PROBABILIY AND STATISTICS I

LECTURE THREE

Data organization

Lecturer: Dr. Emily Roche

INTRODUCTION

This lecture will focus on basic data organization techniques.

Intended learning outcomes

At the end of this lecture, you will be able to organize data into simple arrays, frequency arrays, frequency distribution tables, stem-and-leaf plots and tables

References

These lecture notes should be supplemented with relevant topics from the book listed in the Bibliography at the end of the lecture.

Organizing and condensing data

The process of organizing and condensing data into meaningful arrangement is a first step of data analysis. To make data for comparison and analysis it should be arranged in an orderly sequence on the basis of similarity.

Data can be arranged in t he following forms

- Arrays
- Frequency tables
- Stem and leaf plot
- Tabulations: Simple tables and cross tabulation

- **Arrays**

Arrays are of two types

Simple arrays and frequency arrays

1. **Simple arrays**

A simple array is an arrangement of data in an ascending or descending order.

This is convenient if the number of data items is small. As the number of items increases, the series become too long and unmanageable. Thus there is need to condense the data.

Frequency arrays is one of the methods of condensing the data

2. **Frequency array**

This is a series that is formed on the basis of frequency with which each item of data is repeated in a series.

Steps in constructing frequency array are:

1. Construct a table with three columns
2. Enter the items of data in the first column in order of magnitude such that an item is recorded only once.
3. Prepare the tally sheet in the second column marking one bar for an item. Make block of five tally bars to avoid mistakes in counting. Every fifth bar is shown by crossing the first four bars like /𝍸/.
4. Count the tally bar and record the total number in the third column to get the frequencies

Just like the simple array when the items of data increase, an alternative method of condensing the data is sort. Thus the frequency distribution table.

- **Frequency distribution tables**

This table lists categories of data items along with their corresponding frequencies.

The frequency for a particular category or class is the number of original data items that fall into that class.

The classes or categories are the groupings of a frequency table

The range (R) is the difference between the highest item of data and the lowest item of data.

The class width or class interval is the difference between class boundaries or two consecutive lower class limits.

The class limits are the smallest or the largest numbers that can belong to different classes

The class boundaries are obtained by increasing or decreasing the class limits by the same margin so that there are no gaps between consecutive classes. The amount to be added or subtracted is half of the difference between the upper limit of one class and the lower limit of the succeeding class.

Class marks/mid value/mid points is the average value of two class limits. It is the representative value of a particular class.

To construct a frequency distribution table, the following must be considered:

- Each data item must belong to exactly one class, thus the categories or classes must be mutually exclusive.
- All categories must be included even if the frequency of the given category is zero.
- All categories should have the same width, however, it may be impossible to avoid open – ended categories such as “120 and above”.
- The number of categories should be between 5 and 20.

The process of constructing a frequency distribution table is:

1. Determine the range of the data.

   R = Highest item of data – Lowest item of data

2. Determine the tentative number of classes (k) to the nearest whole number as:

   $$k = 1 + 3.322 \log N$$

   Note: The actual number of classes may be affected by convenience or other subjective factors

3. Find the class width to the nearest whole number by dividing the range by the number of classes.
4. Set up the classes or categories starting with the lowest item of data as the initial lower class limit. Add the class width to the starting point to get the second lower class limit, then list these limits in a vertical column and stop when the class already includes the highest item of data. Enter the upper class limits which can now be easily identified.
5. Determine the frequency for each class from the tally columns and display the results in the last column of the table.

The main limitation of a frequency distribution table is that it does not give an idea of the characteristics of groups. For instance in a class of "30 – 34", it may not be easy to know the distribution of the data items making it impossible to compare characteristics of different groups. From the frequency distribution tables, the frequency of each class is very easy to see but the original data points are lost. This can be solved by using Stem and leaf plot

- **Stem-and-Leaf Plots**

A stem and leaf plot shows the potential patterns in the data items that may not be clearly visible in the original listing of the data items.

Steps of constructing a stem and leaf plot

Example 1:

Construct a stem-and-leaf plot for the following set of data.

28 13 26 12 20 14 21 16 17 22

17 25 13 30 13 22 15 21 18 18

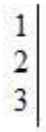
16 21 18 31 15 19

Step 1: Find the least number and the greatest number in the data set.

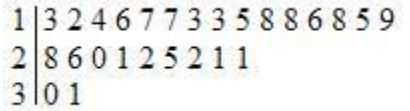
The greatest number is 31 (3 in the tens place)

The smallest number is 12 (1 in the tens place)

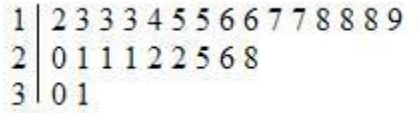
Step 2: Draw a vertical line and write the digits in the tens places in ascending order, meaning from 1 to 3 on the left of the line. The tens digit form the **stems**.



Step 3: Write the units digit to the right of the line. The units' digits form the **leaves**.



Step 4: Rewrite the units' digits in each row in ascending order.



Step 5: Include an explanation/key.

2 | 5 means 25

- **Tabulation: Simple (one way) table and cross tabulation**

A table is a systematic arrangement of statistical data in columns and rows.

Rows are horizontal arrangements while columns are vertical arrangements.

The purpose of a table is to simplify the presentation and to facilitate comparison.

In general, a statistical table consists of the following parts:

i. **Table Number**:

Each table must be given a number to help in distinguishing one table from the other tables.

ii. **Title of the Table**:

Every table should have a short and clear suitable title. The title should be such that one can know the nature of the data contained in the table. The position depends on the chosen writing style.

iii. **Caption**:

This refers to the headings of the columns. It consists of one or more column heads and should be brief, concise and self-explanatory. It is written in the middle of a column in small letters.

iv. **Stub**:

Refers to the headings of rows.

v. **Body**

This is the most important part of a table made up of a number of cells. Cells are formed due to the intersection of rows and column. Items of data are entered into the cells.

vi. **Head Note**:

The head-note (or prefatory note) contains the unit of measurement of the items of data. It is usually placed just below the title or at the right hand top corner of the table.

vii. **Foot Note**

A foot note is given at the bottom of a table to help in clarifying points which may not be clear in the table. It may be keyed to the title or to any column or to any row heading. It is identified by special symbols such as *,+,@,£ etc.

- **Simple (One-Way) Tables**

This type of table displays only one characteristic of the data items.

Example

Table 1. Percentage of employs at management

Base Question	**Percentage**
Product Manager	57%*
Director	12.6%
Product Marketing Manager	24.7%
Program Manager	2.8%
Technical Product Manager	2.8%
Total Counts	215

*percentage to the nearest whole number

- **Cross Tabulation**

In this table, more than one characteristic of the data item is included. It is a good way to compare two subgroups of information. Cross tabs allow you to compare data from two questions to determine if there is a relationship between them.

Table 2. Percentage of employs at management by gender

Base Question	Gender	
	Female	**Male**
Product Manager	57.2%	53.4%
Director	12.6%	14.2%
Product Marketing Manager	24.7%	23.1%
Program Manager	2.8%	1.5%
Technical Product Manager	2.8%	7.7%
Total Counts	215	337

Cross tabs are used most frequently to look at answers to a question among various demographic groups. The intersections of the various columns and rows, commonly called cells, are the percentages of people who answered each of the responses.

Example:

In a certain contest, there were 50 participants. 30 of them were male participants, twenty of them exceled in the contest with 10 of them being female. Present the above information in a table form.

Answer

Contest performance	Gender		
	Male	Female	Total
Exceled	20	**10**	30
Failed to excel	10	10	20
Total	**30**	20	**50**

Individual Assignment

In a sample study about the teas habits, the following data were observed

70% persons were male

80% were tea drinkers

62 % were male tea drinkers

Tabulate the above information

Answer to the individual assignment

Study about tea habits

Tea habits	Gender		
	male	female	total
Tea drinkers	62	**18**	80
None tea drinkers	8	22	20
total	**70**	30	**100**

Bibliography

Gupta, SP (Dr.), (2014). *Statistical methods* (43rd Ed.). Sultan Chand & Sons.

S. C. Gupta and V. K. Kapoor, (2020). *Fundamentals of mathematical Statistics* (12th Ed). Sultan Chand & Sons.