1. 統計學概論
Introduction to Statistics

- 統計學是什麼?
- 以統計學進行決策支援的過程。
- 母體與樣本的概念，母數與統計量的差別和各自的用途。
- 統計量推論母數的統計推論過程。
- 變數資料的各式型態與分類方式。

1.1 統計學是什麼?

- 數據資料(data)與統計數字(figures)充斥在我們的日常生活之中，而這些「數字」，通常是有意義、有內容、有意涵的數字；這意義、內容、意涵，就是資訊(information)，就是知識(knowledge)。

- 統計學(statistics)即是一門研究如何從雜亂無章的數據資料中，粹煉出有用的資訊或知識的學問，以利研究者或實務工作者作成結論，並進一步成為決策依據。

1.1.1 統計學的定義

- 統計分析所研究的對象一般稱之為個體(individuals)或元素 (elements)，可以是人、動物，或是任何其他東西(如：國家、 公司、 藥物、氣候等)，甚至可以是抽象的事物(如：經濟情勢、 市場動態等)。

- 變數(variable)：研究者欲研究的個體所具有的某些特性(feature)、特質(property)或特徵(characteristic)，不同的個體可能有不同的「值」，故稱為變數。變數的「值」不一定是「數」，也可以是「文字」或「符號」。例如，研究個體為「人」，可以考慮的「變數」 有哪些?

- 統計分析是專注在處理各種不同 (研究者感興趣) 的「變數」，而研究者對個體的變數進行測量或觀察，所得到的變數「值」即稱為資料(data，或稱數據)，不同型態的變數資料會有不同的統計方法來處理它們。

1.1.1 統計學的定義(contd.)

定義 1.1 資料 (data)

在實驗(experiment)、調查(survey)或研究(investigation)中，研究者對於欲研究之對象個體，針對其所感興趣的變數，進行測量或蒐集而得到的原始資訊(raw information)。

- 在科學研究中，數據資料依其取得方式的不同，主要可分為下面兩大類：
  - 實驗性資料(experimental data)：主要由自然科學(natural science)與工程(engineering)的研究問題中取得。這類資料的雜訊干擾較?
  - 觀察性資料(observational data，或稱非實驗性資料，non-experimental data)：主要由社會科學(social science)的研究問題中取得。這類資料的雜訊干擾較?
- 整個研究(study)所蒐集的所有資料，稱之為資料集(data set)。

範例1.1

- 以下是NBA球星在2016年的部份基本資料集：

姓名	國籍	所屬球隊	背號	位置	年齡	年資	身高	體重
Tony Parker	France	Spurs	9	Guard	33	14	188	83.9
LeBron James	USA	Cavaliers	23	Forward	30	12	203	113.4
James Harden	USA	Rockets	13	Guard	26	6	195	99.8
Stephen Curry	USA	Warriors	30	Guard	27	6	190	86.2
Jeremy Lin	USA	Hornets	7	Guard	27	5	191	90.7
Dirk Nowitzki	Germany	Mavericks	41	Forward	37	17	213	111.1

我們研究的個體是? 感興趣的變數是? 變數的值是?

範例1.2

- The following data give the lengths of time to failure for 43 light bulbs:

  203, 65, 16, 224, 16, 80, 96, 536, 400, 80, 292, 576, 128, 234, 19, 78, 121, 101, 32, 99, 12, 421, 66, 298, 177, 79, 333, 20, 12, 323, 46, 103, 441, 367, 232, 278, 442, 60, 123, 153, 24, 199, 301

  (1) What is the maximum value? What is the minimum value?

  (2) What can be said or be inferred from these data?

- 這個例子表達了什麼現象或事實?

1.1.1 統計學的定義(contd.)

- 以統計學進行決策支援的過程：

1.1.1 統計學的定義(contd.)

定義 1.2 統計學 (statistics)

(1) 統計學(statistics)是應用數學的一個分支，是一門用以蒐集(collection)、整理(arrangement)、陳示(presentation)、分析(analysis)、解釋(interpretation)和推論(inference)大量資料的學問與技術。

(2)統計學(statistics)是一門決策科學(science of decision-making)，研究如何在面對不確定現象和環境下制定適當的決策。

- 統計學(statistics)可以被視為最傳統也最重要的資料挖掘分析工具之一。
- 你還聽過哪些資料挖掘分析工具?

1.1.2 母體與樣本

定義 1.3 母體 (population)

研究者主觀欲研究對象的全體集合，具有某些研究者感興趣的特徵(characteristics)。

- 母體大小(population size)一般用 N 來表示，我們以 $X_1, X_2, \ldots, X_N$ 來表示這筆母體資料的 N 個觀察數值，而依據母體大小可將母體分為兩類：
  - 有限母體(finite population)
  - 無限母體(infinite population，或稱概念性母體，conceptual population)

範例1.3

研究課題	母體	母體大小
台灣人的健康狀況		$N =$
鼓山國小6年1班學生家庭狀況		$N =$
南亞公司生產之燈泡品質		$N =$
AIDS新藥之治癒率		$N =$

1.1.2 母體與樣本(contd.)

定義 1.4 母數 (parameter)

母數又稱參數或母體參數，是描述某個變數的母體特徵的測量數值。

- 一般而言，除非透過普查(census)的程序，否則母數是一個確實存在，但研究者卻無法得知的數值；我們常用希臘字母 θ 來代表母數。

- 常用的三個重要母數：
  - 描述中央趨勢的特徵：
  - 描述分散趨勢的特徵：
  - 描述母體比例的特徵：

範例1.4

- 台灣人的身高研究中，
  - 「研究的個體」是?
  - 「母體」是?
  - 「母體大小」是?
  - 感興趣的「變數」是?
  - 想要描述此變數的「中央趨勢特徵」，該用什麼母數?
  - 想要描述此變數的「分散趨勢特徵」，該用什麼母數?
  - 想要描述此變數的「比例特徵」，該用什麼母數?
  - 舉個「想要描述此變數的比例特徵」之例子

1.1.2 母體與樣本(contd.)

- 研究母數的兩個主要方法：
  - 普查(census)：

    針對母體全體的每一個個體(individuals)或元素(elements)，將他(牠或它)們的變數進行調查或測量，進而完全掌握母體全體的特徵。

  - 抽樣調查(sample survey)：

    在母體中，以某種抽樣(sampling)的方式，選取部份個體或元素形成樣本(sample)，再對樣本的變數進行調查或測量，進而由樣本的特徵推論出母體的特徵。

- 普查的主要困難點：

1.1.2 母體與樣本(contd.)

定義 1.5 樣本 (sample)

母體的一個子集合(subset)。

- 樣本大小(sample size，或稱樣本數)一般用 n 來表示，我們以 $X_1, X_2, \ldots, X_n$ 來表示這筆樣本資料的 n 個觀察數值。
- 樣本如何能夠代表母體全體?
- 隨機抽樣(random sampling，又稱機率抽樣，probability sampling)：
  母體當中的每一個個體，以事先給定的機率(通常給定的是相同的機率)，決定是否被抽中而納入樣本的抽樣方法，此種樣本稱之為隨機樣本(random sample)。
- 所有的推論統計學理論，都建立在隨機樣本的假設之上，因而機率論(probability theory) 是推論統計學的重要基礎。

1.1.2 母體與樣本(contd.)

定義 1.6 統計量 (statistics)

統計量又稱樣本統計量，是描述某個變數的樣本特徵的測量數值；而統計量最重要的功能，是對於未知的母數進行統計推論(statistical inference)。

- 樣本統計量是隨機樣本 $X_1, X_2, \ldots, X_n$ 的適當「函數組合」，記為$T(X_1, X_2, \ldots, X_n)$，透過這個適當的組合，把隨機樣本中對於母數的訊息(information)包覆住，據以推論未知母數。

- 樣本統計量是隨機變數(random variable)嗎? 亦即它會隨機變動嗎?

- 常用的三個重要樣本統計量：
  - 描述中央趨勢的特徵：
  - 描述分散趨勢的特徵：
  - 描述母體比例的特徵：

1.1.2 母體與樣本(contd.)

定義 1.7 統計推論 (statistical inference)

藉由分析樣本特徵進而歸納和推論出母體特徵的過程，也就是利用樣本統計量來歸納和推論出母數。

- 統計推論主要包含兩種工具：
  - 估計(estimation)：
    以樣本統計量直接估計未知母數可能的數值，包括點估計與區間估計兩種方法；用來估計母數的統計量，稱之為估計量(estimator)。
  - 假說檢定(hypothesis testing ，或稱假設檢定)：
    對未知母數提出一些可能的假說，以樣本統計量來決定這些假說應被廢棄或建立；用來檢定母數的統計量，稱之為檢定統計量(test statistic)。

1.1.2 母體與樣本(contd.)

- 統計分析的基本架構：

1.1.3 統計學之分類

- 依處理問題的不同來區分：
  - 敘述統計學(descriptive statistics)：
    - 研究如何蒐集、整理、陳示、分析、解釋大量資料的工具與技術。
    - 目的：將雜亂無章的資料，轉變為易讀易懂的形式，進而能分析或解釋其性質，無關乎母體與樣本之概念。
  - 推論統計學(inferential statistics)：
    - 研究如何以樣本資料和特性推論出母體特性之方法。
    - 目的：面對不確定狀況或龐大母體問題時，藉由抽樣的理論與方法，以較少的成本，幫助決策者進行決策。

1.2 變數的測量尺度與資料的種類

- 變數的測量尺度
  - 名義尺度(nominal scale)
    - 僅可作為辨識、分類用途之變數，是一種屬質的變數。
  - 順序尺度(ordinal scale)
    - 除了作為辨識、分類用途之外，還可以比較優劣、好壞、高低之順序，是一種屬質的變數。
  - 等距尺度(interval scale)
    - 除了作為辨識、分類用途，可以比較優劣、好壞、高低之外，還能說明「好多少，優多少，高多少」，也就是可以計算差距之大小，必為屬量的變數。
  - 比例尺度(radio scale)
    - 除了作為辨識、分類用途，可以比較優劣、好壞、高低，可以計算差距之大小之外，還能比較差距之倍數，必為屬量的變數。存在一個絕對零點。

1.2 變數的測量尺度與資料的種類(contd.)

- 統計變數的資料依其性質可分為兩大類：
  - 屬質資料(qualitative data，又稱類別資料，categorical data)
    - 僅描述性質(quality)或類別(category)的資料，通常是文字或符號的非數值型態；也有可能是數值型態，但此時只代表了分類號(code)，並無數量的意涵。
    - 屬質資料必為離散型資料(discrete data)。
  - 屬量資料(quantitative data)
    - 描述數量(how many)和多寡(how much)的資料，必為數值型態。
    - 一般可再分為離散型資料(discrete data)與連續型資料(continuous data)。

範例1.5

- The summaries of data, which may be tabular, graphical, or numerical, are referred to as

  (a) inferential statistics (b) descriptive statistics

  (c) statistical inference (d) report generation

- The process of analyzing sample data in order to draw conclusions about the characteristics of a population is called

  (a) descriptive statistics (b) statistical inference

  (c) data analysis (d) data summarization

範例1.6

- Quantitative data
  (a) are always nonnumeric (b) may be either numeric or nonnumeric
  (c) are always numeric (d) are always labels

- Qualitative data
  (a) indicate either how much or how many (b) can not be numeric
  (c) are labels used to identify attributes of elements (d) must be nonnumeric

範例1.7

- The measurement data for the size of women's shoes or cloth are
  (a) the nominal scale (b) the ordinal scale (c) the interval scale
  (d) the ratio scale (e) none of above

- Temperature is an example of a variable which uses
  (a) the ratio scale (b) the interval scale
  (c) the ordinal scale (d) either ratio or interval scale