡/尚可/有點不喜歡/討厭

  * Categorical variable (類別式資料): 性別 (binary), 種族，婚姻狀態，職業別

雙變量之變數型態

- continuous vs. continuous: a, b, c
- continuous vs. binary: d
- binary vs. binary: e
- categorical vs. continuous: f

變數角色

- Response variable (depenent variable) – 通常是感興趣的 "outcome variable"
- Explanatory variable (indepenent variable) – 解釋變數 (亦稱為 covariate)

解釋變數	**因變數 (反應變數)**
父親 IQ (身高，數值型)	孩子 IQ (身高，數值型)
血壓 (數值型)	是否有心臟病 (二元類別)
治療方法 (二元類別)	存活時間 (數值型)

Remark: 有時候不需要區別兩者角色

- 各科成績 (數學, 物理, 化學, 英文, 國文, 歷史)
  - 數學與物理關聯性 > 數學與歷史關連性
- 運動成績 (短跑,跳遠相關性 > 短跑,馬拉松相關性)

常用的統計方法簡介 (本學期會教)

- Regression analysis:
  - response variable: 數值型 (常假設常態)
  - 解釋變數: 型態不拘
  - Example: 高中類別, 是否有補習, 讀書時間, 在校成績 → 學測成績
- Analysis of Variance (ANOVA):
  - response variable: 數值型 (常假設常態)
  - 解釋變數: 類別變數
  - Example: 溫度 (低溫, 常溫, 高溫) → 細菌數目
- Logistic regression analysis:
  - response variable: binary
  - 解釋變數: 可多個且型態不拘
  - Example: 大學排名, 補習與否, 讀書時間, 應考次數 → 考上律師與否
  - Example: 性別, 年齡, 家族史 → 得病與否
- Generalized linear model (logistic regression 為其特例)
  - response variable: 型態離散連續皆可
  - 解釋變數: 可多個且型態不拘
  - Example: 夏季氣溫, 水溝池塘面積, 人口數 → 登革熱人數
- Categorical data analysis (卡方檢定)
  - 未區分 response variable 或是解釋變數
  - 兩個類別式變數 →是否存在關聯性
  - Example:

    大學類別 (前段,中段,後段) 與 公司職位(一般職員, 中階主管, 高階主管)

1.1 Graphical presentation for numerical bivariate data

學習目標:

- 要懂得解讀他人提供的圖 (what are the main features?)
- 能夠把自己的資料透過圖表呈現, 讓人容易了解資料的重要訊息

Scatterplot: plot (X_i, Y_i) $i = 1,...,n$ → 平面上 n 個點

如何解讀 a scatterplot?

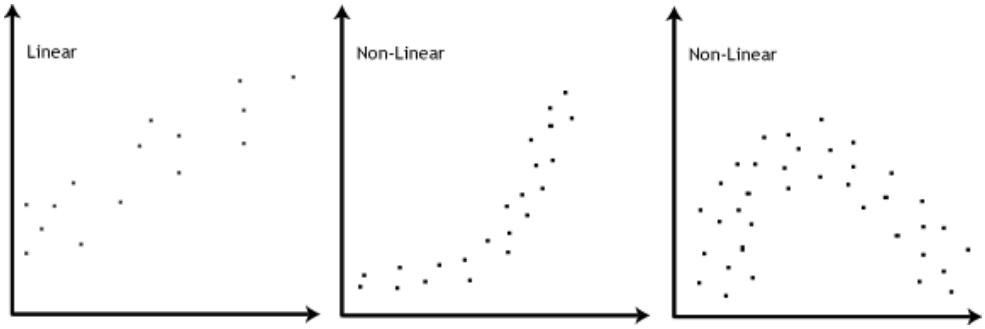
重點 1: 關聯性的方向

正相關 (positive association)

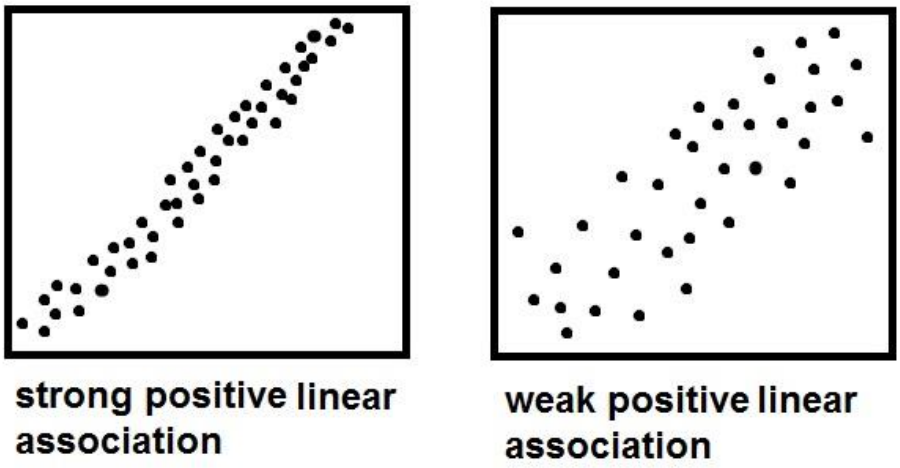
負相關 (negative association)

無相關 (no association)

重點 2: 關聯性的模式 (線性 vs. 非線性)

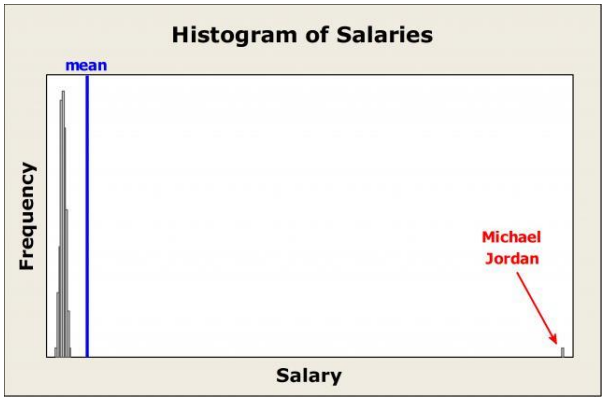


重點 3: 關聯性的強弱與清晰程度

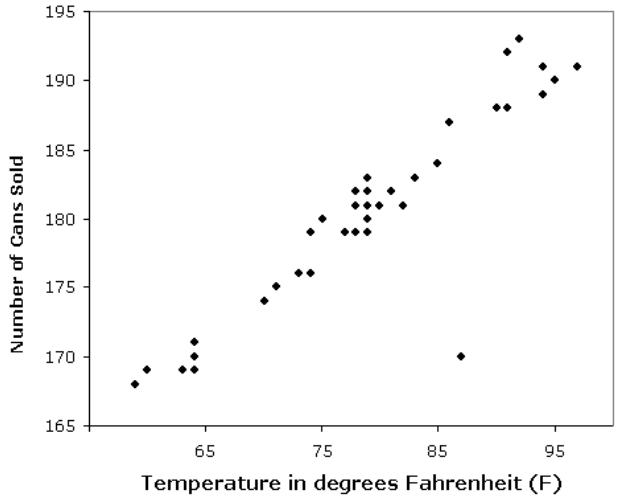


重點 4: 是否有 outlier (與眾不同的點)

Example: 北卡大學 (UNC) 地理系畢業生的薪水分布 → 單維度



example: 冰淇淋公司 → 雙維度



Remarks:

- 學習統計指標時考量
  - 衡量甚麼?
  - 指標的穩健性?

Robustness (穩健性)

- 統計指標是否容易受 outlier (離群值) 影響
- Non-robust measures:

  * mean: $\bar{X} = \sum_{i=1}^{n} X_i / n$

  * variance (S^2) and standard deviation S:

    $$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

  * Pearson's correlation coefficient (definition: later)

- Robust measures:

  * median: 最中間的數 (或兩數平均)

    n = odd → 排序第 $(n+1)/2$ 個數

    n = even → 排序第 $n/2$ 個數與第 $n/2+1$ 個數的平均

  * Interquartile range: Q3-Q1

    Q1 = the first quartile (前四分之一)

    Q3 = the third quartile (前四分之三)

    Q3-Q1 = the range of the middle 50%

  * Kendall's tau and Sparman's rho (definition: later)

機率背景:

單變量: X is a random variable

density: $f(x)$ satisfying $\int_{-\infty}^{\infty} f(x)dx = 1$

descriptive measures:

- mean $E(X) = \int xf(x)dx = \mu_x$

- variance $Var(X) = \int (x-\mu_x)^2 f(x)dx$

- median M satisfying $\Pr(X \le M) = \int_{-\infty}^{M} f(u)du = \frac{1}{2}$

- $Q_1 : \Pr(X \le Q_1) = \int_{-\infty}^{Q_1} f(u)du = \frac{1}{4}$, $Q_3 : \Pr(X \le Q_3) = \int_{-\infty}^{Q_3} f(u)du = \frac{3}{4}$

雙變量: (X,Y) is a pair of random variables

joint density: $f(x,y)$ (聯合機率密度函數) satisfying $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$

if (X,Y) are independent if and only if $f(x,y) = f_x(x)f_y(y)$

三種關聯性的測度

Original data: $(X_i, Y_i)(i = 1,...,n)$

1. Pearson correlation (相關係數) 的定義與意義

$$\rho_{X,Y} = corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Remarks:

1. $Cov(X,Y)$ 稱為 X 與 Y 的 "共變數" (covariance) (有單位)
2. $Cov(X,X) = Var(X) = \sigma_X^2$, $Cov(Y,Y) = Var(Y) = \sigma_Y^2$ (有單位)
3. 可以用科西-舒瓦茲不等式證明 $-1 \le \rho \le 1$ (沒有單位)

Note: 幾何意義

兩個向量 $\vec{a}$, $\vec{b}$, 夾角與內積的關係:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \in [-1,1]$$

內積: $Cov(X,Y)$, $\sqrt{Var(X)}$: $\|\vec{a}\|$, $\sqrt{Var(Y)}$: $\|\vec{b}\|$

Pearson correlation $\rho_{X,Y}$ 的估計

$\rho_{X,Y}$: parameter → unknown constant

Data: (X_i, Y_i) $i = 1,...,n$

估計量: $r_{X,Y} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) / n}{S_X S_Y}$

Remarks:

a. $S_X = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 / n}$ $\quad S_Y = \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 / n}$

b. $-1 \le r \le 1$

Pearson correlation 的解讀與限制

- 提出公式的人: Karl Pearson (與近代統計學發展史頗有淵源)

  Galton (達爾文的表弟) 成立生物統計實驗室, 後由 Karl Pearson 繼承
  Galton → regression analysis
  Pearson → Pearson correlation, 動差估計法 & 卡方分配

- $\rho_{X,Y}$ & $r_{X,Y}$ 均針對連續隨機變數

- 優點:"正負號"代表相關性的"方向";"大小"代表相關性的"強度"
- 缺點: 不能解釋非線性關聯 & 不夠穩健

Example 1: (1,1), (2,3), (3,5), (4,2), (5,6)
以下是 Excel Output (可以自己試試看)

編號	**1**	**2**	**3**	**4**	**5**	**sum**	**avg**	
x	1	2	3	4	5	15	3	
y	1	3	5	2	6	17	3.4	
x-mean	-2	-1	0	1	2			
y-mean	-2.4	-0.4	1.6	-1.4	2.6			
product	4.8	0.4	0	-1.4	5.2	9	1.8	
(x-mean)^2	4	1	0	1	4	10	2	1.414214
(y-mean)^2	5.76	0.16	2.56	1.96	6.76	17.2	3.44	1.854724
Corr	**0.686244**							

換第五個觀察值 → 用來檢視其穩健性

Example 2: (1,1), (2,3), (3,5), (4,2), (100,100)

編號	①	②	③	④	⑤	**sum**	**avg**	
x	1	2	3	4	**100**	110	22	
y	1	3	5	2	**100**	111	22.2	
x-mean	-21	-20	-19	-18	78			
y-mean	-21.2	-19.2	-17.2	-20.2	77.8			
product	445.2	384	326.8	363.6	6068.4	7588	1517.6	
(x-mean)^2	441	400	361	324	6084	7610	1522	39.01282
(y-mean)^2	449.44	368.64	295.84	408.04	6052.84	7574.8	1514.96	38.92249
Corr	**0.999423**							

Remarks:

1. 只改變一個觀測值 (5,6) → (100,100): 相關係數: 0.6862 → 0.99

代表 $\rho_{X,Y}$ 缺乏穩健性 (lack of robustness, not resistant, sensitive to extreme observations)

2. $\rho_{X,Y}$ 無法解釋 "curved relationship": 上課畫圖

https://youtu.be/RaClB0RpQec

2. Kendall's tau 的估計 (補充，不會考) → rank correlation coefficient

$\tau_{X,Y}$ = Pr((i,j) pairs are concordant)- Pr((i,j) pairs are discordant)

"concordance" 定義：兩個變數的大小方向相同

"discordance" 定義：兩個變數的大小方向相反

$\hat{\tau}_{X,Y}$ =(# of concordance – # of discordance)/K where $K=\binom{n}{2}$

計算方法 https://youtu.be/QXliM52_ZWI

- 由樣本任取兩個 (一對) 觀測值，判斷它們是 concordance 或是 discordance
- 再計算兩者比例的差 (分母 = $K=\binom{n}{2}$ = n 取 2 的組合數)

Remarks:

1. τ 為一 rank correlation (只與排序有關)
2. 適用於描述分佈呈偏斜 (skewed) 的變數
3. $-1 \le \tau \le 1$ (沒有單位)
4. 依然只能呈現 linear relationship

Examples 1 & 2 $n=5$, $K=10$

① (1,1)

② (2,3)

③ (3,5)

④ (4,2)

⑤ (5,6)　　5 (100,100)

$\hat{\tau}_{X,Y}=\frac{8}{10}-\frac{2}{10}=0.6$ (⑤ → 5 下表都一樣)

編號配對	con	Dis
(①,②)	✓	
(①,③)	✓	
(①,④)	✓	
(①,⑤)	✓	
(②,③)	✓	
(②,④)		✓
(②,⑤)	✓	
(③,④)		✓
(③,⑤)	✓	
(④,⑤)	✓	
total	8	2

Note: Kendall's tau 比較穩健 (robust, 不易受 outlier 影響)

3. Spearman's rho

Rank data: $(U_i, V_i)(i = 1,...,n)$ ← 排序資料 (值介於 1,…,n)

$$r = \frac{\sum_{i=1}^{n}(U_i - \bar{U})(V_i - \bar{V}) / n}{S_U S_V}$$ ← 還可以再整理

Note: $\sum_{i=1}^{n} U_i = 1 + 2 + ... + n = \frac{(1+n)n}{2}$

編號	①	②	③	④	⑤	sum	avg
x	1	2	3	4	5	15	3
y	1	3	5	2	6	17	3.4
u	1	2	3	4	5	15	3
v	1	3	4	2	5	15	3
u-mean	-2	-1	0	1	2	0	
v-mean	-2	0	1	-1	2	0	
product	4	0	0	-1	4	7	
(u-mean)^2	4	1	0	1	4	10	
(v-mean)^2	4	0	1	1	4	10	
Corr	0.7						

(⑤ → 5 Spearman's rho 都一樣)

- **三種測度比較**
  - Pearson's correlation: 使用原始資料 → 不穩健
  - Spearsman's rho: 使用排序資料 (或說轉到 PR scale) → 穩健
  - Kendall's tau: 兩兩同向反向關係 → 穩健

Remarks:

- "排序" 是建構無母數方法的重要方式
  - Wilcoxon rank test → 取代雙樣本 t 檢定
- 排序統計量也可以改寫為 "兩兩比大小" 的形式

*** Robustness (穩健性):** 不易受少數特異觀測值影響 (not sensitive to outliers)

- univariate: mean vs. median (rank based)

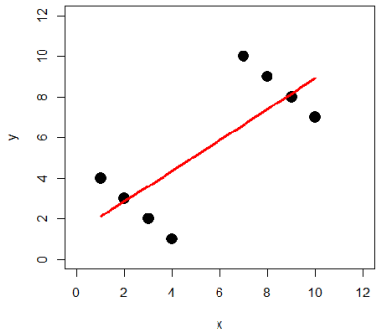
- bivariate: Pearson correlation vs. Kendall's tau or Spearman's rho (rank based)

*** 值得注意的特殊情況** (需要靠畫圖才看得清楚)

- Non-linear relationship (三種指標都無法描繪非線性關係)
- **當兩個關聯性指標的值差很多時，要留意細節了!!**

下圖: tau = 0.14, r = 0.84

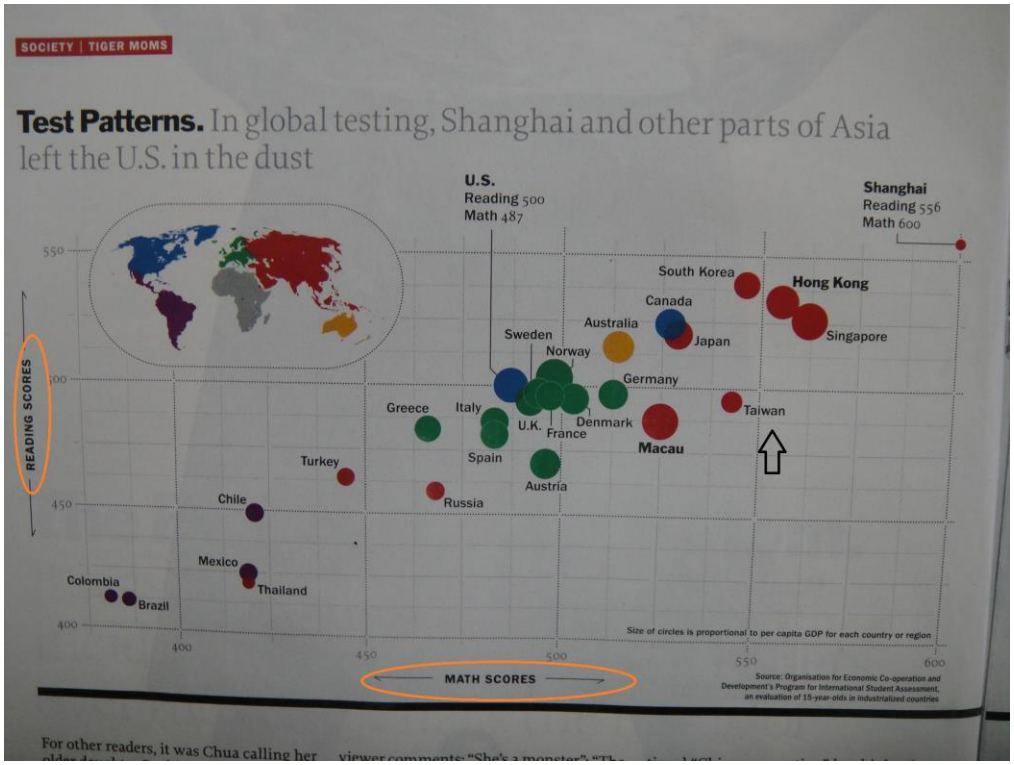
(隱含兩組人混在一起 → mixture)



附錄: PISA Report

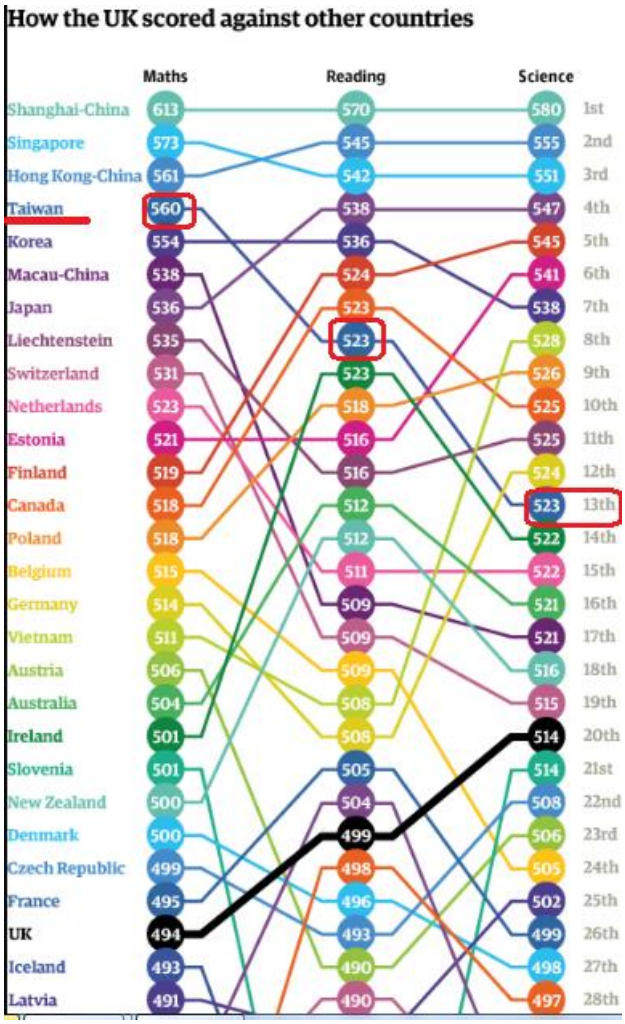
跨國評估學生能力計畫（英語：Programme for International Student Assessment，簡寫：PISA）是一個由 OECD 的對全世界 15 歲學生學習水平的測試計劃。最早開始於 2000 年，每三年進行一次。該計劃旨在發展教育方法與成果。是目前世界上最具影響力的國際學生學習評價項目之一。
2015 年超過 53 萬名學生代表 72 國參與測試。

PISA 各國學生在閱讀與數學的表現 (2010)



東亞國家: 右上, 注意 "上海" 的位置
美國的位置: 中間 ("虎媽的戰歌" by Amy Chua)

2012 年 UK 的檢討 (如何同時呈現三個變數)



PISA 2015：數學、科學全球第 4、閱讀滑落第 23，台灣學生欠實作能力

交大教育所余曉清教授

https://flipedu.parenting.com.tw/article/2977

PISA 2015 與 TIMSS 2015 再看國際數學評量中的臺灣

http://yaucenter.nctu.edu.tw/journal/201701/ch0/%E7%B0%A1%E8%A8%8A2.pdf

- 臺灣一項屢屢為人詬病的現象就是個別差異極為嚴重(換成口語說法即是:「高分的很高，低分的很低」)。
    - ■ 檢視歷屆數學成績，不管是用標準差或用高低分群的分數差距來衡量，均屬世界最高（2009 年、2012 年）或接近最高。
    - ■ 最新的PISA 2015 **臺灣的標準差103 與以色列並列世界第三**，僅次於馬爾它（110）和中國（106），雖已退出第一名寶座，但改善實在有限。
- 臺灣另一項令人憂慮的指標是低成就人數太多。PISA將數學素養水準分成6個等級，每個等級都有明確定義，基準線定在水準2(換句話說，水準1 和未達水準1 類似傳統意義的「不及格」)。

- 臺灣自參加PISA 以來，未達基準線的人數始終大於12%，雖然低於OECD (經濟合作暨發展組織)平均，但在東亞各國卻屬最高。
- 2015 年未達基準的人數佔12.7%，只比上一屆微降0.1%。雖然低於南韓突然竄高的15.5% 和中國的15.8%，但也絕不是好消息。

另一份診斷報告：TIMSSTIMSS 的全稱是「國際數學與科學教育成就趨勢調查」（Trends in International Mathematics and ScienceStudy）

- （黃敏雄）一文即針對臺灣在2003、2007 和2011 的數學成績進行分析，指出了兩個重要現象，其一是從小四到國二的表現劇升，其二是表現差異的擴大。
  - 前者指的是臺灣學生從小四升到國二時，數學表現達到最佳等級（稱為「進階國際標竿」）的比例大幅提高。
  - 關於〈TIMSS 比較〉一文指出的第二個現象，表現差異的擴大，我們可用標準差來觀察學生表現的差距大小，標準差愈小，表示學生的表現愈平均。歷屆以來，臺灣小四的標準差大致都落在較低區段，表示學生素質整齊，但是國二的標準差則居高不下。這個現象從小四到國二的突變現象也是受測各國僅見的。
    - ➔ 2007 年臺灣小四成績的標準差是極佳的69，這批學生到了國二後標準差升到106。
    - ➔ 2011 年臺灣小四成績的標準差是73，這批學生到了國二後標準差升到97。
    - ➔ 作為對照，同樣差距明顯的香港，2007 年四年級的標準差是67，四年後的八年級是84；2011 年四年級的標準差是66，四年後的八年級是78。

數學教育的罪與罰 (摘要) 2014 (依據 2012 年 PISA)

http://scimonth.blogspot.tw/2014/03/blog-post_3583.html

作者／單維彰（作者任教於中央大學數學系）

今年一月，本欄已經闡述了我國 15 歲少年在 2012 年 PISA 國際數學評量，呈現世界第一高的學習成效分散度。更具體地看各段成就分佈可以發現，我國被評定為數學能力落後的 15 歲少年比例過高。影響數學能力的因素當然很多，可能是興趣、天分、用功程度、學校課程的效率等等，而 PISA 特別調查了學生的家庭社經地位，並用統計方法評估 PISA 數學成績的變異性，被家庭社經地位「解釋」的程度，並以百分比呈現。用更淺白（但不盡正確）的話來說，該「解釋度」為 0.15 的意思是，學生的 PISA 數學成績有 15％是由家庭背景決定的。

PISA 官方報告中，特別有一節的標題是「公平性」（Equity），裡面提供一幅散佈圖，如圖。參與 PISA 2012 測驗的每個國家或地區，被賦予兩筆數據 x 和 y，其中 x 是家庭社經地位對該地學生成績變異性之解釋度，y 是該地學生的平均成

績。這兩個數據決定坐標平面上一點(x, y)，PISA 用一個菱形表示該點的位置。圖裡還有鉛直和水平的參考線，分別代表全體的平均成績（水平線）和平均社經解釋度（鉛直線）。

圖：各國 PISA 數學成績的變異性，對照家庭社經地位的「解釋程度」散布圖。（圖片來源： PISA 官方報告）→ 注意 X 軸順序

如果家庭比較有錢，孩子的成績就比較好，比較貧困，成績就比較差，這就是教育機會不公平的現象。注意 x 坐標的方向朝左遞增，也就是越靠右側表示教育機會的公平性越高，而越靠左邊就越不公平。

在上圖中，臺灣頗顯著地座落於第二象限（左上區）。我們的成績頗不錯（高於水平線），但是社經解釋度也偏高（在鉛直線左邊）。因此，PISA 將臺灣歸納為「學習機會不公平」的地區。看看我們的社經解釋度，不但高於世界平均，甚至高於美國（在兩線交點的下方）、英國（在兩線交點的右側），更遠遠高於韓國、日本、香港。

我還要特別提醒一件事。在 2006 年的 PISA 測驗報告中，我們的「不公平」指標和英國還是接近的（參閱民國 97 年 2 月本欄，當時本欄就已經關切此事），兩地都在國際平均附近（那一年，我國的成績分佈標準差是世界第三高）。六年之間，我們「進步」了很多，而英國的解釋度則略微下降，擴大了彼此間的差距。或許，以上的數據陳述，並不讓讀者太過意外；這個現象與趨勢，基本上就是我們許多人的共同生活經驗。當自己的人生被濃縮在一張統計圖表上，感覺並不太好。對照自己的生活經驗，怎能不體會統計的威力？

https://ourworldindata.org/quality-of-education

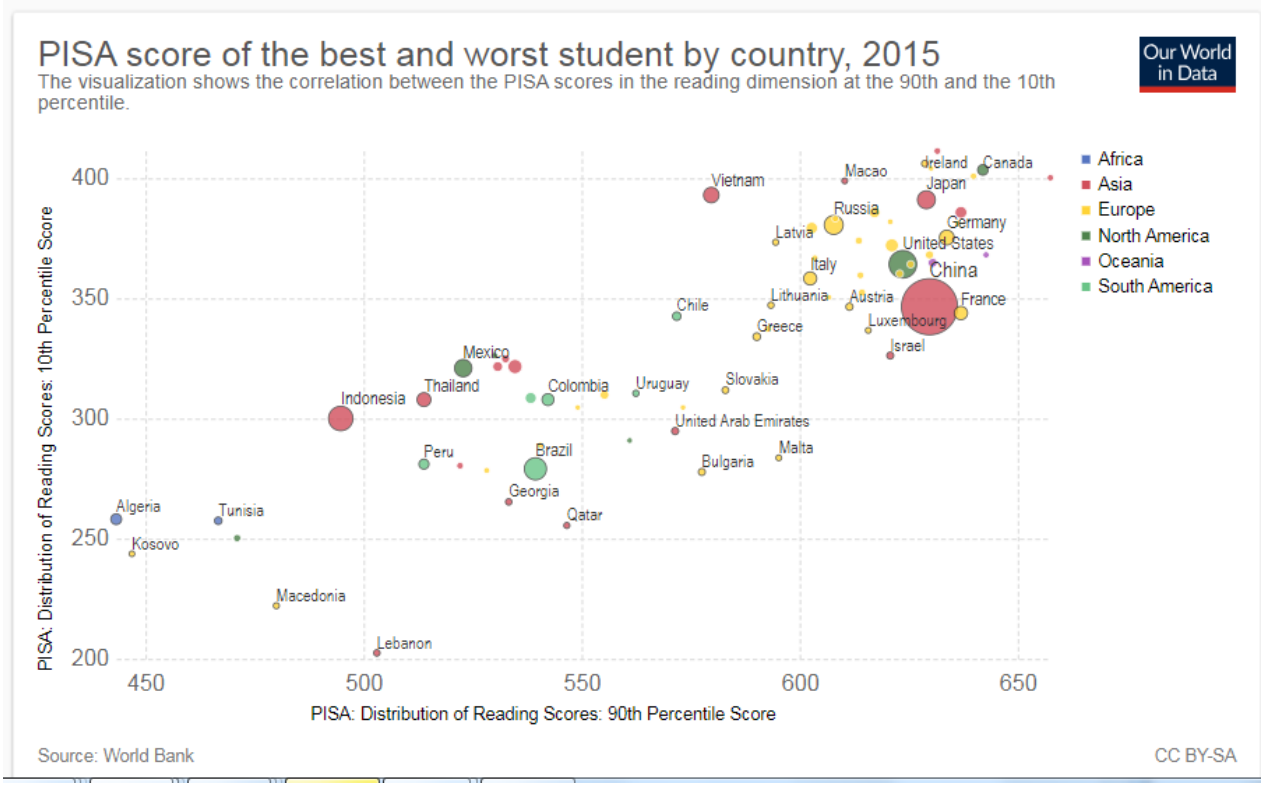
台灣不在內

2015 Plot PR90 vs. PR10

右上方：最好的狀況 → 好的很厲害，弱的也不差

左下方：最不好的狀況 → 好的不怎麼樣，弱的很差

往右下：好的不錯，弱的不佳（代表國內孩子程度差異大，如以色列）

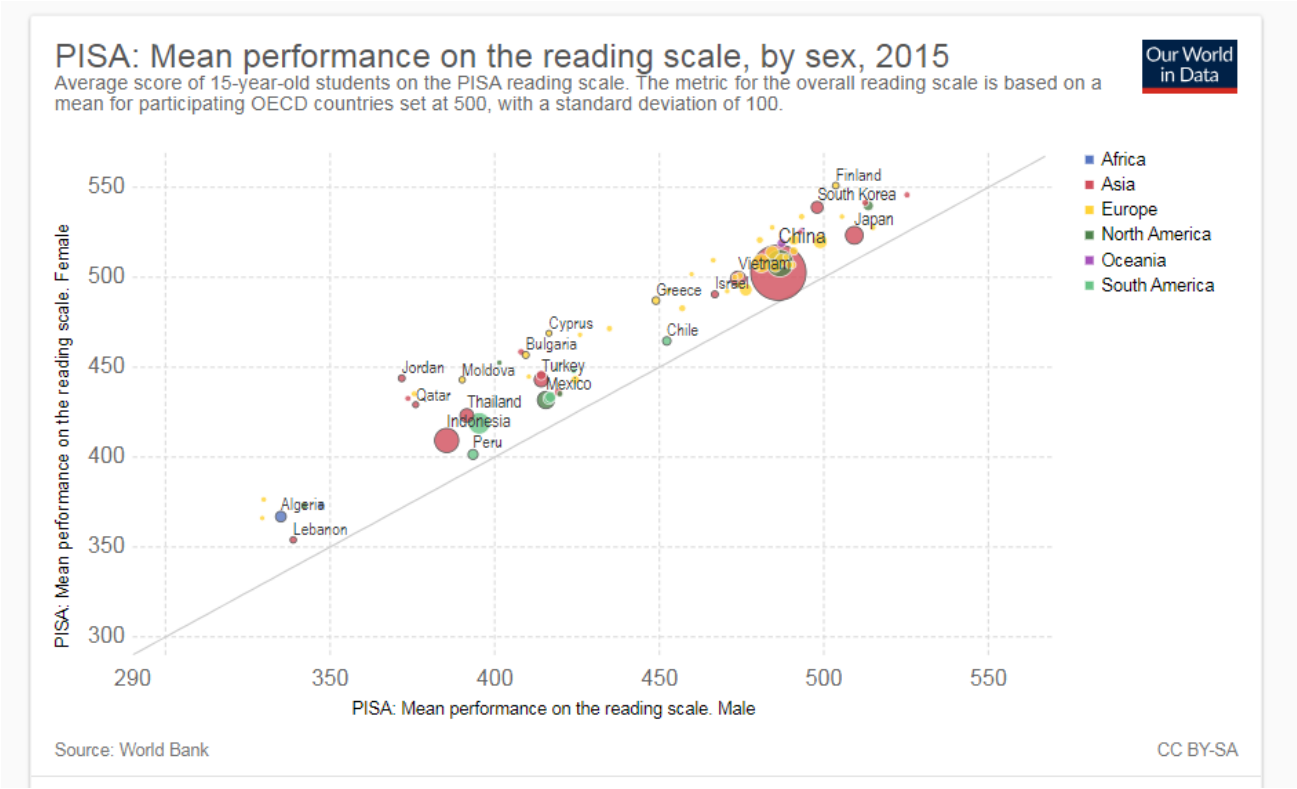


下頁兩個圖：比較男女

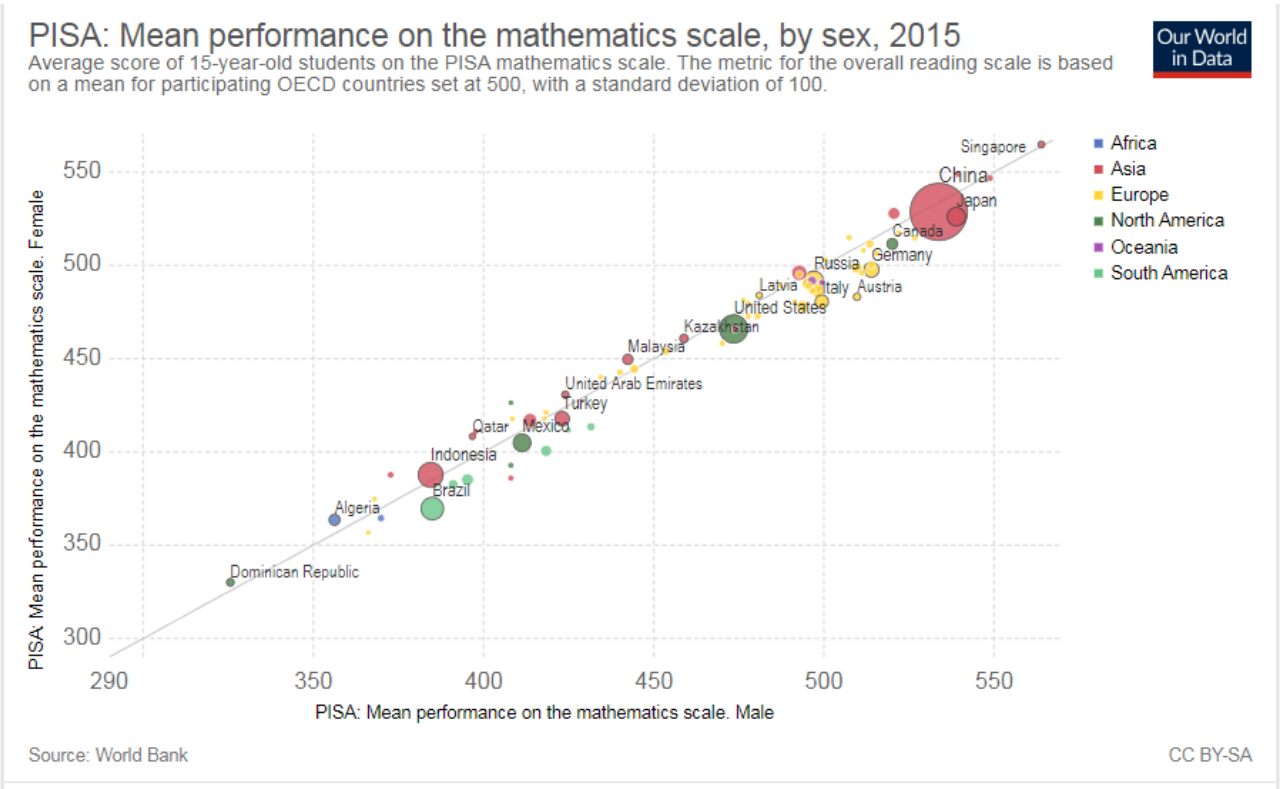
1. Reading: 每個國家都是女明顯優於男

2. Math:超過一半國家男生數學略優於女，但實質差異較不明顯 **2015 Plot**

Reading: male vs. female（45 度線之左：女 > 男）



2015 Plot Math: male vs. female（45 度線之右 男 > 女）



R 補充:

First you need to install R and RStudio

```
> read.csv(file="Data_Ice_Cream.csv",header=TRUE,sep=",") -> icecream
```

The imported file must be located in the working directory

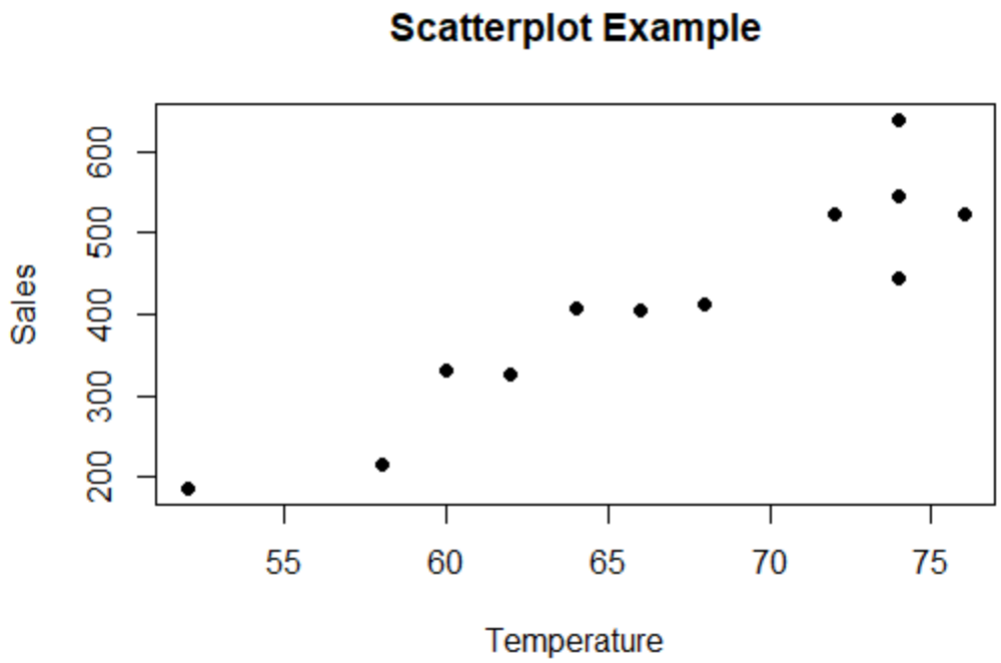
If you don't know where it is, type

```
> getwd()
[1] "C:/Users/Weijing Wang/Documents"
> icecream # display the whole data
   嚜燙ale Temperature
1      185          52
2      215          58
3      332          60
4      325          62
5      408          64
6      406          66
7      412          68
8      522          72
9      445          74
10     545          74
11     640          74
12     522          76
 > names(icecream)
[1] "嚜燙ale"     "Temperature"
```

發現出現亂碼 → 手動改掉

```
> names(icecream)[1] <- "Sales"
> names(icecream)
[1] "Sales"       "Temperature"
> head(icecream) # display the first few lines of the data
  Sales Temperature
1   185          52
2   215          58
3   332          60
4   325          62
5   408          64
6   406          66
```

```
> plot(icecream$Temperature, icecream$Sales, main="Scatterplot
Example",
+      xlab="Temperature", ylab="Sales", pch=19)
```




```
> cor(icecream$Temperature,  icecream$Sales, method="pearson")
[1] 0.9271701
> cor(icecream$Temperature,  icecream$Sales, method="kendall")
[1] 0.8125992
> cor(icecream$Temperature,  icecream$Sales, method="spearman")
[1] 0.9241752
```

Example: Learn how to create variables by yourself

```
> x < c(1,2,3,4,5)
[1] FALSE FALSE FALSE FALSE FALSE
> x <- c(1,2,3,4,5)
> y <- c(1,3,5,2,6)
> cor(x,y,method="pearson")
[1] 0.6862436
> cor(x,y,method="kendall")
[1] 0.6
> cor(x,y,method="spearman")
[1] 0.7
```