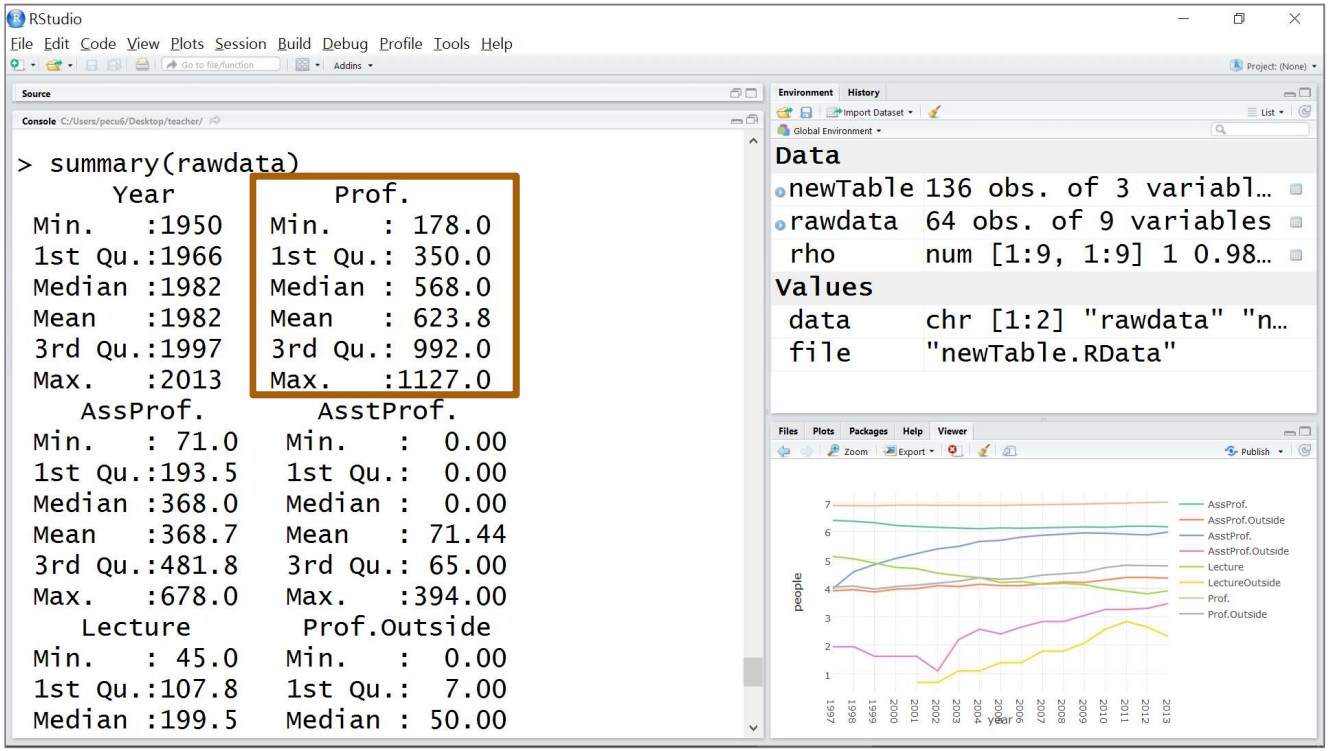
國立臺灣大學
National Taiwan University

# 雙變數分析

國立臺灣大學共同教育中心

蔡芸琤

# 統計敘述


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console C:/Users/pecu6/Desktop/teacher/
> summary(rawdata)
Year Prof.
Min. :1950 Min. : 178.0
1st Qu.:1966 1st Qu.: 350.0
Median :1982 Median : 568.0
Mean :1982 Mean : 623.8
3rd Qu.:1997 3rd Qu.: 992.0
Max. :2013 Max. :1127.0
AssProf. AsstProf.
Min. : 71.0 Min. : 0.00
1st Qu.:193.5 1st Qu.: 0.00
Median :368.0 Median : 0.00
Mean :368.7 Mean : 71.44
3rd Qu.:481.8 3rd Qu.: 65.00
Max. :678.0 Max. :394.00
Lecture Prof.Outside
Min. : 45.0 Min. : 0.00
1st Qu.:107.8 1st Qu.: 7.00
Median :199.5 Median : 50.00
Environment History
Import Dataset
Global Environment
Data
newTable 136 obs. of 3 variabl…
rawdata 64 obs. of 9 variables
rho num [1:9, 1:9] 1 0.98…
Values
data chr [1:2] "rawdata" "n…
file "newTable.RData"
Files Plots Packages Help Viewer
Zoom Export Publish
people
year
AssProf.
AssProf.Outside
AsstProf.
AsstProf.Outside
Lecture
LectureOutside
Prof.
Prof.Outside

# 設計問題

1. 哪一年發生了1127 位正教授的高峰？ (有意義？)
2. 哪一年增聘了最多教授？
3. 增聘最多教授的那一年，學校的經費預算是否有增加？
4. 哪一年減少了最多教授？
5. …

# 討論變數間的相關性，但不包含因果關係

1. 目的：主要衡量兩變數間線性關聯性的高低程度
2. 方法：相關係數
3. 公式說明： http://wiki.mbalib.com/zh-tw/%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0
4. R 語言： https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.html

# 相關係數分析

變項間的相關程度高或低，得到的相關係數只能說明這兩個變項間是正相關、負相關，或者是無關。

相關程度之高低，在正負0.3之間（即0.3至-0.3之間）稱為低度相關；在正負0.3-0.6之間（即指介於0.3至0.6，-0.3至-0.6之間）稱為中度相關；而在正負0.6至0.9之間（即指在0.6至0.9，-0.6至-0.9之間）則稱為高度相關；若是為正負1，即表示完全相關；若是為0，即表示無關。

| 基本函數 | 意義 |
| --- | --- |
| cor() | 計算相關係數 |
| cor.test() | 相關係數分析 |

# 討論變數間的相關性，但不包含因果關係

1. 相關係數值介於 -1 至 1 之間。
2. 相關係數值 =-1：兩變數為完全負相關。
3. -1< 相關係數值 <0：兩變數為負相關。
4. 相關係數值 =0：兩變數為無相關。
5. 0< 相關係數值 <1：兩變數為正相關。
6. 相關係數值 =1：兩變數為完全正相關。

# 數據閱讀

RStudio Source Editor

rho

Filter

| | Year | Prof. | AssProf. | AsstProf. | Lecture | Prof.Outside | AssProf.Outside | AsstProf.Outside | LectureOutside |
|---|---|---|---|---|---|---|---|---|---|
| Year | 1.0000000 | 0.9857888 | 0.84621975 | 0.7485752 | -0.31570368 | 0.9556784 | 0.9755560 | 0.7841485 | 0.6041368 |
| Prof. | 0.9857888 | 1.0000000 | 0.84640694 | 0.7411748 | -0.39576535 | 0.9239644 | 0.9536310 | 0.7261133 | 0.5783464 |
| AssProf. | 0.8462198 | 0.8464069 | 1.00000000 | 0.3132312 | 0.05812416 | 0.7656979 | 0.8070489 | 0.6679175 | 0.2259859 |
| AsstProf. | 0.7485752 | 0.7411748 | 0.31323116 | 1.0000000 | -0.75618717 | 0.7755827 | 0.7411235 | 0.6298275 | 0.8568931 |
| Lecture | -0.3157037 | -0.3957653 | 0.05812416 | -0.7561872 | 1.00000000 | -0.3406727 | -0.3179593 | -0.1832406 | -0.6259633 |
| Prof.Outside | 0.9556784 | 0.9239644 | 0.76569787 | 0.7755827 | -0.34067270 | 1.0000000 | 0.9838035 | 0.8960301 | 0.7329668 |
| AssProf.Outside | 0.9755560 | 0.9536310 | 0.80704889 | 0.7411235 | -0.31795927 | 0.9838035 | 1.0000000 | 0.8141089 | 0.6370524 |
| AsstProf.Outside | 0.7841485 | 0.7261133 | 0.66791754 | 0.6298275 | -0.18324064 | 0.8960301 | 0.8141089 | 1.0000000 | 0.7170201 |
| LectureOutside | 0.6041368 | 0.5783464 | 0.22598586 | 0.8568931 | -0.62596333 | 0.7329668 | 0.6370524 | 0.7170201 | 1.0000000 |

Showing 1 to 9 of 9 entries

# 數據閱讀 (每年師資變化情況)

RStudio Source Editor

rhodiff

Filter

| | Prof. | AssProf. | AsstProf. | Lecture | Prof.Outside | AssProf.Outside | AsstProf.Outside | LectureOutside |
|---|---|---|---|---|---|---|---|---|
| Prof. | 1.00000000 | 0.182781012 | -0.16609075 | -0.08968256 | 0.04257345 | 0.03919189 | 0.04559873 | -0.030178343 |
| AssProf. | 0.18278101 | 1.000000000 | -0.58656357 | 0.24894658 | -0.03096503 | -0.20260016 | 0.17895141 | -0.008136201 |
| AsstProf. | -0.16609075 | -0.586563569 | 1.00000000 | -0.35470678 | 0.02644027 | 0.02561974 | 0.06515885 | -0.101601231 |
| Lecture | -0.08968256 | 0.248946575 | -0.35470678 | 1.00000000 | 0.12750541 | 0.02817573 | 0.25332165 | -0.126145720 |
| Prof.Outside | 0.04257345 | -0.030965034 | 0.02644027 | 0.12750541 | 1.00000000 | 0.76558914 | 0.66250191 | 0.516472720 |
| AssProf.Outside | 0.03919189 | -0.202600155 | 0.02561974 | 0.02817573 | 0.76558914 | 1.00000000 | 0.09987466 | 0.314732145 |
| AsstProf.Outside | 0.04559873 | 0.178951408 | 0.06515885 | 0.25332165 | 0.66250191 | 0.09987466 | 1.00000000 | 0.088347644 |
| LectureOutside | -0.03017834 | -0.008136201 | -0.10160123 | -0.12614572 | 0.51647272 | 0.31473214 | 0.08834764 | 1.000000000 |

Showing 1 to 8 of 8 entries

# 數據閱讀

引起動機：http://news.tvbs.com.tw/life/652591

學生希望師資能再提升，但台灣真的能吸引好的師資嗎？根據了解以鄰近亞洲國家開出的薪資條件，日本、香港都是台灣的2到3倍，新加坡至少3倍起跳，而大陸通常是4到5倍，而這還不包括研究經費，而原本政府編列5年5百億元的預算，台大能分到31億元，但預算卻一直刪減，現在只剩下16億元，讓台灣大學副校長同時也是準教育部次長的陳良基坦言，台灣沒錢也沒條件跟人家搶人才。

Q：臺大師資的吸引力如何？ (少了和其他學校的比較)

1. Prof. 與 Year 的相關係數最高，代表正教授是逐年應聘人數增加的趨勢
2. Prof. 與 Lecture 變化的相關係數是負值，代表正教授增聘時，講師的聘任減少
3. Prof. 與 Lecture Outside 變化的相關係數是負值，代表正教授增聘時，合聘講師減少

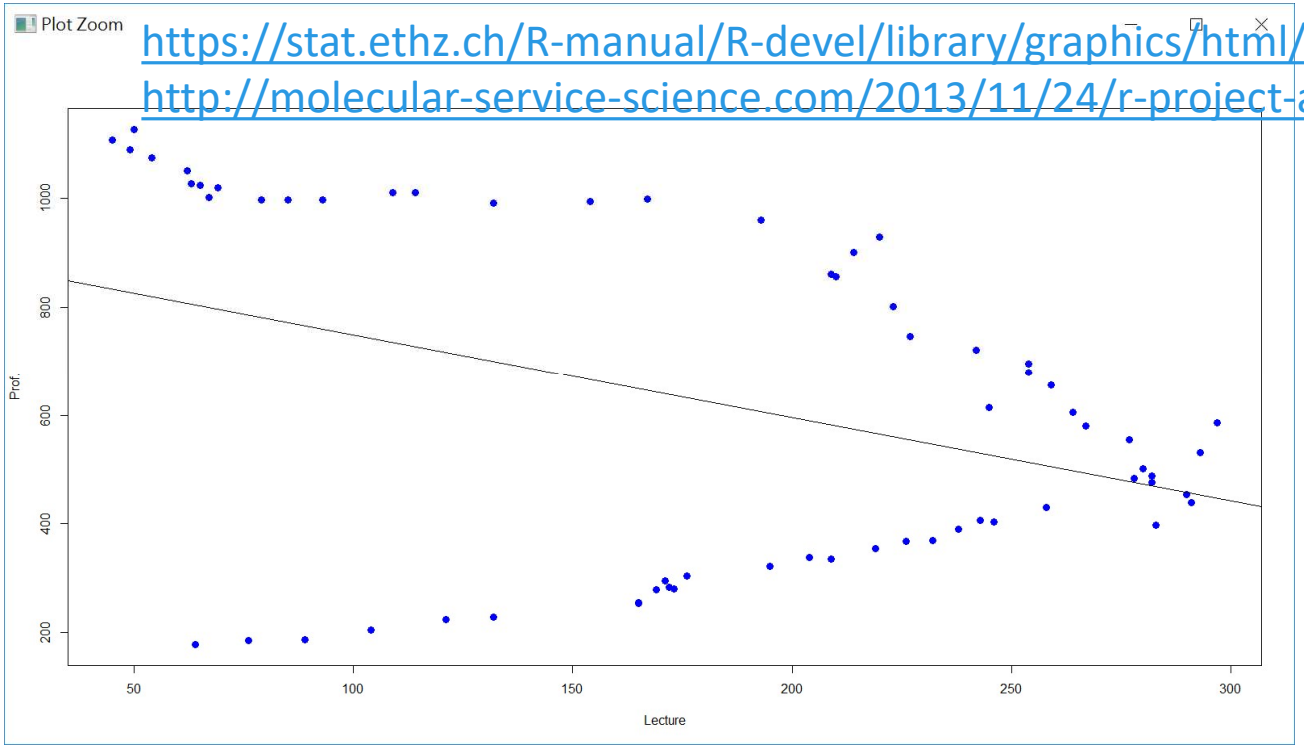
# 討論變數間的相關性，包含因果關係

1. 目的：
   1) 解釋資料過去的現象
   2) 由自變數來預測依變數未來可能產生之數值。
   3) 簡單線性迴歸分析是用一直線來解釋一個自變數 (因, x) 與一個依變數 (果, y) 的關係。
   4) 例如，利率的變化影響股價的漲跌，股價即為依變數，而利率就是自變數。利率的變動是因，股價的波動為果。
2. R 語言： https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html
3. 參考資料：http://molecular-service-science.com/2012/09/12/statistics-regression/

# 回帰分析

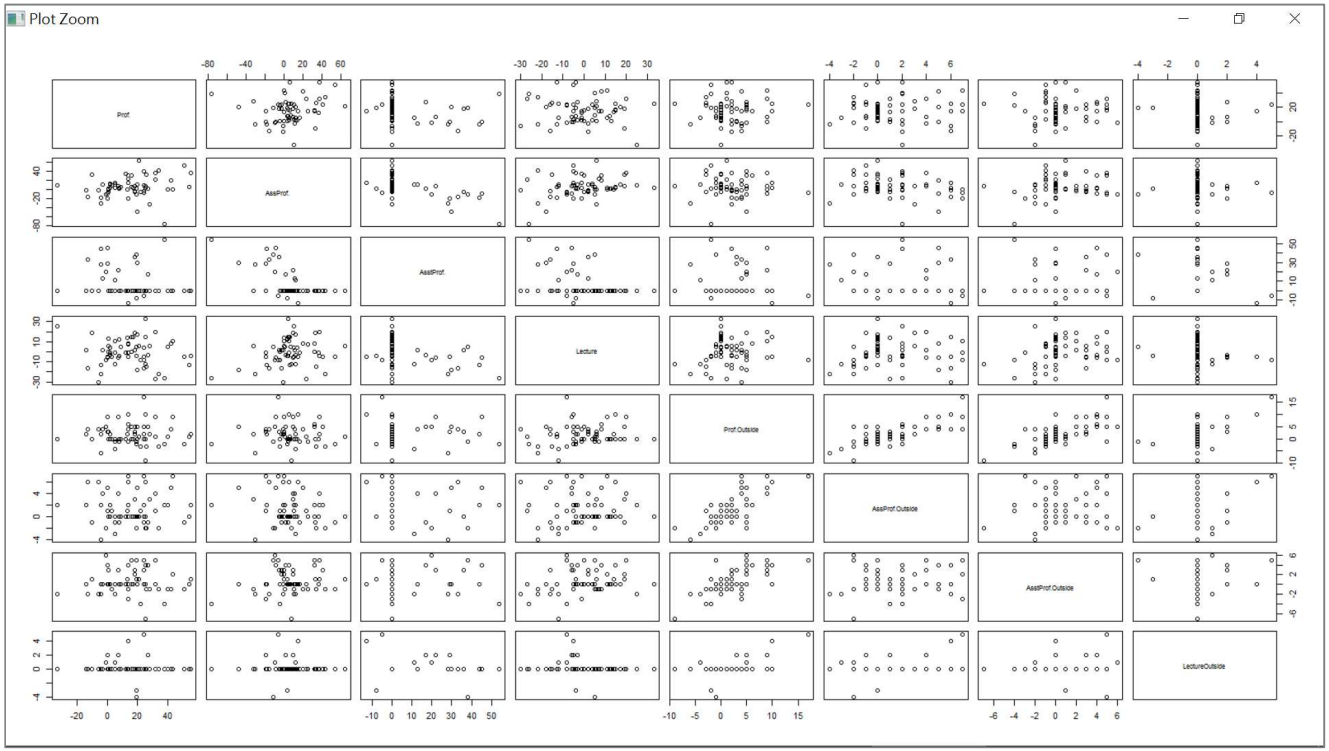
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax}\left( \begin{matrix} W_{1,1}x_1 + W_{1,2}x_1 + W_{1,3}x_1 + b_1 \\ W_{2,1}x_2 + W_{2,2}x_2 + W_{2,3}x_2 + b_2 \\ W_{3,1}x_3 + W_{3,2}x_3 + W_{3,3}x_3 + b_3 \end{matrix} \right)$$

# 繪圖每年正教授與講師人數的關係




https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/abline.html
http://molecular-service-science.com/2013/11/24/r-project-abline-curve-polygon/

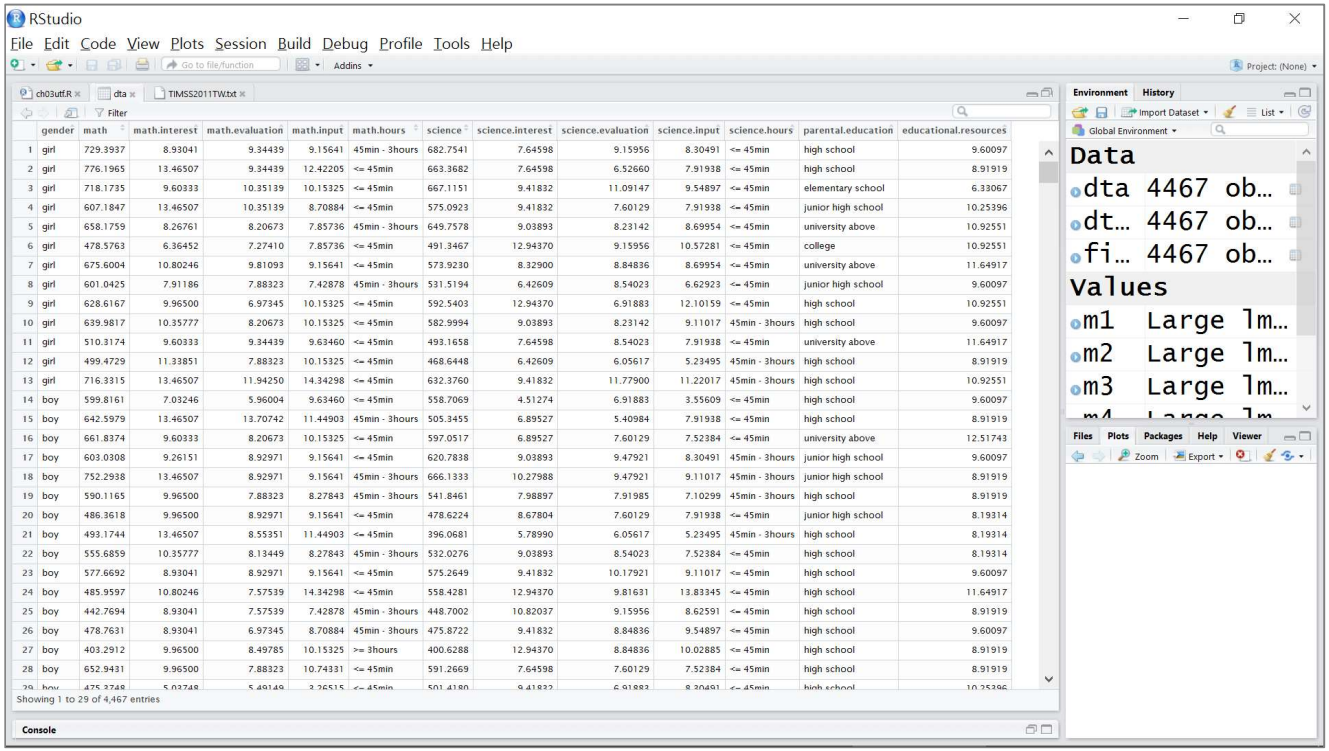
# 繪圖教職變化的關係


Plot Zoom
Prof
AssProf
AssProf
Lecture
Prof.Outside
AssProf.Outside
AssProf.Outside
LectureOutside

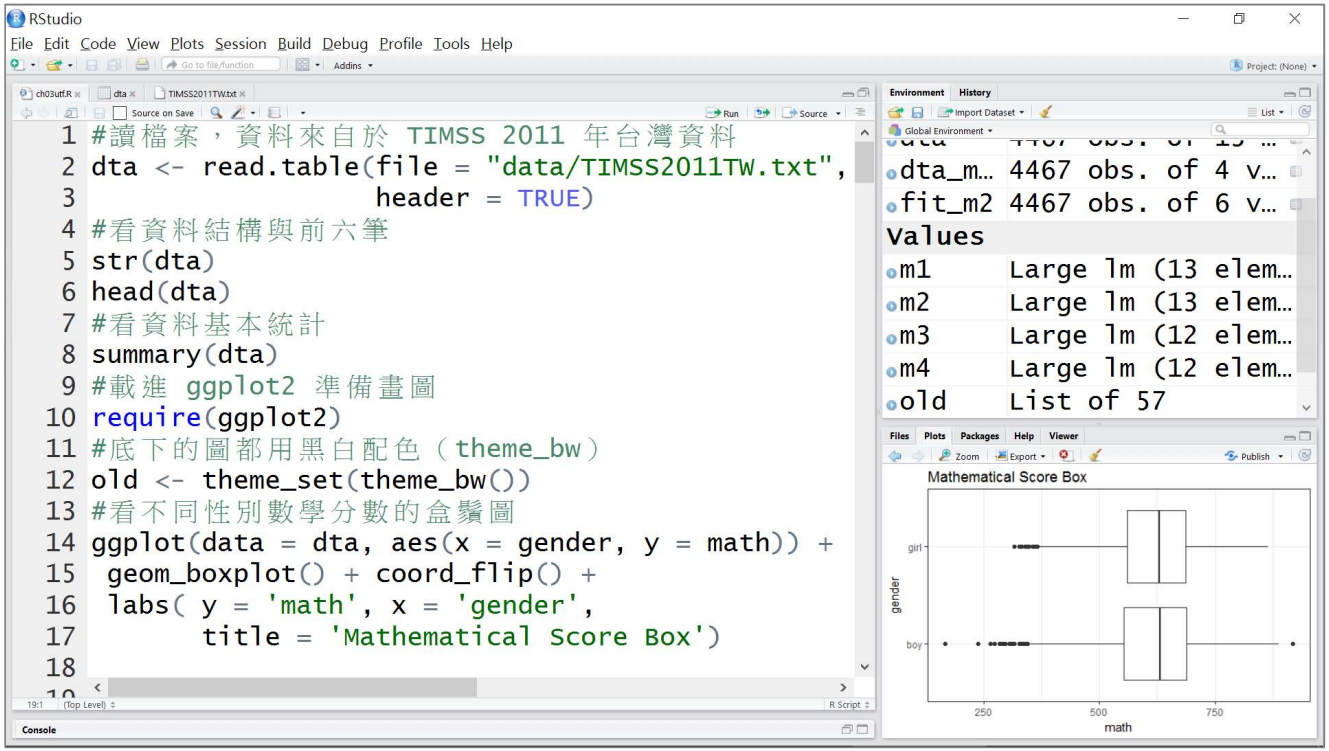
# TIMSS & PIRLS International Study Center

1. 國際數學與科學教育成就趨勢調查
2. 收集台灣 2011 年八年級學生問卷資料
3. 以數學能力為依變數 (果, y)，以性別、數學投入、數學興趣、教育資源與父母教育程度為自變數 (因, x)
4. http://myweb.ncku.edu.tw/~cpcheng/Rbook/03/data/TIMSS2011TW.txt

# TIMSS & PIRLS International Study Center

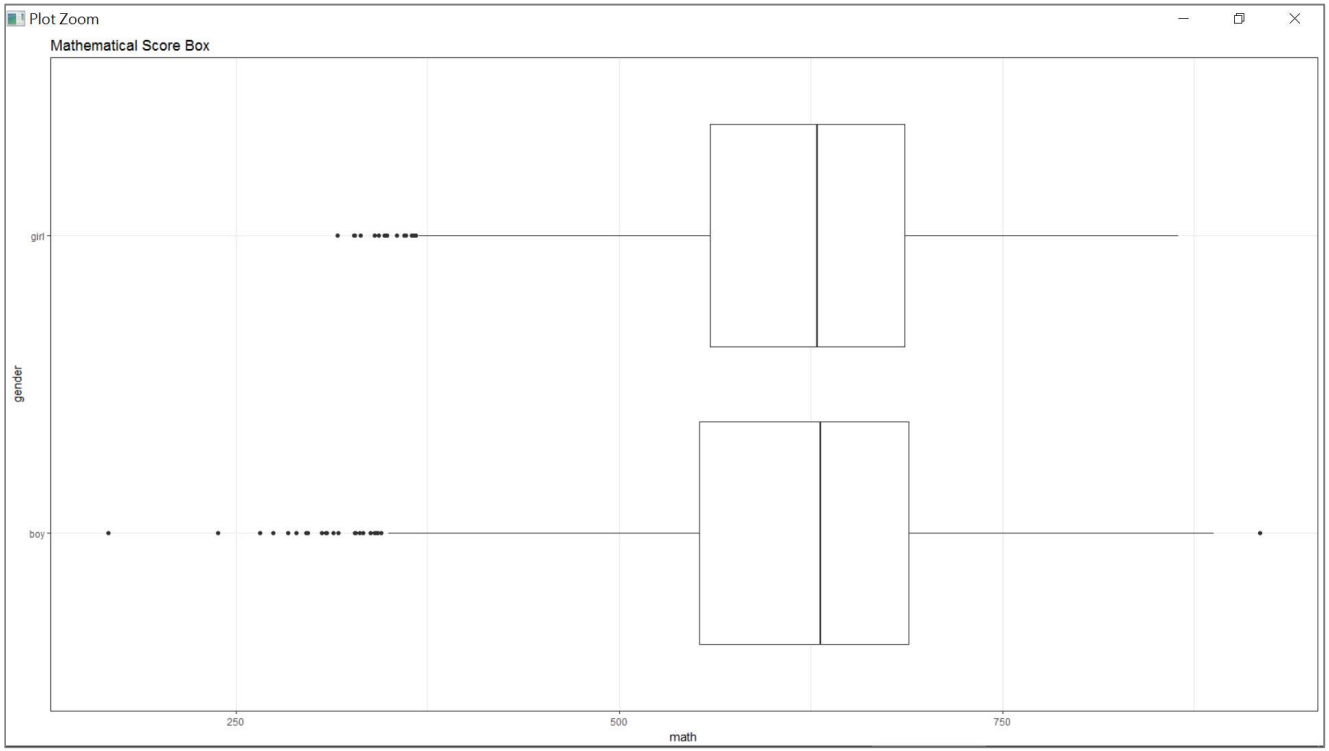


# TIMSS & PIRLS International Study Center




**如果要改成水平方向，可以使用：coord_flip()。**

# TIMSS & PIRLS International Study Center



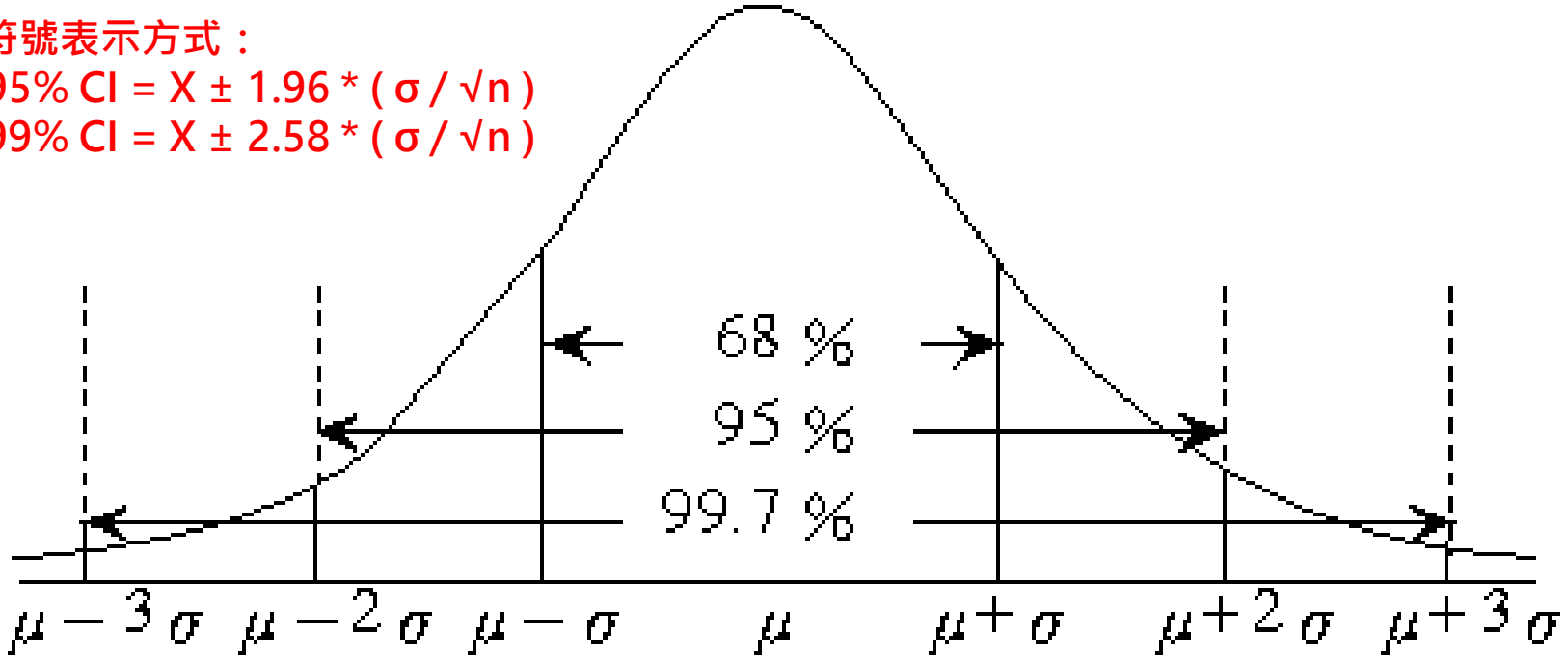

**數學分數因性別差異是不顯著的**

# confidence interval, CI

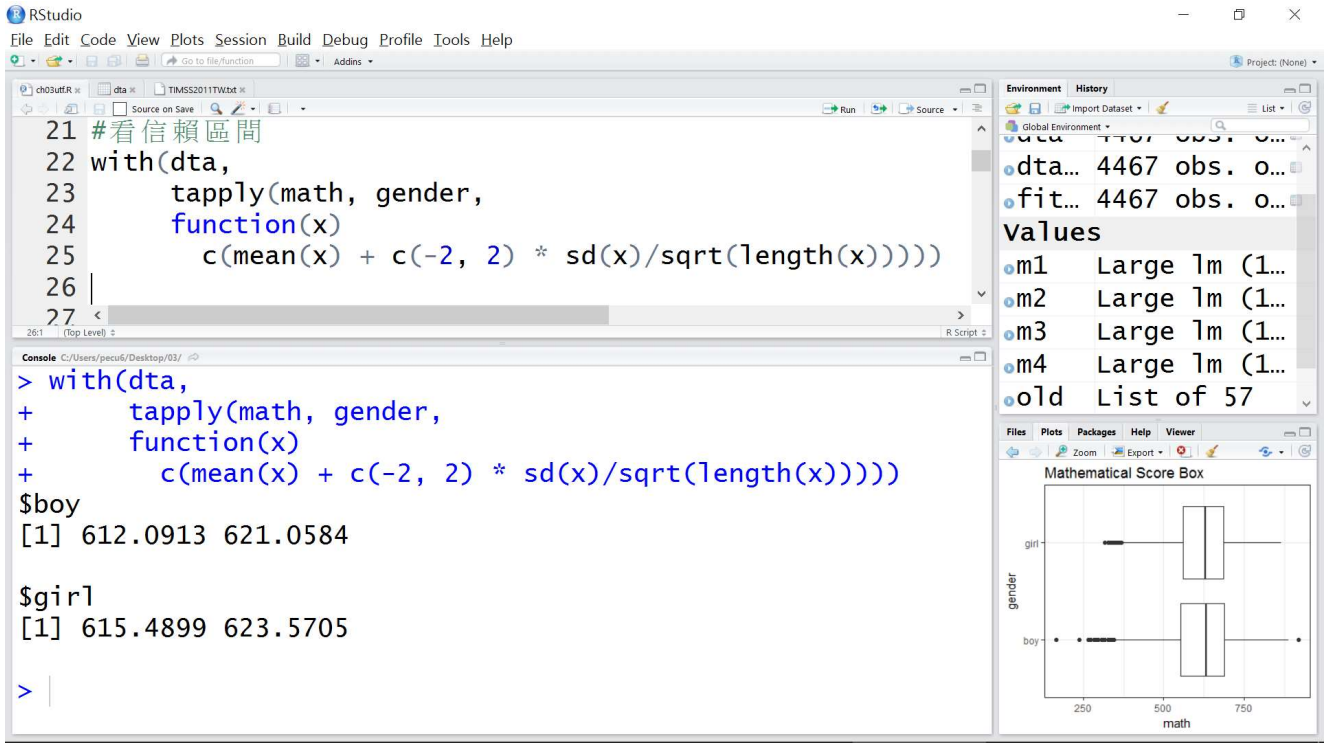
**科學符號表示方式：**

**μ 之 95% CI = X ± 1.96 * ( σ / √n )**

**μ 之 99% CI = X ± 2.58 * ( σ / √n )**

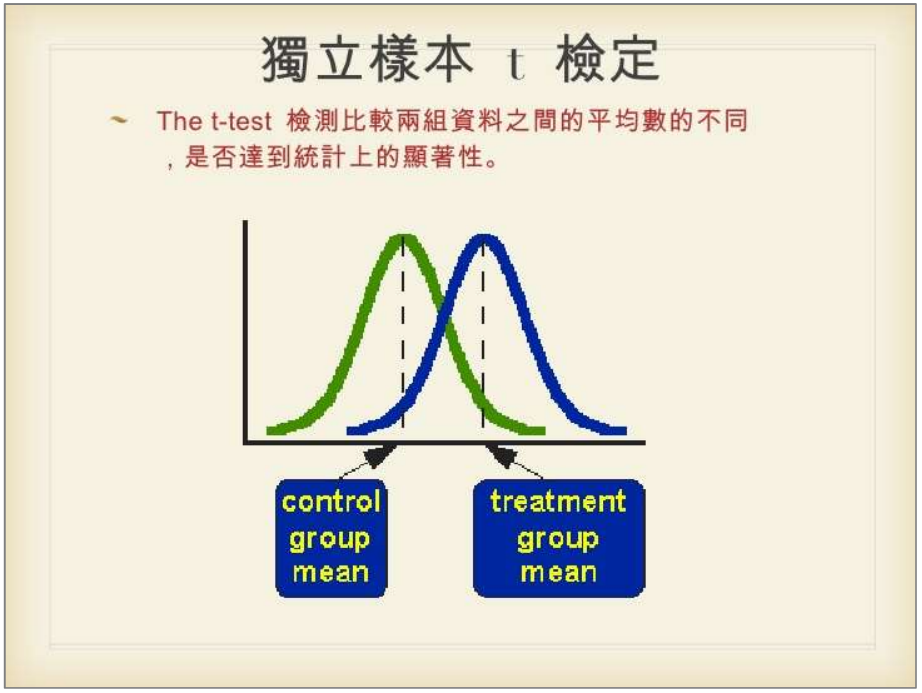


# TIMSS & PIRLS International Study Center

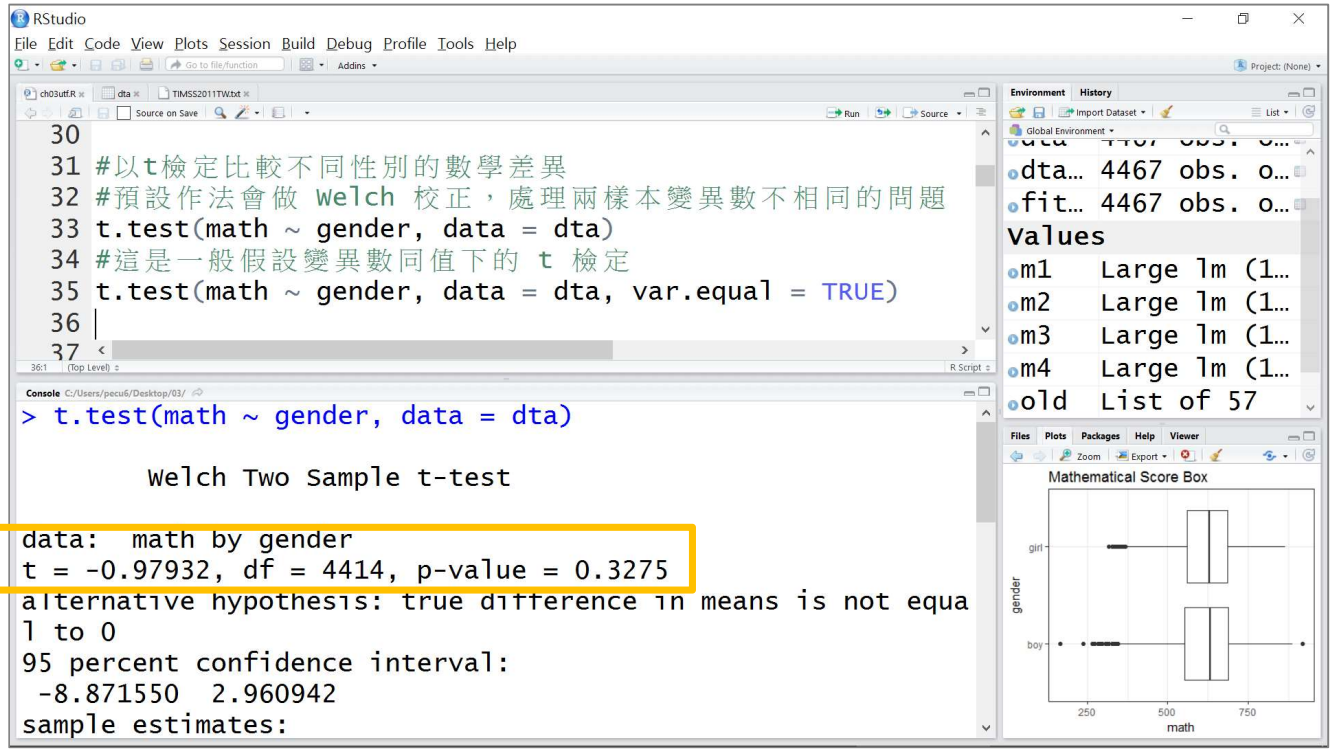


# T-Test

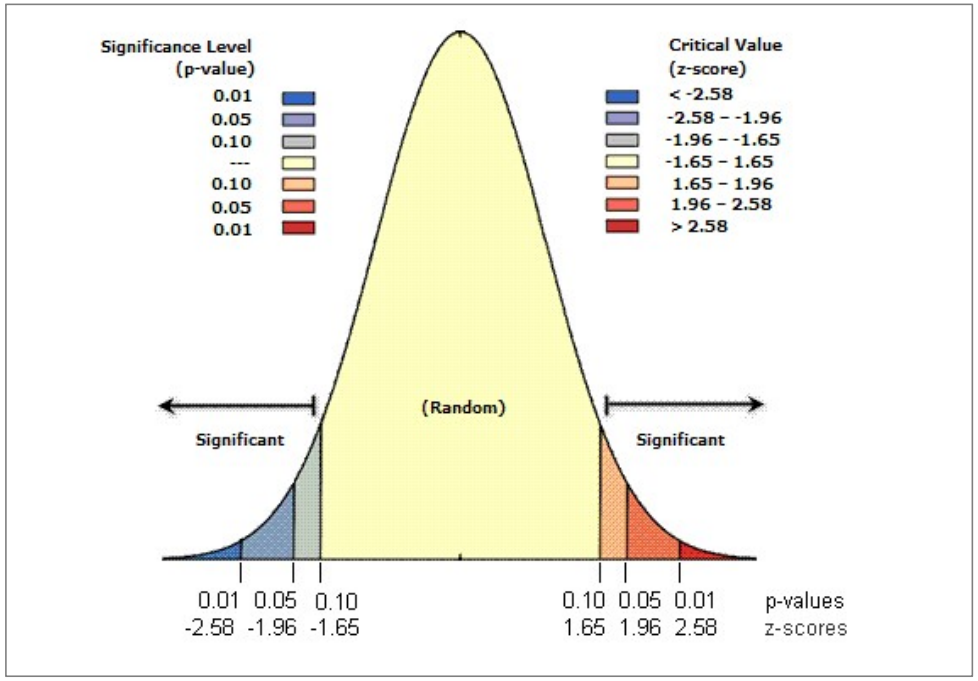
1. T-Test 是對兩樣本平均數 (mean) 差別的顯著性進行檢驗。
2. H0: **μ1 = μ2**
3. T-Test 須知道兩個總體的變異數 (Variances) 是否相等。
4. T-Test 值的計算會因變異數是否相等而有所不同。



# TIMSS & PIRLS International Study Center

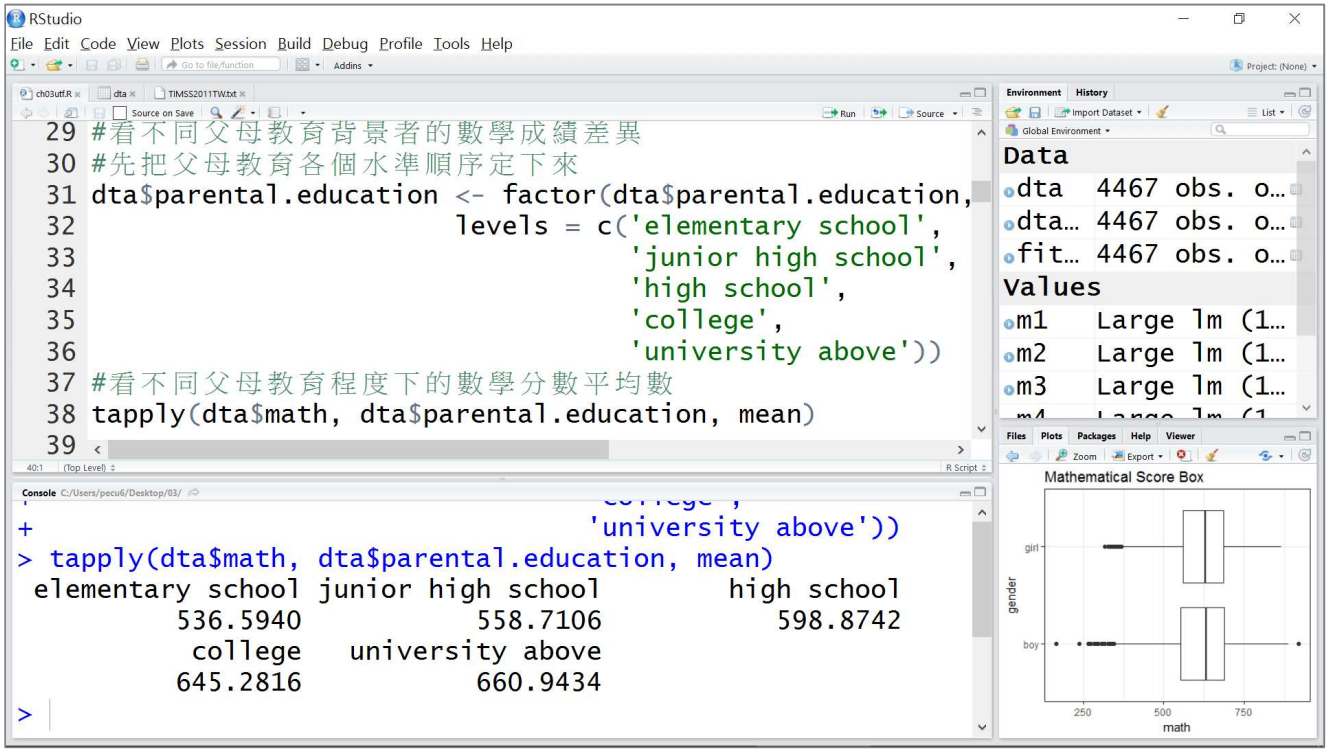


# P-Value

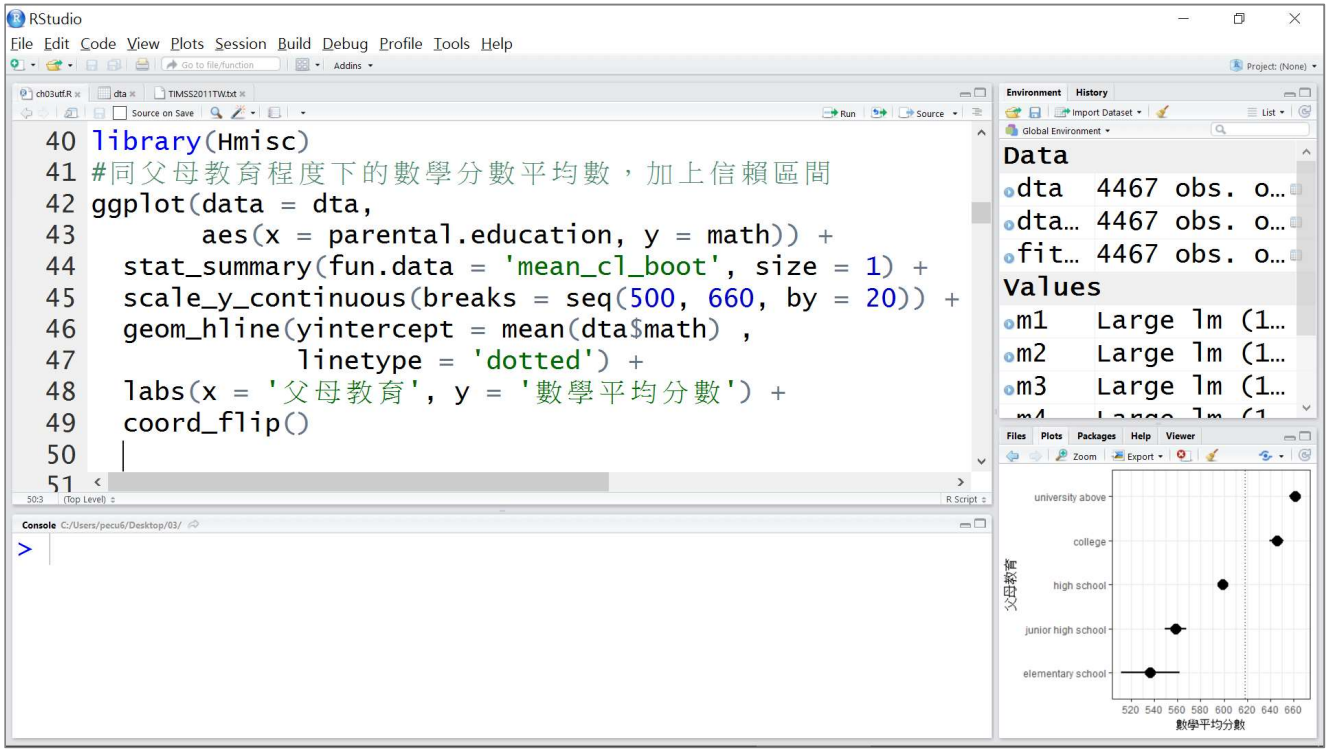



1. 因為 p-value = 0.3275，該值遠大於 1-95% = 0.05。
2. 因為 95% 信賴區間為 (-8.87155, 2.96094)。
3. 這兩者都代表無法否認虛無假設 H0。【接受】

# 父母教育程度與數學成績的關係



# ggplot2




https://www.rdocumentation.org/packages/ggplot2/versions/1.0.1/topics/geom_hline

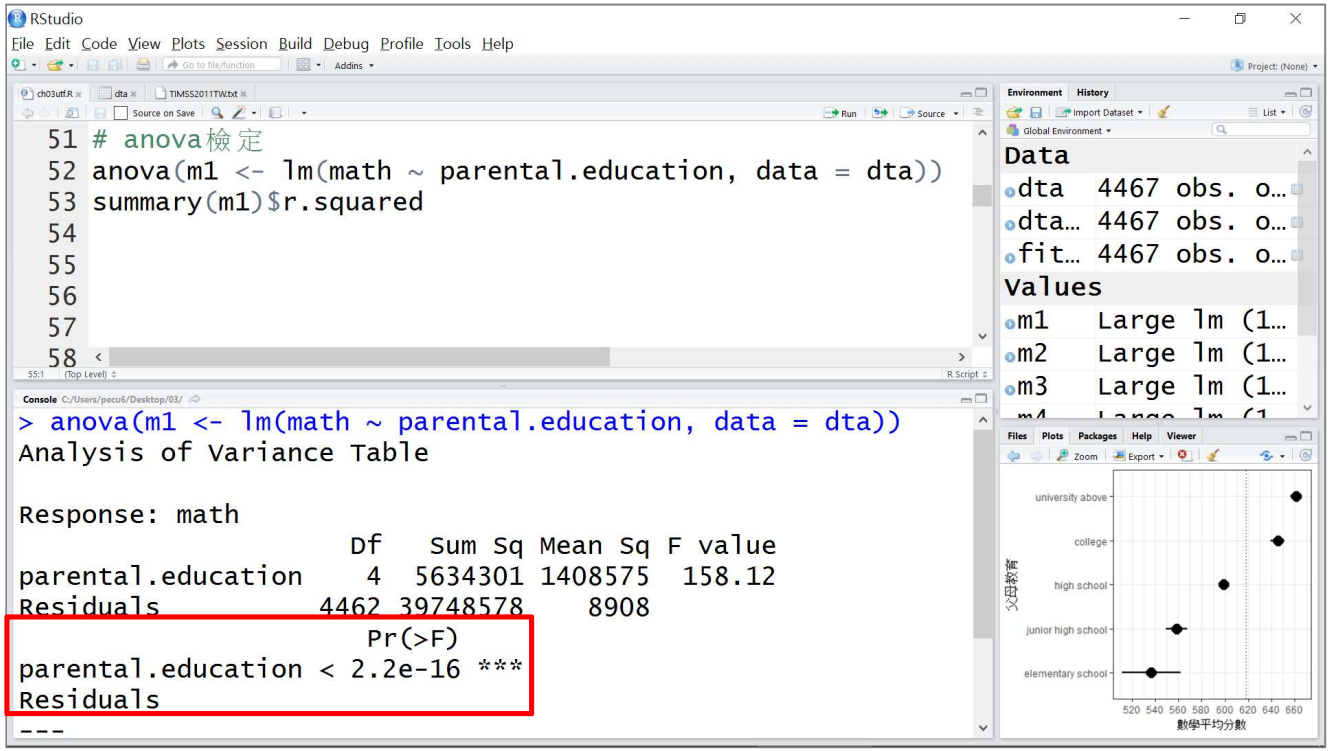
# tapply()

1. tapply() 允許根據某些變數的值，把原始資料分割為若干組。
2. 對每一組資料應用特定的操作。
3. **tapply(dta$math, dta$parental.education, mean)**
4. **tapply(dta$math, dta$parental.education, summary)**
5. 表示將 dta$math 的資料按照 dta$parental.education 的值進行分組，並將分組後資料進行 mean or summary。

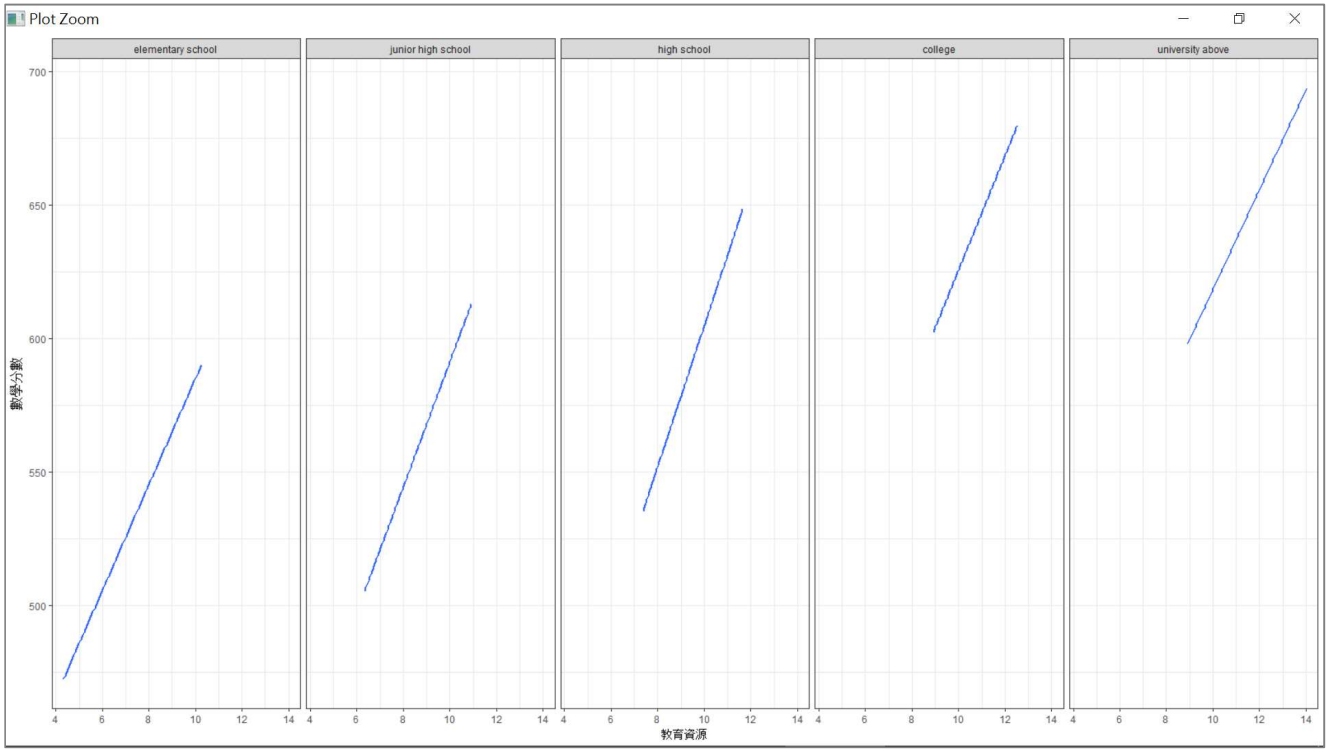
# ANOVA 分析

1. 變異數分析（Analysis of variance，簡稱ANOVA）為資料分析中常見的統計模型。
2. 主要為探討**連續型**（Continuous）資料型態之依變數 (果, y) 與**類別型**資料型態之自變數 (因, x) 的關係。

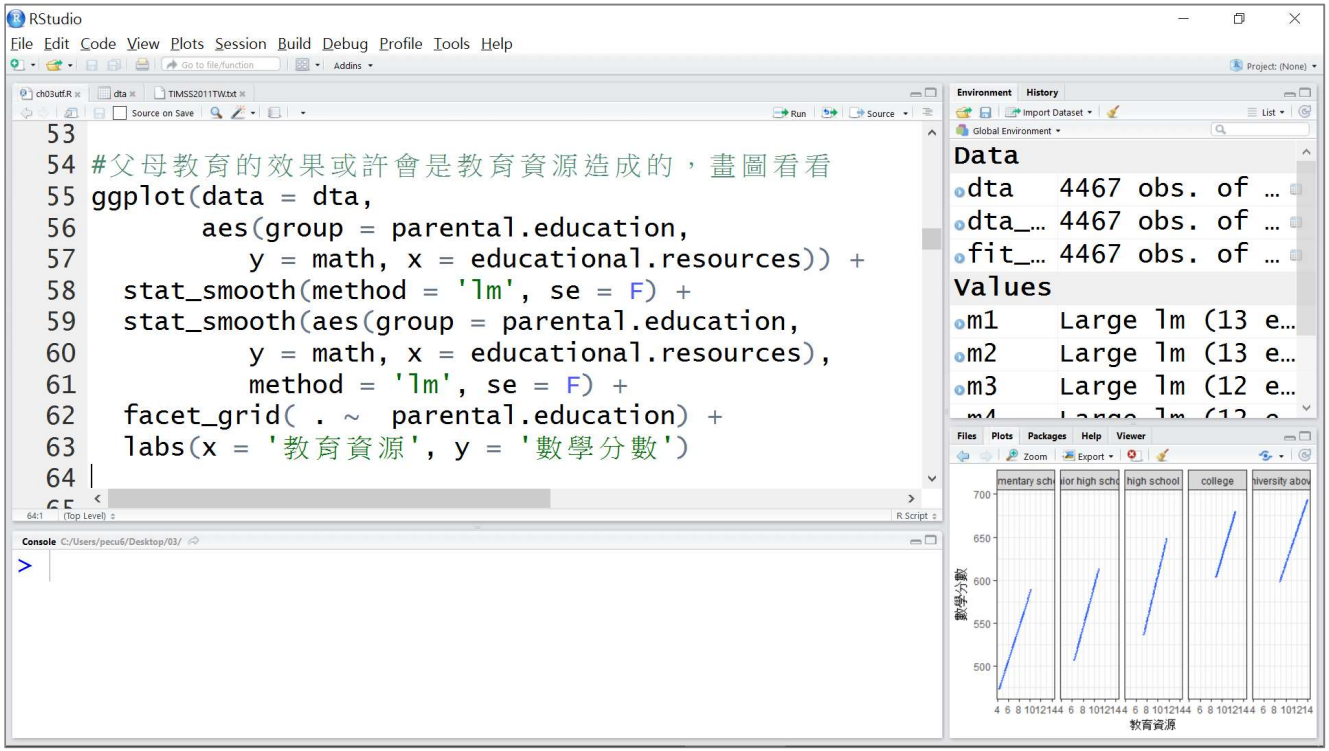
# ANOVA 分析


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
51 # anova檢定
52 anova(m1 <- lm(math ~ parental.education, data = dta))
53 summary(m1)$r.squared
54
55
56
57
58
> anova(m1 <- lm(math ~ parental.education, data = dta))
Analysis of Variance Table
Response: math
Df Sum Sq Mean Sq F value
parental.education 4 5634301 1408575 158.12
Residuals 4462 39748578 8908
Pr(>F)
parental.education < 2.2e-16 ***
Residuals
---
Environment History
Data
dta 4467 obs. o...
dta... 4467 obs. o...
fit... 4467 obs. o...
Values
m1 Large lm (1...
m2 Large lm (1...
m3 Large lm (1...
university above
college
high school
junior high school
elementary school
520 540 560 580 600 620 640 660
數學平均分數

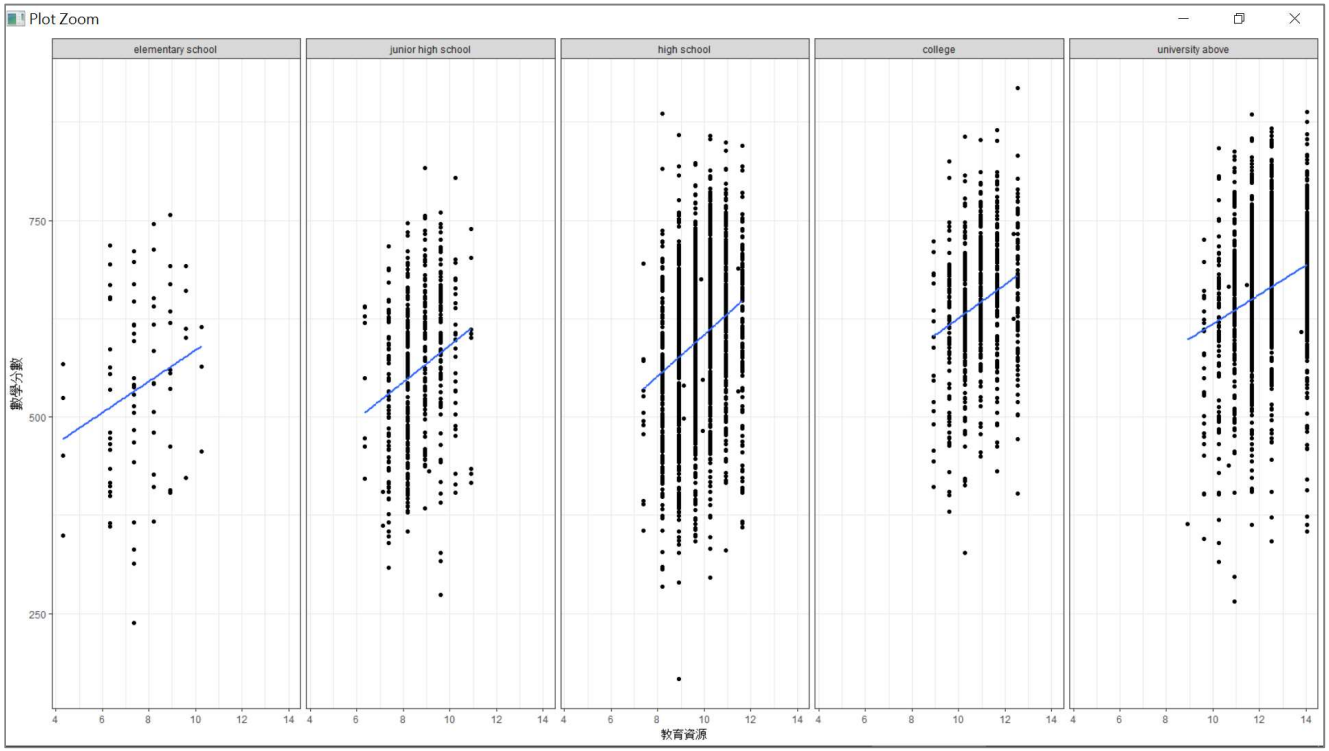
# 父母教育程度與教育資源的影響



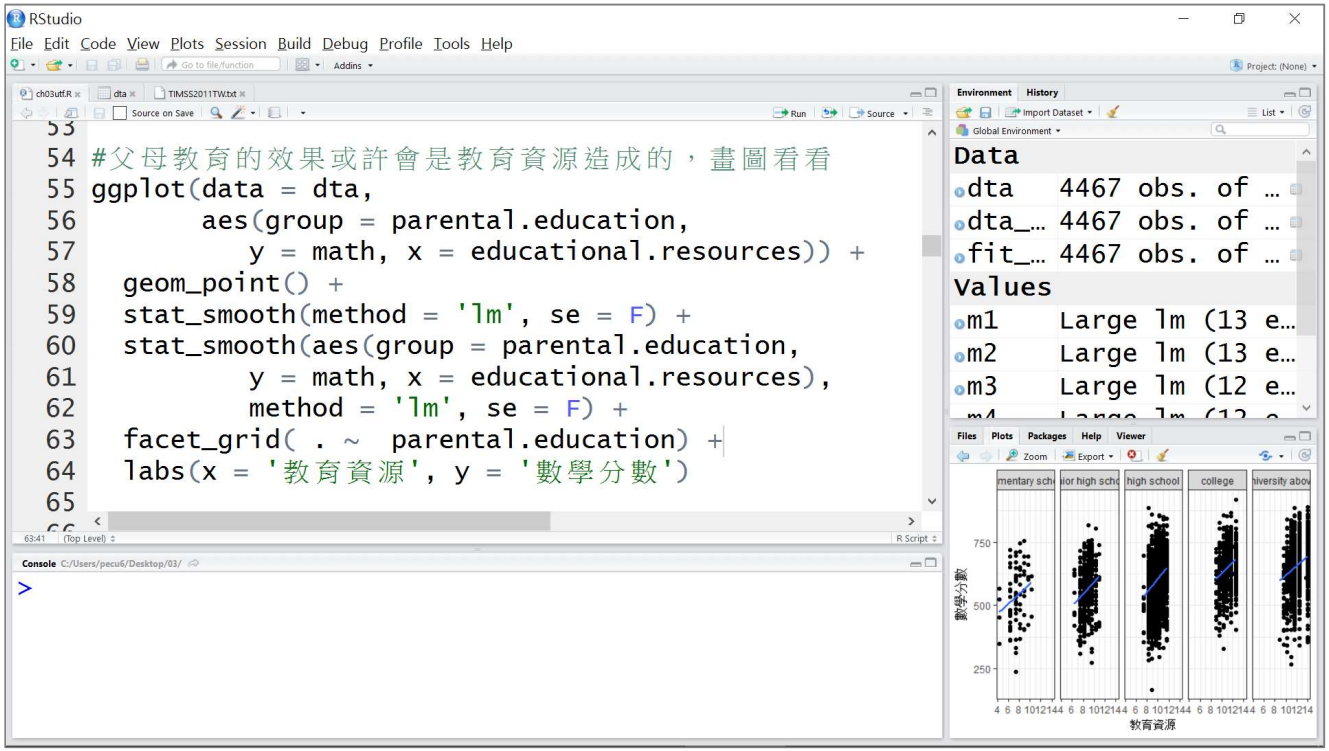
# Linear Regression



# 父母教育程度與教育資源的影響



# Linear Regression



# 課堂練習 (一小時，組內討論)

1. 每一種職位當依變數，其他職位變成自變數進行分析
2. 每一種職位的變化當依變數，其他職位的職位變化當自變數進行分析
3. 各組一起使用自己組內的資料，做一組相關係數與迴歸分析
4. 分別抽四位同學上台 demo 目前的結果
5. https://rpubs.com/flowertear/224429

# HW4 請自行挑選一張表格進行雙變數分析

例如原始資料：
http://acct2013.cc.ntu.edu.tw/acct2013/acct6/6.doc

透過 R 整理資料

完成基本敘述統計分析

完成雙變數分析

指令參考表整理：
http://www3.nccu.edu.tw/~99354011/R%20commands(11.09.13).pdf

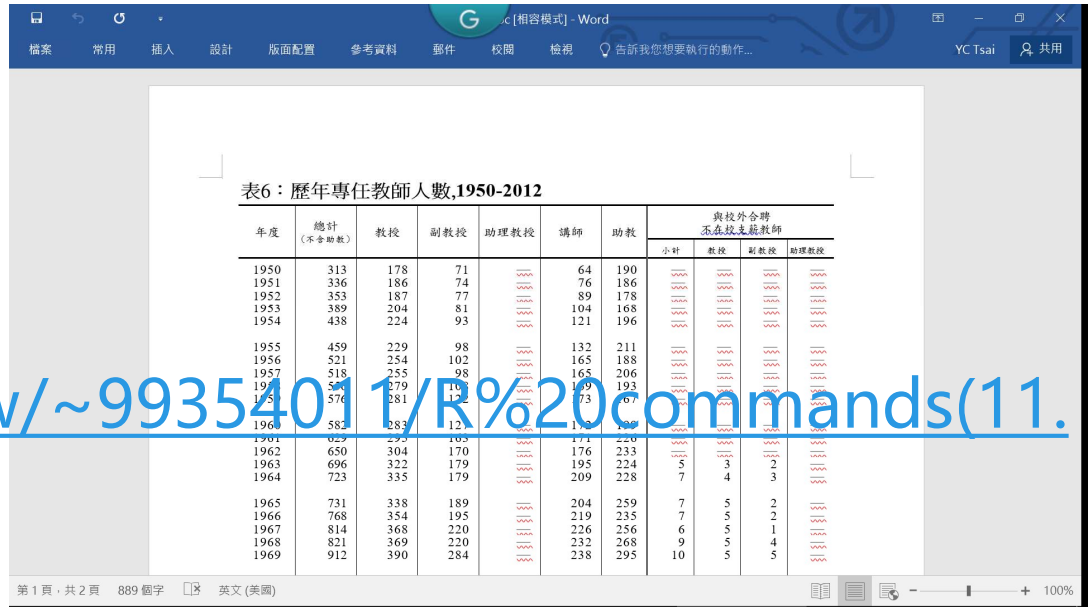


表6：歷年專任教師人數,1950-2012

| 年度 | 總計（不含助教） | 教授 | 副教授 | 助理教授 | 講師 | 助教 | 與校外合聘不在校支薪教師 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 小計 | 教授 | 副教授 | 助理教授 |
| 1950 | 313 | 178 | 71 | ... | 64 | 190 | ... | ... | ... | ... |
| 1951 | 336 | 186 | 74 | ... | 76 | 186 | ... | ... | ... | ... |
| 1952 | 353 | 187 | 77 | ... | 89 | 178 | ... | ... | ... | ... |
| 1953 | 389 | 204 | 81 | ... | 104 | 168 | ... | ... | ... | ... |
| 1954 | 438 | 224 | 93 | ... | 121 | 196 | ... | ... | ... | ... |
| 1955 | 459 | 229 | 98 | ... | 132 | 211 | ... | ... | ... | ... |
| 1956 | 521 | 254 | 102 | ... | 165 | 188 | ... | ... | ... | ... |
| 1957 | 518 | 255 | 98 | ... | 165 | 206 | ... | ... | ... | ... |
| 1958 | 558 | 279 | 107 | ... | 169 | 193 | ... | ... | ... | ... |
| 1959 | 576 | 281 | 122 | ... | 173 | 167 | ... | ... | ... | ... |
| 1960 | 582 | 283 | 123 | ... | 171 | 226 | ... | ... | ... | ... |
| 1961 | 625 | 292 | 163 | ... | 171 | 226 | ... | ... | ... | ... |
| 1962 | 650 | 304 | 170 | ... | 176 | 233 | ... | ... | ... | ... |
| 1963 | 696 | 322 | 179 | ... | 195 | 224 | 5 | 3 | 2 | ... |
| 1964 | 723 | 335 | 179 | ... | 209 | 228 | 7 | 4 | 3 | ... |
| 1965 | 731 | 338 | 189 | ... | 204 | 259 | 7 | 5 | 2 | ... |
| 1966 | 768 | 354 | 195 | ... | 219 | 235 | 7 | 5 | 2 | ... |
| 1967 | 814 | 368 | 220 | ... | 226 | 256 | 6 | 5 | 1 | ... |
| 1968 | 821 | 369 | 220 | ... | 232 | 268 | 9 | 5 | 4 | ... |
| 1969 | 912 | 390 | 284 | ... | 238 | 295 | 10 | 5 | 5 | ... |