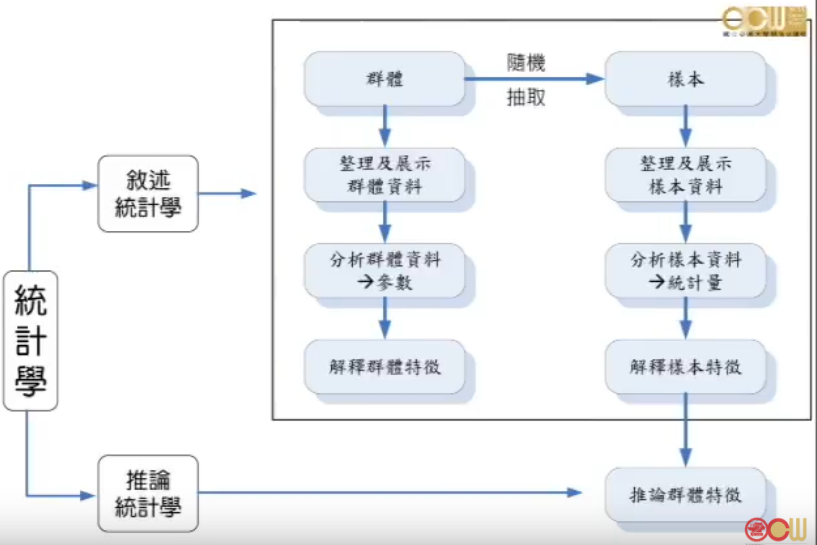
# CH1 Introduction

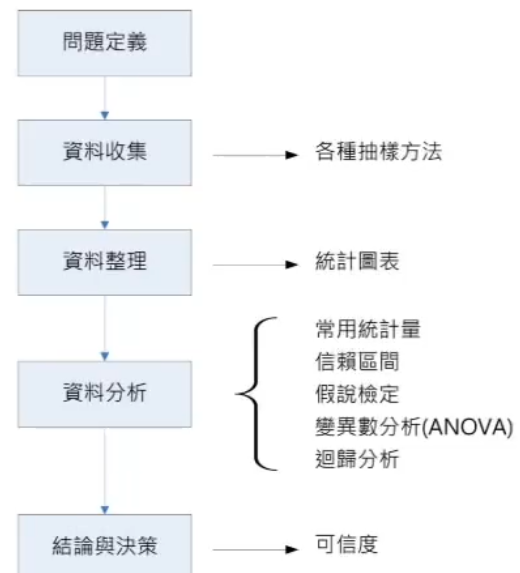
* Statistics: 蒐集、整理、展示、分析、解釋資料，並由樣本推論群體，使在不確定的情況下做成決策的科學方法
* Population: 根據研究目的蒐集個體之資料(data set)
* Sample: Part of population
* Parameter: Feature of population.
* Statistic: Feature of sample.

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.

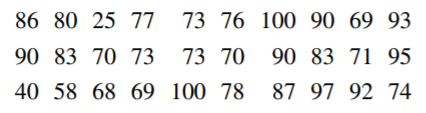
1. Continuous: 連續型

Data obtained through a measuring process.

# CH2 Descriptive Statistics

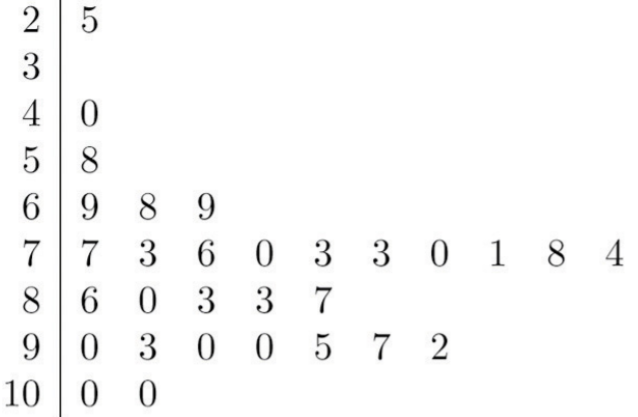
## 2.1 Three popular data displays

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

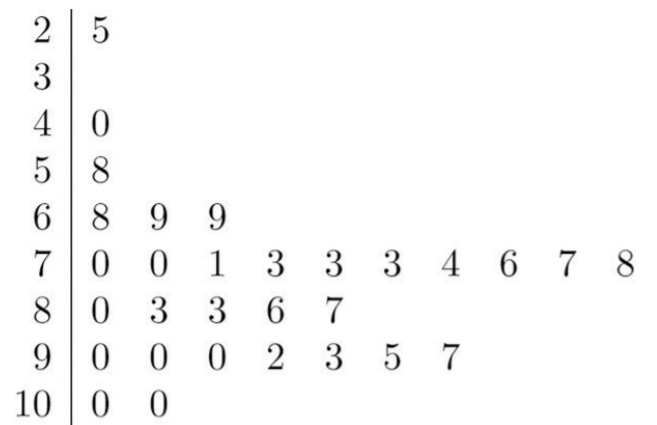


original data set

### Stem and Leaf Diagrams



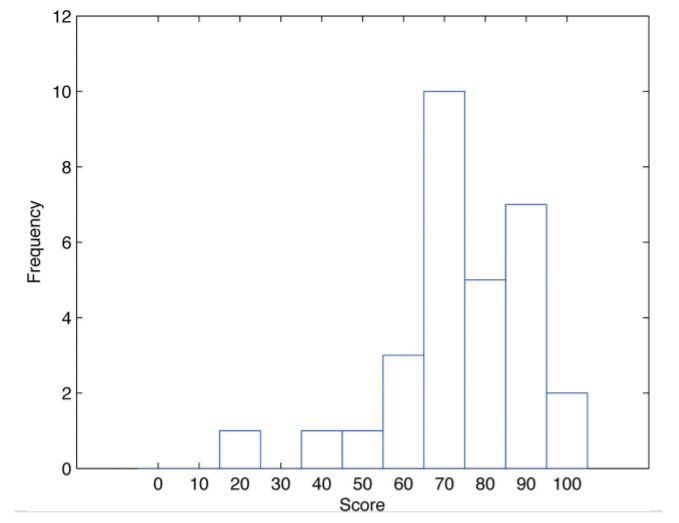
Stem and Leaf Diagram



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.

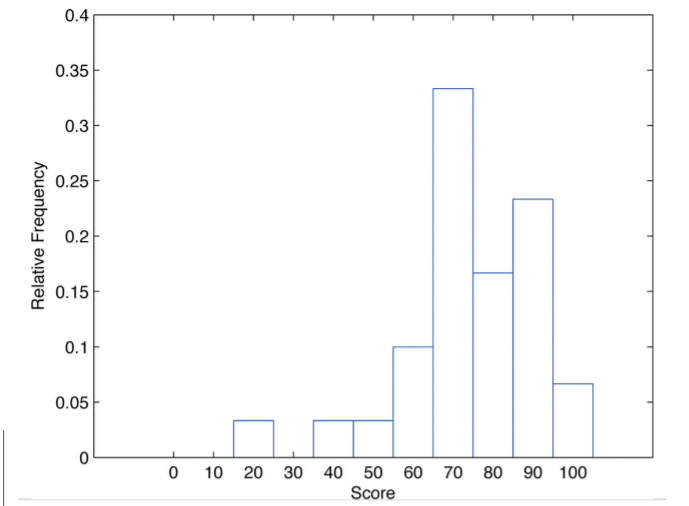
### Frequency Histograms



Frequency Histogram

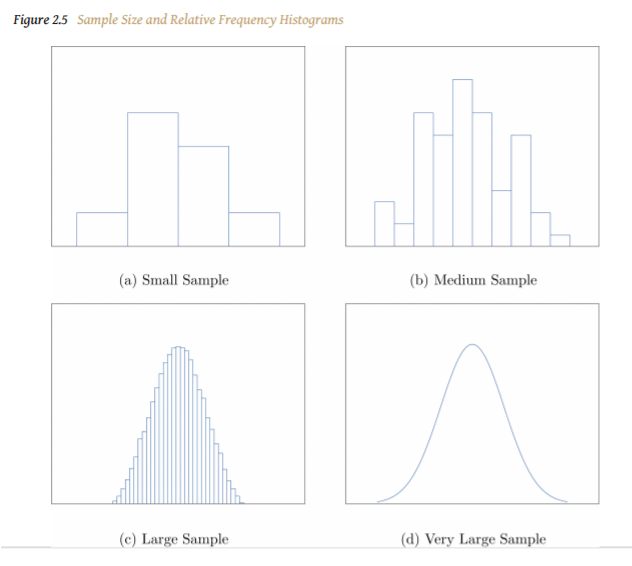
* group the scores on the standard ten-point scale, and count the number of scores in each group.
* gives a sense of data distribution across the range of values that appear.

### Relative Frequency Histograms



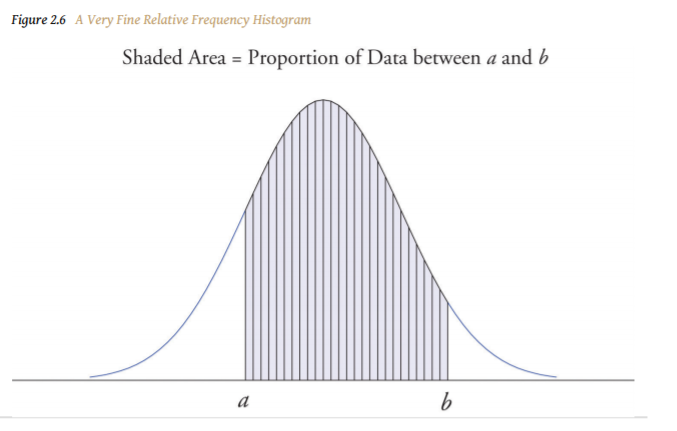
Relative Frequency Histogram

* 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.



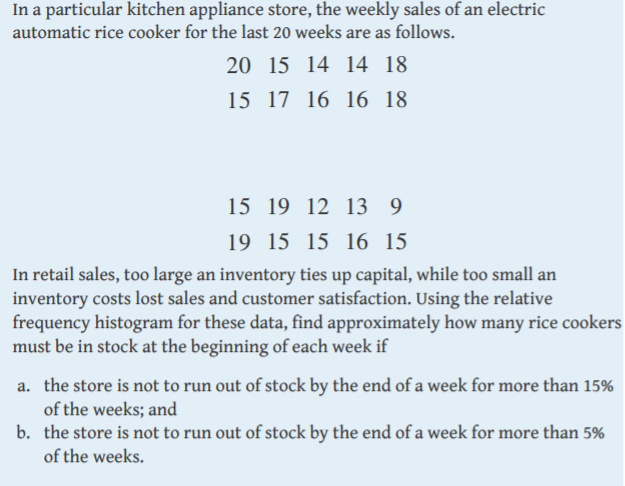
A Very Fine Relative Frequency Histogram

* 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



in

import matplotlib.pyplot as plt

ds = [20, 15, 14, 14, 18, 15, 17, 16, 16, 18, 15, 19, 12, 13, 9,19, 15, 15, 16, 15]

ds\_sorted = sorted(ds)

rf\_ds = [ds.count(i)/ len(ds) for i in range(9,21) ]

sum\_rf\_ds = [sum(rf\_ds[0:i])for i in range (1,13)]

print(ds\_sorted)

print(rf\_ds)

print(sum\_rf\_ds)

plt.bar([i for i in range(9,21)], rf\_ds)

plt.plot([i for i in range(9,21)], sum\_rf\_ds)

plt.show()

out

[9, 12, 13, 14, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 17, 18, 18, 19, 19, 20]

[0.05, 0.0, 0.0, 0.05, 0.05, 0.1, 0.3, 0.15, 0.05, 0.1, 0.1, 0.05]

[0.05, 0.05, 0.05, 0.1, 0.15000000000000002, 0.25, 0.55, 0.7000000000000001, 0.7500000000000001, 0.8500000000000001, 0.9500000000000001, 1.0]

## 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.

## 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:

## 2.4 Relative Position of Data