Lecture 1
09/15/2021

Reference: Syllabus

I. Objectives

The knowledge taught in this course will help you develop the skills of analyzing the data and understand the fundamental knowledge of statistical inference.

II. Prerequisite

* basic probability and calculus

III. Introduction: What is "Statistics"

Population: of interest but unknown

- Too big
  - Number of bird species in Taiwan
- Future
  - Who will be the next president of Taiwan?

Sample (data): obtained from survey or collected via "experiments"

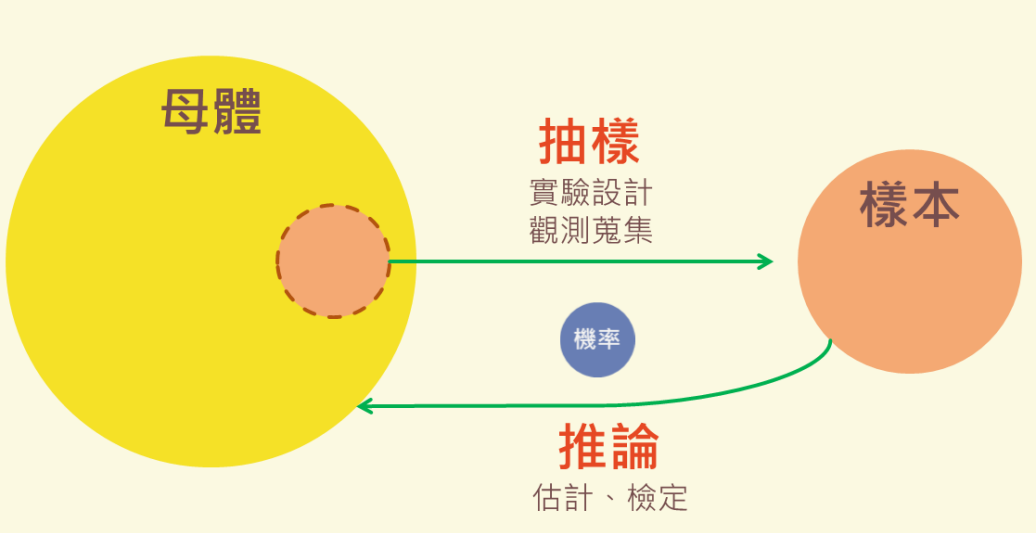
- observational study → researchers do not intervene
  - It is difficult to justify causal relationship.
- experimental data → research can control the confounding variables
  - Causal analysis can be established.

Inference: require the sample to be a good representative of the population

- Parameter estimation (point and interval estimation)
- Hypothesis testing (Hypothesis: the researcher want to test it!)

Probability provides the basis for statistical inference

- Poll (Based on only 1000-1500 people, the result can be used to make inference about the US presidential election)



Historical Development

- Past: it was expensive to collect data
  - Sampling is important
- Now: data are easy to obtain due to technology advancement (internet data, image data , DNA data…)
  - Data are high-dimensional, huge, not well formatted
  - **Big data**（海量資料）

    http://techorange.com/2012/10/05/data-scientists-the-definition-of-sexy/

Topic 1: 資料的描述 (Exploring Data)

Data types → affect the appropriateness of statistical methods

- quantitative (numeric): 可以做算術運算 $(+ - \times \div \sqrt{\ } \log(.))$

  身高 height

  成績 score

  血壓 blood pressure

- qualitative (categorical)

  年級別 (Undergraduate/MS/Ph.D.)

  Urban/rural

  gender (male/female)

  diseased/healthy

How to describe the distribution of data?

- qualitative: pie chart, bar chart
- quantitative:

  * histogram:

  * stem-leaf plot (back-to-back; split the stem …) → useful

Example: # of homeruns (如何比較兩位打擊者的紀錄?)

貝比魯斯 (Babe Ruth): 1920 ~ 1934 n = 15 (Odd)

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

麥奎爾 (McGwire): 1987-1998 n = 12 (Even)

49 32 33 39 22 42 9 9 39 52 58 70

Data: $(X_1,...,X_n)$ → **ordered data** $X_{(1)} \le ... \le X_{(n)}$

由小排到大後:

Ruth (n = 15)

22 25 34 35 41 41 46 46 46 47 49 54 54 59 60

McGwire (n = 12)

9 9 22 32 33 39 39 42 49 52 58 70

Back-to-back stem-leaf plot (in class)

Descriptive Measures

Five number summary

Min – 最小值

Q1 – 最低的 1/4

Median (M) – 一半在 M 之前, 一半在 M 之後

Q3 – 最高的 1/4

Q1 – 最低的 1/4

Max – 最大值

中位數 (median): 最中間的 (或最中間兩個的平均)

$n =$ odd（奇數）→ $M = X_{(\frac{n+1}{2})}$

$n =$ even（偶數）→ $M = \frac{X_{(n/2)} + X_{(n/2+1)}}{2}$

Q1, Q3 的算法不只一種版本, 上課所用的算法如下

Case1: $n =$ 奇數

- Q1: 前半部資料 $X_{(1)},...,X_{(\frac{n+1}{2})}$ 的中位數（“前半部”含中位數）
- Q3: 後半部資料 $X_{(\frac{n+1}{2})},...,X_{(n)}$ 的中位數(“後半部”含中位數)

Case2: $n =$ 偶數（中位數並非資料點,除非有 tie 的情形才發生值剛好一樣）

- Q1: 前半部資料 $X_{(1)},...,X_{(n/2)}$ 的中位數（“前半部”不含中位數）
- Q3: 後半部資料 $X_{(n/2+1)},...,X_{(n)}$ 的中位數(“後半部”不含中位數)

Ruth: n = 15, Min = 22, Max = 60

Median = $\frac{X_{(6)} + X_{(7)}}{2} = 46$

Q1 = $\frac{X_{(4)} + X_{(5)}}{2} = \frac{35+41}{2} = 38$

Q3 = $\frac{X_{(11)} + X_{(12)}}{2} = \frac{49+54}{2} = 51.5$

McGwire: n = 12, Min = 9, Max = 70

Median = $\frac{X_{(6)} + X_{(7)}}{2} = \frac{39+39}{2} = 39$

Q1 = $\frac{X_{(3)} + X_{(4)}}{2} = \frac{22+32}{2} = 27$

Q3 = $\frac{X_{(9)} + X_{(10)}}{2} = \frac{49+52}{2} = 50.5$

Important Features of a Stem-Leaf Plot (讀圖時要注意哪些特徵)

- The shape of the distribution (symmetric or skewed)

- Center or other important locations (Min, Q1, Median, Q3, Max)

- Peak

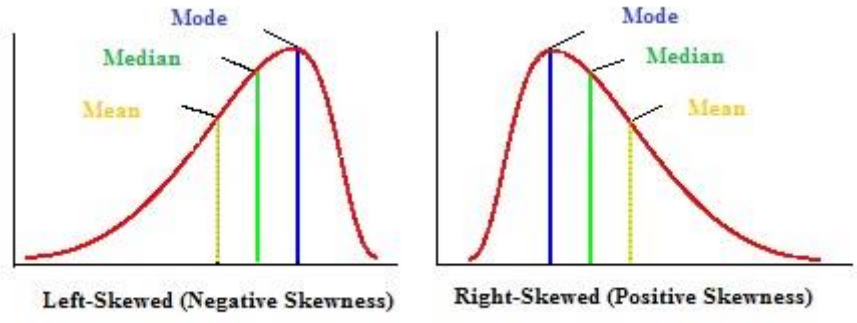
- Possible outliers

Important Measures and Their Practical Applications

1. Shape of the distribution: 看整體分佈的形狀

symmetry (對稱) or skew (偏斜)

"skew to the left" "skew to the right"



Example: The distribution of exam scores (考試分數)

- Too easy: skew to the left (大部分高分，但有少數人很低分)
- Too difficult: skew to the right (大部分低分，但有少數人很高分)
- symmetric (the exam contains both difficult and easy questions)

2. Location parameters: 重要位置 (Center, Peak, Q1, Q3, PR99, 頂標,...)

** Mean or median*

a. 平均數 (mean, average)

$$\bar{X} = (X_1 + ... + X_n) / n$$

b. 中位數 (median)

一半比 median 大, 一半比 median 小

Comparison between mean and median

Properties	*Mean*	*Median*
Data Information	All ☺	Less ☹
Robustness 穩健性	No ☹	Yes ☺

名詞解釋

Robustness: 穩健性 → resistant to outlier (不易受極端值影響)

改良版: Modification

Trimmed mean 切尾均值 (去除 outlier 後再算 mean)

資料智慧化：利用資料科學，將資訊化為創見(電子書): John W.

以下是一些中央計算值，在面對一維異數時，優於平均值：

- 中位數（Median）：沒錯，排在第 50%的資料
- 中樞紐（Midhinge）：第 25%和第 75%的資料取平均
- 三均值（Trimean）：中位數與中樞紐的平均。我喜歡用這個，因為聽起來很厲害。
- 切尾均值（Trimmed/truncated mean）：一樣是平均值，但是把頭尾 N 個資料點剔除，或者頭尾 N 百分比的資料點剔除。在運動界常見這種作法（如體操競賽，剔除最高分與最低分）。如果你剔除頭尾各 25%，把資料靠中間的 50%做平均，那就成了四分位平均（interquartile mean，IQM）。

A. Estimating national bias by modeling the heteroscedasticity of judging errors in gymnastics

Judging in gymnastics is a noisy process and does not rely on comprehensive technical assistance. Athletes are evaluated live by panels of judges, and the final scores aggregate the individual marks given by these judges. This aggregation process typically uses the median or the trimmed mean to remove outliers and improve the accuracy of the overall evaluation. Judges

以下舉數個 mean 不適合的例子

1. 所得新聞: https://www.twreporter.org/a/unequal-distribution-of-income

《所得貧富落差 5 倍… 新竹兩個世界》

新竹市關新里高樓大廈林立，是全國所得中位數最高的里。 記者張念慈／攝影

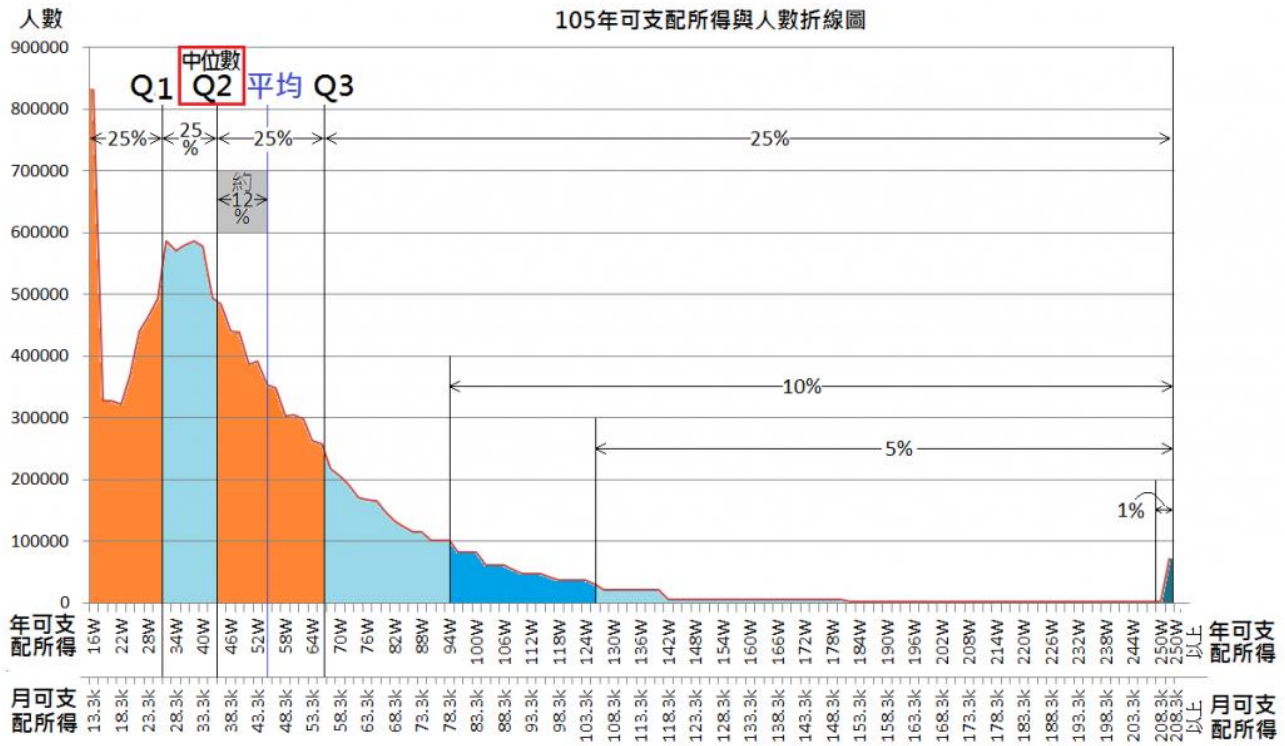
財政部日前公布一〇四年度全台所得總額中位數村里統計分析，新竹縣市包辦全國前十四名，榜首竹市關新里二四二點五萬元，相距不到廿公里「不山不市」的橫山鄉豐鄉村僅四十二點九萬元，是竹縣市最低，差距高達五倍，也讓人感嘆「一個新竹、兩個世界」。

▲新竹市關新里隨處可見高樓大廈，乾淨整潔的街道。（圖／翻攝

https://www.taiwannews.com.tw/ch/news/3356450

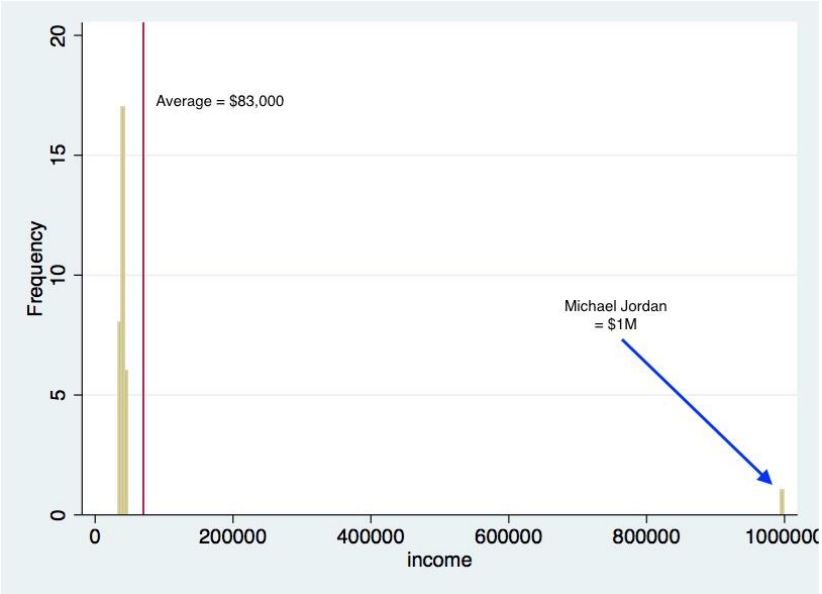
〈時評〉不要再看平均所得、要看中位數所得

By 前太空中心主任 吳作樂 2018/02/05 14:35



2. Example: the most extreme case

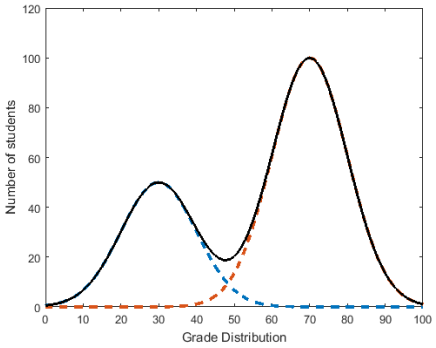
In the case of Michael Jordan and fellow UNC geography graduates, the average is not a good representation of the true center of the data.



** peaks (高峰):*
哪個位置最多人

two peaks → bimodal data (English proficiency scores of Taiwanese students)

雙峰: 70 分 與 30 分



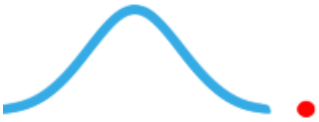
中學階段數學差異化教學研究的初步成果 - 國家教育研究院
https://epaper.naer.edu.tw › edm
歷年來相關的國際評比，我國學生數學學習成就表現經常名列前茅，卻有著學習成就的雙峰現象，且高、低數學學習差距有隨著年級愈益擴大的趨勢，此一現象並非僅 ...

愛學網- -「國際教育心動線」105年第36集學習成就的雙峰現象
https://stv.moe.edu.tw › co_video_content
本院【國際教育心動線】第36集「學習成就的雙峰現象」專輯，由本院課程及教學研究 ... 中學余采玲校長，一起來探討目前我國國民中學階段數學及英語學習成就的雙峰 ...

PISA：台灣數學學習M型化/教育現場/教育趨勢/2015-04-23/即時 ...
https://www.parenting.com.tw › 教育現場 › 教育趨勢
2013年12月3日 - 換言之，測驗對象的高一學生（十年級），數學程度可能只有小學三至四 ... 台灣數學學習落後的學生比例比其他國家高，呈現嚴重的M型雙峰分布。

該給李遠哲、李安一樣的數學課嗎？| 教育家部落格

* *Outliers (離群值, 與眾不同的點)*
find out the reason of outlier
typos or others



資料模型演算判別異常主動發現惡意潛伏
網管人雜誌 - 2019年6月12日
也就是前述分群、分類，以及迴歸分析，透過極端值或稱為離群值（Outlier）分析，檢查當下狀態相較於之前的數值差異，不論過高或過低，皆可判斷為 ...

會計師看時事／內部稽核大數據幫大忙
udn 聯合新聞網 - 2019年4月4日
透過數據分析將原始資料進行全母體查核、發現其中時間順序異常、勾稽不符、離群值、趨勢偏離、比例過高或過低等各種類型風險警訊，提出相關 ...

人工智慧非資安偵測萬靈丹，用對方法才能避開高誤判陷阱
iThome Online - 2018年12月14日
前者需要在事件加上適合的標記，才能學習，而且這些資料本身是有前後脈絡的，不易找出特徵、特色；而後者若是用離群值來判斷，變化也會跟著變 ...

3. Dispersion parameters

Dispersion or variation (離散程度) - 依不同專業解讀其意義

"不患寡而患不均" → 變異大不好

"貧富差距大是社會動亂的根源之一" → 變異大不好

"生物的多樣性卻是物種穩定的基礎" → 變異大好

a. Range = Max – Min

b. Interquartile range (IQR): Q3 – Q1 (中間 50%的範圍)

c. Variance (變異數, S^2) and standard deviation (標準差, S)

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

$$S = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} \quad \text{或是} \quad S = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n} \rightarrow \text{單位與 } X_i \text{ 相同}$$

Variance and standard deviations are not robust

例子: 標準差的意義

[PISA 數學 台學生差異拉大] 2013 年 12 月 3 日

國科會與教育部今天共同公布 PISA 2012 年各國學生表現結果，台灣「數學素養」排名第 4，但個別差異的幅度有持續擴大的趨勢。

國際學生能力評量計畫（PISA）由經濟合作暨發展組織（OECD）主辦的全球學生評量，目的是瞭解 15 歲學生參與社會所需的關鍵知能，自 2000 年起每 3 年進行一次調查。

國際學生能力評估計劃

Programme for International Student Assessment	
簡稱	PISA
成立時間	1997年
目標	世界各地的教育程度比較
總部	OECD總部
地點	法國巴黎 Cedex 16 安德烈·帕斯卡路2號
服務地區	世界
會員	59個政府教育部門

在「數學素養」方面，台灣學生平均分數 560 分，排名第 4，比 2009 年分數進步 17 分。其中台灣學生最擅長的是「空間與形狀」，拿到 592 高分，居世界第 2。

但在優秀的分數表現下，PISA 結果也呈現隱憂。台灣學生數學的個別差異幅度越來越明顯，**2006 標準差 103，2009 年標準差成長為 105，已高居世界第一，2012 標準差更成長為 116，比差距第二大的國家（105）高出許多。**

以性別分析，台灣男、女學生 PISA 數學素養沒有顯著差異，但男學生的個別差異現象比女學生明顯。

教育部次長陳德華表示，台灣學生整體數學表現優異，但對於明顯落後的學生，介入已是「刻不容緩」。教育部推動十二年國教，適性、精緻的補救教學是施政重點，其中關鍵是激起學習動機，避免任何學生放棄數學，而是都能在學習中，獲得成就感。