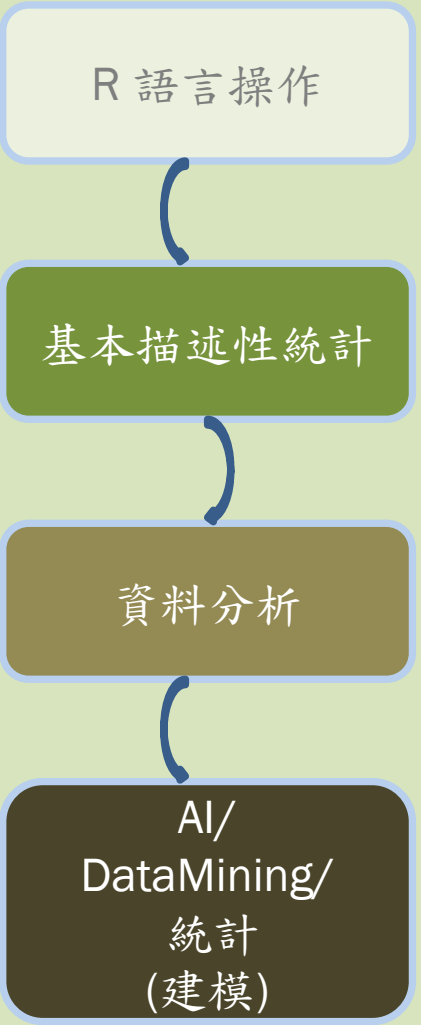
STATISTICS IN R

IOTB_ex64.R

Contents

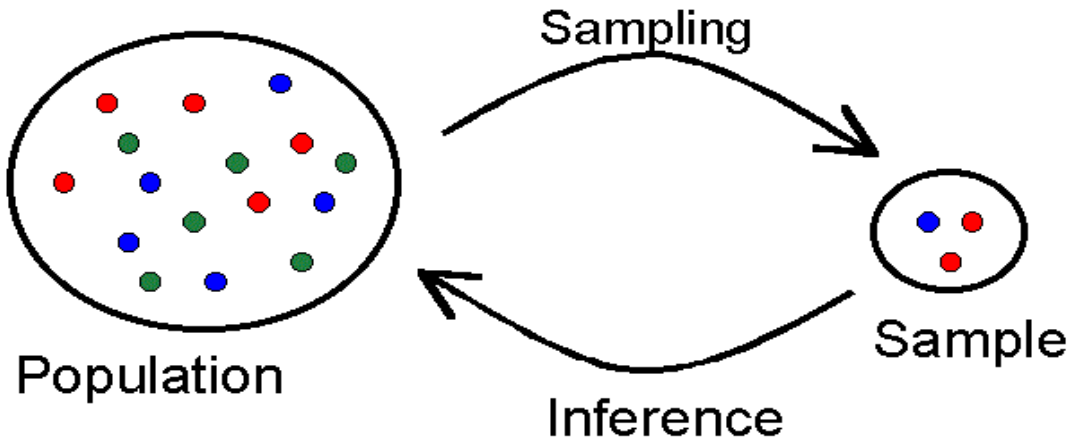
- Basic Concepts
- Data Types in Statistics
- Mean, median and mode
- Other most used statistic
- Correlation and Covariance
- Probability Distributions



Basic Concepts

Population（母體）v.s. Sample（樣本）

- For example, all of University students v.s. the students in Kun-shan University for the investigation in smoking
- Sampling: 抽樣
- Inference: 推論



Parameter(參數or母數)v.s. Statistic(統計量)

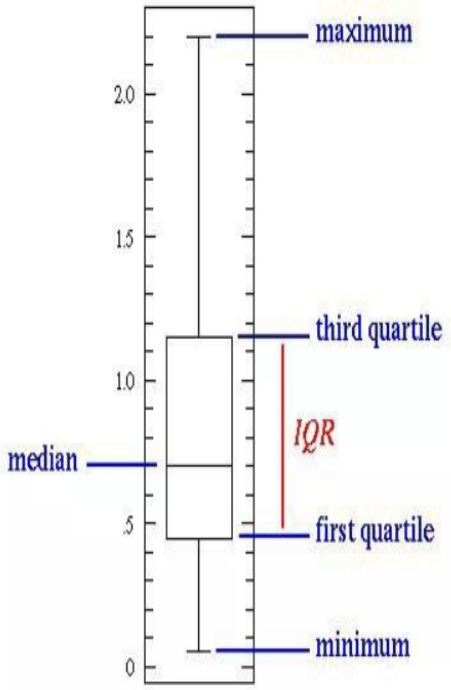
- Parameters are numbers that summarize data for an entire population.
- Statistics are numbers that summarize data from a sample, i.e. some subset of the entire population.
- They are both descriptions of groups, like “50% of dog owners prefer X Brand dog food.”
  The difference between a statistic and a parameter is that statistics describe a sample.
  A parameter describes an entire population. … You calculated what the population was likely to do based on the sample.

統計與資料分析

- 統計，顧名思義即將資料統括起來進行計算的意思，它是對數據進行定量處理的理論與技術。
- 統計分析，常指對收集到的有關數據資料訊進行整理歸類併進行解釋的過程。
- 或者，統計學是一種利用數學理論來進行資料分析的技術，通過統計學我們可以用更富有資訊驅動力和針對性的方式對資料進行操作。
- 在資料分析工作中，利用統計學，我們可以更深入、更細緻地觀察資料是如何進行精確組織的結構，與建模
- 本課程所使用的資料分析方法, 大部份為統計分析
- 統計分析方法的主要內容:統計分析可分為**描述統計**和**推斷統計**。
  - **描述統計**是將研究中所得的資料訊加以整理、歸類、簡化或繪製成圖表，以此描述和歸納數據的特徵及變數之間的關係的一種最基本的統計方法。描述統計主要涉及數據的集中趨勢、離散程度和相關強度，最常用的指標有平均數()、標準差、相關係數等。
  - **推斷統計**指用**機率形式**來決斷數據之間是否存在某種關係，以及用樣本統計值來推測總體特徵的一種重要的統計方法。推斷統計包括總體參數估計和假設檢驗，最常用的方法有Z檢驗、T檢驗、卡方檢驗等。

資料分析中常用的五個統計學基本概念

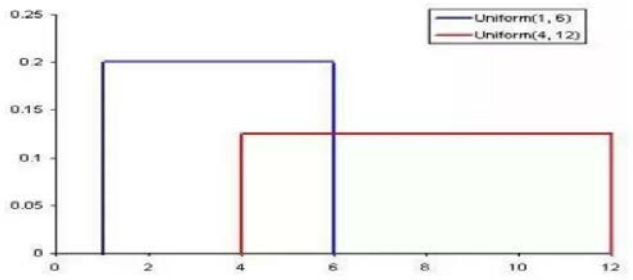
1. **特徵(描述)統計**:特徵統計可能是資料科學中最常用的統計學概念。它是你在研究資料集時，經常使用的統計技術，包括**偏差、方差、四分位數差、平均值、中位數、百分數**等等。理解特徵統計，並且在R程式碼中實現都是非常容易的。例如看下圖(**四分位數差之**箱形圖)：
   - **四分位距**（IQR）。是描述統計學中的一種方法，以確定第三個四分位數和第一個四分位數的差值（即Q1與 Q3的差距）

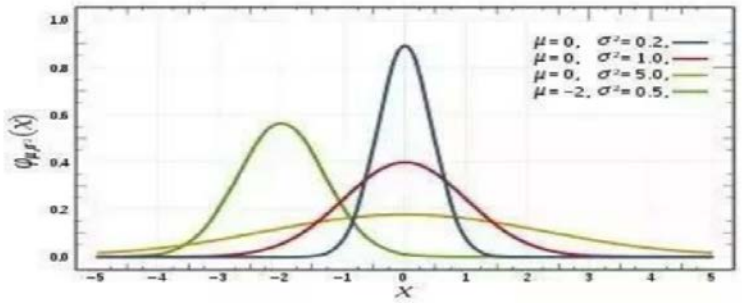


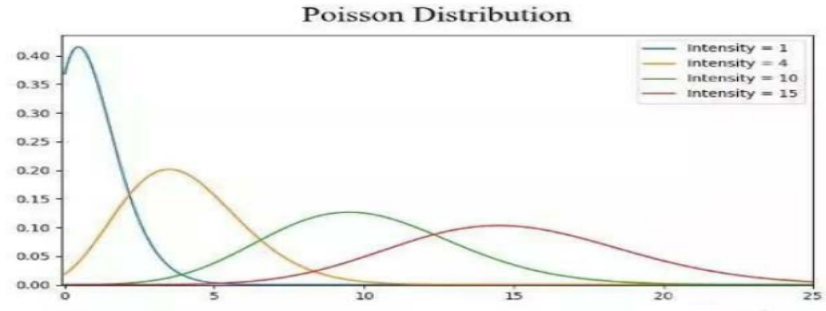
資料分析中常用的五個統計學基本概念

2. **機率分佈:**我們可以將機率定義為一些事件將要發生的可能性大小，以百分數來表示。在資料科學領域中，這通常被量化到0到1的區間範圍內，**其中0表示事件確定不會發生，而1表示事件確定會發生**。那麼，機率分佈就是表示所有可能值出現的機率的函數。請看下圖常見的機率分佈，**均勻分佈、常態分佈、卜瓦松分佈：　：**

   - 均勻分佈是其中最基本的概率分佈方式。**它有一個只出現在一定範圍內的值，而在該範圍之外的都是0**。我們也可以把它考慮為是一個具有兩個分類的變數：0或另一個值。分類變數可能具有除0之外的多個值，但我們仍然可以將其視覺化為多個均勻分佈的分段函數
   - 常態分佈，具體是由它的平均值和標準偏差來定義的。**平均值是在空間上來回變化位置進行分佈的，而標準偏差控制著它的分佈擴散範圍**。與其它的分佈方式的主要區別在於，在所有方向上標準偏差是相同的。因此，我們知道資料集的平均值以及資料的擴散分佈，即它在比較廣的範圍上擴充套件，還是主要圍繞在少數幾個值附近集中分佈。

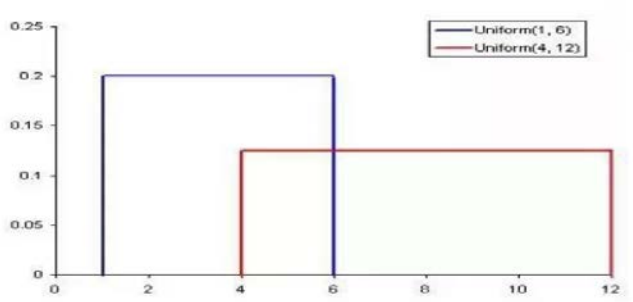


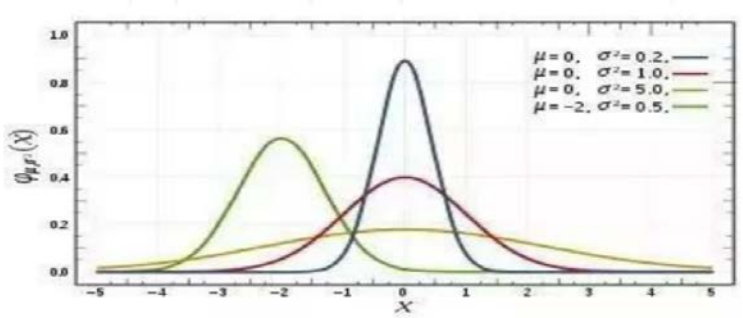


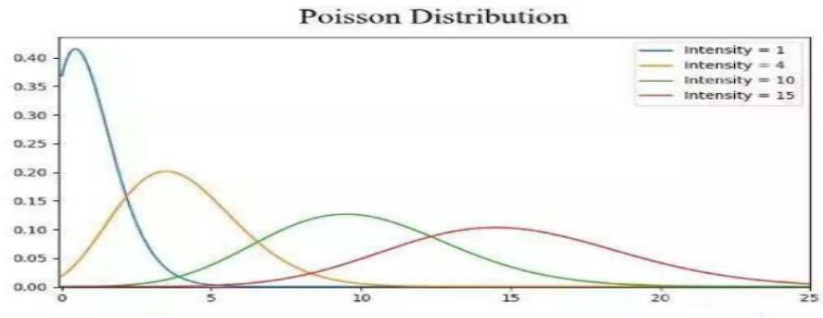


資料分析中常用的五個統計學基本概念

2. **機率分佈:**我們可以將機率定義為一些事件將要發生的可能性大小，以百分數來表示。在資料科學領域中，這通常被量化到0到1的區間範圍內，**其中0表示事件確定不會發生，而1表示事件確定會發生**。那麼，機率分佈就是表示所有可能值出現的機率的函數。請看下圖常見的機率分佈，**均勻分佈、常態分佈、卜瓦松分佈：　：**
   - **卜瓦松分佈**與正態分佈相似，但存在偏斜率。**像常態分佈一樣，在偏斜度值較低的情況下，卜瓦松分佈在各個方向上具有相對均勻的擴散**。但是，當偏斜度值非常大的時候，我們的資料在不同方向上的擴散將會是不同的。在一個方向上，資料的擴散程度非常高，而在另一個方向上，擴散的程度則非常低。

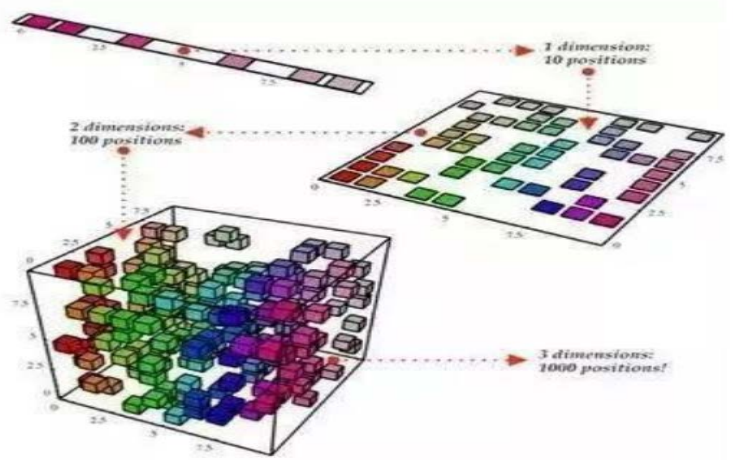






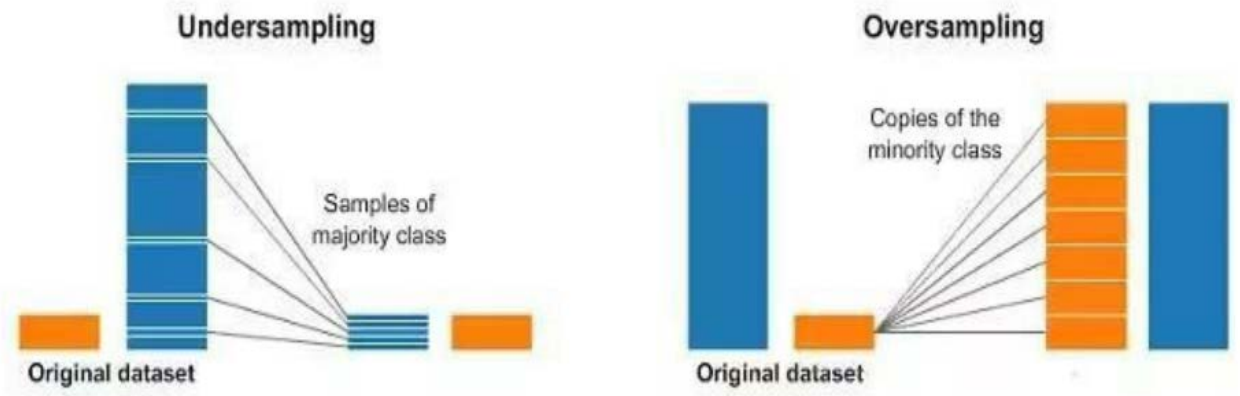
資料分析中常用的五個統計學基本概念

3. **降維：**是降低一個數據集的維數。在資料科學中，這是特徵變數的數量。請看下圖：
   - 下圖中的立方體表示我們的資料集，它有3個維度，總共1000個點。**以現在的計算能力，計算1000個點很容易，但如果更大的規模，就會遇到麻煩了**。然而，僅僅從二維的角度來看我們的資料，比如從立方體一側的角度，可以看到劃分所有的顏色是很容易的。通過降維，我們將3D資料展現到2D平面上，這有效地把我們需要計算的點的數量減少到100個，大大節省了計算量。
   - 另一種方式是我們可以通過特徵剪枝來減少維數。利用這種方法，我們刪除任何所看到的不重要的資料。例如，在研究資料集之後，我們可能會發現，在10個特徵中，有7個特徵與輸出具有很高的相關性，而其它3個則具有非常低的相關性。**那麼，這3個低相關性的特徵可能不值得計算，我們可能只是能在不影響輸出的情況下將它們從分析中去掉。**
   - 用於降維的最常見的統計技術是PCA，
     它本質上建立了特徵的向量表示，表明了它們對輸出的重要性，即相關性。PCA可以用來進行上述兩種降維方式的操作。



資料分析中常用的五個統計學基本概念

4. **過取樣和欠取樣:**過取樣和欠取樣是用於分類問題的技術。例如，我們有1種分類的2000個樣本，但第2種分類只有200個樣本。那麼，過取樣和欠取樣可以應對這種情況。請看下圖：
   - **例如,**在下面圖中的左右兩側，藍色分類比橙色分類有更多的樣本。在這種情況下，我們有2個預處理選擇，可以幫助機器學習模型進行訓練。
   - 欠取樣, 意味著我們將只從樣本多的分類中選擇一些資料，而儘量多的使用樣本少的分類樣本。(Sol)這種選擇應該是為了保持分類的機率分佈。我們只是通過**更少的抽樣來讓資料集更均衡**。
   - 過取樣, 意味著我們將要建立少數分類的副本，(Sol)以便具有與多數分類相同的樣本數量。副本將被製作成保持少數分類的分佈。我們只是在沒有獲得**更多資料的情況下讓資料集更加均衡。**



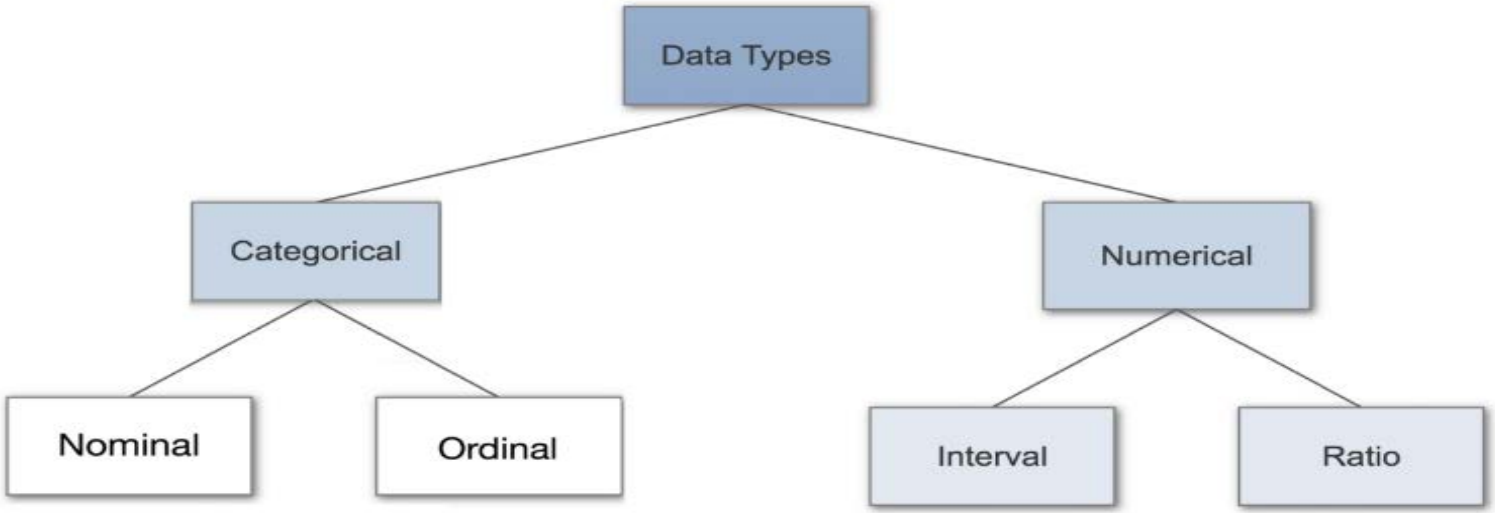
資料分析中常用的五個統計學基本概念

- **貝葉斯統計:** 假設我給了你一個骰子，問你擲出6點的機率是多少，大多數人都會說是六分之一。 但是，如果有人給你個特定的骰子，總能擲出6個點呢?因為頻率分析僅僅考慮之前的資料，而給你作弊的骰子的因素，並沒有被考慮進去。**貝葉斯統計**確實考慮了這一點，我們可以通過貝葉斯法則來進行說明:
  - 完全理解為什麼在我們使用貝葉斯統計的時候，要求首先理解頻率統計失敗的地方。大多數人在聽到“概率”這個詞的時候，頻率統計是首先想到的統計類型。
  - 例如，如果你要擲骰子10000次，並且前1000次全部擲出了6個點，那麼你會非常自信地認為是骰子作弊了。如果頻率分析做的非常好的話，那麼我們會非常自信地確定，猜測6個點是正確的。同時，如果骰子作弊是真的，或者不是基於其自身的先驗概率和頻率分析的，我們也會考慮作弊的因素。

Data Types in Statistics

Data Types in Statistics

- Categorical data – also called as Qualitative（質性）data
- Numerical data – also called as Quantitative（數量）data



Data Types in Statistics

- Data Types are an important concept of statistics, which needs to be understood, to correctly apply statistical measurements to your data and therefore to correctly conclude certain assumptions about it.

- You also need to know which data type you are dealing with to choose the right visualization method. Think of data types as a way to categorize different types of variables. We will discuss the main types of variables and look at an example for each. We will sometimes refer to them as measurement scales.

Categorical Data

- Categorical data represents characteristics. Therefore it can represent things like a person's *gender*, *language* etc.
- Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have mathematical meaning.

Categorical Data – Nominal Data

- Nominal (名義上) values represent **discrete units** and are used to label variables, that have no quantitative value. **Just think of them as "labels". Note that nominal data that has no order.** Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features below:

Examples

What is your Gender?

- ○ Female
- ○ Male

What languages do you speak?

- ○ Englisch
- ○ French
- ○ German
- ○ Spanish

- Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known. Because of that, ordinal scales are usually used to measure non-numeric features like happiness, customer satisfaction and so on.

Categorical Data – Ordinal Data

- Ordinal (順序的) values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that it's ordering matters. You can see an example below:

Example

What Is Your Educational Background?

- 1 - Elementary
- 2 - High School
- 3 - Undegraduate
- 4 - Graduate

Numerical Data

- 1. Discrete Data - We speak of discrete data if its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data **can't be measured but it can be counted**. It basically represents information that can be categorized into a classification. An example is the number of head Example) coin flips.
- 2. Continuous Data - Continuous Data represents measurements and therefore their values **can't be counted but they can be measured**. An example would be the height of a person, which you can describe by using intervals on the real number line. Example

Numerical Data-Continuous Data: Interval Data

- Interval values represent **ordered units that have the same difference**. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see below:

Example

Temperature?

- ○ - 10
- ○ -5
- ○ 0
- ○ + 5
- ○ + 10
- ○ + 15

- The problem with interval values data is

*that they **don't have a „true zero"**.*

That means in regards to our example, that there is no such thing as no temperature.

With interval data, we can add and subtract, but we cannot multiply, divide or calculate ratios.

Because there is no true zero, a lot of descriptive and inferential statistics can't be applied.

Numerical Data-Continuous Data: Ratio Data

- Ratio values are also ordered units that have the same difference. Ratio values are **the same as interval values, with the difference that they do have an absolute zero.** Good examples are height, weight, length etc.

Example

Length (inch)?

- ○ 0
- ◉ 5
- ○ 10
- ○ 15

The difference between interval and ratio scales comes from their ability to dip below zero.
Interval scales hold no true zero and can represent values below zero. ...
Ratio variables, on the other hand, never fall below zero.

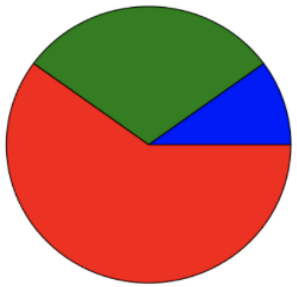
Why Data Types are important?

- Data types are an important concept because statistical methods can only be used with certain data types. You have to analyze continuous data differently than categorical data otherwise it would result in a wrong analysis. Therefore knowing the types of data you are dealing with, enables you to choose the correct method of analysis.

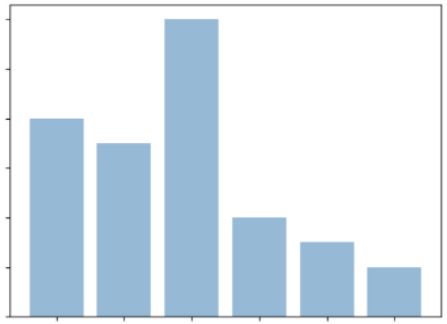
Other talks

- **Nominal Data -** When you are dealing with nominal data, you collect information through the following for your analysis:
  - ***Frequencies****: The Frequency is the rate at which something occurs over a period of time or within a dataset.*
  - ***Proportion****: You can easily calculate the proportion by dividing the frequency by the total number of events. (e.g how often something happened divided by how often it could happen)*
  - ***Percentage.***
  - *Visualisation Methods: To visualise nominal data you can use a pie chart or a bar chart.*

Pie Chart



Bar Chart



Other talks

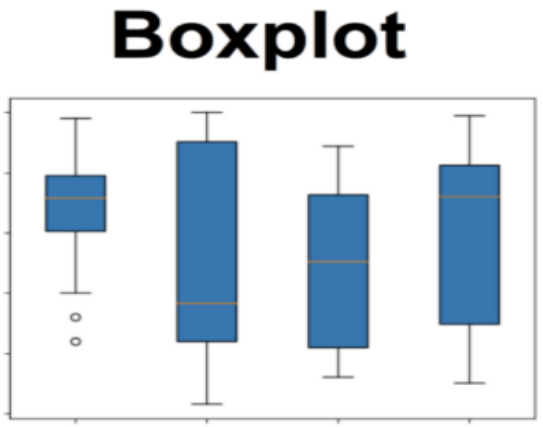
- **Ordinal Data -** When you are dealing with ordinal data, you can use the same methods like with nominal data, but you also have access to some **additional statistical tools**. Therefore you can summarise your ordinal data with frequencies, proportions, percentages. And you can visualise it with **pie** and **bar charts**. Additionally, you can use percentiles(百分位), median, mode and the interquartile range to summarise your data.

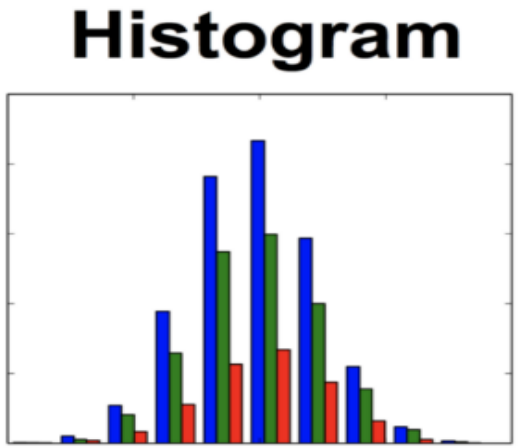
- p.s. interquartile range (IQR) - situated between the first and third quartiles of a distribution.

Other talks

- **Continuous Data** - When you are dealing with continuous data, you can use the most methods to describe and analyze your data. You can summarise your data using percentiles, median, interquartile range, mean, mode, standard deviation, and range.
- To visualise continuous data, you can use a histogram or a box-plot. With a histogram, you can check the central tendency, variability(變化度), modality(樣式、形態), and kurtosis(峰度) of a distribution(分佈). Note that a histogram can't show you if you have any outliers. This is why we also use box-plots (**boxplot** is a method for graphically depicting groups of numerical data through their quartiles (四分位數).

p.s. Boxplot is a graph to display the distribution for a set of data.



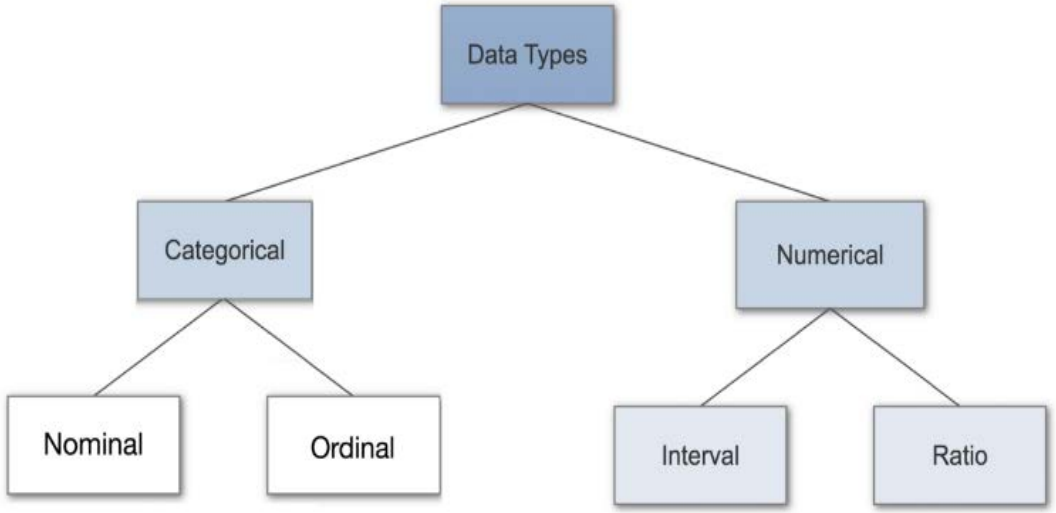


Summary

- In this post, you discovered the different data types that are used throughout statistics. You learned the difference between discrete & continuous data and learned what nominal, ordinal, interval and ratio measurement scales are.
- Furthermore, you now know what statistical measurements you can use at which datatype and which are the right visualization methods. You also learned, with which methods categorical variables can be transformed into numeric variables. This enables you to create a big part of an exploratory analysis on a given dataset.

Exercise: Data Type Identification

- As follows:
  - *A. An employee's age*
  - *B. The number of long-distance calls a month*
  - *C. The time of a long-distance call*
  - *D. Seasons*
  - *E. Employee's position*



What options belong to Categorical data? Numerical data? How about continuous data or discrete data in Numerical type?

mean, median and mode

The Calculations for mean, median and mode

- Statistical analysis in R is performed by using many built-in functions. Most of these functions are part of the R base packages. These functions take R vector as an input along with the arguments and give the result.
- The major measures for **data central tendency（資料集中評估分析）** are focused on mean, median, and mode calculation in which you can find out the data central tendency for your data distribution. However, the results of these calculation just offer possible-reference values for the beginning estimation.
- The functions we are discussing in this section are mean(平均數), median(中位數:代表一個樣本、或機率分布中的一個數值. 對於有限的數集，可以通過把**所有觀察值低高排序後**, 找出正中間的一個作爲中位數。如果觀察值有偶數個，則中位數不唯一，通常取最中間的兩個數值的平均數作爲中位數) and mode(眾數:指一組數據中出現次數最多的數據值).
  - 先將11位同學投中的次數由小到大排序，為

    2 3 3 3 3 ***3*** 4 4 6 6 7

    因為資料筆數(11位同學)為奇數，最中間的位置為第6位，所以中位數(median)為3, 眾數(mode)亦為3. 而mean為4.

Mean (平均數)

- It is calculated by taking the sum of the values and dividing with the number of values in a data series.
- The function **mean()** is used to calculate this in R.
- Syntax: The basic syntax for calculating mean in R is –

```
mean(x, trim = 0, na.rm = FALSE, ...)
```

Following is the description of the parameters used –

- **x** is the input vector.
- **trim** is used to drop some observations from both end of the sorted vector.
- **na.rm** is used to remove the missing values from the input vector.

ex

```
mean(x, trim = 0.1)
```

就是先把x的最大的10%的數和最小的10%的數去掉，然後剩下的數算平均。

```
> x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
> # Find Mean.
> result.mean <- mean(x)
> print(result.mean)
[1] 8.22
```

p.s. no trim parameter on this example

Applying Trim Option in mean()

- When trim parameter is supplied, the values in the vector get sorted and then the required numbers of observations are dropped from calculating the mean.
- When trim = 0.3 (註:30% 在前後端個數), 3 values from each end will be dropped from the calculations to find mean.
- In this case the sorted vector is (~~−21, −5, 2~~, **3, 4.2, 7, 8**, ~~12, 18, 54~~) and the values removed from the vector for calculating mean are (−21,−5, 2) from left and (12,18,54) from right.

```
> x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
> # Find Mean.
> result.mean <-  mean(x,trim = 0.3)
> print(result.mean)
[1] 5.55
```

Applying NA Option in mean()

- If there are missing values, then the mean function returns NA.
- To drop the missing values from the calculation use na.rm = TRUE. which means remove the NA values.

```
> x <- c(12,7,3,4.2,18,2,54,-21,8,-5,NA)
> result.mean <-  mean(x)
> print(result.mean)
[1] NA
> result.mean <-  mean(x,na.rm = TRUE)
> print(result.mean)
[1] 8.22
```

Median (中位數)

- The middle most value in a data series is called the median. The function, **median(),** is used in R to calculate this value.
- Syntax: The basic syntax for calculating median in R is –

```
median(x, na.rm = FALSE)
```

Following is the description of the parameters used –

- **x** is the input vector.
- **na.rm** is used to remove the missing values from the input vector.

```
> x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
> median.result <- median(x)
> print(median.result)
[1] 5.6
```

p.s. Sorted vector: -21, -5, 2, 3, **4.2, 7**, 8, 12, 18, 54 ➔ Median= (4.2+7)/2=5.6
due to even number of occurrences

Mode (眾數)

- The mode is the value that has highest number of occurrences in a set of data. Unlike mean and median, mode can have both numeric and character data.
- R does not have a standard in-built function to calculate mode. So we create a user function to calculate mode of a data set in R. This function takes the vector as input and gives the mode value as output.

Mode calculation

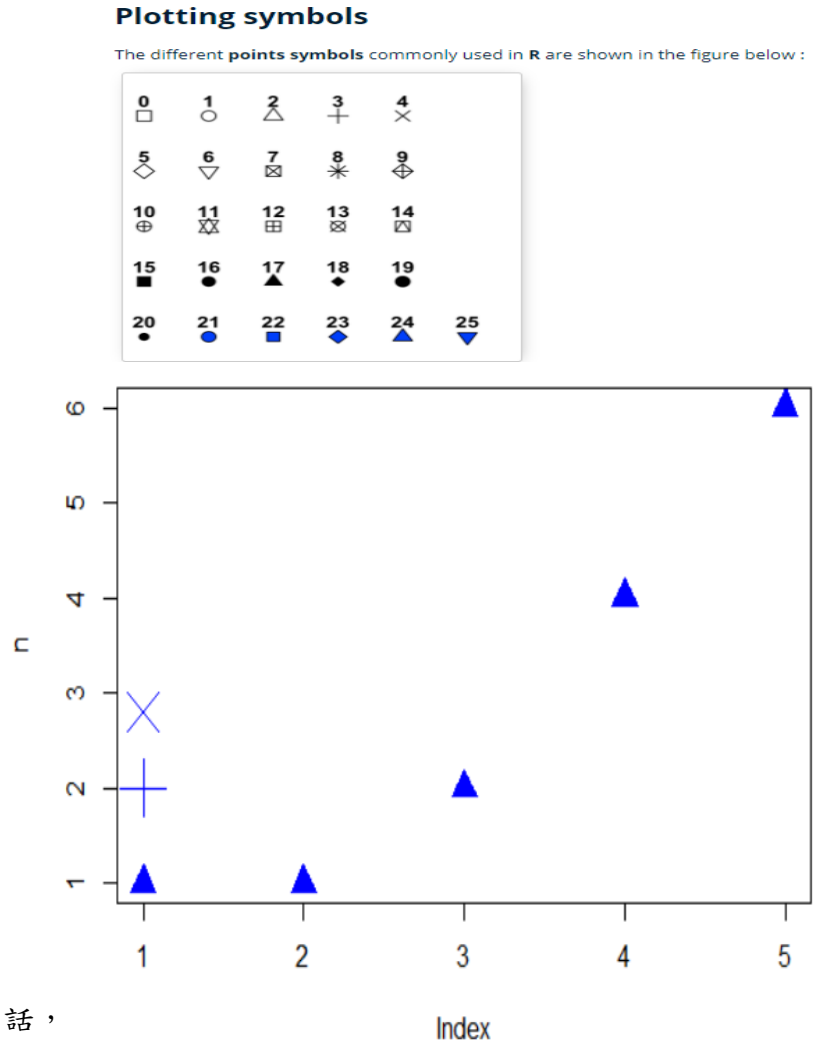
```
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
> # run1: Calculate the mode using the user function.
> v <- c(2,1,2,3,1,2,3,4,1,5,5,3,2,3)
> result <- getmode(v)
> print(result)
[1] 2
>
> #  run2: Calculate the mode using the user function.
> charv <- c("o","it","the","it","it")
> result <- getmode(charv)
> print(result)
[1] "it"
```

Exercise

- This exercise covers all of ideas talked on this Section for data tendency distribution.

```
# exercise - 集中趨勢
n <- c(1,1,2,4,6)
plot(n, pch = 17, col = "blue", cex =2)
```

```
> #總合
> sum(n)
[1] 14
> #平均數 總合除個數
> mean(n)
[1] 2.8
> #中位數:將資料由小到大，位置居中者，就是中位數
> median(n)
[1] 2
> #眾數:一組資料中，出現最多次數的值
> as.numeric(names(table(n)))[which.max(table(n))]
[1] 1
> #畫平均數的點"X"
> points(mean(n), pch = 4, col = "blue", cex = 3)
> #畫中位數的點"+"
> points(median(n), pch = 3, col = "blue", cex = 3)
```



p.s. 圖形中資料點的大小可以使用 cex 參數來指定，其預設值是 1，若指定為 2.5 的話，所有的資料點就會變成原來的 2.5 倍大

Other most used statistic

Other most used statistic functions

- The major measures for **discrete data tendency** are focused on standard deviation(標準差), variation (變異數), range (全距) and Quattile (四分位差) calculation in which you can find out the discrete data tendency for your data distribution. However, the results of these calculation just offer possible-reference values for the beginning estimation.

Standard Deviation(標準差)

- 在機率統計中,最常使用作為測量一組數值的離散程度之用。
- 標準差定義：為**變異數**開主平方根，反映**組內個體**間的離散程度。標準差也可稱標準機差
- 一個**總量**的標準差或一個**隨機變量**的標準差，及一個**子集合**樣品數的標準差之間，有所差別。其公式如下所列。
- 母體(總量)的標準差:

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

μ為平均值（$\overline{x}$）。

- 母體為隨機變數的標準差:並非所有隨機變量都具有標準差，因為有些隨機變量不存在期望值。

$$\sigma = \sqrt{\mathrm{E}((X - \mathrm{E}(X))^2)} = \sqrt{\mathrm{E}(X^2) - (\mathrm{E}(X))^2}$$

  - **期望值:**例如，擲一枚公平的六面骰子，其每次「點數」的期望值是3.5，計算如下：

$$\mathrm{E}(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$
$$= \frac{1+2+3+4+5+6}{6} = 3.5$$

- 樣本的標準差:在真實世界中，找到一個母體的真實的標準差並不實際。大多數情況下，母體標準差是通過隨機抽取一定量的樣本並計算樣本標準差估計的。

從一大組數值$X_1, \cdots, X_N$當中取出一樣本數值組合$x_1, \cdots, x_n : n < N$，常定義其**樣本標準差**：

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standard Deviation(標準差)

- 標準差是一組數值與平均值分散開來的程度的一種測量觀念。 一個較大的標準差，代表大部分的數值和其平均值之間差異較大；一個較小的標準差，代表這些數值較接近平均值。 例如，兩組數的集合{0, 5, 9, 14} 和{5, 6, 8, 9} 其平均值都是7 ，但第二個集合具有較小的標準差。
- R 標準差函數: sd()
- For example, we use the iris dataset which is a built-in dataset (鳶尾花資料集). 從iris的資料集中，取"Sepal.Length"(花萼長度)這個欄位的資料出來.

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1
 1 1 1 1 ...
> iris$Sepal.Length
  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7
 [17] 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4
 [33] 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
 [49] 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1
 [65] 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7
 [81] 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7
 [97] 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2 6.2 6.1
[129] 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
[145] 6.7 6.7 6.3 6.5 6.2 5.9
> sd(iris$Sepal.Length)
[1] 0.8280661
```

Standard Deviation(標準差)

```
> iris$Petal.Length
  [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3
 [18] 1.4 1.7 1.5 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4
 [35] 1.5 1.2 1.3 1.4 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4 4.7
 [52] 4.5 4.9 4.0 4.6 4.5 4.7 3.3 4.6 3.9 3.5 4.2 4.0 4.7 3.6 4.4 4.5 4.1
 [69] 4.5 3.9 4.8 4.0 4.9 4.7 4.3 4.4 4.8 5.0 4.5 3.5 3.8 3.7 3.9 5.1 4.5
 [86] 4.5 4.7 4.4 4.1 4.0 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1 6.0 5.1
[103] 5.9 5.6 5.8 6.6 4.5 6.3 5.8 6.1 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9
[120] 5.0 5.7 4.9 6.7 4.9 5.7 6.0 4.8 4.9 5.6 5.8 6.1 6.4 5.6 5.1 5.6 6.1
[137] 5.6 5.5 4.8 5.4 5.6 5.1 5.1 5.9 5.7 5.2 5.0 5.2 5.4 5.1
> sd(iris$Petal.Length)
[1] 1.765298
```

Variation（變異數）

- 變異數是用於表示一組數值資料中的各數值相對於該組數值資料之平均數的分散程度。計算各數值與平均數的差，取其平方後加總，再除以數值個數，得「變異數」。 變異數開根號後得「標準差」

$$\mathrm{Var}(X) = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- 在所有人當中，要算平均身高，所以抽了三個人出來，身高分別是170、170、170，可得平均為170。若改天，又抽了三人出來，身高分別為160、170、180，也得平均為170。若只比較兩次結果，平均都為170。但還是有差別的！ 所以數學家發明了變異數，用來了解所得數據的差異。要了解一個族群最有用的兩個資訊是平均 μ 與變異數 σ^2。兩個族群平均相同，變異數可以很不同，分布可能差異很大；分布可能很不一樣。 若抽出的 Var(X) =0，表示兩個樣本都一樣！

- The function used in R

```
> sd(iris$Sepal.Length)
[1] 0.8280661
> var(iris$Sepal.Length)
[1] 0.6856935
>
> sd(iris$Petal.Length)
[1] 1.765298
> var(iris$Petal.Length)
[1] 3.116278
```

Range（全距）

- 一組資料的全距是指資料中最大值與最小值的差距。設此組資料最大值為$X_{(n)}$、最小值為$X_{(1)}$，全距以R表示，則 $R = x_{(N)} - x_{(1)}$
- 全距也是離散量數的一種，是計算資料中最大值與最小值的差距，用來表達資料分散的狀況。當全距越大，表示資料的分散狀況越大，反之則越小。但此種評估離勢趨勢較粗略！因考量到全距內,個成員對應值的分佈情況.
- The range() function in R

```
> #range 全距
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
> range(iris$Sepal.Length)
[1] 4.3 7.9
> range(iris$Sepal.Length)[2]- range(iris$Sepal.Length)[1]
[1] 3.6
```

```
> sort(iris$Sepal.Length)
  [1] 4.3 4.4 4.4 4.4 4.5 4.6 4.6 4.6 4.6 4.7 4.7 4.8 4.8 4.8 4.8 4.8 4.9 4.9
 [19] 4.9 4.9 4.9 4.9 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.1 5.1 5.1 5.1
 [37] 5.1 5.1 5.1 5.1 5.1 5.2 5.2 5.2 5.2 5.3 5.4 5.4 5.4 5.4 5.4 5.4 5.5 5.5
 [55] 5.5 5.5 5.5 5.5 5.5 5.6 5.6 5.6 5.6 5.6 5.6 5.7 5.7 5.7 5.7 5.7 5.7 5.7
 [73] 5.7 5.8 5.8 5.8 5.8 5.8 5.8 5.8 5.9 5.9 5.9 6.0 6.0 6.0 6.0 6.0 6.0 6.1
 [91] 6.1 6.1 6.1 6.1 6.1 6.2 6.2 6.2 6.2 6.3 6.3 6.3 6.3 6.3 6.3 6.3 6.3 6.3
[109] 6.4 6.4 6.4 6.4 6.4 6.4 6.4 6.5 6.5 6.5 6.5 6.5 6.6 6.6 6.7 6.7 6.7 6.7
[127] 6.7 6.7 6.7 6.7 6.8 6.8 6.8 6.9 6.9 6.9 6.9 7.0 7.1 7.2 7.2 7.2 7.3 7.4
[145] 7.6 7.7 7.7 7.7 7.7 7.9
```

Quantile-Quartile

- **Quantile**: In statistics and probability, quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. Ex. 1-point cutting for median, 3-point cutting for quartile, 9-point cutting for decile, and so on. Namely, *q*-**quantiles** are values that partition a finite set of values into *(q+1)* subsets of (nearly) equal sizes.
  - ✓ Quartile: A **quartile** is a type of quantile which divides the sorted number of data points into four more or less equal parts, or quarters. The first quartile (Q_1) is defined as the middle number between the smallest number and the median of the data set. It is also known as the lower quartile or the 25th empirical quartile and it marks where 25% of the data is below or to the left of it (if data is ordered on a timeline from smallest to largest). The second quartile (Q_2) is the median of a data set and 50% of the data lies below this point. The third quartile (Q_3) is the middle value between the median and the highest value of the data set. It is also known as the upper quartile or the 75th empirical quartile and 75% of the data lies below this point.

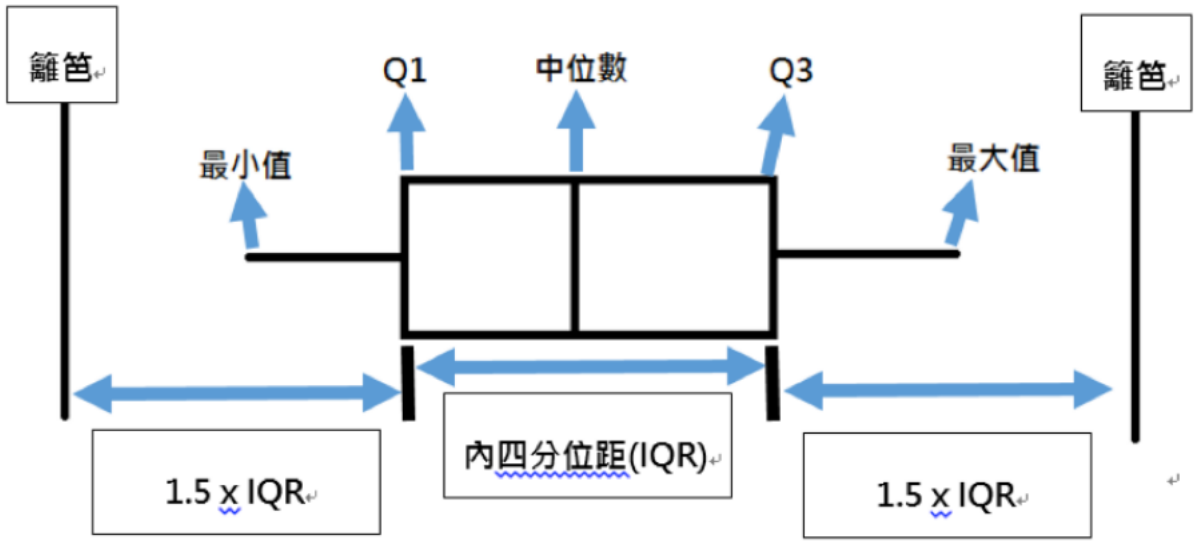
Quantile-Quartile

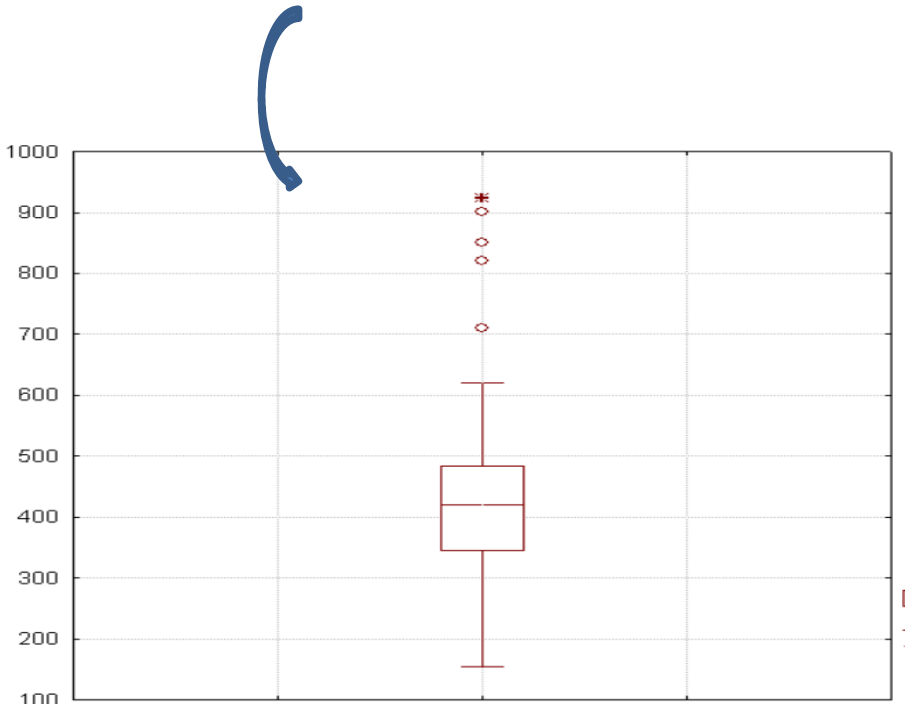
- Quantile: In statistics and probability quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. Ex. 1-point cutting for median, 3-point cutting for quartile, 9-point cutting for decile, and so on. Namely, *q*-**quantiles** are values that partition a finite set of values into (*q+1)* subsets of (nearly) equal sizes.
  - ✓ Quartile- (Q3 –Q1) is called as **IQR** (Interquartile Range). It may be used to characterize the data when there may be extremities that skew the data

Symbol	Names	Definition
Q_1	first quartile lower quartile 25th percentile	splits off the lowest 25% of data from the highest 75%
Q_2	second quartile median 50th percentile	cuts data set in half
Q_3	third quartile upper quartile 75th percentile	splits off the highest 25% of data from the lowest 75%

Quantile-Quartile

- Quantile: In statistics and probability .....
  - ✓ The IQR is often used to find **outliers**（分離物）in data. Outliers (p.s. not acceptable data likely) here are defined as observations that fall below Q1 − 1.5 IQR or above Q3 + 1.5 IQR (p.s. fences).
  - ✓ Example below. Detection of Outliers(離群值的檢測) : Quartile=range (154, 621). range(25%~75%): range(345, 484)=139 (red box). Q3+1.5 IQR=692.5, So, the points for the values greater than 692.5 are called **outliers**. Q3+3 *IQR=901. So, the points for the values greater than 901 are called extreme outliers（極端分離物）.





Quantile-Quartile

- Quantile: In statistics and..... The formula is as follow (where, **p** is percentage):

$$L_p = n \cdot \frac{p}{100}$$

- 情況1：如果L是一個整數，則取第L和第$L+1$的平均值
- 情況2：如果L不是一個整數，則取下一個最近的整數。（比如$L = 1.2$， 則取2）

Quartile

- Quartile:
  - ✓ *ex1. (奇數個數) 6,47,49,15,42,41,7,39,43,40,36. Sorted numbers: 6,7,15,36,39,40,41,42,43,47,49. Q1= 15 (where, 11x 25/100=2.75-choose 3rd), Q2=40 (where, 11x50/100=5.5-choose 6th), Q3=43 (where, 11x75/100=8.25-choose 9th)*
  - ✓ *ex2. (偶數個數) 7,15,36,39,40,41. Q1=15 (where, 6x25/100=1.5-choose 2nd), Q2=37.5(where, 6x50/100=3. Because 3 is an integer, choose (Value-3rd +Value-4th)/2= 37.5) , Q3=40 (where, 6x75/100=4.5-choose 5th)*

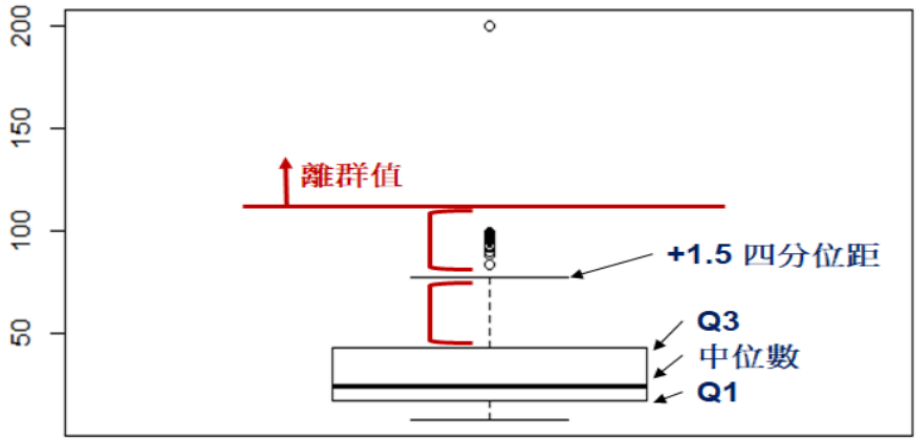
Quartile

- Quantile: In statistics and probability.....
  - Quartile:
    - ✓ *After determining the first and third quartiles and the IQR range as outlined above, then fences are calculated using the following formula:*
    - ✓ *where Q_1 and Q_3 are the first and third quartiles, respectively. The lower fence is the "lower limit" and the upper fence is the "upper limit" of data, and any data lying outside these defined bounds can be considered an **outlier**. Anything below the Lower fence or above the Upper fence can be considered such a case.*（在outlier範圍區的資料, 算是較離散的資料-離散值 on next page)

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR}),$$

Quartile

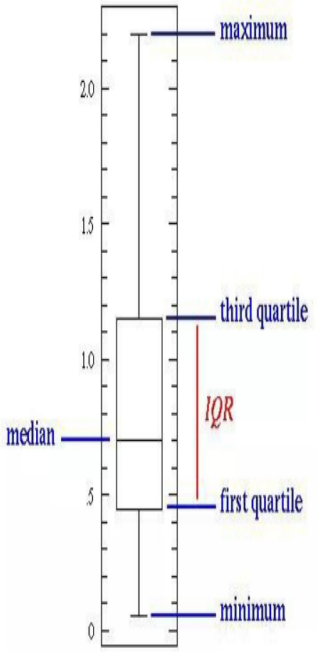
- Quantile: In statistics and probability.....
  - Quartile:
    - ✓ *If the data is in normal distribution, We can remove the data out of Q3 for the unreasonable data. However, if your data is continuous tightly as financial data, please don't remove it in the data cleaning stage.*



箱型圖示意(Box Plot)

Quartile

特徵(描述)統計:特徵統計可能是資料科學中最常用的統計學概念。它是你在研究資料集時，經常使用的統計技術，包括**偏差、方差、四分位數差、平均值、中位數、百分數**等等。理解特徵統計，並且在R程式碼中實現都是非常容易的。例如看下圖(**四分位數差之**箱形圖)：



- **四分位距**（IQR）。是描述統計學中的一種方法，以確定第三個四分位數和第一個四分位數的差值（即Q1與 Q3的差距）
- 當箱形圖很短時，就意味著很多資料點是相似的，因為
  很多值是在一個很小的範圍內分佈;
- 當箱形圖較高時，就意味著大部分的資料點之間的差異很大，
  因為這些值分佈的很廣;
- 如果中位數接近了底部，那麼大部分的資料具有較低的值。
  如果中位數比較接近頂部，那麼大多數的資料具有更高的值。
  基本上，**如果中位線(median)不在框的中間，那麼就表明了是偏斜資料**;
- 如果框上下兩邊的線很長表示資料具有很高的標準偏差和方差，
  意味著這些值被**分散**了，並且**變化非常大**。如果在框的一邊有長線，另一邊的不長，那麼資料可能只在一個方向上變化很大

Exercise

```
> n <- (1:10)
> #標準差
> sd(n)
[1] 3.02765
> #變異數
> var(n)
[1] 9.166667
> sd(n) ^ 2
[1] 9.166667
> #變異係數
> cv <- 100 * sd(n) / mean(n)
> cv
[1] 55.04819
> #全距(最大值減最小值)
> range(n)[2] - range(n)[1]
[1] 9
> #四分位：把資料切分為四等分，中間的三條線就是四分位，Q1=P25,Q2=P50,Q3=75
> Q1 <- quantile(n, 1 / 4)
> Q2 <- quantile(n, 2 / 4)
> Q3 <- quantile(n, 3 / 4)
> Q1
 25%
3.25
> Q2
50%
5.5
> Q3
 75%
7.75
> #IQR = Q3-Q1
> b <- Q3 - Q1 == IQR(n)
> b
 75%
TRUE
> #總結數據(超好用)
> summary(n)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.25    5.50    5.50    7.75   10.00
> #百分位
> quantile(n)
   0%   25%   50%   75%  100%
 1.00  3.25  5.50  7.75 10.00
```

Exercise: Quartile and Quantile

```
> x <- sample(x = 1:200, size = 100, replace = TRUE)
> x
  [1] 186 168  11  36 137  25 195  79  42 178 146  67 141 200  81  38 193
 [18]  48  30  41  74 136  52  90  88 196 100  42  59 174 183 197 128 195
 [35]  43  65 173 126  61  70  15 128  25 196 155 142 104  68  66 118 102
 [52]  72 161 123 125  16 143  55  13 183 134 188 176 100 135  12  95 154
 [69] 143  79 123 178  36 166  93  78 125  17  80  97 178 196 177 166 161
 [86]  61 144 102 119 122  48 191 127 111  24 159 174 179  40 196
> # 複製x
> y <- x
> # 用sample隨機挑選20個元素，把值設為NA
> y[sample(x = 1:200, size = 20, replace = FALSE)] <- NA
> y
  [1] 186  NA  11  36 137  25  NA  79  42 178 146  67 141 200  81  38 193
 [18]  48  30  41  74 136  52  90  88 196 100  42  59 174 183 197 128 195
 [35]  43  65 173  NA  61  NA  NA 128  25 196 155 142 104  68  66 118 102
 [52]  72 161 123 125  16 143  55  NA 183 134 188 176 100  NA  12  95  NA
 [69] 143  79 123 178  36 166  93  78 125  17  80  97 178 196 177 166 161
 [86]  61 144 102 119  NA  48 191 127 111  24 159 174 179  40 196  NA  NA
[103]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
[120]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
[137]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
[154]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
[171]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

```
> # 計算第25和75百分位數
> quantile(x, probs = c(0.25, 0.75))
25% 75%
 64 166
> # 也對y計算同樣的統計量
> quantile(y, probs = c(0.25, 0.75))
Error in quantile.default(y, probs = c(0.25, 0.75)) :
  missing values and NaN's not allowed if 'na.rm' is FALSE
> # 這次設na.rm=TRUE
> quantile(y, probs = c(0.25, 0.75), na.rm = TRUE)
25% 75%
 63 166
> # 計算其它分位數
> quantile(x, probs = c(0.1, 0.25, 0.5, 0.75, 0.99))
   10%    25%    50%    75%    99%
 35.40  64.00 120.50 166.00 197.03
```

Correlation and Covariance

More Properties and Tuples

發票號碼	產品名稱	產品分類	溫度	銷售日期	單價	數量	促銷折扣
N0101	阿里山礦泉水	水	30	2020/6/1	20	2	0
N0201	舒跑	飲料	33	2020/6/1	25	5	0
N0301	可樂	飲料	35	2020/6/5	30	2	85
N0401	椰子餅乾	餅乾	37	2020/7/9	30	3	90
N0302	辣氏泡麵	泡麵	16	2020/12/12	30	3	0
N0909	舒跑	飲料	12	2020/12/27	25	5	0
N0207	椰子餅乾	餅乾	33	2020/6/6	30	1	0

Correlation(相關係數)vs Covariance(共變異數)

- 當我們的資料多於一個變數(屬性)時, 最好的統計量檢查方法是, 是使用相關係數與共變異數, 來檢測變數間的關係
- 相關係數很常用在機器學習或是統計分析上使用，主要衡量兩變數間「線性」關聯性的高低程度。
- 相關係數之計算為最常使用的統計技術。相關性代表兩組數據的符合特性，與兩者的表現現象相互關連。For example, 典型的生醫相關研究, 如血液中血糖濃度與糖化血色素的相關性，生物體年齡與膽固醇程度之關係等。
- 另外, 可能的因素:年齡、早睡、喝牛奶、運動,… ➔ 發生身高長高
  - 問題：在一項人體身高與人體年齡研究中，其相關係數r值為0.97，能否結論人的身高與其年齡大小相關？
  - 解答：雖然r值高於或等於0.97，但是量測誤差與取樣錯誤都有可能導致錯誤結論。此相關性可能來自取樣不周延的緣故。若在嚴謹的試驗條件下，高r值才有可能代表兩者有高相關性。例如以兩只精密的儀器進行血液之生化分析，其量測值則可能高度相關。
- 隨機變數是指變數的值, 無法預先確定. 僅以一定的可能性(機率)取值的量。

Correlation(相關係數)vs Covariance(共變異數)

- 另外,可能的因素:酒精濃度、心臟病、不專心,… ➔發生交通事故
  - 問題：在一項研究中，血液的酒精濃度與交通事故數據之r=0.78，因此能否推導此結論：酒精飲用量引起了交通事故，交通事故是飲用酒精之結果。
  - 解答：相關係數無法提供因果關係。酒精濃度與交通事故有高的相關係數，但是不能得到結論是相互影響。因為交通事故的原因包括道路狀況、駕駛員技術、駕駛身體其他疾病、有否服用藥物等。
- 另外,可能的因素:酵素濃度r 、酵素濃度R^2 ➔發生結晶反應
  - 問題：比較血液中酵素濃度與其結晶反應，其r=0.52，p=0.002。結論是否是酵素濃度影響了52%的結晶反應？
  - 解答：相關係數r並不能用以描述相關性強度。r=0.52，不能用以解釋52%的相關性，或是兩者有52%的關聯性。在迴歸分析中，決定係數 R^2（ps.代表一個回歸模式對於反應值，以此來判斷統計模型的解釋力）以R^2代表，而只有在Pearson's相關計算r才可應用。因此r=0.52, R^2=0.27。,因此若使用迴歸分析,代表有27%的相關性。

Correlation（相關係數）vsCovariance（共變異數）

- **共變異數**（Covariance）用於衡量兩個隨機變數的聯合變化程度。而變異數是共變異數的一種特殊情況
- 共變異數表示的是兩個變數的母體的誤差，這與只表示一個變數誤差的變異數不同。 如果兩個變數的變化趨勢一致，也就是說如果其中一個大於自身的期望值，另外一個也大於自身的期望值，那麼兩個變數之間的共變異數就是正值。 如果兩個變數的變化趨勢相反，即其中一個大於自身的期望值，另外一個卻小於自身的期望值，那麼兩個變數之間的共變異數就是負值。
- 期望值分別為$E(X)=\mu$與$E(Y)=\nu$的兩個具有有限二階矩的實數隨機變數X與Y之間的**共變異數**定義為：

$$\operatorname{cov}(X,Y)=\operatorname{E}((X-\mu)(Y-\nu))=\operatorname{E}(X\cdot Y)-\mu\nu.$$

- 如果X 與Y 是統計獨立的，那麼二者之間的共變異數就是0，這是因為

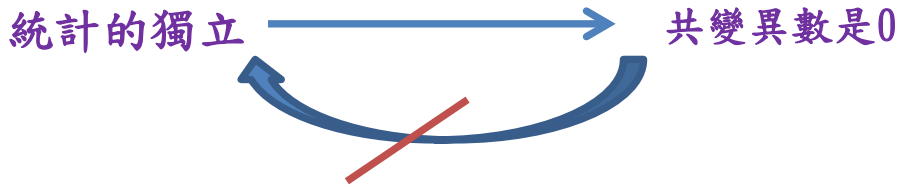
$$E(X\cdot Y)=E(X)\cdot E(Y)=\mu\nu,$$

p.s. 所謂統計的獨立，或稱隨機的獨立. 若要計算二獨立事件交集的機率, 我們只需將二事件個別的機率相乘即可。

但是反過來並不成立，即如果X與Y的共變異數為0，二者並不一定是統計獨立的。

取決於共變異數的相關性η

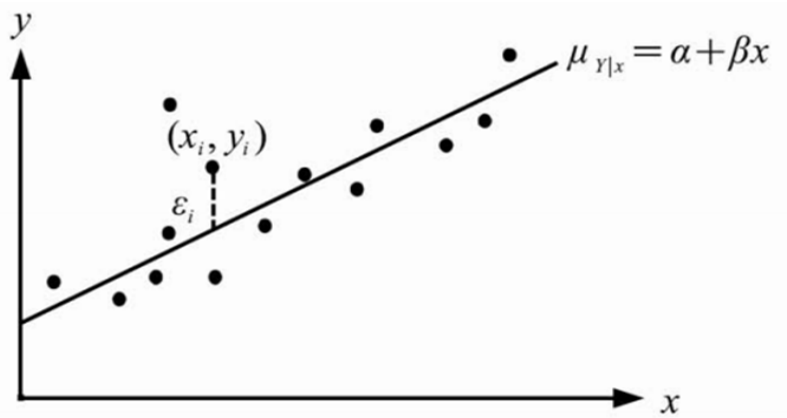
$$\eta=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\cdot\operatorname{var}(Y)}},$$



更準確地說是取決於線性相依性

Correlation（相關係數）vsCovariance（共變異數）

- 兩個隨機變數可能獨立也可能不獨立。所謂獨立，是隨機式的獨立，表知其中之一的值，對另一變數的值毫無影響。如果二隨機變數不獨立，則二變數間便有關係。此關係可能強可能弱。
- 隨機變數之關係的確是有強弱的。如果是獨立，則關係當然是最弱的。 若樣本為水，令 X表其體積，Y表其重量。 顯然X與Y之關係很強。如果取樣多次，將所得的(X,Y)數據畫在座標平面上，則所有的點很可能落在一直線上或直線的附近。這是因為水的重量與體積有線性關係。但是，有些數據不在直線上，可能是因量測的誤差，或水質不純所致。其次，若以X表某人之身高，Y表其體重，X與Y顯然亦有關係，但可能不是那麼強。 共變異數及相關係數，都可用來度量兩隨機變數關係(特別是線性關係)的強弱。
- 相關係數作為一個指標，主要是反應二隨機變數之分佈的線性關係之強度及符號。因此相關係數為0，僅表示二隨機變數之線性關係很低，而非表示二變數機率上無關(即獨立)。
- 而變異數是共變異數的一種特殊情況，即當兩個變量是相同的情況。如果X 與Y是統計獨立的，那麼二者之間的共變異數就是0



Correlation v.s. Covariance

- Correlation（相關係數）與 Covariance（共變異數）：共變異數（又稱協方差）及相關係數（correlation coefficient，或只稱correlation），都可用來度量兩隨機變數關係（特別是線性關係）的強弱
- 一般說的相關係數通常是指「皮爾森相關係數（Pearson's correlation coefficient）」，兩組樣本之間的相關程度，其值介於-1與1之間。但當變數之間是順序尺度（ordinal）時用的則是「斯皮爾曼等級相關係數（Spearman's rank correlation coefficient）」
- 相關係數很常用在機器學習或是統計分析上使用，主要衡量兩變數間「線性」關聯性的高低程度。探討兩個變數（或多變數）間是否存在「線性」關係. 可將線性關係以方程式表示：Linear Regression
- 皮爾森相關係數（Pearson's correlation coefficient）ρ假設有兩個變數(x_i, y_i)

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\sum_{i=1}^{n}(y_i-\mu_y)^2}}$$

$$\rho = \frac{x和y的共變異數}{x的標準差 \times y的標準差}$$

$$\text{共變異數(covariance): } \operatorname{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$$

$$\text{變異數(variance): } \operatorname{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)^2$$

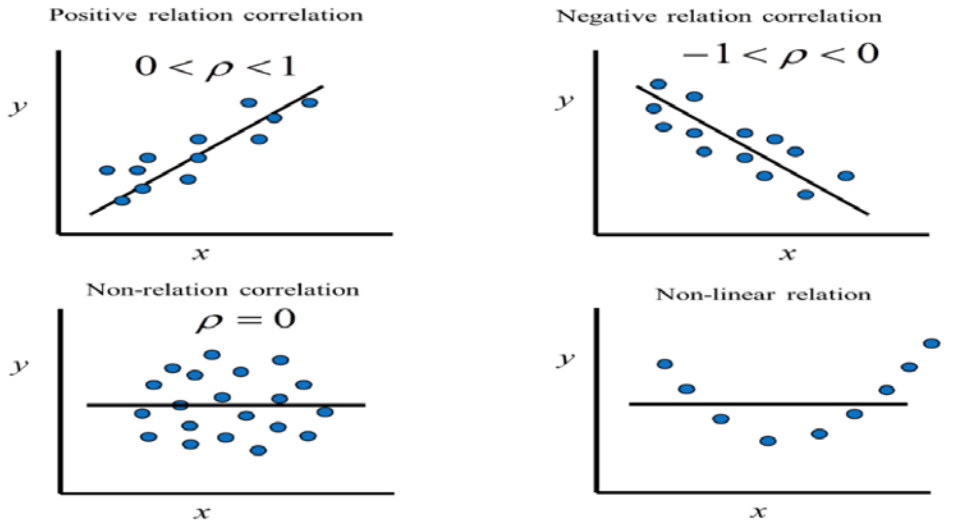
$$\text{標準差(standard deviation): } std(x) = \sqrt{\operatorname{var}(x)}$$

Correlation v.s. Covariance

- 我們從公式看，我們將*x*和*y*減去各自的平均數後相乘最後算總和，這邊如果我們假設變數*x*等於變數*y*時，那這個共變異數不就等於變異數，這時候 ρ 的值不就等於1，所以*x*和y就完全相關 （p. s. *x*=y）。
- 所以共變異數其實就等於在算x和y的相關程度，但此時的相關還是相依在*x*和*y*的尺度上. 如, 假設*x*變數是身高(單位:公分)，y變數是體重(單位:公斤)，*x*的標準差計算出的單位是公分，*y*的標準差計算出的單位是公斤。那共變異數算出來的單位是什麼呢?答案是「公分*公斤」
- 所以今天如果是比較「身高和體重」的相關度高，還是「身高和年齡」的相關度高，如果只看共變異數就是在一個不同單位下比較的方法，很不公平。所以在共變異數我們會除上兩個變數間的標準差，如此一來單位也都被消除了，大家都回到同一個基準線上，值會落在正負1之間。如左下. 右下圖示說明不同範圍內的 ρ 值(皮爾森相關係數)呈現出來的結果

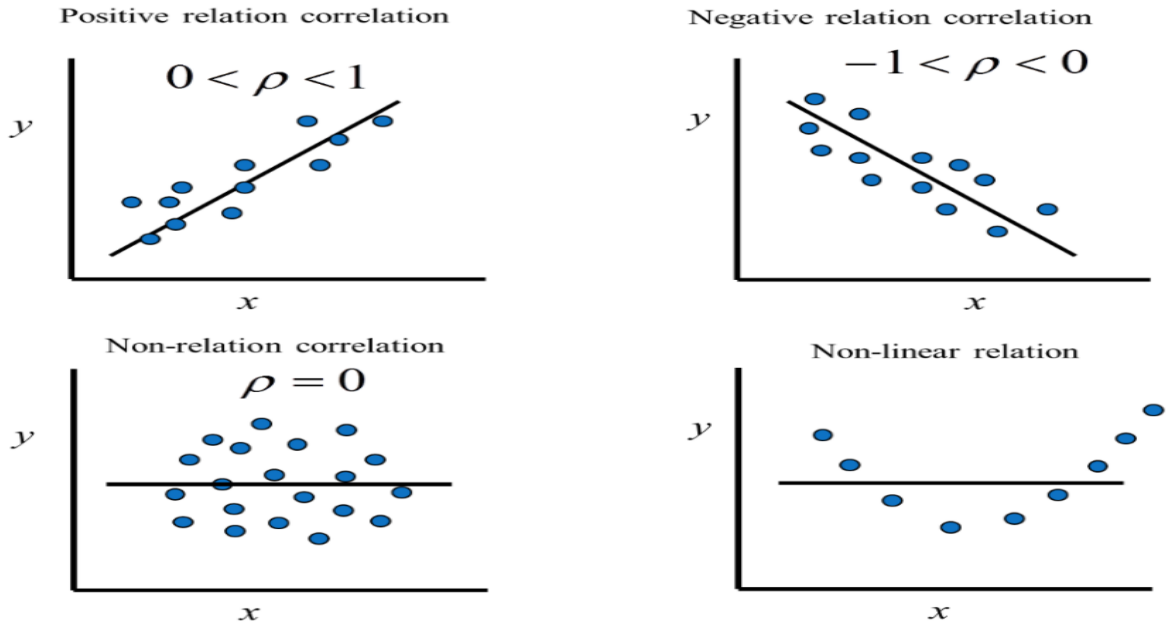
$$\rho = \frac{cov(x,y)\,(單位:公分*公斤)}{std(x)(單位:公分)\times std(y)(單位:公斤)}$$

$$= \frac{cov(x,y)\,(單位:公分*公斤)}{std(x)(單位:公分)\times std(y)(單位:公斤)}$$



Correlation v.s. Covariance

- 「共變異數」是指二個變項離差分數交叉乘積(cross-product)總和之平均數，共變異數可以用來表示二個變項間關聯的強度及方向，當二個變項之相關為零時，其共變異數為零；若相關為正，則共變數大於零；若相關為負，則共變數小於零。但共變數的大小會因測量單位的不同而有所差異，二個相同的變項，以不同測量單位加以表示時，其共變數就會產生極大的差異，因為共變異數有此一特性，所以在表示變項間關係時，習慣上常用相關係數(correlation coefficient)加以表示。



Example: Correlation v.s. Covariance

要評估兩個資產(assets)報酬率的相關性時，就需要計算它們的共變異數和相關係數。

舉例來說，有A和B兩個資產，過去六個月的報酬率分別是

A	B
5.4	3.4
-4.2	-1.2
10.4	7.5
11.2	5.2
-7.8	3
-2	-5

就是將A、B兩者每個月的報酬率和平均報酬率的差相乘，然後除以5。(ps因為是用歷史報酬，要除以n-1)

A	B	A-A 平均	B-B 平均	(A-A 平均)*(B-B 平均)
5.4	3.4	3.23	1.25	4.04
-4.2	-1.2	-6.37	-3.35	21.33
10.4	7.5	8.23	5.35	44.05
11.2	5.2	9.03	3.05	27.55
-7.8	3	-9.97	0.85	-8.47
-2	-5	-4.17	-7.15	29.79
A 平均 2.17	B 平均 2.15			總合 118.29

然後將118.29除以5，等於23.66，這就是A、B兩者的共變異數(Covariance)。

由過去六個月的報酬，可以算出A的平均月報酬是2.17%，B是2.15%(算術平均數)。但可以明顯看出A和B不是齊漲齊跌的。有時候會有一個上揚，另一個卻虧損(第五個月份)。共變異數(Covariance)量測的是，這兩者的月報酬向同一方向偏離各自平均報酬的程度。它的公式為

$$Cov = \frac{1}{n-1} * \sum_{i=1}^{n} (Rai - Raave)*(Rbi - Rbave)$$

Rai 代表A資產各月份的報酬率
Rbi 代表B資產各月份的報酬率
Raave 表A 資產的平均報酬率
Rbave 表B資產的平均報酬率

p.s. use (n-1) because of Bessel's correction

Example: Correlation v.s. Covariance

在計算過程中可以發現，假如AB兩者某個月的報酬都是同時高於平均或同時低於平均，這樣兩者報酬率和平均報酬率的差相乘後會得到一個正數，會讓共變異數的數值變大。假如AB兩者某月份報酬，一個高過平均，一個低於平均，那麼兩者報酬率與平均的差相乘會得到一個負數，會減小共變異數的數值，或使其成為負數。

但問題來了，AB兩資產報酬的共變異數23.66，這樣兩者的報酬率相關性到底如何呢?假如兩者的報酬率都是穩定的，這樣23.66會代表比較高的相關程度，假如兩者波動性都很大，那麼共變異數23.66其實代表的相關程度不高。為了解決這個問題，我們可以把共變異數除以AB兩者報酬率的標準差，就會得到相關係數(Correlation coefficient)。

由共變異數運算, 推展到相關係數運算

算出A的標準差是7.96，B的標準差為4.53。兩者的相關係數 等於23.66/(7.96*4.53)=0.66。

相關係數會是一個介於-1到+1間的值。+1代表兩者完全正相關，-1表完全負相關。在資產配置理論裡，相關程度愈低的資產，將發揮愈大的"互補"功效。

Correlation v.s. Covariance in R

- Check the relationship of the 2 properties **pce** and **psavert** from the dataset **econmics** via **cor()** api. The results come to high correlation in a **negative side**. By the way, a value of 0 for correlation indicates that there is no relationship between the two variables. However, the value closed to 1 means higher correlation in a positive side.

```
> library(ggplot2)
> head(economics)
# A tibble: 6 x 6
  date         pce    pop psavert uempmed unemploy
  <date>     <dbl>  <dbl>   <dbl>   <dbl>    <dbl>
1 1967-07-01  507. 198712    12.6     4.5     2944
2 1967-08-01  510. 198911    12.6     4.7     2945
3 1967-09-01  516. 199113    11.9     4.6     2958
4 1967-10-01  512. 199311    12.9     4.9     3143
5 1967-11-01  517. 199498    12.8     4.7     3066
6 1967-12-01  525. 199657    11.8     4.8     3018
> ?economics
> cor(economics$pce, economics$psavert)
[1] -0.7928546
```

Format

A data frame with 574 rows and 6 variables:

date

Month of data collection

pce

personal consumption expenditures, in billions of dollars, http://research.stlouisfed.org/fred2/series/PCE

pop

total population, in thousands, http://research.stlouisfed.org/fred2/series/POP

psavert

personal savings rate, http://research.stlouisfed.org/fred2/series/PSAVERT/

Correlation v.s. Covariance in R

- Check the relationship of the 2 properties from the dataset **econmics** via a set of statistics instead of **cor()** api.

```
> cor(economics$pce, economics$psavert)
[1] -0.7928546
> ## 計算用來找出相關係數的每個部份
> xPart <- economics$pce - mean(economics$pce)          # mean
> yPart <- economics$psavert - mean(economics$psavert) # mean
> nMinusOne <- (nrow(economics) - 1)
> xSD <- sd(economics$pce)     #Standard Deviation
> ySD <- sd(economics$psavert) #Standard Deviation
>
> # 應用相關係數的公式
> sum(xPart * yPart) / (nMinusOne * xSD * ySD)
[1] -0.7928546
```

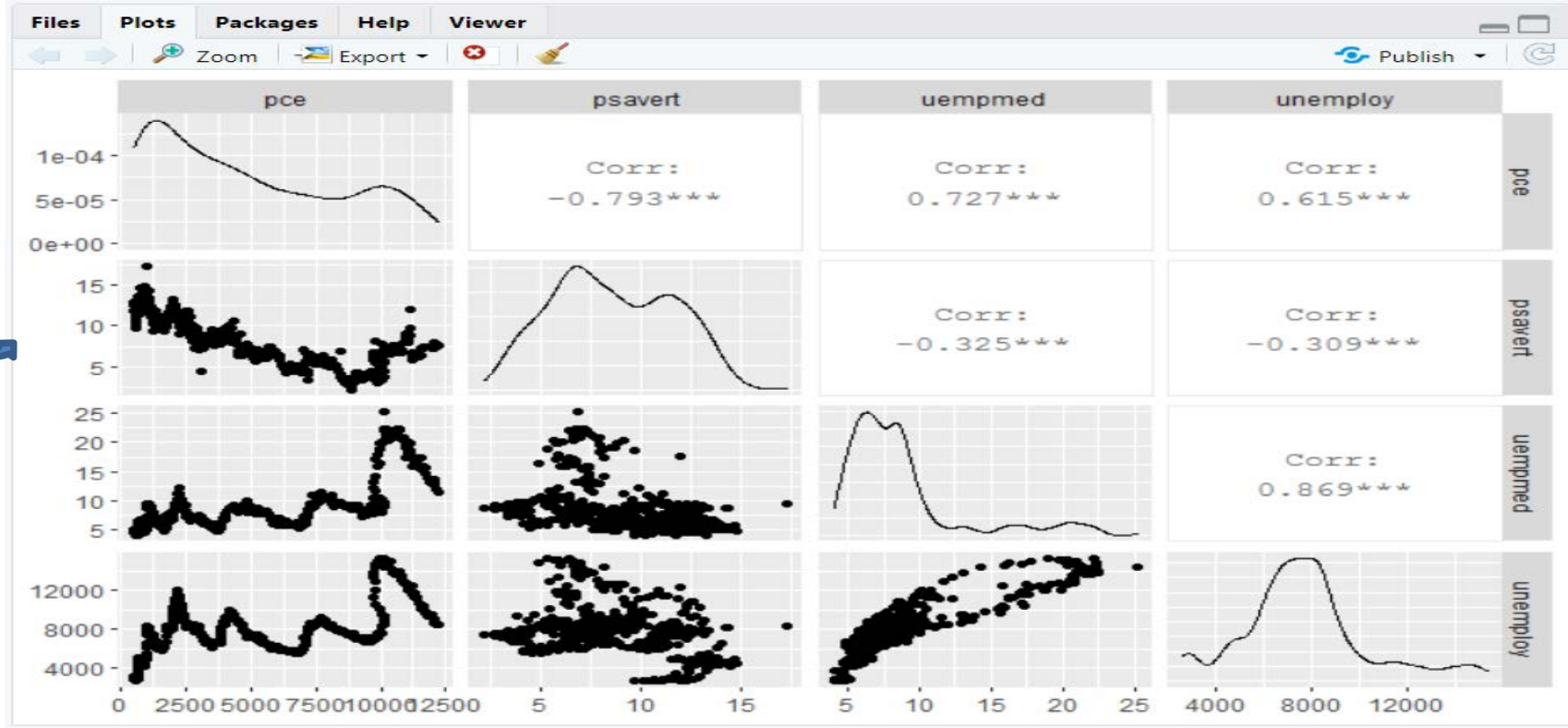
$$r_{xy} \overset{\text{def}}{=} \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Correlation v.s. Covariance in R

- Check the relationship of more than 2 properties from the dataset **econmics** via **cor()** api and ggpairs. Due to the namespace problem, *Ggally:ggpairs* has been used instead of *ggpairs*.

```
> cor(economics[, c(2, 4:6)])
                pce    psavert    uempmed   unemploy
pce       1.0000000 -0.7928546  0.7269616  0.6145176
psavert  -0.7928546  1.0000000 -0.3251377 -0.3093769
uempmed   0.7269616 -0.3251377  1.0000000  0.8693097
unemploy  0.6145176 -0.3093769  0.8693097  1.0000000
```

```
#
install.packages("GGally")
library(GGally)
GGally::ggpairs(economics[, c(2, 4:6)])
```



Correlation v.s. Covariance in R

- 因在統計上，共變異數是一個類似相關係數的統計量．而這統計量，就像變數之間的變異數．見公式如下：

皮爾森相關係數(Pearson's correlation coefficient)ρ假設有兩個變數(x_i, y_i)

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\sum_{i=1}^{n}(y_i-\mu_y)^2}}$$

$$\rho = \frac{x\text{和}y\text{的共變異數}}{x\text{的標準差}\times y\text{的標準差}}$$

共變異數(covariance): $\text{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$

變異數(variance): $\text{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)^2$

標準差(standard deviation): $std(x) = \sqrt{\text{var}(x)}$

Correlation v.s. Covariance in R

```
> cov(economics$pce, economics$psavert)
[1] -8359.069
> cov(economics[, c(2, 4:6)])
                  pce      psavert      uempmed    unemploy
pce      12650851.944 -8359.069071 10618.386190 5774578.978
psavert     -8359.069     8.786360    -3.957847   -2422.805
uempmed     10618.386    -3.957847    16.864531    9431.652
unemploy  5774578.978 -2422.805358  9431.652268 6979948.309
```

Probability Distribution

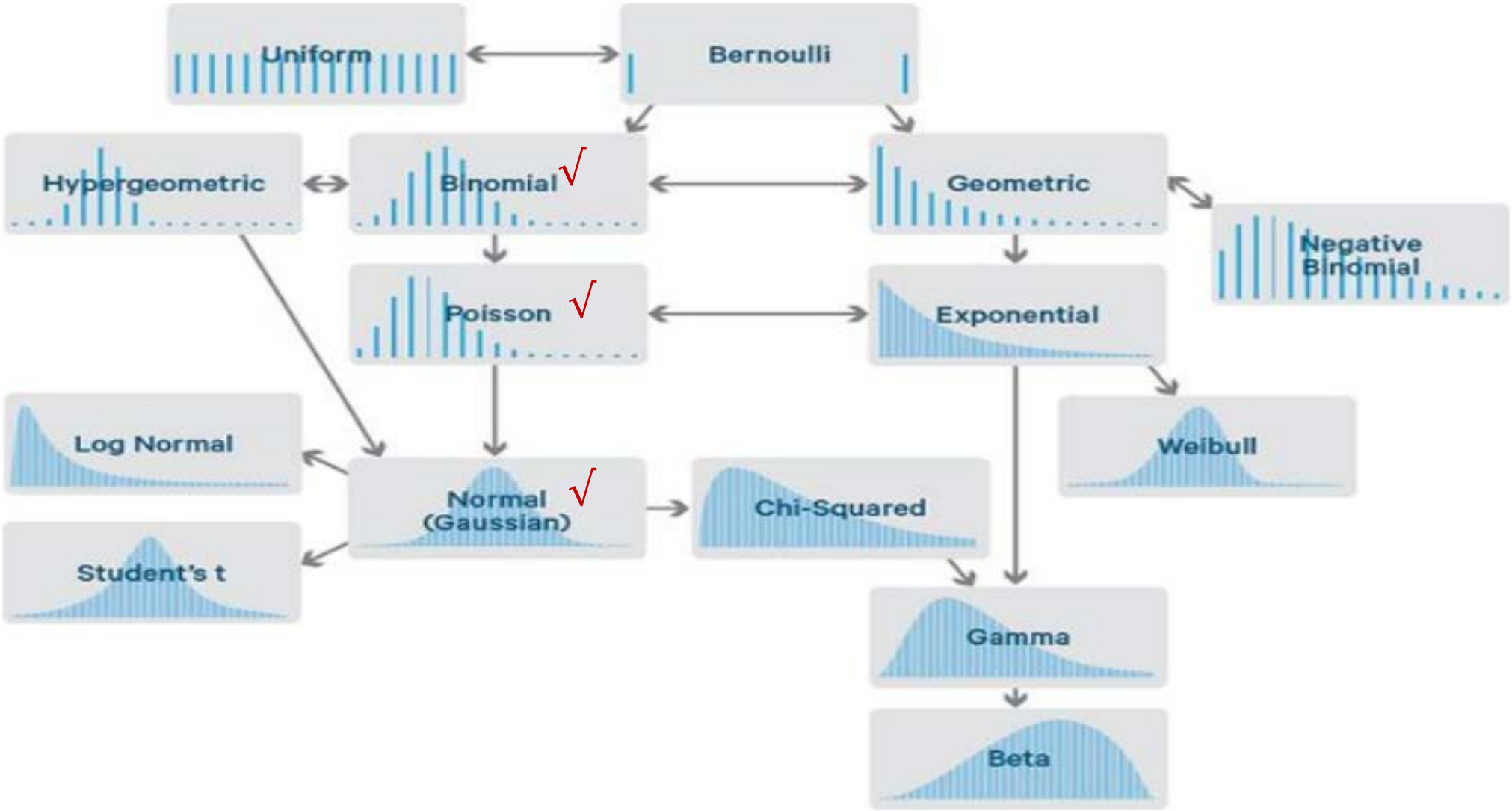
- Normal Distribution -連續分佈
- Binomial Distribution-離散分布
- Poisson Distribution -離散分布

Why Probability Distribution

- 隨機變數是指變數的值, 無法預先確定. 僅以一定的可能性(機率)取值的量。
- Probability distribution: 機率分佈
- 機率分佈是隨機變數的機率函數，其定義域是實數值，值域是機率([0,1])，機率分佈會因隨機變數之性質(Numerical Type: 連續性數值或間斷性數值)分為：連續型分佈以及離散型分佈。
- 以『巧婦難為無米之炊』而言，資料/ 數據分析的『主料』即為資料。當我們對一組資料作分析的時候，一定要明確的是，這組資料只是研究對象（population）中的一部分樣本（sample）。我們只是對一部分樣本進行分析，然後再去推測出整個對象的規律。
- 一般來說, 數據分析中，數據量越多，樣本越大，結果越準確。**那有人會問，既然這樣，為什麼不搜集海量的數據呢？**大部分的工作只是為了找到一個近似的規律，而且過大的數據量會帶來收集費用的飆升、處理難度和時間的增加。因此，數據處理第一步，我們要試著去平衡數據量和處理的耗費（金錢與時間）。(1. 以非巨量資料中估算整體資料的機率分佈.　2. 以機率分布來補足因資料清除後, 所照成的資料不足現象.)
- 數據類型大體分為兩種：數值（如房價）和類別（如品牌，姓名等）。 而數值型數據可細分為離散（不連續）和連續數據。機率分布可以很好的展現數據的內在規律，圖中就總結歸納了大部分的機率分布類型。

Why Probability Distribution

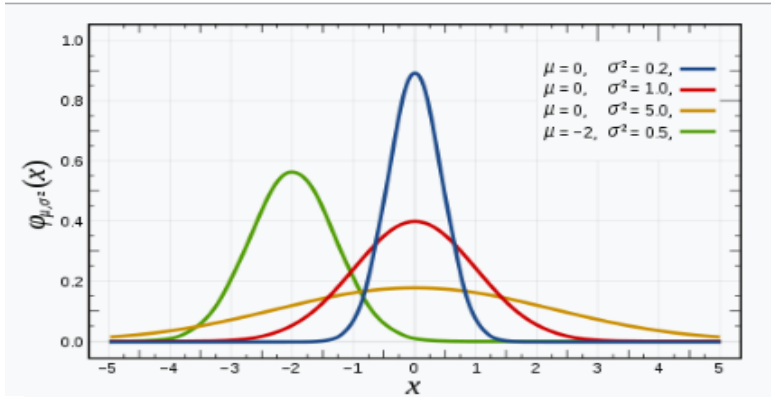
- 大部分的機率分布類型



Normal Distribution

- 常態分布（Normal distribution），是一個非常常見的連續機率分佈。常態分布在統計學上十分重要，經常用在自然和社會科學來代表一個不明的隨機變量。

常態分布



若隨機變量X服從一個位置參數為μ、尺度參數為σ的常態分布，記為：

$$X \sim N(\mu, \sigma^2)$$

則其機率密度函數為

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

常態分布的數學期望值或期望值μ等於位置參數，決定了分布的位置；其變異數σ^2的開平方或標準差σ等於尺度參數，決定了分布的幅度。

- 常態分布的機率密度函數曲線呈鐘形，因此人們又經常稱之為**鐘形曲線**（類似於寺廟裡的大鐘，因此得名）。我們通常所說的**標準常態分布**是位置參數$\mu = 0$，尺度參數$\sigma^2 = 1$的常態分布 （見右圖中紅色曲線）。

p. s. 期望值:為出現各個機率的總和

Normal Distribution -期望值 (Expected value)

在處理有關財務風險的事務時，不免要衡量可能的得與失，「數學期望值」的觀念在此時就顯得特別重要，可以幫助我們思考及判斷出最佳的決策。其定義如下：

若隨機變數 X 的機率分布如下表：

X	x_1	x_2	$\cdots$	x_n
P	p_1	p_2	$\cdots$	p_n

則稱 $E(X) = x_1p_1 + x_2p_2 + \cdots + x_np_n = \sum_{i=1}^{n} x_ip_i$ 為隨機變數 X 的數學期望值。

$$100 \times \frac{1}{10} + 50 \times \frac{4}{10} + 30 \times \frac{5}{10} = 45(\text{元})$$

舉例來說，參加一摸彩遊戲，抽獎箱裡放置了三種分別標有獎金100元、50元、30元的彩券共10張，其中100元有1張、50元有4張和30元有5張。參賽者繳交55元可摸取一張彩券，假設每張彩券被抽出的機率都相同。那麼，平均每張彩券的獎金是：

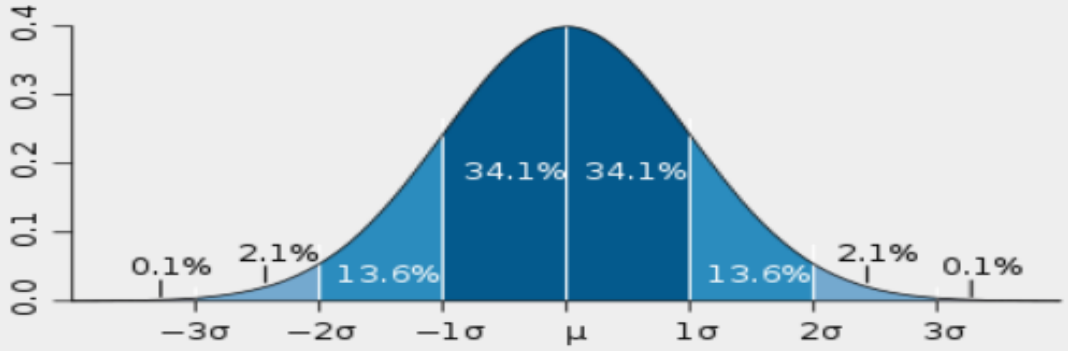
不過，參加一次遊戲就要先付55元，雖然獎金期望值有45元，但參賽者每玩一次預期要虧損10元。

Normal Distribution

- In a random collection of data from independent sources, it is generally observed that the distribution of data is normal.
- Which means, on plotting a graph with the value of the variable in the horizontal axis and the count of the values in the vertical axis we get a bell shape curve.
- The center of the curve represents the mean of the data set. In the graph, fifty percent of values lie to the left of the mean and the other fifty percent lie to the right of the graph. This is referred as normal distribution in statistics.

常態分佈(高斯分佈)：將一連續變項之觀察值發生機率以圖呈現其分布情形，且具有以下特性：

1. 以平均數為中線，構成左右對稱之單峰、鐘型曲線分布。
2. 觀察值之範圍為負無限大至正無限大之間。
3. 變項之平均數、中位數和眾數為同一數值。
4. 標準偏差(standard deviation)：
   68.3%的數值，落在平均數 ± 1個標準差間；
   95.4%的數值，落在平均數 ± 2個標準差間；
   99.7%的數值，落在平均數 ± 3個標準差間。



Normal Distribution

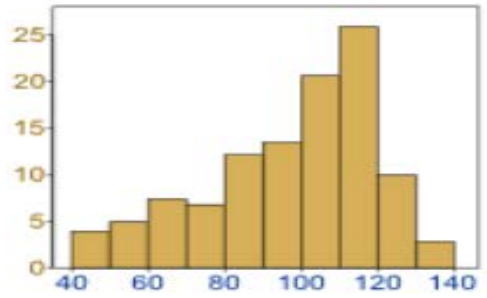
- 有誤差的對稱性分布形式，中央高兩邊低的形式
- 預設常態分布為**標準常態分布**，平均值為 0，標準差為 1
  - **標準差**是一組數值自平均值分散開來的程度的一種測量觀念
  - 一個較大的**標準差**，代表大部分的數值和其平均值之間差異較大；一個較小的**標準差**，代表這些數值較接近平均值。 例如，兩組數的集合{0, 5, 9, 14} 和{5, 6, 8, 9} 其平均值都是7 ，但第二個集合具有較小的**標準差**。 ← Rough description
  - 標準差:變異數σ^2的平方根
  - 補充：要了解一個族群最有用的兩個資訊是平均數μ與變異數σ^2。變異數為對數據的變異程度的衡量，常用來量測**資料分散程度**之指標值，變異數其定義為:每一個觀測值和平均值之間的偏差值的平方值的平均。
  - 兩個族群平均相同，變異數可以很不同，分布可能差異很大；分布可能很不一樣。
- 常態分布是一個非常常見的連續機率分布。
- 常態分布是自然科學與行為科學中的定量現象的一個方便模型。

Normal Distribution

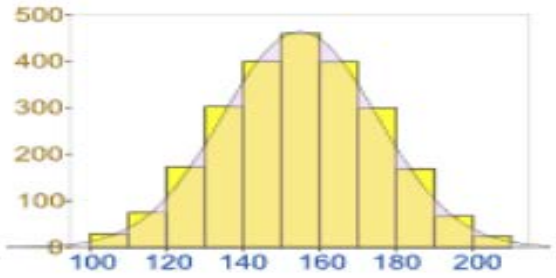
- 應用:常態曲線及分配是一種理論模式，但透過這理論模式，配合平均數及標準差，我們可以對實證研究所得之資料分配，做相當精確之描述及推論。能做到這一點是因常態曲線本身有些重要且已知的特性。
- 常態曲線最重要的特性是：
- **其形狀為左右對稱若鐘形之曲線**。
  - 注意：對稱不一定為常態分布，但常態分布一定為對稱
  - 此曲線只有一個眾數，並與中位數及平均數是三合一的。
  - 其曲線的兩尾是向兩端無限延伸。
  - 曲線之形狀**完全由** μ、σ^2**決定**。
- 因此，雖然實際調查得到的資料，不可能是這種完美的理論模式，但許多實際得到之變項的資料分配是相當接近這種模式，因此可以假定它們的分配是常態的，進而使我們得以運用常態曲線的理論特性。

Normal Distribution & Skeness

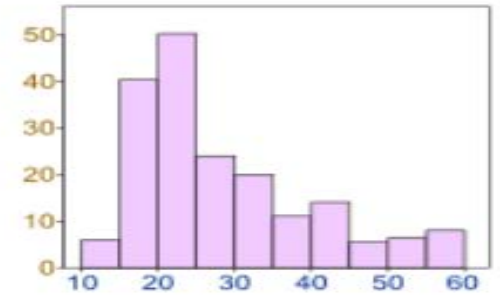
- From Normal Distribution to left/ right Skewnesses . So, do your analysis while using Quartile for your data analysis as compared with mean, median, and mode.



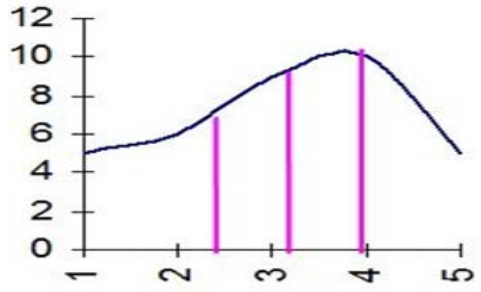
Negative Skew



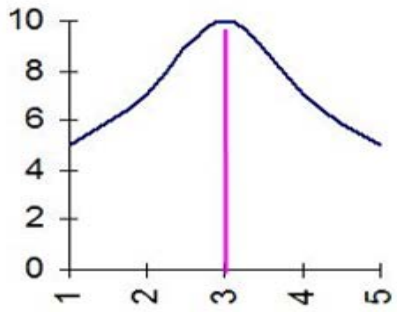
No Skew



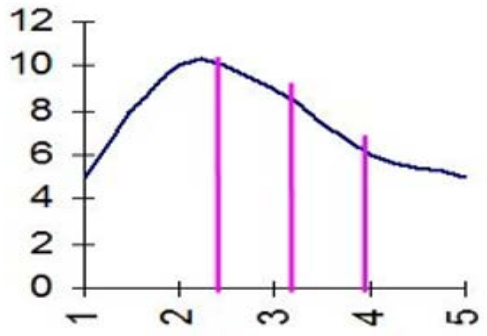
Positive Skew



Mean< Med<Mode



Mean = Median = Mode



Mode < Med< Mean

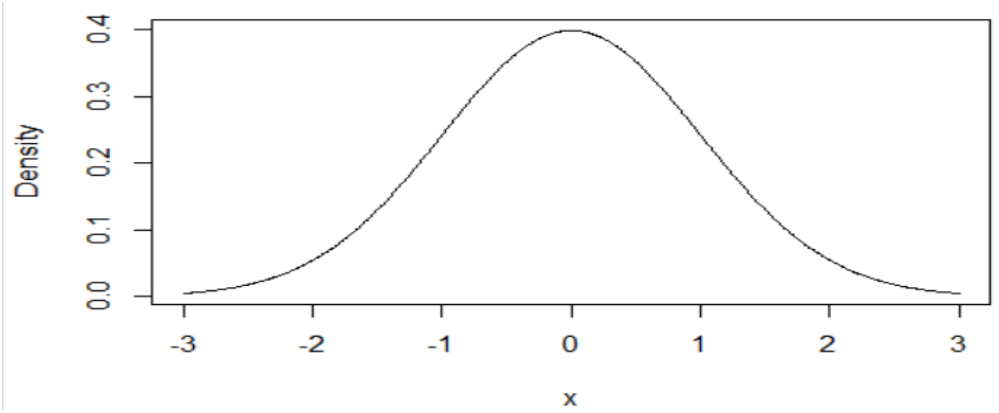
Normal Distribution in R

- R has four built-in functions to generate normal distribution. They are described below.
  - dnorm(x, mean, sd)
  - pnorm(x, mean, sd)
  - qnorm(p, mean, sd)
  - rnorm(n, mean, sd)
- Following is the description of the parameters used in above functions –
  - **x** is a vector of numbers.
  - **p** is a vector of probabilities.
  - **n** is number of observations(sample size).
  - **mean** is the mean value of the sample data. It's default value is zero.
  - **sd** is the standard deviation. It's default value is 1.

dnorm() without mean and standard deviation

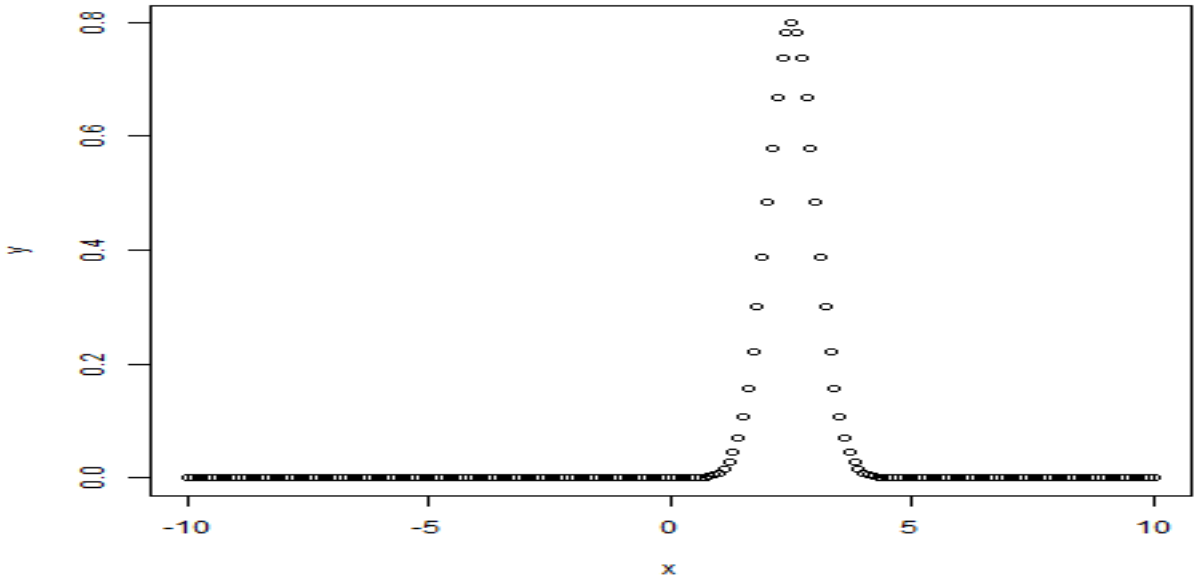
- This function gives height of the probability distribution at each point for a given mean and standard deviation. (回傳常態分布的**機率密度值;**機率密度函數f(x)-是一個描述這個隨機變量的輸出值，在某個確定的取值點x附近的可能性的函數f(x))

```
x <- seq(from = -3, to = 3, by = 0.01)
y <- dnorm(x)
plot(x, y, type = "l", ylab = "Density")
```



dnorm() with mean and standard deviation

```
x <- seq(-10, 10, by = .1)
y <- dnorm(x, mean = 2.5, sd = 0.5)
png(file = "dnorm.png")
plot(x,y)
# Save the file.
dev.off()
```



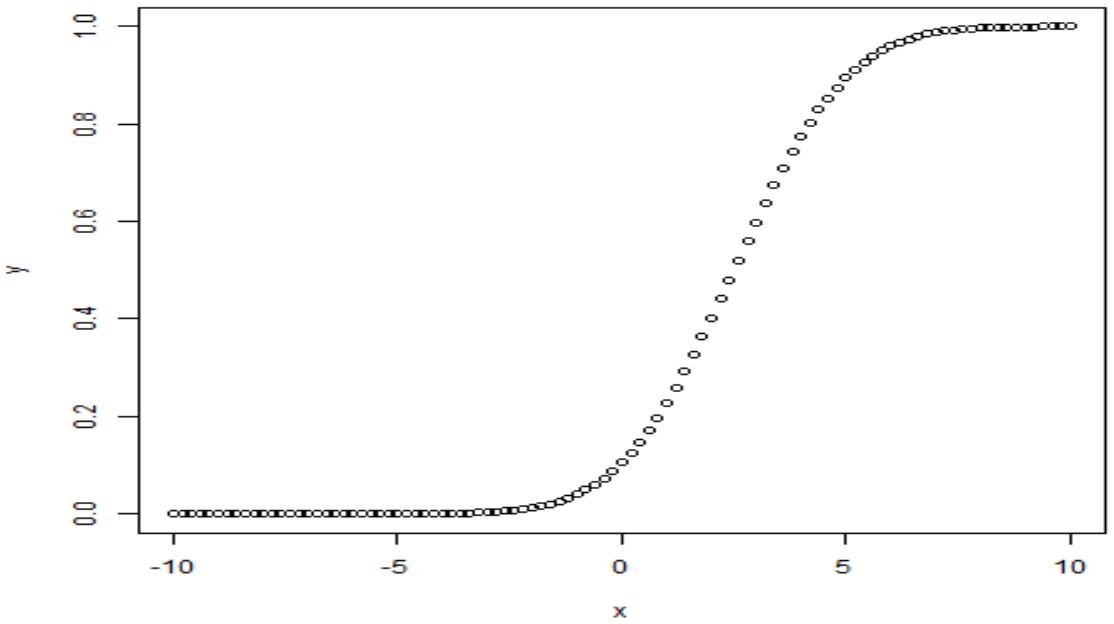
pnorm()

- This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".(回傳對應輸入的常態**累積分布機率值**;累積分布函數F(x)，又叫分布函數，是機率密度函數的積分，能完整描述一個實隨機變量X的機率分布)

```
> pnorm(1.96)
[1] 0.9750021
```

pnorm()

```
> x <- seq(-10,10,by = .2)
> y <- pnorm(x, mean = 2.5, sd = 2)
> png(file = "pnorm.png")
> plot(x,y)
> dev.off()
RStudioGD
        2
```



p. s. pnorm() 該函數給出常態分佈隨機數小於給定數值的機率。
它也被稱爲「累積分佈函數」。面積回推機率

qnorm()

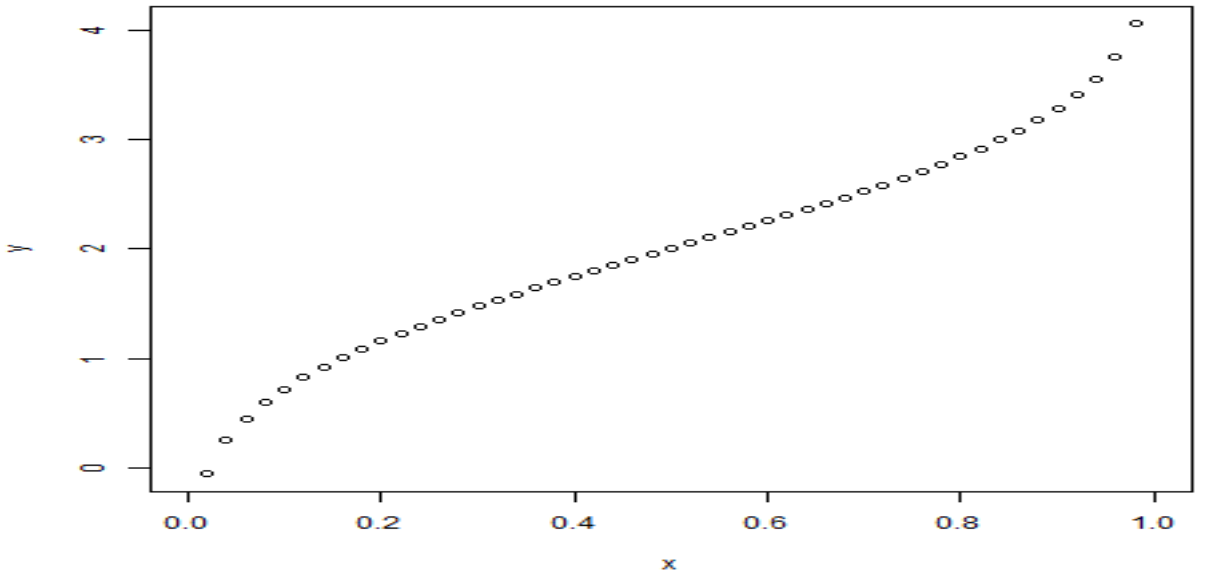
- This function takes the probability value and gives a number whose cumulative value matches the probability value.(回傳對應累積機率值的常態分布輸入;分布函數的反函數，即給定機率p後，求其下分位點.分位數（Quantile），亦稱分位點，是指用分割點（cut point）將一個**隨機變量**的機率分布範圍分為**幾個具有相同機率**的**連續區間**。分割點的數量比劃分出的區間少1，例如3個分割點能分出4個區間)

```
> qnorm(0.975)
[1] 1.959964
```

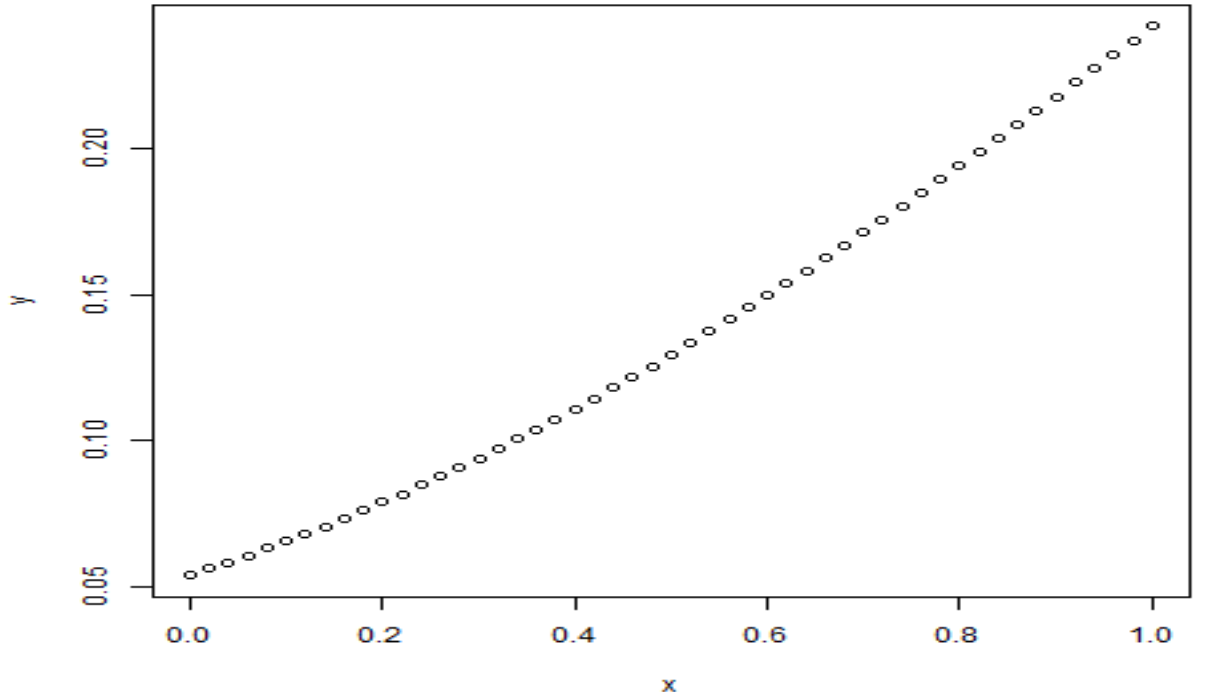
p.s. 機率回推面積

qnorm()

```
x <- seq(0, 1, by = 0.02)
y <- qnorm(x, mean = 2, sd = 1)
png(file = "qnorm.png")
plot(x,y)
dev.off()
```



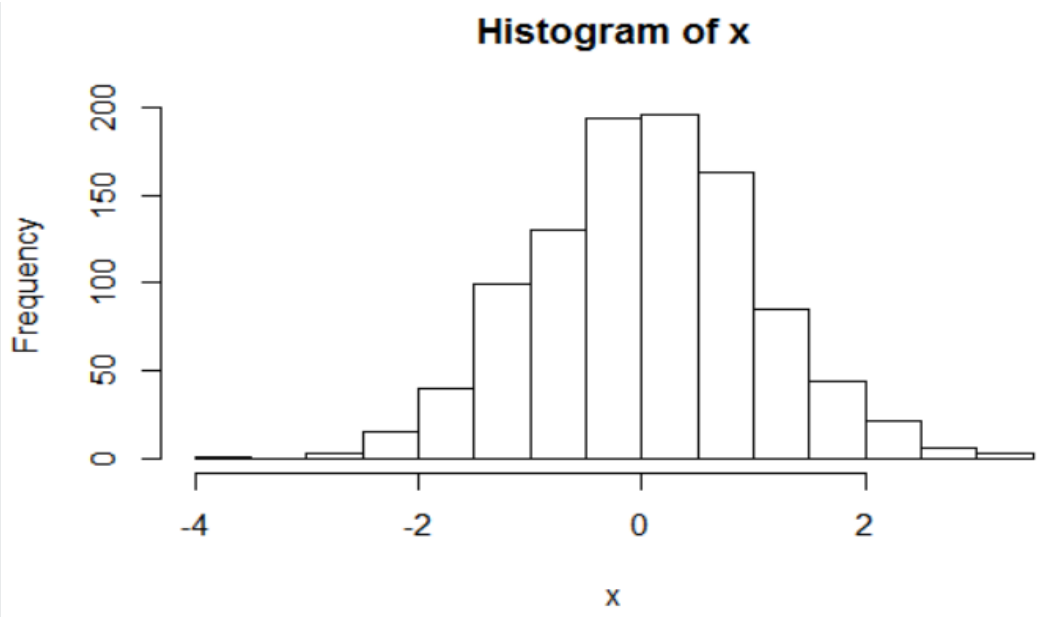

```
x <- seq(0, 1, by = 0.02)
y <- dnorm(x, mean = 2, sd = 1)
png(file = "dnorm.png")
plot(x,y)
dev.off()
```



rnorm()

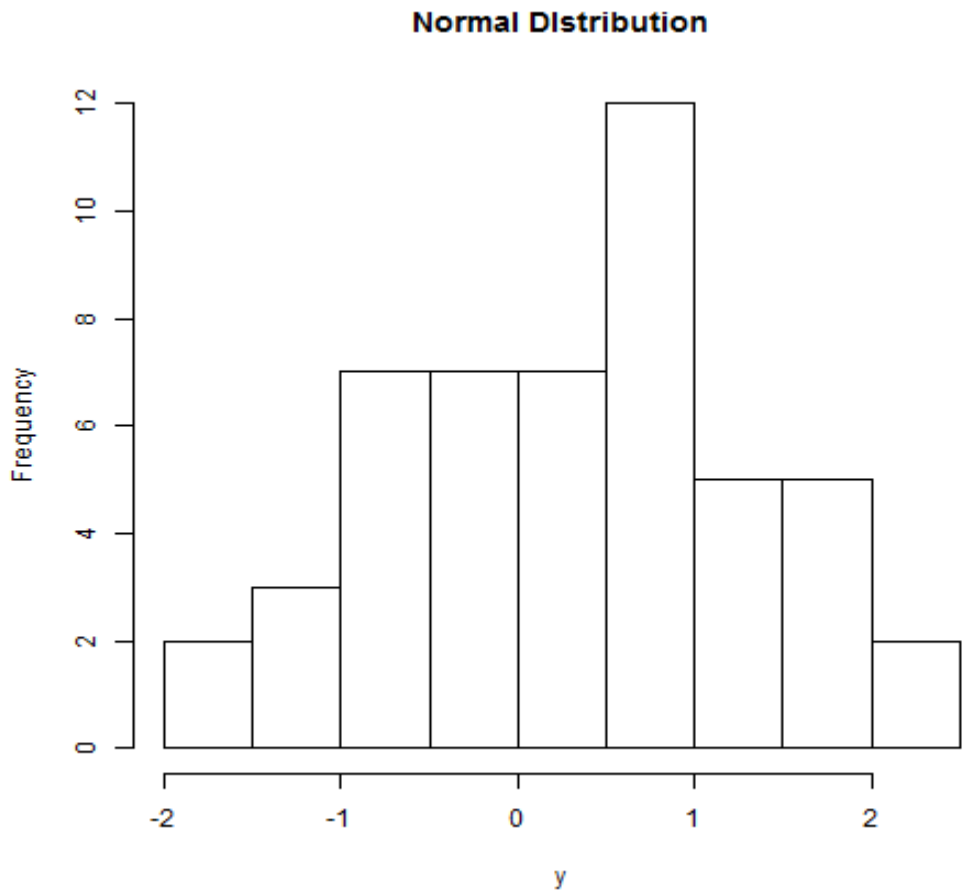
- This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.(回傳 n 個符合常態分布的隨機值;產生相同分布的隨機數)

```
> x <- rnorm(1000)
> hist(x, ylab = "Frequency")
```



rnorm()

```
y <- rnorm(50)
png(file = "rnorm.png")
hist(y, main = "Normal DIstribution")
dev.off()
```



Normal Distribution in R

- R 語言中，對於任何一個機率分布 ，都可以實作出以 d, p, q, r 為字首的四種函數，例如對於常態分布 Normal Distribution（簡寫為 norm）而言，就有 dnorm, pnorm, qnorm, rnorm 等四個函數，

Binomial Distribution in R

- 二項分佈是一種具有廣泛用途的離散型隨機變數的概率分佈，它是由伯努利始創的，所以又叫伯努利分佈。 二項分佈是指統計變數中只有性質不同的兩項群體的概率分佈。所謂兩項群體是按兩種不同性質劃分的統計變數，是二項試驗的結果。 即各個變數都可歸為兩個不同性質中的一個，兩個觀測值是對立的。
- 許多隨機試驗都有一些共同的特性，像丟一個銅板，只丟出正面與反面兩種結 果；抽一支籤，會出現中獎與不中獎兩種結果；候選人支持率的調查中，只有 支持與不支持兩種結果。 只有兩種結果(通常將這兩種結果稱為「成功」及「失敗」)的隨機試驗，稱 為伯努利試驗(Bernoulli Trial)。在伯努利試驗中，如果成功的機率為 p，則失 敗的機率為 1-p。
- 在機率論和統計學中，**二項分布**是n 個獨立的是/非試驗中成功的次數的離散機率分布，其中每次試驗的成功機率為p。這樣的單次成功/失敗試驗又稱為伯努利試驗。

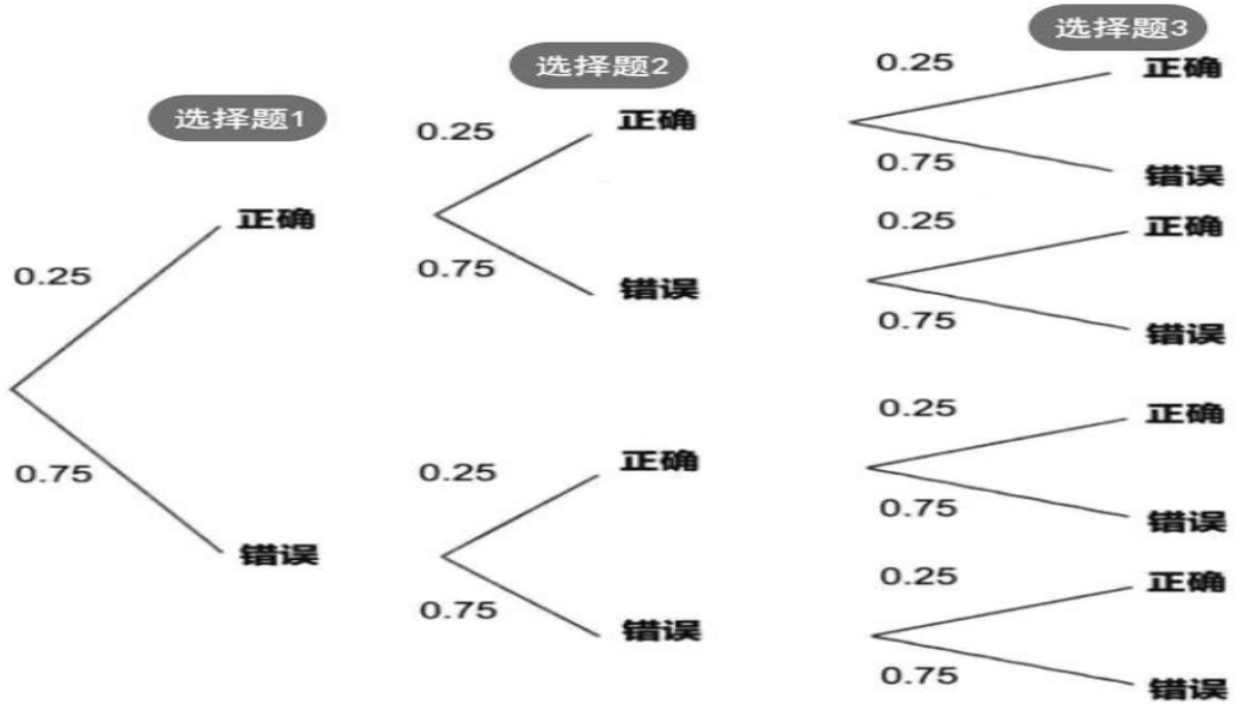
一般地，如果隨機變數X服從參數為n和p的二項分布，我們記$X \sim b(n,p)$或$X \sim B(n,p)$。n次試驗中正好得到k次成功的機率由機率質量函數給出：

$$f(k,n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

對於$k = 0, 1, 2, \ldots, n$，其中$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial Distribution 應用

- 在生活中，有很多的事情都只有兩種結果（outcome），例如考試是否及格、明天是否下雨、丟擲銅板並觀察其結果。當一個試驗只有兩種可能結果（成功與失敗），且兩個結果出現之機率為固定（若成功機率為 p，則失敗機率為 1−p），我們稱這樣的試驗為伯努力試驗（Bernoulli trial）。當我們重複進行多次相同的伯努力試驗（如丟擲一相同硬幣數次），且已知這些試驗之間的結果互相獨立（即這次試驗的結果不影響下次試驗的結果），則稱為二項實驗（binomial experiment）。
- 一個二項分布的實際生活問題，每道選擇題都只有一個正確答案，其他三個為錯誤答案，每道題做題的結果只有正確和錯誤兩種；每道題做對的機率為0.25，做錯的機率為0.75，它們的機率只和為1；每道題相互獨立、互不影響。先用機率樹畫出做題結果的分布情況：



Binomial Distribution 期望值與變異數

- 如果$X \sim B(n, p)$（也就是說，X是服從二項分布的隨機變數），那麼X的期望值為

$$\mathrm{E}[X] = np$$

變異數為

$$\mathrm{Var}[X] = np(1-p).$$

這個事實很容易證明。首先假設有一個伯努利試驗。試驗有兩個可能的結果：1和0，前者發生的機率為p，後者的機率為$1-p$。該試驗的期望值等於$\mu = 1 \cdot p + 0 \cdot (1-p) = p$。該試驗的變異數也可以類似地計算：$\sigma^2 = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p)$.

一般的二項分布是n次獨立的伯努利試驗的和。它的期望值和變異數分別等於每次單獨試驗的期望值和變異數的和：

$$\mu_n = \sum_{k=1}^{n} \mu = np, \qquad \sigma_n^2 = \sum_{k=1}^{n} \sigma^2 = np(1-p).$$

Binomial Distribution in R

- 二項分佈生成隨機變數，不只是單純的生成變數而已，而是生成獨立隨機試驗的成功次數。若要模擬10次試驗的成功次數，而每一次成功機率為0.4，我們使用 rbinom()，parameter 為 n=1（p.s.表示對所有的實驗進行一次），size =10（p.s. 實驗次數為10次），所以 prob= 0.4（p.s.成功機率為0.4）. 底下表示，做了10次試驗，每次試驗成功機率為0.4，而所產生的值，代表10次中有多少次的成功試驗.由於這個值是隨機生成的，因此每次生成的值會不一樣.

```
> rbinom(n = 1, size = 10, prob = 0.4)
[1] 3
> rbinom(n = 1, size = 10, prob = 0.4)
[1] 5
```

- 若將n設為大於1的數，二項分佈會對這 n組試驗生成 n 個成功次數，其中每組試驗的次數則以 size 參數來設定.如下，

```
> rbinom(n = 5, size = 10, prob = 0.4)
[1] 2 5 0 3 3
> rbinom(n = 10, size = 10, prob = 0.4)
 [1] 4 4 5 3 4 4 3 2 3 4
```

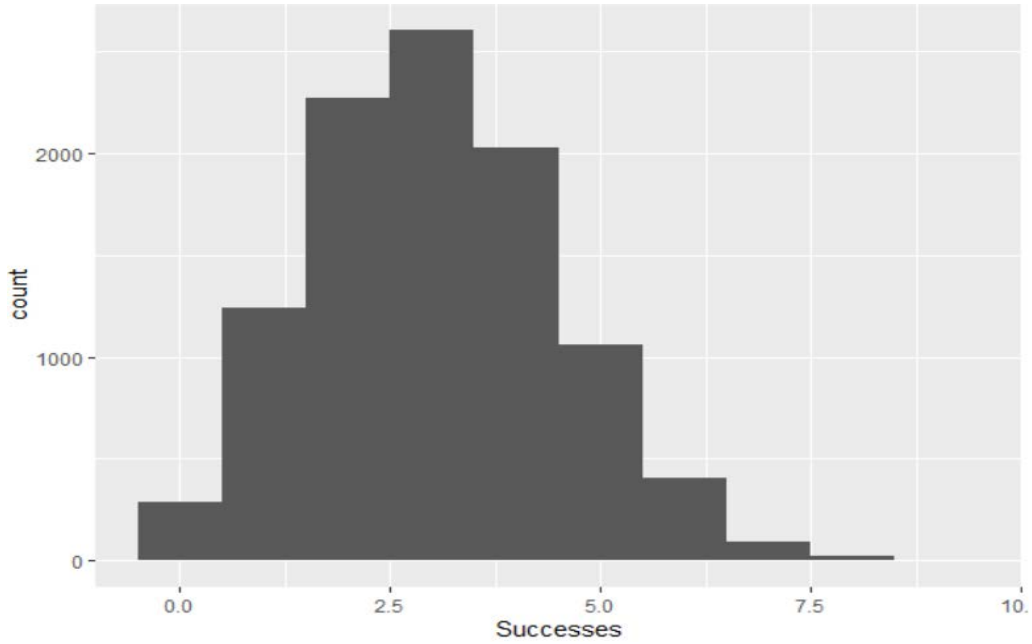
Binomial Distribution in R

- 若將size設為1,則生成的數字變為伯努力隨機變數,則只有兩種可能性的變數值, 1表示成功, 0表示失敗.伯努力隨機試驗為二項分佈試驗的特例?

```
> rbinom(n = 1, size = 1, prob = 0.4)
[1] 1
> rbinom(n = 5, size = 1, prob = 0.4)
[1] 0 1 0 0 0
> rbinom(n = 10, size = 1, prob = 0.4)
 [1] 1 1 1 1 0 0 0 1 1 1
```

- 為了劃出二項分佈試驗的結果,底下隨機生成10,000組的試驗, 每組包含10個試驗,每個試驗的成功率為0.3. 圖中可以看出最常出現的成功次數為3

```
> library(ggplot2)
> binomData <- data.frame(Successes
+     = rbinom(n = 10000, size = 10,prob = 0.3))
> ggplot(binomData, aes(x = Successes)) +
+       geom_histogram(binwidth = 1)
```



p.s. 直方圖的**分箱數**（bins）使用分箱數 30 這個預設值，可以透過調整 binwidth 或 bins 參數來更改，增加 binwidth 與減少 bins 會減少分箱數；減少 binwidth與增加 bins 則會增加分箱數。

Binomial Distribution in R-近似於常態分布

- 當試驗次數被提高,二項分佈結果會近似於常態分布.

```
> # 建立一個持有兩個直排和10,000橫排的data.frame
> # 第一直排為Successes,記錄的是10,000組試驗的成功次數
> # 第二直排記錄每組試驗次數(Size),每組次數為5
> binom5 <- data.frame(Successes=rbinom(n=10000, size=5,
+                                       prob=.3),Size=5)
> dim(binom5)
[1] 10000     2
> head(binom5)
  Successes Size
1         3    5
2         1    5
3         1    5
4         1    5
5         4    5
6         1    5
> # 跟之前的一樣,還是10,000橫排
> # 這次從不同試驗次數的分佈抽取成功次數
> # 現在10,000橫排的Size皆為10
> binom10 <- data.frame(Successes=rbinom(n=10000, size=10,
+                                        prob=.3), Size=10)
> dim(binom10)
[1] 10000     2
> head(binom10)
  Successes Size
1         2   10
2         5   10
3         3   10
4         3   10
5         1   10
6         4   10
> binom100 <- data.frame(Successes=rbinom(n=10000, size=100,
+                                         prob=.3),Size=100)
> binom1000<- data.frame(Successes=rbinom(n=10000,size=1000,
+                                         prob=.3),Size=1000)
```

Binomial Distribution in R-近似於常態分布

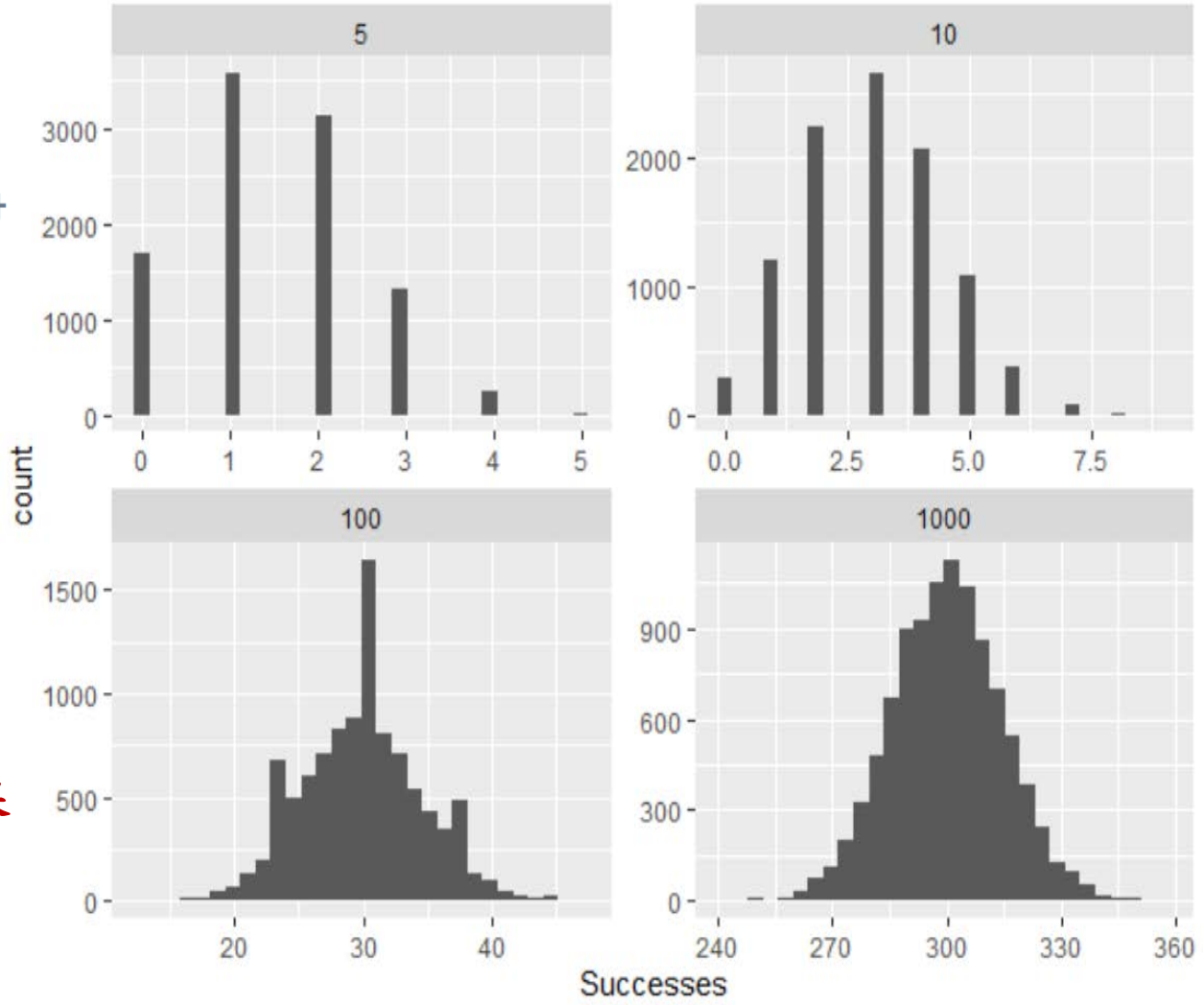
- 當試驗次數被提高, 二項分佈結果會近似於常態分布. 底下會以不同的實驗來生成變數(值)

```
> # 把它們都合併在一起
> binomAll <- rbind(binom5, binom10, binom100, binom1000)
> dim(binomAll)
[1] 40000     2
> head(binomAll, 10)
   Successes Size
1          2    5
2          1    5
3          1    5
4          2    5
5          1    5
6          1    5
7          2    5
8          0    5
9          1    5
10         2    5
> tail(binomAll, 10)
      Successes Size
39991       298 1000
39992       307 1000
39993       297 1000
39994       303 1000
39995       294 1000
39996       286 1000
39997       316 1000
39998       291 1000
39999       309 1000
40000       307 1000
```

Binomial Distribution in R-近似於常態分布

- 當試驗次數被提高, 二項分佈結果會近似於常態分布. 底下會以不同的實驗來生成變數(值). 繪圖如下:

```
# 現在繪圖. 直方圖只需設定x軸
# 圖的分層(拆解)依據為Size的值
# 這些值為5, 10, 100, 1000
ggplot(binomAll, aes(x=Successes)) + geom_histogram() +
       facet_wrap(~ Size, scales="free")
```



p. s. 圖形中證明, 當試驗次數被提高, 二項分佈結果會近似於常態分布

Binomial Distribution-dbinom()+pbinom()

- *n*與*p*為試驗次數和成功機率

一般的二項分布是*n*次獨立的伯努利試驗的和。它的期望值和變異數分別等於每次單獨試驗的期望值和變異數的和：

$$\sigma_n^2 = \sum_{k=1}^{n} \sigma^2 = np(1-p).$$

- 如同 dbinom() 與pbinom()各提供二項分佈的機率密度(某值真正發生的機率)和累積機率

```
> # 10次試驗裏三次成功的機率
> dbinom(x = 3, size = 10, prob = 0.3)
[1] 0.2668279
> # 10次試驗裏三次或更少次成功的機率
> pbinom(q = 3, size = 10, prob = 0.3)
[1] 0.6496107
```

Binomial Distribution-dbinom()+pbinom()+qbinom()

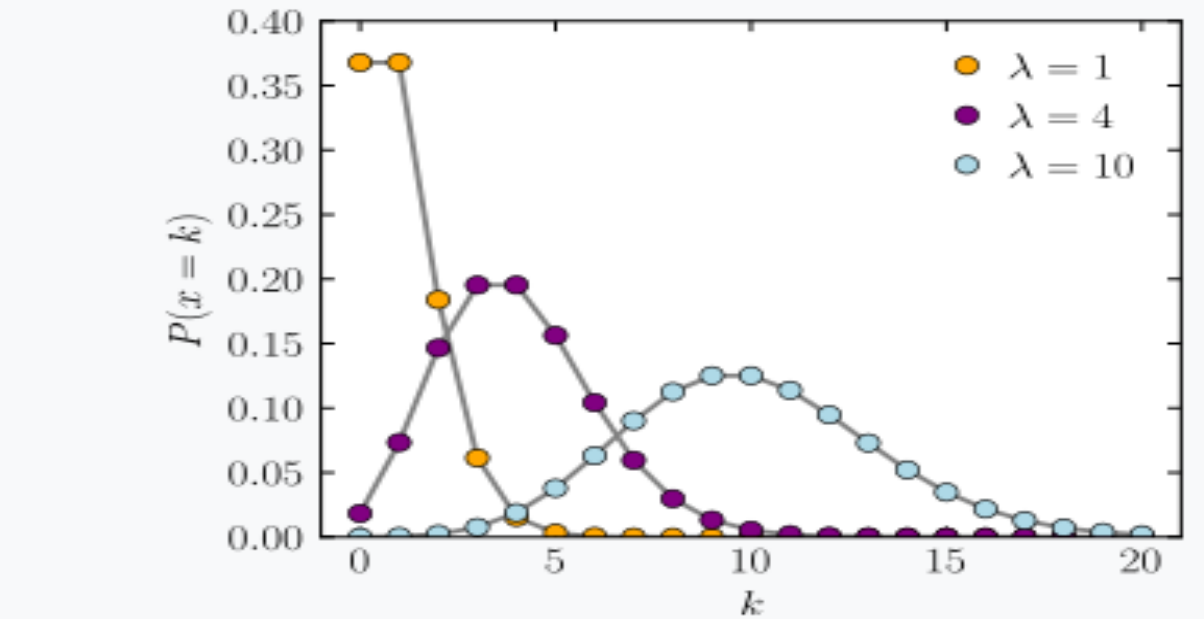
- 如同 dbinom() 與pbinom()各提供二項分佈的機率密度(某值真正發生的機率)和累積機率

```
> # 該兩個函數也可以執行向量化運算
> dbinom(x = 1:10, size = 10, prob = 0.3)
 [1] 0.1210608210 0.2334744405 0.2668279320 0.2001209490 0.1029193452
 [6] 0.0367569090 0.0090016920 0.0014467005 0.0001377810 0.0000059049
> pbinom(q = 1:10, size = 10, prob = 0.3)
 [1] 0.1493083 0.3827828 0.6496107 0.8497317 0.9526510 0.9894079
 [7] 0.9984096 0.9998563 0.9999941 1.0000000

> qbinom(p = 0.3, size = 10, prob = 0.3)
[1] 2
> qbinom(p = c(0.3, 0.35, 0.4, 0.5, 0.6), size = 10, prob = 0.3)
[1] 2 2 3 3 3
```

Poisson Distribution

- 卜瓦松分布適合於描述單位時間內隨機事件發生的次數的機率分布。如某一服務設施在一定時間內受到的服務請求的次數，電話交換機接到呼叫的次數、汽車站台的候客人數、機器出現的故障數、自然災害發生的次數、DNA序列的變異數、放射性原子核的衰變數、雷射的光子數分布等等。如圖，橫軸是索引k，發生次數。該函數只定義在k為整數的時候。連接線是只為了指導視覺的機率。



Poisson Distribution

- 卜瓦松分布的機率質量函數為（機率質量函數是離散隨機變量在各特定取值上的機率）：

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- 卜瓦松分布的累積分佈函數為(ps. 其中λ為期望值與變異數)：

$$M_X(t) = E[e^{tX}] = \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(e^t\lambda)^x}{x!} = e^{\lambda(e^t-1)}$$

Poisson Distribution - 近似於常態分布

- *rpois*、*dpois* 、*ppois* 、*qpois* 可用來生成隨機計數資料, 計算機率密度、累積機率、分位數. 當 λ 值越大, 卜瓦松分布越接近常態分佈. 底下模擬10,000個sample與繪製值方圖,可知

```
> # 從5個不同的泊松分佈各生成10,000個樣本
> pois1 <-rpois(n=10000, lambda=1)
> pois2 <-rpois(n=10000, lambda=2)
> pois5 <-rpois(n=10000, lambda=5)
> pois10<-rpois(n=10000, lambda=10)
> pois20<-rpois(n=10000, lambda=20)
> pois <- data.frame(Lambda.1=pois1, Lambda.2=pois2,
+                    Lambda.5=pois5, Lambda.10=pois10,
+                    Lambda.20=pois20)
```

```
> pois <- melt(data=pois, variable.name="Lambda", value.name="x")
Using Lambda as id variables
> head(pois)
    Lambda Lambda x
1 Lambda.1      x 0
2 Lambda.1      x 1
3 Lambda.1      x 0
4 Lambda.1      x 2
5 Lambda.1      x 0
6 Lambda.1      x 1
```

Poisson Distribution - 近似於常態分布

- *rpois*、*dpois* 、*ppois* 、*qpois* 可用來生成隨機計數資料, 計算機率密度、累積機率、分位數. 當 λ 值越大, 卜瓦松分布越接近常態分佈. 底下模擬10,000個sample與繪製值方圖,可知

```
> library(stringr)
> pois$Lambda <- as.factor(as.numeric
+                          (str_extract(string=pois$Lambda,
+                           pattern="\\d+")))
> head(pois)
  Lambda Lambda x
1      1      x 0
2      1      x 1
3      1      x 0
4      1      x 2
5      1      x 0
6      1      x 1
> tail(pois)
       Lambda Lambda  x
49995      20      x 20
49996      20      x 13
49997      20      x 17
49998      20      x 27
49999      20      x 13
50000      20      x 15
```

Poisson Distribution

- Example,

The **Poisson distribution** is the probability distribution of independent event occurrences in an interval. If λ is the mean occurrence per interval, then the probability of having x occurrences within a given interval is:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad where\ x = 0, 1, 2, 3, ...$$

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

Solution

The probability of having *sixteen or less* cars crossing the bridge in a particular minute is given by the function ppois.

```
> ppois(16, lambda=12)    # lower tail
[1] 0.89871
```

Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the *upper tail* of the probability density function.

```
> ppois(16, lambda=12, lower=FALSE)    # upper tail
[1] 0.10129
```

Answer

If there are twelve cars crossing a bridge per minute on average, the probability of having seventeen or more cars crossing the bridge in a particular minute is 10.1%.

The End