ECON 1203 Notes

What is Statistics

- Descriptive statistics deals with methods of organising, summarising and presenting data in order to extract information
- Inferential statistics are methods used to draw conclusions or inferences about characteristics of populations based on sample data

Key Statistical Concepts

- A **population** is the group of all items of interest to a statistics practitioner - it is generally large and does not have to refer to a group of people
- A **parameter** is a numerical description of a population - it basically represents the information we need
- A **sample** is a set of data drawn from the studied population
- A **statistic** is a numerical description of a sample
- A **statistical inference** is the process of making an estimate, prediction or decision about a population based on sample data
    - As these are not always correct, there are two measures of reliability:
        - **Confidence level**: proportion of times that an estimating procedure will be correct
        - **Significance level**: how frequently the conclusion will be wrong

Consider a study of shoe sizes of men in Australia

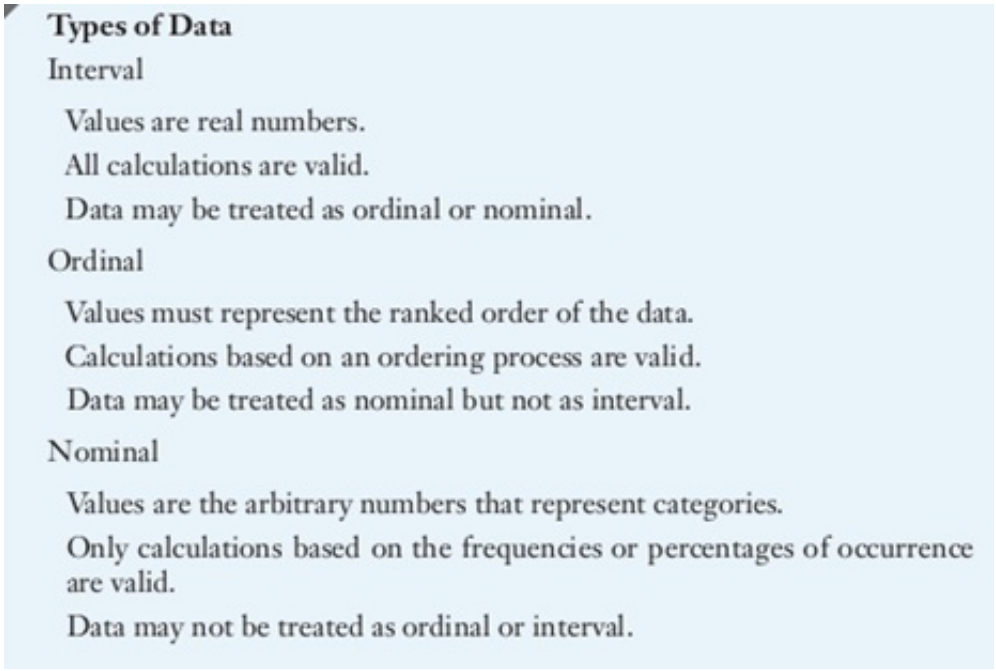
- Population: Shoe sizes of all men in Australia
- Parameter: Average shoe size of all Australian men
- Sample: Shoes sizes of all the men in the class
- Statistic: Average shoe size of all men in the class

Types of Data and Information

- A **variable** is a characteristic of a population or sample. They may be:
    - Quantitative: can be expressed numerically
    - Qualitative: cannot be expressed numerically
    - Discrete: cannot take on all values within its range
    - Continuous: can take any value in their range

For example: Scores and time are quantitative. Gender and nationality are qualitative. Scores, age, shoe sizes are discrete. Height, time, mass are continuous.

- We observe **values** or observations of a variable
- A **data set** contains the observed values of a variable
- There are three types of data:
    - Interval data: real numbers such as heights & weights; these are quantitative
    - Nominal data: categories such as marital status and gender; these are qualitative
    - Ordinal data: order of values has meaning such as [poor, fair, good, very good, excellent]



Types of Data

Interval

Values are real numbers.
All calculations are valid.
Data may be treated as ordinal or nominal.

Ordinal

Values must represent the ranked order of the data.
Calculations based on an ordering process are valid.
Data may be treated as nominal but not as interval.

Nominal

Values are the arbitrary numbers that represent categories.
Only calculations based on the frequencies or percentages of occurrence are valid.
Data may not be treated as ordinal or interval.

- The type of data will determine the appropriate means of analysis. Calculations are permitted on interval data but it would not make sense to perform calculations on the codes which represent nominal data - instead, we can only count and record the frequencies of occurrences in each category

Describing a Set of Nominal Data

We are only able to count the frequency/relative frequency of nominal data.

- A **frequency distribution** presents the categories and their counts in a table
- A **relative frequency distribution** lists the categories and the proportion with which each occurs
- Two graphical techniques can be used to represent the data:
    - **Bar chart**: used to display frequencies
    - **Pie chart**: used to display relative frequencies

Describing a Set of Ordinal Data

There are no specific graphical techniques for ordinal data. Instead, we treat the data as being nominal and use the associated techniques. However, we must arrange the bars or wedges in a bar or pie chart in ascending or descending order

Describing/Graphing the Relationship between Two Nominal Variables

- Graphical and tabular techniques used to summarise singles sets of data are **univariate**
- Techniques used to depict the relationship between two variables are **bivariate**
- A **cross-classification table** is used to describe the relationship between two nominal variables
- We can use multiple bar charts to graph the data between two variables.
    - If two variables are unrelated, the patterns in both bar charts should be the

same
    - If there is a relationship between the two variables, the bar charts will be different

Graphical Techniques to Describe a Set of Interval Data

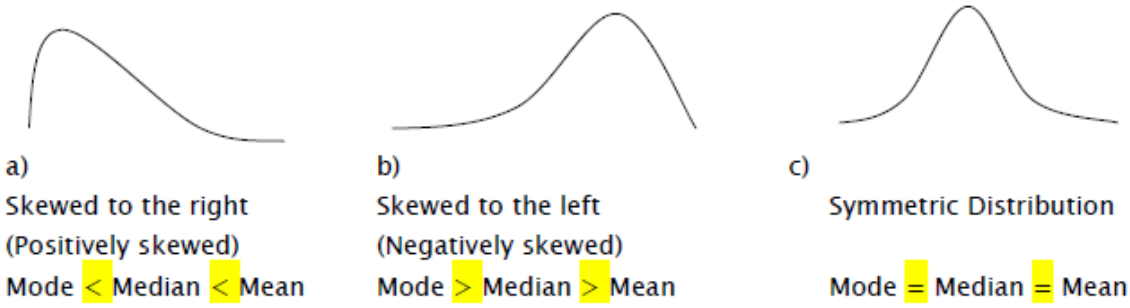
Histogram

The most important graphical method for interval data is the **histogram.** This is created by drawing rectangles whose bases are the intervals and whose heights are the frequencies

- A series of intervals is called a **class (or bin)**. For example: 0-15, 15-30, ...
    - These classes need to be mutually exclusive and exhaustive - there should be no uncertainty about which interval to assign to any observation
- The number of class intervals depends on the number of observations in the data set
    - Sturges's formula: Number of class intervals = 1 + 3.3log(n)
- The class interval widths are determined by finding the difference between the largest and smallest observation and dividing this by the number of classes

Shapes of Histograms

- **Symmetry**: left half is a mirror image of the right half
- Non symmetrical/Skewness: Histogram with a long tail to the right (**positively skewed)** or long tail to the left **(negatively skewed)**
- A **mode** is the observation with the greatest frequency, and the **modal class** is the class with the largest number of observations
    - **Unimodal histograms** have one single peak
    - **Multimodal histograms** have multiple peaks



Stem and leaf display

- The downfall of histograms is that information may be lost when classifying observations. Thus we can use **stem and leaf displays**
    - It is basically a histogram on its side, where the length of each line represents the frequency in the class interval defined by the stems. However, we can see the actual observations

Stem	Leaf
0	000000000111112222223333345555556666667788889999999
1	000001111233333334455555667889999
2	000011111234466677 8999
3	001335589
4	124445589
5	33566
6	3458
7	022224556789
8	334457889999
9	0011222 2233344555999
10	001344446699
11	0124557889

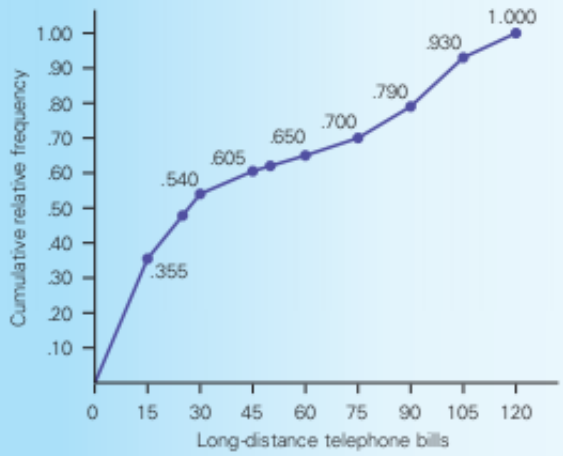
Ogive

- Information in frequency distributions can be converted into a **relative frequency distribution table** and then displayed on an **ogive**
    - We can also establish the **cumulative relative frequency**

TABLE **3.4** Cumulative Relative Frequency Distribution for Example 3.1

CLASS LIMITS	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
0 to 15	71/200 = .355	71/200 = .355
15 to 30	37/200 = .185	108/200 = .540
30 to 45	13/200 = .065	121/200 = .605
45 to 60	9/200 = .045	130/200 = .650
60 to 75	10/200 = .05	140/200 = .700
75 to 90	18/200 = .09	158/200 = .790
90 to 105	28/200 = .14	186/200 = .930
105 to 120	14/200 = .07	200/200 = 1.00

FIGURE **3.8** Ogive for Example 3.1



Describing Time-Series Data

- **Time series** data refers to measurements at different points in time (Births per day)
- **Cross sectional** data refers to measurements at a single point in time (Sydney house prices by suburb)

Line chart

- Time-series data are often graphed on a **line chart** which is a plot of the variable over time
  - X axis: time periods
  - Y axis: value of the variable