程式人《十分鐘系列》

用十分鐘瞭解

機率、統計、還有 R 軟體

陳鍾誠

2016 年 7 月 1 日

大學的時候

- 我就讀的《交大資訊科學系》有一門必修的《機率》課程
- 然後還有一門選修的《統計》課程。

我修了機率

- 但是沒有修統計！

結果

- 我只知道一堆機率分佈
- 但是卻不知道該怎麼用

後來

- 我修了通識開的一門統計課

修這門的原因是因為學校規定要修四門通識課才能畢業 XD

但是那門統計課

- 是採用《通識領域》的上法

結果

- 我還是不知道該怎麼用
- 像是《檢定》之類的方法，我還是不太瞭解！

然後

- 就這樣過了 25 年！

25 年後

- 我也是個大學老師了！

教育部的評鑑委員說

- 你們這個科系的數學課程太少，不符合大學的課程要求！

所以、系主任說

- 請各位老師多開數學課！

於是

- 我決定開一門《機率統計》

這樣

- 我就可以把之前沒學好的《機率統計》學好了！

為了學好《機率統計》

我決定發揮《程式人》的專長

就是用程式的方式學機率統計

於是我決定用 R 軟體

- 來幫助我學機率統計！

我發現

- R 軟體非常的適合用來學機率統計！

所以

- 我想我真的把機率統計學會了！

然後我一邊學一邊教

順便

- 還寫了一本電子書！

ccc.nqu.edu.tw/wd.html#st/home.wd
※ 陳鍾誠 / 電子書 / 機率統計
機率統計 -- 使用 R 軟體
書籍章節
簡報
第 1 章. 機率統計簡介
簡報
第 2 章. 機率的概念
第 3 章. 隨機變數
第 4 章. 機率分布
第 5 章. 期望值與動差生成函數
第 6 章. 聯合分布
第 7 章. 抽樣與敘述統計
第 8 章. 中央極限定理
第 9 章. 平均值的估計與檢定

現在

- 就讓我們來看看

  到底甚麼是《機率和統計》

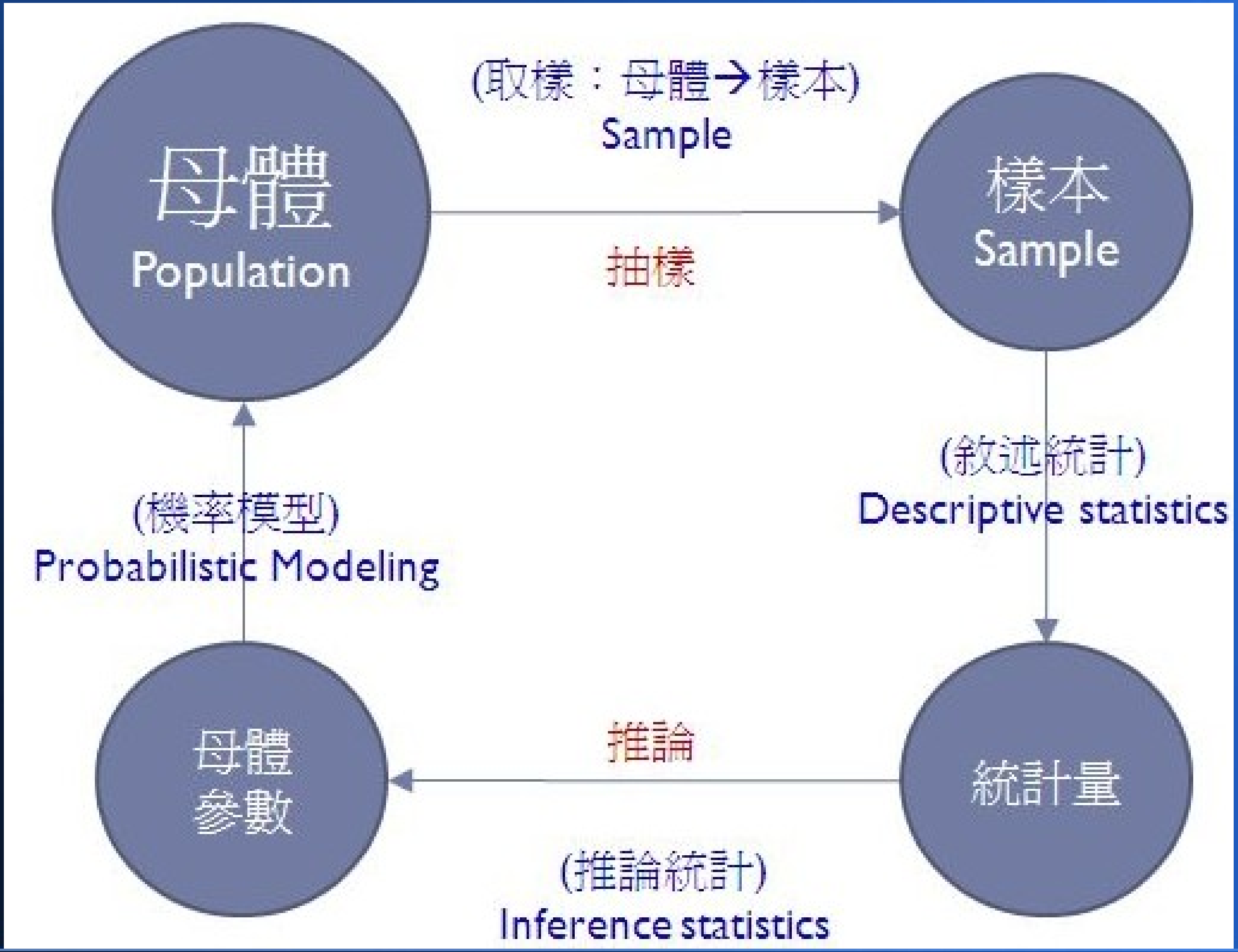
還有

- 如何用 R 軟體來學機率和統計吧！

首先

- 讓我們先看看，機率和統計之間的關係！

如下圖所示



(取樣：母體→樣本)
Sample
母體
Population
抽樣
樣本
Sample
(敘述統計)
Descriptive statistics
(機率模型)
Probabilistic Modeling
母體
參數
推論
統計量
(推論統計)
Inference statistics

機率和統計

- 其實是看待同一件事情的兩種不同角度！

機率

- 是在已知母體的情況下，研究樣本產生的情況！
- 而統計則通常是在母體未知的情況下，透過抽樣來研究母體到底長得甚麼樣？

這樣講

- 或許很難理解！

還是讓我們

- 發揮程式人的本色
- 用程式來解說《機率統計》好了！

假如、我們想從 1 到 100 裏

- 抽出 10 個樣本，可以用下列 R 指令

```
> x = sample(1:100, 10)
> x
 [1] 12 17 50 33 98 77 39 79  7 26
```

在 R 軟體裡

- 內建了很多機率分布，還有對應的抽樣函數
- 像是《均等分布、常態分佈、布瓦松分布、指數分布、二項分布、負二項分布、幾何分布...》等等。

我們可以透過這些函數

- 進行隨機抽樣！

以下是一些抽樣的範例

```
> rbinom(20, 5, 0.5)
 [1] 4 3 3 4 2 4 3 1 2 3 4 3 2 2 2 4 2 3 1 1
> rpois(20, 3.5)
 [1] 2 1 4 2 1 6 3 6 1 3 3 6 6 0 4 2 6 4 6 2
> runif(20, min = 3, max = 8)
 [1] 3.933526 3.201883 7.592147 5.207603 4.897806 3.848298 4.521461 4.437873
 [9] 3.655640 5.633540 6.557995 5.430671 6.502675 5.637283 7.713699 5.841052
[17] 6.859493 5.987991 3.752924 7.480678
> rnorm(20, mean = 5.0, sd = 2.0)
 [1] 6.150209 4.743013 3.328734 5.096294 4.922795 6.272768 4.862825 8.036376
 [9] 4.198432 5.467984 2.046450 6.452511 2.088256 5.349187 3.074408 3.628072
[17] 3.421388 7.242598 3.125895 9.865341
> rexp(20, rate=2.0)
 [1] 0.17667426 0.49729383 0.12786107 0.13983412 0.44683515 1.30482842
 [7] 0.28512544 1.61472266 0.23220649 0.39089780 0.05947224 1.42892610
[13] 0.02555552 0.69409186 0.68228242 0.22542362 0.33590791 0.14684937
[19] 0.34995146 0.80595369
```

rbinom：二項分布，正面機率 0.5 ，每次抽 5 個，會有幾個正面 (共抽 20 組)

rpois：布瓦松分布， lambda=3.5 ，抽 20 次

runif：3 到 8 之間的均等分布，抽 20 個樣本

rnorm：常態分佈 (平均值 5, 標準差 2) ，抽 20 個樣本

rexp：指數分佈 (參數為 2) ，抽 20 個樣本

但是、這樣的抽樣

- 我們只看到一堆數字
- 到底這些數字代表甚麼呢？

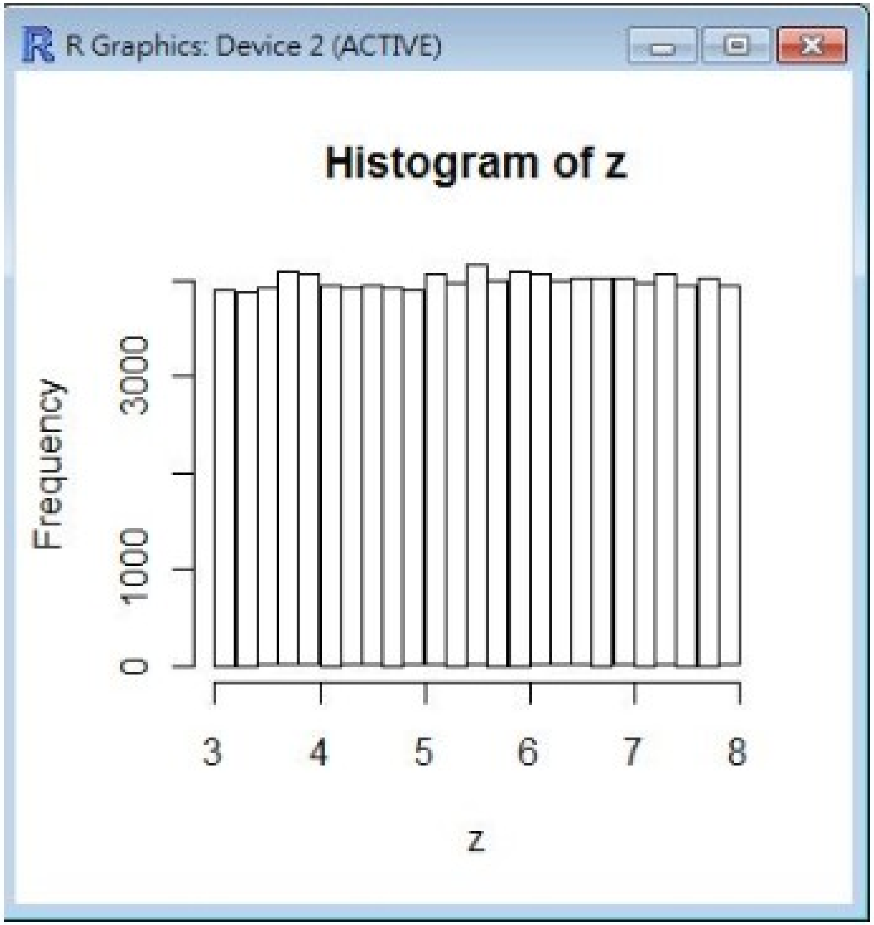
讓我們進一步用程式來畫圖

- 會比較知道這些分布的樣子

舉例而言

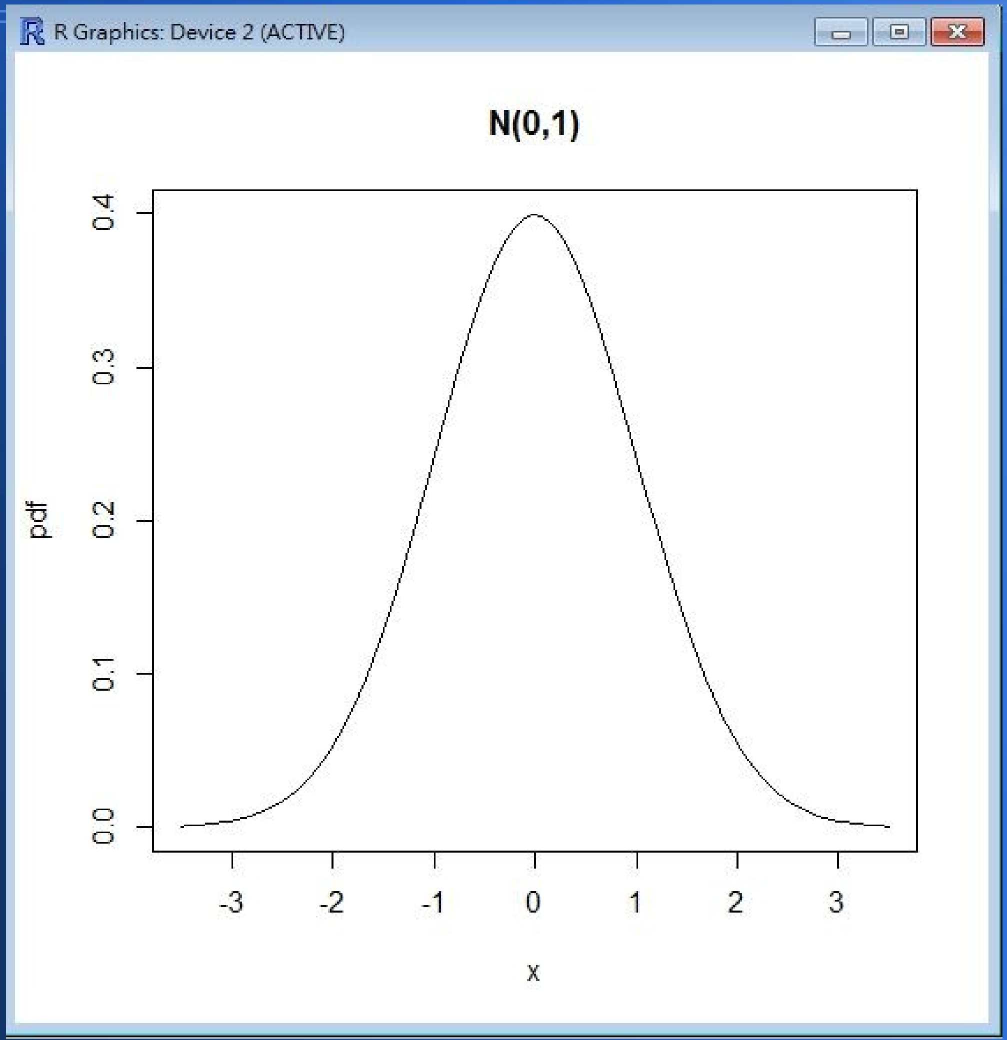
- 右邊是均等分布抽十萬個的次數統計圖(histogram)
- 你可看到每個區域的分布都很均勻。
- 所以才叫均等分布

```
> z = runif(100000, min=3, max=8)
> hist(z)
```



接著讓我們看看常態分佈

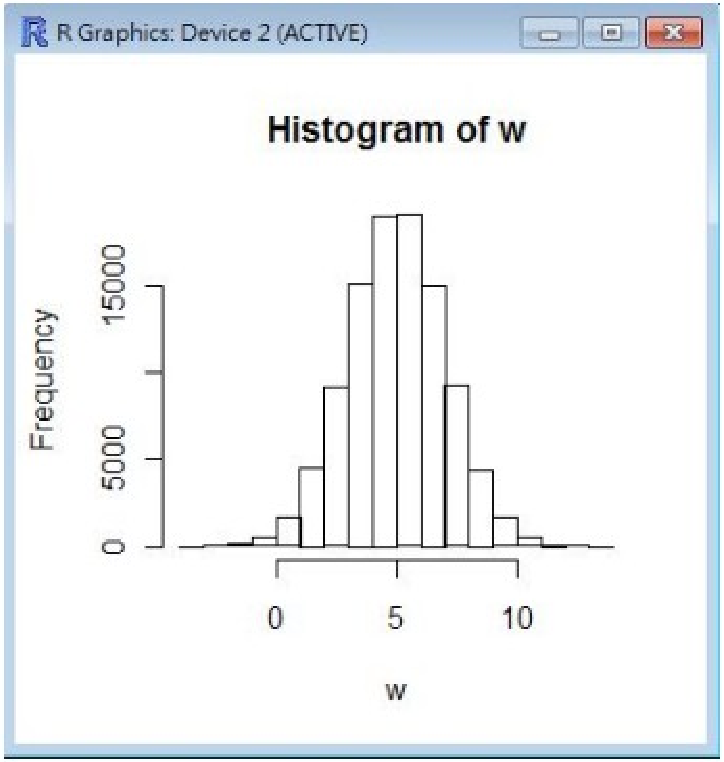
- 右圖是標準常態分佈的機率密度函數 (Probabilistic Density Function, PDF)



如果我們對常態分佈抽樣

- 結果當然也會很常態
- 像是右圖是對平均值 5, 標準差 2 的常態分佈抽十萬個樣本的統計圖形

```
> w = rnorm(100000, mean=5.0, sd=2.0)
> hist(w)
```



一個很直覺的想法是

- 樣本從什麼分布抽出來，《次數統計圖》（histogram）就會長得和該分布很像。
- 基本上這是對的！

如果想大概瞭解

- 該分布到底長得甚麼樣子
- 除了畫圖以外，還可以算出一些數字來讓我們大概理解該分布的外貌！

這些大概的數字

- 就是《敘述統計》裡的那些數字
- 像是《平均值、中位數、標準差、四分位數、最大最小值》等等。

假如我們不知道母體分布

- 那麼《敘述統計》就可以提供一些基本的線索！

但是、統計的力量不止於此

除了《敘述統計》之外

- 更強大的是推論統計！

而《推論統計》的關鍵

- 則是《中央極限定理》！

要理解中央極限定理

- 可以用數學
- 也可以用程式

傳統的統計課程

- 都會教你用《數學》來理解《中央極限定理》

但我是教程式的老師

- 所以打算用程式來教你
  《中央極限定理》！

首先、讓我們寫個 R 程式

```
CLT = function(x)  {
  op<-par(mfrow=c(2,2)) # 設為 2*2 的四格繪圖版
  hist(x, nclass=50)     # 繪製 x 序列的直方圖 (histogram)。
  m2 <- matrix(x, nrow=2, )     # 將 x 序列分為 2*k 兩個一組的矩陣 m2。
  xbar2 <- apply(m2, 2, mean)     # 取每兩個一組的平均值 (x1+x2)/2 放入 xbar2 中。
  hist(xbar2, nclass=50)     # 繪製 xbar2 序列的直方圖 (histogram)。
  m10 <- matrix(x, nrow=10, )     # 將 x 序列分為 10*k 兩個一組的矩陣 m10。
  xbar10 <- apply(m10, 2, mean)     # 取每10個一組的平均值 (x1+..+x10)/10 放入 xbar10 中。
  hist(xbar10, nclass=50)     # 繪製 xbar10 序列的直方圖 (histogram)。
  m20 <- matrix(x, nrow=20, )     # 將 x 序列分為 25*k 兩個一組的矩陣 m25。
  xbar20 <- apply(m20, 2, mean)     # 取每20個一組的平均值 (x1+..+x20)/20 放入 xbar20 中。
  hist(xbar20, nclass=50)     # 繪製 xbar20 序列的直方圖 (histogram)。
}
```

這個程式

- 會畫出 1 個、2 個、10 個、20 個樣本的平均值之分布圖。

然後、我們就可以用下列程式

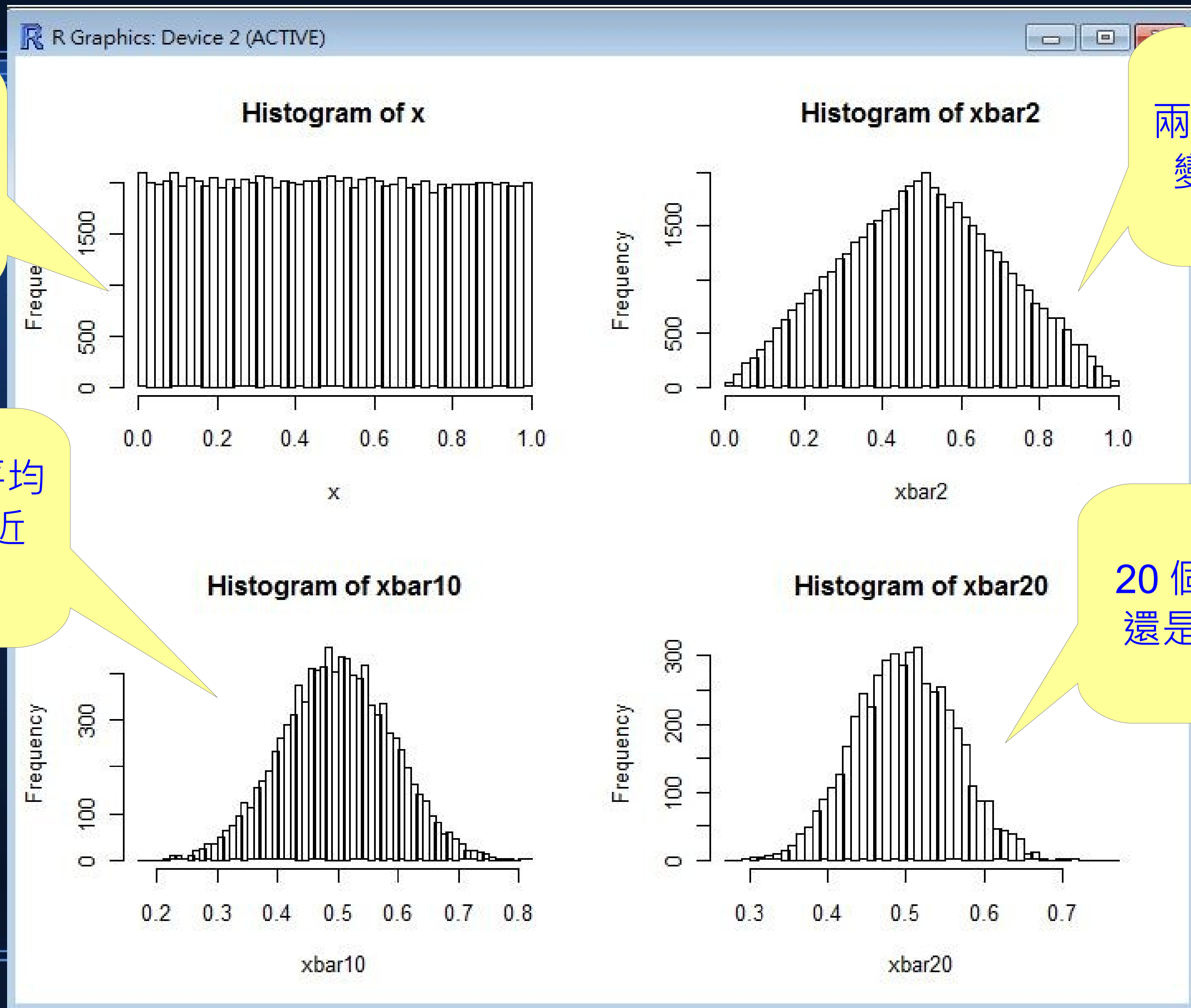
- 來觀察《中央極限定理》到底是甚麼意思

```
CLT(rbinom(100000, 20, 0.5)) # 用參數為 n=20, p=0.5 的二項分布驗證中央極限定理。
CLT(runif(100000)) # 用參數為 a=0, b=1 的均等分布驗證中央極限定理。
CLT(rpois(100000, 4)) # 用參數為 lambda=4 的布瓦松分布驗證中央極限定理。
CLT(rgeom(100000, 0.5)) # 用參數為 n=20, m=10, k=5 的超幾何分布驗證中央極限定理。
CLT(rhyper(100000, 20, 10, 5)) # 用參數為 p=0.5 的幾何分布驗證中央極限定理。
CLT(rnorm(100000)) # 用參數為 mean=0, sd=1 的標準常態分布驗證中央極限定理。
CLT(sample(1:6, 100000, replace=T))    # 用擲骰子的分布驗證中央極限定理。
CLT(sample(0:1, 100000, replace=T))    # 用丟銅板的分布驗證中央極限定理。
```

你會發現、不管母體長什麼樣子

- 只要樣本數愈多，其平均值就會愈來愈接近常態分佈！
- 而且通常 20 個樣本以上就會非常接近常態分佈了。

舉例而言、以下是均等分布的執行結果



單一樣本
是平的

兩個樣本的平均
變成金字塔狀

十個樣本的平均
已經非常接近
常態分佈

20 個樣本平均
還是常態分佈

這種多樣本平均

- 會趨向常態分佈的現象
  - 就是《中央極限定理》

其數學式可以寫成

- 中央極限定理：n 個樣本的平均值會趨向常態分佈

$$\frac{x_1+x_2+\ldots+x_n}{n} = \bar{x} \to N(\mu, \sigma/\sqrt{n})$$

n 個樣本的平均

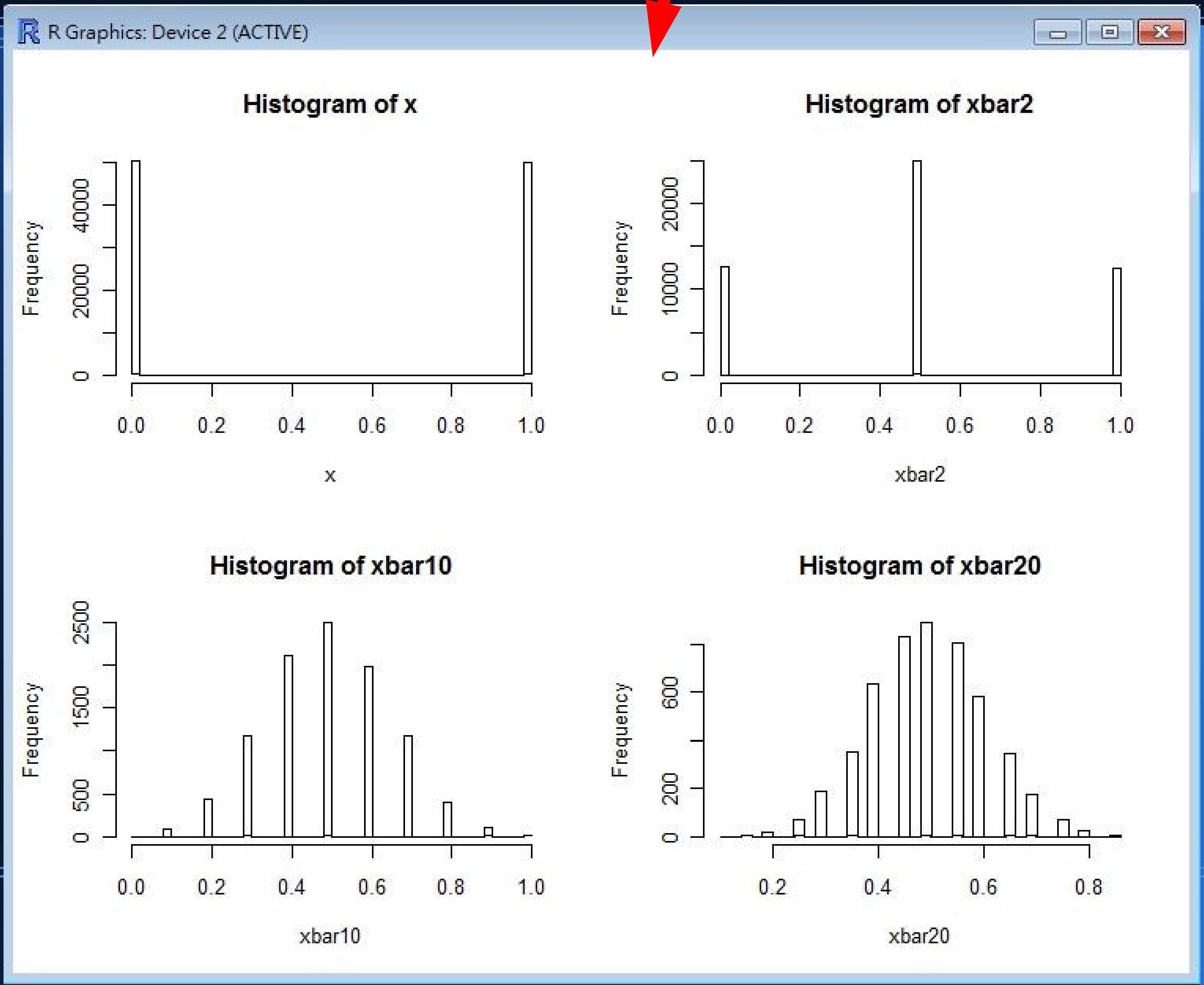
會趨向常態分佈

而且這個常態分佈，還會隨 n 增加而變窄

$$\frac{x_1+x_2+\ldots+x_n}{n} = \bar{x} \to N(\mu, \sigma/\sqrt{n})$$

更精確一點的說，當您從某個母體 X 取出 n 個樣本，則這 n 個樣本的平均數 $\frac{x_1+x_2+\ldots+x_n}{n} = \bar{x}$ 會趨近於以平均期望值 μ 為中心， 以母體標準差 σ 除以 $\sqrt{n}$ 的值 $\sigma/\sqrt{n}$ 為標準差的常態分布。

所以、不管母體是骰子還是銅板

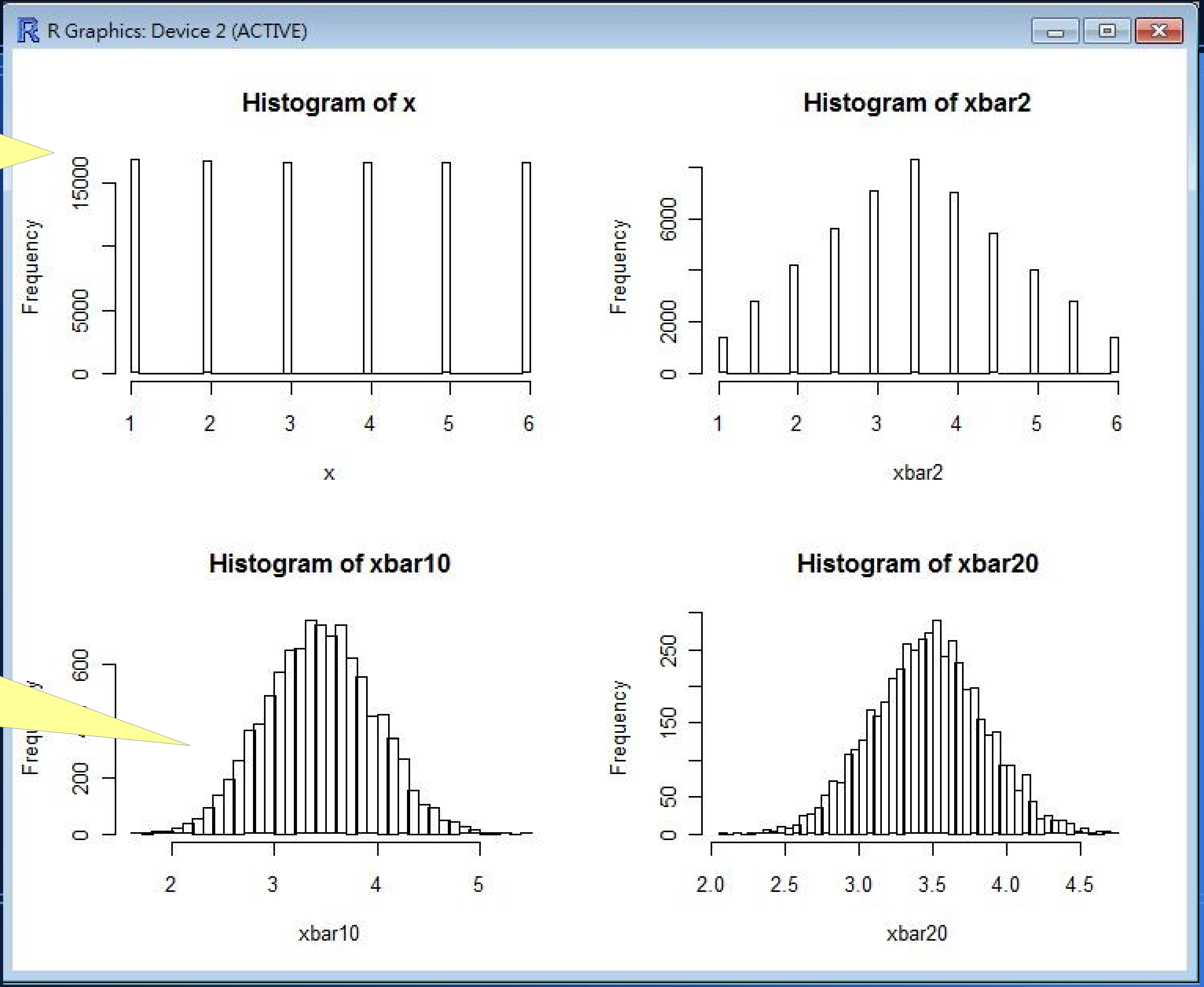


多個樣本的平均值都會趨向常態分佈

骰子

1 到 6 點

十個樣本的平均值
就非常常態了



換句話說

- 只要是前後無關的隨機樣本
- n 個樣本的平均值都會趨向常態分佈
- 只要 n 大一點就行了！
- 而且 n 愈大，標準差就越小

仔細想想

- 你會發現這是一個非常強大的定理！

為甚麼很強大？

因為只要幾十個樣本

- 通常就可以很準確地預測母體的平均值。

不過、這個定理還有一點點缺陷

那個缺陷就是

- 對於非數學化的母體而言
- 我們通常不知道母體的標準差！

這樣、我們就不能套用

- 中央極限定理的公式了！

$$\frac{x_1+x_2+\ldots+x_n}{n}=\bar{x}\rightarrow N(\mu,\sigma/\sqrt{n})$$

σ 未知？

為了處理這個問題

- 英國在酒廠工作的 Willam S. Gosset 於 1908 年提出了《t 分布》，可以用來修正常態 N 分布在 σ 未知時難以套用中央極限定理的問題。

T 分布的想法是

- 改用樣本標準差 S_n 來替代母體標準差 σ

$$S_n^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

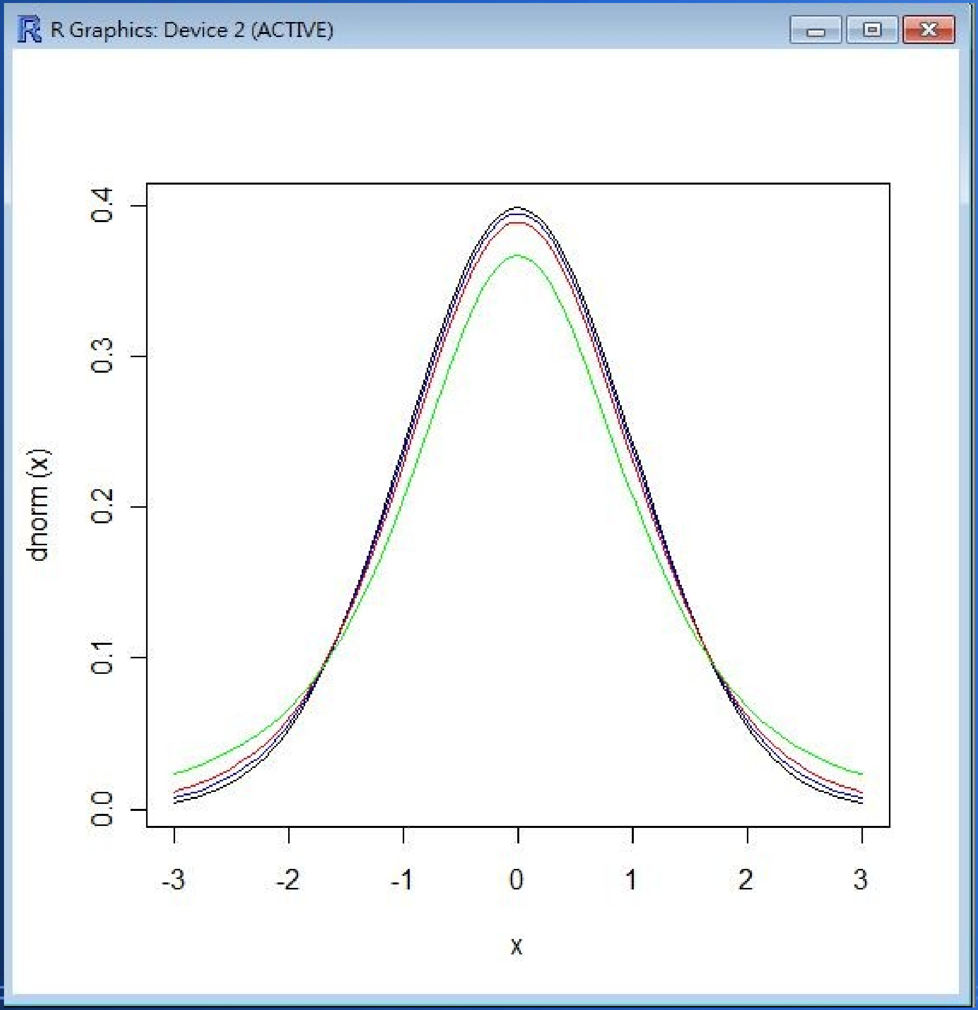
- 於是標準常態分佈 Z 就換成了 T 分布

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

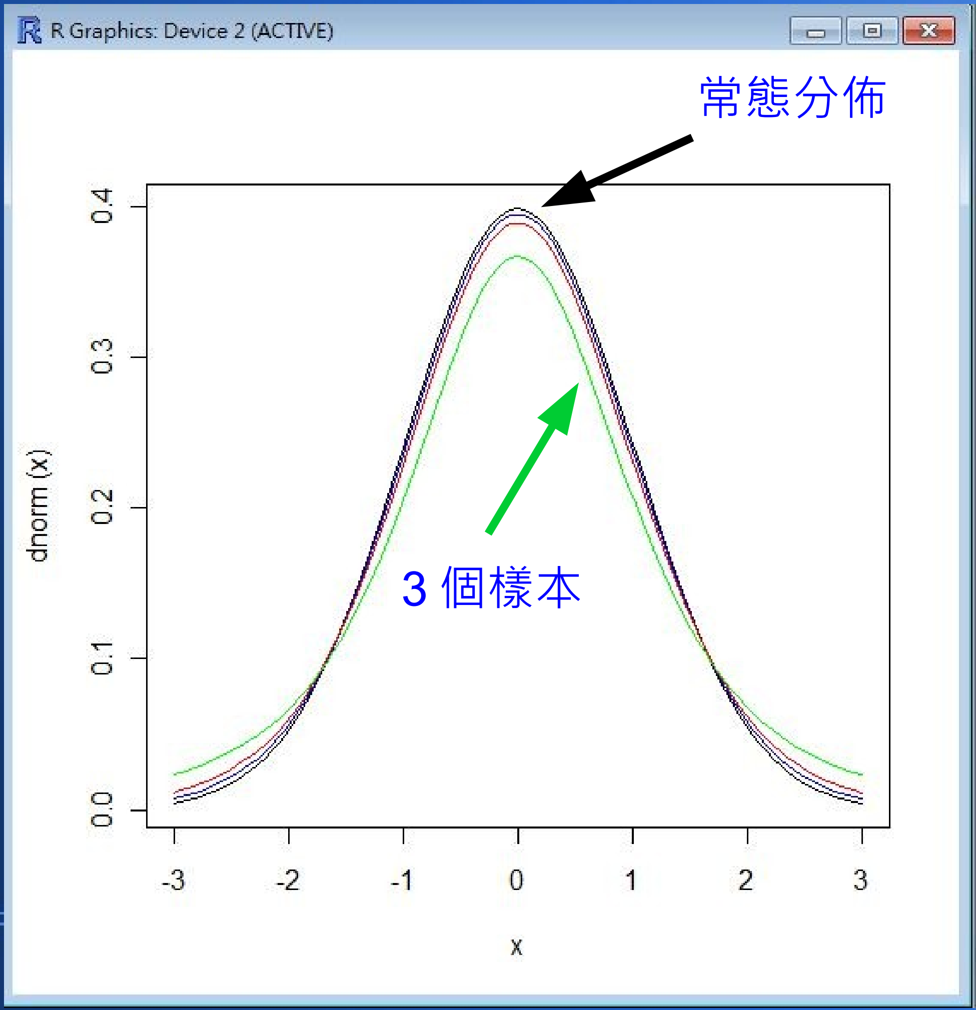
T 分布的樣子如下

- 由於使用了 S_n 樣本標準差
- 所以樣本數 n 不同就會有不同的分佈線



樣本愈多就越接近常態分佈

```
> curve(dnorm, from=-3, to=3, col="black")
> curve(dt(x, df=25), from=-3, to=3, add=T, ylab="T25", col="blue")
> curve(dt(x, df=10), from=-3, to=3, add=T, ylab="T10", col="red")
> curve(dt(x, df=3), from=-3, to=3, add=T, ylab="T3", col="green")
```



有了 T 分布之後

- 既使不知道母體標準差，我們也能很有信心的透過 n 個樣本來估計母體的平均值。
- 因為我們可以用樣本標準差 Sn 取代母體標準差 σ

$$Z=\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}$$

母體標準差

$$T=\frac{\overline{X}_n-\mu}{S_n/\sqrt{n}}$$

樣本標準差

但是有個條件

- 就是樣本之間必須是獨立的
- 抽樣必須要是《隨機抽樣》

只要確保這些條件

- 我們就能使用 t 分布來估計了！

有了中央極限定理和 T 分布

我們還需要什麼大數據呢？

只要幾十個樣本

- 就可以估計的不錯了！

不夠的話

- 就用幾百或上千個樣本
- 縮小平均值 $\overline{X}$ 的標準差就好了！

於是我們可以

- 用抽樣來檢定母體平均數

然後回答下列問題

習題一、以下 x 是某隨機樣本序列，請回答下列問題

```
x = c(46.26534, 49.30766, 53.79364, 53.18000, 48.97584, 51.92664,
44.58280, 62.26655, 54.52493, 55.08502, 56.78329, 45.00972, 46.99871, 43.88388, 52.63184, 53.15600, 48.39374, 51.07595, 47.36923, 52.09186,
46.54074, 54.46617, 47.87038, 42.94228, 48.69307)
```

1. 請問母體平均值 mu 的 95% 信賴區間為何？
2. 請問母體平均值 mu 的 98% 信賴區間為何？
3. 請用 mu=50 檢定該平均值 (a) 請問該檢定的虛無假設為何？(b) 請問該檢定的對立假設為何？(c) 請問顯著性 p-value 是多少？(d) 請問您認為 mu=50 這個虛無假設是否應該被否決？為甚麼？(e) 請問您認為 mu 不等於 50 這個對立假設是否應該被接受？為甚麼？

或者下列問題

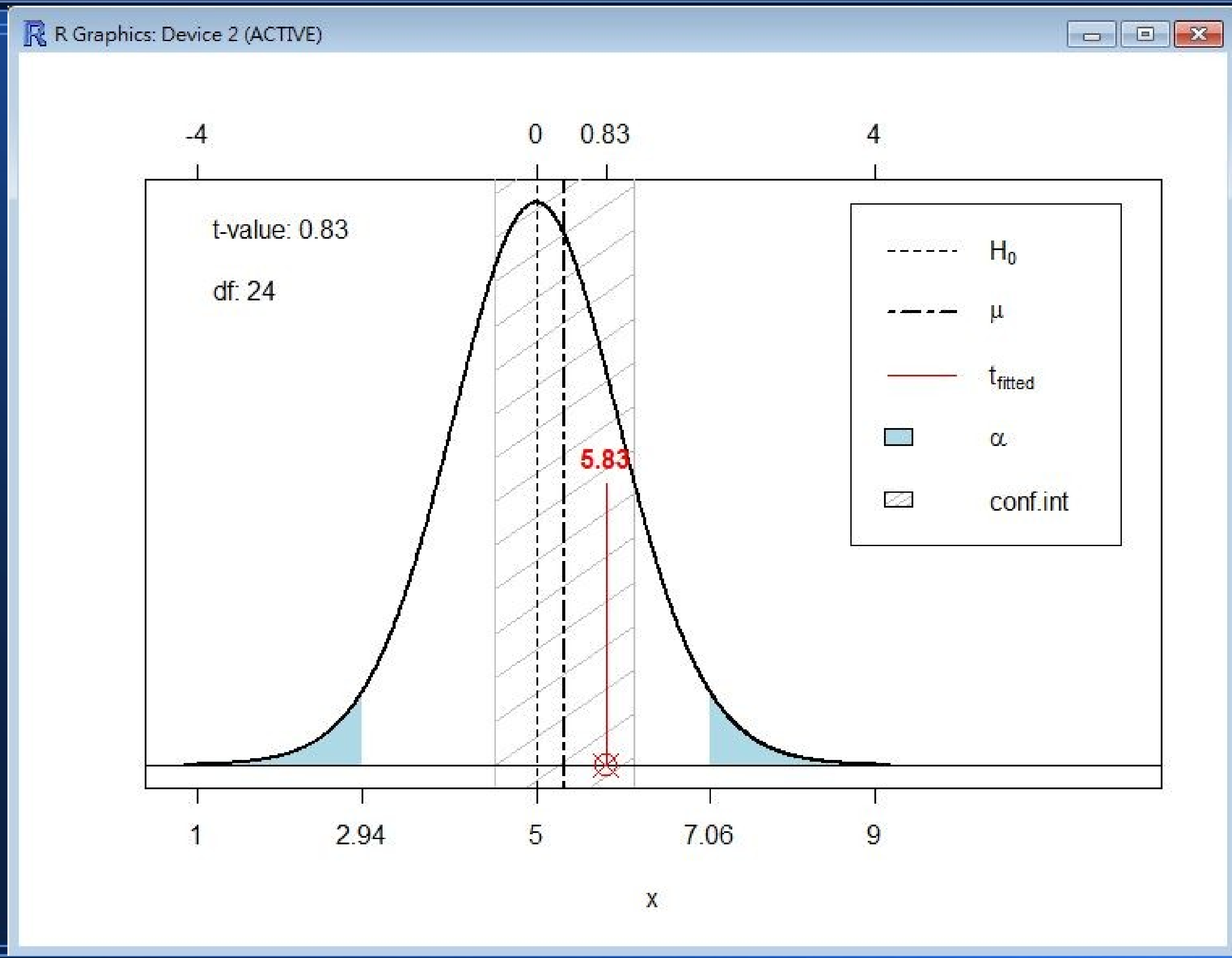
習題二、請用下列方式產生樣本 x，然後回答下列問題

```
mu=runif(1, 0, 10)
sd1 = runif(1, 1,2)
x=rnorm(25, mean=mu, sd=sd1)
x
```

1. 請問母體平均值 mu 的 95% 信賴區間為何？
2. 請問母體平均值 mu 的 98% 信賴區間為何？
3. 請用 mu=5 檢定該平均值 (a) 請問該檢定的虛無假設為何？(b) 請問該檢定的對立假設為何？(c) 請問顯著性 p-value 是多少？(d) 請問您認為 mu=5 這個虛無假設是否應該被否決？為甚麼？(e) 請問您認為 $mu \neq 5$ 這個對立假設是否應該被接受？為甚麼？

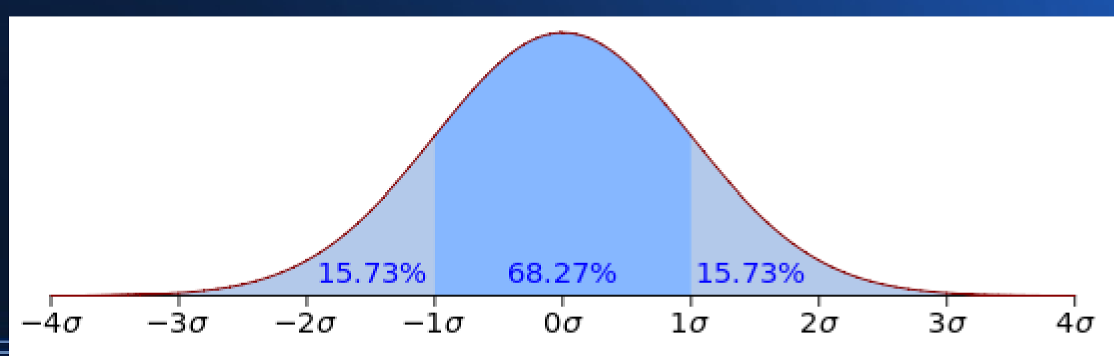
因為落在下圖藍色區域的機率很小

- 所以我們的估計值不太容易落在差很遠的藍色區域



只要在兩個標準差的範圍內就可以達到 95% 的信賴區間

1. $P[-\sigma < X - \mu < \sigma] = 0.68$
2. $P[-2\sigma < X - \mu < 2\sigma] = 0.95$
3. $P[-3\sigma < X - \mu < 3\sigma] = 0.997$
4. $P[-4\sigma < X - \mu < 4\sigma] = 0.99993$
5. $P[-5\sigma < X - \mu < 5\sigma] = 0.9999994$
6. $P[-6\sigma < X - \mu < 6\sigma] = 0.999999998$



然後我們就可以用 t 分布來檢定

```
> t.test(x, mu=8)

        One Sample t-test

data:  x
t = 0.3612, df = 24, p-value = 0.7211
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.205820 9.131145
sample estimates:
mean of x
 8.168483
```

檢定母體平均值 mu 是否為 8

自由度 24 代表有 25 個樣本

95% 信賴區間

x 的樣本平均值為 8.168483

於是我們可以玩玩猜數字遊戲

- 去猜某個分布的母體平均數

mu= μ 值到底是多少？

像是這樣

```
> mu=runif(1, 0, 10)
> sd1=runif(1, 1, 2)
> x=rnorm(25, mean=mu, sd=sd1)
> t.test(x, mu=5)

        One Sample t-test

data:  x
t = -8.5779, df = 24, p-value = 8.985e-09
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 2.901134 3.715254
sample estimates:
mean of x
 3.308194
```

母體平均 mu 值是 0 到 10 之間的一個亂數

母體標準差 sd1 是 1 到 2 之間的一個亂數

用上述參數進行常態分佈抽樣 25 個

然後進行 t 檢定

然後進行 t 檢定，看看 mu 是否為 5

P 值很小，代表 mu 幾乎不可能為 5

95% 信賴區間範圍

樣本平均數 $\bar{x}$ 為 3.308194

既然上述檢定已經告訴我們 $\bar{x}$ 為 3.308194

- 那麼 mu 應該不會離 $\bar{x}$ 太遠
- 以本例而言母體的 mu 為 3.356528

```
> mu
[1] 3.356528
> t.test(x, mu=3.3)

        One Sample t-test

data:  x
t = 0.0415, df = 24, p-value = 0.9672
alternative hypothesis: true mean is not equal to 3.3
95 percent confidence interval:
 2.901134 3.715254
sample estimates:
mean of x
 3.308194
```

透過這種 T 檢定

- 我們就可以很容易的推測母體平均數 mu 值的範圍。

並且可以很有信心

- 因為落在範圍外的機率可以設得很小（像是5%）

但前提是

- 樣本必須要《互相獨立》
- 而且必須是從母體中《隨機抽樣》出來的！

當然、我們不只可以檢定平均值

- 還可以檢定《標準差》（變異數）
  - 只是要改用 F 分布

另外還有

- 單樣本、雙樣本、成對樣本等等檢定方式！

您必須小心選用

- 不同的情況必須採用不同的分布去檢定

4 統計法比較
  - 4.1 單樣本*t*檢驗
  - 4.2 配對樣本*t*檢驗
  - 4.3 獨立雙樣本*t*檢驗
    - 4.3.1 樣本數及變異數相等
    - 4.3.2 樣本數不相等但變異數相等
    - 4.3.3 變異數皆不相等
  - 4.4 簡單線性迴歸之斜率

只要選擇對的檢定方法

- 而且確定樣本的隨機性
- 那麼不需要很多樣本，就可以得到《值得信賴》的信賴區間！

方法很簡單，只要使用下列 R 函數就行了

t.test {stats}

Student's t-Test

Description

Performs one and two sample t-tests on vectors of data.

Usage

```
t.test(x, ...)

## Default S3 method:
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)

## S3 method for class 'formula'
t.test(formula, data, subset, na.action, ...)
```

你可以用單樣本 t 檢定去檢驗

- 某湖水中的平均細菌數量
- 燈泡或機器的平均壽命
- 選舉的投票率或得票率

```
t.test(x, mu=μ, conf.level = 0.95)
```

或者用雙樣本 t 檢定去檢驗

- 兩台機器的加工精度
- 兩種飼料讓豬長大的速率
- 兩個湖水的某種細菌數量

```
t.test(x, y, var.equal=TRUE)
```

或者用成對 t 檢定去檢驗

- 攝氏 70 度與 80 度時某元件斷裂強度是否有差異
- 某班對某主題第二次考試的成績是否比第一次考試進步
- 同一人在服用某維生素後是否比較不容易感冒。

```
t.test(x, y, paired=TRUE)
```

當然、可以檢定的數值

- 並不只限定於平均值
- 還有《標準差》與《比例》等等，也都可以進行檢定。

雖然、還是有些情況

- 少樣本的檢定會偏離太遠！

舉例而言、假如要估計平均財產

- 如果全世界有一百億人，比爾蓋茲的財產佔全世界財產總額的 90%
- 那麼我們的抽樣，只要漏掉比爾蓋茲就會有嚴重的偏離。
- 但是樣本很少卻抽到比爾蓋茲，也一樣會有嚴重的偏離。

像是上述比爾蓋茲的案例

- 我們是很難透過抽樣進行正確估計的
- 還好這種情況並不是很常見！

另外、在統計的實務上

- 會遭遇到很多困難點
- 主要的困難點是《抽樣很難做到完全隨機》

關於這點我們在上次的十分鐘系列

- 《用十分鐘瞭解關於論文的那些事兒》

  當中有提到過！

http://www.slideshare.net/cccmit/ss-63325373

很多碩士論文

- 都會犯下這類因《抽樣不隨機》而導致的統計錯誤

舉例而言

- 當你做問卷調查時，如果用網路問券，那抽到的就只有上網的人，而且為了鼓勵人家來填問卷，動用了親朋好友的關係。
- 這樣就沒辦法做到真正的隨機抽樣，會有嚴重的偏差！

如果你用紙本問卷調查

- 通常要發獎品鼓勵人家來填
- 於是發糖果就一堆小朋友來
- 發日用品又一堆歐巴桑來填
- 最後還是完全走樣！

如果你用電話調查

- 通常一打過去說要調查，人家就掛電話。
- 最後調查到的都是閒閒在家沒事幹想找人聊天的那種
- 這樣的調查仍然離《隨機抽樣》非常的遠！

但是、很多人為了畢業

- 或者為了升等，還是常常去做這種完全不隨機的調查
- 最後還不懂如何設計才能排除亂填或重複的問卷！

以上這些都是

- 抽樣不隨機所導致的統計錯誤。

所以在設計抽樣方法時

- 必須想辦法克服這些問題
- 例如讓《抽樣方法和母體之間》具有獨立性。
- 這樣抽樣就不容易走樣太多！

而且在寫論文時

- 務必清楚地描述，你所採用之抽樣方法，以及和母體之間的關係。
- 這樣就能清楚地讓讀者知道你的研究方法與限制，不會誤導讀者！

當然、還有其他降低抽樣問題的方法

- 這就是進行統計研究前要學習的了！

除了統計檢定之外

- 統計學裡還有一些更進階的內容

像是我們可以用 ANOVA

- Analysis of variance（方差分析、變異數分析）

  來檢定《多組樣本》的《標準差平方》

標準差的平方 = 方差 = 變異數 =variance

然後也可以用相關係數

- 判斷兩組數據的相關程度

甚至用迴歸分析

- 找出兩組數據的相關公式

而這些檢定方法的數學原理

- 主要就是來自《中央極限定理》

以上這些

- 差不多就是大學《機率統計》的主要內容了！

這就是我們

- 今天的十分鐘系列

希望您會喜歡

我們下回見！

Bye Bye!