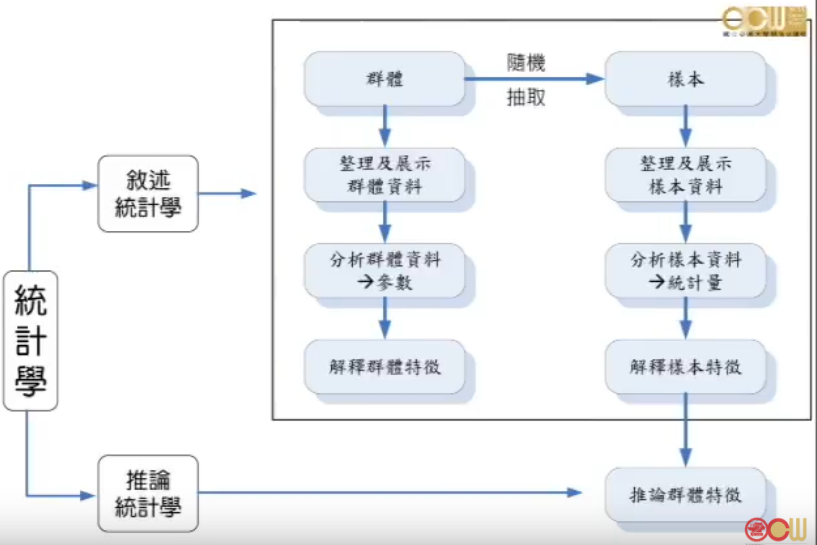
# CH1 Introduction

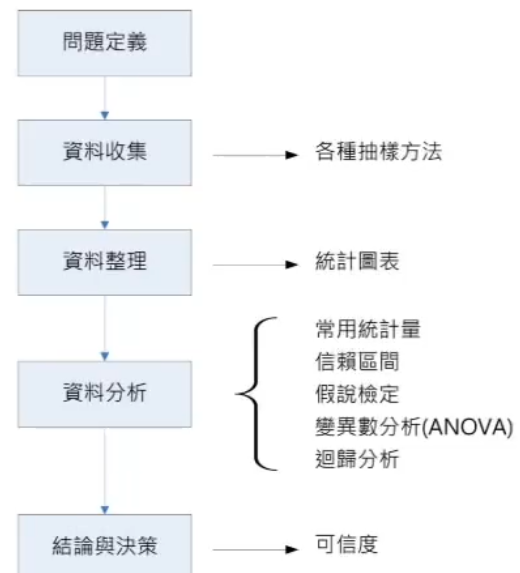
* Statistics: 蒐集、整理、展示、分析、解釋資料，並由樣本推論群體，使在不確定的情況下做成決策的科學方法
* Population: 根據研究目的蒐集個體之資料(data set)
* Sample: Part of population
* Parameter: Feature of population.
* Statistic: Feature of sample.

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.

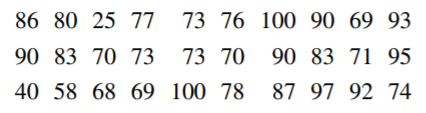
1. Continuous: 連續型

Data obtained through a measuring process.

# CH2 Descriptive Statistics

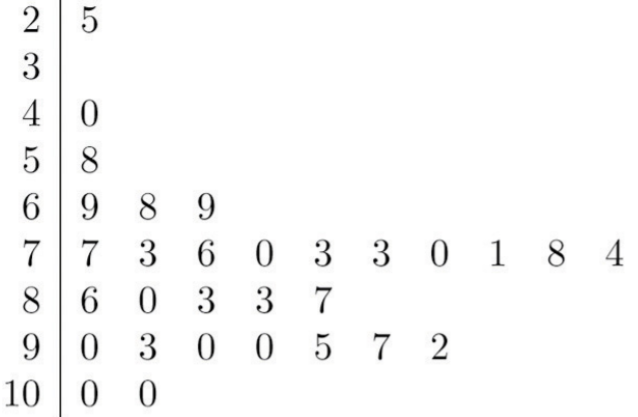
## 2.1 Three popular data displays

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

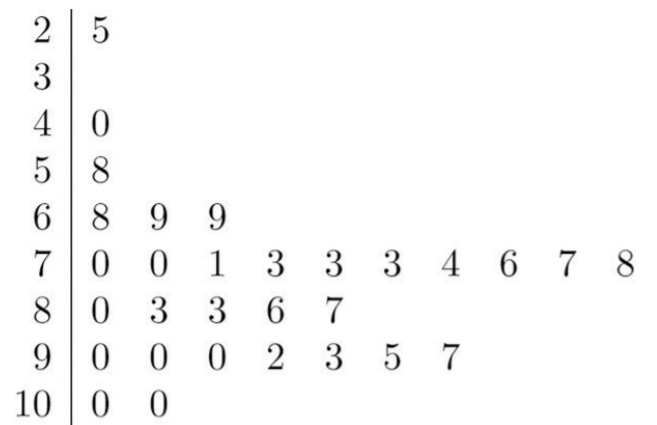


original data set

### Stem and Leaf Diagrams



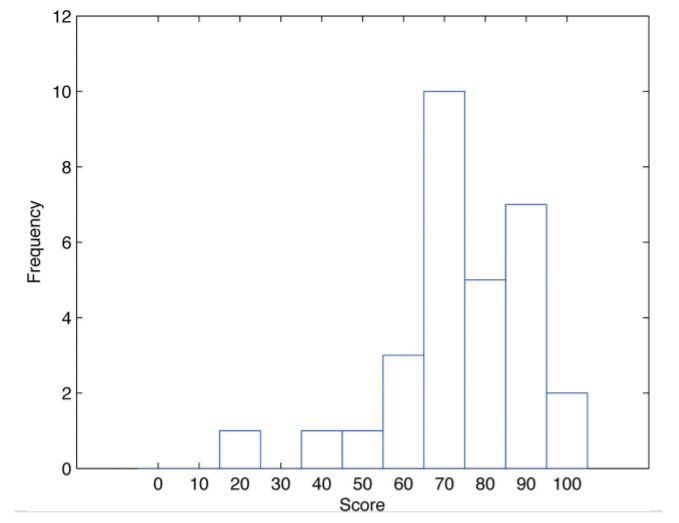
Stem and Leaf Diagram



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.

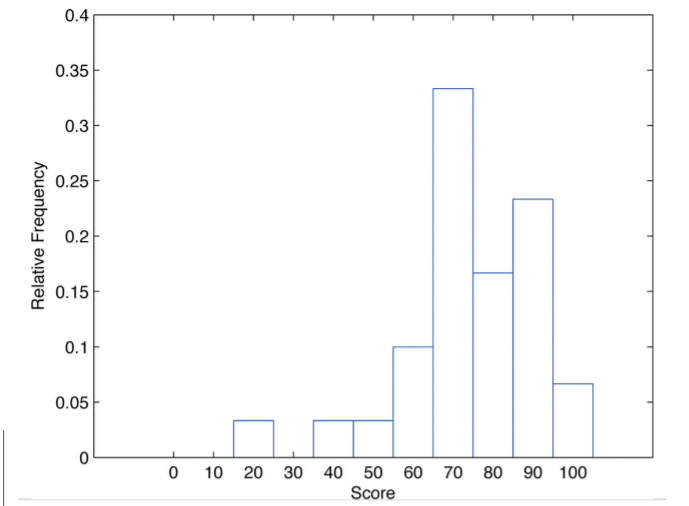
### Frequency Histograms



Frequency Histogram

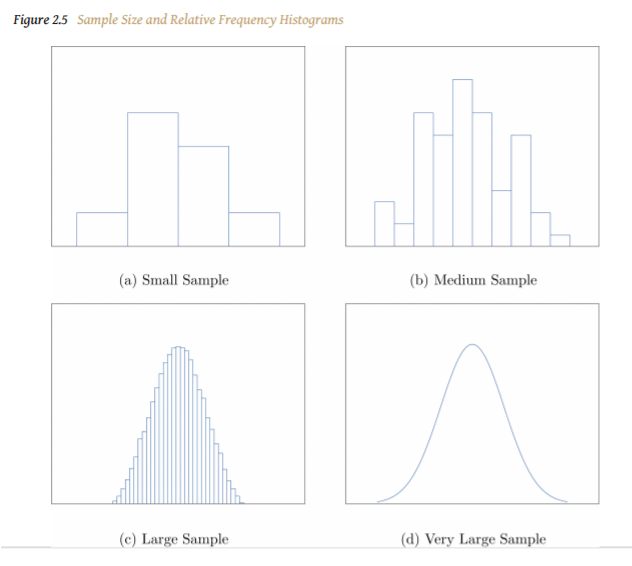
* group the scores on the standard ten-point scale, and count the number of scores in each group.
* gives a sense of data distribution across the range of values that appear.

### Relative Frequency Histograms



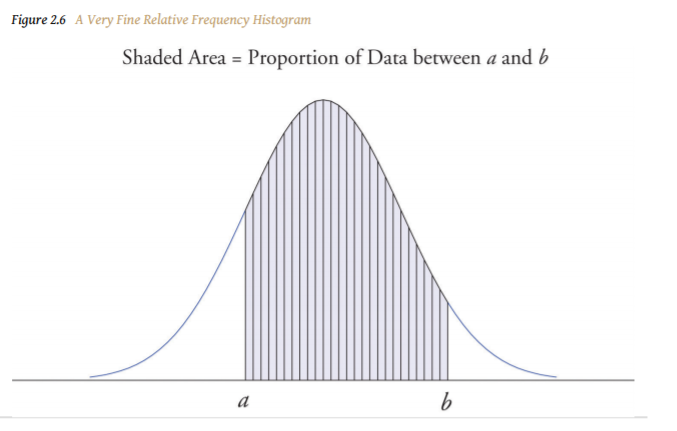
Relative Frequency Histogram

* 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.



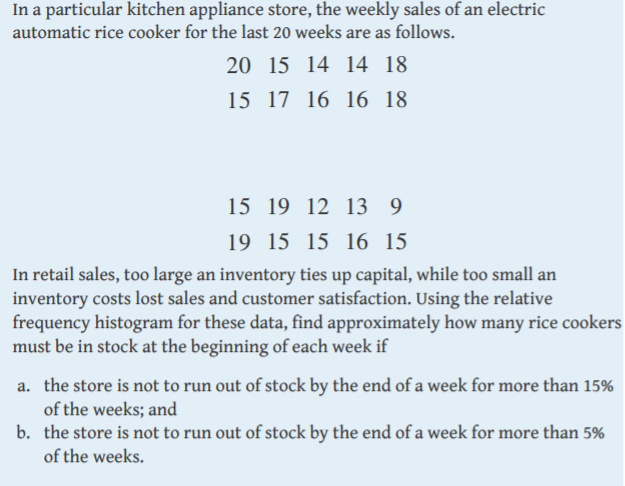
A Very Fine Relative Frequency Histogram

* 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"

Example:

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:

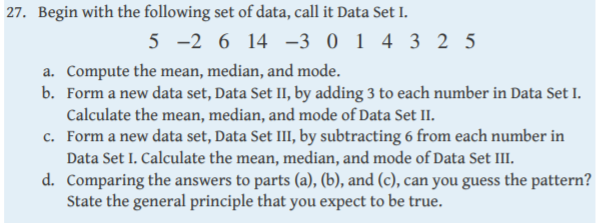
The mean of the tow middle sample data when the data are arranged in numerical order.

### The Mode

Definition:

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

Example:



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:

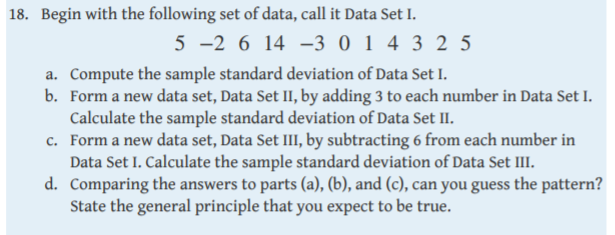
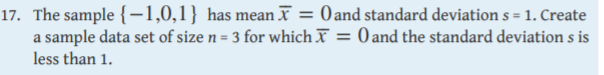
Sample variance:

Sample standard deviation:

Population variance:

Population standard deviation:

Example:



## 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.

## 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

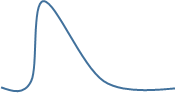
單峰分布，眾數也是約等於

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)

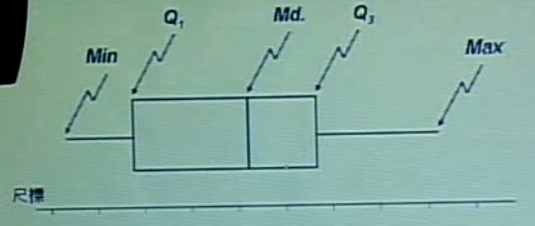
峰度係數 g2 :

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

