1 Introduction to Statistics

What is Statistics

- Statistics is a branch of mathematics in which groups of measurements or observations are studied.
- The subject is divided into two general categories: ***Descriptive statistics*** and ***Inferential statistics***.
- In descriptive statistics one deals with methods used to collect, organize and analyze numerical facts. Its primary concern is to describe information gathered through observation in an understandable and usable manner.
- Similarities and patterns among people, things and events in the world around us are emphasized. Inferential statistics takes data collected from relatively small groups of a population and uses inductive reasoning to make generalizations, inferences and predictions about a wider population.

Throughout the study of statistics certain basic terms occur frequently. Some of the more commonly used terms are defined in the next sections.

2 Populations and Samples

- The collection of everyone or everything that is to be analyzed in a study is called a **population**. As we have seen in the examples above, the population could be enormous in size. There could be millions or even billions of individuals in the population.
  - *Remark: We must not think that the population has to be large. If our group being studied is fourth graders in a particular school, then the population consists only of these students. Depending on the school size, this could be less than a hundred students in our population.*
- To make our study less expensive in terms of time and resources, we only study a subset of the population. This subset is called a **sample**.
- A relatively small group of items selected from a population is a sample. If every member of the population has an equal chance of being selected for the sample, it is called a random sample.
- Samples can be quite large or quite small. In theory one individual from a population constitutes a sample. Many applications of statistics require that a sample have at least 30 individuals.

- A sample is a subset of a population.
- Since it is usually impractical to test every member of a population, a sample from the population is typically the best approach available.

3 Parameters and Statistics

Parameters and Statistics

- The main objective of Statistics as a science is to estimate a population parameter by use of sample statistics.
- What we are typically after in a study is the **parameter**. A parameter is a numerical value that states something about the entire population being studied.
  - For example, we may want to know the mean wingspan of the American bald eagle. This is a parameter, because it is describing all of the population.
- Parameters are difficult if not impossible to obtain exactly. On the other hand, each parameter has a corresponding **statistic** that can be measured exactly.
- Inferential statistics generally require that sampling be **random** although some types of sampling (such as those used in voter polling) seek to make the sample as representative of the population as possible by choosing the sample to resemble the population on the most important characteristics.
- (Important) A statistic is a numerical value that states something about a sample.
- To extend the example above, we could catch 100 bald eagles and then measure the wingspan of each of these. The mean wingspan of the 100 eagles that we caught is a statistic.
- The value of a parameter is a fixed number. In contrast to this, since a statistic depends upon a sample, the value of a statistic can vary from sample to sample.
- Suppose our population parameter has a value, unknown to us, of 100. One sample of size 50 has corresponding statistic with value 95.5. Another sample of size 50 from the same population has corresponding statistic with value 101.1.
- The variability in statistic values is known as **sampling fluctuation**.

Parameter

- This is a numerical characteristic of the population; it is a fixed number with an unknown value.

Statistic

- This is a numerical characteristic of the sample; a value known when the sample is taken but that can change from sample to sample.

Examples of Parameters and Statistics

Below are some more example of parameters and statistics:

- Suppose we study the population of cats in Limerick City. A parameter of this population would be the mean weight of all cats in the city. A statistic would be the mean weight of a sample of 50 of these cats.
- We will consider a study of high school seniors in the United States. A parameter of this population is the standard deviation of grade point averages of all high school seniors. A statistic is the standard deviation of the grade point averages of a sample of 1000 high school seniors.

Mnemonic Device

There is a simple and straightforward way to remember what a parameter and statistic are measuring. All that we must do is look at the first letter of each word. A parameter measures something in a population, and a statistic measures something in a sample.

Summary

- A ***population*** is a collection of data whose properties are analyzed. The population is the complete collection to be studied, it contains all subjects of interest.
- A ***sample*** is a part of the population of interest, a sub-collection selected from a population.
- A ***parameter*** is a numerical measurement that describes a characteristic of a population, while a ***sample statistic*** is a numerical measurement that describes a characteristic of a sample.
- In general, the major use of statistics is to use information from a ***sample*** to infer something about a ***population***.

4 Types of Data

- Data are numbers or measurements that are collected. Data may include numbers of individuals that make up the census of a city, ages of pupils in a certain class, temperatures in a town during a given period of time, sales made by a company, or test scores made by ninth graders on a standardized test.
- Variables are characteristics or attributes that enable us to distinguish one individual from another. They take on different values when different individuals are observed. Some variables are height, weight, age and price. Variables are the opposite of constants whose values never change.

4.1 Types of Data

Data are the facts and figures collected, analyzed and summarized for presentation and interpretation.

- All the data collected in a study is the **data set** for the study.
- There are several types of data and identifying the type of data is vital in determining the statistical method used to describe it.
- Most statistical analysis are specific to a certain data type. Data can be classified as either ***qualitative*** or ***quantitative***.

5 Variables

Variables

The key terms used in data collection can be dened as follows:

- A variable is the phenomenon being measured in the experiment or observational study.
- A variable has two defining characteristics:
  - A variable is an attribute that describes a person, place, thing, or idea.
- The value of the variable can "vary" from one entity to another (randomness).
- Variables can be classified as categorical (or ***qualitative***) or numerical (or ***quantitative***).
  - **Categorical**. Categorical variables take on values that are names or labels.
    - The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., Border collie, German shepherd, Yorkshire terrier) would be examples of categorical variables.
  - **Quantitative**. Quantitative variables are numerical. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

- Quantitative data is always numeric.
- *The distinction between interval data and ratio data will be mentioned in class, but is omitted from syllabus.*
  - Examples: height, weight, age, expenditure.
- Quantitative variables are usually denoted by symbols or letters such as "X". (N.B. Capital letters).
- A **continuous variable** takes any value on a range of real numbers (analogous to 'measuring). Such variables can take any value in a certain range. They are usually measured according to some scale, e.g. age, height, mass.
- A **discrete variable** takes only distinct values, usually often integers (analogous to 'counting). Such variables take values from a set that can be listed (commonly integer values). Such variables are often counted, e.g. number of children, number of subjects passed at leaving certificate
- Note that continuous variables are only measured to a given accuracy. e.g. Age is normally given to the closest year. However, in theory it could be measured much more accurately.
- When a discrete variable takes a very large number of values , e.g. the number of individuals employed by a firm, it may be treated for practical purposes as a continuous variable.

Some examples will clarify the difference between discrete and continouous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and infinity. However, it could not be any number between 0 and infinity.
  We could not, for example, get 2.3 heads. Therefore, the number of heads must be a discrete variable.

5.1 Qualitative Data

- Qualitative data includes labels or names used to identify an attribute of each element.
- Qualitative data may be numeric (e.g. area codes), but usually it is non-numeric.
- Examples: sex, gender, region, colour, socio-economic status.

5.2 Types of Qualitative Data

Qualitative data may be further split into

1. **Nominal classifications**. Such data is defined by a pure classification, in which the order of the classes has no practical interpretation, e.g. Department: 1-Maths, 2-Equine studies, 3-Sociology.

2. **Ordinal classifications**. The order of the classification is important, e.g.

   ```
   1. Non-smoker
   2. Light Smoker
   3. Heavy Smoker,
   ```

   i.e. the higher a number the more an individual smokes.

It is important to distinguish between quantitative variables and classifications using numeric labels, e.g. the mean of a discrete variable has a sensible interpretation, but not the mean of a nominal, or even ordinal, variable.

6 Sampling

- A sampling frame is a list of members of a population. It may be used to choose a sample.
- A sampling frame may be incomplete or inaccurate. For example, the Irish electoral register will be a complete sampling frame for the population of eligible voters in Ireland. However, it will not be a complete sampling frame for the population of adult Irish residents.
- Choice of an inappropriate sampling frame may well lead to systematic errors in estimates obtained from sampling (bias).
- If I tried to measure the mean mobile data usage of the whole Irish population by just observing a sample of Irish students, I would tend to overestimate this population mean.

6.1 Example

- Suppose we wish to do a study of recent immigrants to Ireland and classify their occupation, e.g. managerial, professional, retail, unemployed.
- We may use the register of PPS numbers given to non-nationals in the past 3 years as a sampling frame.
- Such a sampling frame will not be completely accurate as some of these immigrants will have already left Ireland and some immigrants will not have registered.
- A sample from this population is the set of individuals we choose to observe.
- The main variable observed in this study is the ***class of occupation***.

7 Pharmaceutical Company Study

- A pharmaceutical firm might be interested in conducting an experiment (i.e. a clinical trial) to learn about how a new drug affects blood pressure in adult males.
- To obtain data about the effect of the new drug, researchers select a sample of 50 individuals from a list of volunteers.

For the clinical trial example

Population : all adult males.

Unit : any adult male.

Sample : the 50 individuals.

Parameter : The total number of adult male that respond well to the drug.

Sampling frame : the list of volunteers.

Variable : the blood pressure.

8 Accuracy of Estimates - Bias and Precision

- As described above, if we use an inappropriate sampling frame, our estimates of parameters may have a systematic bias. Bias that results from our method of sampling is called sampling bias. Other sources of bias exist (see later).
- Also, we have random errors depending on the sample actually observed. This determines the precision of a study. Consider the hypothetical situation in which we have a large number of small samples.
- The mean heights in these samples will be rather variable, i.e. a small sample leads to low precision.

8.1 Bias and Precision

- Suppose we had many large samples. The mean heights in these samples will be similar to each other (and if the sampling frame is appropriate will also be similar to the population mean).
- In this case we have low bias and high precision (ideal).
- Now suppose we observe the heights of Irish students in order to estimate the mean height of the population of Irish adults.
- When we have a large number of small samples, the sample means will be rather variable and tend to be larger than the mean height of all Irish adults (large bias and low precision).
- Increasing the size of these samples would increase the precision (the sample means would be more similar), but the bias is unaffected (the mean height will still tend to be overestimated as the sampling frame is inappropriate).

- Hence, increasing the sample size will increase the precision of a study.
- However, increasing sample size leaves the bias unaffected.
- Moral: Results may be misleading, even when we have large sample sizes, as inappropriate methods for choosing samples may lead to systematic bias. (Important.)

9 Non-sampling Bias

- This is a form of bias that occurs when certain groups of individuals have a tendency to give inaccurate responses or not give an answer.
- For example, it was noticed that UK political polls systematically underestimated the support of the Conservative party in the 80s and 90s.
- This may well have been due to the fact that Conservative supporters were more likely to hide their preferences than supporters of other parties.
- The wording of a questionnaire, who the interviewer is and what the interviewee perceives to be the "right answer in a given situation may also lead to bias.
- For example, suppose a questionnaire is carried out on how willing people are to pay extra for ecologically friendly goods. Such a survey will tend to overestimate the proportion of individuals willing to pay extra, as it is politically correct to express such a willingness.
- Suppose in survey I there are the options a) not willing to pay more, b) willing to pay 10% more. Suppose in survey II, option b) is replaced by willing to pay 20% more.
- It is likely that survey II would indicate that on average people are willing to pay more for "ecologically friendly goods (this is also an example of why you should not calculate a mean for data categorised in such a way).