應用統計學
Applied Statistics

顏嘉華

Jai-Hua Yen

Email: r06621209@ntu.edu.tw

Dec 6, 2020

目錄

第一章 統計導論

1.1 緒論

統計學的意義: 統計學是對於樣本資料的蒐集、整理、呈現並解釋的一門科學。透過選擇適當的機率模型，對母體未知的特性進行推論。在不確定的條件下，可以做成較有利決策的方法。

1.2 集中趨勢量數

研究者常以某些具體的數值來表示一筆資料或分配的特徵。

定義: 描述樣本資料特徵的數值稱為統計量(statistic)，而描述母體資料特徵的數值稱為參數(parameter)。

常用的集中趨勢量數為算術平均數、中位數、眾數、幾何平均數。

1.2.1 算術平均數

為最簡單且最易了解的集中量數。意義上為將所有的數值資料進行總和後，除上資料的個數。

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，則算術平均數(arithmetic mean, $\bar{x}$)為

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ 。}$$

算術平均數的性質為:

(1) 數值資料中每一筆資料與平均數差的總和為 0，即

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \text{。}$$

(2) 數值資料中每一筆資料與平均數差的平方和為最小，即

$$\min_{a} \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ 。}$$

(3) $y_i = ax_i + b \rightarrow \bar{y} = a\bar{x} + b$。

(4) 容易受極端質影響。

(5) 資料為雙峰時，算術平均數無法代表集中趨勢量數。

1.2.2 中位數

為最簡單且最易了解的集中量數。意義上為將所有的數值資料進行總和後，除上資料的個數。

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，經由排序後得$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，則中位數(median, M_e)為

$$M_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n\text{為奇數} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2}, & n\text{為偶數} \end{cases} 。$$

中位數的性質如下:

(1) 數值資料中每一筆資料與中位數的距離總和為最小，即

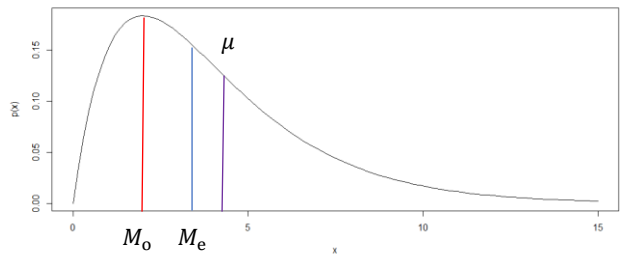
$$\min_a \sum_{i=1}^{n} |x_i - a| = \sum_{i=1}^{n} |x_i - M_e| 。$$

(2) 中位數不受極端值影響，但是受資料個數影響。

1.2.3 眾數

定義: 眾數(mode)指數據中次數最高峰的資料，對屬質資料來說是一種很好的衡量方式。

皮爾森經驗法: 平均數至眾數的距離約等於平均數至中位數距離的三倍。



1.2.4 幾何平均數

幾何平均數多用於計算平均比率和平均速度。

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，且為正數，則幾何平均數(geometric mean, G)為

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i}\ 。$$

幾何平均數的性質如下:

(1) 數值資料中不能有任一個數字為 0，否則為 0。

(2) 當數值資料中有負數時，幾何平均數不能作為衡量標準。

(3) 計算某期間內之平均增加率，可利用幾何平均數，即

$$G = \sqrt[n]{\frac{\text{期間結束時的數值}}{\text{期間開始時的數值}}} - 1$$

(4) 有 n 年每年的報酬率分別為$R_1, R_2, \ldots, R_n$，則 n 年平均報酬率為

$$G = \sqrt[n]{(1+R_1)(1+R_2)\cdots(1+R_n)}。$$

例題: Using the following information to answer Questions (1) and (2)

A statistician collected three data points from an experiment which are normally distributed as follows:

25, 8, 40

(1) What is the arithmetic mean of these data points?

(2) What is the geometric mean of these data points?

Sol:

(1) $\bar{x} = \frac{25+8+40}{3} \cong 24.33$。

(2) $G = \sqrt[3]{25 \times 8 \times 40} = 20$。

1.3 離差量數

離差量數為衡量數值資料分散情形的統計量。

1.3.1 全距

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，經由排序後得$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$，全距(range, R)為

$$R = x_{(n)} - x_{(1)}。$$

全距越大代表分散程度越大，但其缺點如下:

(1) 資料不同單位時無法比較。

(2) 當資料有極端值時，會無法精確反映資料的分散。

1.3.2 四分位距

由於全距容易受到極端值影響，可利用四分位距加以修正。

定義: 四分位差 (interquartile range, IQR)是指將各個變數值按大小順序排列，然後將此數列分成四等份，所得第三個四分位上的值與第一個四分位上的值的差，即

$$IQR = Q_3 - Q_1。$$

1.3.3 平均絕對差

平均偏差為每一個數值與算術平均數距離的總和。

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，則平均絕對差(Mean Absolute Deviation, MAD)為

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} 。$$

1.3.4 變異數

由於平均絕對差牽涉絕對值，不適合代數運算，許多應用也會受到限制。所以以變異數取代。

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，則此筆資料的變異數(variance)為

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ 。}$$

而變異數的平方根稱為此數值資料的標準差(standard deviation)。

由於變異數的單位為平方，對於解釋上較為複雜，因此取其平方根為標準差，使單位與原資料相同，以方便解釋與應用 。

變異數之性質如下:

(1) 變異數必定大於等於 0，若等於 0 表示此筆資料的變量皆相等。

(2) 各數值資料加(或減)一常數，變異數不變。

(3) 各數值資料乘上一常數 a，變異數變為原本a^2倍。

1.3.5 變異係數

由於上述衡量數值資料離散程度的統計量皆具有單位，故無法比較多組資料間的離散程度大小。此時需用無單位的統計量來做比較，在此介紹變異係數。

定義: 假設有 n 筆資料為$x_1, x_2, \ldots, x_n$，則此筆資料的變異係數(coefficient of variation, *CV*)為

$$CV = \frac{S}{\bar{x}} \times 100\% \text{。}$$

例題: The following is a set of data from a sample of $n = 11$ items:

x: 7, 9, 15, 3, 8, 20, 17, 1, 11, 80, 7

(1) Calculate the coefficient of variation (CV) for the variable *x*.

(2) What is the meaning of CV?

(1) $\bar{x} = 16.18$ ，$S = 21.93$。

$CV = \frac{S}{\bar{x}} \times 100\% = \frac{21.93}{16.18} \times 100\% = 135.54\%$。

1.4 偏態與峰態係數

偏態主要是衡量一組數值資料的偏斜情況。峰態則是衡量一組數值資料分配高峰凜旁的次數是高駿或平坦的現象。

定義: 偏態(skewness, β_1)係數為

$$\beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

偏態係數的性質如下:

(1) $\beta_1 = 0$時，表示分配為對稱分配。

(2) $\beta_1 > 0$時，表示分配為右偏(正偏)分配。

(1) $\beta_1 < 0$時，表示分配為左偏(負偏)分配。

定義: 峰態(kurtosis, β_2)係數為

$$\beta_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}}$$

峰態係數的性質如下:

(1) $\beta_2 > 3$時，表示分配為高狹峰(Leptokurtic)分配。

(2) $\beta_2 = 3$時，表示分配為常態峰(Mesokurtic)分配。

(3) $\beta_2 < 3$時，表示分配為低闊峰(Platykurtic)分配。

1.5 統計圖表

1.5.1 莖葉圖

莖葉圖(stem-and-leaf display)為保存原始資料與顯示次數分佈的一種常用圖表。

例題: 一筆成績的資料如下

75 82 60 63 35 80 61 78 71 62 72 68 51 65 53 46

做成莖葉圖如下:

3 | 5

4 | 6

5 | 13

6 | 012358

7 | 1258

8 | 02

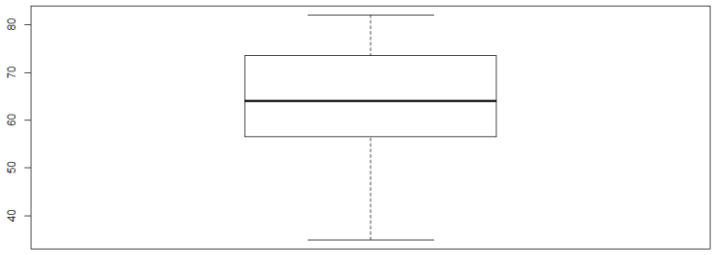
莖葉圖與直方圖類似，但莖葉圖提供兩個優點:

(1) 莖葉圖編制容易。

(2) 在途中仍可保持原始資料。

1.5.2 盒鬚圖

盒鬚圖(box-and-whisker plot)的建立需要計算出最小值、第一個四分位數(Q_1)、第二個四分位數(Q_2)、第三個四分位數(Q_3)及最大值，再將這五個值標示出來。同樣透過上面例題的資料繪製盒鬚圖如下



在編製盒鬚圖時，首先須將離群值(outlier)剔除，所謂離群值是指箱子前後 1.5 倍四分位距外的數值，即在區間$(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$內的數值允予保留。

第二章 機率理論

統計推論為探討利用母體抽出的樣本資訊去推論母體的未知特性，但如果要解釋此推論的可靠度，或者了解推論的風險為何，就必須了解機率的原理。

2.1 集合

集合可是唯一些有意義或具有特定性質事物的集體，為一組或一群的總稱。即許多不同元素(element)所賦予意義而構成的集合。

2.1.1 集合的定義

定義: 與某特定問題有關聯的所有元素所構成的集合，稱之為全集合(universal set)。

定義: 集合內無任一元素者，稱為空集合(empty set; null set)，符號定義為∅。

定義: 集合 A 中的每一元素也是集合 B 的元素，則稱 A 為 B 的部分集合(subset)或子集合，符號定義以$A \subset B$表示。

定義: 兩個集合 A 與 B 其所有元素構成的集合，稱為聯集(union)，符號定義為$A \cup B$。

聯集的意思為兩集合至少有一個發生，即為

$$A \cup B = \{x | x \in A \ or \ x \in B\}。$$

定義: 兩個集合 A 與 B 所共有的元素構成的集合，稱為交集(intersection)，符號定義為$A \cap B$。

聯集的意思為兩集合同時發生，即為

$$A \cap B = \{x | x \in A \; and \; x \in B\}。$$

定義: A 為全集的一個子集合，則將全集合中不屬於 A 元素構成的集合，稱為 A 的餘集合或補集合(complement)，符號定義為A^c。

餘集合的意義代表某個集合不發生，即為

$$A^c = \{x | x \in S \; and \; x \notin A\}。$$

S 代表全集合。

定義: 差集合(different set)是指 A 集合扣除 B 集合所剩餘的部分。符號定義為 $A - B$。

差集合的意義代表某個集合發生但另一個集合不發生，即為

$$A - B = \{x | x \in A \; and \; x \notin B\}。$$

註: $A - B = A \cap B^c$。

定義: 若 A 與 B 兩個集合無共同的元素，則稱 A 與 B 集合分離(disjoint)，或稱為互斥集合(mutually exclusive)。符號定義為$A \cap B = \emptyset$。

2.1.2 集合的運算

交換律(communicative law):

$$A \cup B = B \cup A$$

$$B \cup A = A \cup B$$

結合律(associated law):

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

分配律(distributed law):

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

底摩定律(DeMorgan's law):

$$A^c \cup B^c = (A \cap B)^c$$

$$A^c \cap B^c = (A \cup B)^c$$

2.2 機率

2.2.1 機率基本定義

定義: 在不確定實驗中，求出一個確定結果的過程，稱之為隨機試驗(random experiment)。一個隨機試驗須滿足三個條件，即

(1) 試驗結果所有發生的結果可知。

(2) 在試驗前無法知道結果。

(3) 相同條件下，該試驗可重複進行。

定義: 在隨機試驗中，其所有可能出現的結果所成的集合，稱為該試驗的樣本空間(sample space)，以符號 S 或Ω表示。而樣本空間的每一個元素稱為樣本點。

定義: 出象(outcome)為一隨機試驗中任一可能的結果。

2.2.2 機率的衡量

定義: (古典機率) 設一試驗的樣本空間 S 由 n 個樣本點所組成，且每一個樣本點出線的機會皆相同，則事件 A 發生的機率為

$$P(A) = \frac{\#(A)}{\#(S)}。$$

其中$\#(A)$與$\#(S)$分別表示事件 A 即樣本空間內樣本點的個數。

例題: A poker hand consists of 5 cards. If the cards have distinct consecutive values and are not all the same suit, we say that the hand is a "straight". Assume all possible poker hands are equally likely. What is the probability that one is dealt a straight?

***Sol*:** "straight" 總共有 10 種: A2345, 23456, … , 10JQKA。

五張牌可能為不同花色，但要把同花順的結果扣掉。。令 A 為抽到一個 straight 的事件，所以得知非同花順的機率為

$$P(A)=\frac{\#(A)}{\#(S)}=\frac{10\times 4^5-40}{\binom{52}{5}}=0.00392。$$

定義: (機率公理假設, Axiom of Probability) 設 S 表示某試驗的樣本空間，且 A 為 S 中的一事件，令$P(\cdot)$滿足三個公理(aaiom)，則$P(\cdot)$稱為一機率測度 (probability measure):

(1) 事件 A 發生的機率為$0\le P(A)\le 1$。

(2) S 為樣本空間，則其發生的機率為$P(S)=1$。

(3) 設$A_1,A_2,\cdots$，皆互為互斥的事件，則

$$P(A_1\cup A_2\cup\cdots)=\sum_{i=1}^{\infty}P(A_i)。$$

常見機率的計算公式:

(1) $P(A^c)=1-P(A)$。

(2) $A\subseteq B$，則$P(A)\le P(B)$。

定理: (機率加法定理, Addition Theorem) 設 A、B 為樣本空間的任意兩事件，則

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)。$$

2.3 條件機率與獨立事件

2.3.1 條件機率

有時候我們想知道在事件 A 發生的情況下，事件 B 發生的機率，因此我們需要知道條件機率的概念。

定義: 設 A、B 為樣本空間 S 中任意兩事件，則

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \text{ ，} P(A) > 0。$$

為已知事件 A 發生的條件下，事件 B 發生的條件機率(conditional probability)。

例題: $\mathrm{P(B)} = 0.6$. Suppose that B and C are mutually exclusive and complementary events. Consider another event A such that $P(A|B) = 0.2$。Find $P(A \cup C|B)$.

Sol: $P(A \cup C|B) = \frac{P((A \cup C) \cap B)}{P(B)} = \frac{P(A \cap B)}{P(B)} = \frac{0.12}{0.6} = 0.2$。

註: $P(A \cap B) = P(A|B) \times P(B) = 0.12$。

2.3.2 獨立事件

若兩事件中任一事件發生的機率與另一事件是否發生不相關聯時，則稱兩事件為獨立事件。

定義: 令 A、B 為任意兩事件，若滿足任一條件，即

(1) $P(B|A) = P(B)$，其中$P(A) > 0$。

(2) $P(A|B) = P(A)$，其中$P(B) > 0$。

(3) $P(A \cap B) = P(A) \times P(B)$。

則 A 與 B 稱為獨立事件(independent event)。

定理: 令 A、B 為獨立事件，則其餘事件之間亦獨立，即

(1) A 與B^c為獨立事件。

(2) A^c與B為獨立事件。

(3) A^c與B^c為獨立事件。

例題: Let A and B be independent events. $P(A \cap B) = \frac{1}{10}$, and $P(A \cap B^c) = \frac{1}{5}$. Evaluate $P((A \cup B)^c)$.

Sol: $P(A) = P(A \cap B) + P(A \cap B^c) = \frac{3}{10}$。

$P(B) = \frac{P(A \cap B)}{P(A)} = \frac{1}{3}$ (A and B be independent events)

$P((A \cup B)^c) = P(A^c \cap B^c) = P(A^c)P(B^c) = \left(1 - \frac{3}{10}\right)\left(1 - \frac{1}{3}\right) = \frac{7}{15}$。

定義: 令 A、B、C 為三個任意事件，若同時滿足下列條件，即

(1) $P(A \cap B) = P(A) \times P(B)$

(2) $P(B \cap C) = P(B) \times P(C)$

(3) $P(A \cap C) = P(A) \times P(C)$

(4) $P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$

則稱 A、B、C 為相互獨立事件 (mutually independent)。

註: 若只有(1)、(2)、(3)成立，而(4)不成立時，則稱 A、B、C 為配對獨立 (pairwise independent)。

例題: 擲公正硬幣兩次。令D_1表示第一次為正面的事件，而D_2表示第一次為正面的事件，表示D_3表示兩次結果相同的事件。

(1) D_1，D_2與D_3中任兩個配對是否獨立?

(2) D_1，D_2與D_3三者是否獨立?

Sol: (1) $P(D_1) = \frac{1}{2}$，$P(D_2) = \frac{1}{2}$，$P(D_3) = \frac{1}{2}$。又$P(D_1 \cap D_2) = P(D_1 \cap D_3) = P(D_2 \cap D_3) = \frac{1}{4}$。且$P(D_1 \cap D_2) = P(D_1)P(D_2)$、$P(D_1 \cap D_3) = P(D_1)P(D_3)$和$P(D_2 \cap D_3) = P(D_2)P(D_3)$。亦即$D_1$，$D_2$與$D_3$中兩兩配對獨立。

(2) $P(D_1 \cap D_2 \cap D_3) = \frac{1}{4}$，且知$P(D_1 \cap D_2 \cap D_3) \neq P(D_1)P(D_2)P(D_3) = \frac{1}{8}$。亦即$D_1$，$D_2$與$D_3$不為獨立事件。

2.4 樣本空間分割與貝氏定理

2.4.1 空間分割

定義: 令$A_1, A_2, \ldots, A_n$為樣本空間 S 的部分集合，且滿足下列兩個條件:

(1) $A_1 \cup A_2 \cup \ldots \cup A_n = S$

(2) $A_i \cap A_j = \emptyset$，$\forall i \neq j$

則稱$\{A_1, A_2, \ldots, A_n\}$為樣本空間 S 的一個分割 (partition)。

分割空間的範例:

A_1	A_2	……	A_j	……	A_n

2.4.2 貝氏法則

定理: (機率總和定理, theorem of total probability) 令$\{A_1, A_2, \ldots, A_n\}$為樣本空間 S 的一個分割，且$P(A_i) > 0$，$i = 1,2,\ldots,n$，若$B \subset S$，則

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i) \text{ 。}$$

例題: (Monte Hall Problem) 一遊戲提供三個門讓玩家選擇，獎品坐落在其中一扇門後面。當玩家選擇一扇門後，主持人會開另外兩扇門之中沒有獎品的一門，如此情況下換門得獎的機率為何?

Sol: 令有 A、B、C 三個門。W 為換門得獎的機率。假設今天獎品在 A 門後，則換門得獎的機率為

$$P(W) = P(A)P(W|A) + P(B)P(W|B) + P(C)P(W|C) = \frac{1}{3} \times 0 + \frac{1}{3} \times 1 + \frac{1}{3} \times 1$$

$$= \frac{2}{3} \text{。}$$

定理: (貝氏定理, Bayes' Theorem) 令$\{A_1, A_2, \ldots, A_n\}$為樣本空間 S 的一個分割。若$B \subset S$，$P(B) > 0$，$P(A_i) > 0$，$i = 1,2,\ldots,n$，則

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \text{，} j = 1,2,\ldots,n \text{。}$$

其中$P(A_j)$稱為事前機率 (prior probability)，而$P(A_j|B)$稱為事後機率 (posterior probability)。

事前機率表示某事件自然發生的機率，而事後機率表示獲得額外資訊後修正發生的機率。

例題: A plane is missing, and it is presumed that it was equally likely to have gone down in any of 3 possible regions. Let $1 - \beta_i,\ i = 1, 2, 3$, denote the probability that the plane will be found upon a search of the *i*-th region when the plane is, in fact, in that region. What is the conditional probability that the plane is in the *i*-th region given that a search of region 1 is unsecessful?

Sol: 令$R_i,\ i = 1, 2, 3$，為飛機在區域 i 的事件，且 E 為搜尋第 1 區失敗的事件。

$$\mathrm{P}(R_1|E) = \frac{P(R_1 \cap E)}{P(E)} = \frac{P(E|R_1)P(R_1)}{\sum_{i=1}^{3}(E|R_i)P(R_i)} = \frac{\beta_1 \times \frac{1}{3}}{\beta_1 \times \frac{1}{3} + 1 \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{\beta_1}{\beta_1 + 2} \text{。}$$

$$P(R_j|E) = \frac{P(R_j \cap E)}{P(E)} = \frac{P(E|R_j)P(R_j)}{\sum_{i=1}^{3}(E|R_i)P(R_i)} = \frac{1 \times \frac{1}{3}}{\beta_1 \times \frac{1}{3} + 1 \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{1}{\beta_1 + 2},$$

$j = 2, 3$。

第三章 隨機變數

對於非數值的事件，在分析上會遇到相當大的不便，因此我們可將該試驗中關心的現象數值化。而此種數值化的過程和表現，即為隨機變數的概念。

3.1 隨機變數定義

定義: 設 S 為試驗的樣本空間，則稱以樣本空間 S 為定義域，將其對應到實數的函數為隨機變數。通常以大寫的字母 X 或 Y 表示，即$X: S \rightarrow R$，而隨機變數中每一個變量，皆表示樣本空間中之一種事件。

定義: 隨機變數的所有可能值所成的集合為值域 (range)。若其值域為可數 (countable)或可數無限 (countable infinite)個的集合，則稱該隨機變數為離散隨機變數 (discrete random variable)。若其值域為不可數 (uncountable)集合，則稱該隨機變數為連續隨機變數 (continuous random variable)。

3.2 機率分配

常見的機率分配函數如下:

(1) 機率質量函數

(2) 機率密度函數

(3) 累積分配函數

這三個機率分配函數的介紹與定義如下:

定義: 令 X 為離散隨機變數，若函數$f(x)$滿足:

(1) $0 \le f(x) \le 1$。

(2) $\sum_{x=-\infty}^{\infty} f(x) = 1$。

則稱函數$f(x)$為隨機變數 X 的機率質量函數 (probability mass function; pmf)。

定義: 令 X 為連續隨機變數，若函數$f(x)$滿足:

(1) $0 \le f(x)$。

(2) $\int_{-\infty}^{\infty} f(x)dx = 1$。

則稱函數$f(x)$為隨機變數 X 的機率密度函數 (probability density function; pdf)。

定義: 令 X 為隨機變數，且 x 為任意實數，則

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

稱為隨機變數 X 的累積分配函數 (cumulative distribution function; cdf)。

若 X 為離散隨機變數，則

$$F_X(x) = P(X \le x) = \sum_{t=-\infty}^{x} f(t) \text{ 。}$$

若 X 為連續隨機變數，則

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt \text{ 。}$$

3.3 期望值

定義: 設隨機變數 X 其機率分配為$f(x)$，若

$$\sum_x |x|f(x) < \infty，\int_{-\infty}^{\infty} |x|f(x)dx < \infty\ (absolutely\ convergent)，$$

則隨機變數的期望值 (expected value)為

$$E(X) = \begin{cases} \sum_x xf(x) ，X為離散隨機變數。 \\ \int_{-\infty}^{\infty} xf(x)dx ，X為連續隨機變數。 \end{cases}$$

若 a 與 b 為實數，且$E(X)$存在，則

$$E(aX + b) = aE(X) + b。$$

前面提及的變異數也可透過期望值計算出來:

$$Var(X) = E(X^2) - [E(X)]^2。$$

第四章 常見離散隨機變數

在統計試驗中，獲取資料後為了方便分析進行方法的推導或統計的推論，必須尋找一個適合的機率模型配合該筆資料。此章將介紹幾個常見的離散隨機變數的機率模型。

定義: 離散隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}\frac{1}{n}，x=1,2,\ldots,n\\ 0，o.w.\end{cases}，k\geq 1、k\in N$$

稱隨機變數 X 服從離散均勻分配 (Discrete Uniform Distribution)，符號定義為 $X\sim DU(1,n)$。

隨機試驗: 試驗中的出象有 n 個，且每個出象發生的機率都相同。

期望值: $E(X)=\frac{n+1}{2}$

變異數: $Var(X)=\frac{n^2-1}{12}$

例題: 投擲一個公正的骰子，若 X 為隨機變數表示出現的點數，試問 X 的機率分配為何?

Sol: 因為是公正的骰子，故每一個點出現的機率皆為$\frac{1}{6}$。因此 X 為離散均勻分配，其機率分配為

$$f_X(x)=\begin{cases}\frac{1}{6}，x=1,2,\ldots,6\\ 0，o.w.\end{cases}。$$

定義: 離散隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}p^x(1-p)^{1-x}，x=0,1\\ 0，o.w.\end{cases}，0\le p\le 1$$

稱隨機變數 X 服從柏努力分配 (Bernoulli Distribution)，符號定義為

$X\sim Ber(p)$。

隨機試驗: 試驗中之出象只有互斥的兩種，研究者有興趣的視為「成功」以$x=1$表示，另一種視為「失敗」以$x=0$表示，而成功機率以 p 表示。

期望值: $E(X)=p$

變異數: $Var(X)=p(1-p)$

定義: 離散隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x}，x=0,1,2,\dots,n\\ 0，o.w.\end{cases}，0\le p\le 1，n為正整數。$$

稱隨機變數 X 服從二項分配 (Binomial Distribution)，符號定義為

$X\sim Bin(n,p)$。

隨機試驗: 進行 n 次相互獨立且成功機率 p 相同的柏努力實驗後，令二項分配的隨機變數表示「成功」的總次數。

期望值: $E(X)=np$

變異數: $Var(X)=np(1-p)$

例題: Find the probability of obtaining at least one 6 in four rolls of a fair die.

Sol: 令 X 為一隨機變數代表擲出 6 的次數。所以$X\sim Bin(4,\frac{1}{6})$。因此得出至少擲出一次 6 的機率為

$$\mathrm{P}(\mathrm{X}>0)=1-P(X=0)=1-\binom{4}{0}\left(\frac{1}{6}\right)^{0}\left(\frac{5}{6}\right)^{4}=1-\left(\frac{5}{6}\right)^{4}=0.518。$$

定義: 離散隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}\dfrac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}，x=0,1,2,\ldots,n\\ 0，o.w.\end{cases}$$

稱隨機變數 X 服從超幾何分配 (Hypergeometric Distribution)，符號定義為 $X\sim Hyper(N,m,n)$。

隨機試驗: 一個包含 N 個元素的母體中可以分為兩大類，而研究者有興趣的分類在母體中 m 個視為「成功」，另一類$N-m$則視為「失敗」。自母體中抽出 n 個個體，超幾何隨機變數表示在 n 個個體中「成功」的總次數。試驗屬於取後不放回，即試驗間不獨立。

期望值: $E(X)=n\frac{m}{N}$

變異數: $Var(X)=n\frac{m}{N}\left(1-\frac{m}{N}\right)\left(\frac{N-n}{N-1}\right)$

例題： A purchaser of electrical components buys them in lots of size 10. It is his policy to inspect 3 components randomly from a lot and to accept the lot only if all 3 are non-defective. If 30 percent of the lots have 4 defective components and 70 percent have only 1, what proportion of lots does the purchaser reject?

Sol: 令 A 為購買者接受貨物的事件，則

$$\mathrm{P}(A) = \mathrm{P}(\mathrm{A}|lots\ has\ 4\ defectives) \times \frac{3}{10} + \mathrm{P}(\mathrm{A}|lots\ has\ 1\ defectives) \times \frac{7}{10}$$

$$= \frac{\binom{4}{0}\binom{6}{3}}{\binom{10}{3}} \times \frac{3}{10} + \frac{\binom{1}{0}\binom{9}{3}}{\binom{10}{3}} \times \frac{7}{10} = \frac{54}{100}$$

定義： 離散隨機變數 X 的機率函數為

$$f_X(x) = \begin{cases} (1-p)^{x-1}p\text{，}x = 1, 2, 3\ldots \\ 0\text{，}o.w. \end{cases}\text{，}0 \le p \le 1\text{。}$$

稱隨機變數 X 服從幾何分配 (Geometric Distribution)，符號定義為

$X \sim Geo(p)$。

隨機試驗：連續進行相互獨立且成功機率相同為 p 的柏努力試驗，幾何分配隨機變數表示出現第一次成功所需進行的總試驗次數。

期望值: $E(X) = \frac{1}{p}$

變異數: $Var(X) = \frac{1-p}{p^2}$

定義： 離散隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}\binom{x-1}{r-1}p^r(1-p)^{x-r}，x=r,r+1,r+2\dots \\ 0，o.w.\end{cases}，r\geq 1，0\leq p\leq 1。$$

稱隨機變數 X 服從負二項分配 (Negative Binomial Distribution)，符號定義為 $X\sim NB(r,p)$。

隨機試驗：連續進行相互獨立且成功機率相同為 p 的柏努力試驗，幾何分配隨機變數表示出現第 r 次成功所需進行的總試驗次數。

期望值: $E(X)=\frac{r}{p}$

變異數: $Var(X)=r\frac{1-p}{p^2}$

例題： Suppose that during practice, a basketball player can make a free throw 80% of the time. Furthermore, assume that a sequence of free throw-shooting can be thought of as independent Bernoulli trials. Let X equal the minimum number of free throws that this player must attempt to make a total of 10 shots.

Find the values of $P(X\leq 12)$.

Sol: $P(X\leq 12)=P(X=10)+P(X=11)+P(X=12)$

$$=\binom{10-1}{10-1}(0.8)^{10}(0.2)^{10-10}+\binom{11-1}{10-1}(0.8)^{10}(0.2)^{11-10}$$
$$+\binom{12-1}{10-1}(0.8)^{10}(0.2)^{12-10}=0.5583$$

定義： 設$X(t)$表在 *t* 倍單位時間內「成功」的總次數，若滿足:

(1) 每一時間長度為 *t* 的互斥區間內「成功」總次數為相互獨立

(2) 極小的時間區間內，只可能發生一次成功或失敗。

稱為波瓦松過程 (Poisson Process)。

定義： 離散隨機變數$X(t)$的機率函數為

$$f_X(x)=\begin{cases}\dfrac{e^{-\lambda t}(\lambda t)^n}{n!}, n=0,1,2\cdots, \lambda\geq 0。\\ 0, o.w.\end{cases}$$

稱隨機變數$X(t)$服從波瓦松分配 (Poisson Distribution)，符號定義為

$X(t)\sim Poisson(\lambda t)$。

隨機試驗: 在一個滿足波瓦松過程的定義中，$X(t)$表在 *t* 倍單位時間內「成功」的總次數。在 *t* 倍單位時間內發生成功的平均次數為λt。

期望值: $E(X(t))=\lambda t$

變異數: $Var(X(t))=\lambda t$

例題： Ellen is taking a typing lesson and her assignment is to type a 10-page report in one hour. The number of typing errors she makes per page has a Poisson has a Poisson distribution with mean of $\lambda = 2$.

(1) What is the probability that Ellen does not make any typing error in one randomly selected page?

(2) What is the probability that Ellen making 4 typing errors total in two pages?

Sol: (1) 令隨機變數 X 表示 Ellen 每頁打錯字的個數，則$X \sim Poisson(\lambda)$，且 $\lambda = 2$ 字/頁，則

$$\mathrm{P}(\mathrm{X} = 0) = \frac{e^{-2}(2)^0}{0!} \approx 0.1353$$

(2) 令隨機變數$X(2)$表示 Ellen 兩頁打錯字的個數，

$X(2) \sim Poisson(\lambda t = 2 \times 2 = 4)$，則

$$\mathrm{P}(X(2) = 4) = \frac{e^{-4}(4)^4}{4!} \approx 0.1954$$

第五章 常見連續型隨機變數

在統計試驗中，獲取資料後為了方便分析進行方法的推導或統計的推論，必須尋找一個適合的機率模型配合該筆資料。此章將介紹幾個常見的連續隨機變數的機率模型。

定義: 連續隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b \\ 0, & o.w.\end{cases}, \quad a<b$$

稱隨機變數 X 服從連續型均勻分配 (Uniform Distribution)，符號定義為 $X\sim U(a,b)$。

隨機試驗: 在(a, b)區間中的機率均相同。

期望值: $E(X)=\frac{a+b}{2}$

變異數: $Var(X)=\frac{(b-a)^2}{12}$

定理: (機率積分轉換定理) 連續隨機變數 X 的累積機率函數$F_X(x)$，令隨機變數$Y=F_X(x)$，則$Y\sim U(0,1)$。

此定理使得連續型均勻分配為一個非常重要的分配，為數值模擬的一個重要基礎。

定義: 連續隨機變數 X 的機率函數為

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}，0 < x \\ 0，o.w. \end{cases}，0 < \lambda$$

稱隨機變數 X 服從指數分配 (Exponential Distribution)，符號定義為

$X \sim Exp(\lambda)$。

隨機試驗: 隨機變數 X 表示直到下一次成功發生的間隔時間，單位時間內發生成功的平均次數為λ。

期望值: $E(X) = \frac{1}{\lambda}$

變異數: $Var(X) = \frac{1}{\lambda^2}$

例題: Suppose that the length of a phone call in minutes is an exponential random variable with parameter $\lambda = \frac{1}{10}$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait

(a) more than 10 minutes;

(b) between 10 and 20 minutes.

Sol: 令 X 為電話亭中的人講話的時間，所以

(a) $\mathrm{P}(\mathrm{X} > 10) = 1 - P(X \leq 10) = 1 - \int_0^{10} \frac{1}{10} e^{-\frac{1}{10}x} dx = 1 - (1 - e^{-1}) = e^{-1} = 0.368.$

(b) $\mathrm{P}(10 \leq \mathrm{X} \leq 20) = P(X \leq 20) - P(X \leq 10) = \int_0^{20} \frac{1}{10} e^{-\frac{1}{10}x} dx -$

$\int_0^{10} \frac{1}{10} e^{-\frac{1}{10}x} dx = (1 - e^{-2}) - (1 - e^{-1}) = e^{-1} - e^{-2} = 0.233.$

範例: (機率積分轉換定理)

20.3.2 Example: exponential distribution

If $X \sim \exp(\lambda)$ then the pdf is $f_X(x) = \lambda e^{-\lambda x}$, for $x > 0$, and by integrating we find

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0; \\ 1 - e^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

Putting $y = F_X(x)$ we derive the inverse function as follows:

$$\begin{aligned} y &= 1 - e^{-\lambda x} \\ 1 - y &= e^{-\lambda x} \\ \log(1 - y) &= -\lambda x \\ x &= -\frac{1}{\lambda} \log(1 - y) = F_X^{-1}(y). \end{aligned}$$

So the inversion method generates $X \sim \exp(\lambda)$ by using $-\lambda^{-1} \log(1 - U)$ with $U \sim U(0, 1)$. It is easy to show that if $U \sim U(0, 1)$ then $1 - U \sim U(0, 1)$, so $-\lambda^{-1} \log(U) \sim \exp(\lambda)$.

定義: 連續隨機變數 X 的機率函數為

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & 0 < x \\ 0, o.w. \end{cases}, \quad 0 < \lambda$$

稱隨機變數 X 服從伽瑪分配 (Gamma Distribution)，符號定義為

$X \sim Gamma(\alpha, \lambda)$。

隨機試驗: 隨機變數 X 表示直到第α次成功發生的間隔時間，單位時間內發生成功的平均次數為λ。

期望值: $E(X) = \frac{\alpha}{\lambda}$

變異數: $Var(X) = \frac{\alpha}{\lambda^2}$

Note: $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$。當$\alpha$為整數時，$\Gamma(\alpha) = (\alpha - 1)!$。

例題: 已知某汽車電瓶壽命服從 Gamma 分配，平均壽命為 3 年，標準差為 3 年。請問:

(1) 此電瓶壽命不超過 1.5 年的機率為?

(2) 電瓶壽命可以維持 5 年以上的機率為?

(1) 令隨機變數 X 為汽車電瓶的壽命。$E(X)=\frac{\alpha}{\lambda}=3$，$Var(X)=\frac{\alpha}{\lambda^2}=9$。所以 $\lambda=\frac{1}{3}$，$\alpha=1$。即

$$f(x)=\frac{1}{3}e^{-\frac{x}{3}}，x>0$$

所以，

$$P(X<1.5)=\int_0^{1.5}\frac{1}{3}e^{-\frac{x}{3}}dx=0.3935$$

(2)

$$P(X\geq 5)=\int_5^{\infty}\frac{1}{3}e^{-\frac{x}{3}}dx=0.1889$$

定義: 連續隨機變數 X 的機率函數為

$$f_X(x)=\begin{cases}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}，-\infty<x<\infty\\ 0，o.w.\end{cases}，-\infty<\mu<\infty，\sigma^2>0$$

稱隨機變數 X 服從常態分配 (Normal Distribution)，符號定義為$X\sim N(\mu,\sigma^2)$。

此分配又稱為高斯分配 (Gauss Distribution)。

期望值: $E(X)=\mu$

變異數: $Var(X)=\sigma^2$

此分配的重要性為:

(1) 此分配可以解釋很多自然現象，例如身高、體重、成績等。

(2) 在許多統計推論方法中，假設母體為常態時可推導出優良的統計方法。

(3) 很多不合常態的資料，利用一些數學轉換後，即可成為常態資料。

定義: 連續隨機變數 Z 的機率函數為

$$f_Z(z)=\begin{cases}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, & -\infty<z<\infty\\ 0, & o.w.\end{cases}$$

稱隨機變數 Z 服從標準常態分配 (Standard Normal Distribution)，符號定義為 $Z\sim N(0,1)$。

與常態的關係為: $Z=\frac{X-\mu}{\sigma}$

$E(Z)=0, E(Z^2)=1, E(Z^3)=0, E(Z^4)=3$

期望值: $E(Z)=0$

變異數: $Var(Z)=1$

將常態轉成標準常態的目的: 方便查表。

$$Z_{0.005}=2.576;\ Z_{0.01}=2.326;\ Z_{0.025}=1.96;\ Z_{0.05}=1.645;\ Z_{0.1}=1.282$$

例題: An expert witness in a paternity suit testifies that the length (in days) of human gestation is approximately normally distributed with parameters $\mu = 270$ and $\sigma^2 = 100$. The defendant in the suit is able to prove that he was out of the country during a period that began 290 days before the birth of the child and ended 240 days before the birth. If the defendant was, in fact, the father of the child, what is the probability that the mother could have had the very long or very short gestation indicated by the testimony?

Sol: 令 X 為懷孕的時間。則該名小孩出生在上述的時間的機率為

$$P(X > 290 \text{ or } X < 240) = P(X > 290) + P(X < 240)$$

$$= P\left(\frac{X - 270}{10} > 2\right) + P\left(\frac{X - 270}{10} < -3\right) = 1 - \Phi(2) + 1 - \Phi(3) \approx 0.0241$$

定理: (二項分配以常態分配近似) 設隨機變數 X 服從二項分配，即 $X \sim Bin(n, p)$，則當 n 趨近於無限大時，

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

以標準常態分配為極限，即$Z \sim N(0,1)$。

通常$np \geq 5$且$n(1-p) \geq 5$即可近似，但因為二項分配為離散機率分配，因此當利用常態分配近似其機率時，就必須要考慮連續型校正 (correction for continuity)，亦即事件的機率計算時要修正。

$$P(a \le X \le b) = P\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}} \le Z \le \frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

$$P(X = c) = P\left(\frac{c-\frac{1}{2}-np}{\sqrt{np(1-p)}} \le Z \le \frac{c+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

例題: A recently survey found that 72% of all adults over 50 wear glasses for driving. In a random sample of 100 adults over 50,

(1) what is the mean and standard deviation of the number who wear glasses?

(2) find the probability that more than two of the 100 sampled wear glasses for driving.

Sol: (1) 令隨機變數 X 表示 100 人中戴眼鏡的人數。$X \sim Bin(100, 0.72)$，$\mu == 100 \times 0.72 = 72$，$\sigma = \sqrt{100 \times 0.72 \times 0.28} = 4.49$。

(2) 因$np \ge 5$且$n(1-p) \ge 5$，故以常態分配近似此機率，即

$$\mathrm{P(X > 2)} = P\left(Z > \frac{2 + 0.5 - 100 \times 0.72}{\sqrt{100 \times 0.72 \times 0.28}}\right) \approx P(Z > -15.48) = 1$$

定理: (普瓦松分配以常態分配近似) 設隨機變數 X 服從普瓦松分配，即 $X \sim Poisson(\lambda)$，則當λ趨近於無限大時，

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

以標準常態分配為極限，即$Z \sim N(0,1)$。

因為普瓦松分配為離散機率分配，因此當利用常態分配近似其機率時，就必須要考慮連續型校正，亦即事件的機率計算時要修正。

$$P(a \le X \le b) = P\left(\frac{a - \frac{1}{2} - \lambda}{\sqrt{\lambda}} \le Z \le \frac{b + \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right)$$

定義: 連續隨機變數 X 的機率函數為

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}，0 < x < \infty \\ 0，o.w. \end{cases}，-\infty < \mu < \infty，\sigma^2 > 0$$

稱隨機變數 X 服從對數常態分配 (Log-normal Distribution)，符號定義為 $X \sim logN(\mu, \sigma^2)$。

與常態的關係為: $Y \sim N(\mu, \sigma^2)$，令$X = e^Y$，則$X \sim logN(\mu, \sigma^2)$。

期望值: $E(X) = e^{\mu + \frac{\sigma^2}{2}}$

變異數: $Var(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$

第六章 抽樣和抽樣分配

利用樣本所提供的資訊去估計母體未知參數是統計分析中一個很重要的目的，所以利用之前討論的各種機率模型的性質去發展一些重要的抽樣分配，以便後面推倒統計推論方法。

6.1 解釋名詞

1. 普查 (census): 亦稱為全查，對母體中的所有樣本點全部訪查，無一遺漏。

2. 抽樣調查 (sampling survey): 自所想研究的母體抽取一部分的資料，稱為樣本，再對此樣本進行分析已得到母體未知的特性。

3. 抽樣誤差 (sampling error): 抽樣調查是自母體中抽出部分的資料給予調查，樣本的出現為隨機且不同大小的樣本與抽樣方法獲取的資料皆有差異。因此進行統計推論時，樣本資料所呈現的特性與母體的特性可能會有差異。

4. 非抽樣誤差 (non-sampling error): 為調查人員本身疏忽或者回答者故意引導錯誤所造成的偏差，也有可能是記錄資料或整理資料時發生錯誤，因而導致進行統計推論時所產生的誤差。

6.2 抽樣方法

1. 簡單隨機抽樣 (simple random sampling): 所謂簡單隨機抽樣是指母體中的每一個元素其被選到的機率都是相同的隨機抽樣方法。

2. 分層隨機抽樣 (stratified random sampling): 調查的母體，可依照某個衡量標準，區分成數個不重複的子母體，我們稱之為「層」，且層與層之間有很大的變異性，層內的變異性較小。從每一層中利用簡單隨機抽樣抽出所須比例的樣本數，將各層取得的樣本合在一起即為此抽樣的樣本。

3. 系統抽樣法 (systematic sampling): 將母體的元素排序後，按照某一定的間隔 (可能為時間或空間) 選取一個樣本，直至選滿為止。

6.3 抽樣分配

抽樣分配為發展統計推論的重要基礎，主要的理由為:

(1) 抽樣分配可以測量利用樣本資訊推論母體時，不確定性的大小。

(2) 其機率性質可以用來衡量推論結果的可靠性。

定義： 從母體 (其真實的分配為$f_X(x)$)抽出 n 個隨機變數分別為

$X_1, X_2, \ldots, X_n$，且滿足下列的條件

(1) $X_1, X_2, \ldots, X_n$ 皆為獨立的隨機變數

(2) $X_1, X_2, \ldots, X_n$的機率分配皆相同，同為$f_X(x)$

則稱$(X_1, X_2, \ldots, X_n)$為此母體$f_X(x)$抽出的一組隨機樣本 (random sample)。

以符號上經常以$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} f_X(x)$表示，而 *i.i.d* 表示 independent and identical distribution。

定義： 統計量 (statistic) 為樣本中 n 個隨機變數$(X_1, X_2, \ldots, X_n)$所組成的任意函數，本身不包含未知參數，且為一隨機變數。而統計量的機率分配則稱為抽樣分配 (sampling distribution)。

註: 之前提過的平均數、中位數、眾數、變異數等皆為統計量。

6.3.1 $\bar{X}$的抽樣分配與中央極限定理

定義: 若母體所有資料$(X_1, X_2, \ldots, X_N)$皆為已知，其母體平均數 (population mean) 及母體變異數 (population variance) 分別為

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

定義: 若資料$(X_1, X_2, \ldots, X_n)$表示抽自一母體的一組隨機樣本，則

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{，} S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

$\bar{X}$為樣本平均數 (sample mean)、S^2為樣本變異數 (sample variance)。

定理: 若資料$(X_1, X_2, \ldots, X_n)$表示抽自平均數為μ變異數為σ^2的母體的一組隨機樣本，則

$$E(\bar{X}) = \mu \text{，} Var(\bar{X}) = \frac{\sigma^2}{n}$$

註: 統計量的機率分配稱為抽樣分配，而抽樣分配的標準差稱為標準誤 (standard error)，用來衡量該統計量可能值的分散程度。

例題: 以下是五名學生的統計學考試分數: 30、50、60、60、80，今天從其中抽取兩名學生 (不放回) ，求平均成績的:

(1) 機率分配

(2) 期望值

Sol: (1) 因採不歸還抽樣，有$\binom{5}{2}=10$組可能樣本，即

(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$
(30,50)	40	(50,60)	55
(30,60)	45	(50,80)	65
(30,60)	45	(60,60)	60
(30,80)	55	(60,80)	70
(50,60)	55	(60,80)	70

則$\bar{X}$的機率分配為

$\bar{x}$	40	45	55	60	65	70
$f(\bar{x})$	0.1	0.2	0.3	0.1	0.1	0.2

(2) $E(\bar{X})=40\times0.1+45\times0.2+55\times0.3+60\times0.1+65\times0.1+70\times0.2=56$

定理: (中央極限定理, Central Limit Theorem) 設 X 為一隨機變數，其分配為 $f_X(x)$，不論其形式如何，若平均數μ以及變異數σ^2存在，今天由其中抽取樣本數為 n 的一組隨機樣本，且樣本平均數$\bar{X}$，當 n 充分大時，$\bar{X}$的抽樣分配漸進常態分配，即

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

例題: An anthropologist wishes to estimate the average height of men for a certain race of people. If the population standard deviation is assumed to be 2.5 inches and if she randomly samples 100 men.

(1) Find the probability that the difference between the sample mean and the true population mean will not exceed 0.5 inch.

(2) Suppose that the anthropologist wants the difference between the sample mean and the true population mean to be less than 0.3 inch, with probability 0.9. How many men should she sample to achieve this objective.

Sol: (1) 因$n = 100$，可由中央極限定理得出$\bar{X} \sim N\left(\mu, \frac{(2.5)^2}{100}\right)$。因此

$$\mathrm{P}(|\bar{X} - \mu| \leq 0.5) = P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq \frac{0.5}{2.5/\sqrt{100}}\right) = P(|Z| \leq 2) = 0.9544$$

(2) $\mathrm{P}(|\bar{X} - \mu| < 0.3) = 0.9$

$P\left(|Z| < \frac{0.3}{2.5/\sqrt{n}}\right) = 0.9$ 又 $P(|Z| < 1.645) = 0.9$，所以

$$\frac{0.3}{2.5/\sqrt{n}} = 1.645 \Rightarrow n = \left(\frac{1.645}{0.3}\right)^2 (2.5)^2 = 187.92$$

所以必須抽取 188 人。

6.3.2 兩樣本下平均數差$\bar{X}-\bar{Y}$的抽樣分配

定理: 令$(X_1, X_2, \ldots, X_{n_1})$和$(Y_1, Y_2, \ldots, Y_{n_2})$表示抽自平均數為$\mu_1$、$\mu_2$變異數為$\sigma_1^2$與$\sigma_2^2$的兩獨立母體的隨機樣本，當樣本數$n_1$、$n_2$夠大時，則平均數差$\bar{X}-\bar{Y}$的抽樣分配為近似常態分配，即

$$\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

例題: *A* 廠牌每隻燈管壽命服從於平均數 7.2 月，標準差為 3 月的 Uniform Distribution。*B* 廠牌每隻燈管壽命服從於平均數 7 月，標準差為 4 月的 Exponential Distribution。且 *A*、*B* 兩廠的燈管壽命獨立。今天某人若購買 *A* 廠牌 81 支、*B* 廠牌 100 支。求:

(1) 此人所購買 *B* 廠牌燈管的平均壽命至少為 7.2 月的機率為何?

(2) *A* 廠牌樣本比 *B* 廠牌樣本平均壽命長 0.1 月以上的機率為何?

(3) *A*、*B* 至少有一廠牌的平均壽命超過 7.4 月的機率為何?

Sol: (1) 因$n_B > 30$，可由中央極限定理得出$\bar{X}_B\sim N\left(7, \frac{16}{100}\right)$，所以

$$\mathrm{P}(\bar{X}_B \geq 7.2) = P\left(\frac{\bar{X}_B-\mu_B}{\sigma_B/\sqrt{n_B}} \geq \frac{7.2-7}{4/\sqrt{100}}\right) = P(Z \geq 0.5) = 0.3085$$

(2) 因$n_A > 30$、$n_B > 30$，可由中央極限定理得出$\bar{X}_A-\bar{X}_B\sim N\left(0.2, \frac{9}{81}+\frac{16}{100}\right)$，所以

$$P(\bar{X}_A - \bar{X}_B > 0.1) = P\left(\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} > \frac{0.1 - 0.2}{\sqrt{\frac{9}{81} + \frac{16}{100}}}\right) \approx P(Z > -0.19)$$

$$= 0.5753$$

(3) $\mathrm{P}(\bar{X}_A > 7.4 \text{ or } \bar{X}_B > 7.4) = \mathrm{P}(\bar{X}_A > 7.4\ ) + \mathrm{P}(\bar{X}_B > 7.4) - \mathrm{P}(\bar{X}_A > 7.4\ ) \times$

$\mathrm{P}(\bar{X}_B > 7.4) = P(Z > 0.6) + P(Z > 1) - P(Z > 0.6) \times P(Z > 1) = 0.3895$

6.3.3 樣本比例的抽樣分配

定義： 設$(X_1, X_2, \ldots, X_n)$為抽自$Ber(p)$的一組隨機樣本，則

$$\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n}$$

稱為樣本比例 (sample proportion) 。

註: $E(\hat{P}) = p$，$Var(\hat{P}) = \frac{p(1-p)}{n}$

定理： 設$(X_1, X_2, \ldots, X_n)$為抽自$Ber(p)$的一組隨機樣本，則

(1) 令$Y = \sum_{i=1}^{n} X_i$，則 Y 的抽樣分配為$Bin(n, p)$，即

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}，y = 0,1,2,\ldots,n$$

(2) 當 n 小的時候，$\hat{P}$的抽樣分配為

$$f(\hat{p}) = \binom{n}{n\hat{p}} p^{n\hat{p}} (1-p)^{n-n\hat{p}}，\hat{p} = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$$

(3) 當 n 大時$(n \geq 30)$，或$np \geq 5$且$n(1-p) \geq 5$時，$\hat{P}$的抽樣分配以常態分配為極限，即

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

例題: Assume that 15% of the items produced in an assembly line operation are defective, but that the firm's production manager is not aware of this situation. Assume further that 50 parts are tested by the quality assurance department in order to determine the quality of the assembly operation. Let $\hat{P}$ be the sample proportion defective found by the quality assurance test.

(1) Show the sampling distribution for $\hat{P}$.

(2) What is the probability that the sample proportion will be within ± 0.03 of the population proportion defective?

Sol: (1) 因$n = 50$，因此由中央極限定理可以得知

$$\hat{P} \sim N\left(0.15, \frac{0.15 \times 0.85}{50}\right)$$

(2) $P(|\hat{P} - p| \leq 0.03) = P\left(\left|\frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq \frac{0.03}{\sqrt{\frac{0.15 \times 0.85}{50}}}\right) \approx P(|Z| \leq 0.59) = 0.448$

6.3.4 樣本比例差的抽樣分配

定理: 令$(X_1, X_2, \dots, X_{n_1})$和$(Y_1, Y_2, \dots, Y_{n_2})$表示抽自$Ber(p_1)$，$Ber(p_2)$兩獨立母體的隨機樣本，當樣本數$n_1$、$n_2$夠大時，則平均數差$\hat{P}_1 - \hat{P}_2$的抽樣分配為近似常態分配，即

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

6.4 常態分配下的抽樣分配

定義: 連續隨機變數 X 的機率函數為

$$f_X(x) = \begin{cases} \frac{\left(\frac{1}{2}\right)^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)} x^{\frac{r}{2}-1} e^{-\frac{x}{2}}, & 0 < x \\ 0, & o.w. \end{cases}$$

稱隨機變數 X 服從自由度為 r 的卡方分配 (Chi-square Distribution)，符號定義為$X \sim \chi^2_{(r)}$。

$X \sim \chi^2_{(r)} = Gamma\left(\alpha = \frac{r}{2}, \lambda = \frac{1}{2}\right)$

期望值: $E(X) = r$

變異數: $Var(X) = 2r$

註: 當自由度很大的時候，卡方分配可利用常態分配近似。

定理: 令$Z \sim N(0,1)$，則Z^2為服從自由度為 1 的卡方分配。

定理: 設$(X_1, X_2, \ldots, X_n)$為抽自$N(\mu, \sigma^2)$的一組隨機樣本，令

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

則 Y 服從自由度為 n 的卡方分配。

定理: 設$(X_1, X_2, \ldots, X_n)$為抽自$N(\mu, \sigma^2)$的一組隨機樣本，令

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ ，} S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

則

(1) $\bar{X}$與S^2相互獨立

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

(3) $\frac{(n-1)}{\sigma^2} S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$

定義: 設$Z \sim N(0,1)$，$W \sim \chi^2_{(v)}$，且 Z 與 W 為獨立隨機變數。令

$$T = \frac{Z}{\sqrt{W/v}}$$

則稱隨機變數 T 的機率分配為自由度為v的 t 分配 (student's t distribution)，記做$T \sim t(v)$，且其機率密度函數為

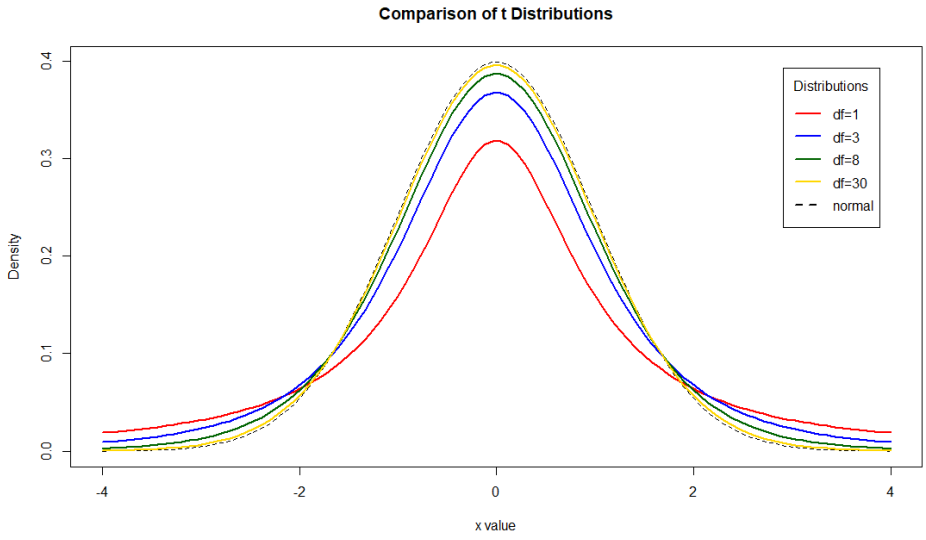
$$f_T(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{\sqrt{v\pi}} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}} \text{ ，} -\infty < t < \infty$$

期望值: $E(T) = 0$，$v > 1$

變異數: $Var(T) = \frac{v}{v-2}$，$v > 2$

t 分配的性質如下:

(1) t 分配與標準常態很相似，皆為對稱於 0 的鈴狀圖形，但因 t 分配的變異數較大，因此 t 分配的兩尾較厚。



(2) 當自由度$v \to \infty$時，t 分配的圖形會與標準常態的分配慢慢重疊，亦即 t 分配可以標準常態分配為極限。

(3) 與標準常態分配相同，t 值也是需要透過查表取得。

定理: 設$(X_1, X_2, \ldots, X_n)$為抽自常態母體$N(\mu, \sigma^2)$的隨機樣本，則

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

為自由度為$n-1$的 t 分配。

定義: 設 U 與 V 為獨立卡方隨機變數，即$U\sim\chi^2_{(v_1)}$，$V\sim\chi^2_{(v_2)}$，今令

$$F=\frac{U/v_1}{V/v_2}$$

則稱隨機變數 F 的機率分配為自由度為v_1和v_2的 F 分配 (F distribution)，記做$F\sim F(v_1,v_2)$，其中v_1為分子 (numerator)，v_2為分子 (denominator)。

F 分配的性質如下:

(1) $F\sim F(v_1,v_2)$，則$\frac{1}{F}\sim F(v_2,v_1)$

(2) 若$T\sim t(v)$，則$T^2\sim F(1,v)$

(3) $F_{1-\alpha}(v_1,v_2)=\frac{1}{F_\alpha(v_2,v_1)}$

定理: 設$(X_1,X_2,\ldots,X_{n_1})$為抽自常態母體$N(\mu_1,\sigma_1^2)$的一組隨機樣本，$(Y_1,Y_2,\ldots,Y_{n_2})$為抽自常態母體$N(\mu_2,\sigma_2^2)$的一組隨機樣本，若兩常態母體獨立，則

$$F=\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$$

為自由度n_1-1和n_2-1的 F 分配。

定理: (分布間的關係) Z、χ^2、T、F 的關係如下:

(1) $Z_{\frac{\alpha}{2}}=\sqrt{\chi^2_\alpha(1)}$

(2) $t_{\frac{\alpha}{2}}(v)=\sqrt{F_\alpha(1,v)}$

(3) $Z_{\frac{\alpha}{2}}=t_{\frac{\alpha}{2}}(\infty)=\sqrt{F_\alpha(1,\infty)}$

(4) $\chi^2_\alpha(v)=vF_\alpha(1,v)$

第七章 區間估計

區間估計就是以一定的機率保證估計包含總體參數的一個值域，即根據樣本指標和抽樣平均誤差推斷總體指標的可能範圍。區間估計表示結果的準確程度，又同時表明這個估計結果的可靠程度，因此其結果較有參考價值。

7.1 區間估計定義

定義: 設$(X_1, X_2, \ldots, X_n)$為由參數θ之母體$f_X(x;\theta)$抽出的一組隨機樣本，令$L(X_1, X_2, \ldots, X_n)$與$U(X_1, X_2, \ldots, X_n)$為兩個統計量，使得

$$P\big(L(X_1, X_2, \ldots, X_n) \le \theta \le U(X_1, X_2, \ldots, X_n)\big) = 1 - \alpha$$

則區間$\big(L(X_1, X_2, \ldots, X_n), U(X_1, X_2, \ldots, X_n)\big)$稱為參數$\theta$的 $100(1-\alpha)\%$的區間估計量 (interval estimator) ，其中$1-\alpha$稱為信賴度。當我們獲取確定的一組樣本資料$(x_1, x_2, \ldots, x_n)$代入上述區間估計量時，則區間

$$\big(L(x_1, x_2, \ldots, x_n), U(x_1, x_2, \ldots, x_n)\big)$$

稱為參數θ的 $100(1-\alpha)\%$的信賴區間 (confidence interval)。

定義: (1) 若$P(L(X_1, X_2, \ldots, X_n) \le \theta) = 1-\alpha$，則區間$(L(X_1, X_2, \ldots, X_n), \infty)$稱為參數$\theta$ 的 $100(1-\alpha)\%$之下的下尾 (one-sided lower)區間估計量。

(2) 若$P\big(\theta \le U(X_1, X_2, \ldots, X_n)\big) = 1-\alpha$，則區間$\big(-\infty, U(X_1, X_2, \ldots, X_n)\big)$稱為參數$\theta$ 的 $100(1-\alpha)\%$之下的上尾 (one-sided upper)區間估計量。

7.2 母體平均數μ的估計

母體平均數μ的估計是很常見的估計問題，由於樣本平均數$\bar{X}$為參數μ的最佳估計式，所以可以利用$\bar{X}$的抽樣分配來發展μ的估計問題。

7.2.1 估計誤差與樣本數問題

定理： 當$(X_1, X_2, \ldots, X_n)$為抽自任意母體，或者常態母體之一組隨機樣本，則樣本平均數$\bar{X}$估計 μ的誤差界限為

$$d = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

又當誤差界線 d 已知時，樣本數 n 為

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{d} \right)^2 \sigma^2$$

證明： 當$(X_1, X_2, \ldots, X_n)$為抽自

(1) 抽自非常態母體，且樣本數大於 30，或者

(2) 抽自常態母體

則樣本平均數$\bar{X}$服從常態分配

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

因此

$$P(|\bar{X} - \mu| \le d) = 1 - \alpha \Longrightarrow P\left(|Z| \le \frac{d}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

又由標準常態分配知

$$P\left(|Z| \le z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

故

$$z_{\frac{\alpha}{2}} = \frac{d}{\sigma/\sqrt{n}} \Longrightarrow d = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Longrightarrow n = \left(\frac{z_{\frac{\alpha}{2}}}{d}\right)^2 \sigma^2$$

例題: We know that the standard deviation of daily output on a production line for steel pipe is $\sigma = 10$ tons. We want to estimate the mean daily output of the production line, μ, to within ± 2.5 tons with 95% confidence. What sample size is required?

Sol: $P(|\bar{X}-\mu| \le 2.5) = 0.95 \Longrightarrow P\left(|Z| \le \frac{2.5}{10/\sqrt{n}}\right) = 0.95$

由 Z 表得知$P(|Z| \le 1.96) = 0.95$，故

$$\frac{2.5}{10/\sqrt{n}} = 1.96 \Longrightarrow n = \left(\frac{1.96}{2.5}\right)^2 10^2 \div 61.47$$

取樣本數$n = 62$。

7.2.2 區間估計

定理: 當$(X_1, X_2, \ldots, X_n)$為抽自非常態母體的一組隨機樣本，若$n > 30$ 且σ^2已知，則母體平均數μ的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

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

$$P\left(-z_{\frac{\alpha}{2}}\le Z\le z_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$\Rightarrow P\left(-z_{\frac{\alpha}{2}}\le\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\le z_{\frac{\alpha}{2}}\right)=1-\alpha$$

$$\Rightarrow P\left(-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\le\bar{X}-\mu\le z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

$$\Rightarrow P\left(\bar{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

註: 當母體變異數未知時，可以樣本變異數S^2估計。則此時的母體平均數μ的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X}-z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\bar{X}+z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$$

例題: A process manager takes a sample of size $n=100$ bottles from the bottling process, and the average content of the bottles is $\bar{x}=350$ c.c.. Assume the population standard deviation is $\sigma=40$ c.c., and significance level $\alpha=0.05$.

(1) Construct the confidence interval for the average content of the process.

(2) How many more bottles are needed if the manager requires the absolute error $E=5$ c.c.?

Sol: (1) μ的 95%信賴區間為

$$\left(\bar{X}-z_{0.025}\frac{\sigma}{\sqrt{n}},\bar{X}+z_{0.025}\frac{\sigma}{\sqrt{n}}\right)$$

$$\Rightarrow \left(350 - 1.96\frac{40}{\sqrt{100}}, 350 + 1.96\frac{40}{\sqrt{100}}\right)$$

$$\Rightarrow (342.16, 357.84)$$

(2) $n = \left(\frac{1.96}{5}\right)^2 40^2 = 245.86$，取$n = 246$。

定理： 當$(X_1, X_2, \ldots, X_n)$為抽自常態母體的一組隨機樣本，無論樣本數為何，且σ^2已知，則母體平均數μ的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

定理： 當$(X_1, X_2, \ldots, X_n)$為抽自常態母體的一組隨機樣本，無論樣本數為何，且σ^2未知，則母體平均數μ的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X} - t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right)$$

證明： 因為$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$，所以

$$P\left(-t_{\frac{\alpha}{2}}(n-1) \le T \le t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

$$\Rightarrow P\left(-t_{\frac{\alpha}{2}}(n-1) \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\frac{\alpha}{2}}(n-1)\right) = 1-\alpha$$

$$\Rightarrow P\left(-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}} \le \bar{X}-\mu \le t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{X} - t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

例題: A random sample of $n = 16$ with $\bar{x} = 50$ and $s = 12$ is taken from a population with $\sigma = 10$.

(1) Find a 95% confidence interval for the unknown population mean if the population is normally distributed.

(2) How would the answer in (1) have differed if the population standard deviation were unknown?

Sol: (1) μ的95%信賴區間為

$$\left(\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}\right)$$

$$\Rightarrow \left(50 - 1.96\frac{10}{\sqrt{16}}, 50 + 1.96\frac{10}{\sqrt{16}}\right) \Rightarrow (45.1, 54.9)$$

(2) μ的95%信賴區間為

$$\left(\bar{X} - t_{0.025}(15)\frac{S}{\sqrt{n}}, \bar{X} + t_{0.025}(15)\frac{S}{\sqrt{n}}\right)$$

$$\Rightarrow \left(50 - 2.131\frac{12}{\sqrt{16}}, 50 + 1.96\frac{12}{\sqrt{16}}\right) \Rightarrow (43.067, 56.393)$$

以下為影響信賴區間寬度的因素:

(1) 母體標準差: σ越大則信賴區間越寬，反之則越窄。

(2) 樣本數的大小: n 越大則信賴區間越窄，反之則越寬。

(3) 信賴係數的大小: $1-\alpha$ 越大時，信賴區間越寬，反之則越窄。

信賴區間的意義: 所謂 95%信賴區間估計，是指如果我們重複抽樣很多次，每次都會得到一個信賴區間，那麼這麼多的信賴區間中，約有 95%的區間會涵蓋真正的μ。但由於實際狀況下，我們只會有抽樣一次的資料且μ未知，因此我們無法得知此區間是否包含μ，但我們可以說此信賴區間有 95%信心會包含μ。

例題: (True or False) Suppose that we obtain a 95% confidence of the mean μ to be (65.5, 68.4). We know that $P(65.5 \leq \mu \leq 68.4) = 0.95$.

Sol: 錯! μ落在此區間內的機率不是 0 就是 1。

7.3 兩獨立母體平均數差$\mu_1 - \mu_2$的估計

自兩個獨立母體各抽取一組隨機樣本$(X_1, \ldots, X_{n_1})$以及$(Y_1, \ldots, Y_{n_2})$，比較兩母體的平均數之間有無差異是常見的統計推論問題 (詳細會於假設檢定中介紹)，在此以$\bar{X} - \bar{Y}$的抽樣分配來發展$\mu_1 - \mu_2$的推論問題。

7.3.1 $\mu_1 - \mu_2$估計誤差

定理: 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自兩個獨立的任意母體$(n_1 > 30，n_2 > 30)$，或抽自兩獨立的常態母體的兩組隨機樣本，則以$\bar{X} - \bar{Y}$估計$\mu_1 - \mu_2$的$100(1-\alpha)\%$誤差界限為

$$d = z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

證明: 假設樣本抽自兩個獨立的任意母體 $(n_1 > 30，n_2 > 30)$，或抽自兩獨立的常態母體，則可知

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

又可知

$$P(|\bar{X} - \bar{Y} - (\mu_1 - \mu_2)| \le d) = 1 - \alpha \Longrightarrow P\left(|Z| \le \frac{d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 1 - \alpha$$

又$P\left(|Z| \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$，故知

$$\frac{d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_{\frac{\alpha}{2}} \Longrightarrow d = z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

7.3.2 $\mu_1 - \mu_2$的信賴區間

定理: 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自兩個獨立的非常態母體$(n_1 > 30，n_2 > 30)$，若σ_1^2與σ_2^2已知，則$\mu_1 - \mu_2$的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

證明: 因$n_1 > 30$，$n_2 > 30$，則由中央極限定理知

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

則

$$P\left(-z_{\frac{\alpha}{2}} \le \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \bar{X} - \bar{Y} - (\mu_1 - \mu_2) \le z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{X} - \bar{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X} - \bar{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

定理: 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自兩個獨立的常態母體的隨機樣本，若σ_1^2與σ_2^2已知，則$\mu_1-\mu_2$的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X}-\bar{Y}-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\bar{X}-\bar{Y}+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

註: 若兩母體的變異數未知，因$n_1>30$，$n_2>30$，可利用樣本變異數S_1^2、S_2^2估計未知的母體變異數，所以$\mu_1-\mu_2$的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X}-\bar{Y}-z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\bar{X}-\bar{Y}+z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

例題: 今隨機抽取 *A* 型燈泡 40 個以及 *B* 型燈泡 50 個進行檢驗，*A* 型燈泡平均壽命為 418 小時，*B* 型燈泡平均壽命為 402 小時。$\sigma_A=26$、$\sigma_B=22$。試求$\mu_A-\mu_B$的 95%信賴區間。

Sol: $\mu_A-\mu_B$的 95%信賴區間為

$$\left(\bar{X}_A-\bar{X}_B-z_{0.025}\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}},\bar{X}_A-\bar{X}_B+z_{0.025}\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}\right)$$

$$\Rightarrow\left(418-402-1.96\sqrt{\frac{26^2}{40}+\frac{22^2}{50}},418-402+1.96\sqrt{\frac{26^2}{40}+\frac{22^2}{50}}\right)$$

$$\Rightarrow(5.895,26.105)$$

定理: 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自$N(\mu_1,\sigma_1^2)$以及$N(\mu_2,\sigma_2^2)$常態母體，若σ_1^2與σ_2^2未知，則

(1) 若$\sigma_1^2=\sigma_2^2=\sigma^2$，則兩母體平均數差$\mu_1-\mu_2$的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X}-\bar{Y}-t_{\frac{\alpha}{2}}(n_1+n_2-2)\times S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\bar{X}-\bar{Y}+t_{\frac{\alpha}{2}}(n_1+n_2-2)\times S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

其中$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$為共同變異數$\sigma^2$的混合或聯合估計量 (pooled estimator) ，且$E\left(S_p^2\right)=\sigma^2$。

(2) 若$\sigma_1^2=k\sigma_2^2$，則兩母體平均數差$\mu_1-\mu_2$的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X}-\bar{Y}-t_{\frac{\alpha}{2}}(n_1+n_2-2)\times S_p^*\sqrt{\frac{k}{n_1}+\frac{1}{n_2}},\bar{X}-\bar{Y}+t_{\frac{\alpha}{2}}(n_1+n_2-2)\times S_p^*\sqrt{\frac{k}{n_1}+\frac{1}{n_2}}\right)$$

其中$S_p^{*2}=\frac{k^{-1}(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$。

(3) 若$\sigma_1^2\neq\sigma_2^2$，則兩母體平均數差$\mu_1-\mu_2$的$100(1-\alpha)\%$區間估計量為

$$\left(\bar{X}-\bar{Y}-t_{\frac{\alpha}{2}}(v)\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}},\bar{X}-\bar{Y}+t_{\frac{\alpha}{2}}(v)\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

其中$v=\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$。一般取不大於它的整數。

例題: 淡水農場以栽種柚子為主，農場主人想了解 A、B 兩種肥料對於柚子收穫量的影響，便自 20 棵柚子樹中抽取 12 棵，施以 A 肥料，另外 8 棵則施以 B 肥料。結果施以 A 肥料的果樹的平均收穫量為 60kg/棵，標準差為 3kg; 施以 B 肥料的果樹的平均收穫量為 64kg/棵，標準差為 6kg。假設μ_1為使用 A 肥料的平均收穫量，μ_2為使用 B 肥料的平均收穫量，若這兩個母體的分配近似常態分配且變異數相等，試求$\mu_1 - \mu_2$的 99%信賴區間。

Sol: 因$S_p^2 = \frac{(12-1)3^2+(8-1)6^2}{12+8-2} = 19.5$

故$\mu_1 - \mu_2$的 99%信賴區間為

$$\left(\bar{X}_A - \bar{X}_B - t_{0.005}(18) \times S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \bar{X}_A - \bar{X}_B + t_{0.005}(18) \times S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

$$\Longrightarrow \left(60 - 64 - 2.878 \times \sqrt{19.5}\sqrt{\frac{1}{12}+\frac{1}{8}}, 60 - 64 + 2.878 \times \sqrt{19.5}\sqrt{\frac{1}{12}+\frac{1}{8}}\right)$$

$$\Longrightarrow (-9.8008, 1.8008)$$

7.4 兩相關母體$\mu_1 - \mu_2$的估計

若自兩母體選取的隨機樣本並非獨立抽取，而為成對抽取時即為相關母體。所謂成對的隨機樣本是指兩樣本的觀察值為成對出現。對於成對抽取的隨機樣本可先將成對觀察的樣本資料相減產生一組新的樣本資料。

	1	**2**	**……**	***i***	**……**	***n***
試驗前(*X*)	X_1	X_2		X_i		X_n
試驗後(*Y*)	Y_1	Y_2		Y_i		Y_n
$\boldsymbol{D = X - Y}$	D_1	D_2		D_i		D_n

今可將$(D_1, D_2, \ldots, D_n)$視為抽自平均數μ_D及變異數為σ_D^2的母體的一組隨機樣本，其中

$$\mu_D = \mu_1 - \mu_2$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

定理: (1) 兩母體為相關母體，且$n > 30$，則μ_D的$100(1-\alpha)\%$區間估計量為

$$\left(\overline{D} - z_{\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}}, \overline{D} + z_{\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}}\right)$$

(2) 兩母體為相關母體，但 n 為小樣本，此時若假設服從常態母體，則μ_D的$100(1-\alpha)\%$區間估計量為

$$\left(\overline{D} - t_{\frac{\alpha}{2}}(n-1)\frac{S_D}{\sqrt{n}}, \overline{D} + t_{\frac{\alpha}{2}}(n-1)\frac{S_D}{\sqrt{n}}\right)$$

例題: 有一個汽車修理行的招牌生意視為各型汽車裝設省油裝置，為了要瞭解這種省油裝置的裝設是否確實有省油的功效，隨機抽出 8 輛有裝設省油裝置的汽車，並記錄它們裝設這種省油裝置前後行駛 100 公里的耗油量如下

	1	2	3	4	5	6	7	8
X_i(裝設前)	3.2	4.6	3.6	5.3	6.2	3.2	3.6	4.5
Y_i(裝設後)	2.9	4.7	3.2	5.0	5.7	3.3	3.4	4.3

設每輛汽車耗油量皆呈常態分配，求一輛汽車裝設這種省油裝置之前與之後行駛 100 公里平均耗油量差的 90%信賴區間。

Sol: 裝設前減裝設後的資料為

	1	2	3	4	5	6	7	8
X_i(裝設前)	3.2	4.6	3.6	5.3	6.2	3.2	3.6	4.5
Y_i(裝設後)	2.9	4.7	3.2	5.0	5.7	3.3	3.4	4.3
$D_i = X_i - Y_i$	0.3	-0.1	0.4	0.3	0.5	-0.1	0.2	0.2

又$\sum_{i=1}^{n} D_i = 1.7$，$\sum_{i=1}^{n} D_i^2 = 0.69$，故知

$$\overline{D} = 0.2125; S_D = 0.2167$$

故知μ_D的 90%信賴區間為

$$\left(\overline{D} - t_{0.05}(7)\frac{S_D}{\sqrt{n}}, \overline{D} + t_{0.05}(7)\frac{S_D}{\sqrt{n}}\right)$$

$$\Longrightarrow \left(0.2125 - 1.895\frac{0.2167}{\sqrt{8}}, 0.2125 + 1.895\frac{0.2167}{\sqrt{8}}\right)$$

$$\Longrightarrow (0.0673, 0.3577)$$

7.5 單一母體比例的估計

在此與推導母體平均數的估計觀念類似

定理: 當$(X_1, X_2, \ldots, X_n)$為抽自$Ber(p)$，且$n > 30$，則樣本比例$\hat{P}$估計 p的

$100(1-\alpha)\%$的誤差界限為

$$d = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

又當誤差界線 d 已知時，樣本數 n 為

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{d}\right)^2 \hat{p}(1-\hat{p})$$

7.5.1 p的信賴區間

定理: 當$(X_1, X_2, \ldots, X_n)$為抽自$Ber(p)$的一組隨機樣本，若$n > 30$，則母體比例p的$100(1-\alpha)\%$區間估計量為

$$\left(\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

例題: At 2011, Department of Business Administration, NCKU, surveyed 2000 web users and asked the about their willingness to pay fees for access to web sites. Of course, 500 were definitely not willing to pay such fees.

(1) Calculate the 95% confidence interval for the proportion definitely unwilling to pay fees.

(2) What is the probability that the population proportion falls into the interval you obtained in (1)?

(3) How large a sample size is necessary to estimate the proportion of interest to within 2% with 95% confidence?

Sol: (1) 因$\hat{p} = 0.25$，所以 p 的 95%信賴區間為

$$\left(\hat{P} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{P} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\Rightarrow \left(0.25 - 1.96\sqrt{\frac{0.25 \times 0.75}{2000}}, 0.25 + 1.96\sqrt{\frac{0.25 \times 0.75}{2000}}\right)$$

$$\Rightarrow (0.231, 0.269)$$

(2) 0 或 1

(3) $n = \left(\frac{z_{0.025}}{d}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.96}{0.02}\right)^2 \times 0.25 \times 0.75 = 1800.75$

故取樣本數$n = 1801$。

7.5.2 兩母體比例差$p_1 - p_2$的估計

定理: 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自兩個獨立的母體$Ber(p_1)$、$Ber(p_2)$的隨機樣本 $(n_1 > 30，n_2 > 30)$，則以$\hat{p}_1 - \hat{p}_2$估計$p_1 - p_2$的$100(1-\alpha)\%$誤差界限為

$$d = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

定理: 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自兩個獨立的母體$Ber(p_1)$、$Ber(p_2)$的隨機樣本 $(n_1 > 30，n_2 > 30)$，則$p_1 - p_2$的$100(1-\alpha)\%$信賴區間為

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

例題: Let $\hat{p}_1$ be the estimated proportion of successes based on sample 1 of size n_1 for population 1 and $\hat{p}_2$ be the estimated proportion of successes based on sample 2 of size n_2 for population 2. Given that $n_1 = n_2 = 100$, $\hat{p}_1 = 0.1$, and $\hat{p}_2 = 0.03$. Find a 95% confidence interval for $p_1 - p_2$.

Sol: $p_1 - p_2$的95%信賴區間為

$$\left(\hat{p}_1-\hat{p}_2-z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\hat{p}_1-\hat{p}_2\right.$$
$$\left.+z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$
$$\Rightarrow\left(0.1-0.03-1.96\sqrt{\frac{0.1\times0.9}{100}+\frac{0.03\times0.97}{100}},0.1-0.03\right.$$
$$\left.-1.96\sqrt{\frac{0.1\times0.9}{100}+\frac{0.03\times0.97}{100}}\right)$$
$$\Rightarrow(0.0024,0.1376)$$

7.6 母體變異數σ^2的估計

定理: 設$(X_1,X_2,\ldots,X_n)$為抽自$N(\mu,\sigma^2)$的一組隨機樣本，則在μ未知下，變異數σ^2的$100(1-\alpha)\%$的區間估計量為

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$$

證明: σ^2的估計問題不像之前利用常態分配可以解決。且因S^2為σ^2的估計式，但是S^2的抽樣分配並非熟悉的抽樣分配，故將其轉換為

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

因此

$$P\left(\chi^2_{1-\frac{\alpha}{2}}(n-1)\le\chi^2\le\chi^2_{\frac{\alpha}{2}}(n-1)\right)=1-\alpha$$
$$\Rightarrow P\left(\chi^2_{1-\frac{\alpha}{2}}(n-1)\le\frac{(n-1)S^2}{\sigma^2}\le\chi^2_{\frac{\alpha}{2}}(n-1)\right)=1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) = 1-\alpha$$

若要求單尾的σ^2的$100(1-\alpha)\%$的區間估計量，則

(1) $\left(\frac{(n-1)S^2}{\chi^2_{\alpha}(n-1)}, \infty\right)$為左尾信賴區間

(2) $\left(-\infty, \frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}\right)$為右尾信賴區間

例題: The capacities of 10 batteries were recorded as follows: 140, 136, 150, 144, 148, 152, 138, 141, 143, 151 (assuming that the population is normal distribution). Find the lower 95% one-sided confidence interval for σ^2.

Sol: $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)}{n}\right]^2 = 32.2333$

所以σ^2的 95%的信賴區間為

$$\left(\frac{(9)S^2}{\chi^2_{0.05}(9)}, \infty\right) \Rightarrow \left(\frac{9 \times 32.2333}{16.919}, \infty\right) \Rightarrow (17.15, \infty)$$

定理: 設$(X_1, X_2, \ldots, X_n)$為抽自$N(\mu, \sigma^2)$的一組隨機樣本，則在μ已知下，變異數σ^2的$100(1-\alpha)\%$的區間估計量為

$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}\right)$$

7.6 母體變異數比$\frac{\sigma_1^2}{\sigma_2^2}$的估計

定理： 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自$N(\mu_1,\sigma_1^2)$以及$N(\mu_2,\sigma_2^2)$的兩組獨立隨機樣本，今若μ_1、μ_2、σ_1^2及σ_2^2未知下，母體變異數比$\frac{\sigma_1^2}{\sigma_2^2}$之$100(1-\alpha)\%$區間估計量為

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)},\frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2}}(n_2-1,n_1-1)\right)$$

證明： 設$(X_1,\ldots,X_{n_1})$以及$(Y_1,\ldots,Y_{n_2})$為分別抽自$N(\mu_1,\sigma_1^2)$以及$N(\mu_2,\sigma_2^2)$的兩組獨立隨機樣本，則

$$\mathrm{F}=\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2}\sim F(n_1-1,n_2-1)$$

因此

$$P\left(F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\le F\le F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right)=1-\alpha$$

$$\Rightarrow P\left(F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\le \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2}\le F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\right)=1-\alpha$$

$$\Rightarrow P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}\le \frac{\sigma_1^2}{\sigma_2^2}\le \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}\right)=1-\alpha$$

$$\Rightarrow P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}\le \frac{\sigma_1^2}{\sigma_2^2}\le \frac{S_1^2}{S_2^2}F_{\frac{\alpha}{2}}(n_2-1,n_1-1)\right)=1-\alpha$$

例題: A taxi company is trying to decide whether to purchase brand A or B tires for its fleet or taxis. To estimate the difference in the two brands, an experiment is conducted using 12 of each brand. The results are:

$$\bar{x}_A = 36300;\ S_A = 5000$$

$$\bar{x}_B = 38100;\ S_B = 6100$$

(1) Compute a 95% confidence interval for $\mu_A - \mu_B$ assuming the populations to be approximately normally distributed with unknown but equal variances.

(2) Construct a 90% confidence interval for $\frac{\sigma_A^2}{\sigma_B^2}$.

(Hint: $F_{0.05}(11,11) = 2.82$)

Sol: (1) 因$S_p^2 = \frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A+n_B-2} = 31105000$，故知$\mu_A - \mu_B$的 95%信賴區間為

$$\left(\bar{x}_A - \bar{x}_B - t_{0.05}(22) \times S_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}, \bar{x}_A - \bar{x}_B + t_{\frac{\alpha}{2}}(22) \times S_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right)$$

$$\Rightarrow \left(36300 - 38100 - 2.074 \times \sqrt{31105000}\sqrt{\frac{1}{12}+\frac{1}{12}}, 36300 - 38100\right.$$

$$\left. + 2.074 \times \sqrt{31105000}\sqrt{\frac{1}{12}+\frac{1}{12}}\right)$$

(2) $\frac{\sigma_A^2}{\sigma_B^2}$的 90%信賴區間為

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{0.05}(11,11)}, \frac{S_1^2}{S_2^2}F_{0.05}(11,11)\right)$$

$$\Rightarrow \left(\frac{(5000)^2}{(6100)^2}\frac{1}{2.82}, \frac{(5000)^2}{(6100)^2}2.82\right)$$

$$\Rightarrow (0.238, 1.895)$$

第八章 假設檢定

假設檢定 (hypothesis testing) 為對母體中某些參數的可能值是對是錯的假設，透過抽樣的樣本資料再經由機率的原理做出拒絕或不拒絕的過程。

8.1 假設檢定定義

定義: 在統計假設檢定中，想要否定的假設稱之為虛無假設 (null hypothesis)，通常以H_0表示。若否定虛無假設而被認為對的假設，則被稱為對立假設 (alternative hypothesis)，通常以H_1表示。

定義: 拒絕域 (reject region) 或稱為臨界域 (critical region)是指其為樣本空間中某些數值的集合，而且會導致否定虛無假設的集合，常以C表示。而拒絕與不拒絕H_0的交接點稱為臨界點 (critical value)。

假設檢定是由抽出的樣本去判斷H_0以及H_1兩種假設的真偽，但因為抽樣可能產生的樣本偏差，進而做出錯誤的決策。而此錯誤分為型 I 誤差 (type I error)以及型 II 誤差 (type II error)。

	H_0為真	H_1為真
不拒絕H_0	正確結論 $(1-\alpha)$	型 II 誤差 (β)
拒絕H_0	型 I 誤差 (α)	正確結論 $(1-\beta)$

定義: 型 I 誤差發生的機率以α表示，稱之為α風險 (α-risk)，以下式表示:

$$\alpha = P\left(\text{拒絕}H_0 \middle| H_0\text{為真}\right) = P\left(\text{落入}C\text{中} \middle| H_0\text{為真}\right)$$

而型 II 誤差發生的機率以β表示，稱之為β風險 (β-risk)，以下式表示:

$$\beta = P\left(\text{不拒絕}H_0 \middle| H_1\text{為真}\right) = P\left(\text{不落入}C\text{中} \middle| H_1\text{為真}\right)$$

定義: 在所有型 I 誤差機率中，發生機率最大者稱為顯著水準 (level of significance)，又稱為檢定的大小 (size of test)，它是在檢定中可容忍的最大型 I 誤差機率，通常以α表示。

例題: 在母體為常態的假設下，有一個假設檢定如下:

$$H_0: \mu \leq 25$$

$$H_1: \mu > 25$$

樣本數$n = 81$，標準差$\sigma = 18$，決策規則為: 如果$\bar{X} \leq 27.56$，則接受H_0，反之則接受H_1。試問:

(1) 若$\mu = 24$，根據上述的決策規則，犯型 I 誤差的機率為何?

(2) 若$\mu = 25$，根據上述的決策規則，犯型 I 誤差的機率為何?

(3) 若$\mu = 29$，根據上述的決策規則，犯型 II 誤差的機率為何?

Sol: (1) $\alpha = \mathrm{P}(\bar{X} > 27.56|\mu = 24) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{27.56-24}{18/\sqrt{81}}\right) = P(Z > 1.78) =$

0.0375

(2) $\alpha = \mathrm{P}(\bar{X} > 27.56|\mu = 25) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{27.56-25}{18/\sqrt{81}}\right) = P(Z > 1.28) = 0.1003$

(3) $\beta = \mathrm{P}(\bar{X} \leq 27.56|\mu = 29) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{27.56-29}{18/\sqrt{81}}\right) = P(Z \leq -0.72) = 0.2358$

檢定的步驟如下:

(1) 建立虛無假設H_0。

(2) 建立對立假設H_1。

(3) 選擇適當的顯著水準α。

(4) 找出此檢定的統計量與拒絕域。

(5) 利用樣本資料計算出統計量的值。

(6) 做結論。

定義: p-value 是指虛無假設H_0為真下，樣本統計量比其觀察到的樣本統計量的值還要極端的機率，亦稱為觀測的顯著水準。它是在給定樣本資訊下，拒絕虛無假設H_0的最小機率水準。

註: $p - \text{value} < \alpha \iff$ 拒絕H_0

8.2 單一母體平均數μ的檢定問題

8.2.1 μ的檢定—以常態分配處理

利用常態分配處理的條件:

(1) 抽自非常態母體，且樣本數大於 30，且σ^2已知，若σ^2未知，可用S^2替代。

(2) 抽自常態母體且σ^2已知。

	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
檢定法 (標準化)	$C = \{Z \mid Z > Z_\alpha\}$	$C = \{Z \mid Z < Z_\alpha\}$	$C = \{Z \mid \|Z\| > Z_{\frac{\alpha}{2}}\}$
信賴區間法	$\mu_0 \in \left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$ 時不拒絕H_0	$\mu_0 \in \left(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$ 時不拒絕H_0	$\mu_0 \in \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$ 時不拒絕H_0
p-value 法	$p-\text{value}$ $= P(\bar{X} > \bar{x} \mid \mu = \mu_0)$ $= P(Z > z)$	$p-\text{value}$ $= P(\bar{X} < \bar{x} \mid \mu = \mu_0)$ $= P(Z < z)$	$p - value$ $= 2P(Z > \|z\|)$

其中: $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

例題: 設$X_1, \ldots, X_{25} \overset{iid}{\sim} N(\mu, \sigma^2 = 64)$，$H_0: \mu = 5$，$H_1: \mu = 7$。

(1) 當顯著水準為 0.05，則拒絕域為何?型二錯誤的機率為何?

(2) 若樣本數不變時，希望將型二錯誤的機率降低至 0.1 時，則型一錯誤的機率會增加為多少?

Sol: (1) $P(\bar{X} > k \mid \mu = 5) = 0.05 \Longrightarrow P\left(Z > \frac{k-5}{8/\sqrt{25}}\right) = 0.05$，故知

$$\frac{k-5}{8/\sqrt{25}} = 1.645 \Longrightarrow k = 5 + 1.645\frac{8}{\sqrt{25}} = 7.632$$

故顯著水準為 0.05 時，拒絕域為$\mathrm{C} = \{\bar{X}|\bar{X} > 7.632\}$，所以型二誤差的機率為

$$\beta = P(\bar{X} \le 7.632|\mu = 7) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le \frac{7.632-7}{8/\sqrt{25}}\right) = P(Z \le 0.395) = 0.65355$$

(2) 因$P(\bar{X} \le k^*|\mu = 7) = 0.1 \Longrightarrow P\left(Z \le \frac{k^*-7}{8/\sqrt{25}}\right) = 0.1$，故知

$$\frac{k^*-7}{8/\sqrt{25}} = -1.282 \Longrightarrow k^* = 4.9488$$

所以型一誤差的機率為

$$\alpha = P(\bar{X} > 4.9488|\mu = 5) = P\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le \frac{4.9488-5}{8/\sqrt{25}}\right) \approx P(Z \le -0.032)$$

$$\approx 0.5120$$

8.2.2 μ的檢定—以 T 分配處理

利用 T 分配處理的條件: 抽自常態母體且σ^2未知。

	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
檢定法 (標準化)	$C = \{t \mid t > t_\alpha(n-1)\}$	$C = \{t \mid t < -t_\alpha(n-1)\}$	$C = \{t \mid \mid t \mid > t_{\frac{\alpha}{2}}(n-1)\}$
信賴區間法	$\mu_0 \in \left(\bar{x} - t_\alpha(n-1)\frac{\sigma}{\sqrt{n}}, \infty\right)$ 時不拒絕H_0	$\mu_0 \in \left(-\infty, \bar{x} + t_\alpha(n-1)\frac{\sigma}{\sqrt{n}}\right)$ 時不拒絕H_0	$\mu_0 \in (\bar{x} - t_{\frac{\alpha}{2}}(n-1)\frac{\sigma}{\sqrt{n}},$ $\bar{x} + t_{\frac{\alpha}{2}}(n-1)\frac{\sigma}{\sqrt{n}})$ 時不拒絕H_0
p-value 法	$\text{p} - \text{value}$ $= P(T > t \mid T \sim t(n-1))$	$\text{p} - \text{value}$ $= P(T < t \mid T \sim t(n-1))$	$\text{p} - \text{value}$ $= 2P(T > \mid t \mid \mid T \sim t(n-1))$

其中: $t = \frac{\bar{x}-\mu_0}{S/\sqrt{n}}$

例題: A chemical process has produced on the average, 800 tons of chemical per day. The daily yields for the past week are 785, 805, 790, 793 and 802 tons. Do these data indicate that the average yield is less than 800 tons, and hence that something is wrong with the process? Test the 5% level of significance. What assumptions must be satisfied in order for the procedure you used to analyze these data to be valid?

Sol: 假設此資料服從常態分配，則

<1> $H_0: \mu \geq 800$

<2> $H_1: \mu < 800$

<3> $\alpha = 0.05$

<4> $C = \{t | t < -t_{0.05} = -2.132\}$

<5> $\bar{x} = 795$，$S \approx 8.3367$，故檢定統計量

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{795 - 800}{8.3367/\sqrt{5}} \approx -1.3411 \notin C$$

<6> 結論: 不拒絕H_0。沒有足夠的證據說明$\mu < 800$。

8.3 兩母體平均數差的檢定問題

8.3.1 兩獨立母體$\mu_1 - \mu_2$的檢定—以常態分配處理

利用常態分配處理的條件:

(1) 抽自非常態母體，但$n_1 > 30$，$n_2 > 30$且σ_1^2、σ_2^2已知時，可用常態處理。

若σ_1^2、σ_2^2未知，可用S_1^2、S_2^2替代。

(2) 抽自常態母體且σ_1^2、σ_2^2已知。

	$H_0: \mu_1 - \mu_2 \leq d_0$ $H_1: \mu_1 - \mu_2 > d_0$	$H_0: \mu_1 - \mu_2 \geq d_0$ $H_1: \mu_1 - \mu_2 < d_0$	$H_0: \mu_1 - \mu_2 = d_0$ $H_1: \mu_1 - \mu_2 \neq d_0$
檢定法 (標準化)	$C = \{Z \mid Z > Z_\alpha\}$	$C = \{Z \mid Z < Z_\alpha\}$	$C = \{Z \mid \|Z\| > Z_{\frac{\alpha}{2}}\}$
p-value 法	$p - value$ $= P(Z > z)$	$p - value$ $= P(Z < z)$	$p - value$ $= 2P(Z > \|z\|)$

其中: $Z = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

例題: A finance professor of NCU reported that the mean of book-to-market (BM) for firms listed in New York Stock Exchange (NYSE) is 0.82, while the mean of OTC stocks is 0.77. The sample consists of 300 NYSE stocks and 400 OTC stocks. Based on historical data, the population standard deviations for the BM can be assumed known at 0.3 for NYSE and 0.2 for OTC.

(1) Do the sample data indicate that NYSE stocks have a higher BM than OTC stocks? Use $\alpha = 0.05$.

(2) What is the p-value?

(3) What is the probability of committing a Type II error when the actual mean difference of BM between NYSE and OTC stocks is 0.0629?

Sol: (1) 令μ_1、μ_2為 NYSE 以及 OTC 的 BM 值。

<1> $H_0: \mu_1 \leq \mu_2$

<2> $H_1: \mu_1 > \mu_2$

<3> $\alpha = 0.05$

<4> $C = \{Z | Z > 1.645\}$

<5> $\bar{x}_1 = 0.82$，$\bar{x}_2 = 0.77$，$\sigma_1^2 = 0.09$，$\sigma_2^2 = 0.04$，故檢定統計量

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{0.82 - 0.77}{\sqrt{\frac{0.09}{300} + \frac{0.04}{400}}} = 2.5 \in C$$

<6> 結論:拒絕H_0。有足夠的證據說明 NYSE 的平均 BM 值顯著高於 OTC。

(2) $p - value = P(Z > 2.5) = 0.0062$

(3) 因檢定的拒絕域為

$$C = \left\{ \bar{X}_1 - \bar{X}_2 | \bar{X}_1 - \bar{X}_2 > z_{0.05} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\} = \{ \bar{X}_1 - \bar{X}_2 | \bar{X}_1 - \bar{X}_2 > 0.0329 \}$$

故型二誤差的機率為

$$\beta = \mathrm{P}(\bar{X}_1 - \bar{X}_2 \le 0.0329 | \mu_1 - \mu_2 = 0.0629)$$

$$= \mathrm{P}\left(\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le \frac{0.0329 - 0.0629}{\sqrt{\frac{0.09}{300} + \frac{0.04}{400}}} \right) = \mathrm{P}(\mathrm{Z} \le -1.5) = 0.0668$$

8.3.2 兩獨立母體$\mu_1 - \mu_2$的檢定—以 T 分配處理

8.3.2.1　當抽自常態母體且$\sigma_1^2 = \sigma_2^2 = \sigma^2$

	$H_0: \mu_1 - \mu_2 \le d_0$ $H_1: \mu_1 - \mu_2 > d_0$	$H_0: \mu_1 - \mu_2 \ge d_0$ $H_1: \mu_1 - \mu_2 < d_0$	$H_0: \mu_1 - \mu_2 = d_0$ $H_1: \mu_1 - \mu_2 \ne d_0$
檢定法 (標準化)	$C = \{t \mid t > t_\alpha(n_1 + n_2 - 2)\}$	$C = \{t \mid t < -t_\alpha(n_1 + n_2 - 2)\}$	$C = \{t \mid \lvert t \rvert > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)\}$
p-value 法	$p - value = P(T > t \mid T \sim t(n_1 + n_2 - 2))$	$p - value = P(T < -t \mid T \sim t(n_1 + n_2 - 2))$	$p - value = 2P(T > \lvert t \rvert \mid T \sim t(n_1 + n_2 - 2))$

其中: $t = \frac{\bar{x} - \bar{y} - d_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

8.3.2.2　當抽自常態母體且$\sigma_1^2 \ne \sigma_2^2$

	$H_0: \mu_1 - \mu_2 \le d_0$ $H_1: \mu_1 - \mu_2 > d_0$	$H_0: \mu_1 - \mu_2 \ge d_0$ $H_1: \mu_1 - \mu_2 < d_0$	$H_0: \mu_1 - \mu_2 = d_0$ $H_1: \mu_1 - \mu_2 \ne d_0$
檢定法 (標準化)	$C = \{t \mid t > t_\alpha(v)\}$	$C = \{t \mid t < -t_\alpha(v)\}$	$C = \{t \mid \lvert t \rvert > t_{\frac{\alpha}{2}}(v)\}$
p-value 法	$P(T > t \mid T \sim t(v))$	$P(T < -t \mid T \sim t(v))$	$2P(T > \lvert t \rvert \mid T \sim t(v))$

其中: $t = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$, $v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$。

例題: A local pizza restaurant and a local branch of a national chain are located across the street from a college campus. The local pizza restaurant advertises that they deliver to the dormitories faster than the national chain. The delivery time of these two restaurants are summarized as follows.

	Local	**Chain**
Mean (minute)	16.7	18.88
Standard deviation	3.0955	2.8662
Observations	12	12

(1) Explain "Type I Error" in this case.

(2) Assuming the probability of type I error is 10%, can you verify the claim of local pizza restaurant? (Assume we know that the population variance of deliver time between two restaurants are same.)

Sol: (1) 設μ_1以及μ_2分別表示當地 pizza 店和連鎖 pizza 店平均配送所花費的時間，因此可寫出此假設為$H_0: \mu_1 \geq \mu_2$; $H_1: \mu_1 < \mu_2$。所以在此型 I 誤差的意思為抽樣後得到的結論為當地 pizza 店在廣告上的宣稱為真，但是事實上連鎖 pizza 店平均花費時間較久。

(2) <1> $H_0: \mu_1 \geq \mu_2$

<2> $H_1: \mu_1 < \mu_2$

<3> $\alpha = 0.1$

<4> $C = \{t | t < -t_{0.1}(22) = -1.32\}$

<5> 因$S_p^2 = \frac{(12-1)(3.0955)^2+(12-1)(2.8662)^2}{12+12-2} = 8.8986$

則檢定統計量為

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{16.7 - 18.88}{\sqrt{8.8986}\sqrt{\frac{1}{12}+\frac{1}{12}}} = -1.79 \in C$$

<6> 結論: 拒絕H_0，亦即當地 pizza 店的平均運送時間顯著較連鎖店的運送時間為快。

8.3.3 兩相關母體$\mu_1 - \mu_2$的檢定

	$H_0: \mu_1 - \mu_2 \leq d_0$ $H_1: \mu_1 - \mu_2 > d_0$	$H_0: \mu_1 - \mu_2 \geq d_0$ $H_1: \mu_1 - \mu_2 < d_0$	$H_0: \mu_1 - \mu_2 = d_0$ $H_1: \mu_1 - \mu_2 \neq d_0$
檢定法 (標準化)	$C = \{t \mid t > t_\alpha(n-1)\}$	$C = \{t \mid t < -t_\alpha(n-1)\}$	$C = \{t \mid \|t\| > t_{\frac{\alpha}{2}}(n-1)\}$
p-value 法	$p - value =$ $P(T > t \mid T \sim t(n-1))$	$p - value =$ $P(T < -t \mid T \sim t(n-1))$	$p - value =$ $2P(T > \|t\| \mid T \sim t(n-1))$

其中: $t = \frac{\bar{D}-d_0}{S_D/\sqrt{n}}$

例題: How does time spent using the computer impact the speed with which you work? A software company ran a study that looked at the effectiveness with which a person uses a mouse. It selected 10 people with the same computer skills and measured the speed with which they moved a mouse at the beginning of a long session of computer use and after two hours of use. The data (in hundredths of a

second) are shown below.

	1	2	3	4	5	6	7	8	9	10
Before	67	64	69	88	72	80	85	116	77	78
After	57	53	71	61	73	50	53	80	63	41

Assuming that the data are normally distributed, set up the hypotheses to test at the 0.05 level of significance whether there was a change in the mean speed with which the people moved the mouse.

Sol: 設μ_1為移動滑鼠平均速度，μ_2為兩個小時後移動的平均速度，則

<1> $H_0: \mu_1 = \mu_2$

<2> $H_1: \mu_1 \neq \mu_2$

<3> $\alpha = 0.05$

<4> $C = \{t | t > t_{0.025}(9) = 2.262\}$

<5> 因兩小時前後的移動速度差異為

	1	2	3	4	5	6	7	8	9	10
Before	67	64	69	88	72	80	85	116	77	78
After	57	53	71	61	73	50	53	80	63	41
$\boldsymbol{D_i}$	10	11	-2	27	-1	30	32	36	14	37

$$\overline{D} = 19.4，S_D = 14.8189$$

故檢定統計量為

$$t = \frac{\overline{D}}{S_D/\sqrt{n}} = \frac{19.4}{14.8189/\sqrt{10}} = 4.13986 \in C$$

<6> 結論: 拒絕H_0，亦即兩小時前後滑鼠移動速度有顯著差異。

8.4 單一母體比例的檢定問題

8.4.1 單一母體 p 的檢定—大樣本下以常態分配處理

大樣本為樣本數大於 30。

	$H_0: p \le p_0$ $H_1: p > p_0$	$H_0: p \ge p_0$ $H_1: p < p_0$	$H_0: p = p_0$ $H_1: p \ne p_0$
檢定法 (標準化)	$C = \{Z \mid Z > Z_\alpha\}$	$C = \{Z \mid Z < Z_\alpha\}$	$C = \{Z \mid \|Z\| > Z_{\frac{\alpha}{2}}\}$
信賴區間法	$p_0 \in \left(\hat{p} - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}, 1\right)$ 時不拒絕H_0	$p_0 \in \left(0, \hat{p} + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}\right)$ 時不拒絕H_0	$p_0 \in \left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}}\right)$ 時不拒絕H_0
p-value 法	$P(Z > z)$	$P(Z < z)$	$2P(Z > \|z\|)$

其中: $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

例題: An automobile manufacturer stated that it will be willing to mass produce electric-powered cars if more than 30% of potential buyers indicate they will purchase the newly designed electric cars. In a sample of 500 potential buyers, indicated that they would buy such a product. At 95% confidence, should the manufacturer produce the new electric powered car?

Sol: 設 p 表示潛在購買者的比例，則

<1> $H_0: p \le 0.3$

<2> $H_1: p > 0.3$

<3> $\alpha = 0.05$

<4> $C = \{Z | Z > 1.645\}$

<5> 故檢定統計量為

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.32 - 0.3}{\sqrt{\frac{0.3(0.7)}{500}}} = 0.9759 \notin C$$

<6> 結論: 不拒絕H_0，亦即無足夠證據說明有 30%的潛在客戶願意購買新車款。

8.4.2 單一母體 p 的檢定—小樣本下以二項分配處理

	$H_0: p \leq p_0$ $H_1: p > p_0$	$H_0: p \geq p_0$ $H_1: p < p_0$	$H_0: p = p_0$ $H_1: p \neq p_0$
檢定法 (標準化)	$C = \{X \mid X \geq k\}$ $\mathrm{P}(X \geq k \mid p = p_0)$ $= \sum_{x=k}^{n} \binom{n}{x} {p_0}^x (1-p_0)^{n-x}$ $\fallingdotseq \alpha$	$C = \{X \mid X \leq k\}$ $\mathrm{P}(X \leq k \mid p = p_0)$ $= \sum_{x=0}^{k} \binom{n}{x} {p_0}^x (1-p_0)^{n-x}$ $\fallingdotseq \alpha$	$C = \{X \mid X \leq k_1 \text{ 或 } X \geq k_2\}$ $\mathrm{P}(X \leq k_1 \mid p = p_0) \fallingdotseq \frac{\alpha}{2}$ $\mathrm{P}(X \geq k_2 \mid p = p_0) \fallingdotseq \frac{\alpha}{2}$
p-value 法	$\mathrm{p-value}$ $= \mathrm{P}(X \geq x \mid p = p_0)$	$\mathrm{p-value}$ $= \mathrm{P}(X \leq x \mid p = p_0)$	$p - value$ $= 2min\{\mathrm{P}(X \geq x \mid p = p_0) \; or \; \mathrm{P}(X \leq x \mid p = p_0)\}$

例題: 甲、乙兩職棒隊一年比賽 20 場，在顯著水準$\alpha = 0.05$下，試問甲隊在這 20 場比賽中至少要贏幾場以上，才能顯示該年甲隊對乙隊獲勝的機會超過五成?

Sol: 設 p 表示甲對乙的勝率，則$H_0: p \leq 0.5$，$H_1: p > 0.5$。令$C = \{X|X \geq k\}$，又

$$\mathrm{P}(X \geq k|p = 0.5) = \sum_{x=k}^{20} \binom{20}{x} 0.5^x (1-0.5)^{20-x} \fallingdotseq \alpha$$

由二項表可以得知$k - 1 = 14$，即$k = 15$。甲至少要贏 15 場才可說勝率超過五成。

8.5 兩獨立母體比例差的檢定

8.5.1 當$d_0 \neq 0$時

	$H_0: p_1 - p_2 \leq d_0$ $H_1: p_1 - p_2 > d_0$	$H_0: p_1 - p_2 \geq d_0$ $H_1: p_1 - p_2 < d_0$	$H_0: p_1 - p_2 = d_0$ $H_1: p_1 - p_2 \neq d_0$
檢定法 (標準化)	$C = \{Z\|Z > Z_\alpha\}$	$C = \{Z\|Z < Z_\alpha\}$	$C = \{Z\|\|Z\| > Z_{\frac{\alpha}{2}}\}$
p-value 法	$P(Z > z)$	$P(Z < z)$	$2P(Z > \|z\|)$

其中: $Z = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

8.5.2 當$d_0 = 0$時

	$H_0: p_1 - p_2 \leq d_0$ $H_1: p_1 - p_2 > d_0$	$H_0: p_1 - p_2 \geq d_0$ $H_1: p_1 - p_2 < d_0$	$H_0: p_1 - p_2 = d_0$ $H_1: p_1 - p_2 \neq d_0$
檢定法 (標準化)	$C = \{Z \mid Z > Z_\alpha\}$	$C = \{Z \mid Z < Z_\alpha\}$	$C = \{Z \mid \|Z\| > Z_{\frac{\alpha}{2}}\}$
p-value 法	$P(Z > z)$	$P(Z < z)$	$2P(Z > \|z\|)$

其中: $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}^*(1-\hat{p}^*)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$，$\hat{p}^* = \frac{x_1+x_2}{n_1+n_2}$。

例題: Surveys have been widely used by politicians around the world as a way of monitoring the opinions of the electorate. Six months ago, a survey was undertaken to determine the degree of support for a national party leader. Of a sample of 800, 46% indicated that they would vote for this politician. This month, another survey of 1100 voter revealed that 56% now support the leader.

(1) At the 1% significance level, can we infer that the national party leader's popularity has increased?

(2) At the 5% significance level, can we infer that the national party leader's popularity has increased by more than 5%?

Sol: (1) 令p_1，p_2分別表示為六個月前以及這個月領導者的支持度，則

<1> $H_0: p_1 \geq p_2$

<2> $H_1: p_1 < p_2$

<3> $\alpha = 0.01$

<4> $C = \{Z \mid Z < -2.326\}$

<5> 因$\hat{p}_1 = 0.46$，$\hat{p}_2 = 0.56$，且

$$\hat{p}^* = \frac{800 \times 0.46 + 1100 \times 0.56}{800 + 1100} = 0.5179$$

故檢定統計量為

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}^*(1-\hat{p}^*)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{0.46 - 0.56}{\sqrt{0.5179(0.4821)}\sqrt{\frac{1}{800}+\frac{1}{1100}}} = -4.31 \in C$$

<6> 結論: 拒絕H_0，亦即這個月領導者的支持度顯著性高於六個月前的支持度。

(2) <1> $H_0: p_1 - p_2 \geq -0.05$

<2> $H_1: p_1 - p_2 < -0.05$

<3> $\alpha = 0.05$

<4> $C = \{Z | Z < -1.645\}$

<5> 檢定統計量為

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{0.46 - 0.56 - (-0.05)}{\sqrt{\frac{0.46(0.54)}{800}+\frac{0.56(0.44)}{1100}}} = -2.1627 \in C$$

<6> 結論: 拒絕H_0，亦即這個月領導者的支持度顯著性高於六個月前的支持度5%。

8.6 常態母體變異數的檢定

8.6.1 單一母體變異數的檢定

8.6.1.1 μ未知下變異數的檢定

	$H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 \geq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$
檢定法 (標準化)	$C = \{\chi^2 \mid \chi^2 > \chi_\alpha^2(n-1)\}$	$C = \{\chi^2 \mid \chi^2 < \chi_{1-\alpha}^2(n-1)\}$	$C = \left\{\chi^2 \mid \chi^2 > \chi_{\frac{\alpha}{2}}^2(n-1) \text{ or } \chi^2 < \chi_{1-\frac{\alpha}{2}}^2(n-1)\right\}$

其中: $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$

8.6.1.2 μ已知下變異數的檢定

	$H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 \geq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$
檢定法 (標準化)	$C = \{\chi^2 \mid \chi^2 > \chi_\alpha^2(n)\}$	$C = \{\chi^2 \mid \chi^2 > \chi_{1-\alpha}^2(n)\}$	$C = \left\{\chi^2 \mid \chi^2 > \chi_{\frac{\alpha}{2}}^2(n) \text{ or } \chi^2 < \chi_{1-\frac{\alpha}{2}}^2(n)\right\}$

其中: $\chi^2 = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$

例題: Suppose the following data are the numbers of gallons of water used in a day to brush teeth by 12 randomly selected people and the data come from a normal distribution. Use these data and a 5% level of significance to test whether the population variance for such water usage is 2.5 gallons.

10 8 13 17 13 15 12 13 15 16 9 7

Sol: <1> $H_0: \sigma^2 = 2.5$

<2> $H_1: \sigma^2 \neq 2.5$

<3> $\alpha = 0.05$

<4> $C = \{\chi^2 | \chi^2 > \chi^2_{0.025}(11)\ or\ \chi^2 < \chi^2_{0.975}(11)\} = \{\chi^2 | \chi^2 > 21.92\ or\ \chi^2 < 3.81575\}$

<5> 因$S^2 \approx 10.424$

故檢定統計量為

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(12-1)10.424}{2.5} = 45.866 \in C$$

<6> 結論: 拒絕H_0，亦即σ^2顯著不等於 2.5。

8.6.2 兩個獨立母體變異數比的檢定

	$H_0: \sigma_1^2 \leq \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 \geq \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$
檢定法	$C = \{F \mid F > F_\alpha(n_1-1, n_2-1)\}$	$C = \{F \mid F < F_{1-\alpha}(n_1-1, n_2-1)\}$	$C = \left\{F \mid F > F_{\frac{\alpha}{2}}(n_1-1, n_2-1)\ or\ F > F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)\right\}$

其中: $F = \frac{S_1^2}{S_2^2}$

例題: Let the observed sample $x_1, x_2, \ldots, x_n$ be taken from $N(\mu_X, \sigma_X^2)$ and the sample $y_1, y_2, \ldots, y_m$ be taken from $N(\mu_Y, \sigma_Y^2)$. Given the data: $n = 25$, $\sum_{i=1}^n x_i = 845, \sum_{i=1}^n x_i^2 = 28678$, and $m = 29$, $\sum_{i=1}^m y_i = 918, \sum_{i=1}^n y_i^2 = 29231$.

Test:

(1) $H_0: \frac{\sigma_X^2}{\sigma_Y^2} = 1$ against a two sided alternative with significance level 0.02.

(2) $H_0: \mu_X = \mu_Y$ against $H_1: \mu_X > \mu_Y$ with significance level 0.01.

Note: (1) $F_{0.01}(24,28) = 2.52$，$F_{0.01}(28,24) = 2.61$

(2) $z_{0.01} = 2.326$

Sol: (1) <1> $H_0: \sigma_X^2 = \sigma_Y^2$

<2> $H_1: \sigma_X^2 \neq \sigma_Y^2$

<3> $\alpha = 0.02$

<4> $C = \left\{F | F > F_{0.01}(24,28) \; or \; F < \frac{1}{F_{0.01}(28,24)}\right\} = \{F | F > 0.383 \; or \; F < 2.52\}$

<5> 因$S_X^2 = 4.875$，$S_Y^2 = 6.127$

故檢定統計量為

$$F = \frac{4.875}{6.127} \fallingdotseq 0.796 \notin C$$

<6> 結論: 不拒絕H_0，亦即無足夠證據證明$\sigma_X^2 = \sigma_Y^2$。

(2) <1> $H_0: \mu_X = \mu_Y$

<2> $H_1: \mu_X \neq \mu_Y$

<3> $\alpha = 0.01$

<4> $C = \{t | t > t_{0.01}(52) = z_{0.01} = 2.326\}$

<5> 因$S_p^2 = \frac{(25-1)4.875+(29-1)6.127}{25+29-2} \fallingdotseq 5.549$

則檢定統計量為

$$t = \frac{\bar{x} - \bar{y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{33.8 - 31.655}{\sqrt{5.549}\sqrt{\frac{1}{25} + \frac{1}{29}}} = 3.337 \in C$$

<6> 結論: 拒絕H_0，亦即μ_X顯著大於μ_Y。

例題: A cigarette manufacturer tests tobacco of two different brands for nicotine content and independent random samples are selected from two normal populations. The data are shown below.

Brand A	22	20	21	19	18	
Brand B	21	29	23	31	20	32

Do these data provide sufficient evidence to indicate that the mean nicotine content for brand B is greater than that for brand A? State the appropriate hypothesis and use a significance level of 0.05.

Sol: 首先檢定兩常態母體的變異數是否相等，因$\bar{x}_A = 20$，$\bar{x}_B = 26$，$S_A^2 = 2.5$，$S_B^2 = 28$，所以

<1> $H_0: \sigma_A^2 = \sigma_B^2$

<2> $H_1: \sigma_A^2 \neq \sigma_B^2$

<3> $\alpha = 0.05$

<4> $C = \left\{F | F > F_{0.025}(4,5) \; or \; F < \frac{1}{F_{0.025}(5,4)}\right\} = \{F | F > 7.39 \; or \; F < 0.1068\}$

<5> 因$S_A^2 = 2.5$，$S_B^2 = 28$，故檢定統計量為

$$F = \frac{2.5}{28} \fallingdotseq 0.0893 \in C$$

<6> 結論: 拒絕H_0，亦即在顯著水準 0.05 下，證明σ_A^2與σ_B^2顯著不相等。

再來檢定B品牌尼古丁的含量是否高於 A 品牌的尼古丁含量，然而因為$\sigma_A^2 \neq \sigma_B^2$，所以用近似 T 分配來做檢定

<1> $H_0: \mu_A \geq \mu_B$

<2> $H_1: \mu_A < \mu_B$

<3> $\alpha = 0.05$

<4> $C = \{t | t < -t_{0.05}(v = 6) = -1.943\}$

而 v 的算法為

$$v = \frac{\left(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}\right)^2}{\frac{\left(\frac{S_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{S_B^2}{n_B}\right)^2}{n_B - 1}} = \frac{\left(\frac{2.5}{5} + \frac{28}{6}\right)^2}{\frac{\left(\frac{2.5}{5}\right)^2}{5 - 1} + \frac{\left(\frac{28}{6}\right)^2}{6 - 1}} = 6.042$$

<5> 檢定統計量為

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}} = \frac{20 - 26}{\sqrt{\frac{2.5}{5} + \frac{28}{6}}} = -2.64 \in C$$

<6> 結論: 拒絕H_0，亦即B品牌尼古丁的含量顯著高於 A 品牌的尼古丁含量。

第九章 變異數分析

在應用中時常需要討論依變數 (dependent variable) 與多個自變數 (independent variable) 之間的關係。前面已經討論過單一或兩個常態母體平均數的檢定，在此討論兩個或兩個以上常態母體平均數是否相等的檢定方法。雖然也可以使用均值差異的 t 值來作檢定，但此方法是不適當的，全部母體均值互相比較的結果犯型 I 誤差比一般設定的顯著水準高很多。例如在此有四個不同母體均值相互比較，全部共有$\binom{4}{2} = 6$對比較結果，若每對母體均值比較的顯著水準設定為$\alpha = 0.05$，則其信賴水準為 0.95。接受虛無假設時，6 對母體均值比較的正確精密度為$(0.95)^6 = 0.735$，則犯型 I 誤差的機率為 26.5%之多，所以要維持測驗的結果顯著水準為α，此方法稱為變異數分析 (Analysis of Variance ; ANOVA) 。

9.1 名詞定義

1. 因子或因素 (factor) : 為一個研究分析中的獨立變數。例如探討不同的栽種方式對於植株的高度是否有影響，栽種方式即為此研究主題的因子。

2. 因子水準 (factor level) : 是指所研究的因子的狀態及特殊形式，例如栽種方式有 A、B、C 三種，則此品牌因子即有三個水準。

3. 一因子與多因子分類: 在變異數分析中，所探討的對象如果只包含一個獨立變數時，稱之為一因子分類，但是如果所探討的對象如果同時包含兩個或兩個

以上獨立變數做分類，則稱之為多因子分類。

3. 處理 (treatment) : 處理的意義隨著研究時所探討的因子多寡而定。在一因子分析中，處理數即為因子的水準數。在多因子分析中，所謂處理數為不同因子其因子水準的組合數。

9.2 變異數分析的基本假設

為了要發展優良的檢定統計方法，必須要對鏡資料做出一些基本假設:

(1) 每一個處理所對應的母體分配，皆要假設服從常態母體。

(2) 每一個常態母體的變異數皆假設相等，亦即母體變異數具有同質性 (homoscedasticity) 。

(3) 來自每個常態母體的樣本資料皆具有隨機性。

註: 只要母體的分配不要太極端偏離常態分配，一般來說變異數分析是穩健 (robust) 的，特別是在觀察資料很多的情況下。

9.3 一因子變異數分析

若研究主題的對象只包含一個獨立變數(因子)稱之為一因子變異數分析 (one way ANOVA) 。

9.3.1 模式的建立與意義

若因子有 k 個水準 (k 個處理)，則可由此 k 個獨立的常態母體各抽出一組隨機

樣本

	母體						
	1	2	…	i	…	k	
	Y_{11}	Y_{21}	…	Y_{i1}	…	Y_{k1}	
	Y_{1j}	Y_{2j}	…	Y_{ij}	…	Y_{kj}	
	Y_{1n_1}	Y_{2n_2}	…	Y_{in_i}	…	Y_{kn_k}	
樣本和	$T_{1\cdot}$	$T_{2\cdot}$	…	$T_{i\cdot}$	…	$T_{k\cdot}$	$T_{\cdot\cdot}$
平均數	$\bar{Y}_{1\cdot}$	$\bar{Y}_{2\cdot}$	…	$\bar{Y}_{i\cdot}$	…	$\bar{Y}_{k\cdot}$	$\bar{Y}_{\cdot\cdot}$

- k 個因子水準皆對應於常態母體，所以知道每組隨機樣本的Y_{ij}皆服從常態分配，即

$$Y_{ij}\sim N(\mu_i,\sigma^2)，i=1,2,\ldots,k，j=1,2,\ldots,n_i$$

因每一個觀察值Y_{ij}與平均數μ_i之間可能有差異，故可將下是寫為

$$Y_{ij}=\mu_i+\varepsilon_{ij}$$

其中$\varepsilon_{ij}\overset{iid}{\sim}N(0,\sigma^2)$，$i=1,2,\ldots,k$，$j=1,2,\ldots,n_i$，此即為一因子變異數分析的模式。

- 一因子變異數分析的模式為

$$Y_{ij}=\mu_i+\varepsilon_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

其中$\varepsilon_{ij}\overset{iid}{\sim}N(0,\sigma^2)$，各個符號的意義為:

(1) Y_{ij}表示第 i 個母體 (或處理) 的第 j 個觀察值。

(2) μ_i為第 i 個母體的平均。

(3) μ為 k 個母體的全體平均數，即所有μ_i的平均。

(4) α_i表示第 i 個處理的處理效果 (treatment effect) ，即$\alpha_i = \mu_i - \mu$，它的限制為$\sum_{i=1}^{k} \alpha_i = 0$ (固定效果模式, fixed effect model)。

(5) ε_{ij}為第 i 個母體 (或處理) 的第 j 個觀察值Y_{ij}與對應的第 i 個處理效果平均數μ_i的差量，它表示隨機誤差。

- 假設檢定的建立

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

$H_1: \mu_i$不全相等，$i = 1,2,\ldots,k$

或是

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

$H_1: \alpha_i$不全為 0，$i = 1,2,\ldots,k$

9.3.2 變異數分析的檢定方法

- 總變異的分割

當試驗者對於同種事務給予不同處理時，事物對於不同處裡就產生不同的反應，試驗後就會有處理誤差 $(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})$，加上不明原因的試驗誤差 $(Y_{ij} - \bar{Y}_{i\cdot})$，兩種偏差和就是試驗所產生的總誤差，也就是試驗最後觀測值與總平均之差

$$Y_{ij} - \bar{Y}_{\cdot\cdot} = (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) + (Y_{ij} - \bar{Y}_{i\cdot})$$

但各觀測值誤差和為 0，所以將其觀測值的誤差平方後再加總，而成為平方和 (sum of squares)，由此平方和求得變異數，以求得進一步研究資料的性質。因此我們可得

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(\bar{Y}_{i.}-\bar{Y}_{..})+\left(Y_{ij}-\bar{Y}_{i.}\right)\right]^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.}-\bar{Y}_{..})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)(\bar{Y}_{i.}-\bar{Y}_{..})$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.}-\bar{Y}_{..})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2$$

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2$稱為總平方和 (Total Sum of Squares; *SST*)，$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.}-\bar{Y}_{..})^2$稱為處理平方和 (Treatment of Sum of Squares; *SSTR*)，而$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2$稱為誤差平方和 (Error of Sum of Squares; *SSE*)，故

$$SST=SSTR+SSE$$

- 自由度的分割

(1) *SST* 的自由度為$N-1$，其原因為總觀測值有$\sum_{i=1}^{k}n_i=N$個。但是$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)=0$，因而受到一個條件限制，故自由度少一個。

(2) *SSTR* 的自由度為$k-1$，因為有 k 個母體，但是$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)=0$，因而受到一個條件限制，故自由度少一個。

定理: 總變異與自由度的分割

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot}-\bar{Y}_{..})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i\cdot}\right)^2$$

變異	代號	自由度
$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2$	*SST*	$N-1$
$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot}-\bar{Y}_{..})^2$	*SSTR*	$k-1$
$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i\cdot}\right)^2$	*SSE*	$N-k$

$N=\sum_{i=1}^{k}n_i$

如果將平方和除上其自由度，即可得出其均方和。因此可得出處理均方和與誤差均方和為

$$MSTR=\frac{SSTR}{k-1}\text{，}MSE=\frac{SSE}{N-k}$$

則此均方和的期望值為

$$E(MSTR)=\sigma^2+\frac{\sum_{i=1}^{k}n_i\alpha_i^2}{k-1}$$

$$E(MSE)=\sigma^2\text{(因此在 ANOVA 中，}MSE\text{為}\sigma^2\text{的估計值)}$$

而處理均方和與誤差均方和的比值即為 Fisher 的 F 值

$$F=\frac{E(MSTR)}{E(MSE)}=\frac{\sigma^2+\frac{\sum_{i=1}^{k}n_i\alpha_i^2}{k-1}}{\sigma^2}=1+\frac{\frac{\sum_{i=1}^{k}n_i\alpha_i^2}{k-1}}{\sigma^2}$$

理論上若 F 值為 1，表示處理效應$\alpha_i=0$，即處理均值間無顯著差異，反之若 F

值大於 1，表示處理效應$\alpha_i \neq 0$，故一般 5%或 1%的 F 值表之 F 值最小值為 1，但在樣本資料有抽樣誤差 (sampling error) ，因此若 F 值大於 1 不一定代表處理均值有差異，要以其不同自由度的 F 值表比較後才能決定。由於在理論上分子必定比分母為大或相等，故 F 值檢定只能為右尾，則

$$F = \frac{MSTR}{MSE} \sim F(k-1, N-k)$$

且當$F > F_\alpha(k-1, N-k)$。變異數分析如下:

變異來源	平方和	自由度	均方和	F
處理	$SSTR$	$k-1$	$MSTR$	$F = \frac{MSTR}{MSE}$
誤差	SSE	$N-k$	MSE	
總變異	SST	$N-1$		

例題: Let $\mu_1, \mu_2,$ and μ_3 be the means of three normal distributions with a common but unknown variance σ^2. In order to test at the $\alpha = 0.05$ significance level, the hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$ against all possible alternative hypotheses, we take a random sample of size four from each of these distributions.

(1) Summarize the numerical results in an ANOVA table.

(2) Determine whether we accept or reject H_0.

The observed values from three distributions are:

Group1: 5, 6, 8, 9

Group2: 10, 11, 11, 12

Group3: 6, 7, 9, 10

Sol: (1) 經由計算得知$SST = 56.6667$、$SSTR = 34.6667$以及$SSE = 22$。故

ANOVA 表為

變異來源	平方和	自由度	均方和	F
處理	34.6667	2	17.3333	$F = 7.091$
誤差	22	9	2.4444	
總變異	56.6667	11		

(2) <1> $H_0: \mu_1 = \mu_2 = \mu_3$

<2> $H_1: \mu_i$不全相等，$i = 1, 2, 3$

<3> $\alpha = 0.05$

<4> $\mathrm{C} = \{F | F > F_{0.05}(2, 9) = 4.26\}$

<5> 計算: $F = 7.091 \in C$

<6> 拒絕H_0。即在 0.05 顯著水準下，$\mu_1, \mu_2,$和μ_3不全相等。

9.4 變異數同質性檢定

檢定各處理的變異數是否同質的檢定為:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1: \sigma_i^2 \neq \sigma_j^2 \left(\text{至少一對處理變異數不同}\right)$$

9.4.1 Hartley 檢定

在$n_1 = n_2 = \cdots = n_k = n$的情況下才可以進行。此檢定所採用的檢定統計量為

$$H = \frac{max\ (S_i^2)}{min\ (S_i^2)}$$

當 H 很大時，表示各樣本變異數差異大，故應拒絕 k 個母體變異數相等。相反地，當 H 很接近 1 時，則表示各樣本變異數差異小，故應不拒絕 k 個母體變異數相等。由此可知，Hartley 檢定的決策法則為

$$H > H_\alpha(k,n)，則拒絕H_0$$

其中$H_\alpha(k,n)$可由 Hartley 表查出。

9.4.2 Bartlett 檢定

其檢定統計量為

$$B = \frac{1}{C}\left[vln(S^2) - \sum_{i=1}^{k} v_i ln(S_i^2)\right]$$

其中$v_i = n_i - 1$，$v = \sum_{i=1}^{k} v_i$，$S^2 = \sum_{i=1}^{k} v_i S_i^2 / v$，

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{v}\right)$$

若$B > \chi_\alpha^2(k-1)$，則拒絕H_0。

9.4.3 Levene 檢定

將各處理觀測值與各處理平均值之差的絕對值 $(Z_{ij} = \left|X_{ij} - \bar{X}_i\right|)$作為新的觀測值進行變方分析。

例題: 今有一農藥進行殺菌的效果測驗，藥劑分為四個濃度: 100，150，200 以及 250ppm。各濃度下菌絲的直徑長度如下

100ppm	150ppm	200ppm	250ppm
28.0	22.0	9.5	2.0
40.0	15.0	10.0	2.4
35.2	24.0	14.5	2.6
34.0	23.0	16.6	4.0
31.0	22.5	8.4	2.0
25.8	18.5	12.0	
	20.0		

在顯著水準 0.05 下，試檢定各濃度下變異數是否相同。

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

$H_1: \sigma_i^2 \neq \sigma_j^2$ (至少一對處理變異數不同)

- Bartlett 檢定

$$v = \sum_{i=1}^{k} v_i = 5 + 6 + 5 + 4 = 20$$

$$S^2 = \sum_{i=1}^{k} v_i S_i^2 / v = 12.2545$$

其檢定統計量為

$$B = \frac{1}{C}\left[v ln(S^2) - \sum_{i=1}^{k} v_i ln(S_i^2)\right] = \frac{9.9917}{1.0852} = 9.20727 > \chi_{0.05}^2(3) = 7.815$$

拒絕H_0，表示四種濃度的變異數並非全部相等。

- Levene 檢定

$$Z_{ij} = |X_{ij} - \bar{X}_i|$$

100ppm	150ppm	200ppm	250ppm
4.33	1.29	2.33	0.60
7.67	5.71	1.83	0.20
2.87	3.29	2.67	0.00
1.67	2.29	4.77	1.40
1.33	1.79	3.43	0.60
6.53	2.21	0.17	
	0.71		

變異來源	平方和	自由度	均方和	F
濃度	33.5536	3	11.1845	$F = 3.5445$
誤差	63.1079	20	3.1554	
總變異	96.6615	23		

因$F = 3.5445 > F_{0.05}(3, 20) = 3.098$，則拒絕$H_0$，表示四種濃度的變異數並非全部相等。

9.5 多重比較法

在一因子 ANOVA 中，當檢定的結果為拒絕H_0，即表示各處理平均數間不全相等。有時需要更進一步探討任兩處理平均數差異的成對比較 (pairwise comparison)，觀察哪幾個處理平均數相等，哪幾個不相等，並找出所有處理平均數大小的排列順序，此為多重比較 (multiple comparison)法的概念。也就要找出$1 - \alpha$信心下，將$\binom{k}{2}$組$\mu_i - \mu_j$的信賴區間皆找出，並稱為聯立信賴區間

(simultaneous confidence intervals)。以下探討幾種常見的多重比較法:

9.5.1 Fisher 最小差異比較法

Fisher 最小差異比較法 (Least Significant Difference, *LSD*)是最早發展的方法。若

$$\left|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}\right| > LSD，1 \le i < j \le k$$

即表示μ_i與μ_j有顯著差異，其中

$$LSD = t_{\frac{\alpha}{2}}(N-k)\sqrt{MSE}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

例題: Given the following data (use $\alpha = 0.1$)

Sample 1	Sample 2	Sample 3
10	6	14
8	9	13
5	8	10
12	13	17
14		16
11		

(1) Construct an ANOVA table.

(2) Use Least Significance Difference (*LSD*) method to compare pairs of treatment mean.

Sol: (1) $SST = 172.933$，$SSTR = 66.933$，故 ANOVA table 為

變異來源	平方和	自由度	均方和	F
處理	66.933	2	33.467	$F = 3.789$
誤差	106	12	8.833	
總變異	172.933	14		

(2) 因$t_{0.05}(12)=1.782$，且$\bar{Y}_{1\cdot}=10$、$\bar{Y}_{2\cdot}=9$、$\bar{Y}_{3\cdot}=14$、$MSE=8.833$

所以

$$|\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}|=|10-9|=1<LSD=t_{0.05}(12)\sqrt{MSE}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$=1.782\sqrt{8.833}\sqrt{\frac{1}{6}+\frac{1}{4}}=3.419$$

即表示μ_1與μ_2無顯著差異。

$$|\bar{Y}_{1\cdot}-\bar{Y}_{3\cdot}|=|10-14|=4>LSD=1.782\sqrt{8.833}\sqrt{\frac{1}{6}+\frac{1}{5}}=3.207$$

即表示μ_1與μ_3有顯著差異。

$$|\bar{Y}_{2\cdot}-\bar{Y}_{3\cdot}|=|9-14|=5>LSD=1.782\sqrt{8.833}\sqrt{\frac{1}{4}+\frac{1}{5}}=3.553$$

即表示μ_2與μ_3有顯著差異。

9.5.2 Sheffe's 多重比較法

設$\{(L_{ij},U_{ij})|1\le i<j\le k\}$為$\{(\mu_i-\mu_j)|1\le i<j\le k\}$的$100(1-\alpha)\%$聯立信賴區間，且

$$\begin{cases}L_{ij}=\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}-\sqrt{(k-1)F_\alpha(k-1,N-k)}\sqrt{MSE}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\\ U_{ij}=\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}+\sqrt{(k-1)F_\alpha(k-1,N-k)}\sqrt{MSE}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\end{cases}$$

如果當$\mu_i-\mu_j$的信賴區間包含 0 時，即表示$\mu_i=\mu_j$，但當$\mu_i-\mu_j$的信賴區間皆為負值時，即表示$\mu_i<\mu_j$。相對來說，當$\mu_i-\mu_j$的信賴區間皆為正值時，即表

示$\mu_i > \mu_j$。

例題: A projector manufacturer is introducing a new product specifically targeted at the home market and wishes to compare the effectiveness of three strategies: Internet stores, home electronics stores, and department stores. The numbers of sales are shown below

Internet stores: 5, 4, 3, 3, 3

Home electronics stores: 9, 7, 8, 6, 5

Department stores: 7, 4, 8, 4, 3

(1) Test the hypothesis that there is no difference between the means of the retailers $(\alpha = 0.05)$.

(2) Use Scheffe's multiple comparison technique to determine which groups differ in mean sales $(\alpha = 0.05)$.

Note: $F_{0.05}(2,12) = 3.8853$

Sol: (1) 令μ_1、μ_2以及μ_3分別表示 internet stores、home electronics stores 以及 department stores 銷售的數量。

<1> $H_0: \mu_1 = \mu_2 = \mu_3$

<2> $H_1: \mu_i$ 不全相等，$i = 1, 2, 3$

<3> $\alpha = 0.05$

<4> $C = \{F|F > F_{0.05}(2,12) = 3.8853\}$

<5> 由 ANOVA 表可知

變異來源	平方和	自由度	均方和	F
處理	28.9333	2	14.46665	$F = 5.425$
誤差	32	12	2.66667	
總變異	60.9333	14		

$F = 5.425 \in C$

<6> 拒絕H_0，即表示個平均數有顯著差異。

(2) Scheffe 95%聯立信賴區間分別為

$\mu_1 - \mu_2$的信賴區間為

$$\left(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - \sqrt{2F_\alpha(2,12)}\sqrt{MSE}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} + \sqrt{2F_\alpha(2,12)}\sqrt{MSE}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

$$\Rightarrow \left(3.6 - 7 - \sqrt{2 \times 3.8853}\sqrt{2.6667}\sqrt{\frac{1}{5}+\frac{1}{5}}, 3.6 - 7 + \sqrt{2 \times 3.8853}\sqrt{2.6667}\sqrt{\frac{1}{5}+\frac{1}{5}}\right)$$

$\Rightarrow (-6.279, -0.521)$

即表示μ_1與μ_2有顯著差異 (即$\mu_1 < \mu_2$)。

$\mu_1 - \mu_3$的信賴區間為

$$\left(3.6 - 5.2 - \sqrt{2 \times 3.8853}\sqrt{2.6667}\sqrt{\frac{1}{5}+\frac{1}{5}}, 3.6 - 5.2 + \sqrt{2 \times 3.8853}\sqrt{2.6667}\sqrt{\frac{1}{5}+\frac{1}{5}}\right)$$

$\Longrightarrow (-4.479, 1.279)$

即表示μ_1與μ_3無顯著差異 (即$\mu_1 = \mu_3$)。

$\mu_2 - \mu_3$的信賴區間為

$$\left(7 - 5.2 - \sqrt{2 \times 3.8853}\sqrt{2.6667}\sqrt{\frac{1}{5}+\frac{1}{5}}, 7 - 5.2 + \sqrt{2 \times 3.8853}\sqrt{2.6667}\sqrt{\frac{1}{5}+\frac{1}{5}}\right)$$

$\Longrightarrow (-1.079, 4.679)$

即表示μ_2與μ_3無顯著差異 (即$\mu_2 = \mu_3$)。

9.5.3 Tukey 多重比較法

設$\{(L_{ij}, U_{ij})|1 \le i < j \le k\}$為$\{(\mu_i - \mu_j)|1 \le i < j \le k\}$的$100(1-\alpha)\%$聯立信賴區間，且

$$\begin{cases} L_{ij} = \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} - \dfrac{1}{\sqrt{2}} q_\alpha(k, N-k)\sqrt{MSE}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}} \\ U_{ij} = \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} + \dfrac{1}{\sqrt{2}} q_\alpha(k, N-k)\sqrt{MSE}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}} \end{cases}$$

如果當$\mu_i - \mu_j$的信賴區間包含 0 時，即表示$\mu_i = \mu_j$，但當$\mu_i - \mu_j$的信賴區間皆為負值時，即表示$\mu_i < \mu_j$。相對來說，當$\mu_i - \mu_j$的信賴區間皆為正值時，即表示$\mu_i > \mu_j$。

換句話說，令HSD (honestly significance difference)為

$$HSD = \frac{1}{\sqrt{2}} q_\alpha(k, N-k)\sqrt{MSE}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

當$\left|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}\right| > HSD$時，即表示第 i 與第 j 個平均數μ_i與μ_j有顯著差異。

例題: 消費者欲對台南市東區的三家牛肉麵加以評比。現有 9 位合格中年男性分配吃這三家的小碗牛肉麵，並於餐後打分數，從 1 分 (最難吃)，到 10 分 (最好吃)。實際實驗方法是從 9 為男性中隨機抽選三位吃 *A* 家牛肉麵、再抽三位試吃 *B* 家牛肉麵、最後三位試吃 *C* 家牛肉麵，分數後打好整理如下表:

店家	***A***	***B***	***C***
分數	2	8	4
	1.5	9	6
	2	8.5	5

(1) 以 ANOVA 分析三家牛肉麵是否依樣好吃。(以顯著水準 5%，*F* 臨界值 $F_{0.05}(2,6) = 5.14$檢驗。)

(2) 以 Tukey 多重比較法區別各家牛肉麵的好吃程度。(以顯著水準 5%，q 臨界值$q_{0.05}(3,6) = 4.34$檢驗。)

Sol: (1) 令μ_1、μ_2以及μ_3分別表示 *A*、*B* 以及 *C* 店的平均分數。

<1> $H_0: \mu_1 = \mu_2 = \mu_3$

<2> $H_1: \mu_i$ 不全相等，$i = 1, 2, 3$

<3> $\alpha = 0.05$

<4> $C = \{F | F > F_{0.05}(2,6) = 5.14\}$

<5> 由 ANOVA 表可知

變異來源	平方和	自由度	均方和	F
處理	66.7222	2	33.3611	$F = 75.062$
誤差	2.6667	6	0.44445	
總變異	69.3889	8		

$F = 75.062 \in C$

<6> 拒絕H_0，即表示三家牛肉麵的好吃程度有顯著差異。

(2) $HSD = q_\alpha(3,6)\sqrt{MSE}\sqrt{\frac{1}{n}} = q_\alpha(3,6)\sqrt{0.44445}\sqrt{\frac{0.44445}{3}} = 1.6705$

$|\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}| = |1.8333 - 8.5| = 6.6667 > HSD$

即表示μ_1與μ_2有顯著差異 (即$\mu_1 < \mu_2$)。

$|\bar{Y}_{1\cdot} - \bar{Y}_{3\cdot}| = |1.8333 - 5| = 3.1667 > HSD$

即表示μ_1與μ_3有顯著差異 (即$\mu_1 < \mu_3$)。

$|\bar{Y}_{2\cdot} - \bar{Y}_{3\cdot}| = |8.5 - 5| = 3.5 > HSD$

即表示μ_2與μ_3有顯著差異 (即$\mu_2 < \mu_3$)。

9.6 二因子變異數分析

一因子變異數分析只針對某一因子做因子分析，在其實存在影響較大的因子未考慮的因子的情況下，會使誤差變異加大，導致檢定效率較差。為了提高檢定的效率，有必要探討多因子變異數分析。而二因子分析又分為無重複試驗與重複試驗的二因子變異數分析。

9.6.1 未重複試驗之二因子變異數分析

在未重複的二因子變異數分析，除了 ANOVA 原本的三個假設外，還需要假設兩因子無交互作用存在，然後再對每處理內觀測一個值，即

		行						總和	平均
		1	2	…	j	…	c		
列	1	Y_{11}	Y_{12}	…	Y_{1j}	…	Y_{1c}	$T_{1\cdot}$	$\bar{Y}_{1\cdot}$
	i	Y_{i1}	Y_{i2}	…	Y_{ij}	…	Y_{ic}	$T_{i\cdot}$	$\bar{Y}_{i\cdot}$
	r	Y_{r1}	Y_{r2}	…	Y_{rj}	…	Y_{rc}	$T_{r\cdot}$	$\bar{Y}_{r\cdot}$
總和		$T_{\cdot 1}$	$T_{\cdot 2}$	…	$T_{\cdot j}$	…	$T_{\cdot c}$	$T_{\cdot\cdot}$	
平均		$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 2}$	…	$\bar{Y}_{\cdot j}$	…	$\bar{Y}_{\cdot c}$		$\bar{Y}_{\cdot\cdot}$

因$Y_{ij} \overset{iid}{\sim} N(\mu_{ij}, \sigma^2)$，$i = 1,2,\ldots,r$，$j = 1,2,\ldots,c$

故知二因子未重複試驗的模式為

$$Y_{ij} = \mu_{ij} + \epsilon_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

其中$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$，而$\alpha_i = \mu_{i\cdot} - \mu$及$\beta_i = \mu_{\cdot j} - \mu$分別為列效果及行效果，且$\sum_{i=1}^{r} \alpha_i = 0$，$\sum_{i=1}^{c} \beta_i = 0$。

定理: 二因子未重複試驗總變異與自由度的分割

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(Y_{ij}-\bar{Y}_{..}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{i.}-\bar{Y}_{..})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\bar{Y}_{.j}-\bar{Y}_{..}\right)^2$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y}_{..}\right)^2$$

變異	代號	自由度
$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(Y_{ij}-\bar{Y}_{..}\right)^2$	*SST*	$rc-1$
$\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{i.}-\bar{Y}_{..})^2$	*SSR*	$r-1$
$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\bar{Y}_{.j}-\bar{Y}_{..}\right)^2$	*SSC*	$c-1$
$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(Y_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y}_{..}\right)^2$	*SSE*	$(r-1)(c-1)$

- 各均方和為

$$MSR=\frac{SSR}{r-1}\text{，}MSC=\frac{SSC}{c-1}\text{，}MSE=\frac{SSE}{(r-1)(c-1)}$$

- 檢定統計量與決策規則:

(1) 列因子 (列效果) 的檢定:

$H_0:\mu_{1.}=\mu_{2.}=\cdots=\mu_{i.}$

$H_1:\mu_{i.}$不全相等，$i=1,2,\ldots,r$

或是

$H_0:\alpha_1=\alpha_2=\cdots=\alpha_i$

H_1: α_i不全為 0，$i = 1,2,\ldots,r$

當$F = \frac{MSR}{MSE} \sim F_\alpha(r-1,(r-1)(c-1))$，則拒絕$H_0$。

(2) 行因子 (行效果) 的檢定:

$H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot j}$

H_1: $\mu_{\cdot j}$不全相等，$j = 1,2,\ldots,c$

或是

$H_0: \beta_1 = \beta_2 = \cdots = \beta_i$

H_1: β_i不全為 0，$j = 1,2,\ldots,c$

當$F = \frac{MSC}{MSE} \sim F_\alpha(c-1,(r-1)(c-1))$，則拒絕$H_0$。

- 未重複二因子變異數分析表

變異來源	平方和	自由度	均方和	F
列間	SSR	$r-1$	MSR	$F_1 = \frac{MSR}{MSE}$
行間	SSC	$c-1$	MSC	$F_2 = \frac{MSC}{MSE}$
誤差	SSE	$(r-1)(c-1)$	MSE	
總變異	SST	$rc-1$		

例題: A company examines three new brands of cereal in selected outlets over a period of 4 months in a market.

Sales achieved (in thousands of dollars) are given in the table

	Brand A	Brand B	Brand C
March	47	52	60
April	56	54	52
May	49	63	55
June	41	44	48

(1) Set out the two-way analysis of variance table.

(2) Test the null hypothesis that the population mean sales are the same for all 3 brands of cereal.

Sol: (1) $SST = 448.25$、$SSR = 230.92$、$SSC = 74$、$SSE = 143.33$

故變異數分析表為

變異來源	平方和	自由度	均方和	F
月份	230.92	3	76.97	$F_1 = 3.22$
品牌	74	2	37	$F_2 = 1.55$
誤差	143.33	6	23.89	
總變異	448.25	11		

(2) <1> $H_0: \mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3}$

<2> $H_1: \mu_{\cdot j}$不全相等，$j = 1,2,3$

<3> $\alpha = 0.05$

<4> $C = \{F|F > F_{0.05}(2,6) = 5.1433\}$

<5> 計算: $F = 1.55 \notin C$

<6> 結論: 不拒絕H_0，即無證據顯示品牌不同隊銷售量有影響。

9.6.2 重複試驗之二因子變異數分析

若任一處理中僅有一個觀測值，則二因子交互作用的影響將與誤差項合併在一起，無法分開進行檢測。因此若重複多個觀測值，二因子交互作用的影響將可與誤差項分開，進而提升檢定效力。並且可由分割出的交互作用影響來檢定二因子間是否存在交互作用。

- 模型建立:

  因每個處理重複觀測 n 筆資料，在此資料中每個處理皆服從常態分配，所以

  $$Y_{ijk} \overset{iid}{\sim} N(\mu_{ij}, \sigma^2)，i = 1,2,\ldots,r，j = 1,2,\ldots,c，k = 1,2,\ldots,n$$

  所以二因子重複試驗的分析模式可以表示成

  $$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

  $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$。其中$\mu$為平均數，$(\alpha\beta)_{ij}$為列因子的第 i 個水準與行因子的第 j 個水準的交互效應，α_i為列因子的第 i 個水準的效應，β_j行因子的第 j 個水準的效應，ϵ_{ijk}為隨機誤差項。而且此模式受到以下幾個限制:

  $$\sum_{i=1}^{r} \alpha_i = 0，\sum_{j=1}^{c} \beta_j = 0，\sum_{i=1}^{r} (\alpha\beta)_{ij} = 0，\sum_{j=1}^{c} (\alpha\beta)_{ij} = 0$$

- 二因子重複試驗的資料

		行						總和	平均
		1	2	…	j	…	c		
列	1	Y_{111} Y_{11n}	Y_{121} Y_{12n}	…	Y_{1j1} Y_{1jn}	…	Y_{1c1} Y_{1cn}	$T_{1\cdot\cdot}$	$\bar{Y}_{1\cdot\cdot}$
	i	Y_{i11} Y_{i1n}	Y_{i21} Y_{i2n}	…	Y_{ij1} Y_{ijn}	…	Y_{ic1} Y_{icn}	$T_{i\cdot\cdot}$	$\bar{Y}_{i\cdot\cdot}$
	r	Y_{r11} Y_{r1n}	Y_{r21} Y_{r2n}	…	Y_{rj1} Y_{rjn}	…	Y_{rc1} Y_{rcn}	$T_{r\cdot\cdot}$	$\bar{Y}_{r\cdot\cdot}$
總和		$T_{\cdot 1\cdot}$	$T_{\cdot 2\cdot}$	…	$T_{\cdot j\cdot}$	…	$T_{\cdot c\cdot}$	$T_{\cdots}$	
平均		$\bar{Y}_{\cdot 1\cdot}$	$\bar{Y}_{\cdot 2\cdot}$	…	$\bar{Y}_{\cdot j\cdot}$	…	$\bar{Y}_{\cdot c\cdot}$		$\bar{Y}_{\cdots}$

定理: 二因子重複試驗總變異與自由度的分割

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(Y_{ijk}-\bar{Y}_{...}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(\bar{Y}_{i..}-\bar{Y}_{...})^2+\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(Y_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2+\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(Y_{ijk}-\bar{Y}_{ij.}\right)^2$$

變異	代號	自由度
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(Y_{ijk}-\bar{Y}_{...}\right)^2$	*SST*	$rcn-1$
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}(\bar{Y}_{i..}-\bar{Y}_{...})^2$	*SSR*	$r-1$
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2$	*SSC*	$c-1$
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2$	*SSI*	$(r-1)(c-1)$
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n}\left(Y_{ijk}-\bar{Y}_{ij.}\right)^2$	*SSE*	$rc(n-1)$

將各平方和除以自由度，即可知各均方和為

$$MSE=\frac{SSE}{rc(n-1)}\text{，}MSR=\frac{SSR}{r-1}\text{，}MSC=\frac{SSC}{c-1}\text{，}MSI=\frac{SSI}{(r-1)(c-1)}$$

- 檢定統計量及決策規則

(1) 列效果的檢定:

$H_0: \mu_{1\cdot\cdot} = \mu_{2\cdot\cdot} = \cdots = \mu_{r\cdot\cdot}$

$H_1: \mu_{i\cdot\cdot}$不全相等，$i = 1,2,\ldots,r$

或是

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$

$H_1: \alpha_i$不全為 0，$i = 1,2,\ldots,r$

當$F = \frac{MSR}{MSE} > F_\alpha\big(r-1, rc(n-1)\big)$時，則拒絕$H_0$。

(2) 行效果的檢定:

$H_0: \mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot} = \cdots = \mu_{\cdot c\cdot}$

$H_1: \mu_{\cdot j\cdot}$不全相等，$j = 1,2,\ldots,c$

或是

$H_0: \beta_1 = \beta_2 = \cdots = \beta_c = 0$

$H_1: \beta_j$不全為 0，$j = 1,2,\ldots,c$

當$F = \frac{MSC}{MSE} > F_\alpha\big(c-1, rc(n-1)\big)$時，則拒絕$H_0$。

(3) 交互效果的檢定:

$H_0: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{rc} = 0$

$H_1: (\alpha\beta)_{ij}$不全為 0，$i = 1,2,\ldots,r$，$j = 1,2,\ldots,c$

當$F = \frac{MSI}{MSE} > F_\alpha\big((r-1)(c-1), rc(n-1)\big)$時，則拒絕$H_0$。

- 二因子重複試驗的 ANOVA 表為

變異來源	平方和	自由度	均方和	F
列間	SSR	$r-1$	MSR	$F_1=\frac{MSR}{MSE}$
行間	SSC	$c-1$	MSC	$F_2=\frac{MSC}{MSE}$
交互作用	SSI	$(r-1)(c-1)$	MSI	$F_3=\frac{MSI}{MSE}$
誤差	SSE	$rc(n-1)$	MSE	
總變異	SST	$rcn-1$		

例題: The partially completed ANOVA for a 3×4 factorial experiment with two replications is shown below:

Source	***S.S***	***d.f.***	***M.S***	***F***
A	0.8			
B	5.3			
$\boldsymbol{A\times B}$	9.6			
Error				
Total	17.0			

(1) Complete the ANOVA table.

(2) Test to determine whether the main effects are warrant. Use $\alpha=0.05$ and interpret the result.

Sol: (1)

Source	***S.S***	***d.f.***	***M.S***	***F***
A	0.8	2	0.4	3.704
B	5.3	3	1.767	16.361
$\boldsymbol{A\times B}$	9.6	6	1.6	14.518
Error	1.3	12	0.108	
Total	17.0	23		

(2) (i) 檢定 A 因子效果

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$

$H_1: \alpha_i$不全為 0，$i = 1,2,3$

因拒絕域$C = \{F|F > F_{0.05}(2,12) = 3.8853\}$

且$F = \frac{MSA}{MSE} = 3.704 \notin C$，亦即無證據顯示 A 因子對資料有影響。

(ii) 檢定 B 因子效果

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1: \beta_j$不全為 0，$j = 1,2,\ldots,4$

因拒絕域$\mathrm{C} = \{\mathrm{F}|\mathrm{F} > F_{0.05}(3,12) = 3.4903\}$

且$F = \frac{MSB}{MSE} = 16.361 \in C$，亦即表示 B 因子對資料有顯著影響。

9.7 試驗設計

一般常見的試驗設計為下列兩種:

(1) 完全隨機設計 (Completely Randomized Design, CRD): 在研究中針對試驗單位採用隨機配置的做法，將所有試驗單位分配在不同處理內以從事試驗的一種設計方法

(2) 隨機集區設計 (Randomized Block Design, RBD): 將不同的試驗單位分為數個集區，而且每個集區內試驗單位的性質皆需要相同。然後在每個集區內隨機指派處理以從事試驗的設計方法。

第十章 迴歸分析

本章則要討論事物間與變數間的相關性，如施肥量與作物產量的關係，氣溫高低與飲料銷售量的關係，廣告投放量與營收額的關係都是。若將研究的變數以一個數學模式連接起來，並觀察其間的關係，即為迴歸分析 (regression analysis) 。

10.1 簡單直線迴歸模式

在直線迴歸模式中，假設獨立變數 X 與依變數 Y 接近直線關係，因此可將此模式表示為

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

其中

(1) Y_i 是依變數的第 i 個觀測值。

(2) X_i 是獨立變數的第 i 個觀測值。

(3) ϵ_i 是隨機誤差項，且$E(\epsilon_i) = 0$，$Var(\epsilon_i) = \sigma^2$。

(4) ϵ_i與ϵ_j無相關 (uncorrelated)，所以對所有 i、j 而言，$Cov(\epsilon_i, \epsilon_j) = 0$。

10.1.1 線性迴歸模式下的參數估計—最小平方法

由於迴歸係數β_0與β_1未知，所以需由樣本資料進行估計。在此只討論使用普通最小平方法 (ordinary least square, *OLS*) 所估計出的參數。

在簡單迴歸模式中，主要是對模式

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

中未知參數β_0與β_1進行估計。也就是利用

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

對母體的迴歸模式進行估計。而如果要尋找$\hat{\beta}_0$與$\hat{\beta}_1$使得$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$為最適合此組樣本的直線，可以利用實際觀測值$(Y_i)$以及預測值$(\hat{Y}_i)$的最小誤差平方和來尋找。即尋找$\hat{\beta}_0$與$\hat{\beta}_1$使得

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i\right)^2$$

為最小。

定理： 在迴歸模式$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$中，$\beta_0$與$\beta_1$的最小平方估計式為

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

例題: Suppose that the height x (inches) and weight y (pounds) of a women basketball player has a relation with equation of the for $Y = \beta_0 + \beta_1 X$.

(1) What are the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively, if we have *n* players' data: $(x_1, y_1), \ldots, (x_n, y_n)$.

(2) The following data were the height and weight of the starting line-ups of a team:

height	68	64	62	65	66
weight	132	108	102	115	128

Compute the estimated regression equation.

(3) If a player's height is 63 inches, what would you estimated her weight be?

Sol: (1) $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$，$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

(2) $\hat{\beta}_1 = 5.5$，$\hat{\beta}_0 = -240.5$，所以最小平方迴歸方程式為$\hat{Y} = -240.5 + 5.5X$

(3) 若某隊員的身高為$\mathrm{X} = 63$，則其體重預測值為

$$\hat{Y} = -240.5 + 5.5 \times 63 = 106$$

定理: 在迴歸模式$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$中，最小平方估計是具備以下的特性:

(1) $\sum_{i=1}^{n} e_i = 0$

(2) $\sum_{i=1}^{n} X_i e_i = 0$

(3) 樣本迴歸線必通過$(\bar{X}, \bar{Y})$

定理: 誤差平方和的計算公式

(1) $SSE = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i$

(2) $SSE = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

註: 需要計算 *SSE* 的目的為估計共同變異數σ^2，而$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}$。

例題: Let the observations for *X* and *Y* are the following table:

X	9	10	13	15	18	13
Y	36	44	48	63	70	45

(1) Find the sample regression line of *Y* on *X*.

(2) Find the unbiased estimator of the variance.

Sol: (1) $\hat{\beta}_1 = 3.7037$，$\hat{\beta}_0 = 2.8519$，所以最小平方迴歸方程式為$\hat{Y} = 2.8519 + 3.7037X$

(2) 因為$SSE = 83.26$，所以σ^2的不偏估計式為

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{83.26}{6-2} = 20.815$$

10.1.2 線性迴歸參數的統計推論

定理: (Gauss-Markov Theorem) 在簡單直線迴歸中，利用最小平方法求出的估計式滿足: (1) 線性，(2) 不偏性，(3) 最小變異。

此估計式又稱為 BLUE (best linear unbiased estimator)。

定理: 在常態迴歸模式下，β_1之$100(1-\alpha)\%$區間估計量為

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}}(n-2)\sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}, \hat{\beta}_1 + t_{\frac{\alpha}{2}}(n-2)\sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)$$

定理: 在常態迴歸模式下，關於β_1的檢定如下:

假設檢定	檢定統計量	拒絕域
$H_0: \beta_1 = b_1$ $H_1: \beta_1 \neq b_1$	$t = \dfrac{\hat{\beta}_1 - b_1}{\sqrt{\dfrac{MSE}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}}$	$C = \left\{t \mid \lvert t\rvert > t_{\frac{\alpha}{2}}(n-2)\right\}$
$H_0: \beta_1 \leq b_1$ $H_1: \beta_1 > b_1$		$C = \{t \mid t > t_{\alpha}(n-2)\}$
$H_0: \beta_1 \geq b_1$ $H_1: \beta_1 < b_1$		$C = \{t \mid t < -t_{\alpha}(n-2)\}$

定理: 在常態迴歸模式下，β_0之$100(1-\alpha)\%$區間估計量為

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}}(n-2)\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}, \hat{\beta}_0 + t_{\frac{\alpha}{2}}(n-2)\sqrt{MSE}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)$$

定理: 在常態迴歸模式下，關於β_1的檢定如下:

假設檢定	檢定統計量	拒絕域
$H_0: \beta_0 = b_0$ $H_1: \beta_0 \neq b_0$	$t = \dfrac{\hat{\beta}_1 - b_1}{\sqrt{MSE}\sqrt{\dfrac{1}{n}+\dfrac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}}$	$C = \left\{t \mid \lvert t\rvert > t_{\frac{\alpha}{2}}(n-2)\right\}$
$H_0: \beta_0 \leq b_0$ $H_1: \beta_0 > b_0$		$C = \{t \mid t > t_{\alpha}(n-2)\}$
$H_0: \beta_0 \geq b_0$ $H_1: \beta_0 < b_0$		$C = \{t \mid t < -t_{\alpha}(n-2)\}$

例題: The British Bankers' Association wanted to look at the relationship between the amount of deposits made (in billons of dollars) and the number of customers that a bank had. Analysts collected data on six different large banks and found the following information. Assuming linear regression model is used.

Bank Name	Deposit (billion dollars)	Customers (billion dollars)
Abbey National	101.7	13.6
Barclays	108.2	10.0
Lloyds	96.9	15.0
National Westminster	113.8	7.5
Woolrich	27.5	4.0
Halifax	77.1	7.6

(1) Write down the model. Which variable is the independent variable? Which variable is the dependent variable?

(2) Find the equation of the regression line for the data.

(3) At the 0.05 level, is the model significant?

Sol: (1) 迴歸模式為$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$，$i = 1,2,\ldots,n$，且 Customers 是獨立變數 X，而 Deposit 為相依變數 Y。

(2) $\hat{\beta}_1 = 4.9056$，$\hat{\beta}_0 = 40.3578$，所以最小平方迴歸方程式為$\hat{Y} = 40.3578 + 4.9056X$。

(3) <1> $H_0: \beta_1 = 0$

<2> $H_1: \beta_1 \neq 0$

<3> $\alpha = 0.05$

<4> $C = \{t||t| > t_{0.025}(4) = 2.776\}$

<5> $SSE = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 = 3070.691$，$MSE = \frac{SSE}{n-2} = 767.67275$，故

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} = \frac{4.9056}{\sqrt{\frac{767.67275}{85.0883}}} = 1.6332 \notin C$$

<6> 結論: 不拒絕H_0，無足夠證據顯示β_1不等於 0。

定理: ($\mu_{Y|X}$的統計推論) 常態迴歸模式下，在$X = X_0$時，$\mu_{Y|X}$的$100(1-\alpha)\%$區間估計量為

$$\left(\hat{Y}_{X_0} - t_{\frac{\alpha}{2}}(n-2)\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \hat{Y}_{X_0} + t_{\frac{\alpha}{2}}(n - 2)\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right)$$

定理: (新觀測值的統計推論) 常態迴歸模式下，當$X = X_h$時，新觀測值Y_{X_h}的$100(1-\alpha)\%$區間估計量為

$$\left(\hat{Y}_{X_h} - t_{\frac{\alpha}{2}}(n-2)\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \hat{Y}_{X_h} + t_{\frac{\alpha}{2}}(n - 2)\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right)$$

例題: The following data give the relationship between house values (*x*) and income (*y*) in thousands of dollars for a sample of twelve in which the *x* values were selected and a single random sample of *y* was selected corresponding to each *x* value.

House Value	54	58	63	66	70	72	75	84	90	100	110	120
Income	13	19	21	15	17	23	25	27	30	35	30	37

(1) Find the equation of the least square regression line.

(2) Assuming that the linear model under normality may be applied here. Find the 95% confidence interval for $\mu_{Y|X}$ for $X = 60$.

(3) Making the same assumptions as in (2), find the 95% prediction interval that would apply to a new observation, y, taken at $X = 60$.

Sol: (1) $\hat{\beta}_1 = 0.3376$，$\hat{\beta}_0 = -2.7309$，所以最小平方迴歸方程式為$\hat{Y} = -2.7309 + 0.3376X$。

(2) $MSE = \frac{SSE}{n-2} = 10.38932$，且$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 4849.667$，又$\hat{Y}_{60} = -2.7309 + 0.3376 \times 60 = 17.5251$，所以得出$\mu_{Y|X=60}$的 95%信賴區間為

$$\left(\hat{Y}_{60} - t_{\frac{\alpha}{2}}(10)\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(60-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}, \hat{Y}_{60} + t_{\frac{\alpha}{2}}(10)\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(60-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)$$

$$\Rightarrow \left(17.5251 - 2.228\sqrt{10.38932}\sqrt{\frac{1}{12} + \frac{(60-80.167)^2}{4849.667}}, 17.5251 + 2.228\sqrt{10.38932}\sqrt{\frac{1}{12} + \frac{(60-80.167)^2}{4849.667}}\right)$$

$$\Rightarrow (14.5887, 20.4615)$$

(3) 當$X = 60$時，新觀測值Y_{60}的 95%預測區間為

$$\left(\hat{Y}_{60} - t_{\frac{\alpha}{2}}(10)\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(60-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}, \hat{Y}_{60} + t_{\frac{\alpha}{2}}(10)\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(60-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)$$

$$\Rightarrow \left(17.5251 - 2.228\sqrt{10.38932}\sqrt{1 + \frac{1}{12} + \frac{(60-80.167)^2}{4849.667}}, 17.5251 +\right.$$

$$2.228\sqrt{10.38932}\sqrt{1+\frac{1}{12}+\frac{(60-80.167)^2}{4849.667}}\Big)$$

$$\Longrightarrow (9.7665, 25.2837)$$

定理: (σ^2的統計推論) 在常態迴歸模式下，共同變異數σ^2的$100(1-\alpha)\%$區間估計量為

$$\left(\frac{SSE}{\chi^2_{\frac{\alpha}{2}}(n-2)}, \frac{SSE}{\chi^2_{1-\frac{\alpha}{2}}(n-2)}\right)$$

例題: Production data for 22 firms in a certain industry produce the following, where *Y* is the output and *X* is the labor hours input:

$$\bar{Y} = 20，\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = 100，\bar{X} = 10，\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$= 60，\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = 30$$

Form a 99% confidence interval for σ^2.

Sol: 因為$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{30}{60} = 0.5$，所以

$$SSE = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2 = 100-(0.5)^2\times 60 = 85$$

所以得出σ^2的 99%信賴區間為

$$\left(\frac{SSE}{\chi^2_{0.005}(20)}, \frac{SSE}{\chi^2_{0.995}(20)}\right) = \left(\frac{85}{39.9968}, \frac{85}{7.43386}\right) = (2.1252, 11.4342)$$

定理: 在常態迴歸模式下，關於σ^2的檢定如下:

假設檢定	檢定統計量	拒絕域
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{SSE}{\sigma_0^2}$	$C = \left\{\chi^2 \middle\vert \chi^2 > \chi^2_{\frac{\alpha}{2}}(n-2) \; or \; \chi^2 < \chi^2_{1-\frac{\alpha}{2}}(n-2)\right\}$
$H_0: \sigma^2 \leq \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$		$C = \{\chi^2 \vert \chi^2 > \chi^2_{\alpha}(n-2)\}$
$H_0: \sigma^2 \geq \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$		$C = \{\chi^2 \vert \chi^2 < \chi^2_{1-\alpha}(n-2)\}$

10.1.3 迴歸分析中的變異數分析

之前討論過使用 *t* 檢定來探討自變數 *X* 與依變數 *Y* 間是否有顯著影響，但沒有說明其影響程度及解釋能力，所以使用變異數分析法來討論 *X* 對 *Y* 的影響。

- 總變異量與自由度的分割

  在此將總變異 (*SST*) 分割成迴歸平方和 (Sum of Squares of Regression, *SSR*)以及誤差平方和 (Sum of Squares of Errors, *SSE*)。*SSR* 又稱已解釋變異，而 *SSE* 又稱未解釋變異。

定理: 在此總變異與自由度的分割如下:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

變異	代號	自由度
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	*SST*	$n-1$
$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$	*SSR*	1
$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	*SSE*	$n-2$

註: $SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$

各個均方和為:

$$MSR = \frac{SSR}{1}\ ,\ SSE = \frac{MSE}{n-2}$$

- 判定係數

  其為自變數 X 對依變數 Y 變異解釋能力的指標。其計算式為

  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

  其性質為:

  (1) $0 \le R^2 \le 1$

  (2) 當R^2越大，X 解釋 Y 的能力就越強。

  (3) R^2為相關係數(r)的平方。

  (4) $\text{SSE} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,(1 - R^2)$

- 假設檢定

  $H_0: \beta_1 = 0$

  $H_1: \beta_1 \neq 0$

  當$F = \frac{MSR}{MSE} > F_\alpha(1, n-2)$則拒絕$H_0$。

  此分析的變異數分析表如下:

變異來源	**平方和**	**自由度**	**均方和**	***F* 值**
迴歸	*SSR*	1	*MSR*	$F = \frac{MSR}{MSE}$
誤差	*SSE*	$n-2$	*MSE*	
總變異	*SST*	$n-1$		

例題: You are given five points with these coordinates:

X	-2	-1	0	1	2
Y	1	2	4	4	3

(1) Estimate the linear regression line.

(2) Construct the ANOVA table for the regression.

Sol: (1) $\hat{\beta}_1 = 0.6$，$\hat{\beta}_0 = 2.8$，所以最小平方迴歸方程式為$\hat{Y} = 2.8 + 0.6X$。

(2) $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 6.8$，$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 = (0.6)^2 \times 10 = 3.6$，$SSE = 3.2$，因此 ANOVA 表為

變異來源	**平方和**	**自由度**	**均方和**	***F* 值**
迴歸	3.6	1	3.6	$F = 3.375$
誤差	3.1	3	1.0067	
總變異	6.8	4		

10.1.4 相關分析

簡單相關分析是討論兩變數 X 與 Y 之間相互的關係，並進一步做推論的方式。衡量兩變數的線性相關的方向與強度最常使用的方式為樣本相關係數，或是皮爾森相關係數。

定義: 令$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$為一組二維的隨機樣本，則兩變數 X 與 Y 之間的樣本共變數為

$$\hat{\sigma}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

樣本相關係數為

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

樣本相關係數的性質為:

(1) R^2稱為判定係數且$R^2 = r^2$。

(2) 樣本相關係數與簡單直線迴歸中斜率$\hat{\beta}_1$之關係密切，為

$$r = \hat{\beta}_1 \frac{S_X}{S_Y}$$

而S_X為 X 的標準差，S_Y為 Y 的標準差。

(3) 樣本相關係數的所在範圍為$0 \le r \le 1$。

(4) 樣本相關係數 r 的相關程度:

- $r = 1 \text{ or } -1$為完全線性相關
- $r = 0$為無線性相關
- $0.7 \le |r| < 1$為高度線性相關
- $0.3 \le |r| < 0.7$為中度線性相關
- $0 < |r| < 0.3$為低度線性相關

(5) 若$U_i = aX_i + b$，$V_i = cY_i + d$，且a、$c > 0$，則$r_{UV} = r_{XY}$。

(6) 有時看似兩變數 X 與 Y 之間有線性相關，不代表兩變數存在著因果關係，因為他們的相關性可能是由另一個變數造成的假性相關。

定理: 母體相關係數ρ的檢定方法:

假設檢定	檢定統計量	拒絕域
$H_0: \rho = 0$ $H_1: \rho \neq 0$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	$C = \left\{t \mid \|t\| > t_{\frac{\alpha}{2}}(n-2)\right\}$
$H_0: \rho \le 0$ $H_1: \rho > 0$		$C = \{t \mid t > t_{\alpha}(n-2)\}$
$H_0: \rho \ge 0$ $H_1: \rho < 0$		$C = \{t \mid t < -t_{\alpha}(n-2)\}$

例題: Under the classical linear regression assumptions, the least squares regression equation estimated from 52 observations is

$$Y_t = 12 + 0.8X_t + e_t \text{，} R^2 = 0.6$$

Also for $\alpha = 0.05,\ t_{0.025}(40) = 2.021,\ t_{0.05}(60) = 2.$ Please use the information above to carry out the test for the existence of a linear relation between *X* and *Y*.

Sol: <1> $H_0: \rho = 0$

<2> $H_1: \rho \neq 0$

<3> $\alpha = 0.05$

<4> $C = \{t||t| > t_{0.025}(50)\}$，$t_{0.05}(60) = 2 \leq t_{0.025}(50) \leq t_{0.025}(40) = 2.021$

<5> $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\sqrt{0.6}\sqrt{52-2}}{\sqrt{1-0.6}} = 8.66 \in C$

<6> 拒絕H_0，即表示 *X* 與 *Y* 之間有顯著線性相關。

10.1.5 殘差分析

由於只有在滿足常態迴歸模式基本假設成立下，上述的統計推論方法才會有效率，利用誤差項對研究問題的資料進行模式基本假設是否成立的分析，此種分析稱為殘差分析 (residual analysis)。

而針對下列問題進行討論:

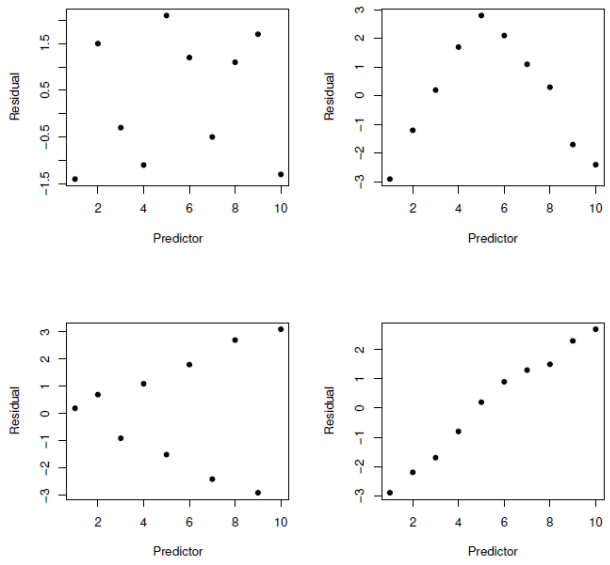
(1) 直線迴歸模式是否成立?

(2) 誤差項是否滿足同質性?

(3) 誤差項是否滿足隨機性?

(4) 誤差項是否滿足常態性?

- 圖形分析



只有左上角的誤差項滿足基本假設 (注意有無明顯規律!)，但是此種方式過於主觀。

- 直線假設的驗證

在簡單迴歸模型下，獨立變數與依變數的關係是否為直線須加以檢定，此方法稱為不適性檢定 (lack of fit test)。需要在每個獨立變數下重複測定數次依變數，資料的形式如下

相異自變數	對應的依變數					平均數	重複數
$\boldsymbol{X_1}$	Y_{11}	$\cdots$	Y_{1j}	$\cdots$	Y_{1n_1}	$\bar{Y}_{1\cdot}$	n_1
$\boldsymbol{X_2}$	Y_{21}	$\cdots$	Y_{2j}	$\cdots$	Y_{2n_2}	$\bar{Y}_{2\cdot}$	n_2
$\boldsymbol{X_i}$	Y_{i1}	$\cdots$	Y_{ij}	$\cdots$	Y_{in_i}	$\bar{Y}_{i\cdot}$	n_i
$\boldsymbol{X_k}$	Y_{k1}	$\cdots$	Y_{kj}	$\cdots$	Y_{kn_k}	$\bar{Y}_{k\cdot}$	n_k

此檢定的做法將殘差變異(*SSE*)分解成純誤差(sum of square of pure error,

SSP)與不適性(sum of square of lack of fit, *SSL*)

$$Y_{ij} - \hat{Y}_i = (Y_{ij} - \bar{Y}_{i\cdot}) + (\bar{Y}_{i\cdot} - \hat{Y}_i)$$

定理: 在此總變異與自由度的分割如下:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_{i\cdot} - \hat{Y}_i\right)^2$$

變異	代號	自由度
$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_i\right)^2$	*SSE*	$n-2$
$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$	*SSP*	$n-k$
$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_{i\cdot} - \hat{Y}_i\right)^2$	*SSL*	$k-2$

各均方和為

$$MSP = \frac{SSP}{n-k}\text{，}MSL = \frac{SSL}{k-2}$$

假設檢定為:

H_0: $\mu_{Y|X_i} = \beta_0 + \beta_1 X_i$

H_1: H_0不成立

若$F = \frac{MSL}{MSP} > F_\alpha(k-2, n-k)$，則拒絕$H_0$。

此檢定的變方分析表如下:

變異來源	平方和	自由度	均方和	F 值
迴歸	SSR	1	MSR	$F=\frac{MSL}{MSP}$
誤差	SSE	$(n-2)$	MSE	
欠合	SSL	$k-2$	MSL	
純誤差	SSP	$n-k$	MSP	
總變異	SST	$n-1$		

例題: 下列有一筆成人年齡與血壓的調查資料，每一年齡測量若干人，在顯著水準 0.05 下，測驗成人年齡與血壓是否有直線關係。如果是，則建立此迴歸模型，並檢定成人年齡對血壓是否有顯著影響。

年齡	20	30	40	50	60	70
血壓	102	120	126	135	150	160
	110	115	119	130	146	155
	108	118	120	136	148	159
		112		128	138	150
					140	

Sol: <1> $H_0: \mu_{Y|X_i}=\beta_0+\beta_1 X_i$

<2> $H_1: H_0$不成立

<3> $\alpha=0.05$

<4> $\mathrm{C}=\{\mathrm{F}|\mathrm{F}>F_{0.05}(4,17)=2.9647\}$

<5> 變方分析表如下:

變異來源	平方和	自由度	均方和	F 值
迴歸	6228.7096	1	6228.7096	$F=1.0420$
誤差	391.0295	(21)	18.6025	
欠合	76.9962	4	19.2491	
純誤差	314.0333	17	18.4725	
總變異	6619.7391	22		

<6> 不拒絕H_0，表示年齡與血壓呈直線關係。

<1> $H_0: \beta_1 = 0$

<2> $H_1: \beta_1 \neq 0$

<3> $\alpha = 0.05$

<4> $\mathrm{C} = \mathrm{F}|\mathrm{F} > F_{0.05}(1,21) = 4.3248$

<5> 變方分析表如下:

變異來源	平方和	自由度	均方和	F 值
迴歸	6228.7096	1	6228.7096	$F = 334.5082$
純誤差	391.0295	21	18.6205	
總變異	6619.7391	22		

<6> 拒絕H_0，表示成人年齡對血壓有顯著影響。$\hat{\beta}_1 = 0.9799$，$\hat{\beta}_0 = 85.5094$，所以最小平方迴歸方程式為$\hat{Y} = 85.5094 + 0.9799X$。

- 異質性檢定

  檢定各處理的變異數是否同質的檢定為:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1: \sigma_i^2 \neq \sigma_j^2 \ (\text{至少一對處理變異數不同})$$

  - Hartley 檢定

  在$n_1 = n_2 = \cdots = n_k = n$的情況下才可以進行。此檢定所採用的檢定統計量為

$$H = \frac{max\ (S_i^2)}{min\ (S_i^2)}$$

  當 H 很大時，表示各樣本變異數差異大，故應拒絕 k 個母體變異數相等。相反地，當 H 很接近 1 時，則表示各樣本變異數差異小，故應不拒絕 k 個

母體變異數相等。由此可知，Hartley 檢定的決策法則為

$$H > H_\alpha(k,n)，則拒絕H_0$$

其中$H_\alpha(k,n)$可由 Hartley 表查出。

- Bartlett 檢定

其檢定統計量為

$$B = \frac{1}{C}\left[v ln(S^2) - \sum_{i=1}^{k} v_i ln(S_i^2)\right]$$

其中$v_i = n_i - 1$，$v = \sum_{i=1}^{k} v_i$，$S^2 = \sum_{i=1}^{k} v_i S_i^2 / v$，

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{v}\right)$$

若$B > \chi_\alpha^2(k-1)$，則拒絕H_0。

- Levene 檢定

將各處理觀測值與各處理平均值之差的絕對值 $(Z_{ij} = \left|X_{ij} - \bar{X}_i\right|)$作為新的觀測值進行變方分析。

- 常態性檢定

詳見無母數方法。

- 隨機性檢定—Durbin-Watson Test

$H_0: \rho = 0$ $H_1: \rho \neq 0$	$H_0: \rho = 0$ $H_1: \rho < 0$	$H_0: \rho = 0$ $H_1: \rho > 0$

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=2}^{n} {e_i}^2}$$

在$n \geq 15$時，可利用DW表查出並判斷自身相關是否存在。

10.2 複迴歸模式

其基本假設與簡單迴歸類似，即

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i，i = 1,2,\ldots,n$$

(1) ϵ_i是隨機誤差項，且$E(\epsilon_i) = 0$，$Var(\epsilon_i) = \sigma^2$。

(2) ϵ_i與ϵ_j無相關 (uncorrelated)，所以對所有i、j而言，$Cov(\epsilon_i, \epsilon_j) = 0$。

(3) 隨機誤差項ϵ_i服從常態分配。

(4) k個自變數$X_1, X_2, \ldots, X_k$，即無線性重合 (multicollinearity)。

10.2.1 二元線性迴歸

其模型為

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i，i = 1,2,\ldots,n$$

定理： 在二元線性迴歸模式中，參數β_0、β_1及β_2的最小平方估計式為

$$\hat{\beta}_1 = \frac{S_{22}S_{1Y} - S_{12}S_{2Y}}{S_{11}S_{22} - S_{12}^2}$$

$$\hat{\beta}_2 = \frac{S_{11}S_{2Y} - S_{12}S_{1Y}}{S_{11}S_{22} - S_{12}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

而

$$S_{11} = \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2$$

$$S_{22} = \sum_{i=1}^{n} (X_{2i} - \bar{X}_2)^2$$

$$S_{1Y} = \sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(Y_i - \bar{Y})$$

$$S_{2Y} = \sum_{i=1}^{n} (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

定理: 在二元線性迴歸模式中，迴歸係數β_1及β_2的$100(1-\alpha)\%$區間估計量為

(1) β_1的$100(1-\alpha)\%$區間估計量為

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}}(n-3) \cdot S(\hat{\beta}_1), \hat{\beta}_1 + t_{\frac{\alpha}{2}}(n-3) \cdot S(\hat{\beta}_1)\right)$$

其中$S(\hat{\beta}_1) = \sqrt{\frac{S_{22}MSE}{S_{11}S_{22}-S_{12}^2}}$，$MSE = \frac{SSE}{n-3}$，$SSE = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}Y_i - \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}Y_i$。

(2) β_2的$100(1-\alpha)\%$區間估計量為

$$\left(\hat{\beta}_2 - t_{\frac{\alpha}{2}}(n-3) \cdot S(\hat{\beta}_2), \hat{\beta}_2 + t_{\frac{\alpha}{2}}(n-3) \cdot S(\hat{\beta}_2)\right)$$

其中$S(\hat{\beta}_2) = \sqrt{\frac{S_{11}MSE}{S_{11}S_{22}-S_{12}^2}}$

定理: 在二元線性迴歸模式中，β_i的檢定為

假設檢定	檢定統計量	拒絕域
$H_0: \beta_i = b_i$ $H_1: \beta_i \neq b_i$	$t = \dfrac{\hat{\beta}_i - b_i}{S(\hat{\beta}_i)}$	$C = \left\{t \mid \lvert t\rvert > t_{\frac{\alpha}{2}}(n-3)\right\}$
$H_0: \beta_i \leq b_i$ $H_1: \beta_i > b_i$		$C = \{t \mid t > t_{\alpha}(n-3)\}$
$H_0: \beta_i \geq b_i$ $H_1: \beta_i < b_i$		$C = \{t \mid t < -t_{\alpha}(n-3)\}$

例題: 設有變數 Y，X_1，X_2資料如下:

Y	X_1	X_2
6	1	0
10	1	1
10	2	0
14	2	1
16	3	0
20	3	1

(1) 估計出此二元迴歸模型

(2) 檢定$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 > 0$

Sol: (1) $S_{YY} = 125.333$、$S_{11} = 4$、$S_{22} = 1.5$、$S_{12} = 0$、$S_{1Y} = 20$、$S_{2Y} = 6$

因此

$$\hat{\beta}_1 = \frac{S_{22}S_{1Y} - S_{12}S_{2Y}}{S_{11}S_{22} - S_{12}^2} = 5$$

$$\hat{\beta}_2 = \frac{S_{11}S_{2Y} - S_{12}S_{1Y}}{S_{11}S_{22} - S_{12}^2} = 4$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2 = 0.6667$$

故知二元最小平方迴歸方程式為$\hat{Y} = 0.6667 + 5X_1 + 4X_2$。

(2) <1> $H_0: \beta_1 = 0$

<2> $H_1: \beta_1 > 0$

<3> $\alpha = 0.05$

<4> $C = \{t | t > t_{0.05}(3) = 2.353\}$

<5> $SSE = 1.333$、$MSE = \frac{SSE}{n-3} = 0.4443$，所以

$$S(\hat{\beta}_1) = \sqrt{\frac{S_{22}MSE}{S_{11}S_{22} - S_{12}^2}} = \sqrt{\frac{1.5 \times 0.4443}{4 \times 1.5 - 0}} = 0.3333$$

檢定統計量為

$$t = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{5}{0.3333} = 15 \in C$$

<6> 結論: 拒絕H_0，亦即β_1顯著大於 0。

10.2.2 多元線性迴歸分析

k個自變數的線性迴歸模型與之前討論的簡單與二元線性迴歸相似，但是其推導較為複雜，所以在此大部分會需要藉由統計軟體的運算作為輔助。其統計推論分為下列四項進行討論:

(1) 個別迴歸係數β_i的檢定

(2) 聯合檢定

(3) 部分項檢定

(4) 參數線性組合檢定

(1) 個別迴歸係數β_i的檢定

定理: 多元常態迴歸模式下，關於個別自變數X_i對相依變數Y是否有顯著影響的檢定

$$H_0: \beta_i = 0，i = 1,2,\ldots,k$$

$$H_1: \beta_i \neq 0$$

其檢定統計量為 t-ratio，即

$$t = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} \sim t(n-k-1)$$

而檢定的拒絕域為

$$C = \left\{ t \middle| |t| > t_{\frac{\alpha}{2}}(n-k-1) \right\}$$

β_i的$100(1-\alpha)\%$區間估計量為

$$\left(\hat{\beta}_i - t_{\frac{\alpha}{2}}(n-k-1) \cdot S(\hat{\beta}_i), \hat{\beta}_i + t_{\frac{\alpha}{2}}(n-k-1) \cdot S(\hat{\beta}_i) \right)$$

(2) 聯合檢定

定理: 多元常態迴歸模式下，關於 k 個自變數對相依變數Y是否有顯著影響的檢定

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \beta_i 不全為\ 0，i = 1,2,\ldots,k$$

其檢定統計量為

$$F = \frac{MSR}{MSE} \sim F(k, n-k-1)$$

而檢定的拒絕域為

$$C = \{F | F > F_\alpha(k, n-k-1)\}$$

其中

變異來源	平方和	自由度	均方和	F 值
迴歸	*SSR*	1	*MSR*	$F = \frac{MSR}{MSE}$
誤差	*SSE*	$n-2$	*MSE*	
總變異	*SST*	$n-1$		

例題: The following regression model has been used to estimate the demand for imports into the US:

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

where Y_t is the value of imported manufactured gods in year t ($billion); X_{1t} is an index of the relative price of domestic goods to imported goods in year *t*; X_{2t} is the value of Gross Domestic Product (GDP) in year *t* ($100 billion).

The following computer output represents a regression using annual data from 1970 to 1986.

The regression equation is

$$Y = 148 + 25.9X_1 + 0.115X_2$$

Predictor	Coefficient	Stdev	t-ratio
Constant	147.82	15.09	9.79
X_1	25.894	5.47	4.73
X_2	0.1494	0.05297	2.17

$s = 10.18$，$R^2 = 64.2\%$，$R^2_{adj} = 59.1\%$

SOURCE	S.S	d.f	M.S
Regression	2600.3	2	1300.2
Error	1449.6	14	103.5
Total	4049.9	16	

Required:

(1) Obtain a 95% confidence interval for the true coefficient of X_{1t}.

(2) Is the coefficient of the relative price index statistically significantly different from zero under 0.05 significance level?

(3) Test the hypothesis that $\beta_1 = 20$ against the alternative that it is greater than 20 under 0.05 significance level.

(4) Do the 2 exploratory variables X_{1t} and X_{2t} (considered together) have a statistically significant effect on the level of imports under 0.05 significance level?

Sol: (1) β_1的 95%信賴區間為

$$\left(\hat{\beta}_1 - t_{0.025}(14)\cdot S(\hat{\beta}_1), \hat{\beta}_1 + t_{\frac{\alpha}{2}}(14)\cdot S(\hat{\beta}_1)\right)$$

$$\Longrightarrow (25.894 - 2.145\cdot 5.47, 25.894 + 2.145\cdot 5.47)$$

(2) <1> $H_0: \beta_1 = 0$

<2> $H_1: \beta_1 \neq 0$

<3> $\alpha = 0.05$

<4> $C = \{t \mid |t| > t_{0.025}(14) = 2.145\}$

<5> 因為

$$t = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{25.894}{5.47} = 4.73 \in C$$

<6> 拒絕H_0，即相對價格指數的係數顯著不等於 0。

(3) <1> $H_0: \beta_1 = 20$

<2> $H_1: \beta_1 > 20$

<3> $\alpha = 0.05$

<4> $C = \{t \mid t > t_{0.05}(14) = 1.761\}$

<5> 因為

$$t = \frac{\hat{\beta}_1 - 20}{S(\hat{\beta}_1)} = \frac{25.894 - 20}{5.47} = 1.0775 \notin C$$

<6> 不拒絕H_0，即表示無證據顯示$\beta_1 > 20$。

(4) <1> $H_0: \beta_1 = \beta_2 = 0$

<2> $H_1: \beta_i$不全為 0，$i = 1, 2$。

<3> $\alpha = 0.05$

<4> $C = \{F | F > F_{0.05}(2,14) = 3.7389\}$

<5> 因為

$$F = \frac{MSR}{MSE} = \frac{1300.2}{103.5} = 12.5623 \in C$$

<6> 不拒絕H_0，即表示兩個自變數對進口需求有顯著影響。

(3) 部分項檢定：聯合檢定是探討 k 個自變數一起考慮對依變數 Y 是否有影響，只是有些時候如果只想要檢定，若只想知道部分自變數是對依變數具有影響時，就必須利用部分項檢定

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_q X_{qi} + \cdots + \beta_q X_{qi} + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_q = 0$，$q \leq k$

$H_1: \beta_i$不全為 0，$i = 1,2,\ldots,q$

此檢定可利用一般線性檢定之，且檢定統計量為

$$F = \frac{[SSE(R) - SSE(F)] \div [df_R - df_F]}{SSE(F) \div df_F}$$

其中$SSE(R)$為遞減模式(reduced model)下的誤差平方和，$SSE(F)$為完整模式(full model)下的誤差平方和。當$F > F_\alpha(df_R - df_F, df_F)$時，拒絕$H_0$。

例題: A regression analysis (involving 45 observations) relating a dependent variable (Y) and two independent variables resulted in the following information.

$$\hat{Y} = 0.408 + 1.3387X_1 + 2X_2$$

The *SSE* for the above model is 49.

When two other independent variables were added to the model, the following information was provided.

$$\hat{Y} = 1.2 + 3.0X_1 + 12X_2 + 4.0X_3 + 8X_4$$

The latter model's *SSE* is 40.

At 95% confidence test to determine if the two added independent variables contribute significantly to the model.

Sol: <1> $H_0: \beta_3 = \beta_4 = 0$

<2> $H_1: \beta_i$不全為 0，$i = 3, 4$

<3> $\alpha = 0.05$

<4> $C = \{F | F > F_{0.05}(2, 40) = 3.2317\}$

<5> 因為

$$F = \frac{[SSE(R) - SSE(F)] \div [df_R - df_F]}{SSE(F) \div df_F} = \frac{[49 - 40] \div [(n-3) - (n-5)]}{40 \div (n-5)} = 4.5$$

$\in C$

<6> 拒絕H_0，即新增X_3與X_4對模型有顯著影響。

10.2.3 複相關與偏相關分析

- 複判定係數: 指 k 個自變數$X_1, X_2, \ldots, X_k$引入迴歸模式中，依變數的總變異被 k 個自變數$X_1, X_2, \ldots, X_k$解釋的比例。

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- 修正判定係數: 一般認為R^2越高模式越好，但事實上自變數美多放一個的同時，誤差平方和必定會降低，因此即使放入的是一個毫無意義的自變數

進去，R^2仍會增加。要修正上述的缺失，就必須考慮修正後的判定係數(adjusted determination coefficient)，來進行檢驗。而修正後的判定係數公式為

$$R_{adj}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

例題: Comparing the multiple coefficient of determination and adjusted multiple coefficient of determination for the same regression model, which of the following is correct?

(A) When an additional independent variable is added to the model, R^2 will always increase.

(B) When an additional independent variable is added to the model, R_{adj}^2 will always increase.

(C) R_{adj}^2 be greater than R^2.

(D) Both (A) and (B) are right.

Sol: (A)

(B) 錯，增加自變數R_{adj}^2不一定會增加。

(C) 錯，$R_{adj}^2 \leq R^2$。

- 偏判定係數與偏相關分析
  - 額外平方和: 在迴歸模式中自變數已經固定下，再增加一個或多個自變數進入此迴歸模式中，其減少的誤差平方和或所增加的迴歸平方和稱為額外平方和 (extra sums of squares)。

    1. 考慮X_1，X_2兩個自變數時: 在X_1固定下，引進一個新的自變數X_2時，其額外平方和為

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = SSR(X_1, X_2) - SSR(X_1)$$

在X_2固定下，引進一個新的自變數X_2時，其額外平方和為

$$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2) = SSR(X_1, X_2) - SSR(X_2)$$

2. 考慮 3 個自變數X_1，X_2，X_3時

$$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$
$$= SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$$

- 偏判定係數: 此意義為在迴歸模式中已有自變數固定下，另一個新的自變數加入，他可增加對 Y 變異解釋的貢獻。當自變數為兩個時，其公式定義為

$$R^2_{Y2\cdot1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = 1 - \frac{SSE(X_1, X_2)}{SSE(X_1)}$$

當自變數為三個時，其公式定義為

$$R^2_{Y3\cdot12} = \frac{SSR(X_1|X_2)}{SSE(X_1, X_2)} = 1 - \frac{SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)}$$

- 偏相關係數: 為偏判定係數的平方根。此意義為在迴歸模式中已有自變數固定下，另一個新的自變數加入，其與依變數的相關程度。當自變數為兩個時，其公式定義為

$$r_{Y2\cdot1} = \pm\sqrt{R^2_{Y2\cdot1}} = \pm\sqrt{\frac{SSR(X_2|X_1)}{SSE(X_1)}} \quad (\text{正負號依據}\hat{\beta}_2)$$

或是

$$r_{Y2\cdot1} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{1 - r^2_{Y1}}\sqrt{1 - r^2_{12}}}，r_{Y1\cdot2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1 - r^2_{Y2}}\sqrt{1 - r^2_{12}}}$$

當自變數為三個時，其公式定義為

$$r_{Y3\cdot12} = \pm\sqrt{R^2_{Y3\cdot12}} = \pm\sqrt{\frac{SSR(X_1|X_2)}{SSE(X_1, X_2)}} \quad (\text{正負號依據}\hat{\beta}_3)$$

例題: The zero-order correlation among variables *X*, *Y*, and *Z* are displayed in the following table.

(1) Calculate the partial correlation coefficient that measures the relationship between *X* and *Y* while controlling for *Z*.

(2) Interpret the result you have obtained in (1).

	Y	***X***	***Z***
Y	1	0.5	-0.3
X		1	-0.47
Z			1

Sol: (1) 已知$r_{X,Y} = 0.5$，$r_{Y,Z} = -0.3$，$r_{X,Z} = -0.47$，故

$$r_{YX\cdot Z} = \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{1 - r_{YZ}^2}\sqrt{1 - r_{XZ}^2}} = \frac{0.5 - (-0.3)(-0.47)}{\sqrt{1 - (-0.3)^2}\sqrt{1 - (-0.47)^2}} = 0.4264$$

(2) 當迴歸模式已有 *Z* 時，引進一個新的自變數 *X*，其與依變數 *Y* 的相關程度為 0.4264。

10.2.4 線性重合

在多元迴歸模式下，若自變數之間有高度線性相關時，則此現象稱之為線性重合或共線性 (multicollinearity)。而共線性對於統計推論有以下幾點影響:

(1) 線性重合下，由於最小平方法下迴歸係數的估計的變異數會較大，此結果會使統計推論時失去應有的檢定效率與精確度。

(2) 線性重合下，會使得個別迴歸係數β_i的檢定結果不顯著，但聯合檢定結果呈現結果顯著的不一致狀況。

(3) 線性重合下，會使重要獨立變數的檢定結果不顯著。

- 檢測方法

  (1) 利用簡單相關係數的概念計算兩變數間的相關性。$|r_{i,j}| \geq 0.7$，則表示獨立變數間具有高度線性相關。

  (2) 變異數膨脹因素法 (variance inflation factors, *VIF*): 將迴歸模式中，其中某個自變數X_j視為依變數，而將其餘k − 1個自變數當成解釋變數來做多元迴歸模式，將此模式的R_j^2 (複判定係數)來判斷，其 *VIF* 定義為

$$VIF_j = \frac{1}{1 - R_j^2} \text{ ，} j = 1,2,\ldots,k$$

  若$R_j^2 \to 0$或是$VIF_j \to 1$，顯示X_j和其他自變數無關。若$R_j^2 \to 1$或是$VIF_j \to \infty$，顯示X_j和其他自變數無關。因此若R_j^2越大代表VIF_j越大，即表示共線性越明顯。通常當$VIF > 10$時表示高度共線性。

- 補救方法

  (1) 將高度共線性的獨立變數刪除即可。

  (2) 大量增加樣本數，可減少共線性的影響程度。

  (3) 利用脊回歸 (ridge regression) 或是主成分分析 (principle component analysis, PCA)。

10.2.5 虛擬變數

迴歸分析中，依變數不僅會受到屬量變數的影響，同時亦會受到一些屬質變數影響。因此需要了解屬質變數在在迴歸模式中扮演的角色，就必須了解虛擬變數 (dummy variable) 方法。

一般虛擬變數的設定通常皆以指標變數 (indicator variable) 值 0 和 1 為主。若獨立變數為屬質變數，且該變數有 c 個水準，則必須選擇$c - 1$個虛擬變數，且其設定值皆為 0 及 1。例如，性別有男性及女性，則我們可以令男性為 1，女性為 0，則此虛擬變數可以寫成

$$D=\begin{cases}1，男性\\0，女性\end{cases}$$

如果區域分為北、中、南三區時，必須設立兩個虛擬變數D_1及D_2，則令

$$D_1=\begin{cases}1，北區\\0，其他\end{cases}，D_2=\begin{cases}1，中區\\0，其他\end{cases}$$

若屬質變數有兩個水準，只需一個虛擬變數即可。因此可將此迴歸模式寫為

$$Y_i=\beta_0+\beta_1 D_i+\varepsilon_i，\varepsilon_i\sim N(0,\sigma^2)$$

此假設檢定為

$$H_0:\beta_1=0;\ H_1:\beta_1\neq 0$$

其檢定統計量為

$$t=\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(D_i-\bar{D})^2}}}\sim t(n-2)$$

當$|t|>t_{\frac{\alpha}{2}}(n-2)$時，拒絕$H_0$。

例題: 若有男、女 5 人的薪資資料如下:

男	40	10	20	25	30
女	20	15	25	5	10

(1) 建立一條迴歸方程式並估計之，以得出不同性別對薪資的影響。

(2) 試取顯著水準 0.05，並利用 t-test 檢定性別對薪資是否有顯著差異。

Sol: (1) 令$D_i=\begin{cases}1，男性\\0，女性\end{cases}$

則迴歸模式可寫為

$$Y_i=\beta_0+\beta_1 D_i+\varepsilon_i，\varepsilon_i\sim N(0,\sigma^2)$$

今將資料整理如下

Y	40	10	20	25	30	20	15	25	5	10
D	1	1	1	1	1	0	0	0	0	0

故知β_0及β_1的最小平方法估計為$\hat{\beta}_1 = 10$，$\hat{\beta}_0 = 15$。

所以性別對於薪資的迴歸方程式為$\hat{Y} = 15 + 10D$。

(2) <1> $H_0: \beta_1 = 0$

<2> $H_1: \beta_1 \neq 0$

<3> $\alpha = 0.05$

<4> $C = \{t||t| > t_{0.025}(8) = 2.306\}$

<5> $MSE = \frac{\sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} D_i Y_i}{n-2} = 93.75$

故知檢定統計量為

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(D_i - \bar{D})^2}}} = \frac{10}{\sqrt{\frac{93.75}{2.5}}} = 1.633 \notin C$$

<6> 結論: 不拒絕H_0，即表示性別不同對於薪資無顯著影響。

參考資源

[1] 程大器。2020。統計學(上)。高點書局。

[2] 程大器。2020。統計學(下)。高點書局。

[3] 沈明來。2004。試驗設計學。九州出版。

[4] Berenson, M., Levine, D., Szabat, K. A., & Krehbiel, T. C. (2012). *Basic business statistics: Concepts and applications*. Pearson higher education AU.

[5] Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.

[6] Ross, S. (2014). *A first course in probability*. Pearson.