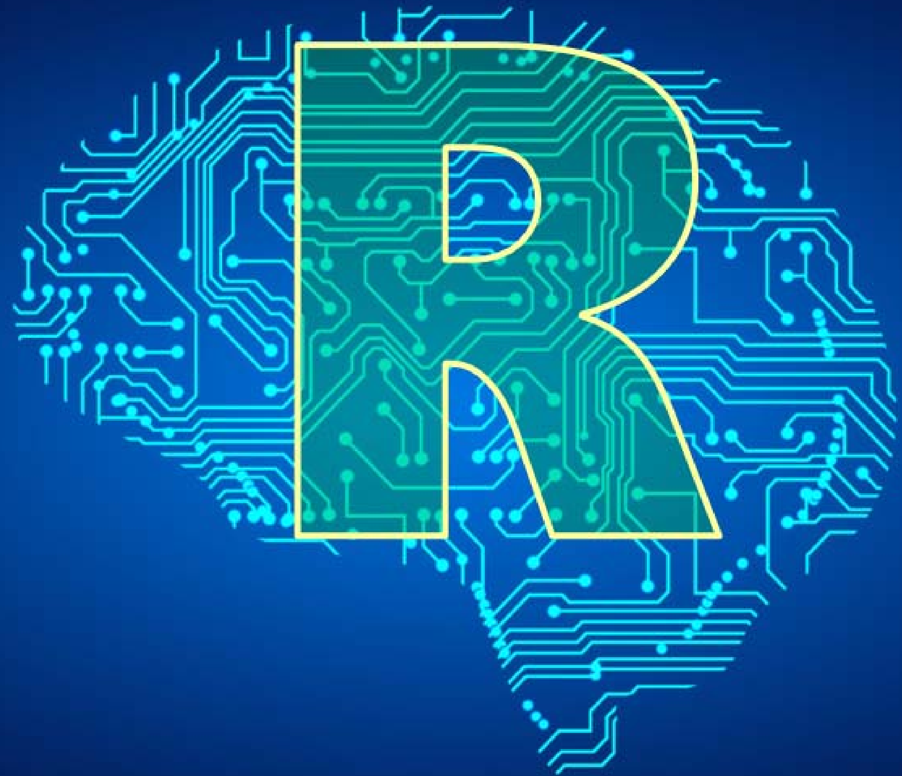
StatR-02

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

機率分佈 - 大綱

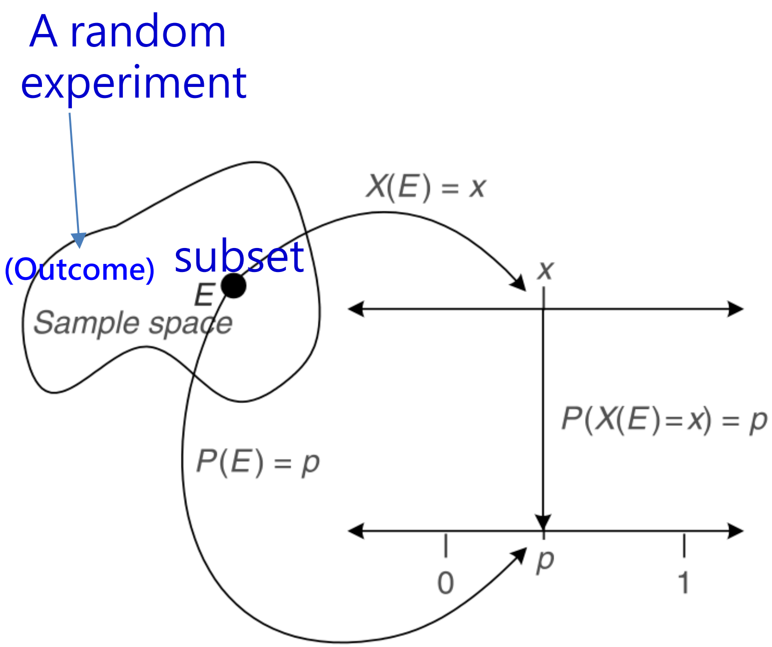
- **主題1**
  - 常見統計名詞
  - 機率分佈 (Probability distribution)
  - 累積機率分配函數 CDF (p)
  - 分位數 Quantiles (q)
- 主題2
  - 常見之分佈(二項式分佈、常態分佈)
  - 以常態機率逼近二項式機率
- 主題3
  - 大數法則 (LLN)
  - 中央極限定理 (CLT)
  - 用R程式模擬算機率

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

機率與隨機變數

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



Source: Statistics and Data with R

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

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

機率質量函數
Probability Mass Function

Formal definition

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

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

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

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

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

***例子.*投擲2顆公正的骰子**

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

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

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

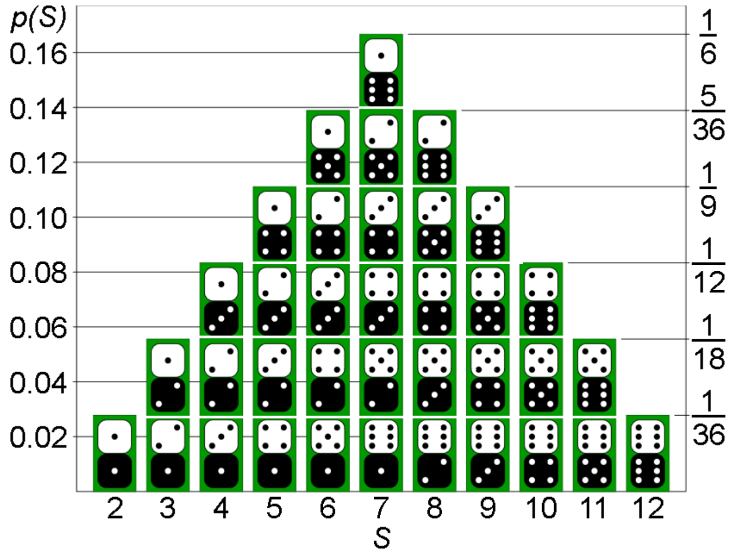
$k = 1,..,6.$

$S = X_1 + X_2$

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

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

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



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

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

機率密度函數
Probability Density Function

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

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

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

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

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

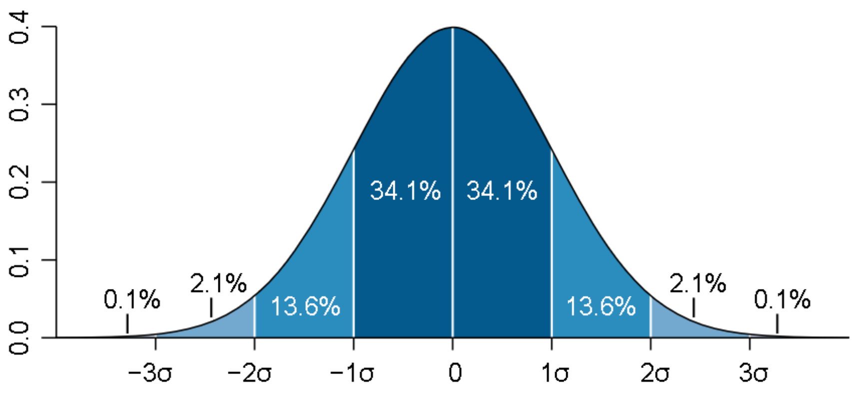
$$\mathrm{P}[a \le X \le b] = \int_a^b f(x)\,dx$$

The probability density of the normal distribution is:

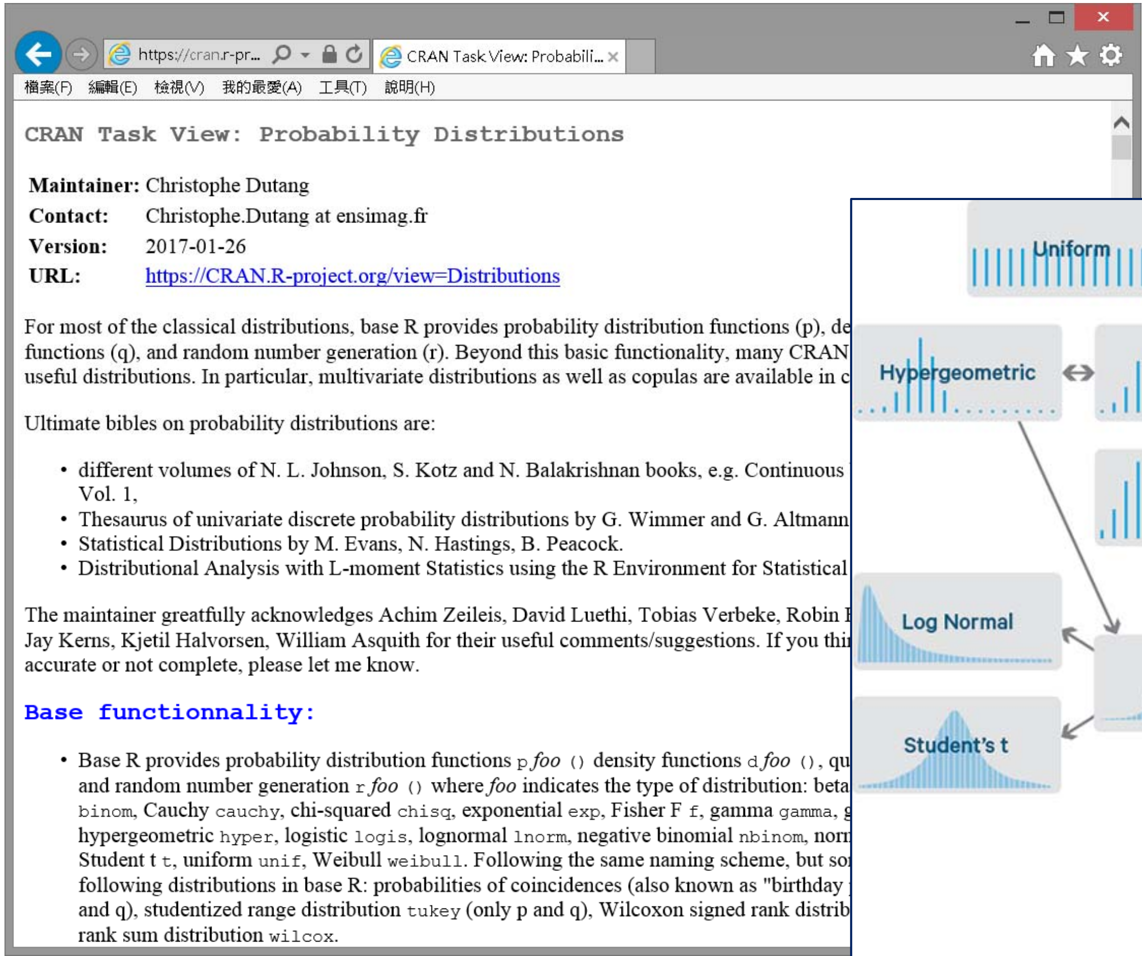
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

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



CRAN Task View: Probability Distribution



CRAN Task View: Probability Distributions

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

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

Ultimate bibles on probability distributions are:

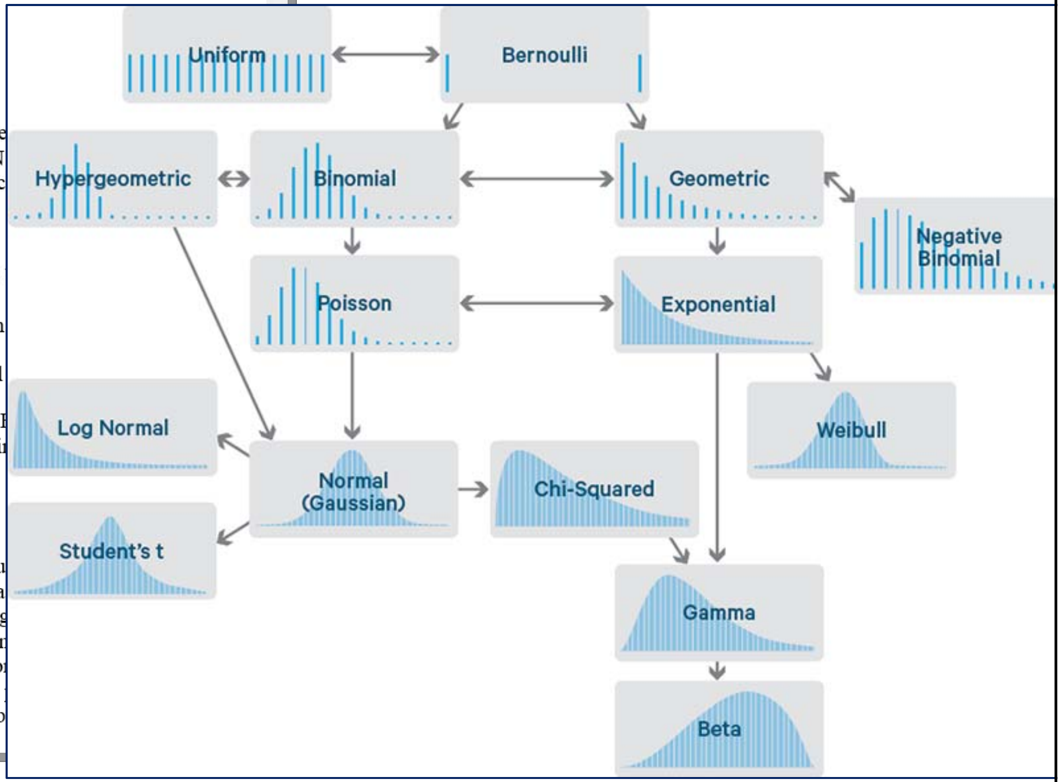
- different volumes of N. L. Johnson, S. Kotz and N. Balakrishnan books, e.g. Continuous Vol. 1,
- Thesaurus of univariate discrete probability distributions by G. Wimmer and G. Altmann
- Statistical Distributions by M. Evans, N. Hastings, B. Peacock.
- Distributional Analysis with L-moment Statistics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

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

Base functionality:

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

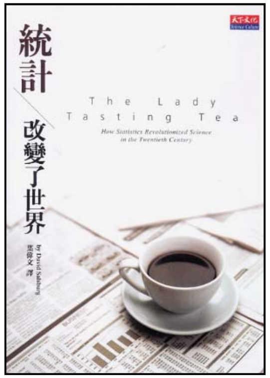
https://cran.r-project.org/web/views/Distributions.html



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

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

機率分佈在統計學中的重要性



統計改變了世界

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

統計革命的起點

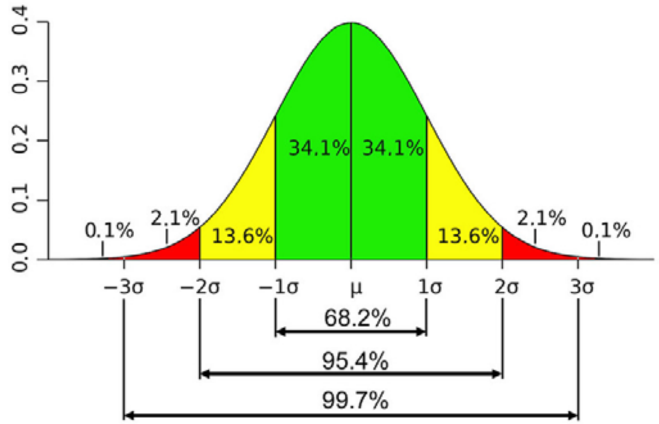
- Karl Pearson (1857-1936), 發表一系列和相關性(correlation)有關的論文,涉及動差、相關係數、標準差、卡方適合度檢定,奠定了現代統計學的基礎。
- 引入了統計模型的觀念: 如果能夠決定所觀察現象的**機率分佈的參數**,就可以了解所觀察現象的本質。

樣本變異數與樣本標準差

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

母體變異數與母體標準差

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



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

常用機率分佈的應用

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

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

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

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

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

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

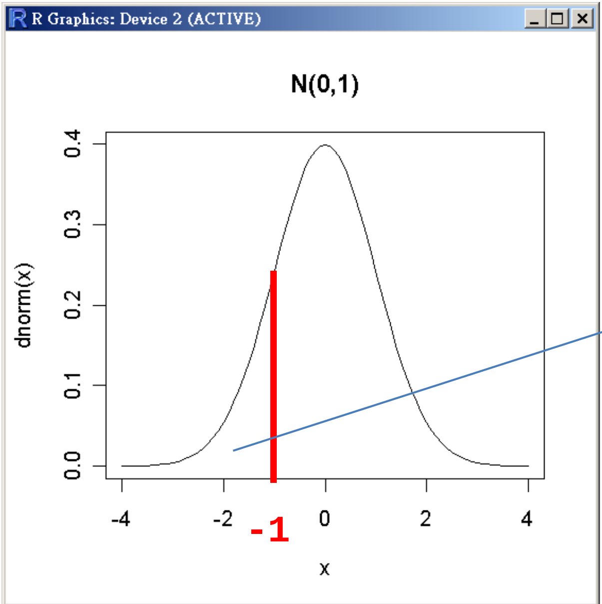
https://en.wikipedia.org/wiki/Probability_distribution

累積機率分配函數 CDF (p)

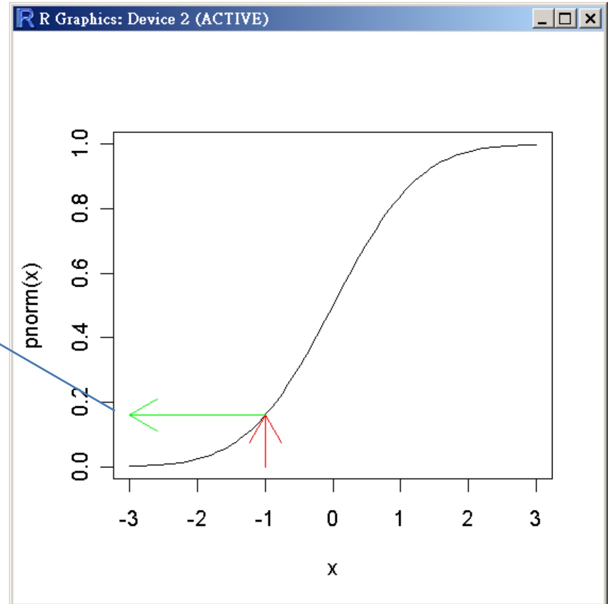
$$F_X(x) = P(X \leq x)$$

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

PDF



0.1586553



CDF

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

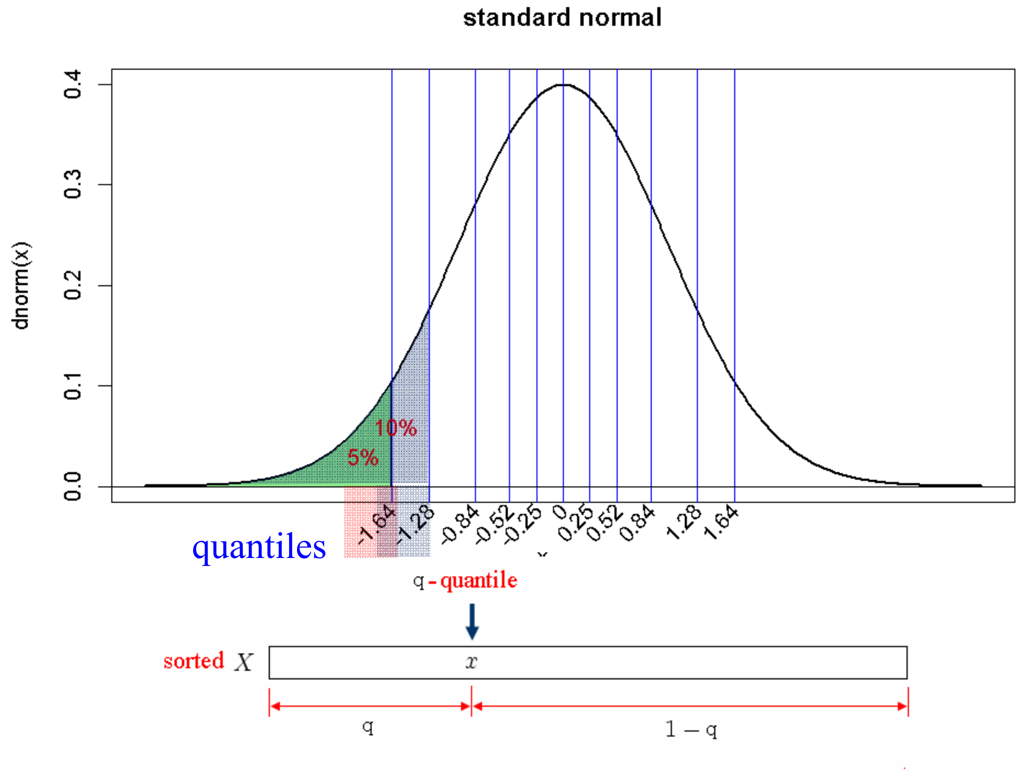
分位數 Quantiles (q)

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

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

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

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



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

常態母體平均數95%的信賴區間

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

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

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

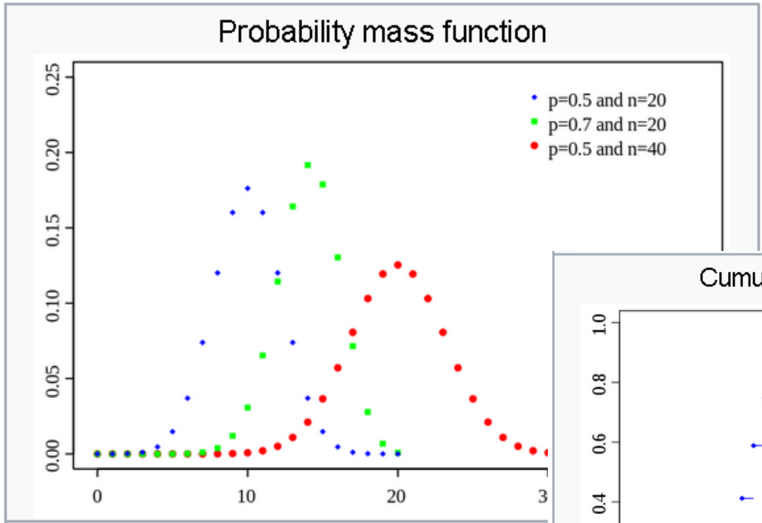
$$\Phi^{-1}(0.975)$$

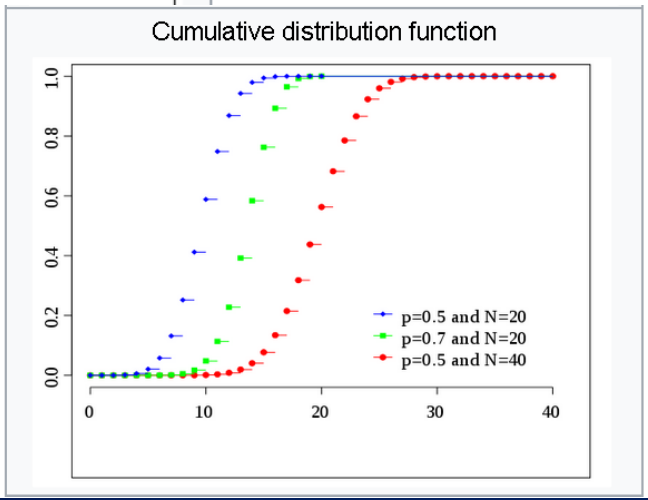
機率分佈 - 大綱

- 主題1
  - 常見統計名詞
  - 機率分佈 (Probability distribution)
  - 累積機率分配函數 CDF (p)
  - 分位數 Quantiles (q)
- 主題2
  - 常見之分佈(二項式分佈、常態分佈)
  - 以常態機率逼近二項式機率
- 主題3
  - 大數法則 (LLN)
  - 中央極限定理 (CLT)
  - 用R程式模擬算機率

二項式分佈 (Binomial)

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





Notation	$B(n, p)$
Parameters	$n \in \mathbf{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — 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。

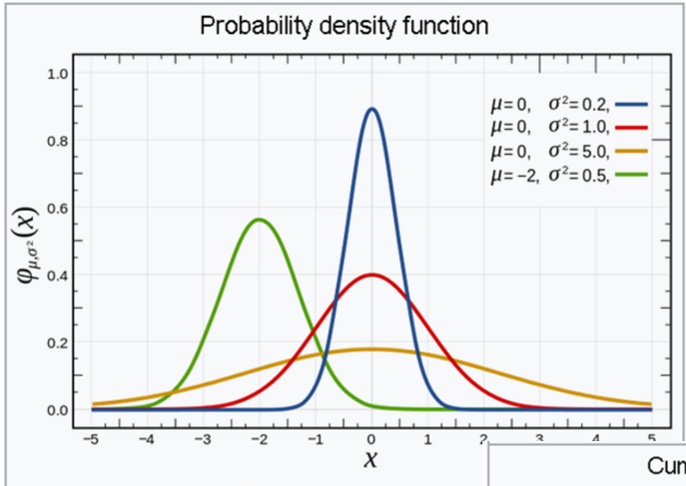
```
> pbinom(3, 10, 0.8)- pbinom(2, 10, 0.8)
[1] 0.000786432
```

- 已知累加機率值為0.1208，求對應的分位數。

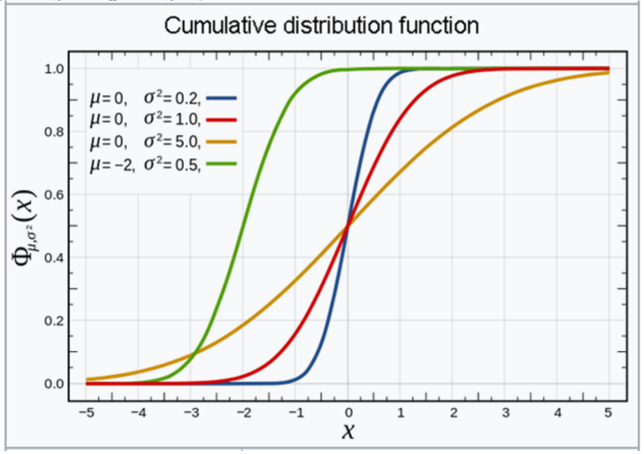
```
> qbinom(0.1208, 10, 0.8)
[1] 6
> pbinom(6, 10, 0.8)
[1] 0.1208739
```

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



The red curve is the *standard normal d*



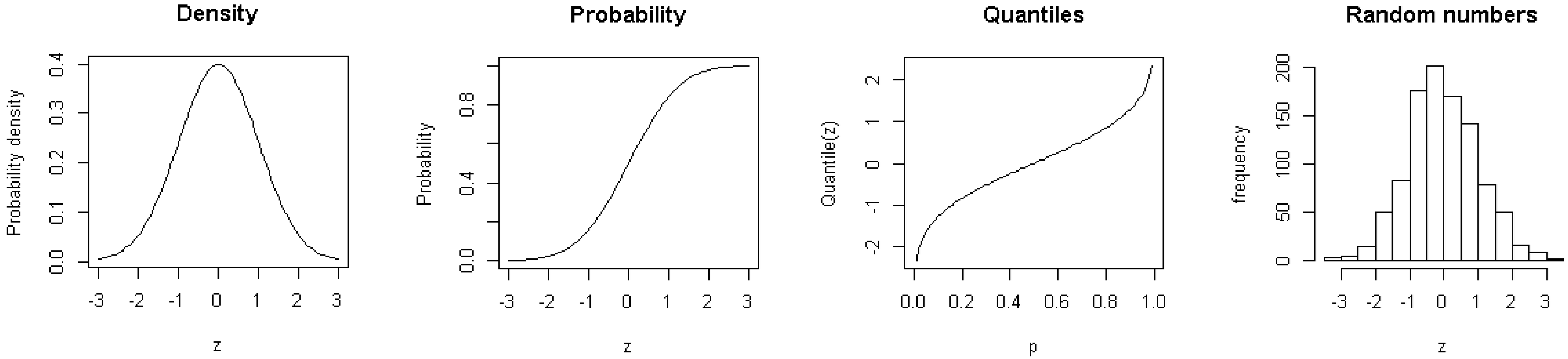
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
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

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

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



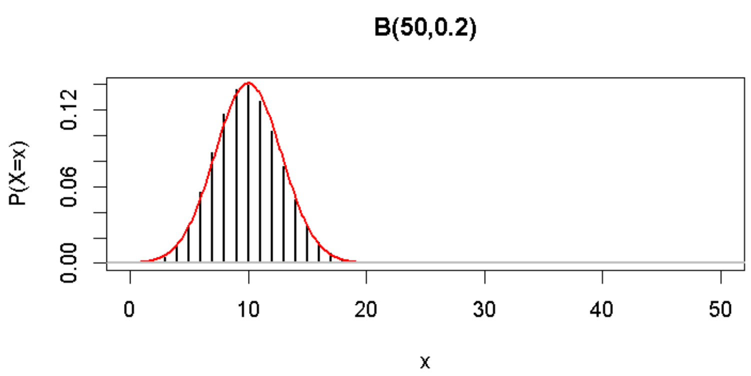

```
> par(mfrow=c(1,4))
> curve(dnorm, -3, 3, xlab="z", ylab="Probability density", main="Density")
> curve(pnorm, -3, 3, xlab="z", ylab="Probability", main="Probability")
> curve(qnorm, 0, 1, xlab="p", ylab="Quantile(z)", main="Quantiles")
> hist(rnorm(1000), xlab="z", ylab="frequency", main="Random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

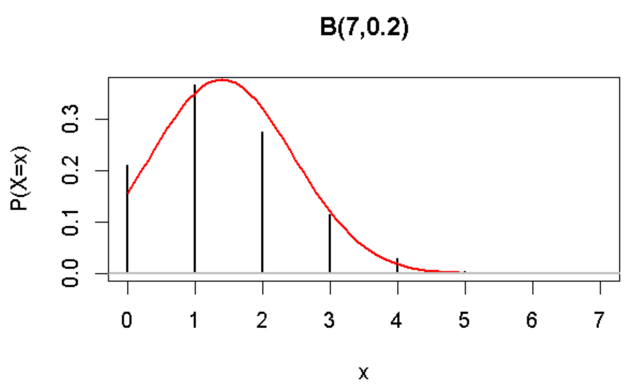
以常態機率逼近二項式機率

The normal approximation to the binomial

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

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





機率分佈 - 大綱

- 主題1
  - 常見統計名詞
  - 機率分佈 (Probability distribution)
  - 累積機率分配函數 CDF (p)
  - 分位數 Quantiles (q)
- 主題2
  - 常見之分佈(二項式分佈、常態分佈)
  - 以常態機率逼近二項式機率
- **主題3**
  - 大數法則 (LLN)
  - 中央極限定理 (CLT)
  - 用R程式模擬算機率

大數法則: The Law of Large Numbers

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

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

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

中央極限定理 (Central Limit Theorem)

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

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

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

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

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

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

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

- 在一般的統計實務上，大部分的應用中均假設當樣本大小為30(含)以上時 $\bar{X}_n$的抽樣分配即近似於常態分配。

- 當母體為常態分配時，不論樣本大小，樣本平均數的抽樣分配仍為常態分配。

應用CLT算機率

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

1.

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

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

2.

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

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

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

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

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

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

3.

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

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

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

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

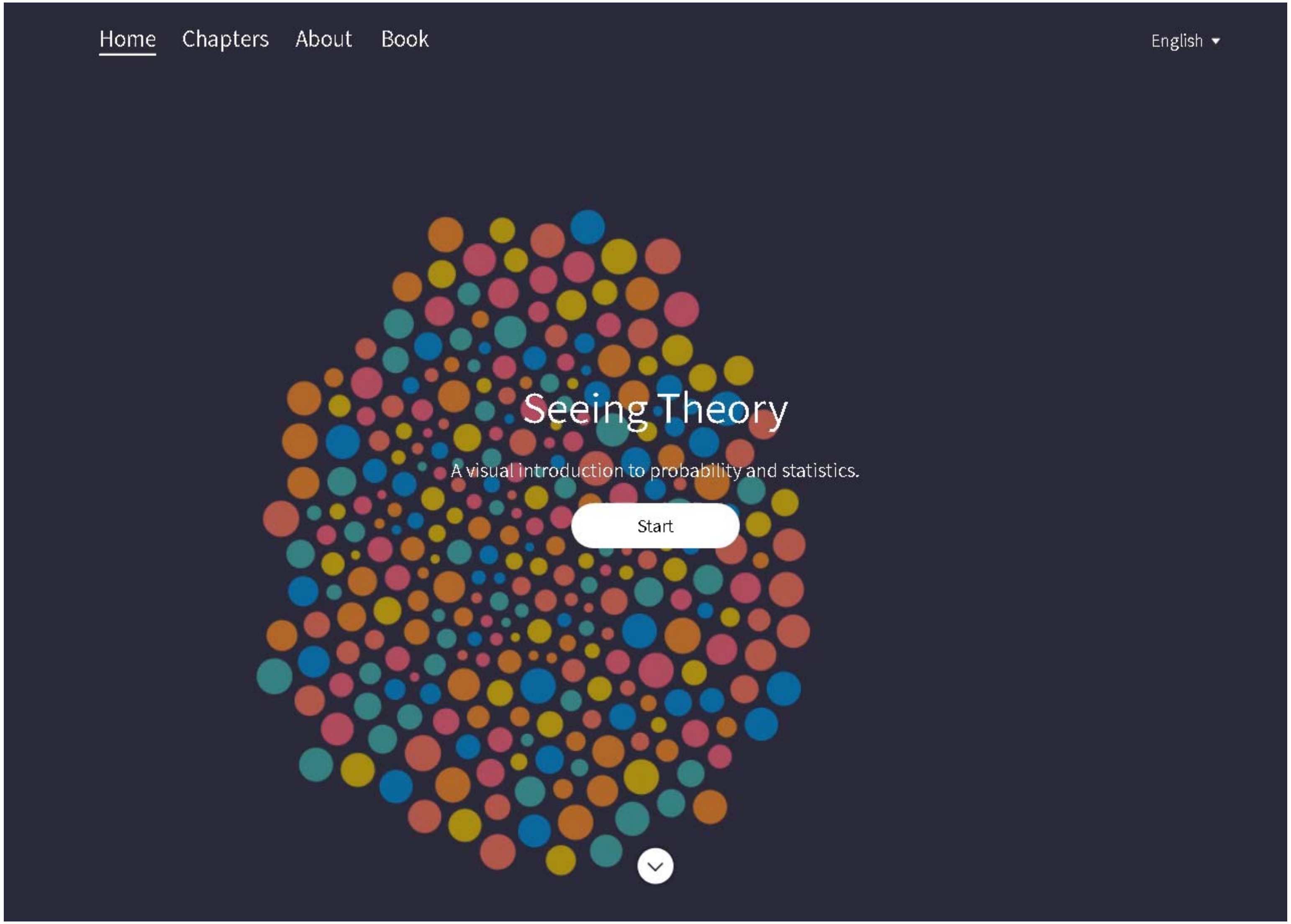
$$= P(Z > -3.16228)$$

$$= 0.99922$$

應用CLT

中央極限定理:
樣本平均之抽樣分佈

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



練習: 用R程式模擬算機率: 我們要生女兒

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

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

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

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

用R程式模擬算機率: 我們要生女兒

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

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

}
```

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