Lecture 2 Data Analysis & Probability

09/22/2021

Topic 1: Measures of Dispersion (variation) - Revisited

a. Variance (變異數 S^2) & Standard deviation (標準差 S): two versions

$$S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / (n-1) \quad \rightarrow \text{an unbiased estimator of } Var(X) = \sigma^2$$

Note: unbiasedness → $E(S^2) = \sigma^2$

$$\tilde{S}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / n \quad \rightarrow \text{the maximum likelihood estimator of } \sigma^2$$

Note: "Maximum likelihood estimation" (MLE) satisfies some optimality property.

Remarks:

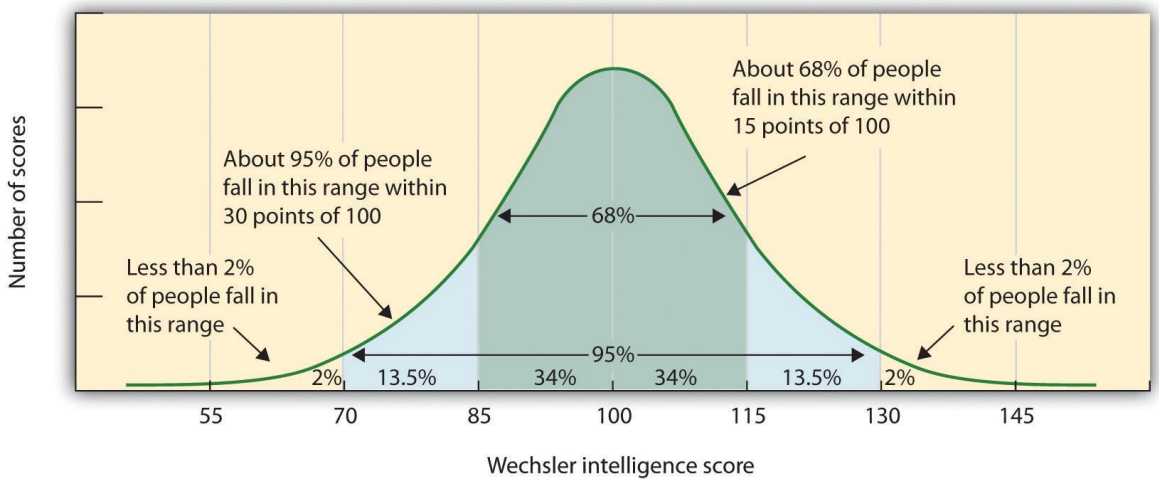
- S or $\tilde{S}$ has the same unit as X_i which makes it easier to interpret.

- Similar to $\bar{X}$, S (or $\tilde{S}$) are not a robust measure of variation.

Example: S (or $\tilde{S}$) is useful (without outliers)

IQ score ~ Normal distribution with mean = 100 and standard deviation = 15

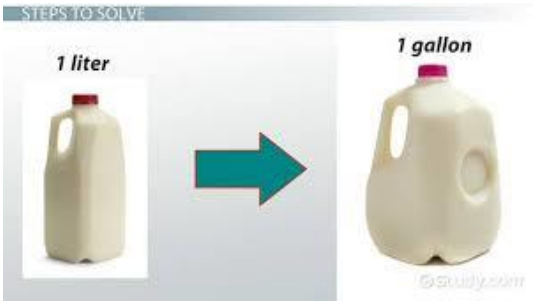
Normal → light tail (it is not likely to get extreme observations)



b. CV: Coefficient of Variation (變異係數)

$$\text{CV} = \frac{S}{\bar{X}} \times 100(\%)$$

CV is "unit less" and hence useful when direct comparison based on the standard deviation is not appropriate.



Ex: *Compare the milk consumption per family in USA and Canada*

Different units（公制 & 英制）

USA: mean = 8 gallons, S = 3 gallons → CV = 37.5

Canada: mean = 12 liters, S = 4 liters → CV = 33.3

Conclusion:

The milk consumption is USA is more variable (每家喝牛奶習慣差異性較大).

Ex: *Evaluate the effect of dog food on the weights of two pieces of dogs*

(*a*) An experiment is conducted to investigate the effect of a new dog food on weight gain in pups during the first 8 weeks of their lives. It is reported that the mean weight gain in a group of Great Dane pups is 30 pounds with a standard deviation of 10 pounds; the mean weight gain in a group of Chihuahua pups is 3 pounds with a standard deviation of 1.5 pounds. Calculate the coefficient of variation for each group. Which group exhibits the greater variability? Why is a direct comparison of standard deviations misleading here?

→ same unit but different scales

Great Dane 大麥町: mean = 30, S = 10 → CV = 10/30*100 = 33.3

Chihuahua 吉娃娃: mean = 3, S = 1.5 → CV = 1.5/3*100 = 50

(*b*) A study of weights of 2-year-old girls in Great Britain yielded a sample mean of 12.74 kilograms with a sample standard deviation of 1.60 kilograms. A similar study in the United States resulted in a sample mean of 29.2 pounds with a sample standard deviation of 2 pounds. Find the coefficient of variation for each group. Which group exhibits greater variability?

Exercise

c. Interquartile range (IQR): Q3 – Q1

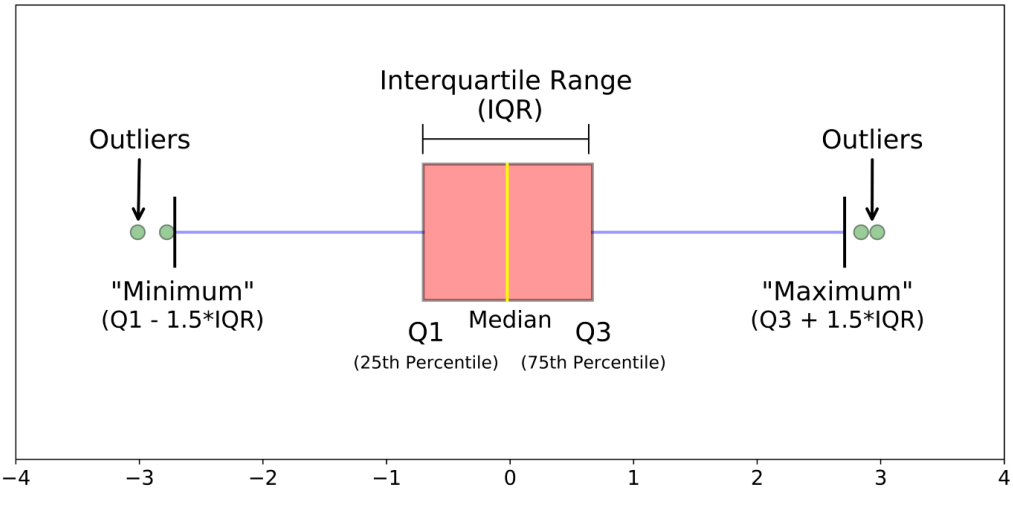
the range of the middle 50%

Definition of Outlier:

An outlier satisfies one of the following condition:

$< Q1 - 1.5IQR$ (outlier on the left)

$> Q3 + 1.5IQR$ (outlier on the right)



d.

Range = Max – Min

- In some applications such as earthquake or flood prevention, extreme values are of great concern.

Estimating maximum magnitude earthquakes in Iraq using ...

由 JN Al-Abbasi 著作 · 1985 · 被引用 31 次 — Extreme value theory is used to estimate maximum magnitude ... The application of Gumbel's theory to earthquake magnitude data dat… 14 頁

A Handbook of Extreme Value Theory and its Applications

Finally, the handbook features the practical applications and techniques and how these can be implemented in financial markets. Extreme Events in Finance: A ...

Suggestions

- 針對諸多統計指標，最重要的是了解每個指標的意義和限制在哪裏
- When you learn a statistical method, it is important to understand the purpose as well as its limitation.

Why do we care about dispersion?

* Biology → Diversity

* Economics: small mean → OK, large standard deviation → NOT OK

"不患寡而患不均"

"All with poverty do not matter, but rather inequality of wealth distribution

* Education: heterogeneity among students

* Education: discrimination of a test (考試鑑別度)

- 升高中: 會考成績最多只能是總積分的 3 分之 1，但在其他評比項目中，如在校公共服務、社團經驗、校外競賽等表現，基本上大多學生都能拿到滿分，**導致最後決定能否錄取的關鍵還是在會考成績。(免試入學=100%考試入學)**
- **大學的申請入學: 二階分數的標準差才是關鍵，**若每個人分數都一樣，標準差為 0，等於不佔權重。分數差異大，標準差大，則影響大。

學系名稱：生物科技學系 校系代碼：013202		指定項目甄試日期：107.4.14 甄試地點：博愛校區賢齊樓國際會議廳 榜示日期：107.4.26 指定項目甄試費：800 元
指定項目內容	審查資料	項目：高中(職)在校成績證明、自傳(學生自述)、讀書計畫(含申請動機)、競賽成果(或特殊表現)證明、個人資料表、其他(詳參下方說明 1)。 說明：1. 個人資料表(請到 http://exam.nctu.edu.tw/教大考生報名(學習資料表)系統填寫，填完確認送出後將 pdf 檔於甄選委員會審查資料上傳系統中上傳。 2. 其他項目說明:任何有利審查資料均可列於其他項目(無則可免)，例如資優班證明文件、社團參與、學生幹部、英文能力等證明。
	甄試說明	1. 考生必須參加團體面談，未參加者不予錄取。 2. 低收入戶考生本系補助交通費(以火車自強號等級為限)。 3. 團體面談相關訊息請參閱網站公告，查詢網址 http://life.nctu.edu.tw。
總成績計算		(國文＊1+英文＊1+數學＊1+社會＊1+自然＊1)/75＊100＊50%＋審查資料＊50%＋團體面談及認識本系＊0%

教育部希望 書面審查佔甄選總成績之比例建議以30%以上為宜

	表定比例	實際比例	勝算比	學測變異數	審查變異數
電機工程學系(甲組)	50.00%	100.00%	9999	0	4.12
電機工程學系(乙組)	50.00%	81.94%	4.54	1.15	23.75
光電工程學系	50.00%	65.38%	1.89	9.81	34.97
資訊工程學系(甲組)	60.00%	68.71%	1.46	7.25	15.53
資訊工程學系(乙組)	60.00%	86.68%	4.34	0.87	16.45
資訊工程學系(APCS組)	50.00%	41.11%	0.70	21.4	10.43
奈米科學及工程學士學位學程	50.00%	69.57%	2.29	1.64	8.58
材料科學與工程學系	44.44%	64.31%	2.25	2.17	11.02
機械工程學系	50.00%	55.63%	1.25	6.17	9.69
土木工程學系	50.00%	56.73%	1.31	4.17	7.17
生物科技學系	50.00%	60.51%	1.53	6.89	16.18

Example of Location Parameters: 學測的五標

College Entrance exam provides 5 location parameters for 5 subjects

- Horizontal from left to right: **PR 88, PR 75, PR 50, PR 25, PR 12**
- Vertical from Top to bottom: *Chinese, English, Math, Social Science, Science, Total*

107學年度學科能力測驗
總級分與各科成績標準一覽表

標準 / 項目	頂標	前標	均標	後標	底標
國文	13	12	11	9	8
英文	14	13	10	6	4
數學	12	10	6	4	3
社會	13	12	10	8	7
自然	12	10	8	6	5
總級分	63	56	45	35	28

108學年度學科能力測驗
各科成績標準一覽表

標準 / 項目	頂標	前標	均標	後標	底標
國文	13	13	11	9	8
英文	14	13	10	5	4
數學	14	12	9	5	4
社會	13	12	10	9	7
自然	13	11	8	6	5

IQR comparison (PR75-PR25)

	國	英	數	社	自
107	3	7	6	4	4
108	4	8	7	3	5

Comment: 比較兩年各科的 IQR, 會覺得 108 學年度更有鑑別度, but ???

108 學年度：五選四之亂（高分沒有用，分佈未能拉開）

數學鑑別度:在前段沒有 → 15 級分 7782 人比去年的 3700 人多了 4 千多人

在後段有 → 均標 – 後標 = 4 級分（較 107 學年拉長了 2 級分）

國文頂標 = 前標 → PR88 = PR75 → 13% 同分

原文網址: 去年數學太簡單大考中心主任下台！今年難到崩潰滿級分砍 1.3 萬人 | ETtoday 生活新聞 | ETtoday 新聞雲 記者謝承恩／台北報導

大學學測成績公布，數學滿級分人數僅 1558 人，比往年少了 1.3 萬人。而去年因為數學考題太簡單，造成數學滿級分人數暴增「嚴重超篩」，大考中心主任張茂桂下台請辭，今年考題難度暴增，考生都哀鴻遍野，連全國教師會解題團隊都形容，今年學測數學科是「史上最難」。

今年學測考試結束後，考生一面倒認為數學試卷相當難，成績結果公布後，果然滿級分人數砍半再砍半，相較去年的 14489 人，今年僅 1558 人，足足少了 1.3 萬。不禁讓人聯想起，去年因為數學考題太過簡單，缺乏鑑別度，相較前年滿級分人數 7782 人，去年躍升至 14489 人，此情況造成「嚴重超篩」，也讓大考中心主任張茂桂請辭下台，藍委甚至要求教育部提檢討報告。

去年大考中心坦言數學只有 1 題有難度，而今年考生吃足苦頭，直接遇上艱深考題，全國高級中等學校教育產業工會也指出，今年學測試題較去年偏難，試題取向類似指考的數甲，會讓社會組學生崩潰。

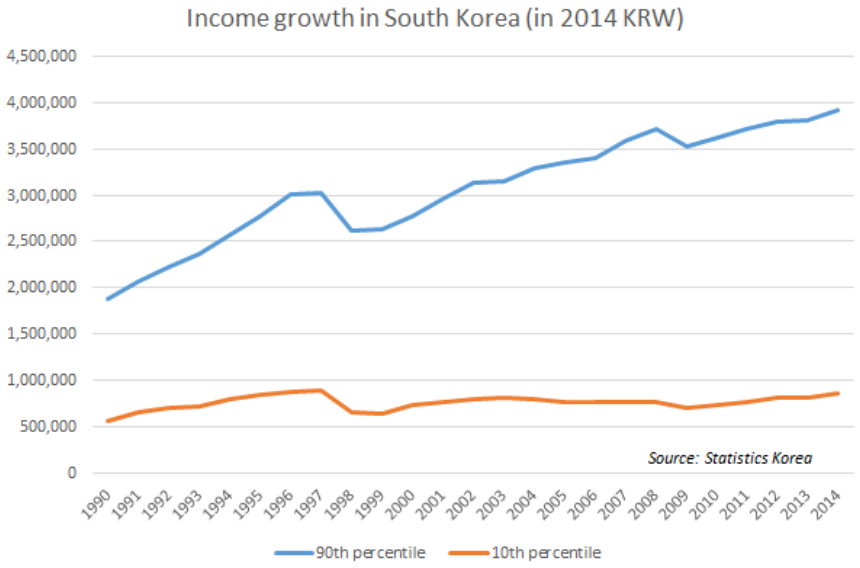
今日成績公布，數學五標全部下修，分別為 11、9、6、4、3 級分，滿級分人數更創近 10 年新低。

根據大考中心資料顯示，今年各科拿到最高 15 級分人數以英文科最多，共 4895 人，其次依序為國文科 4064 人、自然科 4062 人、社會科 2600 人、數學科 1558 人，數學 15 級分人數僅僅 1558 人，為近 10 年新低。至於 0 級分人數，以數學科 44 人最多，其次是社會科 12 人、英文科和自然科為 10 人，國文科 8 人。

110學年度學科能力測驗
各科成績標準一覽表

標準 / 項目	頂標	前標	均標	後標	底標
國文	13	12	11	9	8
英文	13	12	8	5	4
數學	11	9	6	4	3
社會	13	12	10	8	7
自然	13	12	8	6	5

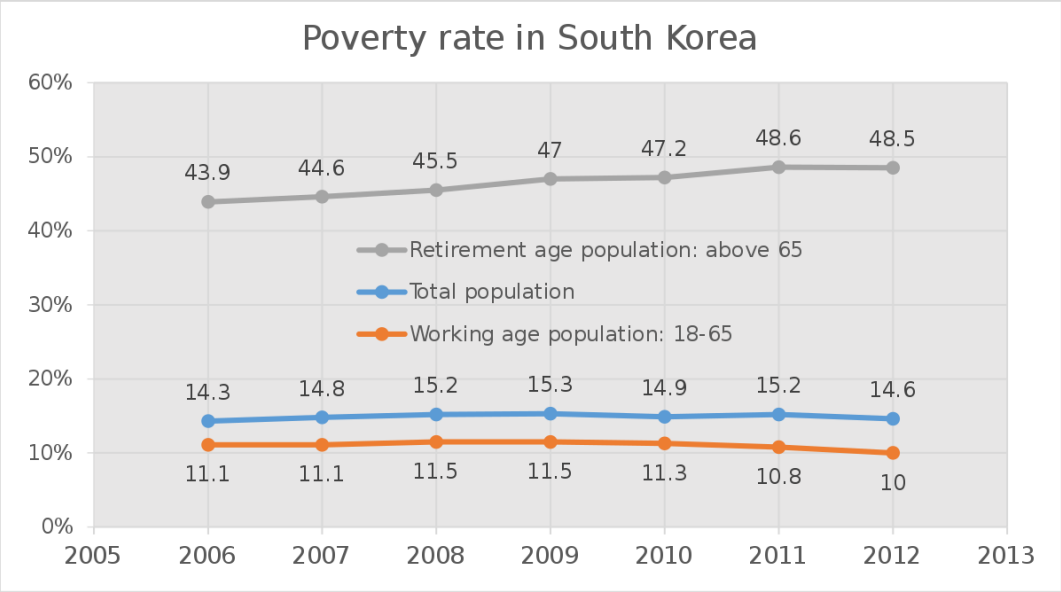
Example: **PR99 – PR1**



The New York Times

'Parasite' and South Korea's Income Gap: Call It Dirt Spoon Cinema

Bong Joon Ho's latest film joins a growing list of movies criticizing South Korean inequality — a problem so pervasive it has given birth to its own slang.



下流老人 [編輯]

維基百科，自由的百科全書

下流老人（日語：かりゅうろうじん）一詞是日本社會學者藤田孝典於其2015年著作《下流老人: 一億総老後崩壊の衝撃》[1]中所提出的。大意為日本近年來出現了大量過著中下階層生活的老人，年金制度即將崩壞、長期照護缺乏人力、高齡醫療缺乏品質、照護條件日益提高、老人居住困難，而且未來會只增不減，若政府不提出有效政策，可能出現「1億人的老後崩壞」。

「下流老人」這個名詞的目的在於說明高齡者的貧窮生活，以及潛藏在其背後的問題，並沒有瞧不起或歧視高齡者。藤田也指出，許多他輔導的下流老人，年輕時也是年薪400萬日元（約120萬台幣）中產階級[2]。

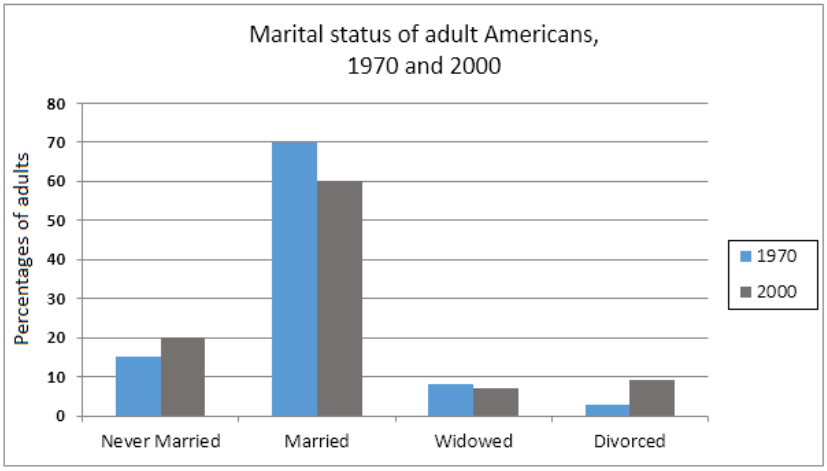
Topic 2: How to describe the distribution of data 何謂分配（分布）?

哪些值(分類)比較多，哪些值(分類)比較少

畫出分配圖（分佈圖)

Categorical Data:

Bar chart

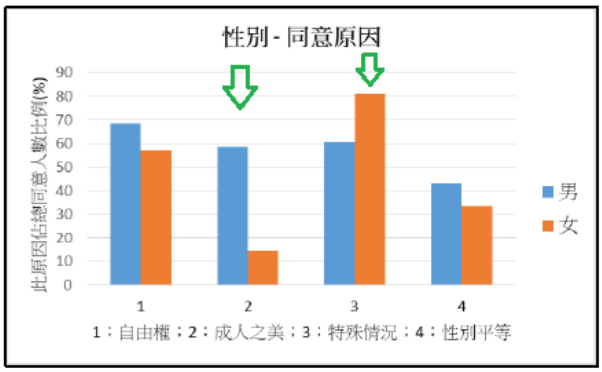


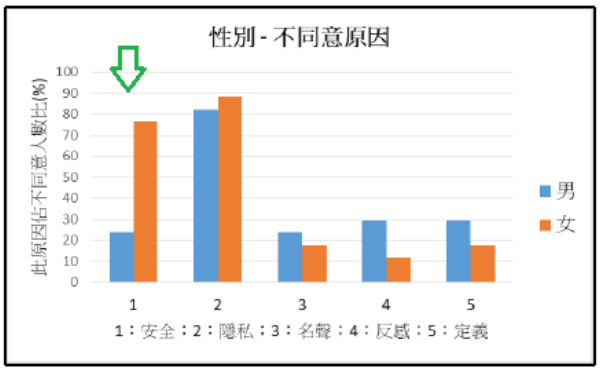
應用統計課程報告: 問卷 (2019 spring by 機械系同學)

Q: 你是否同意異性進入宿舍? (Bar charts)

Pr(同意|男生) = 75%, Pr(同意|女生) = 55%

Pr(同意|有交往經驗) = 72%, Pr(同意|無交往經驗) = 58%



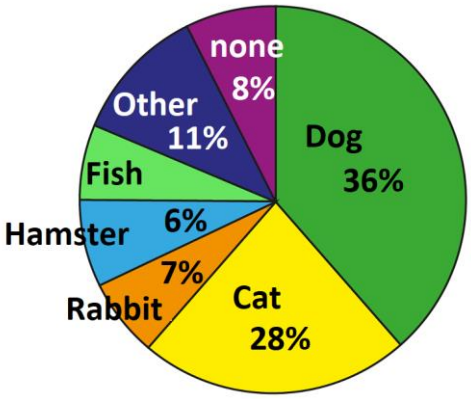


因“成人之美”的原而同意異性進入宿舍男女比例相差懸殊，男性遠高於女性，其顯示在兩互動方面男生似乎較思想開放。

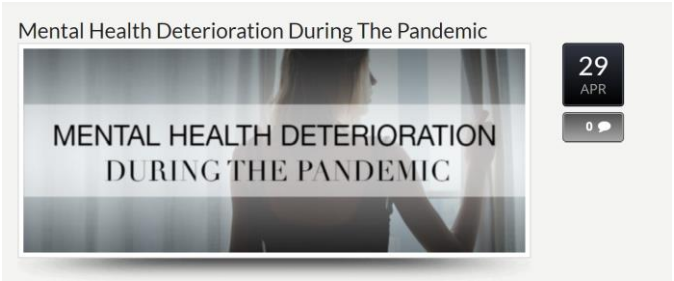
因特殊情況之方便性而同意異進入宿舍的男女比例相差不少，我們推估現象可能源自於女性在特殊狀況下需要幫助的情形較男生常發生（例如：搬重物），故女同意此原因比高。

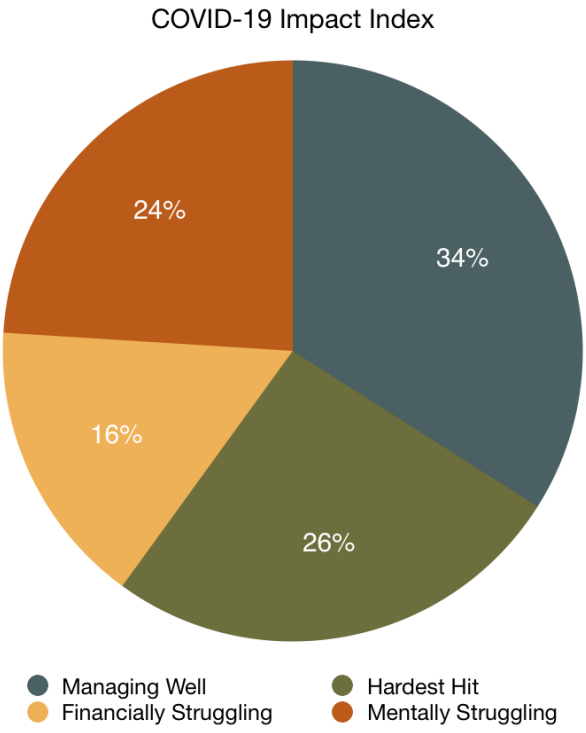
Pie chart

Example: Keeping what kind of pets



Example: How the pandemic affects mental health?





Distribution of numerical data:

Histogram

X-axis: divide data values into several classes (equally classified)

組距: 電腦有預設, 或是自訂

Y-axis: frequency or count in each class

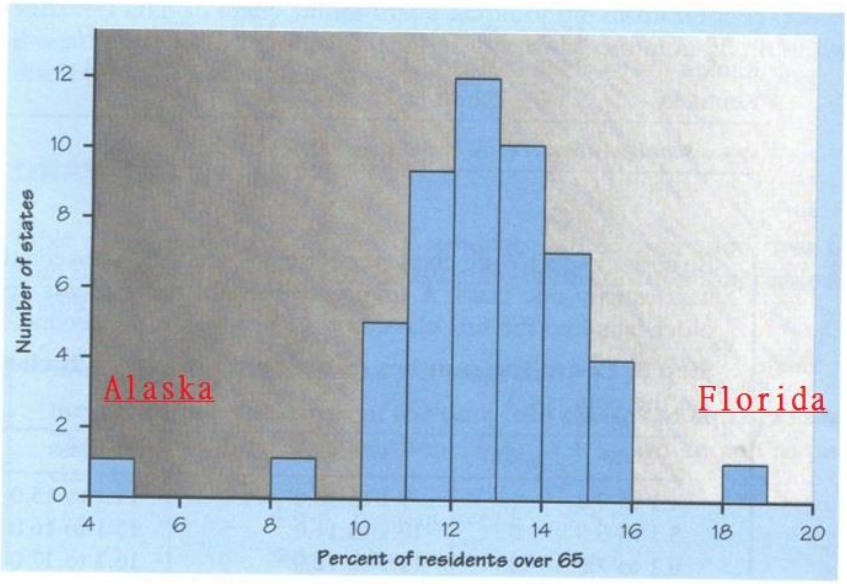


Figure 4-13 A histogram of the distribution of the percent of residents over 65 years of age in the 50 states.

Stem-leaf plot: Re-visited

How to choose a "stem"?

- It is similar to choosing the width of classes for a histogram.
- It depends on the distribution of the data

Example: California earthquake

- stem – digit in one (個位數);
- leaf – first decimal place (小數第一位)

Note: I prefer writing the values in the leaves in an ascending order. Compared with a histogram, a stem-leaf plot contains more detailed information of the data.

EXAMPLE 1.2.1. Consider these observations on the random variable X, the magnitude of a California earthquake as measured on the Richter scale:

1.0	8.3	3.1	1.1	5.1
1.2	1.0	4.1	1.1	4.0
2.0	1.9	6.3	1.4	1.3
3.3	2.2	2.3	2.1	2.1
1.4	2.7	2.4	3.0	4.1
5.0	2.2	1.2	7.7	1.5

The first digits of these numbers are 1, 2, 3, 4, 5, 6, 7, 8. These digits will serve as stems and row labels. See Figure 1.4*a*. We next represent the data graphically by recording the

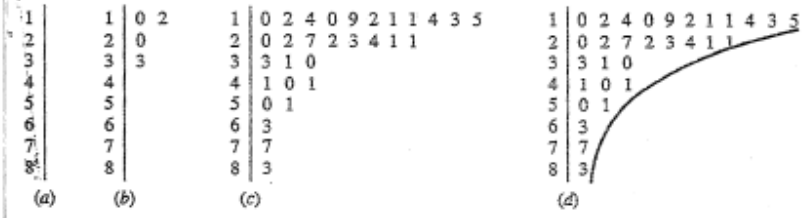


FIGURE 1.4

Stem-and-leaf display for the magnitude of a sample of California earthquakes as measured on the Richter scale: (*a*) choosing stems, (*b*) recording the first four data points, (*c*) the entire data set displayed and (*d*) looking for shape.

Table 1.1 Percent of population 65 years old and over, by state (1996)

State	Percent	State	Percent	State	Percent
Alabama	13.0	Louisiana	11.4	Ohio	13.4
Alaska	5.2	Maine	13.9	Oklahoma	13.5
Arizona	13.2	Maryland	11.4	Oregon	13.4
Arkansas	14.4	Massachusetts	14.1	Pennsylvania	15.9
California	10.5	Michigan	12.4	Rhode Island	15.8
Colorado	11.0	Minnesota	12.4	South Carolina	12.1
Connecticut	14.3	Mississippi	12.3	South Dakota	14.4
Delaware	12.8	Missouri	13.8	Tennessee	12.5
Florida	18.5	Montana	13.2	Texas	10.2
Georgia	9.9	Nebraska	13.8	Utah	8.8
Hawaii	12.9	Nevada	11.4	Vermont	12.1
Idaho	11.4	New Hampshire	12.0	Virginia	11.2
Illinois	12.5	New Jersey	13.8	Washington	11.6
Indiana	12.6	New Mexico	11.0	West Virginia	15.2
Iowa	15.2	New York	13.4	Wisconsin	13.3
Kansas	13.7	North Carolina	12.5	Wyoming	11.2
Kentucky	12.6	North Dakota	14.5		

Source: *Statistical Abstract of the United States, 1997.*

* A poor stem-leaf plot: a lot of stems have no leaves!

(分太細會造成有莖沒有葉)

Example: % of population 65 years old and over

Largest: Florida → 18.5%

Smallest: Alaska → 5.2%

* A poor stem-leaf plot: some leaves are too long!

→ Modification: use double stems (separate 0-4 & 5-9)

Example: Width of head at birth (stem – 個位數拆成兩: 0-4; 5-9)

EXAMPLE 1.2.2. In a study of growth in males these observations are obtained on X, the circumference in centimeters of a child's head at birth.

33.1	34.6	34.2	36.1	34.2	35.6
34.5	35.8	34.5	34.2	34.3	35.2
33.7	36.0	34.2	34.7	34.6	34.3
33.4	34.9	33.8	33.6	35.2	34.6
33.7	34.8	33.9	34.7	35.1	34.2
36.5	34.1	34.0	35.1	35.3	

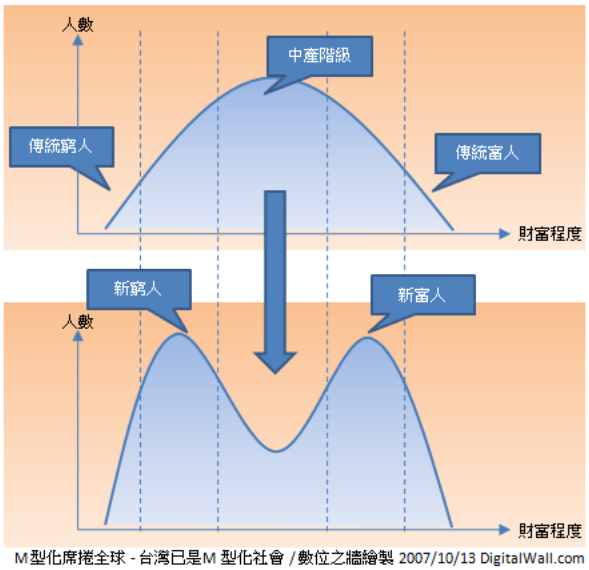
33	1 4
33	7 7 8 9 6
34	1 2 2 0 2 2 3 3 2
34	5 6 9 8 5 7 7 6 6
35	1 2 1 3 2
35	8 6
36	0 1
36	5

FIGURE 1.5
A double stem-and-leaf display giving the circumference in centimeters of a child's head at birth based on the data of Example 1.2.2.

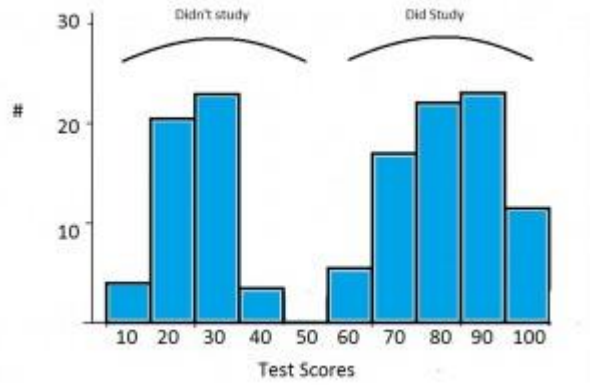
Peaks: local maximum

Example of Bimodal distribution ("double peaks")

- Mixture of two distributions



M型化席捲全球 - 台灣已是M 型化社會 / 數位之牆繪製 2007/10/13 DigitalWall.com



Boxplot: 5-number summary

- Boxplot
  - Box: Q1, Median, Q3 → length of a box = IQR
  - "whisker" (box 旁邊長出來的兩條 "鬚"): more than one definition

Types of box plots [edit]

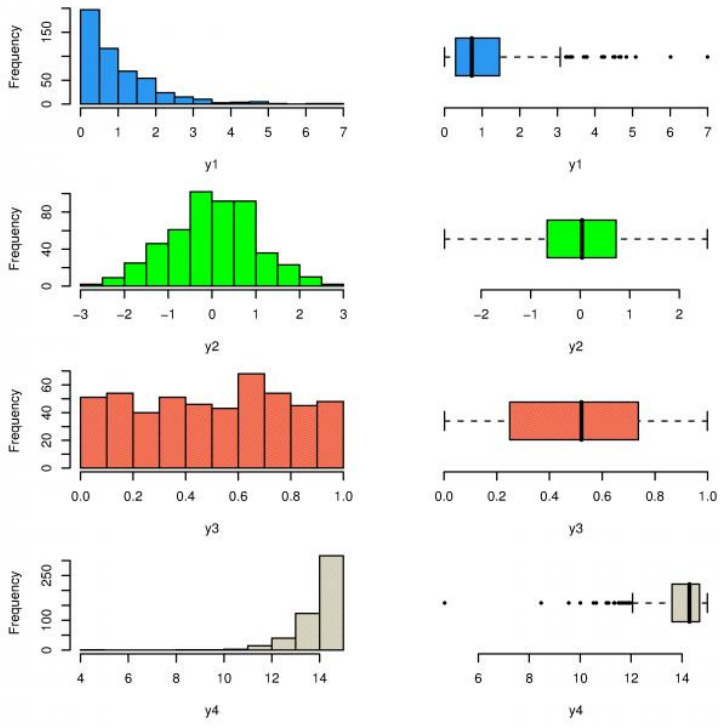
Box and whisker plots are uniform in their use of the box: the bottom and top of the box are always the first and third quartiles, and the band inside the box is always the second quartile (the median). But the ends of the whiskers can represent several possible alternative values, among them:

- the minimum and maximum of all of the data[1] (as in figure 2)
- the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile (often called the Tukey boxplot)[2][3] (as in figure 3)
- one standard deviation above and below the mean of the data
- the 9th percentile and the 91st percentile
- the 2nd percentile and the 98th percentile.

- Inner fence: f1 = Q1 – 1.5 IQR, f3 = Q3 + 1.5 IQR
- Outer fence: F1 = Q1 – 2 * 1.5 IQR, f3 = Q3 + 2 * 1.5 IQR

Note: the two fences are used to identify outliers

Histogram and boxplot:

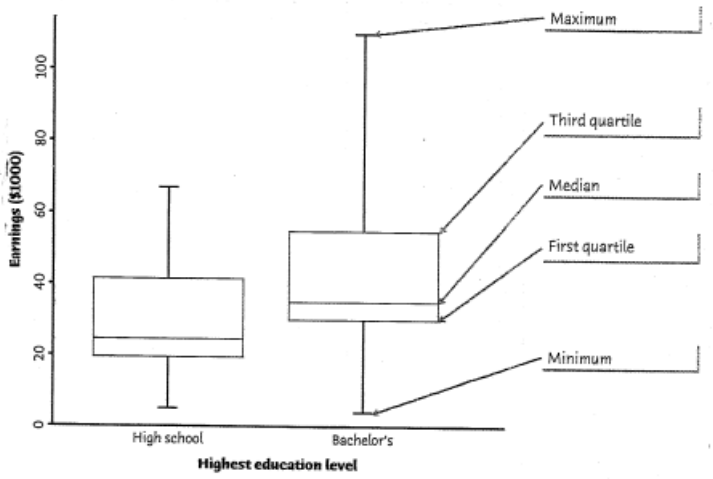


Our textbook:

THE FIVE-NUMBER SUMMARY

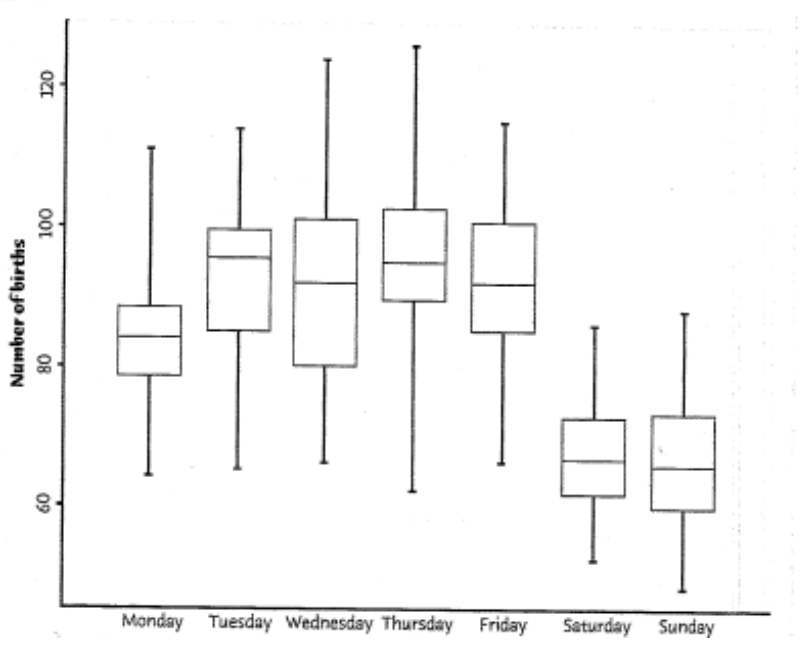
The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the *largest observation, written in order from smallest to largest. In symbols,* the five-number summary is

Minimum Q_1 M Q_3 Maximum



Example: 生產時間是否異常?

2.24 Never on Sunday: also in Canada? Exercise 1.4 (page 9) gives the number of births in the United States on each day of the week during an entire year. The boxplots in Figure 2.5 are based on more detailed data from Toronto, Canada: the number of births on each of the 365 days in a year, grouped by day of the week.[9] Based on these plots, give a more detailed description of how births depend on the day of the week.



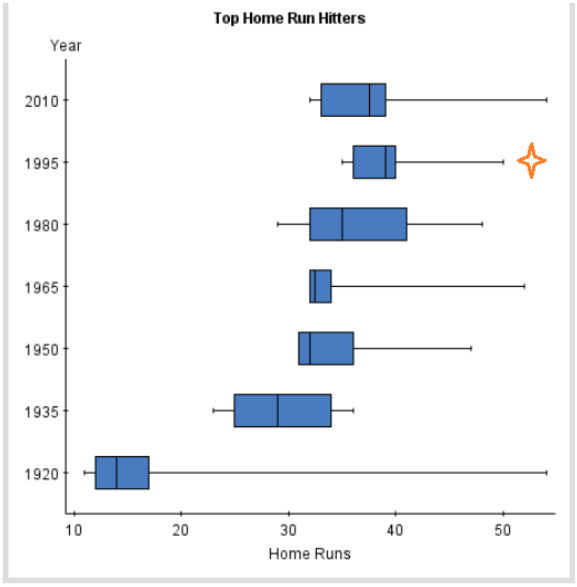
Example: MLB data: Top 10 homerun hitters

https://www.statcrunch.com/5.0/viewreport.php?reportid=29693

Home Run Leaders

Sign in to analyze data!

Row	1920	1935	1950	1965	1980	1995	2010
1	54	36	47	52	48	50	54
2	19	36	37	39	41	40	42
3	17	34	36	34	41	40	39
4	15	31	34	33	38	40	38
5	14	30	32	33	35	39	38
6	14	28	32	32	35	39	37
7	12	26	32	32	33	39	34
8	12	25	31	32	32	36	33
9	11	23	31	32	30	36	33
10	11	23	31	32	29	35	32



Increased offense

During the 1990s, Major League Baseball experienced an increase in offensive output that resulted in some unprecedented home run totals for the power hitters of the decade. While just three players reached the 50-home run mark in any season between 1961 and 1994, many sluggers would start to surpass that number in the mid-90s.

In 1996, Mark McGwire of the Oakland Athletics led the majors with 52 home runs despite missing part of the season. In 1997, both McGwire and the Seattle Mariners' Ken Griffey Jr. threatened the individual record of 61 -- set by Roger Maris in 1961 -- before ending the season with 58 and 56 home runs, respectively.

New MLB Home Run Records Raise Suspicions - Baseball ...

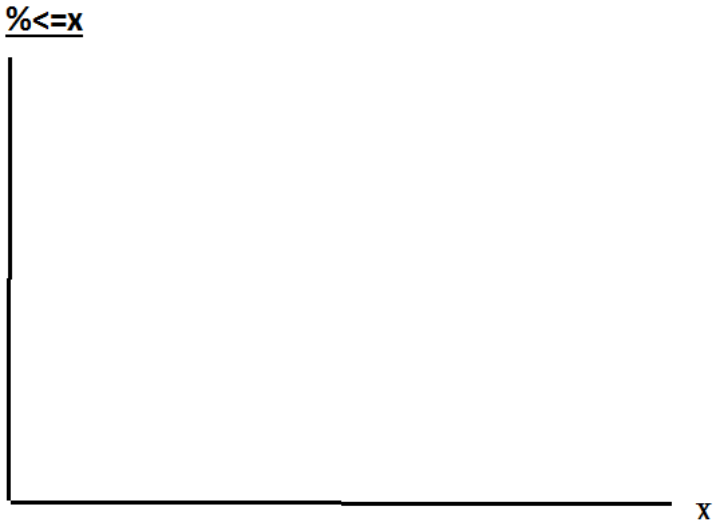
Jul 10, 2019 - New **MLB Home Run** Records Raise Suspicions **Major League Baseball** fans and experts have been in for an exciting, but ... However, since the **90s**, baseball has had a tumultuous relationship with steroid and **drug** use. **MLB** ...

Cumulative Distribution Plot

X-axis: x value, Y-axis: $\% \le x$

Example: 1 5 9 12 18

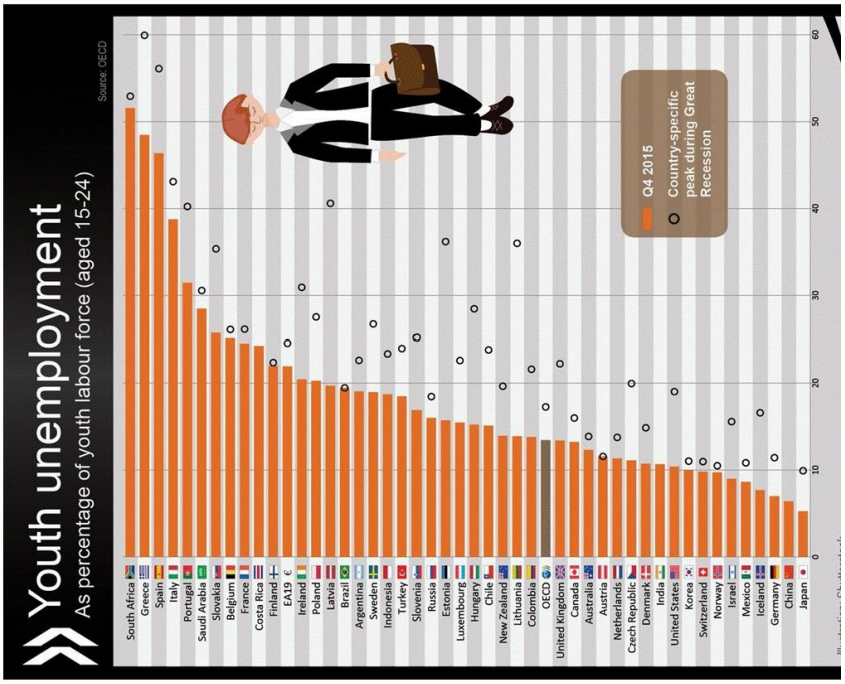
$\% \le 1 = 1/5,\ \% \le 5 = 2/5,\ \% \le 9 = 3/5,\ \% \le 12 = 4/5,\ \% \le 18 = 5/5$

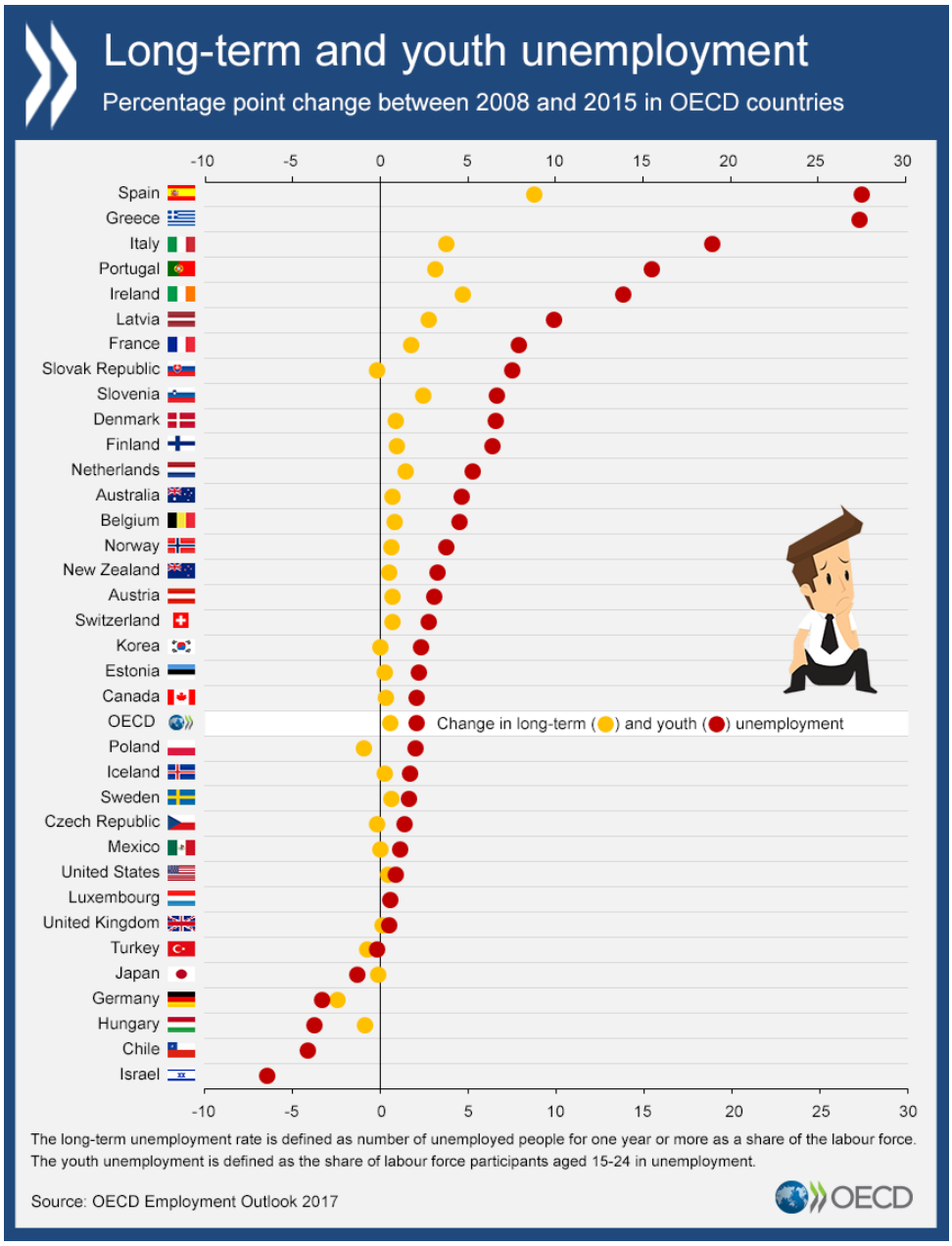


- Step function

- Continuous from the right (right continuous)


Youth unemployment
As percentage of youth labour force (aged 15-24)
Source: OECD
South Africa
Greece
Spain
Italy
Portugal
Saudi Arabia
Slovakia
Belgium
France
Costa Rica
Finland
EA19
Ireland
Poland
Latvia
Brazil
Argentina
Sweden
Indonesia
Turkey
Slovenia
Russia
Estonia
Luxembourg
Hungary
Chile
New Zealand
Lithuania
Colombia
OECD
United Kingdom
Canada
Australia
Austria
Netherlands
Czech Republic
Denmark
India
United States
Korea
Switzerland
Norway
Israel
Mexico
Iceland
Germany
China
Japan
0
10
20
30
40
50
60
Q4 2015
Country-specific peak during Great Recession
Illustration: Shutterstock
Source: OECD calculations based on the OECD Short-Term Labour Market Statistics Database and national labour force surveys.
OECD


Youth unemployment
As percentage of youth labour force (aged 15-24)
Source: OECD
Q4 2015
Country-specific peak during Great Recession


Long-term and youth unemployment
Percentage point change between 2008 and 2015 in OECD countries
-10
-5
0
5
10
15
20
25
30
Spain
Greece
Italy
Portugal
Ireland
Latvia
France
Slovak Republic
Slovenia
Denmark
Finland
Netherlands
Australia
Belgium
Norway
New Zealand
Austria
Switzerland
Korea
Estonia
Canada
OECD
Poland
Iceland
Sweden
Czech Republic
Mexico
United States
Luxembourg
United Kingdom
Turkey
Japan
Germany
Hungary
Chile
Israel
Change in long-term ( ) and youth ( ) unemployment
The long-term unemployment rate is defined as number of unemployed people for one year or more as a share of the labour force.
The youth unemployment is defined as the share of labour force participants aged 15-24 in unemployment.
Source: OECD Employment Outlook 2017
OECD

Topic: Probability Theory

1. Set-based: today
2. Random variables: next week

Set Theory

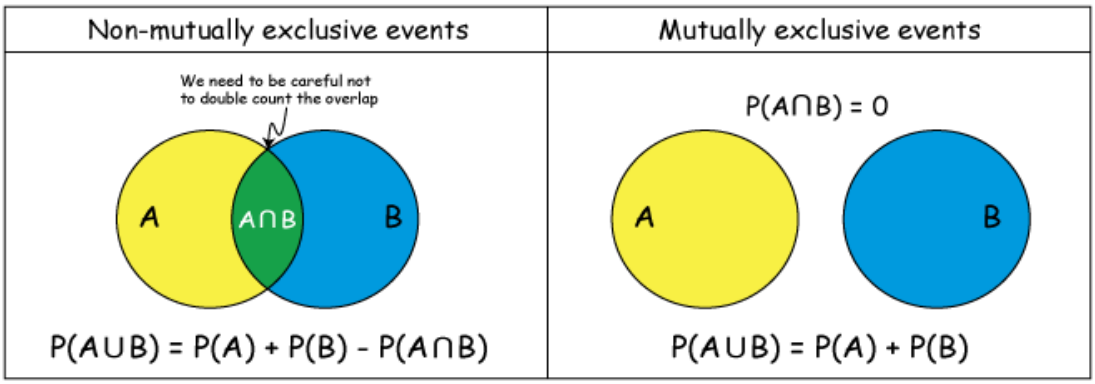
Notations

- 樣本空間 (Sample Space): S
- 事件 (Event)
- 空集合 (Empty set): ϕ

Relationship between two sets:

- 聯集 (union): $A \cup B$
- 交集 (intersection): $A \cap B$
- 互斥 (mutually exclusive): $A \cap B = \phi$

Venn Diagram (以圖形來表示集合關係)



Axioms of probability

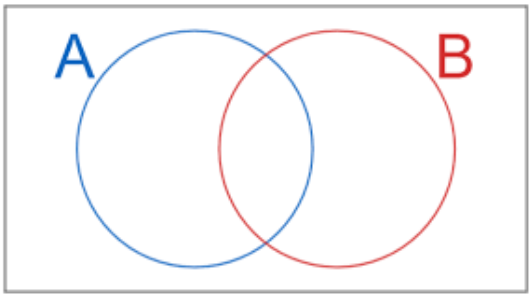
1. $P(S) = 1$
2. $P(A) \geq 0$
3. $A_1, A_2, \ldots$ are mutually exclusive,

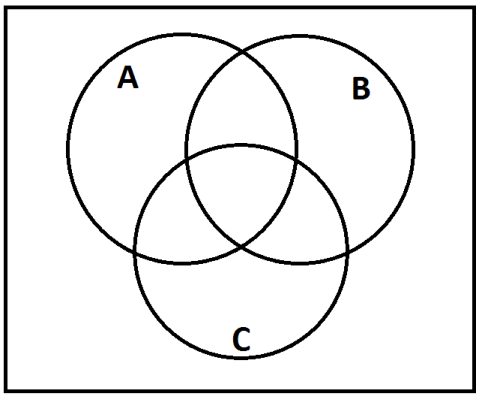
$$P(A_1 \cup A_2 \ldots \cup A_k) = P(A_1) + P(A_2) + P(A_k).$$

Properties:

$$P(A \cup B) = P(B) + P(B) - P(A \cap B)$$

$$P(A \cup B \cup C) = P(B) + P(B) + P(C) - P(A \cap B) - P(C \cap B) - P(A \cap C) + P(A \cap B \cap C)$$





Example:

EXAMPLE 3.2.1. It is thought that 30% of all people in the United States are obese (A_1) and that 3% suffer from diabetes (A_2). Two percent are obese and suffer from diabetes. What is the probability that a randomly selected person is obese *or* suffers from diabetes?

We have been given $P[A_1] = .3$, $P[A_2] = .03$, and $P[A_1 \text{ and } A_2] = .02$. We are asked to find $P[A_1 \text{ or } A_2]$. Applying the general addition rule, we obtain

$$\begin{aligned} P[A_1 \text{ or } A_2] &= P[A_1] + P[A_2] - P[A_1 \text{ and } A_2] \\ &= .30 + .03 - .02 \\ &= .31 \end{aligned}$$

A1: 肥胖, A2: 糖尿病

Find the probability of the union $\Pr(A_1 \cup A_2)$

Example: List all elements in the sample space

Toss a coin 3 times, H : head (正面) T : tail (反面)

HHH
HHT
HTH
HTT
THH
THT
TTH
TTT

Event A: at least two heads (HHH,HHT,HTH,THH)

Event B : no tail (HHH)

Calculation of a probability

Probability = (# of events in A)/(total # in S)

Find $\Pr(A)$, $\Pr(B)$, $\Pr(A \cap B)$, $\Pr(A \cup B)$

$$\Pr(A) = \frac{3}{8} + \frac{1}{8} = \frac{1}{2}$$

$$\Pr(B) = \frac{1}{8}$$

$$\Pr(A \cap B) = \frac{1}{8}$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \frac{1}{2}$$

Definition of a conditional probability (條件機率):

The probability of event A given that B must occur:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad \text{with} \quad \Pr(B) \neq 0 .$$ -

Note: treat B as the new universe (sample space).

Property:

$$\begin{aligned} \Pr(A \cap B) &= \Pr(A \mid B) \Pr(B) \\ &= \Pr(B \mid A) \Pr(A) \end{aligned}$$

Exercise: calculate $\Pr(B \mid A)$

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{1/8}{1/2} = \frac{1}{4}$$

Definition of Independence (集合間的獨立)

A and B are independent if and only if

$\Pr(A \mid B) = \Pr(A)$ (or $\Pr(B \mid A) = \Pr(B)$)

解釋: B 發生與否對 A 的機率沒有影響 (A 發生與否對 B 的機率沒有影響)

The occurrence of B has no influence on the probability of A.

Equivalent definition:

A and B are independent if and only if $\Pr(A \cap B) = \Pr(A)\Pr(B)$

Ex. A = 你的期末考成績超過九十分 (your mid-term score > 90)
B = 十月的交通事故低於平均 (rate of traffic accidents in Oct. < average)
C = 你的期中考成績是全班前 5%. (your final score is top 5%)

A and B are independent, but A and C are dependent.

Important: Determine whether two sets are independent

EXAMPLE 3.2.2. It was recently reported that 18% of all college students at some point in their college careers suffer from depression (A_1), that 2% consider suicide (A_2), and that 19% suffer from depression or consider suicide. What is the probability that a randomly selected college student suffers from depression *and* has considered suicide? What is the probability that a randomly selected student has suffered from depression *but* has not considered suicide?

We know that $P[A_1] = .18$, $P[A_2] = .02$, and $P[A_1 \text{ or } A_2] = .19$. We are asked, first, to find $P[A_1 \text{ and } A_2]$. Applying the general addition rule, we get

$$P[A_1 \text{ or } A_2] = P[A_1] + P[A_2] - P[A_1 \text{ and } A_2]$$

$$P[A_1 \text{ and } A_2] = P[A_1] + P[A_2] - P[A_1 \text{ or } A_2]$$
$$= .18 + .02 - .19$$
$$= .01$$

$P(A_1 \text{ and } A_2^c) = P(A_1) - P(A_1 \cap A_2)$
$= 0.18 - 0.01 = 0.17$

Q: Are A_1 and A_2 independent? No!
$P(A_2 \mid A_1) = 0.01/0.18 = 5.56\% > P(A_2) = 2\%$

EXAMPLE 3.3.2. It is estimated that 15% of the adult population has hypertension, but that 75% of all adults feel that personally they do not have this problem. It is also estimated that 6% of the population has hypertension but does not think that the disease is present. If an adult patient reports thinking that he or she does not have hypertension, what is the probability that the disease is, in fact, present?

A = hypertension is present

B = feel that hypertension is Not present

$P(A) = 0.15 \qquad P(B) = 0.75$

$P(A \cap B) = 0.06$

Q: $P(A|B) = ? = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.06}{0.75}$

$= 8\%$

Q. $P(B^c|A) = \dfrac{P(A \cap B^c)}{P(A)} = \dfrac{0.15 - 0.06}{0.15}$

EXAMPLE 3.5.1. Assume that among the U.S. population as a whole, 55% are overweight (A_1), 20% have high blood pressure (A_2), and 60% are overweight or have high blood pressure. Is the fact that a person is overweight independent of the state of his or her blood pressure? The answer to this question is not obvious. Using the general addition principle yields

$$P[A_1 \text{ and } A_2] = P[A_1] + P[A_2] - P[A_1 \text{ or } A_2]$$

or, in this case,

$$P[A_1 \text{ and } A_2] = .55 + .20 - .60 = .15$$

Thus

$$P[A_2|A_1] = \frac{P[A_1 \text{ and } A_2]}{P[A_1]}$$

$$= \frac{.15}{.55} = \frac{15}{55} = .27$$

Claim: If two sets are "mutually exclusive", they can't be independent

Proof:

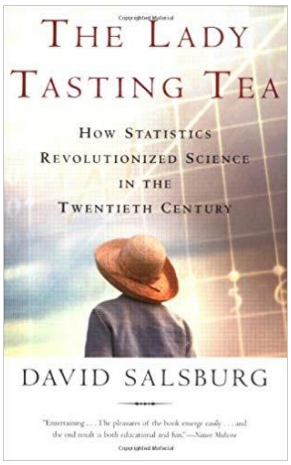
Know that $A \cap B = \phi$ with $\Pr(A) > 0$ and $\Pr(B) > 0$.

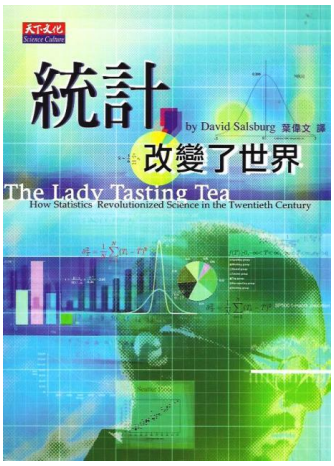
By definition:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = 0 \neq \Pr(A).$$

History of Statistics

參考書籍: 統計改變世界 – 天下文化





補充: History of Statistics (細節見附錄)

Pioneers in Statistics Page 18

- Francis Galton

  Regression analysis

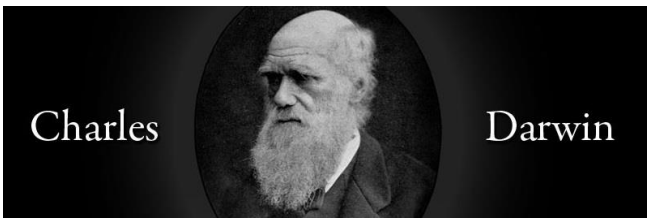
- Karl Pearson

  Method of Moments, Chi-squared test

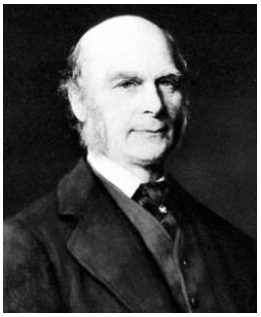
- RA Fisher

  Mathematical statistics, Experimental design, Time Series

Charles Darwin (1809-1882)– Origin of spices, evolution



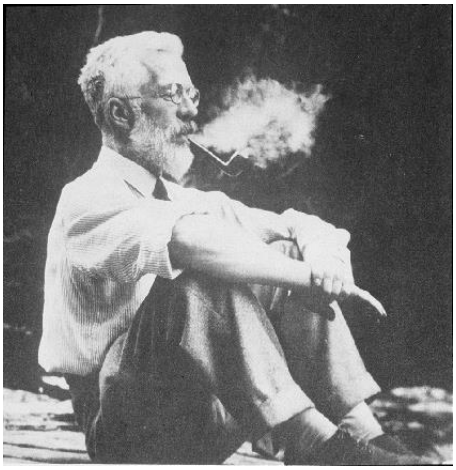
- Francis Galton (1822-1911)–half cousin of Charles Darwin



  - created the statistical concept of correlation and widely promoted regression toward the mean. (迴歸分析)
    - X-axis:: parent's measurement, Y-axis: offspring's measurement
    - Meaning of slope: slope < 1 → regress toward the mediocrity
  - was the first to apply statistical methods to the study of human differences and inheritance of intelligence
  - introduced the use of questionnaires and surveys for collecting data on human communities (問卷設計) and established a biometric lab.
  - was a pioneer in eugenics, coining the term itself and the phrase "nature versus nurture" (優生學)
  - **Karl Pearson (1859-1936)**

  - Galton's student. When Galton died, he left the residue of his estate to the University of London for a Chair in Eugenics. Pearson was the first holder of this chair.
  - Pearson made significant contributions to statistics:
    - Correlation, Chi-square test, method of moment, …
  - Work: "Grammar of Science" (科學的文法)
    - Recommended by Einstein to his friends of the Olympia Academy.
  - A top statistical journal "*Biometrika*" established in 1901 by Francis Galton, Karl Pearson, and Raphael Weldon to promote the study of biometrics.

R.A. Fisher (1890-1962) – The most influential Statistician



- 歷史上最具影響力的統計學家 – 開山祖師的地位

About RA Fisher

- a British statistician and geneticist.
- For his work in statistics, he has been described as "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics".
- In genetics, his work used mathematics to combine Mendelian genetics and natural selection; this contributed to the revival of Darwinism in the early 20th-century revision of the theory of evolution known as the modern synthesis.
- For his contributions to biology, Fisher has also been called "the greatest of Darwin's successors". Fisher also did experimental agricultural research, which has saved millions from starvation.
  - 偉大的遺傳學家
    調停達爾文演化論和孟德爾遺傳率之爭論
    劍橋大學遺傳系主任
  - 統計學的貢獻包含
    古典數理統計
    時間序列: "time domain"
    實驗設計
  - 具有非凡的幾何洞察能力
  - 影響: 統計變得越來越數學了
  - 晚年: 爭論抽煙是否導致肺癌