Summarizing Data

Md. Aminul Islam Shazid

Outline

1. Introduction
2. Tabular Summary

Introduction

Summarizing Data

- Raw data by itself is not useful
- Need to extract insight from data
- Data needs to be summarized to gain insights
- Can summarize in two ways:
  - Tabular summary
  - Graphical summary

Tabular Summary

Tabular Summary

Frequency Table

Frequency Table

- Organizes data from the sample by listing distinct values or classes
- Shows the frequency of each class in the sample
- Can be constructed for categorical or numerical data:
  - For categorical variable, it shows the number of observations in each category
  - For discrete numeric variable, it shows how many times each value has been observed
  - For continuous numeric variable, classes or bins are formed first, then the table shows how many values fall in each class
- Forms the basis for graphical summaries like bar charts and histograms

Example: Frequency Table for Categorical Variable

Suppose we have a discreete variable with four categories: A, B, C and D. The data in the sample is: A, B, A, D, A, A, B, C, C, B, A, A, C, D, D, B, D, C, C, B, C. The Frequency table would be:

Category	Tally	Frequency
A	𝍸	6
B	𝍸	5
C	𝍸	6
D	𝍷	4

The tally column is only used to keep track of the data when counting by hand.

One can add a percentage column in necessary.

Frequency Table for continuous Numeric Variable

- Data is grouped into classes or groups
- Each class has the same class interval (the difference between the upper limit and the lower limit of each class)

Before making the table, need to decide the value of:

- Either the number of classes or groups
- Or, the class interval for each class or group

Frequency Table for continuous Numeric Variable (cont.)

There is no hard and fast rules for setting the classes or the class interval.

It usually depends on the variable and its range (difference between the highest value and the lowest value).

- For example, if the range is 50, then with class interval of 10, there will be 5 classes
- Conversely, if 10 classes are wanted, then the class interval will be 5

Too many or too few classes may fail to reveal the basic shape of the data.

Frequency Table for continuous Numeric Variable (cont.)

Sometimes, the classes are determined through subject-matter considerations.

For example, in case of vitamin D levels:

- Values lower than 10 ng/mL would be classified as deficient
- 10 to 30 is insufficient
- 30 to 100 is sufficient
- Above 100 is considered toxic

Example

Make a frequency table with the following data: 30, 40, 5, 110, 11, 15, 55, 20, 130, 45, 30, 47, 52, 68, 105, 62, 52, 98, 76, 85, 83, 91, 49, 38, 57, 27, 23, 42, 9, 65

Solution:

The range in the sample is = highest value - lowest value = 130 - 5 = 125

Therefore, with 5 classes, the class interval would = 125/5=25

Example (cont.)

Category	Tally	Frequency
5 - 30	𝍸 ||	7
30 - 55	𝍸 𝍸	10
55 - 80	𝍸 |	6
80 - 105	||||	4
105 - 130	|||	3

For each class, the upper limit is excluded and the lower limit is included, except for the last class in which the upper limit is also included. This is called upper limit exclusive and lower limit inclusive.

This is necessary because the variable is continuous, so there is no gap between the values of each class.

Tabular Summary

Contingency Table

Contingency Table

- A contingency table is used to summarize the relationship between two (or more) categorical variables
- Data are arranged in rows and columns, where each cell contains the frequency for a specific combination of categories
- Row totals, column totals, and a grand total are often included
- It helps identify patterns, associations, or possible dependence between variables

Example: Contingency Table

	Lung Cancer		
Smoking	**Yes**	**No**	**Total**
Smoker	45	55	100
Non-smoker	10	90	100
Total	55	145	200

In the above, it can be seen that lung cancer percentage is higher among the smokers compared to those who do not smoke.

Questions?