Samples from: MINITAB BOOK

Quality and Six Sigma Tools using MINITAB Statistical Software: A complete Guide to Six Sigma DMAIC Tools using MINITAB®

Prof. Amar Sahay, Ph.D.

One of the major objectives of this text is to teach quality, data analysis and statistical tools used in the Six Sigma DMAIC (Define, Measure, Analyze, Improve, and Control) process. The chapters in this book provide concepts, understanding, and computer applications of Six Sigma DMAIC tools. The statistical tools used in the DMAIC process are discussed with step-wise MINITAB computer applications. The following are samples from the book randomly selected from different chapters:

CHAPTER 3 Using Statistics to Summarize Data Sets: Concepts and Computer Analysis (Descriptive Statistics_ Numerical Methods)

Chapter Highlights

This chapter deals with the basic tools of data analysis used in Six Sigma. The primary objective of this chapter is to enable you to master the techniques of describing data using numerical methods, and use these methods to compare and draw meaningful conclusions from data. The topics in this chapter will enable you to perform the following analysis using computer:

1. *Calculate and apply the measures of central tendency for both ungrouped and grouped data.*
2. *Calculate the measures of position — percentiles and quartiles, interpret their meaning, and their applications in data analysis.*
3. *Calculate and apply various measures of variation— range, interquartile range, variance, and standard deviation for both grouped and ungrouped data.*
4. *Understand the concept and importance of variation in Six Sigma.*
5. *Compare the mean, median, mode, and standard deviation to draw meaningful conclusions from the data.*
6. *Relate the mean and standard deviation using the Chebyshev's and Empirical rules and understand the importance of Empirical rule in statistics and data analysis.*
7. *Calculate and apply the measures of measures of central tendency, measures of variation, measures of shape (skewness and kurtosis), and measures of position to learn about the data*
8. *Describe the relationship between two variables — covariance and coefficient of correlation.*
9. *Learn the applications of the numerical methods in this chapter as they apply to Six Sigma and Lean Sigma.*

Calculating Descriptive Statistics Using Minitab

This example will demonstrate how to calculate the descriptive statistics of a data set described above. We will use the data file **TVCOMPLIFE.MTB** that shows the life (in hours) of 200 television components. Using MINITAB we will calculate various statistics.

(1) Calculate descriptive statistics of the life data.

To calculate descriptive statistics of the life data, follow the steps in Table 4.10

Table 4.10

Calculating Descriptive Statistics
Open the worksheet **TVCOMPLIFE.MTB**
From the main menu, select **Stat ➢Basic Statistics ➢Display Descriptive Statistics**
For **Variables**, Type **....**
Click **OK**

The following selected statistics will be calculated and displayed (Table 4.11).

Table 4.11

Descriptive Statistics: Life in Hours

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Life in Hours	200	0	324.89	3.19	45.07	203.00	296.25	329.00	359.75

Variable	Maximum
Life in Hours	430.00

(Note: N= number of observations, N*=missing values, SE Mean= standard error of the mean, StDev= sample standard deviation, Minimum= smallest value in the data, Q1 = first quartile, Median= middle value, Q3= third quartile, Maximum= largest value in the data).

(2) Calculate additional statistics and construct graphs using the graph option of the life data.

To calculate additional statistics and graphs of the life data, follow the steps in Table 4.12.

Table 4.12

Calculating Descriptive Statistics
Open the worksheet **TVCOMPLIFE.MTB**
From the main menu select, **Stat ➢Basic Statistics ➢Display Descriptive Statistics**
For **Variables**, type **C1** or select **C1 Life in Hours**
Click the **Statistics** tab and check the …
Click on the Graphs tab and check the boxes for all 4 graph options
Click **OK**

The selected statistics are shown in Table 4.13 and the graphs in Figure 4.5. Note that the graphs are displayed one at a time. You may use the "layout tool" in MINITAB to organize your graphs as shown in Figure 4.5. Table 4.14 has the instructions to organize and edit your graphs.

Table 4.13

Descriptive Statistics: Life in Hours

Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum
Life in Hours	200	0	324.89	3.19	45.07	2031.59	13.87	203.00

Variable	Q1	Median	Q3	Maximum	Range	IQR	Skewness
Life in Hours	296.25	329.00	359.75	430.00	227.00	63.50	-0.14

Variable	Kurtosis
Life in Hours	-0.34

Table 4.14: Organizing and Editing Your Graphs

Use the Layout Tool and Edit Bar to Organize and Edit your Graphs

1. From the main menu, select **Editor ➢ Layout Tool**
2. **Layout Tool** dialog box ….
3. Select the graph from this list and click the right arrow (>) button to move the graph ….
4. Once you select all the graphs you want to group, click **Finish**…
5. All the… will be displayed as a group.
6. Double click on any graph to display the **Edit Bar** dialog box. In this dialog box, click on the circle to the left of **Custom** and you can edit the color and pattern of that graph.

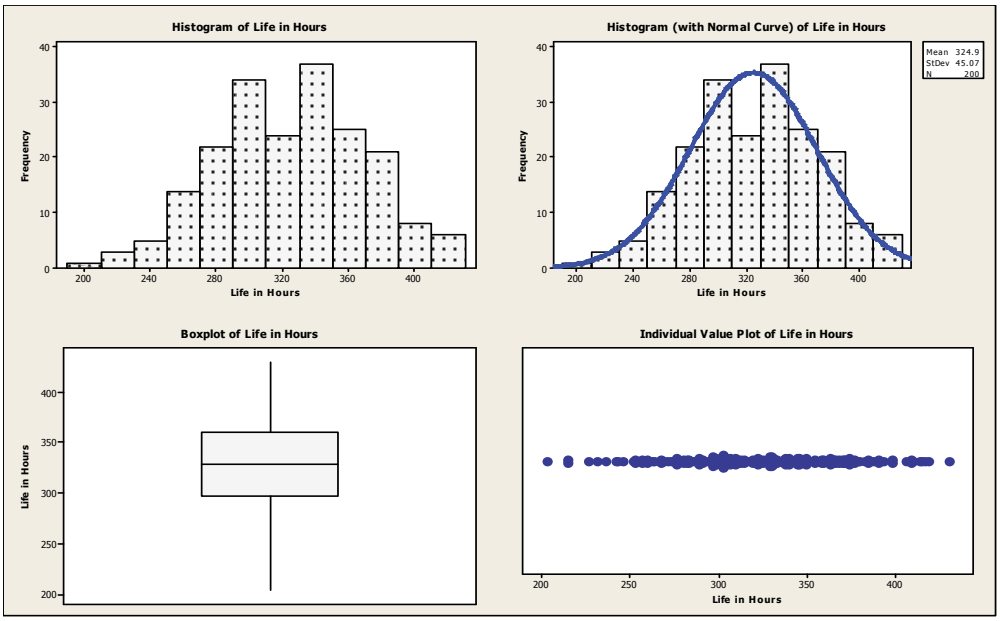


Figure 4.5: Descriptive Statistics Graphs using the Graph Options

(3) Calculate the graphical summary of the data.

This option calculates and displays several statistics along with a histogram with a normal plot, box plot, and 95% confidence intervals for the mean and the median. To produce the graphical summary, follow the steps in Table 4.15.

Table 4.15

Graphical Summary of Data	Open the worksheet **TVCOMPLIFE.MTB** From the main menu, select **Stat ➢Basic Statistics ➧➧** For **Variables**, type**….** Click **OK**

The graphical summary is shown in Figure 4.6.

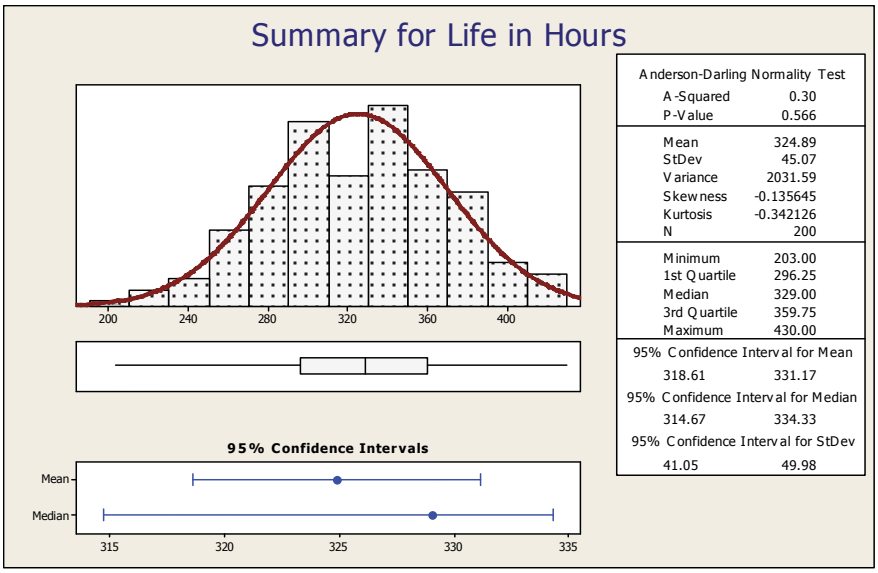


Figure 4.6: Graphical Summary of TV Component Life Data

Calculating Descriptive Statistics of Several Variables Using Minitab

The data file **LENGTH.MTW** contains 4 samples each of size 30 of the length of a certain part (in centimeters) with specification 5.50 ± 0.05. To calculate the descriptive statistics of all the samples, follow the steps in Table 4.16.

Table 4.16

Descriptive Statistics	Open the worksheet **LENGTH.MTB** From the main menu, select **Stat ➢Basic Statistics ➢....** On the left pane, highlight **Sample 1, Sample 2, Sample 3, Sample 4** and click the **Select** box for these variables to appear under **...** Click **OK**

The summary statistics for all four variables are shown in Table 4.17.

Table 4.17

Descriptive Statistics: Sample 1, Sample 2, Sample 3, Sample 4

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Sample 1	30	0	5.4980	0.00938	0.0514	5.3734	5.4646	5.5054	5.5353
Sample 2	30	0	5.4933	0.00913	0.0500	5.3967	5.4488	5.4955	5.5334
Sample 3	30	0	5.4928	0.00781	0.0428	5.4198	5.4628	5.4895	5.5192
Sample 4	30	0	5.5087	0.0118	0.0644	5.3512	5.4664	5.5161	5.5612

Variable	Maximum
Sample 1	5.6144
Sample 2	5.5817
Sample 3	5.5832
Sample 4	5.6361

Calculating Descriptive Statistics for a Specific Row or Column

If your data contains several rows and columns and you want to calculate the descriptive statistics for a particular row or column, use the following command sequence: From the main menu, select

Calc ➢Column statistics

In the **Column Statistics** dialog box, check the statistic you want to calculate. Similarly, to calculate the statistics for a row, select

Calc ➢Row statistics

Describing Data: An example

The data file **COMSTRENGTH. MTW** shows the compressive strength of a sample of 70 concrete specimens in psi (pounds per square inch). The analysis of this data is presented below using the descriptive and numerical methods in MINITAB.

1. ***Open the worksheet COMSTRENGTH.MTW and display the data in session window of MINITAB***. To do this, follow the steps in Table 4.22.

Table 4.22

Displaying Data	Open the worksheet **COMSTRENGTH.MTW** From the main menu, select **....** For **Columns, constants, and matrices to display**, type**...** or select ... Click **OK**

The data shown below will be displayed in the session window.

Data Display : Strength

3160	3560	3760	3010	2360	3210	2660	3410	3060	4310	3310
2460	2660	3060	2110	2910	3910	4210	4160	3210	3060	3310
3160	4310	3310	3260	3610	3710	2960	3460	2810	3410	3110
3310	3660	3010	3160	3060	3210	2510	2710	3660	3510	3310
3160	3410	3610	3310	3910	3060	3460	3810	2860	3160	3560
3760	3010	2360	3210	2660	3410	3060	4310	3310	3310	3160
2660	3010	3410								

2. ***Sort the data and store the sorted value in column C2. Display the sorted data in session window.*** (If the data file has the sorted data, the following steps in Table 4.23 will overwrite the data in the column)

Table 4.23

Sorting Data	Label column C2 **Sorted** From the main menu, select **...** For **Sort column(s)** box, ... In **By column** box type **C1 or Strength** (do not check the Descending box) Click on the circle next to **Column(s) of current worksheet** and type **....** Click **OK**

The data in column C1 will be sorted in increasing order and stored in column C2. You can display the sorted data in session window by following the steps in part (a). The sorted data displayed in session window are shown below.

Sorted (read row wise)

2110	2360	2360	2460	2510	2660	2660	2660	2710	2810	2860
2910	2910	2960	3010	3010	3010	3010	3060	3060	3060	3060
3060	3060	3110	3160	3160	3160	3160	3160	3160	3210	3210
3210	3210	3260	3260	3310	3310	3310	3310	3310	3310	3310
3310	3410	3410	3410	3410	3410	3460	3460	3510	3560	3560
3610	3610	3660	3660	3710	3760	3760	3810	3910	3910	4160
4210	4310	4310	4310							

3. ***Calculate the statistics based on ordered values.***

The statistics based on the ordered values are: minimum, maximum, range, median, quartiles, and interquartile range. These can be very easily calculated using MINITAB. First, we calculate the minimum, maximum, range, median, the first and third quartiles, and the interquartile range (IQR) by following the steps in Table 4.24.

Table 4.24

Descriptive Statistics	Open the worksheet **COMSTRENGTH.MTW** From the main menu, select **Stat ➢....** For **Variables**, type **C1** or ..... Click on **Statistics** tab Check the boxes next to the statistics you want to calculate Click **OK**

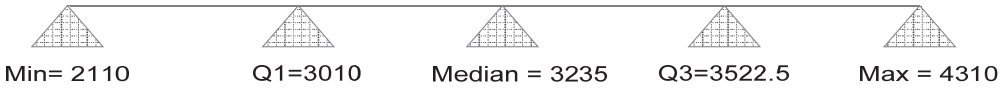
The values of the selected statistics are shown in Table 4.25.

Table 4.25

Descriptive Statistics: Strength

Variable	N	N*	Minimum	Q1	Median	Q3	Maximum	Range	IQR
Strength	70	0	2110.0	3010.0	3235.0	3522.5	4310.0	2200.0	512.5

The positions of the quartiles along with the minimum and maximum values are shown below.



Note: Q1: First quartile, Q2: Median or Second quartile, Q3: Third quartile, IQR: Interquartile Range = Q3 - Q1

From these calculated values, knowledge about the symmetry and skewness can be obtained. Symmetry is a useful concept in the data analysis. For symmetrical data, the "middle" or the average is unambiguously defined. If the data are skewed, another measure (median) should be used to describe the data.

For a symmetrical distribution, the distance between the first quartile, Q1 and the median is same as the distance from the median to the third quartile, Q3. From the calculated value in Table 4.25, we can see that the data is not symmetrical, but is close to symmetry. The distribution can also be checked by plotting a stem-and-leaf, a dot plot, a box plot, or a histogram.

4. ***Construct a stem-and-leaf plot of the data.*** Analyze the graph. What information you can obtain from this plot?

Much useful information can be obtained easily by constructing a stem-and-leaf plot. To construct this plot, follow the steps in Table 4.26.

Table 4.26

Stem-and- Leaf Plot	Open the worksheet **COMSTRENGTH.MTW** From the main menu, select **……** For **Graph variables**, type **C1** or ….. Click **OK**

The stem-and-leaf is shown below in Figure 4.9.

Figure 4.9 shows that there are 70 observations (N=70). The first column gives a cumulative count of the number of observations in that row. For example, the first row has one observation, the first and the second row has 3 observations (one observation in the first and two in the second row). The row that contains the median is enclosed in a parenthesis and this row shows a noncumulative count…….

Stem-and-Leaf Display: Strength
Stem-and-leaf of Strength N = 70
Leaf Unit = 100

```
1    2  1
3    2  33
5    2  45
9    2  6667
14   2  88999
31   3  00000000001111111
(14) 3  22222233333333
25   3  4444444555
15   3  6666777
8    3  899
5    4  1
4    4  2333
```

Figure 4.9: Stem-and-leaf Plot for Strength Data

In Figure 4.9, the second column shows the stem values; and the values to the right of the stem are the leaf values. Usually only the two leading digits in the data ….

Then the resulting two digit numbers are divided into a stem digit (the left-hand digit) and a leaf digit (the right hand digit). Consider the first few numbers in our data (look at the sorted data in part (2) of this analysis above. The numbers are

2110 2360 2360 2460 2510 2660

These are divided into stem and leaf as shown in Table 4.27.

Table 4.27

Value	*Stem*	*Leaf*	*Ignore*
2110	2	1	10
2360	2	3	60
2360	2	3	60
2460	2	4	60
2510	2	5	10
2660	2	6	60

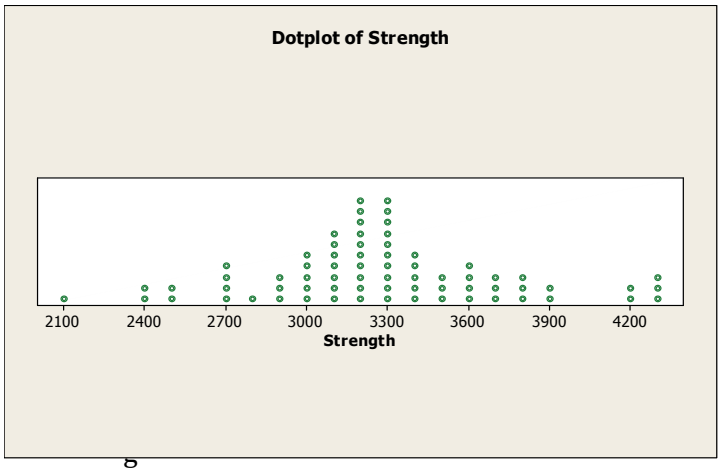
:

:

Continued...

This is because accuracy is lost after chopping the last two digits in the data. You can also see that the distribution seems to be approximately symmetrical or normal.

Figure 4.10 shows a dot plot of data using the ***Graph*** ➢**Dot Plot** option. This plot also provides information about the distribution and spread by plotting individual values.



g
u
re 4.10: Dot Plot for Strength Data

5. ***Calculate the statistics based on averages.***

The simplest statistic is the average of a set of observations or the mean. Other statistics are based on the squares, cubes, or higher exponents. These can be very easily calculated using a statistical package. We have calculated the graphical summary of the data that shows several statistics including the mean, variance, standard deviation, skewness, kurtosis, and others along with some useful graphs. To do such plot, follow the steps in Table 4.28. Figure 4.11 shows the results.

Table 4.28

Graphical Summary	Open the worksheet **COMSTRENGTH.MTW** From the main menu, select **Stat ➢.....** **:** **:** For **Variables**, type or Select **C1 or Strength** Click **OK**

(Note: once the graph is created, you may double click anywhere on the histogram and edit your graph using the "edit bar" dialog box that appears. By selecting the "custom" menu you can change the pattern and color of your graph).

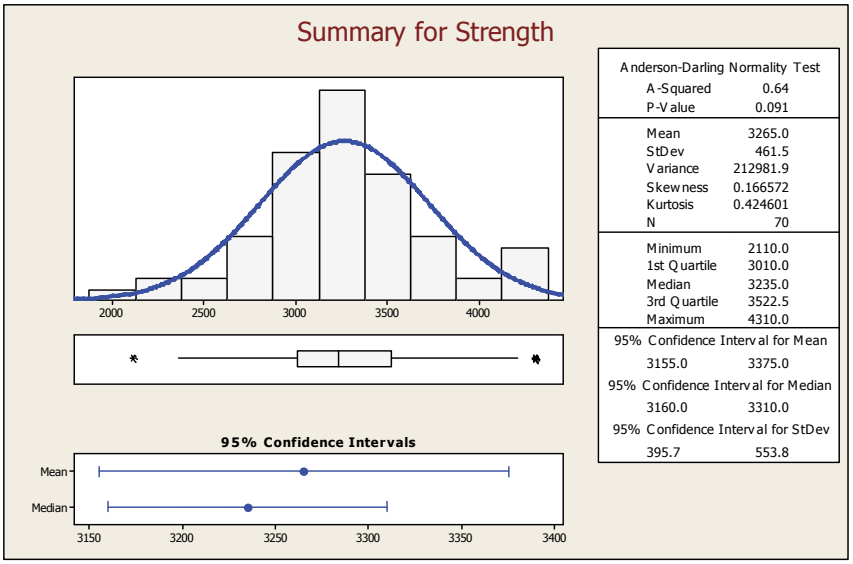


Figure 4.11: Graphical Summary of Strength Data showing the Statistics based on Ordered Values and Based on Averages

In Figure 4.11, the calculated *statistics based on the averages are shown.*

The statistics based on the averages are mean, standard deviation, variance, skewness, and kurtosis.

The statistics based on the ordered values are the minimum, first quartile (25th percentile), median (50th percentile), third quartile (75th percentile), and the maximum.

An investigation of the skewness (0.166572), and the box plot on the left reveals that the data is slightly right skewed. It is not apparent from the ….

H_0: The data follow a normal distribution.
H_1: The data do not follow a normal distribution

Use the p-value (reported under Anderson-Darling Normality Test in Figure 4.11) to test the hypothesis. The calculated p-value from Figure 4.11 is 0.091. The decision rule for conducting the test using p-value approach is given by

If $p \geq \alpha$, reject Ho

If $p < \alpha$, do not reject Ho

For a given α (5% or 0.05 for this case), we find from Figure 4.11 that $p = 0.091 > \alpha = 0.05$ therefore, we do not reject Ho and conclude that the data follow a normal distribution. Note that if you select, $\alpha = 0.10$, the null hypothesis will be rejected. We will discuss the normality tests in other chapters.

6. ***Interpret the confidence intervals in Figure 4.11.***

Figure 4.11 provides the confidence intervals for the mean, median, and standard deviation. The confidence interval can be calculated for any statistic; ….

:

and it provides the reliability of our estimate of a parameter. The narrower the confidence interval is, the more precise the estimate.

Confidence interval for the mean

By default, a 95% confidence interval for the mean is calculated using the following formula:

$$\overline{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

The above formula estimates the unknown population mean (μ) with a 95% confidence level or probability. The calculated confidence interval in Figure 4.11 is

$$3{,}155 \leq \mu \leq 3{,}375$$

which indicates that there is a 95% probability that the true unknown population mean will fall within the range 3155 and 3375 or, there is a 95% chance that ….

Confidence interval for the standard deviation

The confidence interval for the standard deviation is calculated using the following formula

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{(n-1),\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{(n-1),1-\alpha/2}}}$$

In the above equation, *s* is the sample standard deviation and n is the number of observations. $\chi^2_{(n-1),\alpha/2}$ is in general the $(1-\alpha)$ 100th percentile of the chi-square distribution with n degrees of freedom. In Figure 4.11, a 95% confidence interval for the true standard deviation (s) is calculated. This interval is

$$395.7 \leq \sigma \leq 553.8$$

Confidence interval for the median

MINITAB uses nonlinear interpolation to calculate the confidence interval for the median. This method provides a good approximation for both symmetrical and non-symmetrical distributions. A 95% confidence interval for the median in Figure 4.11 shows values between 3160 and 3310. Note that the median is a more reliable measure of central tendency when the data is non-symmetrical.

7. ***Determine the appropriate number of class intervals and the width of the classes for this data. Use the information to construct a frequency histogram.***

The approximate number of classes or the class intervals can be determined using the following formula

$$k = 1 + 3.33 \log n$$

Where, k = number of class intervals, and n= number of observations. There are n=70 observations in our example, therefore

$$k = 1 + 3.33 \log 70 = 7.144$$

The number of classes should be 7 or 8. The width of each class interval can be determined by

$$Width = \frac{Maximum - Minimum}{No.ofClasses} = \frac{4310 - 2110}{8} = 275$$

A histogram with 8 class intervals with a class width of 275 can now be constructed. Note that the above formulas do not provide exact values for the number of classes and the class width. They provide approximate values.

To construct a histogram with 8 class intervals, each having a width of 275, follow the steps in Table 4.29.

:

:

Table 4.29

Histogram	Open the worksheet **COMSTRENGTH.MTW** From the main menu, select **Graph** ➪➪➪➪ Click on **…** then click **OK** For **Graph Variables**, type or select **…** Click **OK**

A default histogram will be displayed. Next, we edit this default histogram to get the desired number of classes and the class width by following the steps in Table 4.30.

Table 4.30

Editing Graphs	**Right click** with the pointer anywhere on the histogram (or double click anywhere on the histogram to display the Edit Bars) In the **Edit Bars** dialog box, click on **….** For the **Interval Type**, select… Click on the circle next to… Click **OK**

Figure 4.12 shows the resulting histogram. Note that the program automatically adjusts the width of the class (300 in this case) for the required class intervals.

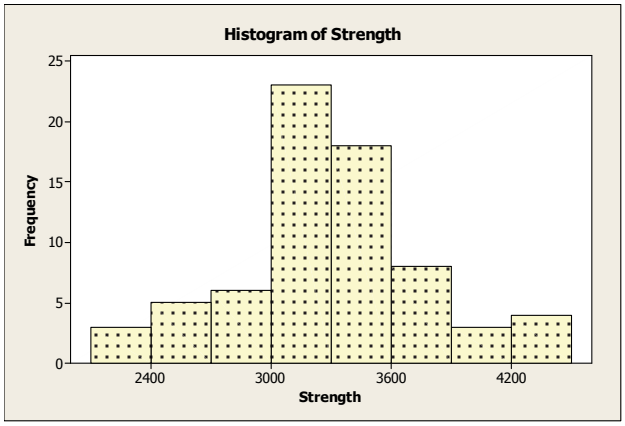


Figure 4.12: Histogram of Strength with 8 Classes

continued…

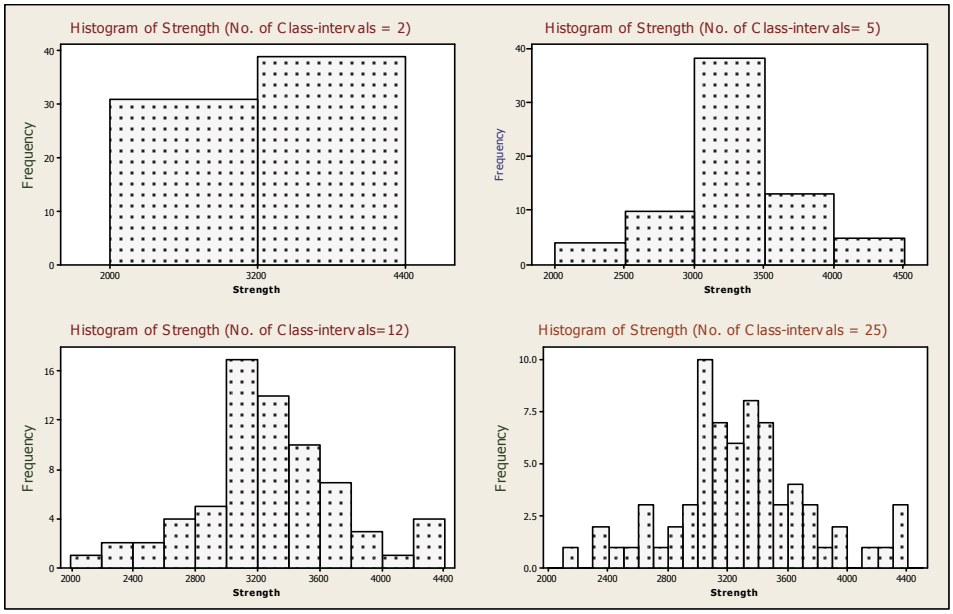


Figure 4.13: Histogram of Strength Data with Different Class Intervals

Figure 4.13 shows that the details in the distribution are lost if too few intervals are used. Similarly, using too many intervals may result into unnecessary details. Therefore, using too few or too many intervals may not display the distribution correctly.

Different Statistics used to describe the Data

Different statistics used to describe the data are summarized in Table 4.31.

Table 4.31 Summarizing Data

Statistics Based on Ordered Values	Minimum, First Quartile, Median, Third Quartile, Maximum, Interquartile Range
Statistics Based on Averages	Mean, Standard Deviation, Variance, Skewness, Kurtosis
Describe a symmetrical (bell-shaped) distribution	Mean and Standard Deviation

Examples of More Graphs and Analysis:

Note: The data file and step-wise instructions for these graphs and analyses are included in the book.

From the box plot, the shape of the data can be determined. Note that Q1, Q2, and Q3 are enclosed in a box. Q2 is the median.

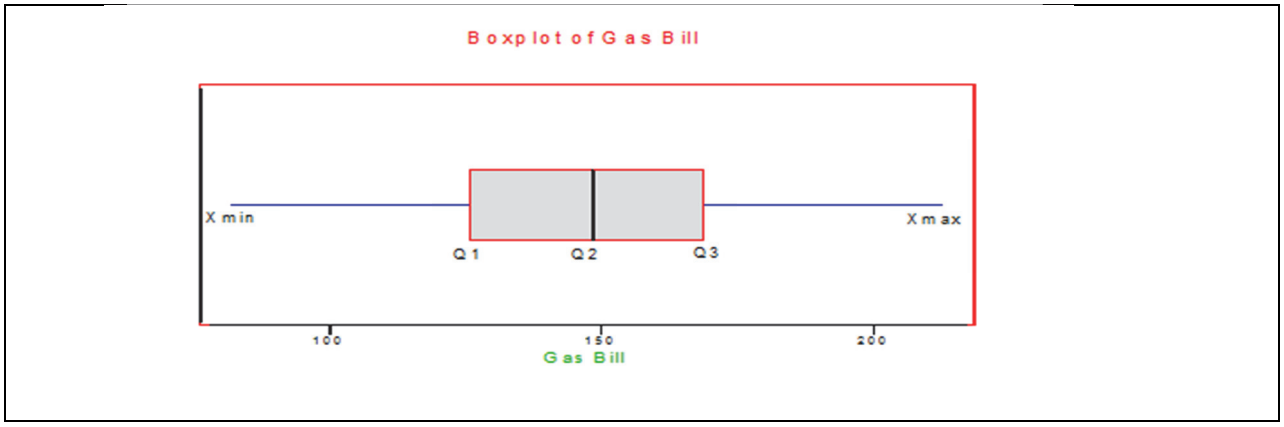
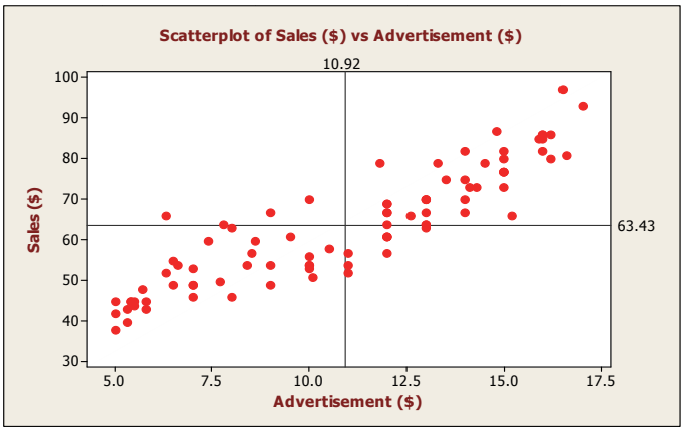


Figure 4.22 shows the scatterplot with reference lines.



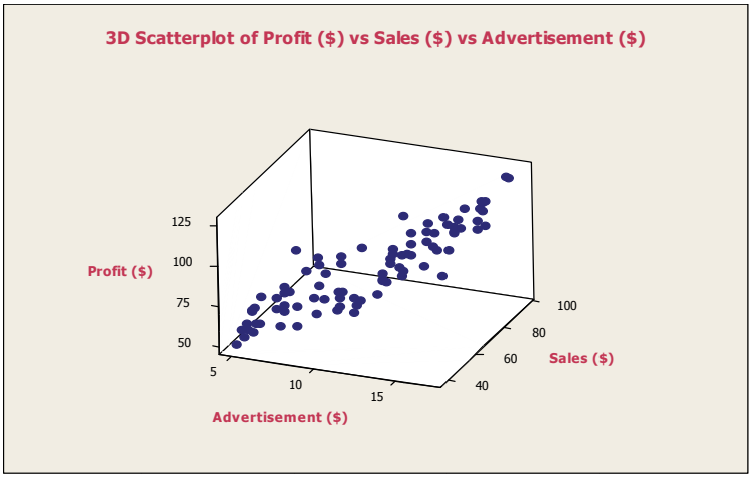


Figure 4.27: A 3D Scatterplot Showing the Relationship between Sales, Advertisement, and Profit

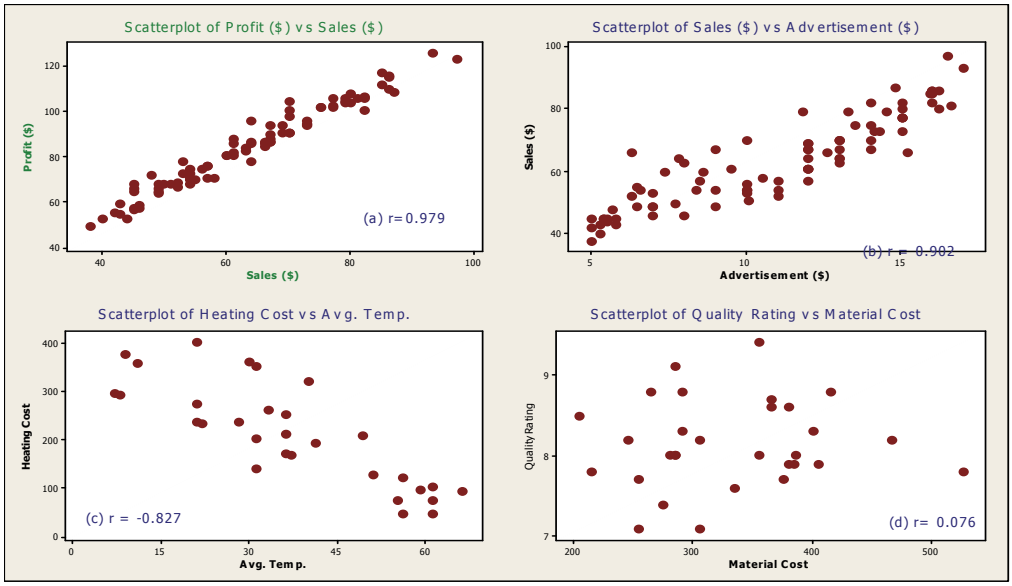


Figure 4.29: Scatterplots with Correlation (r)

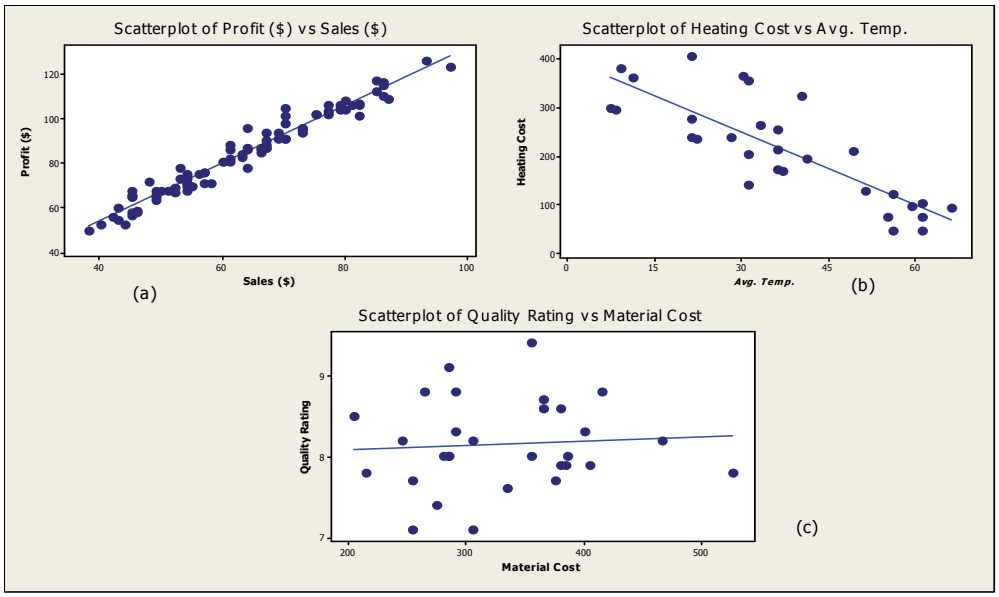


Figure 4.30: Scatterplots with Fitted Regression Lines

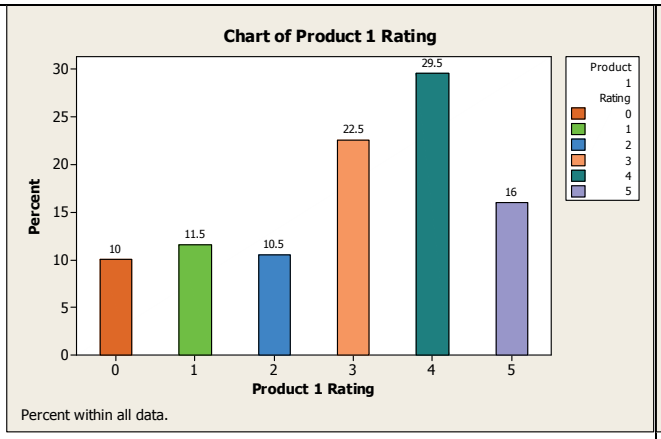


Figure 4.32: Bar Chart of Product 1 Rating

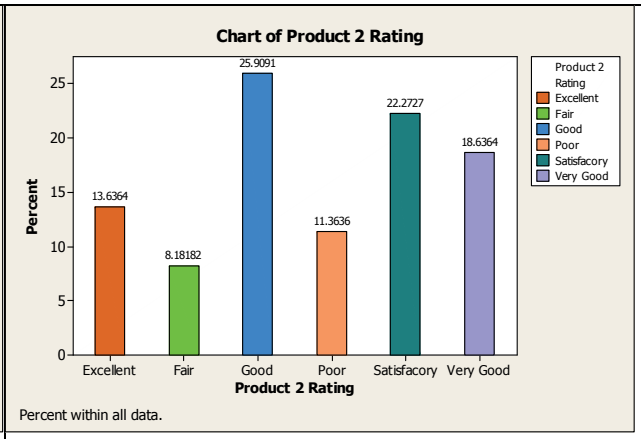


Figure 4.33: Bar Chart of Product 2 Rating

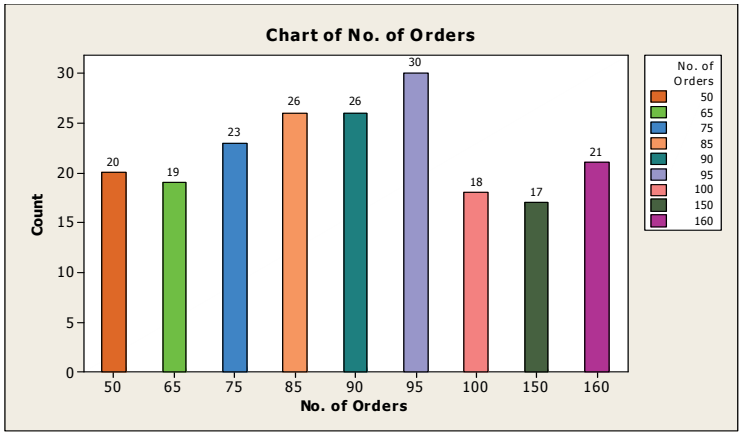


Figure 4.34: Bar Chart for Number of Orders

Table 4.55

Tally for Discrete Variables: No. of Orders

No. of Orders	Count	CumCnt	Percent	CumPct
50	20	20	10.00	10.00
65	19	39	9.50	19.50
75	23	62	11.50	31.00
85	26	88	13.00	44.00
90	26	114	13.00	57.00
95	30	144	15.00	72.00
100	18	162	9.00	81.00
150	17	179	8.50	89.50
160	21	200	10.50	100.00
N=	200			

Cross Tabulation:

A cross table of Rating vs. Supplier as shown in Table 4.57 will be displayed.

Table 4.57

Tabulated statistics: Rating, Supplier

Rows: Rating Columns: Supplier

	A	B	C	D	All
Excellent	17	19	13	11	60
Fair	11	11	7	5	34
Good	27	20	19	7	73
Poor	6	11	11	5	33
All	61	61	50	28	200

You can also create the percentages for rows and columns. To calculate the percentages

---

Example 12

The data in Table 4.18 shows the annual starting salaries for engineering and management majors for the past year. The data was collected to compare the salaries for the two majors.

Table 4.18: Starting Annual Salaries

Engineering majors (salary in thousands of dollars)

37.9 35.9 33.0 37.2 38.7 36.4 37.8 28.6 29.0 40.5 32.6 36.4 36.4 35.9 34.1
42.5 35.6 37.2 40.5 47.2

Management majors (salary in thousands of dollars)

31.7 30.8 25.3 34.0 40.1 34.4 30.0 19.3 27.0 23.4 30.3 28.3 26.5 18.9 35.4
29.4 28.3 14.5 28.3 25.0

(a) Calculate the mean, median, mode, range, variance, standard deviation, coefficient of variation, and interquartile range for the salaries of engineering majors.
The data array (data sorted in increasing order) and the following information are provided.

Engineering Major (sorted)

28.6 29.0 32.6 33.0 34.1 35.6 35.9 35.9
36.4 36.4 36.4 37.2 37.2 37.8 37.9 38.7
40.5 40.5 42.5 47.2

Note that *n = 20.* Also, the sum of the observations and the sum of square of the observations are: $\sum x_i = 733.53$, $\sum x_i^2 = 27250$

Solution: Using the above information, the calculations are shown below.

Mean

$$\bar{x} = \frac{\sum x}{n} = \frac{733.53}{20} = 36.676$$

Median

To calculate the median, use the sorted data above. The number of observations is even. Therefore, the position of the median is obtained by

$$\frac{n+1}{2} = \frac{20+1}{2} = 10.5$$

It means that the median is the average of the 10th and 11th value in the sorted data or,

$$Median = \frac{36.4 + 36.4}{2} = 36.4$$

Mode

Mode is the value that is repeated the maximum number of times. For this data

Mode = 36.4

Range

$$Range = X_{largest} - X_{smallest} = 47.2 - 28.6 = 18.6$$

Variance

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} = \frac{27250 - \frac{(733.53)^2}{20}}{19} = 18.247$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{18.247} = 4.27$$

Coefficient of Variation

$$CV = \frac{s}{\bar{x}} * 100\% = \frac{4.27}{36.676} * 100\% = 11.64\%$$

Interquartile Range

Note: To calculate Q1 and Q3, use the sorted data and the formulas as used in the previous example. For this data, Q1 = 34.505 and Q3 = 38.492. You should verify these answers**.**

$$IQR = Q_3 - Q_1 = 38.492 - 34.505 = 3.987$$

The descriptive statistics for the engineering major data using MINITAB is shown in Table 4.19. Compare the calculated values above with the computer results. Note that the computer printout does not provide the coefficient of variation, interquartile range, or the range. These can be calculated using the information from the computer printout.

Table 4.19: Summary Statistics for Engineering Major using MINITAB

Descriptive Statistics: Engineering

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Engineering	20	36.676	36.419	36.538	4.270	0.955

Variable	Minimum	Maximum	Q1	Q3
Engineering	28.606	47.242	34.505	38.492

(b) Calculate the mean, median, mode, range, variance, standard deviation, coefficient of variation, and interquartile range for the salaries of management majors.

The data array (sorted in increasing order) and the following information are provided.

Management Majors (Sorted)

14.5	18.9	19.3	23.4	25.0	25.3	26.5	27.0	28.3
28.3	28.3	29.4	30.0	30.3	30.8	31.7	34.0	34.4
35.4	40.1							

Use the sorted data and the above information to obtain the required statistics. The answers are given below.

Solution: The calculated statistics for the management data are

$$n = 20, \sum x_i = 561.02, \sum x^2 = 16431$$

Using the above values, the calculated statistics are:

Mean = 28.05

Median = 28.30

Mode = 28.3

Range = 40.11 - 14.49 = 25.62

Variance, $s^2 = 36.48$

Standard deviation, $s = 6.04$

Coefficient of variation,

$$CV = \frac{s}{\bar{x}} * 100\% = \frac{6.04}{28.05} * 100\% = 21.53$$

Interquartile range, IQR = Q3 - Q1 = 31.51 - 25.05 = 6.46

Table 4.20 shows the summary statistics for the management majors.

Table 4.20: Summary Statistics for Management Majors using MINITAB

Descriptive Statistics: Management

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Management	20	28.05	28.30	28.13	6.04	1.35

Variable	Minimum	Maximum	Q1	Q3
Management	14.49	40.11	25.05	31.51

(c) Make a summary table of the calculations.

Table 4.21 provides useful information for comparing the engineering and the management group. From the summary statistics of the two groups, the following conclusions can be drawn.

Table 4.21: Summary Table

	n	Mean	Median	Mode	Range	s^2	s	CV	IQR
Eng.	20	36.68	36.42	36.4	18.6	18.25	4.27	11.64%	3.99
Mgt.	20	28.05	28.30	28.3	25.62	36.48	6.04	21.53%	6.46

- The average salary for engineering majors is higher than the management majors. There is a difference of 8.63 thousand dollars (36.68 - 28.05) or $8630 between the two groups.
- The median salary for the engineering and management majors is 36.42 ($3642) and 28.30 ($2830) respectively. This measure becomes more important if the data are skewed.
- If we compare the variation in terms of range, we find that the management group has larger variation compared to the engineering group. The management group has a range of 25.62 or $25,620, and the engineering group has a range of 18.6 or $18,600. This indicates a larger variation for the management group.
- In terms of standard deviation, the average deviation for engineering salaries is 4.27 or $4270 compared to the 6.04 or $6040 for management salaries. This indicates that the engineering salary has less variation when compared to the management salary data.

- If we compare the coefficient of variation (CV) for the two groups, the management group indicates a larger variation (21.53%) compared to the engineering group (11.64%). Also, for the engineering group the standard deviation (s) is 11.64% of the value of the mean while for the management group, the standard deviation (s) is 21.53% of the sample mean (determined by the coefficient of variations).

(a) Which measure of central tendency (mean, median, or mode) describes the data best? If you were to compare the salaries for the two groups, would you use the mean or the median? What can you tell about the shape of the two groups?

The shape of the distribution can be determined by comparing the mean, median, and mode. The following conclusions regarding the shape can be reached by comparing the mean, median, and mode:

- If Mean = Median = Mode; the data are symmetrical
- If Mean > Median and Mean > Mode; the data are right, or positively skewed
- If Mean < Median and Mean < Mode; the data are left or negatively skewed

If the data are symmetrical, then the mean, median, and mode are all located in the same place (centrally located). For a symmetrical data, the mean, median, or mode describes the data equally well.

If the data are skewed, the median is the best measure of central tendency. For the skewed data (either left or right skewed), the median is always in the center (in between the mean and the mode). See the earlier discussion on comparing the mean, median, and mode. If we compare the two groups of data, we find the following:

Engineering majors

Mean = 36.68
Median = 36.42
Mode = 36.40

In this case, Mean > Median, and Mean > Mode, so strictly speaking the data are slightly right skewed. Note that the mean, median, and mode are very close. If this happens, we may consider the data to be almost symmetrical.

Figure 4.7 shows the histogram with a normal curve superimposed for the engineering major data. From the plot, the data seems to be almost symmetrical. Therefore, the shape may be considered symmetrical so either the mean or the median can be used to describe the data.

Management majors: *The calculated measures of the management majors are shown below.*

Mean = 28.05
Median = 28.30
Mode = 28.30

In this case, Mean < Median and Mean < Mode; so strictly speaking the shape is slightly left skewed. On the other hand, the values of the mean, median, and mode are close therefore,

we may consider the data to be almost symmetrical. Figure 4.8 shows the histogram with the normal curve superimposed for the management data.

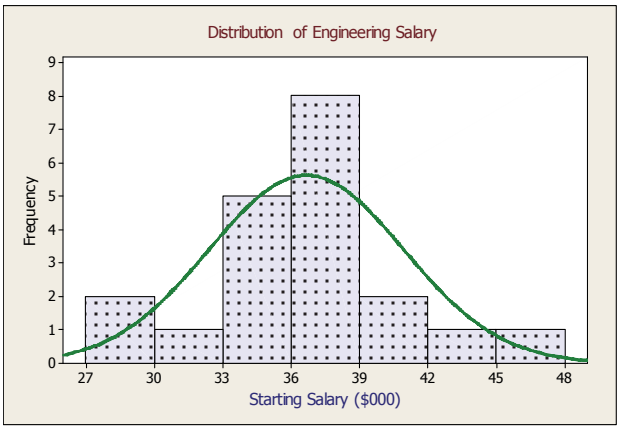


Figure 4.7: Histogram with Normal Curve for Engineering Data

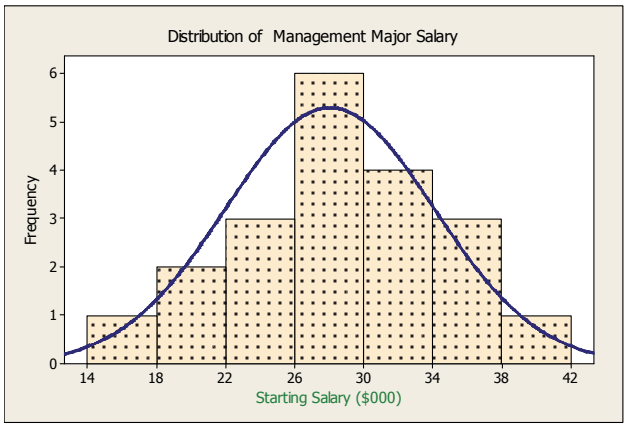


Figure 4.8: Histogram with Normal Curve for Management Data

Since the distribution of the data is approximately symmetrical, either the mean or the median can be used to describe the data. We will see later that there is a distinct advantage if the data have a symmetrical shape. If the data are symmetrical, the measures of central tendency (mean, median, and mode) are more representative of the data.

In this example, both the engineering and the management salary data can be considered symmetrical. Calculating the measures of central tendency and the measures of variation provides useful information that is helpful in analyzing and drawing conclusions from the data.