

Module 2

Describing Data

Objective: At the completion of this module you will learn how to summarize data using statistical tables and graphs and also using various summary statistics.

2.1 Introduction

In Module 1 we stated that the four scales of measurements that occur most often in medical data are nominal, ordinal, discrete and continuous. When such measurements of a variable are taken on the entities of a population or sample, the resulting values are made available to the researcher or statistician as a mass of unordered data. Such measurements that have not been organised, summarised, or otherwise manipulated are called raw data (see appendix, Table 2.5). This unordered data does not convey much information unless the number of observations is extremely small. In this module we discuss several techniques for presenting and summarising raw data so that we can more easily determine the information they contain. Data is usually summarized by presenting it in frequency tables and graphs and by calculating descriptive statistics.

In this module we discuss different methods those are commonly used for summarizing data. This module includes the following topics:

- **Frequency Tables & Graphs**
 - **Frequency Table**
 - **Bar Chart, Pie Chart, Histogram , Scatter plot and Box-plot**
- **Shape of Data**
 - **Symmetric and asymmetric (right skewed, left skewed)**
- **Descriptive Statistics**
 - **Central Tendency**
 - **Mean, Median, Mode and Percentile**
 - **Measures of Dispersion**
 - **Range, Inter-quartile Range and Standard Deviation**

2.2 Frequency Tables and Graphs

Data can be presented in tables or graphs to organize them into a compact and readily comprehensible form. A frequency distribution table gives the number of observations at different values of the variable; frequency is the number of times each observation occurs (repeats) in a data set. For example consider the fasting glucose level of 5 diabetic patients: 107, 145, 237, 145 and 91 mg/100. The number 145 occurs twice, so

the frequency of 145 mg/100 is 2; each of the other observations have a frequency of one. Construction of a frequency table for a small data set (a small number of patients in the study) is fairly straight forward. However, for a large numerical data set as we will see the variable requires grouping of the data into classes or groups called **class intervals**. For categorical data usually grouping is avoided unless the number of categories is very large.

Graphs or charts are another useful way to present the data. Graphs bring out the overall pattern of the data more clearly as compared to frequency tables. The commonly used graphs in statistics are bar chart, pie chart, histogram, scatter plot and box-plot. Bar and pie charts are commonly used for categorical data and for numerical data histogram, scatter plot and box-plot are appropriate. However, the numerical data organised in a group frequency table (you will learn later about group frequency tables) is sometimes presented in a pie chart.

2.2.1 Table and Charts for Categorical Data

Categorical data are usually summarised in frequency tables and diagrams such as Bar Charts and Pie Charts. Let us consider the Honolulu Heart Study data for 100 patients which has been presented in Table 2.5 (see appendix). In fact, this data is selected randomly from the Honolulu Heart Study population of 7683 patients. The variables recorded for each patient are: education level, weight, height, age, smoking status, physical activity at home, blood glucose, serum cholesterol and systolic blood pressure (see Table 2.6 for detailed description of these variables). It should be noted that usually in medical data collection processes many more variables are recorded than those shown in Table 2.5. Suppose we are interested in the number of patients by education levels. For a large data set it is difficult to know, for example, the number of patients with a “Primary Education Level” unless we summarise the data into a frequency table. The data should be arranged in the form of a table, showing the frequency with which each category of education level (none, primary, intermediate, senior high, technical school and university) was mentioned.

Frequency, as mentioned earlier, is the number of cases (observations/patients/subjects) in each category. Columns 1-2 in Table 2.1 present the frequency distribution for various education levels, and the last column shows the relative frequency. The relative frequency is calculated by dividing the frequency of a class by the total number of patients in the sample. For example, the relative frequency for the education level “Senior School” in Table 2.1 is obtained by dividing the corresponding frequency of 9 by the total number of patients (100 patients) which is 0.09 or 9%.

Table 2.1: Frequency for education level for 100 patients

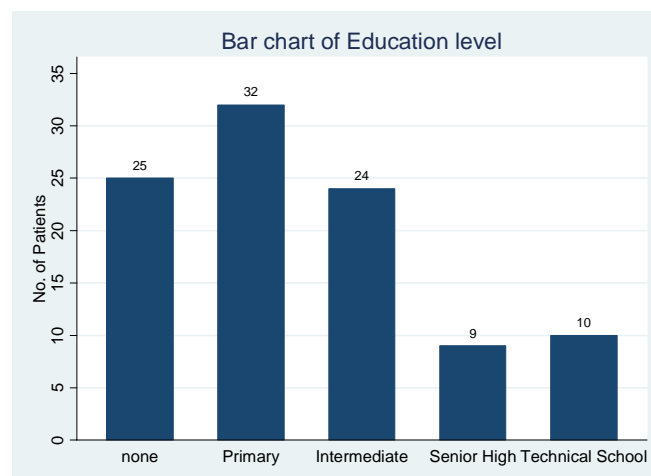
Education Level	Frequency	Relative Frequency (RF)
None	25	$25/100 = 25\%$
Primary	32	$32/100 = 32\%$
Intermediate	24	$24/100 = 24\%$
Senior school	9	$9/100 = 9\%$
Technical School	10	$10/100 = 10\%$
University	0	$0/100 = 0.0\%$
Total	100	$100\% = 1$

Note: Relative frequency lies between 0 and 1 and sum of all relative frequencies must be equal to 1 or 100%.

Bar Charts

The data in Table 2.1 can be presented in a Bar Chart (Bar Diagram) where each bar is proportional in height to the number or percentage of patients in a category. Figure 2.1 shows the bar diagram for the number of patients (equivalently, percentage of patients) in each category of education level in Table 2.1. Clearly, a bar chart is easier to follow than the frequency table- just a quick look at the bar chart gives an idea about the data for which it is created. Figure 2.1 shows that the highest number of patients has primary education and the lowest number of patients has senior high education level. Only 10 patients completed technical school and no one has completed university degree.

Figure 2.1: Bar diagram for patients by education level

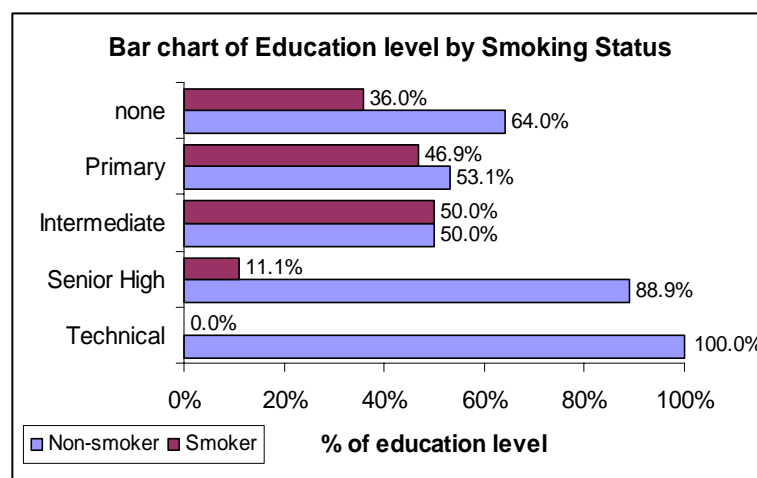


It is always recommended to present the data in percentages particularly when we compare data of different populations, e.g., when comparing the mortality of cardiac surgery patients in all public hospitals in Australia.

Bar charts are appropriate when:

- Comparing various categories of a variable, for example, compare education levels for the Honolulu Heart Study data (see Figure 2.1).
- Comparing categories of one variable by the categories of another variable, for example, compare the smoking status by education level of patients for the Honolulu Heart Study data (see Figure 2.1a).

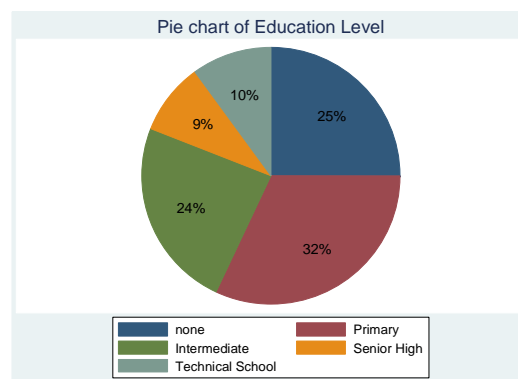
Figure 2.1a: Comparing smoking status by patient’s education level



Pie Chart

Another way of illustrating categorical data, for example data in Table 2.1, is to use a Pie Chart. Here a circle is divided into slices, with the angle each makes at the centre of the circle being proportional to the relative frequency in the category concerned. A pie chart for the relative frequency for education level in Table 2.1 is shown in Figure 2.2. Like bar chart, pie charts are also very often used in medical data presentation.

Figure 2.2: Pie Chart for the education level of patients



2.2.2 Tables and Graphs for Numerical Data

Let us consider the Honolulu Heart Study data presented in Table 2.5. Suppose that we are interested in the cholesterol level of patients, however, the sample data presented in Table 2.5 does not give a good idea about the cholesterol level of patients because there is too much detail. Researchers may ask, for example, is there any pattern in cholesterol level, what is the general picture, how easily one can pick up the maximum and minimum cholesterol levels, are cholesterol levels spread out evenly between the minimum and maximum values, etc.? Answers to the above questions can easily be obtained by summarising the data in frequency tables and diagrams. The data is presented in a group frequency distribution table unless the number of observations in the sample is very small. The general method of constructing a group frequency table is as follows:

- Break the range of measurements into intervals of equal width (called “class interval” or “bins”) and
- Count the number falling in each interval.

Note: Be careful in your choice of the width of class intervals, they should neither be too wide nor too narrow. The number of intervals should be between 5 and 10 – usually each interval has the same width.

A group frequency distribution for the cholesterol level data in Table 2.5 is presented in Table 2.2. The first and second columns shows the class intervals and frequency respectively. The frequency table is constructed by excluding the upper limit from each class interval. This means, for example, if a patient has cholesterol level of 175 mg/100, he/she should be counted in the interval ≥ 175 & < 200 .

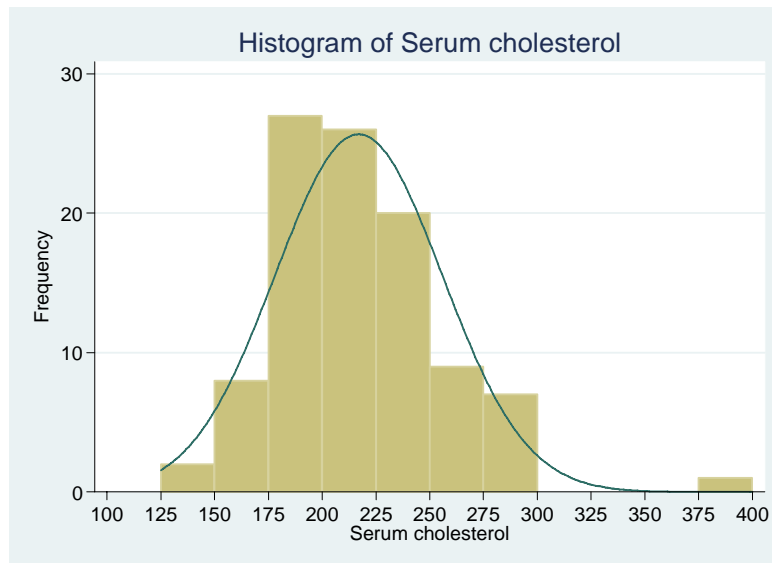
Table 2.2: Group frequency distribution for cholesterol level

Group/Class	Frequency	Cumulative Frequency	Relative Frequency
<150	2	2	0.02 (2%)
≥ 150 & < 175	8	10	0.08 (8%)
≥ 175 & < 200	27	37	0.27 (27%)
≥ 200 & < 225	26	63	0.26 (26%)
≥ 225 & < 250	20	83	0.20 (20%)
≥ 250 & < 275	9	92	0.09 (9%)
≥ 275 & < 300	7	99	0.07 (7%)
≥ 300	1	100	0.01 (1%)
Total	100		1.0 (100%)

The cumulative frequency in column three is the accumulation or sum of frequencies. For example, the cumulative frequency for the 3rd group (≥ 175 & < 200) in the above table is 37, which is the sum of the frequencies of 2, 8 and 27 ($2 + 8 + 27 = 37$). This

number is also equal to the cumulative frequency of the row above this group (which is 10) plus the current frequency of 27, giving 37.

Figure 2.3: Histogram for the frequency distribution of cholesterol level



An arrangement of data as in Table 2.2 is called a grouped frequency distribution table. It brings out the overall pattern of the data more clearly. For example, we may wish to know how many patients there are with a cholesterol level less than 225 mg/100ml, or how many of them have a cholesterol level greater than or equal to 250 mg/100ml. We can bring out the pattern in a grouped frequency distribution even more clearly by presenting the data in Table 2.2 in a Histogram shown in Figure 2.3. Histograms are a block diagram whose blocks are proportional in area to the frequency in each class or group. Histograms also show the distribution or the shape of the data which will be discussed in Section 2.3.

Another useful way to present the numerical data is a box-plot which is also very popular in medical data analyses. It shows a number of summary statistics for the data (these statistics will be discussed in Section 2.4). A box plot for the cholesterol level data in Table 2.5 will be given in Section 2.5 because understanding of box-plot requires the knowledge of measures of central tendency and measures of dispersion which will be discussed in Section 2.4.

2.3. Shape or Distribution of Data

A distribution of a data set may have almost any shape. A well-behaved distribution clearly has a highest frequency and tails off in nearly a smooth fashion on either side of the highest frequency. The distribution is skewed if one tail is dragged out more than the other tail. If the observations are evenly distributed on both sides from the block of the highest frequency, the distribution is symmetric or bell shaped, otherwise asymmetric/skewed. The following three shapes are common in practice:

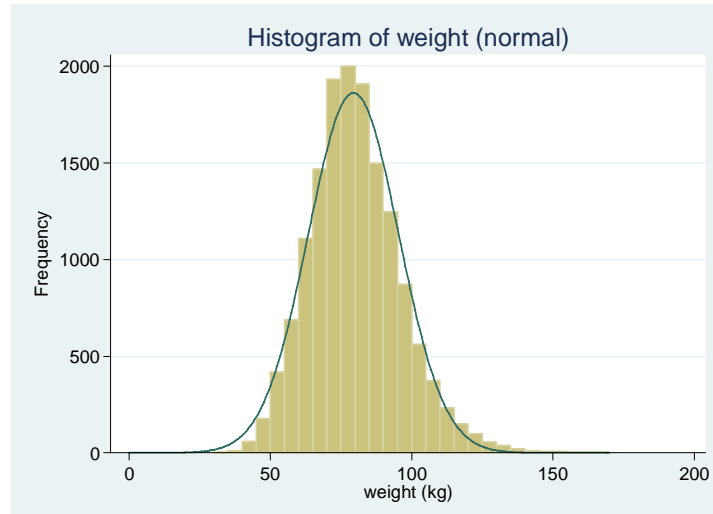
- Symmetric or Bell Shaped
- Positive Skewed or Right Skewed
- Negative Skewed or Left Skewed

We will discuss various shapes or distributions of data in the context of weight, age and post operative length of hospital stay for the cardiac surgery patients in six hospitals in Australia. The data was collected by the Australian Society for Cardiac and Thoracic Surgeons (ASCTS) between 2001 and 2006. Within this time period about 14000 patients underwent cardiac surgery in these six hospitals (name of the hospitals are secret, sorry!).

Symmetric Distribution

Let us present weight of 14000 cardiac surgery patients in a histogram in Figure 2.4. This figure shows that the highest frequency of the data falls in the interval 60-65kg and right hand tail is approximately the same as the left hand tail, so the data is approximately symmetric or the distribution of weight is approximately symmetric or bell shaped. The symmetry/asymmetry is measured from the tallest block in the histogram.

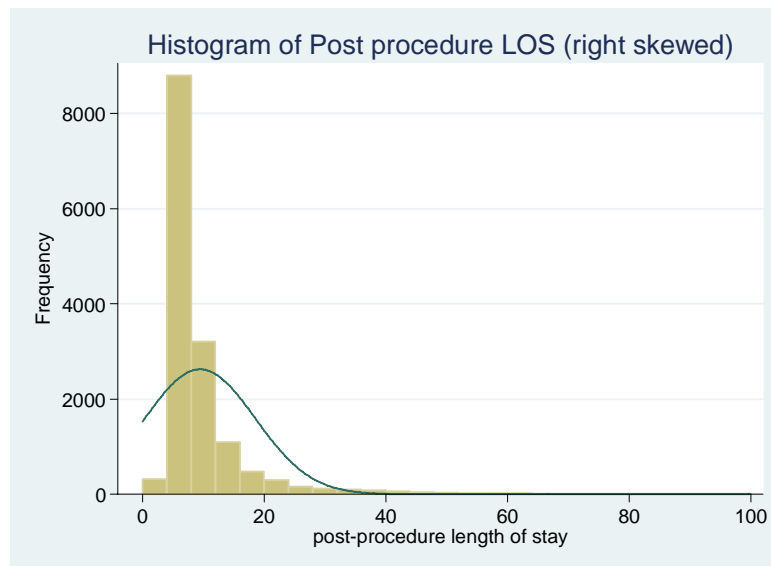
Figure 2.4: Weight Distribution for Cardiac Surgery Patients



Positively Skewed Distribution

Figure 2.5 represents the distribution for the postoperative length of stay of 14000 cardiac surgery patients. The right tail of the histogram is longer than the left tail, so the data is positively skewed or right skewed, more than 8000 patients stayed in hospital less than eight days, very few stayed more than 20 days.

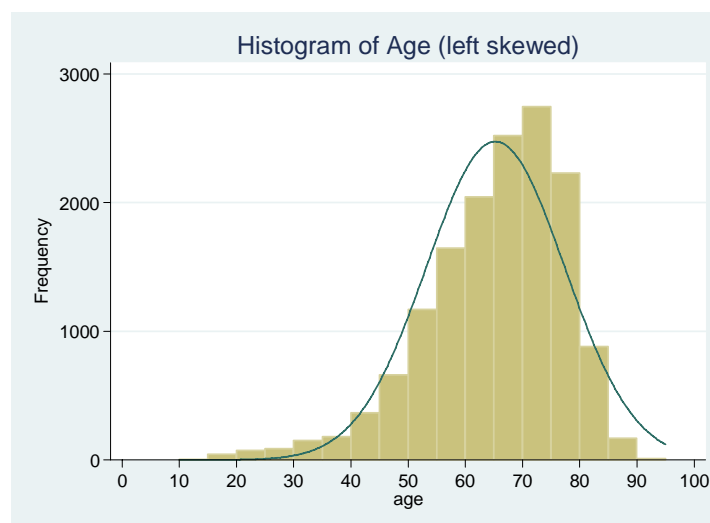
Figure 2.5: Postoperative Length of Hospital Stay for Cardiac Surgery Patients



Negatively Skewed Distribution:

Consider the age of the cardiac surgery patients presented in Figure 2.6. Clearly, the left tail of the histogram is longer than the right tail; so the distribution of age is negatively skewed or left skewed. The figure shows that most of the patients in the data are of age above 60 years. This means compared to younger people older people are more likely to undergo cardiac surgery like Isolated Coronary Artery Bypass Graft (CABG), Valve(s) replacement, CABG + Valve(s), etc.

Figure 2.6: Age Distribution for Cardiac Surgery Patients



2.4 Descriptive Statistics

Although frequency distributions and diagrams serve useful purposes, there are situations that require other types of data summarisation. What we need in many instances is the ability to summarise the data by means of a single number called a descriptive measure. Descriptive measures are usually computed from the data of a sample. Several types of descriptive measures can be computed from a set of data. In this section, however, we limit the discussion to measures of central tendency and measures of dispersion. The main objective of summarising data using descriptive statistics is the calculation of a single number that in some way conveys important information about the data from which it was calculated.

2.4.1 Central Tendency (Averages)

By central tendency we mean the tendency of data/observations to be around the central value of the observations. Measures of central tendency convey information regarding the central value of a set of values or observations. The three most commonly used measures of central tendency are the **mean (average), median and mode**.

Mean

The most familiar and commonly used measures of central tendency is the mean. The mean is obtained by adding all the values in a sample and dividing by the number of values that are added. Consider the serum cholesterol level (SCL) of five patients (ID#s 10-14 in Table 2.5): 240, 209, 210, 171 and 255 mg/100ml. The sample mean SCL is calculated as follows:

$$\text{Mean SCL} = \frac{240 + 209 + 210 + 171 + 255}{5} = \frac{1085}{5} = 217 \text{ mg/100ml}$$

Let us assume that n is the number of observations in a sample, x is the value of the observations and \bar{x} is the sample mean. Then the general formula for calculating sample mean is as follows:

$$\text{Mean: } \bar{x} = \frac{\text{Sum of All Values of Observation}}{\text{Number of Values}} = \frac{\sum x}{n} = \frac{1085}{5}$$

The symbol \sum instructs us to add all values of the variable from the first to the last. When the data is numerical and symmetric or nearly symmetric, the mean is appropriate measure of central tendency. For the sample cholesterol level data above, the value of n is 5, the total number of patients and x is the cholesterol levels: 240, 209, 210, 171 and 255 mg/100ml.

One of the major drawbacks of the mean is that it is highly affected by the extreme values (outliers). For example, let us add one more patient into the above sample whose cholesterol level was recorded 382 mg/100ml (ID#31, Table 2.5). Now the total number of patients in the sample is six. Clearly, the value 382 is very different

than rest of the values in the sample and may be considered as the extreme value. The sum of these cholesterol levels is 1467 and the mean is 244.5 mg/100ml. The mean cholesterol level clearly has been increased to a value that is far from most of the observations in the sample, just because of an additional extreme value (382 mg/100ml) included into the sample. Hence, one should avoid calculation of the mean if the data is asymmetric or contains outliers. In fact, the mean is appropriate when the distribution of the data is symmetric or approximately symmetric.

Median

The median of a set of values or observations is that value which divides the data set into two equal parts. If the number of values is odd, the median will be the middle value **after all values have been arranged in decreasing or increasing order of their magnitude**. When the number of values is even, there is no single middle value. Instead there are two middle values, in this case the median is taken to be the mean of these two middle values **after all values have been arranged in either ascending or descending order of their magnitude**.

Consider the cholesterol level for the five patients considered above, the data in ascending order (smallest to largest) gives: 171, 209, 210, 240 and 255 mg/100ml. Since we have an odd number of observations (the sample size is 5), the median is the middle most value which is 210 mg/100ml.

Let us add the cholesterol level of 382 mg/100ml (extreme value) to the above sample. The observations in ascending order are: 171, 209, 210, 240, 255, 382 mg/100ml. Since the number of observations is six (an even number), the median is the average of the two middle values 210 and 240 mg/100ml. Thus the median is:

$$\text{Median} = \frac{210 + 240}{2} = 225 \text{ mg/100ml.}$$

The median for each of the data set are very similar although the later data set has an extreme value. The median is not affected by the extreme values in a data set because it's calculation does not depends on the observations those are in the tails (lower or upper) of the data set. Median is very useful for asymmetric (right or left skewed) numerical data. For symmetric data the mean and the median are approximately the same.

Mode

The mode of a data set is that value which occurs most frequently. If all the values are different there is no mode; on the other hand, a set of values may have more than one mode. A distribution with one mode is known as uni-modal, with two modes is bimodal and similarly having more than two modes is called multi-modal. The mode is usually useful for categorical data. Mode can also be calculated for numerical data, however, mean and median are more useful for such data.

Let us consider, for example, the frequency distribution in Table 2.3 for different types of cardiac surgery in the ASCTS database as mentioned earlier [Isolated Coronary Artery Bypass Graft (CABG), Valve(s) only, CAGB + Valve(s), and other].

The mode of this data set is Isolated CABG because this surgery types has greater number of cases (1979); the data is uni-modal.

Table 2.3: Cardiac Surgery types in six Public Hospitals in Australia

Surgery type	Isolated CABG	Valve(s) only	CABG + Valve(s)	Others
Number of Cases	1979	367	309	396

Consider the systolic blood pressure (SBP) levels of five patients (ID#s 55-59 in Table 2.5): 134, 124, 124, 114 and 154 mmHg. The mode of this data set is 124, because this number occurs most frequently in the data.

Relationship of Mean, Median and Mode

- For a symmetric distribution, generally, mean \cong median \cong mode (see Figure 2.4)
- For a positively skewed or right tailed data, usually, mean \geq median \geq mode (see Figure 2.5)
- For a negatively skewed or left tailed data, usually, mean \leq median \leq mode (see Figure 2.6).

2.4.2 Percentiles (Quartiles)

Percentile is another type of descriptive measure which is also often used in practice. For a set of observations, the κ th percentile, say Q , is the value of the observation such that κ percent of the smallest observation are less than Q and $(100 - \kappa)$ percent of the largest observations are greater than Q . Thus, for example,

- 25th percentile is the value below which the smallest 25% of observations fall and 75% observations are larger than it.
- 75th percentile is the value above which the 25% of observations falls and 75% fall below.
- Note that the median is the 50th percentile, the value in the middle or the half way point.

The 25th, 50th and 75th percentiles are often referred to as the first, second and third quartiles respectively. The first, second and third quartiles are respectively denoted by Q_1 , Q_2 and Q_3 . The following steps are used for calculating percentiles.

- Arrange the observations from smallest to largest.
- Obtain the place in the list of observations where a percentile lies.
- Place in the List $= (n+1) \times \kappa$, where n is the number of observations (sample size) and κ is the percentile that we want to calculate. For example, for the 25th percentile (first quartile), $\kappa = 0.25$ or 25%.

For calculating places for the percentiles (e.g., 25th, 50th and 75th) for a data set, the following formulas are used.

- Place of the 25th percentile or the 1st Quartile $Q_1 = (n + 1) \times 0.25$
- Place of the 50th percentile or the 2nd Quartile $Q_2 = (n + 1) \times 0.50$
- Place of the 75th percentile or the 3rd Quartile $Q_3 = (n + 1) \times 0.75$

Consider the serum cholesterol level (SCL) data for the first 20 patients (ID#s: 00-19) in Table 2.5 and arrange them in ascending order of magnitude in the following table:

Table 2.4: Cholesterol levels for the first 20 patients from Table 2.5 (arranged in ascending order).

Serial #	SCL	Serial #	SCL
1	147	11	231
2	166	12	232
3	171	13	238
4	189	14	239
5	190	15	240
6	199	16	255
7	199	17	267
8	209	18	268
9	210	19	272
10	223	20	279

For the data in Table 2.4, the places for 25th, 50th and 75th percentiles are respectively as follows:

- $(n + 1) \times 0.25 = 21 \times 0.25 = 5.25 \cong 5$
- $(n + 1) \times 0.5 = 21 \times 0.5 = 11.5$ and
- $(n + 1) \times 0.75 = 21 \times 0.75 = 15.75 \cong 16$.

Thus the 25th percentile or the first quartile (Q_1) is the 5th observation in the above table which is 190 mg/100ml; the 75th percentile or the 3rd quartile (Q_3) is the 16th observation which is 255 mg/100ml. The 50th percentile (median) or the 2nd quartile (Q_2) is the average of 11th and 12th values in the data set in Table 2.4 because 11.5 is

in the middle of the serial #s 11 and 12. Thus the 50th percentile is: $(231+232)/2 = 231.5$ mg/100ml, the median.

2.4.3 Measures of Dispersion or Spread

A measure of dispersion conveys information regarding the amount of variability present in a data set. If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data. The amount of dispersion may be small, when the values, though different, are close together. If the values are widely scattered, the dispersion is greater. Two data sets may have the same mean but different dispersion. The three most commonly used measures of dispersion are the range, inter-quartile range and standard deviation.

Range

One way to measure the variation in a set of values is to compute the range, which is the simplest measure of dispersion. The range of a set of observations is the difference between the largest and the smallest value in the set i.e.,

$$\text{Range} = \text{Largest Observation} - \text{Smallest Observations}$$

Consider, for example, the cholesterol level of 20 patients presented in Table 2.4. The smallest and largest values in this table are 147 mg/100ml and 279 mg/100ml respectively, the range can be computed as

$$\text{Range} = 279 - 147 = 132 \text{ mg/100ml}$$

The usefulness of the range is limited. The fact that it takes into account only two values causes it to be a poor measure of dispersion. Range is highly affected by extreme values because it depends only on the smallest and the largest values in the data set. For example, if we add cholesterol level of 382 mg/100ml (ID# 32, Table 2.5) into the data in table 2.4, then the range will be: $382 - 147 = 235$ mg/100ml which is much higher than 132 mg/100ml. The main advantage in using the range is the simplicity of its computation.

Inter-quartile Range

Similar to range, is the inter-quartile range (IQR), which reflects the variability among the middle 50 percent of the observations in a data set. Thus the inter-quartile range is the difference between the third (75th percentile) and the first (25th percentile) quartiles, that is

$$\text{Inter-quartile Range} = \text{Third Quartile } (Q_3) - \text{First Quartile } (Q_1)$$

Or equivalently,

$$\text{Inter-quartile Range} = 75^{\text{th}} \text{ percentile} - 25^{\text{th}} \text{ percentile}$$

As for example, for the cholesterol level data in Table 2.4, the 25th and 75th percentiles are 190 mg/100ml and 255 mg/100ml respectively and thus the inter-quartile range is

$$\text{Inter-quartile Range} = 255 - 190 = 65 \text{ mg/100ml.}$$

The limitation of the IQR is that it calculates the spread among the middle 50% of the observations and thus is not affected by the extreme values. Like median, the IQR is appropriate for asymmetric numerical data.

Standard Deviation

Standard deviation is the most widely used measure of dispersion for symmetric (or approximately symmetric) numerical data (discrete or continuous). When the values of a set of observations lie close to their mean, the dispersion is less than when they are scattered over a wide range. So it would be naturally interesting if we could measure dispersion relative to the scatter of the values around their mean. Such a measure is known as the variance. In computing the variance of a sample of values, we generally follow the following steps:

- Calculate the mean of the data set
- Subtract the mean from each of the values
- Square the resulting differences
- Add up the squared differences
- The sum of the squared deviations divided by the sample size minus one $[(n - 1)]$ is the sample variance.

We denote the sample variance by s^2 and in notational form it is written as follows.

$$s^2 = \frac{\text{Sum of squared deviations of observations from mean}}{\text{sample size less one}}$$

$$= \frac{\sum (x - \bar{x})^2}{n - 1}$$

Here x is the observation and \bar{x} is the sample mean defined in Section 2.4.1.

To ease the computation let us consider the cholesterol level for the first three patients in Table 2.5 (ID3s 00-02), they are 199, 267 and 272 mg/100ml. The mean of these observations is: $(199+267+272)/3 = 246$ mg/100ml. Thus the sample variance is calculated as follows:

$$s^2 = \frac{(199 - 246)^2 + (267 - 246)^2 + (272 - 246)^2}{3 - 1}$$

$$= \frac{3326}{2} = 1663 \text{ mg}^2 / 100\text{ml}^2$$

Variance has squared units. Therefore, it is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units. To get a measure

of dispersion in original units, we take the square root of the variance. This is called the standard deviation. The sample standard deviation is given by

$$s = \sqrt{s^2}$$

Thus, the standard deviation of the cholesterol levels of the first three patients is obtained by taking the square root of its variance 1663 i.e., standard deviation is $\sqrt{1663} = 40.78$ mg/100ml. This means that in general the cholesterol levels in the data set are 40.78 mg/100ml above or below the mean level of 246 mg/100ml. The standard deviation is the most widely used measure of dispersion in medical research, but can be misleading when there are extreme values (outliers in the data), and or when the distribution is skewed. Therefore, we should check the shape of the data before calculation of standard deviation for it.

2.5 Box Plot

Box-plot is a very important and useful graphical representation for numerical data. It is particularly useful when comparing a numerical variable by the categories of a categorical variable. The serum cholesterol level data in Table 2.5 has been presented in a box-plot in Figure 2.7. The vertical lines in the left and right ends of the graph are the lower whisker (minimum value in the data, excluding outliers) and upper whisker (maximum value in the data, excluding outliers) respectively; the dark thick vertical line within the box is the median value (or 2nd quartile) which divides the data into two equal parts (50%/50%); lower and upper ends of the box are respectively 25th and 75th percentiles. If the median line is in the middle of the box and the difference between the 25th percentile and lower whisker is the same of the difference between upper whisker and the 75th percentile, then the data is symmetric, otherwise, the data is asymmetric. Observations outside of the lower and upper whiskers are known as extreme values or outliers, often such observations are excluded from subsequent data analyses.

Figure 2.7: Box plot for Serum Cholesterol level data



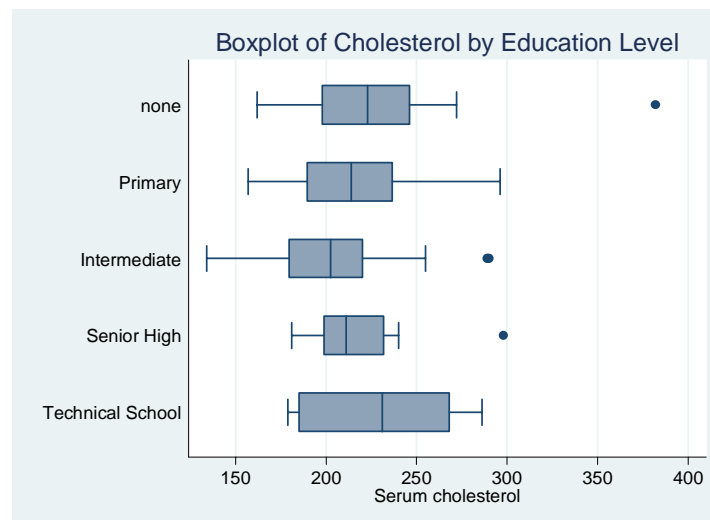
A box-plot provides the following five number summary statistics of a data set. The values shown in the brackets are calculated for the cholesterol level data in Table 2.5 and also evident from Figure 2.7.

- Minimum value (134 mg/100ml)
- Maximum value (298 mg/100ml, we excluded the outlier 382 mg/100ml)
- 25th percentile or the first quartile (188 mg/100ml)
- 50th percentile or the second quartile or median (216 mg/100ml)
- 75th percentile or the third quartile (238.5 mg/100ml)

Figure 2.7 shows that the cholesterol level of Honolulu Heart Study patients is approximately symmetric thus, mean and median values of cholesterol level are approximately the same. This figure also shows that there is an outlier in the data set which is 382 mg/100ml.

The cholesterol level data in Table 2.5 has been compared by education level of the patients in the following box-plot. The patients with intermediate education level have the lowest median cholesterol level and patients with the technical school level education have the highest median cholesterol level. If we ignore the outliers in the data, the spread among the cholesterol levels of the patients with senior high education level is the smallest. The distributions of cholesterol levels for all the categories except senior high are approximately symmetric.

Figure 2.8: Box plot for Serum Cholesterol level by Patient's Education level



Calculation of the Upper and Lower Whiskers:

Upper and lower whiskers of a box-plot can be calculated manually using the following formulas – calculations require the value of median and inter-quartile range for the data set.

- Lower whisker = Median – 1.5 * IQR
- Upper whisker = Median + 1.5 * IQR

A box plot also shows:

- General distribution of a data set (eg, minimum, maximum, shape: symmetric, asymmetric etc.)
- Central tendency (median, percentiles)
- Dispersion: Range of observations (difference between maximum and minimum values) and inter-quartile range (difference between Q_3 and Q_1)

Summary of This Module:

Terms and summary:

- **Statistical Tables:** Frequency table, relative frequency table
- **Charts:**
 - bar chart and pie chart – appropriate for categorical data
 - bar chart is very useful when comparing
 - categories of a categorical variable (e.g., patient's education level) or
 - categories of a categorical variable for each category of another variable (e.g., smoking status by education level).
 - histogram, box-plot and scatter plot – appropriate for continuous data
 - box-plot is very useful when comparing a continuous variable among the categories of a categorical variable.
- **Averages:**
 - mean, median –appropriate for continuous data
 - mode – appropriate for categorical data but can also be calculated for continuous data.
- **Spreads:**
 - Range, IQR – appropriate for categorical data
 - SD – appropriate for continuous data
- **For Numerical Data:**
 - If the distribution of the data is symmetric, report only the mean and standard deviation.
 - If the distribution of the data is skewed, report only the median and inter-quartile range (IQR)
 - If you are comparing summary statistics for two data sets, where one set is symmetric and another set is asymmetric, report median and inter-quartile range.
- **For Categorical Data:**
 - Report proportions

Table 2.5: Honolulu Heart Study Data (original data has 7683 patients)

ID	EDU	WGT	HGT	AGE	SMO	PSY	BG	SCL	SBP	BMI	BMIG
00	1	70	165	61	1	1	107	199	102	26	2
01	1	60	162	52	0	2	145	267	138	23	1
02	1	62	150	52	1	1	237	272	190	28	2
03	1	66	165	51	1	1	91	166	122	24	1
04	1	70	162	51	0	1	185	239	128	27	2
05	2	59	165	53	0	2	106	189	112	22	1
06	1	47	160	61	0	1	177	238	128	18	1
07	2	66	170	48	1	1	120	223	116	23	1
08	3	56	155	54	0	2	116	279	134	23	1
09	1	62	167	48	0	1	105	190	104	22	1
10	2	68	165	49	1	2	109	240	116	25	1
11	1	65	166	48	0	1	186	209	152	24	1
12	1	56	157	55	0	2	257	210	134	23	1
13	1	80	161	49	0	1	218	171	132	31	2
14	2	66	160	50	0	2	164	255	130	26	2
15	2	91	170	52	0	2	158	232	118	31	2
16	2	71	170	48	1	1	117	147	136	25	1
17	3	66	152	59	0	2	130	268	108	29	2
18	1	73	159	59	0	2	132	231	108	29	2
19	2	59	161	52	0	1	138	199	128	23	1
20	1	64	162	52	1	1	131	255	118	24	1
21	2	55	161	52	1	1	88	199	134	21	1
22	1	78	175	50	1	1	161	228	178	25	1
23	1	59	160	54	0	1	145	240	134	23	1
24	2	51	167	48	1	2	128	184	162	18	1
25	2	83	171	55	0	1	231	192	162	28	2
26	1	66	157	49	1	2	78	211	120	27	2
27	2	61	165	51	0	1	113	201	98	22	1
28	1	65	160	53	0	1	134	203	144	25	1
29	2	75	172	49	0	1	104	243	118	25	1
30	2	61	164	49	0	2	122	181	118	23	1
31	1	73	157	53	1	2	442	382	138	30	2
32	1	66	157	52	0	1	237	186	134	27	2
33	1	73	155	48	0	2	148	198	108	30	2
34	1	61	160	53	0	1	231	165	96	24	1
35	2	68	162	50	0	2	161	219	142	26	2
36	1	52	157	50	0	2	119	196	122	21	1
37	3	73	162	50	0	1	185	239	146	28	2
38	1	52	165	61	1	2	118	259	126	19	1
39	1	56	162	53	1	1	98	162	176	21	1
40	2	67	170	48	1	2	218	178	104	23	1
41	1	61	160	47	0	1	147	246	112	24	1
42	2	52	166	62	1	2	176	176	140	19	1
43	1	61	172	56	1	2	106	157	102	21	1
44	2	62	164	55	1	2	109	179	142	23	1
45	1	56	155	57	1	2	138	231	146	23	1
46	1	55	157	50	0	2	84	183	92	22	1
47	2	66	165	48	1	2	137	213	112	24	1
48	1	59	159	51	0	2	139	230	152	23	1
49	2	53	152	53	1	2	97	134	116	23	1

Table 2.5 (continued)

ID	EDU	WGT	HGT	AGE	SMO	PSY	BG	SCL	SBP	BMI	BMIG
50	3	71	173	52	0	2	169	181	118	24	1
51	1	57	152	49	0	1	160	234	128	25	1
52	1	73	165	50	1	1	123	161	116	27	2
53	2	75	170	49	0	2	130	289	134	26	2
54	2	80	171	50	1	2	198	186	108	27	2
55	2	49	157	53	0	1	215	298	134	20	1
56	2	65	162	52	0	1	177	211	124	25	1
57	1	82	170	56	0	2	100	189	124	28	2
58	2	55	155	52	0	2	91	164	114	23	1
59	2	61	165	58	0	1	141	219	154	22	1
60	1	50	155	54	1	2	139	287	114	21	1
61	3	58	160	56	0	1	176	179	114	23	1
62	1	55	166	50	1	2	218	216	98	20	1
63	3	59	161	47	0	2	146	224	128	23	1
64	1	68	165	53	1	1	128	212	130	25	1
65	1	60	170	53	1	2	127	230	122	21	1
66	1	77	160	47	1	1	76	231	112	30	2
67	3	60	155	52	0	1	126	185	106	25	1
68	2	70	164	54	0	1	184	180	128	26	2
69	1	70	165	46	0	1	58	205	128	26	2
70	2	77	160	58	1	1	95	219	116	30	2
71	3	86	160	53	0	2	144	286	154	34	2
72	1	67	152	49	1	2	124	261	126	29	2
73	2	77	165	53	1	1	167	221	140	28	2
74	2	75	169	57	0	2	150	194	122	26	2
75	1	70	165	52	0	2	156	248	154	26	2
76	1	70	165	49	1	1	193	216	140	26	2
77	1	71	157	53	0	1	194	195	120	29	2
78	1	55	162	49	0	2	73	217	140	21	1
79	1	59	165	53	1	2	98	186	114	22	1
80	2	64	159	50	0	2	127	218	122	25	1
81	1	66	160	54	0	1	153	173	94	26	2
82	2	59	165	60	0	2	161	221	122	22	1
83	2	68	165	57	0	1	194	206	172	25	1
84	3	58	160	52	0	1	87	215	100	23	1
85	1	57	154	65	1	1	188	176	150	24	1
86	1	60	160	65	0	2	149	240	154	23	1
87	1	53	162	62	0	1	215	234	170	20	1
88	1	61	159	62	1	2	163	190	140	24	1
89	1	66	154	62	0	1	111	204	144	28	2
90	1	61	152	67	0	2	198	256	156	26	2
91	1	52	152	66	0	2	265	296	132	23	1
92	1	59	155	62	0	2	143	223	140	25	1
93	1	63	155	62	1	1	136	225	150	26	2
94	1	61	165	63	0	2	298	217	130	22	1
95	1	68	155	67	0	2	173	251	118	28	2
96	1	58	170	62	0	1	148	187	162	20	1
97	2	68	160	55	0	1	110	290	128	27	2
98	3	60	159	50	0	2	188	238	130	24	1
99	1	61	160	54	1	1	208	218	208	24	1

Table 2.6: Variable description for the Honolulu Heart Study data

Short variable name	Long name	Statistical Code
ID	ID number	N/A
EDU	Education Level	1= none, 2= primary, 3= intermediate, 4=senior high, 5= technical school, 6= University
WGT	Weight of patients (Kg)	N/A
HGT	Height of patients (Cm)	N/A
Age	Age (in years)	N/A
SMO	Smoking status	0= no 1=yes
PSY	Physical activity at home	1= mostly sitting 2= moderate 3=heavy
BG	Blood glucose (milligrams percent)	N/A
SCL	Serum cholesterol (milligrams percent)	N/A
SBP	Systolic blood pressure (millimetres of mercury)	N/A
BMI	Body mass Index (calculated to the nearest integer)	N/A
BMIG	Body mass index Group	1= "<= 25" (Group 1) 2= "> 25" (Group 2)