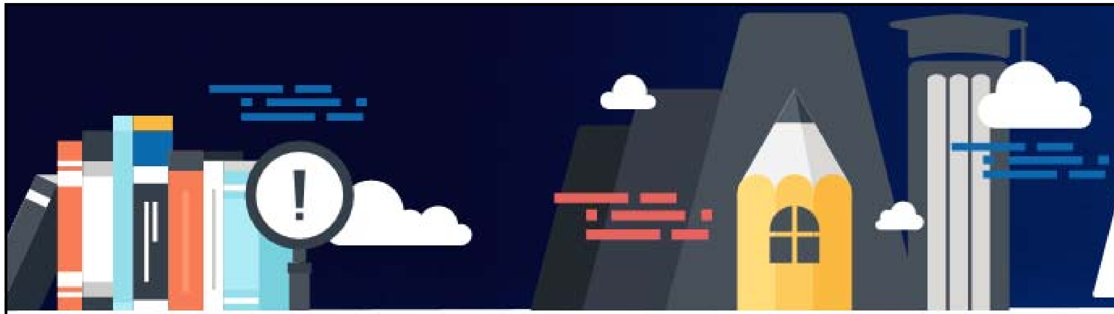


台灣人工智慧學校

經理人週末研修班

敘述統計
與機率分佈

吳漢銘

國立臺北大學 統計學系

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
- 常見統計名詞
- 機率分佈 (Probability distribution)
  - 統計分配之描述、常態分佈)
- 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	↓	C#
6	7	↑	JavaScript
7	6	↓	Visual Basic .NET
8	16	⇈	R
9	10	↑	PHP
10	8	↓	Perl

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

What is Statistics?

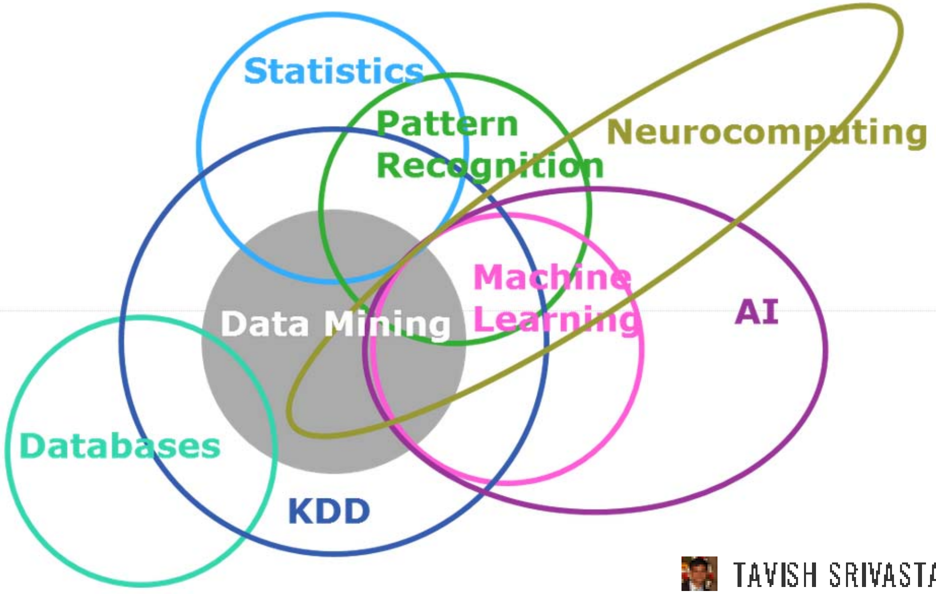
- Merriam-Webster dictionary defines statistics as "a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

- 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/

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

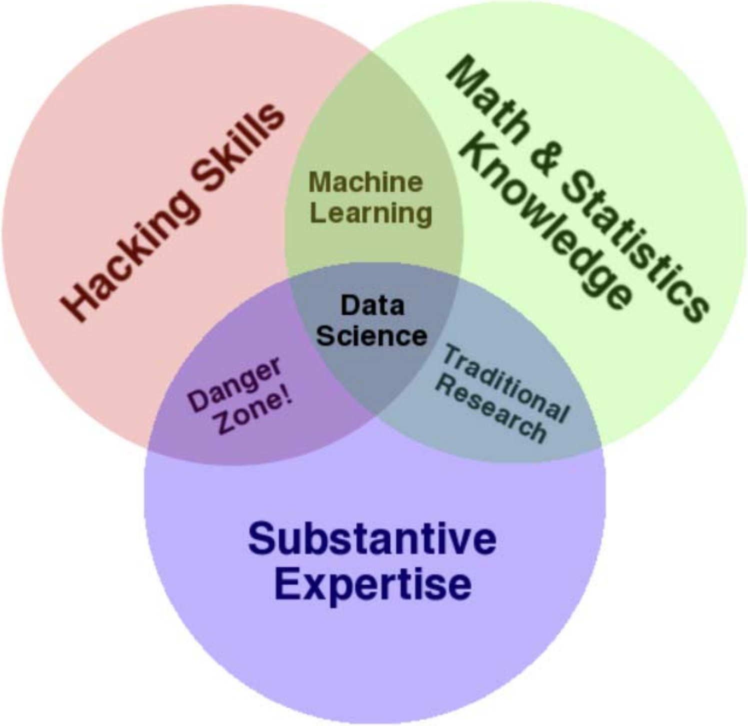
https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

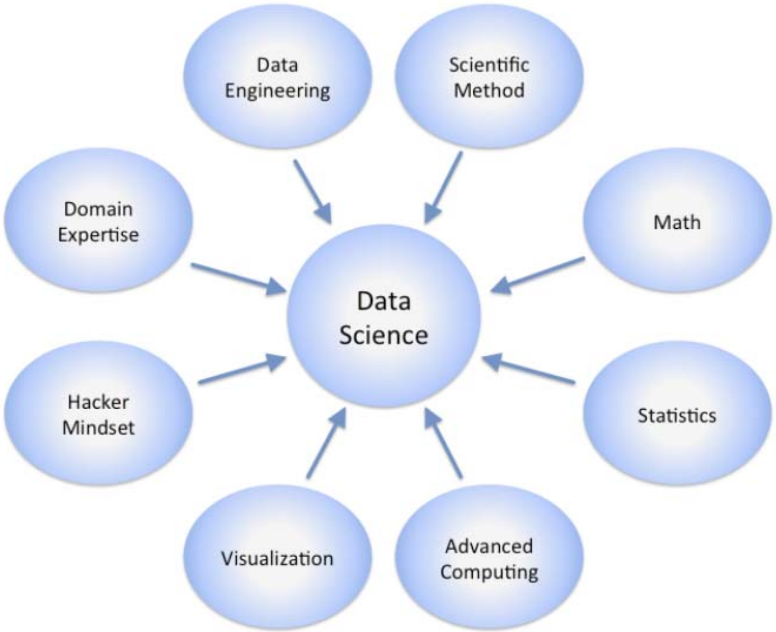
- **Machine Learning** is an algorithm that can learn from data without relying on rules-based programming.
- **Statistical Modelling** is the formalization of relationships between variables in the form of mathematical equations.

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

數據科學 Data Science

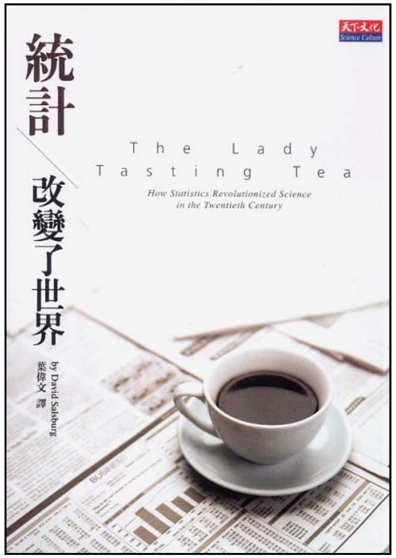
The Data Science Venn Diagram
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

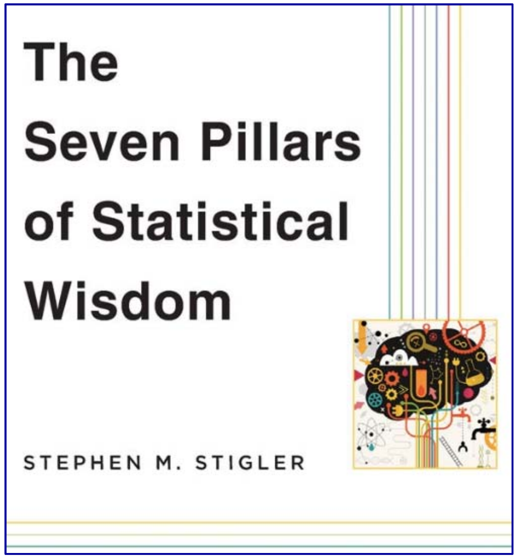




Source: By Calvin.Andrus (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

推薦兩本書





(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

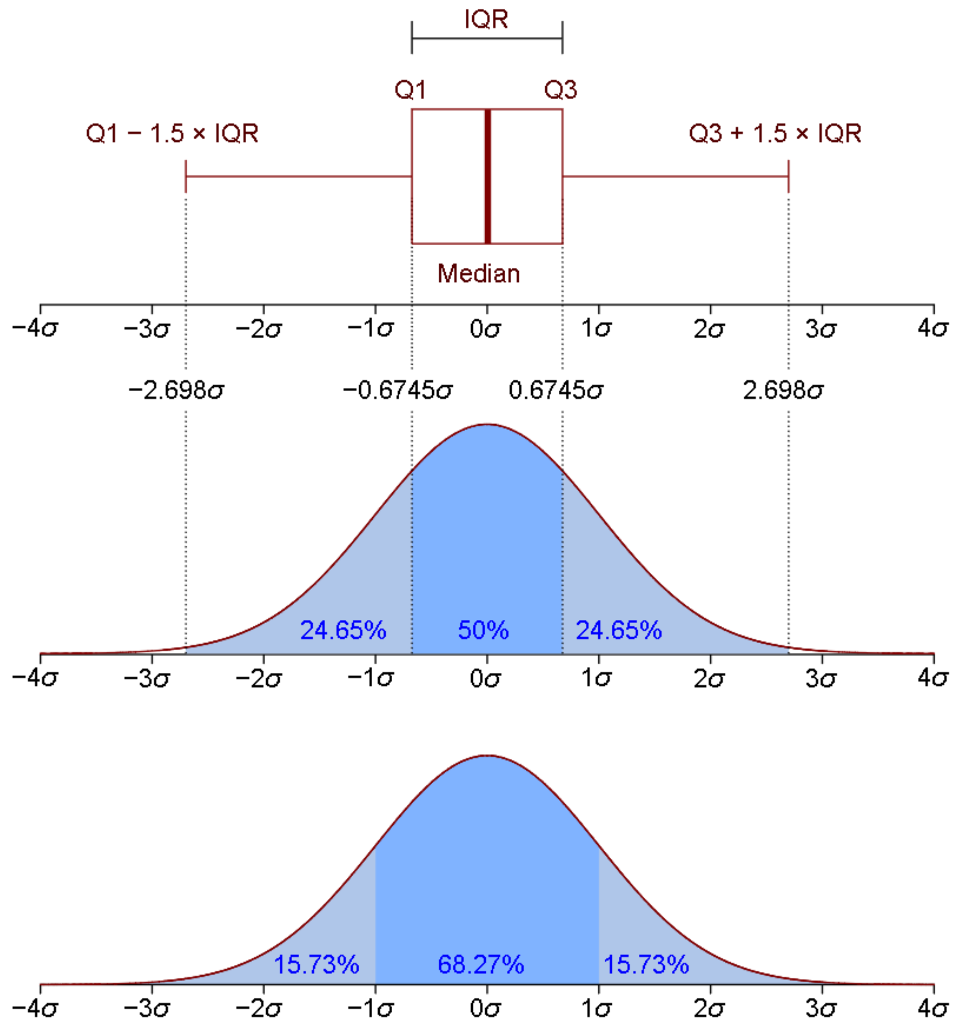
- **資料中心趨勢:**
  平均數(average)
  眾數(mode)
  中位數(median)
- **資料分散程度:**
  四分位數(Quartile)
  全距(range)
  四分位距(interquartile range, IQR)
  百位數(percentile)
  標準差(standard deviation)
  變異數(variance)

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

$n =$ The number of data points

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

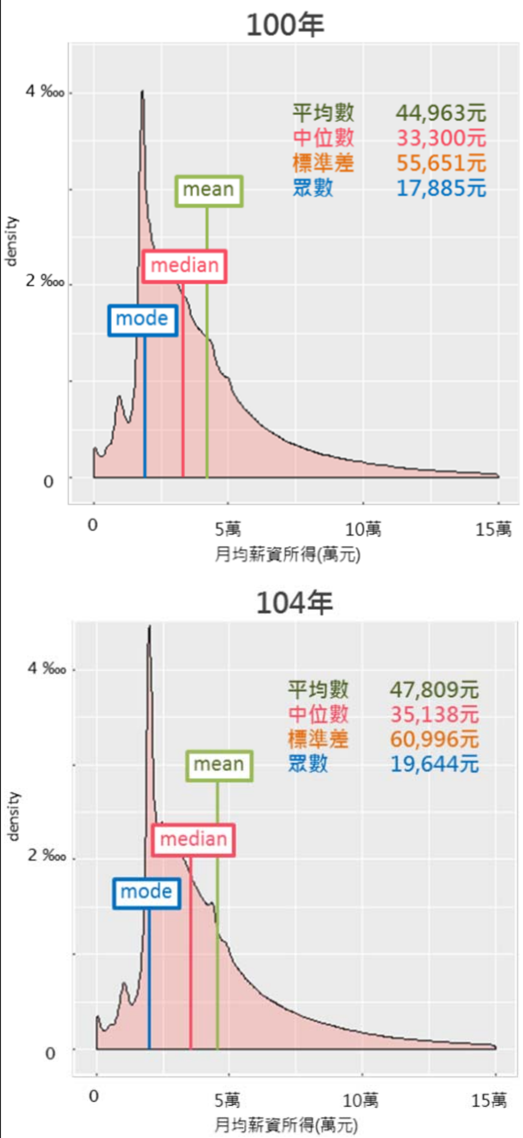
$x_i =$ Each of the values of the data



https://zh.wikipedia.org/wiki/四分位距

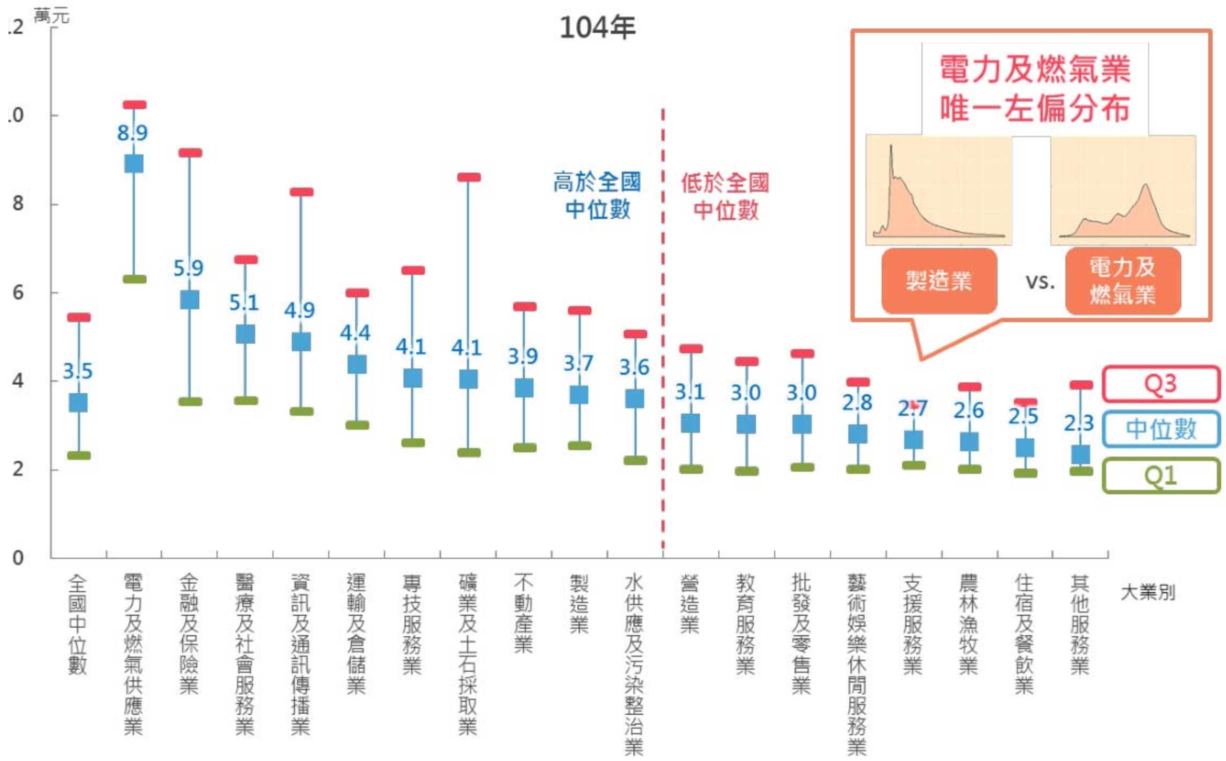
範例: 由財稅大數據探討臺灣近年薪資樣貌

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



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(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
```

```
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)))])
}
```

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

Pearson Correlation Coefficient

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

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

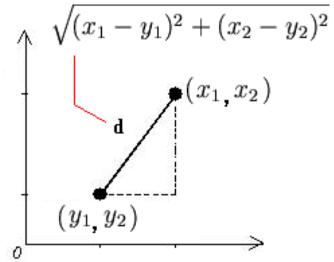
$$y = (y_1, y_2, \cdots, y_n)$$

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

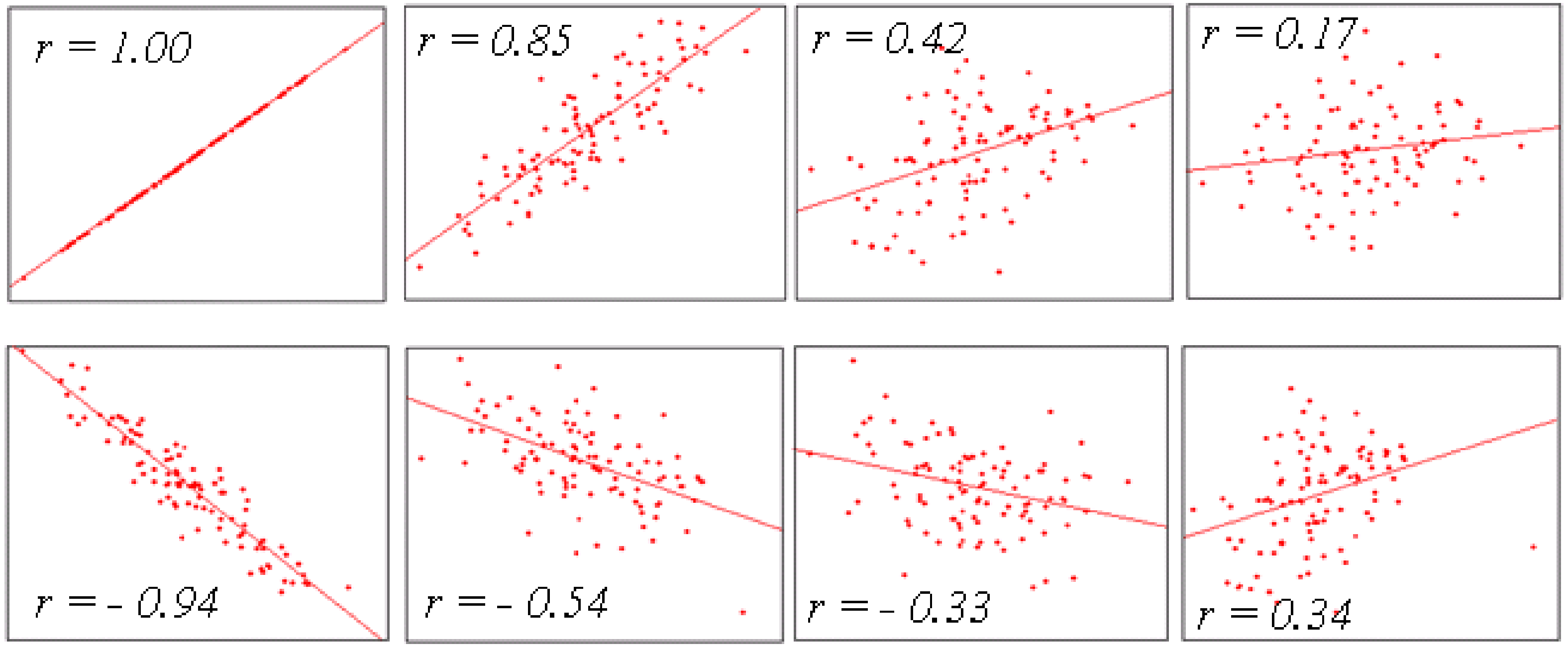
Euclidean Distance



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

- 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	$s(i,j) = \dfrac{\mathrm{cov}(x_i, x_j)}{\sqrt{\mathrm{var}(x_i)\,\mathrm{var}(x_j)}}$
Spearman correlation (r_i is ranked x_i)	$s(i,j) = \dfrac{\mathrm{cov}(r_i, r_j)}{\sqrt{\mathrm{var}(r_i)\,\mathrm{var}(r_j)}}$
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)

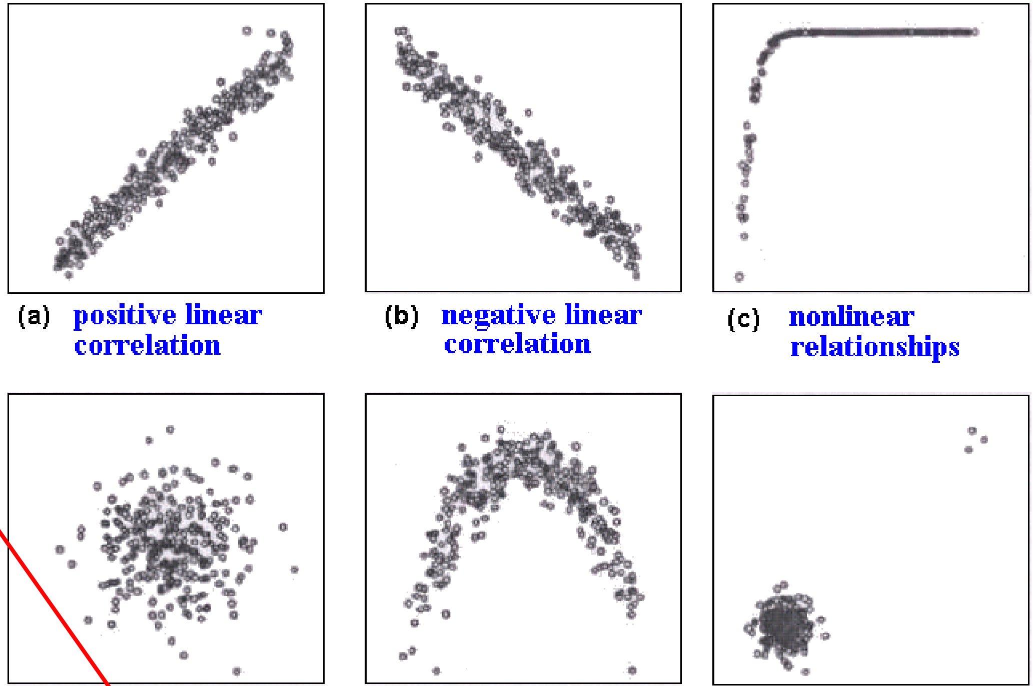
- C: concordant pair: $(x_j - x_i)(y_j - y_i) > 0$
- D: discordant pair: $(x_j - x_i)(y_j - y_i) < 0$
- tie:

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)



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

measures the strength of a linear relationship

measure any monotonic relationship between two variables

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

more robust

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

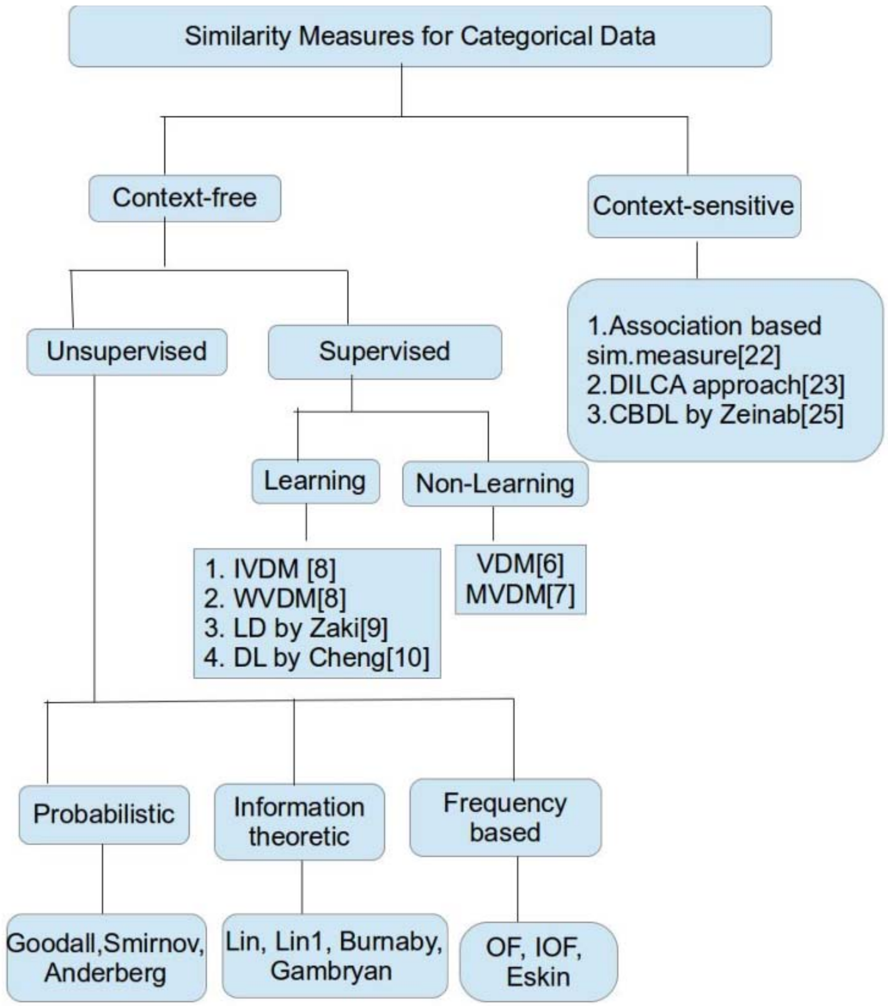
Similarity Measures for Categorical Data

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 $r \le \min\{p, n\}$.
  - 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.

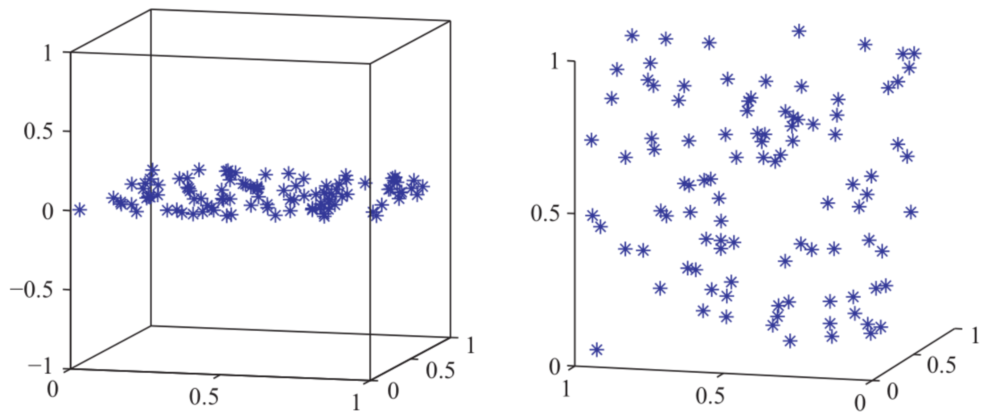
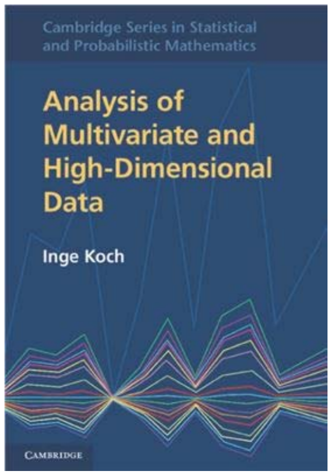


Figure 2.12 Distribution of 100 points in 2D and 3D unit space.



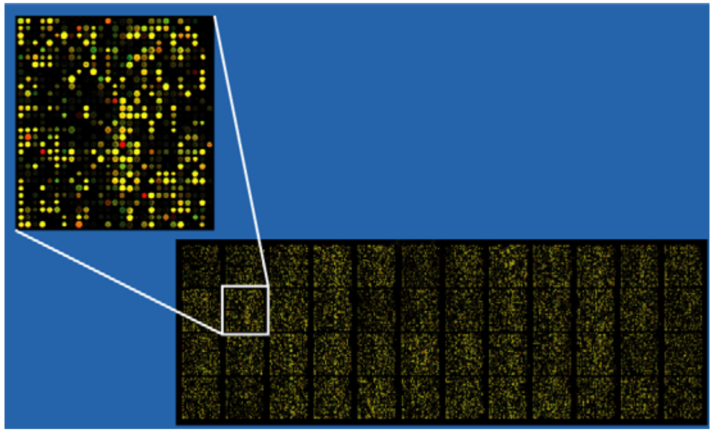
HDLSS examples

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

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

a shrinkage estimator

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

"Small *n*, Large *p*"

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

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

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

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

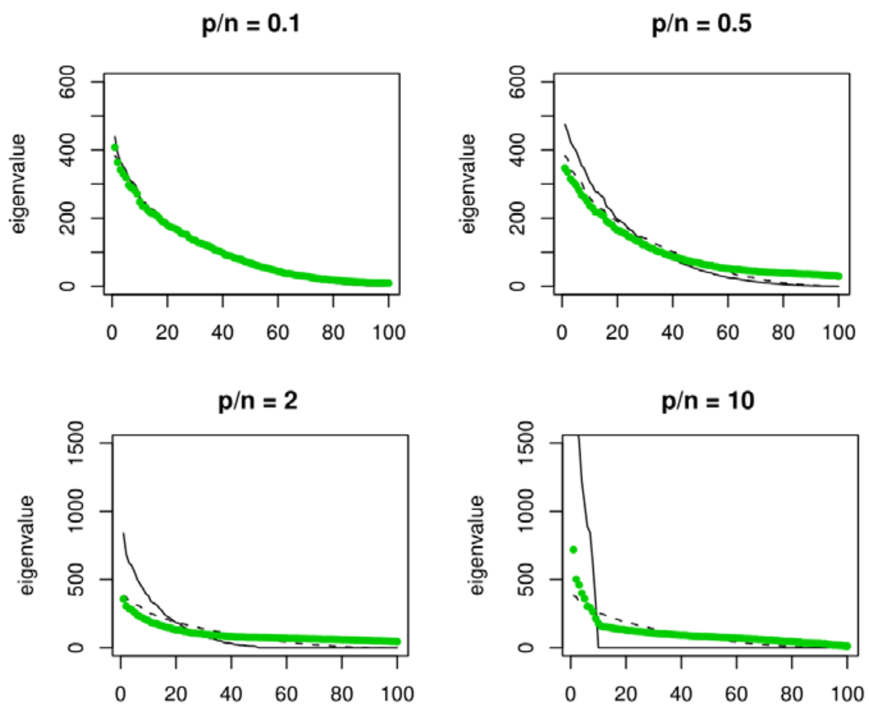


Figure 1: Ordered eigenvalues of the sample covariance matrix S (thin black line) and that of an alternative estimator $S^\star$ (fat green line, for definition see Tab. 1), calculated from simulated data with underlying p-variate normal distribution, for $p = 100$ and various ratios p/n. The true eigenvalues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

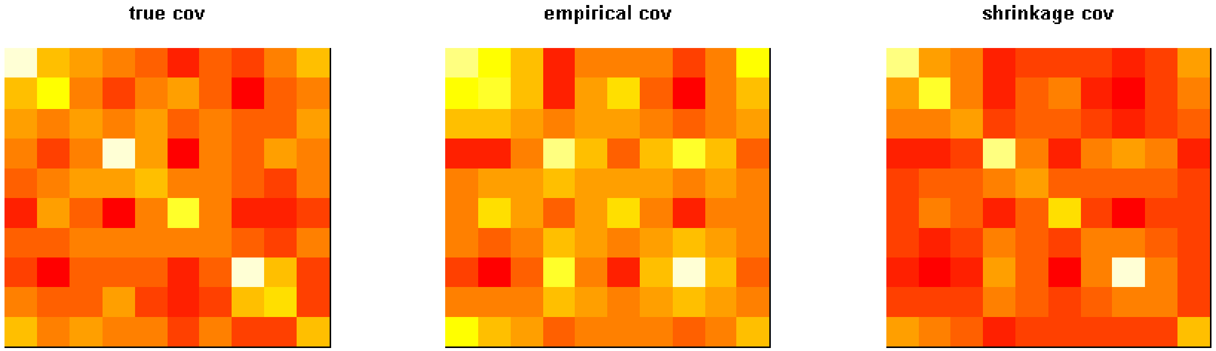
Schäfer, J., and K. Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology . 4: 32.

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
```

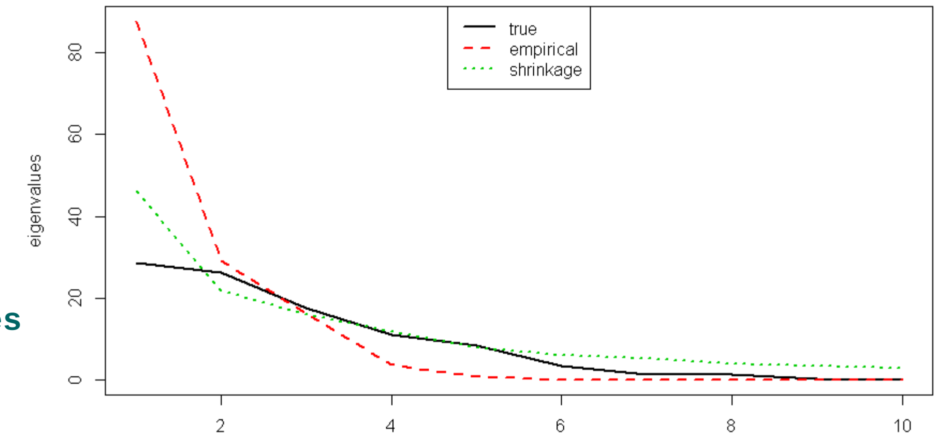
mvrnorm {MASS}:
Simulate from a Multivariate Normal Distribution
`mvrnorm(n = 1, mu, Sigma, ...)`



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}`



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.
- **Event (事件), *E***: an event is a subset of the sample space.
- **Trial (試驗)**: a single performance of an experiment whose outcome is in *S*.

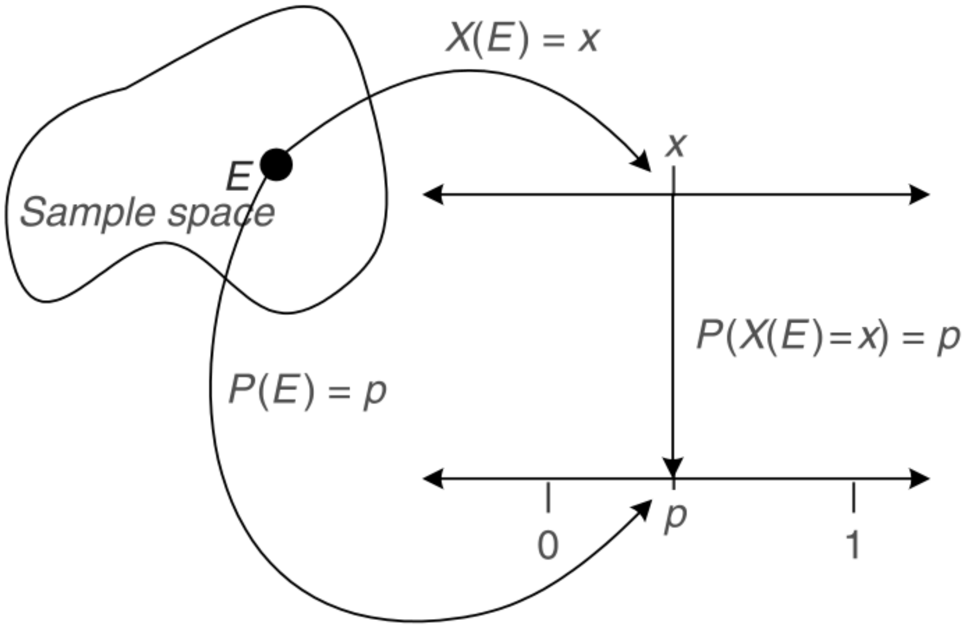
- In the experiment of tossing 4 coins, we may consider tossing each coin as a trial and therefore say that there are 4 trials in the experiment.

- 例子1: 投擲兩硬幣看看正反面之樣本空間 S={HH, HT, TH, TT}.
- 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} .

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

統計分配 (Statistical Distributions)

Four fundamental items can be calculated for a statistical distribution:

- **機率密度函數值(d):** point probability $P(X=x)$ or *probability density function* $f(x)$: `dnorm()`

- **累積機率函數值 (p):** cumulative probability distribution function, $F(x) = P(X<=x)$: `pnorm()`

- **分位數 (q):** the quantiles of the distribution: `qnorm()`
  The inverse of a distribution. That is, given a probability value p, we wish to find the quantile, x, such that $P(X \le x|\theta) = p$.

- **隨機數 (r):** the random numbers generated from the distribution: `rnorm()`

Formal definition

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

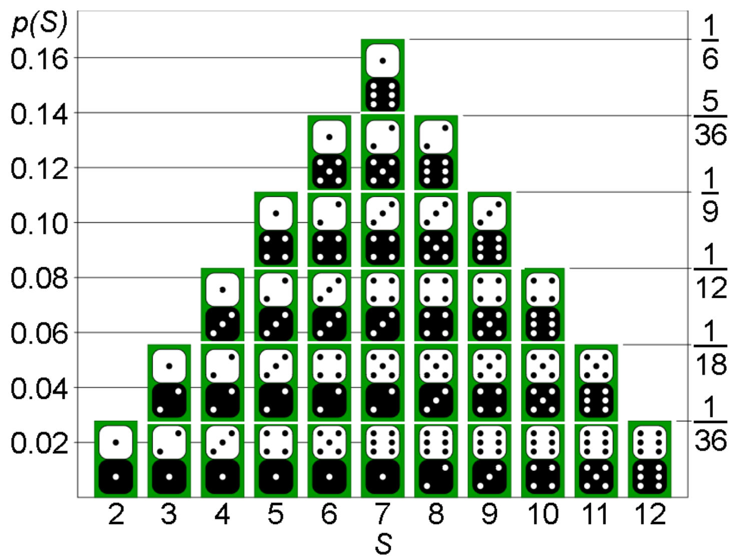
Suppose that $X: S \to A$ ($A \subseteq \mathbf{R}$) is a discrete random variable defined on a sample space S. Then the probability mass function $f_X: A \to [0, 1]$ for X is defined as

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

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:

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

$S = X_1 + X_2$
$X_1 \sim DiscreteUniform\ (1, 6), n=6.$
$X_2 \sim DiscreteUniform(1, 6), n=6.$
$f(X_1 = k) = f(X_2 = k) = 1/6, k = 1,..,6.$
$f(S = s) = p(S = s), s=2, ..., 12.$
$P(S = 2) = 1/36, P(S=3)=2/36, ..., P(S=12)=1/36$
$P(X_1 + X_2 > 9) = 1/12 + 1/18 + 1/36 = 1/6$



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

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

Probability Density Function
機率密度函數

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) If $f(x)$ is the p.d.f. of x, then the probability that x belongs to A, where A is some interval, is given by the integral of $f(x)$ over that interval, that is:

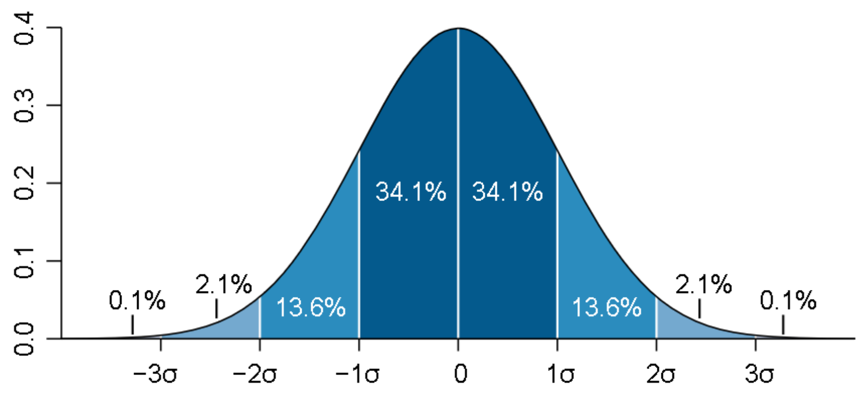
$$P(X \in A) = \int_A f(x)dx \qquad \mathrm{P}[a \le X \le b] = \int_a^b f(x)\,dx$$

The probability density of the normal distribution is:

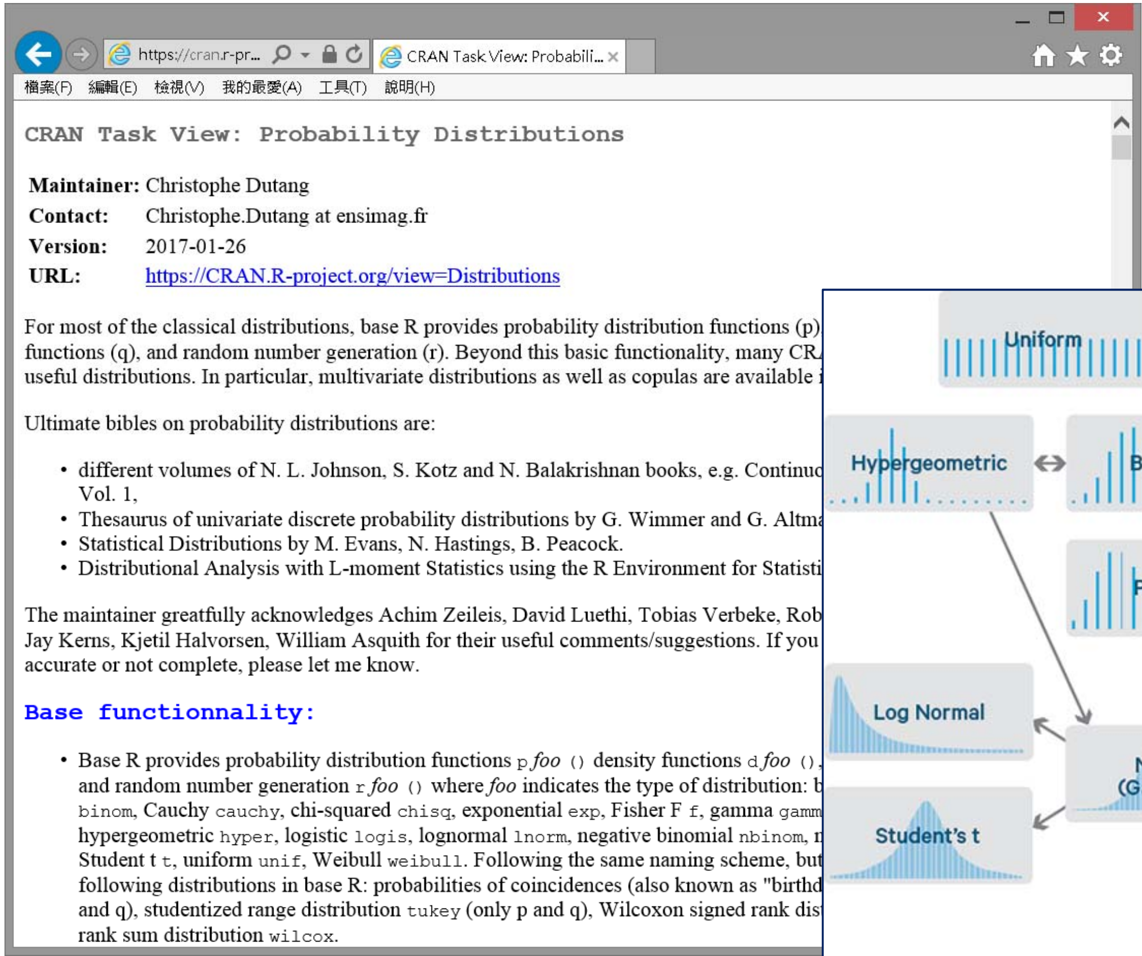
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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



https://cran.r-project.org/ CRAN Task View: Probabili...
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

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) functions (q), and random number generation (r). Beyond this basic functionality, many CR useful distributions. In particular, multivariate distributions as well as copulas are available i

Ultimate bibles on probability distributions are:

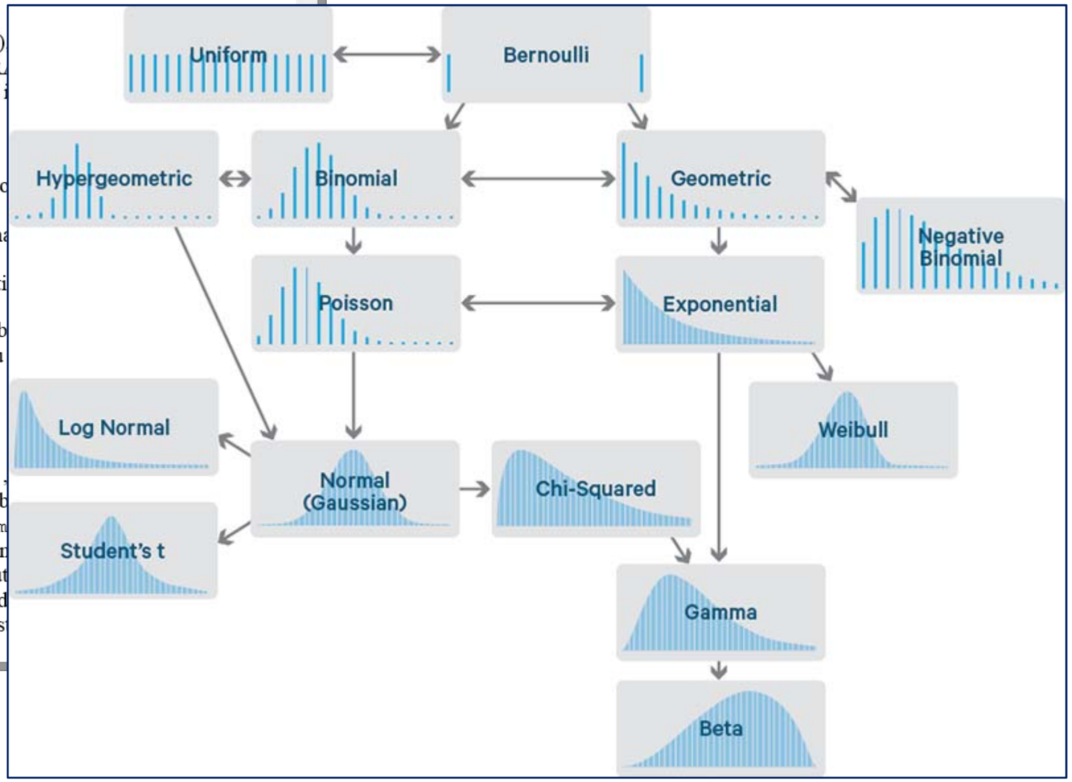
- different volumes of N. L. Johnson, S. Kotz and N. Balakrishnan books, e.g. Continu Vol. 1,
- Thesaurus of univariate discrete probability distributions by G. Wimmer and G. Altm
- Statistical Distributions by M. Evans, N. Hastings, B. Peacock.
- Distributional Analysis with L-moment Statistics using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

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

Base functionality:

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

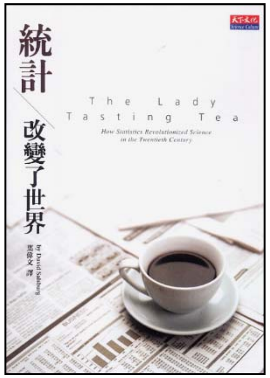
https://cran.r-project.org/web/views/Distributions.html



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

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

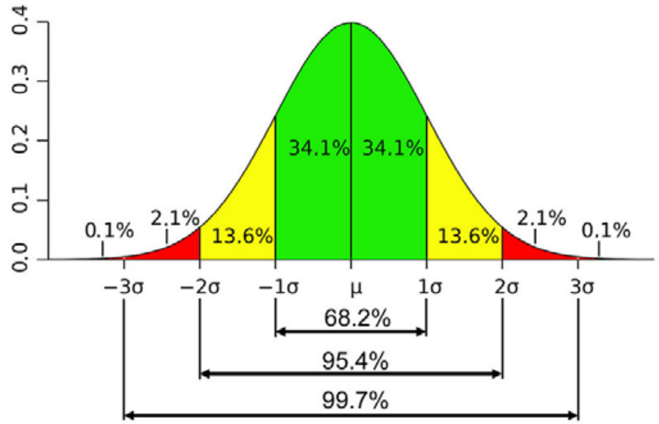
- 1895-1898，發表一系列和相關性(correlation) 有關的論文，涉及動差、相關係數、標準差、卡方適合度檢定，奠定了現代統計學的基礎。
- 引入了統計模型的觀念: 如果能夠決定所觀察現象的**機率分佈的參數**，就可以了解所觀察現象的本質。

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

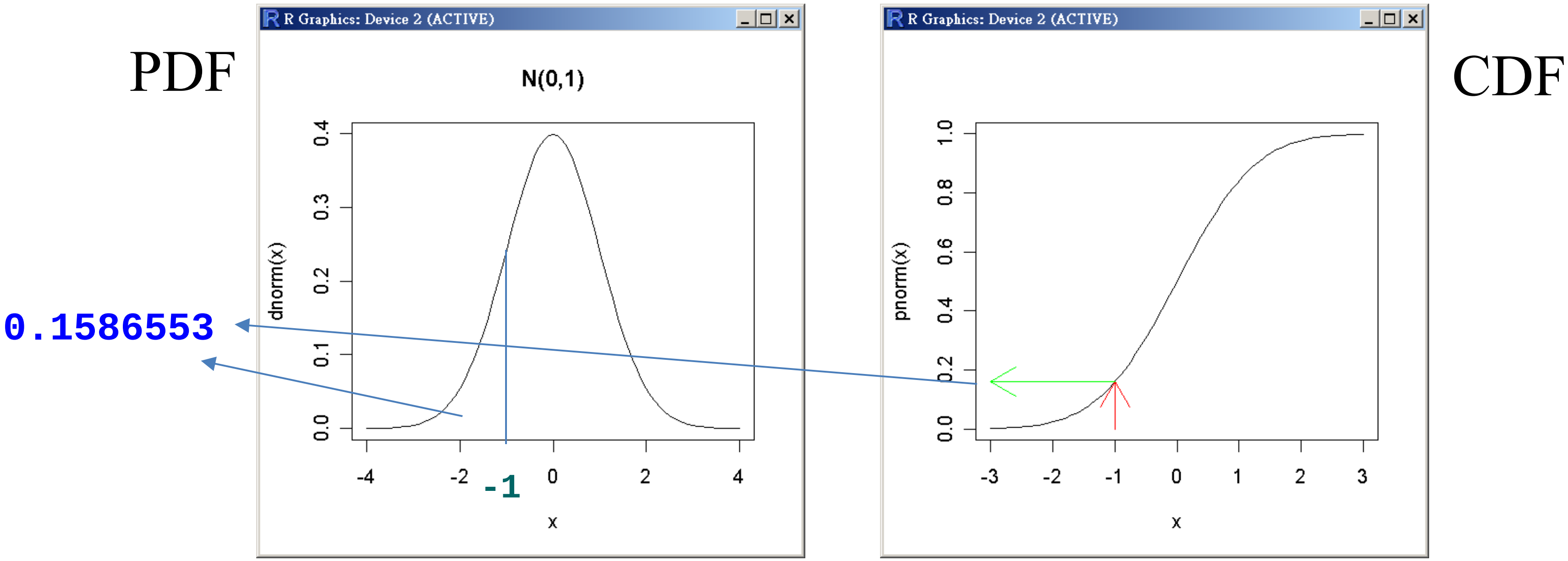
- **Normal distribution**, for a single real-valued quantity that grow linearly (e.g. errors, offsets)
- **Log-normal distribution**, for a single positive real-valued quantity that grow exponentially (e.g. prices, incomes, populations)
- **Discrete uniform distribution**, for a finite set of values (e.g. the outcome of a fair die)
- **Binomial distribution**, for the number of "positive occurrences" (e.g. successes, yes votes, etc.) given a fixed total number of independent occurrences
- **Negative binomial distribution**, for binomial-type observations but where the quantity of interest is the number of failures before a given number of successes occurs.
- **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.
- **F-distribution**, the distribution of the ratio of two scaled chi squared variables; useful e.g. for inferences that involve comparing variances or involving R-squared.

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

累積機率分配函數 CDF (p)

- It is an S-shaped curve showing for any value of x, the probability of obtaining a sample value that is less than or equal to ***x***, ***P(X<= x)***.
- The probability density is the slope of this curve (its derivative) of the cumulative probability function.

```
> 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
```

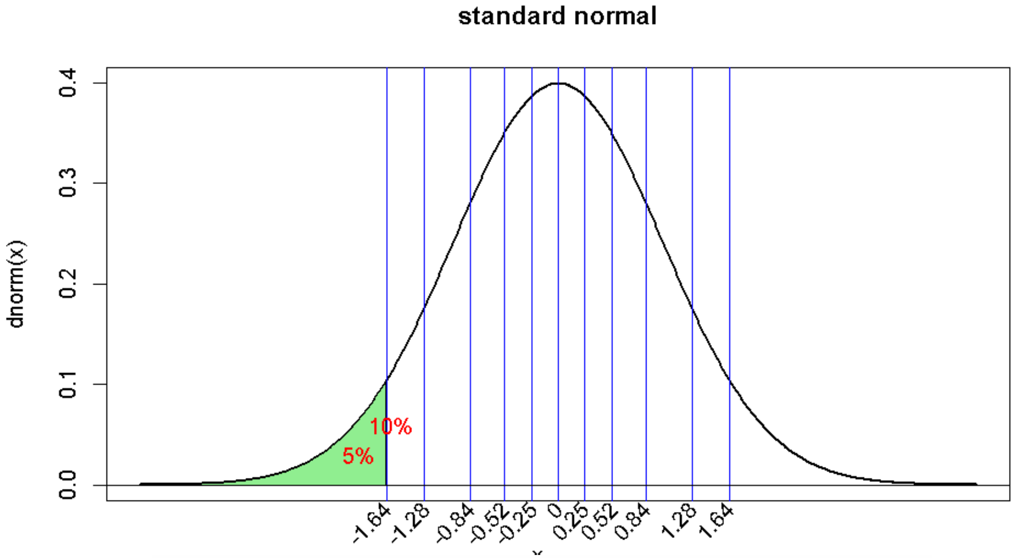


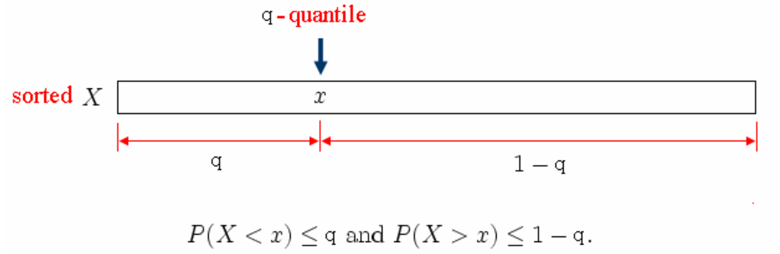
分位數 Quantiles (q)

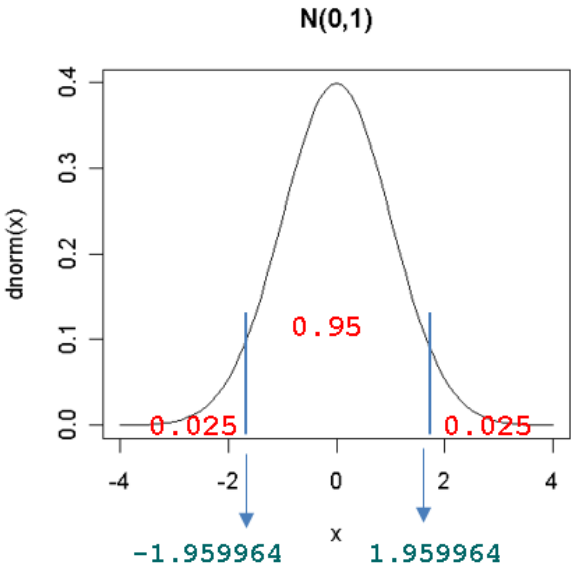
- The quantile function is the inverse of the cumulative distribution function: $F^{-1}(p) = x$.
- We say that q is the $x\%$-quantile if $x\%$ of the data values are $\leq q$.

```
> # 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
```

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





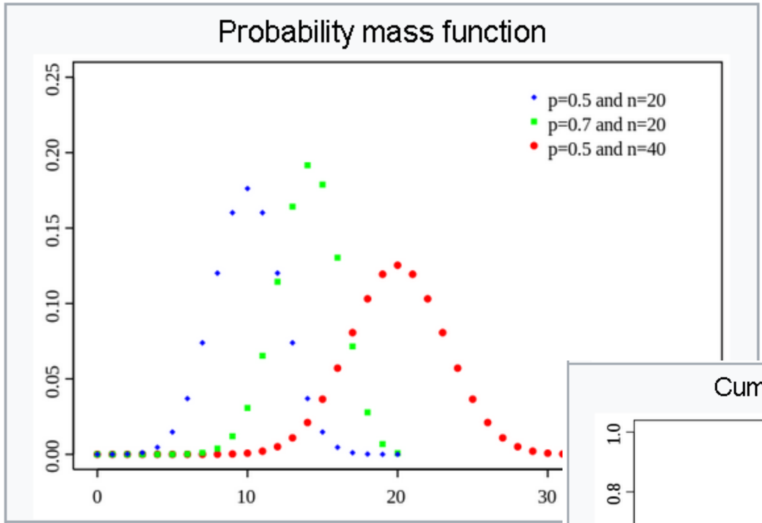


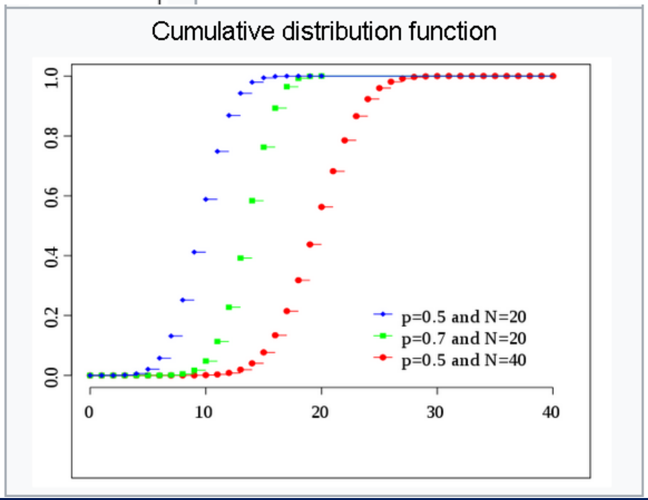
$$\bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}$$

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

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: 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。
- `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	$k \in \{0, \ldots, n\}$ — 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

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。

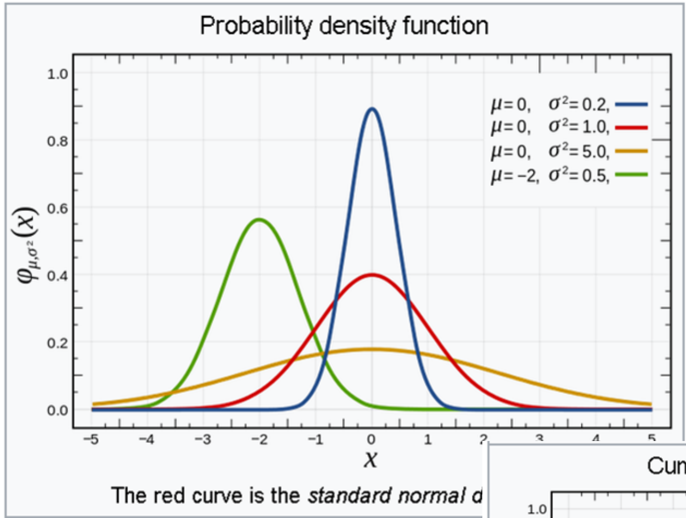
```
> pbinom(3, 10, 0.8)- pbinom(2, 10, 0.8)
[1] 0.000786432
```

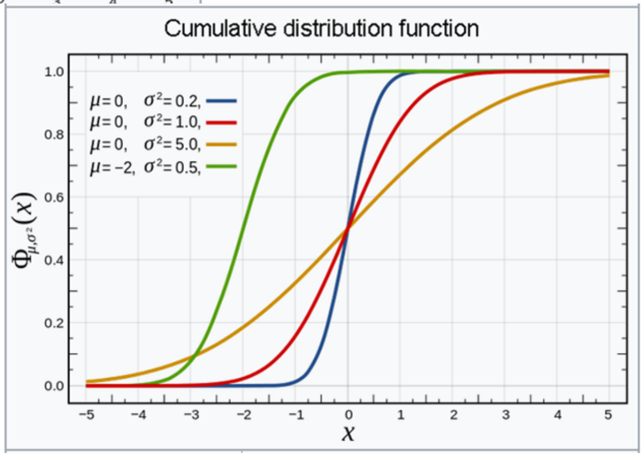
- 已知累加機率值為0.1208，求對應的分位數。

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

常態分佈

- `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





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	$\exp\{\mu t+\frac{1}{2}\sigma^2t^2\}$
CF	$\exp\{i\mu t-\frac{1}{2}\sigma^2t^2\}$
Fisher information	$\begin{pmatrix}1/\sigma^2 & 0\\ 0 & 1/(2\sigma^4)\end{pmatrix}$

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

以常態機率逼近二項式機率

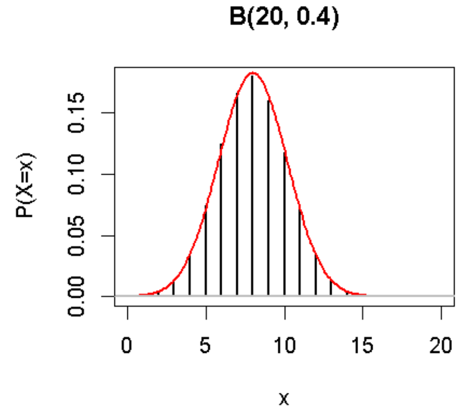
set $n = 20$ and $\pi = 0.4$ and calculate the density of the binomial,

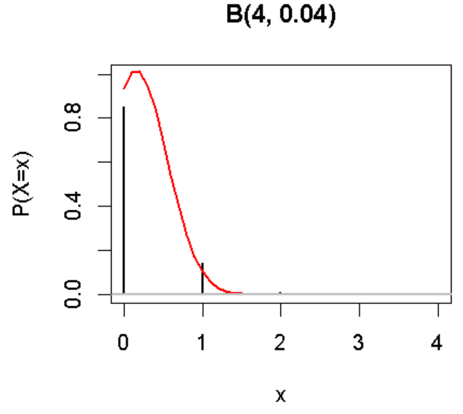
$$P\left(X = x \middle| n, \pi\right) = \binom{n}{x} \pi^x \left(1 - \pi\right)^{n-x}$$

set $\mu = n\pi$ and $\sigma = \sqrt{n\pi\left(1 - \pi\right)}$ and plot the normal density with μ and σ.

set $n = 4$ and $\pi = 0.04$

```
par(mfrow = c(1, 2))
n <- 20 # 4
p <- 0.4 # 0.04
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 = "B(20, 0.4)")
z <- seq(0, n, 0.1)
lines(z, dnorm(z, mu, sigma), col = "red", lwd = 2)
abline(h = 0, lwd = 2, col = "grey")
```





The normal approximation to the binomial Let the number of successes X be a binomial rv with parameters n and π.

Also, let $\mu = n\pi$, $\sigma = \sqrt{n\pi(1 - \pi)}$. Then if $n\pi \geq 5$, $n\left(1 - \pi\right) \geq 5$,

we consider $\phi\left(x \middle| \mu, \sigma\right)$ an acceptable approximation of the binomial.

大數法則: The Law of Large Numbers

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

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

- 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 $\bar{X}_n$ 越接近母體的均數μ 。
- 樣本平均數的這種行為稱為大數法則(law of large numbers)。

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。

- 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，　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。

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

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

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

$$\boxed{\begin{aligned} E(\bar{X}) &= \mu_{\bar{X}} = \mu \\ Var(\bar{X}) &= \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \end{aligned}}$$

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.

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

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

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

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

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

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

3.

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

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

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

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

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

$$= 0.99922$$

課堂練習

```
> z <- (126/210 -0.7)/sqrt(0.001) # 通過人數>126的機率
> z
[1] -3.162278
> 1 - pnorm(z)
[1] 0.9992173
```

寫一「通過人數大於某數的機率」之副程式

- n: 考生總數(n=210)
- X: 通過考生之人數, X~B(210, 0.7)

```
> pass.prob <- function(x, n, mu, sigma2, digit=m){
    xbar <- x/n
    z <- (xbar-mu)/sqrt(sigma2)
    zvalue <- round(z, digit)
    right.prob <- round(1-pnorm(z), digit)
    list(zvalue=zvalue, prob=right.prob)
  }

> pass.prob(126, 210, 0.7, 0.001, 4)
$zvalue
[1] -3.1623

$prob
[1] 0.9992
```

練習2: 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

```r
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

}
```

```r
> 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
```