Chapter2 基本統計方法

Outline

1. 敘述統計
2. 機率分配
3. 抽樣分配
4. 估計
5. 假設檢定
6. 變異數分析

2.1 敘述統計

- 敘述統計（descriptive statistics）的主要目的在於利用統計圖表與數值方法來表達與呈現資料之特性。資料之型態可分為質化型資料（qualitative data）與量化型資料（quantitative data）。
- 質化型資料，主要是依據其性質的不同而加以區分。
- 量化型資料又可分為離散型資料（discrete data）與連續型資料（continuous data）。

3

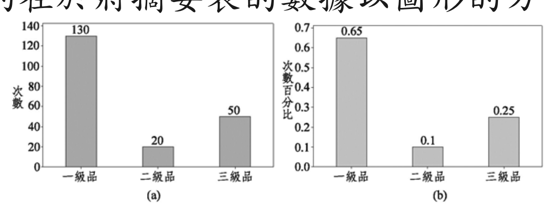
Summary Table & Bar Chart

- Summary Table －Example
  - 收集有200個產品，依照等級做區分並製作摘要表

▼表 2.1　摘要表之例

類別	計次	次數	相對次數
一級品	正正正……	130	0.65
二級品	正……	20	0.1
三級品	正正……	50	0.25
		200	1

- 長條圖之主要目的在於將摘要表的數據以圖形的方式呈現出來。

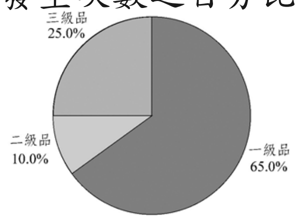


▲圖 2.1　長條圖之例

4

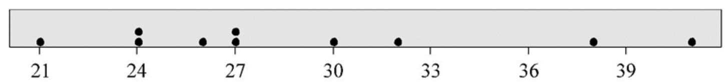
Pie Chart & Dot Plot

- 圓餅圖乃將摘要表用圖形方式顯示出來，其做法主要是將一圓形依照發生次數之百分比進行分割。



▲圖 2.2　圓餅圖之例

- Dot Plot －Example
  - 考慮一筆數據：21, 24, 24, 26, 27, 27, 30, 32, 38, 41



▲圖 2.3　點圖之例

Stem-and-Leaf & Histogram

- 假設某一數值，至少是兩位數，枝葉圖的做法就是將該數值分成兩個部分：「枝」與「葉」。
- 枝葉圖適用於少量資料之描述，當資料量大時我們就須繪製直方圖（histogram）。
- 直方圖的繪製

Step 0. 決定量測變數，蒐集數據，將數據由小至大排列
Step 1. 找出全體數據中之最大值與最小值，計算全距
Step 2. 決定組數
Step 3. 決定組距＝全距／組數
Step 4. 求出各組上、下組界，以及組的中心點
Step 5. 製作次數分配表
Step 6. 繪製直方圖

2	144677
3	028
4	1

▲圖 2.4　枝葉圖之例

Central Tendency

- 集中趨勢主要目的在於描述資料的「中心位置」
  - 樣本平均數 (Sample mean)

    假設某樣本存在n個觀測值$x_1, x_2, \ldots, x_n$，則樣本平均數可計算如下 $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$ (2.1)
  - 中位數 (Median)

    資料從小到大排序後，中位數即為「最中間的數」。換句話說，中位數可將一筆資料切割為相等的兩部分。
  - 眾數 (Mode)即是在一筆資料中，出現次數頻率最多的數。眾數(Mo)可能發生的情況包含單一眾數、無眾數及多重眾數。

Dispersion Tendency(1/2)

- 離散趨勢主要目的在於描述資料的「離散程度」
  - 全距 (Range)

    全距即一組資料中的最大值與最小值的差距

    $$全距 = Max(x_i) - Min(x_i) \quad (2.2)$$
  - 四分位距 (Inter-quartile range)

    將資料由小至大排序並分割成四個等分，每一個切割點稱為四分位數(quartile)。四分位距(Inter-Quartile Range; IQR)即是第一與第三分位數之差，表示如下

    $$IQR = Q_3 - Q_1 \quad (2.3)$$

Dispersion Tendency(2/2)

- 樣本變異數 (Sample variance)：樣本變異數主要是將各個觀測值與樣本平均數之差取平方加總後再取其平均值，表示如下：

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad (2.4)$$

一般而言，樣本變異數s^2越大，樣本變異性就越大。

- 樣本標準差 (Sample standard deviation)

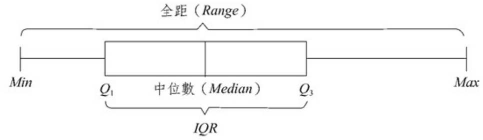
標準差是最常用且最重要的離散趨勢衡量尺度。其表示如下：

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (2.5)$$

使用樣本標準差最大的優點是能以原始的單位作為標準

Box-and-Whisker Plot

- 盒鬚圖（box-and-whisker plot）又稱箱型圖或箱鬚圖，主要目的在利用圖形視覺化資料的集中趨勢、離散趨勢及對稱性。此外也可用以比較兩組以上資料之差異性，以及偵測資料的離群值。
- 盒鬚圖中間的『箱子』是由Q_1及Q_3所組成，因此箱內包含了50%的數值資料，其中*Median*描述了資料中心位置，而*IQR*解釋了中央散佈的情形。此外，『鬚』分別是由箱子延伸至最大及最小值，而*Range*描述了整體資料散佈的情況。



▲圖 2.7　典型盒鬚圖

2.2 Probability Distributions

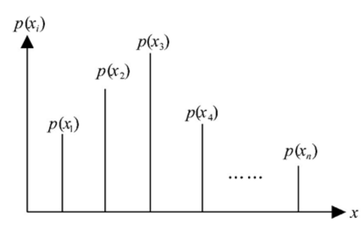
- 機率分配(probability distributions)主要目的在於將品質特性模式化，並解釋其參數行為。
  - 機率分配將觀測之品質特性視為一個隨機變數(Random Variable; R.V.)，因其值在母體中是隨機變動的，因此機率分配即在描述品質特性的某些值在母體中出現的機率分佈情況。
- 機率分配依品質特性的資料型態，可分為
  - 離散型機率分配(discrete probability distributions)與
  - 連續型機率分配(continuous probability distributions)

2.2.1 Discrete Probability Distribution

- 當隨機變數屬於可計數的離散型資料型態，如產品缺點數、書本錯字數、產品不良品數等，其所形成之機率分佈稱為離散型機率分配。
  - 離散型機率分配的形狀類似一釘狀物，其高度即代表隨機變數所對應之機率。當隨機變數X的值為x_i時，其機率表示為

    $$\Pr(X = x_i) = p(x_i)$$

    其中 $0 \le p(x_i) \le 1$，$\sum_{i=1}^{n} p(x_i) = 1$



▲圖 2.10 離散型機率分配

Hypergeometric Distribution

- 超幾何分配（Hypergeometric distribution）常應用於自一批量大小為有限的送驗批中進行抽樣檢驗時，計算貨批之允收機率。
- 假設有一大小為N之有限群體，其中有二類物品，D個為「合格品」而有N - D個為「不合格品」，則超幾何分配的隨機變數（X）代表從N中採不放回方式隨機抽取n個，其中有x個屬於「合格品」之機率分配為：

$$f(X=x)=\Pr(X=x)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad x=0,1,\ldots,\min(n,D) \quad (2.6)$$

- 超幾何分配之平均數與變異數分別為 (2.7)

$$\mu=\frac{nD}{N} \qquad \sigma^2=\frac{nD}{N}(1-\frac{D}{N})(\frac{N-n}{N-1}) \quad (2.8)$$

13

Binomial Distribution

- 伯努力實驗(Bernoulli trial)
  - 即是進行一次實驗，其實驗結果不是「成功」就是「失敗」，每次成功之機率都為p，每次失敗的機率都為$1-p$
- 二項式分配(Binomial distribution)
  - 做了n次獨立的Bernoulli實驗
  - 二項式分配的隨機變數(X)代表n次實驗中成功的次數，其機率分配如下：$f(X=x)=\Pr(X=x)=\binom{n}{x}p^x(1-p)^{n-x} \quad x=0,1,\ldots,n$ (2.9)
  - 母體平均數或稱為期望值為$\mu=E(X)=np$ (2.10)
  - 變異數為$\sigma^2=Var(X)=np(1-p)$ (2.11)

14

Poisson Distribution

- 卜氏分配的隨機變數X表示在某特定單位內事件發生的次數，而所謂某特定單位可以是某特定時間內、某特定面積、某晶圓等。
- 機率分配 $f(X=x)=\Pr(X=x)=\dfrac{\lambda^x e^{-\lambda}}{x!}$ $\quad x=0,1,\ldots,\infty$

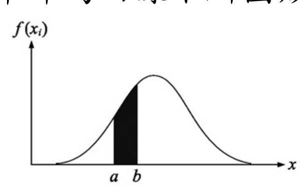
  參數λ表示單位內平均發生之次數，且$\lambda>0$ (2.12)
- 卜氏分配的平均數與變異數相等 (2.13)

  $\mu=E(X)=\lambda \qquad \sigma^2=Var(X)=\lambda$ (2.14)
- 卜氏分配運用層面很廣，常用於等候線理論、品質管制中的缺點數管制圖等。

2.2.2 Continuous Probability Distribution

- 當隨機變數屬於連續型資料型態，其所形成之機率分佈稱為連續型機率分配。
- 不同於離散型機率分配（$f(x)=p(x)$），連續型機率分配的高度（$f(x)$）並不代表隨機變數所對應之機率值（$p(x)$）。
- 連續型機率分配的圖形為一平滑之曲線，假設存在某區間（a到b），則其機率即為曲線下所圍成之面積，表示如下

$$p(a\le x\le b)=\int_a^b f(x)dx$$



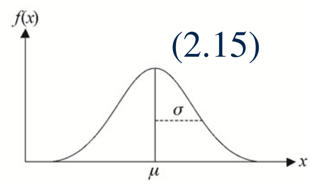
▲圖 2.11 連續型機率分配

Normal Distribution

- 常態分配（Normal distribution）又稱高斯分配（Gaussian distribution）或鐘型分配（Bell shaped distribution）。
- 在推論統計學中，無論是估計、假設檢定、迴歸分析，甚至是變異數分析，無不以常態分配作為基礎。
- 假設存在一連續隨機變數X，則常態分配之機率分配如下：

$$f(X=x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty<x<\infty \qquad (2.15)$$

其中μ為母體平均數，σ為母體標準差



▲圖 2.12　常態分配圖

Standard Normal Distribution

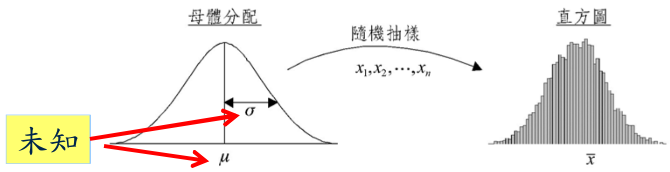
- 欲求常態分配之機率，必須算出函數在某範圍內所圍成之面積，然常態分配之函數相當的複雜，且不易直接進行積分。因此，我們通常的作法就是將常態分配的隨機變數做「標準化」的程序，並使之成為「標準常態分配」
- 所謂標準化的程序，就是針對常態分配之變數(X)進行以下之變數轉換：

$$Z=\frac{X-\mu}{\sigma} \qquad (2.16)$$

  - 轉換過後之變數Z會符合一平均數為0，標準差為1之常態分配($Z \sim N(0, 1)$)。

2.3 Sampling Distribution

- 母體與樣本間之關係



▲圖 2.15　母體與樣本之關係

- 母體分配與抽樣分配之關連性
  - 統計量所形成的機率分配就稱為抽樣分配(sampling distribution)。換言之，抽樣分配就是將樣本統計量視為隨機變數，並考慮所有可能樣本組合而形成之機率分配。

2.3.1 $\overline{X}$ Sampling Distribution

- 假設存在一母體(無論何分配)，其平均數為μ，標準差為σ。隨機自母體抽出一組樣本$x_1, x_2, \ldots, x_n$，則樣本平均數的抽樣分配之平均數與變異數分別如下：

$$\mu_{\overline{X}} = E(\overline{X}) = \mu \qquad (2.17)$$

$$\sigma^2_{\overline{X}} = Var(\overline{X}) = \frac{\sigma^2}{n} \qquad \boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}} \qquad (2.18)$$

- $\overline{X}$ 抽樣分配之標準差是母體標準差的 $1/\sqrt{n}$ 倍。
  $\overline{X}$ 抽樣分配之標準差即代表抽樣誤差(sampling error)。若欲使抽樣誤差為0，則須實行普查。

$$\lim_{n\to\infty} \sigma_{\bar{x}} = \lim_{n\to\infty} \frac{\sigma}{\sqrt{n}} = 0$$

Central Limit Theorem

- 中央極限定理（Central Limit Theorem; CLT）在統計推論中扮演一極重要之角色。

  假設存在一母體(無論分配為何)，其平均數為μ，標準差為σ。今隨機自母體抽出一組樣本$x_1, x_2, \ldots, x_n$，並計算樣本平均數 $\bar{x}$，如果抽樣的樣本數 n 很大 ($n \geqq 30$)，則 $\bar{X}$抽樣分配會接近常態分配。

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- 換言之，當不知道母體分配為何時，只要抽樣的樣本數夠大，則我們可將 $\bar{X}$ 之抽樣分配視為常態分配

2.3.2 *Z* , *t* Distribution

- 自常態母體（$N(\mu, \sigma^2)$）下隨機抽出一組樣本$x_1, x_2, \ldots, x_n$，則其 抽樣分配同樣會符合一常態分配，$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- 標準化 $\bar{X}$，得 $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ 則Z會符合一平均值為0，變異數為1的標準常態分配 Z~$N(0, 1)$。 (2.19)

- 當常態母體μ、σ為已知的情況下，統計量 $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ 會符合標準常態分配。
- 今若母體標準差σ未知，並以樣本標準差s代替σ，則統計量

  $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ 會符合自由度為n-1之t分配，記為 $\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n-1)$ (2.20)

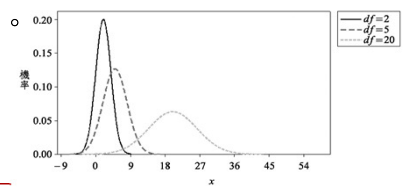
  所謂自由度（Degree of Freedom）是指統計量中隨機變數可以自由變動的數目。

χ^2 Distribution

- 從一常態母體抽出一組樣本$x_1, x_2, \ldots, x_n$，並計算樣本平均數 $\overline{x}$ 與樣本變異數s^2。假設母體平均數μ未知，則統計量 $\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}$ 會符合自由度為n-1之卡方分配

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}=\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1) \qquad (2.21)$$

- 隨著樣本數n的增加，其變異數也會跟著增加。
- 樣本數n越大，χ^2圖形越趨近常態分配。
- 卡方分配常用於機率分配適合度之檢定、獨立性檢定以及對母體變異數σ^2進行推估。



▲圖 2.17 不同自由度下之 χ^2 分配

F Distribution

- 考慮兩獨立常態母體：$X\sim N(\mu_1, \sigma_1^2)$及$Y\sim N(\mu_2, \sigma_2^2)$，自第一個常態母體抽取$n_1$個隨機樣本$x_1, x_2, \ldots, x_{n_1}$，計算出樣本變異數$s_1^2$。另外，自第二個常態母體抽取$n_2$個隨機樣本$y_1, y_2, \ldots, y_{n_2}$，計算出樣本變異數$s_2^2$。則統計量$(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$會符合$F$分配

(n1-1,n2-1)

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}=\frac{\chi^2_{n_1-1}/n_1-1}{\chi^2_{n_2-1}/n_2-1}\sim F(\nu_1,\nu_2) \qquad (2.22)$$

其中ν_1、ν_2分別為分子與分母卡方分配之自由度。

- F分配常用於變異數分析以及對兩獨立母體變異數比值進行推論

2.3.3 Sampling From a Bernoulli Population

- 假設有一隨機變數Y，若y=1代表成功，y＝0 代表失敗，則伯努力機率分配（Bernoulli distribution）如下：

$$f(Y=y)=p^y(1-p)^{1-y} \quad , \; y=0 \;\; or \; 1 \qquad (2.23)$$

- 隨機自伯努力分配母體抽出一組樣本$x_1, x_2, \ldots, x_n$，則樣本之總和（即成功之次數）$X=x_1+x_2+\ldots+x_n$會符合一參數為 n 及 p 之二項式分配。
- 母體參數p 通常為未知，則可利用樣本比例進行估計，如下：

$$\hat{p}=\frac{\text{出現「成功」次數}}{\text{實驗次數}}=\frac{X}{n}=\frac{x_1+x_2+\cdots+x_n}{n}$$

- 抽樣分配之平均數為 $\mu_{\hat{p}}=E(\hat{p})=p$ (2.24)

變異數為 $\sigma_{\hat{p}}^2=Var(\hat{p})=\dfrac{p(1-p)}{n}$ (2.25)

2.4 Estimation

2.4.1 點估計

- 點估計之目的在於利用樣本資料，求得一估計值以表示未知母體參數值。通常以「^」符號表示參數的點估計值。
- 一個好的點估計量（point estimator）須具許多重要之性質，如充分性（sufficient）、不偏性（unbiased）、效率性（efficiency）、一致性（consistence）等；而最重要的性質為不偏性。
- 所謂不偏性指的就是點估計量的期望值必須等於被估計的參數值。假設有一母體參數θ，其點估計值為 $\hat{\theta}$，若$E(\hat{\theta})=\theta$，則 $\hat{\theta}$為一不偏估計量(unbiased estimator)。

Unbiased Estimator

- μ 不偏估計量 $E(\overline{X})=\mu$ (2.28)
  - 樣本平均數 $\overline{x}$ 為母體平均數 μ 的不偏估計量。
- σ^2 不偏估計量 $E(s^2)=\sigma^2$ (2.29)
  - 樣本變異數 s^2 為母體變異數 σ^2 的不偏估計量。
- σ 不偏估計量 (利用樣本標準差) $E(s)=c_4\sigma$ (2.30)
  - s^2 為 σ^2 的不偏估計量，但 s 並非母體標準差 σ 之不偏估計量。
  - 利用 s 估計 σ 之不偏估計量為 $\boxed{\hat{\sigma}=\overline{s}/c_4}$ (重要) (2.31)
- σ 不偏估計量 (利用全距)
  - 隨機變數 $W=R/\sigma$ 稱為相對全距(relative range)，則 W 抽樣分配之期望值為 $E(W)=E(R/\sigma)=d_2$ (2.32)
  - 利用 R 估計 σ 之不偏估計量為 $\boxed{\hat{\sigma}=\overline{R}/d_2}$ (2.34)

2.4.2 Interval Estimation

- 主要目的在於根據樣本資料所得之點估計值，並利用點估計量的抽樣分配建構一區間範圍，以使得參數之真實值落在此區間內之統計方法。
- 我們希望所建構之區間能夠包含真正的母體參數值譬如，利用點估計值 $\overline{x}$ 及 $\overline{X}$ 抽樣分配建構一區間來包含常態母體之平均數 μ，可表示如下：
  - 我們希望估計誤差會小於某個允差 $|\overline{X}-\mu|\le tolerance$
  - 因為 $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ ，將允差設為 $Z\dfrac{\sigma}{\sqrt{n}}$
  - 所以 $|\overline{x}-\mu|\le Z\dfrac{\sigma}{\sqrt{n}}$ 即 $\boxed{\overline{x}-Z\dfrac{\sigma}{\sqrt{n}}\le\mu\le\overline{x}+Z\dfrac{\sigma}{\sqrt{n}}}$

(重要)

Confidence Interval

- 基本上，所建構出之區間只有「包含」與「不包含」真值 。區間越寬，涵蓋真值的機率就越大。因此，建構區間前必須先決定信賴水準(confidence level)，1-α。
- 由 $\Pr\left(-Z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right) = \Pr\left(\overline{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ (2.35)

  所產生之區間 $\boxed{\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$ (2.36)

  稱為未知平均數μ的(1-α)100%信賴區間(confidence interval)
- 常用之信賴水準

信賴水準 (1-α)100%	α	$\alpha/2$	$Z_{\alpha/2}$
90%	0.10	0.05	$Z_{0.05}$ = 1.64
95%	0.05	0.025	$Z_{0.025}$ = 1.96
99%	0.01	0.005	$Z_{0.005}$ = 2.58

(重要)

2.4.3 Single Population Mean

- σ Known－從一常態母體抽出一組樣本$x_1, x_2, \ldots, x_n$，並計算樣本平均數$\bar{x}$，假設母體標準差已知，則μ的(1-α)100% 信賴區間為 $\boxed{\bar{x}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}$ (2.37)
- σ Unknown－自常態母體抽取一組樣本$x_1, x_2, \ldots, x_n$，計算樣本平均數 $\bar{x}$，若樣本數n ≧30（大樣本），且母體標準差σ未知，以s代替，則μ的(1- α)%信賴區間為：

$$\boxed{\bar{x}-Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}} \quad (2.38)$$

- σ Unknown－自常態母體抽取一組樣本$x_1, x_2, \ldots, x_n$，計算樣本平均數 $\bar{x}$，若樣本數n < 30（小樣本），且母體標準差σ未知，則μ的(1- α)%信賴區間為：

$$\boxed{\bar{x}-t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x}+t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}} \quad (2.39)$$

Single Population Variance & Proportion

- 自常態母體抽取一組樣本$x_1, x_2, \ldots, x_n$，計算樣本變異數s^2，則 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

$$Pr\left(\chi^2_{n-1,1-\alpha/2} \le \chi^2 \le \chi^2_{n-1,\alpha/2}\right) = 1-\alpha \quad \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \qquad (2.40)$$

σ的(1- α)% 信賴區間為：

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}} \qquad (2.41)$$

- 由中央極限定理得知，當樣本數n ≧30，則 $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

  因此，$\hat{p}$的(1- α)%信賴區間為

$$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (2.42)$$

2.4.4 Two Population Means: Matched Samples

- 獨立樣本(independent samples)指的是從一母體抽出的樣本不會影響到從另一個母體所抽出的樣本。
- 成對樣本(matched samples)指的是母體分配彼此為相關，或在抽樣時樣本間具有「成對」的情形。
- 假設有兩相關常態母體$X \sim N(\mu_1, \sigma_1^2)$與$Y \sim N(\mu_2, \sigma_2^2)$，其差值仍符合一常態分配。今抽取一組樣本大小為$n$之樣本，計算其差值$d_1, d_2, \ldots, d_n$及樣本平均數$\overline{d}$ 與樣本標準差s_d
  - 若n ≧30 (大樣本)，則兩母體平均數差μ_1-μ_2的 (1-α)100%信賴區間為：
    $$\overline{d} - Z_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}} \le \mu_1 - \mu_2 \le \overline{d} + Z_{\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}} \qquad (2.43)$$
  - 若$n < 30$ (小樣本)，則兩母體平均數差μ_1-μ_2的 (1-α)100%信賴區間為：
    $$\overline{d} - t_{\frac{\alpha}{2}}(n-1)\frac{s_d}{\sqrt{n}} \le \mu_1 - \mu_2 \le \overline{d} + t_{\frac{\alpha}{2}}(n-1)\frac{s_d}{\sqrt{n}} \qquad (2.44)$$

Two Independent Population Means:
(Variance Known & Variance Unknown, Large Sample Size)

- 從兩獨立常態母體（$X\sim N(\mu_1, \sigma_1^2)$與$Y\sim N(\mu_2, \sigma_2^2)$ ）分別抽出一組樣本$x_1, x_2, \ldots, x_{n_1}$和$y_1, y_2, \ldots, y_{n_2}$，並分別計算樣本平均數$\bar{x}$與$\bar{y}$，若母體標準差未知並計算之
- 若兩母體標準差σ_1與σ_2已知，則兩母體平均數差μ_1-μ_2的 (1-α)100% 信賴區間為：

$$(\bar{x}-\bar{y})-Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{x}-\bar{y})+Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \quad (2.45)$$

- 若兩母體標準差σ_1與σ_2未知，且$n_1 \geqq 30$與$n_2 \geqq 30$，則兩母體平均數差μ_1-μ_2的 (1-α)100% 信賴區間為：

$$(\bar{x}-\bar{y})-Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{x}-\bar{y})+Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \quad (2.46)$$

33

Two Independent Population Means:
(Variance Unknown, Small Sample Size, Homogeneity of Variances)

- 從兩獨立常態母體（$X\sim N(\mu_1, \sigma_1^2)$與$Y\sim N(\mu_2, \sigma_2^2)$ ）分別抽出一組樣本$x_1, x_2, \ldots, x_{n_1}$和$y_1, y_2, \ldots, y_{n_2}$ ($n_1 < 30$ 與 $n_2 < 30$)，並分別計算樣本平均數$\bar{x}$、$\bar{y}$ 與樣本標準差s_1、s_2。 若兩母體標準差σ_1與σ_2未知，但具同質性(homogeneity)，即$\sigma_1 = \sigma_2 = \sigma$，則可求出共同樣本標準差(pooled standard deviation)：

$$S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \quad (2.47)$$

- 而兩母體平均數差μ_1-μ_2的 (1-α)100% 信賴區間為：

$$(\bar{x}-\bar{y})-t_{\frac{\alpha}{2}}(n_1+n_2-2)S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le (\bar{x}-\bar{y})+t_{\frac{\alpha}{2}}(n_1+n_2-2)S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \quad (2.48)$$

34

Two Independent Population Means:

(Variance Unknown, Small Sample Size, Heterogeneity of Variances)

- 從兩獨立常態母體（$X\sim N(\mu_1, \sigma_1^2)$與$Y\sim N(\mu_2, \sigma_2^2)$）分別抽出一組樣本$x_1, x_2, \ldots, x_{n_1}$和$y_1, y_2, \ldots, y_{n_2}$ ($n_1 < 30$ 與 $n_2 < 30$)，並分別計算樣本平均數$\bar{x}$、$\bar{y}$ 與樣本標準差s_1、s_2。若兩母體標準差σ_1與σ_2未知，也不具同質性(heterogeneity)，即$\sigma_1 \neq \sigma_2$，則兩母體平均數差μ_1-μ_2的 (1-α)100% 信賴區間為：

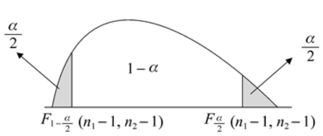
$$(\bar{x}-\bar{y}) - t_{\frac{\alpha}{2}}(v)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1-\mu_2 \le (\bar{x}-\bar{y}) + t_{\frac{\alpha}{2}}(v)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \quad (2.49)$$

而其中自由度為：
$$v = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} \quad (2.50)$$

Two Population Variances

- 從兩獨立常態母體（$X\sim N(\mu_1, \sigma_1^2)$與$Y\sim N(\mu_2, \sigma_2^2)$）分別抽出一組樣本$x_1, x_2, \ldots, x_{n_1}$和$y_1, y_2, \ldots, y_{n_2}$，並分別計算樣本變異數$s_1^2$與$s_2^2$。則

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,\, n_2-1}$$



F分配

- 因為 $p\left(F_{1-\alpha/2, n_1-1, n_2-1} \le F \le F_{\alpha/2, n_1-1, n_2-1}\right) = 1-\alpha$

兩母體變異數比值s_1^2/s_2^2的(1-α)100%信賴區間為：

$$\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} \quad (2.51)$$

Two Population Proportions

- 從兩獨立伯努力分配分別抽出一組樣本$x_1, x_2, \ldots, x_{n_1}$和$y_1, y_2, \ldots, y_{n_2}$。令 $X = x_1+x_2+\ldots+x_{n_1}$，$Y = y_1+y_2+\ldots+y_{n_2}$，樣本比例分別為： $\hat{p}_1 = X/n_1 \quad \hat{p}_2 = Y/n_2$
- 由中央極限定理得知，當樣本數$n \geqq 30$，則

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \hat{p}_2 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- 將 $\hat{p}_1 - \hat{p}_2$進行標準化，得 $Z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$
- $p_1 - p_2$的(1-α)100% 信賴區間為：

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{2.52}$$

2.5 Hypothesis Testing

- 假設檢定（hypothesis testing）就是對有關母體參數進行「假設」，並利用樣本的資訊，「檢定」是否拒絕該假設的一個統計方法。
- 虛無假設 (null hypothesis) H_0

  Ex: “水果冷凍後水分不流失” ，即 H_0：$\mu = 0$
- 對立假設 (alternative hypothesis) H_1或H_a

  Ex: “水果冷凍後水分衰減” ，即 H_1：$\mu < 0$

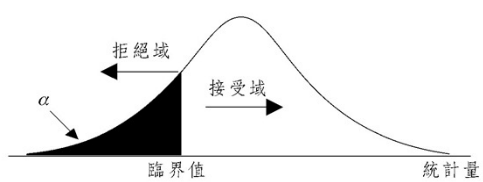
- α: 生產者風險(producer′s risk)
  β: 消費者風險(customer’s risk)
- $1-\beta$: 檢定力 (*power of test*)
  當產品為不良品時, 真正能夠正確判斷它為不良品的機率。

▼表 2.6　母體真正情況與檢定結果

		母體真正狀況	
		H_0 為真	H_0 不為真
決策	拒絕 H_0	型 I 誤差 (α)	裁決正確 ($1-\beta$)
	不拒絕 H_0	裁決正確 ($1-\alpha$)	型 II 誤差 (β)

2.5.2 Testing Procedure

- 當抽出樣本值後，我們會選擇適當的檢定統計量(Test statistic)，並計算其值，若此統計量之值落在接受域(acceptance region)，則不拒絕H_0。反之，若檢定統計量之值不落在接受域，則拒絕H_0。
- 檢定可分為單邊(one-sided)檢定與雙邊(two-sided)檢定，完全取決於對立假設。



H_0：$\mu=0$

H_1：$\mu<0$

▲圖 2.21　臨界值之決定

The Process of Hypothesis Testing

1. 設立虛無假設與對立假設。
2. 選擇檢定統計量。
3. 給定顯著水準α，並獲得檢定的臨界值。
4. 計算檢定統計量之值。
5. 制訂決策（拒絕或不拒絕H_0）。

2.5.3 Hypothesis Testing for a Single Population Mean

- 考慮單一母體平均數等於某一特定值(μ_0)，
  - 雙邊檢定之假設為 $H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0$ (2.57)
  - 右尾檢定之假設為 $H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0$ (2.58)
  - 左尾檢定之假設為 $H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0$ (2.59)
- 1. σ已知，以Z為檢定統計量，並進一步計算檢定值如下
$$Z^* = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad (2.60)$$
- 2. σ未知，樣本數$n \geqq 30$，則以Z為檢定統計量，並計算檢定值
$$Z^* = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad (2.61)$$
- 3. σ未知，樣本數$n < 30$，則以 t 為檢定統計量，並計算檢定值
$$t^* = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \quad (2.62)$$

範例 2.9

單一母體變異數、比例與*P*-value

- 單一母體變異數
  - 雙邊檢定的假設為 H_0：$\sigma = \sigma_0$；H_1：$\sigma \neq \sigma_0$ (2.63)
  - 以χ^2為檢定統計量，並計算檢定值 $\chi^{2*} = \dfrac{(n-1)s^2}{\sigma_0}$ (2.64)
- 單一母體比例
  - 雙邊檢定為 H_0：$p = p_0$；H_1：$p \neq p_0$ (2.65)
  - 以Z為檢定統計量，並計算檢定值 $Z^* = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$ (2.66)
- *P*-value
  - 若*P*-value$<\alpha$，則拒絕虛無假設H_0；若*P*-value$>\alpha$，則不拒絕。
  - 假設Z^*為檢定統計量的值，則*P*-value計算如下

$$P-\text{value} = \begin{cases} 2[1-\Pr(Z<Z^*)] & \text{雙尾檢定:} H_1 : \mu \neq \mu_0 \\ 1-\Pr(Z<Z^*) & \text{上尾檢定:} H_1 : \mu > \mu_0 \\ \Pr(Z<Z^*) & \text{下尾檢定:} H_1 : \mu < \mu_0 \end{cases} \quad (2.67)$$

Hypothesis Testing for Two Population Means

- 兩母體平均數
  - 雙邊檢定之假設為 H_0：$\mu_1 = \mu_2$；H_1：$\mu_1 \neq \mu_2$ (2.68)
  - 右尾檢定之假設為 H_0：$\mu_1 = \mu_2$；H_1：$\mu_1 > \mu_2$ (2.69)
  - 左尾檢定之假設為 H_0：$\mu_1 = \mu_2$；H_1：$\mu_1 < \mu_2$ (2.70)
- 兩母體變異數
  - 雙邊檢定之假設為 H_0：$\sigma_1^2 = \sigma_2^2$；H_1：$\sigma_1^2 \neq \sigma_2^2$ (2.71)
  - 計算樣本標準差s_1與s_2，則F檢定統計量之值為 $F^* = s_1^2 / s_2^2$ (2.73)
- 兩母體比例
  - 雙邊檢定之假設為 H_0：$p_1 = p_2$；H_1：$p_1 \neq p_2$ (2.74)
  - 計算檢定統計量之值 $Z^* = \dfrac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ (2.75)

2.6 Analysis of Variance

- 變異數分析ANOVA為 **AN**alysis **O**f **VA**riance 之縮寫，其目的是用來檢定三個或三個以上**母體平均數**是否相等的假設。
- “變異數分析”這個名詞似乎並不恰當，因為我們要檢定的是母體平均數而非變異數，然而事實上，變異數分析的檢定過程是根據樣本資料的變異分析為基礎的。
- 使用變異數分析檢定多個(a個)母體平均數差異時，其假設如下

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a \quad (2.76)$$

H_1 : 母體平均數並不全相等

2.6.2 Variation

- 變異數分析將總變異（total variation）區分為:
  - 組間變異（between variation）
  - 組內變異（within variation）。
- 總變異=組間變異+組內變異
- 總平方和 ＝ 因子平方和 ＋ 誤差平方和
- $SST = SSF + SSE$

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 \quad (2.77)$$

$$SSF = \sum_{i=1}^{a} n_i(\bar{x}_i - \bar{\bar{x}})^2 \quad (2.78) \qquad SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \quad (2.79)$$

2.6.3 Principle of ANOVA

- 因子變異(between variation)：為可解釋的變異
  隨機變異(within variation)：為不可解釋之變異
- 變異數分析的方法即是利用樣本統計量(F)來比較因子變異和隨機變異的大小，以檢定因子所引起的變異是否大到足以拒絕虛無假設。
- SSF 和 SSE 會受樣本個數多少的影響，因此必須進一步求平均變異。將平方和除以其對應的自由度，所得稱之為均方(mean square)，因此有：

$$MSF = \frac{SSF}{a-1} \qquad MSE = \frac{SSE}{N-a}$$ 其中 N: 所有樣本數 a: 母體個數

- 檢定統計量 $F = \dfrac{MSF}{MSE}$

表2.7 ANOVA

變異來源	自由度 (df)	平方和 (Sum of Square)	均方 (Mean Square)	F*
組間 (between)	a-1	SSF	MSF =SSF / (a-1)	$\frac{MSF}{MSE}$
組內 (within)	N-a	SSE	MSE =SSE / (N-a)	
總和 (total)	N-1	SST		

SSF: sum of squares due to factor
SSE: sum of squares due to error
SST: total sum of squares

補充篇:變異數分析

Reference:
書名：統計學：觀念、方法、應用 3/e
作者：賀力行、林淑萍、蔡明春
出版：前程文化事業有限公司

名詞介紹(1/3)

- 變異數分析是用來檢定兩個以上平均數是否相等或某個變數是否受某些因子所影響之統計方法。

例如：

(1)不同的行銷策略是否會影響產品之銷售量？(不同的行銷策略，其產品之平均銷售量是否相等？)

(2)不同的教育程度與不同的性別對工作滿意度是否有影響？(不同的教育程度與不同的性別之員工，其平均之工作滿意度是否相等)

名詞介紹(2/3)

- 變異數分析常用之名詞：

(1)實驗單位(experiment unit)：實驗所衡量的對象。
例如： 產品、員工為其實驗單位。

(2)因子(factor)：研究者所控制調整的因素。
例如：行銷策略、教育程度為其因子。

名詞介紹(3/3)

(3)處理方法(treatment)：因子之各種水準或類別。例如：不同的行銷策略、不同的教育程度、不同的性別，如不同性別中的男、女為兩種不同的處理方法。
(4)依變數(dependent variable)：實驗單位對不同處理方法的反應變數。例如：銷售量、工作滿意度為其依變數。

單因子變異數分析—完全隨機設計

- 由定理13-3得知，進行變異數分析需滿足以下基本假設條件：

(1)常態母體：各組樣本需取自於常態母體。

(2)變異數具同質性：各組母體變異數需假設相等。而變異數是否具同質性，可利用樣本變異數檢定之。

(3)獨立性：各組樣本彼此獨立。

表13.2 單因子變異數分析表

變異來源	平方和	自由度	均方	f_0 值
處理方法 隨機誤差	*SSB* *SSE*	*k-1* *n-k*	*MSB* *MSE*	$\frac{MSB}{MSE}$
總和	*SST*	*n-1*		

例題 13.1

- 某市場調查公司欲調查市面上四種品牌之相同口味飲料之平均銷售量是否相同，於是由每一品牌隨機選定5個地區作調查，得其每個地區一個月之銷售量如下表(單位：千箱)

品牌			
A	B	C	D
26.5	29.0	26.9	30.5
28.7	27.6	28.3	31.2
25.2	25.4	27.8	29.9
29.3	28.3	26.2	28.1
25.3	29.7	25.8	30.3

例題 13.1(續)

(1)請寫出此問題之假設。

(2)請寫出此問題之變異數分析表。

(3)請根據(2)之結果，以 α=0.05 檢定此四種品牌飲料之平均銷售量是否相等。

解

(1)令 μ_i 表第 i 種品牌銷售量之平均數，則此問題之假設為

H_0: $\mu_A=\mu_B=\mu_C=\mu_D$ & H_1: μ_A、μ_B、μ_C、μ_D不全相等

(2)每種品牌之樣本平均數 $\overline{X}_A$、$\overline{X}_B$、$\overline{X}_C$、$\overline{X}_D$ 及總樣本平均數 $\overline{X}$ 如下：

$$\overline{X}_A=\frac{1}{5}(26.5+28.7+25.2+29.3+25.3)=27$$

$$\overline{X}_B=\frac{1}{5}(29.0+27.6+25.4+28.3+29.7)=28$$

$$\overline{X}_C=\frac{1}{5}(26.9+28.3+27.8+26.2+25.3)=27$$

57

解

$$\overline{X}_D=\frac{1}{5}(30.5+31.2+29.9+28.1+30.3)=30$$

$$\overline{X}=\frac{1}{20}(5\times 27+5\times 28+5\times 27+5\times 30)=28$$

經計算後可得

$$\begin{aligned}SST&=(26.5-28)^2+(28.7-28)^2+(25.2-28)^2+(29.3-28)^2\\&+(25.3-28)^2+(29.0-28)^2+(27.6-28)^2+(25.4-28)^2\\&+(28.3-28)^2+(29.7-28)^2+(26.9-28)^2+(28.3-28)^2\\&+(27.8-28)^2+(26.2-28)^2+(25.8-28)^2+(30.5-28)^2\\&+(31.2-28)^2+(29.9-28)^2+(28.1-28)^2+(30.3-28)^2\\&=65.28\end{aligned}$$

58

解

$$SSB = 5\times(27-28)^2 + 5\times(28-28)^2 + 5\times(27-28)^2 + 5\times(30-28)^2 = 30$$

$$SSE = SST - SSB = 65.28 - 30 = 35.28$$

SSB之自由度為4-1=3，因此 $MSB = \frac{SSB}{k-1} = \frac{30}{3} = 10$

SSE之自由度為 *n*-*k* = 20-4 = 16，因此

$$MSE = \frac{SSE}{n-k} = \frac{35.28}{16} = 2.205$$

由此可得， $f_0 = \frac{MSB}{MSE} = \frac{10}{2.205} = 4.535$

解 **其變異數分析表如下：**

變異來源	平方和	自由度	均方	f_0 值
處理方法	30	3	10	4.535
隨機誤差	35.28	16	2.205	
總和	65.28	19		

(3)因為 $\frac{MSB}{MSE} \sim F(3,16)$ **，因此其拒絕域** $\{f_0 \geq f_{0.05}(3,16)\} = \{f_0 \geq 3.239\}$

而檢定值 $f_0 = 4.535 > 3.239$ **落在拒絕域中，因此拒絕** H_0**，即四種不同品牌飲料之平均銷售量有顯著地差異。**

例題 13.2

某研究人員想瞭解*A*、*B*、*C*三種不同廠牌1800c.c.汽車之耗油率，於是此研究人員蒐集了資料，並以完全隨機設計方式蒐集資料並計算得到以下之變異數分析表，如下表所示。

變異來源	平方和	自由度	均方	f_0 值
處理方法	20	2	?	?
隨機誤差	50	27	?	
總和	70	29		

例題 13.2(續)

(1)請完成此變異數分析表。

(2)請以 α=0.05 來檢定H_0: $\mu_A = \mu_B = \mu_C$ (μ_i 表第 *i* 種品牌汽車平均每公升汽車可行駛之里程數)是否成立

解

(1)因為 $MSB = \frac{SSB}{k-1} = \frac{20}{2} = 10$ $MSE = \frac{SSE}{n-k} = \frac{50}{27} = 1.85$

所以 $f_0 = \frac{MSB}{MSE} = \frac{10}{1.85} = 5.41$

因此其完整變異數分析表如下所示：

變異來源	平方和	自由度	均方	f_0 值
處理方法	20	2	10	5.41
隨機誤差	50	27	1.85	
總和	70	29		

13.2 單因子變異數分析—完全隨機設計(16/16)

解

(2)因為 $\frac{MSB}{MSE} \sim F(2,27)$，因此其拒絕域 $\{f_0 \geq f_{0.05}(2,27) = 3.35\}$

而檢定值 $f_0 = 5.41 > 3.35$ 落在拒絕域中，因此拒絕 H_0，即三種品牌1800c.c.汽車之耗油率有顯著地差異。