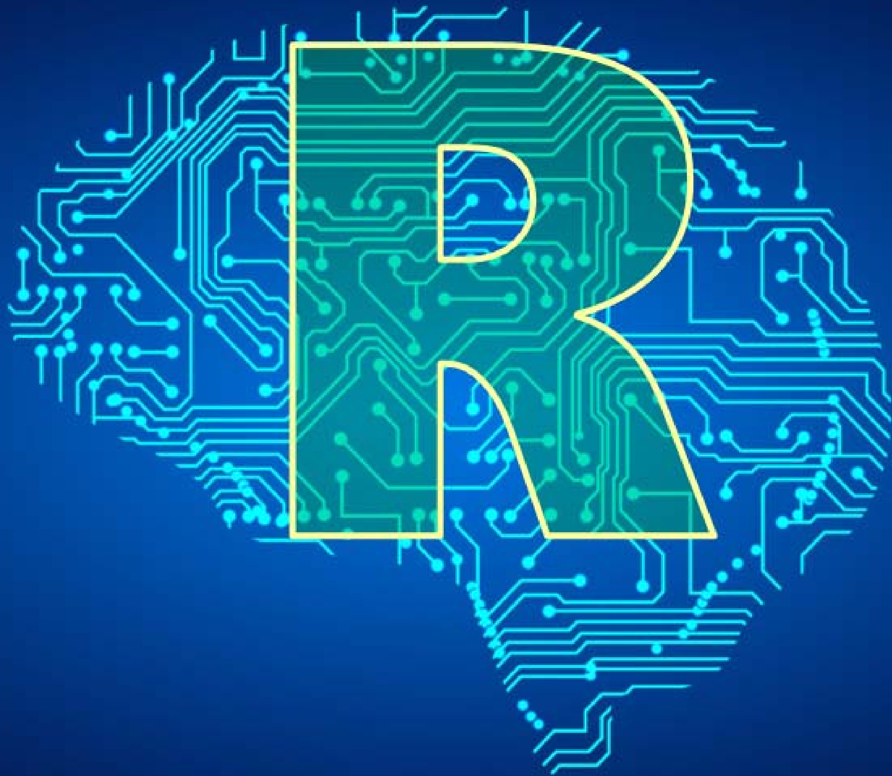
台灣人工智慧學校



機率分佈

吳漢銘
國立臺北大學 統計學系

http://www.hmwu.idv.tw

本章大綱

- 常見統計名詞
- 機率分佈 (Probability distribution)
  - 統計分配之描述、常見之分佈(二項式分佈、常態分佈)、隨機抽樣
- 以常態機率逼近二項式機率
- 大數法則 (LLN)
- 中央極限定理 (CLT)
- 用R程式模擬算機率

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.

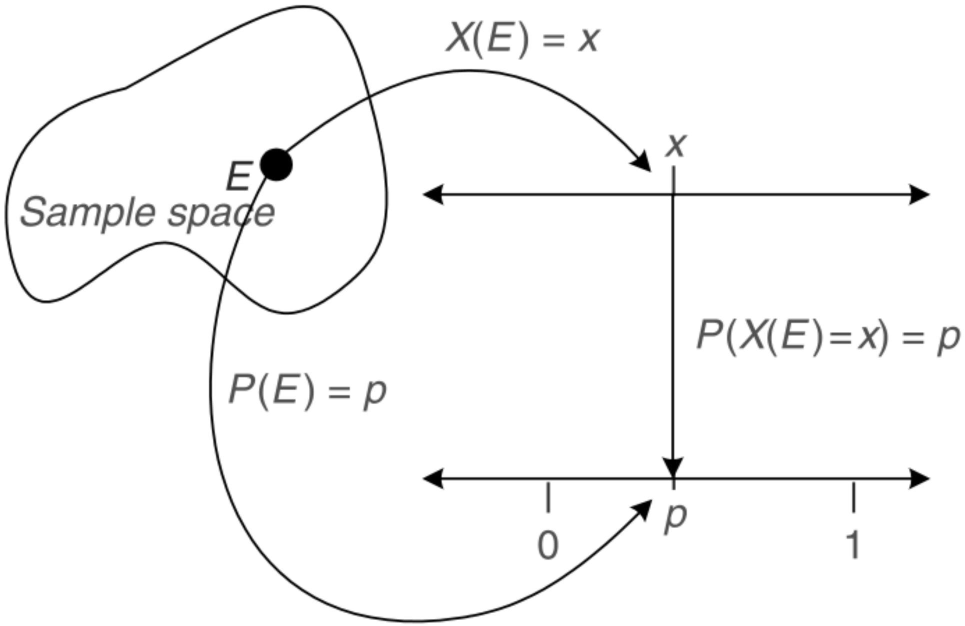
- In the experiment of tossing 4 coins, we may consider tossing each coin as a trial and therefore say that there are 4 trials in the experiment.

- 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 $P(X \leq x|\theta) = p$.

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

Probability Distribution

- The probability distribution is a description of a random phenomenon in terms of the **probabilities of events**.

- A probability distribution is a mathematical function that can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment.

- ***EXAMPLE:*** if the random variable X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair).

NOTE:

- The terms "**probability distribution function**" and "**probability function**" have also sometimes been used to denote the **probability density function**.
- "**probability distribution function**" may be used when the probability distribution is defined as a function over general sets of values, or it may refer to the **cumulative distribution function**.

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

Formal definition

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

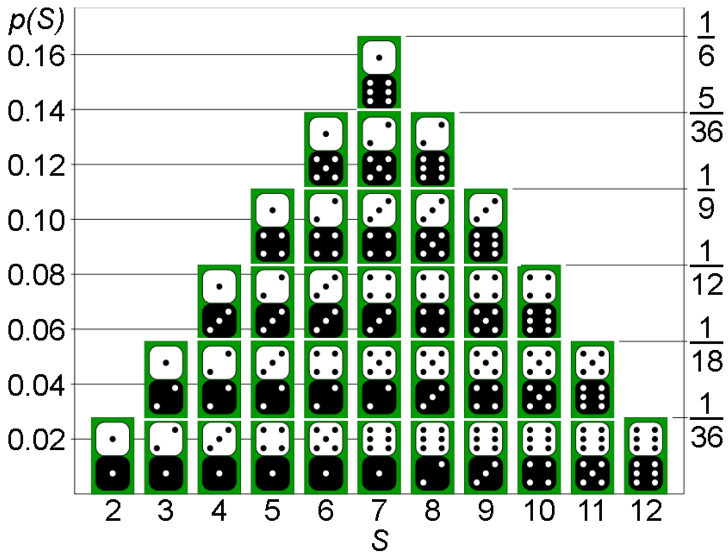
Suppose that X: $S \to A$ ($A \subseteq \mathbf{R}$) is a discrete random variable defined on a sample space S. Then the probability mass function f_X: $A \to [0, 1]$ for X is defined as

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

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:

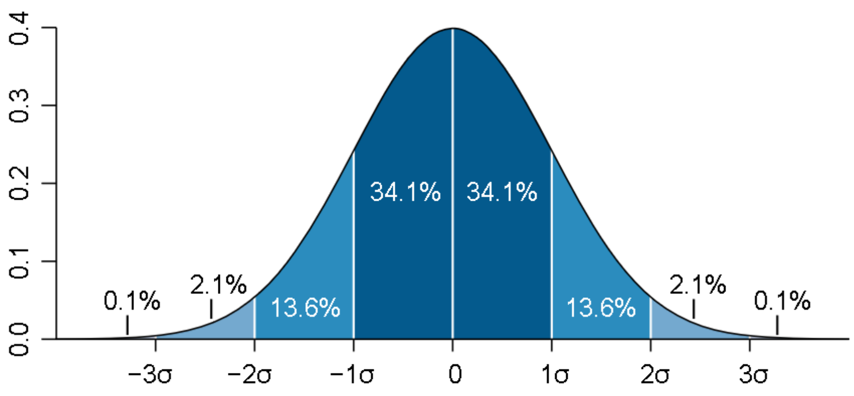
$$P(X \in A) = \int_A f(x)dx \qquad \mathrm{P}[a \le X \le b] = \int_a^b f(x)\,dx$$

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



常用機率分配

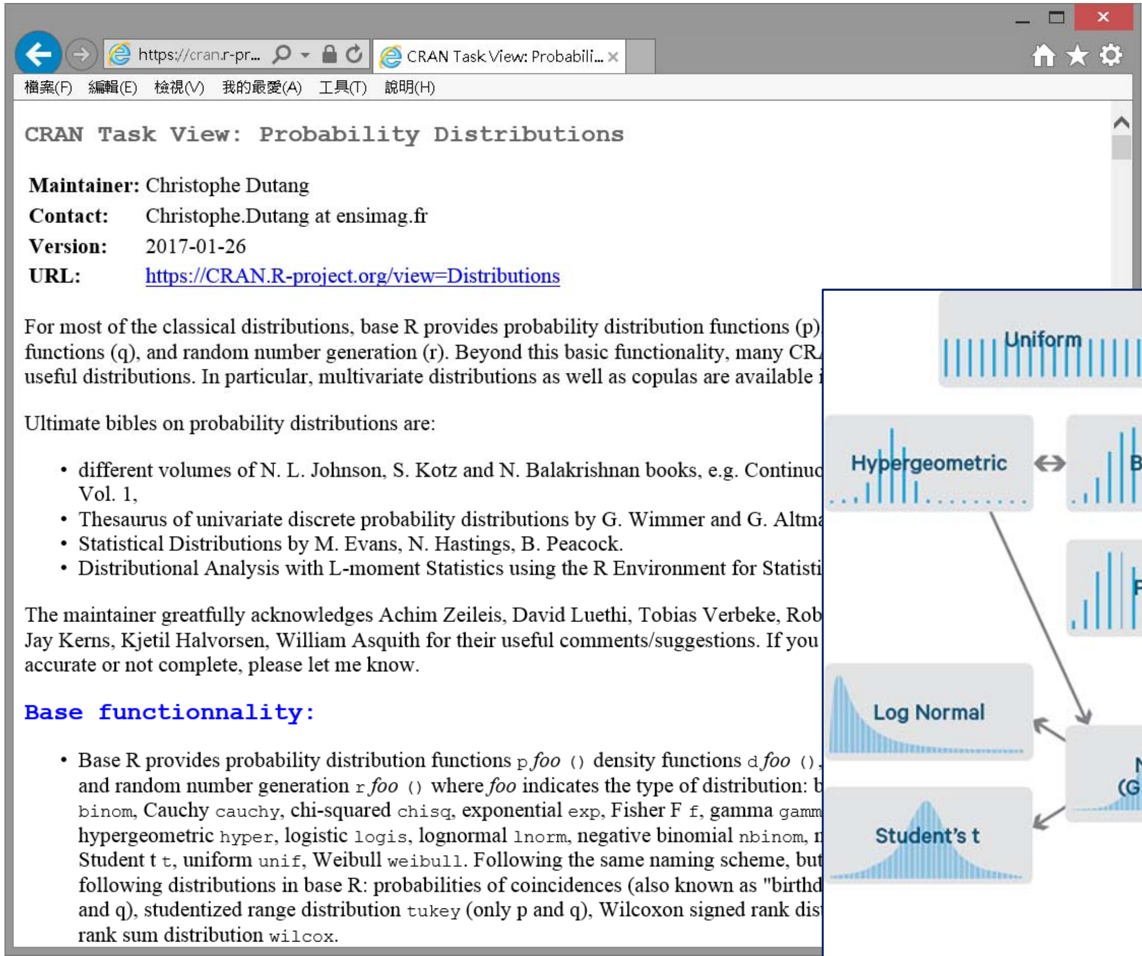
以常態分佈normal為例:

- 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, df1, 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	t df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Wiki Category:Discrete distributions
https://en.wikipedia.org/wiki/Category:Discrete_distributions
Wiki Category:Continuous distributions
https://en.wikipedia.org/wiki/Category:Continuous_distributions

CRAN Task View: Probability Distribution


https://cran.r-pr... 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:

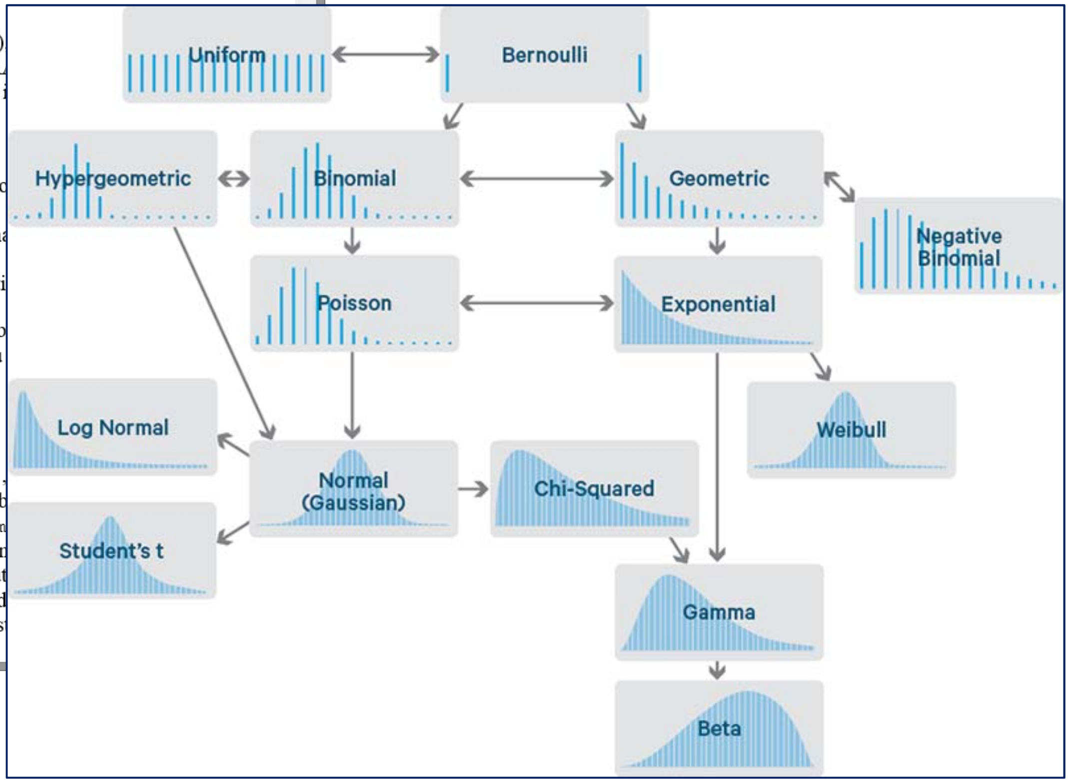
- different volumes of N. L. Johnson, S. Kotz and N. Balakrishnan books, e.g. Continu Vol. 1,
- Thesaurus of univariate discrete probability distributions by G. Wimmer and G. Altma
- Statistical Distributions by M. Evans, N. Hastings, B. Peacock.
- Distributional Analysis with L-moment Statistics using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

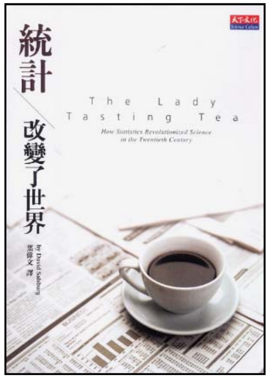
https://cran.r-project.org/web/views/Distributions.html



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

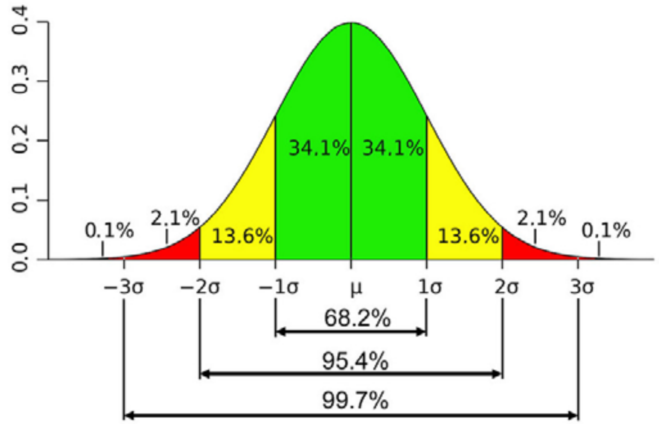
- 1895-1898，發表一系列和相關性(correlation) 有關的論文，涉及動差、相關係數、標準差、卡方適合度檢定，奠定了現代統計學的基礎。
- 引入了統計模型的觀念: 如果能夠決定所觀察現象的**機率分佈的參數**，就可以了解所觀察現象的本質。

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.

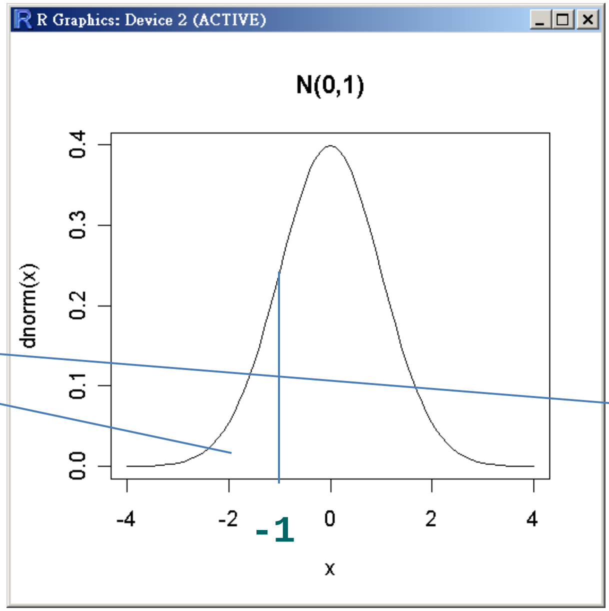
https://en.wikipedia.org/wiki/Probability_distribution

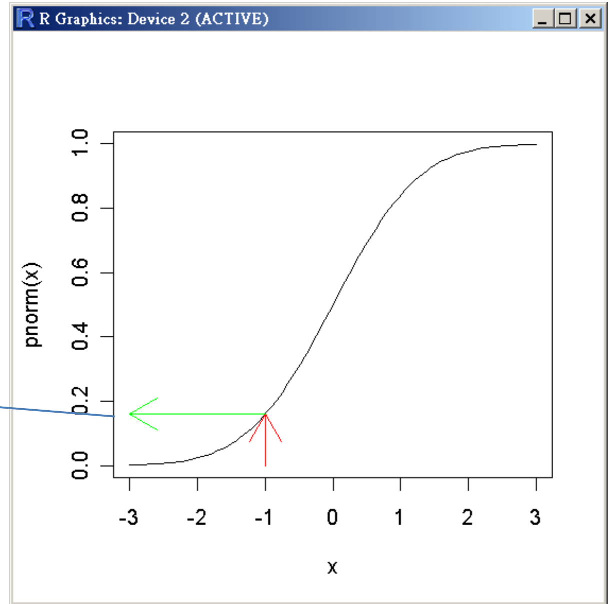
累積機率分配函數 CDF (p)

- 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
```

PDF





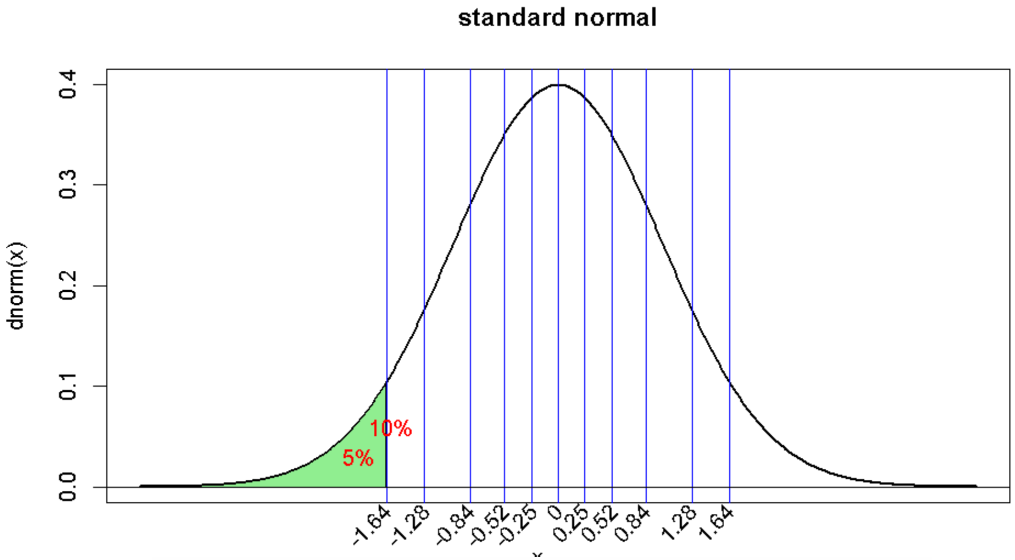
CDF

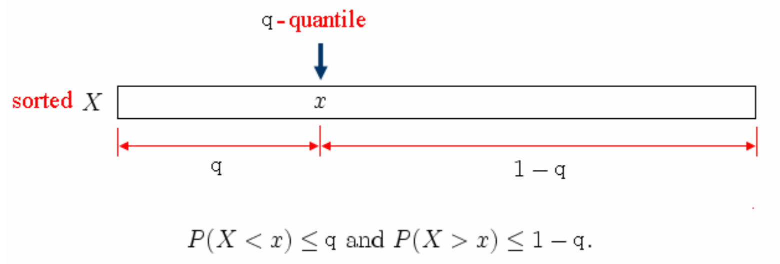
分位數 Quantiles (q)

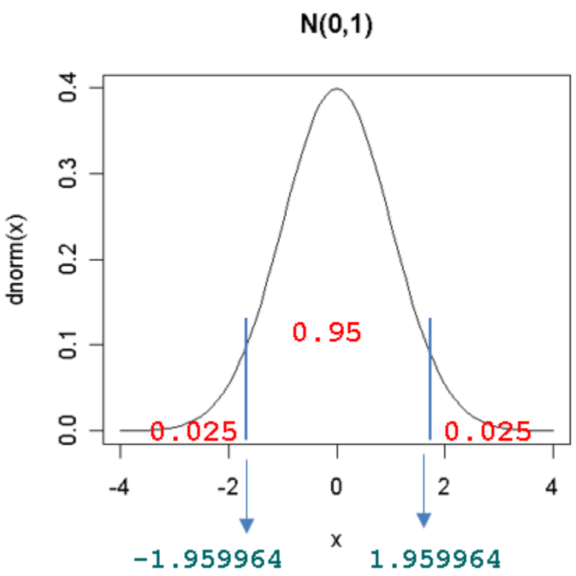
- The quantile function is the inverse of the cumulative distribution function: $F^{-1}(p) = x$.
- We say that q is the $x\%$-quantile if $x\%$ of the data values are $\leq q$.

```
> # 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
```

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







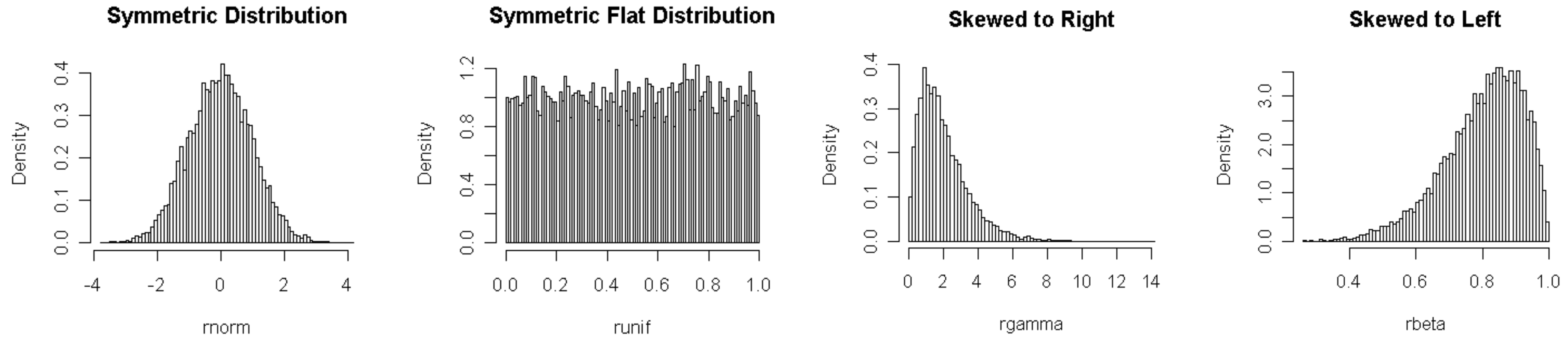
$$\bar{x} + z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}}$$

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

隨機數 Random Numbers (r)

- *Let* X_i is a vector of measurements for the *i*-th object in the sample.
- $(X_1, X_2, \ldots, X_n)$ is said to be a random sample of size n from the common distribution if $X_1, X_2, \ldots, X_n$ 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")
```



隨機抽樣 (Random Sampling)

- 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"
```

隨機抽樣 (Random Sampling)


```
> x <- 1:5
> sample(x) # permutation
[1] 3 1 5 4 2
```

- Clinical trials: randomization: random assign to two groups, total 20 subjects random assigning treatment groups

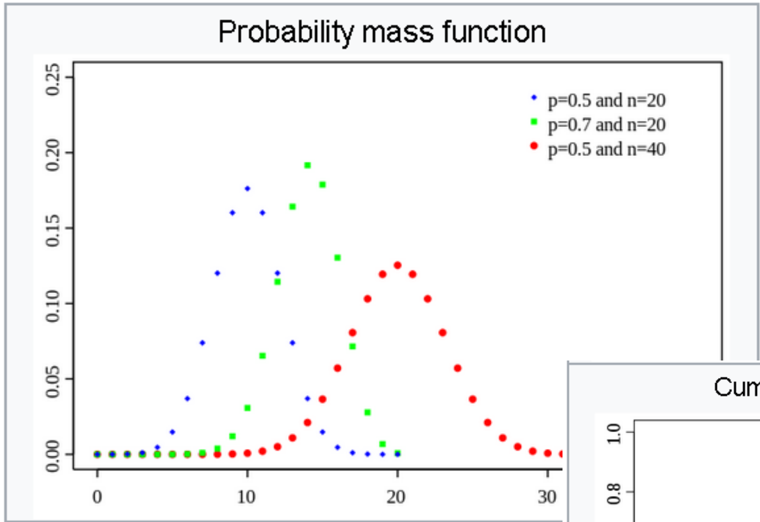
```
> sample(2, size=20, replace=TRUE)
 [1] 2 2 2 1 1 2 2 2 1 2 1 1 2 1 2 1 2 1 1 1
```

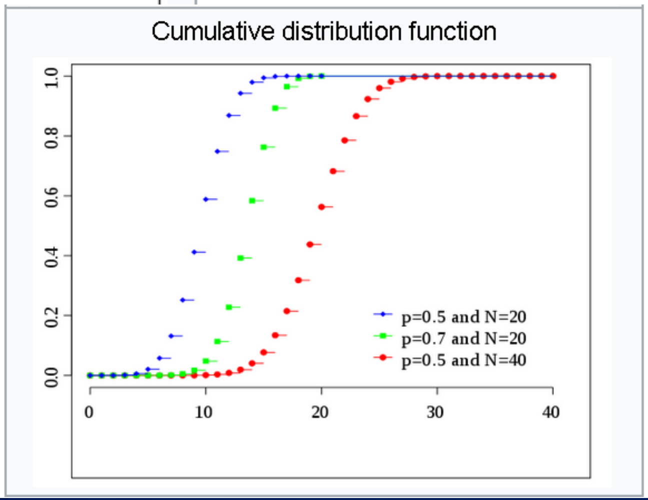
- random choose 10 subjects to group 1

```
> sample(20, size=10, replace=FALSE)
 [1] 10 13 16  8  4 14  7 11  1  5
```

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	$n \in \mathbb{N}_0$ — 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。

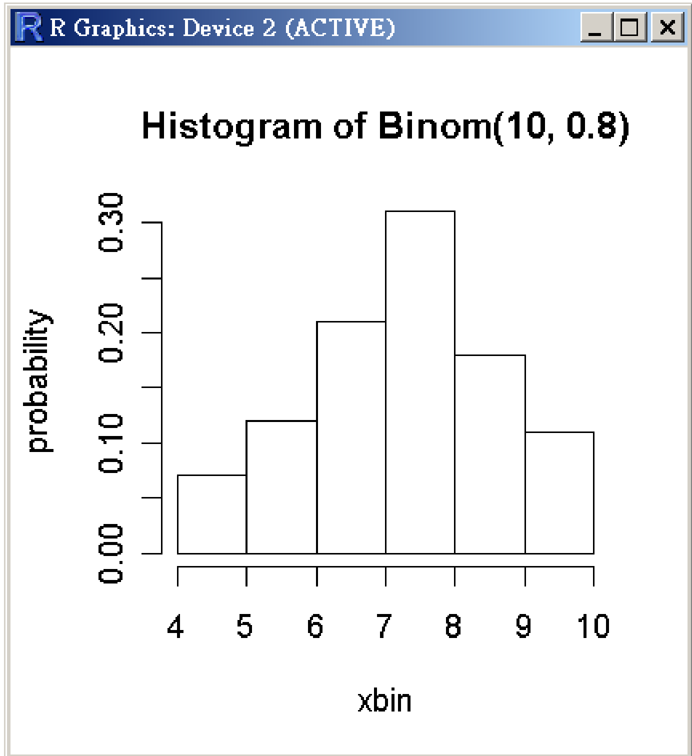
```
> qbinom(0.1208, 10, 0.8)
[1] 6
> pbinom(6, 10, 0.8)
[1] 0.1208739
```

二項式分佈

X~B(10, 0.8)

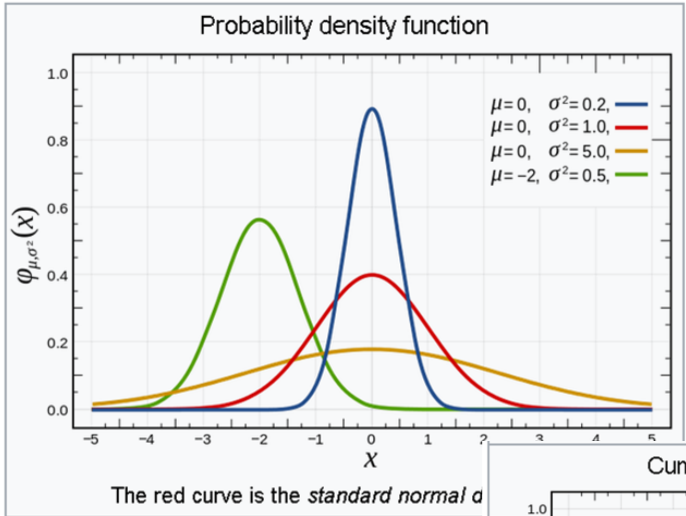
- 產生隨機樣本數100的二項隨機數值，計算其平均數及變異數，並與理論值比較。
- 畫直方圖，x-axis="機率值"，label="probability", title="Histogram of Binom(10, 0.8)"。

```
> n <- 10
> p <- 0.8
> m <- 100
> xbin <- rbinom(m, n, p)
> table(xbin)
xbin
 4  5  6  7  8  9 10
 1  6 12 21 31 18 11
> mu <- n*p; mu
[1] 8
> sigma2 <- n*p*(1-p); sigma2
[1] 1.6
> mean(xbin)
[1] 7.73
> var(xbin)
[1] 1.956667

> hist(xbin, ylab="probability", main="Histogram of
Binom(10, 0.8)", prob=T)
```

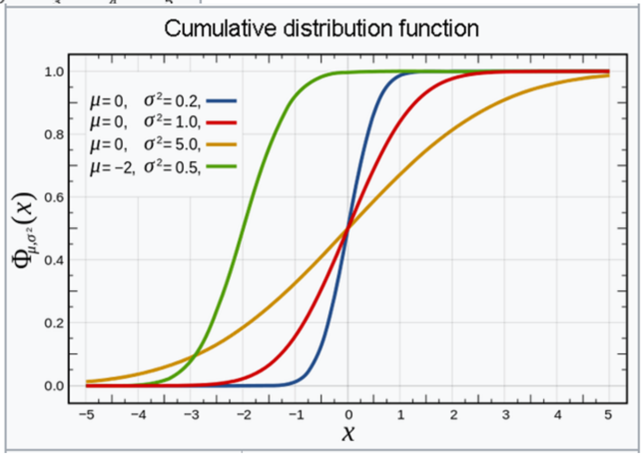


常態分佈

- `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* d



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

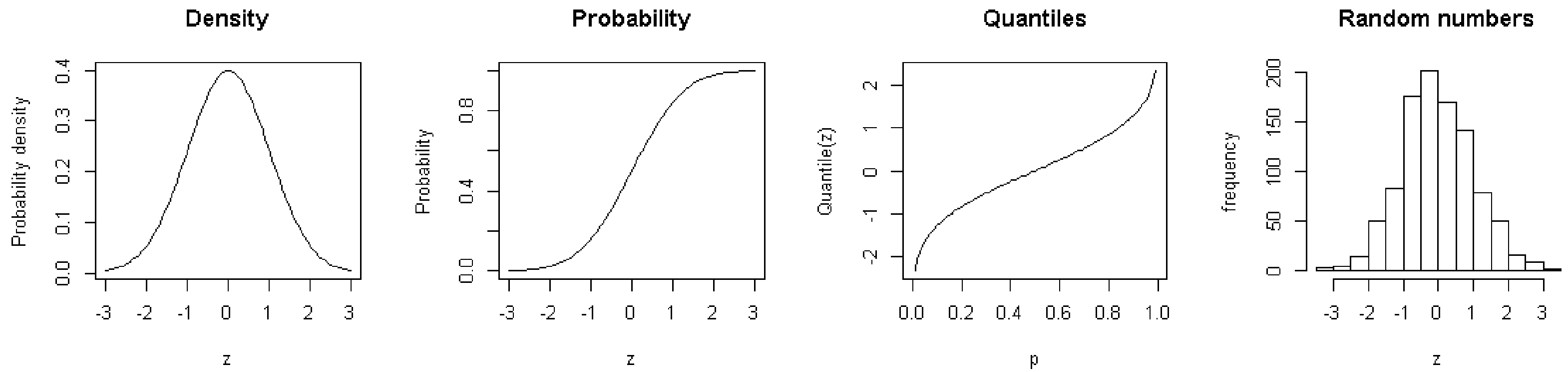
常態分佈 (Normal Distribution)

Z ~N(0, 1)

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

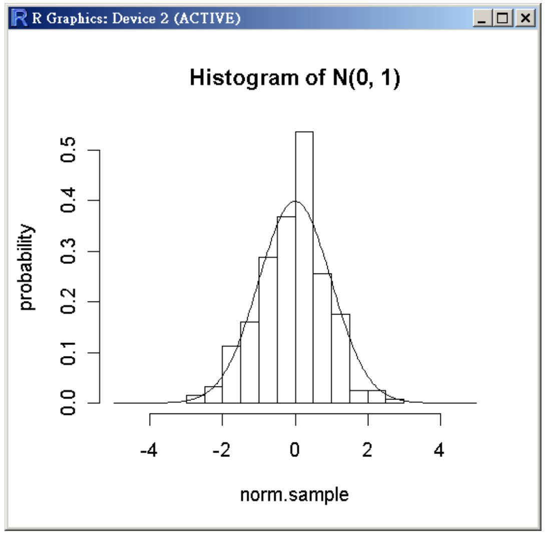
```
> dnorm(10, 10, 2) # X~N(10, 4)
[1] 0.1994711
> pnorm(1.96, 10, 2)
[1] 2.909907e-05
> qnorm(0.975, 10, 2)
[1] 13.91993
> rnorm(5, 10, 2)
[1]  9.043357 11.721717  7.763277  9.563463 10.072386
> pnorm(15, 10, 2) - pnorm(8, 10, 2)  # P(8<=X<=15)
[1] 0.8351351
```

```
> 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")
```



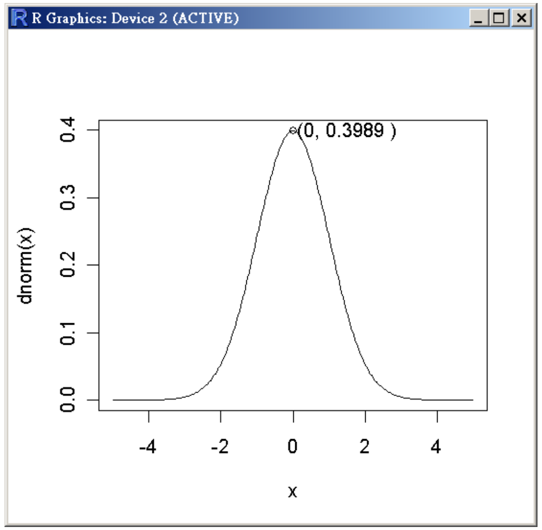
常態分佈

```
> 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, ")"))
```



以常態機率逼近二項式機率

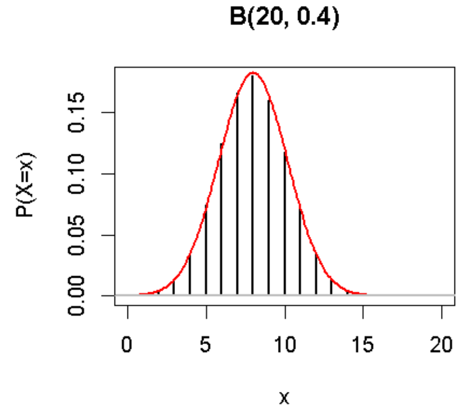
set $n = 20$ and $\pi = 0.4$ and calculate the density of the binomial,

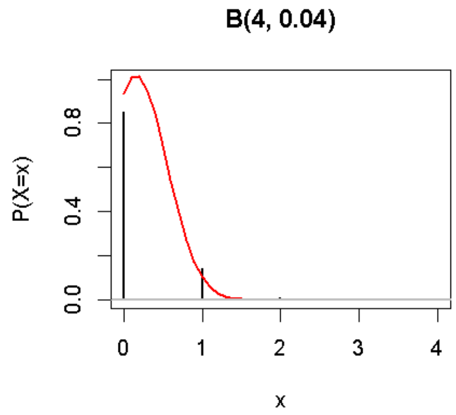
$$P\left(X = x \middle| n, \pi\right) = \binom{n}{x} \pi^{x} \left(1-\pi\right)^{n-x}$$

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)。

特別注意: 本投影片中符號n和m之區別。

Bernoulli試驗

- Bernoulli試驗(伯努利試驗): 擲一公平硬幣一次，可能出現正面或反面。
- 令*X=1*為出現正面, *X=0*為出現反面。
- *X~Binomial(1, 0.5)*。
- 伯努利分佈的平均數 *p*。

$$X_1, X_2, \cdots, \ Binomial(1, 0.5)$$

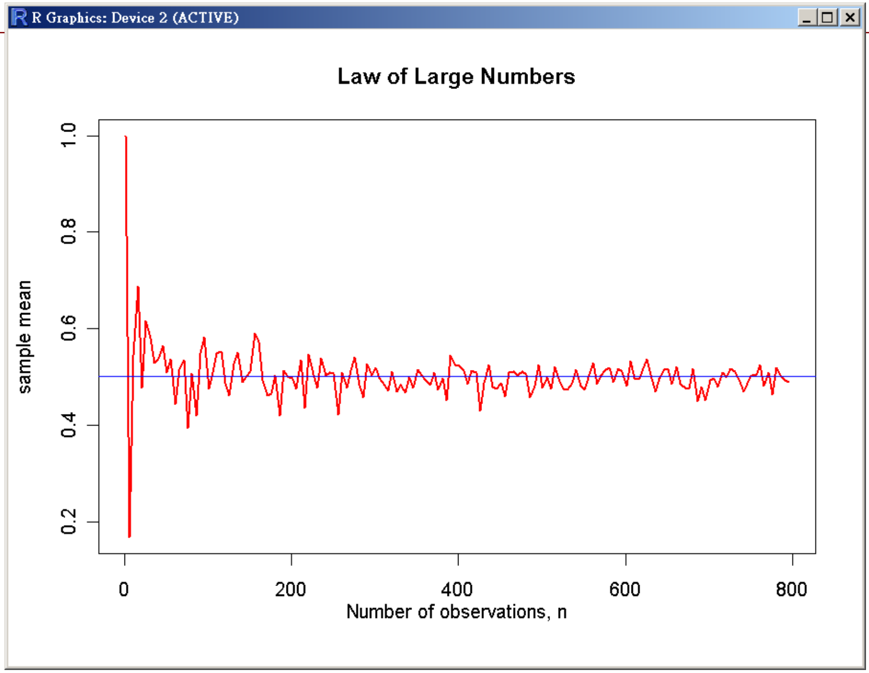
$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$ ：平均正面次數

rbinom(m, size=1, prob)
m: number of observations (樣本數)
size=1: number of trials
prob: probability of success on each trial

m Bernoulli random samples:
rbinom(m, 1, 0.5)

利用Bernoulli試驗說明大數法則

```
sample.size <- seq(from=1, to=800, by=5)
m <- length(sample.size)
xbar <- numeric(m)
for(i in 1:m){
  xbar[i] <- mean(rbinom(sample.size[i], 1, 0.5))
}
plot(sample.size, xbar, xlab="Number of observations, n",
     ylab="sample mean", main="Law of Large Numbers",
     type="l", col="red", lwd=1.5)
abline(h=0.5, col="blue")
```



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 $\sigma^2 > 0$.

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

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

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

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
```

驗証中央極限定理

1. 先做隨機樣本的取樣。

   $X \sim D(\cdot)$ $X_1, X_2, \cdots, X_{m_0} \sim D(\cdot)$ $m = m_0$

2. 計算樣本平均。

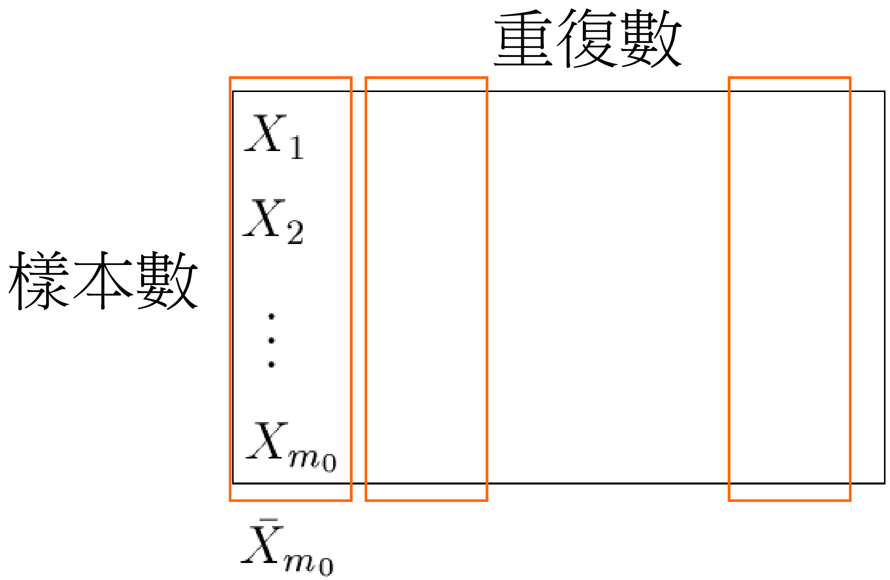
   $$\bar{X}_{m_0} = \frac{1}{m_0}(X_1 + X_2 + \cdots + X_{m_0})$$

3. 重復上述動作數百或數仟次，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,...)的抽樣計算。

範例: Uniform Distribution

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

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



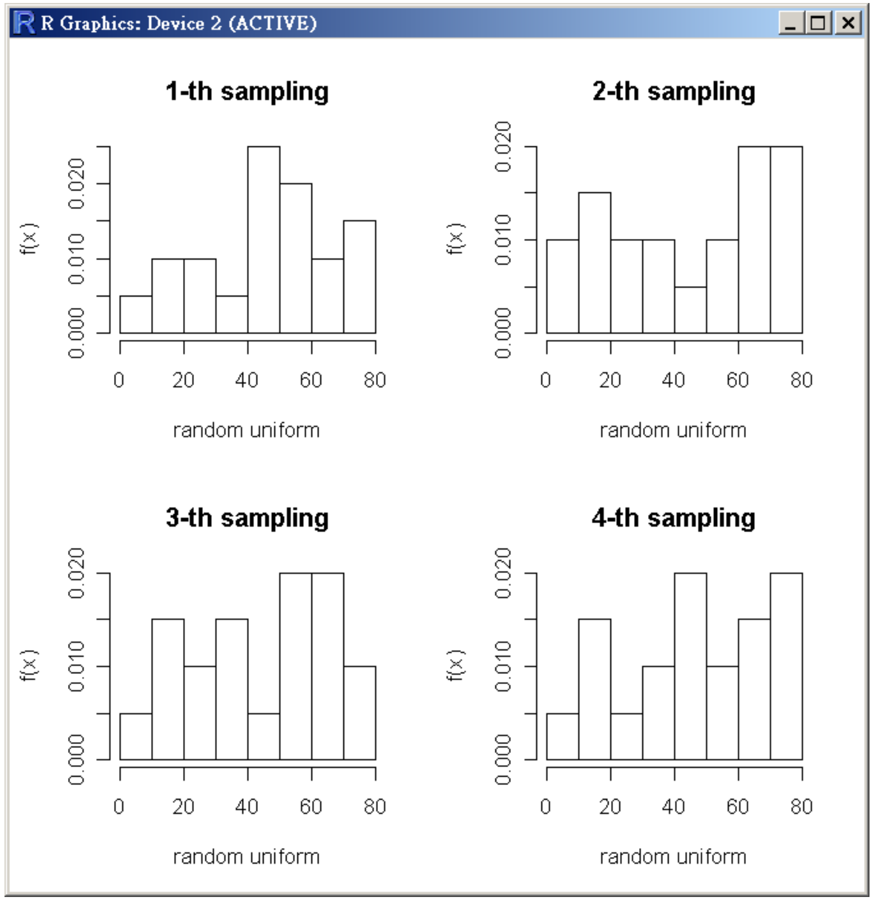

```
umin <- 5
umax <- 80
n.sample <- 20
n.repeated <- 500

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)
```

抽樣樣本之直方圖

```
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)
}
```

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

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

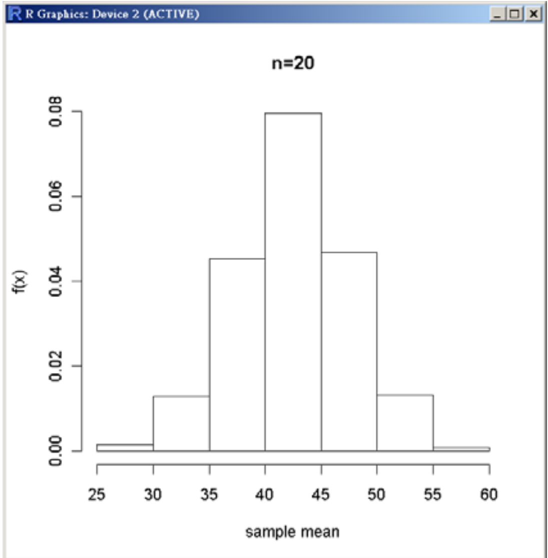


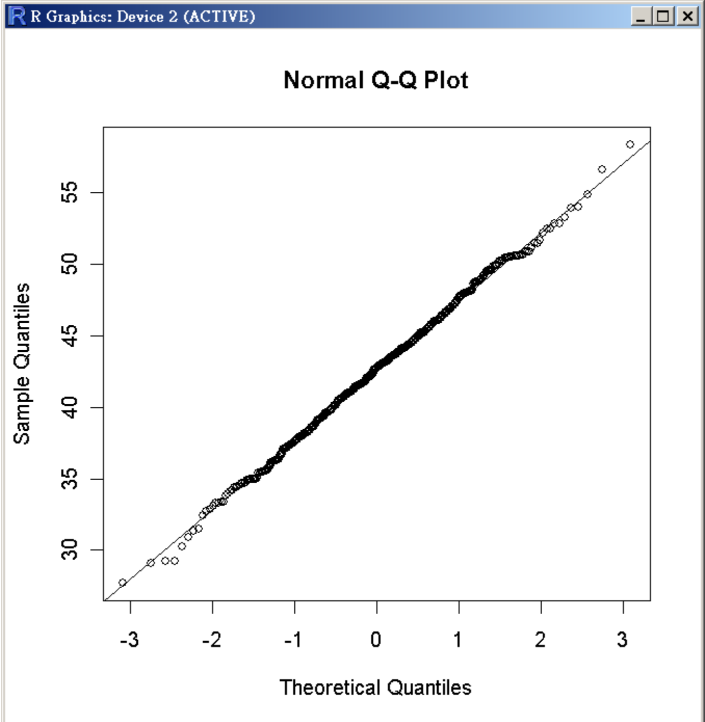
抽樣樣本平均之直方圖&QQplot

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

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

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

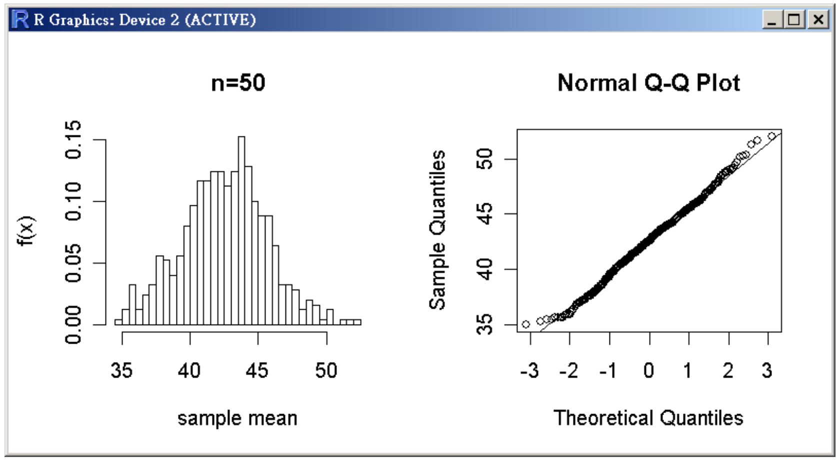





```
> qqnorm(SampleMean)
> qqline(SampleMean)
```

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", pro=T, main=title)
    qqnorm(SampleMean)
    qqline(SampleMean)
}
```

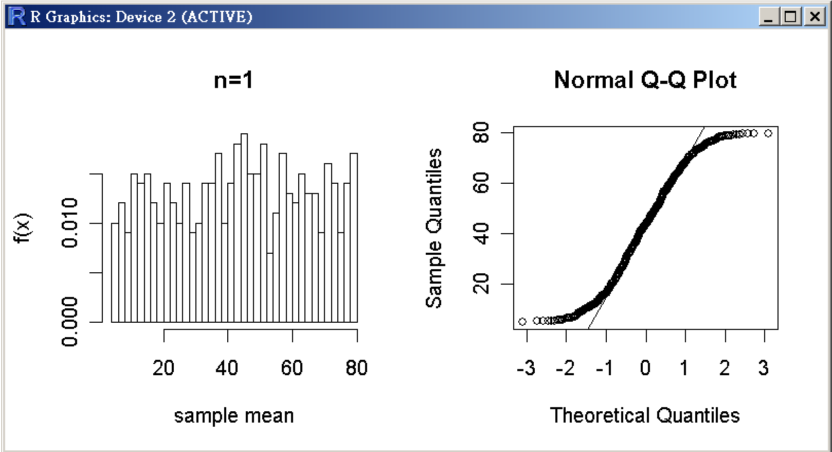



```
CLT.unif(5, 80, 50, 500)
```

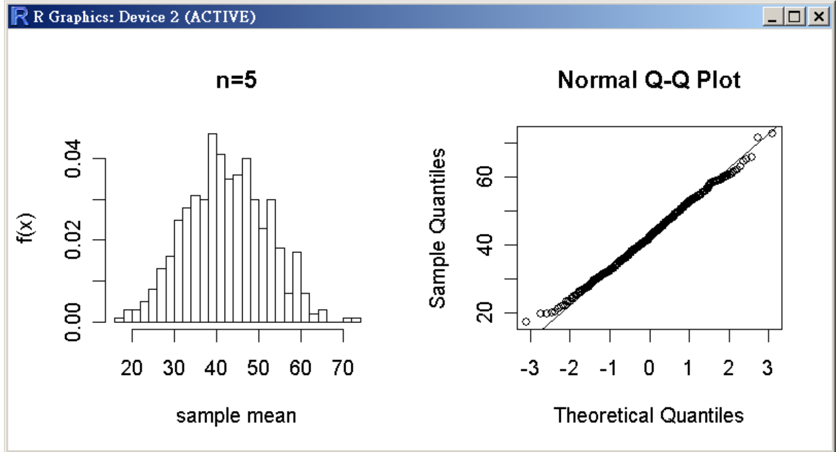
當樣本數*n*愈大時，從樣本平均數的抽樣分配可以得到「中央極限定理」的主要結論。

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

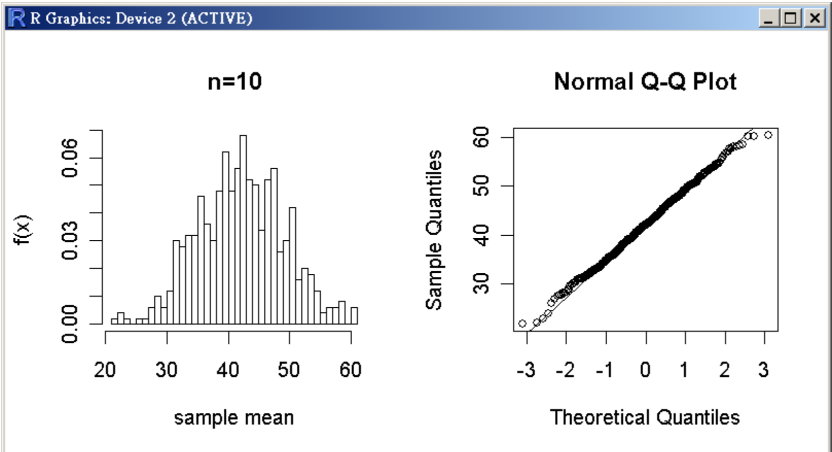
CLT.unif(5, 80, 1, 500)



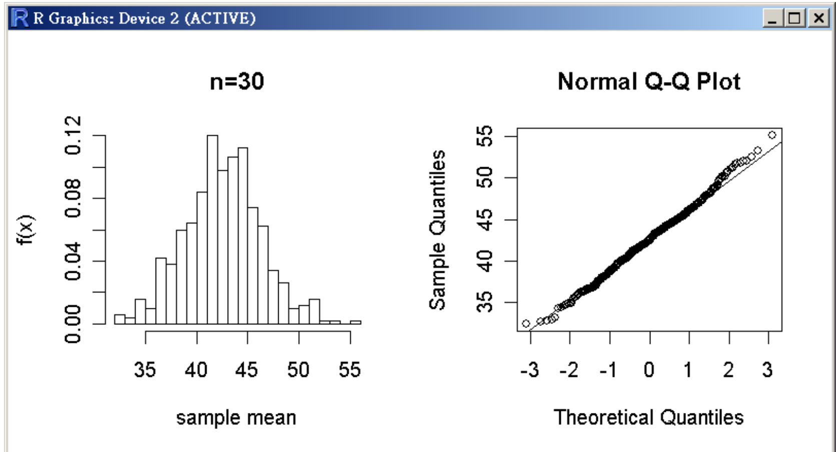
CLT.unif(5, 80, 5, 500)



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



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



練習1: 算大樂透中獎機率

什麼是49選6大樂透

您必須從01~49中任選6個號碼進行投注。開獎時，開獎單位將隨機開出六個號碼加一個特別號，這一組號碼就是該期49選6大樂透的中獎號碼，也稱為「獎號」。您的六個選號中，如果有三個以上（含三個號碼）對中當期開出之六個號碼（特別號只適用於貳獎、肆獎和陸獎），即為中獎，並可依規定兌領獎金。

各獎項的中獎方式如下表：

中獎方式	中獎方式圖示	獎項
與當期六個獎號完全相同者		頭獎
對中當期獎號之任五碼＋特別號		貳獎
對中當期獎號之任五碼		參獎
對中當期獎號之任四碼＋特別號		肆獎
對中當期獎號之任四碼		伍獎
對中當期獎號之任三碼＋特別號		陸獎 NT$1,000
對中當期獎號之任三碼		普獎 NT$400

電腦選號

```
> sample(1:49, 6, replace = FALSE)
[1] 14 45 36 25 38 28
> sample(1:49, 6, replace = FALSE)
[1]  7 25 21 16  8  6
> sample(1:49, 6, replace = FALSE)
[1] 30 17 27 15 19  2
```

```
> set.seed(12345)
> sample(1:49, 6, replace = FALSE)
[1] 36 43 49 41 21  8
> sample(1:49, 6, replace = FALSE)
[1] 16 25 35 46  2  7
> set.seed(12345)
> sample(1:49, 6, replace = FALSE)
[1] 36 43 49 41 21  8
> sample(1:49, 6, replace = FALSE)
[1] 16 25 35 46  2  7
```

資料來源: http://www.taiwanlottery.com.tw

中頭獎機率

大樂透可能出現的號碼組合共有

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6!} = 13{,}983{,}816(\text{種})$$

大樂透的中頭獎機率 $\dfrac{1}{13,983,816}$

彩券玩法

Q9： 投注電腦型彩券時，「電腦快選」會比「自選」號碼容易中獎嗎？

每種遊戲每次開出的獎號都是隨機的，所以「電腦快選」和「自選」號碼的中獎機率都一樣。

資料來源: http://www.taiwanlottery.com.tw

算一下中獎機率

獎項	中獎方式	中獎機率
頭獎	6碼完全相同	$\frac{1}{\binom{49}{6}} = \frac{1}{13983816}$
貳獎	中5碼及特別號	$\frac{\binom{6}{5}}{\binom{49}{6}} = \frac{6}{13983816} = \frac{1}{2330636}$
參獎	中5碼	$\frac{\binom{6}{5}\times\binom{49-6-1}{1}}{\binom{49}{6}} = \frac{252}{13983816} = \frac{1}{554913}$
肆獎	中4碼及特別號	$\frac{\binom{6}{4}\times\binom{49-6-1}{1}}{\binom{49}{6}} = \frac{630}{13983816} = \frac{1}{22196.5}$
伍獎	中4碼	$\frac{\binom{6}{4}\times\binom{49-6-1}{2}}{\binom{49}{6}} = \frac{12915}{13983816} = \frac{1}{1082.8}$
陸獎	中3碼及特別號	$\frac{\binom{6}{3}\times\binom{49-6-1}{2}}{\binom{49}{6}} = \frac{17220}{13983816} = \frac{1}{812.1}$
普獎	中3碼	$\frac{\binom{6}{3}\times\binom{49-6-1}{3}}{\binom{49}{6}} = \frac{229600}{13983816} = \frac{1}{60.9}$

```
> 1 / choose(49, 6)
[1] 7.151124e-08
> choose(6, 5) / choose(49, 6)
[1] 4.290674e-07
> (choose(6, 5)*choose(49-6-1, 1)) / choose(49, 6)
[1] 1.802083e-05
```

你知道嗎?

- 若1注50元，要花7億才可買遍所有號碼組合。
- 被雷擊幾率幾何？因地而異

(1) 全世界每年因雷擊造成的傷亡人數超過1萬，按世界人口數量為70億計算，雷擊機率大約為70萬分之一。

(2) 美聯邦應急管理局估計當前美國人平均遭雷擊的機率為60萬分之一，

(3) 中國國際防雷論壇公布中國人遭雷擊的機率大約為33萬分之一。

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

```
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) # random number
        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
```