

敘述統計 機率分佈與模擬

吳漢銘

國立政治大學 統計學系



<https://hmwu.idv.tw>

- 敘述統計
- 排列組合
- 隨機變數
- 機率分佈
 - 機率質量函數/機率密度函數(d)、累積分佈函數(p)分位數(q)、隨機數(r)
 - 常見之分佈(二項式分佈、常態分佈)、隨機抽樣
- QQplot
- 中央極限定理 (Central Limit Theorem, CLT)
- 用R程式模擬算機率

敘述統計



請舉幾個應用敘述統計於商業、工業、製造業和生物醫學的例子。



當然可以！敘述統計在多個領域中都有廣泛的應用，以下是一些應用敘述統計於商業、工業、製造業和生物醫學的例子：

商業：

1. 市場研究：利用敘述統計對消費者調查數據進行分析，了解消費者的喜好、購買行為和市場趨勢。
2. 銷售分析：分析各個產品或服務的銷售數據，以確定熱銷產品、低銷產品和季節性銷售變化。
3. 客戶滿意度調查：通過對客戶滿意度數據進行敘述統計分析，企業可以找到改進服務品質的方向。

工業：

1. 品質控制：在生產過程中，利用敘述統計分析產品的品質標準。
2. 生產效率：通過分析生產線的運行速度、產量等數據，採取措施改進。
3. 能源消耗分析：對工業生產中的能源消耗進行敘述統計分析，以優化能源利用。

製造業：

1. 供應鏈管理：通過分析供應商的交貨時間、產品質量等數據，可對供應鏈進行優化。
2. 存貨管理：利用敘述統計分析庫存水平，以確保適時補充並減少過剩庫存造成的損失。
3. 員工生產力：通過對員工生產力數據進行敘述統計分析，找出影響生產力的因素，並制定相應的培訓和激勵措施。

生物醫學：

1. 臨床試驗：在藥物研究和開發過程中，利用敘述統計分析試驗結果，以確定藥物的療效和安全性。
2. 流行病學研究：對疾病發病率、死亡

資料分析的基本流程



到底要用哪一種處理/分析方法?

統計方法? 大數據? 機器學習? 深度學習? AI?

Source: 資料科學的統計實務, 溫政堯譯, 旗標出版社, 2021/11/01。



今日新聞
NOWNEWS

60.9k 人追蹤

☆ 追蹤

善用房產科技！靠數位工具突破限制

生活消費中心/綜合報導

2022年11月22日

房地產科技PropTech (Property Technology)，結合房地產銷售結合AI人工智慧、大數據、機器學習等數位科技，藉此突破時間和空間限制，提升銷售效率，包括增加消費者和物件的媒合效率，國外已有許多房地產業者逐漸向科技靠攏，而國內房仲龍頭信義房屋近年也致力於PropTech發展，以科技賦能輔助，希望為買賣雙方建構出更多元、貼心的服務。

<http://shorturl.at/kzOU7>

<http://shorturl.at/oGY04>

3.1k 人追蹤

☆ 追蹤

yahoo! 新聞

大數據分析交通違規事件 違停居首

【記者張淑珠 / 台中報導】

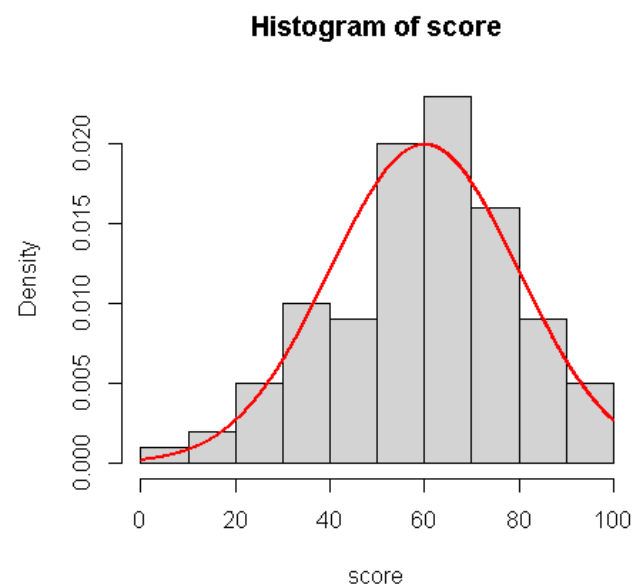
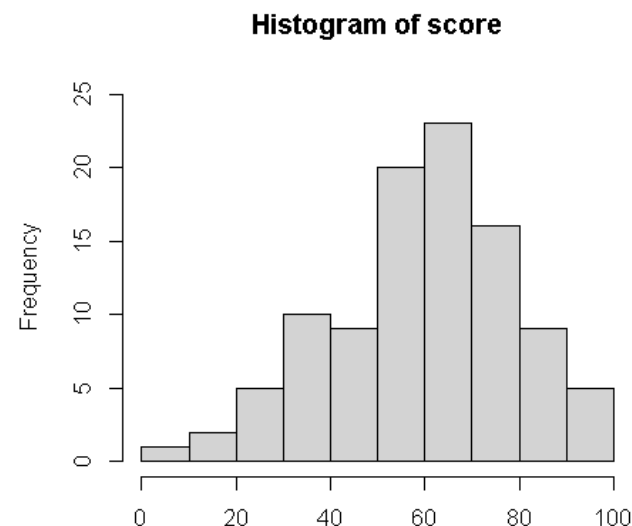
2023年2月7日 週二 下午8:47

在重大違規部分，黃士哲說，依據無照駕駛違規態樣分析，實際未領有駕照有59%，其中未滿18歲占14%、滿18歲占45%；進一步分析，未領有駕照、且未滿18歲所駕駛的車種有95%是機車。對此，交通事件裁決處自110年6月起，針對未滿18歲違規駕駛案，每月挑檔提供清冊，交由警察局進行後續個案關懷協助，警察局另將學籍在台中市的無照駕駛列管案件，每月通報教育局轉各校，並啟動校園關懷輔導。

此外，針對酒後駕車，經大數據分析顯示，男性為主要的酒駕違規者，且在各年齡層的每十萬人違規人數中，以成年男性 (25-64歲) 酒駕情形最為嚴重，年輕男性 (18-24歲) 次之。交通事件裁決處表示，將強力配合法務部行政執行署執行滯欠酒

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。





統計、資料探勘、機器學習、資料科學

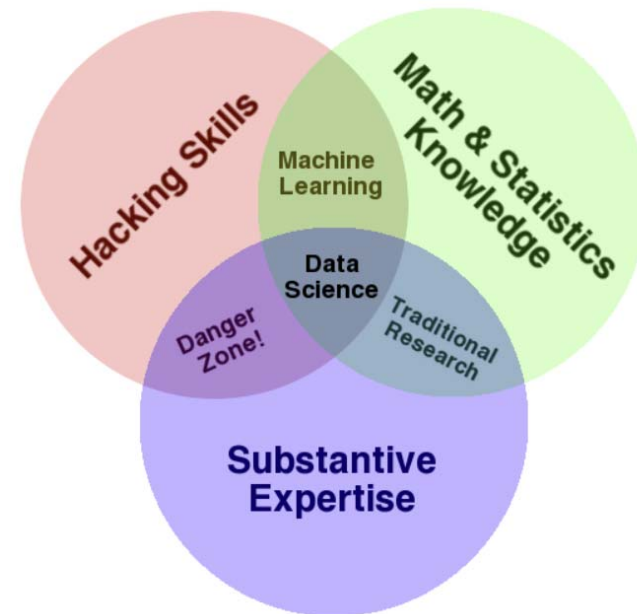
6/48

- **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

The Data Science Venn Diagram

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



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>



資料的型態 (Types of Data Scales)

7/48

- **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 40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**)。

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)

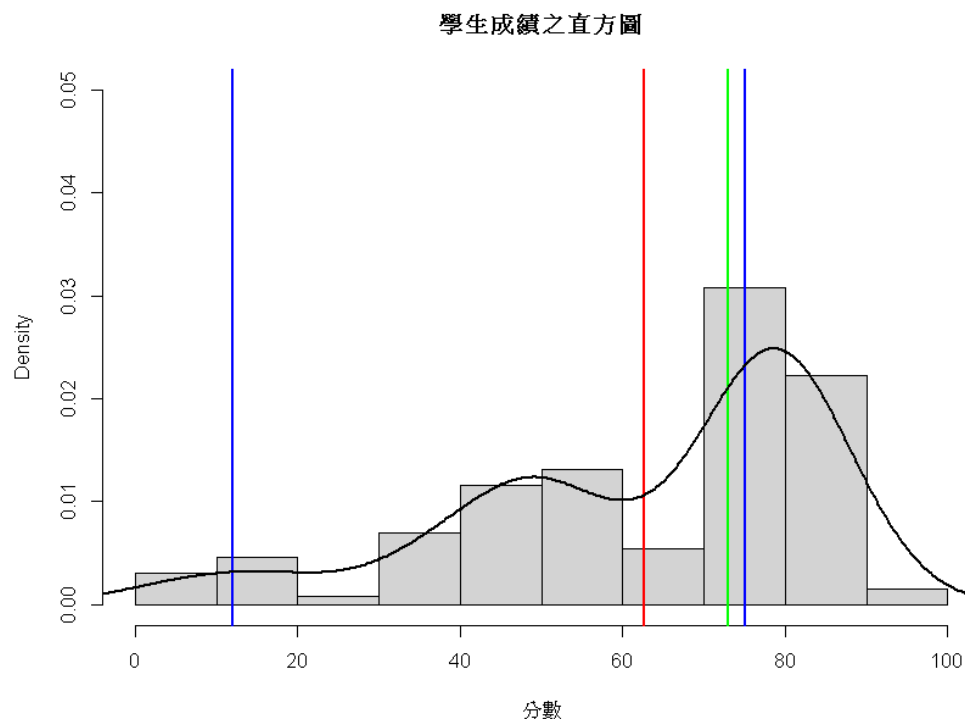
標準差(standard deviation)

變異數(variance)

Median absolute deviation

$$\tilde{X} = \text{median}(X)$$

$$\text{MAD} = \text{median}(|X_i - \tilde{X}|)$$



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

n = The number of data points

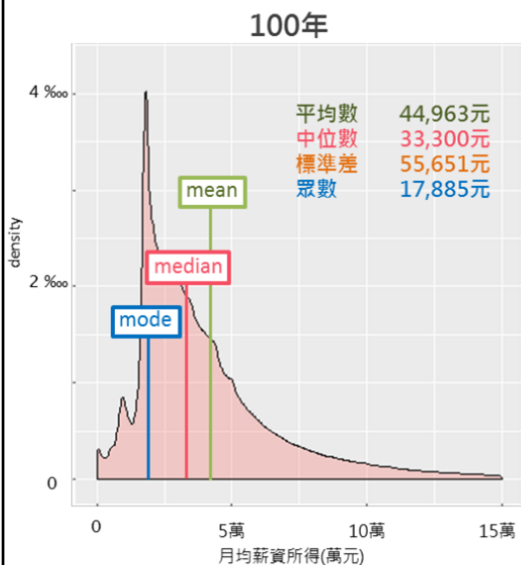
\bar{x} = The mean of the x_i

x_i = Each of the values of the data

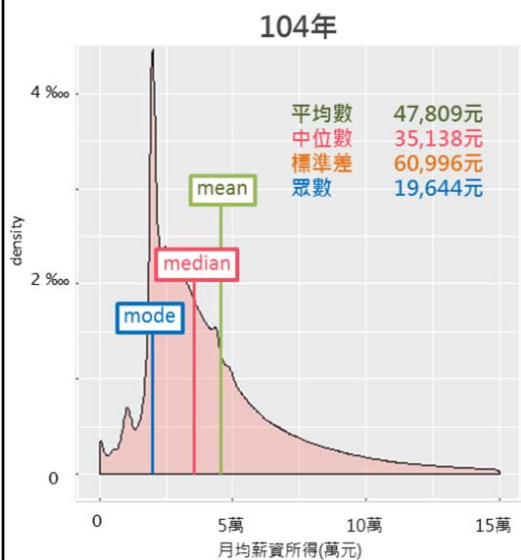
- 偏態係數
- 峰態係數

範例: 由財稅大數據探討臺灣近年薪資樣貌

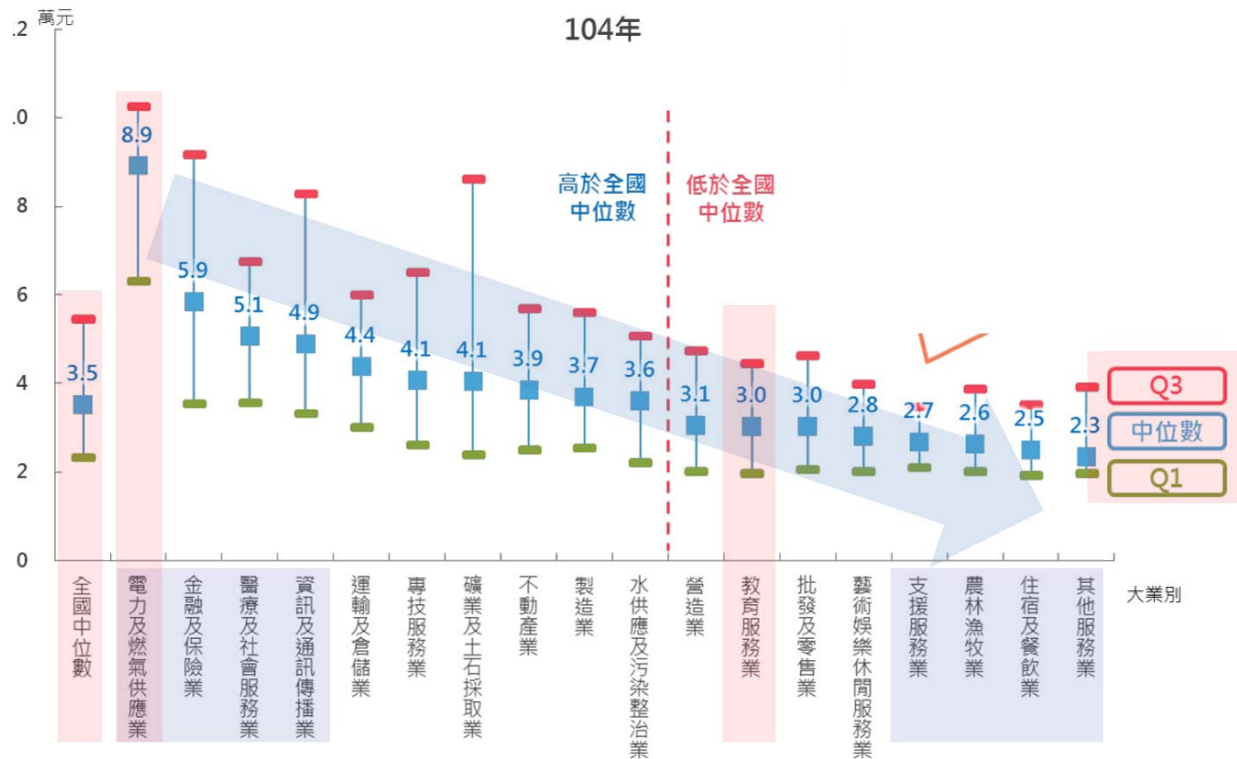
月均薪資所得機率分布圖



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



月均薪資所得中位數 - 按大業別分



玩玩看~薪情平臺



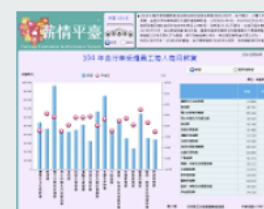
薪情互動



製造業四大產業
薪資趨勢



男女薪資差異



各業薪資概況

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

我的薪資合理嗎？行政院「薪情平台」，一看就知道自己的薪水等級

<https://buzzorange.com/techorange/2019/04/10/salary-distribution/>

<https://hmwu.idv.tw>



要如何介紹機率分佈(probability distribution)給一般人或數學能力較不好的人了解?



要對一般人教授統計學和資料科學的知識，可以不講述機率分佈嗎?



要向
解釋
機率
訴我
機事
以摸
子時
同，



對於
些更
等。
然而
許多
更全
在教
例如
瞭解



一般來說，機率分佈(probability distribution)應用在哪些場域?



機率分佈在許多場域中都有廣泛的應用，以下是一些常見的應用場域：

1. 統計學與資料科學：在統計學和資料科學中，機率分佈被用來建立概率模型、估計參數、進行假設檢驗等。例如，使用正態分佈進行回歸分析、用泊松分佈分析事件的發生率等。
2. 金融與風險管理：在金融領域，機率分佈被用來分析和預測金融市場的波動、評估投資組合的風險，以及建立風險模型等。
3. 工程與品質控制：在工程領域，機率分佈可以用來分析設備的失效概率、預測維修需求，以及進行可靠性分析等。
4. 電信與網絡流量分析：在電信和網絡領域，機率分佈用於分析和預測通信網絡的流量，並優化網絡資源分配。
5. 生物統計學與醫學研究：在生物統計學和醫學研究中，機率分佈用於分析基因變異、進行臨床試驗分析，以及研究疾病的流行病學等。
6. 機器學習與人工智能：在機器學習和人工智能領域，機率分佈被用來建立概率模型，如隱馬爾可夫模型、貝葉斯網絡等，並用於對未知數據進行推斷和預測。

7.

- 瞭解和掌握機率分佈有助於在各種情況下更好地分析數據、預測未來事件和做出決策。在現代社會中，機率分佈在幾乎所有學科和領域中都發揮著重要作用，因此瞭解機率分佈對於解決現實問題具有很高的價值。
- 8.

主於一
析方法
。
分佈在
囊學生
概念。
主深入

```
> x <- letters[1:3]
> library(combinat)
> permn(x)
[[1]]
[1] "a" "b" "c"

[[2]]
[1] "a" "c" "b"

[[3]]
[1] "c" "a" "b"

[[4]]
[1] "c" "b" "a"

[[5]]
[1] "b" "c" "a"

[[6]]
[1] "b" "a" "c"
> # the number of
distinct arrangements
of 3 out of 5
> nCm(5, 3)
[1] 10
```

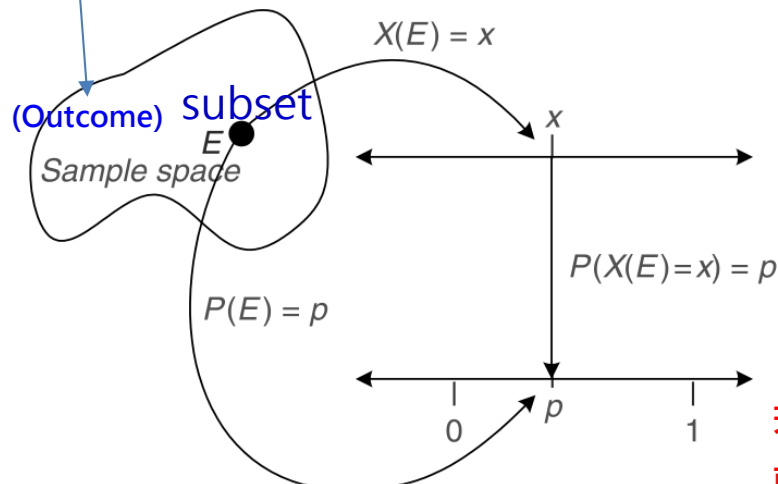
- `combn{utils}`: Generate All Combinations of `n` Elements, Taken `m` at a Time.
- Usage: `combn(x, m, FUN = NULL, simplify = TRUE, ...)`
- Arguments:
 - `x`: vector source for combinations, or integer `n` for `x <- seq_len(n)`.
 - `m`: number of elements to choose.
 - `FUN`: function to be applied to each combination; default NULL means the identity, i.e., to return the combination (vector of length `m`).

```
> combn(5, 3)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    2    2    2    3
[2,]    2    2    2    3    3    4    3    3    4    4
[3,]    3    4    5    4    5    5    4    5    5    5
> combn(5, 3, min)
[1] 1 1 1 1 1 1 2 2 2 3
> choose(5, 3)
[1] 10
```

- 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 $S = \{\text{sick}, \text{healthy}, \text{dead}\}$. The following are all events: $\{\text{sick}\}$, $\{\text{healthy}\}$, $\{\text{dead}\}$, $\{\text{sick}, \text{healthy}\}$, $\{\text{sick}, \text{dead}\}$, $\{\text{healthy}, \text{dead}\}$, $\{\text{sick}, \text{healthy}, \text{dead}\}$, $\{\text{none of the above}\}$.
- **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」。

- **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.

A random experiment



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: $P(X=1) = 0.5$, $P(X=0) = 0.5$.

若將隨機變數每種情況的機率值都列出來, 就是其機率分佈。

編註: 所謂機率分佈, 就是每個可能出現的值(例如: 1、2、3、4、5、6), 各有多少機率(例如: $\frac{1}{6}$ 、 $\frac{1}{6}$ 、 $\frac{1}{6}$ 、 $\frac{1}{6}$ 、 $\frac{1}{6}$ 、 $\frac{1}{6}$), 把它們全部列出來, 如上表所示。



統計分配 (Statistical Distributions)

15/48

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, $F(x) = P(X \leq x)$:
`pnorm()`

- 分位數 (q)

- the quantiles of the distribution: `qnorm()`

Often, we are interested in the inverse of a distribution. That is, given a probability value p , we wish to find the quantile, x , such that $P(X \leq x | \theta) = p$.

- 隨機數 (r)

- the random numbers generated from the distribution:
`rnorm()`

以常態分佈normal為例:

機率密度(分配)函數: **dnorm()**

累積機率(分配)函數: **pnorm()**

分位數: **qnorm()**

隨機數: **rnorm()**

Distribution	R name	additional arguments
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df2, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's	t	df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Probability Mass Function, PMF

Formal definition

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

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

$X_1 \sim \text{DiscreteUniform}(1, 6)$.

$X_2 \sim \text{DiscreteUniform}(1, 6)$.

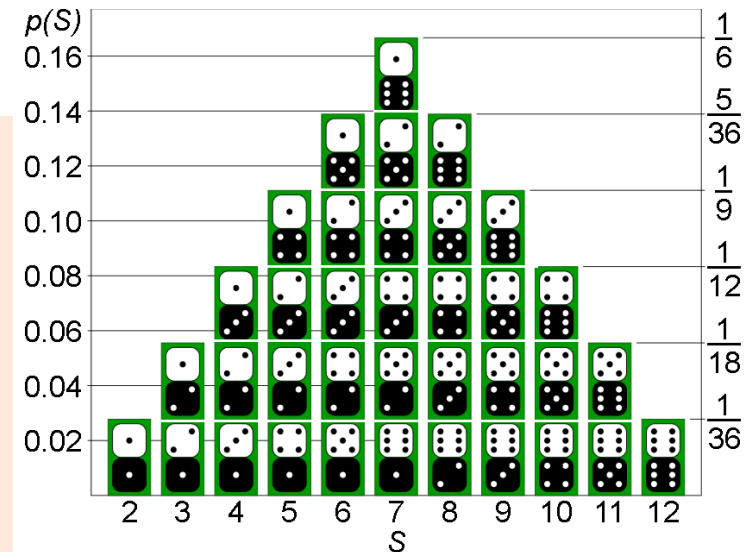
$f_{X_1}(k) = f_{X_2}(k) = P(X_1 = k) = P(X_2 = k) = 1/6,$
 $k = 1, \dots, 6.$

$S = X_1 + X_2$

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

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

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



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

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

Probability Density Function, PDF

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:

$$P(X \in A) = \int_A f(x)dx$$

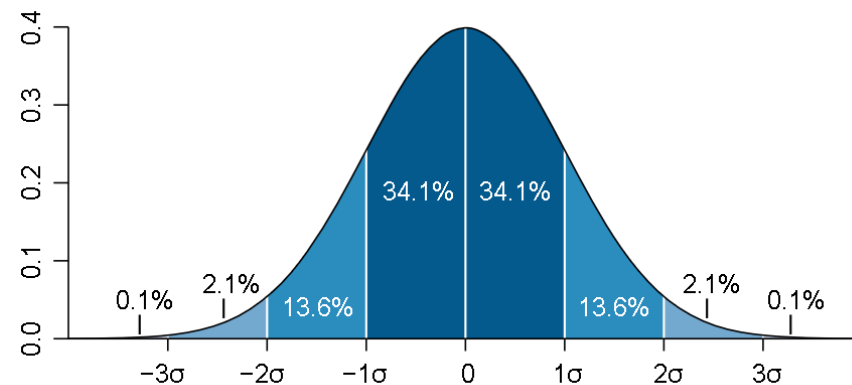
$$P[a \leq X \leq b] = \int_a^b f(x) dx$$

The **probability density** of the normal distribution is:

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

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**

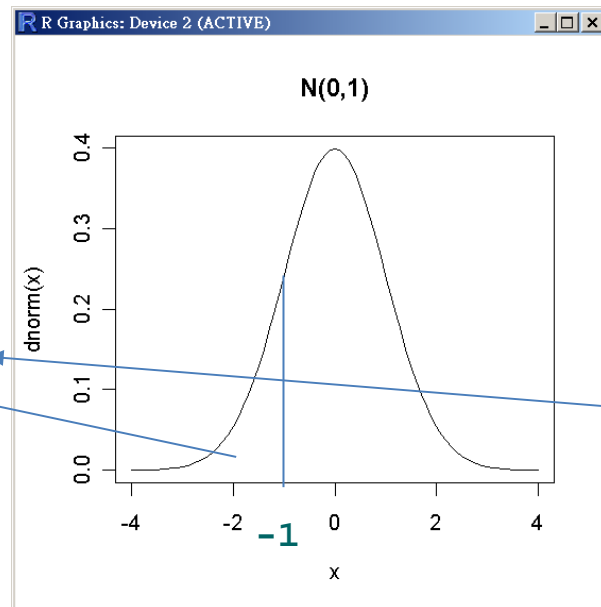


Cumulative Probability Function, CDF

- 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 , $P(X \leq x)$.
- The probability density is the slope of this curve (its derivative) of the cumulative probability function.

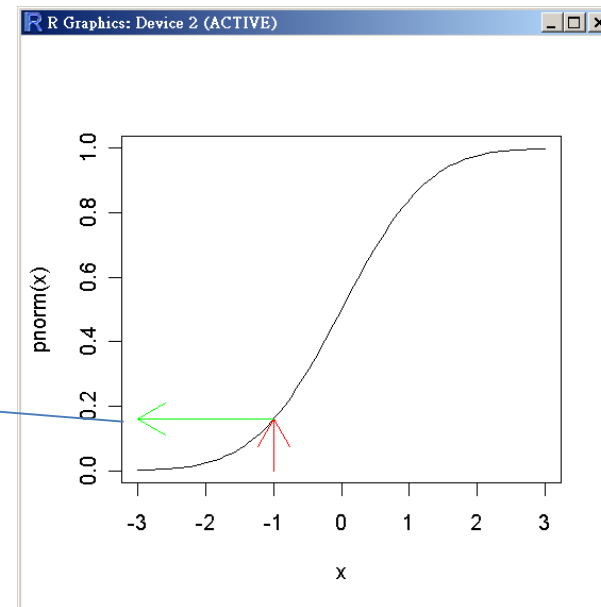
```
> curve(pnorm(x), -3, 3)
> arrows(-1, 0, -1, pnorm(-1), col="red")
> arrows(-1, pnorm(-1), -3, pnorm(-1), col="green")
> pnorm(-1)
[1] 0.1586553
```

PDF



0.1586553

CDF



CRAN Task View: Probability Distribution

20/48

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

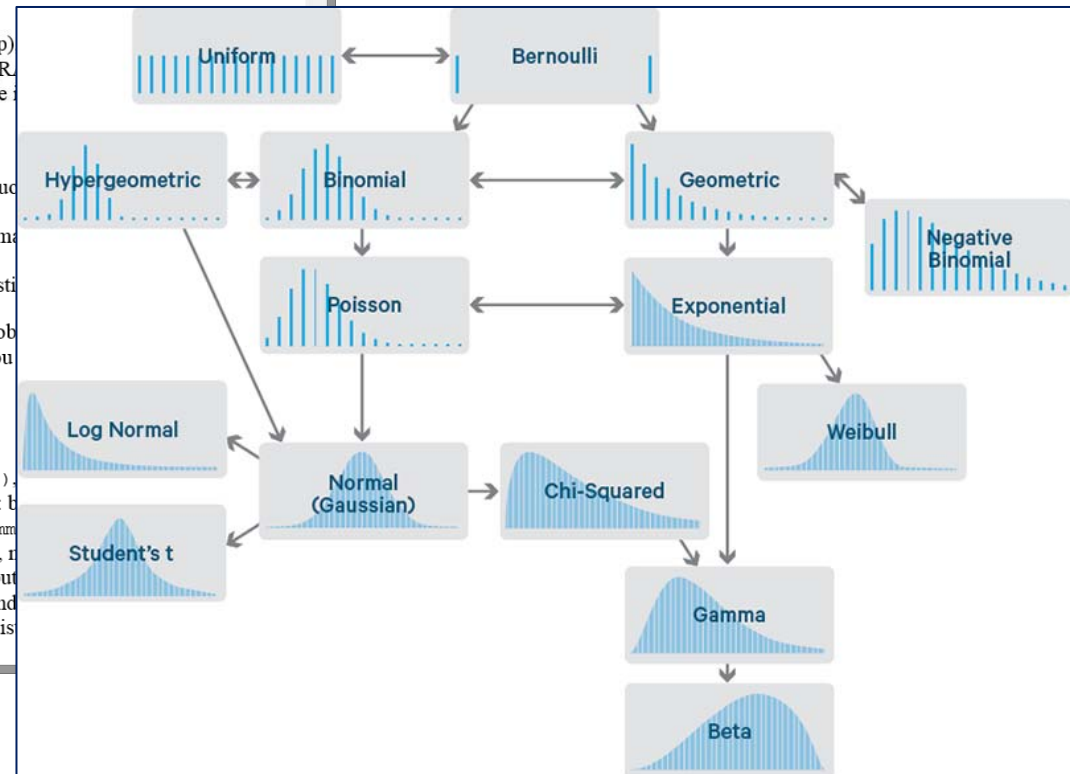
Ultimate bibles on probability distributions are:

- 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 Statist

The maintainer greatly 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 fonctionnalité:

- Base R provides probability distribution functions `pfoo()` density functions `dfoo()`, and random number generation `rfoo()` where `foo` indicates the type of distribution: `binom`, `Cauchy` `cauchy`, `chi-squared` `chisq`, `exponential` `exp`, `Fisher` `F` `f`, `gamma` `gamma`, `hypergeometric` `hyper`, `logistic` `logis`, `lognormal` `lnorm`, `negative binomial` `nbinom`, `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>

<https://hmwu.idv.tw>

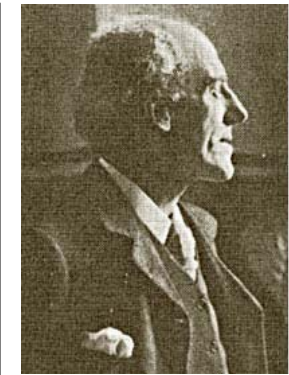
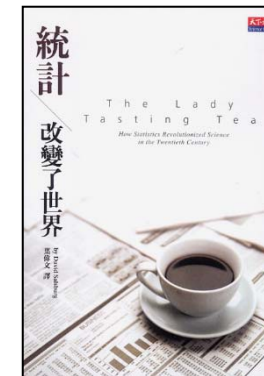
機率分佈在統計學中的重要性

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。

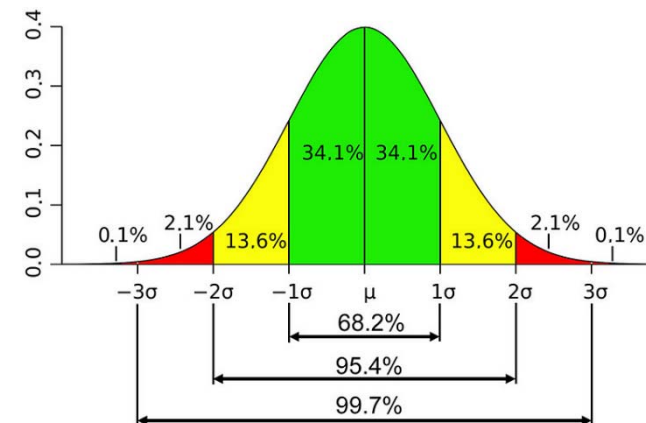


樣本變異數與樣本標準差

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

母體變異數與母體標準差

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



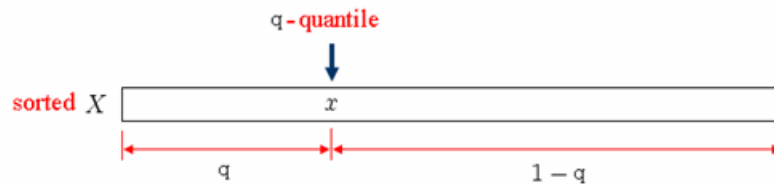
Schweizer, B. (1984), **Distributions Are the Numbers of the Future**, in Proceedings of The Mathematics of Fuzzy Systems Meeting, eds. A. di Nola and A. Ventre, Naples, Italy: University of Naples, 137–149. (The present is that future.)



- **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

- 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 $\leq q$.



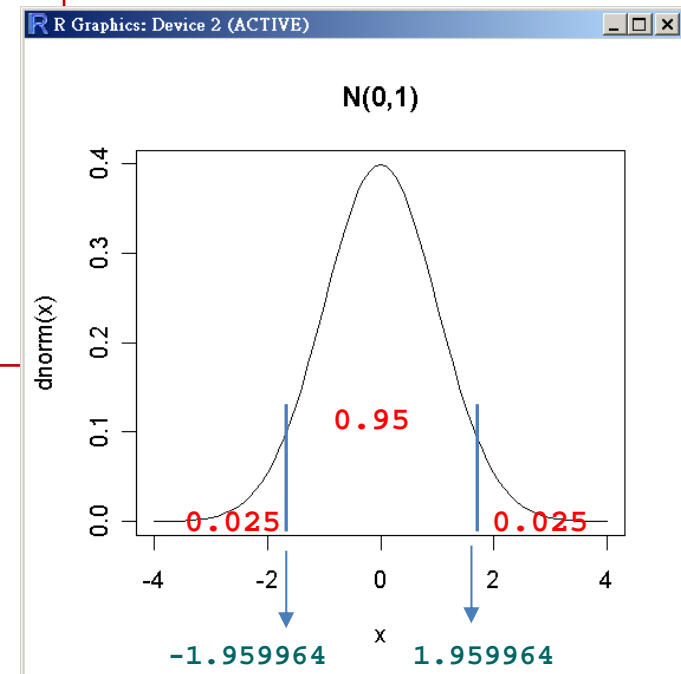
$$P(X < x) \leq q \text{ and } P(X > x) \leq 1 - q.$$

```
> # 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
```

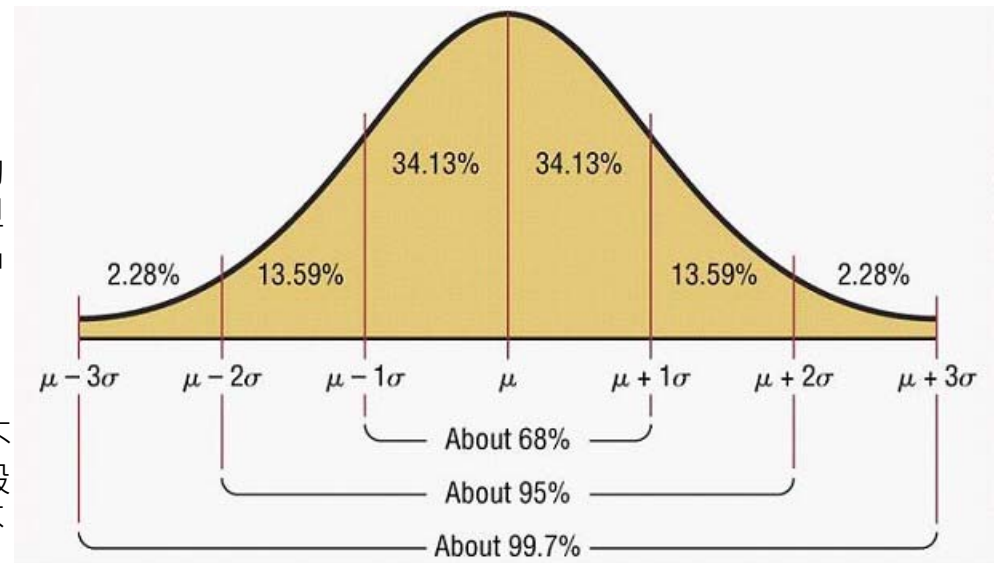
$$\Phi^{-1}(0.975)$$

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

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



- 標準差是用來描述母體的「變異性」(variation)或「不一致程度」(inconsistency)。
- 過去習慣用「平均數」作為其績效表現，例如平均成本、平均產能、平均交貨時間、平均工資等，但卻忽略了變異程度。現今用「標準差」來衡量產品之品質分佈的變異狀況。
- 六標準差管理是1986年Motorola發展出來的管理方法。符合 6σ 就代表每生產出一百萬個產品，其不良品必須低於3.4個(良率99.99966%)。轉換成一般服務業的用語，就是每一百萬人次的客人，其中不滿意服務的不可以超過四人。



Source: <https://www.usastock88.com/2013/04/BBand.html>

- 在美國而言(Harry 「1978)，平均99%的品質水準相當於 (假設資料呈現常態分佈)：
 - (1)每小時有2萬件郵件遺失，這相當於1%的誤投率。
 - (2)每天供應的自來水有15分鐘是不適合飲用的。
 - (3)每星期有5千例外科誤診。
 - (4)每個月有7小時停電。
- 對良率的要求需更嚴格的場景:
 - (1) 某航空公司宣稱其飛安率可高達99.73%。(每起落一百萬個架次，約有2,700次失事)
 - (2) 某醫院聲明其婦產科接生新生嬰兒的平均成功率是99.9%。
- 六標準差是一個利用統計、問題排除和問題預防等工具，將客戶滿意度提升至99.999%完美層級的品質改善商業策略。

中華六標準差管理學會
<https://www.6sigmai.org/>

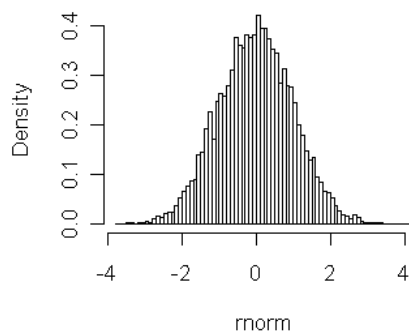
隨機數 Random Numbers (r)

25/48

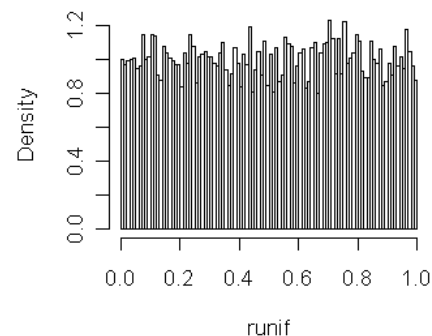
- Let X_i is a vector of measurements for the i -th object in the sample.
- (X_1, X_2, \dots, X_n) is said to be a random sample of size n from the common distribution if X_1, X_2, \dots, X_n as independent copies of an underlying measurement vector. (an n -tuple of identically-distributed independent random variables).

```
> par(mfrow=c(2,2))
> hist.sym <- hist(rnorm(10000),nclas=100,freq=FALSE,
+ main="Symmetric Distribution", xlab="rnorm")
> hist.flat <- hist(runif(10000),nclas=100,freq=FALSE,
+ main="Symmetric Flat Distribution", xlab="runif")
> hist.skr <- hist(rgamma(10000,shape=2,scale=1),freq=FALSE, nclas=100,
+ main="Skewed to Right", xlab="rgamma")
> hist.skl <- hist(rbeta(10000,8,2),nclas=100,freq=FALSE,
+ main="Skewed to Left", xlab="rbeta")
```

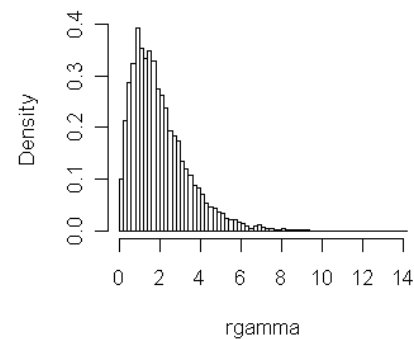
Symmetric Distribution



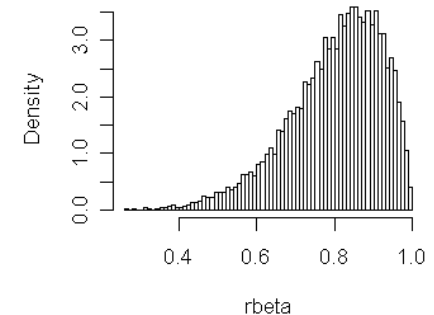
Symmetric Flat Distribution



Skewed to Right



Skewed to Left



- The concepts of randomness and probability are central to statistics.

```
> sample(x, size, replace = FALSE, prob = NULL)
```

- sampling without replacement

```
> sample(1:40, 5)
```

```
[1] 12 38 2 3 7
```

- sampling with replacement

```
> sample(1:40, 5, replace=TRUE)
```

```
[1] 35 4 4 16 22
```

- Simulate 10 coin tosses (fair coin-tossing)

```
> sample(c("H", "T"), 10, replace=T)
```

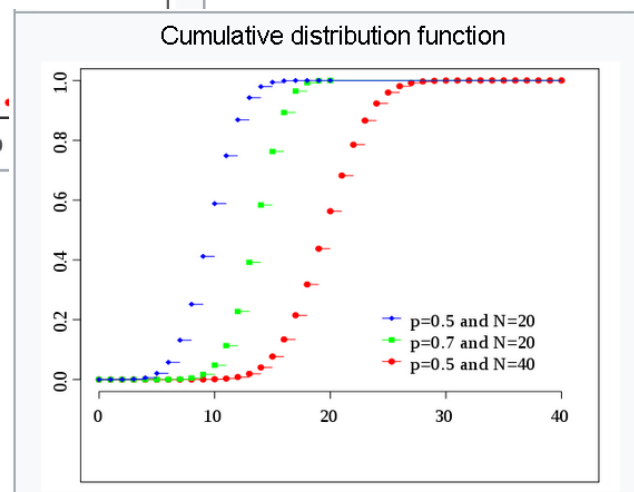
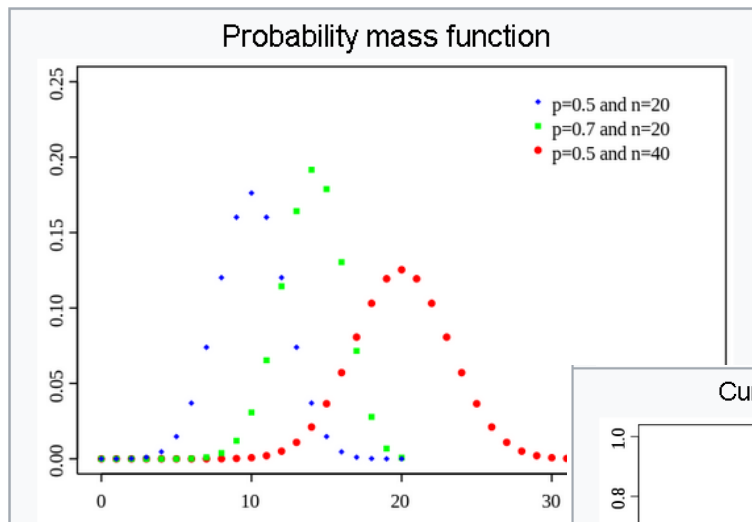
```
[1] "T" "T" "T" "H" "H" "H" "T" "H" "T" "H"
```

```
> sample(c("succ", "fail"), 10, replace=T, prob=c(0.9, 0.1))
```

```
[1] "succ" "succ" "succ" "fail" "fail" "fail" "succ" "succ" "succ" "succ"
```

二項式分佈 (Binomial)

- $X \sim B(n, p)$ 表示 n 次伯努利試驗中 (size) · 成功結果出現的次數。
- 例: 擲一枚公正銅板十次, 那麼擲得正面次數就服從 $n = 10$ 、 $p = 1/2$ 的二項分佈。
- `dbinom(x, size, prob)` # 機率公式值 $P(X=x)$
- `pbinom(q, size, prob)` # 累加至 q 的機率值 $P(X \leq q)$
- `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	$\frac{1}{2} \log_2 (2\pi e np(1-p)) + O\left(\frac{1}{n}\right)$ 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

$$X \sim B(10, 0.8)$$

- 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 $P(X \leq 3) - P(X \leq 2)$ ，並和 $P(X=3)$ 相比較。

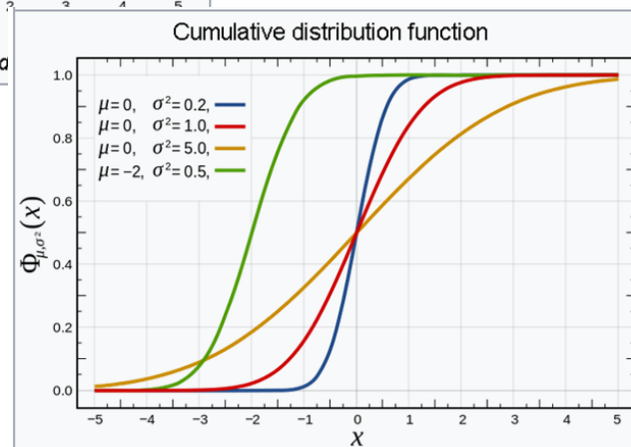
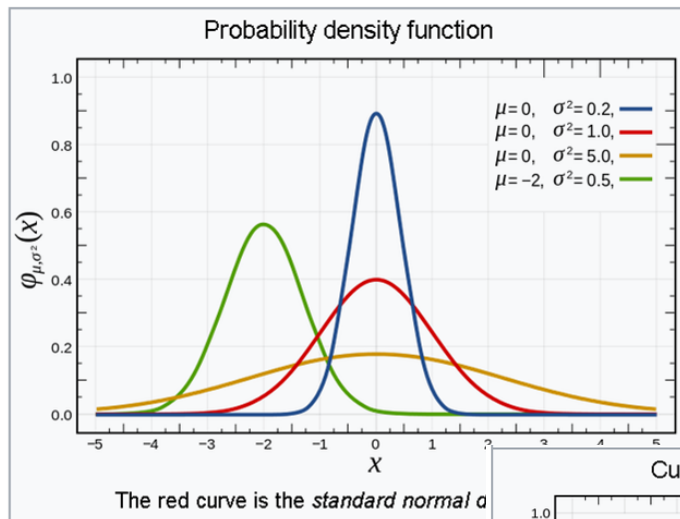
```
> pbinom(3, 10, 0.8) - pbinom(2, 10, 0.8)  
[1] 0.000786432
```

- 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 (Normal Distribution)

- `dnorm(x, mean, sd)` # 機率密度函數值 $f(x)$
- `pnorm(q, mean, sd)` # 累加機率值 $P(X \leq x)$
- `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	$\exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}$
CF	$\exp\left\{i\mu t - \frac{1}{2}\sigma^2 t^2\right\}$
Fisher information	$\begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$

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

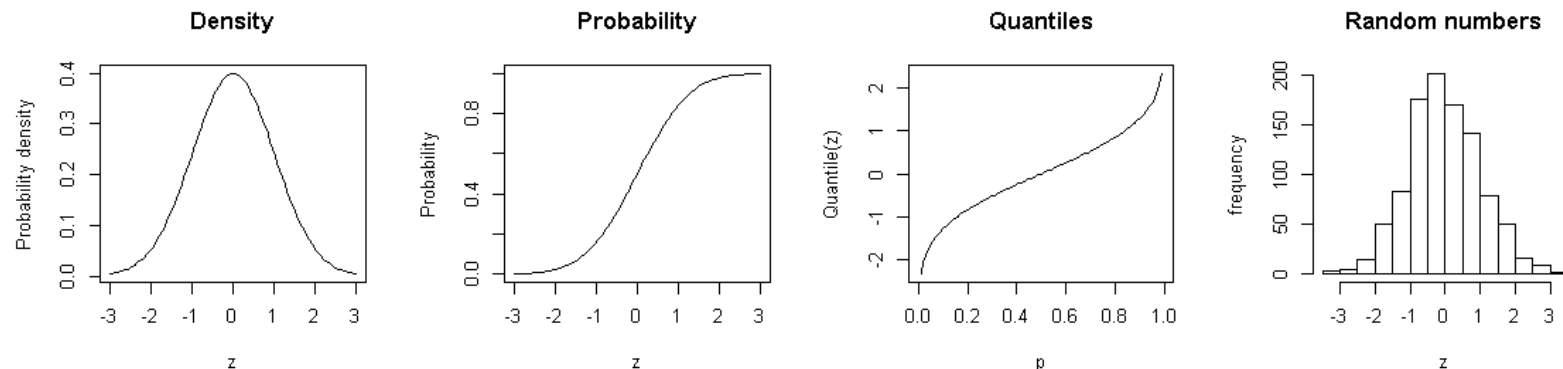
常態分佈 (Normal Distribution)

$Z \sim N(0, 1)$

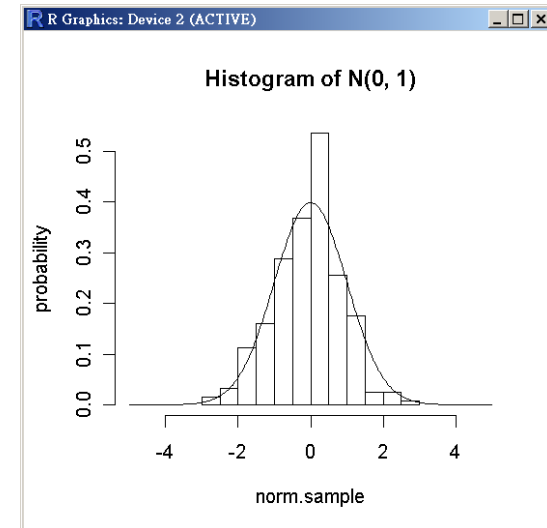
```
> dnorm(0)
[1] 0.3989423
> pnorm(-1)
[1] 0.1586553
> qnorm(0.975)
[1] 1.959964
```

```
> dnorm(10, 10, 2) #  $X \sim N(10, 4)$ 
[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) #  $P(8 \leq X \leq 15)$ 
[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")
```

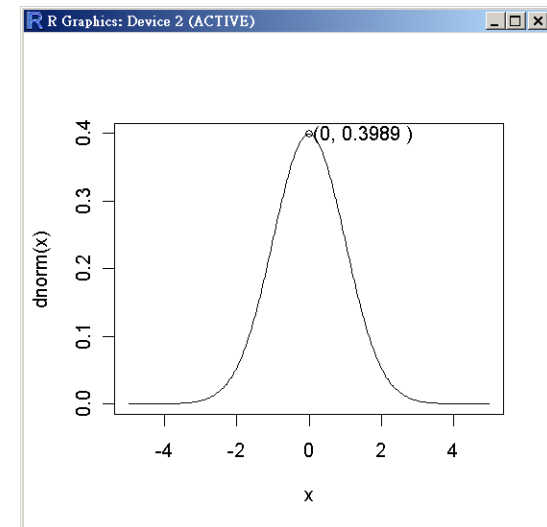


```
> norm.sample <- rnorm(250)
> summary(norm.sample)
> hist(norm.sample, xlim=c(-5, 5), ylab="probability",
+ main="Histogram of N(0, 1)", prob=T)
> x <- seq(from=-5, to=5, length=300)
> lines(x, dnorm(x))
```



標出最頂點的座標

```
> x <- seq(from=-5, to=5, length=300)
> plot(x, dnorm(x), type="l")
> points(0, dnorm(0))
> height <- round(dnorm(0), 4); height
> text(1.5, height, paste("(0,", height, ")"))
```





Quantile-Quantile Plots : Testing for Normality Graphically

32/48

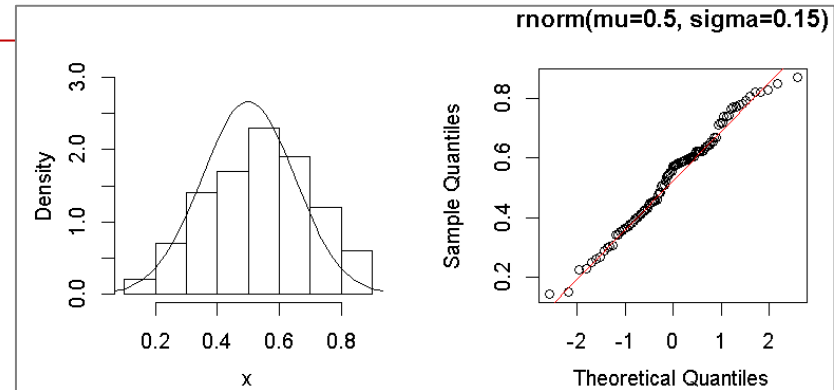
- The quantile-quantile (Q-Q) plot is used to determine if two data sets come from populations with a common density.
- Q-Q plots are sometimes called **probability plots**, especially when data are examined against a theoretical density.
- **qqnorm()**: produces a normal QQ plot of the values in sample
- **qqline()**: adds a line which passes through **the first and third quartiles**.
 - Use the diagonal line would not make sense because the first axis is scaled in terms of the theoretical quantiles of a $N(0,1)$ distribution.
 - Using the first and third quartiles to set the line gives a robust approach for estimating the parameters of the normal distribution, when compared with using the empirical mean and variance, say.
 - Departures from the line (except in the tails) are indicative of a lack of normality.
- **qqplot()**: qqplot produces a QQ plot of two datasets.

qqnorm, qqline, qqplot

```

> par(mfrow = c(1, 2))
> set.seed(12345);
> n <- 100; mu <- 0.5; sigma <- 0.15
> x <- rnorm(n, mu, sigma)
> hist(x, freq=FALSE, ylim=c(0, 3), main="")
> y <- seq(0, 1, length = n)
> lines(y, dnorm(y, mu, sigma), type = 'l')
> qqnorm(x, main = "rnorm(mu=0.5, sigma=0.15)");
> qqline(x)

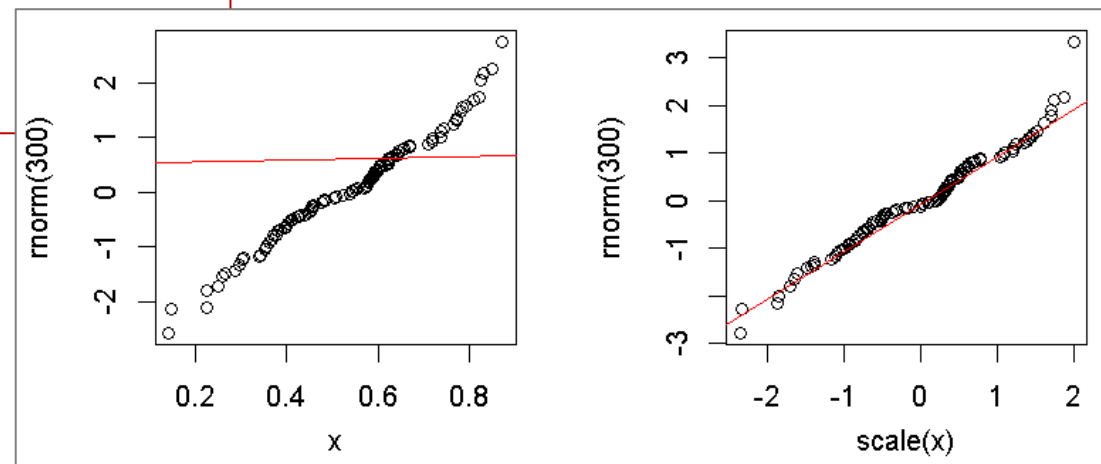
```



```

> qqplot(x, rnorm(300))
> qqline(x, col = 2)
> qqplot(scale(x), rnorm(300))
> qqline(scale(x), col = 2)

```

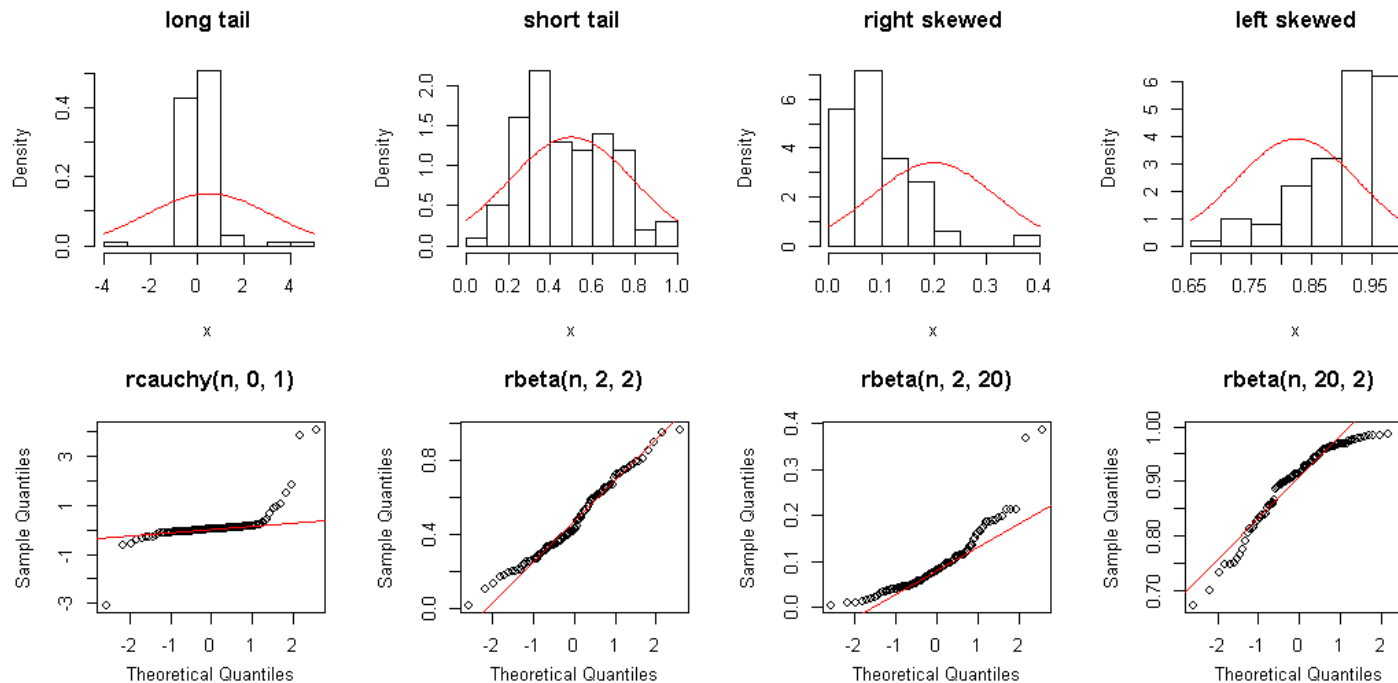


```

par(mfcol=c(2,4))
n <- 100
x <- rcauchy(n, 0, 0.1) # a long tail density
hist(x, freq=FALSE, main="long tail")
curve(dnorm(x, mean=mean(x), sd=sd(x)), add=TRUE, col="red")
qqnorm(x, main = "rcauchy(n, 0, 1)");
qqline(x, col="red")
# different shapes of distributions
x <- rbeta(n, 2, 2) # a short tail density
x <- rbeta(n, 2, 20) # a right skewed density
x <- rbeta(n, 20, 2) # a left skewed density
    
```

Try

- Normal
- Symmetric, Non-Normal, Short-Tailed
- Symmetric, Non-Normal, Long-Tailed
- Symmetric and Bimodal
- Bimodal Mixture of 2 Normals
- Skewed (Non-Symmetric) Right
- Skewed (Non-Symmetric) Left
- Symmetric with Outlier





實作 Quantile-Quantile Plots

(X_1, X_2, \dots, X_n)

1. 計算樣本平均數及樣本變異數。

$$\bar{X} = \frac{\sum X_i}{n}$$

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

2. 將隨機樣本標準化並排序。

$$d_{(1)}, d_{(2)}, \dots, d_{(n)}$$

$$d_{(i)} = \frac{X_i - \bar{X}}{S}$$

3. 查出 n 個標準常態值: (將標準態分配, 區分成 $n+1$ 區塊, 最左及最右區塊的機率分別為 $1/2n$, 中間的 $n-1$ 區塊, 機率分別為 $1/n$)。

$$q_{(1)} = z_{\frac{1}{2n}}, q_{(2)} = z_{\frac{3}{2n}}, \dots, q_{(n)} = z_{\frac{2n-1}{2n}}$$

$$q_{(i)} = z_{\frac{2i-1}{2n}}$$

$$P(Z < q_{(i)}) = \frac{2i-1}{2n}$$

4. 畫散佈圖: x軸: 排序的標準化樣本, y軸: 標準常態值。

$$(d_{(i)}, q_{(i)})$$

5. 加入一條由 $(q_{(i)}, q_{(i)})$ 產生的標準常態直線。
(或加入一條通過第25%及第75% quantiles 的直線)

$$(q_{(i)}, q_{(i)})$$

```
> qqnorm(iris[,1])
> qqline(iris[,1])

> qqnorm(scale(iris[,1]))
> qqline(scale(iris[,1]))

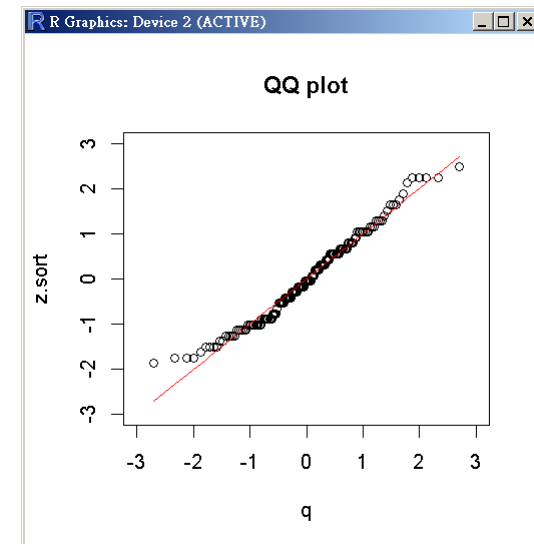
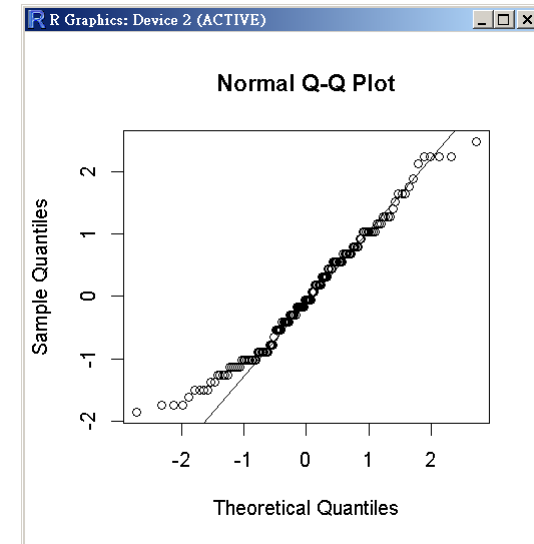
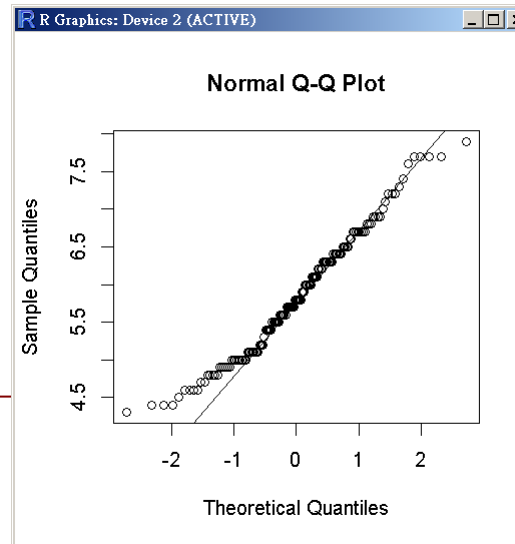
> my.qqplot(iris[,1])
```

```
my.qqplot <- function(x){
  x.mean <- mean(x)
  x.var <- var(x)
  n <- length(x)

  z <- (x-x.mean)/sqrt(x.var)
  z.mean <- mean(z)
  z.var <- var(z)
  z.sort <- sort(z)

  k <- 1:n
  p <- (k-0.5)/n
  q <- qnorm(p)

  plot(q, z.sort, xlim=c(-3, 3), ylim=c(-3, 3))
  title("QQ plot")
  lines(q, q, col=2)
}
```



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

$$S_n = X_1 + \dots + X_n$$

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

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

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

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

- 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。

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	-0.84	0.77	1.60		0.55

mean	0.07	-0.04	0.44	0.31	...	-0.21
------	------	-------	------	------	-----	-------

範例: 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。

$X \sim B(n, p)$ 二項式分布
<http://shorturl.at/iLUV5>

$$1. \quad \mu = E(X) = p = 0.7$$

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

$$2. \quad X_1, X_2, \dots, X_{210}: \\ X_i = 1 : \text{success} \\ X_i = 0 : \text{fail} \\ \bar{X}_{210} = \frac{X_1 + \dots + X_{210}}{210} \\ \mu_{\bar{X}} = \mu = 0.7 \\ \sigma_{\bar{X}}^2 = \frac{\sigma^2}{210} = 0.001$$

$$3. \quad P(X_1 + X_2 + \dots + X_{210} > 126) \\ = P(\bar{X} > \frac{126}{210}) \\ = P(\bar{X} > 0.6) \\ = P(Z > \frac{0.6 - 0.7}{\sqrt{0.001}}) \\ = P(Z > -3.16228) \\ = 0.99922$$

應用CLT

```
> 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, $X \sim B(210, 0.7)$

```
> 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

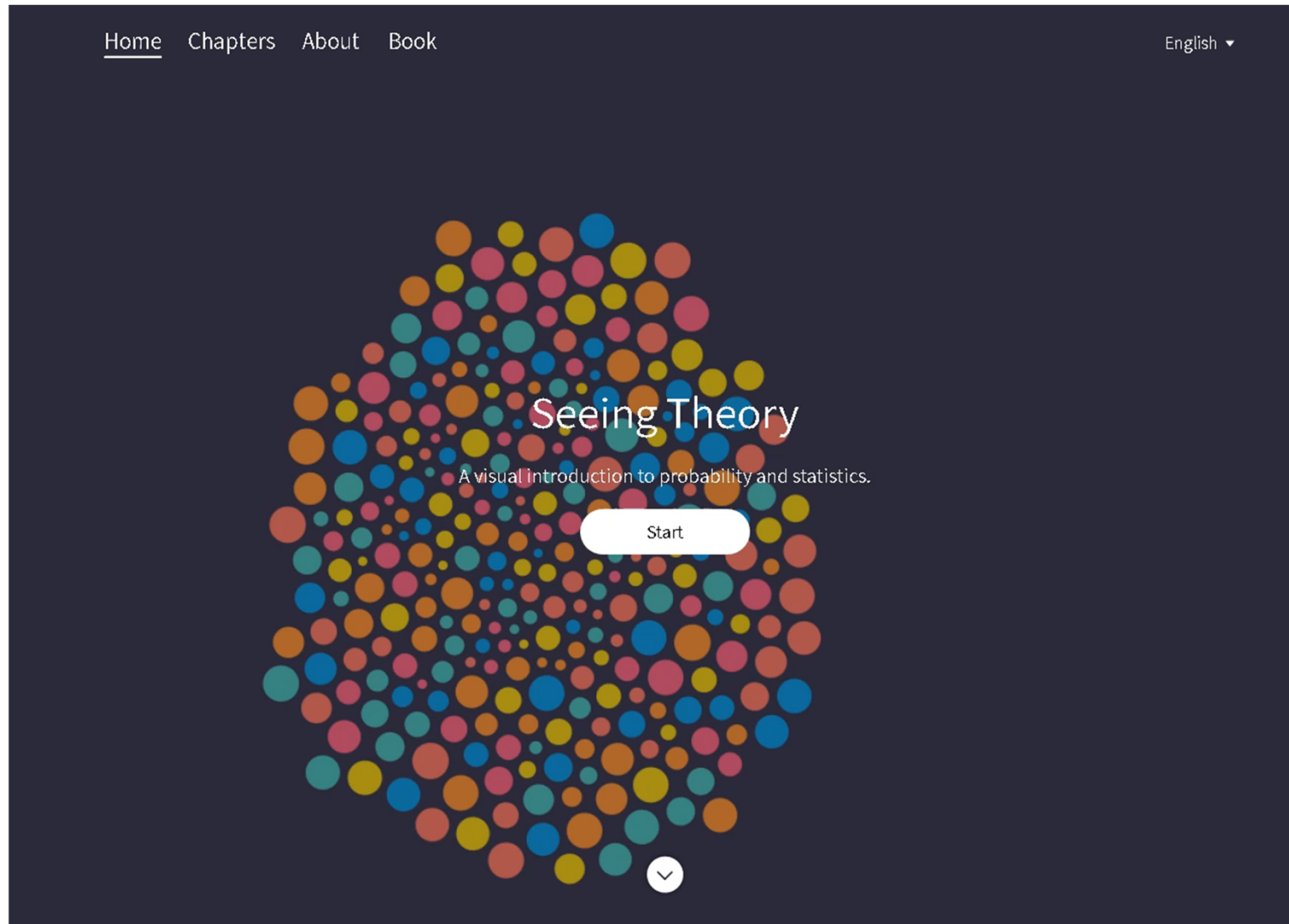
$prob
[1] 0.9992
```



中央極限定理： 樣本平均之抽樣分佈

40/48

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



1. 先做隨機樣本的取樣。

$$X \sim D(\cdot)$$

$$X_1, X_2, \dots, X_{m_0} \sim D(\cdot)$$

$$m = m_0$$

2. 計算樣本平均。

$$\bar{X}_{m_0} = \frac{1}{m_0}(X_1 + X_2 + \dots + X_{m_0})$$

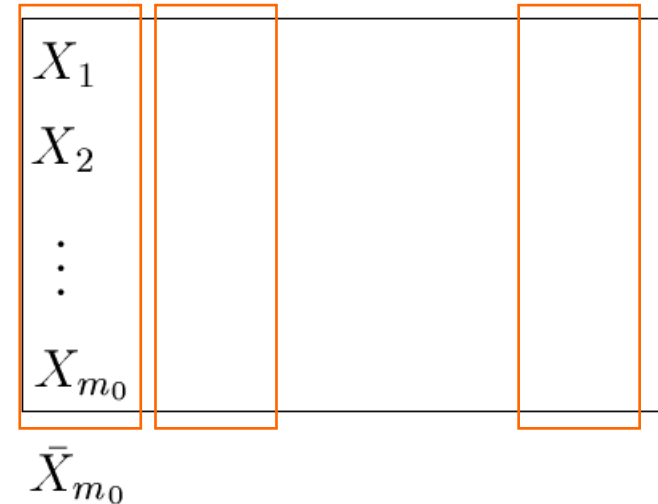
3. 重復上述動作數百或數千次，得到抽樣平均的分佈。
4. 描繪出抽樣平均之抽樣分配直方圖。
5. 畫出相對應的qqplot。
6. 再做各種不同樣本數($m_0=1, 5, 15, 30, \dots$)的抽樣計算。

$$X_1, X_2, \dots \sim U(5, 80)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

樣本數

重復數



```

umin <- 5
umax <- 80
n.sample <- 20
n.repeated <- 500

RandomSample <- matrix(0, n.sample, n.repeated)
for(i in 1:n.repeated){
  rnumber <- runif(n.sample, umin, umax)
  RandomSample[,i] <- as.matrix(rnumber)
}
dim(RandomSample)

```

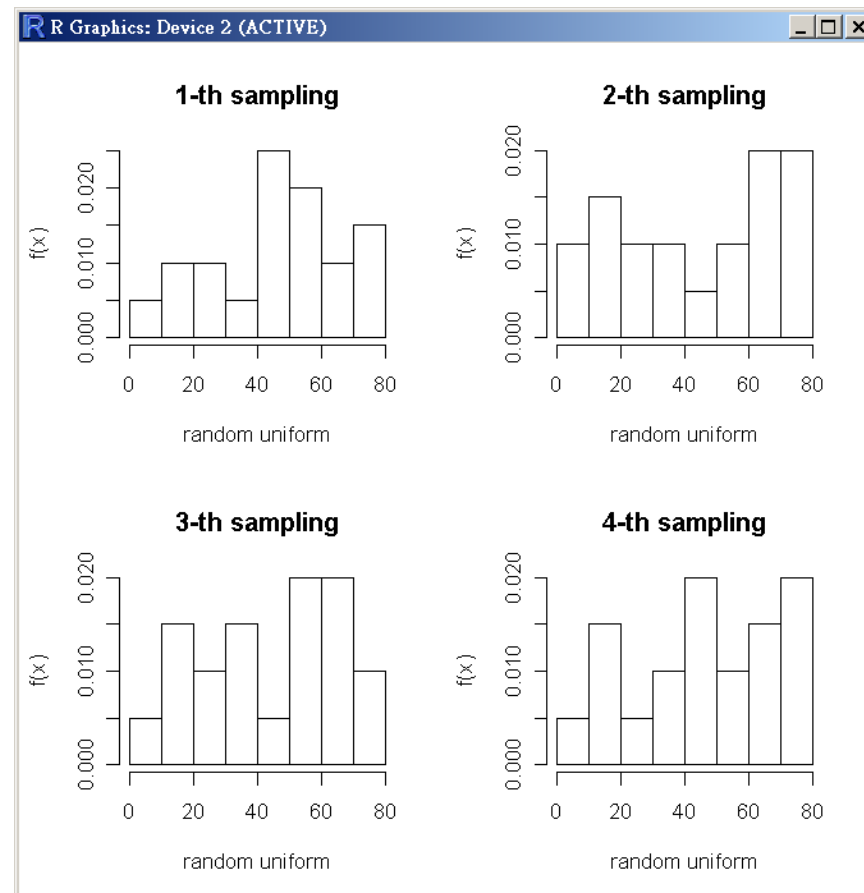
```

par(mfrow=c(2,2))
for(i in 1:4){
  title <- paste(i,"-th sampling", sep="")
  hist(RandomSample[,i], ylab="f(x)", xlab="random uniform", pro=T, main=title)
}

```

$$X_1, X_2, \dots \sim U(5, 80)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$



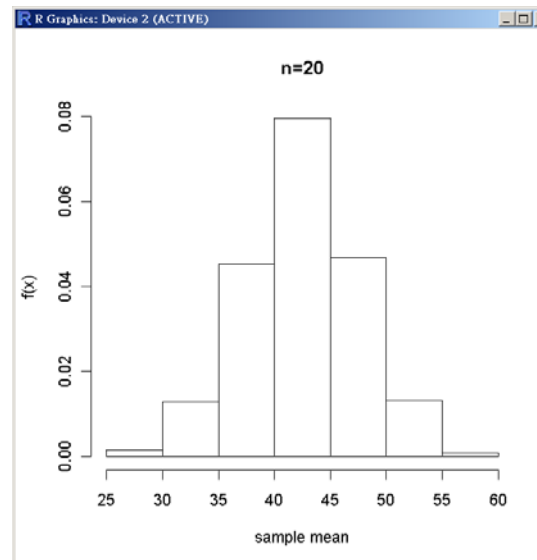
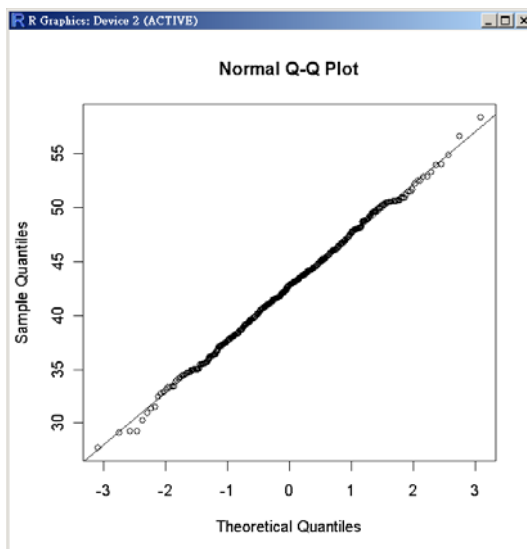
抽樣樣本平均之直方圖&QQplot

```
> SampleMean <- apply(RandomSample, 2, mean)
> hist(SampleMean, ylab="f(x)", xlab="sample mean", pro=T, main="n=20")
```

$$X_1, X_2, \dots \sim U(5, 80)$$

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

```
myfun <- function(x){
  mu <- (umin + umax)/2
  s2 <- (umax - umin)^2/12
  z <- (mean(x) - mu)/sqrt(s2/n)
  z
}
```



Notation	$U(a, b)$ or $\text{unif}(a, b)$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a, b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$
Mean	$\frac{1}{2}(a + b)$
Median	$\frac{1}{2}(a + b)$
Mode	any value in (a, b)
Variance	$\frac{1}{12}(b - a)^2$

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

```
> qqnorm(SampleMean)
> qqline(SampleMean)
```

重複不同的樣本數

```

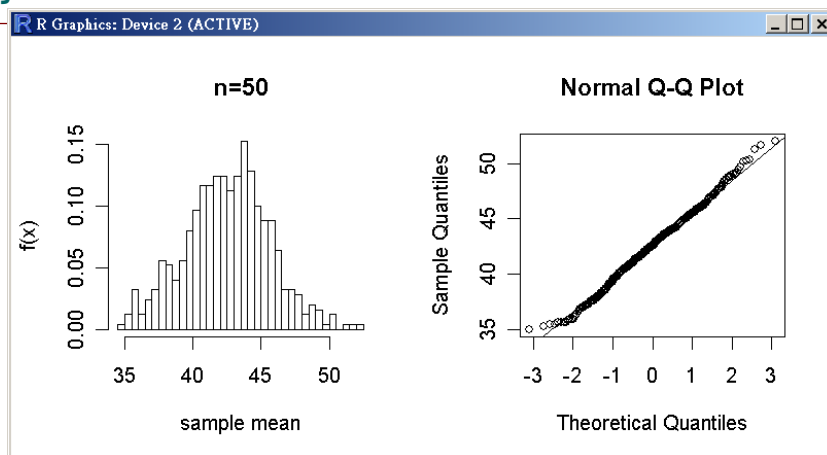
CLT.unif <- function(umin, umax, n.sample, n.repeated){
  RandomSample <- matrix(0, n.sample, n.repeated)
  for(i in 1:n.repeated){
    rnumber <- runif(n.sample, umin, umax)
    RandomSample[,i] <- as.matrix(rnumber)
  }
  SampleMean <- apply(RandomSample, 2, mean)
  par(mfrow=c(1,2))
  title <- paste("n=",n.sample, sep="")
  hist(SampleMean, breaks=30, ylab="f(x)", xlab="sample mean", freq=F,
main=title)
  qqnorm(SampleMean)
  qqline(SampleMean)
}

```

```

myfun <- function(x){
  mu <- (umin + umax)/2
  s2 <- (umax - umin)^2/12
  z <- (mean(x) - mu)/sqrt(s2/n)
  z
}

```

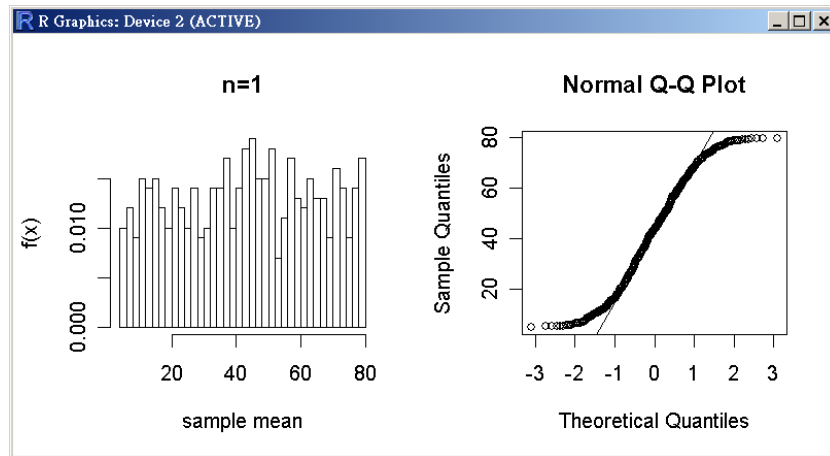


`CLT.unif(5, 80, 50, 500)`

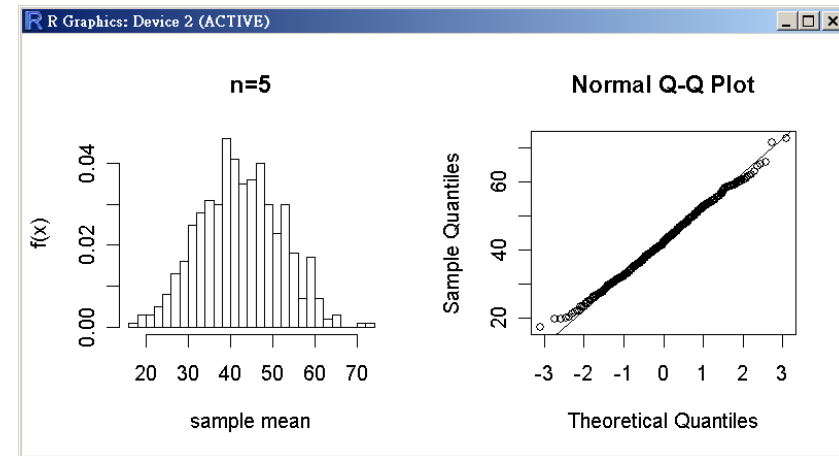
當樣本數 n 愈大時，從樣本平均數的抽樣分配可以得到「中央極限定理」的主要結論。

CLT.unif(umin, umax, n.sample, n.repeated)

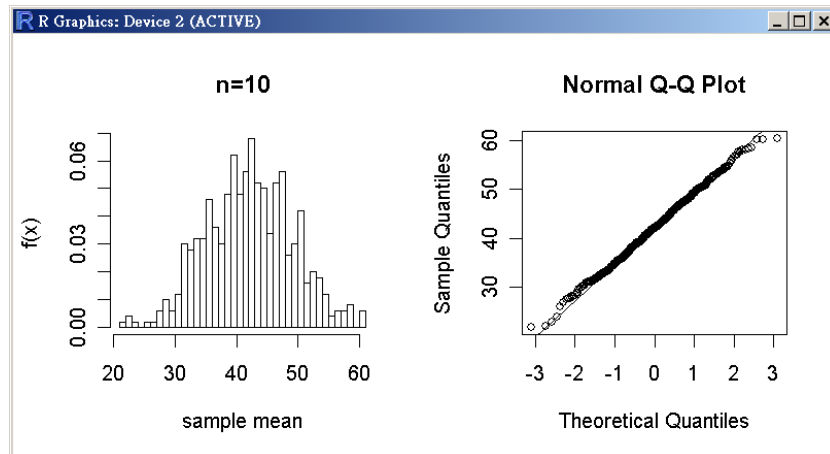
CLT.unif(5, 80, 1, 500)



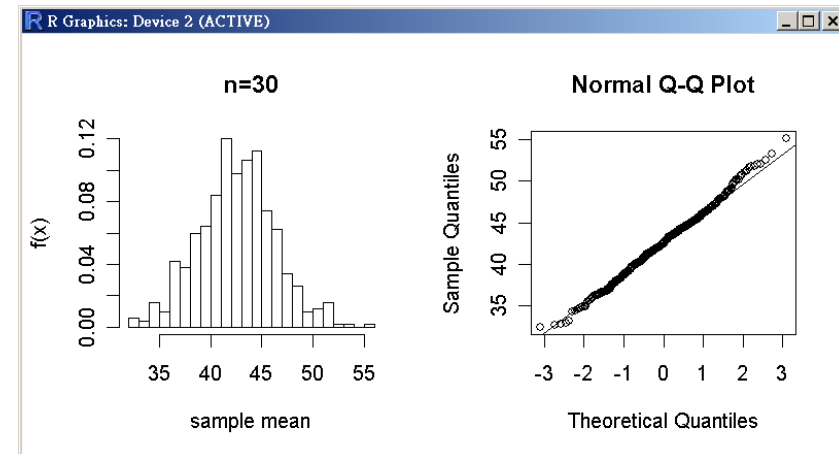
CLT.unif(5, 80, 5, 500)



CLT.unif(5, 80, 10, 500)



CLT.unif(5, 80, 30, 500)



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1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
```