Data Management

MDM4U

Qinghao Hu

October 10, 2025

Contents

Chapter 3

One Variable Statistics

3.1 Variables and Data

3.1.1 Definitions

Categorical variables represent data that are generally grouped into categories, and are also known as qualitative variables

Ordinary variables are categorical variables whose data has a natural order but the difference between values cannot be determined or is not meaningful

Nominal variables describe names, labels, or categories that have no natural order

Quantitative variables describe data values that are numerical, and are also known as numerical variables

Continuous variables are numerical variables which can assume an infinite number of values in a given interval

Descrete variables are numerical variables that only take on a finite number of possible values in a given interval

Primary data are data that are collected by the statisticians who are analyzing the data, from first-hand sources such as surveys or experiments

Secondary data are data that the statisticians who are analyzing the data did not participate in the first hand data collection process (ie Surveys or experiments)

Microdata contains records for each individual surveyed

Aggregate or summary data are data that are combined or summarized in such a way that the individual microdata can no longer be determined.

Data gathered from a **cross** sectional study considers individuals from different groups at the same time

Data gathered from a **longitudinal** study considers how the characteristics of a specific sample changes over time

An **index** is a continous variable such that it is an arbitrarity defined number that provides a measure of scale. It is used to related the values of a variable to a base level

The **consumer** price index, CPI, provides a broad picture of the cost of living in Canada by comparing the cost of a wide variety of consumer goods, such as food, clothing, fuel, heating cost, transportation, shelter, and recreation

Health officials use the **body** mass index to determine whether a person is overweight. The BMI is calculated by dividing a person's mass in kilograms by the square of their height in meters

3.2 One Variable Graphs

3.2.1 Some Definitions

A **Frequency** bar (column) graph is a visual display of data in which quantities are represented by bars of equal width, typically used with categorical or discrete data

A **CIRCLE** graph or **PIE** chart contains a circle divided into sectors whose areas are proportional to the categories represented. It is used to show how each category is compared to the whole

A **PICTOGRAPH** is a graph that uses pictures or symbols to represent categorical quantities. It's advantage is being visually appealing, hence it is the most often used graphical format. However, it may be difficult to present exact values when using the format, depending on the data given

A **STEN** and **LEAF** plot can be created easily to see the distribution of a set of numerical data. However, its appearance is not as scientific as a histogram.

A **HISTOGRAM** is used to represent numerical data or data organized using intervals. The bars of a histogram are attached and each bar is placed between two intervals endpoints. The area of each bar is proportional to the frequency of data in the interval. Typically, 5 15 intervals/bins of equal length are used and every piece of data must fall into exactly one bin. The width of each bin is the **bin width**

Values of a continuous varialbe can be grouped into intervals in the form of $(a,b]$ such that this interval includes all values from a to b, including a but excluding b.

A bin width of **5** units witht he first bin being $[10, 15)$ is reasonable if a set of continuous data has 26 values, a minimum of 12