R 語言與統計(一)

丘祐瑋
David Chiu

環境資訊頁面

- 所有課程補充資料、投影片皆位於
  - https://github.com/ywchiu/cdc_course

關於統計

統計是什麼

- 日常生活中常需要根據不完整的資訊做決定
- 統計可以把不確定的程度量化，用精確的方式來表達，掌握不確定的程度
- 統計學的目的
  - 分析數據，將資料做出摘要
  - 做出更好的決定
  - 辨識出能提升做每件事的效果
  - 評估決策或事項的效用

敘述性統計 v.s. 推論性統計

- 敘述性統計
  - 有系統的歸納數據，了解數據的輪廓
  - 對數據樣本做敘述性陳述，例如：平均數、標準差、 計次頻率、百分比
  - 對數據資料的圖像化處理，將資料摘要變為圖表

- 推論性統計
  - 資料模型的建構
  - 從樣本推論整體資料的概況
  - 相關、迴歸、 單因子變異數、因素分析

描述數據

成人危險行為因子監測系統

- 自1984年起，美國疾病管制局以電腦電話訪問輔助系統建立危險行為因子監測系統（Behavioral Risk Factor Surveillance System, BRFSS）
- 該調查一年約電話訪問35萬名受訪者，收集具各州代表性的18歲以上成年人的各項健康相關危險因子的資訊。

資料集

- 從2000年調查中的取樣20,000個隨機樣本 cdc.Rdata
- 使用load 讀取cdc.Rdata

```
load("cdc.Rdata")
```

觀察資料

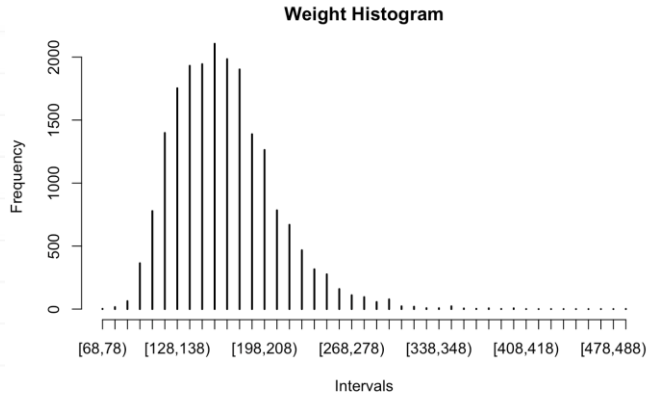
- 觀察變數

```
names(cdc)
```

- 變數內容
  - *genhlth* 受測者自評身體健康狀態
  - *Exeranyvariable* 受測者是否於過去一個月有運動(1) 或沒有 (0)
  - *hlthplan* 受測者是否有健康保險
  - *smoke100* 受測者一生中是否有抽超過一百根菸

繪製直方圖

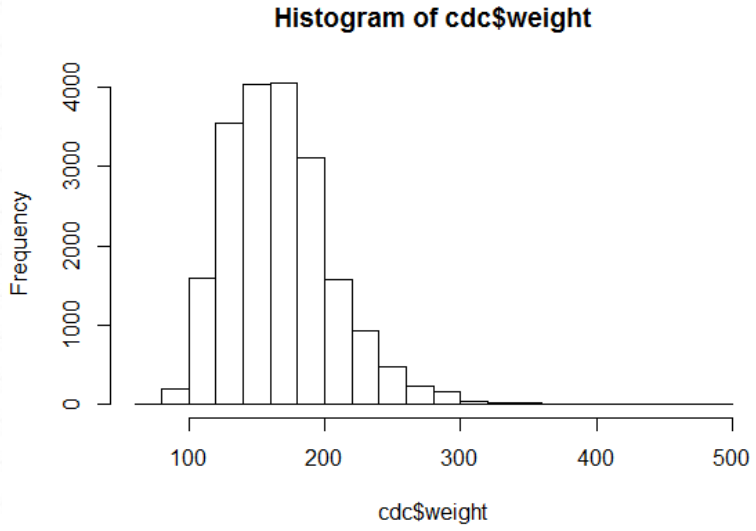
```
range(cdc$weight)
bins <- seq(min(cdc$weight),max(cdc$weight),by=10)
intervals <- cut(cdc$weight,bins,right=FALSE)
intervals
table(intervals)
plot(table(intervals), type ="h", main = "Weight Histogram", xlab = "Intervals",
ylab = "Frequency" )
```



連續型資料分布(直方圖)

- 直方圖
  - 將數據分成數組，依次數將連續型資料繪製成分布圖
  - 探索資料密度(density)

```
hist(cdc$weight)
```

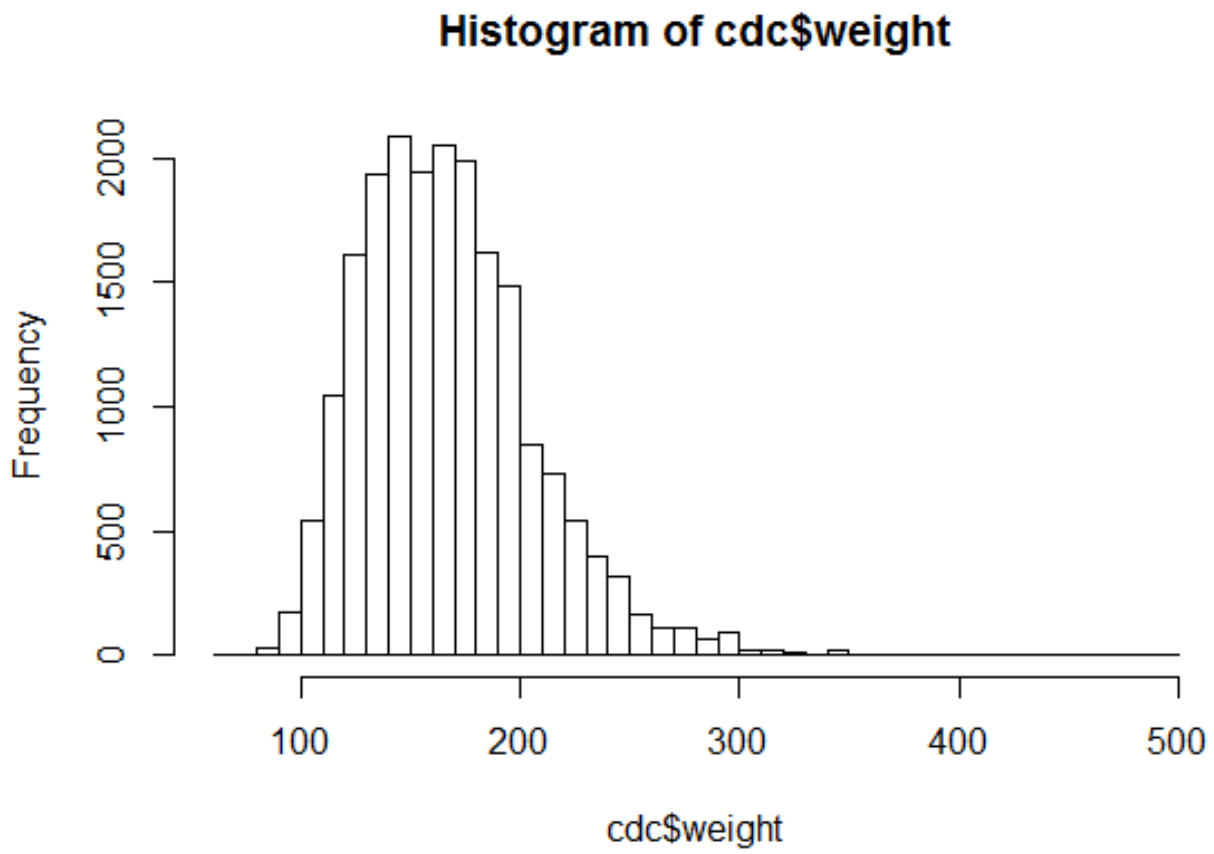


增加分組組數

```
hist(cdc$weight, breaks=50)
```

注意事項：

1. 數據較少，少分幾個組數
2. 數據多，增加組數



產生莖葉圖

直方圖無法看到每一個點，因此**杜奇**使用莖葉圖綜合數據，同時保留每一個點

```
stem(cdc$weight)
```

stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

找出資料的中心

- 找出數值的**中心**
  - 平均營業額
  - 平均氣溫

- 以平均值作為基準，找出遠離平均值的數據，是找出問題，做出改善的關鍵
  - 今日業績大於平均值（預估營收，因素）
  - 今日業績小於平均值（檢討行動或決策）

找出中心的方法

- 平均值
  - 數據總和 / 觀測值的個數
- 中位數
  - 由小到大排序，最中間的數值
  - 如果觀測值為偶數，中位數就是最中間兩個值加總的平均
  - 中位數不受離群值影響 (e.g. 薪水)
- 眾數
  - 出現最多的數值
  - e.g. 找出賣最好的商品

平均值、中位數與眾數

平均值：數據的中間水平

依分數重新排列

眾數：出現最多次的數值

拿整後出現最多次的分數是40分

眾數=40

中位數：中間位置的數值

此例為第三個數值

中位數=45

使用 R 探索平均值，中位數與眾數

- 使用mean 函式
  - mean(cdc$weight)
- 使用median 函式
  - median(cdc$weight)
- 使用table 跟 which.max (或使用sort)
  - names(which.max(table(cdc$weight)))
  - sort(table(cdc$weight), decreasing=TRUE)

基礎統計函式 (類別資料)

- 使用table 產生類別資料的統計資訊

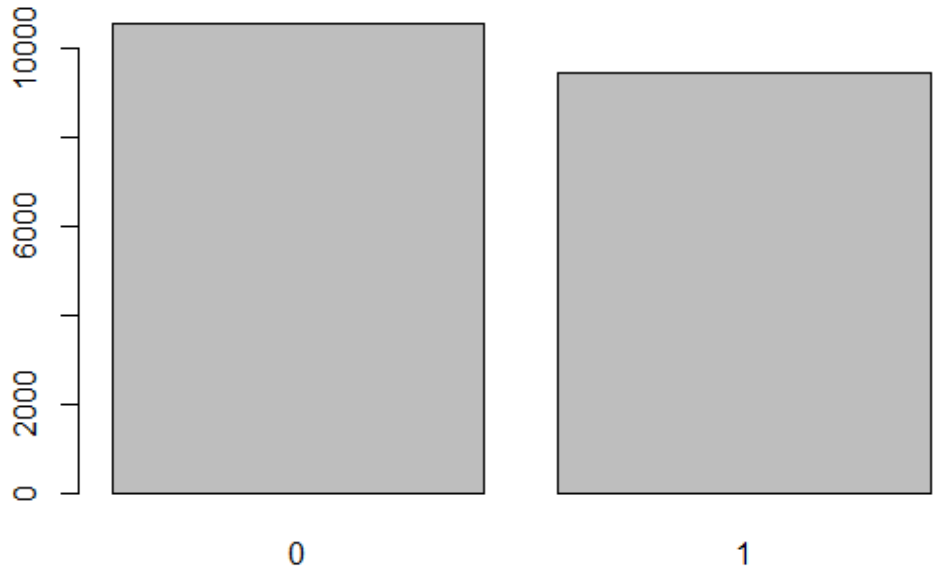
```
table(cdc$smoke100)
```

- 使用table 產生類別資料的相對比例

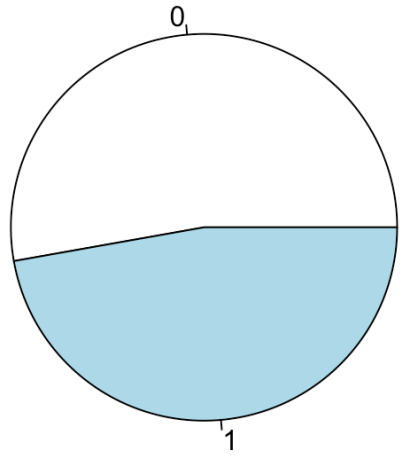
```
table(cdc$smoke100) / length(cdc$smoke100)
```

使用barplot 與 pie 繪製統計圖

```
barplot(table(cdc$smoke100))
```




```
pie(table(cdc$smoke100))
```



針對多維資料產生統計

- 使用table 產生統計資料

```
gender_smokers = table(cdc$gender, cdc$smoke100)
```

- 使用mosaicplot 繪製多維統計

```
mosaicplot(gender_smokers)
```

離度的量度

- 想知道數據與中心點散佈的有多遠

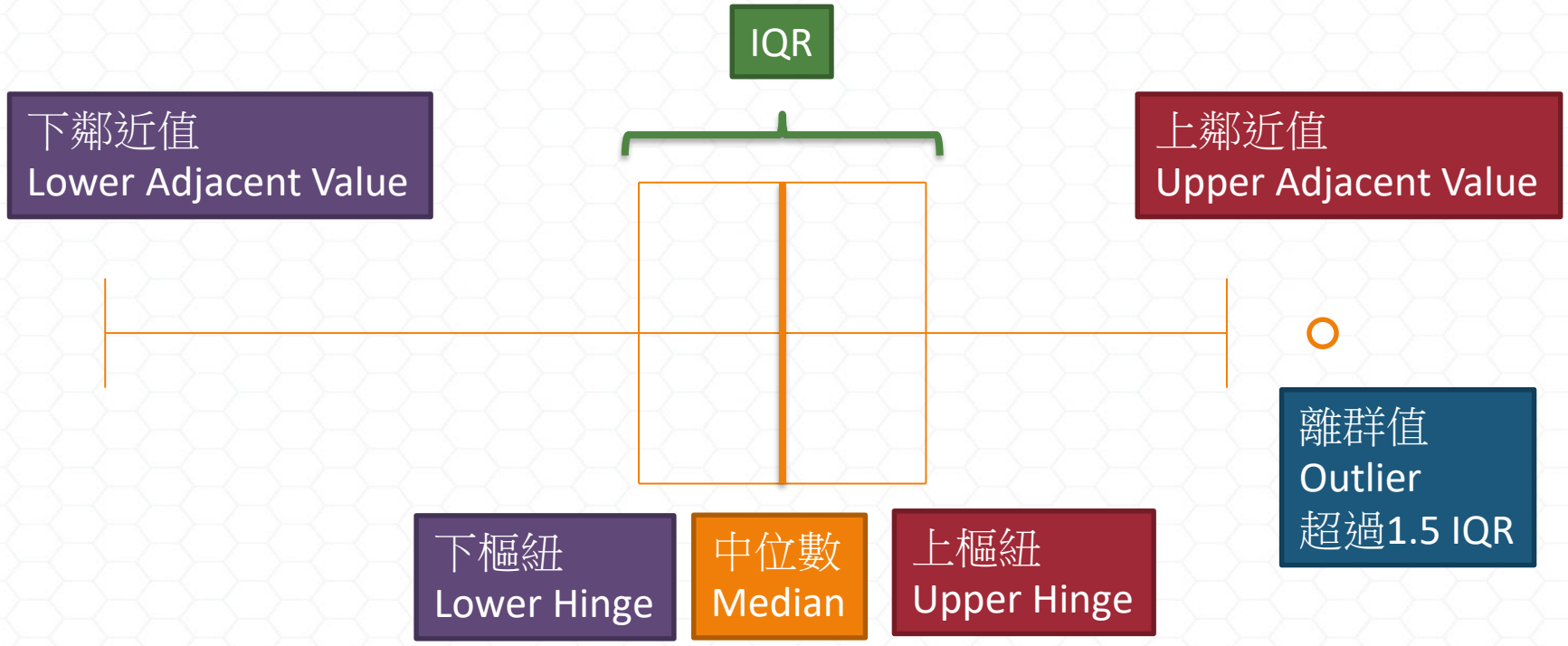
- 內四分位距 (IQR)
  - 把數據由大到小作排序
  - 以中位數為界，把資料分為高低兩組
  - 低組的中位數就是第一四分位數
  - 高組的中位數就是第三四分位數

IQR (範例)

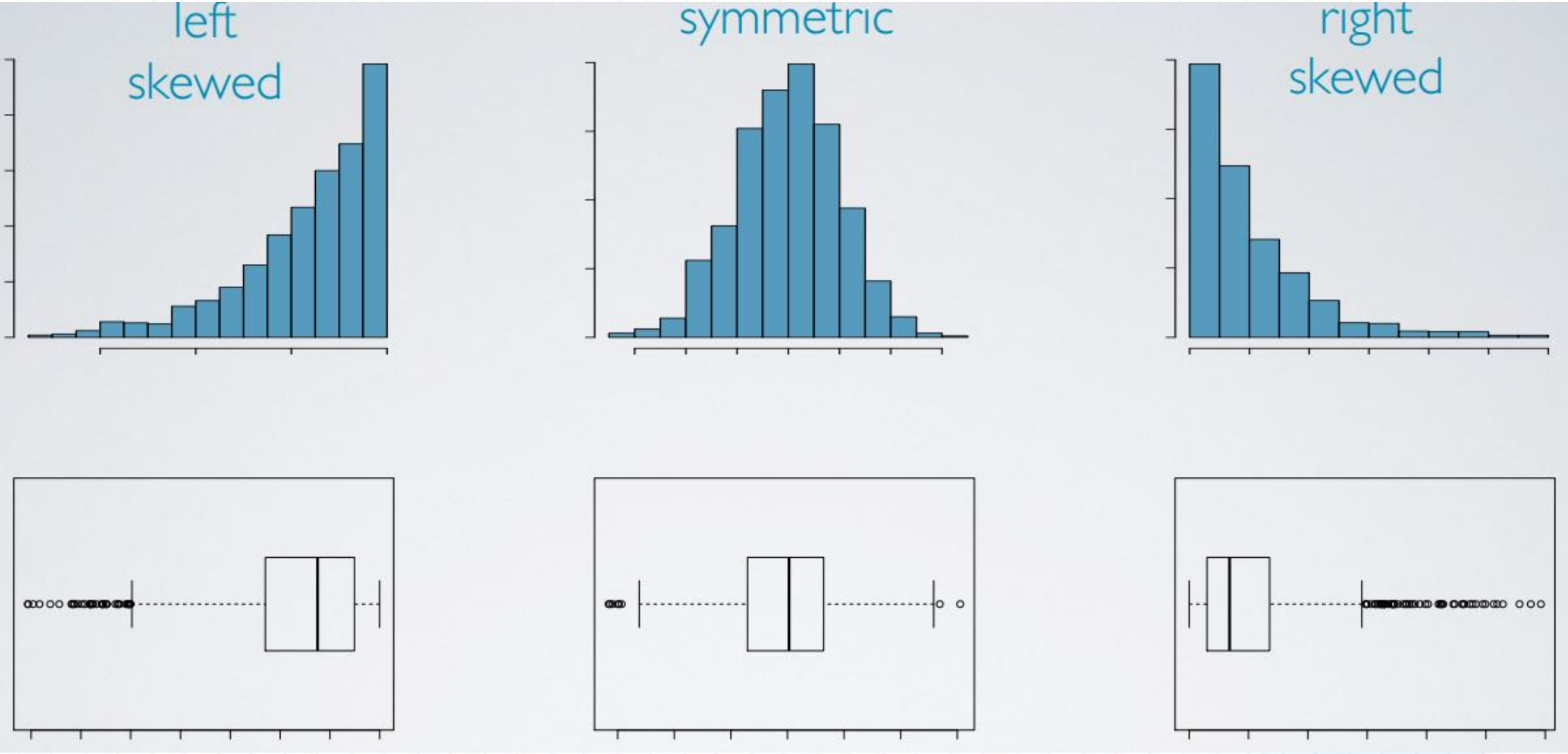
```
a = c(150, 155, 160, 162, 168, 171, 173, 175, 178, 182, 185)
stem(a)
quantile(a,0.75)
quantile(a,0.25)
IQR(a)
```

```
15 | 05
16 | 028
17 | 1358
18 | 25
```

箱型圖 – 表示IQR 的方式



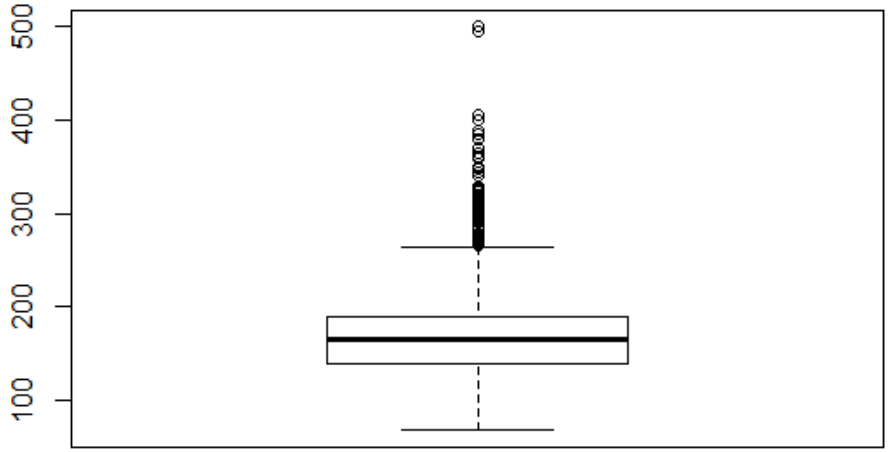
箱型圖與資料分布



Boxplot

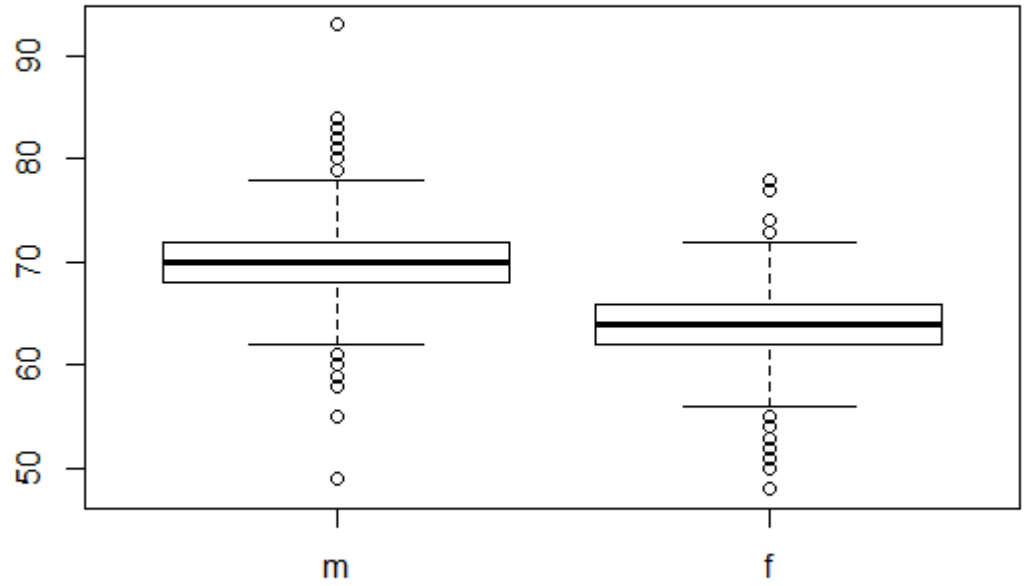
- 使用boxplot 繪製資料

```
boxplot(cdc$weight)
```



分組繪製資料

- boxplot(cdc$height ~ cdc$gender)

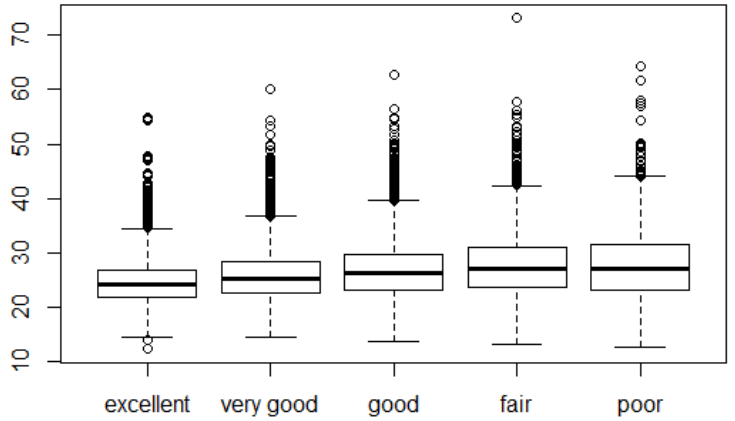


產生BMI 統計圖

- BMI 計算公式 (磅/英吋^2) * 703

$$BMI = \frac{weight(lb)}{height(in)^2} * 703$$

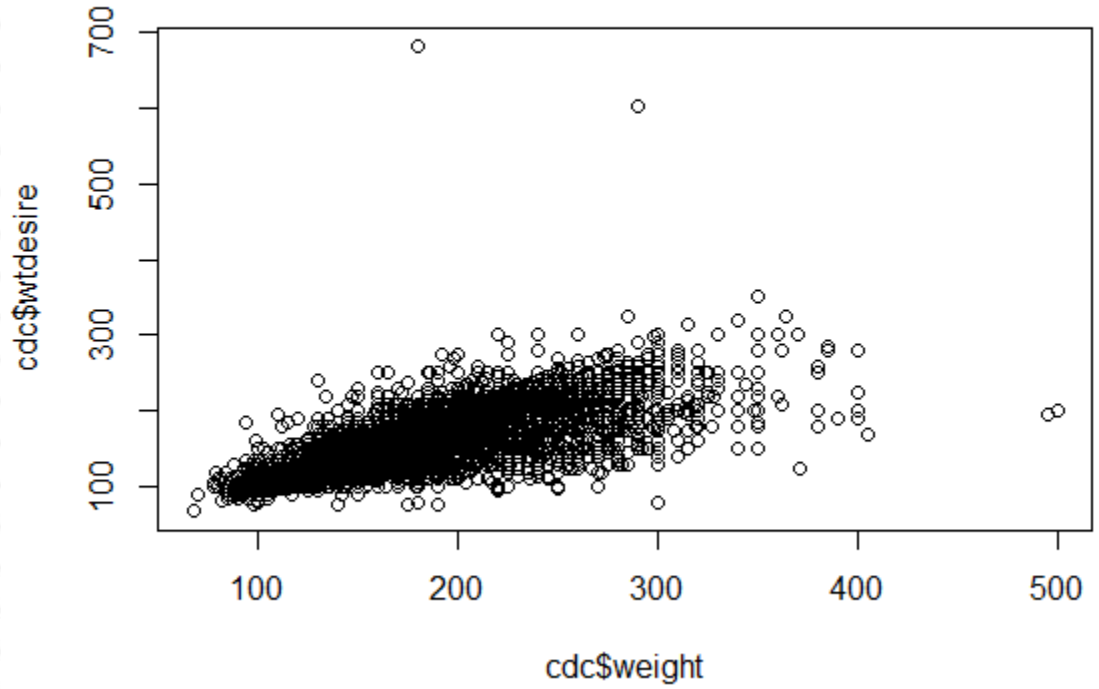
```
bmi = (cdc$weight/cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth)
```



繪製資料散佈圖 (Scatter Plot)

- 理解不同組資料之間的關係 (e.g.線性)

```
plot(cdc$weight, cdc$wtdesire)
```



標準差

- 銷售額大於/小於平均銷售額多少，算合理範圍？
- 該如何估算合理的變動幅度？
- 實際應用：
  - 比較組內數據的變動幅度(每日合理的營業波動幅度)
  - 比較不同組數據的變動幅度 (不同店的營業波動幅度)
  - 比較不同單位的變動幅度 (來客數變動與營業額變動)

標準差

- 標準差是以平均數來度量資料的離散程度
  - IQR 是以中位數為基準
  - 可以想成是距離平均數的平均距離 (距離的平方作為距離)
- 變異數 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - 變異數越大，資料越分散
- 標準差 = s
  - 非負整數值
  - 與度量單位有相同單位

變異數與標準差

■ 標準差

```
sd(cdc$weight)
sqrt(var(cdc$weight))
```

■ 變異數

```
sd(cdc$weight) ^ 2
var(cdc$weight)
```

IQR v.s. 標準差

```
contender1 <-
c(8.4,8.6,8.8,9,9,9.2,9.7,10.1,10.4,10.3,10.5,10.6,11.0,11.1,11.4,11.7,11.9,12.3,12.8,13,13,14.2,14.4,14.6)
contender2 <-
c(9.8,9.8,9.9,10.1,10.1,10.2,10.2,10.3,10.3,10.7,10.8,10.8,11,11.1,11.2,11.2,11.3,11.6,11.7,11.7,11.8,11.8,1
1.9,11.9)
summary(contender1)
summary(contender2)
IQR(contender1)
IQR(contender2)
combined <- cbind(contender1,contender2)
boxplot(combined)
sd(contender1)
sd(contender2)
```

機率

機率

- 事情發生的可能性

- 基本定義
  - 隨機實驗: 產生結果的過程
    - e.g. 紀錄銅板丟出的結果
  - 基本結果: 所有可能結果
    - e.g.正面與反面
  - 樣本空間: 基本結果的集合
    - e.g. {正面, 反面}

機率的特質

- 機率的特質
  - 機率一定是非負的
  - 所以的基本結果總和為一

隨機實驗模擬

■ 產生各式樣本

```
sample(1:10)
sample(1:10, size = 5)
sample(c(0,1), 10, replace = TRUE)
sample.int(20, 12)
```

■ 投一百次硬幣

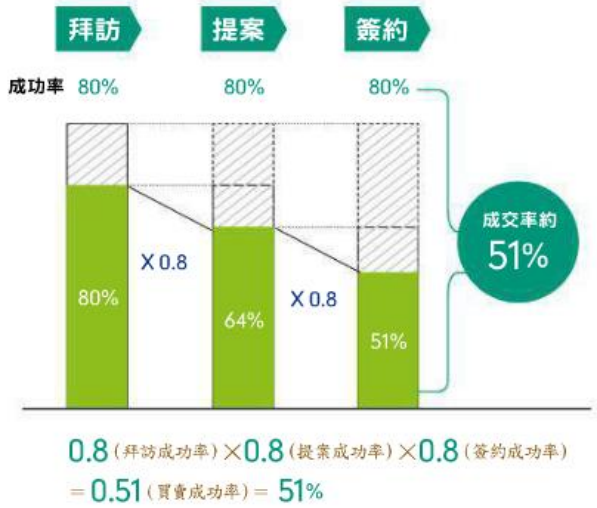
```
coins <- c("heads", "tails")
fair_coin <- sample(coins, size = 100, replace = TRUE)
table(fair_coin)
```

不公平的硬幣投擲結果

```
outcomes <- c(“heads”, “tails”)
unfair_coin <- sample(outcomes, size = 100,
                      replace = TRUE, prob = c(0.3,0.7))
table(unfair_coin)
```

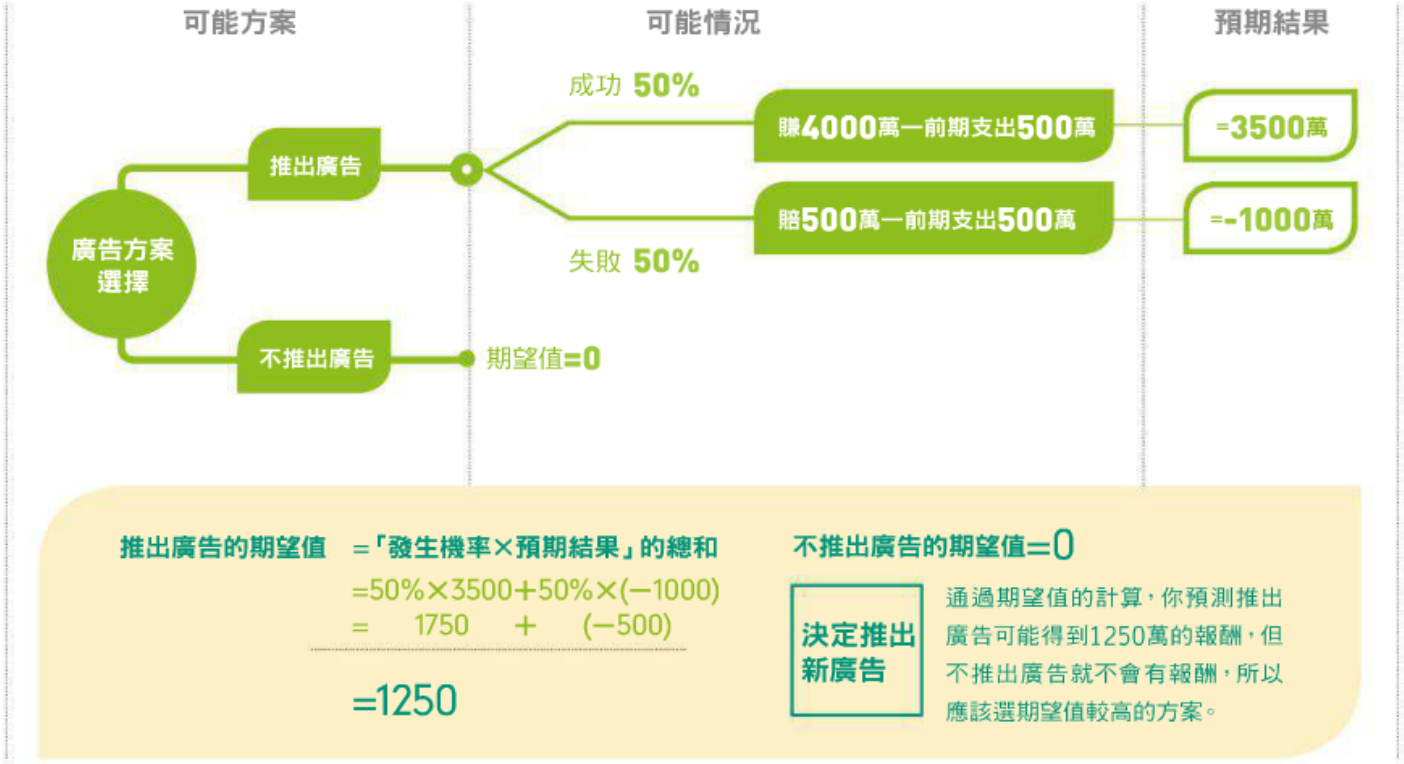
獨立事件

- 獨立事件
  - 一事件的發生不會影響到另一事件發生的機率
  - 可以互相相乘
- 如何驗證兩事件是獨立
  - P(A | B) = P(A), 那A跟B 是獨立



期望值

- 機率的延伸
- 做出報酬率高的決定

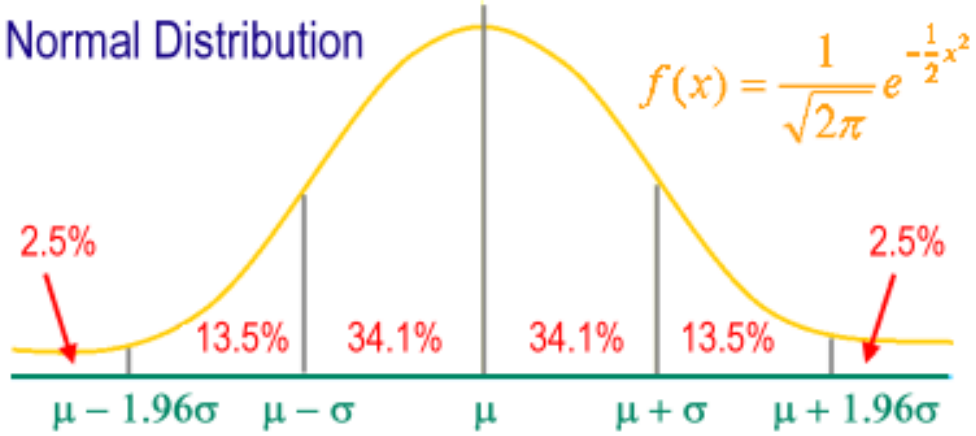


常態分布

常態分佈

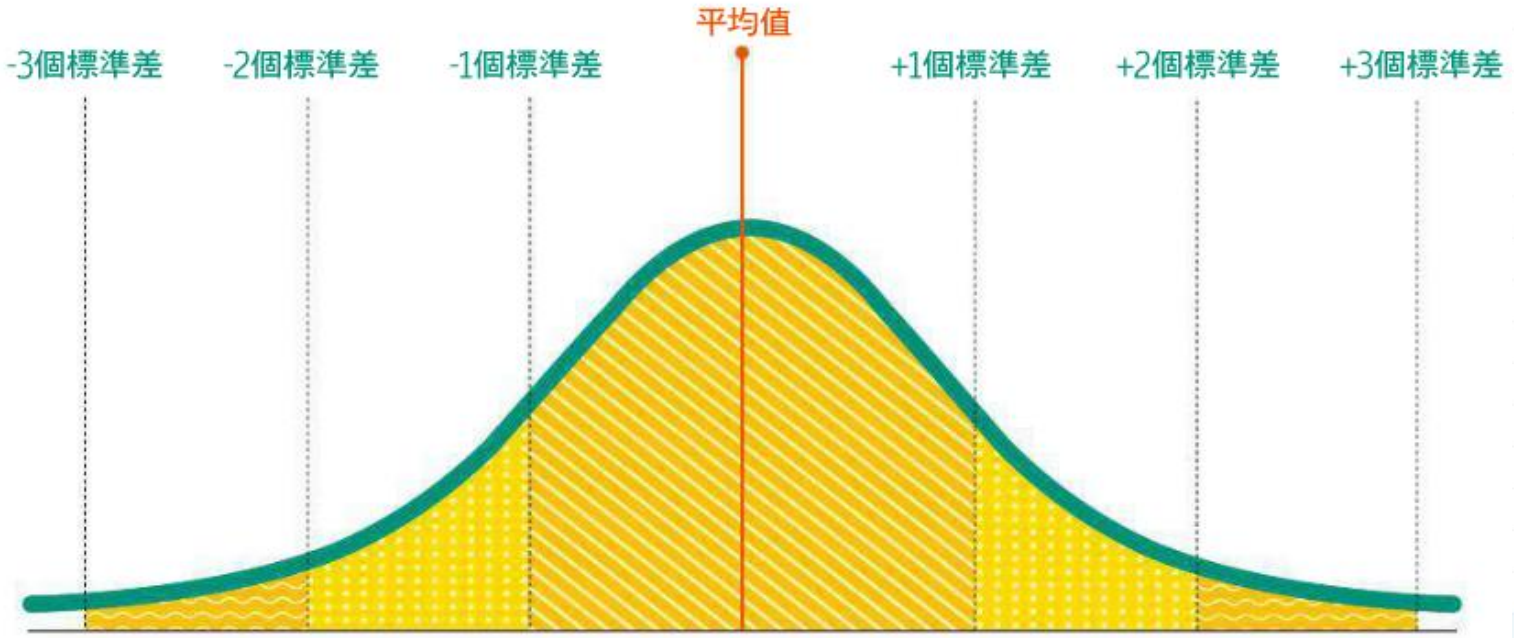
- 標準常態分佈：以平均值 (mean) 為中心，標準差 (standard deviation) 為座標軸之基本單位所繪之常態分佈圖。 形狀為覆鐘形的對稱圖形
  - 只要數據來自於隨機獨立的樣本, 則最終都會形成常態分布

  - m ± 1s含有整個樣本群之 68.26% 的個體。
  - m ± 2s含有整個樣本群之 95.44% 的個體。
  - m ± 3s含有整個樣本群之 99.74% 的個體。
  - 95% 個體落在 m ± 1.96s 之間。
  - 99% 的個體落在 m ± 2.58s 之間。



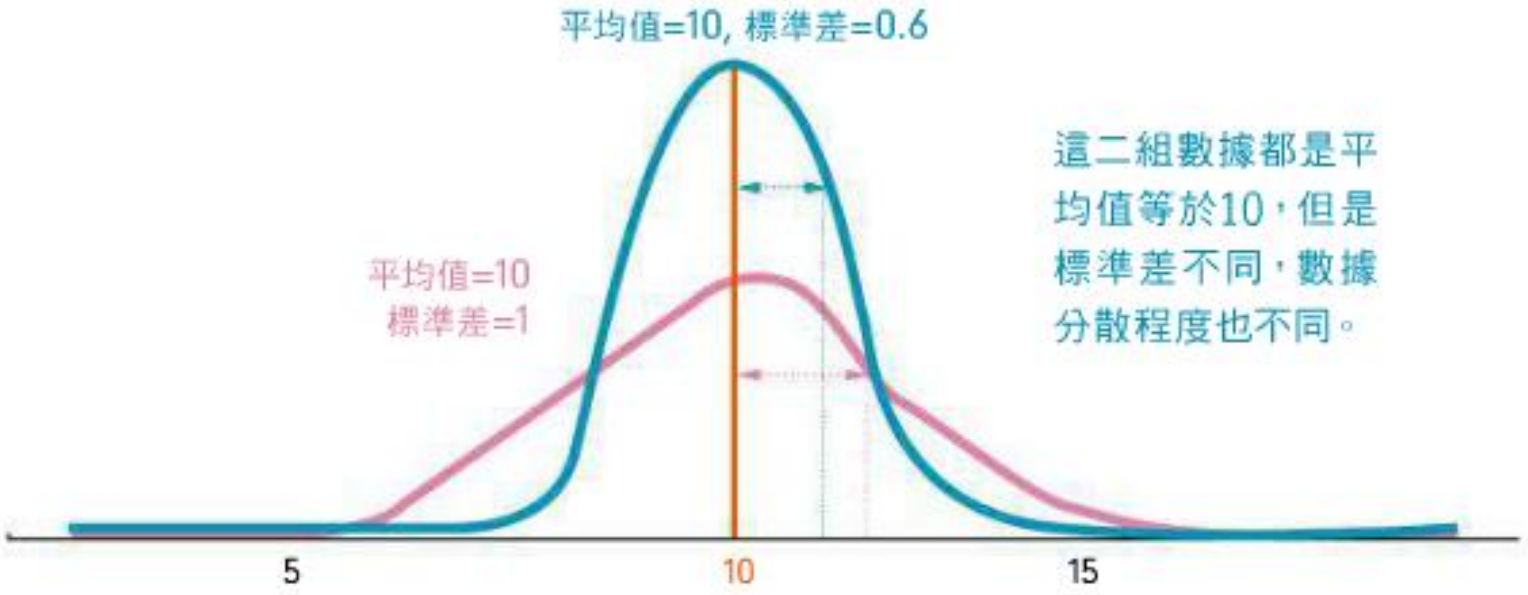
常態分佈 -掌握主流意見、留意極端現象

- 大部分的數據都會遵從常態分布 (身高，體重，分數，智力)
- 百分之68.2 的資料會位於+-一個標準差的區間之內
- 六個標準差代表 99.99% 產品與服務都符合目標值



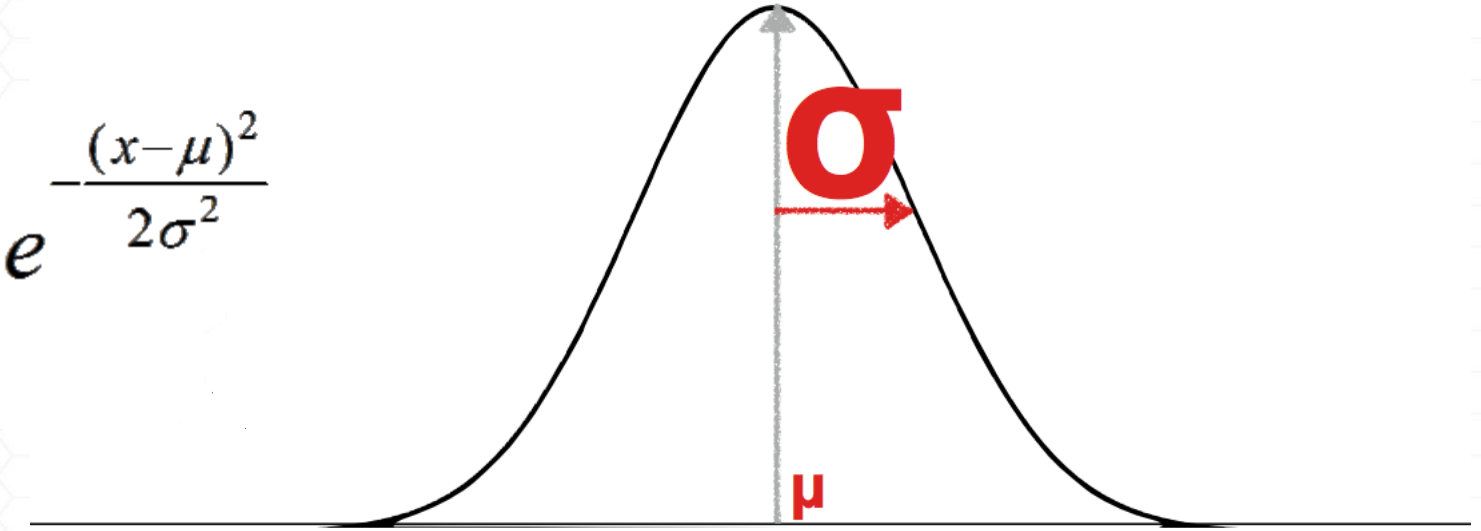
同一平均值 不同標準差

- 直方圖 (histogram) 呈現單一主峰 (single peak) 的左右對稱圖形，即平均值 (mean, m) 在正中央。
- 越接近平均值的數值出現的頻率越高，越遠離平均值的數值出現的頻率越低
- 同一平均值，不同標準差
- 資料的分散程度就會不相同



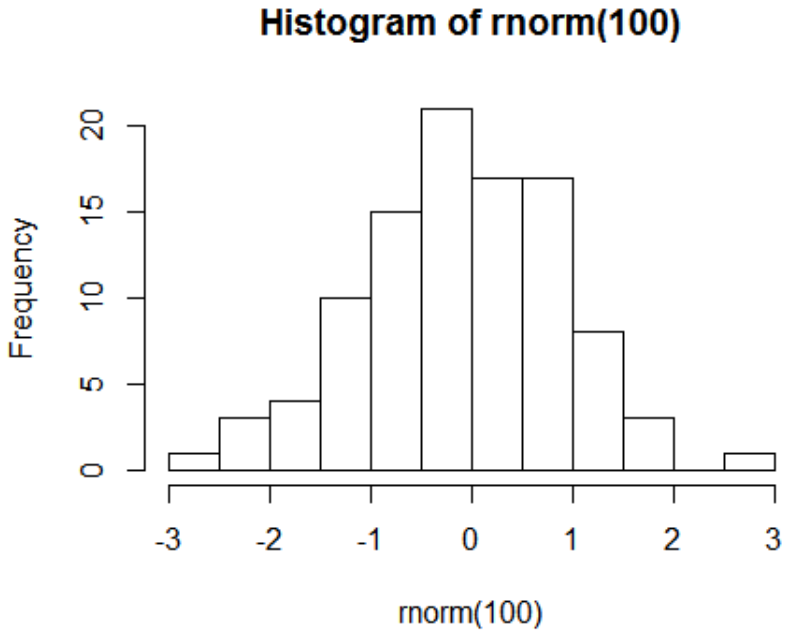
只要知道平均與標準差就可以算出所有機率

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



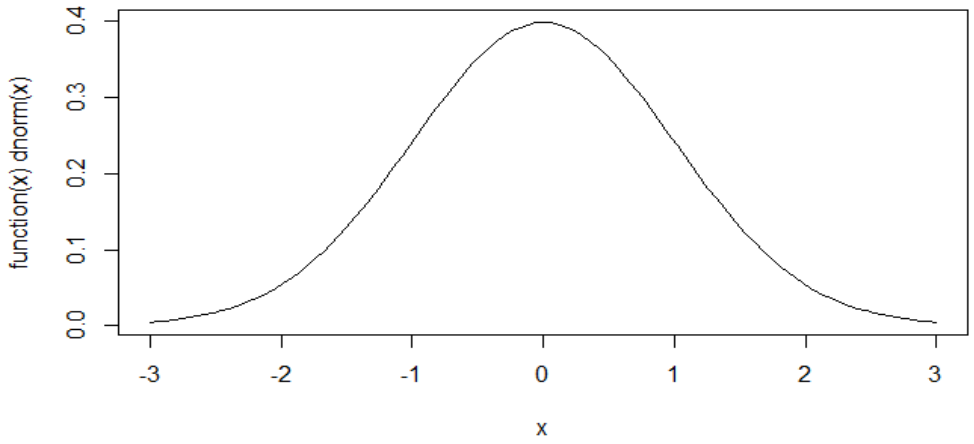
使用R 產生常態分布

- 使用rnorm
  - rnorm(100)
  - hist(rnorm(100))



檢視分布曲線的高度 (在該點的機率)

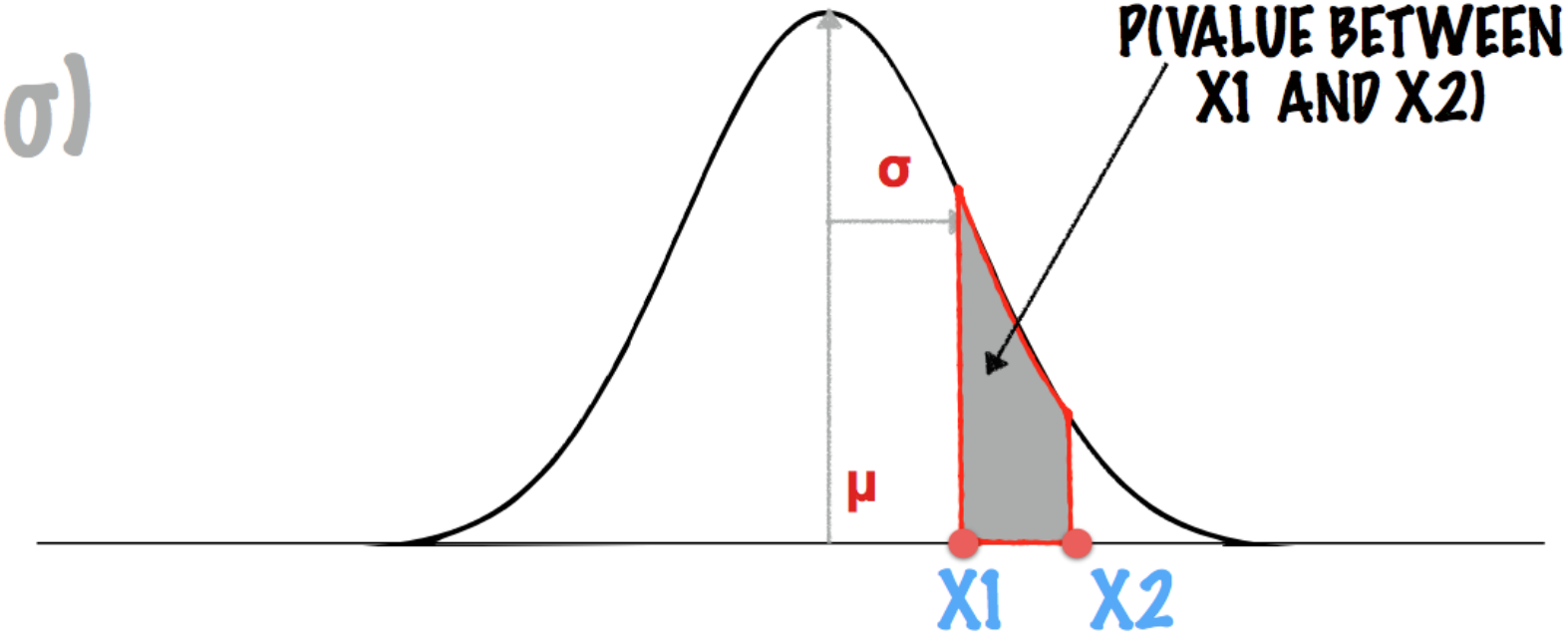
```
dnorm(0)
[1] 0.3989423

dnorm(0,mean=3,sd=5)
[1] 0.06664492

curve(dnorm,-3,3)
```



計算點與點之間發生的機率

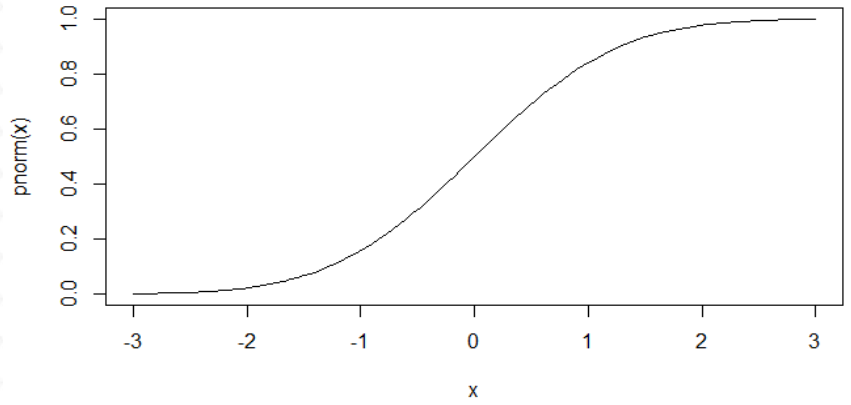
- 曲面下的面積代表任兩點之間的機率

$$f(x) = F(x, \mu, \sigma)$$



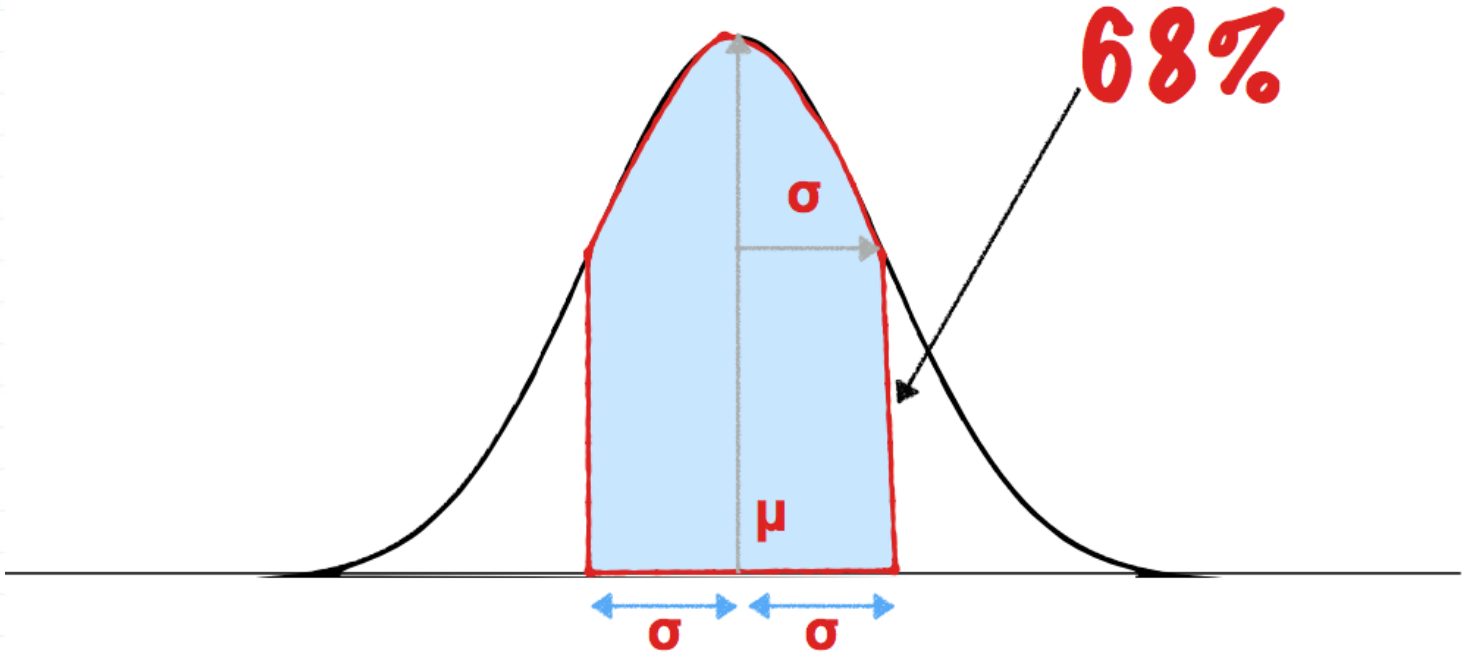
檢視分布曲線的面積

```
pnorm(1.5)
[1] 0.9331928

pnorm(1.5, lower.tail=FALSE)
[1] 0.0668072

curve(pnorm(x), -3,3)
```



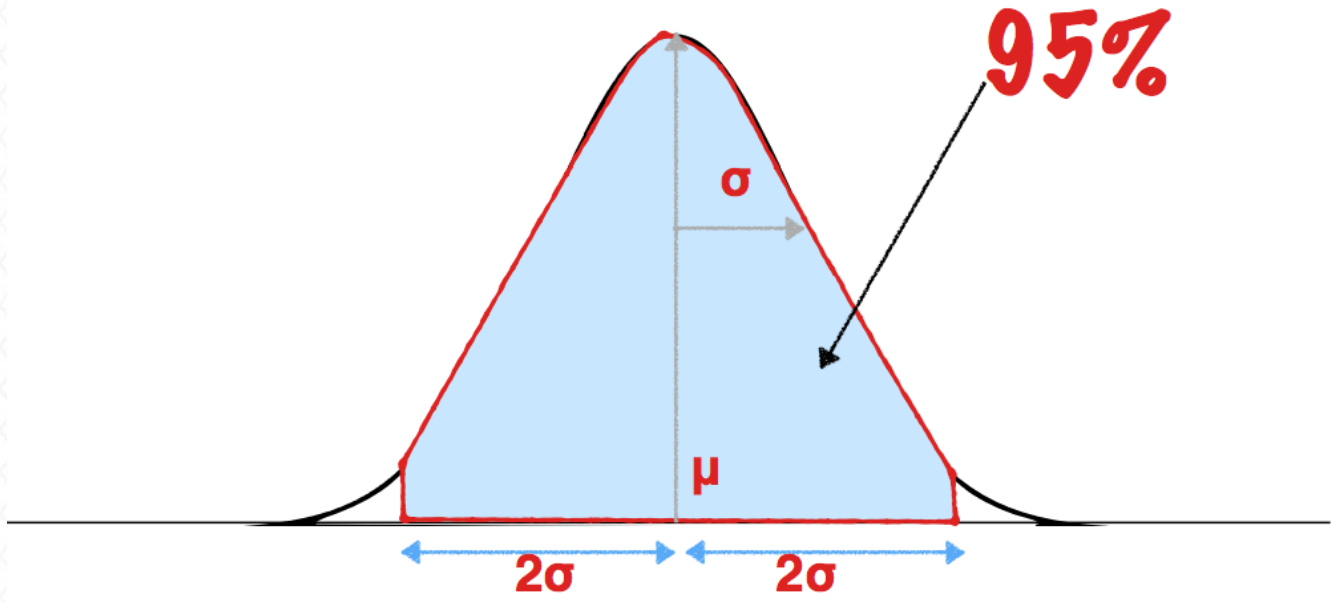
計算一個正負一標準差之間的機率

```
pnorm(1) - pnorm(1, lower.tail=FALSE)
```



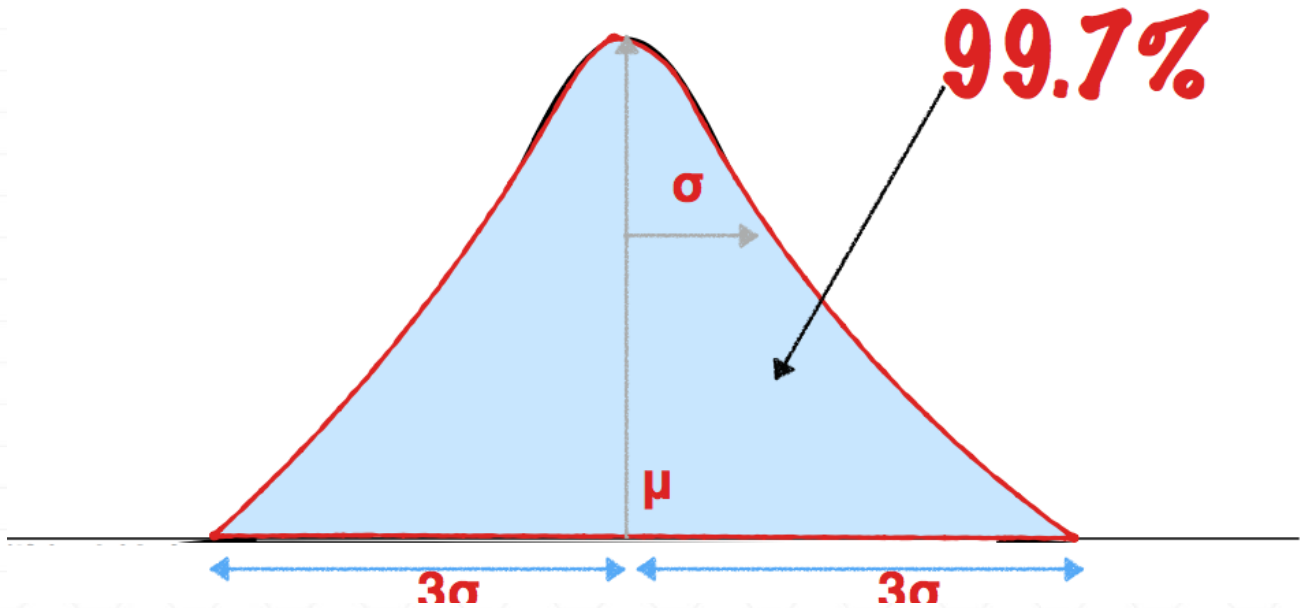
計算一個正負二標準差之間的機率

```
pnorm(2) - pnorm(2, lower.tail=FALSE)
```



計算一個正負三標準差之間的機率

```
pnorm(3) - pnorm(3, lower.tail=FALSE)
```



常態分布下的機率值

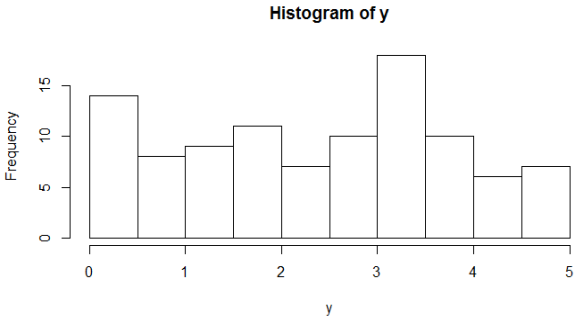
- P(距離平均值正負一標準差) = 68%
- P(距離平均值正負二標準差) = 95%
- P(距離平均值正負三標準差) = 99%

- 可以用來
  - 檢驗分布是否為常態分布
  - 找出離群值

除了常態分佈以外

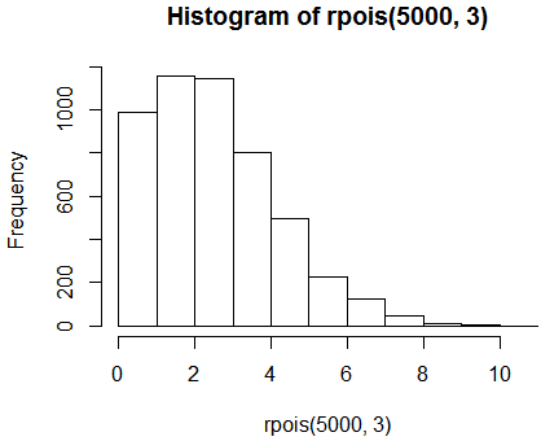
- 均勻分佈

```
set.seed(50)
y = runif(100,0,5)
hist(y)
```



- poisson 分佈

```
hist(rpois(5000,3))
```



檢視是否為常態分佈

■ Shapiro 檢定

```
shapiro.test(sample(cdc$weight,5000))

        Shapiro-Wilk normality test

data:  sample(cdc$weight, 5000)
W = 0.9612, p-value < 2.2e-16
```

如果 p < 0.05，則推翻虛無假設，表示為非常態分佈；
如果是 p > 0.05，則接受虛無假設，表示為常態分佈

取樣分佈

平均數的抽樣分佈

- 生產手機的廠商，希望不用檢查每一隻手機，就知道平均手機的耗電量
  - 隨機取出 n 支手機，度量耗電量

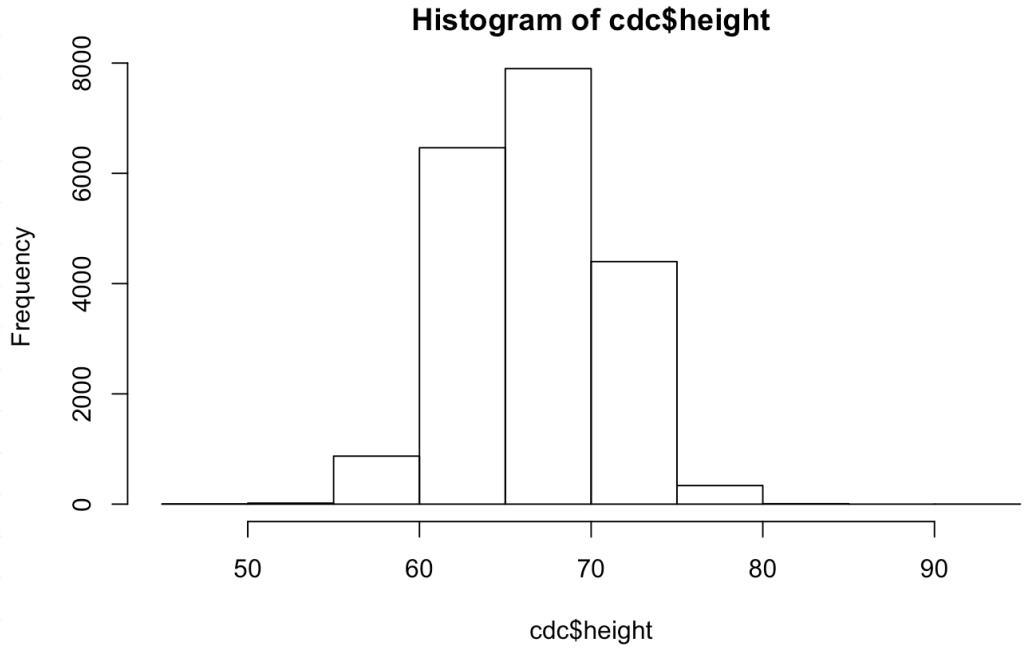
- 抽樣分佈
  - 對特定樣本數 N，算術平均數的機率分佈
  - 對任意樣本數，取樣的算術平均數亦為常態分佈
  - **中央極限定理**

取樣分佈(Sampling Distribution)

- **中央極限定理 Central Limit Theorem**
  - 從任何一個母群(算術平均數為μ，標準差為σ)中取大小為N之樣本，當N足夠大時，取樣的算術平均數會接近常態分佈
  - 要知道平均數的分佈，只要知道母體平均數與標準差即可
  - 不同種的分佈（常態、均勻或布瓦松分配）只要平均數和標準差相同，平均數的分佈都為常態分佈

列出身高的分佈

```
hist(cdc$height)
```



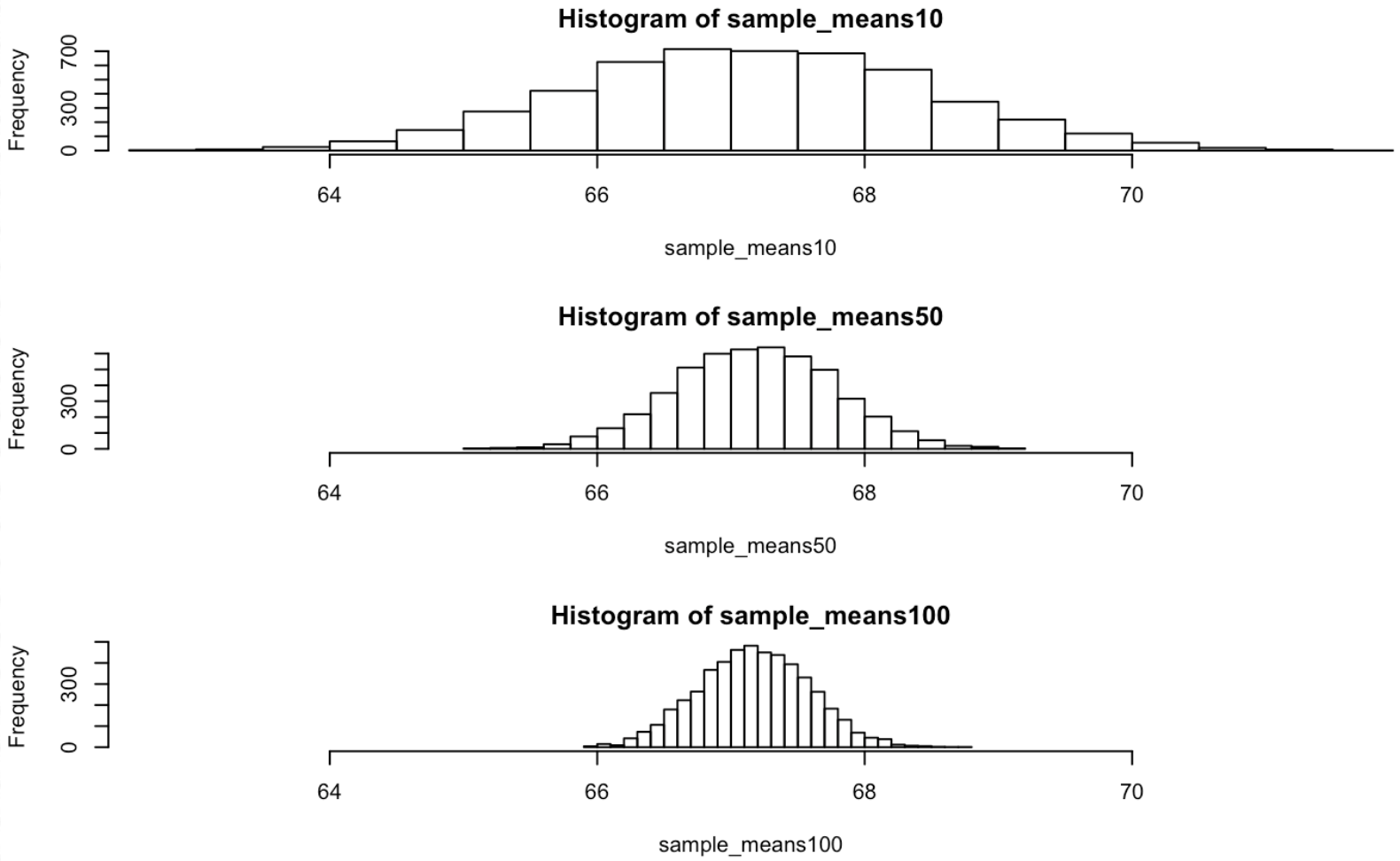
資料取樣

```
sample_means10  <- rep(NA, 5000)
sample_means50  <- rep(NA, 5000)
sample_means100 <- rep(NA, 5000)


for (i in 1:5000) {
    samp <- sample(cdc$height, 10)
    sample_means10[i] = mean(samp)
    samp <- sample(cdc$height, 50)
    sample_means50[i] = mean(samp)
    samp <- sample(cdc$height, 100)
    sample_means100[i] = mean(samp)
}
```

觀察不同取樣的資料分佈

```
par(mfrow = c(3, 1))

xlimits <- range(sample_means10)

hist(sample_means10, breaks = 20, xlim = xlimits)
hist(sample_means50, breaks = 20, xlim = xlimits)
hist(sample_means100, breaks = 20, xlim = xlimits)
```

不同取樣值產生的直方圖



THANK YOU