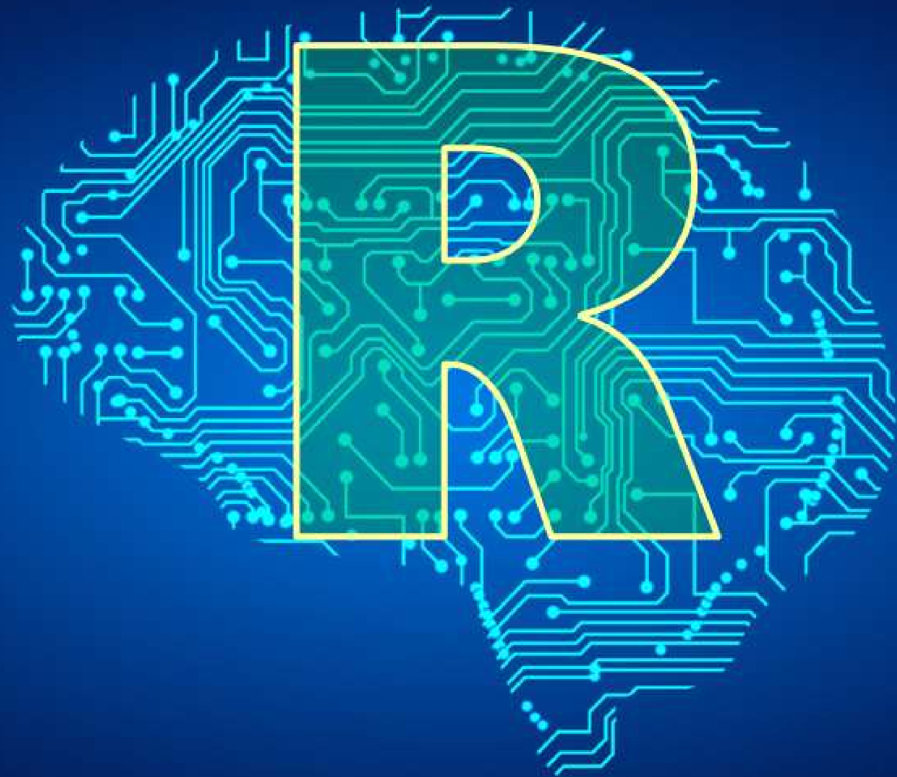
台灣人工智慧學校



敘述統計

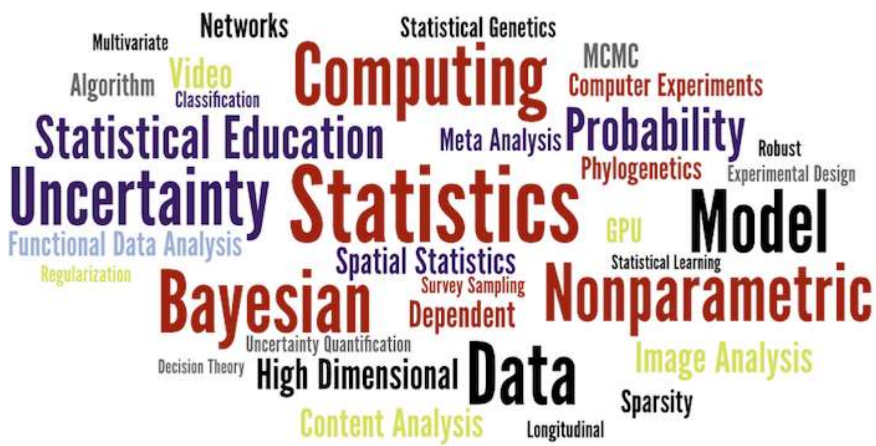
吳漢銘
國立臺北大學 統計學系

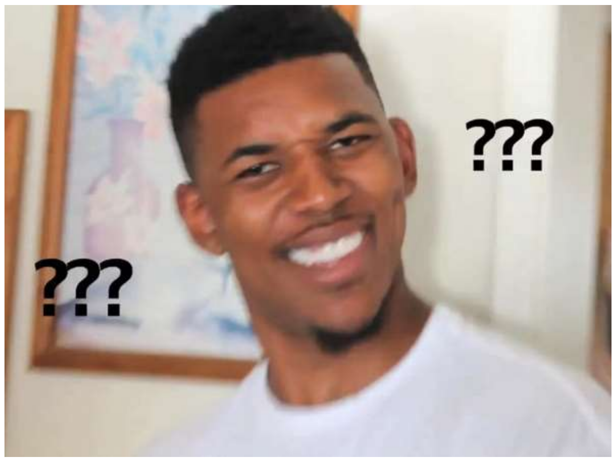
敘述統計 - 大綱

- 主題1
  - 為什麼學習機率統計? 為什麼要使用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
  - 範例:「由財稅大數據探討臺灣近年薪資樣貌」
- 主題2
  - 距離及相似度量測指標
  - 相關係數: Pearson's rho、Spearman's rho、Kendall's tau
  - 小樣本數高維度資料問題(HDLSS Problem)

為什麼要學習機率統計?

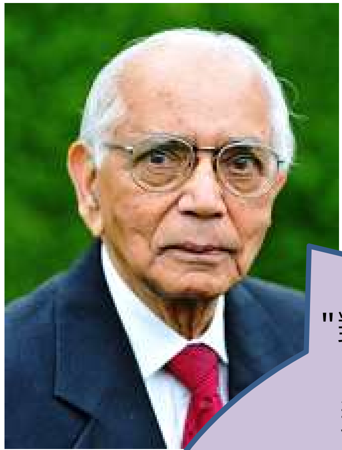
- 為什麼要學機率**統計**?
- 學**統計**，一定要學機率嗎?
- 數學不好，機率**統計**可以學的好嗎?
- 分析資料，一定要學**統計**嗎?
- 我要成為一位資料科學家，一定要學**統計**嗎?





大師們對統計的看法

C.R. Rao (1920-):



統計與真理：
怎樣運用偶然性

"對統計學的一知半解，
常常造成不必要的上當受騙;
對統計學的一概排斥，
往往造成不必要的愚昧無知"

"在終極的分析中，一切知識都是歷史;
在抽象的意義下，一切科學都是數學;
在理性的基礎上，所有的判斷都源於統計學
"

"統計學是人類探求真理必不可少的工具"

科學家不能離開統計而研究
政治家不能離開統計而施政
企業家不能離開統計而執業
軍事家不能離開統計而謀略

馬寅初(1882-1982)
經濟學家、教育家、人口學家。
曾任北京大學校長。

"事實上，無論是做人工智慧，還是做商業數據分析，如果能夠對統計學有系統的理解，那麼，他對於機器學習的研究和應用便會如虎添翼，登堂入室。"

吳喜之教授
(中國人民大學統計學院教授)

成不了AI高手？因為你根本不懂數據！
https://kknews.cc/tech/e8ykpyn.html

科學事實與**統計思維** (程開明, 中國統計, 2015年第12期, 24-26.)
http://www.slstjj.gov.cn/index/ShowArticle.asp?ArticleID=1856

我所理解的**統計思維**
http://blog.sciencenet.cn/blog-242272-1047853.html

為什麼要使用R做為資料分析工具?

http://www.r-project.org
https://www.rstudio.com/

- 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.

寫程式是資料分析的必要技能
https://medium.com/datainpoint/9ee15b58cc
Python or R, what should you learn first?
https://read01.com/0ePnyD.html#.Wu66C3--kZY
Why I use R for Data Science – An Ode to R
https://www.r-bloggers.com/why-i-use-r-for-data-science-an-ode-to-r-2/
選擇R開發數據分析平台的 4 個不錯的理由
https://read01.com/660M4g.html
做數據分析必須學R語言的4個理由
https://read01.com/yyREB2.html
Hadley Wickham：一個改變了R的人
https://read01.com/Mmy64J.html
Hadley Wickham: "R is ... tailored to the problems of data science"

TIOBE 全球程式語言排名

TIOBE Index for January 2018

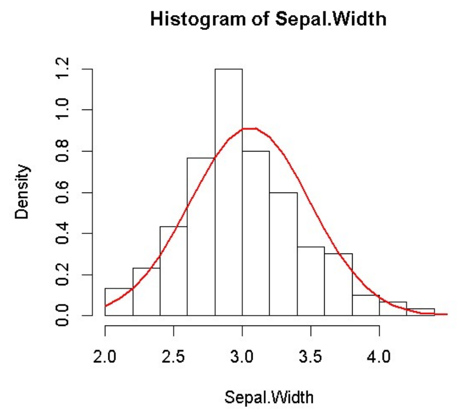
January Headline: Programming Language C awarded Language of the Year 2017

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)

什麼是統計?

- **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**: 對所收集到樣本的摘要結果。
  - **推論統計**: 考慮隨機性之下，根據樣本的特性去推論母體的參數(例如: 估計母體平均數、推論母體的分佈)。

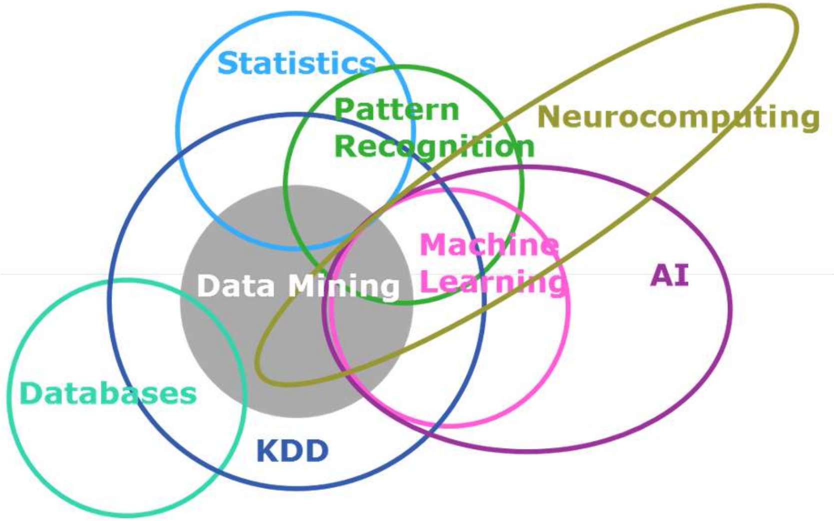
- 統計研究領域的分類: 數理統計、工業統計、商用統計、生物統計、社會統計、貝氏統計、空間統計等等。

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

統計模型、資料探勘、機器學習

- **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.

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

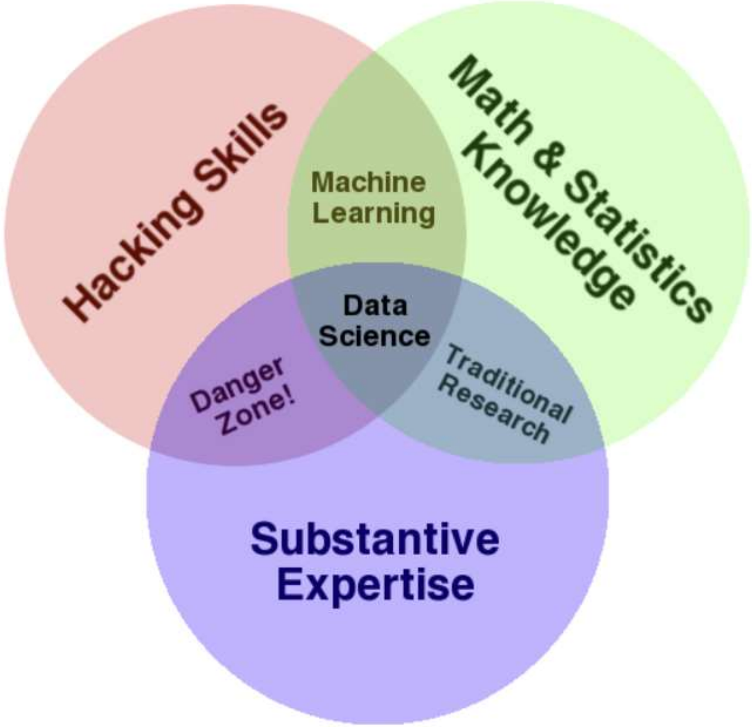
https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

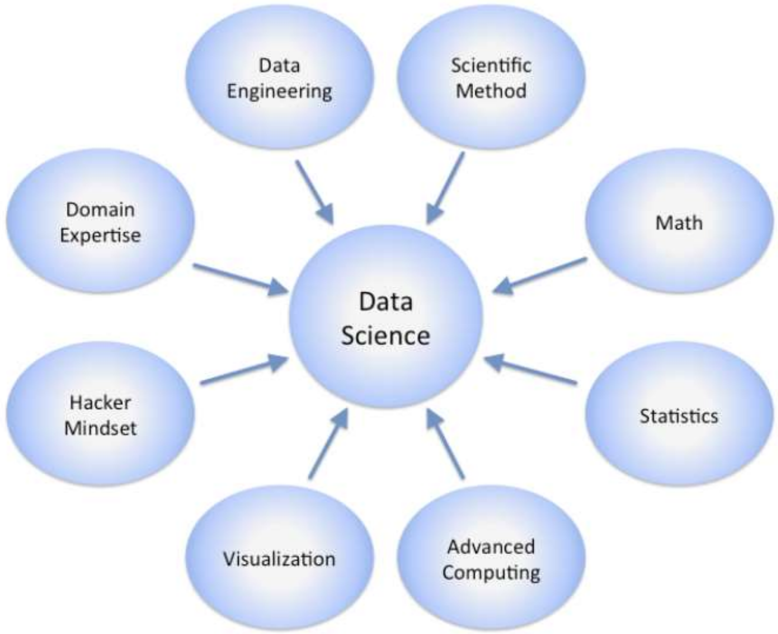
機器學習和統計模型的差異
http://vvar.pixnet.net/blog/post/242048881
為什麼統計學家、機器學習專家解決同一問題的方法差別那麼大?
https://read01.com/EBPPK7.html
機器學習與統計學是互補的嗎？
https://read01.com/ezQ3K.html

資料科學 Data Science

The Data Science Venn Diagram

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

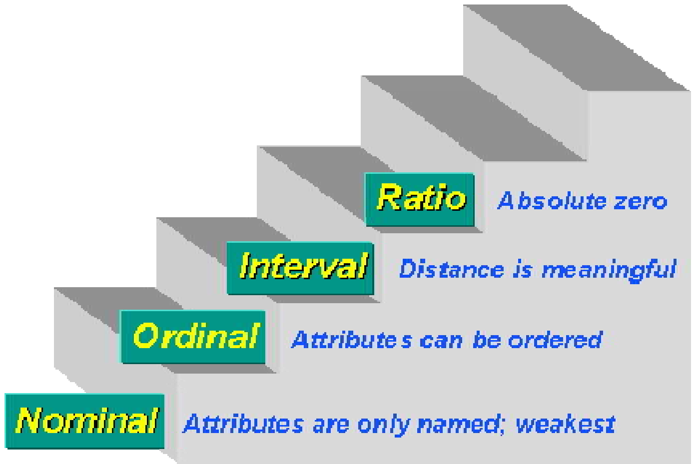




Source: By Calvin.Andrus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Types of Data Scales

- Nominal (名目變數), Categorical (類別資料), discrete: 性別、種族、宗教信仰、交通工具、音樂類型... (qualitative 屬質)。
- Ordinal (順序): 精通程度、同意程度、滿意程度、教育程度。
- Interval — Distances between values are meaningful, but **zero point** is not meaningful. (例如:華氏溫度)(不能說：80 度是4 0度的兩倍熱)。
- Ratio (Continuous Data 連續型資料)— Distances are meaningful and a zero point is meaningful: 年收入、年資、身高、... (quantitative 計量)。



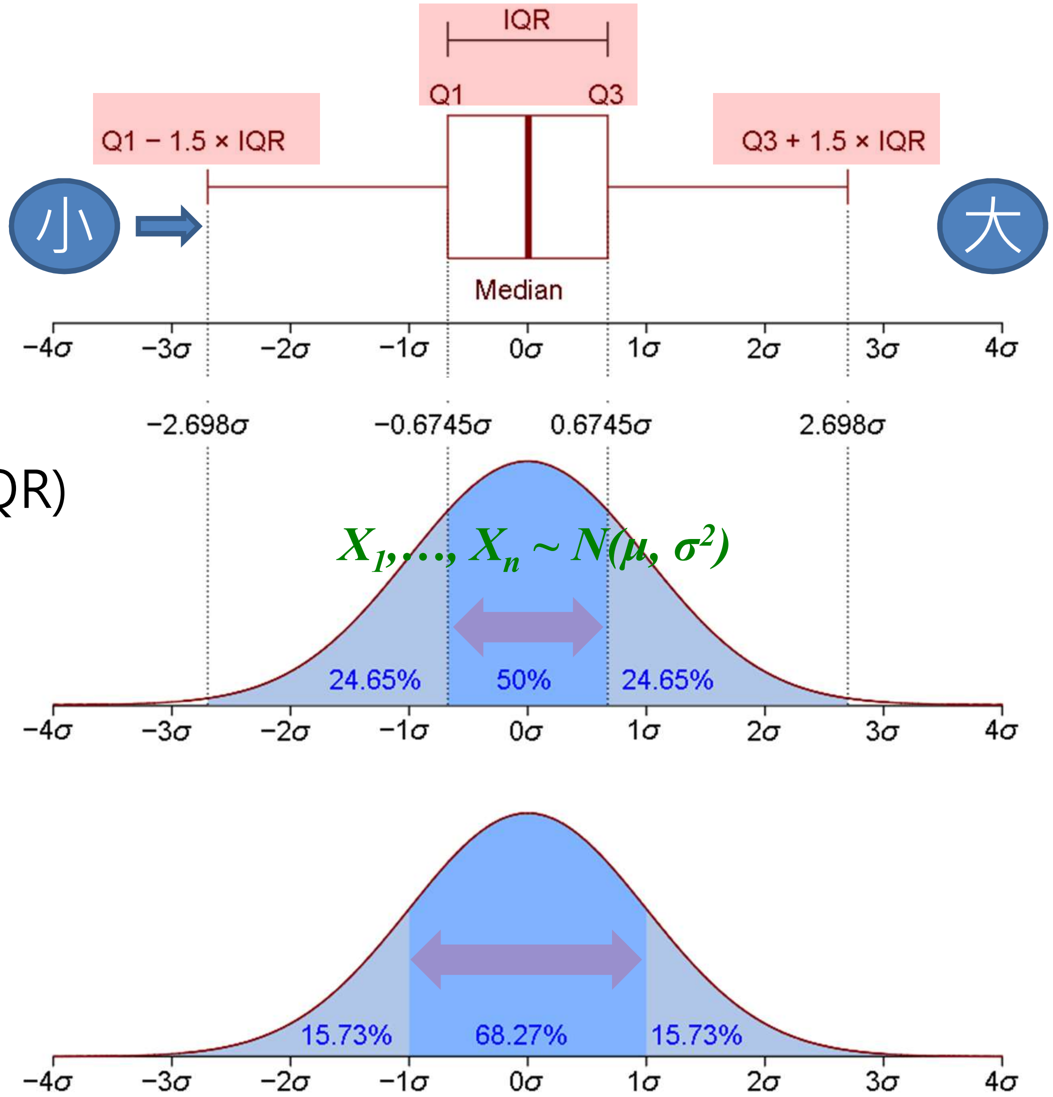
https://socialresearchmethods.net/kb/measlevl.php

資料描述: 中心趨勢、分散程度

- **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)

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

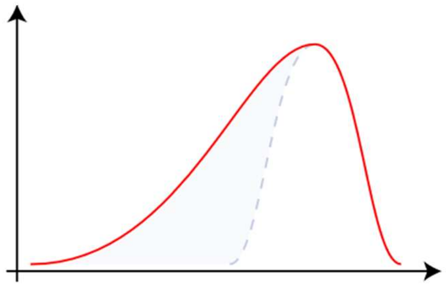
n = The number of data points
$\bar{x}$ = The mean of the x_i
x_i = Each of the values of the data



https://zh.wikipedia.org/wiki/四分位距

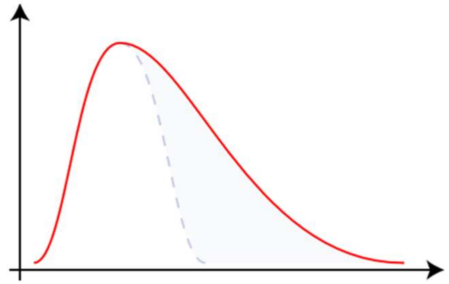
資料描述: 偏態係數

偏態(skewness)係數:



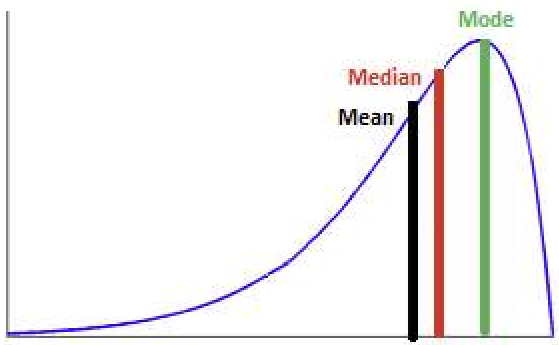
Negative Skew

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

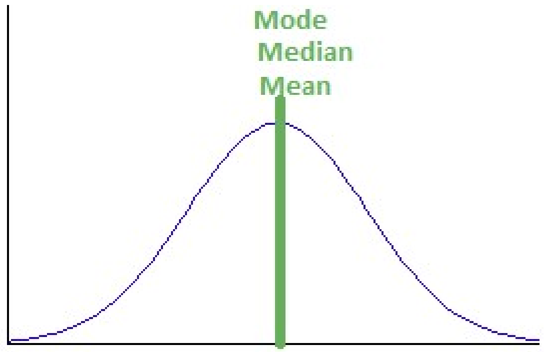


Positive Skew

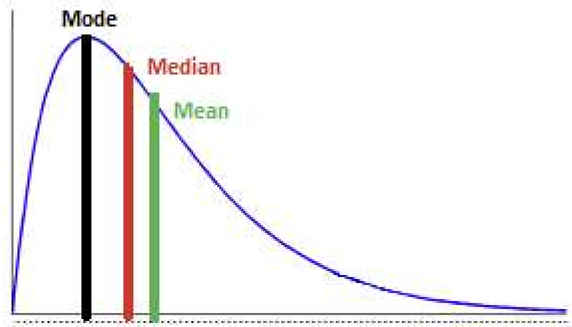
小於0：左偏分配 等於0：對稱分配 大於0：右偏分配



Negatively Skewed



Symmetric

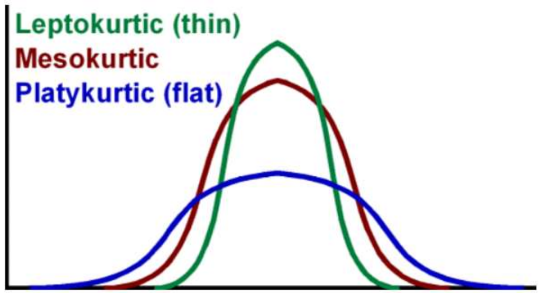


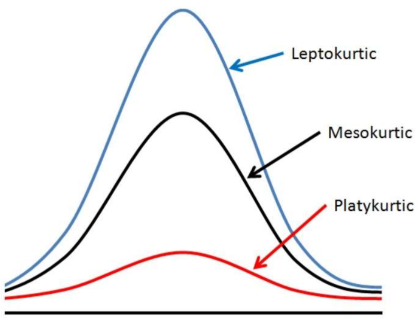
Positively Skewed

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

https://en.wikipedia.org/wiki/Skewness

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)。

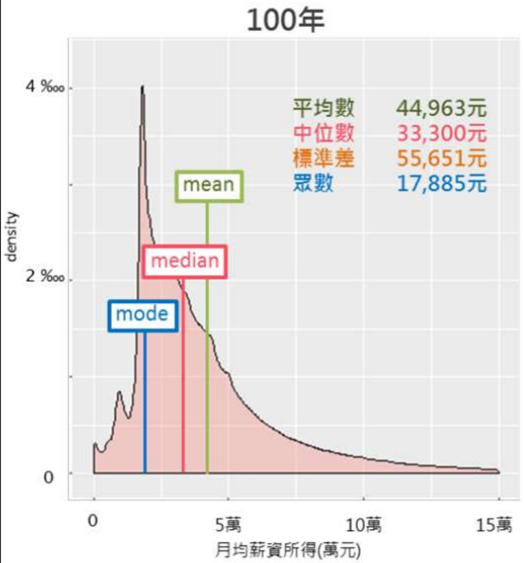
常用的樣本峰度係數的計算式有以下三項:

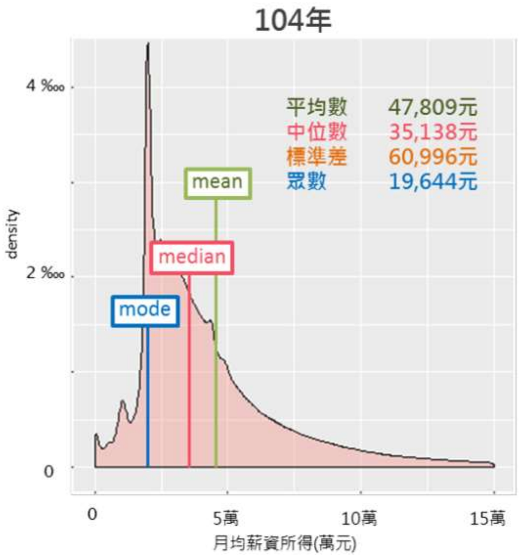
- The typical definition used in many older textbooks: $g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3$
- Used in SAS and SPSS: $G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$
- Used in MINITAB and BMDP: $b_2 = (g_2 + 3)(1 - \frac{1}{n})^2 - 3$

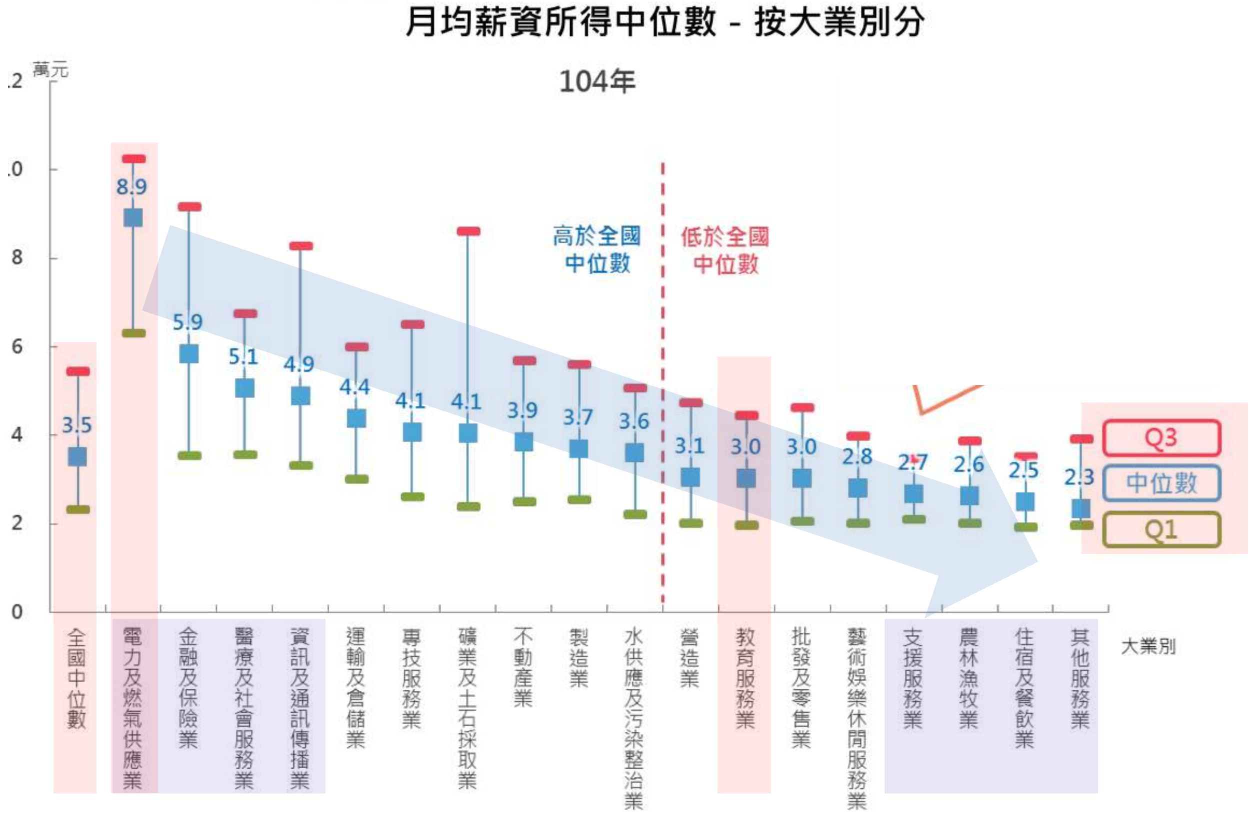
範例: 由財稅大數據探討臺灣近年薪資樣貌

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







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)

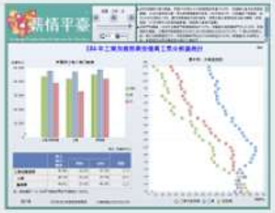
熱誠・公正・效率・精確

統計資訊網　答客問

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

薪 情 互 動

製造業四大產業薪
況



男女薪資差異

各業薪情概況

敘述統計 - 大綱

- 主題1
  - 為什麼學習機率統計? 為什麼要使用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
  - 範例:「由財稅大數據探討臺灣近年薪資樣貌」
- 主題2
  - 距離及相似度量測指標
  - 相關係數: Pearson's rho、Spearman's rho、Kendall's tau
  - 小樣本數高維度資料問題(HDLSS Problem)

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

Correlation Matrix
(p by p)

DR, ...

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

Pearson Correlation Coefficient

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

$$x = (x_1, x_2, \cdots, x_n)$$

$$y = (y_1, y_2, \cdots, y_n)$$

Euclidean Distance

$$s_u = (u_1, u_2, \cdots, u_p)$$

$$s_v = (v_1, v_2, \cdots, v_p)$$

$$d_{uv} = \sqrt{\sum_{k=1}^{p}(u_k - v_k)^2}$$

Data Matrix (tidy form)

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

Distance matrix
(n by n)

clustering algorithms, ...

相關係數

Pearson correlation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

All indices range from -1 to +1

Spearman rank correlation

$$\rho_R(X,Y) = \frac{Cov(R_X,R_Y)}{\sqrt{Var(R_X)Var(R_Y)}}$$

Kendall's tau

$$\tau(X,Y) = \frac{1}{\binom{p}{2}} \sum_{i \neq j}^{n} \text{sign}\left[(x_i - x_j)(y_i - y_j)\right]$$

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: $(x_j - x_i) = 0$

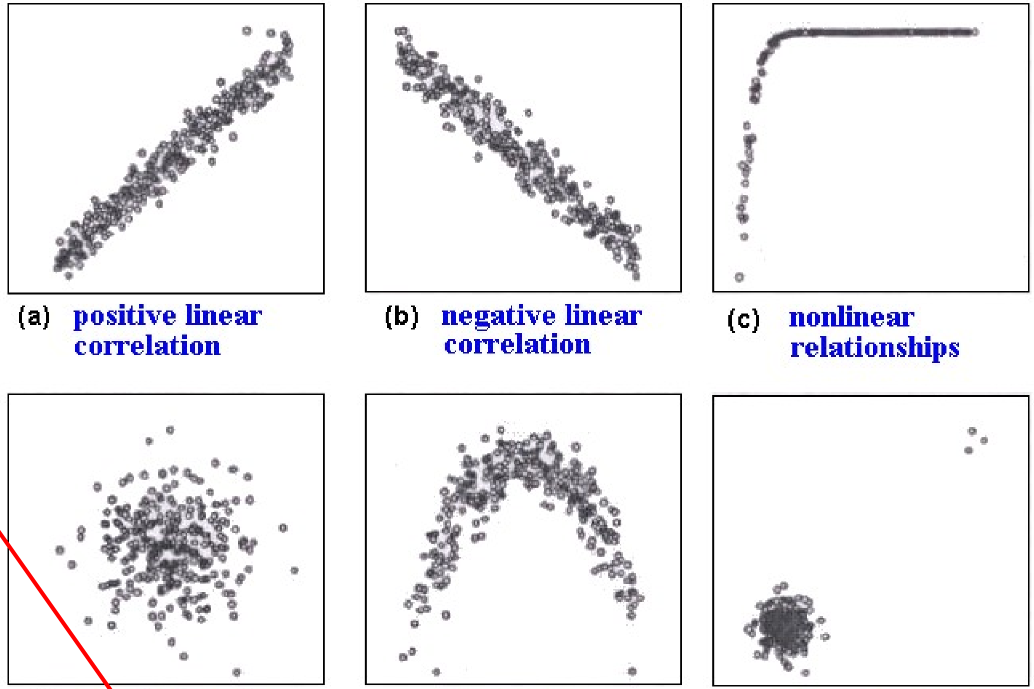
E_x : extra x pair in y's: $(y_j - y_i) = 0$

x	y
.	.
x_i	y_i
.	.
x_j	y_j
.	.

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

Pearson's rho、Spearman's rho、Kendall's tau

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

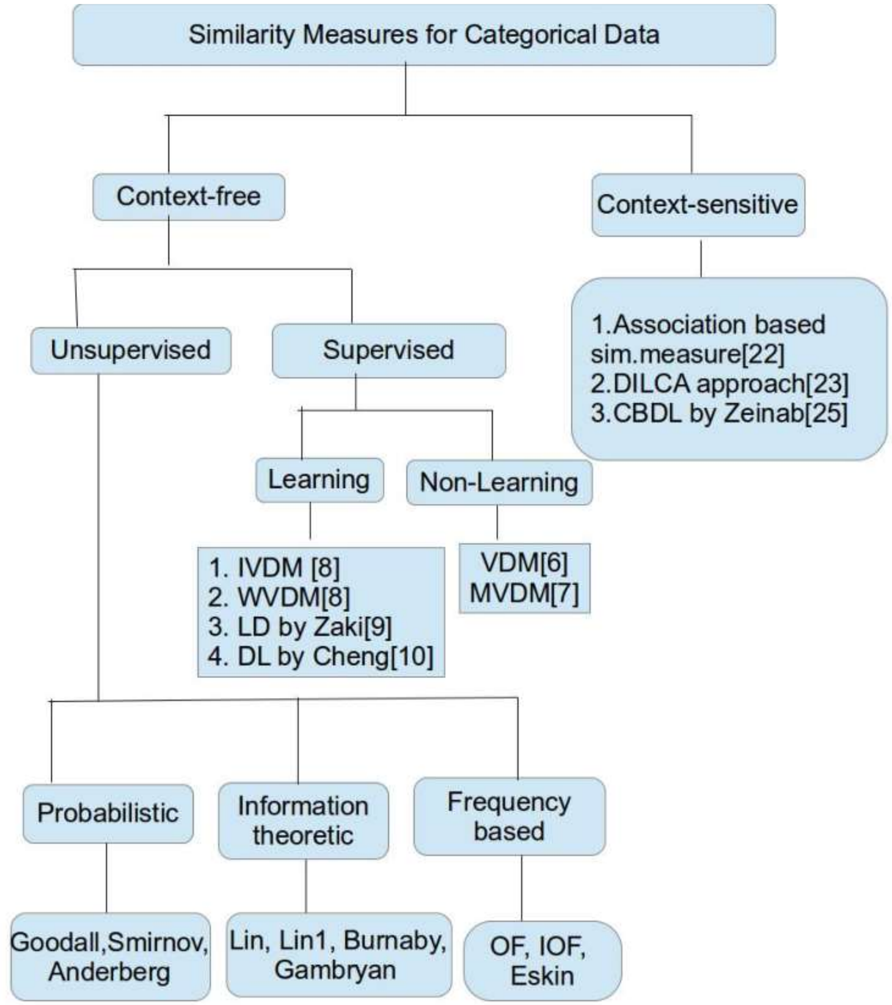
Similarity Measures for Categorical Data

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.

High-dimensional data (HDD)

- 高維度資料的三種類型:
  - p is large but smaller than n;
  - p is large and larger than n:
    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 $r \leq min\{p, n\}$.

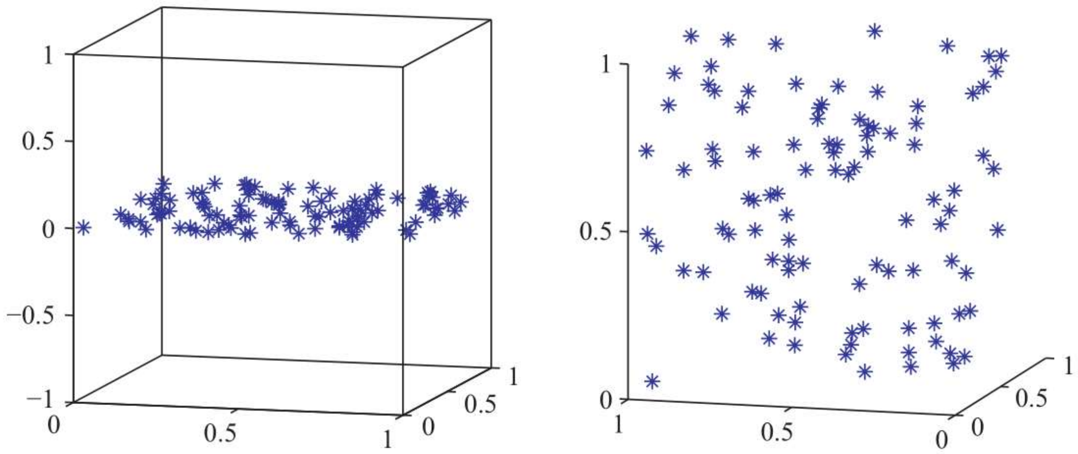
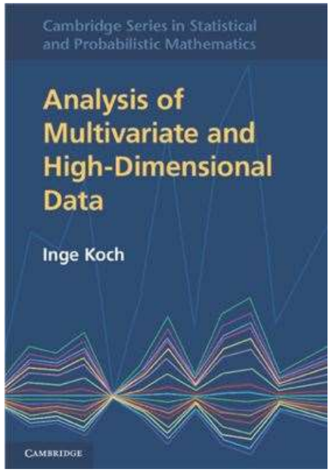


Figure 2.12 Distribution of 100 points in 2D and 3D unit space.



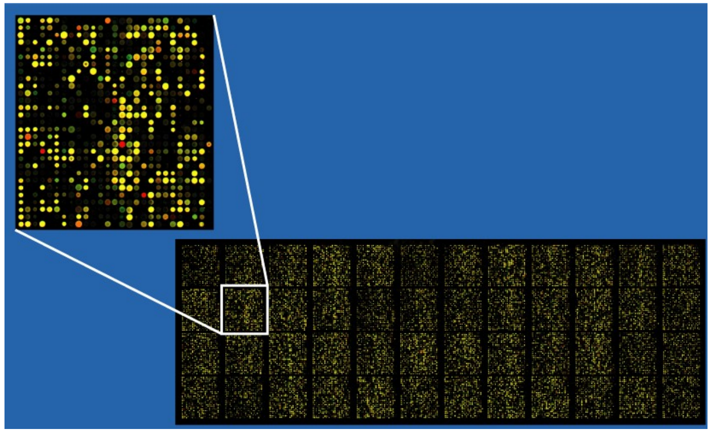
HDLSS examples

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:
  - 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

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

a shrinkage estimator

$$\hat{\mathbf{\Sigma}}_{\mathrm{LW}} = \alpha_1 \mathbf{I} + \alpha_2 \mathbf{S}.$$

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.

"Small *n*, Large *p*"

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

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

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

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

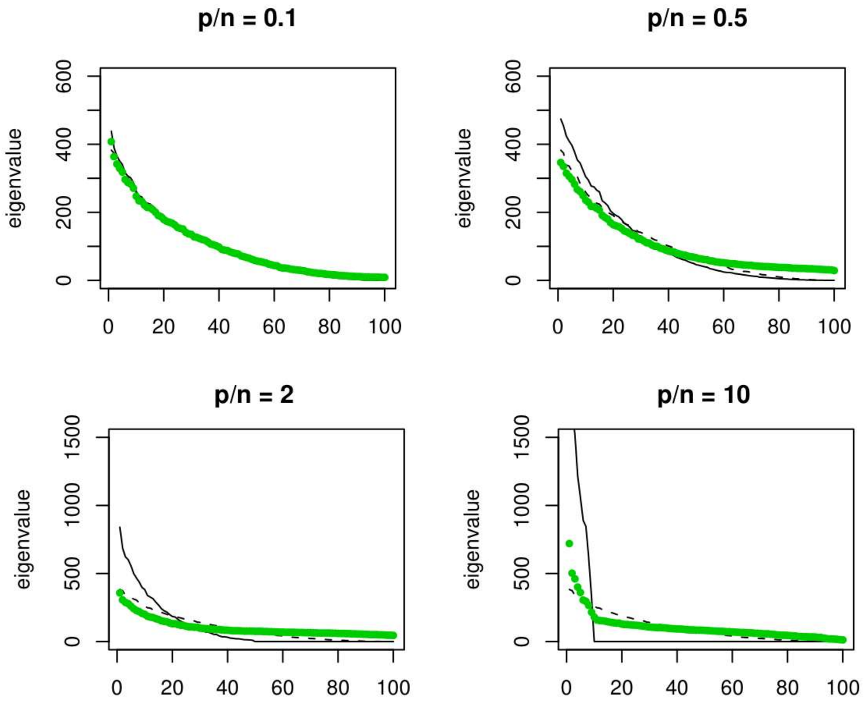


Figure 1: Ordered eigenvalues of the sample covariance matrix S (thin black line) and that of an alternative estimator $S^\star$ (fat green line, for definition see Tab. 1), calculated from simulated data with underlying p-variate normal distribution, for $p = 100$ and various ratios p/n. The true eigenvalues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

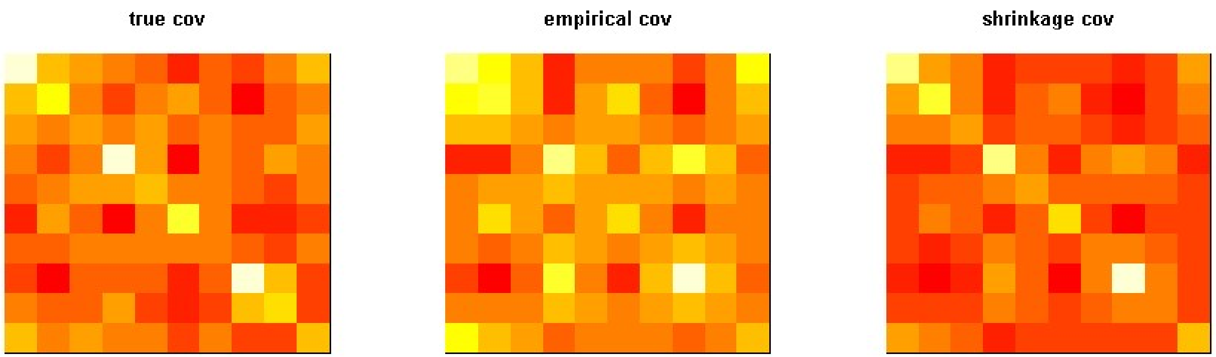
google: Penalized/Regularized/Shrinkage Methods

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")
```

corpcor: Efficient Estimation of Covariance and (Partial) Correlation

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



Compare Eigenvalues

Shrinkage estimation of covariance matrix:
- `cov.shrink {corpcor}`
- `shrinkcovmat.identity {ShrinkCovMat}`
- `covEstimation {RiskPortfolios}`

rank: the number of singular values D[i] > tol
condition: the ratio of the largest and the smallest singular value

```
> # 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)
```