2. Descriptive Statistics

2.1 Data Summarization: Graphical Methods

Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data. Different graphical techniques are available for different types of data.

Summarizing Qualitative Data

The only allowable calculation on qualitative data is to count the frequency of each category of a variable. When the raw data can be naturally categorized in a meaningful manner, we can display frequencies by

- Bar charts – emphasize frequency of occurrences of the different categories.
- Pie chart – emphasize the proportion of occurrences of each category.

Bar Charts

A bar chart can be used to depict any level of measurement: nominal, ordinal, interval, or ratio.

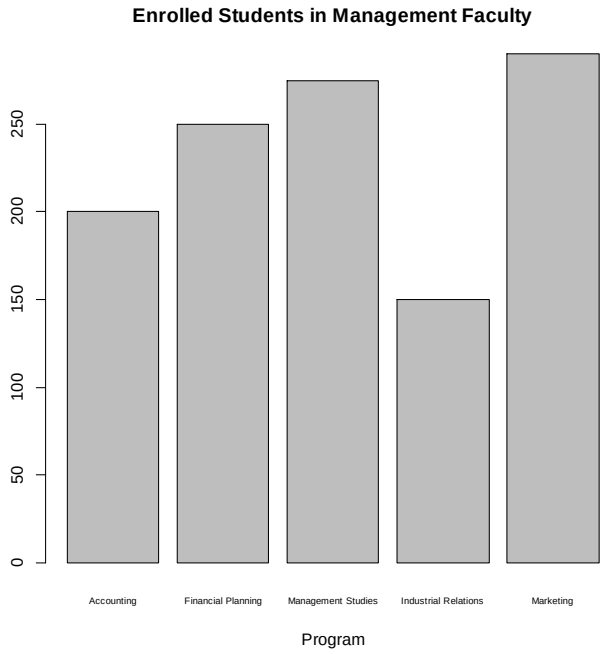
Example: The number of students enrolled in a management faculty for a particular year is given below according to the subject streams they have selected.

Program	Number of Students
Accounting	200
Financial Planning	250
Management Studies	275
Industrial Relations	150
Marketing	290

The qualitative variable contains five categories: Accounting, Industrial Relations, Financial Planning, Marketing, and Management Studies. The frequency (number of students) for each category is given. As the variable is qualitative, we select a bar chart to depict the data.

```
program<-c("Accounting","Financial Planning","Management Studies","Industrial
Relations","Marketing")
students<-c(200,250,275,150,290)
bar<-cbind(program,students)

barplot(students, main="Enrolled Students in Management
Faculty",xlab="Program",names.arg =program,cex.names=0.6)
```

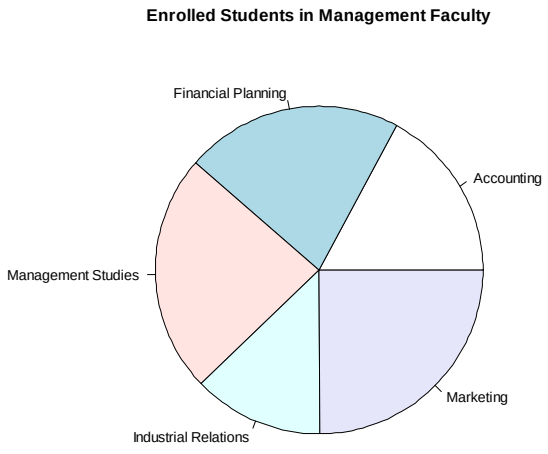


With this chart, it is easy to see that the highest enrollment is in Marketing and the lowest is in the Industrial Relations program.

Pie Chart

A pie chart, like a bar chart, is also used to summarize qualitative data. The pie chart is a circle, subdivided into a number of slices that represent the various categories. The size of each slice is proportional to the percentage corresponding to the category it represents.

```
pie(students, labels = program, main="Enrolled Students in Management Faculty")
```



Because the areas of the sectors, or "slices," represent the relative frequencies of the categories, we can quickly compare them. With this chart also, it is easy to see that the highest enrollment is in Marketing and the lowest is in the Industrial Relations program.

Summarizing Quantitative Data

There are several different ways of displaying quantitative data as they are given in numbers. This section introduced the basic methods of descriptive statistics used for organizing a set of numerical data in tabular form and presenting it graphically.

Some of the main descriptive methods for quantitative data include,

- Stem and leaf display
- Frequency Tables
- Histogram
- Box plot
- Scatter Diagrams

Stem and leaf display

A basic stem-and-leaf display contains two columns separated by a vertical line. The left column contains the stems and the right column contains the leaves. The first step in constructing a stem and leaf display is to decide how to split each observation (weight) into two parts: a stem and a leaf.

Example:

The weights in pounds of a group of workers are as follows:

173	183	162	168	154
165	177	179	158	180
171	160	145	186	164
175	151	171	182	166
188	169	175	162	157

For this example, define the first two digits of an observation to be its stem and the third digit to be its leaf. Thus, the weights are split into a stem and a leaf and can be displayed as follows:

```
weights<-c(173,183,162,168,154,165,177,179,158,180,171,160,145,186,164,175,
151,171,182, 166,188,169,175,162,157)

stem(weights)
```

The decimal point is 1 digit(s) to the right of the |

```
14 | 5
15 | 1478
16 | 02245689
17 | 1135579
18 | 02368
```

Frequency Tables

Frequency distributions are used to organize and present frequency counts in a summary form so that the information can be interpreted more easily. A frequency distribution of data can be shown in a table or graph. Some common methods of showing frequency distributions include frequency tables and histograms.

The hardest, and most important, step in constructing a grouped frequency distribution is choosing the number and width of the classes. Constructing a stem and leaf display first is often helpful. For the above example, the stem and leaf display suggests using five classes, each with a width of 10 pounds. The number (or frequency) of weights falling into each class is then recorded as shown in the table that follows. Care must be taken to define the classes in such a way that each measurement belongs to exactly one class.

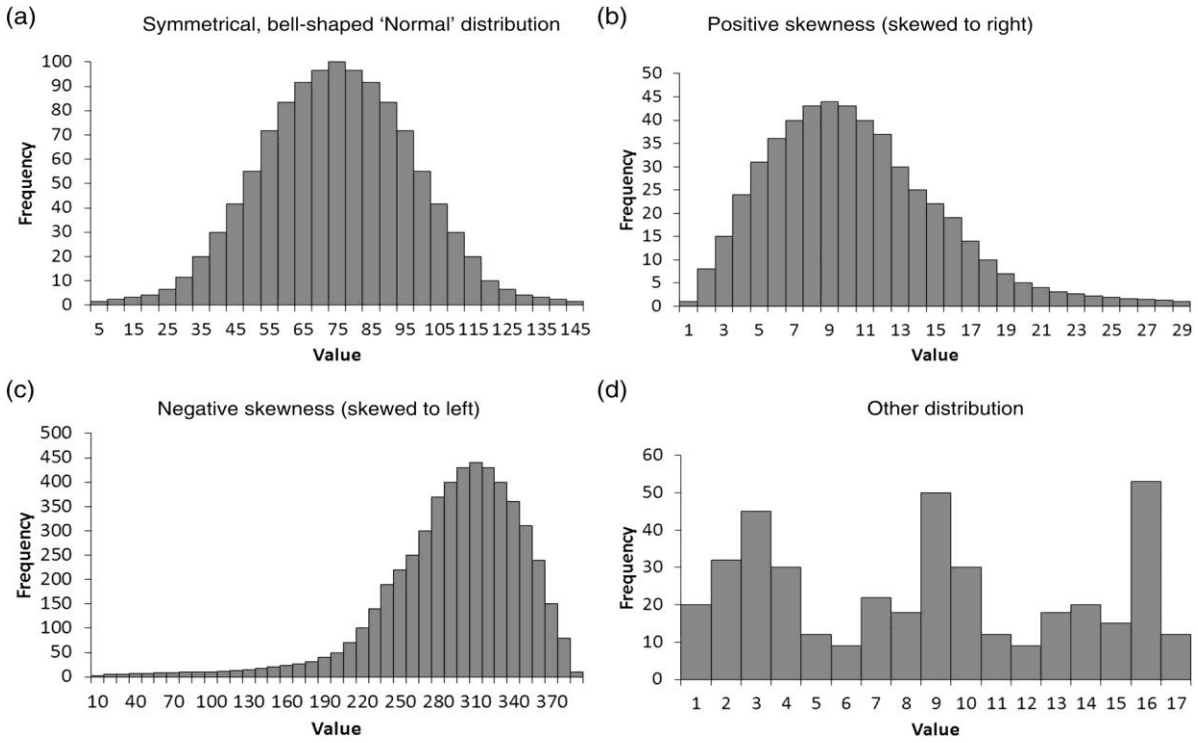
Class Limits	Frequency
140 up to 150	1
150 up to 160	4
160 up to 170	8
170 up to 180	7
180 up to 190	5
Total	25

Histogram:

Histograms are another useful graphical tool for quantitative data. Usually a histogram would be drawn with following features.

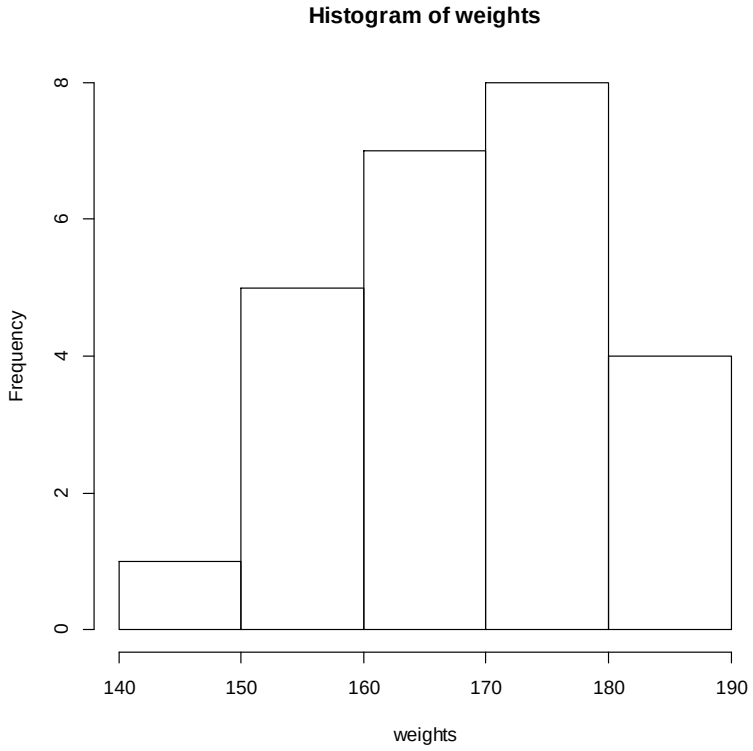
- The height of the column shows the frequency for a specific range of values.
- Columns are usually of equal width; however, a histogram may show data using unequal ranges (intervals) and therefore have columns of unequal width.
- The values represented by each column must be mutually exclusive and exhaustive. Therefore, there are no spaces between columns and each observation can only ever belong in one column.

Histogram can be used to comment on the shape of the distribution as well. Following images represent different types of histograms with different distribution shapes.



The following figure represents a histogram for the above weight data. Notice that if a statistical package is used to draw a histogram, it automatically decides the width of the classes, but some of them have options to change the widths as needed by the user.

```
hist(weights)
```



Box plot

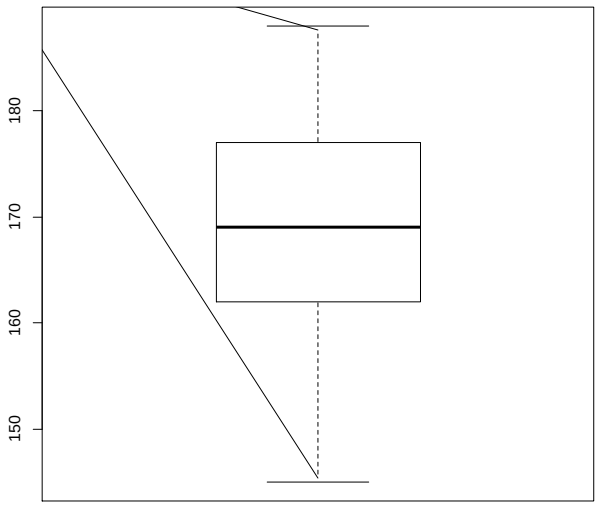
Boxplot is a convenient way of graphically depicting groups of numerical data through their five-number summaries: the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum). A boxplot may also indicate which observations, if any, might be considered outliers.

Steps of drawing a boxplot:

1. Calculate the median and the quartiles (the lower quartile is the 25th percentile and the upper quartile is the 75th percentile).
2. Plot a symbol at the median (or draw a line) and draw a box (hence the name--box plot) between the lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.
3. Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.
4. Calculate the following points:
5. L = lower quartile - 1.5*IQ
   U = upper quartile + 1.5*IQ
6. The line from the lower quartile is drawn from the lower quartile to the smallest point that is greater than L. Likewise, the line from the upper quartile is drawn to the largest point smaller than U.
7. Points less than L or greater than U are drawn as small circles.

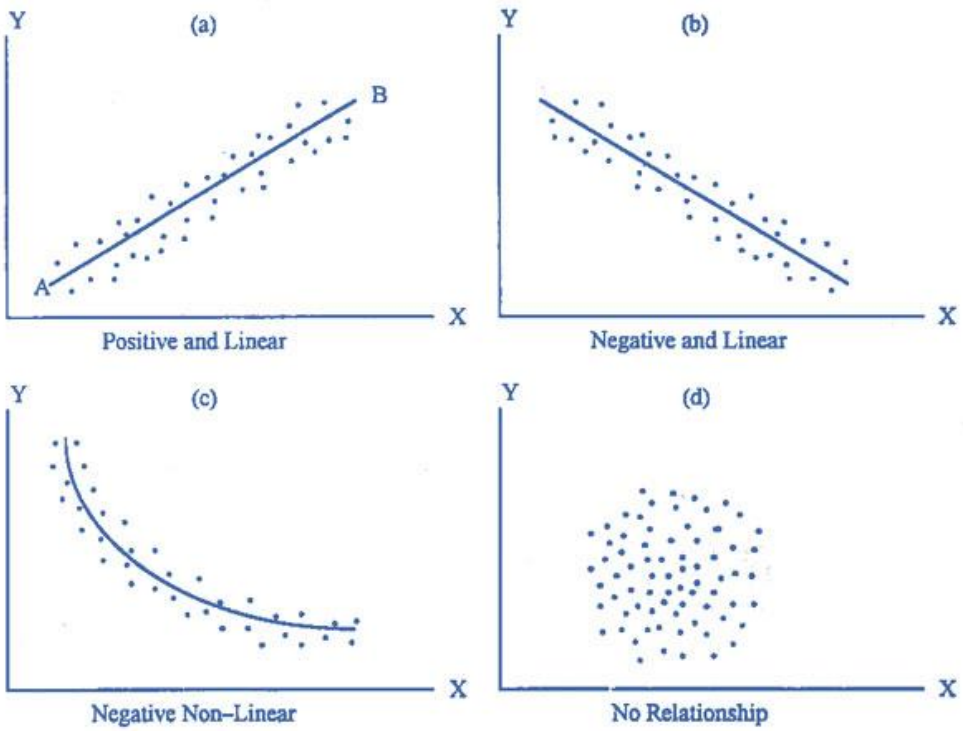
Thus the box plot identifies the middle 50% of the data, the median, and the extreme points. The following figure illustrates the boxplot for *weight* data.

```
boxplot(weights)
```



Scatter Diagrams

A scatter diagram is a tool for analyzing relationships between two quantitative variables. The plot displays as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis drawing a point for each pair. Scatter diagrams will generally show the possible correlations between the variables as displayed below.



plot(variable1, variable2)

So far in the descriptive methods, the graphical and tabular methods are introduced and then the focus would be on the numerical summaries of the data.

2.2 Data Summarization: Numerical Methods

2.2.1 Measures of Location

A fundamental task in many statistical analyses is to estimate a location parameter for the distribution, that is to find a typical or central value that best describes the data. There are five main location measures in statistics.

- Mean
- Median
- Mode
- Percentiles
- Quartiles

If the measures are computed for data from a sample, they are called sample statistics. If the measures are computed for data from a population, they are called population parameters. A sample statistic is referred to as the point estimator of the corresponding population parameter.

Mean:

Perhaps the most important measure of location is the mean, or average value for a variable. The sample mean $\bar{X}$ is the point estimator of the population mean μ. Sample mean for n observations $x_1, x_2, \ldots, x_n$ can be calculated as,

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: Apartment Rents

Seventy efficiency apartments were randomly sampled in a small college town. The monthly rent prices for these apartments are listed below in ascending order.

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{34{,}356}{70} = 490.8$$

Median:

The median of a data set is the value in the middle when the data items are arranged in ascending order. Whenever a data set has extreme values, the median is the preferred measure of central location. The median is the measure of location most often reported for annual income and property value data. A few extremely large incomes or property value scan inflate the mean.
For an odd number of observations, the median is the middle value when arranged in ascending or descending order and if even number of observations the median is the average of two middle values.

Example: Apartment Rents

Averaging the 35th and 36th data values:
Median = (475 + 475)/2 = 475

Mode :
The mode of a data set is the value that occurs with greatest frequency. The greatest frequency can occur at two or more different values. If the data have exactly two modes, the data are bimodal. If the data have more than two modes, the data are multimodal.

Example : Apartment Rents

450 occurred most frequently (7 times)
Mode = 450

Percentiles:

A percentile provides information about how the data are spread over the interval from the smallest value to the largest value. Admission test scores for colleges and universities are frequently reported in terms of percentiles. The p^{th} percentile of a data set is a value such that at least p percent of the items take on this value or less and at least (100 - p) percent of the items take on this value or more.

Compute index i, the position of the p^{th} percentile. $i = (p/100)*n$.

If i is not an integer, round up. The p^{th} percentile is the value in the i^{th} position. If i is an integer, the p^{th} percentile is the average of the values in positions i and i +1.

Example: Apartment Rents

90^{th} Percentile:

$i = (p/100)n = (90/100)70 = 63$
Averaging the 63rd and 64th data values:
90th Percentile = (580 + 590)/2 = 585
At least 90%of the item stake on a value of 585 or less.

Quartiles:

Quartiles are specific percentiles. First quartile is the 25^{th} percentile, second quartile is the 50^{th} percentile (or the median) and third quartile is the 75^{th} percentile.

Example : Apartment Rents

Third quartile = 75th percentile
$i = (p/100)n = (75/100)70 = 52.5 = 53$
Third quartile = 525

2.2.2 Measures of Variability

In a data set, it is often desirable to consider measures of variability (dispersion), as well as measures of location. For example, suppose that you are a purchasing agent for a large manufacturing firm that you regularly place orders with two different suppliers. In choosing between the two suppliers we might consider not only the average delivery time for each, but also the variability in delivery time for each.

There are several measures of variability we can calculate for a data set to see the dispersion.

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

Range:

The range of a data set is the difference between the largest and smallest data values. It is the simplest measure of variability and is very sensitive to the smallest and largest data values.

Example : Apartment Rents

Range = largest value - smallest value

Range = 615 - 425 = 190

Interquartile Range:

The interquartile range of a data set is the difference between the third quartile and the first quartile which represents the range for the middle 50% of the data. It overcomes the sensitivity to extreme data values.

Example : Apartment Rents

3rd Quartile ($Q3$) = 525

1st Quartile ($Q1$) = 445

Interquartile Range = $Q3$ - $Q1$ = 525 - 445 = 80

Variance:

The variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (x_i) and the mean ($\bar{X}$ for a sample, μ for a population).

The sample variance denoted by s^2, is the average of the squared differences between each data value and the mean and can be computed as follows:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Example : Apartment Rents

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(425-490.8)^2 + (430-490.8)^2 + \cdots + (615-490.8)^2}{70-1} = 2996.16$$

Standard Deviation:

The standard deviation of a data set is the positive square root of the variance. It is measured in the same units as the data, making it more easily interpreted than the variance.

The sample standard deviation is computed as,

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

Example : Apartment Rents

$s = \sqrt{s^2} = \sqrt{2996.16} = 54.74$

Coefficient of Variation:

The coefficient of variation indicates how large the standard deviation is in relation to the mean. The sample coefficient of variation is computed as,

$$\frac{s}{\bar{x}} * 100\%$$

Example: Apartment Rents

$$\frac{54.74}{490.8} * 100\% = 11.15\%$$

```
apts<-c(425,430,430,435,435,435,435,435,440,440,440,440,440,445,445,445,445,445,
450,450,450,450,450,450,450,460,460,460,465,465,465,470,470,472,475,475,
475,480,480,480,480,485,490,490,490,500,500,500,500,510,510,515,525,525,
525,535,549,550,570,570,575,575,580,590,600,600,600,600,615,615)

summary(apts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  425.0   446.2   475.0   490.8   522.5   615.0
```

Exercise: sheep weight

A random sample of 5 sheep in UK is taken and their weights were measured and recorded below in Kilograms (kg).

84.5, 72.6, 75.7, 94.8, 71.3

Suppose that, for each of the sheep in the above example, their heights at the shoulder are also measured. The heights (cm) are

86.5, 71.8, 77.2, 84.9, 75.4

Create a data frame consisting of 2 variables and 5 observations. Suppose that a third variable consisting of measurements of the length of the sheeps' backs becomes available. The values (in cm) are

130.4, 100.2, 109.4, 140.6, 101.4

Add the new variables to the same data frame and calculate the summary statistics of all 3 variables.