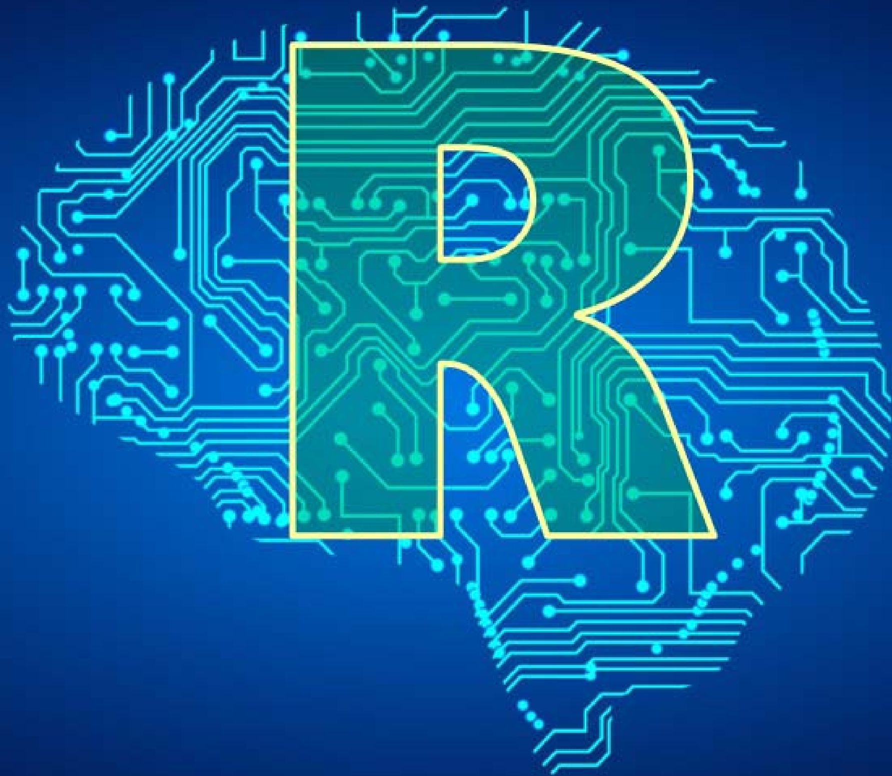
台灣人工智慧學校



敘述統計

吳漢銘
國立臺北大學 統計學系

http://www.hmwu.idv.tw

本章大綱

- 資料分析工具: R
- 傳統統計
  - 敘述性統計
  - 推論統計
- 統計/資料探勘/數據科學/資料科學
- 描述資料: 中心趨勢，分散程度
- 相關係數
- 共變異數矩陣
- HDLSS Problem

為什麼要使用R做為資料分析工具?

Why R?

- R is a high-quality, cross-platform, flexible, widely used open source, free language for statistics, graphics, mathematics, and data science.
- R contains more than 5,000 algorithms (>10,000 packages) and millions of users with domain knowledge worldwide.

http://www.r-project.org

https://www.rstudio.com/

TIOBE 全球程式語言排名

TIOBE Index for January 2018

January Headline: Programming Language C awarded Language of the Year 2017

Jan 2018	Jan 2017	Change	Programming Language
1	1		Java
2	2		C
3	3		C++
4	5	^	Python
5	4	v	C#
6	7	^	JavaScript
7	6	v	Visual Basic .NET
8	16	^^	R
9	10	^	PHP
10	8	v	Perl

http://www.tiobe.com/tiobe-index/
(共243種程式語言)

What is Statistics?

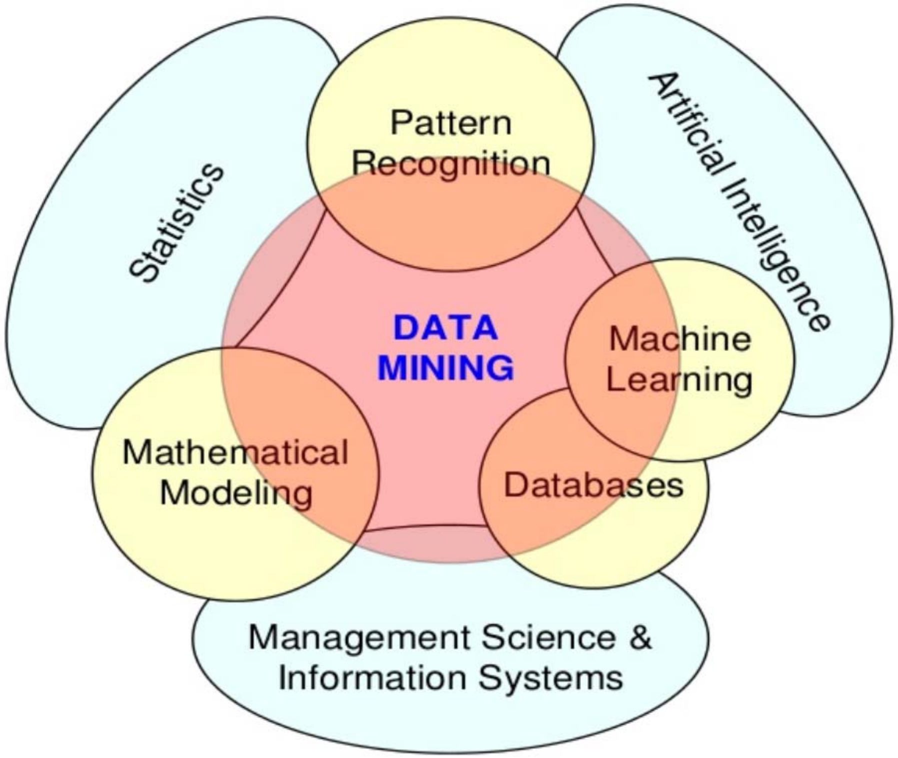
- Merriam-Webster dictionary defines statistics as "a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

- 傳統統計(歷史源自17世紀), 分兩類:
  - 敘述統計 (Descriptive statistics):
  - 推論統計(Inferential statistics): It uses patterns in the sample data to draw inferences (estimation, hypothesis testing) about the population represented, accounting for randomness.

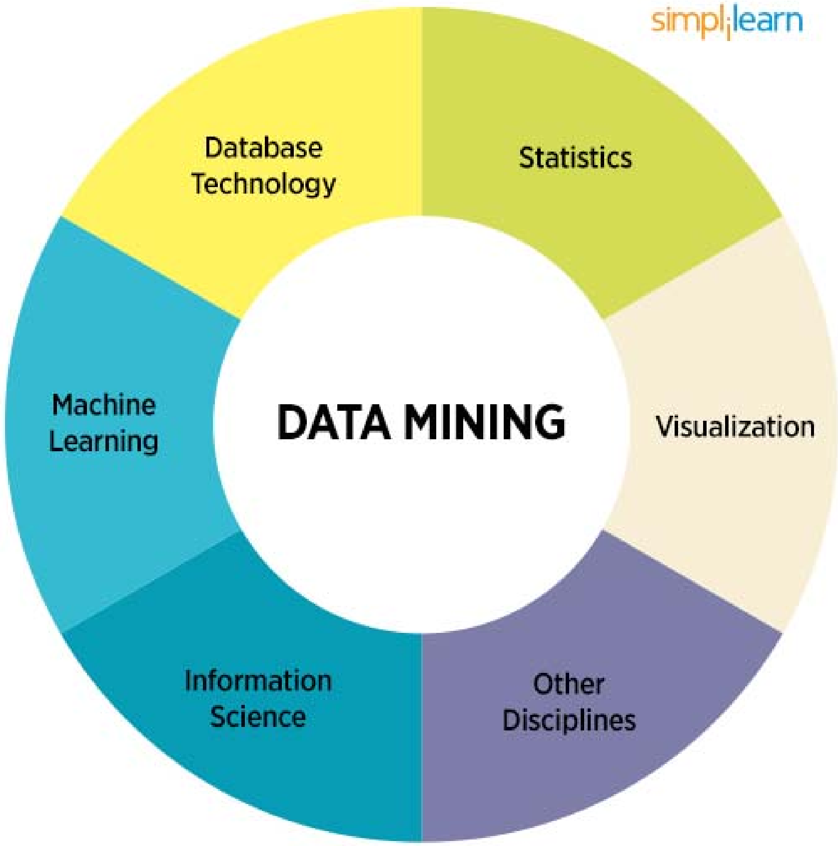
- 統計研究領域的分類: 數理統計、工業統計、商用統計、生物統計等等。

http://www.theusrus.de/blog/some-truth-about-big-data/

Data Mining Diagrams

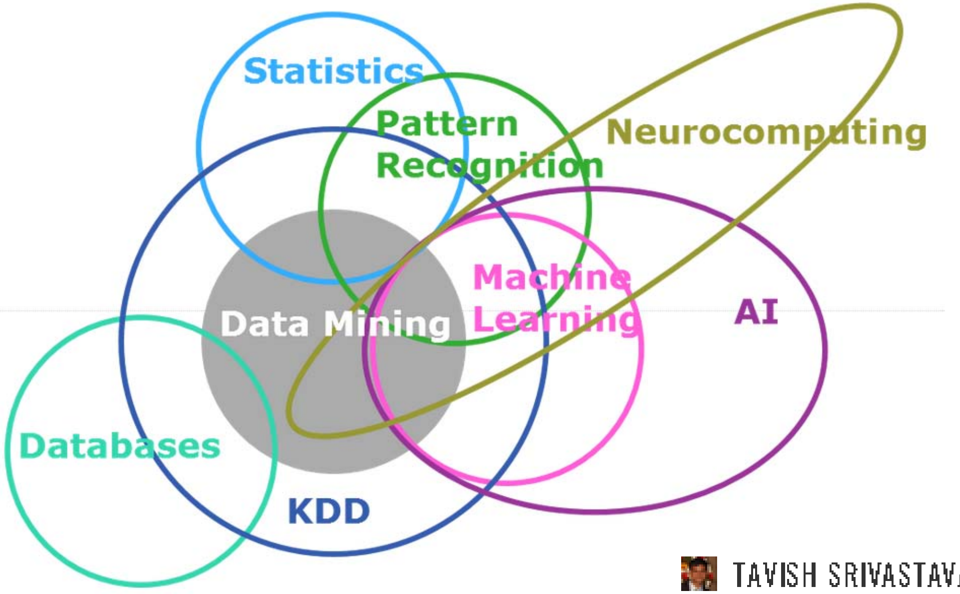


Source: Published on Nov 26, 2014
Language Technologies for Geomatics: From Intelligence to Agility
Published in: Technology
http://www.slideshare.net/VisionGEOMATIQUE2014/gagnon-20141112vision



Source:
http://www.simplilearn.com/data-mining-vs-statistics-article

Difference between Machine Learning & Statistical Modeling



Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering

TAVISH SRIVASTAVA , JULY 1, 2015

https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

- **Machine Learning** is an algorithm that can learn from data without relying on rules-based programming.
- **Statistical Modelling** is the formalization of relationships between variables in the form of mathematical equations.

機器學習和統計楷型的差異
http://vvar.pixnet.net/blog/post/242048881
為什麼統計學家、機器學習專家解決同一問題的方法差別那麼大?
https://read01.com/EBPPK7.html
深度 | 機器學習與統計學是互補的嗎？
https://read01.com/ezQ3K.html

Statistics, Data Mining and Big Data

	Statistics	Data Mining	Big Data
Structure	structured	structured	unstructured
Size	small	large	very large
Generation	planned	transactional	behavioral
Aim	understand	optimize business	generate business
Privacy Issues	non	minor	huge
Founded On	concepts & theory	technology & tool	technology & tools
Marketing	bad	good	perfect

Source: http://www.theusrus.de/blog/some-truth-about-big-data/

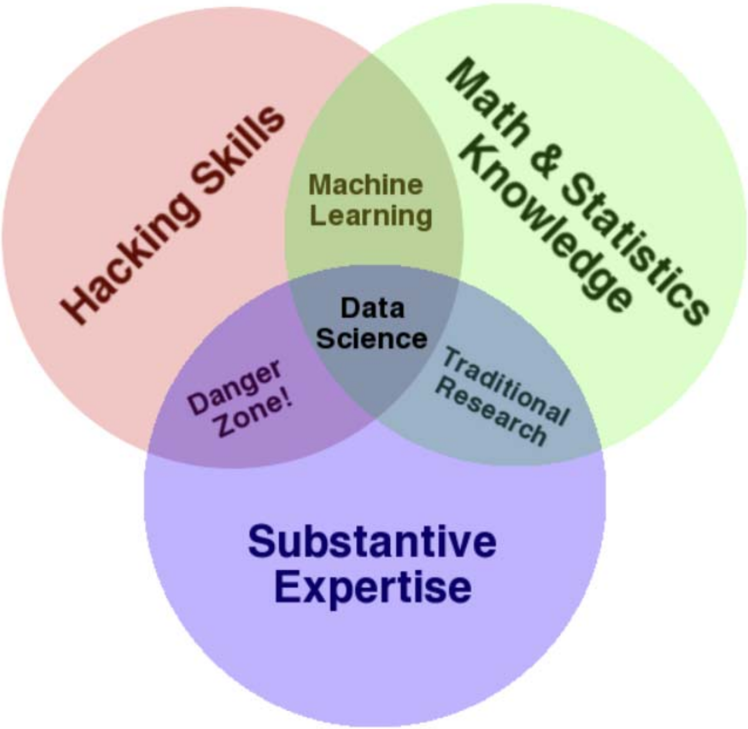
小數據與大數據的區別

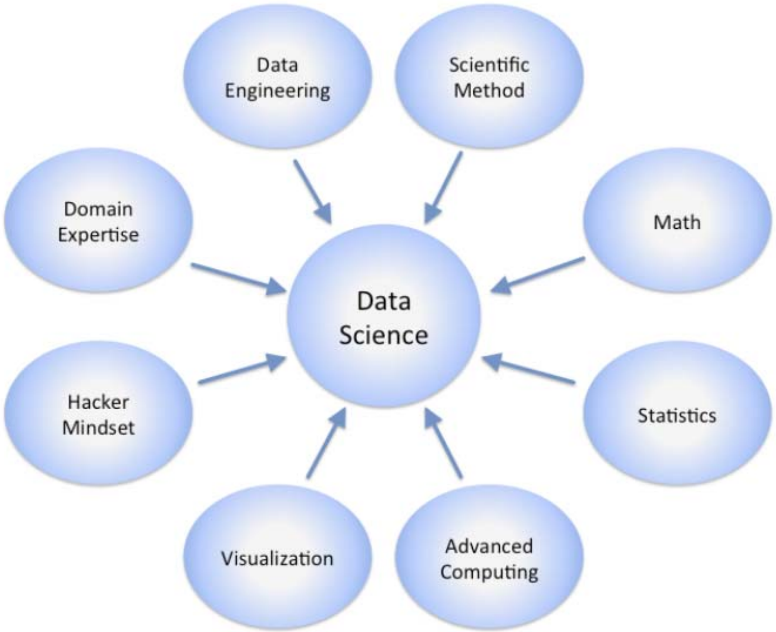
- **調查資料**
  - 抽樣的
  - 樣本反饋的
  - 主觀的
  - 結果的
  - 結構化的
  - 斷點的
- **監測資料**
  - 全樣的
  - 監測紀錄
  - 客觀的
  - 過程的
  - 非結構化的
  - 連續的

數據科學 Data Science

The Data Science Venn Diagram

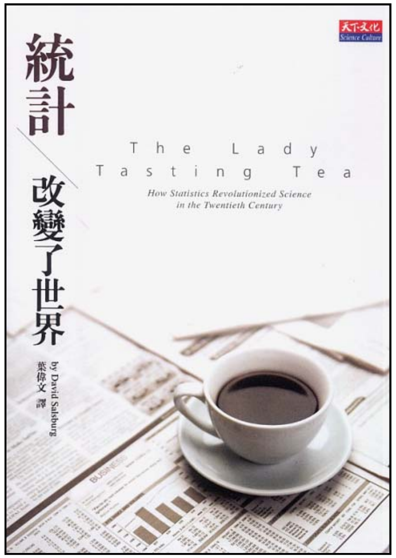
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

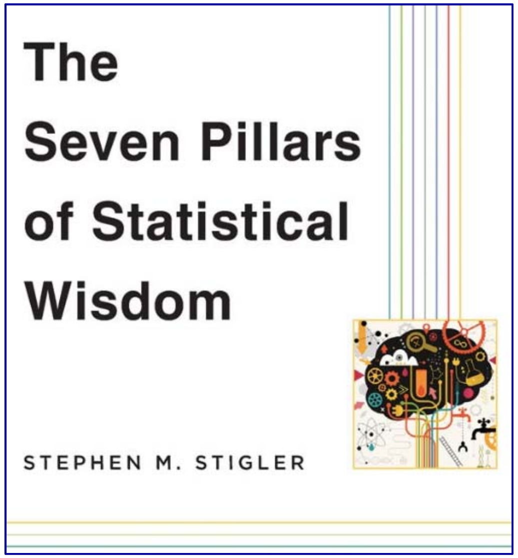




Source: By Calvin.Andrus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

推薦兩本書





1 **AGGREGATION** From Tables and Means to Least Squares

2 **INFORMATION** Its Measurement and Rate of Change

3 **LIKELIHOOD** Calibration on a Probability Scale

4 **INTERCOMPARISON** Within-Sample Variation as a Standard

5 **REGRESSION** Multivariate Analysis, Bayesian Inference, and Causal Inference

6 **DESIGN** Experimental Planning and the Role of Randomization

7 **RESIDUAL** Scientific Logic, Model Comparison, and Diagnostic Display

(March 7, 2016)

趙民德，1999，「統計已死，統計萬歲！」第八屆南區統計研討會演說稿

趙民德
台灣

趙民德，國立台灣大學數學系畢業、美國加州大學柏克萊分校統計博士。在美國求學及工作多年後，1982年回台灣籌設中央研究院統計學研究所，該所於1987年正式成立，並正名為統計科學研究所。國內統計學有今日的發展，以及能在世界佔一席之地，功不可沒。

在文學成就上，名家王鼎鈞以「詩的精緻，劇的張力，散文的鋪陳」肯定其業餘小說家的地位。

統計有沒有死？會不會萬歲？

只要有米倉，就會有老鼠；只要有數據，就會發展處理數據的方法。
但是不是叫做統計學、或者叫做computer science 的data mining，就要看這一代的統計人如何因應變局。

Types of Data Scales

- Categorical (類別資料), discrete, or nominal (名目變數) — Values contain no ordering information: 性別、種族、教育程度、宗教信仰、交通工具、音樂類型... (qualitative 屬質)

- Ordinal (順序) — Values indicate order, but no arithmetic operations are meaningful (e.g., "novice", "experienced", and "expert" as designations of programmers participating in an experiment); 非常同意，同意，普通，不同意，非常不同意; 優，佳，劣。

- Interval — Distances between values are meaningful, but zero point is not meaningful. (e.g., degrees Fahrenheit)

- Ratio (Continuous Data 連續型資料)— Distances are meaningful and a zero point is meaningful (e.g., degrees K, 年收入、年資、身高、... (quantitative 計量)

- **Ordinal** methods cannot be used with nominal variable
- **Nominal** methods can be used with nominal, ordinal variables.

資料描述

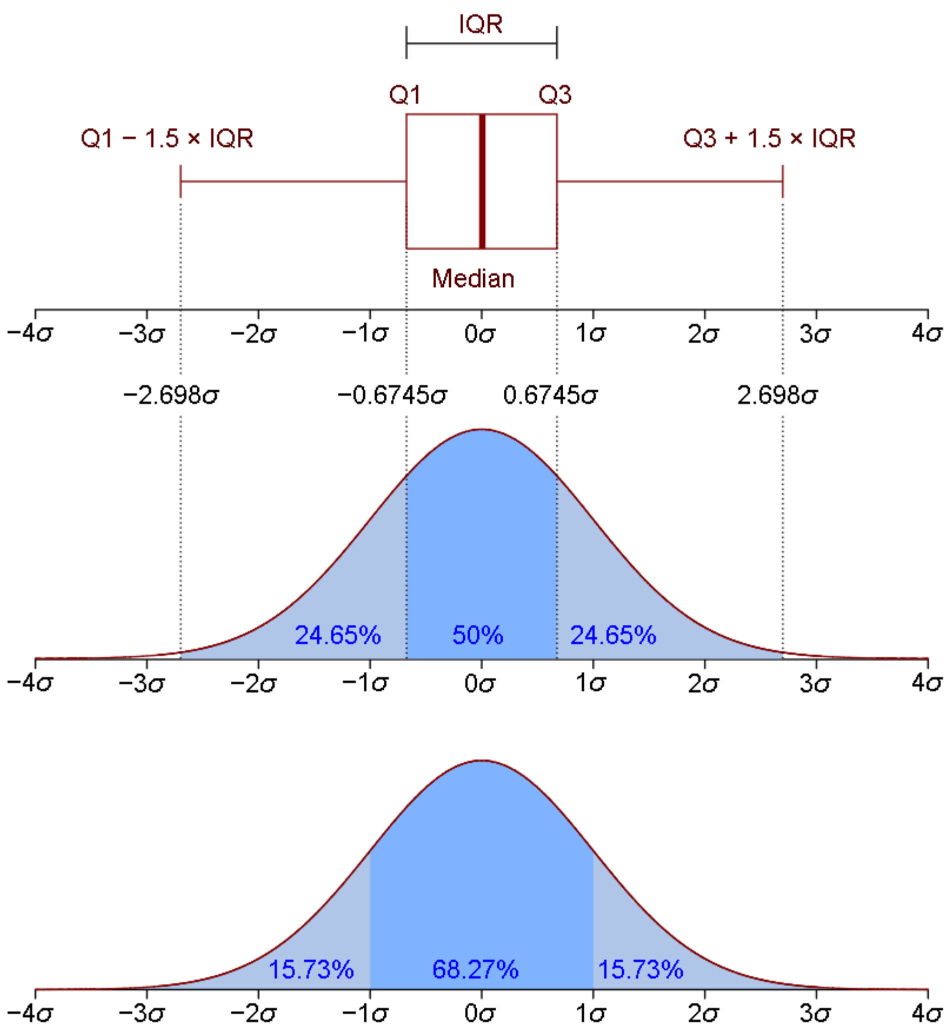
- **資料中心趨勢:**
  平均數(average)
  眾數(mode)
  中位數(median)
- **資料分散程度:**
  四分位數(Quartile)
  全距(range)
  四分位距(interquartile range, IQR)
  百位數(percentile)
  標準差(standard deviation)
  變異數(variance)

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$n =$ The number of data points

$\bar{x} =$ The mean of the x_i

$x_i =$ Each of the values of the data



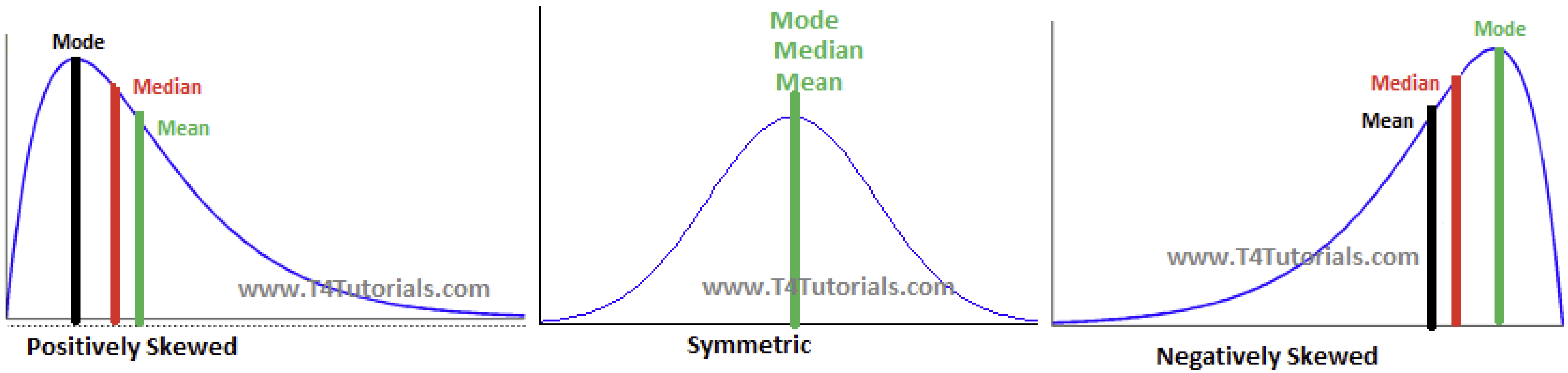
https://zh.wikipedia.org/wiki/四分位距

資料描述:偏態係數

- 偏態(skewness):

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}^{\,3}}$$

大於0：右偏分配
等於0：對稱分配
小於0：左偏分配

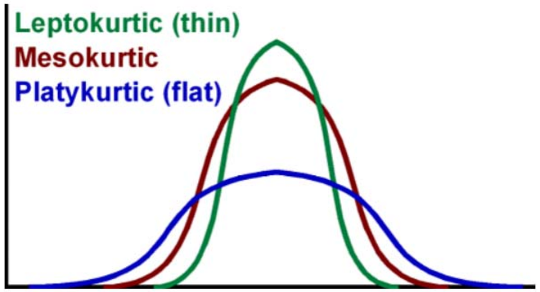


http://www.t4tutorials.com/data-skewness-in-data-mining/

資料描述:峰態係數

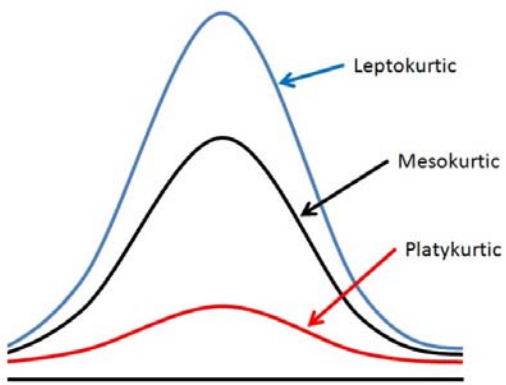
峰度係數 k_c(coefficient of kurtosis) 為一測量峰度高低的量數，可以反映資料的分佈形狀。峰度係數一般是與常態分配作比較而言, 該資料分配是否比較高聳或是扁平的形狀。其判別如下:

- 若 $k_c > 0$, 表示資料分布呈高狹峰 (lepto kurtosis)。
- 若 $k_c = 0$, 表示資料分布呈常態峰 (normal kurtosis)。
- 若 $k_c < 0$, 表示資料分布呈低潤峰 (platy kurtosis)。



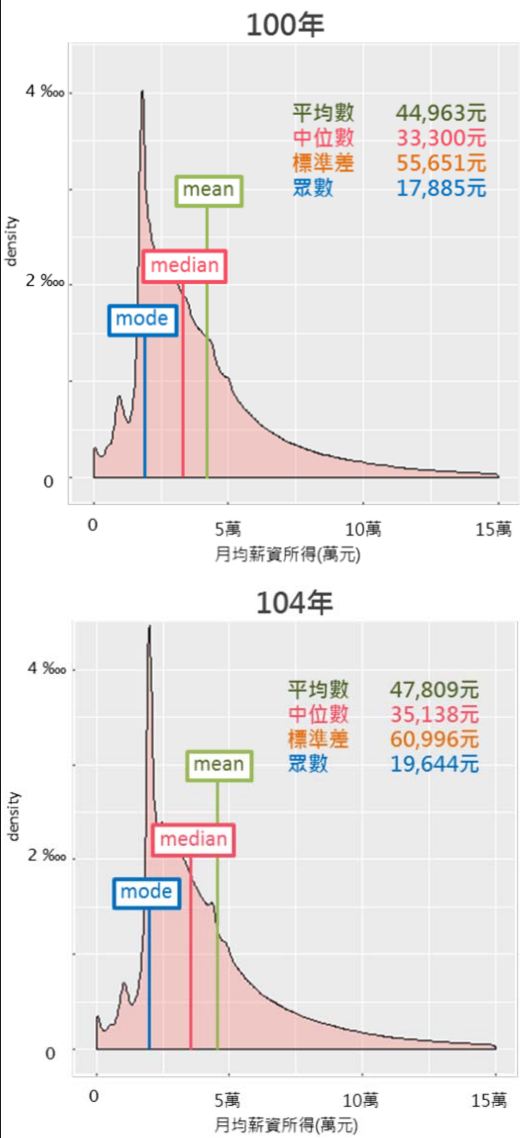
常用的樣本峰度係數的計算式有以下三項:

- The typical definition used in many older textbooks: $g_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)^2} - 3$
- Used in SAS and SPSS: $G_2 = \dfrac{n-1}{(n-2)(n-3)}[(n+1)g_2+6]$
- Used in MINITAB and BMDP: $b_2 = (g_2+3)(1-\frac{1}{n})^2 - 3$



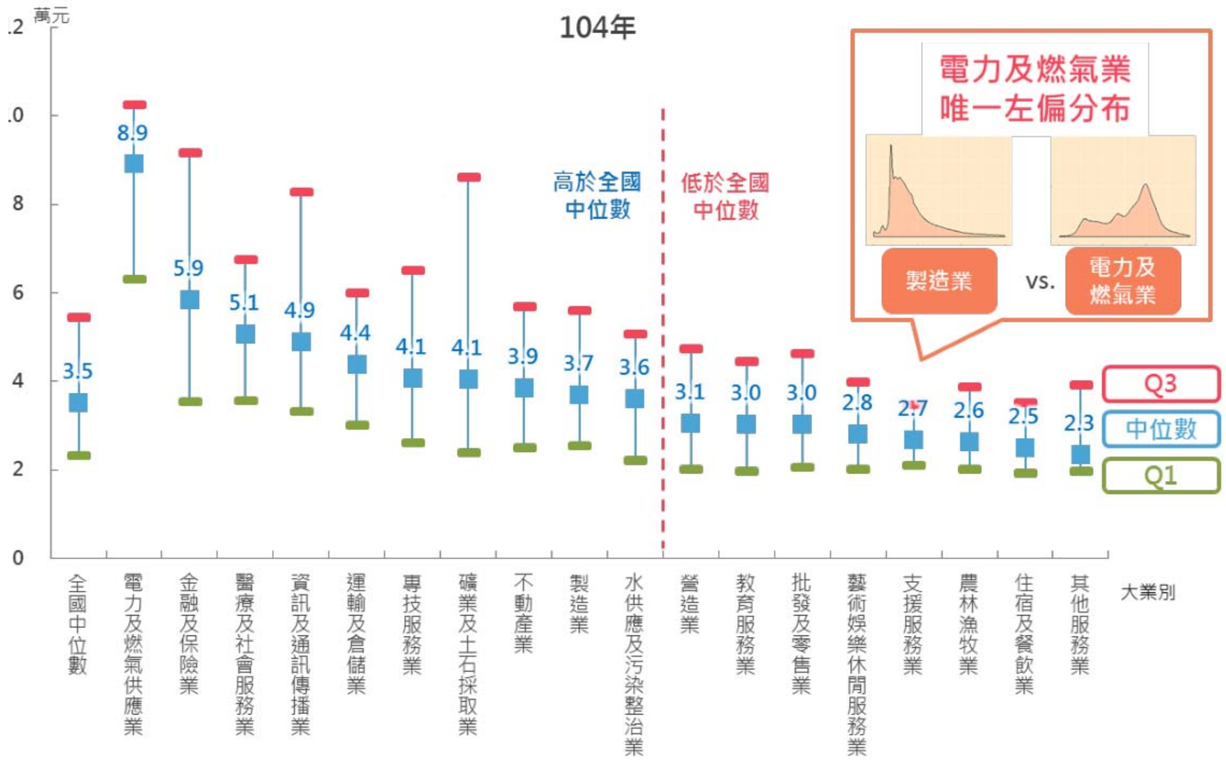
範例: 由財稅大數據探討臺灣近年薪資樣貌

圖 3 月均薪資所得機率分布圖



由財稅大數據探討臺灣近年薪資樣貌 財政部統計處 106年8月
https://www.mof.gov.tw/File/Attach/75403/File_10649.pdf

圖 8 月均薪資所得中位數 - 按大業別分



玩玩看~薪情平臺

行政院主計總處
Directorate-General of Budget, Accounting and Statistics, Executive Yuan, R.O.C. (Taiwan)

熱誠・公正・效率・精確

統計資訊網　答客問

https://earnings.dgbas.gov.tw/

薪 情 互 動

製造業四大產業
況

男女薪資差異

各業薪情概況

R程式練習: 加權算術平均數

有某班學生之微積分成績明細紀錄於資料檔 (score2015.txt) 中，其中成績以 60 分為及格，100 分為滿分，成績空白以零分計。學期總成績計算方法如下: (i) 配分比例為: 小考成績佔 40%(各次小考平均配分)、期中考佔 25%、期末考佔 25%、助教實習課佔 10%，出席次數分數為額外加分，每出席一次，加 2 分 (滿分 18 分); 成績紀錄共 8 項。(ii) 小考成績刪除其中最低分一次。

學號	性別	姓名	小考1	小考2	小考3	小考4	助教	期中考	期末考	出席次數
920541081	女	高婕嘉	0	0	0	36	35	26	25	6
920660451	女	倪儒子	30	0			19	28	0	4
921190391	女	曾翔家	35	35	20	9	19	83	24	6
921530877	女	宋良楹	33	65	60	64	52	69	69	6
921537146	女	吳潔品	35	58	100	77	47	100	84	6
921451012	女	洪銘學	35	13	20	29	55	44	40	8
922030257	女	林雅潔	55	31	40	31	80	74	47	8
922030448	女	朱新太	10	20			49	38	0	7
922030497	女	洪苡彥	50	41	75	86	69	89	59	8
922739223	女	洪文依	78	78	80	88	100	88	84	8

提示: 小考刪除最差一次之後的計分方式, 舉例如下: 若有三次小考分為 60, 30, 90 。配分為 5%, 6%, 7% 。原始得分為 60*0.05 + 30*0.06 + 90*0.07 = 11.1 若刪除最差一次成績後, 所得分數為: (60*0.05 + 90*0.07)*(5+6+7)/(5+7)= 13.95

想想看: 如何決定權重? 維度縮減方法 (e.g., PCA)

R程式練習

```
> score2015.orig <- read.table("score2015.txt", header=T, sep = "\t")
> dim(score2015.orig)
[1] 80 12
> head(score2015.orig)
  座號      學號 性別   姓名 小考1 小考2 小考3 小考4 助教 期中考 期末考 出席次數
1    1 920541081   女 高婕嘉     0     0     0    36   35     26     25        6
2    2 920660451   女 倪儒子    30     0    NA    NA   19     28      0        4
...
6    6 921451012   女 洪銘學    35    13    20    29   55     44     40        8
> summary(score2015.orig[, 3:ncol(score2015.orig)])
 性別         姓名         小考1           小考2           小考3
 女:60   王彥珮 : 1   Min.   : 0.00   Min.   : 0.0   Min.   :  0.00
 男:20   王淳昀 : 1   1st Qu.:25.25   1st Qu.:10.0   1st Qu.: 20.00
          王銘軒 : 1   Median :40.00   Median :30.0   Median : 40.00
          朱新太 : 1   Mean   :40.00   Mean   :28.9   Mean   : 47.76
          何竣育 : 1   3rd Qu.:50.25   3rd Qu.:40.0   3rd Qu.: 80.00
          余馨繁 : 1   Max.   :90.00   Max.   :80.0   Max.   :100.00
          (Other):74   NA's   :4       NA's   :7      NA's   :13
     小考4             助教            期中考           期末考          出席次數
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :1.0
 1st Qu.: 36.00   1st Qu.: 35.00   1st Qu.: 32.00   1st Qu.: 23.75   1st Qu.:7.0
 Median : 67.00   Median : 59.50   Median : 68.50   Median : 50.00   Median :8.0
 Mean   : 56.75   Mean   : 56.24   Mean   : 57.56   Mean   : 46.71   Mean   :7.7
 3rd Qu.: 81.00   3rd Qu.: 75.25   3rd Qu.: 80.25   3rd Qu.: 69.50   3rd Qu.:9.0
 Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :9.0
 NA's   :15
>
> table(score2015.orig["出席次數"])
 1  2  3  4  5  6  7  8  9
 1  1  2  3  3  7  4 21 38
```

R程式練習

```
> score2015 <- score2015.orig
> score2015[is.na(score2015)] <- 0
> colMeans(score2015[, 5:11])
  小考1    小考2    小考3    小考4     助教   期中考   期末考
38.0000 26.3750 40.0000 46.1125 56.2375 57.5625 46.7125
> apply(score2015[, 5:11], 1, mean)
 [1] 17.4285714 11.0000000 32.1428571 58.8571429 71.5714286 33.7142857 51.1428571
 [8] 16.7142857 67.0000000 85.1428571 31.2857143 65.5714286 19.8571429 88.7142857
...
[78]  3.4285714 19.2857143 23.1428571
> apply(score2015[, 5:11], 2, sd)
   小考1     小考2     小考3     小考4      助教    期中考    期末考
23.29883 22.83478 36.26939 35.13014 27.04391 31.00708 30.71848
> x <- score2015[,"小考1"]
> min(x)
[1] 0
> max(x)
[1] 90
> sum(x)
[1] 3040
> mean(x)
[1] 38
> mean(x)
[1] 38
> mean(x, trim=0.1)
[1] 37.45312
> median(x)
[1] 40
```

```
Mode <- function(x, na.rm = FALSE) {
  if(na.rm) x = x[!is.na(x)]
  ux <- unique(x)
  ifelse(length(x)==length(ux),
         "no mode",
         ux[which.max(tabulate(match(x, ux)))])
}
```

```
> Mode(x)
[1] 50
> quantile(x)
  0%  25%  50%  75% 100%
   0   20   40   50   90
> quantile(x, prob= seq(0, 100, 10)/100)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
 0.0  4.5 14.6 27.4 33.6 40.0 45.0 50.0 55.0 68.2 90.0
> range(x)
[1]  0 90
> sd(x)
[1] 23.29883
> var(x)
[1] 542.8354
```

Distance and Similarity Measure

Cov	x1	x2	x3	x4		x p
x1	1.00	0.48	0.10	-0.10		-0.28
x2	0.48	1.00	0.41	0.22		-0.23
x3	0.10	0.41	1.00	0.36		-0.05
x4	-0.10	0.22	0.36	1.00		0.10
x p	-0.28	-0.23	-0.05	0.10		1.00

Proximity Matrix

Data Matrix

x y

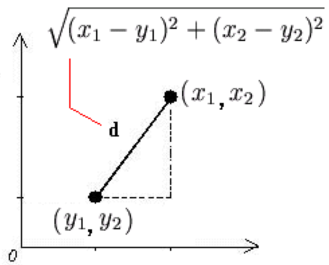
Data	x1	x2	x3	x4	•••	x p
subject01	-0.48	-0.42	0.87	0.92		-0.18
subject02	-0.39	-0.58	1.08	1.21		-0.33
subject03	0.87	0.25	-0.17	0.18		-0.44
subject04	1.57	1.03	1.22	0.31		-0.49
subject05	-1.15	-0.86	1.21	1.62		0.16
subject06	0.04	-0.12	0.31	0.16		-0.06
subject07	2.95	0.45	-0.40	-0.66		-0.38
subject08	-1.22	-0.74	1.34	1.50		0.29
subject09	-0.73	-1.06	-0.79	-0.02		0.44
subject10	-0.58	-0.40	0.13	0.58		0.02
subject11	-0.50	-0.42	0.66	1.05		0.06
subject12	-0.86	-0.29	0.42	0.46		0.10
subject13	-0.16	0.29	0.17	-0.28		-0.55
subject14	-0.36	-0.03	-0.03	-0.08		-0.25
subject15	-0.72	-0.85	0.54	1.04		0.24
subject16	-0.78	-0.52	0.26	0.20		0.48
subject17	0.60	-0.55	0.41	0.45		-0.66
⋮						
subject n	-2.29	-0.64	0.77	1.60		0.55
mean	0.07	-0.04	0.44	0.31	•••	-0.21

Pearson Correlation Coefficient

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$\boxed{\begin{aligned} x &= (x_1, x_2, \cdots, x_n) \\ y &= (y_1, y_2, \cdots, y_n) \end{aligned}}$$

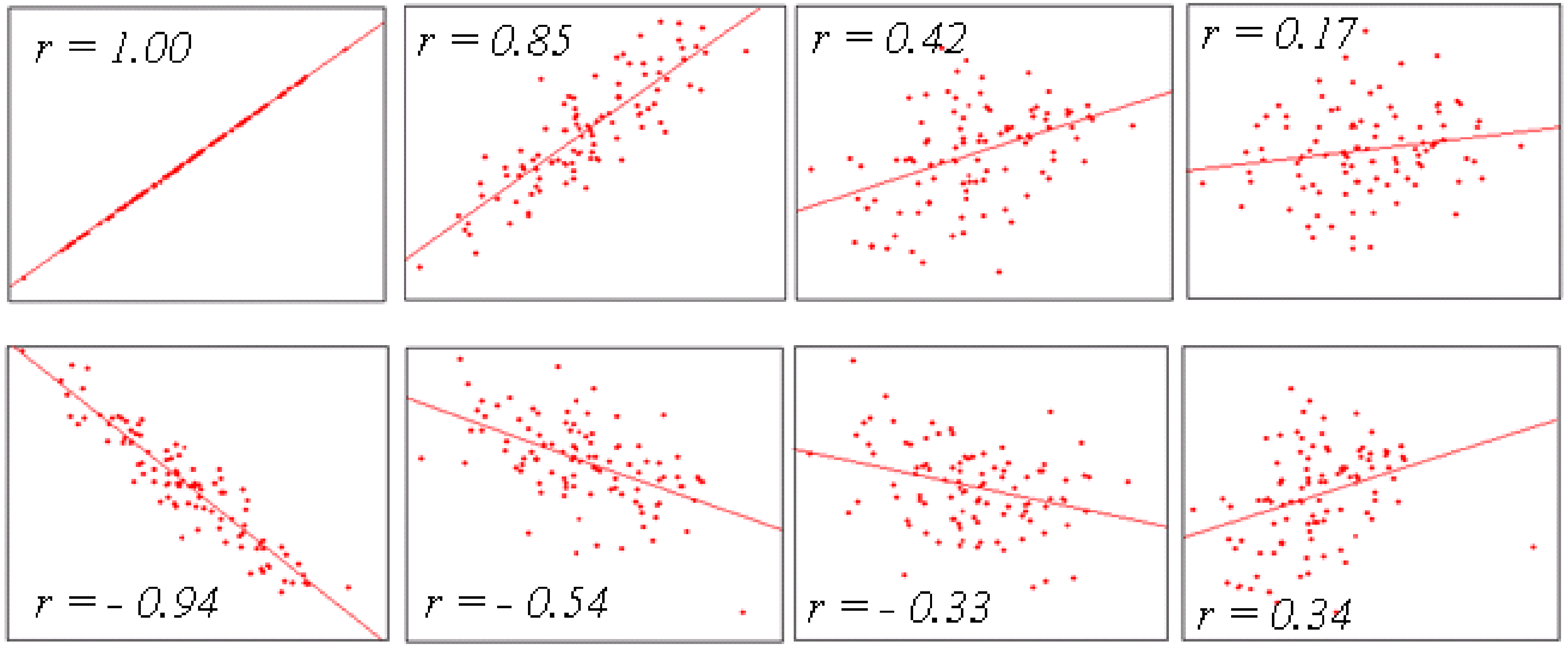
Euclidean Distance



$$d_{xy} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
- Other transformations (Chatfield and Collins 1980, Section 10.2)

Pearson Correlation Coefficient




```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
     method: one of "euclidean", "maximum", "manhattan", "canberra", "binary"
or "minkowski" distance measure.
cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

More Similarity Measures (1/4)

Dissimilarity/Similarity Measure for Quantitative Data

Similarity	Formula
Pearson correlation	$s(i,j) = \dfrac{\operatorname{cov}(x_i, x_j)}{\sqrt{\operatorname{var}(x_i)\operatorname{var}(x_j)}}$
Spearman correlation (r_i is ranked x_i)	$s(i,j) = \dfrac{\operatorname{cov}(r_i, r_j)}{\sqrt{\operatorname{var}(r_i)\operatorname{var}(r_j)}}$
Kendall's Tau	$s(i,j) = \dfrac{1}{\binom{p}{2}} \sum_{k \neq k'} sign\,[(x_{ik} - x_{ik'})(x_{jk} - x_{jk'})]$

All indices range from -1 to +1

Kendall's tau

Two pairs of observation (x_i, y_i) and (x_j, y_j)

- C: concordant pair: $(x_j - x_i)(y_j - y_i) > 0$
- D: discordant pair: $(x_j - x_i)(y_j - y_i) < 0$
- tie:

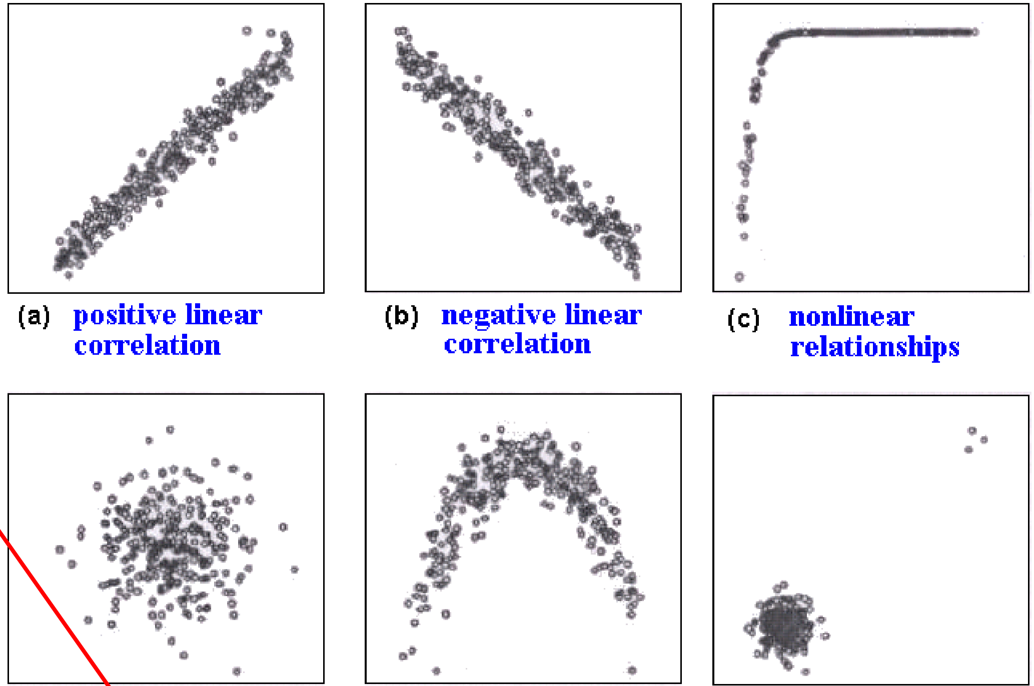
E_y: extra y pair in x's: $(x_j - x_i) = 0$

E_x : extra x pair in y's: $(y_j - y_i) = 0$

$$\tau = \frac{C - D}{\sqrt{C + D - E_y}\sqrt{C + D - E_x}}$$

More Similarity Measures (2/4)

measures the strength of a linear relationship



(a) positive linear correlation
(b) negative linear correlation
(c) nonlinear relationships
(d) no relationship
(e) nonlinear relationships
(f) no relationship with outliers

measure any monotonic relationship between two variables

Data	Pearson's rho	Spearman's rho	Kendall's tau
(a)	0.98	0.98	0.87
(b)	-0.98	-0.98	-0.87
(c)	0.50	0.99	0.98
(d)	-0.02	-0.03	-0.02
(e)	-0.06	-0.02	-0.02
(f)	0.68	0.00	0.00

non-monotonic, fail to detect the existence of a relationship

more robust

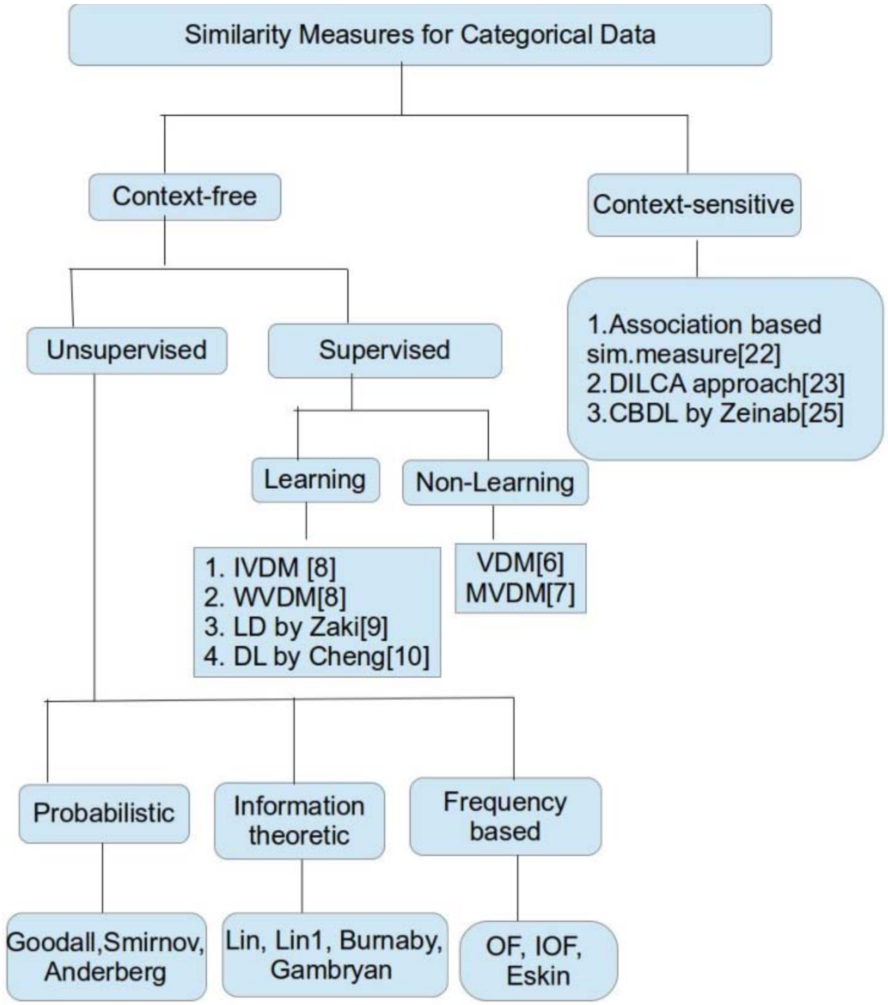
Similarity Measures for Categorical Data

Table 1. Commonly used similarity coefficients for binary data.

Binary Data		Object B: 1	Object B: 0	
Object A	1	a	b	$(a+b)$
	0	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	$(a+b+c+d)$

Similarity	Formula
Braun	$\frac{a}{\max(a+b,\ a+c)}$
Dice	$\frac{2a}{2a+b+c}$
Hamman	$\frac{a+d-(b+c)}{a+b+c+d}$
Jaccard	$\frac{a}{a+b+c}$
Kappa	$\left(1+\frac{(b+c)(a+b+c+d)}{2ad-2bc}\right)^{-1}$
Kulczynskl	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
Ochiai	$\frac{a}{\sqrt{((a+b)(a+c))}}$
Phi	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
Rao	$\frac{a}{a+b+c+d}$
Rogers	$\frac{a+d}{a+2b+2c+d}$
simple match	$\frac{a+d}{a+b+c+d}$
Simpson	$\frac{a}{\min(a+b,\ a+c)}$
Sneath	$\frac{a}{a+2b+2c}$
Yule	$\frac{ad-bc}{ad+bc}$

Taxonomy of Categorical Data Similarity Measures



2014, A survey of distance/similarity measures for categorical data, 2014 International Joint Conference on Neural Networks (IJCNN), 1907-1914.

Sample Variance-Covariance Matrix
Correlation Matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{pmatrix}$$

$s_j^2 = (1/n)\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ is the variance of the j-th variable

$s_{jk} = (1/n)\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ is the covariance between the j-th and k-th variables

$\bar{x}_j = (1/n)\sum_{i=1}^n x_{ij}$ is the mean of the j-th variable

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

eigen-decomposition

$$A\mathbf{v} = \lambda\mathbf{v}.$$

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}}$$

High-dimensional data (HDD)

- Three different groups of HDD:
  - p is large but smaller than n;
  - p is large and larger than p: the high-dimension low sample size data (HDLSS); and
  - the data are functions of a continuous variable d: the functional data.
- In high dimension, the space becomes emptier as the dimension increases
  - when p > n, the rank r of the covariance matrix S satisfies r ≤ min{p, n}.
  - For HDLSS data, one cannot obtain more than n principal components.
  - Either PCA needs to be adjusted, or other methods such as ICA or Projection Pursuit could be used.

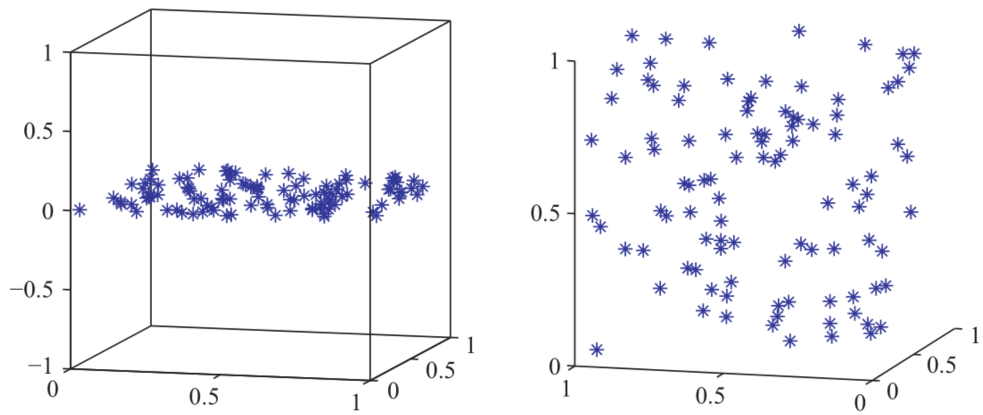
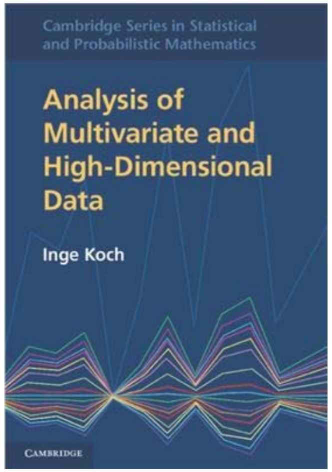


Figure 2.12 Distribution of 100 points in 2D and 3D unit space.



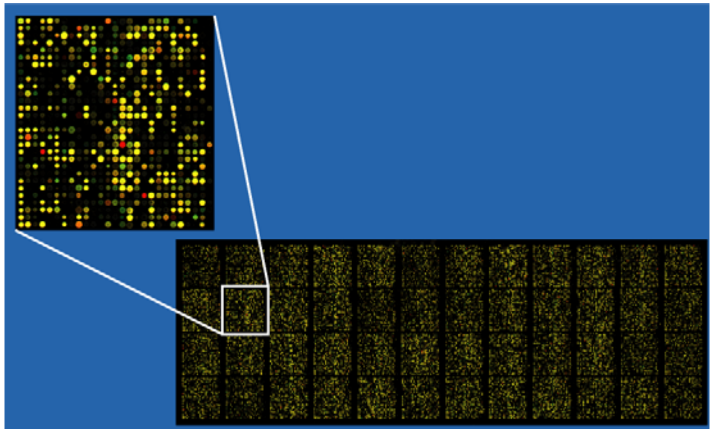
HDLSS examples

Sungkyu Jung and J. S. Marro, 2009, PCA Consistency In High Dimension, Low Sample Size Context, The Annals of Statistics 37(6B), 4104–4130.

- Examples:
  - in face recognition (images) we have many thousands of variables (pixels), the number of training samples defining a class (person) is usually small (usually less than 10).
  - Microarray experiments is unusual for there to be more than 50 repeats (data points) for several thousand variables (genes).
- The covariance matrix will be singular, and therefore cannot be inverted. In these cases we need to find some method of estimating a full rank covariance matrix to calculate an inverse.

Face recognition using PCA

https://www.mathworks.com/matlabcentral/fileexchange/45750-face-recognition-using-pca



https://zh.wikipedia.org/wiki/DNA微陣列

Efficient Estimation of Covariance: a Shrinkage Approach

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j),$$

$$\boxed{\text{a shrinkage estimator} \quad \hat{\boldsymbol{\Sigma}}_{\text{LW}} = \alpha_1 \mathbf{I} + \alpha_2 \mathbf{S}.}$$

"Small *n*, Large *p*"

Covariance and Correlation Estimators $S^\star$ and $R^\star$:

$$s^\star_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ r^\star_{ij}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases}$$

$$r^\star_{ij} = \begin{cases} 1 & \text{if } i = j \\ r_{ij}\min(1, \max(0, 1-\hat{\lambda}^\star)) & \text{if } i \neq j \end{cases}$$

$$\text{with} \quad \hat{\lambda}^\star = \frac{\sum_{i\neq j}\widehat{\text{Var}}(r_{ij})}{\sum_{i\neq j} r_{ij}^2}$$

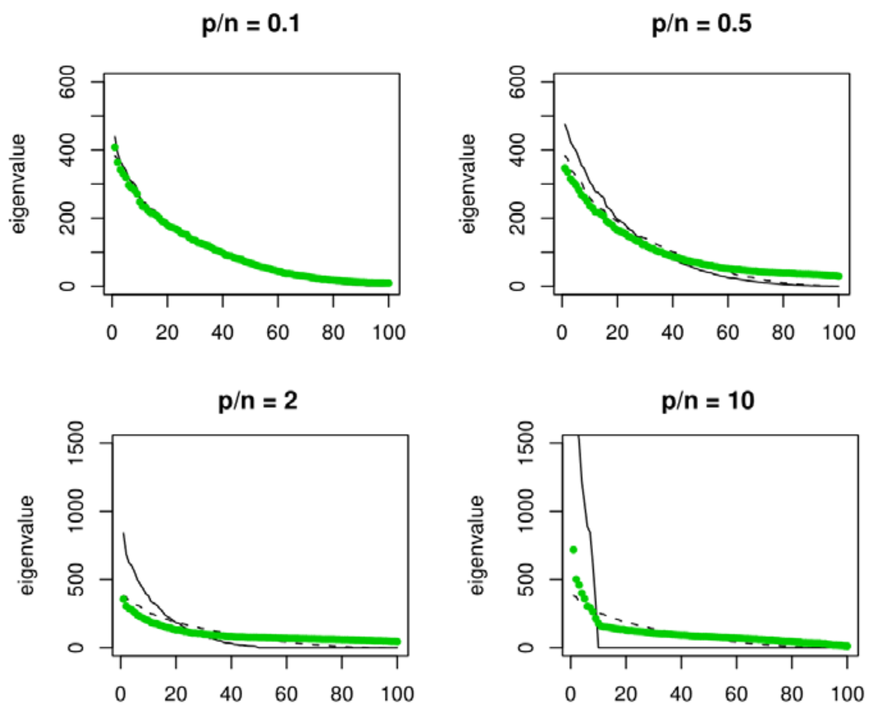


Figure 1: Ordered eigenvalues of the sample covariance matrix S (thin black line) and that of an alternative estimator $S^\star$ (fat green line, for definition see Tab. 1), calculated from simulated data with underlying p-variate normal distribution, for $p = 100$ and various ratios p/n. The true eigenvalues are indicated by a thin black dashed line.

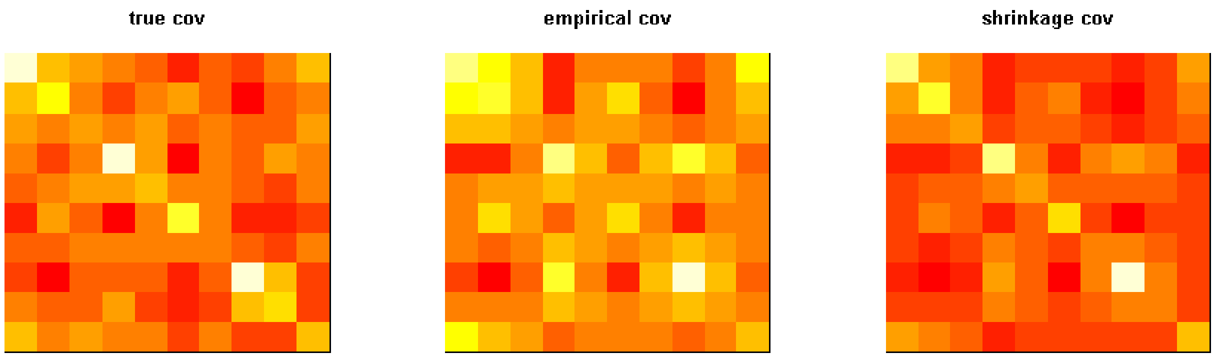
Schäfer, J., and K. Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology . 4: 32.

google: Penalized/Regularized/Shrinkage Methods

Example Script from corpcor Package

```
> library("corpcor")
>
> n <- 6 # try 20, 500
> p <- 10 # try 100, 10
> set.seed(123456)
> # generate random pxp covariance matrix
> sigma <- matrix(rnorm(p * p), ncol = p)
> sigma <- crossprod(sigma) + diag(rep(0.1, p)) #  t(x) %*% x
>
> # simulate multivariate-normal data of sample size n
> x <- mvrnorm(n, mu=rep(0, p), Sigma=sigma)
> # estimate covariance matrix
> s1 <- cov(x)
> s2 <- cov.shrink(x)
Estimating optimal shrinkage intensity lambda.var (variance vector): 0.4378
Estimating optimal shrinkage intensity lambda (correlation matrix): 0.6494
> par(mfrow=c(1,3))
> image(t(sigma)[,p:1], main="true cov", xaxt="n", yaxt="n")
> image(t(s1)[,p:1], main="empirical cov", xaxt="n", yaxt="n")
> image(t(s2)[,p:1], main="shrinkage cov", xaxt="n", yaxt="n")
>
> # squared error
> sum((s1 - sigma) ^ 2)
[1] 4427.215
> sum((s2 - sigma) ^ 2)
[1] 850.2443
```

`mvrnorm {MASS}`:
Simulate from a Multivariate Normal Distribution
`mvrnorm(n = 1, mu, Sigma, ...)`



Compare Eigenvalues

```
> # compare positive definiteness
> is.positive.definite(sigma)
[1] TRUE
> is.positive.definite(s1)
[1] FALSE
> is.positive.definite(s2)
[1] TRUE
>
> # compare ranks and condition
> rc <- rbind(
+   data.frame(rank.condition(sigma)), data.frame(rank.condition(s1)),
+   data.frame(rank.condition(s2)))
> rownames(rc) <- c("true", "empirical", "shrinkage")
> rc
          rank condition          tol
true        10 256.35819 6.376444e-14
empirical    5       Inf 1.947290e-13
shrinkage   10  15.31643 1.022819e-13
>
>
>
> # compare eigenvalues
> e0 <- eigen(sigma, symmetric = TRUE)$values
> e1 <- eigen(s1, symmetric = TRUE)$values
> e2 <- eigen(s2, symmetric = TRUE)$values
>
>
> matplot(data.frame(e0, e1, e2), type = "l", ylab="eigenvalues", lwd=2)
> legend("top", legend=c("true", "empirical", "shrinkage"), lwd=2, lty=1:3, col=1:3)
```

Shrinkage estimation of covariance matrix:

- `cov.shrink {corpcor}`
- `shrinkcovmat.identity {ShrinkCovMat}`
- `covEstimation {RiskPortfolios}`