TYPES OF DATA

The methods for displaying and analyzing data depend upon the type of data being used. In this section, we will define and provide examples of the two major types of data: qualitative and quantitative. Quantitative data can be continuous or discrete. Chapter 11 will give more information about the related topic of measurement systems. We collect data to characterize populations and to estimate parameters, which are numerical or categorical characteristics of a population probability distribution.

In order to describe types of data, we need to be familiar with the concept of variables. The term "variable" is used to describe a quantity that can vary (i.e., take on various values), such as age, height, weight, or sex. Variables can be characteristics of a population, such as the age of a randomly selected individual in the U.S. population. They can also be estimates (statistics) of population parameters such as the mean age of a random sample of 100 individuals in the U.S. population.

Qualitative Data

Variables that can be identified for individuals according to a quality are called qualitative variables. These variables place individuals into categories that do not have numerical values. When the observations are not ordered, they form a nominal scale. (A dichotomous scale—true/false, male/female, yes/no, dead/alive—also is a nominal scale.) Many qualitative variables cannot be ordered (as in going from worst to best). Occupation, marital status, and sex are examples of qualitative data that have no natural ordering. The term nominal refers to qualitative data that do not have a natural ordering.

Some qualitative data can be ordered in the manner of a preference scale (e.g., strongly agree, agree, disagree, strongly disagree). Levels of educational attainment can be ordered from low to moderate to high: less than a high school education might be categorized as low; education beyond high school but without a four year bachelor's degree could be considered moderate; a four year bachelor's degree might be considered high; and a degree at the masters, Ph.D., or M.D. level considered very high. Although still considered qualitative, categorical data that can be ordered are called ordinal.

Qualitative data can be summarized and displayed in pie charts and bar graphs, which describe the frequency of occurrence in the sample or the population of particular values of the characteristics. These graphical representations will be described in Section 3.3. For ordinal data with the categories ordered from lowest to highest, bar graphs might be more appropriate than pie charts. Because a pie chart is circular, it is more appropriate for nominal data.

Quantitative Data

Quantitative data are numerical data that have a natural order and can be continuous or discrete. Continuous data can take on any real value in an interval or over the whole real number line. Continuous data can be classified as interval. Continuous data also can be summarized with box-

and-whisker plots, histograms, frequency polygons, and stem-and-leaf displays. Examples of continuous data include variables such as age, height, weight, heart rate, blood pressure, and cholesterol level.

Discrete data take on only a finite or countable (equivalent to the set of integers) number of values. Examples of discrete data are the number of children in a household, the number of visits to a doctor in a year, or the number of successful ablation treatments in a clinical trial. Often, discrete data are integers or fractions. Discrete data can be described and displayed in histograms, frequency polygons, stem-and leaf displays, and box-and-whisker plots.

If the data can be ordered, and we can identify ratios with them, we call the data ratio data. For example, integers form a quantitative discrete set of numbers that are ratio data; we can quantify 2 as being two times 1, 4 as two times 2, and 6 as three times 2. The ability to create ratios distinguishes quantitative data from qualitative data. Qualitative ordinal data can be ordered but cannot be used to produce ratios. We cannot say, for example, that a college education is worth twice as much as a high school education.

Continuous interval data can be used to produce ratios but not all ratio data are continuous. For example, the integers form a discrete set that can produce ratios, but such data are not interval data because of the gaps between consecutive integers.

FREQUENCY TABLES

A frequency table provides one of the most convenient ways to summarize or display grouped data. Before we construct such a table, let us consider the following numerical data. Table 3.1 lists 120 values of body mass index data from the 1998 National Health Interview Survey. The body mass index (BMI) is defined as [Weight (in kilograms)/Height (in meters) squared]. According to established standards, a BMI from 19 to less than 25 is considered healthy; a BMI from 25 to less than 30 is regarded as overweight; a BMI greater than or equal to 30 is defined as obese. Table 3.1 arranges the numbers in the order in which they were collected.

In constructing a frequency table for grouped data, we first determine a set of class intervals that cover the range of the data (i.e., include all the observed values). The class intervals are usually arranged from lowest numbers at the top of the table to highest numbers at the bottom of the table and are defined so as not to overlap. We then tally the number of observations that fall in each interval and present that number as a frequency, called a class frequency. Some frequency tables include a column that represents the frequency as a percentage of the total number of observations; this column is called the relative frequency percentage. The completed frequency table provides a frequency distribution.

TABLE 3.1. Body Mass Index for a Sample of 120 U.S. Adults

27.4	31.0	34.2	28.9	25.7	37.1	24.8	34.9	27.5	25.9
23.5	30.9	27.4	25.9	22.3	21.3	37.8	28.8	28.8	23.4
21.9	30.2	24.7	36.6	25.4	21.3	22.9	24.2	27.1	23.1
28.6	27.3	22.7	22.7	27.3	23.1	22.3	32.6	29.5	38.8
21.9	24.3	26.5	30.1	27.4	24.5	22.8	24.3	30.9	28.7
22.4	35.9	30.0	26.2	27.4	24.1	19.8	26.9	23.3	28.4
20.8	26.5	28.2	18.3	30.8	27.6	21.5	33.6	24.8	28.3
25.0	35.8	25.4	27.3	23.0	25.7	22.3	35.5	29.8	27.4
31.3	24.0	25.8	21.1	21.1	29.3	24.0	22.5	32.8	38.2
27.3	19.2	26.6	30.3	31.6	25.4	34.8	24.7	25.6	28.3
26.5	28.3	35.0	20.2	37.5	25.8	27.5	28.8	31.1	28.7
24.1	24.0	20.7	24.6	21.1	21.9	30.8	24.6	33.2	31.6

Source: Adapted from the National Center for Health Statistics (2000). Data File Documentation, National Health Interview Survey, 1998 (machine readable data file and documentation, CD-ROM Series 10, No 13A), National Center for Health Statistics, Hyattsville, Maryland.

Although not required, a good first step in constructing a frequency table is to rearrange the data table, placing the smallest number in the first row of the leftmost column and then continuing to arrange the numbers in increasing order going down the first column to the top of the next row. (We can accomplish this procedure by sorting the data in ascending order.) After the first column is completed, the procedure is continued starting in the second column of the first row, and continuing until the largest observation appears in the rightmost column of the bottom row.

We call the arranged table an ordered array. It is much easier to tally the observations for a frequency table from such an ordered array of data than it is from the original data table. Table 3.2 provides a rearrangement of the body mass index data as an ordered array.

In Table 3.2, by inspection we find that the lowest and highest values are 18.3 and 38.8, respectively. We will use these numbers to help us create equally spaced intervals for tabulating frequencies of data. Although the number of intervals that one may choose for a frequency distribution is arbitrary, the actual number should depend on the range of the data and the number of cases. For a data set of 100 to 150 observations, the number chosen usually ranges from about five to ten. In the present example, the range of the data is 38.8 – 18.3 = 20.5. Suppose we divide the data set into seven intervals. Then, we have 20.5 ÷ 7 = 2.93, which rounds to 3.0. Consequently, the intervals will have a width of three. These seven intervals are as follows:

1. 18.0 – 20.9

2. 21.0 – 23.9

3. 24.0 – 26.9

4. 27.0 – 29.9

5. 30.0 – 32.9

6. 33.0 – 35.9

7. 36.0 – 38.9

TABLE 3.2. Body Mass Index Data for a Sample of 120 U.S. Adults: Ordered Array (Sorted in Ascending Order)

18.3	21.9	23.0	24.3	25.4	26.6	27.5	28.8	30.9	34.8
19.2	21.9	23.1	24.3	25.6	26.9	27.5	28.8	30.9	34.9
19.8	21.9	23.1	24.5	25.7	27.1	27.6	28.9	31.0	35.0
20.2	22.3	23.3	24.6	25.7	27.3	28.2	29.3	31.1	35.5
20.7	22.3	23.4	24.6	25.8	27.3	28.3	29.5	31.3	35.8
20.8	22.3	23.5	24.7	25.8	27.3	28.3	29.8	31.6	35.9
21.1	22.4	24.0	24.7	25.9	27.3	28.3	30.0	31.6	36.6
21.1	22.5	24.0	24.8	25.9	27.4	28.4	30.1	32.6	37.1
21.1	22.7	24.0	24.8	26.2	27.4	28.6	30.2	32.8	37.5
21.3	22.7	24.1	25.0	26.5	27.4	28.7	30.3	33.2	37.8
21.3	22.8	24.1	25.4	26.5	27.4	28.7	30.8	33.6	38.2
21.5	22.9	24.2	25.4	26.5	27.4	28.8	30.8	34.2	38.8

Table 3.3 presents a frequency distribution and a relative frequency distribution (%) of the BMI data.

A cumulative frequency (%) table provides another way to display a frequency distribution. In a cumulative frequency (%) table, we list the class intervals and the cumulative relative frequency (%) in addition to the relative frequency (%). The cumulative relative frequency or cumulative percentage gives the percentage of cases less than or equal to the upper boundary of a particular class interval. The cumulative relative frequency can be obtained by summing the relative frequencies in a particular row and in all the preceding class intervals. Table 3.4 lists the relative frequencies and cumulative relative frequencies for the body mass index data.

TABLE 3.3. Body Mass Index (BMI) Data (n = 120)

Class Interval for BMI Levels	Frequency (f)	Cumulative Frequency (cf)	Relative Frequency (%)
18.0–20.9	6	6	5.00
21.0–23.9	24	30	20.00
24.0–26.9	32	62	26.67
27.0–29.9	28	90	23.33
30.0–32.9	15	105	12.50
33.0–35.9	9	114	7.50
36.0–38.9	6	120	5.00
Total	120	—	100.00

TABLE 3.4. Relative Frequency Table of BMI Levels

Class Interval for BMI Levels	Relative Frequency (%)	Cumulative Relative Frequency (%)
18.0–20.9	5.00	5.00
21.0–23.9	20.00	55.00
24.0–26.9	26.67	51.67
27.0–29.9	23.33	75.00
30.0–32.9	12.50	87.50
33.0–35.9	7.50	95.00
36.0–38.9	5.00	100.00
Total	100.00	100.00

A histogram presents the same information as a frequency table in the form of a bar graph. The endpoints of the intervals are displayed as the x-axis; on the y-axis the frequency is represented, shown as a bar with the frequency as the height. We call a histogram a relative frequency histogram if we replace the frequency on the y-axis with the relative frequency expressed as a percent.

Table 3.5 summarizes Section 3.2 by providing guidelines for creating frequency distributions of grouped data.

TABLE 3.5. Guidelines for Creating Frequency Distributions from Grouped Data

1. Find the range of values—the difference between the highest and lowest values.
2. Decide how many intervals to use (usually choose between 6 and 20 unless the data set is very large). The choice should be based on how much information is in the distribution you wish to display.
3. To determine the width of the interval, divide the range by the number of class intervals selected. Round this result as necessary.
4. Be sure that the class categories do not overlap!
5. Most of the time, use equally spaced intervals, which are simpler than unequally spaced intervals and avoid interpretation problems. In some cases, unequal intervals may be helpful to emphasize certain details. Sometimes wider intervals are needed where the data are sparse.

Graphic presentation of frequency distributions

Frequency distributions summarize the number of times each possible value appears in a dataset. Presenting this information visually can be incredibly helpful for understanding the data's central tendency, spread, and shape. Here are some common graphical methods for presenting frequency distributions:

For continuous data:

- **Histogram:** This uses vertical bars of varying heights to represent the frequency of data points falling within specific intervals. It's ideal for continuous data and reveals the overall shape of the distribution (symmetrical, skewed, multimodal etc.).
- **Frequency Polygon:** This connects data points at the midpoints of each interval with straight lines, forming a closed polygon. It provides a smooth visual representation of the distribution similar to a histogram but may obscure details for highly uneven distributions.
- **Line Plot:** For smaller datasets, individual data points can be plotted directly on a number line, with dots positioned at their corresponding values. This offers a clear view of each data point but becomes impractical for large datasets.

For discrete data:

- **Bar Chart:** This uses vertical bars of varying heights to represent the frequency of each different category or value. It's suitable for comparing the occurrence of specific categories and works well for both small and large datasets.
- **Pie Chart:** This divides a circle into slices proportional to the frequency of each category, visually showing the relative contribution of each category to the whole. However, pie charts can be challenging to interpret accurately for more than 4-5 categories.

Frequency Histograms

As we mentioned previously, a frequency histogram is simply a bar graph with the class intervals listed on the x-axis and the frequency of occurrence of the values in the interval on the y-axis. Appropriate labeling is important. For the BMI data described earlier, Figure 3.1 provides an appropriate example of a frequency histogram.

Figure 3.1. Frequency histogram of the BMI data.

Figure 3.2 provides a graph, called a relative frequency histogram, of the same data as in Figure 3.1 with the height of the y-axis represented by the relative frequency (%) rather than the actual frequency. By comparing Figures 3.1 and 3.2, you can see that the shapes of the graphs are similar.

Here the magnitude of the relative frequency is determined strictly by the height of the bar; the width of the bar should be ignored. For equally spaced class inter, the height of the bar multiplied by the width of the bar (i.e., the area of the bar) also can represent the proportion of the cases in the given class.

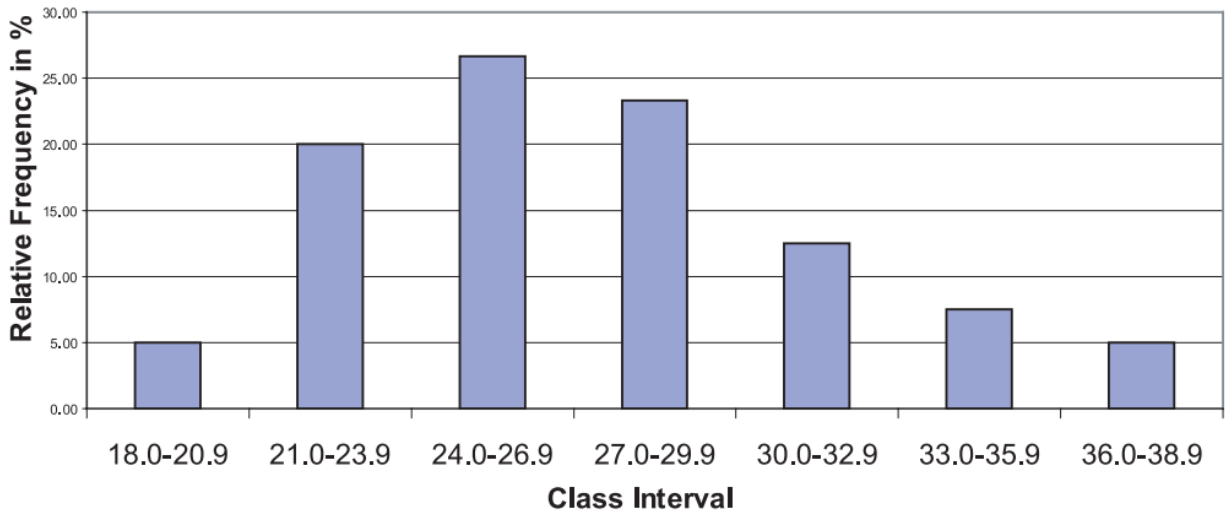

Figure 3.2. Relative frequency histogram for the BMI data.

Bar Diagram

This is known as dimensional diagram also. Bar diagram is most useful for categorical data. A bar is defined as a thick line. Bar diagram is drawn from the frequency distribution table representing the variable on the horizontal axis and the frequency on the vertical axis. The height of each bar will be corresponding to the frequency or value of the variable. However , width of the rectangles is immaterial but proper and uniform spacing should be between different bars. It is different from the histogram when both the height and width of the bar are important and even bars are placed adjacent to one another without any gap.

Example: In a study on causes of strikes in mills. Hypothetical data are given below.

Causes of strikes :	Economic	Personal	Political	Rivalry	Others
Occurrence of strikes:	45	13	25	7	10

Let us take the above example to demonstrate the construction of a bar diagram.

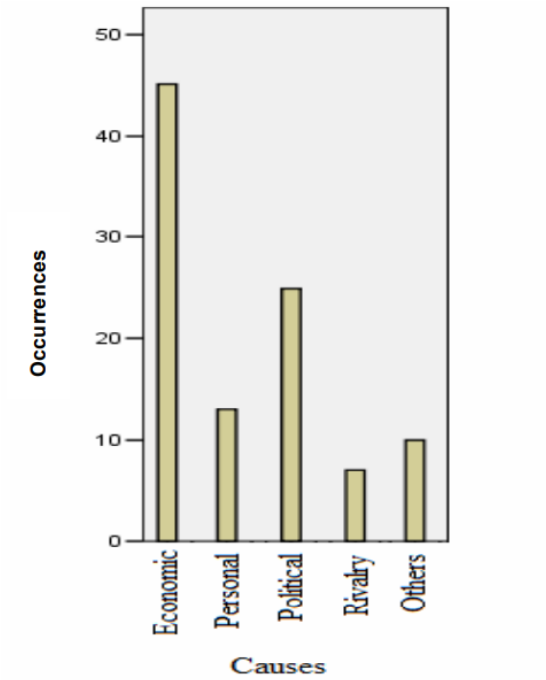

Fig. 4.4: Bar diagram

Pie Diagram

It is also known as angular diagram. A pie chart or diagram is a circle divided into component sectors corresponding to the frequencies of the variables in the distribution. Each sector will be proportional to the frequency of the variable in the group. A circle represent 3600. So 360 angle is divided in proportion to percentages. The degrees represented by the various component parts of given magnitude can be obtained by using this formula.

$$\text{Degree of any component part} = \frac{\text{Component Value}}{\text{Total Value}} \times 360°$$

After the calculation of the angles for each component, segments are drawn in the circle in succession corresponding to the angles at the center for each segment. Different segments are shaded with different colour, shades or numbers.

Table 4.11: 1000 software engineers pass out from a institute X and they were placed in four different company in 2009.

Company	Placement
A	400
B	200
C	300
D	100

Pie Diagram Representing Placement in four different company.

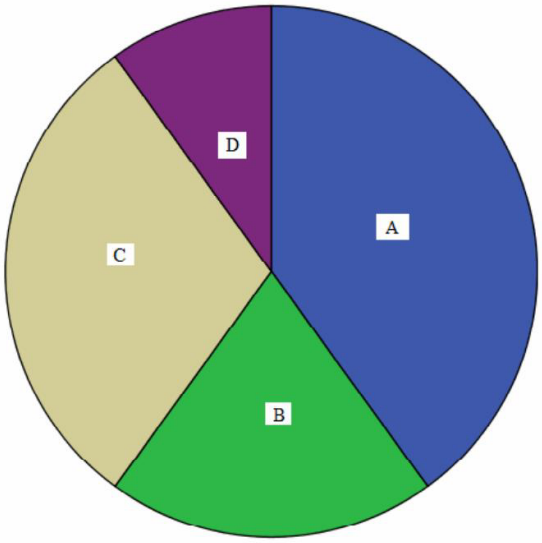

Fig.4.7: Pie diagram representing placement in four companies

Merits of Graphical Presentation of Frequencies:

• **Enhanced understanding:** Visualizing data through graphs often makes it easier to grasp patterns and relationships within the data compared to raw numbers. The human brain excels at processing visual information, allowing for quicker identification of trends, outliers, and central tendencies. Imagine you have data on the frequency of different eye colors in a population. A simple bar chart like this can instantly reveal which color is most common, allowing even someone with no statistical background to grasp the main trend.

• **Comparative analysis:** Graphs facilitate effective comparison of frequencies across different categories or groups. Bar charts, histograms, and line graphs are particularly useful for showing how various values are distributed or how they change over time.

• **Communication and engagement:** Visual representations are generally more engaging and impactful than numerical tables. They can easily capture the attention of viewers and effectively communicate information to a wider audience, even those with limited statistical knowledge.

• **Identification of patterns and trends:** Visualizations can reveal subtle patterns and trends that might be overlooked in numerical data. For example, a histogram might highlight a bimodal distribution that wouldn't be immediately apparent from just looking at the numbers. A histogram of income levels in a city might reveal a bimodal distribution, suggesting two distinct income groups.

• **Flexibility and customization:** Various types of graphs exist, each suited for different purposes. Choosing the right graph type can maximize the clarity and effectiveness of the data presentation.

Demerits of Graphical Presentation of Frequencies:

• **Loss of precision:** Graphs often provide an overview of the data rather than exact values. While trends and patterns are readily apparent, precise numerical readings might be difficult to extract directly from the graph, requiring consultation with the underlying data.

• **Misinterpretation and subjectivity:** While visualizations are generally understood, their interpretation can be subjective and influenced by individual biases or preconceived notions. Choosing the wrong graph type or using misleading scales can lead to misinterpretations of the data.

A pie chart can be visually impactful, but if not carefully designed, it can be misleading. For example, using 3D effects or distorting slice sizes can exaggerate certain values.

• **Complexity and clutter:** Overly complex graphs with too many elements can become cluttered and difficult to read. It's important to strike a balance between informative graphics and visual simplicity to avoid overwhelming the viewer.

• **Limited applicability:** Not all types of data are well-suited for graphical representation. Complex relationships or highly nuanced data might be better represented through tables or mathematical models.

• **Accessibility challenges:** Individuals with visual impairments or color blindness might face difficulties in interpreting certain types of graphs. Choosing accessible color palettes and providing alternative representations (e.g., audio descriptions) can improve inclusivity.

Overall, graphical presentations of frequencies offer a valuable tool for data analysis and communication. However, it's essential to be aware of both their merits and limitations to ensure effective interpretation and communication of the information they convey. Choosing the right type of graph, ensuring clarity and accuracy, and considering accessibility can maximize the impact of graphical representations of frequencies.