

Lecture 2 Data Analysis & Probability

09/23/2020

Topic 1: Measures of Dispersion (variation) - Revisited

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

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

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

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

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

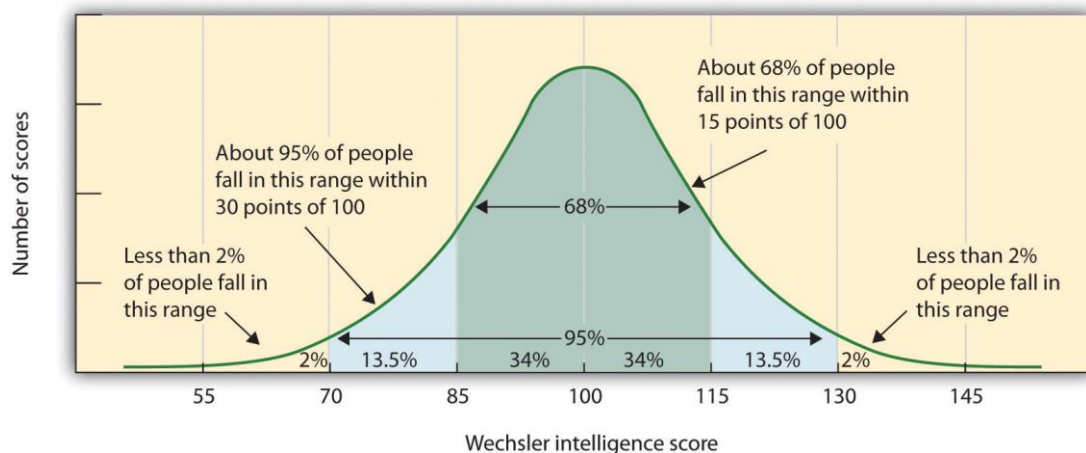
Remarks:

- 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}) is not a robust measure of variation.

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

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

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



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

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

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 in 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:

$$\begin{aligned} &< Q1 - 1.5IQR \quad (\text{outlier on the left}) \\ &> Q3 + 1.5IQR \quad (\text{outlier on the right}) \end{aligned}$$

d. range = Max – Min

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

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		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
校系代碼：013202		
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%+ <u>審查資料*50%</u> +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*

項目 \ 標準	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

項目 \ 標準	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

台大未來不再有「N冠王」，[清大電機正取40卻備取了392人](https://www.thenewslens.com)- The ...

<https://www.thenewslens.com> > 教育 ▾

2019年5月4日 - 標籤: 大學甄選, 學測, 台大, 清大, 台灣大學, 教育部, 清大電機, 交大電機, ... 今年學測改採「5選4」, 以至於4科滿級分人數增多, 台大今年共4039人次 ...

家長批大學申請備取名額過多招聯會回應了| 文教新訊| 文教 ...

<https://udn.com> > 文教, 文教新訊 ▾

2019年5月11日 - 招聯會請清大、交大、長庚大學各一系代表說明該系甄選與錄取情況。 ... 五選四上陸第一年, 未來各校系也將依108年實施狀況, 觀察學生入學後 ...

提高鑑別度，[台大教務長：明年部分科系申請入學擬加考筆試](https://www.parenting.com.tw) ...

<https://www.parenting.com.tw> > 教育現場, 教改追蹤 ▾

2019年7月3日 - 今年起大學申請入學實施學測5選4新制, 同分人數過多, 特別是台大、清大、交大這類的頂尖大學理工科系, 許多科系只採計數學和自然兩科成績, ...

恢復聯考- Posts | Facebook

<https://www.facebook.com> > Pages > Other > Community > 恢復聯考, Posts ▾

新聞標題: 交大為自辦數理檢定國教盟: 反映學測鑑別度不足©新聞來源: 聯合晚報©新聞 ... 我猜清大的目的是把台灣這一分段的學生全都混在一樣分數 2019-06-11 聯合報記者章凱閏今年度大學申請入學首次採5選4, 卻因部分考科難度偏易, 發生 ...

學測綁指考台大醫擦身過 兩年後綁得更緊

2020-08-11 15:30 聯合報 / 記者鄭惠仁／台南即時報導

日前指考放榜，台南一中有名學生考 451.9 分，超過台大醫學系錄取分數 442 分，卻因學測國文未達前標，無緣登記，另上高醫醫學系。

學測綁指考已多年，不少考生都有類似遭遇，另從 2022 年度開始，指考改為分科測驗，國、英、數乙都不考，採計學測分數。因此，老師無不提醒學生，好好準備學測，免得後悔。

台南一中今年沒有學生上台大醫成熱門話題，其中有名學生雖然指考拿了高分，遠超過台大最低錄取分數，不過，台大另規定醫學系的錄取標準，學測國文必須達前標，這位考生因為沒有達前標，無法登記台大，讓同學、師長為他抱屈。

台南一中教務主任蔡明輝說，這種學測綁指考的情形已有多年，各校學系因錄取規定不同，每年也都會類似情形發生；學生疑惑的是，指考不就是要給學測沒考好的學生另個機會，卻又要用學測分數綁指考，已失去指考目的，而且各校規定也不同，失去公平意義。蔡明輝說，學生會有這樣的抱怨難免，也希望被聽得到；不過，2022 年開始，這種現象會愈來愈明顯，還是要有心理準備。

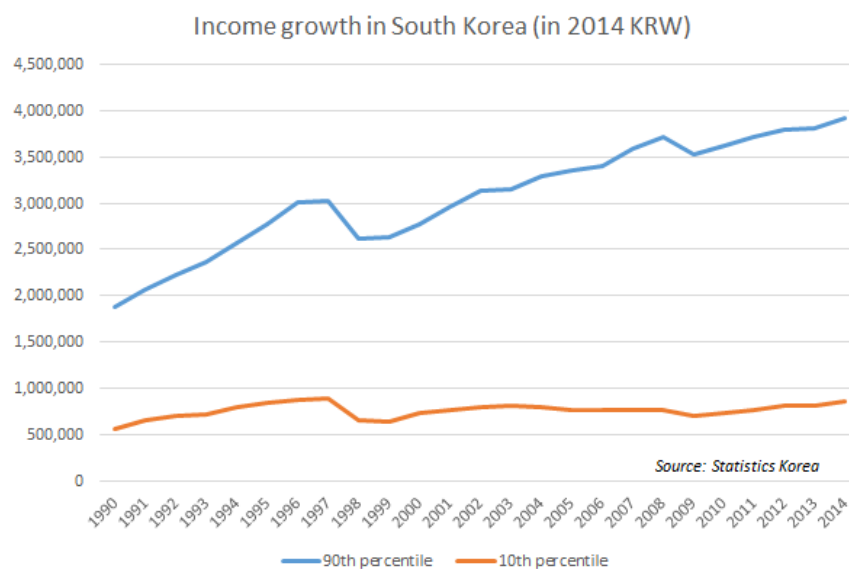
台南女中教務主任胡瑞原表示，從 2022 年起，學測仍考國、英、數、社、自，其中數學分數學乙、數學甲，但指考改為分科測驗後，國、英、數乙不考，以學測的分數來計分，但加大級距，從學測每科 15 級分，提升為 45 級分，除降低競爭性，也更為公平。

胡瑞原說，各校作法不一，採計的標準也會不同，但也有學校擔心，若採取新措施，反而收不到學生，寧願把門檻降低。不過，最重要的仍是要好好準備，把學測考好，避免因學測分數未如預期，而無法上理想學校的遺憾。

Example: PR99 – PR1

1%vs.99% · 所得差距創历史新高

34年來，金字塔頂端1%的高所得者，收入與其他99%逐漸拉大。
2011年創历史新高，1077萬對比78萬。

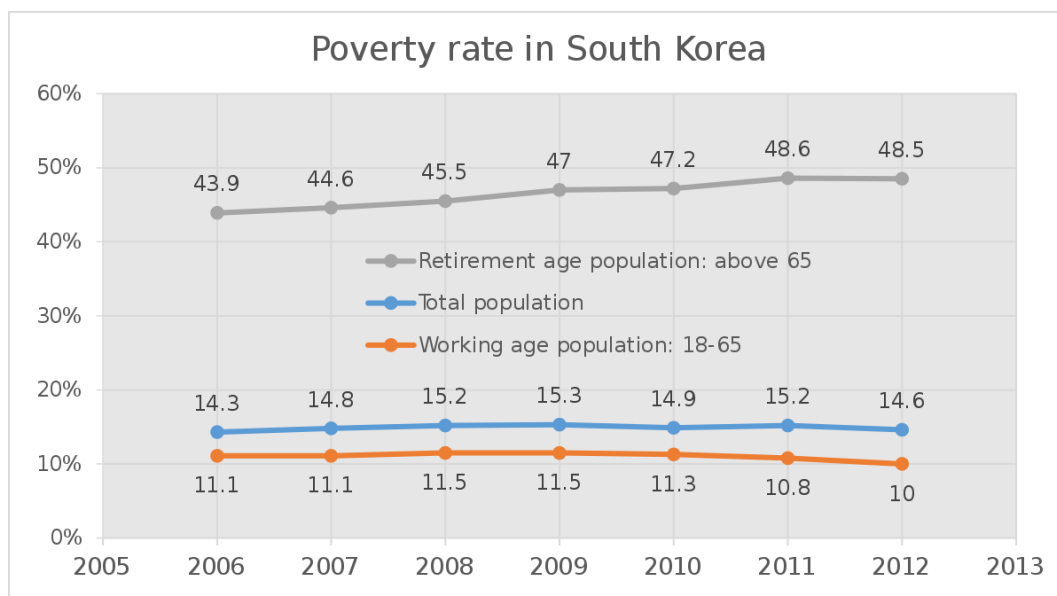


The New York Times

'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：一億總老後崩壞的衝擊》^[1]中所提出的。大意为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]。



下流老人

NT\$221

博客來



今周刊：拒當下流老人

NT\$99

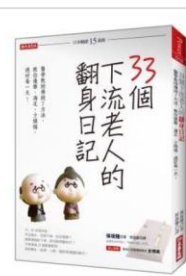
博客來



續·下流老人：政府養不起你、家人養不...

NT\$221

博客來



33個下流老人的翻身日記：醫學教授傅...

NT\$221

金石堂網路書店

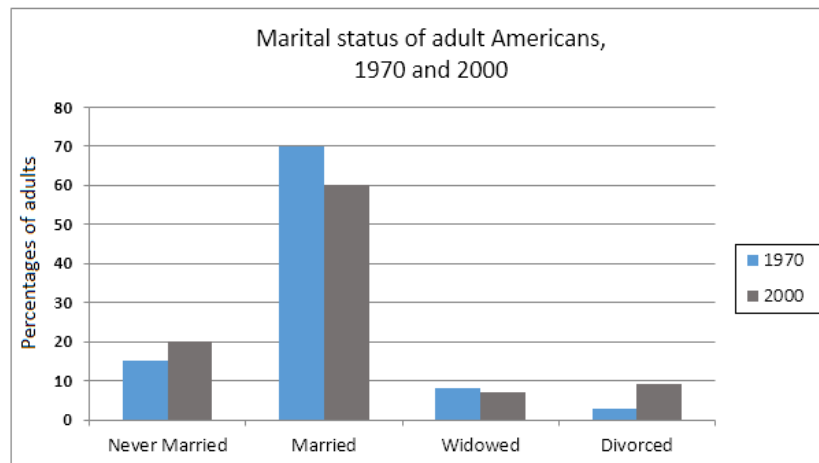
Topic 2: How to describe the distribution of data 何謂分配 (分布)?

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

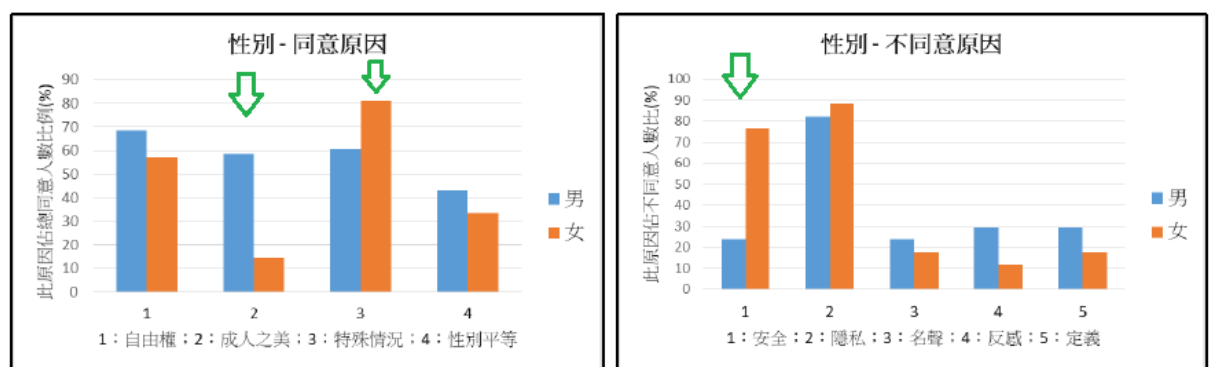


應用統計課程報告：問卷 (2019 spring by 機械系同學)

Q: 你是否同意異性進入宿舍? (Bar charts)

$\Pr(\text{同意}|\text{男生}) = 75\%$, $\Pr(\text{同意}|\text{女生}) = 55\%$

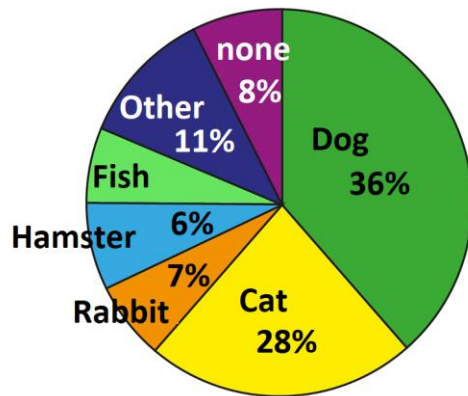
$\Pr(\text{同意}|\text{有交往經驗}) = 72\%$, $\Pr(\text{同意}|\text{無交往經驗}) = 58\%$



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



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

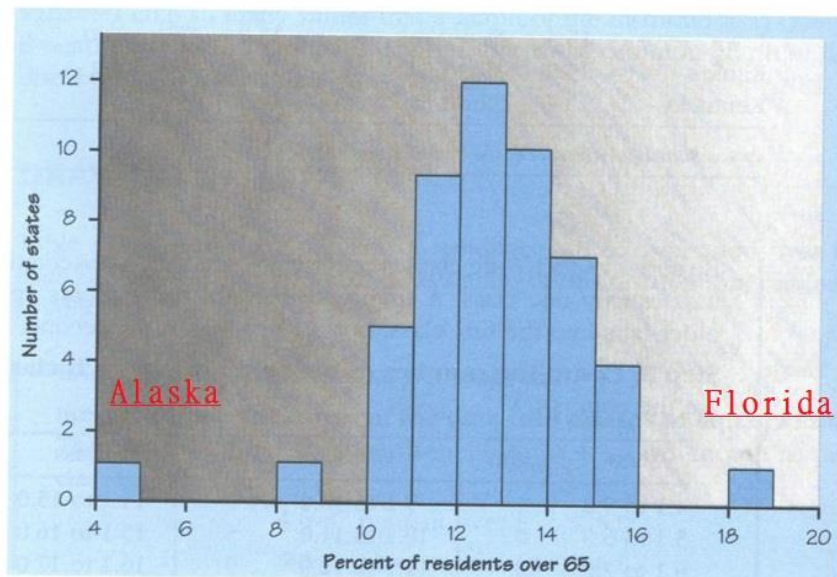


Figure 4-13 A histogram of the distribution of the percent of residents over 65 years of age in the 50 states.

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 1.4a. We next represent the data graphically by recording the

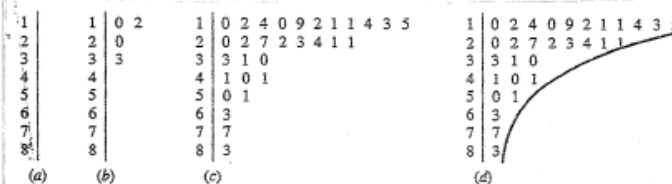


FIGURE 1.4

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.

* 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%

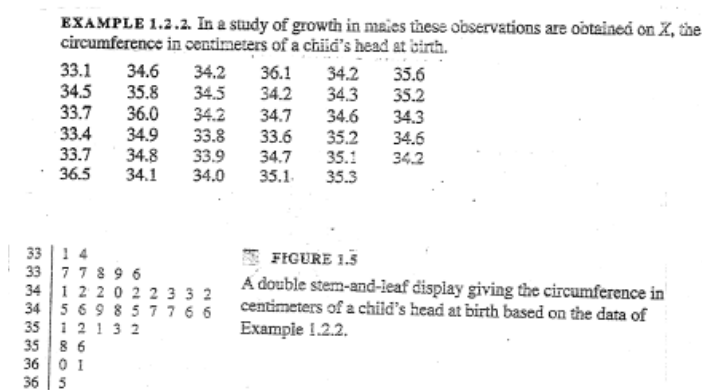
and over, by state (1996)

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: some leaves are too long! → **Modification: use double stems (separate 0-4 & 5-9)**

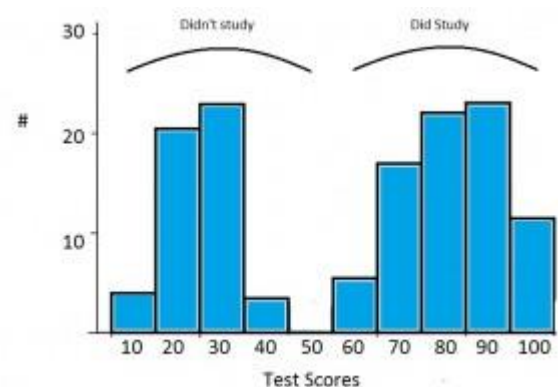
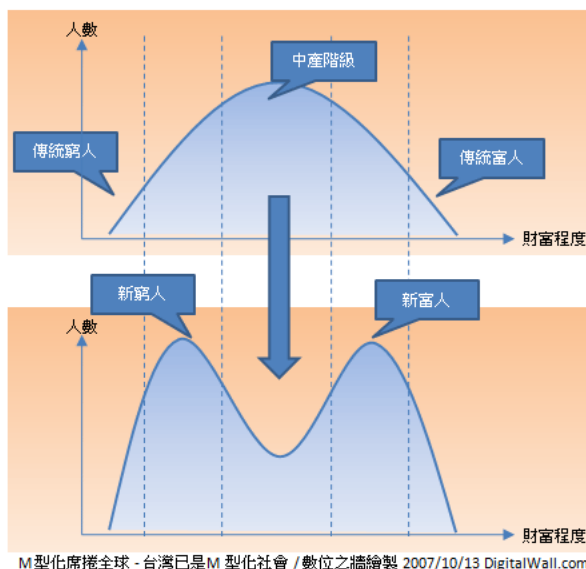
Example: Width of head at birth (stem – 個位數拆成兩: 0-4; 5-9)



Peaks: local maximum

Example of Bimodal distribution (“double peaks”)

- Mixture of two distributions



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

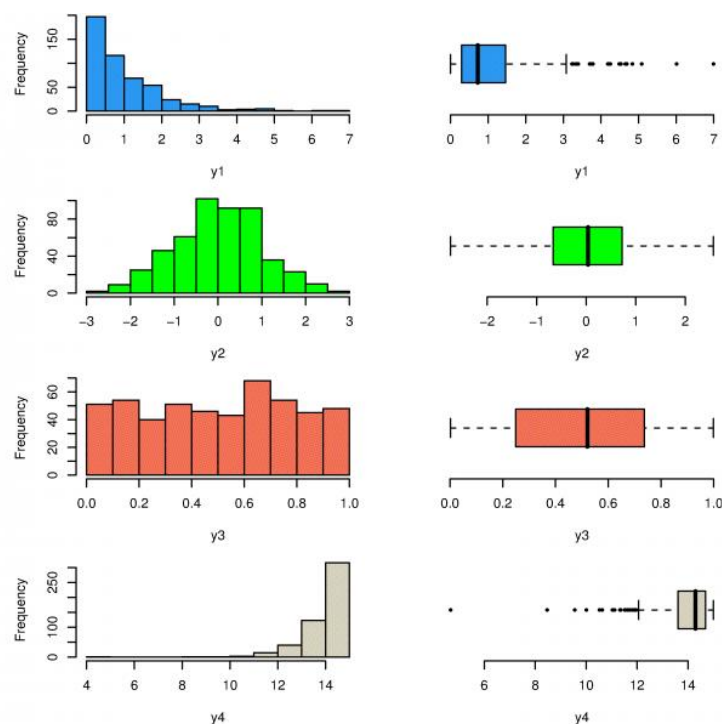
Types of box plots [\[edit\]](#)

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: $f1 = Q1 - 1.5 \text{ IQR}$, $f3 = Q3 + 1.5 \text{ IQR}$
 - Outer fence: $F1 = Q1 - 2 * 1.5 \text{ IQR}$, $f3 = Q3 + 2 * 1.5 \text{ IQR}$

Note: the two fences are used to identify outliers

Histogram and boxplot:

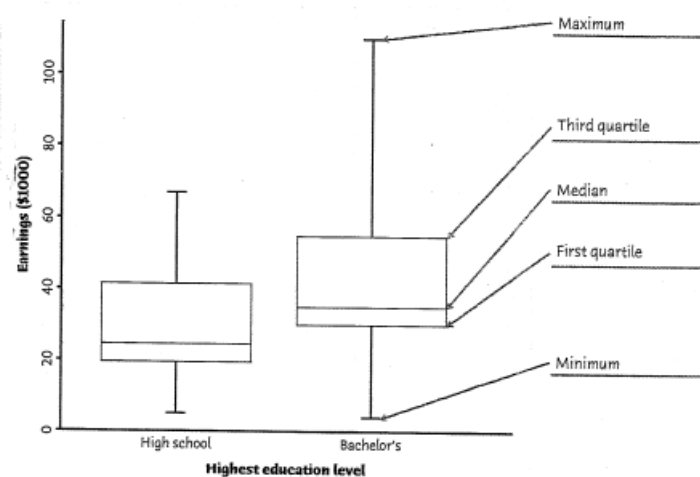


Our textbook:

THE FIVE-NUMBER SUMMARY

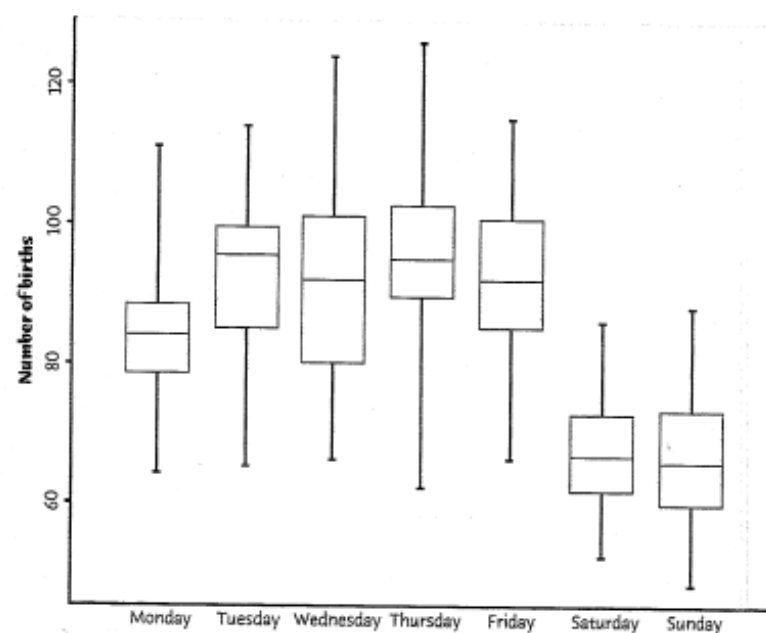
The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from *smallest to largest*. In symbols, the five-number summary is

Minimum Q_1 M Q_3 Maximum



Example: 生產時間是否異常?

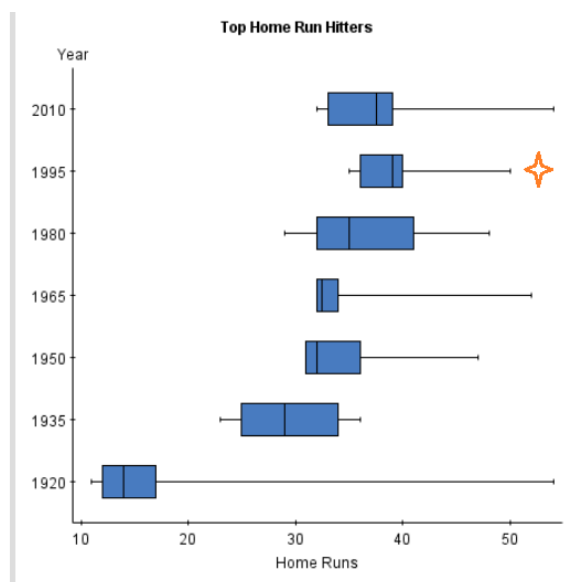
2.24 Never on Sunday: also in Canada? Exercise 1.4 (page 9) gives the number of births in the United States on each day of the week during an entire year. The boxplots in Figure 2.5 are based on more detailed data from Toronto, Canada: the number of births on each of the 365 days in a year, grouped by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



The Steroids Era

Updated: December 5, 2012, 4:23 PM ET

RECOMMEND (0) TWEET (0) COMMENTS (0)
EMAIL PRINT



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 Suspicion - Baseball ...

Jul 10, 2019 - New **MLB Home Run** Records Raise Suspicion **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: $\% \leq x$

Example: 1 5 9 12 18

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



- Step function
- Continuous from the right (right continuous)

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

Topic 3: Probability Theory

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

Set Theory

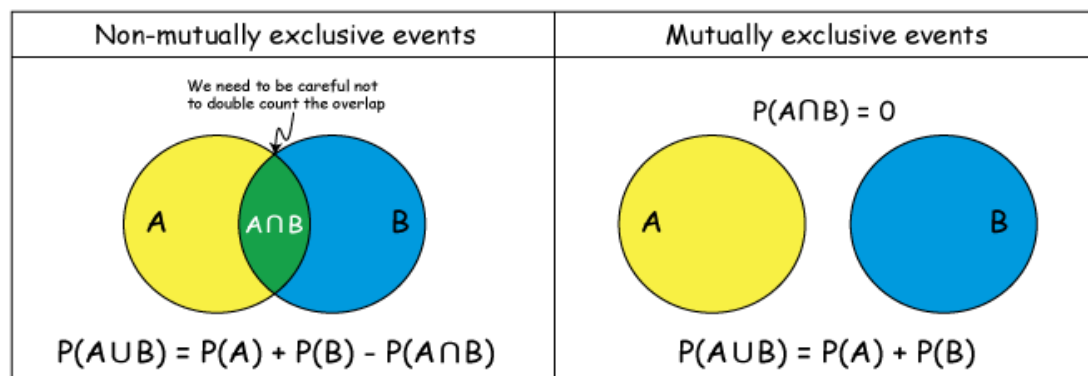
Notations

- 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

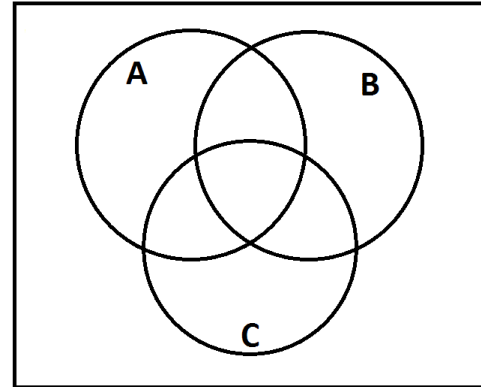
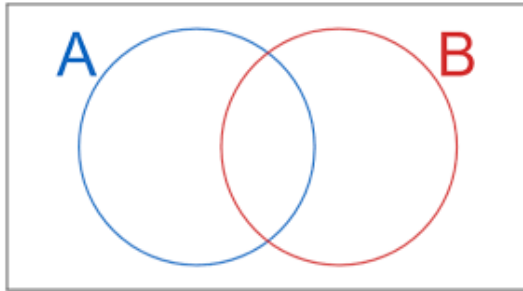
1. $P(S) = 1$
2. $P(A) \geq 0$
3. A_1, A_2, \dots are mutually exclusive,

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

Properties:

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

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



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:

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

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

Property:

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

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

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

Definition of Independence (集合間的獨立)

A and B are independent if and only if

$$\Pr(A | B) = \Pr(A) \quad (\text{or} \quad \Pr(B | A) = \Pr(B))$$

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

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

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

Q: Are A_1 and A_2 independent? No!

$$P(A_2 | A_1) = 0.01 / 0.18 = 5.56\% > P(A_2) = 2\%$$

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?

$$\begin{aligned}
 A &= \text{hypertension is present} \\
 B &= \text{feel that hypertension is Not present} \\
 P(A) &= 0.15 \quad P(B) = 0.75 \\
 P(A \cap B) &= 0.06 \\
 Q: P(A|B) &= ? = \frac{P(A \cap B)}{P(B)} = \frac{0.06}{0.75} \\
 &= 8\% \\
 Q: P(B^c|A) &= \frac{P(A \cap B^c)}{P(A)} = \frac{0.15 - 0.06}{0.15}
 \end{aligned}$$

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

$$\begin{aligned}
 P[A_2|A_1] &= \frac{P[A_1 \text{ and } A_2]}{P[A_1]} \\
 &= \frac{.15}{.55} = \frac{15}{55} = .27
 \end{aligned}$$

Claim: If two sets are “mutually exclusive”, they can’t be independent

Proof:

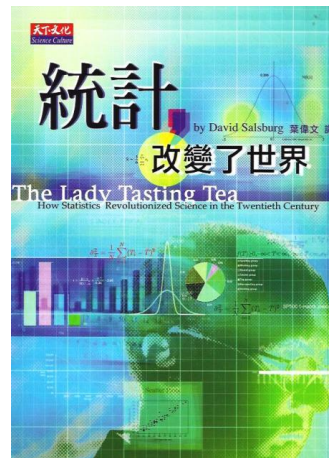
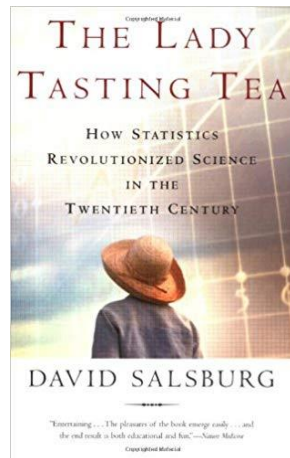
Know that $A \cap B = \emptyset$ with $\Pr(A) > 0$ and $\Pr(B) > 0$.

By definition:

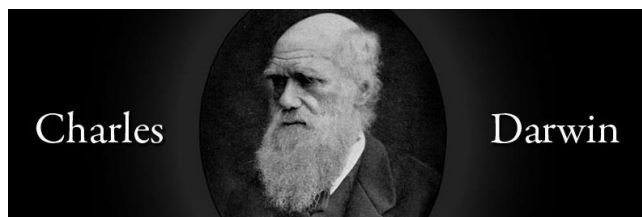
$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = 0 \neq \Pr(A).$$

History of Statistics

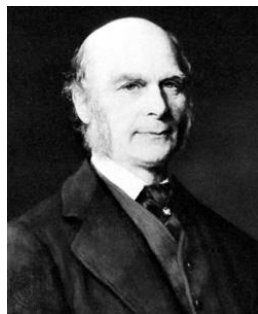
參考書籍：統計改變世界 - 天下文化



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



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



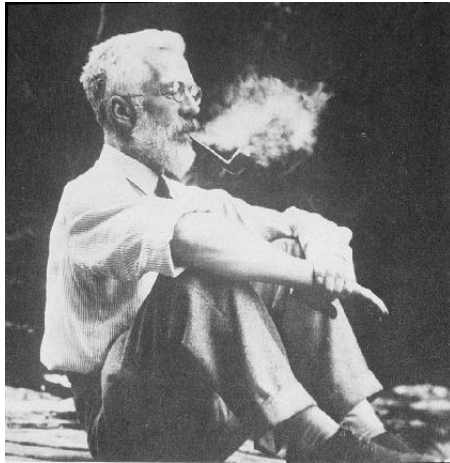
- created the statistical concept of correlation and widely promoted regression toward the mean. (迴歸分析)
 - X-axis:: parent's measurement, Y-axis: offspring's measurement
 - Meaning of slope: slope < 1 → regress toward the mediocrity
- was the first to apply statistical methods to the study of human differences 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



- 歷史上最具有影響力的統計學家 – 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