Basic Statistics

Created by LaTeX

雨宮宮＃ 7617,July 2022

資料來源：國立交通大學管理學院工業工程與管理學系唐麗英教授

Contents

1 Introduction

1.1 Basic Statistical Concepts

1.1.1 統計學 (Statistics)

統計學是在資料分析的基礎上，研究測定、收集、整理、歸納和分析反映數據資料，使在不確定的情況下做成決策的科學方法。

1.1.2 群體 (population)

根據研究目的，研究範圍內所有個體 (object) 之資料，這些資料組成的所有資料檔 (data set) 稱為群體。

1.1.3 樣本 (sample)

群體的一部份。

1.1.4 實驗個體 (experimental unit)

有研究興趣之個體 (人、事、物等) 稱為實驗單位。

1.1.5 參數 (parameter)

由群體資料所計算之表徵值。

1.1.6 統計量 (statistic)

由樣本資料所計算之表徵值。

1.1.7 統計學的目的 (The objective of Statisics)

由樣本資料推論母體參數。

1.1.8 統計學的範圍

分為敘述統計 (Descriptive Statisics) 及推論統計 (Inferential Statisics) 兩部分。
敘述統計：包含如何蒐集數據、展示數據及找出可描述數據特徵之值的方法。
推論統計：包含如何由樣本資訊來推論群體，並估計該推論之可信度大小的方法。

1.1.9 解決統計問題的五大步驟

問題定義 → 資料收集 → 資料整理 → 資料分析 → 結論與決策

1.2 Types of Random Variables

隨機變數 (Types of Random Variables)

1.2.1 質變數、定性變數或類別變數 (Qualitative R.V.)

隨機變數的各結果不以數量表示，而依其特性之類別表之。如：性別、住居地、職業等。

1.2.2 量變數、數值變數 (Quantitative R.V.)

隨機變數各結果可以數量表示。如身高等。

1.2.2.1 離散型 (Discrete)

經由計數的方式取得資料

1.2.2.2 連續型 (Continuous)

經由量測的方式取得資料。

2 Descriptive Statistics

2.1 Graphs

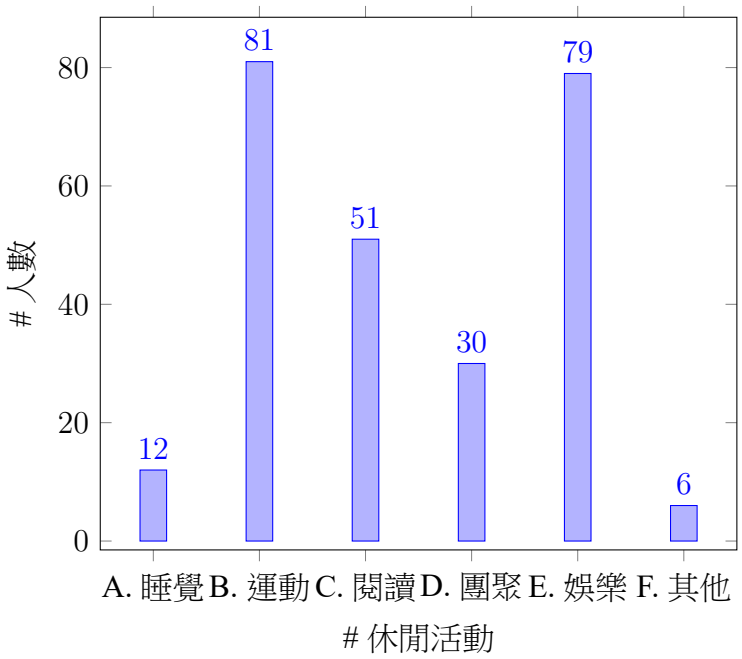
2.1.1 定性資料如何以圖來表示

利用長條圖 (Bar Graph)、柏拉圖 (Pareto Diagram)、圓餅圖 (Pie Chart)。

2.1.1.1 長條圖

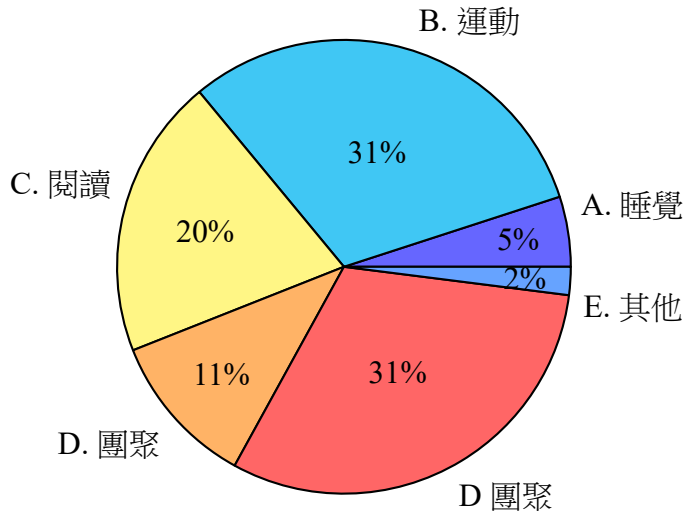
長條圖是用來比較及對照不同時期或類別間的差異。今有一假日快樂中學學生們假日從事之休閒活動類別資料，可化成長條圖：

A. 睡覺	12
B. 運動	81
C. 閱讀	51
D. 團聚	30
E. 娛樂	79
F. 其他	6



2.1.1.2 圓餅圖

圓餅圖是用來顯示一個單一總和量如何攤分於各種群體中。承上例：



2.1.1.3 柏拉圖

柏拉圖根據「關鍵的少數和次要的多數」的原理而製作，其結構為兩個縱坐標和一個橫坐標，合併長條圖及折線圖所構成。左側縱坐標表示頻率，右側縱坐標則表示累計頻率（以百分比表示），橫坐標表示影響質量的各種因素之名稱，按影響大小順序排列，直方形高度表示相應的因素的影響程度（即出現頻率為多少），上方之折線則表示累計頻率線（又稱柏拉圖曲線）。

柏拉圖一般用以在大量的事項中找到最常出現的事項。在品質管理中，多半是代表最常出現缺陷的來源、最常出的缺陷種類，或是最常見的客戶抱怨原因等。若可以改善最常出現的前幾個項目，就可以大幅降低缺陷的累積頻率，這也是柏拉圖的目的，又稱 ABC 理論。

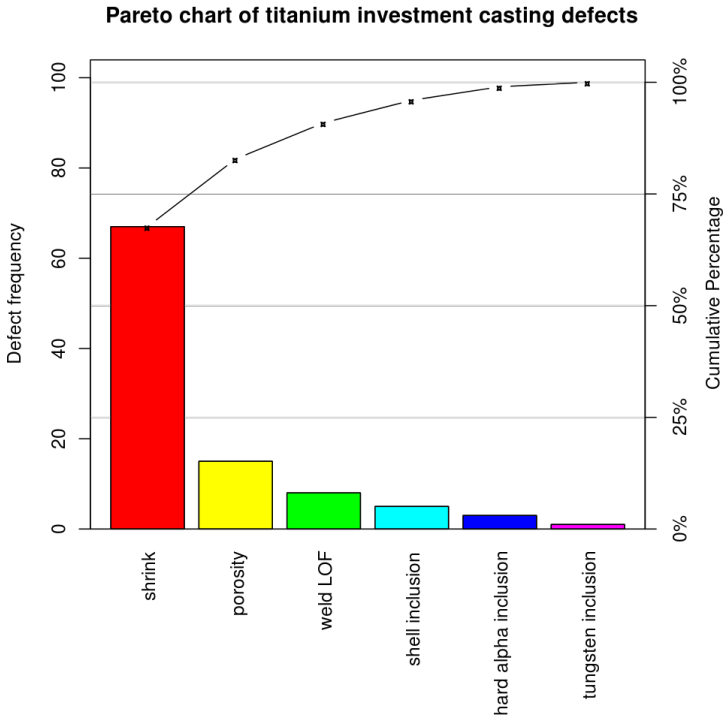


Figure 1: 柏拉圖示例

2.1.2 定量資料如何以圖來表示

可利用點圖 (Dot Diagram) 或直方圖 (Histogram)。

2.1.2.1 點圖

點圖可利用以顯現資料之分布型態。

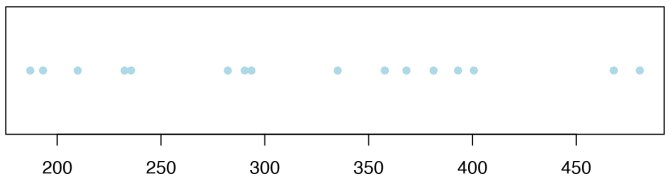


Figure 2: 點圖示例

2.1.2.2 直方圖

直方圖示次數分布的圖形表示，是以直立的條狀或矩形所構成。

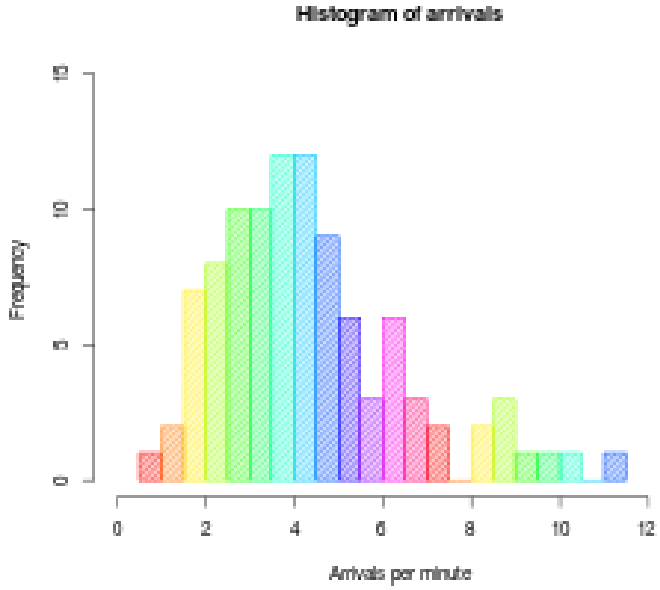


Figure 3: 直方圖示例

直方圖之分組數一般以 2 的冪次方法則為建議。

e.g. n=40，則分組數為 5 或 6 組

$$\because 2^5 < n < 2^6$$

2.2 Statistic and Parameter

統計量 (Statistic)

2.2.1 原始數據特徵值之計算

原始連續性數據分析之特徵主要可分為以下四大類：

1. 集中趨勢 (Central Tendency of Location)

2. 離中趨勢 (Dispersion)

3. 偏態 (Skewness)

4. 峰度 (Kurtosis)

2.2.1.1 集中趨勢 (Central Tendency of Location)

「集中趨勢指標」是表示一組數據中央點位置所在的一個指標。最常用的集中趨勢指標：平均數、中位數、眾數。

- 平均數 (μ or $\overline{X}$)
  可分為群體平均和樣本平均，其中
  (1) 群體平均數 (μ) 為
  $$\mu = \frac{\Sigma X_i}{N}$$
  (2) 樣本平均數 ($\overline{X}$) 為
  $$\overline{X} = \frac{\Sigma X_i}{n}$$
  其中 N 表示群體大小，n 表示樣本大小

- 中位數 (M_d)
  將一組數據由小至大排序後，最中間的那一個數值稱為中位數。

- 眾數
  在一組數據中，出現次數最多者稱之。

平均數對離群值非常敏感，而中位數或眾數對離群值較不敏感。因此，當資料中有離群值時，則使用中位數或眾數，否則，使用平均數。

2.2.1.2 離中趨勢 (Dispersion)

「離中趨勢」是表示一組是距間差異大小或數值變化的一個量數。三個主要量測離中趨勢之量數：全距 (Range)、變異數及標準差 (Variance and Standard Deviation) 及變異係數 (Coefficient of Variation)。

1. 全距 (R)：全距是用來衡量一組數距差異最簡單的方法，公式為：

$$R = \text{最大值} - \text{最小值}$$

用全距之缺點：

當一組數據中有離群值出現或資料量太大 ($n > 10$) 時，全距並非一個一個很好的衡量數距變一的量數，因其無法解釋最小與最大值之間，數據分布的情形。

2. 變異數與標準差 (Variance and Standard Deviation)

可分為群體變異數與樣本變異數、群體標準差、樣本標準差。其中

(1) 群體變異數 (σ^2) 為

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

(2) 樣本變異數 (S^2) 為

$$S^2 = \frac{1}{n-1} \times \sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

(3) 群體標準差 (σ) 為

$$\sigma = \sqrt{\sigma^2}$$

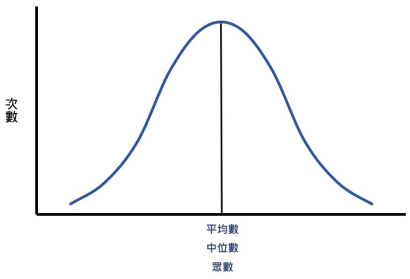
(4) 樣本標準差 (S) 為

$$S = \sqrt{S^2}$$

2.2.1.3 偏態 (Skewness)

「偏態」是用來說明一組數據分布的形態。單峰分布有三種之偏態：

1. 對稱：平均數 = 中位數



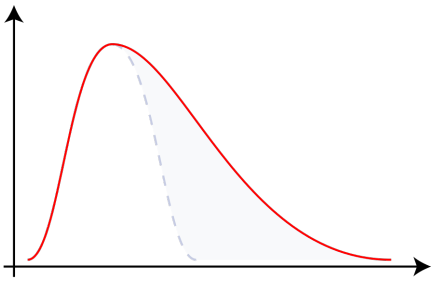
2. 右偏，正偏：平均數 > 中位數

3. 左偏，負偏：平均數 < 中位數

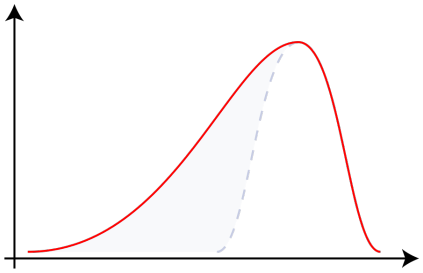
2.2.1.4 偏態係數

樣本偏態係數如下：

$$g_1 = \frac{1}{S^3} \times \left[\sum_{i=1}^{n}(X_i - \overline{X})^3 \times \frac{1}{n-1}\right] = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^3/(n-1)}{S^3}$$



Positive Skew



Negative Skew

偏態係數 = 0 表示樣本分布是對稱的。

偏態係數 >0 表示樣本分布是偏右的。

偏態係數 <0 表示樣本分布是偏左的。

2.2.1.5 峰度 (Kurtosis)

在統計學中衡量實數隨機變數機率分布的峰態。峰度高就意味著變異數增大是由低頻度的大於或小於平均值的極端差值引起的。峰度係數如下：

$$g_2 = \frac{1}{S^4} \times \sum_{i=1}^{n} (X_i - \overline{X})^4 \times \frac{1}{n-1} - 3 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^4 / (n-1)}{S^4} - 3$$

常態分佈峰度係數 = 0

2.3 Box-Whisple Plot

2.3.1 盒鬚圖是什麼？

盒鬚圖 (Box Plot) 是一種圖形表示法，此圖可同時標出資料之集中趨勢、離中趨勢、偏態、最小值、最大值等。

- Q1：第一四分位數或第 25 百分位數。
- Q2：第二四分位數或中位數 (M_d)。
- Q3：第三四分位數或第 75 百分位數。

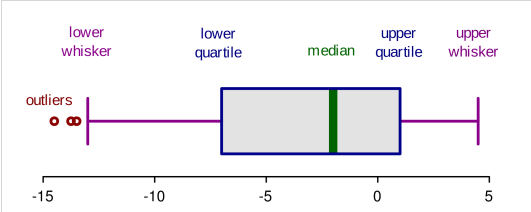


Figure 4: 盒鬚圖示例

- $Q_3 - Q_1$ = 第 75 百分位數 − 第 25 百分位數 = IR
- IR = interquartile range = 四分位距

2.3.2 盒鬚圖的主要功用

從視覺上即可有效地找出資料之主要表徵值。

2.3.3 盒鬚圖之其他功用

1. 可同時比較數組資料。

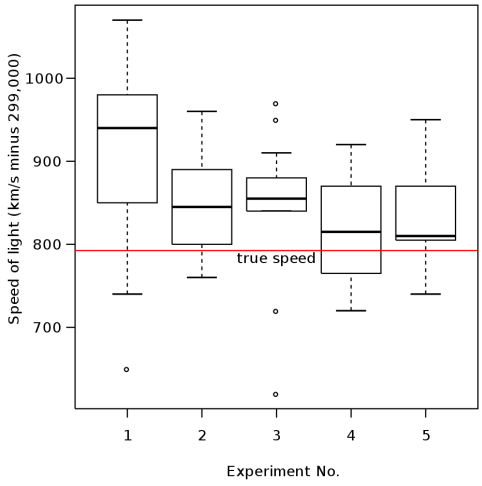


Figure 5: 同時比較數組盒鬚圖資料示例

2. 可辨認出離群值 (Outliers)。

(1) 離群值：遠大於或遠小於同一筆數據中之其他值之數據。

(2) 辨認

(a) 超過盒鬚圖之盒 $1.5(Q_3 - Q_1)$ 至 $3(Q_3 - Q_1)$ 距離內之值可當作可能之離群值（適度離群值 mild outlier）。

(b) 超過盒鬚圖之盒 $3(Q_3 - Q_1)$ 距離外之值可當作非常可能之離群值（極端離群值 extreme outlier）。

3 Probability

3.1 Introduction and The Role of Probability in Statistics

3.1.1 實驗 (Experience)

實驗是指一個可記錄一些觀察體量測值的過程 (Process)，例：

- 擲一個銅板一百次
- 擲一個骰子十次
- 量測某物一百次

3.1.2 樣本空間 (Sample Space,S)

一個實驗的所有可能出現的結果之集合稱為樣本空間。

3.1.3 事件 (Event)

實驗的結果稱為事件。

3.1.4 事件 A 的機率

$$P(A) = \frac{n(A)}{n(S)}$$

其中，n(A) 表 A 事件中的元素個數；n(B) 表樣本空間中之元素個數。

3.1.5 事件 A 與 B 之關係

1. 相依

   事件 A 的發生會受事件 B 的影響，反之亦然。

2. 獨立

   事件 A 的發生與事件 B 的發生無任何關係或彼此不會互相影響。

3. 互斥

   若事件 A 與事件 B 不可能同時發生，則兩事件互斥。

3.1.6 機率三原理

1. $0 \leq$ P(A) ≤ 1(對樣本空間任一事件 A)
2. P($\emptyset = 0$)，P(S)$= 1$
3. 若 $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, \ldots, \mathrm{A}_K$ 互為互斥事件，則

$$P(\mathrm{A}_1 \cup \mathrm{A}_2 \cup \mathrm{A}_3 \cup \ldots \mathrm{A}_K) = P(\mathrm{A}_1) + P(\mathrm{A}_2) + P(\mathrm{A}_3) + \cdots + P(\mathrm{A}_K)$$

- $P(A') = 1 - P(A)$

3.1.7 條件機率 (Conditional Probability)

The conditional probability $P(A|B)$ is the probability of event A given that B occurs.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

,where we assume that $P(B) \neq 0$

3.1.8 貝氏定理 (Bays' Theorem)

An application of the cinditional probability.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})}$$

Note:$P(B) = P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})$ — This is called "total probability".

3.1.9 獨立事件 (Independent)

If A and B are independent,then

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

3.2 Probability Distributions

機率分布隨機變數的兩種型式：

1. 離散型隨機變數 (Discrete R.V.) 定義：離散型隨機變數為應用在計數值的隨機變數。如：缺陷品的數目等
2. 連續型隨機變數 (Continuous R.V.) 連續型隨機變數是應用於連續值的隨機變數。例：重量、長度等

3.2.1 離散型隨機變數 (Discrete R.V.)

3.2.1.1 離散型隨機變數之機率分布

離散型隨機變數之機率分布，是以圖或表來表示隨機變數 X 的每一可能值之相關機率。

$$p(x) = P(X = x), (\forall x)$$

3.2.1.2 p(x) 之特性

1. $0 \leq p(x) \leq 1$
2. $\sum_{\text{all x}} p(x) = 1$

3.2.1.3 找出離散型隨機變數之機率分布之方法

1. 建立一表列出離散型隨機變數 X 的所有可能值
2. 計算出每一 X 之相對機率 $p(x)$

3.2.1.4 離散型隨機變數之期望值

設 X 為一離散型隨機變數，其機率分配為 $p(x)$，則 X 的期望值為

$$E(X) = \mu = \sum_{\text{all x}} x \cdot p(x)$$

3.2.1.5 離散型隨機變數的變異數與標準差

設 X 為一離散型隨機變數，其機率分配為 $p(x)$，則

- X 的期望值為 $E(x) = \mu$
- X 的變異數為 Var(X) $= \sigma^2 = E\left[(x-\mu)^2\right] = \sum_{\text{all x}} x^2 p(x) - \mu^2$
- X 的標準差為 St.D.(X) $= \sigma = \sqrt{\sigma^2}$

3.2.2 連續型隨機變數 (Continuous R.V.)

For a continuous R.V. X,the role of the probability function is taken by a probability density function,$f(x)$.

3.2.2.1 The Density Function for a Continous R.V.

If X is a continuous R.V., then the probability that X takes on any particular value is 0:

$$P(X = t) = 0$$

If X is a continuous R.V., then

$$P(a \leq X \leq b) = P(a \leq X < b) = (a < X \leq b) = (a < X < b)$$

Note:This is not true for a discrete R.V.

3.2.2.2 The poperties of the probability density function,$f(x)$

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

If X is a continuous R.V.,with a density function $f(x)$, then for any $a < b$ the probability that X falls in the interval (a, b) is the area under the density function between a and b:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$$

3.2.2.3 The Expected Value of a Continous Random Variable

If X is a continuous R.V. with the density function $f(x)$,the Expected Vaule of X,is

$$E(x) = \mu_x = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

$E(x)$ is a weighted average of all possible value of X with each value weighted by it associated probability.

Let X be a continuous R.V. with the density function $f(x)$, and let $g(x)$ be any function of X. Then the expected value of $g(x)$ is

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx$$

3.2.2.4 連續型隨機變數的變異數和標準差

設 X 為一連續型隨機變數，其機率分配為 $f(x)$，則

- X 的期望值為 $E(x) = \mu$
- X 的變異數為 Var(X) $= \sigma^2 = E\left[(x - \mu)^2\right]$
- X 的標準差為 St.D.(X) $= \sigma = \sqrt{\sigma^2}$

3.2.3 (Cumulative) Distribution Function (c.d.f)

3.2.3.1 The (Cumulative) Distribution Function(簡稱 c.d.f. or d.f.) 累加函數

The distribution function(c.d.f.) of a random variable X is defined to be

$$F_X(t) = P(X \leq t) for -\infty \leq t \leq \infty$$

3.2.3.2 Poperties or Requirements of $F(x)$

1. If $a < b$, then $F(a) \leq F(b)$
2. $\lim_{t \to -\infty} F_X(t) = 0$
3. $\lim_{t \to \infty} F_X(t) = 1$
4. $F_X(t)$ is a right continuous function.

We may use the c.d.f., $F_X(t)$, to evaluate the probability that X lies in a particular interval.

4 Discrete Probability Distributions

常見的離散型機律分布

- 白努力分布 (Bernoulli Probability Distribution)
- 二項分布 (Binomial Probability Distribution)
- 超幾何分布 (Hypergeometric Probability Distribution)
- 波瓦松分布 (Poisson Probability Distribution)
- 負二項分布 (Negative Binomial Probability Distribution)
- 幾何分布 (Geometric Probability Distribution)

4.1 Binomial Probability Distribution

4.1.1 二項實驗

一個實驗必須滿足以下四個條件，才能稱為二項實驗：

1. 某一實驗獨立、重複的試行 n 次。

2. 每一試行皆產生兩種結果：成功 (Success) 或失敗 (Failure)。

3. 每一試行成功的機率皆為 p，失敗的機率為 (1-p) 或 q。

4. 我們對試行 n 次中，成功 X 次之機率有興趣。

4.1.2 二項機率分布

在 n 次獨立的二項實驗試行中，出現 x 次成功的機率為

$$p(x) = C_x^n p^x q^{n-x}, x = 0, 1, 2, \ldots, n$$

其中

- n 代表全部的試行數
- x 表在 n 次試行中成功的次數
- C_x^n 表在 n 次試行中取 x 成功次數的組合數
- p 表每一試行成功的機率
- $q = 1 - p$ 表每一試行失敗的機率

4.1.3 二項隨機變數的平均數與變異數

- 平均數 $E(x) = \mu_x = np$
- 變異數 Var$(x) = \sigma_x^2 = npq$

4.2 Bernoulli Probability Distribution

白努力隨機變數 (Bernoulli R.V.)

1. 當 $X \sim (n = 1，p)$ 之二項分布，則稱 X 為白努力 (Bernoulli) 隨機變數。

2. 一個服從二項分布 (n, p) 之隨機變數 Y 是 n 個白努力隨機變數之和。

4.3 Hypergeometric Probability Distribution

4.3.1 超幾何隨機變數 (Hypergeometric R.V.)

The experiment consists of randomly drawing n elements without replacement from a set of N elements,a of which are S's (for Success) and (N-a) of which are F's (for Failure).

The hypergeometric random variable X is the number of S's in the draw of n elements.

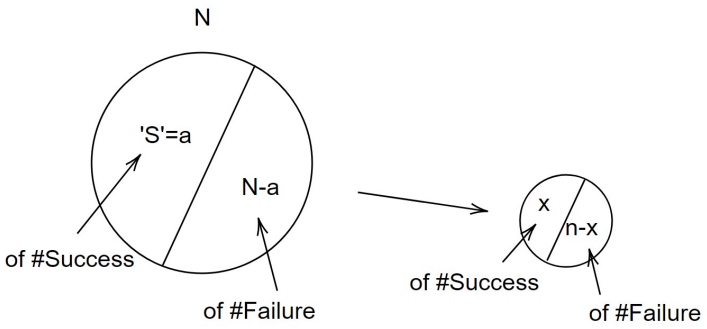


Figure 6: The experiment of Hypergeometric R.V.

4.3.2 超幾何機率分布

The probability function of hypergeometric random variable X is

$$p(x) = \frac{C_x^a \cdot C_{n-x}^{N-a}}{C_n^N} = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}, x = 0, 1, 2, \ldots, a$$

其中

- N = total number of elements (群體總數)
- a = Number of S's in the N elements (群體中成功的個數)
- n = Number of elements drawn (從群體中抽取 n 個)
- x = Number of S's drawn in the n elements (抽取 n 個中成功的個數)

4.3.3 以二項分布近似超幾何分布

In general,it can be shown that the hypergeometric distribution approaches binomial distribution with

$$p = \frac{a}{N} \text{ ,when } N \to \infty$$

A good rule of thumb is to use the binomial distribution as approximation to the hypergeometric distribution if $n \leq \frac{N}{10}$.

4.4 Poisson Probability Distribution

波瓦松分布是用來形容在某一特定時間或面積內稀有事件之機率。

- 波瓦松隨機變數的例子：

1. 幾周內保險公司收到的要保信數
2. 幾分鐘內經過剪票口的旅客數
3. 一段短時間內轉接的電話次數
4. 一段時間內地震發生次數

4.4.1 波瓦松機率分布

假設事件是隨機且彼此獨立的發生，單位時間的平均次數為 μ，而 X 表示一段時間事件發生的次數，則波瓦松機率密度函數如下：

$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} = \frac{\mu^x e^{-\mu}}{x!} \text{ ,for } x = 0, 1, 2, \dots$$

其中

- $\mu =$ 波瓦松分布事件在某一特定時間 (或面積) 內發生的平均數
- $\lambda =$ 單位時間 (或面積) 內發生的平均數
- $t =$ 特定之時間 (或面積)
- $e = 2.718$

4.5 Negative Binomial Probability Distribution

4.5.1 負二項實驗

一個實驗必須滿足下列各條件，才能稱為負二項實驗。

1. 某一實驗獨立、重複的試行 y 次。
2. 每一試行均產生兩種結果：成功 (Success) 或失敗 (Failure)。
3. 每一試行成功的機率皆為 p，失敗的機率為 (1-p) 或 q。
4. 我們對出現第 r 次成功所經歷之事行剌數 y 有興趣。

4.5.2 負二項機率分布

The probability distribution for a negative binomial random variable Y is given by

$$p(y) = \binom{y-1}{r-1} p^r q^{y-r}, (y = r, r+1, r+2, \dots)$$

其中

- $p =$ Probability of success on a single Bernoulli trail
- $q = 1 - p$
- $y =$ Number of trails until the r^{th} success is observed

4.5.3 負二項隨機變數的平均數與變異數

- 平均數 $E(x) = \mu = \dfrac{r}{p}$
- 變異數 Var$(X) = \sigma^2 = \dfrac{rq}{p^2}$

4.6 Geometric Probability Distribution

4.6.1 幾何分布是負二項分布的特例

The geometric distribution is a special case of the negative binomial distribution.It deals with the number of trails required for a single success.Thus,the geometric distribution is negative binomial distribution where the number of successes (r) is equal to 1.

4.6.2 幾何機率分布

For the special case $r = 1$,the probability distribution of Y is Known as a geometric probability distribution.(Y 表示出現第 1 次所經歷的試驗次數)

$$p(y) = pq^{y-1}, (y = 1, 2, \dots)$$

其中

- $p =$ Probability of success ona single Bernoulli trail
- $q = 1 - p$
- $y =$ Number of trails until the first success is observed

4.6.3 幾何隨機變數的平均數與變異數

- 平均數 $E(x) = \mu = \dfrac{1}{p}$
- 變異數 Var$(X) = \sigma^2 = \dfrac{q}{p^2}$

5 Continuous Probability Distributions

常見的連續型機律分布

- 常態分布 (Normal Distribution)
- 對數常態分佈 (Lognormal Distribution)
- 齊一分布 (Uniform Distribution)
- 珈瑪分布 (Gamma Distribution)
- 指數分布 (Exponential Distribution)
- 韋伯分布 (Weibull Distribution)
- 貝塔分布 (Beta Distribution)

5.1 Normal Distribution

5.1.1 何謂常態分布 (Normal Distribution)

自然界所觀察到許多連續型隨機變數常成鐘形分布，如下圖所示。此鐘形分布又稱常態分佈（或稱高斯分布 (Gaussian distribution)）。

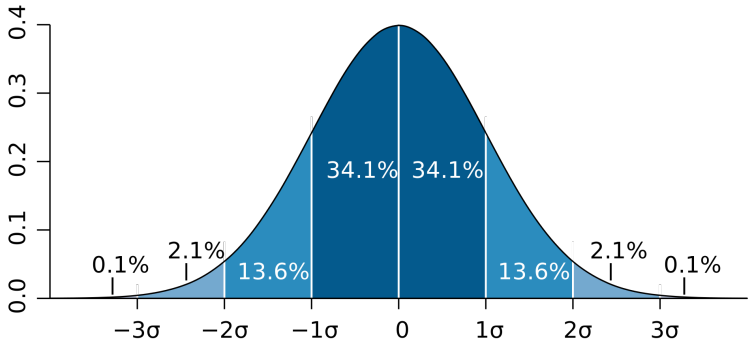


Figure 7: 常態分佈及其各區段所佔面積比例

5.1.2 常態機率分布

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

其中

- π = Mathematical constant approximated by 3.1416
- e = Mathematical constant approximated by 2.718
- μ =Population mean or the true mean
- σ^2 =Population variance

It is denoted by $N(\mu, \sigma)$

5.1.3 常態曲線

原始資料之 μ 與 σ 值不同時，其常態曲線之變化亦不同。

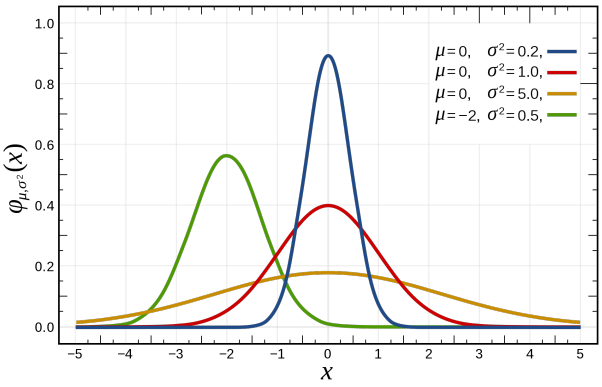


Figure 8: 每一常態分佈，可以用 $N(\mu, \sigma)$ 表之

5.1.4 $N(\mu, \sigma)$ 的特性

1. 對稱於 μ。
2. 隨機變數 x 之值可由 $-\infty$ 至 ∞。
3. 鐘形分布。
4. 曲線下面積為 1。
5. 集中趨勢的三個量數（平均數、中位數及眾數）是一致的。

5.1.5 μ 與 σ 如何影響常態曲線

- μ–位置參數 (Location parameter)
- σ–變異參數 (Dispersion parameter)

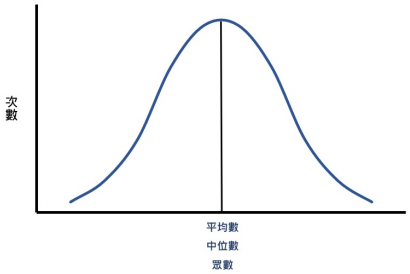


Figure 9: 平均數 = 中位數 = 眾數

5.1.6 標準常態分佈

平均數為 0、變異數為 1 之常態分佈稱為標準常態分佈，以 $N(0,1)$ 表之。

5.1.7 標準常態分佈之積分值

標準常態分佈 $N(0,1)$ 之值可透過查表取得。

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6311	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Figure 10: 常態分佈值表

5.1.8 一般常態分佈之機率

將其標準化 (Standardize)，轉換成標準常態分佈後，再求其機率。轉換公式如下：

$$Z = \frac{X - \mu}{\sigma}$$

其中 $Z \sim N(0,1)$

5.1.9 檢查數據是否常態分佈

1. 利用直方圖

   只要出現鐘形分布圖形，即判定成常態分佈。

2. 利用常態機率圖 (Normal Probability Plot)

   只要圖形成直線，即判定數據成常態分布。

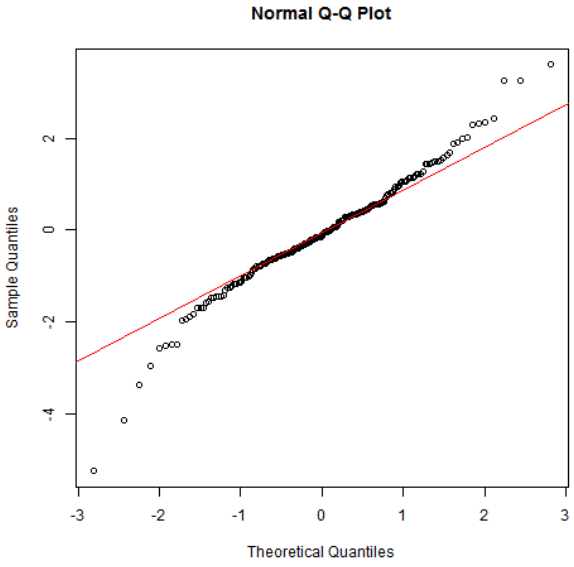


Figure 11: 常態機率圖 (Normal Probability Plot)

3. 利用統計檢定

   只要顯著度 p-value<0.05，即判定數據成常態分布

   - 卡方適配度檢定 (Chi-Sqare Goodness-of-fit Test)
   - K-S 檢定 (Kolmogorov-Smirnov Test)
   - A-D 檢定 (Anderson-Daring Test)

5.2 Lognormal Distribution

The log-normal distribution occurs when the logarithm of a random variable has a normal distribution.

X is called a log-normal random variable if and only if

$$f(x) = \frac{1}{\sqrt{2\pi}\beta} x^{-1} e^{-(\ln x - \mu)^2/2\beta^2}, (x > 0, \beta > 0)$$

$$f(x) = 0, (\text{otherwise})$$

Where $\ln x$ is the natural logarithm of X.

5.2.1 The mean and Variance of Log-normal Distribution

The mean of log-normal distribution is:

$$\mu = e^{\alpha + \frac{1}{2}\beta^2}$$

The variance of log-normal distribution is:

$$\sigma^2 = e^{2\alpha + \beta^2}(e^{\beta \exp 2} - 1)$$

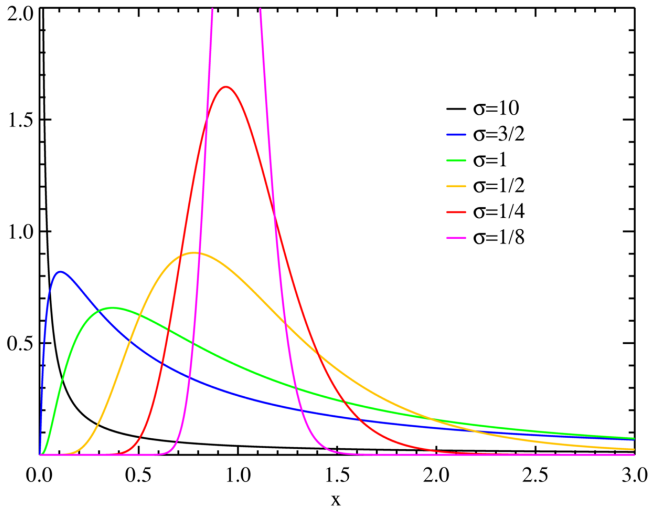


Figure 12: Plot of the Lognormal

5.3 Uniform Distribution

The continuous random variable X is called a uniform random variable if and only if X is uniformly distributed over the interval (a, b) ,i.e.,the density for X is

$$f(x) = \frac{1}{b-a}, a < X < b$$

$$f(x) = 0, (\text{otherwise})$$

The mean for the Uniform R.V. is:

$$E(x) = \mu = \frac{a+b}{2}$$

The variance for the Uniform R.V. is:

$$\text{Var}(X) = \sigma^2 = \frac{(b-a)^2}{12}$$

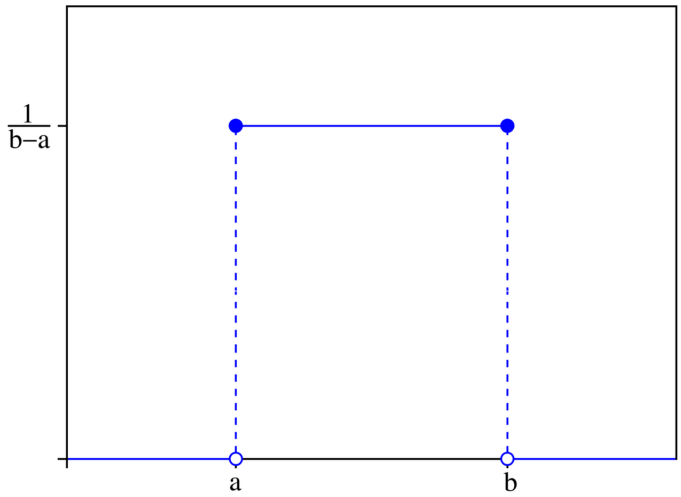


Figure 13: Uniform distribution

5.4 Gamma Distribution

Several important probability densities (such as Exponential($\alpha = 1$), Weibull) are special cases of the gamma distribution.

X is called a Gamma random variable if and only if

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{x/\beta}, x > 0, \alpha > 0, \beta > 0$$

$$f(x) = 0, \text{otherwise}$$

Where $\Gamma(\alpha)$ is the value of the gamma function

- Gamma Function

$$\Gamma(\alpha) = \int_0^x x^{\alpha-1} e^{-x}\, dx, \text{where } \alpha > 0$$

5.4.1 Properties of the Gamma Function

1. $\Gamma(\alpha) < \infty,$ if $\alpha < 0$
2. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1),$ if $\alpha > 1$
3. $\Gamma(\alpha) = (\alpha - 1)!,$ if α is a postive integer

5.4.2 The Mean and Variance for the Gamma R.V.

The mean for the Gamma R.V. is:

$$E(X) = \alpha\beta$$

The variance for the Gamma R.V. is:

$$\text{Var}(X) = \alpha\beta^2$$

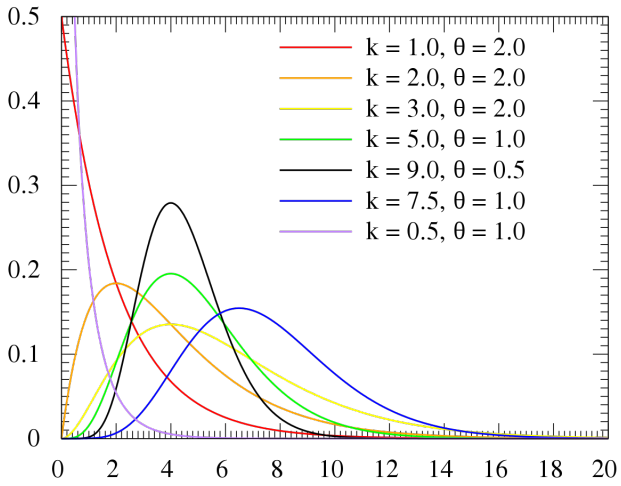


Figure 14: Probability density plots of gamma distributions($k = \alpha, \theta = \beta$)

6 Bivariate Probability Distribtions & Sampling Distribtions

雙變量機率分布 (Bivariate Probability Distribtions)

6.1 Random Vector

Suppose that S is the sample space associated with an experiment.

Let $X = X(\omega)$ and $Y = Y(\omega)$ be two functions each assigning a real number to every point ω of S. Then (X, Y) is called a two-dimensional random vector (or we say that X, Y are jointly distributed random variables).

6.2 Bivariate Probability Distribtions for Discrete R.V.

6.2.1 Joint Probability Mass Function for Discrete R.V.

Suppose that X and Y are discrete random variables defined on the same probability space and that they take on values $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$, respectively.Their jointly probability mass function $P(X, Y)$is:

$$P(X, Y) = P(X = x, Y = y)$$

The joint probability mass function must satisfy the following conditions:

1. $P(X, Y) = P(X = x, Y = y) \leq 0. \forall (x, y) \in \text{R}$
2. $\sum_{\text{all}_x} \sum_{\text{all}_y} P(X = x, Y = y) = 1$

6.2.2 Find the Joint Probability Mass Function for the Discrete X and Y

Construct a table listing each value that the R.V. X and Y can assume.Then find $p(x, y)$ for each combination of $P(X, Y)$.

6.2.2.1 Example

Toss a fair coin 3 times.Let X be the number of heads on the first toss and Y the total number of heads observed for the three tosses.Find the joint probability mass function of (X, Y).

		Y				P(X=x)
		0	1	2	3	
X	0	1/8	2/8	1/8	0	4/8
	1	0	1/8	2/8	1/8	4/8
P(Y=y)		1/8	3/8	3/8	1/8	1

Figure 15: Example table for Joint Probability Mass Function

6.2.3 Find the Marginal Probability Function feom the table

Since $P_Y(y)$ and $P_X(x)$ are located in the row and column ”margins”,these distribution are called marginal probability function.

1. To find $P_Y(y)$,sum down the appropriate column of the table.

2. To find $P_X(x)$,sum across the appropriate row of the table.

6.3 Bivariate Probability Distribtions for Continous R.V.

6.3.1 Joint Probability Mass Function for Continous R.V.

Suppose that X and Y are jointly distribution random variables.Their jointly probability density function is a piecewise continuous function of two variables, $f(X,Y)$,such that for any "reasonable" two-dimensional set A:

$$P((X,Y) \in A) = \int_A \int f(x,y)\,dy\,dx$$

The joint probability density function must satisfy the following conditions:

1. $f(x,y) \leq 0, \forall (x,y) \in \mathrm{R}$
2. $\int_{\mathrm{all}_y} \int_{\mathrm{all}_x} f(x,y)\,dx\,dy = 1$

6.3.2 The Marginal density function for X and Y

1. $f_Y(y) = \int_{\mathrm{all}_y} f(x,y)\,dy$
2. $f_X(x) = \int_{\mathrm{all}_x} f(x,y)\,dx$

In the discrete case,the marginal mass function was found by summing the joint probability mass function over the other variable;in the continuous case,it is found by integration.

6.4 The Expected Values and Covariance for Jointly Distribution R.V.

6.4.1 The Expection Value of any Two Varible X and Y

If X and Y are independent, then

$$E\left[g(x)\cdot h(y)\right] = E\left[g(x)\right]\cdot E\left[h(y)\right]$$

$$E\left[XY\right] = E\left[X\right]\cdot E\left[Y\right]$$

6.4.2 The Covariance of any Two Varible X and Y

$\mathrm{Cov}(X,Y) = E\{\left[X - E(X)\right]\left[Y - E(Y)\right]\} = E(XY) - E(X)E(Y)$

If X and Y are independent,then $\mathrm{Cov}(X,Y) = 0$

6.5 Independence and Conditional Distributions

6.5.1 Independent Random Variables

X and Y are independent random variables if and only if

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

X and Y are independent random variables if and only if

$$F_{XY}(X, Y) = F_X(X) \cdot F_Y(Y)$$

6.5.2 Conditional Distributions

1. If X and Y are discrete random variables, then

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P_{XY}(x, y)}{P_Y(y)}, P_Y(y) \neq 0$$

   (a) $P_{XY}(x, y) = P(x|y)P_Y(y)$

   (b) $P_X(x) = \sum_{\text{all}_y} P(x|y)P_Y(y)$

2. If X and Y are continuous random variables, then

$$f(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}, f(Y) \neq 0$$

   (a) $f_{XY}(x, y) = f(x|y)f_Y(y)$

   (b) $f_X(x) = \int_{-\infty}^{\infty} f(x|y)f_Y(y)\, dy$

6.6 Covariance and Correlation

Two measures of association between two random variables:

6.6.1 Covariance

The covariance of any two random variables X and Y is

$$\text{Cov}(X, Y) = E\{[X - E(X)]\,[Y - E(Y)]\} = E\,[(X - \mu_X)(Y - \mu_Y)]$$

6.6.2 Correlation

The correlation of any two random variables X and Y is

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}, \text{provide that } \sigma_x < \infty \text{ and } \sigma_Y < \infty$$

Note:

1. $-1 <= \rho(X,Y) \leq 1$
2. X and Y are said to be postively correlated if $\rho(X,Y) > 0$.
3. X and Y are said to be negatively correlated if $\rho(X,Y) < 0$.
4. X and Y are said to be uncorrelated if $\rho(X,Y) = 0$.

6.6.3 Theorems of Covariance and Correlation

1. $\text{Cov}(X,Y) = E(XY) - \mu_X \mu_Y$
2. If X and Y are independent, then $\text{Cov}(X,Y) = \rho(X,Y) = 0$ If $\text{Cov}(X,Y) = 0$, X and Y may not be independent (i.e.,X and Y may be independent).
3. Suppose that X is a random variable such that $0 \leq \sigma_X^2 \leq \infty$, and that $Y = aX + b$ for some constant a and b, where $a \neq 0$.If $a > 0$,then $\rho(X,Y) = +1$.If $a < 0$,then $\rho(X,Y) = -1$.
   That is, if Y is a linear function of X, then X and Y must be correlated and $|\rho(X,Y)| = 1$.
4. $\text{Cov}(X+Y) = \text{Cov}(X) + \text{Cov}(Y) + 2\text{Cov}(X,Y)$(provide that $\text{Var}(X), \infty$ and $\text{Var}(Y) < \infty$)
   (a) $\text{Cov}(aX, bY) = ab \cdot \text{Cov}(X,Y)$
   (b) $\text{Var}(aX + bY + c) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)$
   (c) $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(X) - \text{Cov}(X,Y)$
5. If $X_1, X_2, \ldots, X_n$ are random variables and $\text{Var}(X_i) < \infty$, for $i = 1, 2, \cdots n$, then

$$\text{Var}(\sum X_i) = \sum_{n=1}^{n} \text{Var}(X) + 2 \sum \sum_{i<j} \text{Cov}(X_i, Y_j)$$

6.7 Sampling Distribtions

Sampling Distribtions 抽樣分布

The probability distribution of a statistic that results when random sample of size n are repeatedly

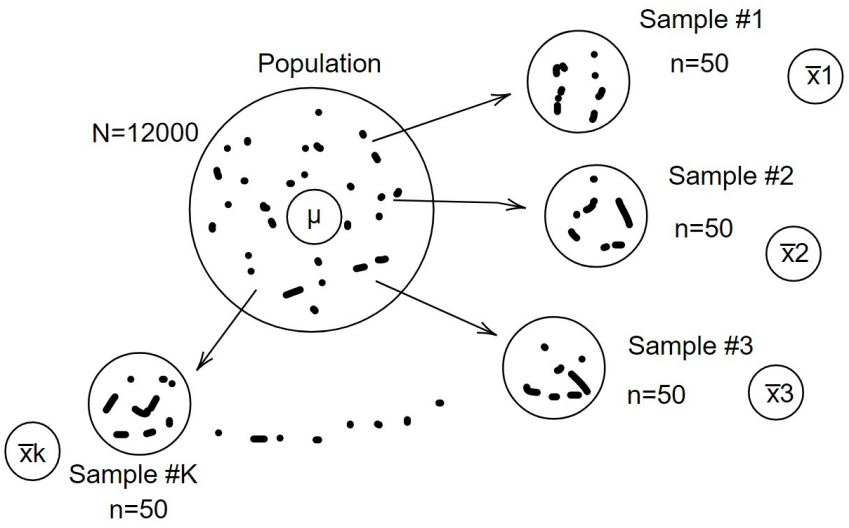


Figure 16: Sampling

from a given population is called the sampling distribution of the statistic.

If to simulated distribution of $\overline{X}$ based on independent random samples from any distributions is close to Normal distribution.Note that:

1. The values of $\overline{X}$ tend to cluster about the mean of any distributions.
2. As the sample size n increases, there is less variation in the sampling distribution of $\overline{X}$ and the shape of the sampling distribution $\overline{X}$ tends toward the shape of the Normal distribution.

6.8 The Sampling Distribution of the Sample Mean and Standard Deviation

6.8.1 The Central Limit Theorem(C.L.T.)

If random sample of n observations are drawn from a population with mean μ and standard deviation σ, then when n is large($n \geq 30$),the sampling distribution of $\overline{X}$ is approximately normally distributed with $\mu_X = \mu$, and $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}), \text{if } n \geq 30$$

The approximation will become more and more accurate as n becomes large.

If the population is normal, then the distribution of the sample mean $\overline{X}$ will always be normal,regardless of the sample size (n).

6.9 The Sampling Distribution of the Sample Mean

If X is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2 (when the σ is unknown.) ,then the sample statistic

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

has a T-distribution with degrees of freedom(d.f. 自由度)$\nu = n - 1$.

Note: T-distribution is also called "Student's T-distribution".

6.9.1 Degree of Freedom

We use the degrees of freedom as a measure of sample information.

For example, we say that the T-distribution has degrees of freedom $n - 1$.

Why?

There are n degrees of freedom or independent pieces of information in the random sample of size n from the normal distribution.

In calculating $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$,we do not know σ and need to use the sample data to estimate σ. When the data (the values in the sample) are used to compute the mean $\overline{X}$ for obtaining $S^2 = \sum_{i=1}^{n} \frac{(x_i - x)^2}{n-1}$,there is 1 less degree of freedom in the information used to estimate σ^2.

6.9.2 T-Distribution Table

When $d.f. \geq 29$ or $n \geq 30$,T-distribution is very close to Z-distribution.

6.10 The Sampling Distribution of the Sample Propotion

p：Population Propotion

$\hat{p}$：Sample Propotion $= \frac{x}{n}$ = 成功次數/總試驗次數

When the sample size n is large, the sampling distribution of $\hat{p}$ is approximately normal with mean p and standard deviation $\sqrt{\frac{pq}{n}}$.

$$\hat{p} \sim N(p, \sqrt{\frac{pq}{n}})$$

	P						
one-tail	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792

Figure 17: T-Distribution Table

6.11 The Sampling Distributions Related to the Normal Distribution

6.11.1 Chi-Square Distribution

χ^2-Distribtion(卡方分布)

If S^2 is the variance of a random sample of size n taken from Normal population having the variance σ^2, then

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

has a (Greek letter,Chi) distribution with the $d.f. = \nu = n - 1$.

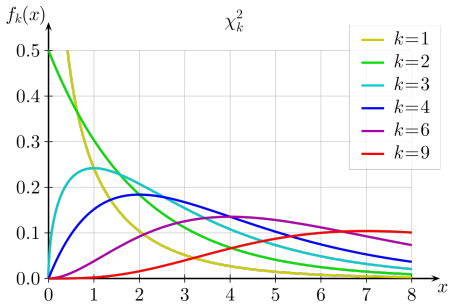


Figure 18: Chi-Square Distribution

6.11.1.1 Chi-Square Distribution Table

	Significance level (α)							
Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Figure 19: Chi-Square Distribution Table($k = \nu = d.f.$)

6.11.2 T-Distribution

Definition:Let Z be a standard Normal random variable and χ^2 be a Chi-Sqare random variable with degrees of freedom ν, then

$$T = \frac{\mathbb{Z}}{\sqrt{\chi^2/\nu}}$$

has a T distribution with ν numerator d.f.(分子自由度) and ν denominator d.f.(分母自由度).

If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2,then the sample statistic

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

has a T-distribution with degrees of freedom (d.f.) $\nu = n - 1$.

6.11.3 F-Distribution

Definition:Let χ_1^2 and χ_2^2 be two independent chi-sqare random variables with ν_1 and ν_2 degrees of freedom, respectively, then

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

has a F distribution with ν_1 numerator d.f. and ν_2 denominator d.f.

If S_1^2 and S_2^2 are the variances of a random sample of size n_1 and n_2 taken from a normal population having the same variances,then

$$F = \frac{\chi_1^2}{\chi_2^2}$$

has a F-distribution with $d.f. = (\nu_1, \nu_2) = (n_1 - 1, n_2 - 1)$.

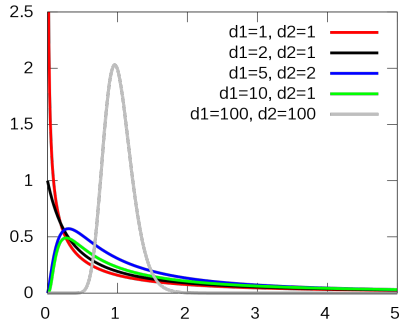


Figure 20: F-distribution($d1 = \nu_1, d2 = \nu_2$)

7 Estimation Using Confidence Intervals

Two types of estimation:

1. Point Estimation
2. Interval Estimation

7.1 Point Estimators and their Properties

A point estimator of a population parameter is a rule (or formula) that tells you how to calculate a single number baesd on sample data.The resulting number is called a point estimate of the parameter.

- μ : $\overline{X}$
- σ : s
- P : $\hat{p}$

7.1.1 Evalute the Goodness of a Point Estimator

- Unbiasedness(不偏性)
- Efficiency(有效性)
- Consistency(一致性)
- Sufficientcy(充分性)

7.1.2 Unbiased Estimator

An estimator of a population parameter is said to be unbiased if the mean of its sampling distribution is equal to the parameter. Otherwise the estimator is said to be biased.
That is, an estimator is unbiased if

$$E(\text{Sample estimator}) = \text{Population parameter}$$

For example: $\overline{X}$ and S and $\hat{p}$ are unbiased estimators.

7.1.3 An Efficient Estimator

A point estimator $\hat{\theta}_1$ is said to be a more efficient unbiased estimate of θ than $\hat{\theta}_2$ if

1. $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimates of θ.
2. the variance of the sampling distribution of $\hat{\theta}$ is less than that of $\hat{\theta}_2$.

7.1.4 The Maximum Error of Estimate for μ

When we use $\overline{X}$ to estimate μ, the maximum error of estimate can be expressec as follows:

$$E = Max|\overline{X} - \mu| = Z_{\frac{1-\alpha}{2}}(\frac{\sigma}{\sqrt{n}}) \text{ or } Z_{\frac{1-\alpha}{2}}(\frac{S}{\sqrt{n}})(\text{ for large sample })$$

$$= T_{\frac{\alpha}{2}}(\frac{S}{\sqrt{n}})(\text{ for small sample })$$

7.2 Interval Estimation

Def:

An interval estimate of a population parameter us a rule that tells you how to calculate two numbers based on sample data.

7.2.1 Confidence Coefficient 信賴係數

Def:

The probability that a confidence interval will enclose the estimated parameter is called the confidence coefficient.

7.2.1.1 Interval Estimation of μ

1. (a) σ 已知,large sample($n \geq 30$)

$$\overline{X} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

(b) σ 已知,$n < 30$

假設母體之分布呈常態

2. (a) σ 未知，不論 n 之大小，假設母體呈常態

$$\overline{X} \pm T \cdot \frac{S}{\sqrt{n}}$$

(b) σ 未知，但 $n \leq 30$，用 $S \simeq \sigma$

$$\overline{X} \pm Z \cdot \frac{\sigma}{\sqrt{n}}$$

The CLT guarantees that $\overline{X}$ is approximately normally distribution regardless of the distribution of the sampled population.

The value of z selected for constucting such a confidence interval is called the critical value(臨界值) of the distribution.

7.2.1.2 Estimation of Population Propotion

Recall:

$$\hat{p} \sim N(\mu_{\hat{p}} = P, \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}})$$

Estimation of Population Propotion:

$$\hat{p} \pm Z \cdot \sqrt{\frac{pq}{n}}$$

The sample size n is sufficiently large so that the approximation is valid.As a rule of thumb,the codition of a "sufficiently large" sample size will be satisfied if $n\hat{p} \geq 4$ and $n\hat{q} \geq 4$

7.2.1.3 Confidence Interval for σ^2

Recall:

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

Estimation of σ^2:

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

Assumption:

母體呈常態

(1-α) is called the confidence coefficient. 信賴係數

(1-α)100% is called the confidence level. 信賴水準

7.3 Choosing the Sample Size

7.3.1 Choosing the Sample Size,n

1. Choosing n for estimating μ correct to within E units with confidence $(1-\alpha)$

$$n = \left[\frac{z_{1-\alpha/2}\sigma}{E}\right]^2$$

Note:

The population standard deviation σ will usually have to be approximated by S.

2. Choosing n for estimating P correct to within E units with confidence $(1-\alpha)$

$$n = p \cdot q \cdot \left[\frac{z_{1-\alpha/2}}{E}\right] \text{ ,where } q = 1-p$$

Note:

This technique requires previous estimates of p and q. If none are available, use $p = q = 0.5$ for a conservative of n.

7.4 Estimation of the Difference Between Two Means:Independent Samples

7.4.1 Inerval Estimation of $(\mu_1 - \mu_2)$

1. Large-Sample $(1-\alpha)100\%(n_1 \geq 30, n_2 \geq 30)$

$$(\overline{y}_1 - \overline{y}_2) \pm Z \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note:

We have used the sample variance S_1^2 and S_2^2 as approximations to the corresponding population parameters.

Assumption:

The two random samples are selected in an independent manner from the target population. That is, the choice of elements in one sample does not affect, and is not affected by the choice of elements in the other sample.

The sample size n_1 and n_2 are sufficiently large for the central limit theorem to apply.

2. $(1-\alpha)100\%$($n_1 \leq 30$ or $n_2 \leq 30$)

   (a) assume $\sigma_1^2 = \sigma_2^2$

   $$(\overline{X}_1 - \overline{X}_2) \pm t \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

   其中 $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \dfrac{\sum_{i=1}^{n}(X_1 i - \overline{X}_1)^2 + \sum_{i=1}^{n}(X_2 i - \overline{X}_2)^2}{n_1+n_2-2}$

   (b) assume $\sigma_1^2 \neq \sigma_2^2$

   $$(\overline{y}_1 - \overline{y}_2) \pm t_\nu \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

   where

   $$\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Assumption:

Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.

The random samples are selected in independent manner from the two populations.

7.4.1.1 Let (LCL,UCL) represent a $(1-\alpha)100\%$ confidence interval for $(\theta_1 - \theta_2)$

1. If LCL > 0 and UCL > 0, conclude $\theta_1 > \theta_2$.
2. If LCL < 0 and UCL < 0, conclude $\theta_1 < \theta_2$.
3. If LCL < 0 and UCL > 0,(i.e.,the interval includes 0), conclude no evidence of a difference between θ_1 and θ_2.

7.4.1.2 Estimation of the Difference Between Two Population Means:Matched Pairs

算差額然後當成一組資料，結束。

7.4.2 Inerval Estimation of $(p_1 - p_2)$

Large-Sample$(1-\alpha)100\%$ Confidence Interval for $(p_1 - p_2)$

$$(\hat{p}_1 - \hat{p}_2) \approx Z \pm \sqrt{\frac{\hat{p}_1\hat{p}_1}{n_1} + \frac{\hat{p}_2\hat{p}_2}{n_2}}$$

where $\hat{p}_1$ and $\hat{p}_2$ are the sample proortions with the characteristic of interest.

Note:

We have followed the usual procedure of substituing the sample value $\hat{p}_1$ and $\hat{q}_1$, $\hat{p}_2$ and $\hat{q}_2$ for the corresponding population values required for $\sigma(\hat{p}_1 - \hat{p}_2)$.

Assumption:

The samples are sufficiently large that the approximation is valid.As a general rule of thumb,we will require that ,$n_1\hat{q}_1 \geq 4$,$n_2\hat{p}_2 \geq 4$,$n_2\hat{q}_2 \geq 4$.

7.4.3 Inerval Estimation of $\frac{\sigma_1^2}{\sigma_2^2}$

A $(1-\alpha)100\%$ Confidence Inerval for the Two Population Variance,$\frac{\sigma_1^2}{\sigma_2^2}$

$$\frac{S_1^2}{S_2^2}(\frac{1}{F_{1-\alpha/2(\nu_1,\nu_2)}}) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}F_{1-\alpha/2(\nu_2,\nu_1)}$$

where $F_{\alpha/2(\nu_1,\nu_2)}$ is the value of F that locates an area $\alpha/2$ in the upper tail of the F-distribution with $\nu_1 = (n_1 - 1)$ numerator and $\nu_2 = (n_2 - 1)$ denominator degrees of freedom,and $F_{\alpha/2(\nu_2,\nu_1)}$ is the value of F that locates an area $\alpha/2$ in the upper tail of the F-distribution with $\nu_2 = (n_2 - 1)$ numerator and $\nu_1 = (n_1 - 1)$ denominator degrees of freedom.

Assumption:

Both of the populations from which the samples are selected have relative frequency distributions that are approximately normal.

The random samples are selected in an independent manner from the two populations.

7.5 Choosing the Sample Size for Estimating

7.5.1 Choosing the Sample Size for Estimating $(\mu_1 - \mu_2)$

Choosing the sample size for estimating the difference $(\mu_1 - \mu_2)$ between two population means correct to within H units with probability $(1-\alpha)$

$$n_1 = n_2 = (\frac{Z_{\alpha/2}}{H})^2(\sigma_1^2 + \sigma_2^2)$$

where n_1 and n_2 are the numbers of observations sampled from each of the two populations,and σ_1^2 and σ_2^2 are the variances of two populations.

Note:

The population variances σ_1^2 and σ_2^2 will usually have to be approximated.

7.5.2 Choosing the Sample Size for Estimating $(p_1 - p_2)$

Choosing the sample size for estimating the difference $(p_1 - p_2)$ between two population propotion correct to within H units with probability $(1 - \alpha)$

$$n_1 = n_2 = (\frac{Z_{\alpha/2}}{H})^2(p_1q_1 + p_2q_2)$$

where p_1 and p_2 are the proportions for populations 1 and 2,respectively,and n_1 and n_2 are the numbers of observations sampled from each of the two populations.

8 Hypothesis Tests of a Single Population

8.1 Concepts of Hypothesis Testing

8.1.1 統計檢定之目的

假說檢定主要是利用樣本資訊來測試一個或多個群體的參數值。

8.1.2 Steps for Testing a Hypothesis

1. 設立虛無假說 (null hypothesis,H_0) 和對立假說 (alternative hypothesis,H_1 or H_a)。
2. 指定顯著水準 (level of significance α)。
3. 決定適當之檢定統計量 (test statistic)。
4. 決定棄卻域 (rejection region)。
5. 下結論—推翻虛無假說 (reject H_0) 或不推翻虛無假說 (fail to reject H_0) 並將此結論按題意引申。

8.1.3 設立假說

假說是關於一個或多個群體參數值的一段敘述。

假說又分為以下兩種：

- 對立假說 (Alternative Hypothesis,H_1)
  研究者想要蒐集證據支持之假說。

- 虛無假說 (Null Hypothesis,H_0)
  研究者所欲蒐集證據推翻之假說。虛無假說為對立假說之相反。

注意：

1. 永遠先設立 H_1，再以其相反之敘述設立 H_0。
2. ”=” 只能放在 H_0 中。

依據 H_1 中之符號，檢定又可分為以下兩類：

- 單尾檢定（或單邊檢定）One-sided Test
  如果對立假設中有”<” 或”>’ 出現，則此種統計檢定為單邊檢定。
  Note:
  如果”<” 出現在”H_1”，則稱為左尾檢定。
  如果”>” 出現在”H_1”，則稱為右尾檢定。

- 雙尾檢定（或雙邊檢定）Two-sided Test
  如果對立假設中出有 $\neq$ 出現，則此種統計檢定稱為雙邊檢定。

8.1.4 指定顯著水準

作統計檢定時，有兩種不可避免的誤差：

1. 型一誤差 (Type I Error)
   當 H_0 是對的，檢定結果卻判其為錯的而推翻 H_0。
   $\alpha = P$（犯型一誤差）

2. 型二誤差 (Type II Error)
   當 H_0 是錯的，檢定結果卻判其為對的而不推翻 H_0。
   $\beta = P$（犯型二誤差）

- 檢定力 (Power of a test)
  當 H_0 是錯的，檢定結果亦判其為錯的而推翻 H_0(The probability of rejecting H_0 when it is actually false.)

$$\text{Power} = 1 - \beta$$

- 決策表

決策	True State of Nature 真實情況	
	H_0 是對的	H_0 是錯的
推翻 H_0	型一誤差	正確
不推翻 H_0	正確	型二誤差

Figure 21: 決策表

注意：

α 與 β 是衡量一個統計檢定好壞的指標，一個決策者必須平衡此兩種形式的誤差。當我們的檢定結果是推翻 H_0 時，我們只可能犯型一誤差；當我們的檢定結果是不推翻 H_0 時，我們只可能犯型二誤差。我們不可能同時犯型一誤差，又犯型二誤差。

α 與 β 有反向之關係 (當 α 增加時，β 減少；當 α 減少時，β 增加)。欲使 α 與 β 同時變小的唯一方法是：增加 n。

8.1.5 決定適當之檢定統計量

關於一個群體參數的檢定統計量

1. 當 σ 已知，檢定 μ：

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

其中 μ_0 為 H_0 下的 μ 值。

2. 當 σ 未知，檢定 μ：

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}\text{，(自由度)} = n - 1$$

其中 μ_0 為 H_0 下的 μ 值。

3. 檢定 P：

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

其中 p_0 為 H_0 下的 P 值。

4. 檢定 σ^2：

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

其中 σ_0^2 為 H_0 之下的 σ^2 值。

8.1.6 決定棄卻域

- 棄卻域 Rejection Region(R.R.)

  棄卻域是推翻虛無假設的檢定統計量計算值之值域。

- 臨界值 Critical Value

  在一個檢定的棄卻域中，其棄卻域的邊界值稱為臨界值。

  ※ 棄卻域的範圍是依據檢定為單尾或雙尾及所預先規定的顯著水準值 α 而定。

決定棄卻域之法則：

- 在 H_1 中有符號"<"(或">") 則為單邊檢定，其棄卻域為標準話檢定統計量的抽樣分配的最低（或最高）尾部。臨界值的左邊（或右邊）區域為 α。
- 在 H_1 中有符號"$\neq$" 則為雙邊檢定，其棄卻域為尾端兩部分。標準化檢定統計量的抽樣分配的每尾端區位為 $\alpha/2$。

8.2 P-Value and Power of a Test

8.2.1 表示統計結果的另一個方法：**p-value**

一般在進行統計假設檢定時，必須在獲得資料和計算檢定統計量之前，先選定顯著水準 α。檢定的棄卻域則是依據所選的 α 值而定，因此不論檢定統計量之值多大或多小，關於 H_0 的決策法則如下：

- 若檢定統計量之值落於棄卻域內，則拒絕 H_0。（即此檢定結果顯著）
- 若檢定統計量之值落於棄卻域外，則不拒絕 H_0。（即此檢定結果不顯著）

此指定之顯著水準 (α) 值是我們下結論時的一個可靠的評估依據，然而，這種檢定方式有一個缺點，就是不易評估檢定結果的顯著程度；也就是說，當檢定統計量落域棄卻域時，我們無法評估資料與虛無假社不符合的程度有多嚴重。

Def：p-value

- P-value 值亦稱為觀察的或樣本資料估算出來的顯著水準。P 值是評估樣本資料與虛無假設之間不符合程度的一個指標。

$$\text{p-value} = p(\text{reject } H_0 | H_0 \text{ is true}) = p(\text{棄卻域} | H_0 \text{ is true})$$

- 利用 p-value 來決定是否拒絕虛無假設 H_0

  1. 選擇所能容忍的最大顯著水準 α。
  2. 假如檢定統計量的 P-value 小於 α，則推翻 H_0。否則，不推翻 H_0
     - 若 p-value > 0.10，則稱統計結果不顯著。

- 若 $0.05 <$ p-value ≤ 0.10，則稱統計結果趨於顯著。
- 若 $0.01 <$ p-value ≤ 0.05，則稱檢定結果顯著（$*$）。
- 若 $0.001 <$ p-value ≤ 0.01，則稱檢定結果為非常顯著（$**$）。
- 若 p-value ≤ 0.001，則稱檢定結果有很高的顯著性（$***$）。

9 Hypothesis Tests of Two Populations

9.1 Testing the Difference Between Two Population Means: Matched Pairs

兩配對群體平均數之檢定順序：

1. $H_0 : \mu_d = \delta_0 (H_0 : \mu_d \geq \delta_0$ or $H_0 : \mu_d \leq \delta_0)$
   $H_1 : \mu_d \neq \delta_0 (H_1 : \mu_d > \delta_0$ or $H_1 : \mu_d < \delta_0)$，其中 $\mu_d = (\mu_1 - \mu_2)$
2. α
3. 檢定統計量：$t = \dfrac{\overline{D} - \delta_0}{S_D/\sqrt{n}}$，其中 $\overline{D} = \displaystyle\sum_{i=1}^{n} \frac{D}{n}$，$S_D = \sqrt{\dfrac{\sum D^2 - (\sum D)^2/n}{n-1}}$，自由度 $n-1$；n 為成對數。
4. 棄卻域：查 t-表
5. 結論

假設：

1. 成對樣本隨機抽自成對群體。
2. 成對群體服從常態分配。

9.2 Testing the Difference Between Two Population Means: Independent Samples

9.2.1 σ_1^2 與 σ_2^2 已知

檢定步驟：

1. $H_0 : \mu_1 - \mu_2 = \delta_0 (H_0 : \mu_1 - \mu_2 \geq \delta_0$ or $H_0 : \mu_1 - \mu_2 \leq \delta_0)$
   $H_1 : \mu_1 - \mu_2 \neq \delta_0 (H_1 : \mu_1 - \mu_2 > \delta_0$ or $H_1 : \mu_1 - \mu_2 < \delta_0)$
2. 定 α 值

3. 計算檢定統計量：$Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

4. 棄卻域：查 Z-表

5. 下結論

Note:

此檢定之統計假設為：當 $n_1 < 30$ 或 $n_2 < 30$ 時虛假設兩樣本是隨機抽自兩個獨立之常態母體。

9.2.2 σ_1^2 與 σ_2^2 未知但可假設 $\sigma_1^2 = \sigma_1^2$

檢定步驟：

1. H_0：$\mu_1 - \mu_2 = \delta_0$(H_0：$\mu_1 - \mu_2 \geq \delta_0$ or H_0：$\mu_1 - \mu_2 \leq \delta_0$)
   H_1：$\mu_1 - \mu_2 \neq \delta_0$($H_1$：$\mu_1 - \mu_2 > \delta_0$ or H_1：$\mu_1 - \mu_2 < \delta_0$)

2. 定 α 值

3. 計算檢定統計量：$t = \dfrac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$，其中

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n}(X_1 i - \overline{X}_1)^2 + \sum_{i=1}^{n}(X_2 i - \overline{X}_2)^2}{n_1 + n_2 - 2}$$

4. 棄卻域：查 t-表，自由度 $\nu = n_1 + n_2 - 2$

5. 下結論

Note:

此檢定之統計假設為：當 $n_1 < 30$ 或 $n_2 < 30$ 時虛假設兩樣本是隨機抽自兩個獨立之常態母體。

9.2.3 σ_1^2 與 σ_2^2 未知但假設 $\sigma_1^2 \neq \sigma_1^2$

檢定步驟：

1. H_0：$\mu_1 - \mu_2 = \delta_0$(H_0：$\mu_1 - \mu_2 \geq \delta_0$ or H_0：$\mu_1 - \mu_2 \leq \delta_0$)
   H_1：$\mu_1 - \mu_2 \neq \delta_0$($H_1$：$\mu_1 - \mu_2 > \delta_0$ or H_1：$\mu_1 - \mu_2 < \delta_0$)

2. 定 α 值

3. 計算檢定統計量：$t = \dfrac{(\overline{X}_1 - \overline{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

4. 棄卻域：查 t-表，自由度 $\nu = \dfrac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$

5. 下結論

Note:

此檢定之統計假設為：當 $n_1 < 30$ 或 $n_2 < 30$ 時虛假設兩樣本是隨機抽自兩個獨立之常態母體。

9.3 Testing the Difference Between Two Population Proportions

檢定步驟：

1. $H_0 : P_1 - P_2 = 0 (H_0 : P_1 - P_2 \geq 0$ or $H_0 : P_1 - P_2 \leq 0)$
   $H_0 : P_1 - P_2 \neq 0 (H_0 : P_1 - P_2 > 0$ or $H_0 : ¶_1 - P_2 < 0)$

2. 定 α 值

3. 計算檢定統計量：$t = \dfrac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\frac{\hat{P}_0(1-\hat{P}_0)}{n_1} + \frac{\hat{P}_0(1-\hat{P}_0)}{n_2}}}$

4. 棄卻域：查 Z-表

5. 下結論

9.4 Testing the Ratio of Two Population Variances

當兩個樣本隨機抽自兩個獨立常態母體且 $\sigma_1^2 = \sigma_2^2$，則 σ_1^2/σ_2^2 呈 F 分配，其自由度 $d.f. = (\nu_1, \nu_2) = (n_1 - 1, n_2 - 1)$。檢定步驟：

1. $H_0 : \sigma_1^2 = \sigma_2^2 (H_0 : \sigma_1^2 \geq \sigma_2^2$ or $H_0 : \sigma_1^2 \leq \sigma_2^2)$
   $H_1 : \sigma_1^2 \neq \sigma_2^2 (H_1 : \sigma_1^2 > \sigma_2^2$ or $H_1 : \sigma_1^2 < \sigma_2^2)$

2. 定 α 值

3. 計算檢定統計量：$F = S_1^2/S_2^2$

4. 棄卻域：查 F 表，自由度 $= (\nu_1, \nu_2) = (n_1 - 1, n_2 - 1)$

5. 下結論

Note:

此檢定之統計假設為：當 $n_1 < 30$ 或 $n_2 < 30$ 時虛假設兩樣本是隨機抽自兩個獨立之常態母體。

10 Simple Regression Analysis and Correlation Analysis

10.1 Simple Regression Analysis

回歸的分類依自變數的數量不同可分為簡單迴歸和複迴歸。在進行迴歸分析之前，需先瞭解變數間呈何種關係，才能適配一個適當之數學方程式或迴歸模式。而建立迴歸分析有以下三個步驟：

1. 建立迴歸方程式（迴歸模式）
2. 評估迴歸方程式與資料之適配度
3. 應用迴歸方程式：
    - 解釋 X 與 Y 之關係
    - 估計 y（給定一個 $X = X_0$ 值）

10.1.1 散佈圖 Scatter Diagram

我們可利用散佈圖來決定兩變數的關係。將 X 變數標示於一維座標圖的橫坐標，Y 變數標示於縱座標，並將各 (X, Y) 各對應點繪在 $X - Y$ 二維座標上，以觀察點之變化，此圖形即稱為散佈圖。

10.1.2 兩變數間之關係

1. 正相關 (positive relationship)
   假如 X 增加，則 Y 增加；或 X 減少，而 Y 減少，稱為 X 與 Y 有正相關。
2. 負相關 (negative relationship)
   假如 X 增加，則 Y 減少；或 X 減少，而 Y 增加，稱為 X 與 Y 有負相關。
3. 無相關 (no relationship)
   在散佈圖中之點大部分與水平軸平行，看不出任何特殊圖形。

10.1.3 收集迴歸分析的資料時之注意事項

1. 迴歸分析的資料必須先能代表所研究的系統或問題。
2. 在作迴歸分析之前須先確定資料不含離群值。

10.1.4 迴歸分析之功用

1. 描述資料
2. 估計參數
3. 預測與估計應變數之值
4. 控制應變數之值

10.1.5 決定簡單直線迴歸模式

由散佈圖大致可看出自變數與應變數間的關係。自變數與應變數間最簡單的關係即為直線關係。真實之簡單直線迴歸模式如下

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

其中 Y_i 表第 i 個觀測值；
X_i 表對應第 i 個觀測值之自變數之值；
β_0 為截距；
β_1 為斜率，表自變數每增加一單位時，應變數 Y 之改變量；
ϵ_i 表隨機誤差。

10.1.5.1 決定或觀測值最適配之直線迴歸模式

利用最小平方法 (Least Squares Method)。

- 何謂最小平方法
  - 簡單線性迴歸分析主要是找到一條最適配 data 的直線。
  - 所謂最適配是指所找出之直線方程式所得預測 Y 值 ($\hat{Y}$) 與真實 Y 值的差異最小。

樣本直線迴歸式：

$$\hat{Y}_i = b_0 + b_i X_i$$

其中 $\hat{Y}_i$ 為在特定之 X_i 值下 Y 估計值的平均。

b_0 與 b_1 分別為 β_0 與 β_1 之不偏估計值。利用最小平方法可找出 b_0 與 b_1 之公式，此公式可使 $\sum(Y_i - \hat{Y}_i)^2$ 為最小。

$$b_1 = \frac{SS_{XY}}{SS_X}, b_0 = \overline{Y} - b_1\overline{X}$$

其中 $SS_{XY} = \sum(X_i - \overline{Y})(Y_i - \overline{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n}$；$SS_X = \sum(X_i - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$

- 殘差 (Residual)e_i

$$e_i = Y_i - \hat{Y}_i(\sum e_i = 0)$$

10.1.5.2 定義迴歸分析中的幾個變異量

在迴歸模式中，為了要知道自變數預測應變數的能力，必須要知道下列三個變異數的衡量值：

1. 總變異量 $= SST = \sum(Y_i - \overline{Y})^2 = \sum Y_i^2 - (\sum Y_i)^2/n = SS_y$ (Total variation)
2. 迴歸可解釋的變異量 $= SSR = \sum(\hat{Y}_i - \overline{Y})^2 = bSS_{xy} = (SS_{xy})^2/SS_X$ (Explained Variation)
3. 其他因素解釋的變異量 $= SSE = \sum(Y_i - \hat{Y}_i)^2 = SST - SSR$ (Unexplained Variation)

10.1.5.3 判斷迴歸方程式是否顯著

1. 由圖形判定（限簡單迴歸模式）

   資料點與迴歸方程式越接近越有用。

2. 判定係數 r^2(Coefficient of Determination)

   判定係數是用來衡量自變數 (X) 所能解釋應變數 (Y) 之變異量占 Y 總變異量的百分比。

   $$r^2 = \frac{\text{迴歸式可解釋的變異量}}{\text{總變異量}} = \frac{SSR}{SST}(0 \le r^2 \le 1), \text{r 值越近 1 越好}$$

   - 相關係數 r(Correlation Coefficient) 是用來衡量兩個隨機變數 X 與 Y 間值限關係的方向與強弱。r 可由 $\pm\sqrt{r^2}$ 求得，"+" 或"-" 符號則與斜率 b_1 同。

     (a) $-1 \le r \le 1$

     (b) $r = 0$ 並不一定表示 Y 與 X 間沒有關係，僅表示 Y 與 X 間無線性關係。

3. 統計檢定—t test 及 F test

   在線性模式 $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ 中，為了檢定 $H_0 : \beta_1 = 0$，首先須假設誤差項 ϵ_i 彼此獨立且服從平均數為 0 和變異數為 σ^2 的常態分配（亦即 $\epsilon_i \sim N(0, \sigma^2)$）。

- $H_0 : \beta_1 = 0$ 的兩種檢定

(a) ANOVA F test:

可利用 ANOVA 的 F 檢定來檢定 X 與 Y 之間是否有顯著之線性關係。

ANOVA 之檢定程序：

決策	True State of Nature 真實情況	
	H_0 是對的	H_0 是錯的
推翻 H_0	型一誤差	正確
不推翻 H_0	正確	型二誤差

Figure 22: 決策表

假設：$\epsilon_i \sim NID(0, \sigma^2)$

i. $H_0 : \beta_1 = 0$（X 與 Y 之間沒有線性關係；或對預測 Y 而言，迴歸模式無法提供有用之資訊）

$H_1 : \beta_1 \neq 0$（X 與 Y 之間有線性關係，即斜率不為 0；或在預測 Y 上，迴歸模式有用）

ii. 設定 α 值

iii. 檢定值：$F = \dfrac{MSR}{MSE}$

iv. 棄卻域：查 F-表，自由度 $= (1, n-2)$ 或計算 p-值

v. 下結論

(b) t test

如前章所述

10.1.6 作迴歸分析時應注意事項

1. 利用迴歸模式估計 y 時，所給定之 x 值必須在樣本之 x 範圍內，y 之估計值才會準確。
2. 迴歸式並不表示自變數與應變數間一定有因果關係。其因果關係可能經由第三變數或其他理論依據而成立。

10.2 Statistical Inference:Hypothesis Tests and Confidence Intervals

假設：$\epsilon_i \sim NID(0, \sigma^2)$，則 b_0 與 b_1 為 β_0 與 β_1 之不偏估計式，即 $E(b_0) = \beta_0, E(b_1) = \beta_1$.

$$V(b_0) = \sigma^2 (\frac{\sum_{i=1}^{n} x_i^2}{n}) \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \right]$$

$$S_{b_0} = \sqrt{MSE(\sum_{i=1}^{n} \frac{x_i^2}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2)}\ ;$$

$$V(b_1) = \sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$S_{b_1} = \sqrt{\frac{MSE}{\sum(x_i - \overline{x})^2}} = \sqrt{\frac{MSE}{SS_x}}$$

b_1 之抽樣分佈：$b_1 \sim N(\beta_1, \sigma_{b_1})$

b_0 之抽樣分佈：$b_0 \sim N(\beta_0, \sigma_{b_0})$

$\hat{\sigma}^2 = MSE$；$\hat{\sigma}^2 = S_e = (MSE)^{\frac{1}{2}}$

$S_e = \sqrt{MSE}$ 稱作 Y 之估計標準誤。

1. 在特定之 $X = x_0$ 值下，$E(Y|x_0) = \mu_Y|x_0$ 之 $(1-\alpha)100\%$ 估計區間 (confidence interval for predictions)

   $$\hat{y}_0 \pm t_{n-2,\alpha/2}\sqrt{\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]}S_e = \hat{y}_0 \pm t_{n-2,\alpha/2}\sqrt{\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_x}\right]}\sqrt{MSE}$$

   其中 $\hat{y}_0 = b_0 + b_1 x_0$。

2. 在特定之 $X = x_0$ 值下，預測第 $n+1$ 筆新觀察值 y_{n+1} 之 $(1-\alpha)100\%$ 預測區間 (prediction interval)

   $$\hat{y}_{n+1} \pm t_{n-2,\alpha/2}\sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]}S_e$$

   $$= \hat{y}_{n+1} \pm t_{n-2,\alpha/2}\sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_x}\right]}\sqrt{MSE}$$

   其中 $\hat{y}_0 = b_0 + b_1 x_0$。

10.3 Correlation Analysis

群體相關係數：$\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$

樣本相關係數：$r = \dfrac{\sum_{i=1}^{n}(X - \bar{x})(Y - \bar{y})}{\sqrt{\sum(X - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \dfrac{S_{xy}}{S_x S_y}$

關於 r 之假說檢定步驟：

假設：$\epsilon_i \sim NID(0, \sigma^2)$

1. $H_0 : \rho = 0$

   $H_1 : \rho \neq 0$

2. 設定 α 值

3. 檢定值：$t = \dfrac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$

4. 棄卻域：查 t-表，自由度 $= n - 2$ 或計算 p-值

5. 下結論

11 Multiple Regression Analysis

11.1 The Multiple Regression Model

在實際應用迴歸分析時，要考慮的自變數通常不只一個。複迴歸分析的主要目的是探究兩個或兩個以上的自變數 $(x_1, \ldots, x_k)$ 對應變數 (y) 的影響，進而建構一個 y 與 $(x_1, \ldots, x_k)$ 之關係式或迴歸方程式。

- 複迴歸分析之標準假設 (Standard Multiple Regression Assumptions)

  1. 應變數 Y 為隨機變數，自變數 $X_j, j = 1, ...k,$ 為預先選定之變數。
  2. x_{ji} 是隨機變數 X_j 之觀測值，為固定數值。
  3. $\hat{Y} = E(Y)$ 為 X_j 之線性函數。
  4. $\epsilon_i \sim NID(0, \sigma^2)$。
  5. $\text{Cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$，即任兩個誤差項不相關。
  6. $\text{Cov}(X_j, \epsilon_i) = 0, i = 1, ...n, j = 1, ...k$，即 X_j 與 ϵ_i 彼此不相關。
  7. 自變數彼此間無完全之線性關係。
  8. 樣本數 n 必須大於 $k + 1$，否則無足夠資訊求解迴歸係數 (共有 $k + 1$ 個迴歸係數)

- 有兩個自變數之群體複迴歸模式:$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$

- 有兩個自變數之樣本複迴歸模式

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2$$

  其中 b_0 為截距，$b_0 = \hat{\beta}_0$
  b_1 為當 x_2 為固定常數時，x_1 為對 y 之斜率，$b_1 = \hat{\beta}_1$
  b_2 為當 x_1 為固定常數時，x_2 為對 y 之斜率，$b_2 = \hat{\beta}_2$

- 利用最小平方法可得 b_0，b_1 及 b_2 公式

$$b_1 = \frac{SS_y(r_{x_1 y} - r_{x_1 x_2} r_{x_2 y})}{SS_{x_1}(1 - r_{x_1 x_2}^2)}$$

$$b_2 = \frac{SS_y(r_{x_2 y} - r_{x_1 x_2} r_{x_1 y})}{SS_{x_2}(1 - r_{x_1 x_2}^2)}$$

$$b_0 = \overline{y} - b_1 \overline{x}_1 - b_2 \overline{x}_2$$

11.2 Explanatory Power of a Multiple Regression Model

複迴歸模型的解釋能力：R^2 與 Adjusted R^2

- 判定係數 (Coefficient of Determination):R^2
- Adjusted R^2

  當樣本數 n 太小或自變數個數太多時，會高估真實之 R^2。此表示在複迴歸模型中加入許多 Y 無太大關聯之自變數時，R^2 會膨脹，因此不再能代表迴歸模式真正的解釋能力。在此情況，統計學家建議使用 Adjusted R^2(調整的判定係數) 來取代 R^2。

  Adjusted R^2 之公式如下：

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

  其中 n 為樣本數，k 為自變數的個數。

  ※Adjusted R^2 值越近 1 表示迴歸式越顯著。

- $R = \sqrt{R^2}$：表複迴歸模式中 r 與 $\hat{Y}$ 的簡單相關係數，R 越接近 1 表迴歸式越顯著

11.2.1 複迴歸模式誤差項變異之估計

ϵ_i 為一隨機變數，Var(ϵ) =Var(Y_i) = σ^2。

σ^2 之不偏估計式為

$$S_E = SSE(n-k-1) = MSE$$

S_e 稱作 Y 之估計標準誤。

11.3 Confidence Intervals and Hypothesis Tests for Individual Regression Coefficients

假設有兩個自變數之群體複迴歸模式如下：

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

樣本複迴歸模式如下：

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2$$

b_0，b_1，b_2 分別為 β_0，β_1 及 β_2 之不偏估計式，

$$S_{b_1}^2 = \frac{S_e^2}{(n-1)SS_{x_1}^2(1-r_{x_1x_2}^2)}$$

$$S_{b_2}^2 = \frac{S_e^2}{(n-1)SS_{x_2}^2(1-r_{x_1x_2}^2)}$$

11.3.1 群體迴歸係數 β_j 之 $(1-\alpha)100\%$ 之信賴區間 **(Confidence Interval for β_j)**

群體迴歸係數 β_j 之 $(1-\alpha)100\%$ 之信賴區間為

$$b_j \pm t_{(n-k-1,\alpha/2)} \cdot S_{b_j}$$

假設：$\epsilon_i \sim NID(0, \sigma^2)$

11.3.2 群體迴歸係數 β_j 之假說檢定

Hypothesis Testing for β_j:

1. $H_0 : \beta_j = \beta_j^*$
   $H_1 : \beta_j \neq \beta_j^*$
2. 設定 α 值
3. 檢定值：$t = \dfrac{b_j - \beta_j^*}{S_{b_j}}$
4. 棄卻域：查 t-表，自由度 $= n - k - 1$
5. 下結論
   (假設:$\epsilon_i \sim NID(0, \sigma^2)$)

12 Analysis of Variance,ANOVA

12.1 One-Way ANOVA:Randomized of Design for Single Factor

- 主要目的：
  檢定多個群體平均數 ($\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$) 間是否有差異之統計方法。
- 單因子變異數分析之執行步驟：
  1. 設立虛無假說 (H_0) 及對立假說 (H_1)
     $H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ v.s. H_1 : 至少有一處不同。
  2. 指定顯著水準 (level of significance,α)
  3. 決定適當之檢定樣本量 (test statisic):ANOVA 之 F 檢定

     $$F = \frac{MS_G}{MS_W}$$

     其中 MS_G 表平均組間變異，MS_W 表平均組內變異。

群體					
	1	2	3	i…	K
平均數	μ_1	μ_2	μ_3	…	μ_K
變異數	σ_1^2	σ_2^2	σ_3^2	σ_i^2	σ_K^2

樣本組					
	1	2	…	i	K
樣本數 n_i	n_1	n_2	…	n_3	n_k
樣本和 $y_{i.}$	$y_{1.}$	$y_{2.}$	…	$y_{i.}$	$y_{K.}$
樣本平均數	$\bar{y}_{1.}$	$\bar{y}_{2.}$	…	$\bar{y}_{i.}$	$\bar{y}_{K.}$

4. 決定棄卻域：查 F-表，自由度 $d.f. = (K-1, n-K)$，其中 K 表組數，$n = n_1 + n_2 + \cdots + n_K$ 表各組樣本數之總和。
5. 下結論：推翻虛無假說或不推翻虛無假說，並將此結論案題意引申。

- 單因子變異數分析：

1. 計算 ANOVA 之 F 檢定統計量
2. 如 F 檢定之結果顯著，則進行後續檢定（多重比較法）。

12.1.1 單因子變異數分析之計算

1. 符號：
2. 公式：

   總變異 = 組間變異 + 組內變異

   (a) 總變異 $= SS_T = \sum_i \sum_j (y_{ij} - \overline{y}_{..})^2 = \sum_i \sum_j y_{ij}^2 - CM$ 其中 $CM = y_{..}^2/n, (n = n_1 + n_2 + ... + n_K)$，$y_{..} = \sum_i \sum_j y_{ij}, \overline{y}_{..} = \dfrac{\sum_i \sum_j y_{ij}}{n}$

   (b) 組間變異 $= SSG = \sum_i n_i(\overline{y}_{i.} - \overline{y}_{..})^2 = \sum_i \dfrac{y_{i.}^2}{n_i} - CM$

   (c) 組內變異 $= SSW = \sum_i \sum_j (\overline{y}_{ij} - \overline{y}_{i.})^2 = \sum_i \sum_j y_{ij}^2 - \dfrac{y_{i.}^2}{n_i} = SS_T - SS_G$

3. 單因子變異數分析表 (One-Way ANOVA table)：

變異來源	平方和 SS	自由度	均方 MS	F
組間 (between Groups)	SS_G	K-1	$MS_G = SS_G/(K-1)$	$F = MS_G/MS_W$
組內 (withins Groups)	SS_W	n-K	$MS_W = SS_W/(n-K)$	
總變異	SS_T	n-1		

12.1.2 多重比較法 (Multiple-Comparsion Procedure)

如變異數分析結果顯著，則可利用多重比較法來找出哪一組或哪幾組和其他組間有顯著差異。常用的多重比較法為最小顯著差異法 (Minimum Significance Difference(MSD) Method) 或 Tukey 法。

- 最小顯著差異法 (MSD Method)(或 Tukey 法) 之執行步驟

  1. 計算所有可能之任兩組之平均數差異：$|\overline{y}_{i.} - \overline{y}_{j.}|, i \neq j$。
  2. 計算 MSD Method 之臨界值 MSD

  $$MSD = Q_{\alpha,(K,n-K)}\sqrt{\frac{MSW}{n^*}}$$

  其中 $Q_{\alpha,(K,n-K)}$ 值可查表後得，$MSE = S_p^2$ 由變異數分析表得之，n 為每組樣本數，如每組的樣本數不同，則 n 以 n^* 代之（各組樣本數之調和平均數）。

  3. 比較所有 $|\overline{y}_{i.} - \overline{y}_{j.}|$ 與臨界值 MSD 值，若 $|\overline{y}_{i.} - \overline{y}_{j.}| > MSD$，則表第 i 組與第 j 組的平均數間有顯著差異。