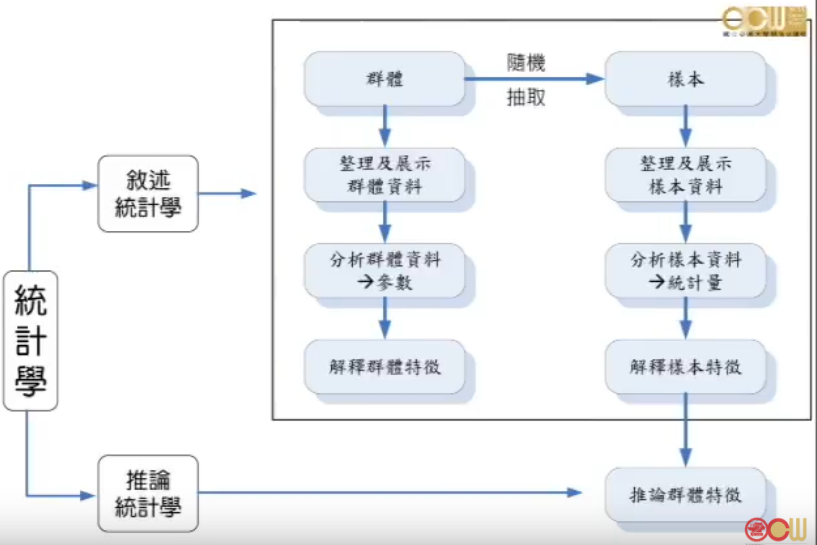
# CH1 Introduction

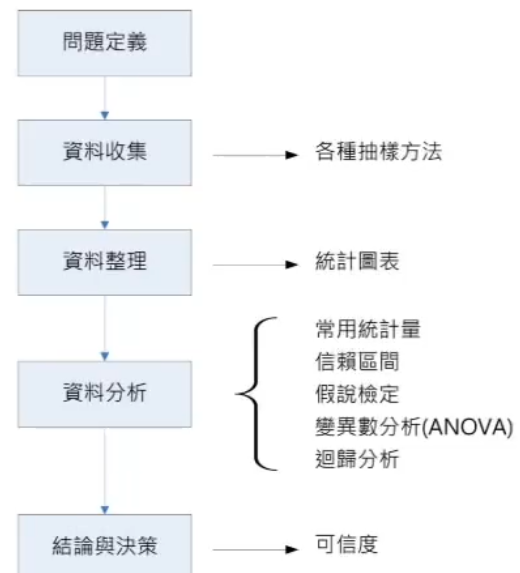
* Statistics: 蒐集、整理、展示、分析、解釋資料，並由樣本推論群體，使在不確定的情況下做成決策的科學方法
* Population: 根據研究目的蒐集個體之資料(data set)
* Sample: Part of population
* Parameter: Feature of population.
* Statistic: Feature of sample.

The objective of Statistics: 由sample 推論parameter.

* Descriptive Statistics: 敘述統計，如何蒐集、展示、及找出可描述data feature的方法
* Inferential Statistics: 推論統計，由sample 推論population，並估計該推論之可信度大小

5

* 5 steps to solve statistics problem.



* Random Variable: population 中你感興趣的Feature.

Types of Random Variables:

* + Qualitative RV:

類別變數，結果以類別表示。

ex: 文具有那些? 1.寫字用 2.塗改用 3.量測用

* + Quantitative RV:

數值變數，結果可以數量表示。

1. Discrete: 離散型

Data obtained through a counting process.

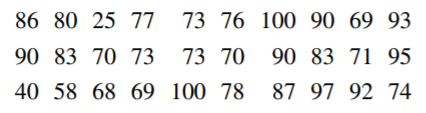
1. Continuous: 連續型

Data obtained through a measuring process.

# CH2 Descriptive Statistics

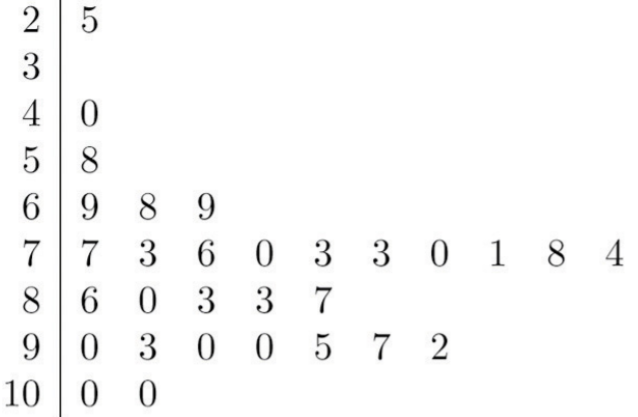
## 2.1 Three popular data displays

To learn to interpret the meaning of three graphical representations of sets of data: stem and leaf diagrams, frequency histograms, and relative frequency histograms.

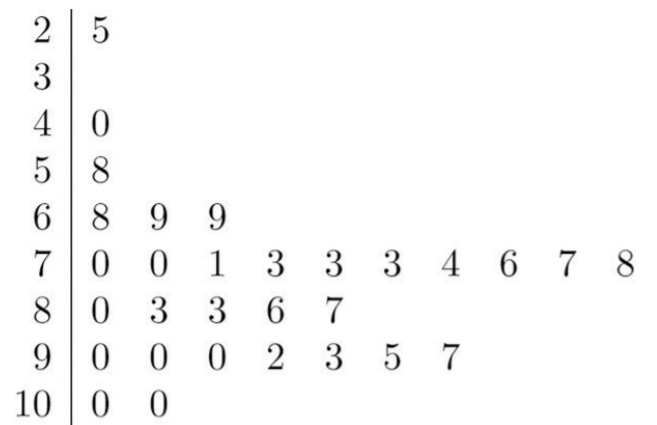


original data set

### Stem and Leaf Diagrams



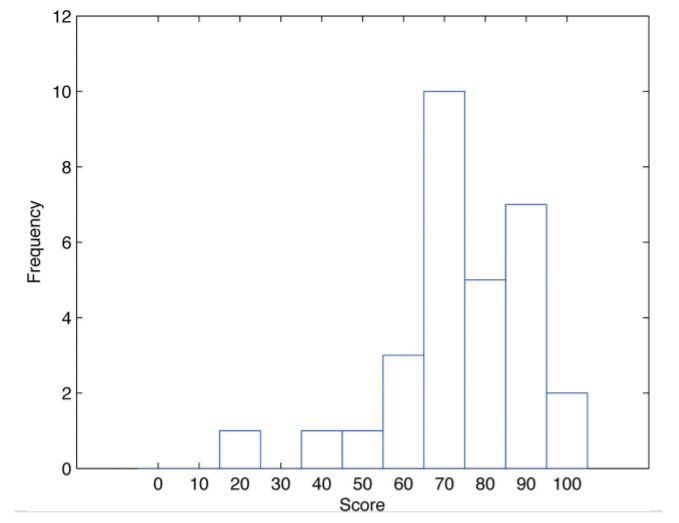
Stem and Leaf Diagram



Ordered Stem and Leaf Diagram

* The general purpose of a stem and leaf diagram is to provide a quick display of how the data are distributed across the range of their values
* All of the original data can be recovered from the stem and leaf diagram.

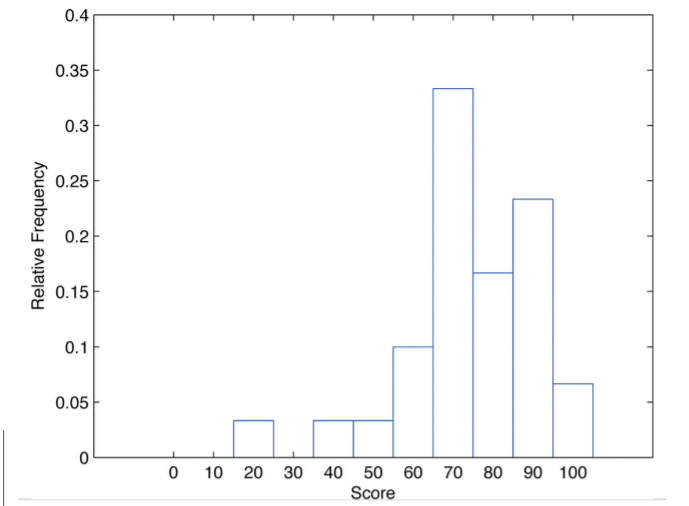
### Frequency Histograms



Frequency Histogram

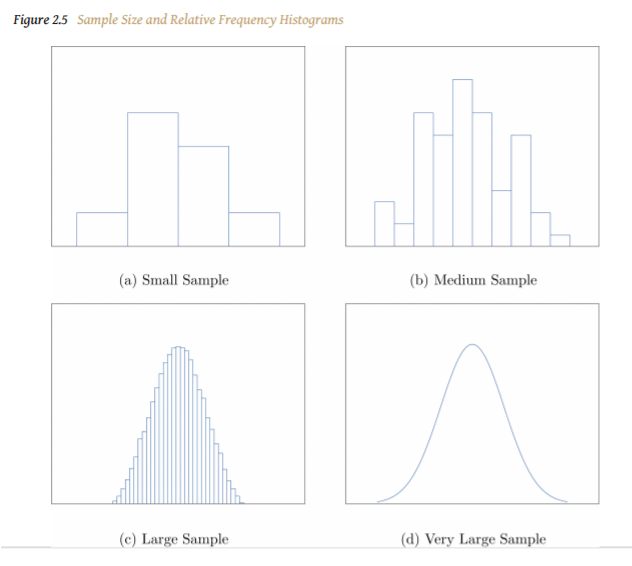
* group the scores on the standard ten-point scale, and count the number of scores in each group.
* gives a sense of data distribution across the range of values that appear.

### Relative Frequency Histograms



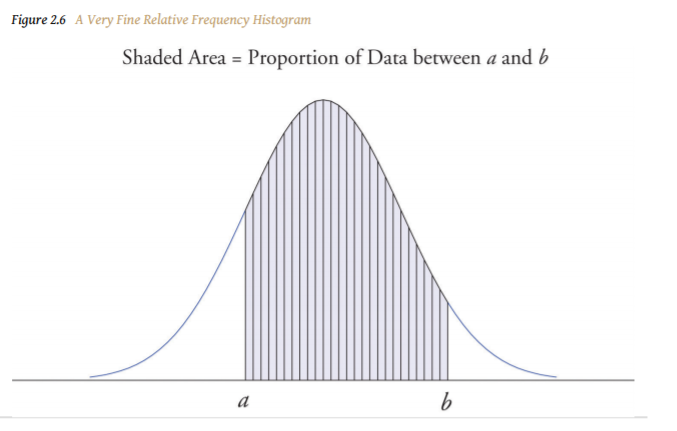
Relative Frequency Histogram

* Classes are selected, the relative frequency of each class is noted, the classes are arranged and indicated in order on the horizontal axis, and for each class a vertical bar, whose length is the relative frequency of the class



Sample Size and Relative Frequency Histograms

* The relative frequency histogram is important because the labeling on the vertical axis reflects what is important visually: the relative sizes of the bars.



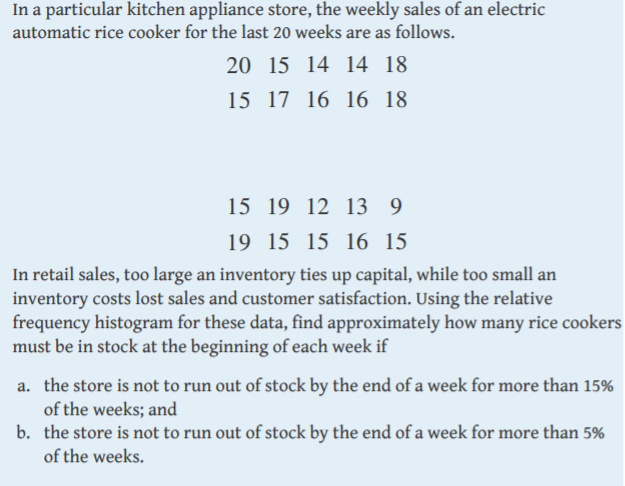
A Very Fine Relative Frequency Histogram

* for any two numbers a and b, the proportion of the data that lies between the two numbers a and b is the area under the curve that is above the interval (a,b) in the horizontal axis.

### KEY TAKEAWAYS

* Graphical representations of large data sets provide a quick overview of the nature of the data.
* A population or a very large data set may be represented by a smooth curve. This curve is a very fine relative frequency histogram in which the exceedingly narrow vertical bars have been omitted.
* When a curve derived from a relative frequency histogram is used to describe a data set, the proportion of data with values between two numbers a and b is the area under the curve between a and b, as illustrated in Figure 2.6 "A Very Fine Relative Frequency Histogram"

Exercises:

Ans: a. 19, b. 20

## 2.2 Measures of Central Location

* To learn the concept of the “center” of a data set.
* To learn the meaning of each of three measures of the center of a data set—the mean, the median, and the mode—and how to compute each one.

### The Mean

Definition:

The **sample mean** of a set of n sample data is the number x ⎯⎯ defined by the

formula: (可能有抽樣誤差)

The **population mean** of a set of N population data is the number μ defined by the formula:

Mean: the balance point.

### The Median

Definition:

The **sample median** x~

* For a set has **odd** number of sample data:

The middle sample data when the data are arranged in numerical order.

* For a set has **even** number of sample data:

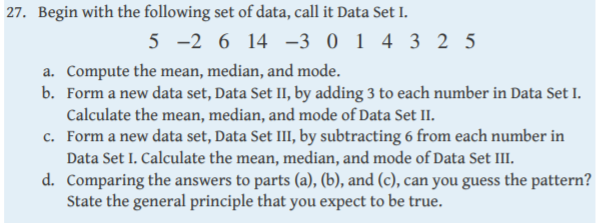
The mean of the tow middle sample data when the data are arranged in numerical order.

### The Mode

Definition:

The **mode** of a set of data values is the value that appears most often.

Exercises:



Ans: a. 3.182, 3, 5;

b. 6.182, 6, 8

c. -2.818, -3, -1

## 2.3 Measures of Variability

* To learn the concept of the variability of a data set.
* To learn how to compute three measures of the variability of a data set: the range, the variance, and the standard deviation.

### Range

Definition:

The range of a data set is the number R defined by the formula

where xmax is the largest measurement in the data set and xmin is the smallest.

太大會喪失資訊

有離群值會喪失資訊

### The Variance and the Standard Deviation

Definition:

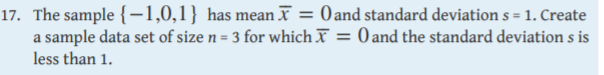
Sample variance:

Sample standard deviation:

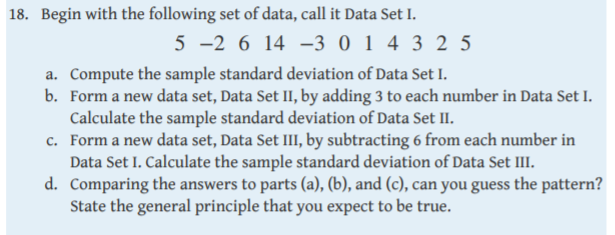
Population variance:

Population standard deviation:

Exercises:



Ans: see Exercises.ipynb



Ans: a = 4.622

b = 4.622

c = 4.622

d = For stdev 'S' = square root of S^2, S^2 = sum(xi - x\_mean)^2/n-1,

varies by (xi - x\_mean) & n, since n doesn't change,

sum(xi - x\_mean) dominate the variety of S^2,

but add & subtract won't change the result of sum(xi - x\_mean),

(cause they vary linearly by add & subtract, they get bigger and

smaller together)

so the variances S^2 doesn't change, S won't change.

## 2.4 Relative Position of Data

* To learn the concept of the relative position of an element of a data set.
* To learn the meaning of each of two measures, the percentile rank and the z-score, of the relative position of a measurement and how to compute each one.
* To learn the meaning of the three quartiles associated to a data set and how to compute them.
* To learn the meaning of the five-number summary of a data set, how to construct the box plot associated to it, and how to interpret the box plot.

### Percentiles

Definition:

Given an observed value x in a data set, x is the Pth percentile of the data if the percentage of the data that are less than or equal to x is P. The number P is the percentile rank of x.

### Quartiles

Definition:

1. The second quartile Q2 of the data set is its median.
2. Define two subsets:
   1. the lower set: all observations that are strictly less than Q2 ;
   2. the upper set: all observations that are strictly greater than Q2 .
3. The first quartile Q1 of the data set is the median of the lower set.
4. The third quartile Q3 of the data set is the median of the upper set.

### InterQuartile Range

Definition:

The interQuartile Range (IQR) is the quantity

IQR = Q3 − Q1

### Z-score

Definition:

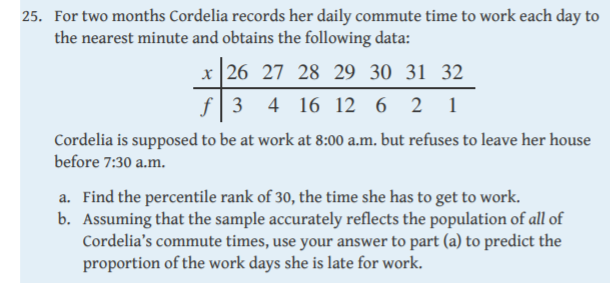
The z-score of an observation x is the number z given by the computational formula.

z = (x – x̅) / S or = (x - μ) / σ

x = x̅ + s\*z or = μ + σ\*z

The z-score indicates how many standard deviations an individual observation x is from the center of the data set, its mean. If z is negative then x is below average. If z is 0 then x is equal to the average. If z is positive then x is above average.

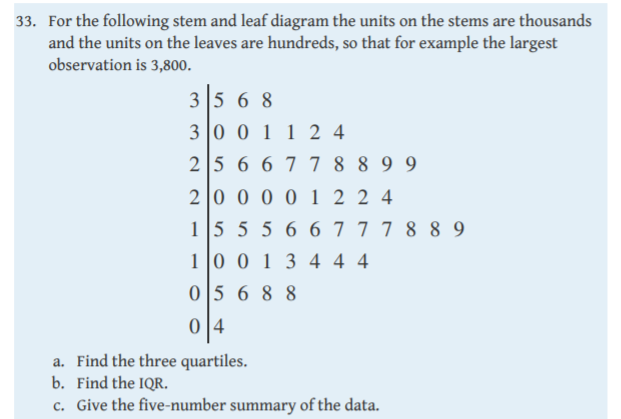
Exercises:



Ans:

a = 0.93

b = 0.07



Ans:

a = 12.5th = 1450.0

50th = 2000

75th = 2800.0

b = IQR = 1350.0

c = five-number summary

12.5th = 1450.0

50th = 2000

75th = 2800.0

xmin = 400

xmax = 3800

## 2.5 The Empirical Rule and Chebyshev’s Theorem

* To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the Empirical Rule and Chebyshev’s Theorem.
* To use the Empirical Rule and Chebyshev’s Theorem to draw conclusions about a data set.

### The Empirical Rule

Definition:

If a data set has an approximately bell-shaped relative frequency histogram, then:

1. Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints x̅ ± s for samples and with endpoints μ ± σ for populations.
2. approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints x̅ ± 2s for samples and with endpoints μ ± 2σ for populations.
3. approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints x̅ ± 3sfor samples and with endpoints μ ± 3σ for populations.

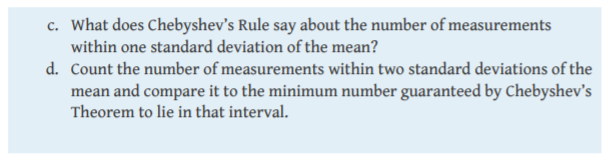
### Chebyshev’s Theorem

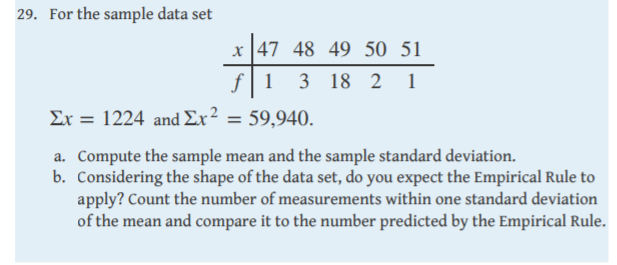
Definition:

For any numerical data set,

1. at least 3/4 of the data lie within two standard deviations of the mean, that is, in the interval with endpoints x̅ ± 2sfor samples and with endpoints μ ± 2σ for populations.
2. at least 8/9 of the data lie within three standard deviations of the mean, that is, in the interval with endpoints x̅ ± 3sfor samples and with endpoints μ ± 3σ for populations.
3. at least 1 – (1 / k^2) of the data lie within k standard deviations of the mean, that is, in the interval with endpoints x̅ ± ks for samples and with endpoints μ ± kσ for populations, where k is any positive whole number that is greater than 1.

Exercises:





Ans:

a = ds\_maen = 48.960000, ds\_stdev = 0.734847

b = yes, cause the shape is similar to bell-shape

in range mean +- 1d : [48.225, 49.695], there are 18 samples.

For Empirical Rule, in mean +- 1d have 68% of samples = 17.0

c = Chebyshev’s Theorem dose not define 1 stdev

d = There are 23 samples

CT guaranteed 18.75, hence at least 19

# CH3 Basic Concepts of Probability

## 3.1 Sample Spaces, Events, and Their Probabilities

* To learn the concept of the sample space associated with a random experiment.
* To learn the concept of an event associated with a random experiment.
* To learn the concept of the probability of an event.

### Sample Spaces and Events

Definition:

* Sample Spaces:

一個實驗的所有可能出現的結果之集合

ex: 丟一個骰子: S = {1, 2, 3, 4, 5, 6}

丟一個銅板兩次: S = {++, +-, -+, --}

* Event:

實驗的結果(測量的結果)，Sample Space 的 subspace

ex: 丟骰子出現奇數，event A = {1, 3, 5}

### Probability of outcome:

0 = 不可能, 1 = 確定

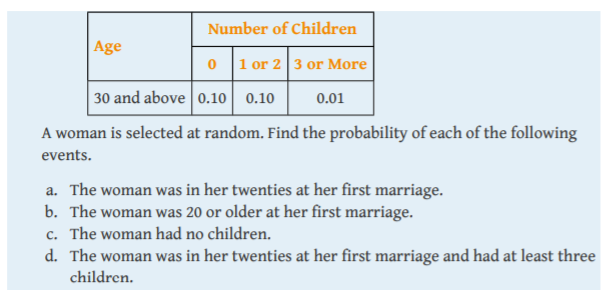
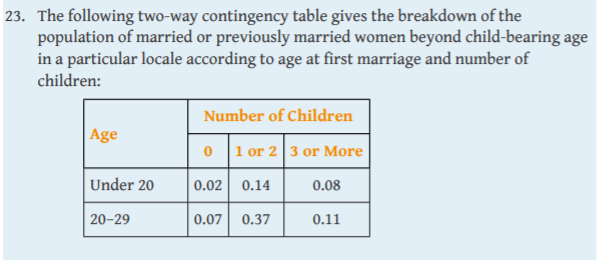
number between 0 and 1. The probabilities of all the outcomes add up to 1.

### Probability of event:

P(A) = event A中元素發生的次數合 / Sample Space 中元素發生的次數合

= sum of the probabilities of the outcomes in A.

Exercises:



Ans: a = 0.55, b = 0.76, c = 0.19, d = 0.11

## 3.2 Complements, Intersections, and Unions

* To learn how some events are naturally expressible in terms of other events.
* To learn how to use special formulas for the probability of an event that is expressed in terms of one or more other events.

### Complements 補集

Definition:

A is an event in sample space S,

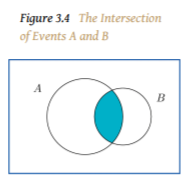
A 的 complement = Ac = 除了A有的元素外，S其他所有的元素

P(Ac) = 1 – P(A)

### Intersection 交集

Definition:

The collection of all outcomes that are elements of both of the sets A and B.

 A ∩ B

屬於A 且 屬於B， and

### Mutually Exclusive 獨立

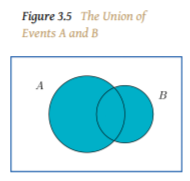
Definition:

A and B have no elements in common.

P(A ∩ B) = 0

### Union 聯集

Definition:

The collection of all outcomes that are elements of one or the other of the sets A and B, or of both of them.

A ∪ B

屬於A 或者 屬於B， or

### Additive Rule of Probability

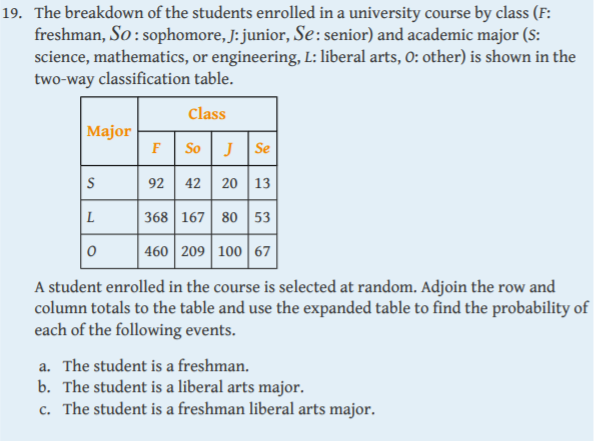
Definition:

P (A ∪ B) = P (A) + P (B) − P (A ∩ B)

P(A或者B) = P(A) + P(B) – P(A和B)

Exercises:

C:\Users\user\AppData\Local\Microsoft\Windows\INetCache\Content.Word\3.2ex.1-2.png



Ans:

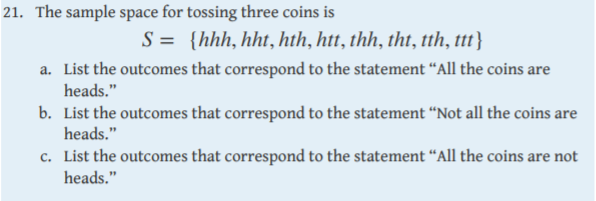
a = 0.551

b = 0.400

c = 0.220

d = 0.730

e = 0.600



Ans:

a = ['hhh']

b = ['hht', 'hth', 'htt', 'thh', 'tht', 'tth', 'ttt']

c = ['ttt']

## 快記

### 常用統計量或指標

#### 原始數據特徵值之計算

類別型特徵

量化方法，主要為百分比

連續型數據分析之特徵主要可分為4類(如何量化)

1. 集中趨勢(Central Tendency of Location)
2. 離中趨勢(Dispersion)
3. 偏態(Skewness)
4. 峰態(Kurtosis)

##### 集中趨勢(Central Tendency of Location)

集中趨勢指標，表示一組數據中央點位置所在

mean, median, mode

當有離群值時，用 mode and median，否則用mean。

##### 離中趨勢(Dispersion)

一組數據間差異大小或數值變化的一個量數

range, variance & standard deviation, coefficient of variation

range: 全距

R = Max – Min

coefficient of variation: 變異係數

##### 偏態(Skewness)

對稱:

平均數 約等於 中位數

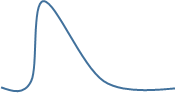
單峰分布，眾數也是約等於

正偏、右偏:

平均數 大於 中位數

負偏、左偏:

平均數 小於 中位數



偏態係數:

樣本偏態係數 g1 :

g1 ≅ 0 對稱

g1 > 0 右偏，越大偏越多

g1 < 0 左偏，越負偏越多

##### 峰態(Kurtosis)

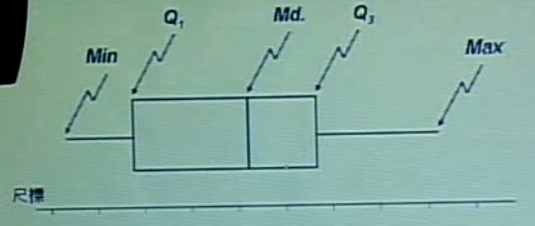
峰度係數 g2 :

常態分布: g2 = 0 (意思是前面那串算出來會是3，所以扣掉)

g2 > 0 高峽峰(又高又細又尖)

g2 < 0 低闊峰(又矮又寬又平)

##### 盒鬚圖Box Plot :



Q1 : 第1四分位數、25%數

Q2 : 第2四分位數、中位數

Q3 : 第3四分位數、75%數

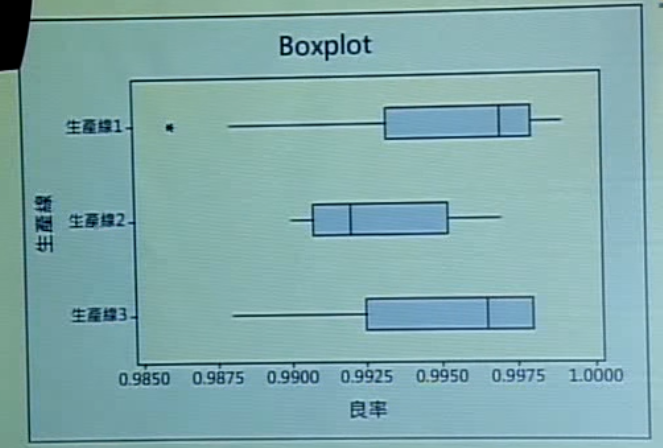
功能:

從視覺上即可有效找出資料之主要表徵值

可同時標出資料的 集中趨勢、離中趨勢、偏態、最小最大值

如何辨認 界值、離群值(Outliers)?

1. 超過 1.5( Q3 – Q1 ) ~ 3 (Q3 – Q1) ，可能的離群值
2. 超過 3 (Q3 – Q1) ，非常可能的離群值



### 機率

#### relationship in event A & B

1. 相依: 會互相影響
2. 獨立: 無任何關係或不互相影響
3. 互斥: A, B不可能同時發生

#### 機率原理

* 0 ≤ P(A) ≤ 1
* P(0) 空集合 = 0 , P(S) = 1
* A1, A2, …, Ak 互斥

P(A1∪A2∪…∪Ak) = P(A1) + P(A2) +…+P(Ak) ∪: or

ex:

1張黑桃，1張K，1張黑桃K，1張黑桃或1張K

Ans: 13/52, 4/52, 1/52, (13+4-1)/52[A集合+B集合-聯集]

3個小孩只有1個女生，只有1個女生且是老大

#(S) = 8 (number of element)

A = {GBB,BGB,BBG}, #(A) = 3, P(A) = 3/8

B = {GBB}, P(B) = 1/8

兩個骰子出現不同點數的機率?

等於 1 – 出現相同點數的機率

1 – 1/6\*1/6\*6 = 1 – 1/6 = 5/6

#### Conditional Probability:

P(A|B) : B已經發生時，A發生的機率，where P(B) nonzero

P(A|B) = P(A∩B) / P(B) ∩: and

#### Bay’s Theorem:

P(A|B) = A̅ =A沒發生

P(B) = P(B|A)\*P(A)+P(B|A ̅)\*P(A ̅)

= A發生且B發生\*A發生 + A沒發生且B\*A沒發生

ex:

晴天有太陽 = 0.8 雨天有太陽 = 0.4 晴天的機率 = 0.6

今天是有太陽且是晴天的機率

P(晴天|有太陽) =

晴天有太陽 = 0.8 = P(有太陽|晴天)

雨天有太陽 = 0.4 = P(有太陽|雨天) aka 不是晴天

晴天的機率 = 0.6 = P(晴天)

不是晴天的機率 = 1- 0.6 = 0.4 aka 雨天

P(晴天|有太陽) =