Applied Statistics I

(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

Outline of course unit

1. Collecting data
2. Summarizing data
3. Measures of central tendency
4. Measures of variation
5. Joint distributions of data
6. Linear regression
7. Statistical applications with probability models
8. Statistical software packages

References

- Basic Business Statistics Concept and Applications, Mark L. Berenson (519.5BER).
- Statistics concept and applications, Harray Jrankan, Steven C. Althoen (519.5FRA).
- Sampling Theory, William G. Cochram.
- Applied Statistics for Public Administration, Jeffrey L Bradney.
- www.maths.ruh.ac.lk/∼pubudu

Introduction

- **Statistics** is the study of the collection, organization, and interpretation of data.
- **Mathematical statistics**, which is concerned with the theoretical basis of the subject.
- **Applied statistics** is concerned about application of the subject.

Real world application of statistics

- **Weather Forecasts**
  Computer models are built using statistics that compare prior weather conditions with current weather to predict future weather.
- **Medical Studies**
  Scientists must show a statistically valid rate of effectiveness before any drug can be prescribed.

Real world application of statistics

Cont...

- **Quality Testing**
  Company uses statistics to test just a few, called a sample, of what they make.
- **Stock Market**
  Stock analysts also use statistical computer models to forecast what is happening in the economy.

Why study statistics?

- Statisticians are in demand in all sectors of society, ranging from government, to business and industry, to universities and research labs.
- Today large amounts of data being collected in different fields for different purposes and an understanding of statistics is needed to make sense of it.
- Statistical methods and analyses are often used to communicate research findings and to support hypotheses and give credibility to research methodology and conclusions.

Collecting data

What is data?

- Data is a collection of facts, such as values or measurements.
- It can be numbers, words, measurements, observations or even just descriptions of things.
- There are two main types of data: **quantitative** and **qualitative**.

What is data?
Quantitative data

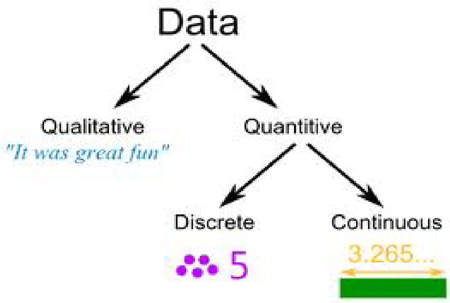
- **Quantitative** data are anything that can be expressed as a numerical value.
- There are two types of quantitative data: **discrete** and **continuous**.

1. Discrete data can only take specific numerical values.
   Eg: number of employee, number of sisters.
2. Continuous data can take any numerical value.
   Eg: weight, height, length.

What is data?
Qualitative data

- **Qualitative** data is data that is not given numerically.
- Qualitative data describe items in terms of some quality or categorization.
  Eg: type of car, favourite color, hair color.

What is data?
Summary



Example

Decide what type of data each of the following would give.

1. Mass of an object
2. Favourite cricket team
3. Price of pencil
4. Distance from home to university
5. Day of the week

Example

Solution

1. Mass of an object ⇒ Continuous quantitative
2. Favourite cricket team ⇒ Qualitative
3. Price of pencil ⇒ Discrete quantitative
4. Distance from home to university ⇒ Continuous quantitative
5. Day of the week ⇒ Qualitative

Collecting primary data

There are many methods of collecting primary data and the main methods include:

- observation
- questionnaires
- interviews
- case-studies

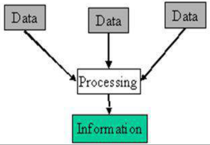
Collecting primary data

Cont...

- One problem with data collection is knowing how much to collect.
- Data collection and processing inevitably costs money and collecting unnecessary data is wasteful.
- In principle there is an optimal amount of data which should be collected for any purpose.
- The optimal amount of data is not usually calculated, but is suggested in the light of previous experience.

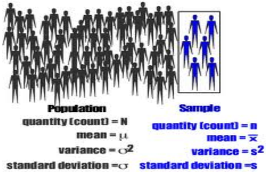
Data versus information

Data	**Information**
Data is a collection of raw facts	Information is the outcome derived after processing the data
May or may not be meaningful	Information is always meaningful
Understanding is difficult	Understanding is easy
Data must be processed to understand	Information is already in understandable form
Data may not be in the order	Information should be in the order



Population versus sample

- The term **population** is used in statistics to represent all possible measurements that are of interest to us in a particular.
- The term **sample** refers to a portion of the population.



Finite and infinite population

- A population is called finite if it is possible to count its individuals.

  Eg: The number of vehicles crossing a bridge every day, the number of births per years.

- Sometimes it is not possible to count the units contained in the population. Such a population is called infinite or uncountable.

  Eg: The number of germs in the body of a patient of malaria is perhaps something which is uncountable.

Hypothetical population

A population that does not exist really but exists only in minds of statisticians is called the hypothetical population.

Homogeneous/Non homogeneous populations

A population that is well mixed with respect to the characteristics we measure is called homogeneous population.

Eg : Taste of a curry.

Non homogeneous population is one that is not mixed well.

Eg : Population of Sri Lanka.

Remark
A population that is homogeneous is uniform in composition or character; one that is heterogeneous lacks uniformity in one of these qualities.

Census versus sample

- A **census** is when you collect data for every member of the group.
- That means study on whole population is considered as a census.
- A **sample** is when you collect data just for selected members of the group.

Census versus sample

Example

- There are 200 students in the second year physical science batch.
- You can ask everyone (all 200) what their age is. That is a **census**.
- Or you could just choose the students that are in the canteen this afternoon. That is a **sample**.

Why do we need sampling?

- Sampling saves money.
- Sampling saves a lot of time and energy.
- It provides information that is almost as accurate as that obtained from a complete census.

Representative and non representative samples

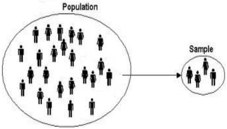
- The sample that represents the corresponding population well is termed as a representative sample. If not it is a non representative sample.
- In statistical sampling, people gather data from a small group and try to extrapolate the results to make generalizations about a larger group.

Approaches to take samples

1. Probability sampling
2. Non probability sampling

Probability sampling

- A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample.
- And also this probability can be accurately determined.

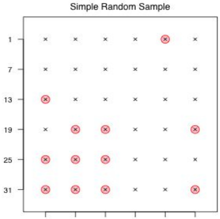


Types of probability sampling

1. Simple random sample
2. Stratified random sample
3. Cluster sample
4. Systematic sample

[1] Simple random sample

- If each unit of the population has an equal chance of being selected for the sample it is called simple random sample.
- The population should be homogeneous and the list of all the items in the population should be available.
- A simple random sample is usually selected by without replacement.



[1] Simple random sample

Cont...

The following methods are used for the selection of a simple random sample:

(i) Lottery method

(ii) Using a random number table

(iii) Using the computer

[1] Simple random sample

(i) Lottery method

- All the units of the population are numbered from 1 to N.
- These numbers are written on the small slips of paper.
- The slips are thoroughly mixed and a slip is picked up.
- Again the population of slips is mixed and the next unit is selected.
- In this manner, the number of slips equal to the sample size n is selected.

[1] Simple random sample

(ii) Using a random number table

- Suppose the size of the population is 80 and we have to select a random sample of 8 units.
- The units of the population are numbered from 1 to 80.
- We read two-digit numbers from the table of random numbers.
- We can take a start from any columns or rows of the table.
- Any number above 80 will be ignored and if any number is repeated, we shall not record it if sampling is done without replacement.

[1] Simple random sample

(iii) Using the computer

- The facility of selecting a simple random sample is available on the most of the statistical software packages.
- We can use statistical software like SPSS, Minitab, R for that purpose.

[1] Simple random sample

Advantages of simple random sampling

- It provides us with a sample that is highly representative of the population being studied.
- It allows us to make generalizations from the sample to the population.

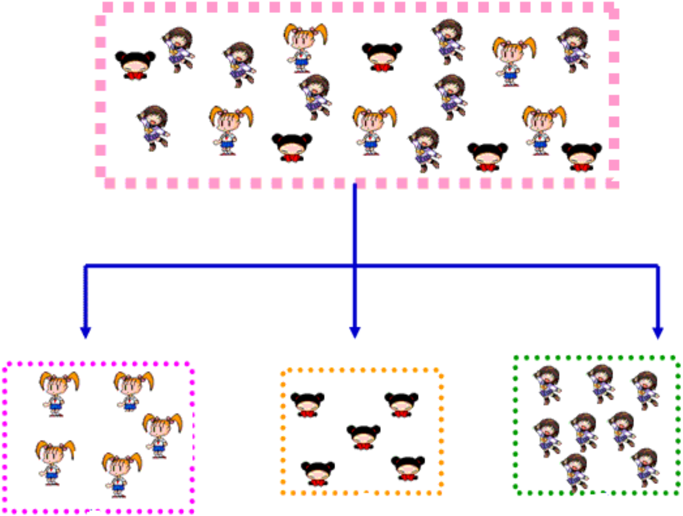
[1] Simple random sample

Disadvantages of simple random sampling

- One of the most obvious limitations of simple random sampling method is its need of a complete list of all the members of the population.

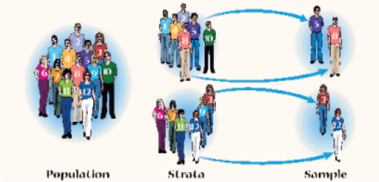
[2] Stratified random sample

Why do we need stratified random sample?



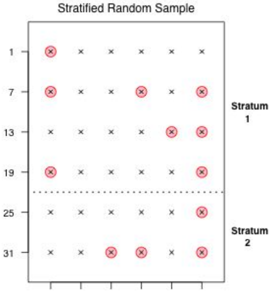
[2] Stratified random sample

- When the population is not homogeneous, the simple random sampling technique is not appropriate.
- We may use stratified random sampling technique, if the population can be separated into subpopulations, which are homogeneous.
- We call these subpopulations as **strata**.
- The sample taken is termed as **stratified random sample**.



[2] Stratified random sample
Cont...

- To select a simple random sample from a stratum we usually use the proportional allocation method.
- That is we select the size of samples proportional to the sizes of strata.



[2] Stratified random sample

Example

Suppose that in a company there are the following staff.

Category	Number of persons
male, full time	90
male, part time	18
female, full time	9
female, part time	63

We are asked to take a sample of 40 staff, stratified according to the above categories.

[2] Stratified random sample

Example⇒Solution

The total number of staff =180.

Calculate the percentage in each group.

$$\text{male, full time} = \frac{90}{180} \times 100 = 0.5 \times 100 = 50\%.$$

$$\text{male, part time} = \frac{18}{180} \times 100 = 0.1 \times 100 = 10\%.$$

$$\text{female, full time} = \frac{9}{180} \times 100 = 0.05 \times 100 = 5\%.$$

$$\text{female, part time} = \frac{63}{180} \times 100 = 0.35 \times 100 = 35\%.$$

[2] Stratified random sample

Example⇒Solution⇒Cont...

This tells us that of our sample of 40,

50% should be male, full time. 50% of 40 is 20.

10% should be male, part time. 10% of 40 is 4.

5% should be female, full time. 5% of 40 is 2.

35% should be female, part time. 35% of 40 is 14.

[2] Stratified random sample

Advantages of stratified random sampling

- It provides greater precision than a simple random sample of the same size.
- Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money.
- A stratified sample can guard against an "unrepresentative" sample (e.g., an all-male sample from a mixed-gender population).

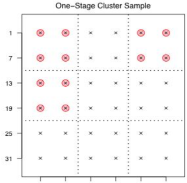
[2] Stratified random sample

Disadvantages of stratified random sampling

- It may require more administrative effort than a simple random sample.

[3] Cluster sample

- The entire population of interest is divided into groups, or clusters, and a random sample of these clusters is selected.
- Each cluster must be mutually exclusive and together the clusters must include the entire population.
- After clusters are selected, then all units within the clusters are selected.
- This differs from stratified sampling, in which some units are selected from each group.



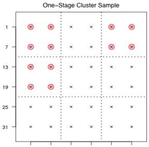
[3] Cluster sample

Cont...

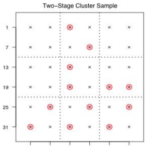
- When all the units within a cluster are selected, the technique is referred to as **one-stage cluster** sampling.

- If a subset of units is selected randomly from each selected cluster, it is called **two-stage cluster** sampling.

- Cluster sampling can also be made in three or more stages: it is then referred to as **multistage cluster** sampling.



(a)



(b)

[3] Cluster sample

Example

Suppose we want to measure the Mathematics knowledge of grade 5 students. If we compare two grade 5 classes in Sri Lanka, the knowledge of students of two such classes may not be much different. However the knowledge of students within a class may vary. The one such class (similar subgroup) is termed as a cluster. By investigating a few cluster, we can estimate the knowledge of students in the whole population.

[3] Cluster sample

Advantages of cluster sampling

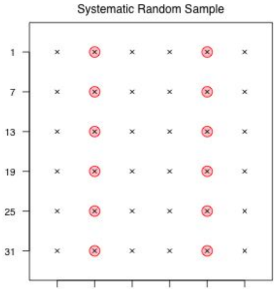
- It is very feasible when you are dealing with large population.
- Here traveling as well as listing efforts will be greatly reduced. So it is economical.

[3] Cluster sample

Disadvantages of cluster sampling

- If the group in population that is chosen as a cluster sample has a biased opinion then the entire population is inferred to have the same opinion. This may not be the actual case.
- The other probabilistic methods give fewer errors than cluster sampling.

[4] Systematic sample



- Methodology for sampling in which units are selected from the population at a regular interval.
- First, the statistician has to select an integer k, which is approximately equal to the ratio of population size and sample size.

[4] Systematic sample

Cont...

- If the population size is unknown we can guess k. Guessing the value of k does not have much effect on sampling.
- Now select any integer i from 1 to k.
- Then select i^{th} unit of the population to the sample.
- Thereafter select every k^{th} unit till we select the desired sample size.
- Thus the sample contains $i, i+k, i+2k, ..., i+(n-1)k$ units of the above serial numbers.

[4] Systematic sample
Example

If there are 120 names on the list. How we obtain a systematic sample of 20 names?

Solution

Let $k = 120/20 = 6$.

Therefore, we would randomly select a number between 1 and 6.

Primary unit selected $= 2$.

After selecting this primary unit, we would include every 6^{th} unit in the sample.

Secondary units in the sample are $8, 14, 20, 26, 32, 38, 44, 50, 56, 62, 68, 74, 80, 86, 92, 98, 104, 110, 116$.

[4] Systematic sample

Advantages of systematic sampling

- The principal advantage of this technique is its simplicity.
- Apart from that the population will be evenly sampled.

[4] Systematic sample

Disadvantages of systematic sampling

- This sampling technique is not necessary if the population items are naturally arranged in a periodic order.

Multistage sampling techniques

- Given a population it is possible to select a sample at one stage or at number of stages.
- In one stage sampling we directly get the measurements/observations from the selected sample.
- In the second stage sampling technique we select a sample from the sample we selected at the first stage.

Multistage sampling techniques

Example

- Suppose we have selected cluster sample at the first stage.
- If it is not possible to study the whole cluster (due to lack of resource/time limits), we can select another sample as second stage sampling by using an appropriate sampling technique.
- This sampling scheme is referred to as two stage sampling. The process can be extended to multistage sampling.

Non-probability sampling

- Non-probability sampling is a sampling technique where the samples are gathered in a process that does not give all the individuals in the population chances of being selected.
- It is not a product of a randomized selection processes.
- The sample may or may not represent the entire population accurately.

Non-probability sampling

Types of non-probability sampling

1. Convenience sampling
2. Consecutive sampling
3. Quota sampling
4. Judgmental sampling
5. Snowball sampling

Thank You