STATISTICS IS THE GRAMMAR OF SCIENCE

PROBABILITY AND STATISTICS

LECTURE – 6

MEASURES OF POSITION

PREPARED BY
HAZBER SAMSON
FAST NUCES ISLAMABAD

MEASURES OF POSITION

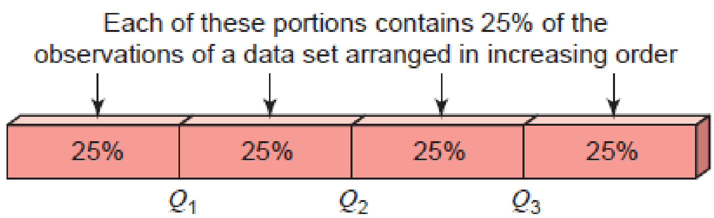
A **measure of position** determines the position of a single value in relation to other values in a sample or a population data set. There are many measures of position; however, in this section, we shall study only the following

- Quartiles
- Deciles
- Percentiles
- Percentile Rank
- Box and Whisker Plot

QUARTILES

Quartiles are three summary measures that divide a ranked data set into four equal parts. These three measures are the $first\ quartile\ denoted\ by\ Q_1$, the $second\ quartile\ denoted\ by\ Q_2$, and the $third\ quartile\ denoted\ by\ Q_3$. The second quartile is the same as the median of a data set. The first quartile is the value of the middle term among the observations that are less than the median, and the third quartile is the value of the middle term among the observations that are greater than the median. Note that the data should be ranked in increasing order before the quartiles are determined.

FIGURE 3.1 QUARTILES



Quartiles can be calculated using the following formulas

$$Q_1 = \frac{1}{4}(n+1)^{th}\ value$$

$$Q_2 = \frac{2}{4}(n+1)^{th}\ value \qquad Note\ that\ Q_2 = Median$$

$$Q_3 = \frac{3}{4}(n+1)^{th}\ value$$

INTER QUARTILE RANGE

The difference between the between the third quartile and first quartiles gives the interquartile range. It is denoted by IQR and is given by $IQR = Q_3 - Q_1$

EXAMPLE-1 Find quartiles and Interquartile range for the following data

21,25,35,37,39,40,41,44,46,48,49

SOLUTION $Q_1 = \frac{1}{4}(n+1)^{th} value = \frac{1}{4}(11+1)^{th} value = 3rd\ value = 35$

$$Q_2 = \frac{2}{4}(n+1)^{th} value = \frac{1}{2}(11+1)^{th} value = 6th\ value = 40$$

$$Q_3 = \frac{3}{4}(n+1)^{th} value = \frac{3}{4}(11+1)^{th} value = 9th\ value = 46$$

$$IQR = Q_3 - Q_1 = 46 - 35 = 11$$

EXAMPLE-2 Weekly TV-Viewing Times The A. C. Nielsen Company publishes information on the TV-viewing habits of Americans in Nielsen Report on Television. A sample of 20 people yielded the weekly viewing times, in hours, displayed in Table 3.1. Determine and interpret the quartiles and inter quartile range for these data.

TABLE 3.1

Weekly TV-viewing times

25	41	27	32	43
66	35	31	15	5
34	26	32	38	16
30	38	30	20	21

SOLUTION First, we arrange the data in Table 3.1 in increasing order:

5 15 16 20 21 25 26 27 30 **30** **31** 32 32 34 35 38 38 41 43 66

$$Q_1 = \frac{1}{4}(n+1)^{th} value = \frac{1}{4}(20+1)^{th} value = 5{,}25th\ value$$

$$= 5th\, value + 0.25\ (6th\, value - 5th\ value)$$

$$= 21 + 0.25\ (25 - 21) = 22$$

$$Q_2 = \frac{2}{4}(n+1)^{th} value = \frac{1}{2}(20+1)^{th} value = 10.5th\ value$$

$$= 10th\, value + 0.5\ (11th\, value - 10th\ value)$$

$$= 30 + 0.5\ (31 - 30) = 30.5$$

$$Q_3 = \frac{3}{4}(n+1)^{th} value = \frac{3}{4}(20+1)^{th} value = 15.75th\ value$$

$$= 15th\, value + 0.75\ (16th\, value - 15th\ value)$$

$$= 35 + 0.75\ (38 - 35) = 37.25$$

$$IQR = Q_3 - Q_1 = 37.25 - 22 = 15.25$$

In summary, the three quartiles for the TV-viewing times in Table 3.1 are

$Q_1 = 22\ hours,\ Q_2 = 30.5\ hours,\ Q_3 = 37.25\ hours\ \ and\ IQR = 15.25$

Interpretation We see that 25% of the TV-viewing times are less than 22 hours, 25% are between 22 hours and 30.5 hours, 25% are between 30.5 hours and 37.25 hours, and 25% are greater than 37.25 hours. IQR shows that the middle 50% of the TV-viewing times are spread out over a15.25-hour interval, roughly.

DECILES Deciles are the summary measures that divide a ranked data set into ten equal parts. Nine measures will divide any data set into ten equal parts. These Nine measures are the first decile denoted by (D_1), the second decile denoted by (D_2) ,…,ninth decile denoted by (D_9) . The data should be ranked in increasing order before the deciiles are determined. The quartiles are defined as follows.

Deciles can be calculated using the following formulas

$$D_1 = \frac{1}{10}(n+1)^{th} value$$

$$D_2 = \frac{2}{10}(n+1)^{th} value$$

…

$$D_9 = \frac{9}{10}(n+1)^{th} value$$

PERCENTILES Percentiles are the summary measures that divide a ranked data set into hundred equal parts. Ninety-nine measures will divide any data set into hundred equal parts. These Ninty-nine measures are the first percentile denoted by (P_1), the second percentile denoted by (P_2) ,…,ninty-ninth percentile denoted by (P_{99}) . The data should be ranked in increasing order before the percentiles are determined. The percentiles are defined as follows.

Percentiles can be calculated using the following formulas

$$P_1 = \frac{1}{100}(n+1)^{th} value$$

$$P_2 = \frac{2}{100}(n+1)^{th} value$$

…

$$P_{99} = \frac{99}{100}(n+1)^{th} value$$

EXAMPLE-3 Find D_4 *and* P_{37} for the following data

$$21,25,35,37,39,40,41,44,46,48,49$$

SOLUTION $D_4 = \frac{4}{10}(n+1)^{th} value = \frac{4}{10}(11+1)^{th}\ value = 4.8th\ value = 38.6$

$$P_{37} = \frac{37}{100}(n+1)^{th} value = \frac{37}{100}(11+1)^{th}\ value = 4.44th\ value = 37.88$$

PERCENTILE RANK

Percentile Rank of x_i gives the percentage of values in the data set that are less than x_i. It is given by the following formula

$$Percentile\ Rank\ of\ x_i = \frac{Number\ of\ values\ less\ than\ x_i}{Total\ number\ of\ values} \times 100$$

***EXAMPLE-3*:** In a university exam, students get marks as follows

$$21, 25, 35, 37, 39, 40, 41, 44, 46, 48, 49.$$

Find the percentile rank of the value 37.

SOLUTION: Percentile Rank of 37

$$Percentile\ Rank\ of\ 37 = \frac{Number\ of\ values\ less\ than\ 37}{Total\ number\ of\ values} \times 100$$

$$Percentile\ Rank\ of\ 37 = \frac{3}{11} \times 100 = 27.27\%$$

$$\Rightarrow 27.27\%\ values\ are\ less\ than\ 37.$$

BOX AND WHISKER PLOT

We convert five number Summary into a useful diagram, called Box and Whisker Plot or Box Plot.

DEFINITION A plot that shows the center, spread and skewness of data set is called Box and Whisker Plot. It is constructed by drawing the Quartiles (Q_1, Q_2, Q_3) and Smallest/Largest values in the data set between the lower and upper fences.

OUTLIER outliers are unusual values which differ greatly in magnitude from the majority of the data values. For decision of large and small we make lower and upper fences. A value lesser than lower fence or greater than upper fence will be considered as an outlier.

STEPS INVOLVED IN CONSTRUCTING BOX AND WHISKER PLOT

Step-1: First arrange the data in increasing order.

Step-2: Find Q_1, Q_2, Q_3 *and* IQR

Step-3: Find Lower and Upper Fences using the formulas

$$Lower\ Fence = Q_1 - 1.5 \times IQR$$
$$Upper\ Fence = Q_3 + 1.5 \times IQR$$

Step-4: Determine Smallest and Largest values in the given data with in fences

Step-5: Draw a horizontal line and mark the data values according to a suitable scale. Now above the horizontal line draw a box with is left side at the position of Q_1 and with the right side at the position of Q_3. Draw a vertical line at the position of median.

Step-6: Draw a line to the left of box at the point of smallest value and draw a line to the right of box at the point of largest value within the two fences. The two lines that join the box to these values are called whiskers.

Step-7: Identify the outliers if any

EXAMPLE-5: The following data are the incomes (in thousands rupees) for a sample of 12 households. $75, 69, 84, 112, 74, 104, 81, 90, 94, 144, 79, 98$.
Construct a box and whisker plot for these data.

SOLUTION: Step-1: Rearranging data in increasing order we get

$$69,\ 74,\ 75,\ 79,\ 81,\ 84,\ 90,\ 94,\ 98,\ 104,\ 112,\ 144$$

For these data

$$Q_1 = 77,\ \ Q_2 = 87,\ \ Q_3 = 101$$
$$IQR = 101 - 77 = 24$$

Step 2. Find the points that are 1.5 × IQR below Q_1 and 1.5 × IQR above Q_3. These two points are called the **lower** and the **upper inner fences**, respectively.

$$1.5 \times \text{IQR} = 1.5 \times 24 = 36$$
$$\text{Lower inner fence} = Q_1 - 36 = 77 - 36 = 41$$
$$\text{Upper inner fence} = Q_3 + 36 = 101 + 36 = 137$$

Step 3. Determine the smallest and the largest values in the given data set within the two inner fences. These two values for our example are as follows:

Smallest value within the two inner fences = 69
Largest value within the two inner fences = 112

Step 4. Draw a horizontal line and mark the income levels on it such that all the values in the given data set are covered. Above the horizontal line, draw a box with its left side at the position of the first quartile and the right side at the position of the third quartile. Inside the box, draw a vertical line at the position of the median. The result of this step is shown in Figure 3.13.

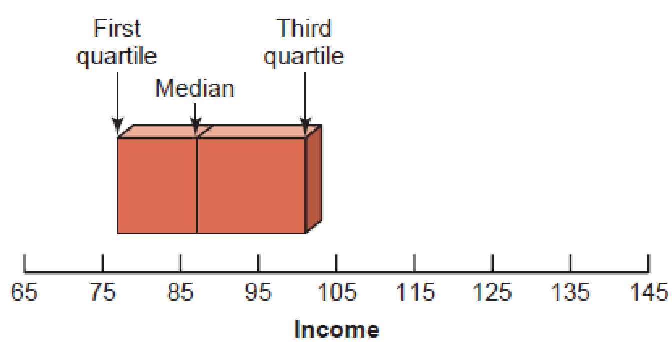


Figure 3.13

Step 5. By drawing two lines, join the points of the smallest and the largest values within the two inner fences to the box. These values are 69 and 112 in this example as listed in Step 3. The two lines that join the box to these two values are called **whiskers**. A value that falls outside the two inner fences is shown by marking an asterisk and is called an outlier. This completes the box-and-whisker plot, as shown in Figure 3.14.

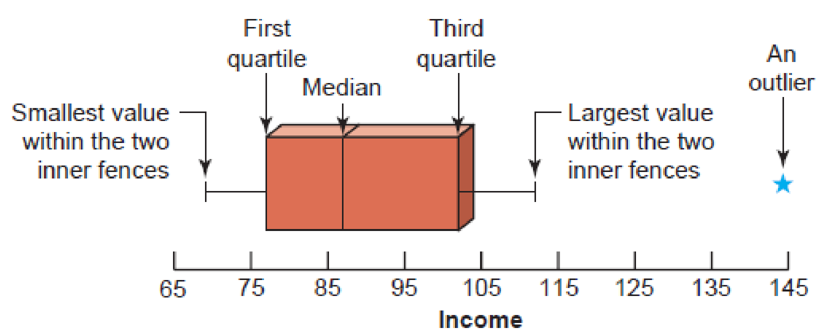


Figure 3.14

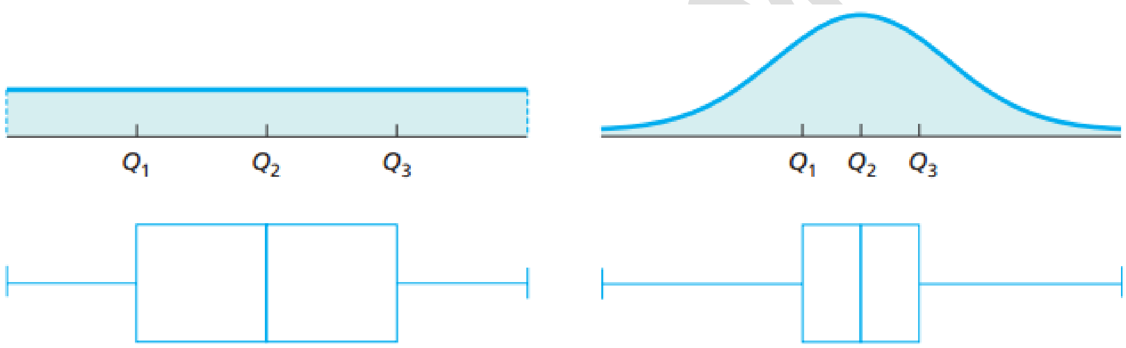
USING BOXPLOT TO IDENTIFY THE SHAPE OF THE DISTRIBUTION

You can also use a boxplot to identify the approximate shape of the distribution of a data set.

Information Obtained from a Boxplot

1. *a.* If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a.* If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.

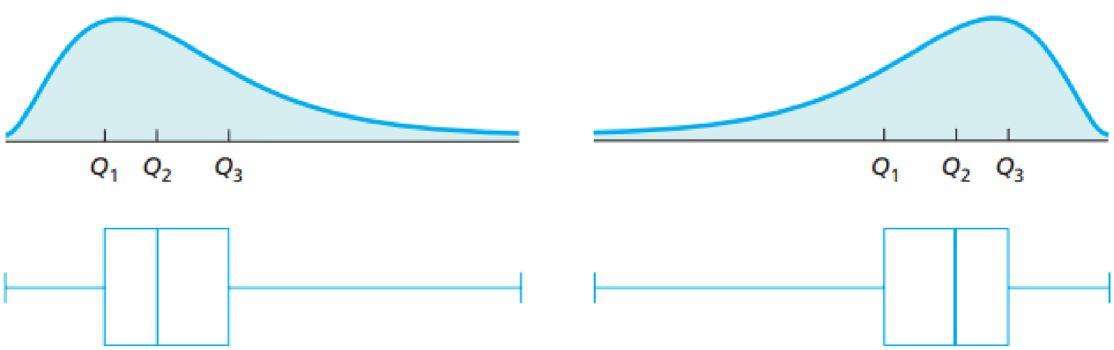
FIGURE 3.2 SYMMETRIC DISTRIBUTIONS



(a) Uniform (b) Bell shaped

In figure 3.2 both graphs showed symmetric distributions.

FIGURE 3.3 SKEWED DISTRIBUTIONS



(c) Right skewed (d) Left skewed

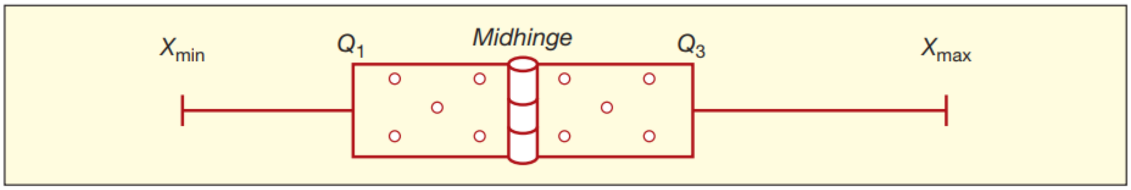
In figure 3.3 both graphs showed skewed distributions. Graph (c) shows positively skewed distribution while Graph (d) shows negatively skewed distribution

MIDHINGE

Quartiles can be used to define an additional measure of center that has the advantage of not being influenced by outliers. The **midhinge** is the average of the first and third quartiles:

$$\text{Midhinge} = \frac{Q_1 + Q_3}{2}$$

The name "midhinge" derives from the idea that, if the "box" were folded at its halfway point, it would resemble a hinge:



Since the midhinge is always exactly *halfway* between Q_1 and Q_3 while the median Q_2 can be *anywhere* within the "box," we have a new way to describe skewness:

$$\begin{aligned}
&If\ \ Midhinge = Median \ \Rightarrow distribution\ is\ symmetric \\
&If\ \ Midhinge < Median \ \Rightarrow distribution\ is\ negatively\ skewed \\
&If\ \ Midhinge > Median \ \Rightarrow distribution\ is\ positively\ skewed
\end{aligned}$$

EXAMPLE-6: The following data are the incomes (in thousands rupees) for a sample of 12 households. $75, 69, 84, 112, 74, 104, 81, 90, 94, 144, 79, 98$. Using Midhinge, Check whether the data is symmetric or skewed, if skewed identify type of skewness.

SOLUTION: Rearranging data in increasing order we get

$$69, 74, 75, 79, 81, 84, 90, 94, 98, 104, 112, 144$$

For these data

$$\begin{aligned}
&Q_1 = 77,\ \ Q_2 = 87,\ \ Q_3 = 101 \\
&IQR = 101 - 77 = 24 \\
&Here\ \ Midhinge = \frac{Q_1 + Q_3}{2} = \frac{77 + 101}{2} = \frac{178}{2} = 89 \\
&Also\ \ Median = Q_2 = 87 \\
&Midhinge > Median \\
&\Rightarrow Data\ is\ positively\ skewed.
\end{aligned}$$