M1

台灣人工智慧學校

經理人週末研修班

敘述統計
與機率分布

吳漢銘

國立臺北大學 統計學系

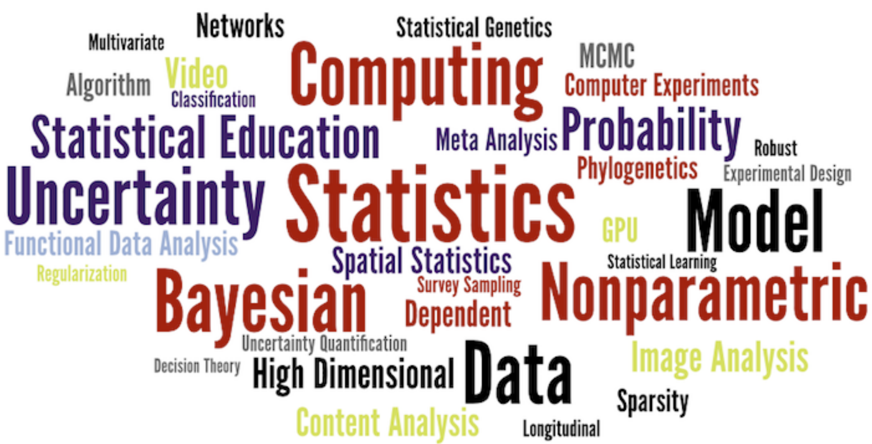
http://www.hmwu.idv.tw

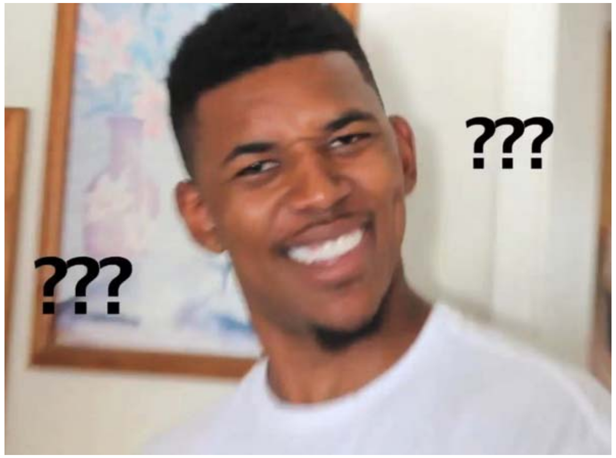
敘述統計與機率分布- 大綱

- 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) [進階選讀]
- 主題3
  - 常見統計名詞
  - 機率分佈 (Probability distribution)
  - 累積機率分配函數 CDF (p)
  - 分位數 Quantiles (q)
- 主題4
  - 常見之分佈(二項式分佈、常態分佈)
  - 以常態機率逼近二項式機率 [進階選讀]
- 主題5
  - 大數法則 (LLN)
  - 中央極限定理 (CLT)
  - 用R程式模擬算機率

為什麼要學習機率統計?

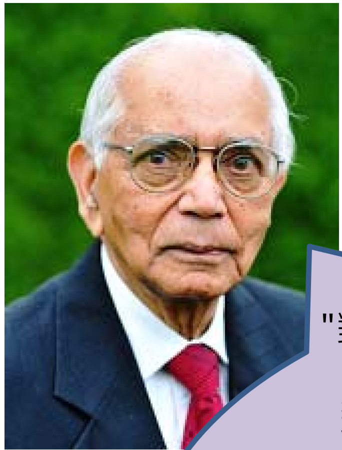
- 為什麼要學機率**統計**?
- 學**統計**，一定要學機率嗎?
- 數學不好，機率**統計**可以學的好嗎?
- 分析資料，一定要學**統計**嗎?
- 我要成為一位資料科學家，一定要學**統計**嗎?





大師們對統計的看法

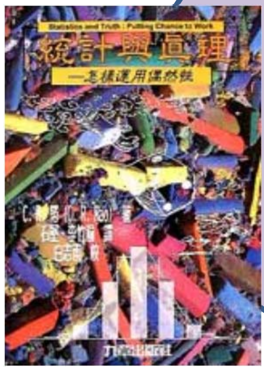
C.R. Rao (1920-):



"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"



統計與真理：
怎樣運用偶然性

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?

The R Project for Statistical Computing

[Home]

Download

CRAN

R Project

Getting Started

R is a free software environment for statistical computing and graphics. It c
variety of UNIX platforms, Windows and MacOS. To **download R**, please c
CRAN mirror.

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"

Power BI

TIOBE 全球程式語言排名

TIOBE Index for January 2018

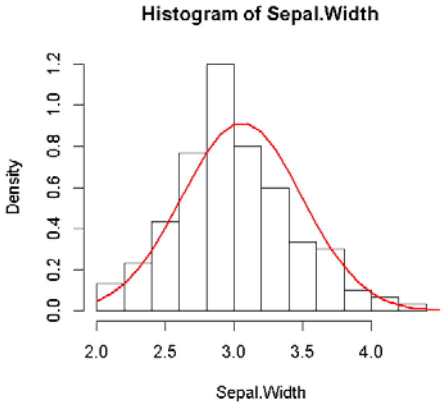
January Headline: Programming Language C awarded Language of the Year 2017

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."



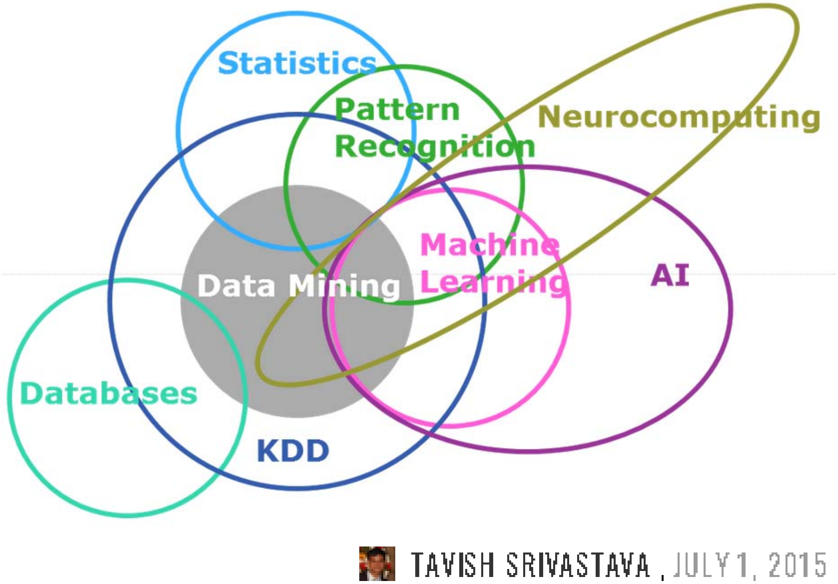
- 傳統統計(歷史源自17世紀), 分兩類:
  - **敘述統計**: 對所收集到樣本的摘要結果。
  - **推論統計**: 考慮隨機性之下，根據樣本的特性去推論母體的參數(例如: 估計母體平均數、推論母體的分佈)。
- 統計研究領域的分類: 數理統計、工業統計、商用統計、生物統計、社會統計、貝氏統計、空間統計等等。

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



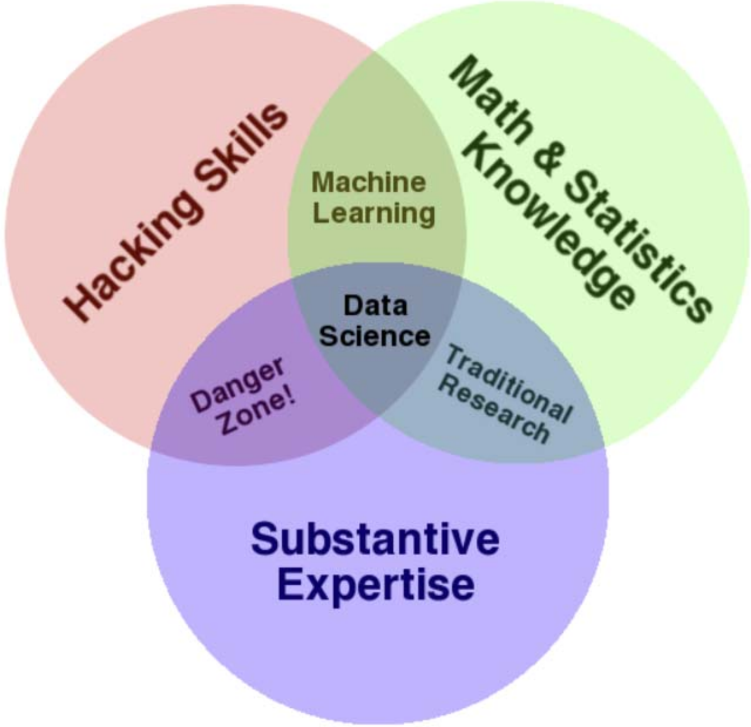
https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

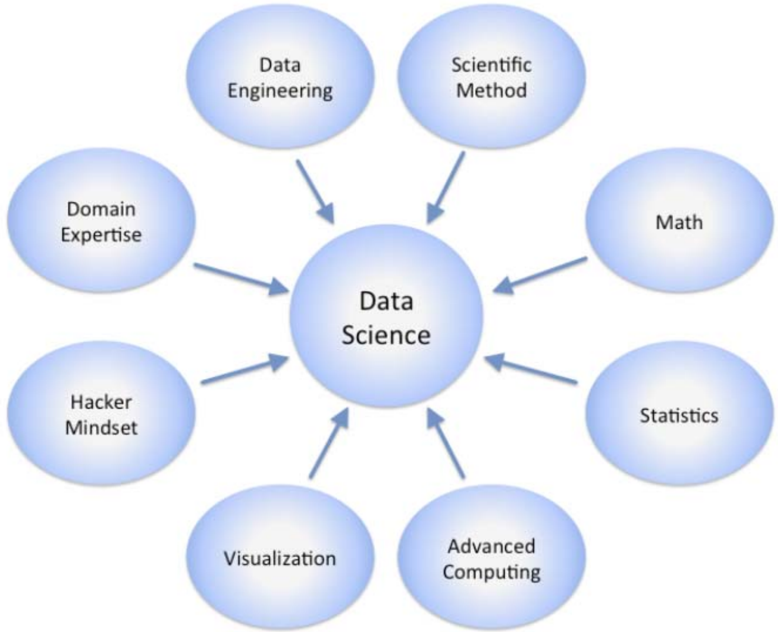
機器學習和統計模型的差異
http://vvar.pixnet.net/blog/post/242048881
為什麼統計學家、機器學習專家解決同一問題的方法差別那麼大?
https://read01.com/EBPPK7.html
機器學習與統計學是互補的嗎？
https://read01.com/ezQ3K.html

資料科學 Data Science

The Data Science Venn Diagram

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

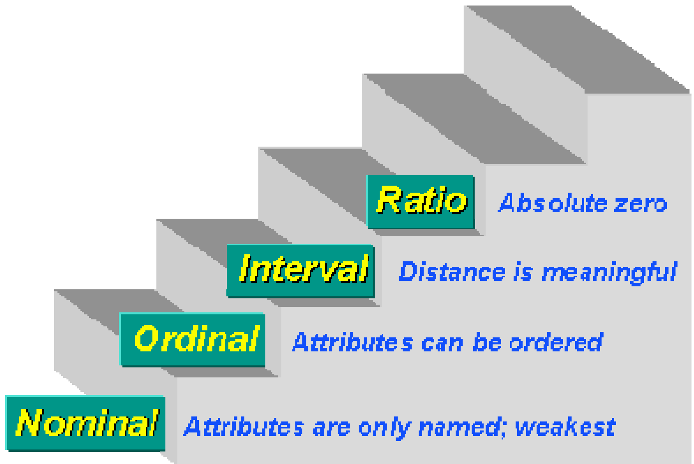




Source: By Calvin.Andrus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

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

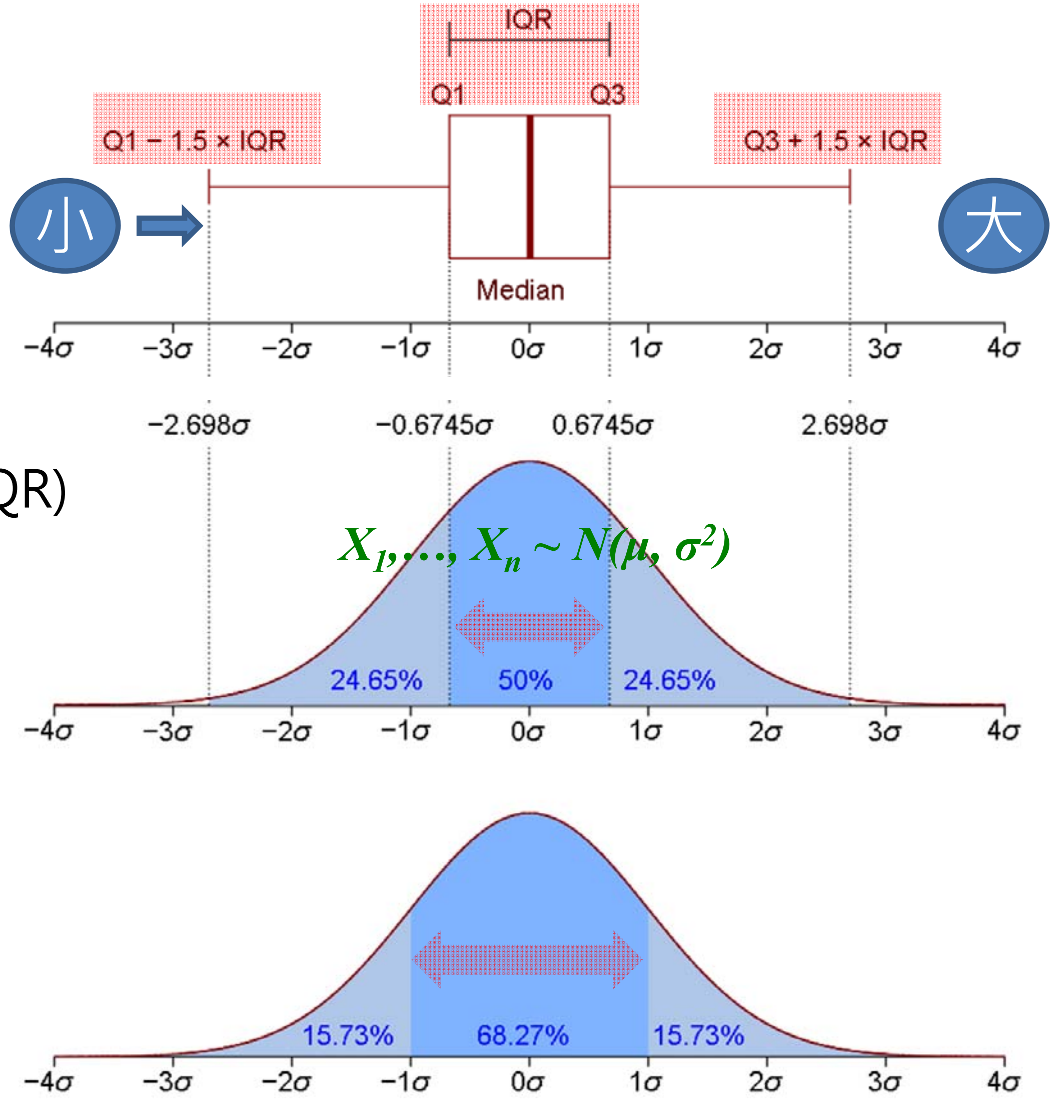
- **資料中心趨勢:**
  平均數(average)
  眾數(mode)
  中位數(median)
- **資料分散程度:**
  四分位數(Quartile)
  全距(range)
  四分位距(interquartile range, IQR)
  百位數(percentile)
  標準差(standard deviation)
  變異數(variance)

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

$n =$ The number of data points

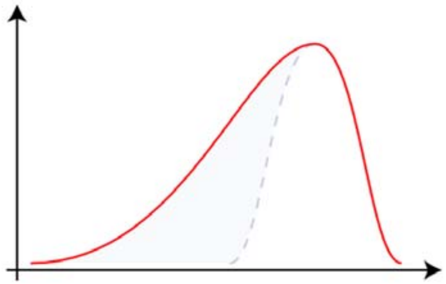
$\bar{x} =$ The mean of the x_i

$x_i =$ Each of the values of the data



https://zh.wikipedia.org/wiki/四分位距

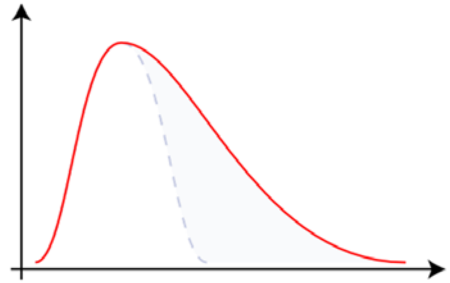
資料描述: 偏態係數



Negative Skew

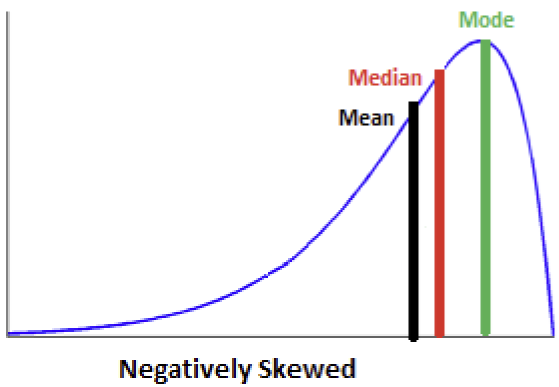
偏態(skewness)係數:

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

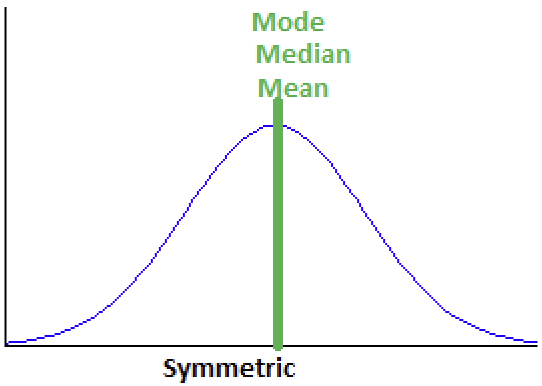


Positive Skew

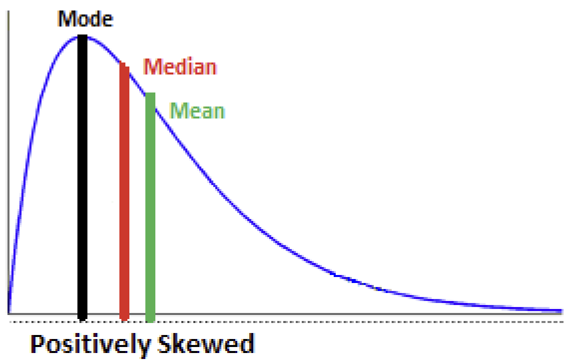
小於0：左偏分配



等於0：對稱分配



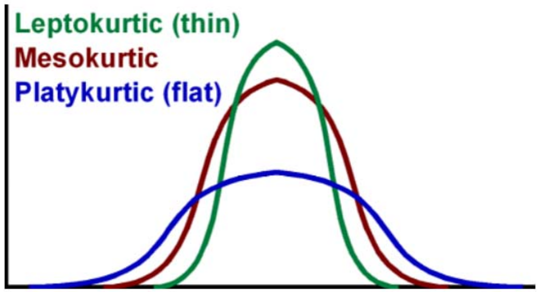
大於0：右偏分配

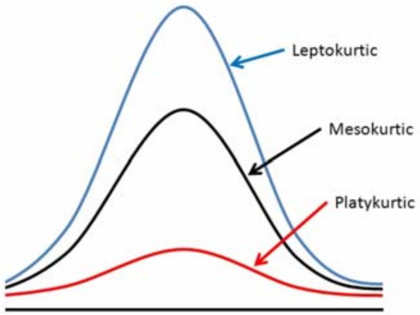


http://www.t4tutorials.com/data-skewness-in-data-mining/

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:

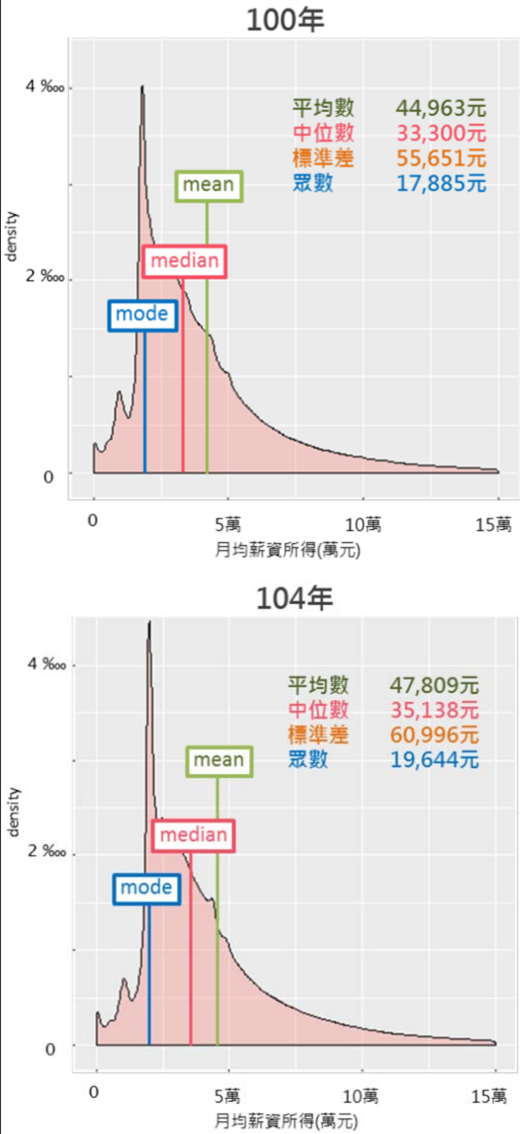
- 若 $k_c > 0$, 表示資料分布呈高狹峰 (lepto kurtosis)。
- 若 $k_c = 0$, 表示資料分布呈常態峰 (normal kurtosis)。
- 若 $k_c < 0$, 表示資料分布呈低濶峰 (platy kurtosis)。

常用的樣本峰度係數的計算式有以下三項:

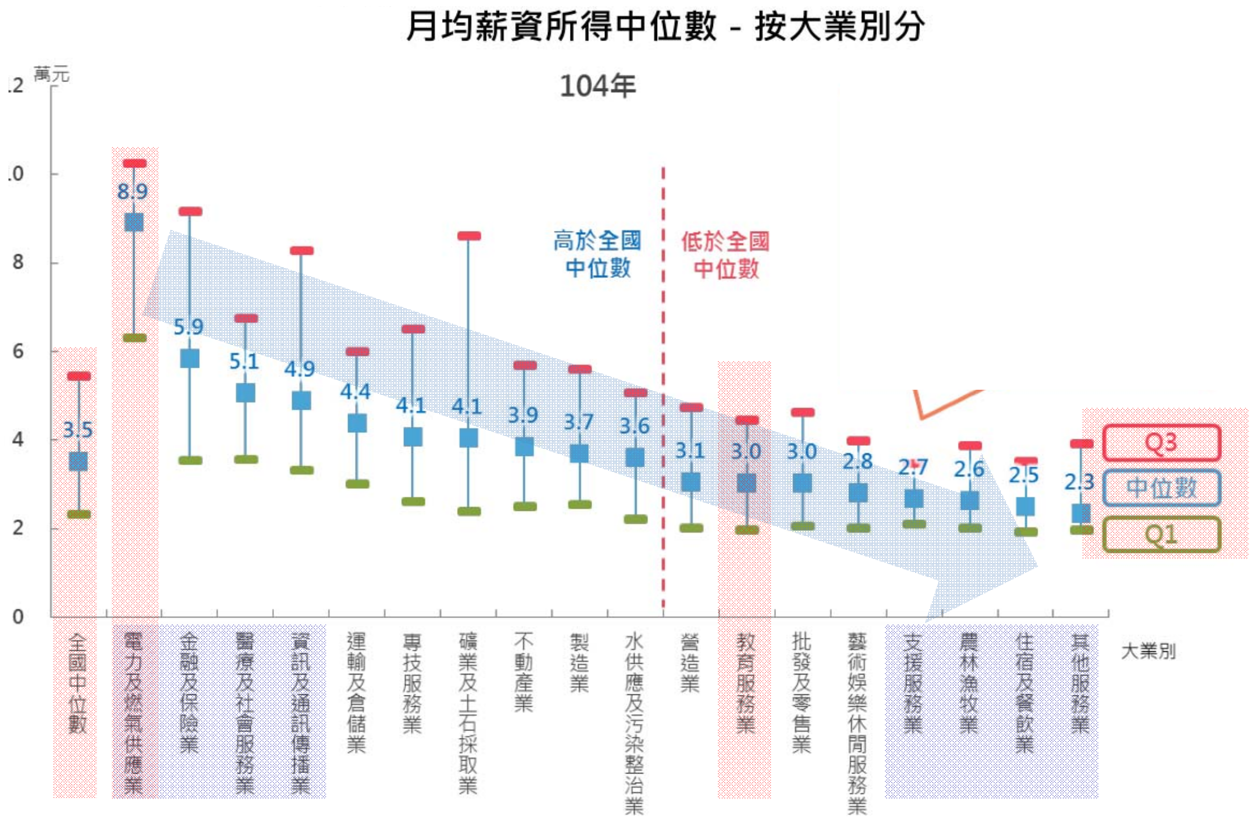
- The typical definition used in many older textbooks: $g_2 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3$
- Used in SAS and SPSS: $G_2 = \dfrac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$
- Used in MINITAB and BMDP: $b_2 = (g_2 + 3)(1 - \frac{1}{n})^2 -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

範例: 由財稅大數據探討臺灣近年薪資樣貌

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

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

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, ...

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

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	-1.08	-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

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

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

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

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

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

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

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:

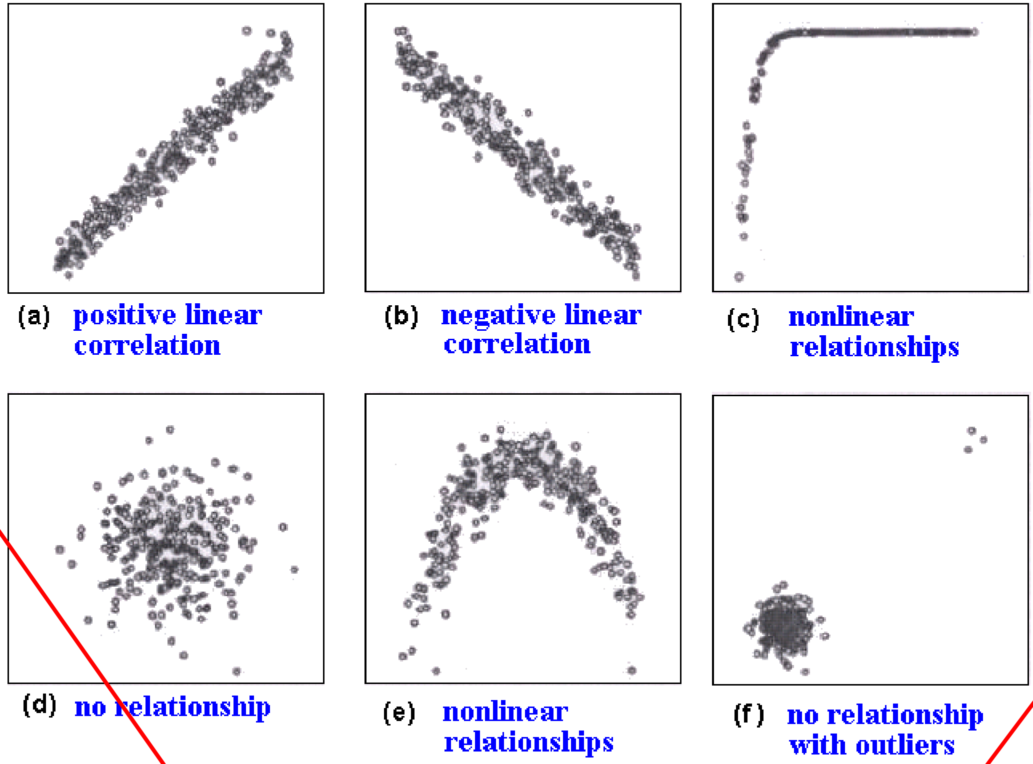
E_y: extra y pair in x's: $(x_j - x_i) = 0$

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

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

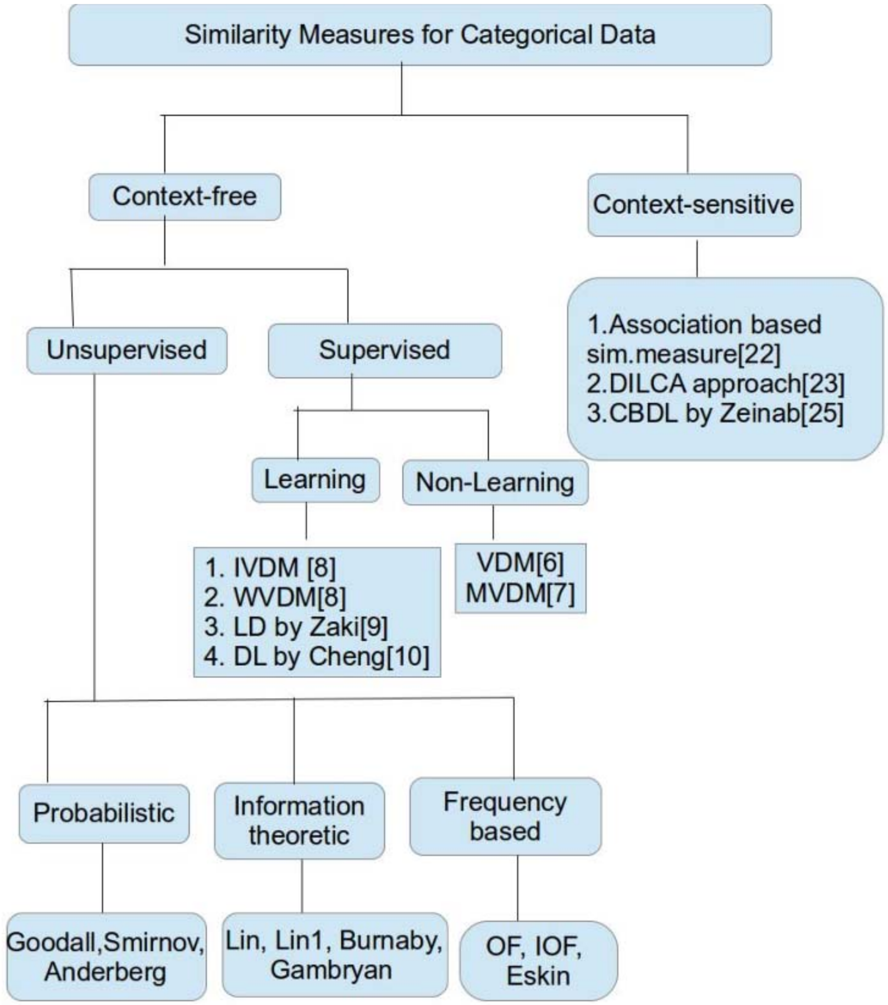
Similarity Measures for Categorical Data

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



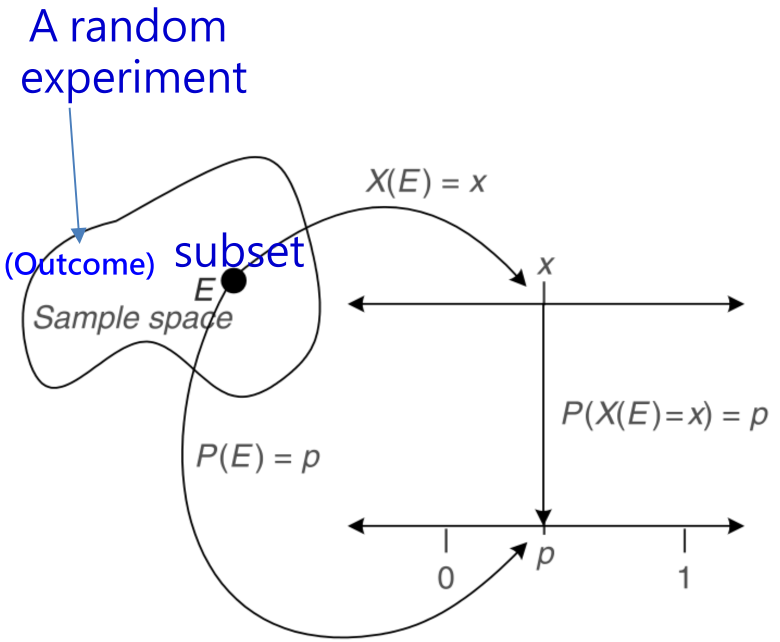
2014, A survey of distance/similarity measures for categorical data,
2014 International Joint Conference on Neural Networks (IJCNN), 1907-1914.

常見統計名詞

- A **random experiment (隨機實驗)** is a process by which we observe something uncertain. After the experiment, the result of the random experiment is known.
- **Outcome (結果)**: An outcome is a result of a random experiment.
- **Sample space (樣本空間), *S***: the set of all possible outcomes.
  - 例子1: 投擲兩硬幣, 正(Head)反(Tail)面之樣本空間 S={HH, HT, TH, TT}.
- **Event (事件), *E***: an event is a subset of the sample space.
  - 例子2: In the context of an experiment, we may define the sample space of observing a person as *S* = {sick, healthy, dead} . The following are all events: {sick} , {healthy} , {dead} , {sick, healthy} , {sick, dead} , {healthy, dead} , {sick, healthy, dead} , {none of the above} .
- **Trial (試驗)**: a single performance of an experiment whose outcome is in *S* .
  - 例子3: 投擲4枚硬幣的隨機實驗中，每投擲一次硬幣皆是一次「試驗」。

機率與隨機變數

- **Probability (機率)**: the probability of event **E**, **P(E)**, is the value approached by the relative frequency of occurrences of **E** in a long series of replications of a random experiment.
(The frequentist view)
- **Random variable (隨機變數)**: A function that assigns real numbers to events, including the null event.



Source: Statistics and Data with R

Probability Distribution (機率分佈):
是以數學函數的方式來表示隨機實驗中不同的可能結果(即樣本空間之每個元素)發生的可能性(機率)。

例子: 假如令隨機變數 X 表示是投擲一枚公平硬幣的結果: X=1為正面，X=0為反面，
則X的機率分佈是:
P(X=1) = 0.5, P(X=0) = 0.5.

機率質量函數 Probability Mass Function

Formal definition

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

Suppose that $X: S \to A$ ($A \subseteq \mathbf{R}$) is a discrete random variable defined on a sample space S. Then the probability mass function $f_X: A \to [0, 1]$ for X is defined as

$$f_X(x) = \Pr(X = x) = \Pr(\{s \in S : X(s) = x\}).$$

Thinking of probability as mass helps to avoid mistakes since the physical mass is conserved as is the total probability for all hypothetical outcomes x:

$$\sum_{x \in A} f_X(x) = 1$$

***例子.*投擲2顆公正的骰子**

$X_1 \sim DiscreteUniform\ (1, 6).$

$X_2 \sim DiscreteUniform(1, 6).$

$f_{X1}(k) = f_{X2}(k) = P(X_1 = k) = P(X_2 = k) = 1/6,$

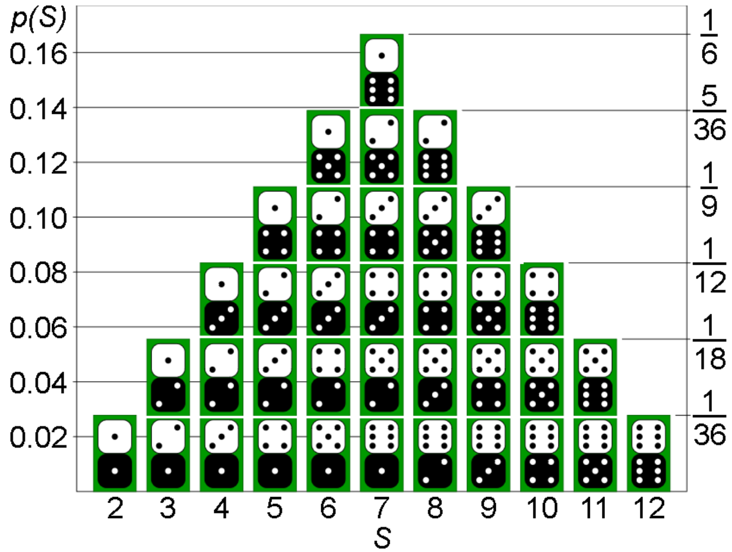
$k = 1,..,6.$

$S = X_1 + X_2$

$f_S(s) = p(S = s),\ s=2,\ ...,\ 12.$

$P(S = 2) = 1/36,\ P(S=3)=2/36,\ ...,\ P(S=12)=1/36$

$P(X_1 + X_2 > 9) = 1/12 + 1/18 + 1/36 = 1/6$



pmf ($p(S)$) specifies the probability distribution for the sum S of counts from two dice.

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

機率密度函數
Probability Density Function

Definition. The **probability density function** (**"p.d.f."**) of a continuous random variable X with support S is an integrable function $f(x)$ satisfying the following:

(1) $f(x)$ is positive everywhere in the support S, that is, $f(x) > 0$, for all x in S

(2) The area under the curve $f(x)$ in the support S is 1, that is: $\int_S f(x)dx = 1$

(3) The probability that x belongs to A, where A is some interval, is given by the integral of $f(x)$ over that interval.

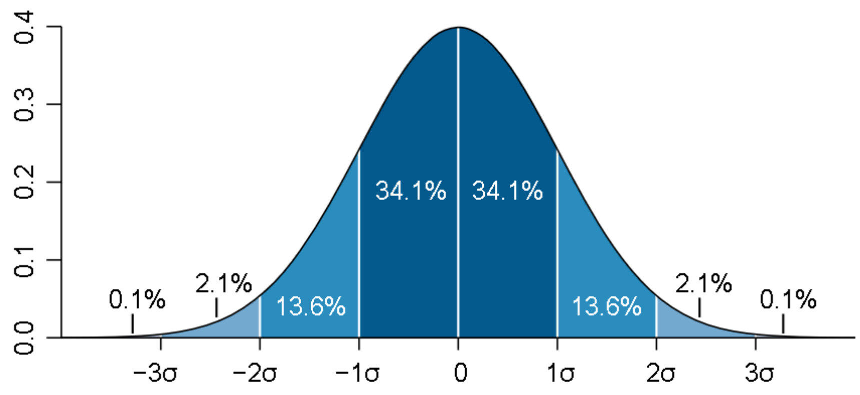
$$P(X \in A) = \int_A f(x)dx \qquad \mathrm{P}[a \le X \le b] = \int_a^b f(x)\,dx$$

The probability density of the normal distribution is:

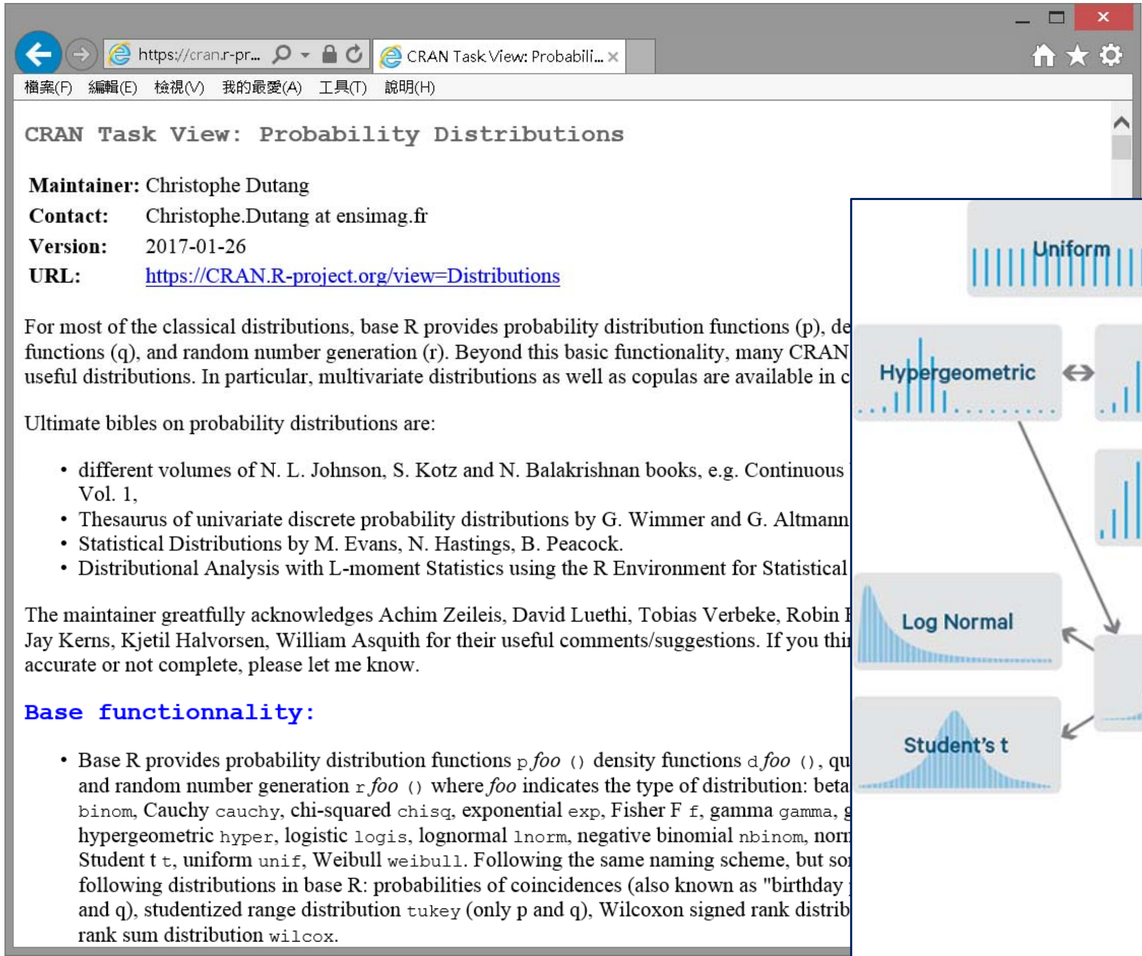
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

- μ is the mean or expectation of the distribution (and also its median and mode).
- σ is the standard deviation
- σ^2 is the variance



CRAN Task View: Probability Distribution



CRAN Task View: Probability Distributions

Maintainer: Christophe Dutang
Contact: Christophe.Dutang at ensimag.fr
Version: 2017-01-26
URL: https://CRAN.R-project.org/view=Distributions

For most of the classical distributions, base R provides probability distribution functions (p), de
functions (q), and random number generation (r). Beyond this basic functionality, many CRAN
useful distributions. In particular, multivariate distributions as well as copulas are available in c

Ultimate bibles on probability distributions are:

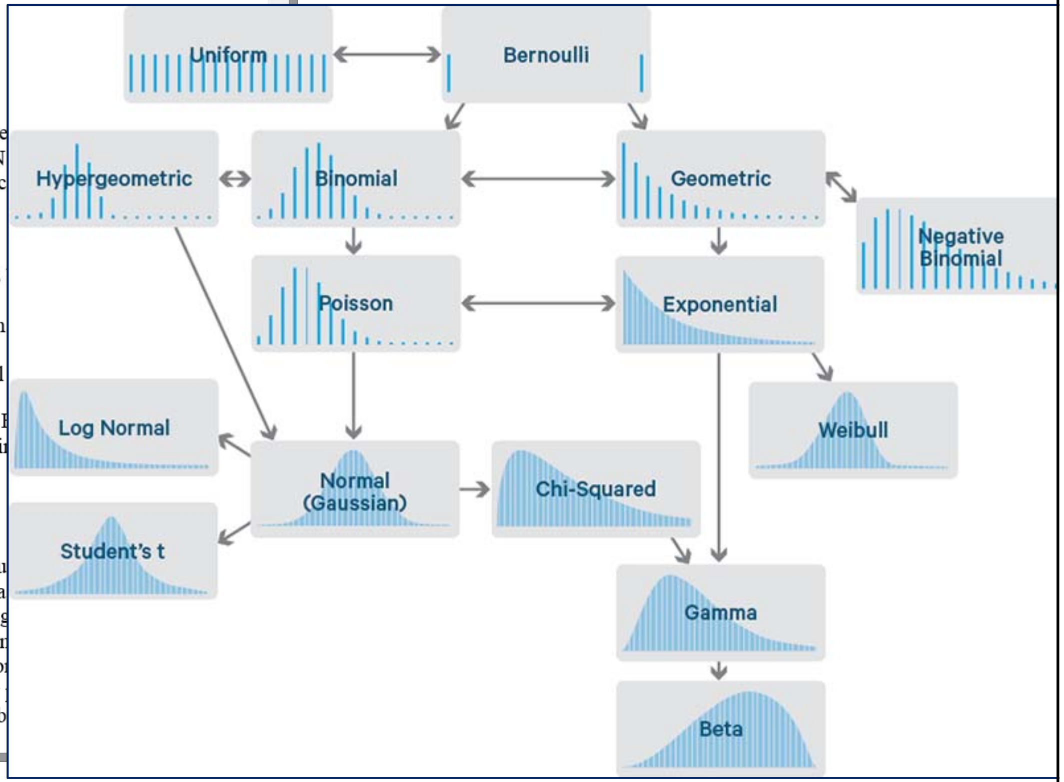
- different volumes of N. L. Johnson, S. Kotz and N. Balakrishnan books, e.g. Continuous Vol. 1,
- Thesaurus of univariate discrete probability distributions by G. Wimmer and G. Altmann
- Statistical Distributions by M. Evans, N. Hastings, B. Peacock.
- Distributional Analysis with L-moment Statistics using the R Environment for Statistical

The maintainer greatfully acknowledges Achim Zeileis, David Luethi, Tobias Verbeke, Robin I
Jay Kerns, Kjetil Halvorsen, William Asquith for their useful comments/suggestions. If you thi
accurate or not complete, please let me know.

Base functionnality:

- Base R provides probability distribution functions `p`*foo* `()` density functions `d`*foo* `()`, qu
and random number generation `r`*foo* `()` where *foo* indicates the type of distribution: beta
`binom`, Cauchy `cauchy`, chi-squared `chisq`, exponential `exp`, Fisher F `f`, gamma `gamma`, g
hypergeometric `hyper`, logistic `logis`, lognormal `lnorm`, negative binomial `nbinom`, norn
Student t `t`, uniform `unif`, Weibull `weibull`. Following the same naming scheme, but so
following distributions in base R: probabilities of coincidences (also known as "birthday
and q), studentized range distribution `tukey` (only p and q), Wilcoxon signed rank distrib
rank sum distribution `wilcox`.

https://cran.r-project.org/web/views/Distributions.html



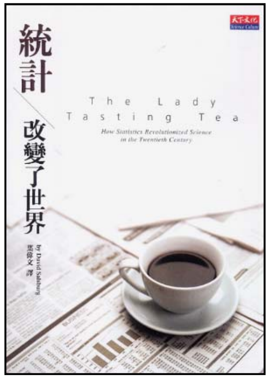
http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

Univariate Distribution Relationships: http://www.math.wm.edu/~leemis/chart/UDR/UDR.html
Wiki Category:Discrete distributions: https://en.wikipedia.org/wiki/Category:Discrete_distributions
Wiki Category:Continuous distributions: https://en.wikipedia.org/wiki/Category:Continuous_distributions

機率分佈在統計學中的重要性

統計改變了世界

- 十九世紀初:「機械式宇宙」的哲學觀
- 二十世紀: 科學界的統計革命。
- 二十一世紀: 幾乎所有的科學已經轉而運用統計模式了。



統計革命的起點

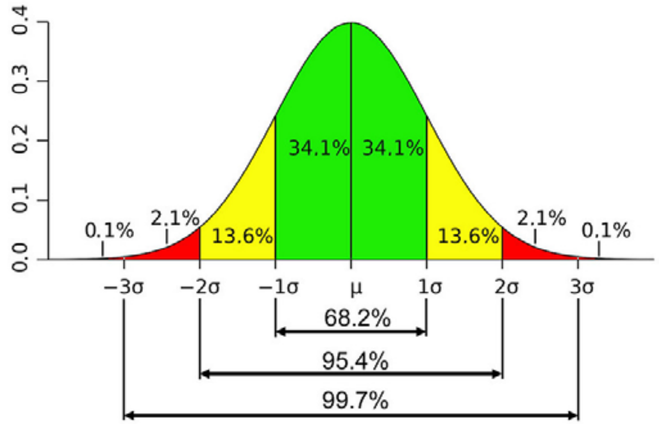
- Karl Pearson (1857-1936)，發表一系列和相關性(correlation)有關的論文，涉及動差、相關係數、標準差、卡方適合度檢定，奠定了現代統計學的基礎。
- 引入了統計模型的觀念: 如果能夠決定所觀察現象的**機率分佈的參數**，就可以了解所觀察現象的本質。

樣本變異數與樣本標準差

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

母體變異數與母體標準差

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$



Schweizer, B. (1984), Distributions Are the Numbers of the Future, in Proceedings of The Mathematics of Fuzzy Systems Meeting, eds. A. di Nola and A. Ventre, Naples, Italy: University of Naples, 137–149. (The present is that future.)

常用機率分佈的應用

- **Normal distribution**, for a single real-valued quantity that grow linearly (e.g. errors, offsets) ($X \sim N(\mu, \sigma^2)$)

- **Log-normal distribution**, for a single positive real-valued quantity that grow exponentially (e.g. prices, incomes, populations) ($log(X) \sim N(\mu, \sigma^2)$)

- **Discrete uniform distribution**, for a finite set of values (e.g. the outcome of a fair die) ($X \sim Unif(\{a, b\})$)

- **Binomial distribution**, for the number of "positive occurrences" (e.g. successes, yes votes, etc.) given a fixed total number of independent occurrences. ($X \sim B(n, p)$)

- **Negative binomial distribution**, for binomial-type observations but where the quantity of interest is the number of failures (r) before a given number of successes (k) occurs. ($X \sim NB(r, p)$)

- **Chi-squared distribution**, the distribution of a sum of squared standard normal variables; useful e.g. for inference regarding the sample variance of normally distributed samples. ($X \sim \chi^2_{(d)}$)

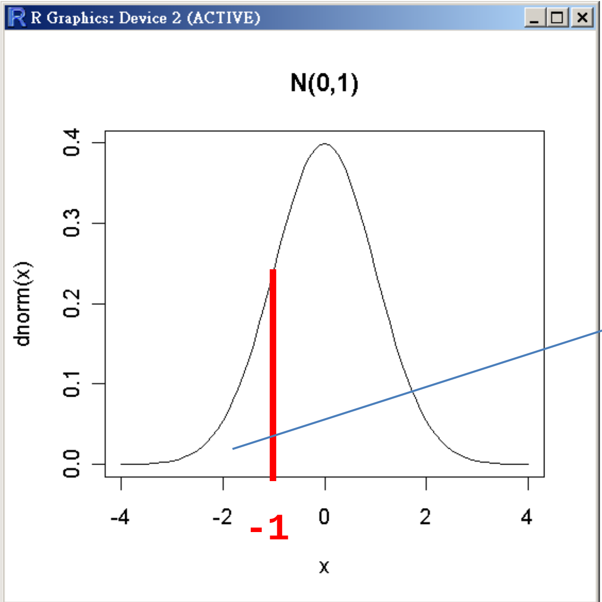
https://en.wikipedia.org/wiki/Probability_distribution

累積機率分配函數 CDF (p)

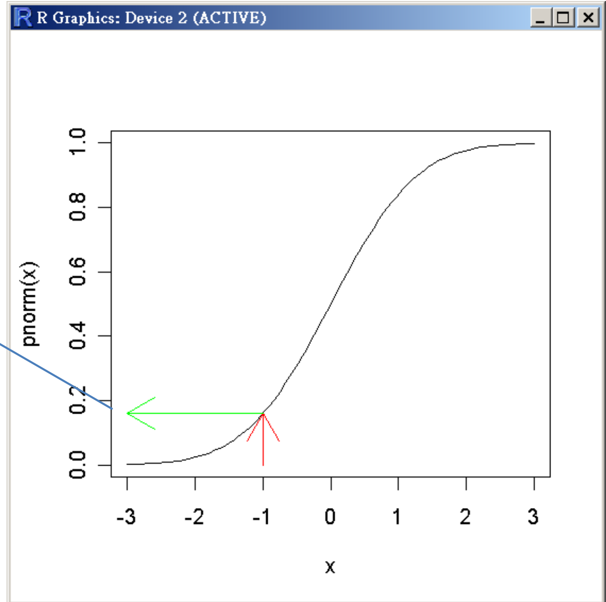
$$F_X(x) = P(X \leq x)$$

- The probability of obtaining a sample value that is less than or equal to x.

PDF



0.1586553



CDF

```
> curve(pnorm(x), -3, 3)
> arrows(-1, 0, -1, pnorm(-1), col="red")
> arrows(-1, pnorm(-1), -3, pnorm(-1), col="green")
> pnorm(-1)
[1] 0.1586553
```

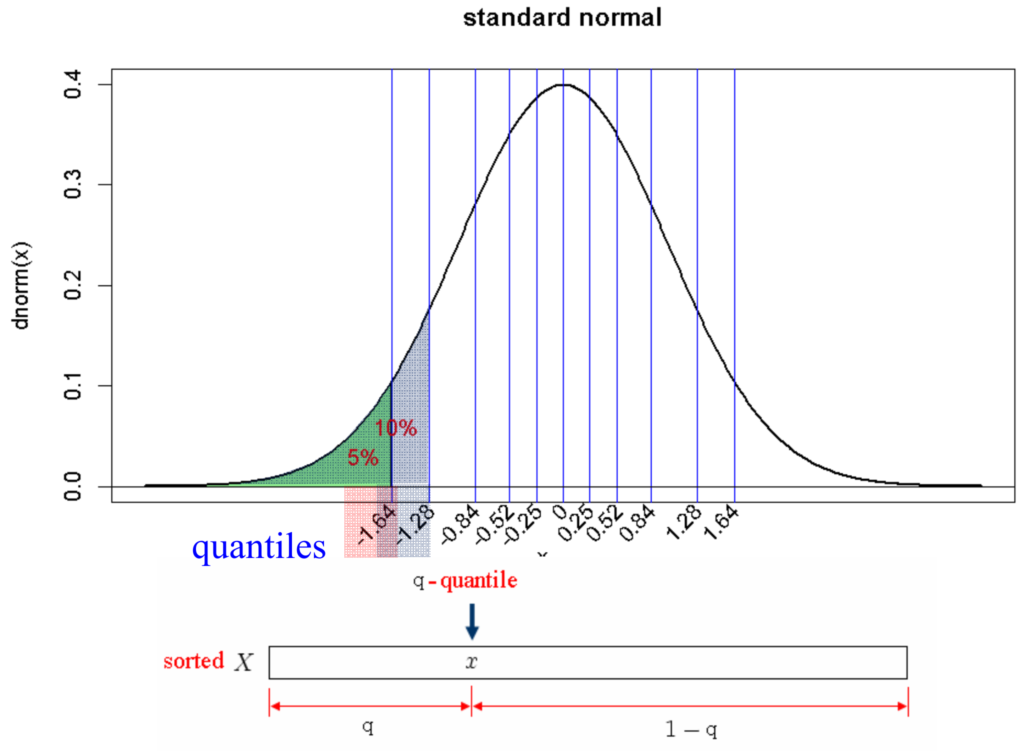
分位數 Quantiles (q)

$$F_X(x) = P(X \leq x) = p$$

- The quantile function is the inverse of the cumulative distribution function.

$$F_X^{-1}(p) = x$$

- We say that x is the q %-quantile if q% of the data values are $\leq x$.



$P(X < x) \leq q$ and $P(X > x) \leq 1 - q$.

常態母體平均數95%的信賴區間

$$\bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}$$

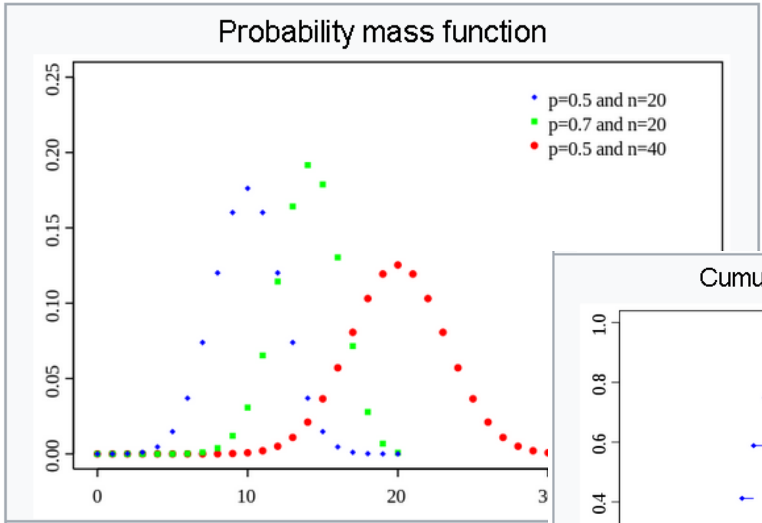
$$P(z_{0.025} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{0.975}) = 0.95$$

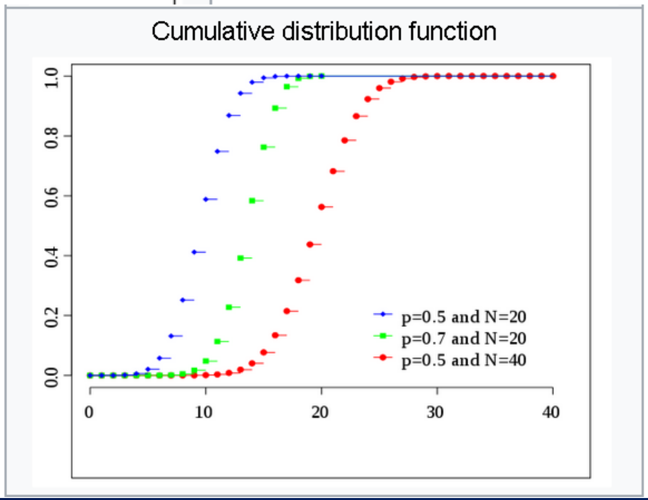
```
> # 2.5% quantile of N(0, 1)
> qnorm(0.025)
[1] -1.959964
> # the 50% quantile (the median) of N(0, 1)
> qnorm(0.5)
[1] 0
> qnorm(0.975)
[1] 1.959964
```

$$\Phi^{-1}(0.975)$$

二項式分佈 (Binomial)

- *X~B(n, p)*表示*n*次**伯努利試驗**中(size)，成功結果出現的次數。
- 例子: 擲一枚骰子十次，那麼擲得4的次數就服從***n* = 10**、***p* = 1/6**的二項分布*X~B(10, 1/6)*。
- `dbinom(x, size, prob)  # 機率公式值 P(X=x)`
- `pbinom(q, size, prob)  # 累加至q的機率值 P(X <= q)`
- `qbinom(p, size, prob)  # 已知累加機率值，對應的機率點。`
- `rbinom(n, size, prob)  # 隨機樣本數=n的二項隨機變數值。`



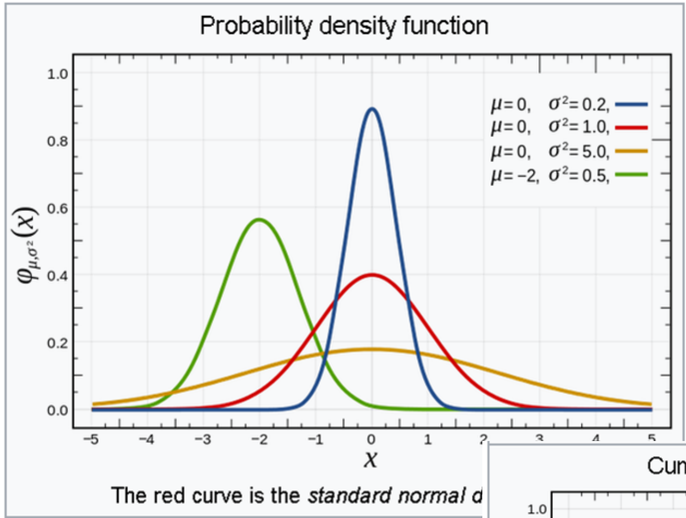


Notation	$B(n, p)$
Parameters	$n \in \mathbf{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \dots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$
Ex. kurtosis	$\frac{1-6p(1-p)}{np(1-p)}$
Entropy	$\frac{1}{2}\log_2\left(2\pi e\, np(1-p)\right) + O\left(\frac{1}{n}\right)$ in shannons. For nats, use the natural log in the log.
MGF	$(1-p+pe^t)^n$
CF	$(1-p+pe^{it})^n$
PGF	$G(z) = [(1-p)+pz]^n$.
Fisher information	$g_n(p) = \frac{n}{p(1-p)}$ (for fixed n)

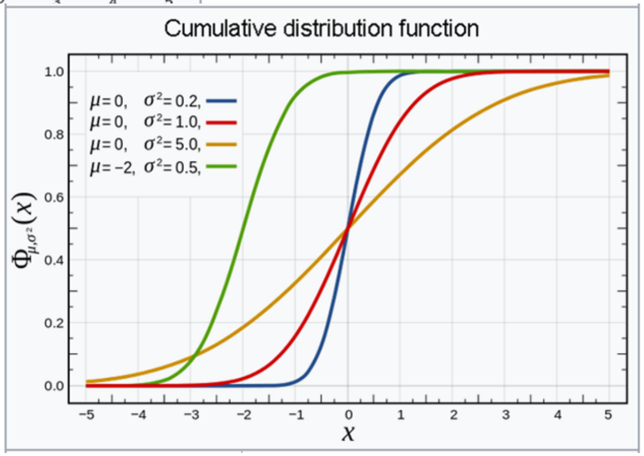
https://en.wikipedia.org/wiki/Binomial_distribution

常態分佈

- **`dnorm(x, mean, sd)`**#機率密度函數值 f(x)
- **`pnorm(q, mean, sd)`**#累加機率值P(X<= x)
- **`qnorm(p, mean, sd)`**#累加機率值p 對應的分位數
- **`rnorm(n, mean, sd)`**#常態隨機樣本



The red curve is the *standard normal*



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2}\ln(2\sigma^2\pi e)$
MGF	$\exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$
CF	$\exp\{i\mu t - \frac{1}{2}\sigma^2 t^2\}$
Fisher information	$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$

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

大數法則: The Law of Large Numbers

- 由具有有限(finite)平均數μ的母體隨機抽樣，隨著樣本數n的增加，樣本平均數 $\bar{X}_n$ 越接近母體的平均數μ 。

If $X_1, X_2, \cdots,$ an infinite sequence of i.i.d. random variables with finite expected value $E(X_1) = E(X_2) = \cdots = \mu < \infty$, then

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to \mu \quad \text{as} \quad n \to \infty$$

中央極限定理 (Central Limit Theorem)

- 由一具有平均數μ，標準差σ的母體中抽取樣本大小為n的簡單隨機樣本，當樣本大小n夠大時，**樣本平均數**的**抽樣分配**會近似於常態分配。

$X_1, X_2, X_3, \cdots$ be a set of n independent and identically distributed random variables having finite values of mean μ and variance $\sigma^2 > 0$.

$$S_n = X_1 + \cdots + X_n$$

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \to N(0,1) \quad \text{as} \quad n \to \infty$$

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

- 在一般的統計實務上，大部分的應用中均假設當樣本大小為30(含)以上時 $\bar{X}_n$的抽樣分配即近似於常態分配。
- 當母體為常態分配時，不論樣本大小，樣本平均數的抽樣分配仍為常態分配。

應用CLT算機率

- 於某考試中，考生之通過標準機率為0.7，以隨機變數表示考生之通過與否 (X=1表示通過) (X=0表示不通過)，其機率分配為 P(X=1)=0.7, P(X=0)=0.3。
  1. 計算母體平均數及變異數。
  2. 假如有210名考生，計算「平均通過人數」的平均數及變異數。
  3. 計算通過人數> 126的機率。

1.

$$\mu = E(X) = p = 0.7$$

$$\sigma^2 = Var(X) = p(1-p) = 0.21$$

2.

$$X_1, X_2, \cdots, X_{210}:$$

$$X_i = 1 : \text{success}$$

$$X_i = 0 : \text{fail}$$

$$\bar{X}_{210} = \frac{X_1 + \cdots + X_{210}}{210}$$

$$\mu_{\bar{X}} = \mu = 0.7$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{210} = 0.001$$

3.

$$P(X_1 + X_2 + \cdots + X_{210} > 126)$$

$$= P(\bar{X} > \frac{126}{210})$$

$$= P(\bar{X} > 0.6)$$

$$= P(Z > \frac{0.6 - 0.7}{\sqrt{0.001}})$$

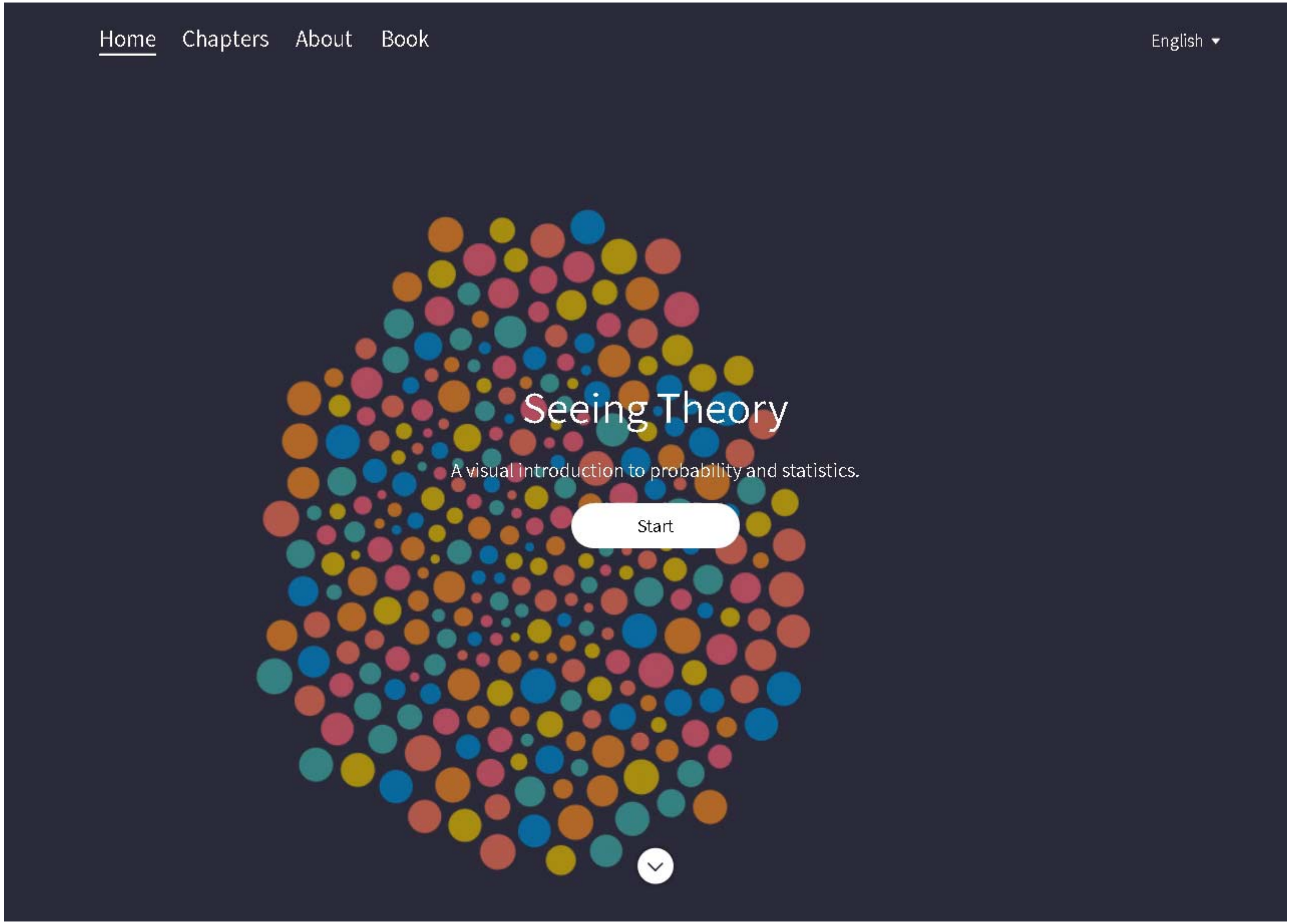
$$= P(Z > -3.16228)$$

$$= 0.99922$$

應用CLT

中央極限定理:
樣本平均之抽樣分佈

https://students.brown.edu/seeing-theory/



練習: 用R程式模擬算機率: 我們要生女兒

一對夫婦計劃生孩子生到有女兒才停，或生了三個就停止。他們會擁有女兒的機率是多少？

- **第I 步：機率模型**
  - 每一個孩子是女孩的機率是0.49 ，是男孩的機率是0.51。
    各個孩子的性別是互相獨立的。
- **第2 步：分配隨機數字。**
  - 用兩個數字模擬一個孩子的性別: 00, 01, 02, ..., 48 ＝ 女孩; 49, 50, 51, ..., 99 ＝ 男孩
- **第3 步：模擬生孩子策略**
  - 從表A當中讀取一對一對的數字，直到這對夫婦有了女兒，或已有三個孩子。

6905	16	48	17	8717	40	9517	845340	648987	20
男女	女	女	女	男女	女	男女	男男女	男男男	女
+	+	+	+	+	+	+	+	−	+

- 10次重複中，有9次生女孩。會得到女孩的機率的估計是9/10=0.9。
- 如果機率模型正確的話，用數學計算會有女孩的真正機率是0.867。(我們的模擬答案相當接近了。除非這對夫婦運氣很不好，他們應該可以成功擁有一個女兒。)

用R程式模擬算機率: 我們要生女兒

```
girl.born <- function(n, show.id = F){

  girl.count <- 0
  for (i in 1:n) {
    if (show.id) cat(i,": ")
    child.count <- 0
    repeat {
        rn <- sample(0:99, 1, replace=T)
        if (show.id) cat(paste0("(", rn, ")"))
        is.girl <- ifelse(rn <= 48, TRUE, FALSE)
        child.count <- child.count + 1
        if (is.girl){
          girl.count <- girl.count + 1
          if (show.id) cat("女+")
          break
        } else if (child.count == 3) {
          if (show.id) cat("男")
          break
        } else{
          if (show.id) cat("男")
        }
    }
    if (show.id) cat("\n")
  }
  p <- girl.count / n
  p

}
```

```
> girl.p <- 0.49 + 0.51*0.49 + 0.51^2*0.49
> girl.p
[1] 0.867349
>
> girl.born(n=10, show.id = T)
1 : (73)男(18)女+
2 : (23)女+
3 : (53)男(74)男(64)男
4 : (95)男(20)女+
5 : (63)男(16)女+
6 : (48)女+
7 : (67)男(51)男(44)女+
8 : (74)男(99)男(25)女+
9 : (47)女+
10 : (81)男(41)女+
[1] 0.9
> girl.born(n=10000)
[1] 0.8674
```

進階選讀

二項式分佈

X~B(10, 0.8)

- 利用二項分配理論公式，計算機率公式值 P(X=3)。

```
> factorial(10)/(factorial(3)*factorial(7))*0.8^3*0.2^7
[1] 0.000786432
```

- 利用R函數，計算機率值 P(X=3)。

```
> dbinom(3, 10, 0.8)
[1] 0.000786432
```

- 計算P(X<= 3)- P(X<= 2) ，並和P(X=3)相比較。

```
> pbinom(3, 10, 0.8)- pbinom(2, 10, 0.8)
[1] 0.000786432
```

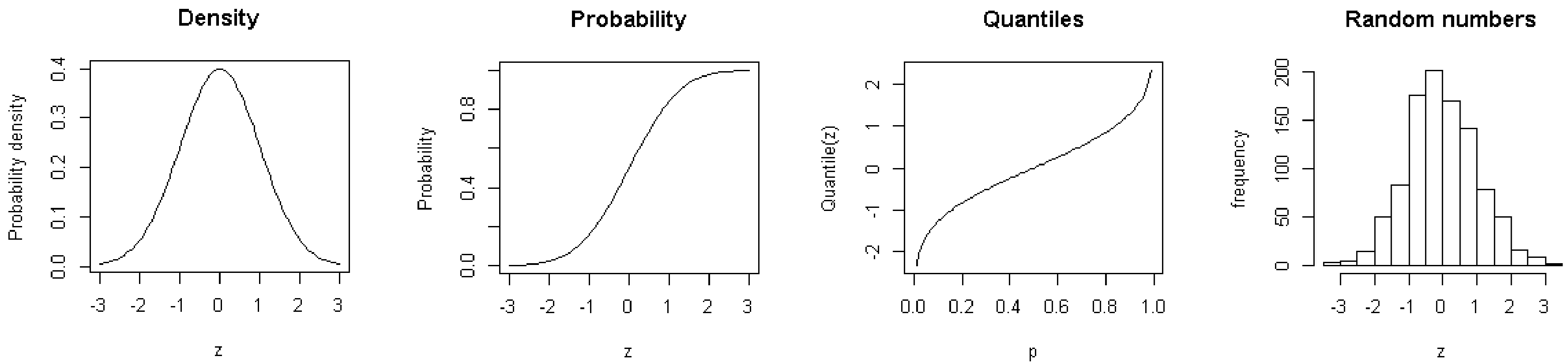
- 已知累加機率值為0.1208，求對應的分位數。

```
> qbinom(0.1208, 10, 0.8)
[1] 6
> pbinom(6, 10, 0.8)
[1] 0.1208739
```

常態分佈

```
> #Z ~ N(0, 1)
> dnorm(0)
[1] 0.3989423
> pnorm(-1)
[1] 0.1586553
> qnorm(0.975)
[1] 1.959964
```

```
> dnorm(10, 10, 2) # X~N(10, 4)
[1] 0.1994711
> pnorm(1.96, 10, 2)
[1] 2.909907e-05
> qnorm(0.975, 10, 2)
[1] 13.91993
> rnorm(5, 10, 2)
[1]  9.043357 11.721717  7.763277  9.563463 10.072386
> pnorm(15, 10, 2) - pnorm(8, 10, 2)  # P(8<=X<=15)
[1] 0.8351351
```



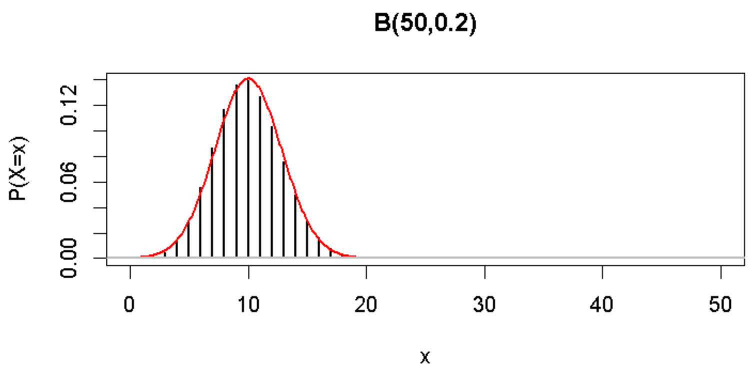

```
> par(mfrow=c(1,4))
> curve(dnorm, -3, 3, xlab="z", ylab="Probability density", main="Density")
> curve(pnorm, -3, 3, xlab="z", ylab="Probability", main="Probability")
> curve(qnorm, 0, 1, xlab="p", ylab="Quantile(z)", main="Quantiles")
> hist(rnorm(1000), xlab="z", ylab="frequency", main="Random numbers")
```

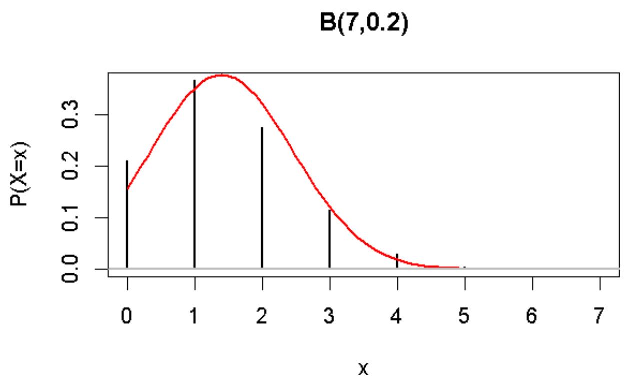
以常態機率逼近二項式機率

The normal approximation to the binomial

Let the number of successes X be a binomial r.v. with parameters n and p. Also, let $\mu = np$, and $\sigma = \sqrt{np(1-p)}$. Then if $np \geq 5$, $n(1-p) \geq 5$, we consider $\phi(x|\mu, \sigma)$ an acceptable approximation of the binomial.

```
n <- 50 # n = 7
p <- 0.2
mu <- n * p
sigma <- sqrt(n * p * (1 - p))
x <- 0:n
plot(x, dbinom(x, n, p), type = 'h', lwd = 2, xlab = "x", ylab = "P(X=x)",
     main = paste0("B(",n,",",p,")"))
z <- seq(0, n, 0.1)
lines(z, dnorm(z, mu, sigma), col = "red", lwd = 2)
abline(h = 0, lwd = 2, col = "grey")
```





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 $r \leq min\{p, n\}$.

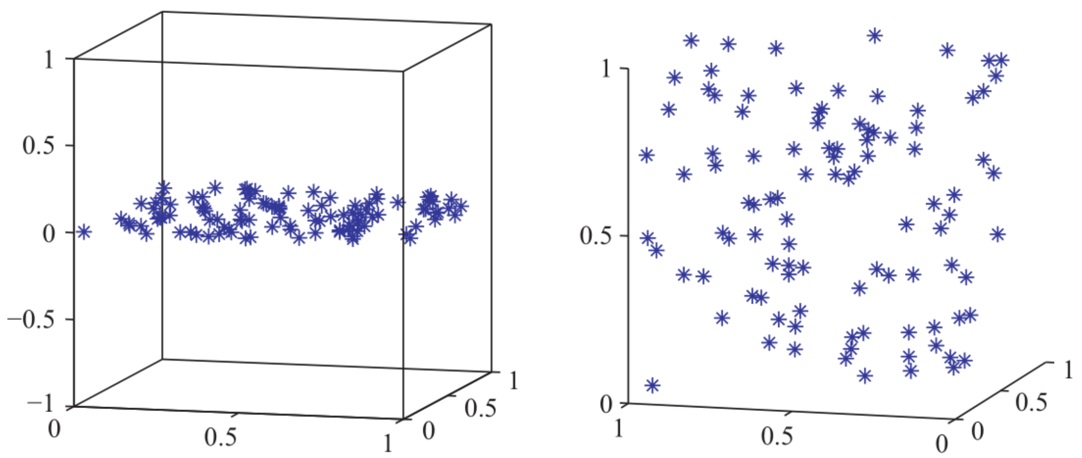
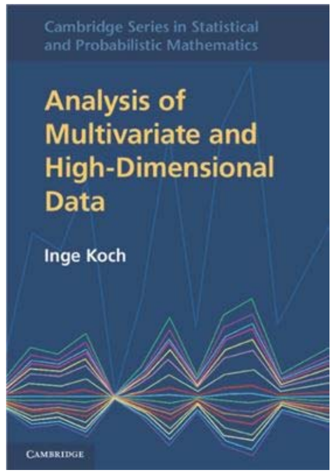


Figure 2.12 Distribution of 100 points in 2D and 3D unit space.



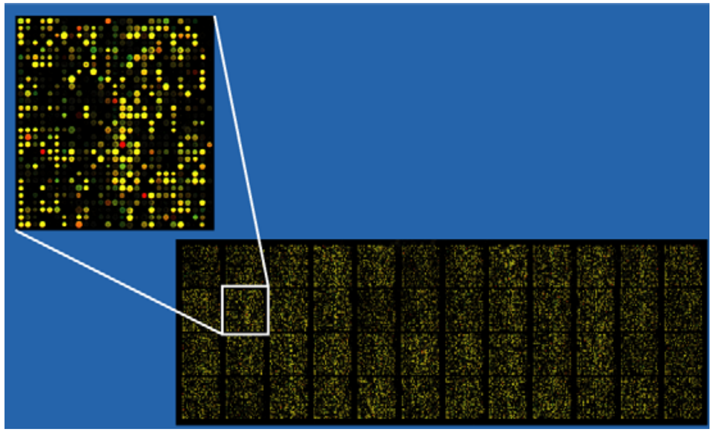
HDLSS examples

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

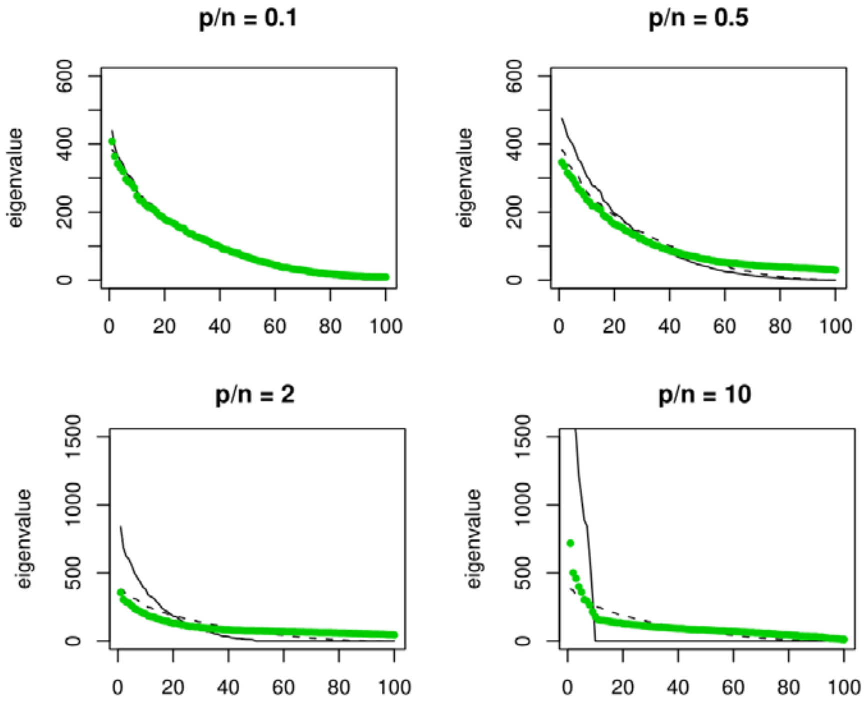


Figure 1: Ordered eigenvalues of the sample covariance matrix S (thin black line) and that of an alternative estimator $S^\star$ (fat green line, for definition see Tab. 1), calculated from simulated data with underlying p-variate normal distribution, for $p = 100$ and various ratios p/n. The true eigenvalues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

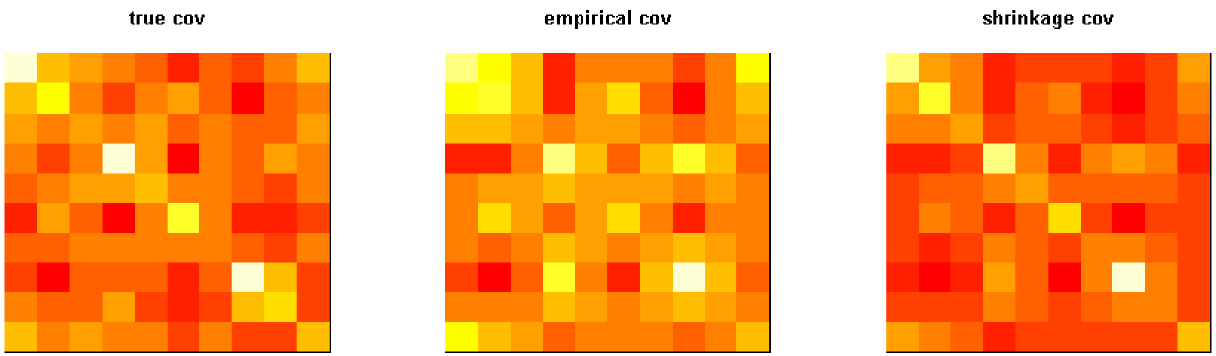
google: Penalized/Regularized/Shrinkage Methods

Example Script from corpcor Package

corpcor: Efficient Estimation of Covariance and (Partial) Correlation

```
> 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")
```

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)
```