R 語言與統計(二)

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

點估計與信賴區間

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

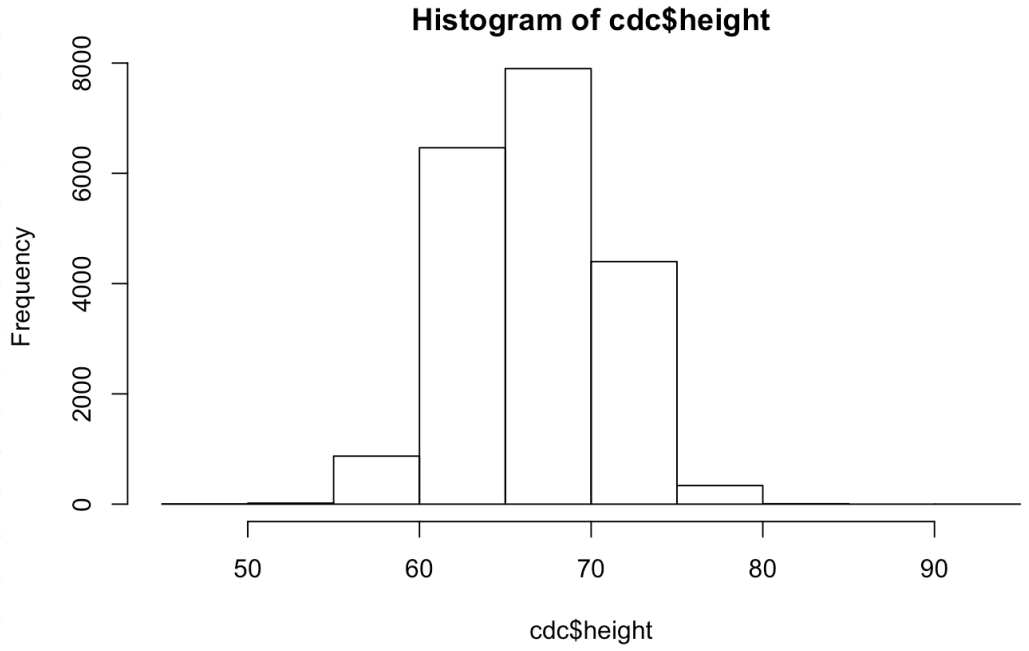
- 抽樣分佈
  - 對特定樣本數 N，算術平均數的機率分佈
  - 對任意樣本數，取樣的算術平均數亦為常態分佈
  - **中央極限定理**

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)
}
```

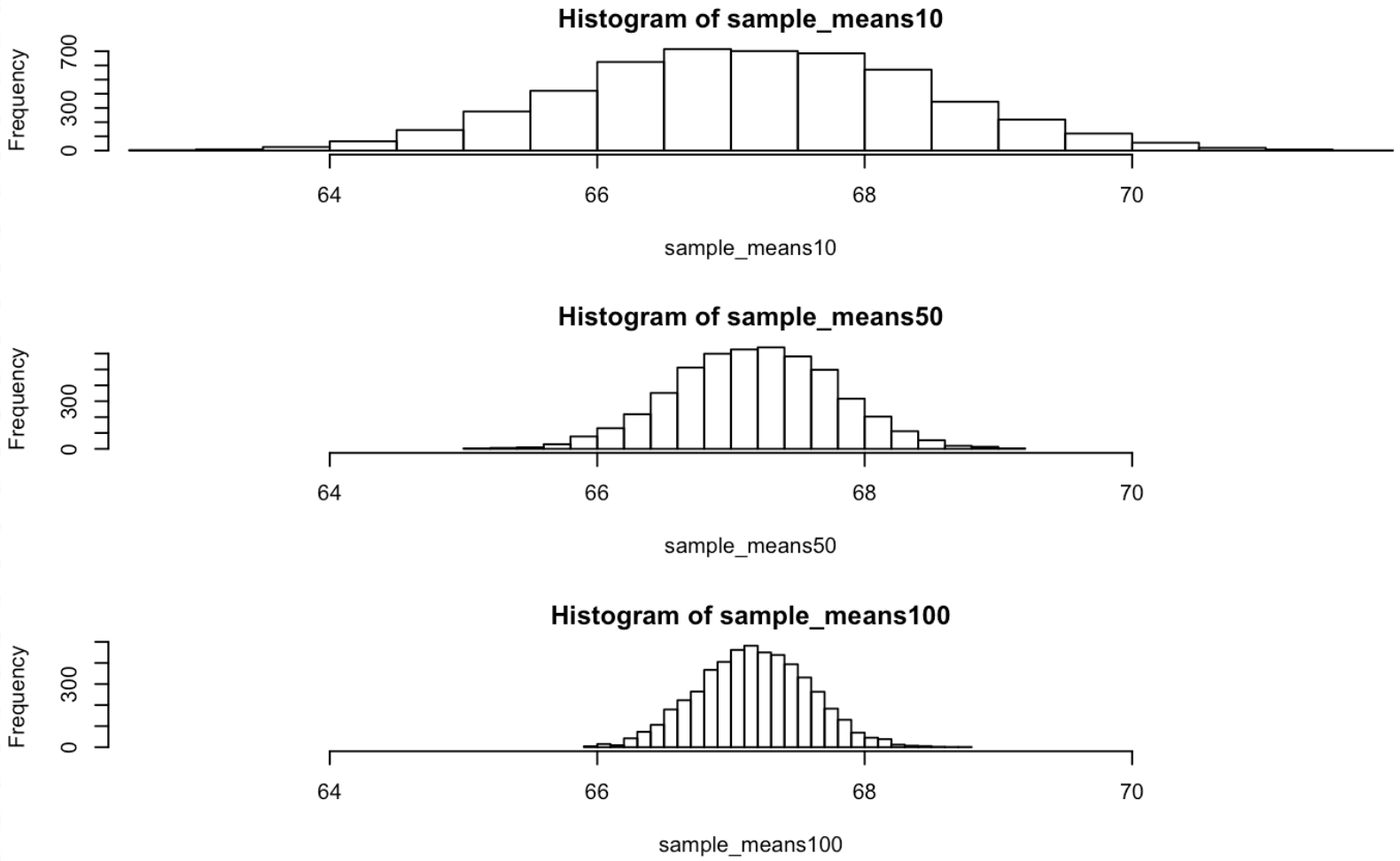
觀察不同取樣的資料分佈

```
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



信賴區間(Confidence Interval)

- 為想了解所有民眾的身高分布，通常會從兩個基本資訊著手
  - 一般民眾大概身高多高?
  - 人與人之間的身高變異數有多大?

但如果人口很多, 沒有辦法一一調查呢?

抽樣調查

- 誤差：推論會產生誤差，要得到比較小的誤差，樣本數就比較大
- 信心水準：有多大的信心可以用樣本推論母體，通常設定在95或99%

95%信心水準

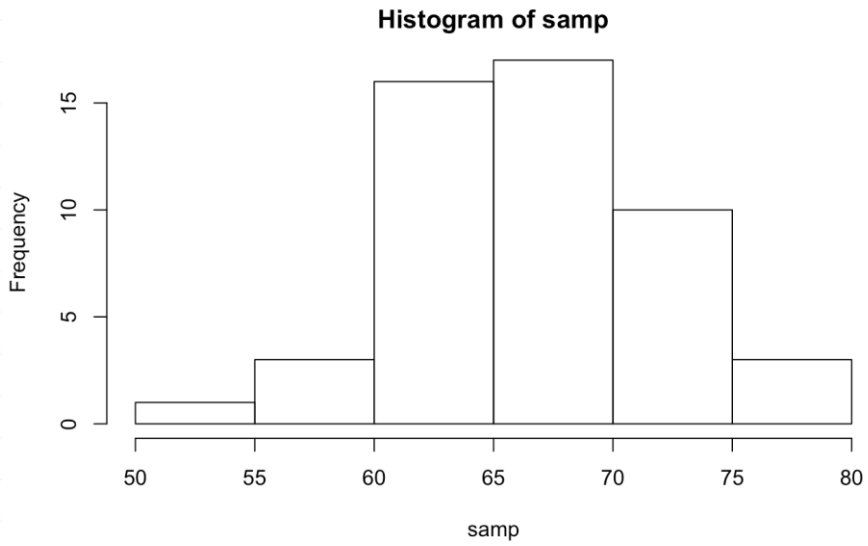
- 有95%的機率，會產生一個包含p在內的區間 => 95%的時候，p會落在以觀測值$\hat{p}$為中心的區間

- 假設從母體中隨機抽出1,000名消費者，如果發現80%消費者對A產品滿意在95%信心水準下，誤差範圍為正負3%
  - 代表針對該產品重複進行100次調查，有95次消費者對A產品的滿意比例介於77% ~ 83%

隨機從身高資料中取出50個樣本

```
population <- cdc$height

set.seed(123)
samp <- sample(population, 50)
sample_mean <- mean(samp)
hist(samp)
```



取樣樣本的平均值，與母體的平均值應該相去不遠

標準誤差 Standard Error

- 標準誤差即樣本統計量的標準差
  - 樣本的平均數和總體真實平均數的偏離程度

- 每次抽樣都能得到一個樣本均值
  - 如果你抽樣的樣本數N很大，那麼相對你的樣本均值和總體均值的偏離程度就比較小(如果你的N接近無窮大那麼偏離程度趨近0)
  - 而反之, 如果樣本數N很小, 它和真正總體的均值的偏離程度是相對較大

Z-Score

- 是藉由從單一（原始）分數中減去母體的平均值，再依照母體（母集合）的標準差分割成不同的差距

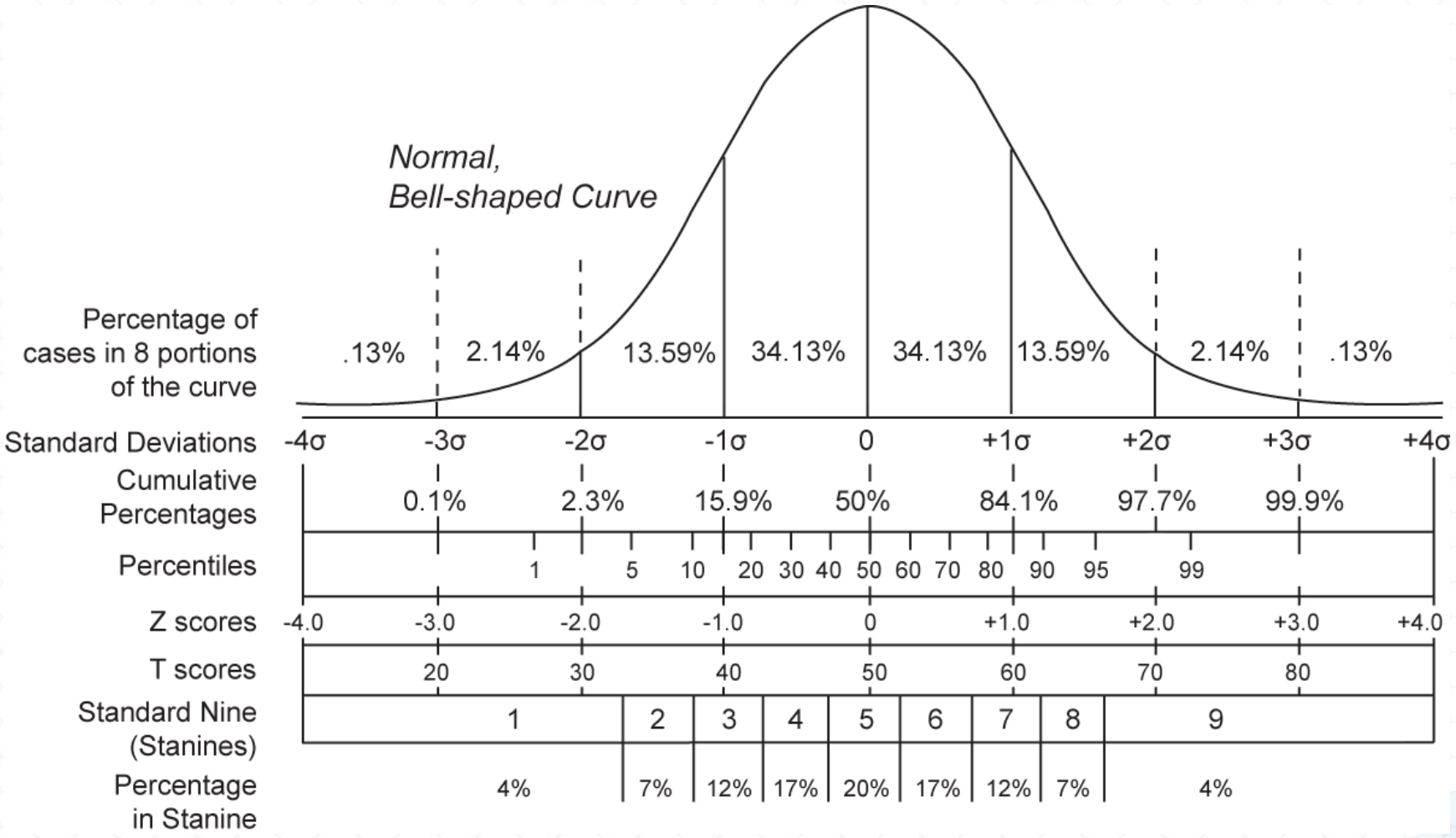
$$z = \frac{x - \mu}{\sigma}$$

其中 $\sigma \neq 0$。

其中

- x 是需要被標準化的原始分數
- μ 是母體的平均值
- σ 是母體的標準差

常態分布 與 Z Scores



信賴區間(Confidence Interval)

- 要正確估計母體參數是不可能的，但是可以假設母體參數應該落在一定的區間，稱為信賴區間(confidence interval)
- 而產生信賴區間需要信心水準(confidence level)，或者是誤差(margin of error)
- 點估計加減誤差便是區間估計
  - 信賴區間=點估計+-誤差
  - 誤差=critical value × sde
  - 而critical value(z值)來自於α=1-（信賴區間/100)
  - z值對應α，α/2分屬於z值分佈的兩端

推算信賴區間參數

- 95% 信賴區間，代表有95%的區域位於-1.96~1.96 之間
  - P(X>1.96) = 0.025
  - P(X<1.96) = 0.975
  - P(-1.96<X<1.96) = 0.95

- qnorm 是pnorm 的反向推導

```
qnorm(0.975)
pnorm(1.96)
```

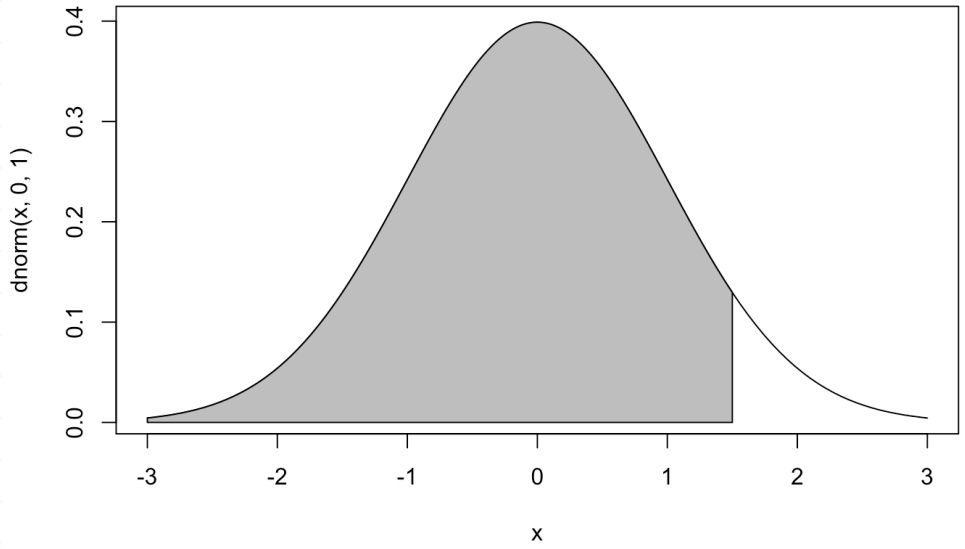
pnorm & qnorm

- pnorm 函數可求出某一分位點的累積幾率值
  - e.g.：標準常態分位點z = 1.5 的累積機率值
  - pnorm(1.5, 0, 1)

- qnorm 求出累積機率為0.9331928 所對應的標準常態分位點
  - qnorm(0.9331928, 0, 1)

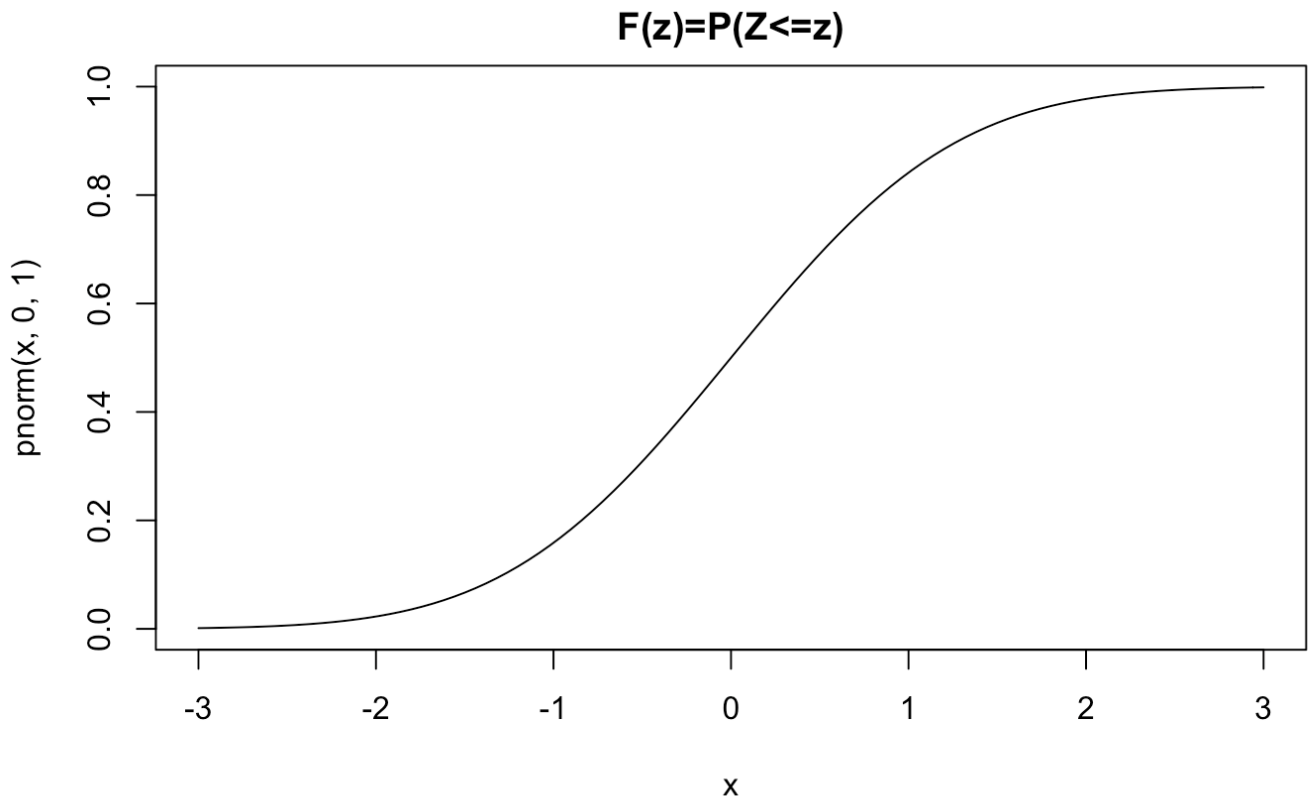
繪製機率密度曲線與累積分配曲線

```
curve(dnorm(x, 0, 1), xlim = c(-3, 3))
cord.x <- c(-3, seq(-3, 1.5, 0.01), 1.5)
cord.y <- c(0, dnorm(seq(-3, 1.5, 0.01)), 0)
polygon(cord.x, cord.y, col = "grey")
```



pnorm v.s. Z-Score 圖

```
curve(pnorm(x, 0, 1), xlim = c(-3, 3), main = "F(z)=P(Z<=z)")
```



推算95%信賴區間的上下界

```
sde <- sd(samp)/sqrt(50)
lower <- sample_mean - 1.96 * sde
upper <- sample_mean + 1.96 * sde

lower
upper
```

多變量分析

單變量分析 v.s. 多變量分析

- 單變量分析
  - 針對單一變數進行簡單的量化分析
  - mean, median, var
- 多變量分析
  - 旨在了解變數之間的相互關係
  - 例如氣溫跟感冒人數的關係
  - cov, cor

cov 和 cor

- cor
  - 計算列與列之間的相關係數
  - 顯示兩個隨機變數之間線性關係的強度和方向
  - http://zh.wikipedia.org/wiki/%E7%9B%B8%E5%85%B3
- cov
  - 計算列與列之間的共變異數
  - 衡量兩個變量的總體誤差
  - http://zh.wikipedia.org/wiki/%E5%8D%8F%E6%96%B9%E5%B7%AE

共變異數

- 用來衡量兩變量之間的密切程度與方向

- 共變異數(covariance)：
  - 測量兩個數值變數間的線性關係
  - 計算列與列之間的共變異數
  - 衡量兩個變量的總體誤差
  - 如果兩個變量的波動幅度都很大，看不出相關程度

$$Cov(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{n}$$

相關係數

- 相關係數(Correlation Coefficient)
  - 計算列與列之間的相關係數
  - 顯示兩個隨機變數之間線性關係的強度和方向
  - 將共變異數除以兩者的標準差
  - 數值介於 -1 ~ 1 之間 (通常絕對值>0.7為高度正相關)

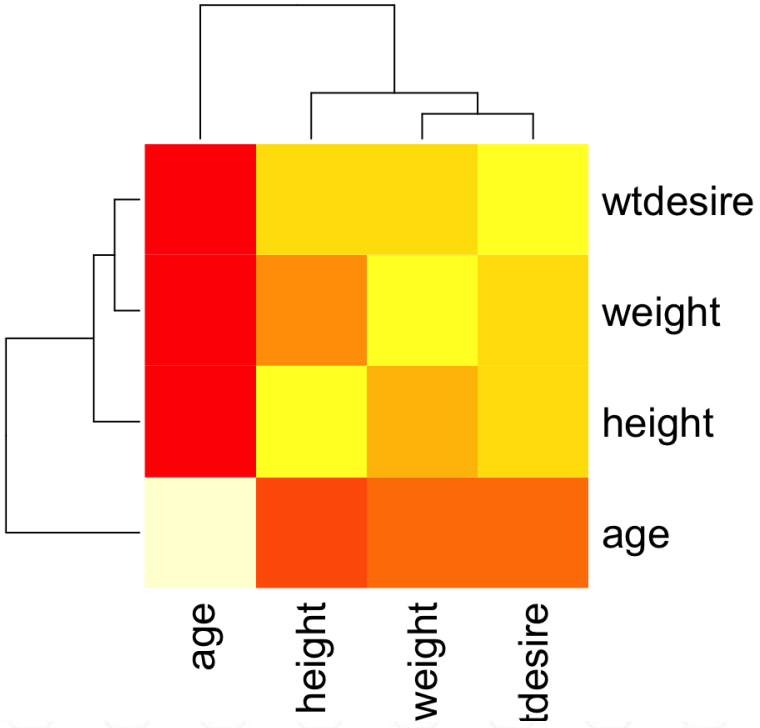
$$Cor(x,y) = \frac{Cov(x,y)}{\sigma(x)\sigma(y)}$$

使用cov 和cor

```
numeric_dataset <- cdc[,c('height', 'weight', 'wtdesire', 'age')]
cor(numeric_dataset)
cov(numeric_dataset)
```

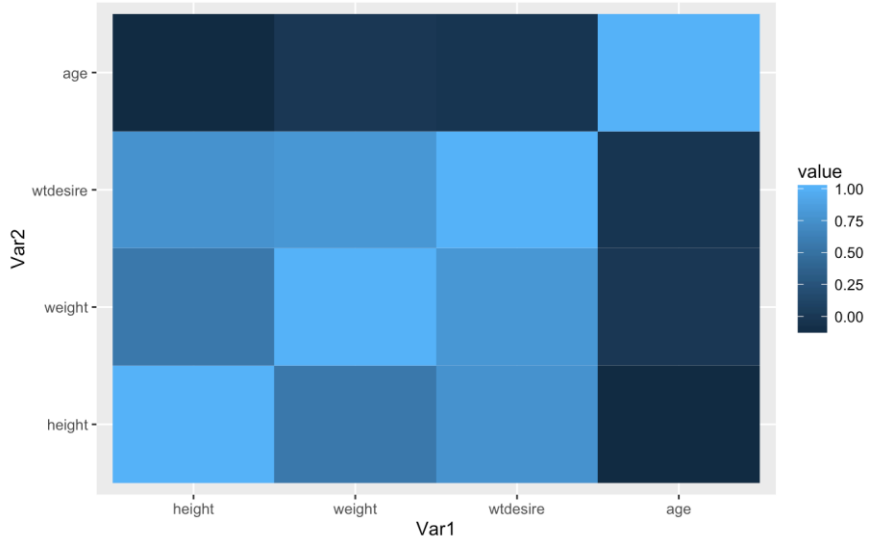
使用Heatmap 繪製相關係數

```
heatmap(cor(numeric_dataset))
```



使用qplot 繪製cor圖

```
library(reshape2)
library(ggplot2)
co <- cor(numeric_dataset)
qplot(x=Var1, y=Var2, data=melt(co), fill=value, geom="tile")
```



線性迴歸

回歸分析

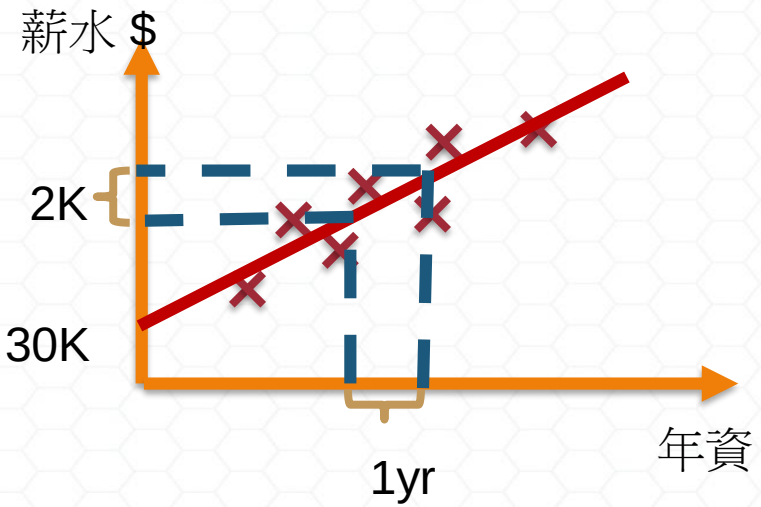
- 線性回歸是研究**單一依變項**(dependent variable)與一個或以上**自變項**(independent variable)之間的關係

- 線性回歸有兩個主要用處：
  - **預測**指的是用已觀察的變數來預測依變項
  - **因果分析**則是將自變項當作是依變項發生的原因
- **Francis Galton 在1886 年發表論文Regression Towards Mediocrity in Hereditary Stature，認為孩子身高跟父親的身高成正相關，而身高的變異數不會隨時間而增加。**

簡單線性回歸

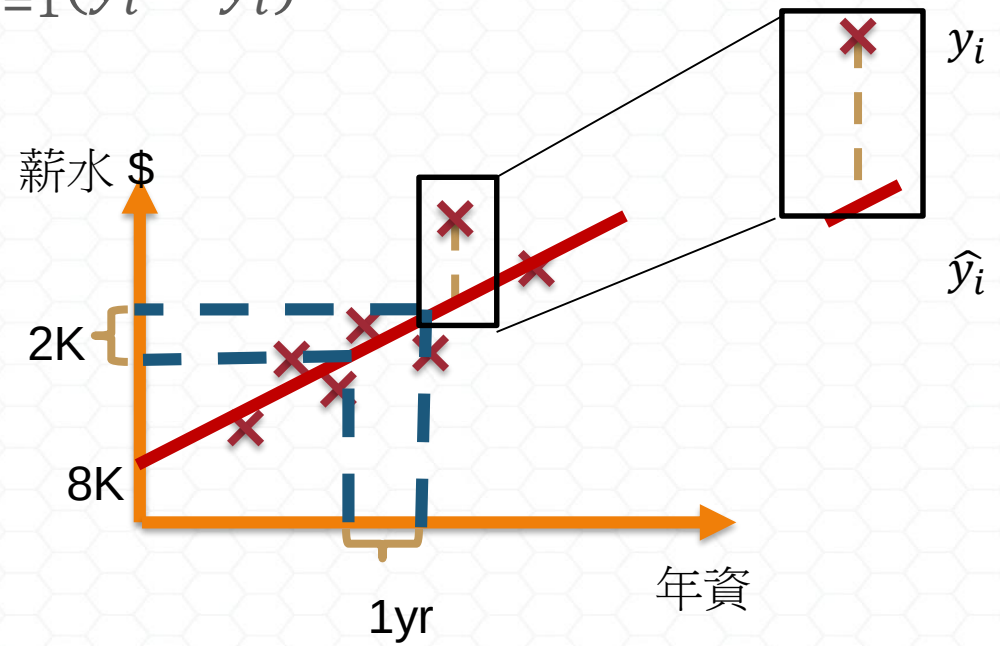
- 數學模型
  - $y = \beta_1 x + \beta_0$
  - y 是依變數
  - x 是自變數
  - β_1 是回歸係數
  - β_0是截距

薪資 $= \beta_1$年資 $+ \beta_0$
β_0 = 30k
β_1 = 2k/年



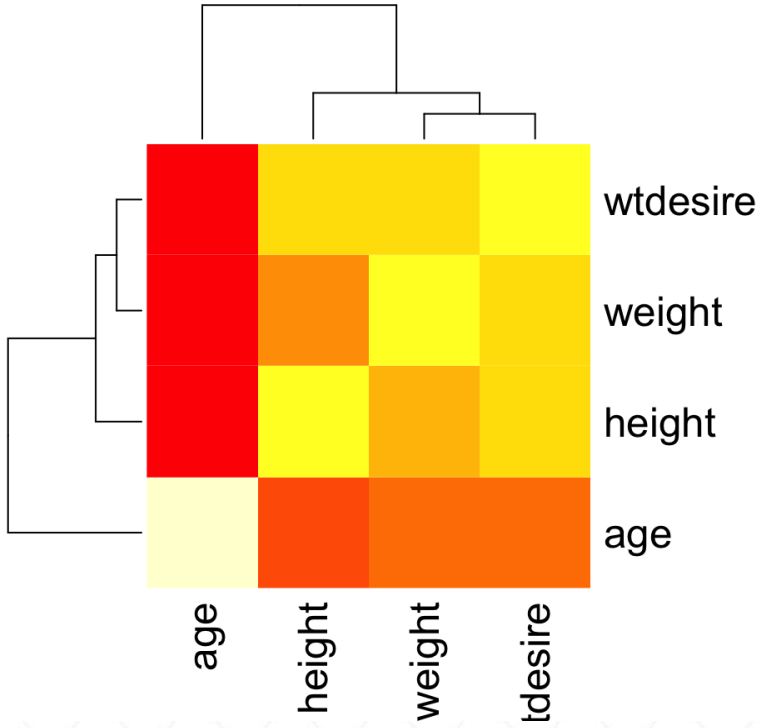
最小平方估計法 - OLS

- 找出殘差平方和最小的一條線
  - 殘差$e_i = (y_i - \hat{y}_i)$
  - 殘差平方和 $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$



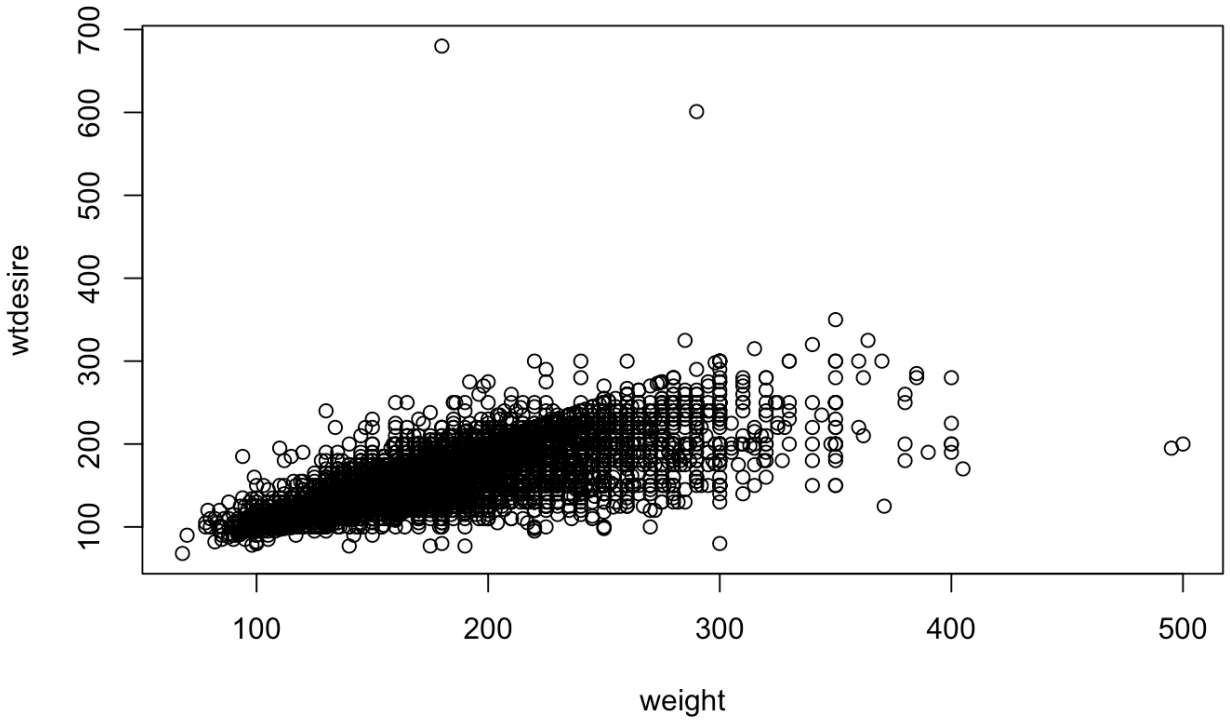
以cor 檢視相關係數

```
co <- cor(numeric_dataset)
heatmap(co)
```



以plot 檢視兩變數的關係

```
plot(wtdesire ~ weight, data = numeric_dataset)
```

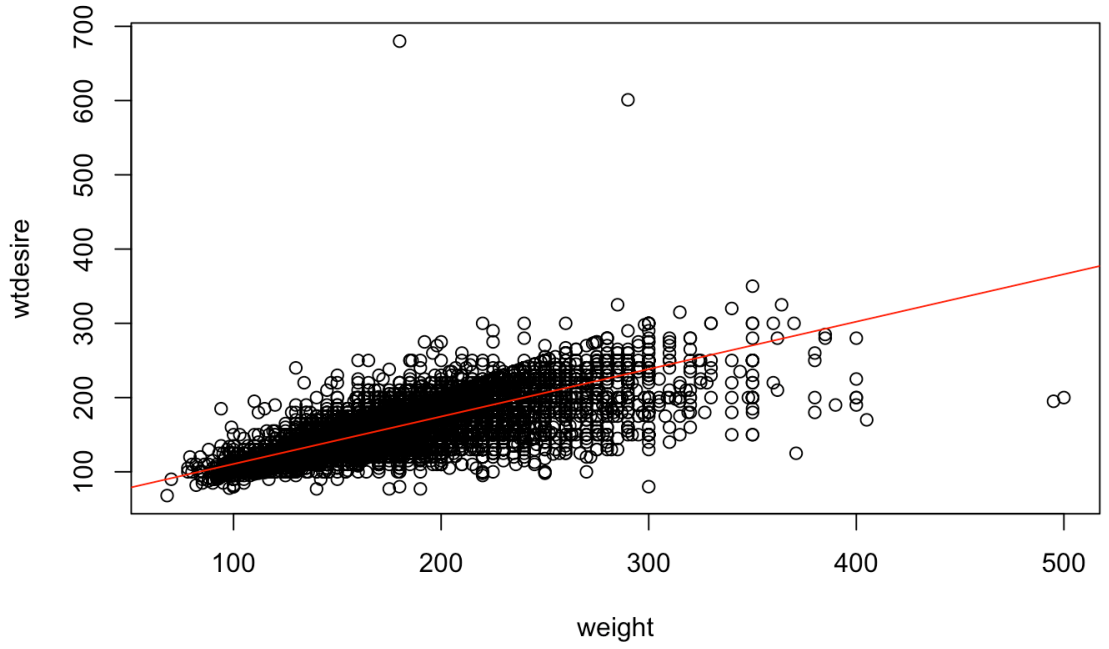


使用lm做回歸分析

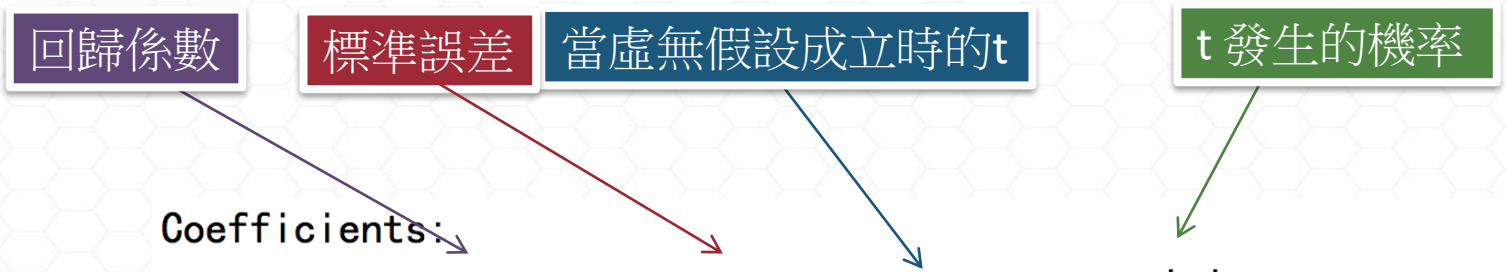
```
fit <- lm(wtdesire ~ weight, data = numeric_dataset)
fit
summary(fit)
```

增添回歸線

```
plot(wtdesire ~ weight, data = numeric_dataset)
abline(fit, col='red')
```



參數估計




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.664015   0.590782   78.99   <2e-16 ***
weight       0.639014   0.003388  188.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.21 on 19998 degrees of freedom
Multiple R-squared:  0.6401,    Adjusted R-squared:  0.6401
F-statistic: 3.556e+04 on 1 and 19998 DF,  p-value: < 2.2e-16
```

參數估計(續)

- 假設顯著性標準是0.01
- 推翻虛無假設的標準是 p 值< 0.01

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.664015   0.590782   78.99   <2e-16 ***
weight       0.639014   0.003388  188.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.21 on 19998 degrees of freedom
Multiple R-squared:  0.6401,    Adjusted R-squared:  0.6401
F-statistic: 3.556e+04 on 1 and 19998 DF,  p-value: < 2.2e-16
```

- t = 188.59, P(>t) =2e-16

驗證兩者關係顯著

基本統計檢定

假設檢定

- 兩種假設： 虛無假設(null hypothesis)與對立假設(alternative hypothesis)
- 進行假設檢定是為了驗證實驗結果為顯著(Significant)。但為了驗證對立假設正確，即實驗變數之間是有相關的，則必須推翻虛無假設(即實驗變數之間毫無關聯)。
- 非A即B 的驗證: 一個為真(對立假設)，另一個即非真(虛無假設)

檢定步驟

- 兩種假設： 虛無假設(null hypothesis)與對立假設(alternative hypothesis)
- 檢定步驟
  - 寫出所有假設
    - H_0原始假設: 觀測值是隨機結果
    - H_1對立假設: 存在一些因素影響結果
  - 檢查統計量
    - 檢驗數量與分佈
  - 檢驗P值
    - 觀察到極端機率有多少，P值越小越不利於原始假設
  - 比較P值與顯著性水準
    - e.g. <0.05 代表顯著

精確二項檢定(binom.test)

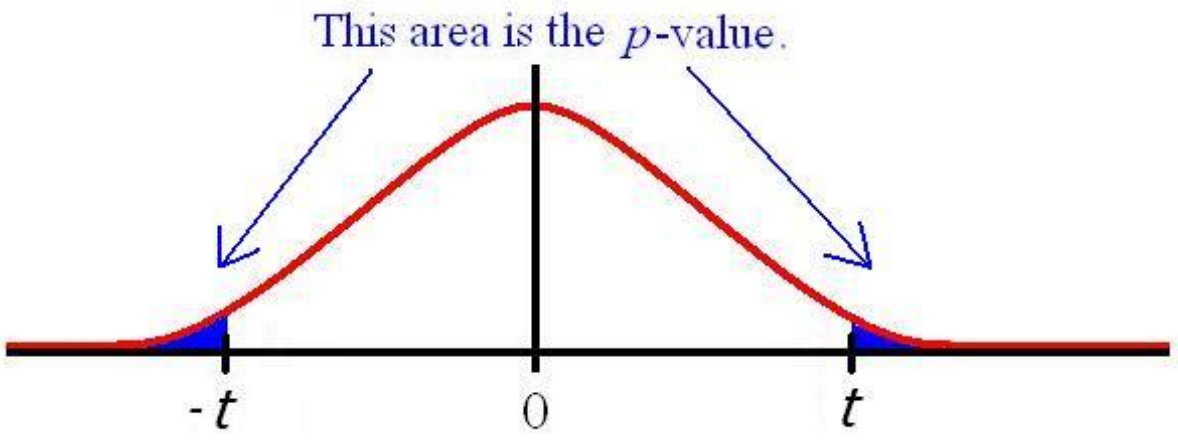
- 假設在一個擲骰子賭局中，賭徒擲到六才能獲勝，要是有一賭徒於315場賭局中，獲勝92場，推斷該賭徒使用的骰子是否公平?

$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$

binom.test(x=92, n=315, p=1/6)

Student's t test

- 用於樣本含量較小（例如n<30），總體標準差σ未知的常態分佈資料。它是用T分佈理論來推斷差異發生的概率，從而判定兩個平均數的差異是否顯著
- T檢驗是William Sealy Gosset了觀測釀酒質量而發明的， Gosset在Guinness釀酒廠擔任統計學家。戈斯特於1908年在Biometrika上公佈T檢驗，但因其老闆認為其為商業機密而被迫使用筆名（Student）



One Sample T-test

- A 商家所生產的霜淇淋引起沙門氏菌病的爆發, 檢驗人員隨機抽取了9個樣本中的沙門氏菌: 量測量分別為 0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418 (來自常態分佈)

- 是否有任何證據顯示該商家所生產之霜淇淋的沙門氏菌的平均量大於 0.3 MPN/g?

檢驗步驟

- 檢定假設 H_0: $\mu = 0.3$, H_1: $\mu > 0.3$.
- 設定 α 檢定水準值 (e.g. 0.05)
- 基準量 $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, 服從自由度為n-1 的t分佈, 其中 $\bar{x}$ 是樣本平均, μ_0 是母體平均, s是樣本標準差對母體標準差的比例, n 是樣本個數
- 計算對應 p-value
- 決定是否推翻虛無假設

實際求出p-value

- 求解是否可以推翻虛無假設

```
x <- c(0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418)
t.test(x, alternative="greater", mu=0.3)
```

- P-value: 0.02927 < 0.05
  - 接受H_1, 拒絕 H_0.
  - 代表很有可能該商家所生產之霜淇淋的沙門氏菌的平均量大於 0.3 MPN/g

Two-sample T-test

- 現有一新藥物, 為了能確定該藥物對患者有效, 我們量測了 6名被給予藥物的受試者(治療組)與另外被給予安慰劑組的6名受試者(對照組) 對刺激的反應時間 (ms)

- 該如何證明被給予藥物的受試者的平均反應時間少於對照組?

檢驗步驟

- 檢定假設
  - H_0: $\mu 1 - \mu 2 = 0$
  - H_1: $\mu 1 - \mu 2 < 0$
- 設定 α 檢定水準值 (e.g. 0.05)

實際求出p-value

■ 假設對照組與實驗組標準差相同

```
Control <- c(91, 87, 99, 77, 88, 91)
Treat <- c(101, 110, 103, 93, 99, 104)
t.test(Control,Treat,alternative="less", var.equal=TRUE)
```

■ 假設對照組與實驗組標準差不同

```
t.test(Control,Treat,alternative="less")
```

Paired T-test

- 當欲比較之實驗組與對照組皆來自相同環境時，如每個試驗單位可分前後期來比較: 宜採用成對t檢定法

- e.g. 想要比較兩種不同藥物對同一組病人是否有治療效果上的差異

實際求出p-value

- 假設受測組為同一組人, 給予兩種不同藥劑

```
druga <- c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
drugb <- c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
t.test(druga,drugb,alternative="greater", paired=TRUE)
```

卡方檢驗

- 如何檢驗兩組類別資料的分佈是否相同?
- 卡方檢定為處理分類並計次資料的統計方法。通常是以觀察次數 (observed frequency, O) 及期望次數 (expected frequency, E) 的比較來進行檢定

卡方檢驗的應用

- 「適合度檢定」
  - 檢查某變數是否依循某比例呈現結果分佈，稱為「適合度檢定」(test of goodness of fit)
  - e.g. 驗證AB基因型之病原菌後代出現AA, AB, BB的基因型是否符合1：2：1之比例。

- 「獨立性檢定」
  - 檢定兩變數間是否相互獨立，稱為「獨立性檢定」(test of independent)，「獨立性檢定」用以判定兩變數獨立與否，亦即判定兩變數是否有關聯，故又稱「關聯性檢定」(test of association)。
  - e.g. 驗證抽菸與得肺癌是否有關聯。

卡方檢驗範例

- 例如，有45名受試者被問及他們更喜歡哪一種篩選測試;10個測試者更喜歡測試A, 15個喜歡測試B, 20個喜歡測試C

- 我們希望驗證3個組別對篩選測試的喜好相同

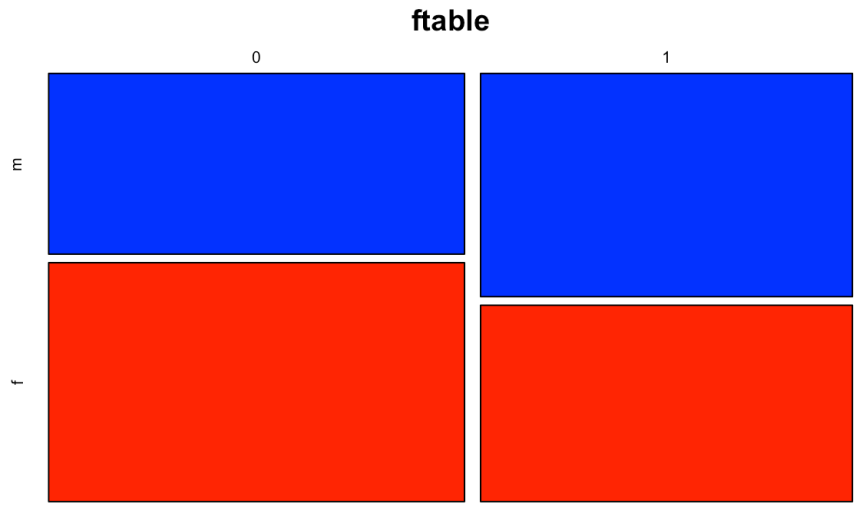
Preference	Observed Frequency	Expected Proportion Under the Null
Test A	10	0.333
Test B	15	0.333
Test C	20	0.333

檢驗 P Value

```
obsfreq <- c(10,15,20)
nullprobs <- c(.333,.333,.334)
chisq.test(obsfreq,p=nullprobs)
```

使用table 產生類別資料統計

```
ftable <- table(cdc$smoke100, cdc$gender)
mosaicplot(ftable, col=c('blue', 'red'))
```



進行卡方檢驗

```
chisq.test(ftable)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  ftable
X-squared = 204.6, df = 1, p-value < 2.2e-16
```

列聯表與獨立性檢驗

- 我們想要比較測試陽性的受試者在三組測試中的陽性結果

	Group 1	Group 2	Group 3
Test Positive	20 (40%)	5 (33.3%	40 (50%0
Test Negative	30	10	40

進行卡方檢定

```
obsfreq <- matrix(c(20,30, 5,10, 40,40),nrow=2,ncol=3)
chisq.test(obsfreq)
```

```
        Pearson's Chi-squared test

data:  obsfreq
X-squared = 2.1378, df = 2, p-value = 0.3434
```

THANK YOU