Six Sigma Quality: Concepts & Cases- Volume I
STATISTICAL TOOLS IN SIX SIGMA DMAIC PROCESS WITH MINITAB® APPLICATIONS

Chapter 4

Using Statistics To Summarize Data Sets: Concepts & Computer Analysis

© Amar Sahay, Ph.D.
Master Black Belt

Chapter Outline

Numerical Methods of Describing Data

- *Measures of central tendency or measures of location,*
- *Measures of position,*
- *Measures of variation or dispersion, and*
- *Measures of shape.*

Calculating Descriptive Statistics

Use MINITAB to verify the Empirical Rule

Application of the Empirical Rule

Use Random Number Generator, Descriptive Statistics, and Graphs to Check if the Random Number Generator Produces a Uniform Distribution

Describing Data: An Example

- *Data Display*
- *Sorting the Data*
- *Calculate the Statistics based on Ordered Values*
- *Construct a stem-and-leaf plot of Data*
- *Calculate the Statistics based on Averages*
- *Interpret the Confidence Intervals*
- *Confidence Interval for the Mean*
- *Confidence Interval for the Standard Deviation*
- *Confidence Interval for the Median*
- *Determine Appropriate Class-intervals*
- *Summarizing Data*

Relating Continuous Variables: Scatterplots and Correlation

- *Constructing a Simple Scatterplot*
- *Adding Reference Lines to the Scatterplot*
- *Scatterplot with a Categorical Variable*
- *A 3D Scatterplot*

Correlation

- *Calculating Coefficient of Correlation (r)*
- *Scatterplot with Regression*

Describing Categorical Variables

- *Creating a Simple Tally*
- *Bar Chart for Product 1 Rating*
- *Tally and Bar Chart for Product 2 Rating*
- *Another Example of Tally*

Cross Tabulation: Two-Way Table

- *Cross Tabulation with Two and Three Categorical Variables*

This chapter deals with the numerical methods of describing and analyzing data. Numerical methods include several statistical measures that are used to describe the data. In the previous chapter several graphical techniques and their importance in describing and summarizing data were discussed in detail. This chapter deals with the measures that are used to describe and summarize numerical variables. Although visual representation of data such as, the charts and graphs are very helpful in summarizing, visualizing the pattern, and drawing conclusions from the data; in many cases, additional numerical measures are needed to describe the data. In this chapter we will investigate the numerical measures that involve computations to describe, summarize, and draw meaningful conclusions from the data. We also discuss numerous computer applications and data analysis concepts in this chapter.

NUMERICAL METHODS OF DESCRIBING DATA

In this chapter we investigate the numerical measures that involve computations to describe, summarize, and draw meaningful conclusions from the data. We also discuss numerous computer applications and data analysis concepts in this chapter.

The numerical methods of describing data can be divided into following categories:
(1) Measures of central tendency or measures of location,
(2) Measures of position,
(3) Measures of variation or dispersion, and
(4) Measures of shape.

All these measures are discussed in detail in this chapter.

:
:

MEASURES OF POSITION

Other important measures in describing the data are percentile and quartiles. These are known as ***measures of position*** and are described below.

Percentiles and Quartiles

The p^{th} percentile of a data set is a value, such that at least p percent of the values are less than or equal to this value, and at least (100-p) percent of the values are greater than or equal to this value.

:

:

Q_1 = first quartile or 25th percentile

Q_2 = 2nd quartile or 50th percentile or the median

Q3 = 3rd quartile or 75th percentile

The first quartile or Q_1 is the value such that 25% of the observations are below Q_1 and 75% of the values are above Q_1. The other quartiles can be interpreted in a similar way. Using the formula below, we can determine the percentile and quartile for any data set.

Calculating Percentiles and Quartiles

To find a percentile or quartile

- Arrange the data in increasing order
- Find the location of the percentile using the following formula

$$L_p = (n+1)\frac{P}{100} \qquad (4.4)$$

Where, L_p = location of the percentile

n = total number of observations

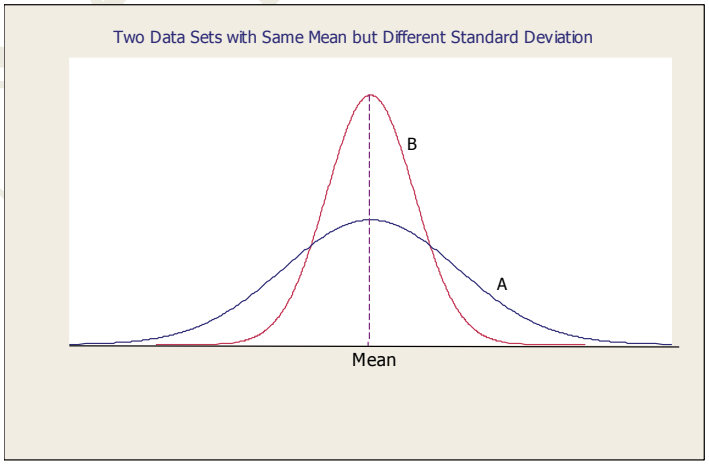
P = desired percentile

:

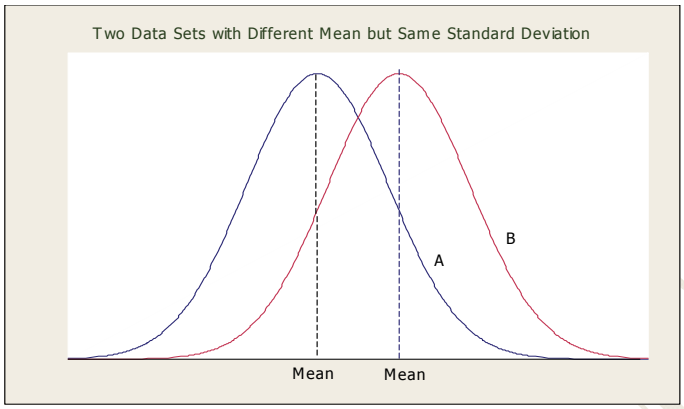
MEASURES OF VARIATION

;

:



Data Sets A and B with Same Mean but Different Variation



Data Sets A and B with Same Variation but Different Mean

The variability in the data is measured using the following measures. These are known as the ***measures of variation***.

(1) Range
(2) Variance
(3) Standard Deviation
(4) Coefficient of Variation
(5) Interquartile Range

What do the variance (s^2) and the standard deviation (s) tell us?

The variance and standard deviation measure the average deviation (or the scatter) around the mean. The variance is the average of squared distances from the mean. In calculating the variance, the computation results in squared units, such as dollar squared, inch squared, etc. This makes the interpretation difficult. Therefore, for practical purposes we calculate the standard deviation by taking the square root of the variance. Taking the square root of the variance results in the original unit of data (it is no more dollars squared or inch squared but, dollars or inches). In other words, the variance is the measure of variation affected by the units of measurement whereas; the standard deviation is measured in the same unit as the data.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad (4.6)$$

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \qquad (4.7)$$

Note: it is critical to understand the concepts of variance and standard deviation in data analysis because measuring and reducing variability is one of the major goals of Six Sigma projects.

Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum \left(x_i - \overline{x}\right)^2}{n-1}} \quad \text{or,}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} \qquad (4.8)$$

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 **Data ➢Display Data** For **Columns, constants, and matrices to display**, type **C1** or sel **Strength** 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
:
:
:
  260  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 **Data ➢Sort** For **Sort column(s)** box, type **C1** or select **C1 Strength** **:** **:** 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
:
:

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 ➢Basic Statistics ➢Display Descriptive Statistics** For **Variables**, …… : : Check the boxes next to the statistics you want to calculate Click **OK**

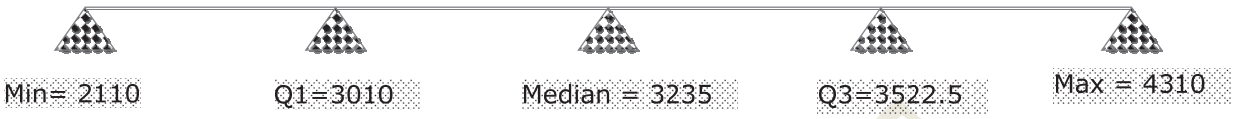
The values of the selected statistics are shown in Table 4.25.

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 4.20, 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. 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 parenthesis and this row shows a noncumulative count. In Figure 4.9, the row indicated by (14) contains the median. The number 14 means there are 14 observations in that row. Once the median row is found, the cumulative count begins from the bottom row.

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

:

The plot in Figure 4.9 provides several useful statistics. We can see that the minimum value is approximately 2100, and the maximum is 4300. Since there are 70 values, the median is half way between the 35 and 36 values. The 35^{th} and 36^{th} values from the stem and leaf are both 3200, so the median is 3200. These values are different from the exact values calculated in Table 4.20. 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 below 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.

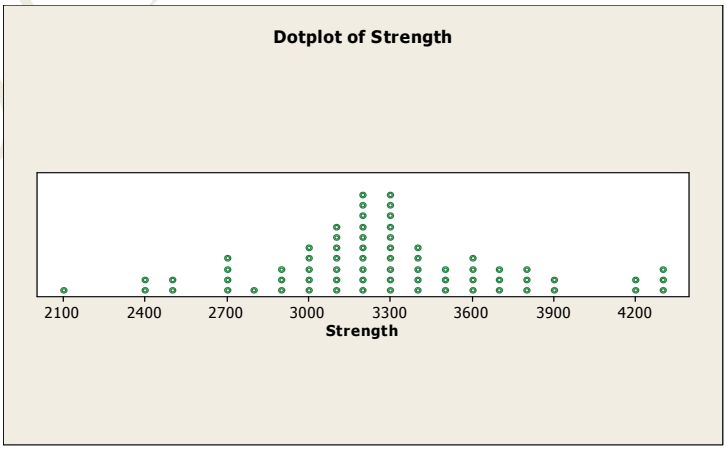


Figure 4.10: Dot Plot for Strength Data

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**
:	
:	
	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).

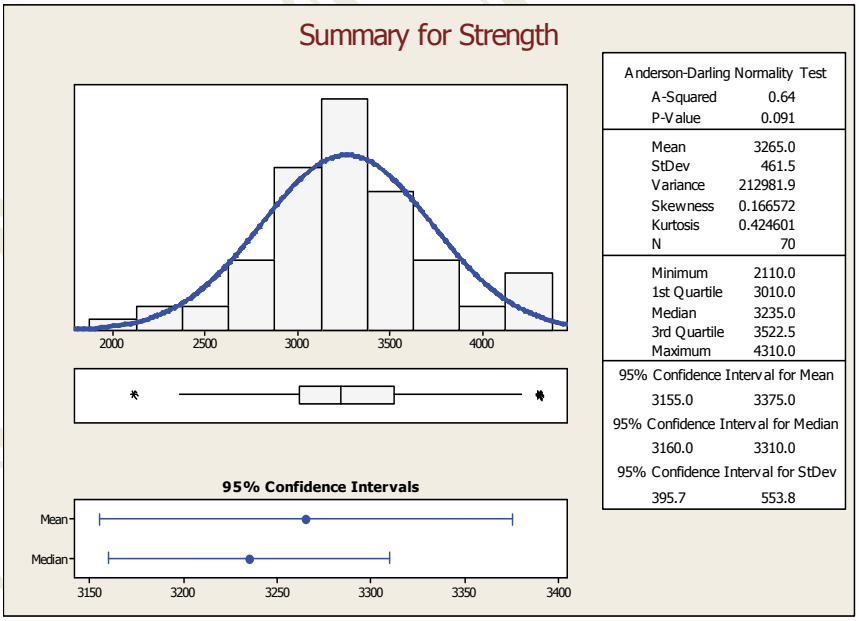


Figure 4.11: Graphical Summary of Strength Data showing the Statistics based on Ordered Values and Based on Averages

In Figure 4.11, the calculated *statistics based on the averages are 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 histogram (which shows that the data is approximately symmetrical). For practical purposes, we can conclude that the data is approximately symmetrical. **However, the Anderson Darling normality test can be used to test if the data are symmetrical. We test the following hypothesis:**

H_0: The data follow a normal distribution.

H_1: The data do not follow a normal distribution

Use the p-value approach to test the hypothesis. The calculated p-value from Figure 4.10 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 an observation will fall within the range 3155 and 3375 or, there is a 95% chance that the true mean is included between 3155 and 3375.

Confidence interval for the standard deviation

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

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

Where, *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

$$395.7 \le \sigma \le 553.8$$

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 **……..** **:** **:** Click on For **Graph Variables**, type or select **C1 Strength** 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.

:

:

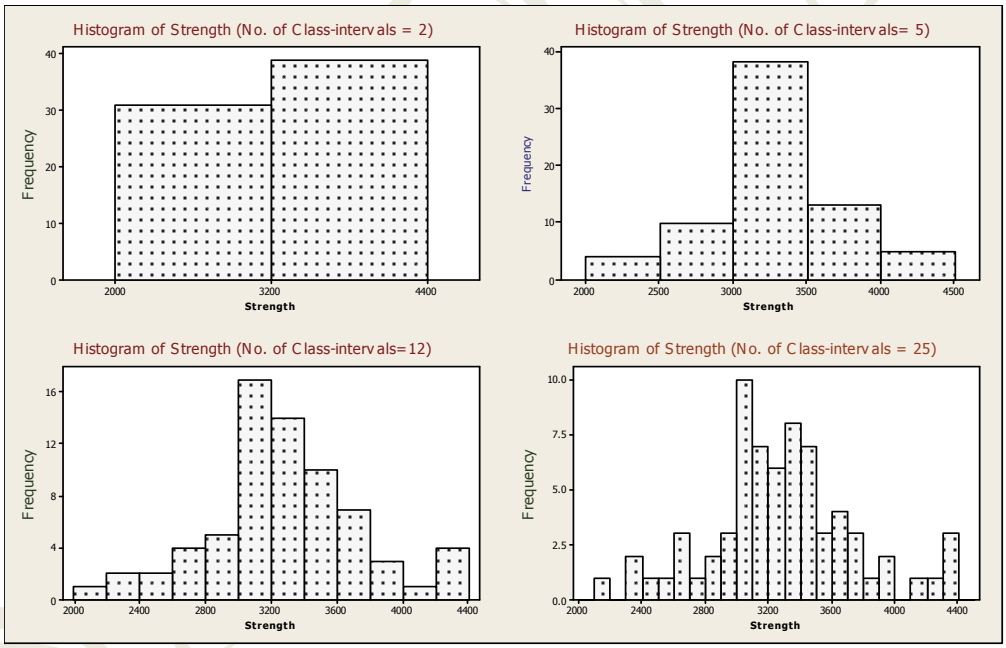


Figure 4.13: Histogram of Strength Data with Different Class Intervals

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

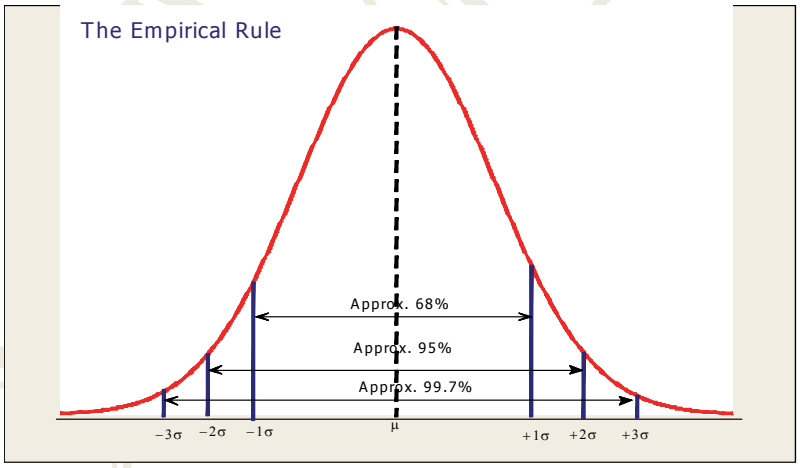
RELATIONSHIP BETWEEN THE MEAN AND THE STANDARD DEVIATION

Chebyshev's Theorem

Within k standard deviation of the mean at least $(1-\frac{1}{k^2})$ ***percent of the values occur where, k is given by***

$$k=\frac{x-\bar{x}}{s} or, k=\frac{x-\mu}{\sigma}$$

The Empirical Rule



CALCULATING COEFFICIENT OF CORRELATION (r)

The data file **SALES&AD.MTW** contains the data for advertisement dollars, sales, and profit for a company.

:

:

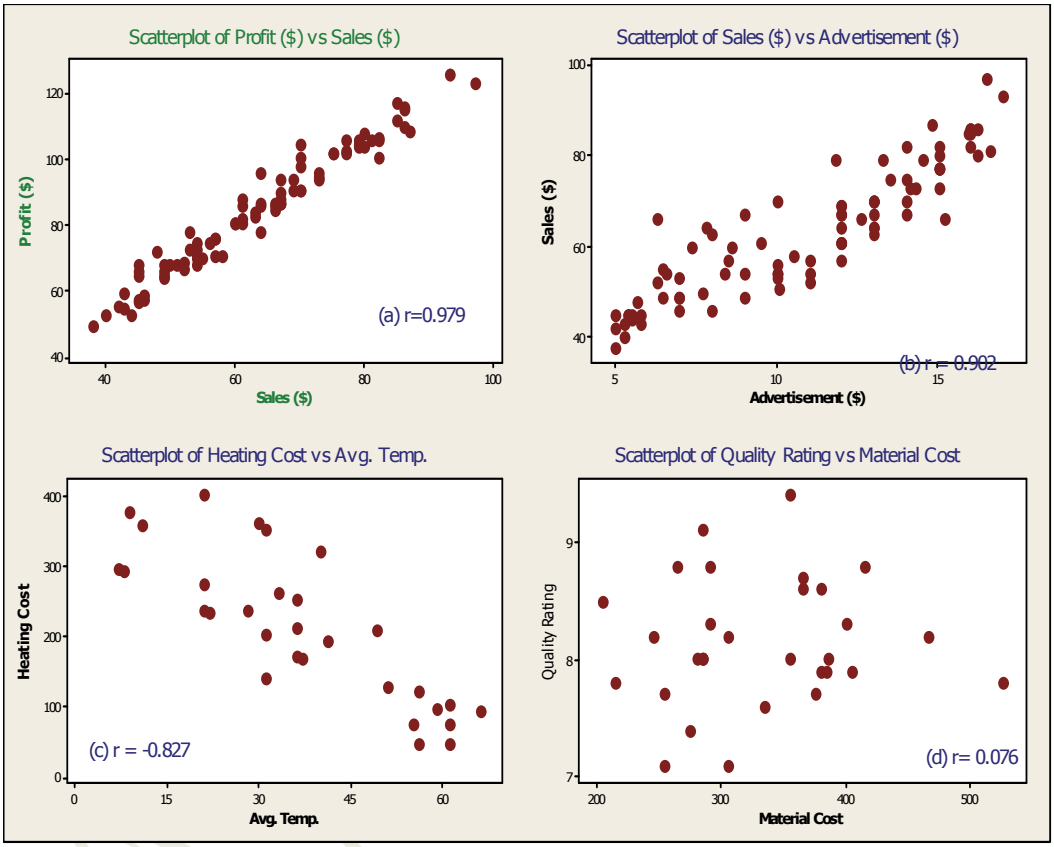


Figure 4.29: Scatterplots with Correlation (r)

Figure 4.29, (a) and (b) shows strong positive correlation; (c) shows a negative correlation while (d) shows a weak correlation.

SCATTERPLOT WITH REGRESSION

This option in MINITAB allows fitting a line or curve to the scatter plot. It is useful to see the best fitting line or a curve to the data.

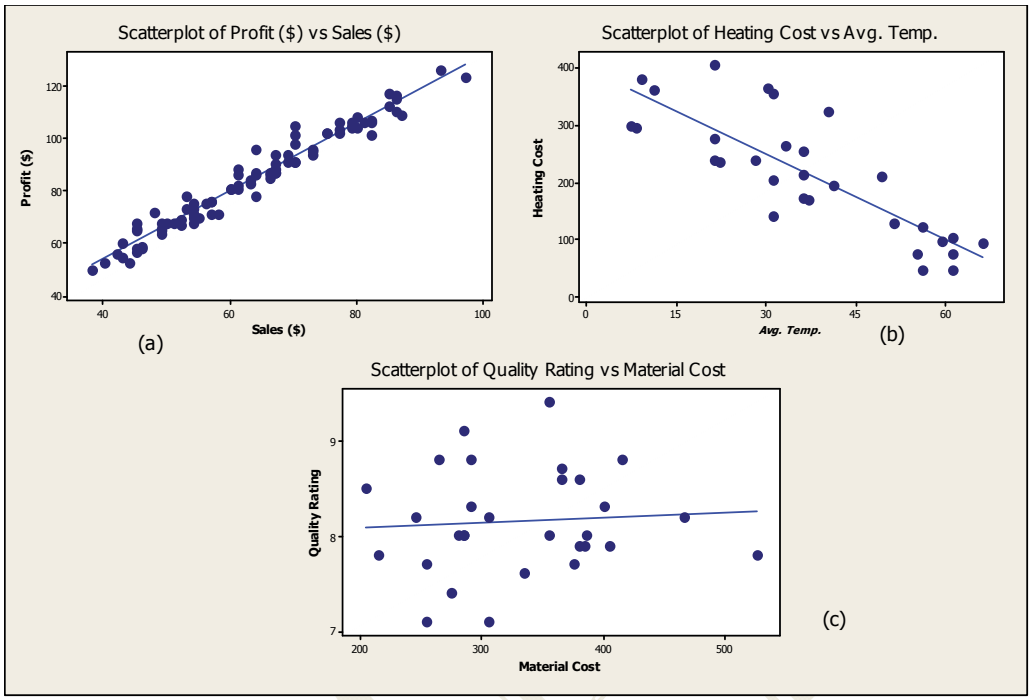
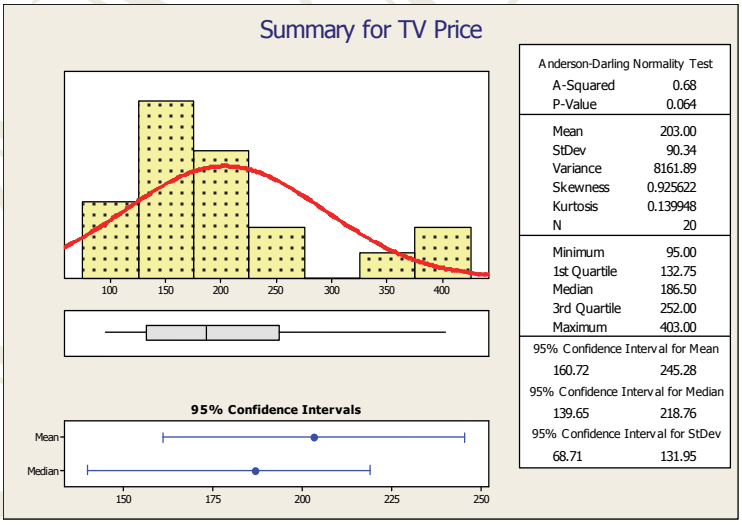


Figure 4.30: Scatterplots with Fitted Regression Lines

MEASURES OF SHAPE: SKEWNESS AND KURTOSIS



More details with examples on how to perform data analysis using computer can be found in Chapter 4.

To buy chapter 4 or Volume I of Six Sigma Quality Book, please click on our products on the home page.