Advanced Statistical Approaches to Quality

A. Ben Hamza

CONCORDIA INSTITUTE FOR INFORMATION SYSTEMS ENGINEERING

Concordia University
Engineering and
Computer Science

Concordia University
2018

Contents

Chapter

Chapter

Chapter 1

Introduction to Statistical Quality Control

Quality means doing it right when no one is looking.

Henry Ford

1.1 What is Quality Control?

Quality can mean different things to different people. Quality is often used to signify excellence of a manufactured product or a service received. From the manufacturing standpoint quality is simply conformance to specifications. Quality is also defined as meeting the requirements of the customer. Exceeding customer expectations has been extremely effective in building a loyal customer base. It is estimated to be five to seven times more costly to attain a new customer that it is to retain a current one, so it makes a lot of sense to go that extra step [3].

There are various definitions attributed to historical leaders in the field of quality. Juran described quality as "fitness for use".

Quality control is a process employed to ensure a certain level of quality in a product or service, which may include whatever actions a business deems necessary to provide for the control and verification of certain characteristics of a product or service. The basic goal of quality control is to ensure that the products, services, or processes provided meet specific requirements and are dependable, satisfactory, and fiscally sound.

Essentially, quality control involves the examination of a product, service, or process for certain minimum levels of quality. The goal of a quality control team is to identify products or services that do not meet a companys specified standards of quality. If a problem is identified, the job of a quality control team or professional may involve stopping production temporarily. Depending on the particular service or product, as well as the type of problem identified, production or implementation may not cease entirely.

Usually, it is not the job of a quality control team or professional to correct quality issues. Typically, other individuals are involved in the process of discovering the cause of quality issues and fixing them. Once such problems are overcome, the product, service, or process continues production or implementation as usual.

1.2 What is Statistical Quality Control?

Statistical quality control (SQC) is a term used to describe the activities associated with ensuring that goods and services satisfy customer needs. SQC uses statistical analysis based on measurements taken from a process or from a sample of products or services, to make decisions regarding the quality of goods and services. The statistical methods of SQC may be divided into two main categories: Statistical process control (SPC) and acceptance sampling. SPC refers to the use of statistical methods to measure and control the performance of a process to ensure that the output meets customer needs. Acceptance sampling is a methodology of taking samples from lots of materials or products and inspecting the items to determine if the items meet customer requirements. SPC may be used to to help control almost all processes that can measured or monitored to ensure that the process performs within limits.

Deming, a statistician who gained fame by helping Japanese companies improve quality after the World War II, believed that quality and productivity increase as variability decreases and, because all things vary, statistical methods of quality control must be used to measure and gain understanding of the causes of the variation. Many companies, particularly those in the auto industry, have adopted Deming's philosophy and approach to quality.

The causes of variations in a product quality characteristic may be broadly classified into two main categories: common causes of variation (variation due to the system itself) and special causes of variation (variation due to factors external to the system).

- *Chance* or *common* causes: when only these causes of variations are present in a process, the process is considered to be **stable** or **in-control**. Examples of such causes include atmospheric pressure or temperature changes in the production area, worker fatigue, and fluctuations caused by hiring, training, and supervisory policies and practices. Common causes of variation are the responsibility of management.

- *Assignable* or *special* causes: when these causes of variations are present in a process, variation will be excessive and the process is considered to be **unstable** or **out-of-control**. Examples of such causes include tampering or unnecessary adjusting the process when it is inherently stable, using a wrong tool or an incorrect procedure. Special causes of variation are the responsibility of workers and engineers.

1.3 A Brief History of Quality Control

The quality movement can trace its roots back to medieval Europe, where craftsmen began organizing into unions called guilds in the late 13th century. Until the early 19th century, manufacturing in the industrialized world tended to follow this craftsmanship model. The factory system, with its emphasis on product inspection, started in Great Britain in the mid-1750s and grew into the Industrial Revolution in the early 1800s.

In the early 20th century, manufacturers began to include quality processes in quality practices. After the United States entered World War II, quality became a critical component of the war effort: Bullets manufactured in one state, for example, had to work consistently in rifles made in another. The armed forces initially inspected virtually every unit of product; then to simplify and speed up this process without compromising safety, the military began to use sampling techniques for inspection, aided by the publication of military-specification standards and training courses in Walter Shewharts statistical process control techniques.

The birth of total quality in the United States came as a direct response to the quality revolution in Japan following World War II. The Japanese welcomed the input of Americans Joseph M. Juran and W. Edwards Deming and rather than concentrating on inspection, focused on improving all organizational processes through the people who used them. By the 1970s, U.S. industrial sectors such as automobiles and electronics had been broadsided by Japans high-quality competition. The U.S. response, emphasizing not only statistics but approaches that embraced the entire organization, became known as total quality management (TQM). By the last decade of the 20th century, TQM was considered a fad by many business leaders. But while the use of the term TQM has faded somewhat, particularly in the United States, its practices continue.

In the few years since the turn of the century, the quality movement seems to have matured beyond Total Quality. New quality systems have evolved from the foundations of Deming, Juran and the early Japanese practitioners of quality, and quality has moved beyond manufacturing into service, healthcare, education and government sectors.

1.4 What is Statistical Process Control?

A process is the transformation of a set of inputs, and may include customer services, productions systems, and administration activities. In each area or function of an organization there will be many processes taking place. Each process may be analyzed by an examination of the inputs and outputs. This will determine the action necessary to improve quality [2]. In order to produce an output which meets the requirements of the customer, it is necessary to define, monitor and control the inputs to the process.

Statistical process control (SPC) is a procedure in which data is collected, organized, analyzed, and interpreted so that a process can be maintained at its present level of quality or improved to a higher level of quality. SPC requires that the process be improved continuously by reducing its variability. SPC refers to a number of different methods for monitoring and assessing the quality of manufactured goods. Combined with methods from the Design of Experiments, SPC is used in programs that define, measure, analyze, improve, and control development and production processes. These programs are often implemented using "Design for Six Sigma" methodologies.

1.5 REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, 2005.

[2] J. S. Oakland, *Statistical Process Control*, Butterworth-Heinmann, 2003.

[3] G. M. Smith, *Statistical Process Control and Quality Improvement*, Prentice Hall, 2003.

Chapter 2

Descriptive Statistics

> Numerical quantities focus on expected values, graphical summaries on unexpected values.
>
> *John Tukey*

Statistics is the science of collecting, organizing, summarizing and analyzing data in order to draw conclusions. Data are typically collected from a **population**, which is defined as a collection of units. Then, we define a **sample** as a subset of units selected from a population. A central problem in statistics is to obtain information about a population from a sample. Statistics is often divided into two branches: **descriptive statistics** and **inferential statistics**. Descriptive statistics focus on the collection, analysis, presentation, and description of a set of data, whereas inferential statistics focus on making decisions about a large set of data (i.e. population) from a subset of the data (i.e. sample). Both descriptive and inferential statistics can be used to analyze the population and sample data. Numerical descriptive measures computed from the sample data are often called **statistics**, while numerical descriptive measures of the population are called **parameters**. The population parameters are typically unknown. For example, the unknown mean μ and variance σ^2 of the population are usually estimated using sample statistics.

2.1 Descriptive Statistics

One important use of descriptive statistics is to summarize a collection of data in a clear and understandable way. For example, the manufacturing manager's job is to present results effectively so that the appropriate action can be taken. The most powerful tool for expressing the results of any study is the graph. By presenting a graph, the manufacturing manager need not explain it in words. Graphical methods are better suited than numerical methods for identifying patterns in the data. It is usually easier for people to grasp the results from a graph than from any other form of analysis. Numerical approaches are more precise and objective. Since the numerical and graphical approaches compliment each other, it is wise to use both.

Descriptive statistics are very important, as if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore allow us to present the data in a more meaningful way which allows simpler interpretation of the data. Typically, there are two general types of numerical descriptive measures computed from the sample data, namely measures of central tendency and measures of spread. Suppose, then, that our data set is of the form $x_1, \ldots, x_n$, where each x_i is a number.

2.1.1 Measures of Central Tendency

The measures of central tendency or location are ways of describing the central position of a frequency distribution for a group of data. We can describe this central position using a number of statistics, including the mean, median, and mode. Because we almost always think of our data set as a sample, we will refer to these statistics as the sample mean, sample median, and sample mode. Hence, we will use the terms sample and data set interchangeably when dealing with descriptive statistics.

Let $x_1, \ldots, x_n$ be a sample of n measurements or observations. Denote by $x_{(1)}, \ldots, x_{(n)}$ the ordered sample, sorted in ascending order, where $x_{(1)}$ are $x_{(n)}$ are the smallest and largest data values, respectively.

Definition 2.1

- *The sample mean or arithmetic average $\bar{x}$ is given by*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1.1}$$

- *The sample median, denoted by Q_2, is given by*

$$Q_2 = \begin{cases} x_{(k+1)} & \text{if } n = 2k+1 \text{ (odd number)} \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{if } n = 2k \text{ (even number)} \end{cases} \tag{2.1.2}$$

- *The sample mode is the value that occurs most frequently in a data set.*

The sample median is the middle value in an ordered list of an odd number of observations. If the number is even, the median is defined as the average of the two middle values.

It is important to note that when there are multiple values occurring equally frequently in the data set, then the mode is chosen as the smallest of those values.

Example 2.1 *The following data represent the number of defectives observed each day over a 15-day period for a manufacturing process. Calculate the sample mean, sample median, and sample mode of the defect data.*

Day	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Defects	*10*	*10*	*6*	*12*	*6*	*9*	*16*	*20*	*11*	*10*	*11*	*11*	*9*	*12*	*11*

Solution: The sample mean, sample median, and sample mode are given by

$$\bar{x} = 10.9333, \quad Q_2 = 11, \quad \text{mode} = 11.$$

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> mean(X)
>> median(X)
>> mode(X)
```

Python code

```
>> import numpy as np
>> from scipy.stats import mode
>> X = np.array([10, 10, 6, 12, 6, 9, 16, 20, 11, 10, 11, 11, 9, 12, 11])
>> Xbar = np.mean(X) #sample mean
>> Q2 = np.median(X) #sample median
>> print('Xbar={0:.2f}, Q2={1:.2f}'.format(Xbar, Q2))
>> M = mode(X) #sample mode
>> print("The mode is {} with a count of {}".format(M.mode[0], M.count[0]))
```

2.1.2 Measures of Spread

The measures of central tendency only locate the center of the data; they do not provide information on how the data are spread. The measures of spread help us summarize how spread out the data are. To describe this spread, a number of statistics are available to us, including the range, quartiles, variance and standard deviation. Variation is very important in quality control because it determines the level of conformance of the production process to the set standards. For instance, if we are manufacturing tires, an excessive variation in the depth of the treads of the tires would imply a high rate of defective products.

Definition 2.2

- *The sample range, R, is given by* $R = x_{(n)} - x_{(1)}$
- *The sample variance,* s^2*, is given by*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.1.3}$$

- *The sample standard deviation is the positive square root of the sample variance. That is,* $s = \sqrt{s^2}$.

The range is the simplest of all measures of variability. It is the difference between the highest and the lowest values of a data set. Sample variances and standard deviations are used as estimators of a populations variance and standard deviation. Using $n-1$ in the denominator instead of n yields a better estimate of the population. It is worth pointing out that the smaller the sample standard deviation, the closer the data are scattered around the mean. If the sample standard deviation is zero, then all the data observed are equal to the mean.

Example 2.2 *Calculate the sample range, sample variance, and sample standard deviation for the defect data given in Example 2.1.*

Solution: The sample range, sample variance, and sample standard deviation for the defect data are

$$R = 14, \quad s^2 = 12.0667, \quad s = 3.4737.$$

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> range(X)
>> var(X)
>> std(X)
```

Python code

```
>> import numpy as np
>> from scipy.stats import mode
>> X = np.array([10, 10, 6, 12, 6, 9, 16, 20, 11, 10, 11, 11, 9, 12, 11])
>> R = np.ptp(X)  #sample range: ptp (peak to peak)
>> print('R={0:.2f}'.format(R))
>> var = np.var(X, ddof=1) #sample variance
>> s = np.std(X, ddof=1) #sample standard deviation
>> print('var = {0:.2f}, s = {1:.2f}'.format(var, s))
```

2.1.3 Graphical Representation of Data

When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description or graphical description. Graphical representations can make data easy to interpret by just looking at graphs. Dot plot, bar chart, histogram, stem-and-leaf graph, Pareto chart, box plot, and scatter plots are types of graphs commonly used in statistics.

Dot plot:

A dot plot is a plot that displays a dot or a point for each value in a data set along a number line. If there are multiple occurrences of a specific value, then the dots will be stacked vertically. The frequency or the frequency count for a data value is the number of times the value occurs in the data set. A dot plot gives us, for example, information about how far the data are scattered and where most of the observations are concentrated.

Example 2.3 *The following data gives the number of defective motors received in 20 shipments*

```
8   12   10   16   6   25   21   15   17   5   26   21   29   8   10   21   10   17   15   13
```

Construct the dot plot for this data.

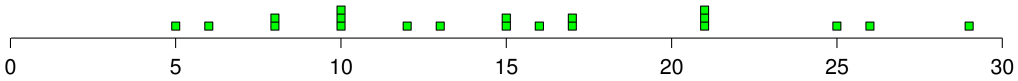


FIGURE 2.1: Dot plot for the data on defective motors that are received in 20 shipments.

Solution: Figure 2.1 shows the dot plot for the on on defective motors. We see that 70% of the time, the number of defective motors was between 8 and 21.

Example 2.4 *Construct the dot plot for the data given in Example 2.1.*

Solution: Figure 2.2 shows the dot plot for the defective data set. Observe that since there are multiple occurrences of specific observations, the dots are stacked vertically. The number of dots represents the frequency count for a specific value. For instance, the value of 10 occurred 3 times since there are 3 dots stacked above the value of 10.

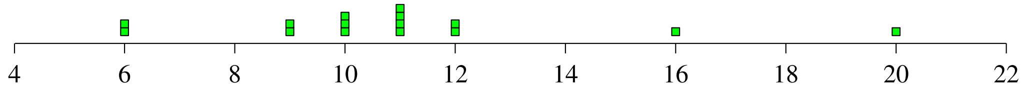


FIGURE 2.2: Dot plot for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

Bar chart:

A bar chart presents each category of an attribute type variable as a bar whose length is the frequency or percentage of values falling into a category. Bar charts permit the visual comparison of data by displaying the magnitude of each category as a vertical (or horizontal) bar.

Example 2.5 *Display the bar chart for the data given in Example 2.1.*

Solution: Figure 2.3 shows the bar chart for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> bar(X,'g','EdgeColor',[1 0.5 0.5]);
>> xlabel('Days','fontsize',14,'fontname','times');
>> ylabel('Defects','fontsize',14,'fontname','times');
```

Python code

```
>> import numpy as np
>> import matplotlib.pyplot as plt
>> import seaborn as sns; sns.set()
>> X = np.array([10, 10, 6, 12, 6, 9, 16, 20, 11, 10, 11, 11, 9, 12, 11])
>> plt.bar(np.arange(len(X)), X)
>> plt.xlabel('Days')
>> plt.ylabel('Defects')
```

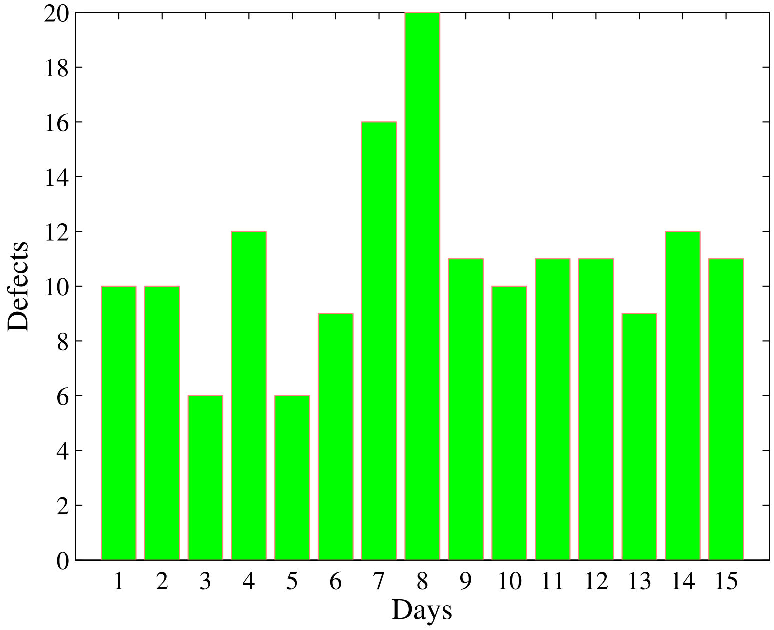


FIGURE 2.3: Bar chart for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

Histogram:

A histogram is a special bar chart for measurement data. Its purpose is to graphically summarize the distribution of a data set. The construction of a histogram starts with the division of a frequency distribution into equal classes, and then each class is represented by a vertical bar. A histogram gives an estimate of the shape of the distribution of the population from which the sample was taken. Histograms and relative-frequency histograms provide effective visual displays of data organized into frequency tables. In these graphs, we use bars to represent each class, where the width of the bar is the class width. For histograms, the height of the bar is the class frequency, whereas for relative-frequency histograms, the height of the bar is the relative frequency of that class. The relative frequency for any class is obtained by dividing the frequency for that class by the total number of observations.

$$\text{Relative frequency} = \frac{\text{Frequency for class}}{\text{Total number of observations}} \tag{2.1.4}$$

Example 2.6 *Construct the histogram for the data given in Example 2.1.*

Solution: Figure 2.4 shows the relative frequency histogram for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> a = min(X):1:max(X);
>> ncount = histc(X,a);
>> relativefreq = ncount/length(X);
>> bar(a, relativefreq,1)
>> xlabel('Defects')
>> ylabel('Relative frequency')
```

Python code

```
>> import numpy as np
>> import matplotlib.pyplot as plt
>> import seaborn as sns; sns.set()
>> X = np.array([10, 10, 6, 12, 6, 9, 16, 20, 11, 10, 11, 11, 9, 12, 11])
>> plt.hist(X, bins=len(X), density=True)
>> plt.xlabel('Defects')
>> plt.ylabel('Relative frequency')
```

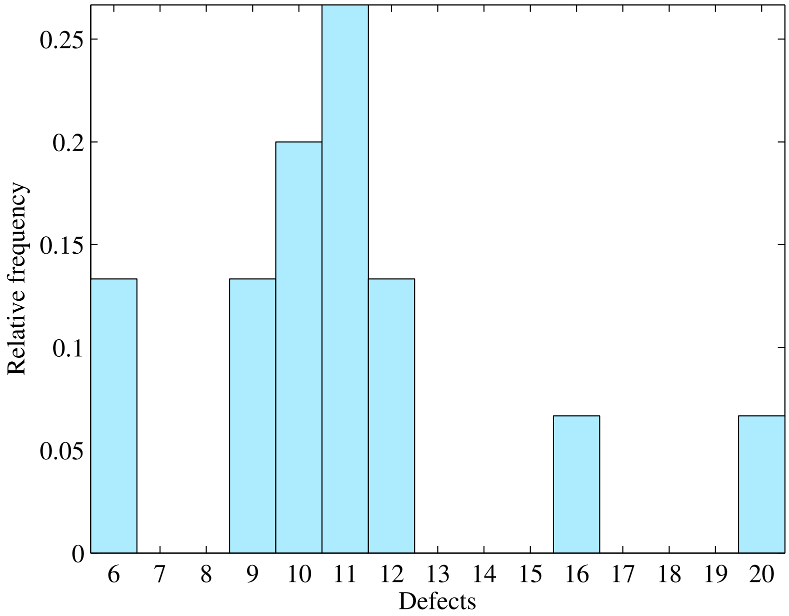


FIGURE 2.4: Histogram for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

Stem-and-Leaf Diagram:

A stem-and-leaf plot or diagram is a data plot that uses part of a data value as the stem to form groups or classes and part of the data value as the leaf: A stem-and-leaf plot has an advantage over a grouped frequency distribution, since a stem-and-leaf plot retains the actual data by showing them in graphic form. The stem-and-leaf diagram is essentially composed of two parts: the stem, which is on the left side of the diagram, and the leaf on the right. It resembles a histogram that has been turned on its side.

The first step in creating a stem-and-leaf diagram is to reorganize the data in ascending (or descending) order.

Example 2.7 *Construct the stem-and-leaf diagram for the data given in Example 2.1.*

Solution: Figure 2.5 shows the stem-and-leaf diagram for the data on the number of defectives observed each day over a 15-day period for a manufacturing process. The ordered data are: 6, 6, 9, 9, 10, 10, 10, 11, 11, 11, 11, 12, 12, 16, 20. The numbers that start with 0 (i.e. single digits) have 6, 6, 9 again, and 9 as the second digits. The numbers that start with 1 have 0, 0, 0, 1, 1, 1, 1, 2, 2, 6 as the second digits. Finally, the only number that starts with 2 has 0 as the second digit. The stem-and-leaf diagram shows that most of the data are clustered between 10 and 16.

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> sort(X)
ans =
     6     6     9     9    10    10    10    11    11    11    11    12    12    16    20
>> StemLeafPlot(X)

 0 | 6 6 9 9
```

```
0 | 6 6 9 9
1 | 0 0 0 1 1 1 1 2 2 6
2 | 0
```

FIGURE 2.5: Stem-and-leaf diagram for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

```
1 | 0 0 0 1 1 1 1 2 2 6
2 | 0
```

Pareto Chart:

A Pareto chart or diagram is a special type of bar chart in which the categories of an attribute type variable are shown on the x-axis, the frequencies in each category (listed from largest to smallest frequency) are shown on the left side y-axis, and the cumulative percentage of frequencies are shown on the right side y-axis. The key idea behind Pareto chart is to separate the "vital few" from the "trivial many". A Pareto chart is a type of bar chart in which the horizontal axis represents categories of interest. When the bars are ordered from largest to smallest in terms of frequency counts for the categories, a Pareto chart can help you determine which of the categories make up the critical few and which are the insignificant many. A cumulative percentage line helps you judge the added contribution of each category.

Example 2.8 *Construct the Pareto chart for the data given in Example 2.1.*

Solution: Figure 2.6 shows the Pareto chart for the data on the number of defectives observed each day over a 15-day period for a manufacturing process. Observe that the categories have been ordered from the highest frequency to the lowest frequency.

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> pareto(X)
```

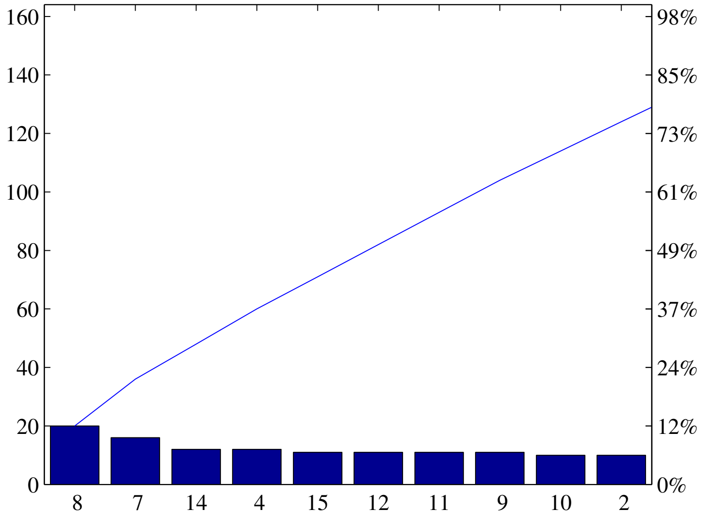


FIGURE 2.6: Pareto chart for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

Box Plot:

The box plot (also called box-and-whisker plot) is an alternative graphical representation method to either the stem-and-leaf plot or the histogram. Recall that the median Q_2 (also called second quartile) divides the lower 50% of a sample from its upper 50%. The first quartile, denoted Q_1, divides the bottom 25% of the sample from the top 75%. The third quartile, denoted Q_3, divides the bottom 75% of the sample from the top 25%. These 3 quartiles, in conjunction with the smallest and largest values of the sample, form the so-called **five-number summary**.

A box plot provides an excellent visual summary of many important aspects of a distribution, and it is a convenient way of graphically depicting groups of numerical data through their five-number summaries: the smallest data value (sample minimum $x_{(1)}$), first quartile (Q_1), median (Q_2), third quartile (Q_3), and largest data value (sample maximum $x_{(n)}$). A box plot may also indicate which observations, if any, might be considered outliers (marked with a '+' on the graph), as shown in Figure 2.7. The advantage of using the box plot is that when it is used for multiple variables, not only does it graphically show the variation between the variables but it also shows the variations within the ranges.

The purpose of a box plot is not only to show how the data are spread but also to make obvious the presence of outliers, i.e. observations (or measurements) that are unusually large or small relative to the other values in a data set. Outliers typically are attributable to one of the following causes:

- The measurement is observed, recorded, or entered into the computer incorrectly.
- The measurement comes from a different population.
- The measurement is correct, but represents a rare (chance) event.

To determine the presence of outliers, we first need to find the interquartile range (IQR), which is defined as: IQR $= Q_3 - Q_1$. The IQR measures the vertical distance of the box, i.e. the distance between the first and third quartiles. An outlier is defined as any observation away from the closest quartile by more than 1.5(IQR), i.e. any measurement smaller that the lower inner fence $LIF = Q_1 - 1.5(\text{IQR})$ or greater than the upper inner fence $UIF = Q_3 + 1.5(\text{IQR})$ is considered an outlier. Thus, a box plot helps provide information on whether outliers exist in the data. The vertical dashed lines are referred to as whiskers, and they extend to the most extreme observation inside the inner fences. If one whisker is longer than the other, then the distribution of the data is probably skewed in the direction of the longer whisker. As can be seen in Figure 2.7, the largest measurement inside the fence is the second-largest observation.

The box plot can be used to describe the shape of a data distribution by looking at the position of the median line compared to Q_1 and Q_3, the lower and upper ends of the box. If the median is close to the middle of the box, the distribution is fairly symmetric, providing equal-sized intervals to contain the two middle quarters of the data. If the median line is to the left of center, the distribution is skewed to the right; if the median is to the right of center, the distribution is skewed to the left. Also, for most skewed distributions, the whisker on the skewed side of the box tends to be longer than the whisker on the other side. If the data is a matrix, then there is one box plot per column.

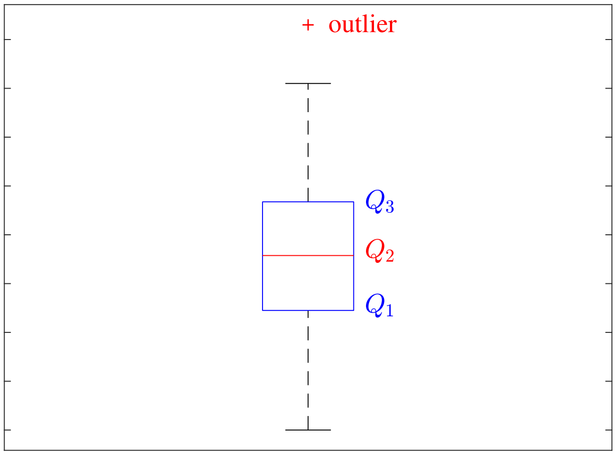


FIGURE 2.7: Illustration of box plot.

Example 2.9 *The following data give the number of persons who take the bus during the off-peak time schedule from Central Station to South Shore in Montreal:*

```
12 12 12 14 15 16 16 16 16 17 17 18 18 18 19 19 20 20 20 20
20 20 20 20 21 21 21 22 22 23 23 23 24 24 25 26 26 28 28 28
```

(i) Calculate the mean, median, and mode for the data.

(ii) Determine the five-number summary for the data.

(iii) Construct the box plot for the data.

Solution: From the box plot plot in Figure 2.8, we can see that the data are symmetric and do not contain any outliers.

MATLAB code

```
>> X = [12 12 12 14 15 16 16 16 16 17 17 18 18 18 19 19 20 20 20 20 ...
        20 20 20 20 21 21 21 22 22 23 23 23 24 24 25 26 26 28 28 28];
>> mean(X)
ans =
    20
>> median(X)
ans =
    20
>> mode(X)
ans =
    20
>> fivenumbersummary(X)
ans =
    min: 12
    max: 28
     Q1: 17
     Q2: 20
     Q3: 23
>> boxplot(X);
```

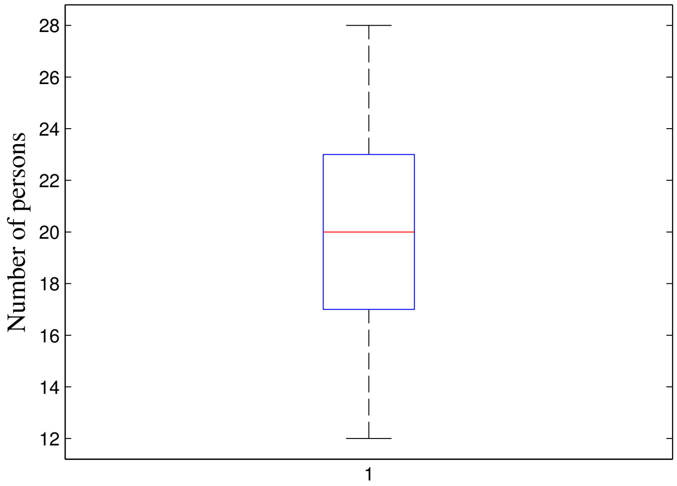


FIGURE 2.8: Box plot for the data on the number of persons who take the bus during the off-peak time.

Example 2.10 *Compute the five-number summary and construct the box plot for the data given in Example 2.1.*

Solution: Figure 2.9 shows the box plot for or the data on the number of defectives observed each day over a 15-day period for a manufacturing process. The interquartile range is $\text{IQR} = Q_3 - Q_1 = 11.75 - 9.25 = 2.5$. So any observation smaller than $Q_1 - 1.5(\text{IQR}) = 5.5$ or greater than $Q_3 + 1.5(\text{IQR}) = 15.5$ is considered as an outlier. Therefore, the observations 16 and 20 (i.e. the numbers of defectives observed on days 7 and 8) are outliers. Before removing the outliers from the data set, we need to make a concerted effort to find the cause of the outliers. An investigation may discover that these measurements were correctly recorded or they represent defects that correspond to exceptional parts being manufactured. Notice also that the lower whisker is longer that upper one, indicating that the number of defectives are negatively skewed.

MATLAB code

```
>> X = [10 10 6 12 6 9 16 20 11 10 11 11 9 12 11];
>> fivenumbersummary(X)
>> boxplot(X);
>> ylabel('Defects','fontsize',14,'fontname','times');
```

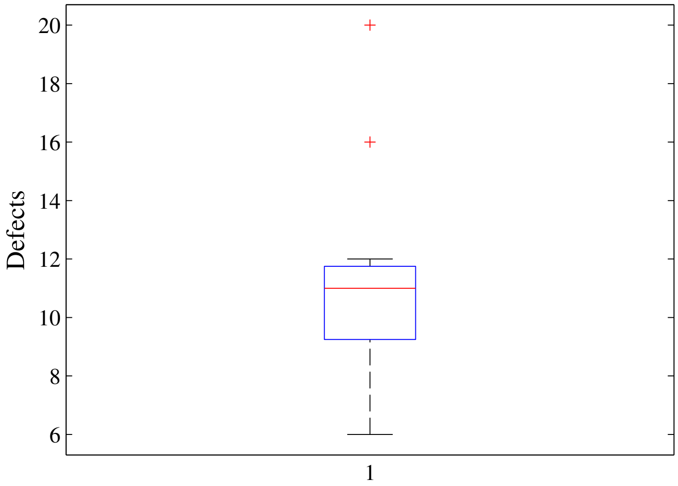


FIGURE 2.9: Box plot for the data on the number of defectives observed each day over a 15-day period for a manufacturing process.

Example 2.11 *The following data represent the temperature of two rooms: one was built with metal door and window frames and the other one without any metal.*

With metal	*52*	*81*	*83*	*79*	*89*	*89*	*98*	*96*	*98*	*99*	*95*	*99*	*99*	*99*	*101*
Without metal	*58*	*62*	*65*	*71*	*59*	*60*	*99*	*60*	*96*	*93*	*87*	*89*	*92*	*85*	*81*

(i) Construct the box plot.

(ii) Is there a difference between the two rooms and the level of temperature variability within the two rooms?

Solution:

(i) Figure 2.10 shows the box plot for the data of the temperature of the rooms.

(ii) The graphs show that there is a large disparity between the two groups, and for the room with metal the heat level is predominantly below the median. For the room with metal, the observation number 52 is an outlier.

For the room without metal, the temperatures are more evenly distributed, albeit most of the observations are below the median. Notice that the upper whisker is relatively longer that lower one, indicating that the temperatures are positively skewed.

MATLAB code

```
>> data = [ 52  81  83  79  89  89  98  96  98  99  95  99  99  99  101
            58  62  65  71  59  60  99  60  96  93  87  89  92  85   81];
>> char = {'With metal','Without metal'};
>> boxplot(data',char);
```

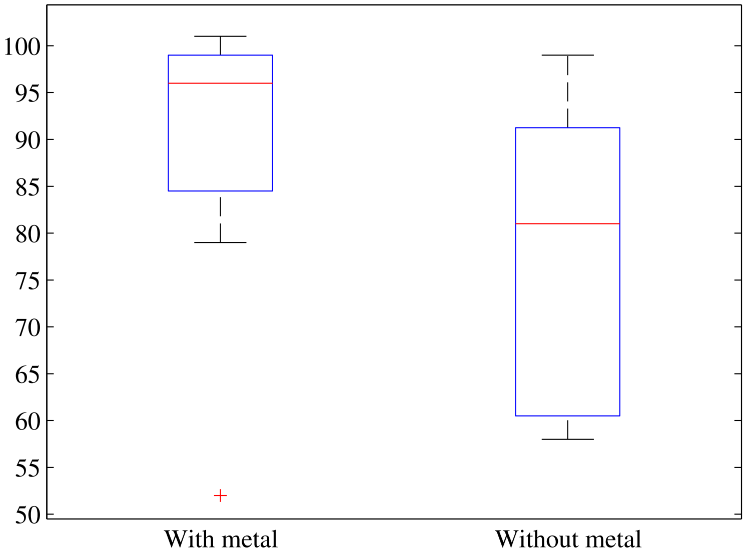


FIGURE 2.10: Box plot for the data on the temperature of the rooms.

Normal Probability Plot:

The normal probability plot is a graphical technique for assessing whether or not a data set is approximately normally distributed. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. The shape in the normal probability plot can be described as follows:

- For a skewed distribution, we should see a systematic non-linear tendency.
- For a heavy tailed distribution, we should see the data clumping away from the line. This usually results in a systematic deviation away from the line in a form of an S.
- For a normal distribution, we should expect to see a linear tendency. It can have weak deviations in the tail, but the overall tendency is linear.

Example 2.12 *Suppose that seventeen randomly selected workers at a detergent factory were tested for exposure to a Bacillus subtillis enzyme by measuring the ratio of forced expiratory volume (FEV) to vital capacity (VC).*

```
0.61 0.70 0.76 0.84 0.63 0.72 0.78 0.85 0.64 0.73 0.82 0.85 0.67 0.74 0.83 0.87 0.88
```

Is it reasonable to conclude that the FEV to VC (FEV/VC) ratio is normally distributed?

Solution: FEV is the maximum volume of air a person can exhale in one second; VC is the maximum volume of air that a person can exhale after taking a deep breath. Figure 2.11 shows the normal probability plot for FEV/VC. The plot has the sample data displayed with the plot symbol '+'. Superimposed on the plot is a line joining the first and third quartiles of each column of X (a robust linear fit of the sample order statistics.) This line is extrapolated out to the ends of the sample to help evaluate the linearity of the data. The tendency appears to be linear, hence it is reasonable to believe that FEV/VC is normally distributed.

MATLAB code

```
>> X = [.61 .70 .76 .84 .63 .72 .78 .85 .64 .73 .82 .85 .67 .74 .83 .87 .88];
>> normplot(X);
```

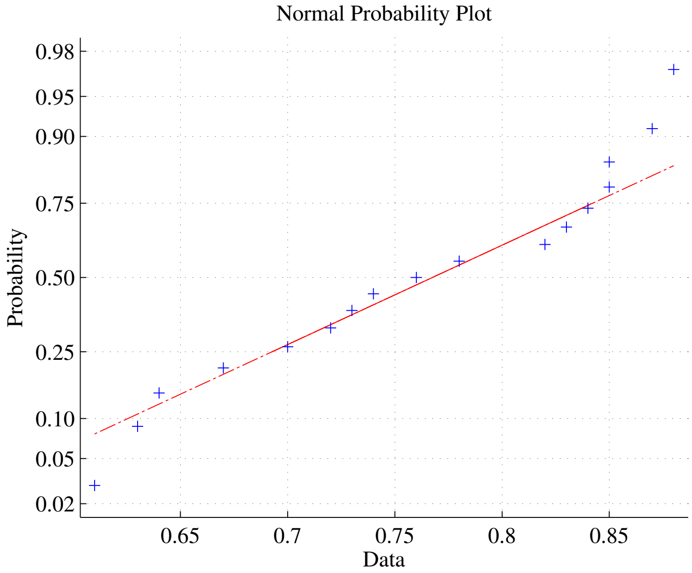


FIGURE 2.11: Normal probability plot for FEV/VC.

Scatter Plot:

In simple correlation and regression studies, data are collected on two quantitative variables to determine whether a relationship exists between the two variables. Let $x_1, \dots, x_n$ and $y_1, \dots, y_n$ be two samples of size n, and denote by $\bar{x}$, s_x, $\bar{y}$ and s_y their sample mean standard deviations, respectively. An important first step in studying the relationship between the variables x_i and y_i is to graph the data. A **scatter plot** is a graph in which each plotted dot represents an observed pair (x_i, y_i) of values (see Figure 2.12). The value of x_i is plotted with respect to the horizontal axis, and the value of y_i is plotted with respect to the vertical axis. The scatter plot provides a visual impression of the nature of the relation between the x and y values in a bivariate data set. The variable along the vertical axis is called the dependent variable, and the variable along the horizontal axis is called the independent variable. Our visual impression of the closeness of the scatter to a linear relation can be quantified by calculating the correlation coefficient. Figure 2.12 shows the correspondence between the appearance of a scatter plot and the value of the correlation coefficient.

Definition 2.3

- *The sample covariance is defined as*

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2.1.5}$$

- *The sample correlation coefficient is defined as*

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{2.1.6}$$

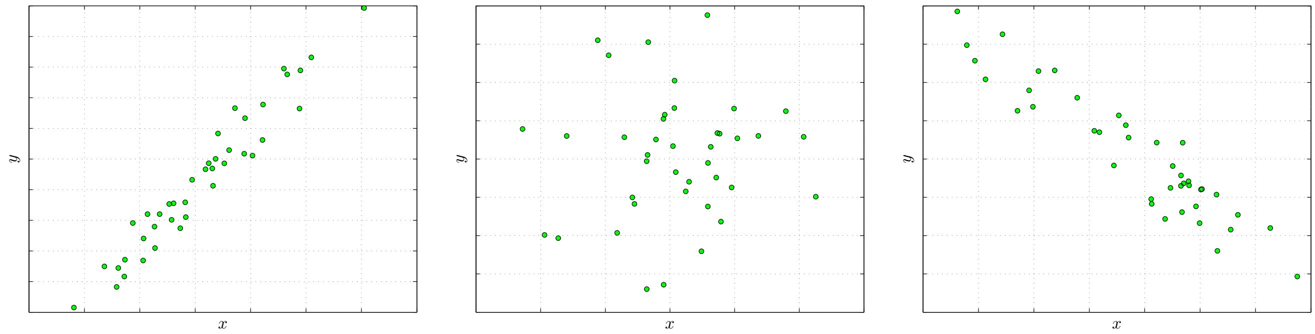
The sample correlation coefficient measures the strength and direction of a relationship between two variables using sample data. The range of the correlation coefficient is from -1 to l.

Example 2.13 *The following data relate the high temperature (°F) reached on a given day and the number of cans of soft drinks sold from a particular vending machine in front of a grocery store. Data were collected for 15 different days.*

Temperature	*70*	*75*	*80*	*90*	*93*	*98*	*72*	*75*	*75*	*80*	*90*	*95*	*98*	*91*	*98*
Quantity	*30*	*31*	*40*	*52*	*57*	*59*	*33*	*38*	*32*	*45*	*53*	*56*	*62*	*51*	*58*

(i) Calculate the sample covariance and sample correlation coefficient.

(ii) Construct the scatter plot.



(a) $r_{xy} = 0.95$ (b) $r_{xy} = 0.20$ (c) $r_{xy} = -0.95$

FIGURE 2.12: Correspondence between the values of correlation coefficient and the amount of scatter.

Solution:

(i) The sample covariance and sample correlation coefficient between temperature and quantity are given by

$$s_{xy} = 116.2619 \qquad \text{and} \qquad r_{xy} = 0.979.$$

(ii) Figure 2.13 shows the scatter plot for the soft drinks data. We can observe from the plot that the number of cans of soft drinks sold increases as the temperature increases, and that there seems to be a linear trend for this association.

MATLAB code

```
>> X = [70 75  80  90  93  98  72  75  75  80  90  95  98  91  98];
>> Y = [30 31  40  52  57  59  33  38  32  45  53  56  62  51  58];
>> cov(X,Y)
>> corrcoef(X,Y)
>> scatter(X,Y,'ko','MarkerFaceColor',[.49 1 .63]); grid on;
>> xlabel('Temperature','fontsize',14,'fontname','times');
>> ylabel('Number of cans','fontsize',14,'fontname','times');
```

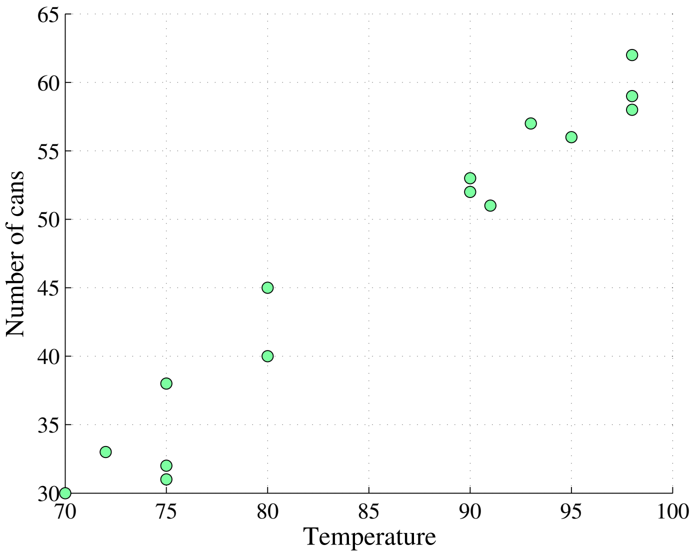


FIGURE 2.13: Scatter plot for the soft drinks data.

2.2 Problems

❶ An engineer wants to measure the bias in a pH meter. She uses the meter to measure the pH in 13 neutral substances and obtains the following data:

```
6.90  7.00  7.03  7.01  6.97  7.00  6.95  7.00  6.99  7.04  6.97  7.07  7.04
```

a) Calculate the sample range and mean.

b) Calculate the sample variance and standard deviation.

c) Check the assumption of normality for the pH data using normplot and boxplot. Identify any outliers.

❷ An entrepreneur faces many bureaucratic and legal hurdles when starting a new business. The following data shows the time, in days, to complete all of the procedures required to start a business in 25 developed countries:

```
23 4 29 67 44 47 24 40 23 116 44 33 27 60 46 61 11 23 62 31 44 77 14 65 42
```

It is also reported that the overall average time to start a business in all developed countries is 30 days.

a) Construct the dot plot.

b) Construct the stem-and-leaf diagram.

c) Calculate the sample range, mean, and median of the time-to-start data.

d) Calculate the first quartile, second quartile, and the interquartile range.

e) Calculate the sample variance and standard deviation.

f) Check the assumption of normality for the time-to-start data using normplot and boxplot. Identify any outliers.

2.3 REFERENCES

[1] D.C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, 6th Edition, 2009.

[2] I. Bass and B. Lawton, *Lean Six Sigma using SigmaXL and Minitab*, McGraw-Hill Professional, 1st Edition, 2009.

Chapter 3

Random Variables and Probability Distributions

Life is a school of probability.

Walter Bagehot

Probability theory is a branch of mathematics that deals with calculating the likelihood of a given event's occurrence, which is expressed as a number between zero and one. An event with a probability of zero can be considered an impossibility, whereas an event with a probability of one can be considered a certainty. The fundamental ingredient of probability theory is an experiment that can be repeated, at least hypothetically, under essentially identical conditions and that may lead to different outcomes on different trials. An experiment is the process by which one observation is obtained. An example of an experiment would be the sorting out of defective parts from a production line. A probability distribution assigns a probability to each of the possible outcomes of a random experiment. A random variable is a function that associates a unique numerical value with every outcome of an experiment. The value of the random variable will vary from trial to trial as the experiment is repeated.

3.1 Probability

3.1.1 Sample Space and Events

Definition 3.1

- *A **random experiment** is a procedure or an operation whose outcome is uncertain and cannot be predicted in advance.*
- *A **sample space**, denoted by $\mathcal{S}$, is the set of all possible outcomes of a random experiment.*
- *An **event** is a subset of the sample space $\mathcal{S}$. That is, an event is an outcome or a collection of outcomes of a random experiment.*

Example 3.1

- *A classic example that is used often to illustrate concepts of probability theory, is the experiment of flipping a coin. The outcomes of this experiment are head and tail. So, the sample space is $\mathcal{S} = \{H, T\}$, where H represents the event 'head' and T represents the event 'tail'.*
- *Tossing two coins and observing the up face on each is random experiment with sample space $\mathcal{S} = \{HH, HT, TH, TT\}$ of four events.*
- *Rolling a single die is a random experiment with sample space $\mathcal{S} = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. Rolling an even number (⚁, ⚃or ⚅) is an event, and rolling an odd number (⚀, ⚂or ⚄) is also an event.* ■

3.1.2 Axioms of Probability

Given a random experiment and a sample space $\mathcal{S}$, the objective of probability is to assign to each event A a number $P(A)$, called the probability of the event A., which will give a precise measure of the chance that A will occur. $P(A)$ is defined as the ratio of the number of ways event A can occur to the number of possible outcomes:

$$P(A) = \frac{\text{Number of ways event } A \text{ can occur}}{\text{Total number of possible outcomes (sample space)}}. \tag{3.1.1}$$

A probability must obey the following rules or axioms:

- **Axiom 1:** For any event A, $0 \leq P(A) \leq 1$
- **Axiom 2:** $P(\mathcal{S}) = 1$
- **Axiom 3:** If $A_1, A_2, \ldots$ is a collection of mutually exclusive events then

$$P(A_1 \cup A_2 \cup \ldots) = \sum_i P(A_i).$$

Example 3.2 • *What is the probability of each outcome when a loonie is tossed?*

- *A number from 1 to 9 is chosen at random. What is the probability of choosing an odd number?*
- *What is the probability of choosing a vowel from the alphabet?*

Solution:

- $P(\text{head}) = 1/2$ and $P(\text{tail}) = 1/2$
- $P(\text{odd}) = 5/9$
- $P(\text{vowel}) = 5/26.$

Example 3.3 *Determine the probability of the following events when rolling a single die:*

- *Probability of rolling a 4*
- *Probability of rolling an odd number*
- *Probability of rolling an even number*
- *Probability of rolling a 7*

Solution: Since the sample space is $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$, the total number of outcomes is equal to 6.

- There is only one number 4 on the die. Thus, $P(4) = 1/6$
- There are 3 odd numbers $(1, 3, 5)$. Thus, $P(\text{odd number}) = 3/6 = 1/2$
- There are 3 even numbers $(2, 4, 6)$. Thus, $P(\text{even number}) = 3/6 = 1/2$
- The die does not contain a number 7. Thus, $P(7) = 0$.

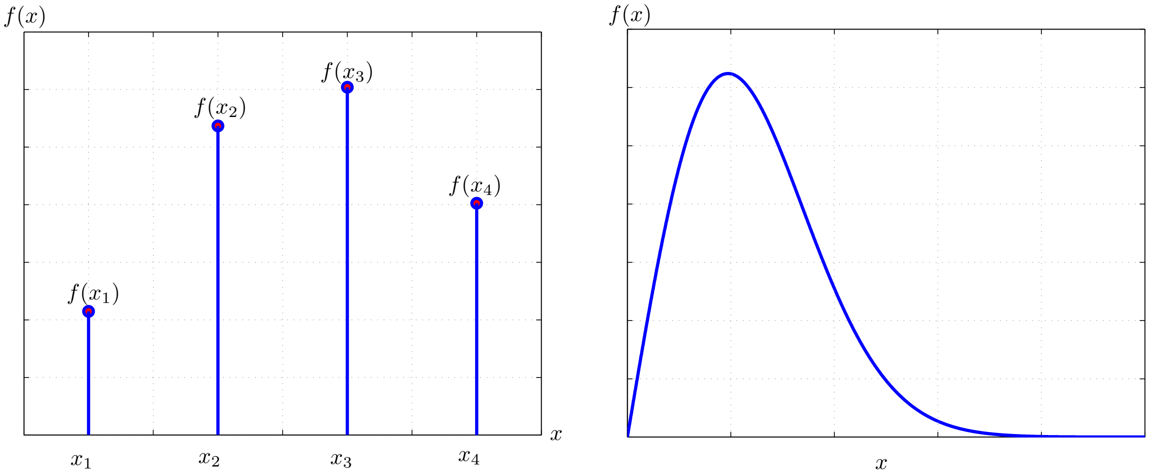
3.2 Random Variables

In many situations, we are interested in numbers associated with the outcomes of a random experiment. For example, testing cars from a production line, we are interested in variables such as average emissions, fuel consumption, or acceleration time.

Definition 3.2 *A **random variable** $X : \mathcal{S} \to \mathbb{R}$ is a function that maps every outcome in the sample space of a random experiment to a real number.*

A **discrete random variable** X is a random variable that has a finite number of possible values in a discrete set, whereas a **continuous random variable** X is a random variable that takes its values in a interval of numbers. Random variables are usually denoted by capital letters X, whereas the values of the variables are usually denoted by lower case letters x.

The probability distribution of a random variable X tells us what the possible values of X are and how probabilities are assigned to those values. Figure 3.1 displays an example of discrete and continuous distributions.



(a) Discrete distribution

(b) Continuous distribution

Figure 3.1: Probability distributions.

Definition 3.3 *The probability distribution or probability mass function (pmf) of a discrete random variable X is defined for every value x of X by $f(x) = P(X = x)$, and it lists the values and their probabilities:*

Value of X	x_1	x_2	x_3	$\ldots x_k$
$P(X = x_i)$	$f(x_1)$	$f(x_2)$	$f(x_3)$	$\ldots f(x_k)$

The conditions $f(x_i) \geq 0$ and $\sum_i f(x_i) = 1$ are required for any pmf.

- *The cumulative distribution function (cdf) $F(x)$ is defined by*

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i) \tag{3.2.1}$$

- *The expected value of X, denoted by $E(X)$ or μ, is*

$$E(X) = \mu = \sum_i x_i f(x_i) \tag{3.2.2}$$

- *The variance of X, denoted by $Var(X)$ or σ^2, is*

$$Var(X) = \sigma^2 = \sum_i (x_i - \mu)^2 f(x_i) = E(X^2) - [E(X)]^2 \tag{3.2.3}$$

For a discrete random variable X, we have:

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) = F(b) - F(a) && (3.2.4)\\ P(a \leq X \leq b) &= P(a < X \leq b) + P(X = a) = F(b) - F(a) + f(a) && (3.2.5)\\ P(a < X < b) &= P(a < X \leq b) - P(X = b) = F(b) - F(a) - f(b). && (3.2.6) \end{aligned}$$

Example 3.4 *Let a random variable X denote the number of defective parts produced per eight-hour shift by a machine. Experience shows that it produces between 0 and 4 (inclusive) with the following probabilities:*

$X = x$	*0*	*1*	*2*	*3*	*4*
$f(x) = P(X = x)$	*0.10*	*0.40*	*0.25*	*0.20*	*0.05*

(i) Calculate the mean and standard deviation of X.

(ii) Calculate the probability that X falls in the interval $(\mu - 2\sigma, \mu + 2\sigma)$.

Solution: The probability mass function is shown in Figure 3.2.

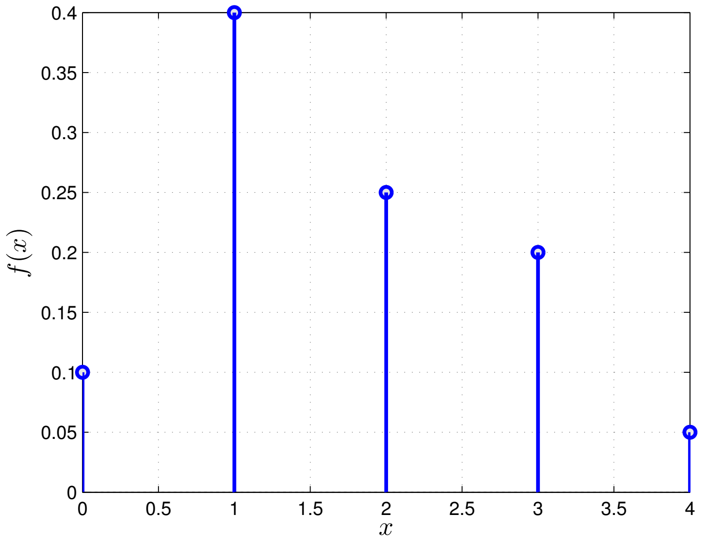


FIGURE 3.2: Probability mass function for the number of defective parts produced per eight-hour shift by the machine.

MATLAB code

```
>> x = 0:4;
>> fx = [0.1 0.4 0.25 0.2 0.05];
>> h=stem(x,fx,'LineWidth',2); % Visualize the pmf
>> set(get(h,'BaseLine'),'LineStyle',':')
>> set(h,'MarkerFaceColor',[.9 .9 .9]); grid
>> xlabel('x','fontsize',14,'fontname','times');
>> ylabel('f(x)','fontsize',14,'fontname','times');
```

(i) The mean and standard deviation of X are given by

$$\mu = \sum_i x_i f(x_i) = (0)(0.1) + (1)(0.4) + (2)(0.25) + (3)(0.2) + (4)(0.05) = 1.70$$

$$\sigma = \sqrt{\sum_i (x_i - \mu)^2 f(x_i)} = 1.05$$

(ii) Calculate the probability that X falls in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ is given by

$$\begin{aligned} P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= P(-0.4 \leq X \leq 3.8) \\ &= P(X=0) + P(X=1) + P(X=2) + P(X=3) \\ &= 0.95. \end{aligned}$$

In other words, the machine produces the number of defective parts within 2 standard deviations of the mean with probability 95%. ■

Definition 3.4 *The probability distribution or probability density function (pdf) of a continuous random variable X is a nonnegative function $f(x)$ such that*

$$P(a \leq X \leq b) = \int_a^b f(x)dx \qquad \textit{and} \qquad \int_{-\infty}^{\infty} f(x)dx = 1 \tag{3.2.7}$$

- *The cumulative distribution function (cdf) $F(x)$ is defined by*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy \tag{3.2.8}$$

- *The expected value of X, denoted by $E(X)$ or μ, is*

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)dx \tag{3.2.9}$$

- *The variance of X, denoted by $Var(X)$ or σ^2, is*

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = E(X^2) - [E(X)]^2 \tag{3.2.10}$$

For a continuous random variable X, probabilities correspond to areas under the curve $f(x)$. Also, for any single value a, we have $P(X = a) = 0$. In addition,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b). \tag{3.2.11}$$

Let a and b be two constants. The expected value and variance of a random variable X satisfy the following properties:

- $E(aX + b) = aE(X) + b$
- $Var(aX + b) = a^2 Var(X)$

Example 3.5 *Let X denote the width in mm of metal pipes from an automated production line. If X has the probability density function*

$$f(x) = \begin{cases} 10\, e^{-10(x-5.5)} & \textit{if } x \geq 5.5 \\ 0 & \textit{if } x < 5.5. \end{cases}$$

determine:

(i) $P(X < 5.7)$

(ii) $P(X > 6)$

(iii) $P(5.6 < X \leq 6)$

(iv) the cumulative distribution function $F(x)$

Solution:

(i)

$$\begin{aligned} P(X < 5.7) &= \int_{5.5}^{5.7} 10\, e^{-10(x-5.5)} dx \\ &= -e^{-10(x-5.5)}\Big|_{5.5}^{5.7} \\ &= 1 - e^{-2} = 0.8647 \end{aligned}$$

(ii)

$$\begin{aligned} P(X > 6) &= \int_{6}^{\infty} 10\, e^{-10(x-5.5)} dx \\ &= -e^{-10(x-5.5)}\Big|_{6}^{\infty} \\ &= e^{-5} = 0.0067 \end{aligned}$$

(iii)

$$\begin{aligned} P(5.6 < X \leq 6) &= \int_{5.6}^{6} 10\, e^{-10(x-5.5)} dx \\ &= -e^{-10(x-5.5)}\Big|_{5.6}^{6} \\ &= e^{-1} - e^{-5} = 0.3611 \end{aligned}$$

(iv) For $x < 5.5$, $F(x) = 0$. For $x \geq 5.5$, we have

$$\begin{aligned} F(x) &= \int_{5.5}^{x} 10\, e^{-10(t-5.5)} dt \\ &= -e^{-10(x-5.5)}\Big|_{5.5}^{x} \\ &= 1 - e^{-10(x-5.5)} \end{aligned}$$

Example 3.6 *Let X denote the time in milliseconds for a chemical reaction to complete. Assume that the cumulative distribution function of X is given by*

$$F(x) = \begin{cases} 1 - e^{-0.05x} & \textit{if } x \geq 0 \\ 0 & \textit{if } x < 0. \end{cases}$$

(i) What is the probability density function of X?

(ii) What is the probability that a reaction completes within 40 milliseconds?

Solution:

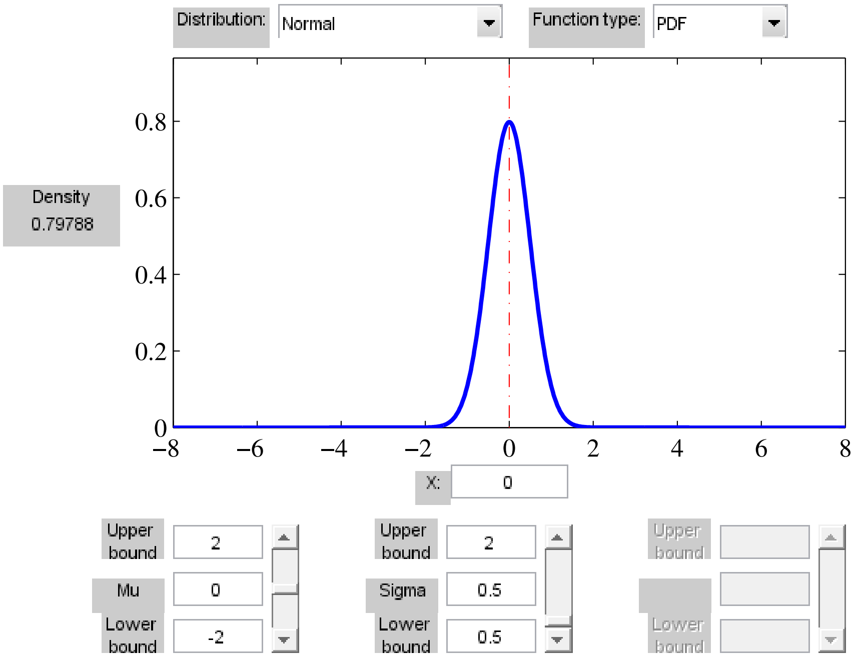
(i) The probability density function is given by

$$f(x) = F'(x) = \begin{cases} 0.05\, e^{-0.05x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

(ii) The probability that the reaction completes within 40 milliseconds is

$$P(X \leq 40) = F(40) = 1 - e^{-2} = 0.8647.$$

Using MATLAB, type: ≫ disttool to display interactive plots of the pdfs and cdfs of various continuous and discrete probability distributions, as shown in Figure 3.3.



Distribution: Normal

Function type: PDF

Density 0.79788

X: 0

Upper bound 2

Mu 0

Lower bound -2

Upper bound 2

Sigma 0.5

Lower bound 0.5

Upper bound

Lower bound

FIGURE 3.3: disttool: graphical interface for exploring the effects of changing parameters on the plot of a cdf or pdf.

3.3 DISCRETE PROBABILITY DISTRIBUTIONS

A distribution is said to be *discrete* if it is built on discrete random variables. All the possible outcomes when pulling a card from a stack are finite because we know in advance how many cards are in the stack and how many are being pulled. A random variable is said to be discrete when all the possible outcomes are countable.

3.3.1 BINOMIAL DISTRIBUTION

The binomial distribution assumes an experiment with n identical trials, each trial having only two possible outcomes considered as success or failure and each trial independent of the previous ones. The binomial distribution is probably the most commonly used discrete distribution.

Definition 3.5 *A random variable X is said to have a binomial distribution, denoted $X \sim bino(n, p)$, if its probability mass function is*

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad (3.3.1)$$

where

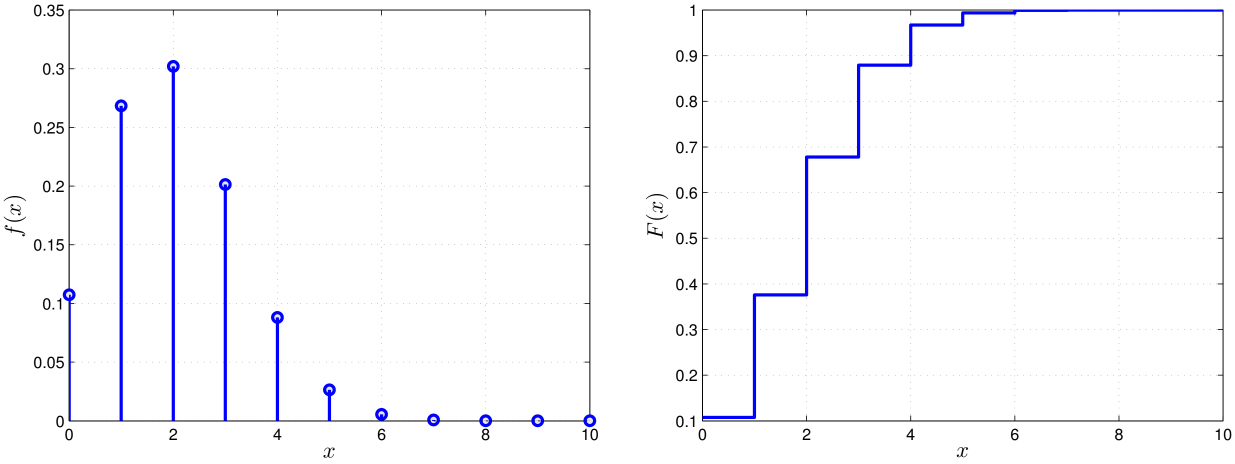
$$\begin{aligned} n &= \text{number of trials} \\ x &= \text{number of successes in } n \text{ trials} \\ p &= \text{probability of success on any one trial} \\ \binom{n}{x} &= \frac{n!}{x!(n-x)!} \end{aligned}$$

The mean and variance of $X \sim bino(n, p)$ are $E(X) = np$ and $Var(X) = np(1-p)$. Moreover, the cdf is given by

$$F(x) = P(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}. \qquad (3.3.2)$$

Figure 3.4 shows the pdf and cdf of the binomial distribution $X \sim bino(n, p)$ with parameters $n = 10$ and $p = 0.2$.

```
p = 0.2; % Probability of success for each trial
n = 10; % Number of trials
x = 0:n; % Outcomes
fx = pdf('bino',x,n,p); % pmf
stem(x,fx,'LineWidth',2) % Visualize the pmf
figure;
Fx = cdf('bino',x,n,p); % cdf
stairs(x,Fx,'LineWidth',2); grid % Visualize the cdf
```



(a) (b)

FIGURE 3.4: Binomial distribution $X \sim bino(n, p)$ with $n = 10$ and $p = 0.2$: (a) pdf, and (b) cdf.

Example 3.7 *A biased coin is tossed 6 times. The probability of heads on any toss is 0.3. Let X denote the number of heads that come up. Calculate:*

(i) $P(X = 2)$

(ii) $P(X = 3)$

(iii) $P(1 < X \leq 5)$

(iv) the mean and variance of X.

Solution: If we call heads a success then X follows a binomial distribution with parameters $n = 6$ and $p = 0.3$.

(i) $P(X = 2) = \binom{6}{2}(0.3)^2(0.7)^4 = 0.3241$

(ii) $P(X = 3) = \binom{6}{3}(0.3)^3(0.7)^3 = 0.1852$

(iii) $P(1 < X \leq 5) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 0.5791$

(iv) $E(X) = np = (6)(0.3) = 1.8$ and $Var(X) = np(1 - p) = (6)(0.3)(0.7) = 1.26$

These probabilities can be computed using MATLAB as follows:

MATLAB code

```
>> pdf('bino',2,6,.3)
ans =
    0.3241
>> pdf('bino',3,6,.3)
ans =
    0.1852
>> cdf('bino',5,6,.3)-cdf('bino',1,6,.3)
```

```
ans =
    0.5791
```

Example 3.8 *The probability that a randomly selected technician will finish his or her project successfully is 0.8. Let X be the number of technicians among a randomly selected group of 10 technicians who will finish their projects successfully.*

(i) Plot the probability mass function of X.

(ii) Calculate the probability that exactly one technician will finish his/her project successfully.

(iii) Calculate the probability that at least three technicians will finish their project successfully.

(iv) Calculate the probability that at most five technicians will finish their project successfully.

(v) Calculate the probability that between four and six (inclusive) technicians will finish their project successfully.

(vi) Calculate the mean and standard deviation of X.

Solution: The random variable X follows a binomial distribution with parameters $n = 10$ and $p = 0.8$.

(i) The graphical representation of the probability distribution is shown in Figure 3.5.

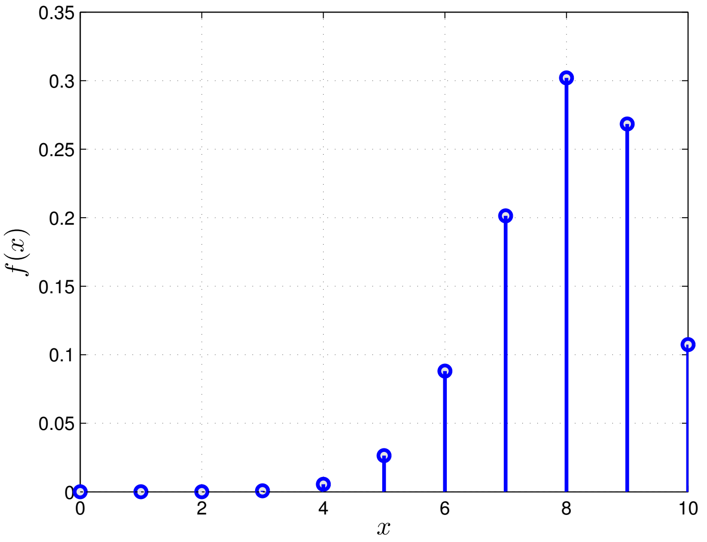


FIGURE 3.5: Probability mass function of $X \sim bino(n, p)$ with $n = 10$ and $p = 0.8$.

(ii) $P(X = 1) = \binom{10}{1}(0.8)^1(0.2)^9 \approx 0$

(iii) $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - (P(X = 0) + P(X = 1) + P(X = 2)) = 0.9999$

(iv) $P(X \leq 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 0.0328$

(v) $P(4 \leq X \leq 6) = P(X = 4) + P(X = 5) + P(X = 6) = 0.12$

(vi) $\mu = np = (10)(0.8) = 8$ and $\sigma = \sqrt{np(1-p)} = \sqrt{(10)(0.8)(0.2)} = 1.2649$

Example 3.9 *The probability that the Food and Drug Administration (FDA) will approve a new drug is 0.6. Suppose that five new drugs are submitted to FDA for its approval.*

(i) Plot the probability mass function of X.

(ii) Calculate the probability that exactly three drugs are approved.

(iii) Calculate the probability that at most three drugs are approved.

(iv) *Calculate the probability that at least three drugs are approved.*

(v) *Calculate the probability that between two and four (inclusive) drugs are approved.*

(vi) *Calculate the mean and standard deviation of X.*

Solution: Let X be the number of dugs among the five new drugs submitted to FDA will be approved. The random variable X follows a binomial distribution with parameters $n = 5$ and $p = 0.6$.

(i) The graphical representation of the probability distribution is shown in Figure 3.6.

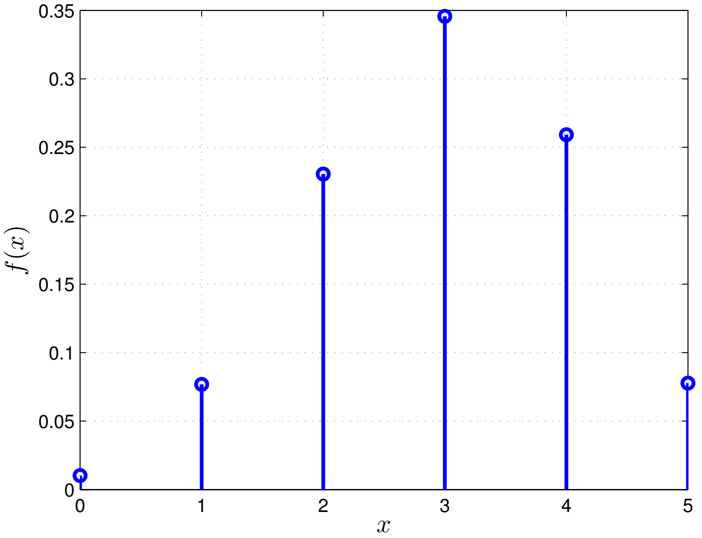


FIGURE 3.6: Probability mass function of $X \sim bino(n, p)$ with $n = 5$ and $p = 0.6$.

(ii) $P(X = 3) = \binom{5}{3}(0.6)^3(0.4)^2 = 0.3456$

(iii) $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0.6630$

(iv) $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - (P(X = 0) + P(X = 1) + P(X = 2)) = 0.6826$

(v) $P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) = 0.8352$

(vi) $\mu = np = (5)(0.6) = 3$ and $\sigma = \sqrt{np(1-p)} = \sqrt{(5)(0.6)(0.4)} = 1.0954$

Example 3.10 *Each item produced by a certain process is independently defective with probability 0.02.*

(i) *What is the probability that a batch of 60 such items contains 0 defective items?*

(ii) *What is the probability that a batch of 60 such items contains at most 1 defective item?*

Solution: Let X denote the number of defective items in the batch. X follows a binomial distribution with parameters $n = 60$ and $p = 0.02$

(i)

$$P(0 \text{ defectives}) = P(X = 0) = \binom{60}{0}(0.02)^0(1 - 0.02)^{60} = 0.2976$$

(ii) X follows a binomial distribution with parameters $n = 60$ and $p = 0.02$. Thus,

$$\begin{aligned} P(\text{at most 1 defective}) &= P(X \leq 1) \\ &= P(X = 0) + P(X = 1) \\ &= \binom{60}{0}(0.02)^0(0.98)^{60} + \binom{60}{1}(0.02)^1(0.98)^{59} \\ &= 0.6619. \end{aligned}$$

Example 3.11 *A machine produces soda bottles, and 95 percent of all bottles produced pass an audit. What is the probability of having only 2 bottles that pass audit in a randomly selected sample of 10 bottles?.*

Solution: Let X denote the number bottles that pass audit. X follows a binomial distribution with parameters $n = 10$ and $p = 0.95$. The probability of having only 2 bottles that pass audit in a randomly selected sample of 10 bottles is given by

$$P(X = 2) = \binom{10}{2}(0.95)^2(1 - 0.95)^8 \approx 0.$$

In other words, the probability of having only two good bottles out of 10 is zero.

Example 3.12 *Suppose each customer who visits a car dealership independently returns his or her questionnaire with probability of 0.78. What is the probability that out of 64 customers there are at most 36 returns?.*

Solution: Let X denote the number of returns. X follows a binomial distribution with parameters $n = 64$ and $p = 0.78$. The probability that out of 64 customers there are at most 36 returns is given by

$$P(X \leq 36) = \sum_{i=0}^{36} \binom{64}{i}(0.78)^i(0.22)^{64-i} \approx 0.00009.$$

3.3.2 Poisson Distribution

In many practical situations we are interested in measuring how many times a certain event occurs in a specific time interval or in a specific length or area, where the average of such an event occurring is known. For instance:

- the number of phone calls received at an exchange or call center in an hour;
- the number of customers arriving at a toll booth per day;
- the number of flaws on a length of cable;
- the number of cars passing using a stretch of road during a day.

The Poisson distribution plays a key role in modeling such problems. For instance, a Quality Control manager may want to know the probability of finding a defective part on a manufactured circuit board.

Definition 3.6 *A random variable X is said to have a Poisson distribution, denoted $X \sim poiss(\lambda)$, if its probability mass function is*

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{3.3.3}$$

where the parameter $\lambda > 0$ is the number of occurrences per unit time or space.
The mean and variance of $X \sim poiss(\lambda)$ are equal: $E(X) = Var(X) = \lambda$

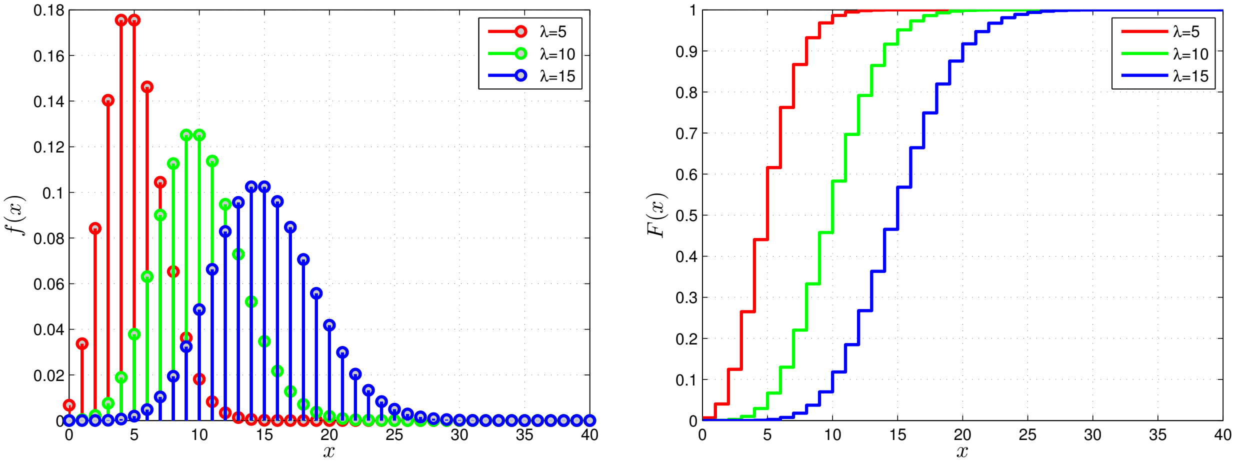
The expected value of a Poisson-distributed random variable is equal to the parameter λ and so is its variance. Figure 3.7 shows the pdf and cdf of the Poisson distribution $X \sim poiss(\lambda)$ for different values of λ.

```
lambda = 5; % parameter lambda
x = 0:40;
fx = pdf('poiss',x,lambda); % pmf
stem(x,fx,'LineWidth',2); % Visualize the pmf
figure;
Fx = cdf('poiss',x,lambda); % cdf
stairs(x,Fx,'LineWidth',2); grid % Visualize the cdf
```

Example 3.13 *The number of flaws in a fibre optic cable follows a Poisson distribution. The average number of flaws in 50m of cable is 1.2.*

(i) What is the probability of exactly three flaws in 150m of cable?

(ii) What is the probability of at least two flaws in 100m of cable?



(a) (b)

FIGURE 3.7: Poisson distribution $X \sim poiss(\lambda)$ for different values of λ: (a) pdf, and (b) cdf.

(iii) What is the probability of exactly one flaw in the first 50m of cable and exactly one flaw in the second 50m of cable?

Solution:

(i) The mean number of flaws in 150m of cable is 3.6. So the probability of exactly three flaws in 150m of cable is

$$P(X=3)=\frac{e^{-3.6}(3.6)^3}{3!}=0.212$$

(ii) The mean number of flaws in 100m of cable is $(2)(1.2)=2.4$. Let X be the number of flaws in 100m of cable.

$$\begin{aligned}P(X\geq 2) &= 1-P(X\leq 1)\\ &= 1-\big(P(X=0)+P(X=1)\big)\\ &= 1-\left(\frac{e^{-2.4}(2.4)^0}{0!}+\frac{e^{-2.4}(2.4)^1}{1!}\right)\\ &= 1-(0.0907+0.2177)=0.6916\end{aligned}$$

(iii) Now let X denote the number of flaws in a 50m section of cable. Then we know that

$$P(X=1)=\frac{e^{-1.2}(1.2)^1}{1!}=0.3614$$

As X follows a Poisson distribution, the occurrence of flaws in the first and second 50m of cable are independent. Thus the probability of exactly one flaw in the first 50m and exactly one flaw in the second 50m is: $(0.3614)(0.3614)=0.1306$.

Example 3.14 *The number of breakdowns of a machine is a random variable having the Poisson distribution with $\lambda = 2.2$ breakdown per month.*

(i) Plot the probability mass function of the distribution over a period of 12 months.

(ii) Calculate the probability that the machine will work during any given month with no breakdown.

(iii) Calculate the probability that the machine will work during any given month with one breakdown.

(iv) Calculate the probability that the machine will work during any given month with two breakdowns.

(v) Calculate the probability that the machine will work during any given month with at least two breakdowns.

(vi) *Calculate the probability that the machine will work with four breakdowns in two months.*

(vii) *Calculate the probability that the machine will work with five breakdowns in two and half months.*

Solution: Let X denote the number of breakdowns per month. From the given information, $X \sim poiss(\lambda)$ with $\lambda = 2.2$.

(i) The probability mass function of the distribution over a period of 12 months is shown in Figure 3.8.

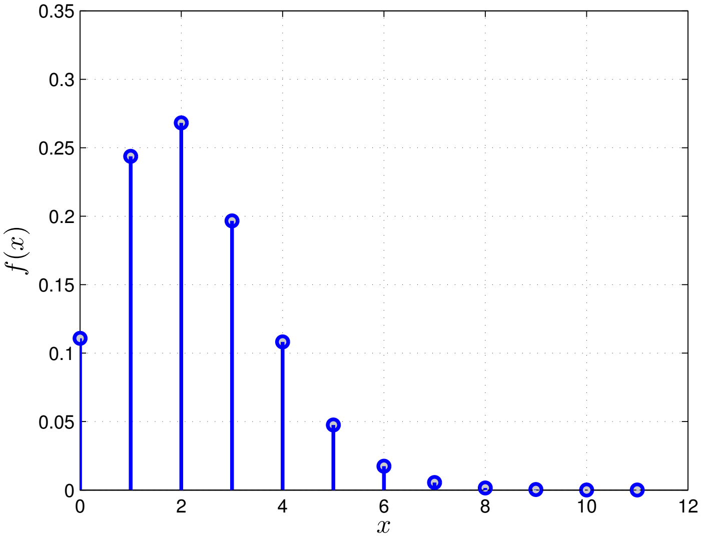


FIGURE 3.8: Probability mass function of the distribution over a period of 12 months.

(ii) The probability that the machine will work during any given month with no breakdown is

$$P(X = 0) = \frac{e^{-2.2}(2.2)^0}{0!} = 0.1108$$

(iii) The the probability that the machine will work during any given month with one breakdown is

$$P(X = 1) = \frac{e^{-2.2}(2.2)^1}{1!} = 0.2438$$

(iv) The the probability that the machine will work during any given month with two breakdowns is

$$P(X = 2) = \frac{e^{-2.2}(2.2)^2}{2!} = 0.2681$$

(v) The the probability that the machine will work during any given month with at least two breakdowns is

$$P(X \geq 2) = 1 - P(X < 2) = 1 - (P(X = 0) + P(X = 1)) = 0.6454$$

(vi) The mean number of breakdowns in two months is $\lambda = (2)(2.2) = 4.4$. The probability that the machine will work with four breakdowns in two months is

$$P(X = 4) = \frac{e^{-4.4}(4.4)^4}{4!} = 0.1917$$

(vii) The mean number of breakdowns in two and half months is $\lambda = (2.5)(2.2) = 5.5$. The probability that the machine will work with five breakdowns in two and half months is

$$P(X = 5) = \frac{e^{-5.5}(5.5)^5}{5!} = 0.1714.$$

Example 3.15 *The number of visitors to a web server per minute follows a Poisson distribution. If the average number of visitors per minute is 4, what is the probability that:*

(i) There are two or fewer visitors in one minute?

(ii) There are exactly two visitors in 30 seconds?

Solution: The average number of visitors in a minute is $\lambda = 4$.

(i) The probability that there are two or fewer visitors in one minute is

$$\begin{aligned} P(X \leq 2) &= P(X=0)+P(X=1)+P(X=2) \\ &= \frac{e^{-4}(4)^0}{0!}+\frac{e^{-4}(4)^1}{1!}\frac{e^{-4}(4)^2}{2!} \\ &= 0.0183+0.0733+0.1465 \\ &= 0.2381 \end{aligned}$$

(ii) If the average number of visitors in 1 minute is 4, the average in 30 seconds is $\lambda = 2$. Thus, the probability that there are exactly two visitors in 30 seconds is

$$P(X=2) = \frac{e^{-2}(2)^2}{2!} = 0.2707$$

Example 3.16 *A product failure has historically averaged 3.84 occurrences per day. What is the probability of 5 failures in a randomly selected day?*

Solution: The average failure occurrences per day is $\lambda = 3.84$. The probability of 5 failures in a randomly selected day is given by

$$P(X=5) = \frac{e^{-3.84}(3.84)^5}{5!} = 0.1495.$$

MATLAB code

```
>> lambda = 3.84; x = 5;
>> P = pdf('poiss',x,lambda)
P =
    0.1495
```

Example 3.17 *It is believed that the number of bookings taken per hour at an online travel agency follows a Poisson distribution. Past records indicate that the hourly number of bookings has a mean of 15 and a standard deviation of 2.5. Comment on the suitability of the Poisson distribution for this example?*

Solution: If the number of hourly bookings at this travel agent did follow a Poisson distribution, we would expect that the mean and variance should be equal. However, in this case

$$E(X) = 15, \quad Var(X) = (2.5)^2 = 6.25$$

This suggests that the Poisson distribution is not appropriate for this case.

3.4 Continuous Probability Distributions

Most experiments in business operations have sample spaces that do not contain a finite, countable number of simple events. A distribution is said to be *continuous* when it is built on continuous random variables, which are variables that can assume the infinitely many values corresponding to points on a line interval. An example of a random variable would be the time it takes a production line to produce one item. In contrast to discrete variables, which have values that are countable, the continuous variables values are measurements.

3.4.1 Uniform Distribution

The simplest continuous probability distribution is called *uniform distribution*, as shown in Figure 3.9. It provides a model for continuous random variables that are evenly (or randomly) distributed over a certain interval.

Definition 3.7 *A random variable X is said to have a uniform distribution, denoted $X \sim unif(a,b)$, on the interval $[a,b]$ if its probability density function is*

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise.} \end{cases} \tag{3.4.1}$$

The mean and variance of $X \sim unif(a,b)$ are $E(X) = \dfrac{a+b}{2}$ and $Var(X) = \dfrac{(b-a)^2}{12}$. Moreover, the cdf is given by

$$F(x) = P(X \le x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \ge b \end{cases} \tag{3.4.2}$$

All outcomes over the uniform distribution's entire range are equally likely. To throw a die is an example of a uniform distribution, where any number of $1, 2, \dots, 6$ has an equal probability to show up. Figure 3.10 displays the pdf and cdf of a uniform distribution with parameters $a = 0$ and $b = 1$.

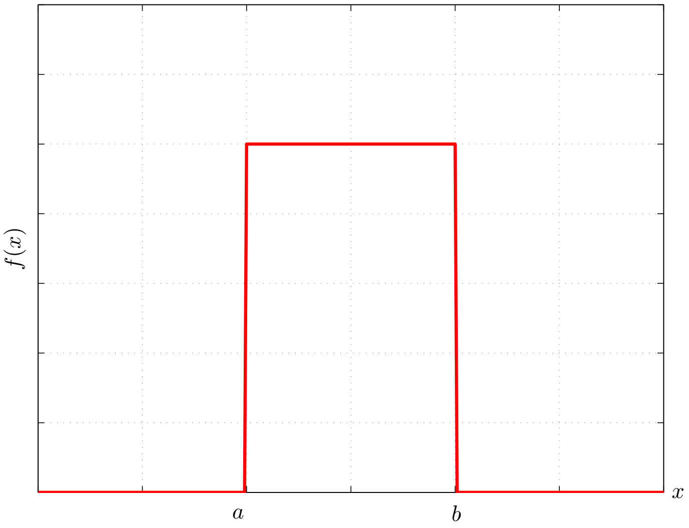


FIGURE 3.9: Illustration of the uniform pdf on the interval $[a,b]$.

```
a = 0; % parameter a
b = 1; % parameter b
x =-3:.01:3;
fx = pdf('unif',x,a,b); % pdf
plot(x,fx,'LineWidth',2); % Visualize the pdf
figure;
Fx = cdf('unif',x,a,b); % cdf
plot(x,Fx,'LineWidth',2); grid % Visualize the cdf
```

Example 3.18 *Suppose a delay in starting production due to an unexpected mechanical failure is anywhere from 0 to 30 minutes.*

(i) Calculate the probability that production will be delayed by less than 10 minutes.

(ii) Calculate the probability that production will be delayed by more than 10 minutes.

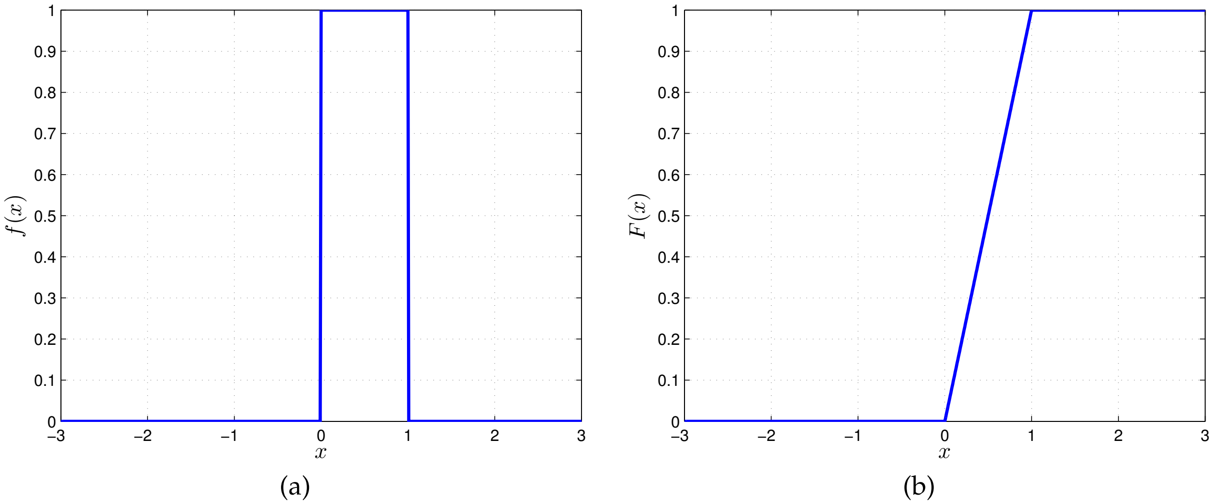


FIGURE 3.10: Uniform distribution $X \sim unif(a, b)$ with $a = 0$ and $b = 1$: (a) pdf; (b) cdf.

(iii) Calculate the probability that production will be delayed between 12 and 22 minutes.

(iv) Calculate the mean and standard deviation of the distribution.

Solution: Let X denote the the time by which production will be delayed. From the given information, we can see that $X \sim unif(a, b)$ with $a = 0$ and $b = 30$. Using the cdf of the uniform distribution, we have:

(i) $P(X \leq 10) = F(10) = \frac{10}{30} = \frac{1}{3}$

(ii) $P(X \geq 10) = 1 - F(10) = 1 - \frac{10}{30} = 1 - \frac{1}{3} = \frac{2}{3}$

(iii) $P(12 \leq X \leq 22) = F(22) - F(12) = \frac{22}{30} - \frac{12}{30} = \frac{1}{3}$

(iv) The mean and standard deviation of X are given by

$$\mu = \frac{a+b}{2} = \frac{0+30}{2} = 15$$

and

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{30-0}{\sqrt{12}} = 8.6603.$$

The pdf and cdf of X are depicted in Figure 3.11.

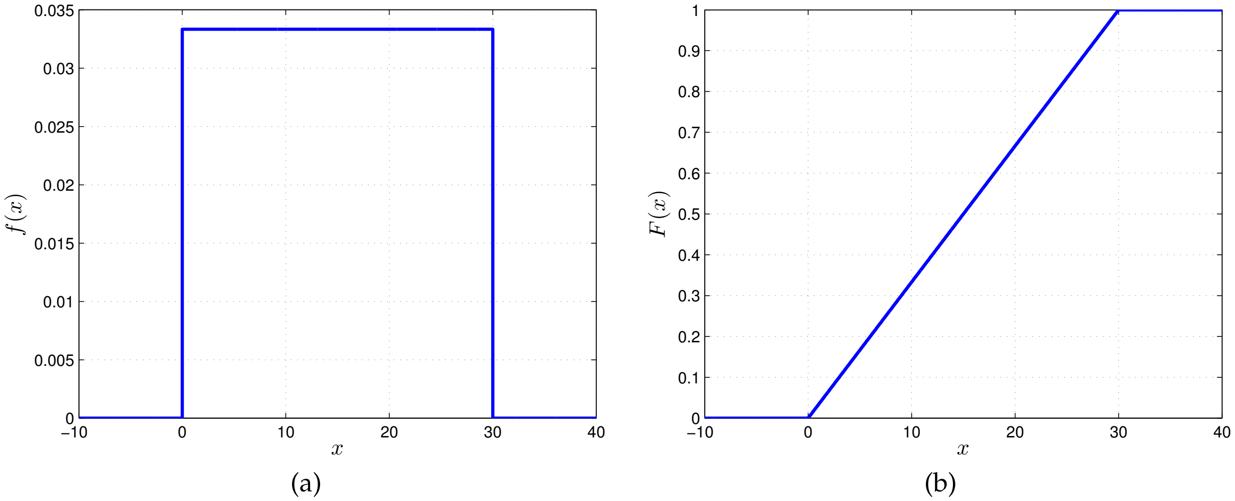
Example 3.19 *A quality control inspector for Tamuda Company, which manufactures aluminum water pipes, believes that the product has varying lengths. Suppose the pipes turned out by one of the production lines of Tamuda Company can be modeled by a uniform distribution over the interval 29.50 to 30.05 feet. Calculate the mean and standard deviation of the length of the aluminum water pipe.*

Solution: Let X denote the length of the aluminum water pipe. From the given information, we can see that the random variable X is uniformly distributed over the interval $(29.50, 30.05)$, i.e. $X \sim unif(29.50, 30.05)$. The mean and standard deviation of X are given by

$$\mu_X = E(X) = \frac{30.05 + 29.50}{2} = 29.775 \text{ feet}$$

and

$$\sigma_X = \sqrt{Var(X)} = \frac{30.05 - 29.50}{\sqrt{12}} = 0.1588 \text{ feet.}$$



(a)

(b)

FIGURE 3.11: $X \sim unif(a,b)$ with parameters $a = 0$ and $b = 30$: (a) pdf; (b) cdf.

3.4.2 NORMAL DISTRIBUTION

The normal distribution is probably the most important distribution in the study of quality control. It is often used to describe the behavior of a process and make estimates and inferences concerning the process.

Definition 3.8 *A random variable X is normally distributed with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if its probability density function is*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty \tag{3.4.3}$$

A normal distribution is symmetric and has a bell-shaped density curve with a single peak, as shown in Figure 3.12. The mean μ is where the peak of the density occurs, and the standard deviation σ indicates the spread or girth of the bell curve. Many common attributes and physical characteristics such as test scores, heights, weights, etc., tend to follow a normal distribution. Also, errors in measurement or production processes can often be approximated by a normal distribution. The graph of the normal distribution depends on two factors: the mean μ and the standard deviation σ. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curves. This can be visualized using the MATLAB Probability Distribution Function Tool (`>> disttool`) shown in Figure 3.3.

In Fig. 3.13(left), the shaded area under the curve is equal to the probability $P(a \leq X \leq b)$.

If $X \sim N(\mu, \sigma^2)$, then the random variable $Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$. The distribution of the random variable Z is called the *standard normal distribution*, and its cdf is denoted by Φ. The value $\Phi(z) = P(Z \leq z)$ is the shaded area under the standard norm curve as shown in Figure 3.13(right). This value is tabulated and is used to calculate probabilities for any normal random variable. Since the normal pdf is symmetric about zero, it follows that $\Phi(-z) = 1 - \Phi(z)$. Figure 3.14 shows the pdf and cdf of the standard normal distribution.

```
mu = 0; % mean
sigma = 1; % standard deviation
z =-3:.1:3;
fz = pdf('norm',z,mu,sigma); % pmf
plot(z,fz,'LineWidth',2); % Visualize the pdf
figure;
Fz = cdf('norm',z,mu,sigma); % cdf
plot(z,Fz,'LineWidth',2); grid % Visualize the cdf
```

The characteristics of the normal probability distribution are shown in Figure 3.15, where

- 68.26% of values of a normal random variable are within +/- 1 standard deviation of its mean.

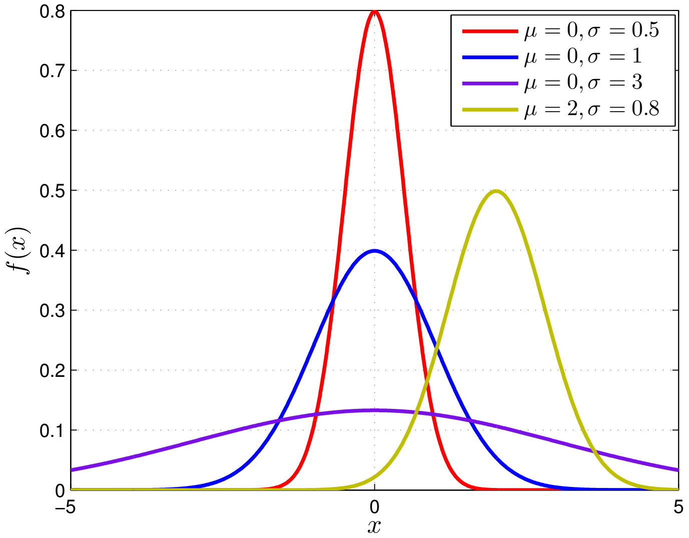


FIGURE 3.12: Probability density functions of the normal distribution. The blue curve is the standard normal distribution.

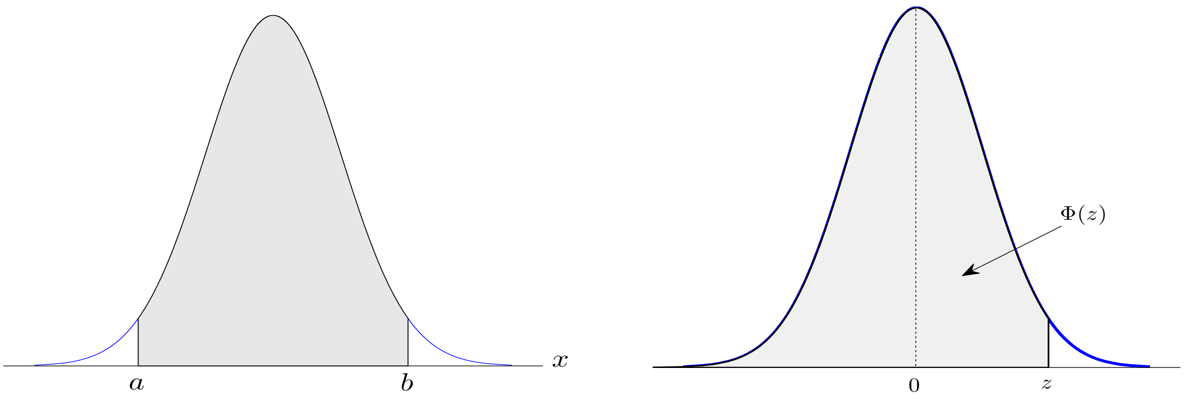


FIGURE 3.13: Left: Probability $P(a \leq X \leq b)$, where $X \sim N(\mu, \sigma^2)$. Right: Illustration of $\Phi(z)$.

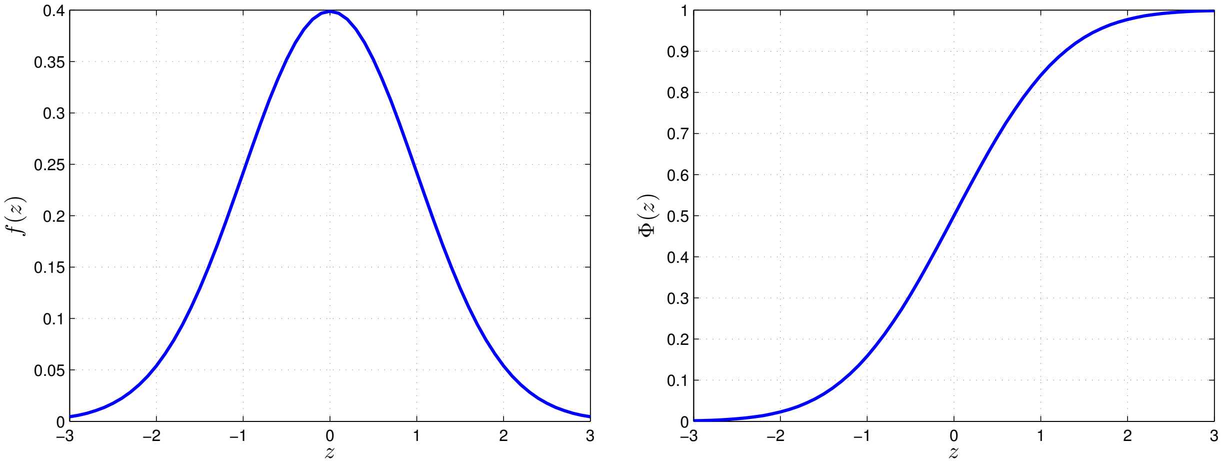
- 95.44% of values of a normal random variable are within +/- 2 standard deviations of its mean.

- 99.72% of values of a normal random variable are within +/- 3 standard deviations of its mean.

These percentage values can be calculated using the standard normal distribution table, or simply using the following MATLAB commands:

MATLAB code

```
>> cdf('normal',1,0,1)-cdf('normal',-1,0,1)
ans =
    0.6827
>> cdf('normal',2,0,1)-cdf('normal',-2,0,1)
ans =
    0.9545
>> cdf('normal',3,0,1)-cdf('normal',-3,0,1)
ans =
    0.9973
```



(a) (b)

FIGURE 3.14: Standard normal distribution $Z \sim N(0,1)$: (a) pdf; (b) cdf.

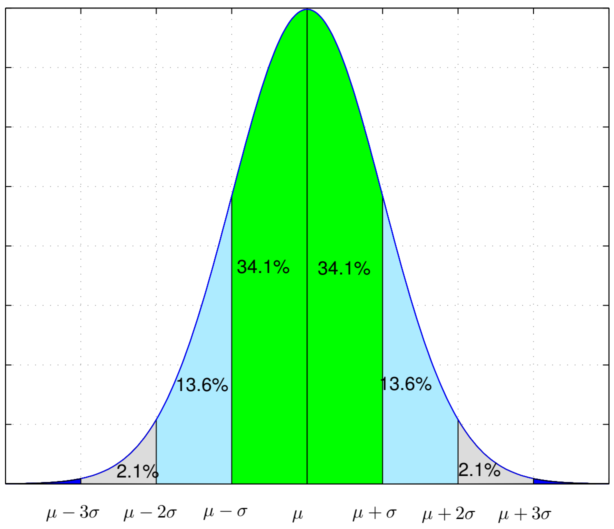


FIGURE 3.15: Characteristics of the normal probability distribution.

For $\alpha \in (0,1)$, denote by $z_{\alpha/2}$ the upper $100\alpha/2$ percentage point of the standard normal distribution as shown in Figure 3.16. That is,

$$P(Z > z_{\alpha/2}) = 1 - P(Z \leq z_{\alpha/2}) = 1 - \Phi(z_{\alpha/2}) = \alpha/2$$

Thus,

$$\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2} \quad \Rightarrow \quad z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where Φ^{-1} denotes the inverse cdf of Φ.
For $\alpha = 0.05$, the value of $z_{\alpha/2}$ can be computed using MATLAB as follows:

MATLAB code

```
>> alpha = 0.05;
>> icdf('normal',1-alpha/2,0,1)
ans =
    1.96
```

Similarly, the lower $100\alpha/2$ percentage point of the standard normal distribution is such that

$$\alpha/2 = P(Z < -z_{\alpha/2}) = \Phi(-z_{\alpha/2}).$$

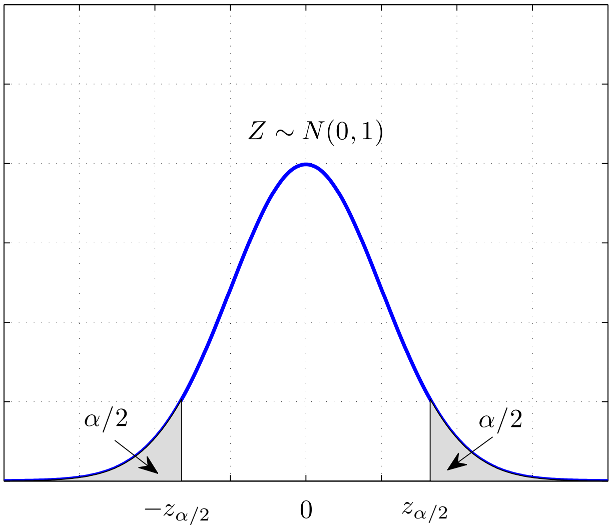


FIGURE 3.16: Upper and lower $100\alpha/2$ percentage points of the standard normal distribution.

Six Sigma Methodology and Normal Distribution:

Six Sigma is a business management strategy that seeks to improve the quality of process outputs by identifying and removing the causes of defects and minimizing variability in manufacturing and business processes. It is based on the statistical concept of 6σ, measuring a process at only **3.4 defects per million opportunities** (DPMO), where the standard deviation σ is the measure of variation in the process. DPMO is a measure of process performance and it is defined as

$$DPMO = \frac{1{,}000{,}000 \times \text{number of defects}}{\text{number of units} \times \text{number of opportunities per unit}} \tag{3.4.4}$$

where a defect is defined as a measurable quality characteristic of the process or its output that is not conforming to specifications or customer requirements.
The term "six sigma process" comes from the notion that if one has six standard deviations between the process mean and the nearest specification limit, as shown in Figure, practically no items will fail to meet specifications. This is based on the calculation method employed in process capability studies, which will be covered in Chapter 5.

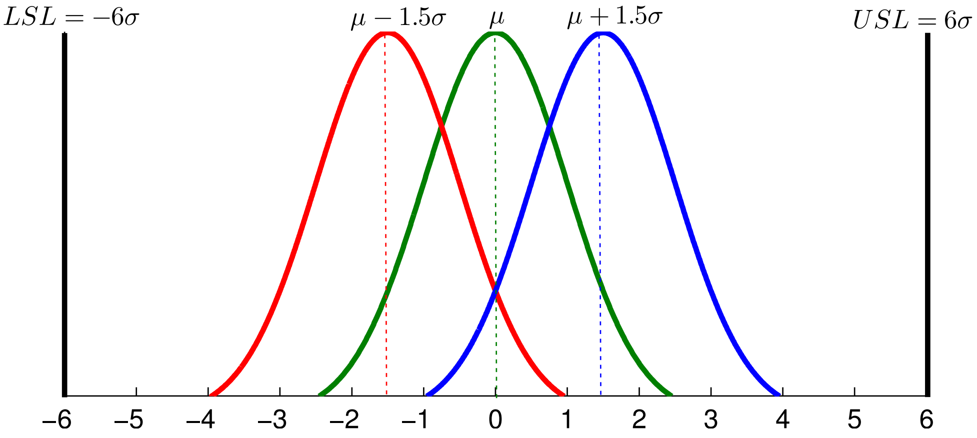


FIGURE 3.17: Statistical assumptions of the Six Sigma model.

Given the process standard deviation σ, we may calculate the DPMO as follows:

$$DPMO = 1000000\left[1 - \Phi\left(\frac{k\sigma - 1.5\sigma}{\sigma}\right)\right] = 1000000[1 - \Phi(k - 1.5)] \tag{3.4.5}$$

where the 1.5σ shift is subjective but some experts use this as conversion from long to short term performance estimates, and $k = 1, 2, \ldots, 6$. For example, $k = 6$ corresponds to 6σ, which yields $DPMO = 1000000[1 - \Phi(6 - 1.5)] \approx 3.4$. So the 3.4 DPMO of a Six Sigma process in fact corresponds to 4.5σ, namely 6σ minus the 1.5σ shift introduced to account for long-term variation

Example 3.20 *Suppose a quality characteristic of a product is normally distributed with mean $\mu = 18$ and standard deviation $\sigma = 1.5$. The specification limits given by the customer are $(15, 21)$. Determine what percentage of the product meets the specifications set by the customer.*

Solution: Let the random variable X denote the quality characteristic of interest. Then, X is normally distributed with mean $\mu = 18$ and standard deviation $\sigma = 1.5$. We are interested in finding the probability

$$\begin{aligned} P(15 \leq X \leq 21) &= P\left(\frac{15-\mu}{\sigma} \leq Z \leq \frac{21-\mu}{\sigma}\right) \\ &= P\left(\frac{15-18}{1.5} \leq Z \leq \frac{21-18}{1.5}\right) \\ &= P(-2 \leq Z \leq 2) \\ &= \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0.9545 \end{aligned}$$

That is, the percentage of product that will meet the specifications set by the customer is 95.44%.

Example 3.21 *The weekly profits of a large group of stores are normally distributed with a mean of $\mu = 1200$ and a standard deviation of $\sigma = 200$. What is the percentage of the stores that make \$1500 or more a week?*

Solution: The percentage of the stores that make \$1500 or more a week is given by the probability

$$P(X \geq 1500) = P\left(Z \geq \frac{1500-\mu}{\sigma}\right) = P\left(Z \geq \frac{1500-1200}{200}\right) = P(Z \geq 1.5) = 1 - \Phi(1.5) = 0.0668$$

That is, 6.68% of the stores make more than \$1500 week.

Example 3.22 *The mean number of defective parts that come from a production line is $\mu = 10.5$ with a standard deviation of $\sigma = 2.5$. What is the probability that the number of defective parts for a randomly selected sample will be less than 15?*

Solution:

$$P(X < 15) = P\left(Z < \frac{15-\mu}{\sigma}\right) = P\left(Z < \frac{15-10.5}{2.5}\right) = P(Z < 1.8) = \Phi(1.8) = 0.9641.$$

Example 3.23 *The actual volume of soup in 500ml jars follows a normal distribution with mean 500ml and variance 16ml. If X denotes the actual volume of soup in a jar, calculate*

(i) $P(X > 496)$

(ii) $P(X < 498)$

(iii) $P(492 < X < 506)$

Solution: From the given information, $X \sim N(\mu, \sigma^2)$ with $\mu = 500$ and $\sigma = 4$.

(i)

$$P(X > 496) = P\left(Z > \frac{496-500}{4}\right) = P(Z > -1) = 1 - \Phi(-1) = \Phi(1) = 0.8413$$

(ii)

$$P(X < 498) = P\left(Z < \frac{498-500}{4}\right) = P(Z < -0.5) = \Phi(-0.5) = 1 - \Phi(0.5) = 0.3085$$

(iii)

$$P(492 < X < 506) = P\left(\frac{492-500}{4} < Z < \frac{506-500}{4}\right) = \Phi(1.5) - \Phi(-2) = 0.9104$$

Example 3.24 *In the previous example, suppose that the mean volume of soup in a jar is unknown but that the standard deviation is 4. If only 3% of jars are to contain less than 492ml what should the mean volume of soup in a jar be?*

Solution: We want the value of μ for which

$$P\left(Z < \frac{492 - \mu}{4}\right) = 0.03$$

From the standard normal distribution table, we have $P(Z < -1.88) = 1 - \Phi(1.88) = 0.03$. Thus,

$$\frac{492 - \mu}{4} = -1.88 \quad \Rightarrow \quad \mu = 492 + (4)(1.88) = 499.52.$$

3.4.3 Normal Approximation to the Binomial Distribution

The normal distribution can be used to approximate binomial probabilities when there is a very large number of trials and when both np and $n(1-p)$ are large. A rule of thumb is to use this approximation when both $np \geq 5$ and $n(1-p) \geq 5$. If both are greater than 5, then the approximation should be good.

Definition 3.9 *Let $X \sim bino(n, p)$. If n is large relative to p, then the random variable*

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \tag{3.4.6}$$

is approximately a standard normal random variable.

Figure 3.18 shows normal approximations to the binomial distribution $bino(10, 0.5)$ and $bino(30, 0.5)$.

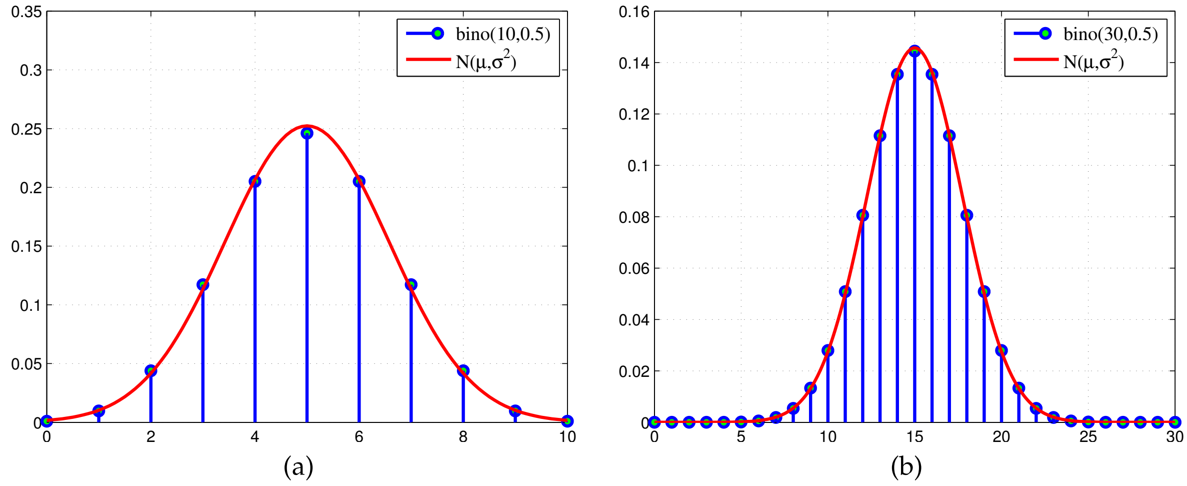


FIGURE 3.18: Normal approximation to the binomial distribution: (a) $bino(10, 0.5)$; (b) $bino(30, 0.5)$.

Continuity Correction for the Normal Approximation to the Binomial Distribution

To improve the accuracy of the approximation, we usually use a correction factor to take into account that the binomial random variable is discrete while the normal is continuous. The basic idea is to treat the discrete value x as the continuous interval from $x - 0.5$ to $x + 0.5$. Let $X \sim bino(n, p)$ and $Z \sim N(0, 1)$. Then,

$$P(X \leq x) \approx P\left(Z < \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) \tag{3.4.7}$$

$$P(X \geq x) \approx P\left(Z > \frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right) = 1 - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right) \tag{3.4.8}$$

$$P(x_1 \leq X \leq x_2) \approx \Phi\left(\frac{x_2 + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x_1 - 0.5 - np}{\sqrt{np(1-p)}}\right) \tag{3.4.9}$$

Example 3.25 *12% of the memory cards made at a certain factory are defective. If a sample of 150 cards is selected randomly, use the normal approximation to the binomial distribution to calculate the probability that the sample contains:*

(i) at most 20 defective cards

(ii) between 15 and 23 defective cards

(iii) exactly 17 defective cards

(iv) at least 19 defective cards

Solution: From the given information, we have $n = 150$ and $p = 0.12$. Both $np = 18 \geq 5$ and $n(1-p) = 132 \geq 5$ are satisfied, so we can use the normal approximation to the binomial. Using the continuity correction yields

(i)

$$P(X \leq 20) \approx \Phi\left(\frac{20 + 0.5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{20 + 0.5 - 18}{\sqrt{18(1-0.12)}}\right) = \Phi(0.63) = 0.7357$$

(ii)

$$P(15 \leq X \leq 23) \approx \Phi\left(\frac{23 + 0.5 - 18}{\sqrt{18(1-0.12)}}\right) - \Phi\left(\frac{15 - 0.5 - 18}{\sqrt{18(1-0.12)}}\right) = \Phi(1.38) - \Phi(-0.88) = 0.7268$$

(iii)

$$P(17 \leq X \leq 17) \approx \Phi\left(\frac{17 + 0.5 - 18}{\sqrt{18(1-0.12)}}\right) - \Phi\left(\frac{17 - 0.5 - 18}{\sqrt{18(1-0.12)}}\right) = \Phi(-0.13) - \Phi(-0.38) = 0.0963$$

(iv)

$$P(X \geq 19) \approx 1 - \Phi\left(\frac{19 - 0.5 - 18}{\sqrt{18(1-0.12)}}\right) = 1 - \Phi(0.55) = 0.45.$$

3.4.4 Exponential Distribution

The exponential distribution is commonly used to model waiting times between occurrences of rare events, lifetimes of electrical or mechanical devices. It is primarily used in reliability applications. The exponential distribution closely resembles the Poisson distribution. The Poisson distribution is built on discrete random variables and describes random occurrences over some intervals, whereas the exponential distribution is continuous and describes the time between random occurrences. Examples of an exponential distribution are the time between machine breakdowns and the waiting time in a line at a supermarket.

Definition 3.10 *A random variable X is said to have an exponential distribution, denoted $X \sim exp(\lambda)$, with positive parameter $\lambda > 0$ if its probability density function is*

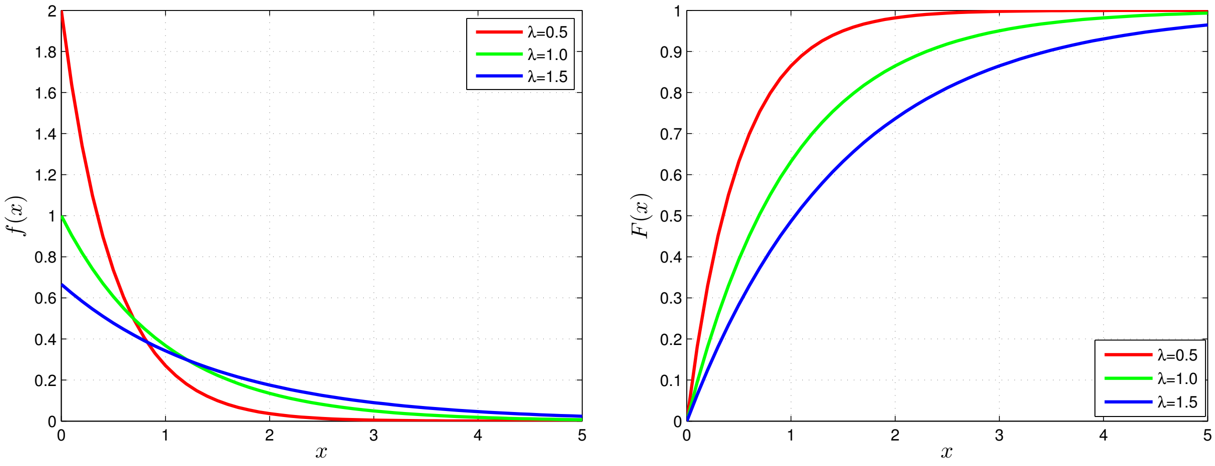
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.4.10}$$

The mean and variance of $X \sim exp(\lambda)$ are $E(X) = \dfrac{1}{\lambda}$ and $Var(X) = \dfrac{1}{\lambda^2}$. Moreover, the cdf is given by

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}. \tag{3.4.11}$$

The parameter λ of the exponential distribution is often called the *rate parameter*. The pdf and cdf of the exponential distribution $X \sim exp(\lambda)$ for various values of the rate parameter λ are shown in Figure 3.19.

```
lambda = 0.5; % parameter
x = 0:.1:5;
fx = pdf('exp',x,lambda); % pdf
plot(x,fx,'LineWidth',2); % Visualize the pdf
figure;
Fx = cdf('exp',x,lambda); % cdf
plot(x,Fx,'LineWidth',2); grid % Visualize the cdf
```



(a)

(b)

FIGURE 3.19: Exponential distribution $X \sim exp(\lambda)$ for various values of λ: (a) pdf; (b) cdf.

Example 3.26 *Jobs are sent to a printer at an average of 3 jobs per hour.*

(i) What is the expected time between jobs?

(ii) What is the probability that the next job is sent within 5 minutes?

Solution: Job arrivals represent rare events, thus the time X between them is exponentially distributed with rate 3 jobs/hour i.e. $\lambda = 3$.

(i) $E(X) = 1/\lambda = 1/3$ hours or 20 minutes

(ii) Using the same units (hours), we have 5 min=1/12 hours. Thus, the probability that the next job is sent within 5 minutes is

$$P(X < 1/12) = F(1/12) = 1 - e^{-(3)(1/12)} = 0.2212.$$

Example 3.27 *Suppose that the time in months between line stoppages on a production line follows an exponential distribution with $\lambda = 1/2$.*

(i) What is the probability that the time until the line stops again will be more than 15 months?

(ii) What is the probability that the time until the line stops again will be less than 20 months?

(iii) What is the probability that the time until the line stops again will be between 10 and 15 months?

(iv) Calculate the mean μ and standard deviation σ. Then, calculate the probability that the time until the line stops will be between $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$.

Solution: The time X between them line stoppages is exponentially distributed with $\lambda = 1/2$.

(i) The probability that the time until the line stops again will be more than 15 months is

$$P(X > 15) = e^{-15\lambda} = e^{-(15)(1/2)} = e^{-7.5} \approx 0$$

(ii) The probability that the time until the line stops again will be less than 20 months is

$$P(X < 20) = 1 - e^{-(20)(1/2)} \approx 1$$

(iii) The probability that the time until the line stops again will be between 10 and 15 months is

$$P(10 \leq X \leq 15) = F(15) - F(10) = e^{-(10)(1/2)} - e^{-(15)(1/2)} \approx 0.0062$$

(iv) The mean and the standard deviation are given by $\mu = \sigma = 1/\lambda = 2$. Thus, the probability that the time until the line stops will be between $(\mu - 3\sigma)$ and $(\mu + 3\sigma)$ is

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = P(-4 \leq X \leq 8) = P(0 \leq X \leq 8) = F(8) = 1 - e^{-(8)(1/2)} \approx 1.$$

3.4.5 Chi-Squared Distribution

The chi-squared distribution with n degrees of freedom is the distribution of a sum of the squares of n independent standard normal random variables. It is typically used in hypothesis testing and in determining confidence intervals.

Definition 3.11 *Suppose that $Z_1, Z_2, \ldots, Z_n \sim N(0,1)$. Then, the random variable $X = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ is said to have a chi-squared distribution, denoted $X \sim \chi^2(n)$, with n degrees of freedom. The density function of X is*

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} e^{-x/2} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4.12}$$

where the gamma function is defined by the integral $\Gamma(v) = \int_0^\infty e^{-t} t^{v-1} dt$.
The mean and variance of $X \sim \chi^2(n)$ are $E(X) = n$ and $Var(X) = 2n$

The pdf and cdf of the χ^2-distribution for various values of degrees of freedom n are shown in Figure 3.20. The distribution is asymmetric and as the degrees of freedom increase, the χ^2-curve approaches a normal distribution. The mean of the χ^2-distribution is the degrees of freedom and the standard deviation is twice the degrees of freedom. This implies that the χ^2-distribution distribution is more spread out, with a peak farther to the right, for larger than for smaller degrees of freedom.

```
n = 5; % degree of freedom
x = 0:.1:30;
plot(x,pdf('chi2',x,n),'LineWidth',2); % Visualize the pdf
figure;
plot(x,cdf('chi2',x,n),'LineWidth',2); % Visualize the cdf
```

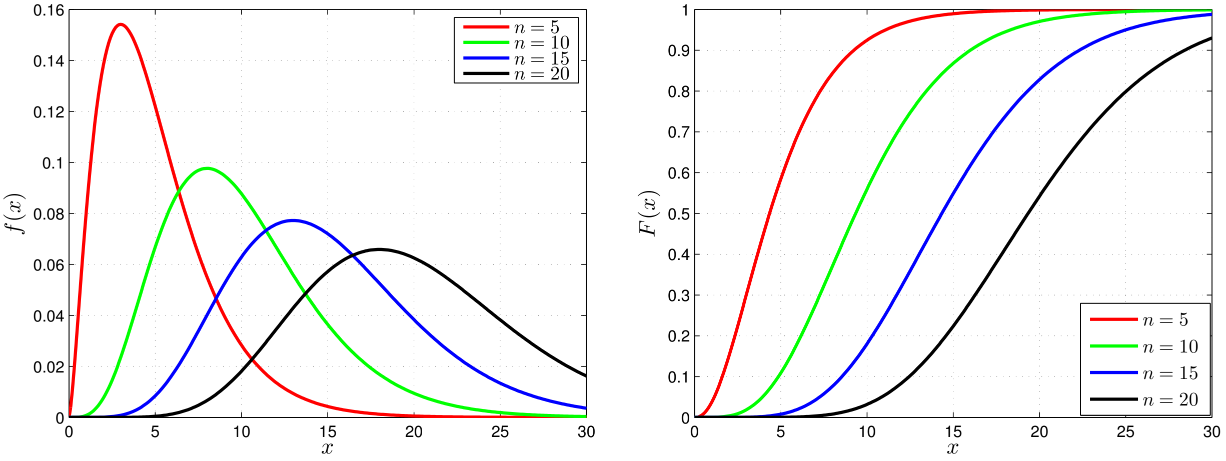
For $\alpha \in (0,1)$, let $\chi^2_{\alpha,n}$ be the percentage point of the χ^2 distribution as shown in Figure 3.21. That is,

$$P(X > \chi^2_{\alpha,n}) = 1 - P(X \leq \chi^2_{\alpha,n}) = 1 - F(\chi^2_{\alpha,n}) = \alpha$$

Thus, $\chi^2_{\alpha,n} = F^{-1}(1-\alpha)$ where F^{-1} denotes the inverse cdf of the $\chi^2(n)$ distribution.
For $\alpha = 0.05$ and $n = 5$, the value of $\chi^2_{\alpha,n}$ is computed using:

MATLAB code

```
>> alpha = 0.05; n = 5;
>> icdf('chi2',1-alpha,n)
ans =
    11.0705
```



(a) (b)

FIGURE 3.20: χ^2-distribution $X \sim \chi^2(n)$ for various values of n: (a) pdf; (b) cdf.

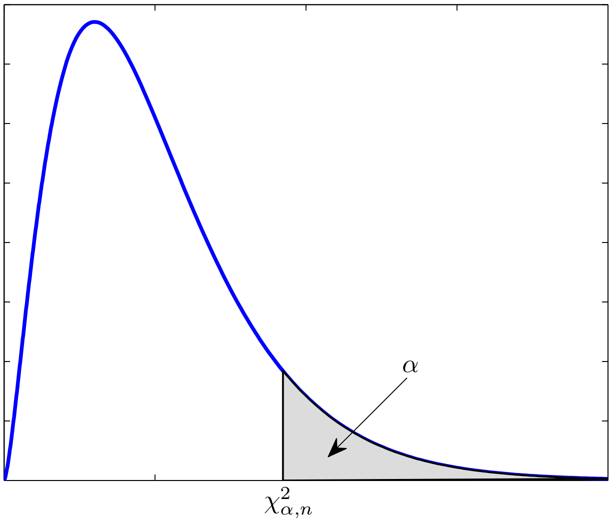


FIGURE 3.21: The percentage point of the χ^2-distribution.

3.4.6 STUDENT'S t-DISTRIBUTION

The Student's t-distribution (or simply, t-distribution) is a continuous probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown. It is typically used to develop hypothesis tests and confidence intervals.

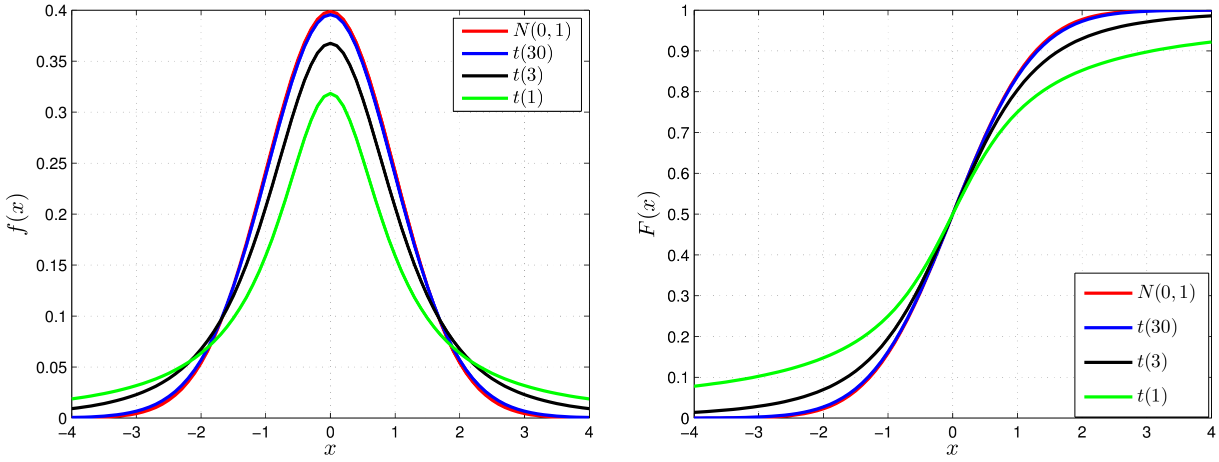
Definition 3.12 *Let $Z \sim N(0,1)$ and $Y \sim \chi^2(n)$ be independent random variables. Then, the random variable $X = \frac{Z}{\sqrt{Y/n}}$ is said to have a t distribution, denoted $X \sim t(n)$, with n degrees of freedom. The density function of X is*

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, \qquad -\infty < x < \infty \tag{3.4.13}$$

The mean and variance of $X \sim t(n)$ are $E(X) = 0$ and $Var(X) = \dfrac{n}{n-2}$

The pdf and cdf of the t-distribution for various values of degrees of freedom n are shown in Figure 3.22. As the number of degrees of freedom grows, the t-distribution approaches the normal distribution (red plot) with mean 0 and variance 1. Note that the degrees of freedom of the t-distribution can be a real number.

```
n = 3; % degree of freedom
x = -4:.1:4;
plot(x,pdf('t',x,n),'LineWidth',2); % Visualize the pdf
figure;
plot(x,cdf('t',x,n),'LineWidth',2); % Visualize the cdf
```



(a)

(b)

FIGURE 3.22: t-distribution $X \sim t(n)$ for various values of n: (a) pdf; (b) cdf.

For $\alpha \in (0,1)$, let $t_{\alpha,n}$ be the percentage point of the t-distribution as shown in Figure 3.23. That is,

$$P(X > t_{\alpha,n}) = 1 - P(X \leq t_{\alpha,n}) = 1 - F(t_{\alpha,n}) = \alpha$$

Thus, $t_{\alpha,n} = F^{-1}(1-\alpha)$ where F^{-1} denotes the inverse cdf of the t distribution. From the symmetry of the t-distribution about the origin, it follows that $-t_{\alpha,n} = t_{1-\alpha,n}$.
For $\alpha = 0.05$ and $n = 5$, the value of $t_{\alpha,n}$ is computed using:

MATLAB code

```
>> alpha = 0.05; n = 5;
>> icdf('t',1-alpha,n)
ans =
    1.4759
```

3.4.7 F-DISTRIBUTION

The F-distribution is formed from the ratios of two chi-squared random variables, and it arises in the testing of whether two observed samples have the same variance. This distribution is typically used to develop hypothesis tests and confidence intervals. The most common application of the F-distribution is in standard tests of hypotheses in analysis of variance and regression.

Definition 3.13 *Let $Y_1 \sim \chi^2(n_1)$ and $Y_2 \sim \chi^2(n_2)$ be independent random variables distributed as chi-squared with n_1 and n_2 degrees of freedom, respectively. Then, the random variable $X = \frac{Y_1/n_1}{Y_2/n_2}$ is said to have an F distribution, denoted $X \sim F(n_1, n_2)$, with n_1 and n_2 degrees of freedom. The density function of X is*

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)(n_1^{n_1/2} n_2^{n_2/2}) x^{(n_1-2)/2}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)[n_2 + n_1 x]^{(n_1+n_2)/2}}, \qquad x > 0 \tag{3.4.14}$$

The mean and variance of $X \sim F(n_1, n_2)$ are $E(X) = \dfrac{n_2}{n_2 - 2}$ for $n_2 > 2$ and $Var(X) = \dfrac{n_2^2(2n_2 + 2n_1 - 4)}{n_1(n_2-2)^2(n_2-4)}$ for $n_2 > 4$

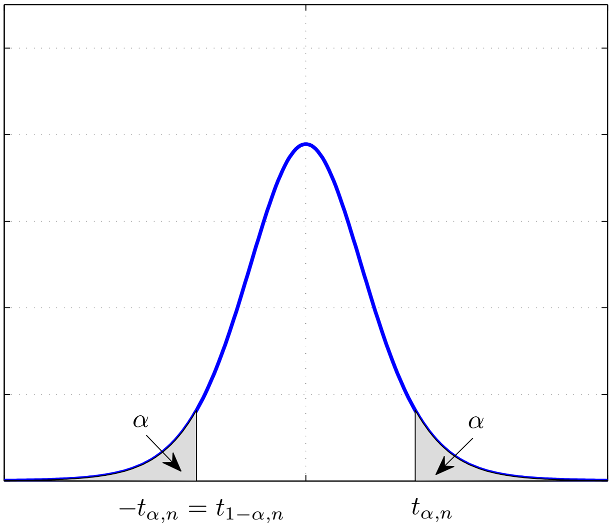


FIGURE 3.23: The percentage point of the t-distribution.

The pdf and cdf of the F-distribution for degrees of freedom $n_1 = 50$ and $n_2 = 10$ are shown in Figure 3.24. When describing an F-distribution, the number of degrees of freedom associated with the standard deviation in the numerator of the F random variable is always stated first. Thus, $F(50, 10)$ would refer to an F-distribution with $n_1 = 50$ and $n_2 = 10$ degrees of freedom; whereas $F(10, 50)$ would refer to an F-distribution with $n_1 = 10$ and $n_2 = 50$ degrees of freedom. Note that the curve represented by $F(50, 10)$ would differ from the curve represented by $F(10, 50)$.

The MATLAB function `cdf('F',x,n1,n2)` returns values of the cdf of any specified F distribution. For example, if $X \sim F(2, 27)$, then $P(X > 2.5)$ is computed as follows:

MATLAB code

```
>> x = 2.5; n1 = 2; n2 = 27;
>> 1-cdf('F',x,n1,n2)
ans =
    0.1009
```

```
n1 = 50;
n2 = 10;
x = 0:.01:5;
plot(x,pdf('F',x,n1,n2),'LineWidth',2); % Visualize the pdf
figure;
plot(x,cdf('F',x,n1,n2),'LineWidth',2); % Visualize the cdf
```

For $\alpha \in (0, 1)$, let F_{α,n_1,n_2} be the percentage point of the F distribution as shown in Figure 3.25. That is,

$$P(X > F_{\alpha,n_1,n_2}) = 1 - P(X \leq F_{\alpha,n_1,n_2}) = 1 - F(F_{\alpha,n_1,n_2}) = \alpha$$

Thus, $F_{\alpha,n_1,n_2} = F^{-1}(1 - \alpha)$ where F^{-1} denotes the inverse cdf of the F distribution. For $\alpha = 0.05$, $n_1 = 11$, and $n_2 = 9$, the value of F_{α,n_1,n_2} is computed using:

MATLAB code

```
>> alpha = 0.05; n1 = 11; n2 = 9;
>> icdf('F',1-alpha,n1,n2)
ans =
    2.1171
```

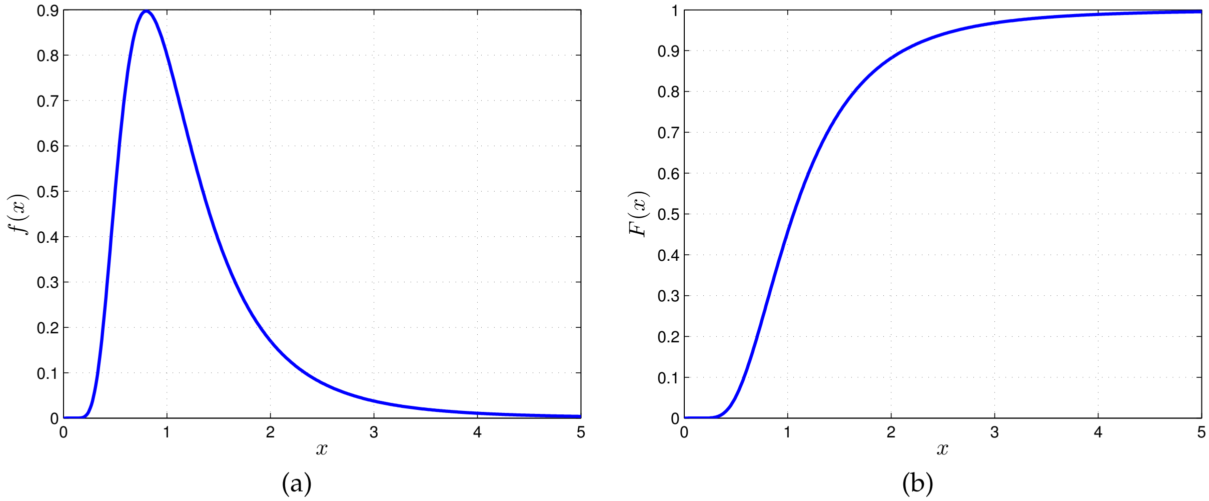


FIGURE 3.24: F-distribution $F \sim F(n_1, n_2)$ with $n_1 = 50$ and $n_2 = 10$: (a) pdf; (b) cdf.

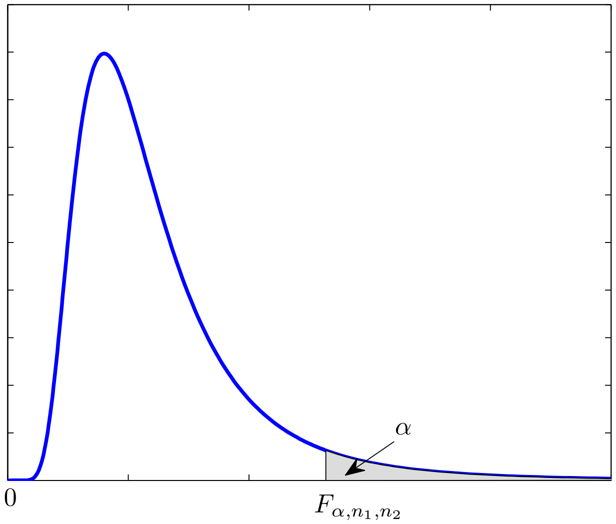


FIGURE 3.25: The percentage point of the F-distribution.

3.5 INFERENTIAL STATISTICS

Inferential statistics are used to draw inferences about a population from a sample. The term population refers to all possible measurements or outcomes that are of interest to us in a particular study, while the term sample refers to a portion of the population that is representative of the population from which it was selected, as shown in Figure 3.26. Consider an experiment in which 10 subjects who performed a task after 24 hours of sleep deprivation scored 12 points lower than 10 subjects who performed after a normal night's sleep. Is the difference real or could it be due to chance? How much larger could the real difference be than the 12 points found in the sample? These are the types of questions answered by inferential statistics. There are two main methods used in inferential statistics: estimation and hypothesis testing. In estimation, the sample is used to estimate a parameter and a confidence interval about the estimate is constructed. In the most common use of hypothesis testing, a null hypothesis is put forward and it is determined whether the data are strong enough to reject it. For the sleep deprivation study, the null hypothesis would be that sleep deprivation has no effect on performance.

FIGURE 3.26: Illustration of population vs. sample.

3.5.1 Random Sampling

Definition 3.14 *A* ***random sample*** $X_1, \ldots, X_n$ *of size n is a collection of n random variables that are independent and identically distributed (i.i.d).*

- *The sample mean is defined as*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3.5.1}$$

- *The sample variance is defined as*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{3.5.2}$$

- *The sample standard deviation S is the positive square root of the sample variance*

In descriptive statistics, note that the sample mean, for instance, was defined as an arithmetic average of a fixed set of numbers. In inferential statistics, however, the sample mean is defined as the average of random variables. Thus, $\overline{X}$ is also a random variable. Assume that $X_1, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, that is $E(X_i) = \mu$ and $var(X_i) = \sigma^2$ for $i = 1, \ldots, n$. Then, the mean $\mu_{\overline{X}}$ and variance $\sigma^2_{\overline{X}}$ of the sample mean $\overline{X}$ are given by

$$\mu_{\overline{X}} = E(\overline{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu \tag{3.5.3}$$

and

$$\sigma^2_{\overline{X}} = var(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}\cdot \tag{3.5.4}$$

Example 3.28 *The GPAs of all students enrolled at a large university have an approximately normal distribution with mean of 3.02 and a standard deviation of 0.29. Let $\overline{X}$ the mean GPA of a random sample of 25 students selected from this university. Find the mean and standard deviation of $\overline{X}$.*

Solution: From the given information, we have $n = 25$, $\mu = 3.02$, and $\sigma = 0.29$.
The mean $\mu_{\overline{X}}$ and variance $\sigma^2_{\overline{X}}$ of $\overline{X}$ are

$$\mu_{\overline{X}} = E(\overline{X}) = \mu = 3.02$$

and

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.29}{\sqrt{25}} = 0.058.$$

Definition 3.15 *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be two random samples of size n from a population with mean μ and variance σ^2, and denote by S_X and S_Y their respective sample standard deviations.*

- *The sample covariance is defined as* $S^2_{XY} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$
- *The sample correlation coefficient is defined as* $r_{XY} = \dfrac{S^2_{XY}}{S_X S_Y}$

3.5.2 Central Limit Theorem for Sample Means

Theorem 3.16 *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Then for n large, the standardized random variable*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{\bullet}{\sim} N(0,1) \tag{3.5.5}$$

where $\overset{\bullet}{\sim}$ means "is approximately distributed as." Thus, the sample mean $\overline{X}$ has approximately a normal distribution with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, i.e. $\overline{X} \overset{\bullet}{\sim} N(\mu, \sigma^2/n)$.

An important point of this result is that the variance of $\overline{X}$ decreases as the sample size increases. Basically for a large sample size n, the central limit theorem states that the sample mean $\overline{X}$ from a population is approximately normally distributed with a mean $\mu_{\overline{X}}$ equal to the population mean μ, even if the population is not normally distributed, and that the variance $\sigma^2_{\overline{X}}$ of the sample mean is n times smaller than the population variance σ^2. But, how large is "large enough"? As a rough rule of thumb, many statisticians say that a sample size of 30 is large enough. The random variable Z is referred to as a *test statistic* and its value $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is called *z-score*.

Therefore, the Central Limit Theorem states that for sufficiently large sample sizes ($n \geq 30$), regardless of the shape of the population distribution, if samples of size n are randomly drawn from a population that has a mean μ and a standard deviation σ, the samples' means $\overline{X}$ are approximately normally distributed. If the populations are normally distributed, then the samples' means are normally distributed regardless of the sample sizes. The implication of this theorem is that for sufficiently large populations, the normal distribution can be used to analyze samples drawn from populations that are not normally distributed, or whose distribution characteristics are unknown.

Regardless of its shape, the sampling distribution of $\overline{X}$ always has a mean identical to the mean of the sampled population and a standard deviation equal to the population standard deviation σ divided by $\sqrt{n}$. Consequently, the spread of the distribution of sample means is considerably less than the spread of the sampled population.

Figure 3.27 displays the probability density function of the sample mean of 30 uniformly distributed random variables $X_i \sim unif(0,1)$, $i = 1, \ldots, 30$, and the corresponding sampling distribution $N(\mu, \sigma^2/n)$ with $\mu = 1/2$ and $\sigma^2 = 1/12$. That is, the sampling distribution of $\overline{X}$ is $N(1/2, 1/360)$. Note how close the approximation to the normal distribution is.

Example 3.29 *Assume that the weights of all packages of a certain brand of chocolate cookies have a mean of 32 ounces and a standard deviation of $\sigma = 0.3$ ounce. Find the probability that the mean weight of a random sample of 40 packages of this brand of cookies will be between 31.8 and 31.9 ounces.*

Solution: From the given information, we have $n = 40$, $\mu = 32$, and $\sigma = 0.3$. Since the sample size $n = 40$ is large, we can apply the central limit theorem for sample means to find $P(31.8 \leq \overline{X} \leq 31.9)$, which is given by

$$P(31.8 \leq \overline{X} \leq 31.9) = P\left(\frac{31.8 - 32}{0.3/\sqrt{40}} \leq Z \leq \frac{31.9 - 32}{0.3/\sqrt{40}}\right) = P(-4.21 \leq Z \leq -2.11) = 0.0175.$$

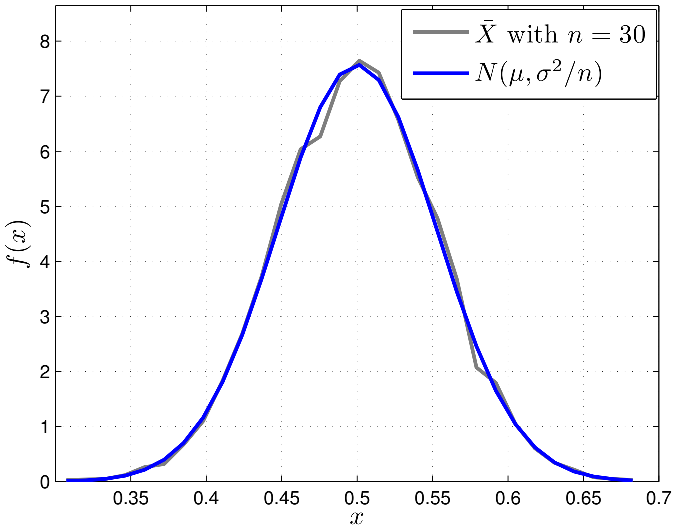


FIGURE 3.27: Illustration of the central limit theorem using 30 uniformly generated random variables.

Example 3.30 *A tire manufacturer claims that its tires will last an average of 60,000 miles with a standard deviation of 3,000 miles. Sixty-four tires were placed on test and the average failure miles for these tires was recorded. What is the probability that the average failure miles will be more than 59,500 miles?.*

Solution: We have $\bar{x} = 59500, \mu = 60000, \sigma = 3000$, and $n = 64$. Since the sample size $n = 64$ is large, we can apply the central limit theorem for sample means to find $P(\overline{X} > 59,500)$. The z-score is given by

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{59500 - 60000}{3000/\sqrt{64}} = -1.33$$

Thus, $P(\overline{X} > 59500) = P(Z > -1.33) = 1 - \Phi(-1.33) = \Phi(1.33) = 0.9082$.

3.5.3 Central Limit Theorem for Sample Proportions

A sampling distribution of a sample proportion is a distribution obtained by using the proportions computed from random samples of a specific size obtained from a population.

Suppose that random samples of size n are selected from a population (distribution) in which the true proportion of the attribute of interest is p. Then, provided that $np > 5$ and $n(1-p) > 5$, the sampling distribution of the sample proportion $\hat{p}$ is approximately normally distributed with mean $\mu = p$ and standard deviation $\sqrt{p(1-p)/n}$

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \overset{\bullet}{\sim} N(0,1). \tag{3.5.6}$$

Example 3.31 *It is estimated that approximately 53 percent of university students graduate in 4 years or less. This figure is affected by the fact that more students are attending university on a part-time basis. If 500 students on a large campus are selected at random, what is the probability that between 50 and 60 percent of them will graduate in 4 years or less?.*

Solution: We have $n = 500, p = 0.53$, and $\sqrt{p(1-p)/n} = 0.0223$. Since $np = 265 > 5$ and $n(1-p) = 235 > 5$, we can apply the central limit theorem for sample proportions to find $P(0.5 \leq \hat{p} \leq 0.6)$. The z-scores are $z = (0.5 - 0.53)/0.0223 = -1.35$ and $z = (0.6 - 0.53)/0.0223 = 3.14$. Thus,

$$P(0.5 \leq \hat{p} \leq 0.6) = P(-1.35 \leq Z \leq 3.14) = \Phi(3.14) - \Phi(-1.35) = 0.91.$$

Example 3.32 *Almudaina Corporation manufactures USB drives. The machine that is used to make these USBs is known to produce 6% defective USBs. The quality control inspector selects a sample of 100 USBs every week and inspect them for being good or defective. If 8% or more of the USBs in the sample are defective, the process is stopped and the machine is readjusted. What is the probability that based on a sample of 100 USBs, the process will be stopped to readjust the machine?.*

Solution: From the given information, we have $n = 100$, and $p = 0.06$. Since $np = 6 > 5$ and $n(1-p) = 94 > 5$, we can apply the central limit theorem for sample proportions to find $P(\hat{p} > 0.08)$, which is given by

$$P(\hat{p} > 0.08) = P\left(Z > \frac{0.08 - 0.06}{\sqrt{(0.06)(1-0.06)/100}}\right) = P(Z > 0.8422) = 1 - \Phi(0.8422) = 0.1999.$$

3.6 Multivariate Distributions

There are many situations in which we are interested in how two or more random variables are related. We start with a brief review of some important concepts from matrix theory.

3.6.1 Concepts from Matrix Algebra

Matrices play a central role in multivariate statistics. A matrix is a rectangular array of elements arranged in rows and columns. An $m \times n$ matrix with m rows and n columns is given by

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{1,2} & \cdots & a_{1n} \\ a_{21} & a_{2,2} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{3.6.1}$$

where a_{ij} are the elements or entries of the matrix A.

A column vector is an $m \times 1$ matrix that consists of a single column of m elements:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

whereas a row vector is an $1 \times m$ matrix that consists of a single row of m elements:

$$\boldsymbol{x} = (x_1,\ x_2, \ldots, x_m).$$

Square Matrix:

A *square matrix* is an $n \times n$ matrix with the same number of rows and columns.

Transpose of a Matrix:

Let A be $m \times n$ matrix. The transpose of A, denoted by A', is obtained by interchanging its rows and columns. The order (or size) of A' is $n \times m$. For example, if

$$A = \begin{pmatrix} 1 & 3 & 4 \\ 7 & 0 & 1 \end{pmatrix}$$

then

$$A' = \begin{pmatrix} 1 & 7 \\ 3 & 0 \\ 4 & 1 \end{pmatrix}.$$

Symmetric Matrix:

When $A' = A$, the matrix is called *symmetric*. That is, a symmetric matrix is a square matrix, in that it has the same number of rows as it has columns, and the off-diagonal elements are symmetric (i.e. $a_{ij} = a_{ji}$ for all i and j). For example,

$$A = \begin{pmatrix} 1 & 9 & 5 \\ 9 & 2 & -4 \\ 5 & -4 & 3 \end{pmatrix}$$

is a symmetric matrix.

Diagonal Matrix:

A special case of a symmetric matrix is the *diagonal matrix*, which is an $n \times n$ matrix whose non-diagonal entries are zeros. A diagonal matrix is trivially symmetric. For example,

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

is a diagonal matrix.

An *identity matrix* is a diagonal matrix of order n which has 1's on the diagonal and 0's on the off-diagonal

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Matrix Inverse:

A $n \times n$ square matrix A is called *invertible* or *non-singular* if there exists a matrix B such that: $AB = BA = I_n$. The matrix B is called the *inverse* of A and is denoted by A^{-1}. For example, the inverse of

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

is given by

$$A^{-1} = \begin{pmatrix} -1 & 1 \\ 1.5 & -0.5 \end{pmatrix}$$

MATLAB code

```
>> A = [1 2; 3 4]
>> inv(A)
ans =
   -2.0000    1.0000
    1.5000   -0.5000
```

Eigenvalues and Eigenvectors of a Square Matrix:

An eigenvalue and eigenvector of an $n \times n$ square matrix A are a scalar λ_i and a nonzero vector $\boldsymbol{x}_i$ that satisfy

$$A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i \qquad i = 1, \ldots, n \tag{3.6.2}$$

which can expressed in matrix form as follows:

$$AV = VD \tag{3.6.3}$$

where D is a diagonal matrix whose diagonal elements are the eigenvalues, and V is a matrix whose columns are the eigenvectors. These matrices can be easily computed using the MATLAB function `eig`:

MATLAB code

```
>> A = [1 2; 3 4]
>> [V,D]=eig(A)
V =
   -0.8246   -0.4160
    0.5658   -0.9094

D =
   -0.3723         0
         0    5.3723
```

In this output, the columns of V correspond to eigenvectors of A, with eigenvalues given on the diagonal of D. For instance, in our example, A has eigenvector $\begin{pmatrix} -0.8246 \\ 0.5658 \end{pmatrix}$ with eigenvalue -0.3723, and eigenvector $\begin{pmatrix} -0.4160 \\ -0.9094 \end{pmatrix}$ with eigenvalue 5.3723.

3.6.2 Bivariate Distributions

Let X and Y be two random variables, then the probability distributions may be summarized as follows:

Discrete case:

- The joint pmf of X and Y is defined as $f_{XY}(x,y) = P(X = x, Y = y)$
- The marginal pmfs of X and of Y are $f_X(x) = \sum_y f(x,y)$ and $f_Y(y) = \sum_x f(x,y)$

Continuous case:

- The joint cdf of X and Y is defined as $F_{XY}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dxdy$, where $f(x,y)$ is the joint pdf also called bivariate probability density.
- The marginal pdfs are $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$

We say that the random variables X and Y are *independent* if for every pair of x and y values, the joint pmf or pdf is equal to the product of the marginal pmfs or pdfs: $f(x,y) = f_X(x)f_Y(y)$. Otherwise, the random variables X and Y are said to be *dependent*. Note that the independence property can be generalized to any number of random variables.

When two random variables are not independent, it is frequently of interest to measure how they are related to one another. In other words, information about the value of one random variable helps determine the value of the other. This information is often measured using the covariance or correlation between the two random variables.

Definition 3.17 *The covariance between X and Y, denoted cov(X,Y) or σ_{XY}, is defined as*

$$\mathrm{cov}(X,Y) = \sigma_{XY} = \mathrm{E}(X - \mu_X)\mathrm{E}(Y - \mu_Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{3.6.4}$$

where μ_X and μ_Y are the expected values of X and Y, respectively.
The correlation coefficient between X and Y, denoted $\rho(X,Y)$, is defined as

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{3.6.5}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

Note that the correlation coefficient $\rho(X,Y)$ remains unaffected by a change of units, and therefore it is dimensionless. Also, it can be shown that $-1 \leq \rho \leq 1$. The correlation coefficient measures the linear relationship between X and Y. The strongest possible relationship corresponds to $\rho = 1$, while the strongest negative relationship corresponds to $\rho = -1$. When $\rho = 0$, the random variables X and Y are said to be *uncorrelated*, but *not necessarily independent*. Two variables could be uncorrelated yet highly dependent because there is a strong nonlinear relationship. Any value $|\rho| < 1$ indicates only that the relationship is not completely linear, but there may still be a very strong nonlinear relationship.

3.6.3 Multivariate Normal Distribution

A multivariate random variable is a row-vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ (or column-vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$) whose p components are scalar-valued random variables. The expected value or mean of the random vector X is equal to $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu} = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_p))$, and the covariance matrix of X is a $p \times p$ symmetric non-negative definite matrix, denoted $\Sigma = (\sigma_{ij})$, whose (i,j)-th element σ_{ij} is the covariance between the i-th and the j-th random variables, that is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} \tag{3.6.6}$$

Note that the diagonal element $\sigma_{ii} = \sigma_i^2$ represents the variance of the i-th random variable X_i, while the off-diagonal element $\sigma_{i,j}$ represents the covariance between the i-th and the j-th random variables X_i and X_j, where $i, j = 1, \ldots, p$.

Definition 3.18 *A p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ is said to have a multivariate normal distribution, denoted $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if the pdf of $\boldsymbol{X}$ is*

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3.6.7}$$

where $|\Sigma|$ and Σ^{-1} are the determinant and inverse matrix of Σ, respectively.

In the 2-dimensional (bivariate) case, the pdf of a random vector (X, Y) is

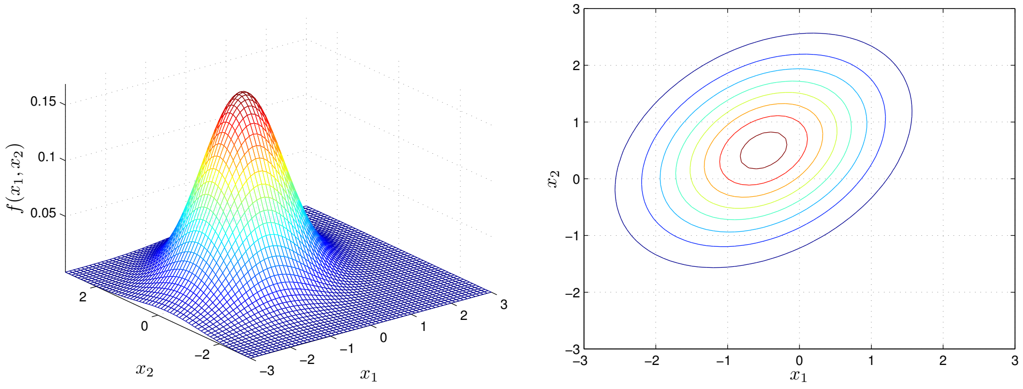
$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right] - \frac{2\rho(x_1-\mu_1)^2}{\sigma_1\sigma_2}\right),$$

where ρ is the correlation coefficient between X_1 and X_2. In this case,

$$\boldsymbol{\mu} = (\mu_1, \mu_2) \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Figure 3.28 displays the surface and contour plots of a bivariate normal pdf with

$$\boldsymbol{\mu} = (1/2, -1/2) \quad \text{and} \quad \Sigma = \begin{pmatrix} 1 & 1/3 \\ 1/3 & 1 \end{pmatrix}.$$



(a) (b)

FIGURE 3.28: Bivariate normal pdf: (a) surface plot, and (b) contour plot.

3.7 Problems

❶ Let a random variable X denote the number of medication errors for a patient at a hospital. Experience shows that it occurs between 0 and 3 (inclusive) with the following probabilities:

$X = x$	0	1	2	3
$f(x) = P(X = x)$	0.90	0.07	0.02	0.01

(i) Plot the probability mass function.

(ii) Calculate the mean μ and variance σ^2 of the random variable X.

(ii) Calculate the probability that X falls in the interval $(\mu - 6\sigma, \mu + 6\sigma)$.

❷ A machine produces soda bottles, and 98.5 percent of all bottles produced pass an audit. What is the probability of having only 2 bottles that pass audit in a randomly selected sample of 7 bottles?

❸ A product failure has historically averaged 3.84 occurrences per day. What is the probability of 5 failures in a randomly selected day?

❹ The average number of accidents occurring in a manufacturing plant over a period of one year is equal to two. Find the probability that during any given year five accidents will occur.

❺ The mean daily milk production of a herd of Greenville cows has a normal distribution with $\mu = 70$ pounds and $\sigma = 13$ pounds.

a) What is the probability that the milk production for a cow chosen at random will be less than 60 pounds?

b) What is the probability that the milk production for a cow chosen at random will be greater than 90 pounds?

c) What is the probability that the milk production for a cow chosen at random will be between 60 pounds and 90 pounds?

❻ A large drug company has 100 potential new prescription drugs under clinical test. About 20% of all drugs that reach this stage are eventually licensed for sale. What is the probability that at least 15 out of the 100 drugs will be eventually licensed?

❼ The reliability of an electrical fuse is the probability that a fuse, chosen at random from production, will function under its designed conditions. A random sample of 1000 fuses was tested and $x = 27$ defectives were observed. Calculate the approximate probability of observing 27 or more defectives, assuming that the fuse reliability is 0.98.

3.8 REFERENCES

[1] D.C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, 6th Edition, 2009.

[2] I. Bass and B. Lawton, *Lean Six Sigma using SigmaXL and Minitab*, McGraw-Hill Professional, 1st Edition, 2009.

Chapter 4

Confidence Intervals and Hypothesis Testing

> The only relevant test of the validity of a hypothesis is comparison of its predictions with experience.
>
> *Milton Friedman*

A confidence interval is an interval that provides an estimated range of values which is likely to include, with a certain level of confidence, an unknown population parameter of interest. This estimated range of values is calculated from a given set of sample data. Confidence intervals are one way to represent how "good" an estimate is; the larger a confidence interval for a particular estimate, the more caution is required when using the estimate. Hypothesis testing, on the hand, is a method of drawing inferences about a population based on statistical evidence from a sample. A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. The purpose of hypothesis testing is to assess the validity of a statistical hypothesis made about a population.

4.1 Point and Interval Estimation

Point estimation involves the use of sample data to calculate a single value (statistic) which serves as an estimate of an unknown population parameter, while interval estimation involves the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter.

Let $X_1, \ldots, X_n$ be a random sample from a population with an unknown parameter θ. A **point estimator** of θ is the statistic $\hat{\Theta} = \varphi(X_1, \ldots, X_n)$, where φ is a given function. A numerical value $\hat{\theta}$ of the statistic $\hat{\Theta}$, calculated from the observed sample data $X_1 = x_1, \ldots, X_n = x_n$, is called the **point estimate** of the parameter θ. For example, the sample mean $\overline{X}$ is an estimator of a normal population mean μ, while the numerical value $\hat{\mu} = \bar{x}$ calculated from a observed sample data is the point estimate of μ. A good estimator must satisfy three conditions:

- **Unbiased:** The expected value of the estimator must be equal to the parameter

- **Consistent:** The value of the estimator approaches the value of the parameter as the sample size increases

- **Relatively Efficient:** The estimator has the smallest variance of all estimators which could be used.

Definition 4.1 *Assume that $X_1, \ldots, X_n$ is a random sample from a population with an unknown parameter θ. Then,*

- *The point estimator $\hat{\Theta}$ is an unbiased estimator for the parameter θ if $E(\hat{\Theta}) = \theta$.*

- *The minimum variance unbiased estimator (MVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter θ.*

- *The standard error of a point estimator $\hat{\Theta}$ is given by $s.e.(\hat{\Theta}) = \sqrt{var(\hat{\Theta})}$.*

A biased point estimator is an estimator such that $E(\hat{\Theta}) = \theta + \text{bias}$. For example, both the sample mean $\overline{X}$ and sample variance S^2 are unbiased estimators for the population mean μ and the population variance σ^2, respectively, i.e.

$E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$. However, the sample standard deviation S is a biased estimator of the population standard deviation σ, i.e. $E(S) \neq \sigma$.

The MVUE is the most efficient estimator. An efficient estimator $\hat{\Theta}$ will produce an estimate closer to the true parameter μ, and it can be shown that the sample mean $\overline{X}$ is MVUE for the population mean μ. On the other hand, an **interval estimate** refers to a range of values used to estimate the population parameter. Such an interval estimate is obtained by making use of the probability distribution of the point estimator.

Example 4.1 *In a sample of five measurements, the diameter of a sphere was recorded by a scientist as 6.33, 6.37, 6.36, 6.32, and 6.37 centimeters (cm).*

(i) Determine an unbiased and efficient estimate of the population mean.

(ii) Determine an unbiased and efficient estimate of the population variance.

(iii) Give an unbiased and inefficient estimate of the population mean.

Solution: From the given information, the sample size is $n = 5$.

(i) The unbiased and efficient estimate of the the population mean is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{5}(6.33 + 6.37 + 6.36 + 6.32 + 6.37) = 6.35$$

(ii) The unbiased and efficient estimate of the population variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0.00055.$$

(iii) The median Q_2 is one example of an unbiased and inefficient estimate of the population mean. By ordering the data, we obtain $Q_2 = 6.36$.

MATLAB code

```
>> X = [6.33 6.37 6.36 6.32 6.37];
>> mean(X), var(X), median(X)
```

4.2 Sampling Distributions

Any function of the random sample observations or elements is called a **statistic**. The sample mean $\overline{X}$, the sample variance S^2, and the sample standard deviation S are examples of a statistic. Generally, a statistic is used to estimate the value of a population parameter. For example, $\overline{X}$ is a statistic that serves as an estimate of the population mean μ. The probability distribution of a statistic is called a **sampling distribution**.

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be a random sample from a *normal* population with mean μ and variance σ^2. According to the central limit theorem, the sampling distribution of $\overline{X}$ is $N(\mu, \sigma^2/n)$. Therefore, the sampling distribution of the statistic

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{\bullet}{\sim} N(0,1) \tag{4.2.1}$$

approximately follows a standard normal distribution. Figure 4.1 displays probability density functions of the sample mean from a standard normal population for various sample sizes. The red curve for the sample size of $n = 1$ corresponds to $N(0,1)$.

It can be shown that the expected value of the sample variance is equal to the population variance, that is $E(S^2) = \sigma^2$. Moreover, sampling distribution of the statistic

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1) \tag{4.2.2}$$

has a χ^2-distribution with $n - 1$ degrees of freedom. That is, the sampling distribution of the sample variance is chi-squared: $S^2 \sim [\sigma^2/(n-1)]\chi^2(n-1)$.

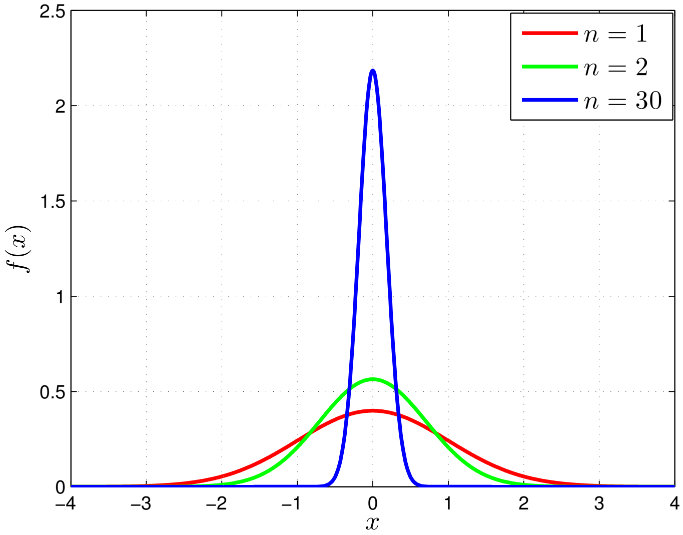


FIGURE 4.1: Sampling pdfs of the sample mean $\overline{X}$ for different values of the sample size.

Example 4.2 *The time it takes a central processing unit to process a certain type of job is normally distributed with mean 20 seconds and standard deviation 3 seconds. If a sample of 15 such jobs is observed, what is the probability that the sample variance will exceed 12?*

Solution: From the given information, we have $n = 15$ and $\sigma^2 = 9$. Thus,

$$\begin{aligned} P(S^2 > 12) &= P\left((n-1)\frac{S^2}{\sigma^2} > (14)\frac{12}{9}\right) \\ &= P(\chi^2(14) > 18.67) \\ &= 1 - P(\chi^2(14) \leq 18.67) = 0.1779, \end{aligned}$$

where $\chi^2(14)$ is a χ^2-distribution with $n - 1 = 14$ degrees of freedom. ■

MATLAB code

```
>> n = 15; sigma2 = 9; x = 12;
>> chisq = (n-1)*x/sigma2;
>> P = 1-cdf('chi2',chisq,n-1)
```

Example 4.3 *The Bravo Widget Company claims that their widgets last 5 years, with a standard deviation of 1 year. Assume that their claims are true. If you test a random sample of 9 Acme widgets, what is the probability that the standard deviation in your sample will be less than 0.95 years?*

Solution: From the given information, we have $n = 9$ and $\sigma = 1$. Thus,

$$\begin{aligned} P(S < 0.95) &= P\left((n-1)\frac{S^2}{\sigma^2} < (8)\frac{(0.95)^2}{(1)^2}\right) \\ &= P(\chi^2(8) < 7.22) = 0.4869, \end{aligned}$$

where $\chi^2(8)$ is a χ^2-distribution with $n - 1 = 8$ degrees of freedom. ■

If in Eq. (4.2.1) we replace σ by the sample standard deviation S, then the sampling distribution of the statistic

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{S/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-1}\chi^2(n-1)}} \sim t(n-1) \tag{4.2.3}$$

has a t-distribution with $n-1$ degrees of freedom.

Example 4.4 *Eureka Corporation manufactures light bulbs. The CEO claims that an average Eureka light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?*

Solution: From the given information, we have $n = 15$, $\mu = 300$, $\bar{x} = 260$ and $s = 50$. Thus,

$$\begin{aligned} P(\overline{X} \le 290) &= P\left(\frac{\overline{X}-\mu}{S/\sqrt{n}} \le \frac{290-300}{50/\sqrt{15}}\right) \\ &= P(t(14) \le -0.7746) = 0.2257, \end{aligned}$$

where $t(14)$ is a t-distribution with $n-1 = 14$ degrees of freedom. ■

MATLAB code

```
>> n = 15; mu = 300; xbar = 290; s = 50;
>> x = (xbar-mu)/(s/sqrt(n));
>> P = cdf('t',x,n-1)
```

Let $X_1, \dots, X_{n_1} \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \dots, Y_{n_2} \sim N(\mu_Y, \sigma_Y^2)$ be two samples from two normal populations. Then, the sampling distribution of the statistic

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n_1-1, n_2-1) \tag{4.2.4}$$

has an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, where S_X^2 and S_Y^2 are the sample variances of the corresponding populations.

Example 4.5 *Consider two independent samples: the first of size 10 from a normal population having variance 4 and the second of size 5 from a normal population having variance 2. Compute the probability that the sample variance from the second sample exceeds the one from the first.*

Solution: From the given information, we have $n_1 = 10$, $n_2 = 5$, $\sigma_X^2 = 4$ and $\sigma_Y^2 = 2$. Thus,

$$\begin{aligned} P(S_Y^2 > S_X^2) &= P\left(\frac{S_X^2}{S_Y^2} < 1\right) \\ &= P\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} < \frac{\sigma_Y^2}{\sigma_X^2}\right) \\ &= P(F(9,4) < 1/2) = 0.1782, \end{aligned}$$

where $F(9,4)$ is an F-distribution with $n_1 - 1 = 9$ and $n_1 - 1 = 4$ degrees of freedom. ■

MATLAB code

```
>> n1 = 10; n2 = 5; sigmax2 = 4; sigmay2 = 2;
>> x = sigmay2/sigmax2;
>> P = cdf('F',x,n1-1,n2-1)
```

4.3 Confidence Intervals

A **confidence interval** is an interval estimate with a specific *confidence level*, $(1-\alpha)\%$, where $\alpha \in (0,1)$. The *confidence coefficient*, denoted $1-\alpha$, is the probability that the interval estimate will contain the population parameter θ. More specifically, we want to construct a $100(1-\alpha)\%$ confidence interval $\ell \le \theta \le u$ such that

$$P(\ell \le \theta \le u) = 1 - \alpha \tag{4.3.1}$$

where ℓ and u are called *lower and upper confidence limits*, respectively. These confidence limits are calculated from the observed sample data.

When a one-sided specification is employed, then only a one-sided confidence limit is needed. In this case, a lower-confidence interval on θ is $\ell \leq \theta$ such that $P(\ell \leq \theta) = 1 - \alpha$ and an upper-confidence interval on θ is $\theta \leq u$ such that $P(\theta \leq u) = 1 - \alpha$. A typical value of the the confidence level $(1 - \alpha)$% is 95%, which means that if all samples of the same size were selected, 95% of them include the population parameter θ somewhere within the confidence interval, and 5% would not.

4.3.1 Confidence Interval on the Population Mean when the Variance is Known

Let $X_1, \ldots, X_n$ be a random sample of size n (≥ 30) from a normal population with an unknown parameter mean μ and known parameter variance σ^2. Since the sample size is large, it follows from the central limit theorem that the statistic $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, and as illustrated in Figure 4.2 we can write

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha,$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point of the standard normal distribution.

Rearranging the terms inside the parentheses yields

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

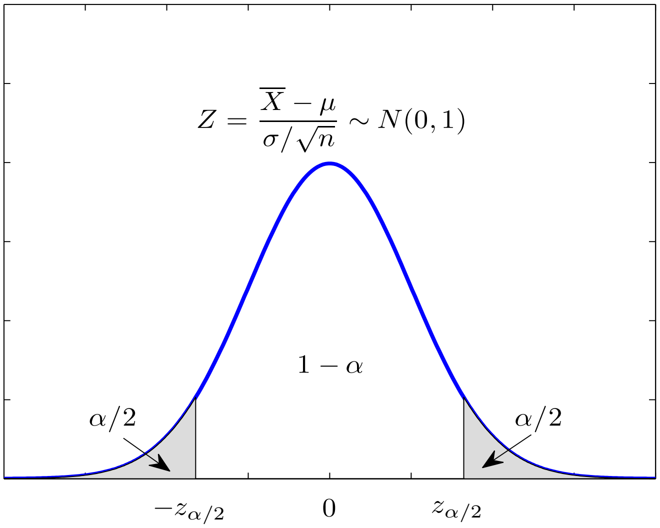


FIGURE 4.2: Illustration of confidence interval on the normal mean when the variance is known.

Definition 4.2 *Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$ distribution with an unknown parameter mean μ and known parameter variance σ^2.*

- *A $100(1-\alpha)\%$ confidence interval on μ is given by*

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{4.3.2}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on μ is given by*

$$\mu \le \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}} \tag{4.3.3}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on μ is given by*

$$\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} \le \mu \tag{4.3.4}$$

where $\bar{x}$ is the observed sample mean.

Example 4.6 *A survey was conducted of companies that use solar panels as a primary source of electricity. The question that was asked was this: How much of the electricity used in your company comes from the solar panels? A random sample of 55 responses produced a mean of 45 megawatts. Suppose the population standard deviation for this question is 15.5 megawatts.*

(i) Find the 95% confidence interval for the mean.

(ii) Find the 95% upper-confidence interval for the mean.

(iiii) Find the 95% lower-confidence interval for the mean.

Solution: From the given information, we have $\bar{x} = 45$, $\sigma = 15.5$, and $n = 55$.
For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$, and $z_{\alpha} = z_{0.05} = 1.64$.

(i) The 95% confidence interval is given by

$$\begin{gathered}\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ 45 - 1.96\frac{15.5}{\sqrt{55}} \le \mu \le 45 + 1.96\frac{15.5}{\sqrt{55}} \\ 40.9 \le \mu \le 49.1\end{gathered}$$

Thus, we can be 95% sure that the mean will be between 40.9 and 49.1 megawatts. In other words, the probability for the mean to be between 40.9 and 49.1 will be 0.95.

$$P(40.9 \le \mu \le 49.1) = 0.95.$$

MATLAB code

```
>> xbar = 45; sigma = 15.5; n = 55; alpha = 1-0.95;
>> zalpha2 = icdf('norm',1-alpha/2,0,1);
>> LC = xbar - zalpha2*sigma/sqrt(n);
>> UC = xbar + zalpha2*sigma/sqrt(n);
```

(ii) The 95% upper-confidence interval is given by

$$\mu \le \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}} = 45 + 1.64\frac{15.5}{\sqrt{55}} \;\Rightarrow\; \mu \le 48.44$$

(iii) The 95% lower-confidence interval is given by

$$\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} = 45 - 1.64\frac{15.5}{\sqrt{55}} \le \mu \;\Rightarrow\; 41.56 \le \mu.$$

4.3.2 Confidence Interval on the Population Mean when the Variance is Unknown

When the sample size is small, we cannot apply the central limit theorem. Thus, we either assume that the sampled population is normally distributed, or we need to verify that the sample data is approximately normally distributed using for example the normal probability plot or the box plot. Therefore, when the population is normal and sample size is small, the statistic $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ follows a t-distribution with $n-1$ degrees of freedom. Hence, as illustrated in Figure 4.3 we can write

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = 1-\alpha,$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the t-distribution with $n-1$ degrees of freedom. It follows that

$$P\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

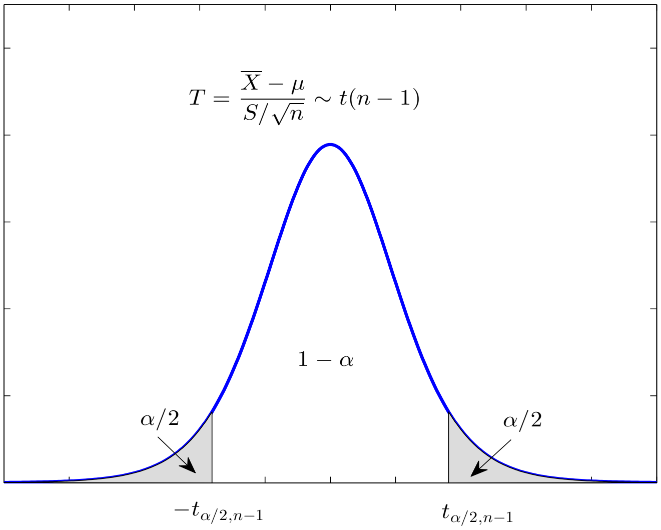


FIGURE 4.3: Illustration of confidence interval on the normal mean when the variance is unknown.

Definition 4.3 *Let $X_1, \ldots, X_n$ be a random sample from a normal population with an unknown parameter variance σ^2.*

- *A $100(1-\alpha)\%$ confidence interval on μ is given by*

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{4.3.5}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on μ is given by*

$$\mu \leq \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}} \tag{4.3.6}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on μ is given by*

$$\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} \leq \mu \tag{4.3.7}$$

where $\bar{x}$ and s are the observed sample mean and sample standard deviation, respectively.

Example 4.7 *A random sample of size 25 of a certain kind of lightbulb yielded an average lifetime of 1875 hours and a standard deviation of 100 hours. From past experience it is known that the lifetime of this kind of bulb is normally distributed. Find the a 99% confidence interval for the population mean.*

Solution: From the information given, we have $n = 25$, $\bar{x} = 1875$, $s = 100$. For 99% confidence level, we have $\alpha = 1 - 0.99 = 0.01$. Also, the population is assumed to be normally distributed. The 99% confidence interval is given by

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
$$1875 - 2.7969\frac{100}{\sqrt{25}} \leq \mu \leq 1875 + 2.7969\frac{100}{\sqrt{25}}$$
$$1819.1 \leq \mu \leq 1930.9$$

We can be 99% sure that the mean will be between 1819.1 and 1930.9. In other words, the probability for the mean to be between 1819.1 and 1930.9 will be 0.99. That is,

$$P(1819.1 \leq \mu \leq 1930.9) = 0.99.$$

MATLAB code

```
>> n=25; xbar = 1875; s = 100; alpha = 1-0.99;
>> talpha2 = icdf('t',1-alpha/2,n-1)
>> LC = xbar - talpha2*s/sqrt(n)
>> UC = xbar + talpha2*s/sqrt(n)
```

Example 4.8 *A manager of a car rental company wants to estimate the average number of times luxury cars would be rented a month. She takes a random sample of 19 cars that produces the following number of times the cars are rented in a month:*

```
3  7  12  5  9  13  2  8  6  14  6  1  2  3  2  5  11  13  5
```

(i) *Check the assumption of normality for the number of times the cars are rented in a month.*

(ii) *Find the 95% confidence interval to estimate the average.*

(iii) *Find the 95% upper-confidence interval to estimate the average.*

(iv) *Find the 95% lower-confidence interval to estimate the average.*

Solution: The sample size is $n = 19$. Thus, the sample mean and sample standard deviation of the data are: $\bar{x} = 127/19 = 6.68$ and $s = 4.23$. For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, $t_{\alpha/2,n-1} = t_{0.025,18} = 2.101$, and $t_{\alpha,n-1} = t_{0.05,18} = 1.734$.

(i) According to the normal probability plot shown in Figure 4.4, there does not seem to be a severe deviation from normality for this data. This is evident by the fact that the data appears to fall along a straight line.

MATLAB code

```
>> X = [3 7 12 5 9 13 2 8 6 14 6 1 2 3 2 5 11 13 5];
>> normplot(X);
```

(ii) The 95% confidence interval is given by

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
$$6.68 - 2.101\frac{4.23}{\sqrt{19}} \leq \mu \leq 6.68 + 2.101\frac{4.23}{\sqrt{19}}$$
$$4.65 \leq \mu \leq 8.72$$

We can be 95 percent sure that the mean will be between 4.65 and 8.72. In other words, the probability for the mean to be between 4.65 and 8.72 will be 0.95.

$$P(4.65 \leq \mu \leq 8.72) = 0.95.$$

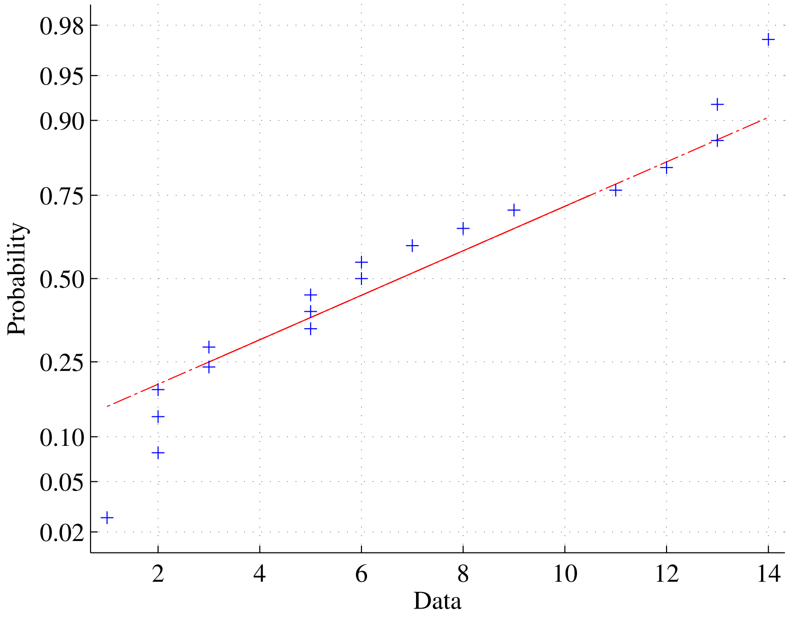


FIGURE 4.4: Normal probability plot for the number of times the cars are rented in a month.

MATLAB code

```
>> X = [3 7 12 5 9 13 2 8 6 14 6 1 2 3 2 5 11 13 5];
>> xbar = mean(X); s = std(X); n = 19; alpha = 1-0.95;
>> talpha2 = icdf('t',1-alpha/2,n-1)
>> LC = xbar - talpha2*s/sqrt(n)
>> UC = xbar + talpha2*s/sqrt(n)
```

(iii) The 95% upper-confidence interval is given by

$$\mu \leq \bar{x} + t_{\alpha,n}\frac{s}{\sqrt{n}} = 6.68 + 1.734\frac{4.23}{\sqrt{19}} \;\Rightarrow\; \mu \leq 8.37$$

(iv) The 95% lower-confidence interval is given by

$$\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} = 6.68 - 1.734\frac{4.23}{\sqrt{19}} \leq \mu \;\Rightarrow\; 5 \leq \mu.$$

4.3.3 Confidence Interval on the Population Variance

In quality control, in most cases the objective of the auditor is not to find the mean of a population but rather to determine the level of variation of the output. For instance, they would want to know how much variation the production process exhibits about the target to see what adjustments are needed to reach a defect-free process.

When the population is normally distributed, the statistic $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$ follows a $\chi^2(n-1)$ distribution with $n-1$ degrees of freedom, and as illustrated in Figure 4.5 we can construct the the $100(1-\alpha)\%$ confidence interval for σ^2 as follows

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha,$$

where $\chi^2_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the $\chi 2$-distribution with $n-1$ degrees of freedom. It follows that

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

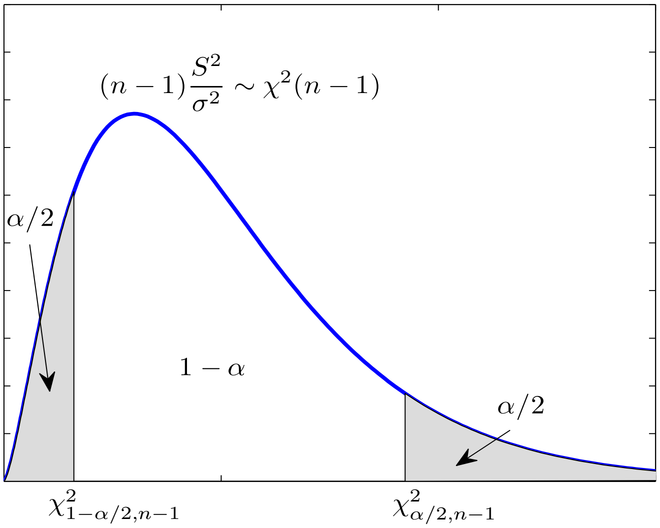


FIGURE 4.5: Illustration of confidence interval on the normal variance.

Definition 4.4 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be a random sample from a normal population with mean μ and variance σ^2.*

- *A $100(1-\alpha)\%$ confidence interval on σ^2 is given by*

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \tag{4.3.8}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on σ^2 is given by*

$$\sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \tag{4.3.9}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on σ^2 is given by*

$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \le \sigma^2 \tag{4.3.10}$$

where s^2 is the observed sample variance.

Example 4.9 *A sample of 9 screws was taken out of a production line and the sizes of the diameters in millimeters are as follows:*

```
13  13  12  12.55  12.99  12.89  12.88  12.97  12.99
```

(i) Find the 95% confidence interval to estimate the population variance.

(ii) Find the 95% upper-confidence interval to estimate the population variance.

(iii) Find the 95% lower-confidence interval to estimate the population variance.

Solution: The sample size is $n = 9$. Thus, the sample variance of the data is: $s^2 = 0.11$. For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,8} = 17.53$, $\chi^2_{1-\alpha/2,n-1} = 2.18$, $\chi^2_{\alpha,n-1} = 15.51$, and $\chi^2_{1-\alpha,n-1} = 2.73$.

(i) The 95% confidence interval is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \Rightarrow 0.05 \le \sigma^2 \le 0.41$$

Thus, we can be 95% sure that the population variance will be between 0.05 and 0.41. In other words, the probability for the population variance to be between 0.05 and 0.41 will be 0.95.

$$P(0.05 \leq \sigma^2 \leq 0.41) = 0.95.$$

MATLAB code

```
>> X = [13 13 12 12.55 12.99 12.89 12.88 12.97 12.99];
>> n = length(X); s2 = var(X); alpha = 1-0.95;
>> LC = (n-1)*s2/icdf('chi2',1-alpha/2,n-1)
>> UC = (n-1)*s2/icdf('chi2',alpha/2,n-1)
```

(ii) The 95% upper-confidence interval is given by

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \Rightarrow \sigma^2 \leq 0.33$$

(iii) The 95% lower-confidence interval is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2 \Rightarrow 0.06 \leq \sigma^2$$

Example 4.10 *The time taken by a worker in a car manufacturing company to finish a paint job on a car is normally distributed with mean μ and variance σ^2. A sample of 15 paint jobs is randomly selected and assigned to that worker, and the time taken by the worker to finish the job is jotted down. These data yield a sample standard deviation of 2.5 hours.*

(i) Find the 95% confidence interval to estimate the population standard deviation.

(ii) Find the 95% upper-confidence interval to estimate the population standard deviation.

(iii) Find the 95% lower-confidence interval to estimate the population standard deviation

Solution: From the given information, we have $n = 15$ and $s = 2.5$. For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, $\chi^2_{\alpha/2,n-1} = \chi^2_{0.025,14} = 26.1189$, $\chi^2_{1-\alpha/2,n-1} = 5.6287$, $\chi^2_{\alpha,n-1} = 23.6848$, and $\chi^2_{1-\alpha,n-1} = 6.5706$.

(i) The 95% confidence interval is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \Rightarrow 3.3501 \leq \sigma^2 \leq 15.5453$$

We can be 95% sure that the population variance will be between 3.3501 and 15.5453. Therefore, by taking the square root of the lower and upper confidence limits for σ^2, the 95% confidence interval of the population standard deviation σ is $(1.8303, 3.9427)$.

(ii) The 95% upper-confidence interval is given by

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \Rightarrow \sigma^2 \leq 13.3168$$

Thus, the 95% upper-confidence interval of the population standard deviation σ is $(0, 3.6492)$.

(iii) The 95% lower-confidence interval is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2 \Rightarrow 3.6944 \leq \sigma^2$$

Thus, the 95% lower-confidence interval of the population standard deviation σ is $(1.9221, +\infty)$.

4.3.4 Confidence Interval on the Population Proportion

Consider a population of items, each of which independently meets certain standards with some unknown probability p, and suppose that a sample of size n was taken from this population. If X denote the number of the n items that meet the standards, then $X \sim bino(n, p)$. Thus, for n large and both $np \geq 5$ and $n(1-p) \geq 5$, it follows that

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\bullet}{\sim} N(0,1)$$

where $\overset{\bullet}{\sim}$ means "is approximately distributed as." Since $\hat{p} = X/n$ is a point estimate of p, it follows that $\sqrt{n\hat{p}(1-\hat{p})} \approx \sqrt{np(1-p)}$ and

$$\frac{n\hat{p} - np}{\sqrt{n\hat{p}(1-\hat{p})}} \sim N(0,1)$$

Thus,

$$P\left(-z_{\alpha/2} \leq \frac{n\hat{p} - np}{\sqrt{n\hat{p}(1-\hat{p})}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

or, equivalently,

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right) \approx 1 - \alpha$$

Definition 4.5 *If $\hat{p}$ is the proportion of observations in a random sample of size n, then*

- *A $100(1-\alpha)\%$ confidence interval on p is given by*

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \tag{4.3.11}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on p is given by*

$$p \leq \hat{p} + z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n} \tag{4.3.12}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on p is given by*

$$\hat{p} - z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \tag{4.3.13}$$

Example 4.11 *The fraction of defective integrated circuits produced in a photolithography process is being studied. A random sample of 300 circuits is tested, revealing 13 defectives. Find a 95% two-sided confidence interval on the fraction of defective circuits produced by this particular tool.*

Solution: From the given information, we have $n = 300$ and $x = 13$. Thus, the point estimate of the proportion is $\hat{p} = x/n = 13/300 = 0.0433$. Since both $n\hat{p} = 13 \geq 5$ and $n(1-\hat{p}) = 287 \geq 5$ are satisfied, $\hat{p}$ is approximately normal. For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, and $z_{\alpha/2} = z_{0.025} = 1.96$. Thus, the 95% confidence interval is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \Rightarrow 0.02 \leq p \leq 0.07$$

We can be 95% sure that the fraction of defective circuits will be between 0.02 and 0.07. In other words, the probability for the proportion to be between 0.02 and 0.07 will be 0.95. ■

MATLAB code

```
>> n = 300; x = 13; phat = x/n; alpha = 1-0.95;
>> zalpha2 = icdf('norm',1-alpha/2,0,1);
>> LC = phat - zalpha2*sqrt(phat*(1-phat)/n)
>> UC = phat + zalpha2*sqrt(phat*(1-phat)/n)
```

Example 4.12 *A random sample of 400 computer chips is taken from a large lot of chips and 50 of them are found defective. Find a 90% two-sided confidence interval for the proportion of defective chips contained in the lot.*

Solution: From the given information, we have $n = 400$ and $x = 50$. Thus, the point estimate of the proportion of defective chips contained in the lot is $\hat{p} = x/n = 50/400 = 0.125$. Since both $n\hat{p} = 50 \geq 5$ and $n(1-\hat{p}) = 350 \geq 5$ are satisfied, $\hat{p}$ is approximately normal. For 90% confidence level, we have $\alpha = 1 - 0.9 = 0.1$, and $z_{\alpha/2} = z_{0.025} = 1.6449$. The 90% confidence interval is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \Rightarrow 0.0978 \leq p \leq 0.1522$$

We can be 90% sure that the proportion of defective chips contained in the lot will be between 0.0978 and 0.1522. ■

4.3.5 Confidence Interval on the Difference in Populations Means

Just as in the analysis of a single population, to estimate the difference between two populations the researcher would draw samples from each population.

Assume that $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ are two independent samples from two independent normal populations, as depicted in Figure 4.6. Let $\overline{X}, \overline{Y}, S_1^2$ and S_2^2 be the sample means and sample variances, respectively. Then,

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{X} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \Longrightarrow \overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

To construct the confidence interval on the difference in means $\mu_1 - \mu_2$, we consider two particular cases:

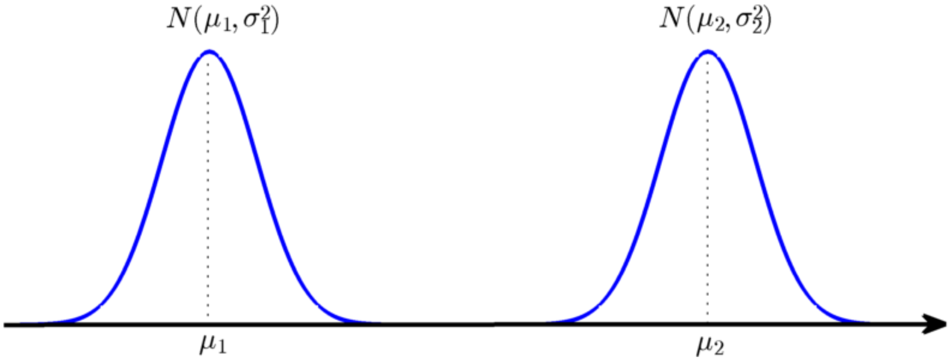


FIGURE 4.6: Illustration of two independent distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.

Case 1: If σ_1 and σ_2 are known, then the statistic

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Thus,

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

or, equivalently,

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Definition 4.6 *Let $\bar{x}$ and $\bar{y}$ be the observed sample means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variances σ_1^2 and σ_1^2, respectively. Then,*

- *A $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ is given by*

$$\overline{x} - \overline{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{x} - \overline{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{4.3.14}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on σ^2 is given by*

$$\mu_1 - \mu_2 \leq \overline{x} - \overline{y} + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{4.3.15}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on σ^2 is given by*

$$\overline{x} - \overline{y} - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \tag{4.3.16}$$

where s^2 is the observed sample variance.

Example 4.13 *The variances of two populations 1 and 2 are 16 and 9, respectively. A sample of 25 items was taken from Population 1 with a mean of 50, and a sample of 22 items was taken from Population 2 with a mean of 45. Construct a 95% two-sided confidence interval for the difference in population means.*

Solution: We have $n_1 = 25, n_2 = 22, \sigma_1 = 4$, and $\sigma_2 = 3$. For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, and $z_{\alpha/2} = z_{0.025} = 1.96$. Thus, the 95% two-sided confidence for the difference in population means is given by

$$\overline{x} - \overline{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{x} - \overline{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Rightarrow 2.99 \leq \mu_1 - \mu_2 \leq 7.01.$$

Hence, the probability for the difference in population means to be between 2.99 and 7.01 will be 0.95. ■

Case 2: If σ_1 and σ_2 are unknown and equal to σ^2, then

$$\overline{X} \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right) \quad \text{and} \quad \overline{X} \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right) \Longrightarrow \overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

and

$$(n_1 - 1)\frac{S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1) \quad \text{and} \quad (n_2 - 1)\frac{S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

$$\Longrightarrow (n_1 - 1)\frac{S_1^2}{\sigma^2} + (n_2 - 1)\frac{S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

or, equivalently,

$$(n_1 + n_2 - 2)\frac{S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

where $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is called the pooled variance.

Since $Z = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim N(0,1)$ and $V = (n_1 + n_2 - 2)\dfrac{S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$, it follows that

$$\frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

Therefore,

$$P\left(-t_{\alpha/2,n_1+n_2-2} \leq \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}} \leq t_{\alpha/2,n_1+n_2-2}\right) = 1-\alpha$$

Definition 4.7 *Let $\bar{x}$ and $\bar{y}$ be the observed sample means of independent random samples of sizes n_1 and n_2 from two independent normal populations with unknown variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, respectively. Then,*

- *A $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ is given by*

$$\overline{x}-\overline{y}-t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2 \leq \overline{x}-\overline{y}+t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4.3.17}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on σ^2 is given by*

$$\mu_1-\mu_2 \leq \overline{x}-\overline{y}+t_{\alpha,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{4.3.18}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on σ^2 is given by*

$$\overline{x}-\overline{y}-t_{\alpha,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2 \tag{4.3.19}$$

where s_p is the observed pooled standard deviation.

Example 4.14 *The variances of two populations are assumed to be equal. A sample of 15 items was taken from Population I with a mean of 50 and a standard deviation of 3, and a sample of 19 items was taken from Population II with a mean of 47 and a standard deviation of 2.*

(i) Calculate the pooled sample variance.

(ii) Construct a 95% two-sided confidence for the difference between the two population means.

Solution: From the given information, we have $n_1 = 15, n_2 = 19, \bar{x} = 50, \bar{y} = 47, S_1 = 3$, and $S_2 = 2$.

(i) The value of the pooled sample variance is given by

$$s_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{(15-1)(3^2)+(19-1)(2^2)}{19+15-2} = 6.19.$$

(ii) For 95% confidence level, we have $\alpha = 1-0.95 = 0.05$, and $t_{\alpha/2,n_1+n_2-2} = t_{0.025,32} = 2.04$. Thus, the 95% two-sided confidence for the difference between the two population means is given by

$$\overline{x}-\overline{y}-t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2 \leq \overline{x}-\overline{y}+t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$1.25 \leq \mu_1-\mu_2 \leq 4.75.$$

Hence, a 95% two-sided confidence for $\mu_1-\mu_2$ is $(1.25, 4.75)$. ∎

MATLAB code

```
>> n1 = 15; n2 = 19; alpha = 0.05;
>> talpha2 = icdf('t',1-alpha/2,n1+n2-2)
```

Example 4.15 *A pharmaceutical company sets two machines to fill 15oz bottles with cough syrup. Two random samples of $n_1 = 16$ bottles from machine 1 and $n_2 = 12$ bottles from machine 2 are selected. The two samples yield the following sample statistics:*

$$\overline{X} = 15.24 \qquad S_1^2 = 0.64$$

$$\overline{Y} = 14.96 \qquad S_2^2 = 0.36$$

(i) Calculate the pooled sample variance.

(ii) Construct a 95% two-sided confidence for the mean difference of the amount of cough syrup filled in bottles by the two machines.

Solution: From the given information, we have $n_1 = 16, n_2 = 12, \bar{x} = 15.24, \bar{y} = 14.96, S_1^2 = 0.64$, and $S_2^2 = 0.36$.

(i) The value of the pooled sample variance is given by

$$s_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(16-1)(0.64) + (12-1)(0.36)}{16 + 12 - 2} = 0.5215$$

(ii) For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, and $t_{\alpha/2, n_1+n_2-2} = t_{0.025,26} = 2.0555$. Thus, the 95% two-sided confidence for the difference between the two population means is given by

$$\overline{x} - \overline{y} - t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \overline{x} - \overline{y} + t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$-0.2869 \le \mu_1 - \mu_2 \le 0.8469.$$

Hence, a 95% two-sided confidence for $\mu_1 - \mu_2$ is $(-0.2869, 0.8469)$. ■

4.3.6 Confidence Interval on the Ratio of Variances

Assume that $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ are two independent samples from two independent normal populations. Let S_1^2 and S_2^2 be the sample variances, respectively. Then the statistic

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Denote by $f_{\alpha/2,n_1-1,n_2-1}$ and $f_{1-\alpha/2,n_1-1,n_2-1}$ the upper and lower $\alpha/2$ percentage points of the $F(n_1 - 1, n_2 - 1)$ distribution, as shown in Figure 4.7.

Thus,

$$P\left(f_{1-\alpha/2,n_1-1,n_2-1} \le \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \le f_{\alpha/2,n_1-1,n_2-1}\right) = 1 - \alpha$$

or, equivalently,

$$P\left(f_{1-\alpha/2,n_1-1,n_2-1}\frac{S_1^2}{S_2^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le f_{\alpha/2,n_1-1,n_2-1}\frac{S_1^2}{S_2^2}\right) = 1 - \alpha$$

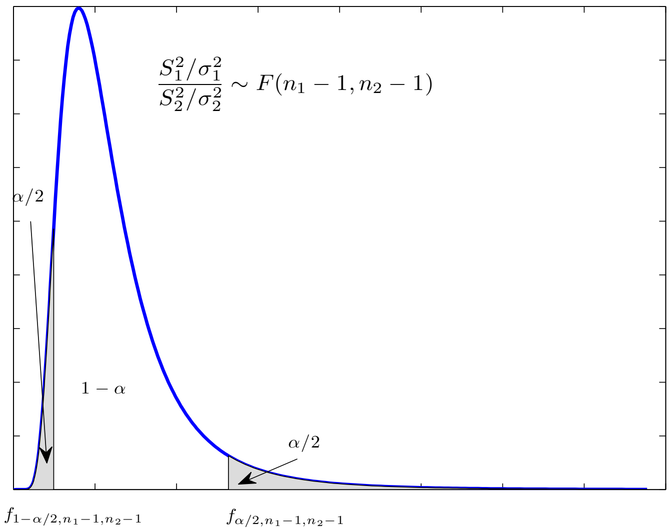


FIGURE 4.7: Illustration of confidence interval on the ratio of variances.

Definition 4.8 *Let s_1^2 and s_2^2 be the observed sample variances of two independent random samples of sizes n_1 and n_2 from two independent normal populations with unknown variances σ_1^2 and σ_2^2, respectively. Then,*

- *A $100(1-\alpha)\%$ confidence interval on σ_1^2/σ_2^2 is given by*

$$f_{1-\alpha/2,n_1-1,n_2-1}\frac{s_1^2}{s_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq f_{\alpha/2,n_1-1,n_2-1}\frac{s_1^2}{s_2^2} \tag{4.3.20}$$

- *A $100(1-\alpha)\%$ upper-confidence interval on σ_1^2/σ_2^2 is given by*

$$\frac{\sigma_1^2}{\sigma_2^2} \leq f_{\alpha,n_1-1,n_2-1}\frac{s_1^2}{s_2^2} \tag{4.3.21}$$

- *A $100(1-\alpha)\%$ lower-confidence interval on σ_1^2/σ_2^2 is given by*

$$f_{1-\alpha,n_1-1,n_2-1}\frac{s_1^2}{s_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \tag{4.3.22}$$

Example 4.16 *The variances of two populations are assumed to be equal. A sample of 15 items was taken from Population I with a standard deviation of 3, and a sample of 19 items was taken from Population II with a standard deviation of 2. Construct a 95% two-sided confidence for the ratio of variances.*

Solution: We have $n_1 = 15, n_2 = 19, S_1 = 3$, and $S_2 = 2$. For 95% confidence level, we have $\alpha = 1 - 0.95 = 0.05$, $f_{\alpha/2,n_1-1,n_2-1} = f_{0.025,14,18} = 2.70$, and $f_{1-\alpha/2,n_1-1,n_2-1} = 0.35$. Thus, the 95% two-sided confidence for the ratio of variances is given by

$$f_{1-\alpha/2,n_1-1,n_2-1}\frac{s_1^2}{s_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq f_{\alpha/2,n_1-1,n_2-1}\frac{s_1^2}{s_2^2} \Rightarrow 0.78 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 6.07.$$

MATLAB code

```
>> n1 = 15; n2 = 19; alpha = 0.05;
>> falpha1 = icdf('f',1-alpha/2,n1-1,n2-1) %f_{alpha/2,n1-1,n2-1}
>> falpha2 = icdf('f',alpha/2,n1-1,n2-1) %f_{1-alpha/2,n1-1,n2-1}
```

4.4 Hypothesis Testing

A **statistical hypothesis** is a statement or claim about a set of parameters of one or more populations. It is called a hypothesis because it is not known whether or not it is true. A **Hypothesis test** is the decision-making procedure about the hypothesis. In statistics terms, the hypothesis that we try to establish is called an **alternative hypothesis** H_1, while its contradiction is called a **null hypothesis** H_0.

Consider a population with unknown parameter θ. Basically, there are three ways to set up the null and alternatives hypothesis:

1. **Two-tailed** hypothesis test

$$\begin{aligned} H_0: &\quad \theta = \theta_0 \\ H_1: &\quad \theta \neq \theta_0 \end{aligned} \tag{4.4.1}$$

2. **Upper-tailed** hypothesis test

$$\begin{aligned} H_0: &\quad \theta = \theta_0 \\ H_1: &\quad \theta > \theta_0 \end{aligned} \tag{4.4.2}$$

3. **Lower-tailed** hypothesis test

$$\begin{aligned} H_0: &\quad \theta = \theta_0 \\ H_1: &\quad \theta < \theta_0 \end{aligned} \tag{4.4.3}$$

The two-tailed test is also called two-sided test, whereas the upper-and lower-tests are also referred to as one-sided tests.

Example 4.17 *A manufacturer of a certain brand of rice cereal claims that the average saturated fat content does not exceed 1.5 grams per serving. State the null and alternative hypotheses to be used in testing this claim.*

Solution: The manufacturers claim should be rejected only if μ is greater than 1.5 milligrams and should not be rejected if μ is less than or equal to 1.5 milligrams. Thus, we test

$$\begin{aligned} H_0 &: \quad \mu = 1.5 \\ H_1 &: \quad \mu > 1.5. \end{aligned}$$

Example 4.18 *A real estate agent claims that 60% of all private residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test.*

Solution: If the test statistic were substantially higher or lower than $p = 0.6$, we would reject the agents claim. Hence, we should make the hypotheses:

$$\begin{aligned} H_0 &: \quad p = 0.6 \\ H_1 &: \quad p \neq 0.6 \end{aligned}$$

The alternative hypothesis implies a two-tailed test with the critical region divided equally in both tails of the distribution of the test statistic. ■

The goal of any hypothesis test is to make a decision; in particular, we will decide whether to reject the null hypothesis in favor of the alternative hypothesis H_1. Although we would like to be able to always make a correct decision, we must remember that the decision will be based on the sample data. When a test is done, there are four possible outcomes, as summarized in Table 4.1.

	H_0 is true	H_0 is false
Reject H_0	**Type I Error**	Correct Decision
Do not reject H_0	Correct Decision	**Type II Error**

TABLE 4.1: Possible outcomes for a hypothesis test.

From Table 4.1, we can observe that there are two ways of making a mistake when doing a hypothesis test. Thus, we may make one of the following two types of errors:

Type I error: is the error of rejecting H_0 when it is true. The probability of making a Type I error, denote by α, is given by

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true})$$

Type II error: is the error of accepting H_0 when it is false. The probability of making a Type II error, denote by β, is given by

$$\beta = P(\text{Type II error}) = P(\text{Accept } H_0 \text{ when } H_0 \text{ is false})$$

The Type I error and Type II error are related. A decrease in the probability of one generally results in an increase in the probability of the other. The probability, α, is also called the **significance level** for the hypothesis test. If the significance level is fixed, then the rejection of H_0 is done with a fixed degree of confidence in the decision. Because we specify the level of significance before performing the hypothesis test, we basically control the risk of making a Type I error. Typical values for α are 0.1, 0.05, and 0.01. For example, if $\alpha = 0.1$ for a test, and the null hypothesis is rejected, then one will be 90% certain that this is the correct decision.

After the hypotheses are stated, the next step is to design the study. An appropriate statistical test will be selected, the level of significance will be chosen, and a plan to conduct the study will be formulated. To make an inference for the study, the statistical test and level of significance are used. Once the level of significance is selected, a critical value for the appropriate test is selected from a table in the Appendix.

A hypothesis testing procedure consists of four main steps:

Step 1: Specify the null and alternative hypotheses, H_0 and H_1, and the significance level, α

Step 2: Determine an appropriate test statistic and compute its value using the sample data

Step 3: Specify the rejection region

Step 4: Make the appropriate conclusion by deciding whether H_0 should be rejected.

4.4.1 Tests on the Mean of a Normal Population with Known Variance

Let $X_1, \ldots, X_n$ be a sample of size n from a $N(\mu, \sigma^2)$ population with unknown mean μ and known variance σ^2. In testing the population mean μ, there are three ways to structure the hypothesis test:

Two-tailed		Upper-Tailed		Lower-Tailed	
$H_0 :$	$\mu = \mu_0$	$H_0 :$	$\mu = \mu_0$	$H_0 :$	$\mu = \mu_0$
$H_1 :$	$\mu \neq \mu_0$	$H_1 :$	$\mu > \mu_0$	$H_1 :$	$\mu < \mu_0$

Recall that the sample mean $\overline{X} \sim N(\mu, \sigma^2/n)$. Under the assumption that the null hypothesis is true (i.e. $H_0 : \mu = \mu_0$), it follows that the test statistic

$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

has a standard normal distribution. Thus, this hypothesis test is called z**-test**.

A **critical region** or rejection region is the set of all values such that the null hypothesis is rejected. We can then determine a critical region based on the computed test statistic.

Let z_0 be the numerical value, calculated from the sample, of the test statistic Z_0. Then, for a selected significance level α, the critical regions are (see Figure 4.8).

Two-tailed	Upper-Tailed	Lower-Tailed
If $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$ reject H_0	If $z_0 > z_\alpha$ reject H_0	If $z_0 < -z_\alpha$ reject H_0

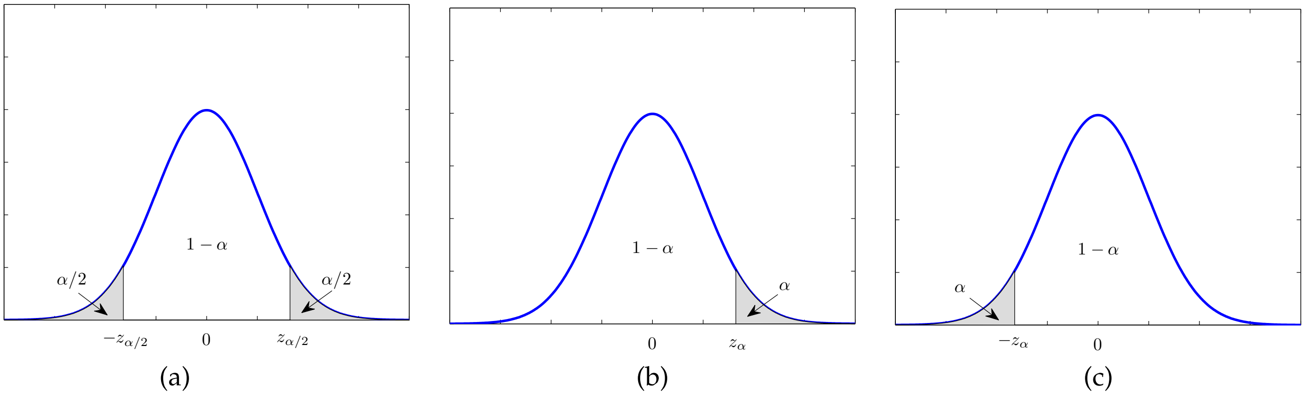


FIGURE 4.8: Critical region for the z-test alternative hypothesis: (a) $\mu \neq \mu_0$; (b) $\mu > \mu_0$; (c) $\mu < \mu_0$.

Usually, α is specified in advance before any samples are drawn so that results will not influence the choice for the level of significance. To conclude a statistical test, we compare our α-value with the ***p*-value**, which is the probability of observing the given sample result under the assumption that the null hypothesis is true. The p-value is computed using sample data and the sampling distribution. If the p-value is less than the significance level α, then we reject the null hypothesis. For example, if $\alpha = 0.05$ and the p-value is 0.03, then we reject the null hypothesis. The converse is not true. If the p-value is greater than α, then we have insufficient evidence to reject the null hypothesis.

- If p-value $\leq \alpha$, we reject the null hypothesis and say the data are statistically significant at the level α.
- If p-value $> \alpha$, we do not reject the null hypothesis.

Denote by $\Phi(\cdot)$ the cdf of the $N(0,1)$ distribution. Then, the z-test may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$

Test Statistic (z-test): $Z_0 = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$\lvert z_0 \rvert > z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

***p*-values:**

Two-tailed	Upper-Tailed	Lower-Tailed
$2[1 - \Phi(\lvert z_0 \rvert)]$	$1 - \Phi(z_0)$	$\Phi(z_0)$

Example 4.19 *The CEO of a large financial corporation claims that the average distance that commuting employees travel to work is 32 km. The commuting employees feel otherwise. A sample of 64 employees was randomly selected and yielded a mean of 35 km. Assuming a population standard deviation of 5 km,*

(i) Test the CEO's claim at the 5% level of significance.

(ii) Calculate the p-value for this test.

(iii) Test the CEO's claim using a confidence interval with a 95% confidence coefficient.

Solution: From the given information, we have $n = 64$, $\bar{x} = 35$, $\mu_0 = 32$, $\sigma = 5$, and $\alpha = 0.05$.

(i) **Step 1:** This is a two-tailed test, since the employees feel that the CEO's claim is not correct, but whether they feel that the average distance is less than 32 km or more than 32 km is not specified. Thus,

$$\begin{aligned} H_0 &: \mu = 32 \\ H_1 &: \mu \neq 32 \end{aligned}$$

Step 2: Since σ is known, the appropriate test statistic is the z-test and its value is given by

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{35 - 32}{5/\sqrt{64}} = 4.8$$

Step 3: The rejection region is $|z_0| > z_{\alpha/2}$, where $z_{\alpha/2} = z_{0.025} = 1.96$.

Step 4: Since $|z_0| = 4.8 > z_{\alpha/2} = 1.96$, we reject the null hypothesis H_0. There is sufficient sample evidence to refute the CEO's claim. The sample evidence supports the employees' claim that the average distance commuting employees travel to work is not equal to 32 km at the 5 percent level of significance. That is, there is a significant difference between the sample mean and the postulated value of the population mean of 32 km.

(ii) The p-value is equal to $2[1 - \Phi(|z_0|)] = 2[1 - \Phi(4.8)] = 0$, which is less than the significance level $\alpha = 0.05$; therefore we reject the null hypothesis.

(iii) A confidence interval with a 95% confidence coefficient implies $\alpha = 0.05$. The 95% confidence interval is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$35 - 1.96\frac{5}{\sqrt{64}} \leq \mu \leq 35 + 1.96\frac{5}{\sqrt{64}}$$

$$33.775 \leq \mu \leq 36.225$$

This interval clearly does not contain 32, the value of μ under the null hypothesis. Thus, we reject the null hypothesis. ■

Example 4.20 *A random sample of 36 pieces of copper wire produced in a plant of a wire manufacturing company yields the mean tensile strength of 950 psi. Suppose that population of tensile strengths of all copper wires produced in that plant are distributed with mean μ and standard deviation $\sigma = 120$ psi. Test the statistical hypothesis:*

$$H_0 : \mu = 980 \quad \textit{versus} \quad H_1 : \mu < 980$$

at the 99% level of confidence. Then, calculate the p-value.

Solution: From the given information, we have $n = 36$, $\bar{x} = 950$, $\mu_0 = 980$, $\sigma = 120$, and $\alpha = 0.01$.

Step 1: This is a lower-tailed test. Thus,

$$\begin{aligned} H_0 &: \mu = 980 \\ H_1 &: \mu < 980 \end{aligned}$$

Step 2: Since σ is known, the appropriate test statistic is the z-test and its value is given by

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{950 - 980}{120/\sqrt{36}} = -1.5$$

Step 3: The rejection region is $z_0 < -z_\alpha$, where $z_\alpha = z_{0.01} = 2.3263$.

Step 4: Since $z_0 = -1.5 \nless -z_\alpha = -2.3263$, we do not reject the null hypothesis H_0.

The p-value is equal to $\Phi(z_0) = \Phi(-1.5) = 1 - \Phi(1.5) = 0.0668$, which is greater than the significance level $\alpha = 0.01$; therefore, there is insufficient evidence to reject the null hypothesis. ■

4.4.2 Tests on the Mean of a Normal Population with Unknown Variance

Let $X_1, \ldots, X_n$ be a sample of size n from a $N(\mu, \sigma^2)$ population with unknown mean μ and unknown variance σ^2. In testing the population mean μ, there are three ways to structure the hypothesis test:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$

Under the assumption that the null hypothesis is true (i.e. $H_0 : \mu = \mu_0$), it follows that the test statistic

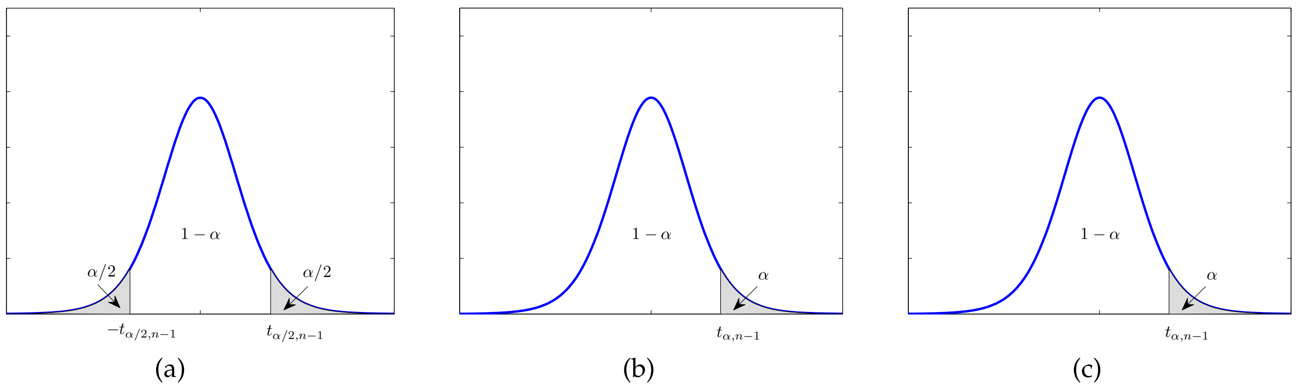
$$T_0 = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

has a t-distribution with $n - 1$ degrees of freedom. Thus, this hypothesis test is called ***t*-test**. When the population standard deviation σ is not known, we typically use the t-test either (i) when the sample size is large (i.e. $n \geq 30$) or (ii) when the sample size is small (i.e. $n < 30$) and the population from which the sample is selected is approximately normal.

A **critical region** or rejection region is the set of all values such that the null hypothesis is rejected. Let t_0 be the numerical value, calculated from the sample, of the test statistic T_0. Then, for a selected significance level α, the critical regions are (see Figure 4.9):

Two-tailed	Upper-Tailed	Lower-Tailed
If $t_0 < -t_{\alpha/2,n-1}$ or $t_0 > t_{\alpha/2,n-1}$	If $t_0 > t_{\alpha,n-1}$	If $t_0 < -t_{\alpha,n-1}$
reject H_0	reject H_0	reject H_0

Denote by $F(\cdot)$ the cdf of the $t(n-1)$ distribution. Then, the t-test may be summarized as follows:



(a) (b) (c)

FIGURE 4.9: Critical region for the t-test alternative hypothesis: (a) $\mu \neq \mu_0$; (b) $\mu > \mu_0$; (c) $\mu < \mu_0$.

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$

Test Statistic (t-test): $T_0 = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$\lvert t_0 \rvert > t_{\alpha/2,n-1}$	$t_0 > t_{\alpha,n-1}$	$t_0 < -t_{\alpha,n-1}$

***p*-values:**

Two-tailed	Upper-Tailed	Lower-Tailed
$2[1 - F(\lvert t_0 \rvert)]$	$1 - F(t_0)$	$F(t_0)$

Example 4.21 *The Atlas Electric Institute has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Then, calculate the p-value. Assume the population of kilowatt hours to be normal.*

Solution: From the information given, we have $n = 12$, $\bar{x} = 42$, $\mu_0 = 46$, $s = 11.9$, and $\alpha = 0.05$.

Step 1: The hypotheses are:

$$\begin{aligned} H_0 &: \mu = 46 \\ H_1 &: \mu < 46 \end{aligned}$$

Step 2: Since σ is unknown and the population of kilowatt hours is assumed to be normally distributed, the appropriate test statistic is the t-test and its value is given by

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{42 - 46}{11.9/\sqrt{12}} = -1.1644$$

Step 3: The rejection region is $t_0 < -t_{\alpha,n-1}$, where $t_{\alpha,n-1} = t_{0.05,11} = 1.7959$.

Step 4: Since $t_0 = -1.16 \nless -t_{\alpha,n-1} = -1.7959$, we do not reject the null hypothesis H_0. We conclude that the average number of kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

The p-value is equal to $F(t_0) = F(-1.1644) = 0.1344$, which is greater than the significance level $\alpha = 0.05$; therefore we have insufficient evidence to reject the null hypothesis. ■

Example 4.22 *Grand Auto Corporation produces auto batteries. The company claims that its top-of-the-line Never Die batteries are good, on average, for at least 65 months. A consumer protection agency tested 45 such batteries to check this claim. It found that the mean life of these 45 batteries is 63.4 months and the standard deviation is 3 months. Using 1% significance level, can you conclude that the company's claim is true?. Then, calculate the p-value. Assume the population of life of batteries to be normal.*

Solution: From the given information, we have $n = 45$, $\bar{x} = 63.4$, $\mu_0 = 65$, $s = 3$, and $\alpha = 0.01$.

Step 1: This is a lower-tailed test:

$$\begin{aligned} H_0 &: \mu \geq 65 \qquad \text{(The mean life of batteries is at least 65 months)} \\ H_1 &: \mu < 65 \qquad \text{(The mean life of batteries is less than 65 months)} \end{aligned}$$

Step 2: Since σ is unknown and $n \geq 30$, the appropriate test statistic is the t-test and its value is given by

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{63.4 - 65}{3/\sqrt{45}} = -3.5777$$

Step 3: The rejection region is $t_0 < -t_{\alpha,n-1}$, where $t_{\alpha,n-1} = t_{0.01,44} = 2.4141$.

Step 4: Since $t_0 = -3.5777 < -t_{\alpha,n-1} = -2.4141$, we reject the null hypothesis H_0. We conclude that the mean life of such batteries is less than 65 months.

The p-value is equal to $F(t_0) = F(-3.5777) = 0$, which is less than the significance level $\alpha = 0.01$; therefore we reject the null hypothesis. ■

Example 4.23 *A tool assembling company believes that a worker should take no more than 30 minutes to assemble a particular tool. A sample of 16 workers who assembled that tool showed that the average time was 33 minutes with a sample standard deviation of 6 minutes. Test at the 5% level od significance if the data provide sufficient evidence to indicate the validity of the company's belief. Then, calculate the p-value. Assume that the assembly times are normally distributed.*

Solution: From the information given, we have $n = 16$, $\bar{x} = 33$, $\mu_0 = 30$, $s = 6$, and $\alpha = 0.05$.

Step 1: This is a upper-tailed test:

$$\begin{aligned} H_0 &: \mu \leq 30 \qquad \text{(The mean assembly time is no more than 30 minutes)} \\ H_1 &: \mu > 30 \qquad \text{(The mean assembly time is more than 30 minutes)} \end{aligned}$$

Step 2: Since σ is unknown and the assembly times are assumed to be normally distributed, the appropriate test statistic is the t-test and its value is given by

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{33 - 30}{6/\sqrt{16}} = 2$$

Step 3: The rejection region is $t_0 > t_{\alpha,n-1}$, where $t_{\alpha,n-1} = t_{0.05,15} = 1.7531$.

Step 4: Since $t_0 = 2 > t_{\alpha,n-1} = 1.7531$, we reject the null hypothesis H_0. We conclude that the mean assembly time is more than 30 minutes.

The p-value is equal to $F(t_0) = 1 - F(2) = 0.032$, which is less than the significance level $\alpha = 0.05$; therefore we reject the null hypothesis. ■

Example 4.24 *A city health department wishes to determine if the mean bacteria count per unit volume of water at a lake beach is within the safety level of 200. A researcher collected 10 water samples of unit volume and found the bacteria counts to be*

```
175   190   205   193   184   207   204   193   196   180
```

(i) Check the assumption of normality for the bacteria counts.

(ii) Do the data strongly indicate that there is no cause for concern at 5% level of significance?

Solution: From the given information, we have $n = 10$, $\bar{x} = 192.7$, $s = 10.812$, $\mu_0 = 200$, and $\alpha = 0.05$.

(i) Because the sample size is small, we must be willing to assume that the population distribution of bacteria counts is normally distributed. As shown in Figure 4.10, the normal probability plot and boxplot indicate that the measurements constitute a sample from a normal population. The normal probability plot appears to be reasonably straight. Although the boxplot is not perfectly symmetric, it is not too skewed and there are no outliers.

MATLAB code

```
>> X = [175 190 205 193 184 207 204 193 196 180];
>> subplot(1,2,1); normplot(X); subplot(1,2,2); boxplot(X);
```

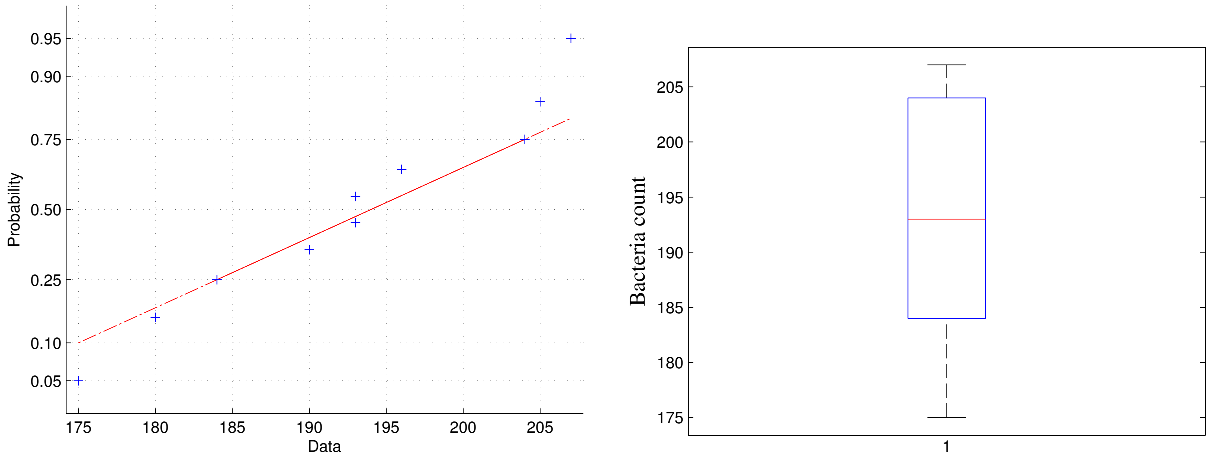


FIGURE 4.10: Normal probability and box plots for the bacteria counts.

(ii) **Step 1:** Let μ denote the current (population) mean bacteria count per unit volume of water. Then, the statement "no cause for concern" translates to $\mu < 200$, and the researcher is seeking strong evidence in support of this hypothesis. So the formulation of the null and alternative hypotheses should be

$$\begin{aligned} H_0 &: \mu = 200 \\ H_1 &: \mu < 200 \end{aligned}$$

Step 2: Since σ is unknown, the appropriate test statistic is the t-test and its value is given by

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{192.7 - 200}{10.812/\sqrt{10}} = -2.1351$$

Step 3: The rejection region is $t_0 < -t_{\alpha,n-1}$, where $t_{\alpha,n-1} = t_{0.05,9} = 1.8331$.

Step 4: Since $t_0 = -2.1351 < -t_{\alpha,n-1} = -1.8331$, we reject the null hypothesis H_0. On the basis of the data obtained from these 10 measurements, there does seem to be strong evidence that the true mean is within the safety level.

The p-value is equal to $F(t_0) = F(-2.1351) = 0.0308$, which is less than the significance level $\alpha = 0.05$; therefore we reject the null hypothesis. There is strong evidence that the mean bacteria count is within the safety level. ■

MATLAB code

```
>> X = [175 190 205 193 184 207 204 193 196 180];
>> mu0 = 200; alpha = 0.05;
>> [h,p,ci,stats]=ttest(X,mu0,alpha,'left')
```

4.4.3 Tests on the Variance of a Normal Population

Let $X_1, \dots, X_n$ be a sample of size n from a $N(\mu, \sigma^2)$ population with unknown variance σ^2. In testing the population variance σ^2, there are three ways to structure the hypothesis test:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$

Under the assumption that the null hypothesis is true (i.e. $H_0 : \sigma^2 = \sigma_0^2$), it follows that the test statistic

$$X_0^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

has a χ^2-distribution with $n-1$ degrees of freedom. Thus, this hypothesis test is called **χ^2-test**.

Let χ_0^2 be the numerical value, calculated from the sample, of the test statistic X_0^2. Then, for a selected significance level α, the critical regions are:

Two-tailed	Upper-Tailed	Lower-Tailed
If $\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$ or $\chi_0^2 > \chi^2_{\alpha/2,n-1}$	If $\chi_0^2 > \chi^2_{\alpha,n-1}$	If $\chi_0^2 < \chi^2_{1-\alpha,n-1}$
reject H_0	reject H_0	reject H_0

Then, the χ^2-test may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$

Test Statistic: $\chi_0^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$ or $\chi_0^2 > \chi^2_{\alpha/2,n-1}$	$\chi_0^2 > \chi^2_{\alpha,n-1}$	$\chi_0^2 < \chi^2_{1-\alpha,n-1}$

***p*-values:**

Two-tailed	Upper-Tailed	Lower-Tailed
$2\min(P(\chi^2_{\alpha,n-1} < \chi_0^2), 1 - P(\chi^2_{\alpha,n-1} < \chi_0^2))$	$P(\chi_0^2 > \chi^2_{\alpha,n-1})$	$P(\chi_0^2 < \chi^2_{1-\alpha,n-1})$

Example 4.25 *A manufacturer of car batteries claims that the life of the companys batteries is approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Then, calculate the p-value. Use a 0.05 level of significance.*

Solution: From the given information and data, we have $n = 10$, $s^2 = (1.2)^2 = 1.44$, $\sigma_0 = 0.9$, and $\alpha = 0.05$.

Step 1: This is an upper-tailed test, that is

$$\begin{aligned} H_0 &: \sigma^2 = 0.81 \\ H_1 &: \sigma^2 > 0.81 \end{aligned}$$

Step 2: The appropriate test statistic is the χ^2-test and its value is given by

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(10-1)(1.2)^2}{(0.9)^2} = 16$$

Step 3: The rejection region is $\chi_0^2 > \chi^2_{\alpha,n-1}$, where $\chi^2_{\alpha,n-1} = \chi^2_{0.05,9} = 16.919$.

Step 4: Since $\chi_0^2 = 16 > \chi^2_{\alpha,n-1} = 16.919$, we do not reject the null hypothesis H_0.

The p-value is equal to $1 - F(\chi_0^2) = 1 - F(16) = 0.0669$, where F is the cdf of the χ^2-distribution with 9 degrees of freedom. Since the p-value is greater than the significance level $\alpha = 0.05$, we have insufficient evidence to reject the null hypothesis. ■

Example 4.26 *The production manager of a light bulb manufacturer believes that the lifespan of the* 14*W bulb with light output of 800 lumens is 6000 hours. A random sample of 25 bulbs produced the sample mean of 6180 hours and sample standard deviation of 178 hours. Test at the 5% level of significance that the population standard deviation is less than 200 hours. Then, calculate the p-value for the test. Assume that the lifespan of these bulbs is normally distributed.*

Solution: From the given information and data, we have $n = 25$, $s = 178$, $\sigma_0 = 200$, and $\alpha = 0.05$. The lifespan of the the light bulbs is assumed to be normally distributed.

Step 1: This is a lower-tailed test, that is

$$\begin{aligned} H_0 &: \sigma^2 = (200)^2 \\ H_1 &: \sigma^2 < (200)^2 \end{aligned}$$

or equivalently

$$\begin{aligned} H_0 &: \sigma = 200 \\ H_1 &: \sigma < 200 \end{aligned}$$

Step 2: The appropriate test statistic is the χ^2-test and its value is given by

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(25-1)(178)^2}{(200)^2} = 19.0104$$

Step 3: The rejection region is $\chi_0^2 < \chi^2_{1-\alpha,n-1}$, where $\chi^2_{1-\alpha,n-1} = \chi^2_{1-0.05,24} = 13.8484$.

Step 4: Since $\chi_0^2 = 19.0104 \not< \chi^2_{1-\alpha,n-1} = 13.8484$, we do not reject the null hypothesis H_0. The supervisor can conclude that the standard deviation of the lifespan of a light bulb is 200.

The p-value is equal to $F(\chi_0^2) = F(19.0104) = 0.2486$, which is less than the significance level $\alpha = 0.05$; therefore, there is insufficient evidence to reject the null hypothesis. ■

4.4.4 Tests on a Proportion

Consider a population of items, each of which independently meets certain standards with some unknown probability p, and suppose that a sample of size n was taken from this population. In testing the proportion p, there are three ways to structure the hypothesis test:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
$H_1 : p \neq p_0$	$H_1 : p > p_0$	$H_1 : p < p_0$

If X denote the number of the n items that meet the standards, then $X \sim bino(n,p)$. Thus, for n large and under the assumption that the null hypothesis is true (i.e. $H_0 : p = p_0$), it follows that the test statistic

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$$

has a approximate standard normal distribution.

Let z_0 be the numerical value, calculated from the sample, of the test statistic Z_0. Then, for a selected significance level α, the critical regions are (see Figure)

Two-tailed	Upper-Tailed	Lower-Tailed
If $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$	If $z_0 > z_\alpha$	If $z_0 < -z_\alpha$
reject H_0	reject H_0	reject H_0

Denote by $\Phi(\cdot)$ the cdf of the $N(0,1)$ distribution. Then, the proportion test may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
$H_1 : p \neq p_0$	$H_1 : p > p_0$	$H_1 : p < p_0$

Test Statistic: $Z_0 = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$\lvert z_0\rvert > z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

***p*-values:**

Two-tailed	Upper-Tailed	Lower-Tailed
$2[1-\Phi(\lvert z_0\rvert)]$	$1-\Phi(z_0)$	$\Phi(z_0)$

Example 4.27 *A builder claims that heat pumps are installed in 70% of all homes being constructed today in the city of Granada, Spain. Would you agree with this claim if a random survey of new homes in this city showed that 8 out of 15 had heat pumps installed? Then, calculate the p-value. Use a 0.10 level of significance.*

Solution: From the given information, we have $n = 15$, $p_0 = 0.7$, $x = 8$, and $\alpha = 0.10$.

Step 1: This is a two-tailed test on proportion:

$$\begin{aligned} H_0 &: \quad p = 0.7 \\ H_1 &: \quad p \neq 0.7 \end{aligned}$$

Step 2: The appropriate test statistic is the z-test and its value is given by

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{8 - (15)(0.7)}{\sqrt{(15)(0.7)(1-0.7)}} = -1.4086$$

Step 3: The rejection region is $|z_0| > z_{\alpha/2}$, where $z_{\alpha/2} = z_{0.05} = 1.6449$.

Step 4: Since $|z_0| = 1.4086 \not> z_{\alpha/2} = 1.6449$, we do not reject the null hypothesis H_0. We conclude that there is insufficient reason to doubt the builders claim.

The p-value is equal to $2[1-\Phi(|z_0|)] = 2[1-\Phi(1.4086)] = 0.1590$, which is greater than the significance level $\alpha = 0.05$; therefore we have insufficient evidence to reject the null hypothesis. ■

Example 4.28 *A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Then, calculate the p-value. Use a 0.05 level of significance.*

Solution: From the given information, we have $n = 100$, $p_0 = 0.6$, $x = 70$, and $\alpha = 0.05$.

Step 1: This is an upper-tailed test on proportion:

$$\begin{aligned} H_0 &: p = 0.6 \\ H_1 &: p > 0.6 \end{aligned}$$

Step 2: The appropriate test statistic is the z-test and its value is given by

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{70 - (100)(0.6)}{\sqrt{(100)(0.6)(1-0.6)}} = 2.0412$$

Step 3: The rejection region is $z_0 > z_\alpha$, where $z_\alpha = z_{0.05} = 1.6449$.

Step 4: Since $z_0 = 2.0412 > z_\alpha = 1.6449$, we reject the null hypothesis H_0. We conclude that the new drug is superior.

The p-value is equal to $1 - \Phi(z_0) = 1 - \Phi(2.0412) = 0.0206$, which is smaller than the significance level $\alpha = 0.05$; therefore we reject the null hypothesis. ■

Example 4.29 *Direct Mailing Company sells computers and computer parts by mail. The company claims that at least 90% of all orders are mailed within 72 hours after they are received. The quality control department at the company often takes samples to check if this claim is valid. A recently taken sample of 150 orders showed that 129 of them were mailed within 72 hours. Using 2.5% significance level, do you think the company's claim is true?*

Solution: From the given information, we have $n = 150$, $p_0 = 0.9$, $x = 129$, and $\alpha = 0.025$.

Step 1: This is an upper-tailed test on proportion:

$$\begin{aligned} H_0 &: p \geq 0.9 \qquad \text{(The company's claim is true)} \\ H_1 &: p < 0.9 \qquad \text{(The company's claim is false)} \end{aligned}$$

Step 2: The appropriate test statistic is the z-test and its value is given by

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{129 - (150)(0.9)}{\sqrt{(150)(0.9)(1-0.9)}} = -1.6330$$

Step 3: The rejection region is $z_0 < -z_\alpha$, where $z_\alpha = z_{0.025} = 1.96$.

Step 4: Since $z_0 = -1.6330 \not< -z_\alpha = -1.96$, we do not reject the null hypothesis H_0. We conclude that the company's claim is true.

The p-value is equal to $\Phi(z_0) = \Phi(-1.6330) = 1 - \Phi(1.6330) = 0.0512$, which is greater than the significance level $\alpha = 0.025$; therefore we have insufficient evidence to reject the null hypothesis. ■

4.4.5 Tests on the Difference in Means

Assume that $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ are two independent samples from two independent normal populations. Let $\overline{X}, \overline{Y}, S_1^2$ and S_2^2 be the sample means and sample variances, respectively. In testing the mean difference $\mu_1 - \mu_2 = \Delta_0$, there are three ways to structure the hypothesis test:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$H_1 : \mu_1 - \mu_2 > \Delta_0$	$H_1 : \mu_1 - \mu_2 < \Delta_0$

► **Case I) When σ_1 and σ_2 are known:**

Under the assumption that the null hypothesis is true (i.e. $H_0 : \mu_1 - \mu_2 = \Delta_0$), the test statistic

$$Z_0 = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

has a standard normal distribution.

Let z_0 be the numerical value, calculated from the sample, of the test statistic Z_0. Then, for a selected significance level α, the critical regions are

Two-tailed	Upper-Tailed	Lower-Tailed
If $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$ reject H_0	If $z_0 > z_\alpha$ reject H_0	If $z_0 < -z_\alpha$ reject H_0

and the z-test for the difference in means may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$H_1 : \mu_1 - \mu_2 > \Delta_0$	$H_1 : \mu_1 - \mu_2 < \Delta_0$

Test Statistic (z-test): $Z_0 = \dfrac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$\lvert z_0 \rvert > z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

***p*-values:**

Two-tailed	Upper-Tailed	Lower-Tailed
$2[1 - \Phi(\lvert z_0 \rvert)]$	$1 - \Phi(z_0)$	$\Phi(z_0)$

Example 4.30 *A random sample of size $n_l = 36$ selected from a normal distribution with standard deviation $\sigma_1 = 4$ has a mean $\bar{x} = 75$. A second random sample of size $n_2 = 25$ selected from a different normal distribution with a standard deviation $\sigma_2 = 6$ has a mean $\bar{y} = 85$. Is there a significant difference between the population means at the 5 percent level of significance?. Then, calculate the p-value.*

Solution: From the given information, we have $n_1 = 36$, $n_2 = 25$, $\bar{x} = 75$, $\bar{y} = 85$, $\Delta_0 = 0$, $\sigma_1 = 4$, $\sigma_2 = 6$, and $\alpha = 0.05$.

Step 1: Since we want to determine whether there is a difference between the population means, this will be a two-tailed test. Hence,

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

Step 2: Since σ_1 and σ_2 are known, the appropriate test statistic is the z-test and its value is given by

$$z_0 = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{75 - 85}{\sqrt{\dfrac{16}{36} + \dfrac{36}{25}}} = -7.2846$$

Step 3: The rejection region is $|z_0| > z_{\alpha/2}$, where $z_{\alpha/2} = z_{0.025} = 1.96$.

Step 4: Since $|z_0| = 7.2844 > z_{\alpha/2} = 1.96$, we reject the null hypothesis H_0. We can conclude that the means are significantly different from each other.

The p-value is equal to $2[1 - \Phi(|z_0|)] = 2[1 - \Phi(7.2846)] = 0$, which is less than the significance level $\alpha = 0.05$; therefore we reject the null hypothesis. ■

► **Case II) When σ_1 and σ_2 are unknown and equal to σ^2:**

Under the assumption that the null hypothesis is true (i.e. $H_0 : \mu_1 - \mu_2 = \Delta_0$), the test statistic

$$T_0 = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} = \frac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, and $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is the pooled variance.

Let t_0 be the numerical value, calculated from the sample, of the test statistic T_0. Then, for a selected significance level α, the critical regions are

Two-tailed	Upper-Tailed	Lower-Tailed
If $t_0 < -t_{\alpha/2,,n_1+n_2-2}$ or $t_0 > t_{\alpha/2,,n_1+n_2-2}$ reject H_0	If $t_0 > t_{\alpha,,n_1+n_2-2}$ reject H_0	If $t_0 < -t_{\alpha,,n_1+n_2-2}$ reject H_0

Denote by $F(\cdot)$ the cdf of the $t(n_1 + n_2 - 2)$ distribution. Then, the t-test for the difference in means may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$H_1 : \mu_1 - \mu_2 > \Delta_0$	$H_1 : \mu_1 - \mu_2 < \Delta_0$

Test Statistic (t-test): $T_0 = \dfrac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{1/n_1 + 1/n_2}}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$|t_0| > t_{\alpha/2,n_1+n_2-2}$	$t_0 > t_{\alpha,n_1+n_2-2}$	$t_0 < -t_{\alpha,n_1+n_2-2}$

***p*-values:**

Two-tailed	Upper-Tailed	Lower-Tailed
$2[1 - F(|t_0|)]$	$1 - F(t_0)$	$F(t_0)$

Example 4.31 *An experiment was performed to compare the abrasive wear of two different laminated materials* X *and* Y*. Twelve pieces of material* X *were tested by exposing each piece to a machine measuring wear. Ten pieces of material* Y *were similarly tested. In each case, the depth of wear was observed. The samples of material* X *gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material* Y *gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material* X *exceeds that of material* Y *by more than 2 units? Assume the populations to be approximately normal with equal variances.*

Solution: From the given information, we have $n_1 = 12$, $n_2 = 10$, $\bar{x} = 85$, $\bar{y} = 81$, $\Delta_0 = 2$, $s_1 = 4$, $s_2 = 5$, and $\alpha = 0.05$.

Step 1: Let μ_1 and μ_2 represent the population means of the abrasive wear for material X and material Y, respectively. This is an upper-tailed test. Thus,

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 2 \\ H_1 &: \mu_1 - \mu_2 > 2 \end{aligned}$$

Step 2: Since the standard deviations of the populations are unknown, the appropriate test statistic is the t-test and its value is given by

$$t_0 = \frac{\bar{x} - \bar{y} - \Delta_0}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{85 - 81 - 2}{4.4777\sqrt{1/12 + 1/10}} = 1.0432$$

where the pooled variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)(4)^2 + (10 - 1)(5)^2}{12 + 10 - 2} = 20.05$$

Step 3: The rejection region is $t_0 > t_{\alpha,n_1+n_2-2}$, where $t_{\alpha,n_1+n_2-2} = t_{0.05,20} = 1.7247$.

Step 4: Since $t_0 = 1.0432 \ngtr t_{\alpha,n_1+n_2-2} = 1.7247$, we do not reject the null hypothesis H_0. We are unable to conclude that the abrasive wear of material X exceeds that of material Y by more than 2 units.

The p-value is equal to $1 - F(t_0) = 1 - F(1.0432) = 0.1547$, which is greater than the significance level $\alpha = 0.05$; therefore we have insufficient evidence to reject the null hypothesis. ■

Example 4.32 *One process of making green gasoline, not just a gasoline additive, takes biomass in the form of sucrose and converts it into gasoline using catalytic reactions. This research is still at the pilot plant stage. At one step in a pilot plant process, the product volume (liters) consists of carbon chains of length 3. Nine runs were made with each of two catalysts and the product volumes measured:*

catalyst 1:	*1.86*	*2.05*	*2.06*	*1.88*	*1.75*	*1.64*	*1.86*	*1.75*	*2.13*
catalyst 2:	*0.32*	*1.32*	*0.93*	*0.84*	*0.55*	*0.84*	*0.37*	*0.52*	*0.34*

Is the mean yield with catalyst 1 more than 0.80 liters higher than the yield with catalyst 2? Test with $\alpha = 0.05$.

Solution: From the given information and data, we have $n_1 = 9$, $n_2 = 9$, $\bar{x} = 1.8867$, $\bar{y} = 0.67$, $\Delta_0 = 0.80$, $s_1 = 0.1642$, $s_2 = 0.3366$, and $\alpha = 0.05$.

As shown in Figure 4.11, the normal probability plots and box plot indicate that the measurements constitute a sample from a normal population.

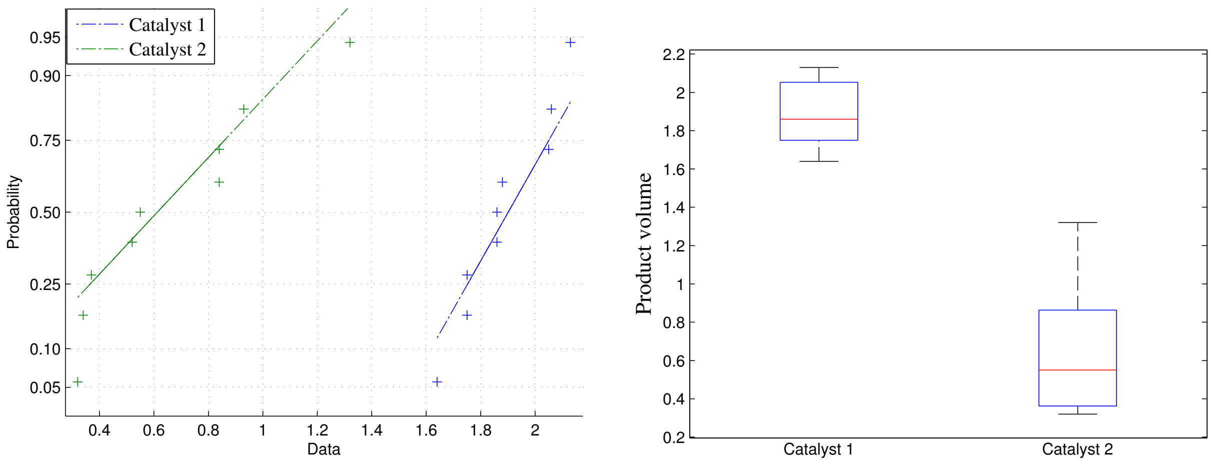


FIGURE 4.11: Normal probability and box plots for the product volumes.

Step 1: Let μ_1 and μ_2 represent the population means for catalyst 1 and catalyst 2, respectively. This is an upper-tailed test. Thus,

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0.80 \\ H_1 &: \mu_1 - \mu_2 > 0.80 \end{aligned}$$

Step 2: Since the standard deviations of the populations are unknown, the appropriate test statistic is the t-test and its value is given by

$$t_0 = \frac{\bar{x} - \bar{y} - \Delta_0}{s_p\sqrt{1/n_1 + 1/n_2}} = \frac{1.8867 - 0.67 - 0.80}{0.2648\sqrt{1/9 + 1/9}} = 3.3381$$

where the pooled variance is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(9-1)(0.1642)^2 + (9-1)(0.3366)^2}{9 + 9 - 2} = 0.0701$$

Step 3: The rejection region is $t_0 > t_{\alpha, n_1+n_2-2}$, where $t_{\alpha, n_1+n_2-2} = t_{0.05,16} = 1.7459$.

Step 4: Since $t_0 = 3.3381 > t_{\alpha, n_1+n_2-2} = 1.7459$, we reject the null hypothesis H_0. We conclude that the mean product volume from catalyst 1 is more than 0.80 liters higher than catalyst 2.

The p-value is equal to $1 - F(t_0) = 1 - F(3.3381) = 0.0021$, which is smaller than the significance level $\alpha = 0.05$; therefore we reject the null hypothesis. ■

4.4.6 Tests on the Equality of Variances

The quality of any process depends on the amount of variability present in the process, which we measure in terms of the variance of the quality characteristic. For example, if we have to choose between two similar processes, we would prefer the one with smaller variance. Any process with smaller variance is more dependable and more predictable. In fact, one of the most important criteria used to improve the quality of a process or to achieve 6σ quality is to reduce the variance of the quality characteristic in the process. In practice, comparing the variances of two processes is common.

Assume that $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2)$ are two independent samples from two independent normal populations. Let S_1^2 and S_2^2 be the sample variances. In testing the equality of variances $\sigma_1^2 = \sigma_2^2$, there are three ways to structure the hypothesis test:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_1 : \sigma_1^2 > \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$

Under the assumption that the null hypothesis is true (i.e. $H_0 : \sigma_1^2 = \sigma_2^2$), the test statistic

$$F_0 = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Let f_0 be the numerical value, calculated from the sample, of the test statistic F_0. Then, for a selected significance level α, the critical regions are

Two-tailed	Upper-Tailed	Lower-Tailed
If $f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$ or $f_0 > f_{\alpha/2, n_1-1, n_2-1}$ reject H_0	If $f_0 > f_{\alpha, n_1-1, n_2-1}$ reject H_0	If $f_0 < f_{1-\alpha, n_1-1, n_2-1}$ reject H_0

and the f-test for the equality of variances may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_1 : \sigma_1^2 > \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$

Test Statistic (f-test): $F_0 = \dfrac{S_1^2}{S_2^2}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$f_0 > f_{\alpha/2,n_1-1,n_2-1}$ or $f_0 < f_{1-\alpha/2,n_1-1,n_2-1}$	$f_0 > f_{\alpha,n_1-1,n_2-1}$	$f_0 < f_{1-\alpha,n_1-1,n_2-1}$

Example 4.33 *In testing for the difference in the abrasive wear of the two materials in the previous example, we assumed that the two unknown population variances were equal. Were we justified in making this assumption? Use a 0.10 level of significance.*

Solution: From the given information, we have $n_1 = 12$, $n_2 = 10$, $s_1 = 4$, $s_2 = 5$, and $\alpha = 0.10$.

Step 1: Let σ_1^2 and σ_2^2 represent the population variances of the abrasive wear for material X and material Y, respectively. This is a two-tailed test. Thus,

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

Step 2: Since the standard deviations of the populations are unknown, the appropriate test statistic is the f-test and its value is given by

$$f_0 = \frac{s_1^2}{s_2^2} = 0.64$$

Step 3: The rejection region is $f_0 > f_{\alpha/2,n_1-1,n_2-1}$ or $f_0 < f_{1-\alpha/2,n_1-1,n_2-1}$, where $f_{\alpha/2,n_1-1,n_2-1} = 3.1025$ and $f_{1-\alpha/2,n_1-1,n_2-1} = 0.3453$.

Step 4: Since $f_0 = 0.64 \not> f_{\alpha/2,n_1-1,n_2-1} = 3.1025$, we do not reject the null hypothesis H_0. We conclude that there is insufficient evidence that the variances differ.

Example 4.34 *Suppose the following is the sample summary of samples from two independent processes:*

$$n_1 = 21 \qquad S_1^2 = 24.6$$
$$n_2 = 16 \qquad S_2^2 = 16.4$$

We assume that the quality characteristics of the two processes are normally distributed $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Test at the 5% level of significance the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$, and find the p-value for the test.

Solution: From the given information, we have $n_1 = 21$, $n_2 = 16$, $s_1^2 = 24.6$, $s_2^2 = 16.4$, and $\alpha = 0.05$.

Step 1: This is a two-tailed test. Thus,

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

Step 2: Since the standard deviations of the populations are unknown, the appropriate test statistic is the f-test and its value is given by

$$f_0 = \frac{s_1^2}{s_2^2} = 1.50$$

Step 3: The rejection region is $f_0 > f_{\alpha/2,n_1-1,n_2-1}$ or $f_0 < f_{1-\alpha/2,n_1-1,n_2-1}$, where $f_{\alpha/2,n_1-1,n_2-1} = 2.7559$ and $f_{1-\alpha/2,n_1-1,n_2-1} = 0.3886$.

Step 4: Since $f_0 = 1.50 \not> f_{\alpha/2,n_1-1,n_2-1} = 2.7559$ and $f_0 = 1.50 \not< f_{1-\alpha/2,n_1-1,n_2-1} = 0.3886$, we do not reject the null hypothesis H_0. We conclude that there is insufficient evidence that the variances differ. ■

4.5 Problems

❶ The total weight of a filled tire can dramatically affect the performance and safety of an automobile. Some transportation officials argue that mechanics should check the tire weights of every vehicle as part of an annual inspection. Suppose the weight of a 185/60/14 filled tire is normally distributed with standard deviation 1.25 pounds. In a random sample of 15 filled tires, the sample mean weight was 18.75 pounds. Find a 95% confidence interval for the true mean weight of 185/60/14 tires.

❷ An electro Pneumatic hammer has an advertised impact force of 2.2 joules. In a random sample of 23 hammers, the impact force for each tool was carefully measured (in joules), and the resulting data are as follows:

```
2.16 1.69 2.30 2.08 1.72 2.17 2.25 2.06 2.00 2.29 2.15 2.49
2.12 2.17 1.93 2.39 2.22 2.26 2.14 1.92 2.06 2.09 2.08
```

a) Check the assumption of normality for the impact force data.

b) Find the 95% two-sided confidence interval for the true mean impact force for this type of pneumatic hammer.

c) Using the confidence interval constructed in part b), is there any evidence to suggest that the true mean impact force is different from 2.2 joules as advertised? Justify your answer.

❸ Adobeware dishes are made from clay and are fired, or exposed to heat, in a large kiln. Large fluctuations in the kiln temperature can cause cracks, bumps, or other flaws (and increase cost). With the kiln set at 800°C, a random sample of 19 temperature measurements (in °C) was obtained. The sample variance was 17.55.

a) Find the 95% two-sided confidence interval for the true population variance in temperature of the kiln when it is set to 800°C. Assume that the underlying distribution is normal.

b) Quality control engineers have determined that the maximum variance in temperature during firing should be 16°C. Using the confidence interval constructed in part a), is there any evidence to suggest that the true temperature variance is greater than 16°C? Justify your answer.

❹ A successful company usually has high brand name and logo recognition among consumers. For example, Coco-Cola products are available to 98% of all people in the world, and therefore may have the highest logo recognition on any company. A software firm developing a product would like to estimate the proportion of people who recognize the Linux penguin logo. Of the 952 randomly selected consumers surveyed, 132 could identify the product associated with the penguin.

a) Is the distribution of the sample proportion, $\hat{p}$, approximately normal? Justify your answer.

b) Find the 95% two-sided confidence interval for the true proportion of consumers who recognize the Linux penguin.

c) The company will market a Linux version of their new software if the true proportion of people who recognize the logo is greater that 0.10. Is there any evidence to suggest that the true proportion of people who recognize the logo is greater than 0.10? Justify your answer.

❺ An engineer wants to measure the bias in a pH meter. She uses the meter to measure the pH in 15 neutral substances ($pH = 7.0$) and obtains the following data:

```
7.04  7.0  7.03  7.01  6.97  7.00  6.95  7.00  6.99  7.04  6.97  7.07  7.04  6.97  7.08
```

a) Check the assumption of normality for the pH meter data.

b) Is there sufficient evidence to support the claim that the pH meter is not correctly calibrated at the 5% level of significance

c) Find the 95% two-sided confidence interval to estimate the mean. Comment on your result.

❻ A quality control supervisor in a cannery knows that the exact amount each can contains will vary, since there are certain uncontrollable factors that affect the amount of fill. Suppose regulatory agencies specify that the standard deviation of the amount of fill should be less that 0.1 ounce. The quality control supervisor sampled 10 cans and measured the amount of fill in each. The resulting data measurements are:

$$7.96, 7.90, 7.98, 8.01, 7.97, 7.96, 8.03, 8.02, 8.04, 8.02$$

Does this information, at the 0.05 level of significance, provide sufficient evidence to indicate that the standard deviation of the fill measurements is less than 0.1 ounce? Then, calculate the p-value.

❼ The management of a luxurious hotel is concerned with increasing the return rate for hotel guests. One aspect of first impressions by guests relates to the time it takes to deliver the guest's luggage to the room after check-in to the hotel. A random sample of 20 deliveries on a particular day were selected in Wing A of the hotel and a random sample of 20 deliveries were selected in Wing B.

Wing A:	10.70, 9.89, 11.83, 9.04, 9.37, 11.68, 8.36, 9.76, 13.67, 8.96, 9.51, 10.85, 10.57, 11.06, 8.91, 11.79, 10.59, 9.13, 12.37, 9.91
Wing B:	7.20, 6.68, 9.29, 8.95, 6.61, 8.53, 8.92, 7.95, 7.57, 6.38, 8.89, 10.03, 9.30, 5.28, 9.23, 9.25, 8.44, 6.57, 10.61, 6.77

a) Is the normality assumption of the data satisfied? Justify your answer.

b) Was there a difference in the mean delivery time in the two wings on the hotel?. Test with $\alpha = 0.05$.

c) Determine whether the variance in luggage delivery time is the same for Wing A and Wing B of the hotel at the $\alpha = 0.05$ level of significance.

d) Assume that delivery times to Wing A and Wing B are two independent normal populations with unknown variances σ_1^2 and σ_2^2, respectively. Construct a 90% two-sided confidence interval on the ratio of the two standard deviations σ_1/σ_2. Comment on your result.

4.6 REFERENCES

[1] D.C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, 6th Edition, 2009.

[2] I. Bass and B. Lawton, *Lean Six Sigma using SigmaXL and Minitab*, McGraw-Hill Professional, 1st Edition, 2009.

[3] B.C. Gupta and H.F. Walker, *Applied statistics for the Six Sigma Green Belt*, ASQ Quality Press, 2004.

CHAPTER 5

STATISTICAL PROCESS AND QUALITY CONTROL

Almost all quality improvement comes via simplification of design, manufacturing,, layout, processes, and procedures.

Tom Peters

Statistical quality control (SQC) is a term used to describe the activities associated with ensuring that goods and services satisfy customer needs. SQC uses statistical analysis based on measurements taken from a process or from a sample of products or services, to make decisions regarding the quality of goods and services. The statistical methods of SQC may be divided into two main categories: Statistical process control (SPC) and acceptance sampling. SPC refers to the use of statistical methods to measure and control the performance of a process to ensure that the output meets customer needs. Acceptance sampling is a methodology of taking samples from lots of materials or products and inspecting the items to determine if the items meet customer requirements. A process may include customer services, productions systems, and administration activities. SPC may be used to help control almost all processes that can be measured or monitored to ensure that the process performs within limits.

5.1 STATISTICAL PROCESS CONTROL

Statistical process control allows engineers to understand and monitor process variation through **control charts**. The causes of process variation in a product quality characteristic may be broadly classified into two main categories: **common causes** of variation (variation due to the system itself) and **assignable causes** of variation (variation due to factors external to the system).

The concept of control charts was first introduced by Walter A. Shewhart of Bell Telephone Laboratories during the 1920's. For this reason, statistical control charts are also known as Shewhart control charts. A control chart is a graphical method to quickly spot assignable cause of variation of a process. Variation is present in any process; deciding when the variation is natural and when it needs correction is the key to quality control. A control chart displays a quality characteristic that has been measured or computed from a sample versus the sample number or time. The sample values to be used in a quality control effort are divided into subgroups with a sample representing a subgroup. A control chart contains a center line (CL) that represents the average value of the quality characteristic when the process is in control. Two other horizontal lines, called the upper control limit (UCL) and the lower control limit (LCL), are also shown on the chart. These control limits are chosen so that if the process is in control, nearly all of the sample points will fall between them. In general, as long as the points plot within the control limits, the process is assumed to be **in-control**, and no action is necessary. However, a point that plots outside of the control limits is interpreted as evidence that the process is **out-of-control**, and investigation and corrective action are required to find and eliminate the assignable cause or causes responsible for this behavior. The sample points on the control chart are usually connected with straight-line segments so that it is easier to visualize how the sequence of points has evolved over time. Figure 5.1 illustrates the concept of a control chart, where the process is found to be out of control due to the sample number 15 which falls outside the control limits.

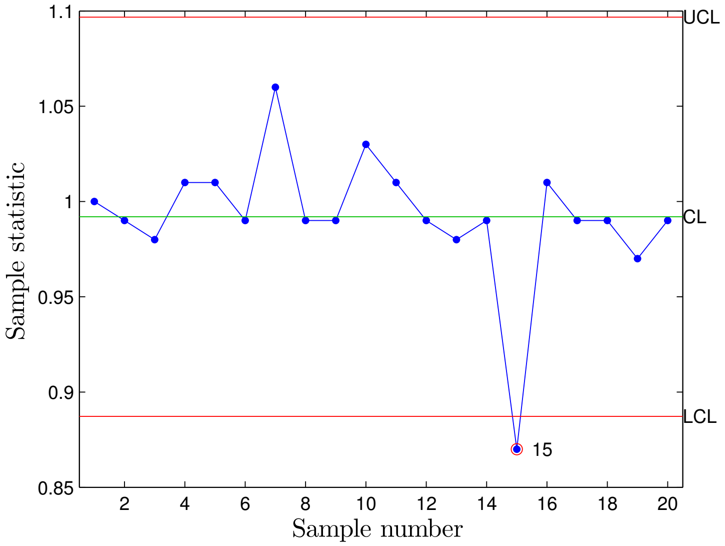


FIGURE 5.1: Illustration of a control chart.

5.1.1 Hypothesis Testing and Control Charts

There is a close connection between control charts and hypothesis testing. A control chart may be formulated as hypothesis test:

$$
\begin{aligned}
H_0 &: \quad \text{process is in-control}\\
H_1 &: \quad \text{process is out-of-control}
\end{aligned}
\tag{5.1.1}
$$

Control limits are established to control the probability of making the error of concluding that the process is out of control when in fact it is not. This corresponds to the probability of making a Type I error if we were testing the null hypothesis that the process is in control. On the other hand, we must be attentive to the error of not finding the process out of control when in fact it is (Type II error). Thus, the choice of control limits is similar to the choice of a critical region. When a point plots within the control limits, the null hypothesis is not rejected; and when a point plots outside the control limits, the null hypothesis is rejected.

Definition 5.1 *Let Θ be a sample statistic that measures some quality characteristic of interest, with mean μ_Θ and standard deviation σ_Θ. The upper control limit, center line, and lower control limit are given by*

$$
\begin{aligned}
UCL &= \mu_\Theta + 3\sigma_\Theta\\
CL &= \mu_\Theta\\
LCL &= \mu_\Theta - 3\sigma_\Theta
\end{aligned}
\tag{5.1.2}
$$

The $3\sigma_\Theta$ limits imply that there is a probability of only 0.0026 of a sample statistic to fall outside the control limits if the process is in-control.

Control chart are broadly classified into control charts for variables and control charts for attributes. Variable control charts are used for quality characteristics that are measured on a continuous scale such as length, temperature, weight, and time. Attribute control charts are used for quality characteristics in discrete (count) data, such as number of defects. Attribute control charts are further divided into two main classes: attributes control charts for defective units, and attribute control charts for defects per unit.

5.1.2 Rules for Determining Out-Of-Control Points

The control chart is an important tool for distinguishing between the common causes of variation that are due to the process and special causes of variation that are not due to the process. Only management can change the process. One of the main goals of using a control chart is to determine when the process is out-of-control so that necessary actions may be taken. The simplest rule for detecting the presence of an assignable (or special) cause of variation is one or more plotted points falling outside the control limits UCL and LCL. *Assignable causes* are special causes of variation that are ordinarily not part of the process, and should be corrected as warranted. *Common causes*, on the other hand, are inherent in the design of the system and reflect the typical variation to be expected. An *unstable* (or out-of-control) process exhibits variation due to both assignable and common causes. Improvement can be achieved by identifying and removing the assignable cause(s). A *stable* process is one that exhibits only common-cause variation, and can be improved only by changing the design of the process. Attempts to make adjustments to a stable process, which is called tampering, results in more variation in the quality of the output. Control charts are used to detect the occurrence of assignable causes affecting the quality of process output. Figure depicts a control chart in which the area between UCL and LCL is subdivided into bands, each of which is $1\sigma_\Theta$ wide.

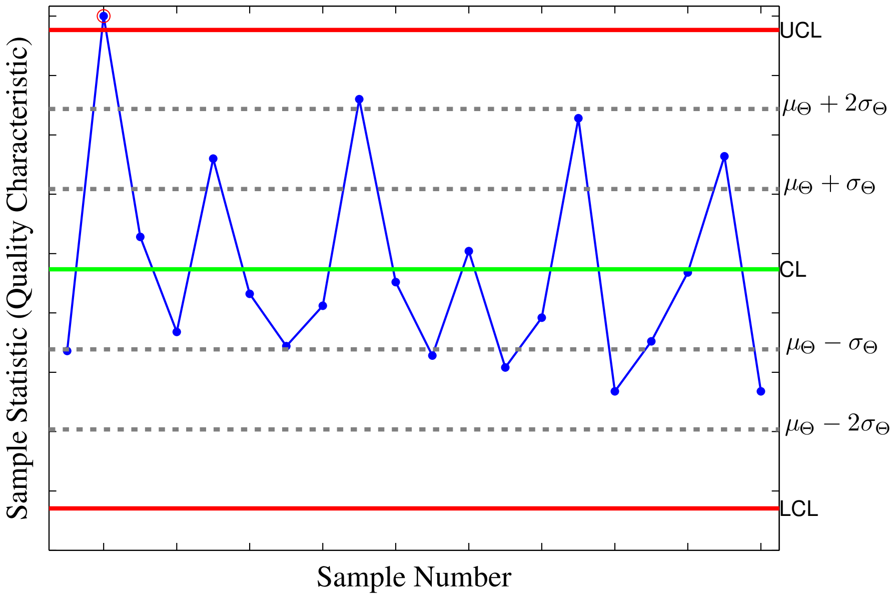


FIGURE 5.2: Illustration of control chart bands, each of which is $1\sigma_\Theta$ wide.

The rules for determining out-of-control points in a control chart may be summarized in five main rules (refereed to as **Western Electric rules**). That is, a process is considered out-of-control (unstable) if:

Rule 1: A point falls outside the upper and lower control limits, i.e. above UCL or below LCL

Rule 2: Two out of three consecutive points fall above $\mu_\Theta + 2\sigma_\Theta$ or below $\mu_\Theta - 2\sigma_\Theta$

Rule 3: Four out of five consecutive points fall above $\mu_\Theta + 1\sigma_\Theta$ or below $\mu_\Theta - 1\sigma_\Theta$

Rule 4: Eight or more consecutive points fall above μ_Θ or below μ_Θ

Rule 5: Eight or more consecutive points move upward (increasing) or downward (decreasing) in value.

5.2 Control Charts for Variables

Control charts for variables are used to study a process when a characteristic is a measurement; for example, temperature, cost, revenue, processing time, area, and waiting time. Variable charts are typically used in pairs. One chart studies the variation in a process, and the other chart studies the variation in the process mean. A chart that studies the process variability must be examined before the chart that studies the process mean. This is due to the fact that the chart that studies the process mean assumes that the process variability is stable over time. One of the most commonly used pairs of charts is the $\overline{X}$-chart and the R-chart. Another pair is the $\overline{X}$-chart and the s-chart. In this section, we discuss in detail these two pairs of charts.

5.2.1 Control Charts for the Mean and Range

Control Chart for the Mean ($\overline{X}$-chart):

An $\overline{X}$-chart is a control chart plotting the sample means vs the sample number. Denote by μ and σ the process mean and standard deviation, respectively.

Considering the sample mean $\overline{X}$ as the sample statistic Θ yields $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When the parameters μ and σ are unknown, we usually estimate them on the basis of preliminary samples (subgroups), taken when the process is thought to be in control. Suppose m preliminary samples are available, each of size n. Denote by $\overline{X}_i$ and R_i the sample mean and range of the i-th sample, respectively. An unbiased estimator of μ is obtained by averaging m sample (subgroup) means when the process is in control

$$CL = \overline{\overline{X}} = \frac{1}{m}\sum_{i=1}^{m} \overline{X}_i \tag{5.2.1}$$

where $\overline{\overline{X}}$ (pronounced "X-double-bar") is the average of the $\overline{X}_i$'s.

On the other hand, it can be shown that the mean and standard deviation of relative range $W = R/d_2$ are $\mu_W = d_2$ and $\sigma_W = d_3$, where d_2 and d_3 are constants that depend on the sample size n. That is, the relative range is an unbiased estimator of σ. Since $R = W\sigma$, it follows that $\mu_R = d_2\sigma$ and $\sigma_R = d_3\sigma$. An unbiased estimator of μ_R is the average of the sample ranges given by

$$\overline{R} = \frac{1}{m}\sum_{i=1}^{m} R_i \tag{5.2.2}$$

where $R_i = X_{(n),i} - X_{(1),i}$ is the range of the ith-sample (subgroup). Here $X_{(n),i}$ and $X_{(1),i}$ denote the largest and smallest observations, respectively, in the sample.
Thus, an unbiased estimator of σ is

$$\hat{\sigma} = \frac{\overline{R}}{d_2} \tag{5.2.3}$$

The upper and control limits are located at a distance of $3\sigma_{\overline{X}} = 3\sigma/\sqrt{n}$ above and below the center line. An estimator of this distance is given by

$$3\hat{\sigma}/\sqrt{n} = \frac{3(\overline{R}/d_2)}{\sqrt{n}} = \frac{3}{d_2\sqrt{n}}\overline{R} = A_2\overline{R} \tag{5.2.4}$$

where A_2 is a constant that depends on the sample size n (see Appendix Table V).

Control chart for the mean ($\overline{X}$-chart):
The upper control limit, center line, and lower control limit of the $\overline{X}$-chart are given by

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{r} \\ CL &= \bar{\bar{x}} \\ LCL &= \bar{\bar{x}} - A_2\bar{r} \end{aligned} \tag{5.2.5}$$

Example 5.1 *A control chart for $\overline{X}$ is to be set up for an important quality characteristic. The sample size is $n = 4$, and $\bar{x}$ and r are computed for each of 25 preliminary samples. The summary data are*

$$\sum_{i=1}^{25} \bar{x}_i = 7657 \qquad \sum_{i=1}^{25} \bar{r}_i = 1180$$

(i) Find the control limits for $\overline{X}$-chart

(ii) *Assuming the process is in control, estimate the process mean and standard deviation.*

Solution: From the given information, the number of samples is $m = 25$ and the sample size is $n = 4$.

(i) The grand mean and the average range are given by

$$\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i = \frac{7657}{25} = 306.28 \qquad \bar{r} = \frac{1}{m}\sum_{i=1}^{m} r_i = \frac{1180}{25} = 47.20$$

The value of A_2 for samples of size 4 is $A_2 = 0.729$. Therefore, the control limits of the $\overline{X}$-chart are

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{r} = 306.28 + (0.729)(47.20) = 340.69 \\ CL &= \bar{\bar{x}} = 306.28 \\ LCL &= \bar{\bar{x}} - A_2\bar{r} = 306.28 + (0.729)(47.20) = 271.87 \end{aligned}$$

(ii) The estimates of the process mean and standard deviation are

$$\hat{\mu} = \bar{\bar{x}} = 306.28 \qquad \hat{\sigma} = \frac{\bar{r}}{d_2} = \frac{47.20}{2.059} = 22.92$$

Control Chart for the Range (R-chart):

In quality control, we want to control not only the mean value of some quality characteristic but also its variability. A range chart (or simply R-chart) is a control chart plotting the sample ranges vs the sample number, and it monitors the variation of a quality characteristic. Considering the sample range R as the statistic Θ yields $\mu_R = d_2\sigma$ and $\sigma_R = d_3\sigma$, where d_2 and d_3 are constants (see Appendix Table V) that depend on the sample size n. An unbiased estimator of μ_R is the mean $\overline{R}$ of the ranges of m samples

$$CL = \overline{R} = \frac{1}{m}\sum_{i=1}^{m} R_i$$

The upper and control limits are located at a distance of $3\sigma_R = 3d_3\sigma$ above and below the center line. Since $\overline{R}/d_2$ as an unbiased estimator of σ, it follows that the upper and control limits may be expressed as

$$UCL = \overline{R} + 3\frac{d_3}{d_2}\overline{R} = \left(1 + 3\frac{d_3}{d_2}\right)\overline{R} = D_4\overline{R} \tag{5.2.6}$$

and

$$LCL = \overline{R} - 3\frac{d_3}{d_2}\overline{R} = \left(1 - 3\frac{d_3}{d_2}\right)\overline{R} = D_3\overline{R} \tag{5.2.7}$$

where D_3 and D_4 are constants that depend on the sample size n (see Appendix Table V).

Control chart for the range (R-chart):
The upper control limit, center line, and lower control limit for the R-chart are given by

$$\begin{aligned} UCL &= D_4\bar{r} \\ CL &= \bar{r} \\ LCL &= D_3\bar{r} \end{aligned} \tag{5.2.8}$$

The R-chart highlights the changes in the process variability and shows better results when analyzed in conjunction with the $\overline{X}$-chart. It is likely that a sample with the same mean may not reveal a shift in the process at all. Thus, it is necessary to analyze both the $\overline{X}$-and R-chart together to decide whether the process is in-control or out-of-control.

Example 5.2 *Samples of size 5 are collected from a process every hour. After 30 samples have been collected, we calculate the value of the average range* $\bar{r} = 2.5$. *Find the control limits for the R-chart.*

Solution: From the given information, the number of samples is $m = 30$ and the sample size is $n = 5$. From Appendix Table V, the values of D_3 and D_4 are $D_3 = 0$ and $D_4 = 2.115$. Thus, the upper control limit, center line, and lower control limit for the R-chart are given by

$$\begin{aligned} UCL &= D_4\bar{r} = (2.114)(2.5) = 5.29 \\ CL &= \bar{r} = 2.5 \\ LCL &= D_3\bar{r} = 0 \end{aligned}$$

Example 5.3 *The data provided in Table 5.1 have been obtained by measuring four consecutive units on a assembly line every 30 minutes until the 20 subgroups (samples) are obtained. Each subgroup has 5 observations. Construct the $\overline{X}$- and R-charts. Is the process under statistical control? Explain.*

TABLE 5.1: Assembly Data in Subgroups (Samples) Obtained at Regular Intervals.

Sample	Data				
Number	X_1	X_2	X_3	X_4	X_5
1	29	51	75	62	42
2	97	73	75	99	56
3	46	60	68	76	57
4	40	61	66	41	59
5	66	76	70	76	52
6	58	61	45	41	78
7	53	38	71	57	42
8	48	71	37	62	60
9	86	65	68	72	74
10	77	55	43	50	63
11	54	50	41	66	46
12	58	48	49	87	59
13	83	57	35	25	52
14	87	43	57	39	47
15	76	48	76	77	80
16	51	32	53	52	54
17	47	48	65	56	47
18	70	45	60	73	44
19	42	67	78	95	59
20	39	82	54	35	32

Solution: The upper control limit, center line, and lower control limit for the R-chart are

$$\begin{aligned} UCL &= D_4\bar{r} = 74.219 \\ CL &= \bar{r} = 35.10 \\ LCL &= D_3\bar{r} = 0 \end{aligned}$$

where for a sample of size $n = 5$, Appendix Table V gives $D_3 = 0$ and $D_4 = 2.114$.

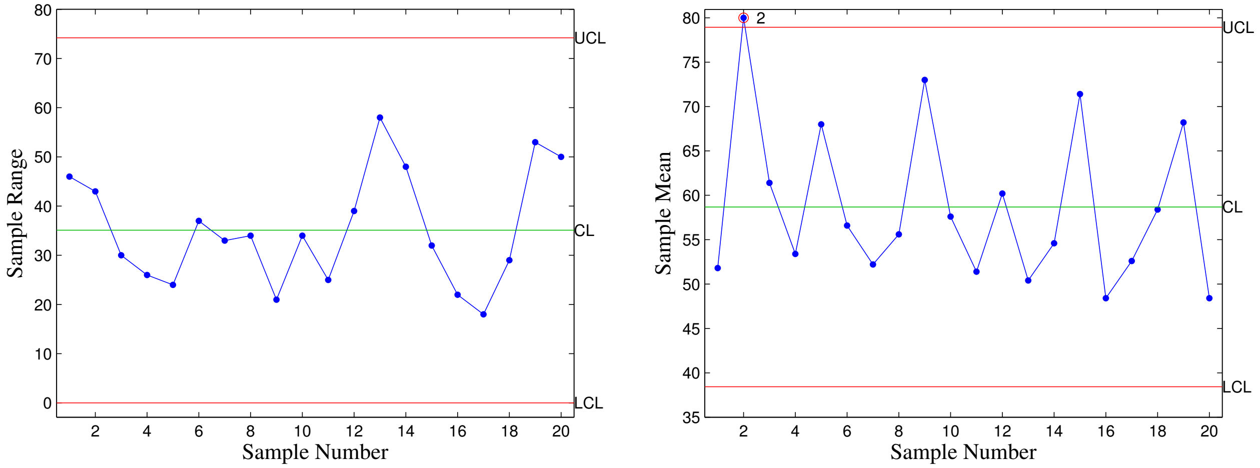
```
load assembly.mat %load assembly data
[stats,plotdata]=controlchart(X,'chart',{'r','xbar'},'sigma','range'); %R- and Xbar-charts
```

First we examine the R-chart for signs of special variation. The R-chart is shown in Figure 5.3(a), where all samples appear to be in-control. None of the points on the R-chart is outside the control limits, and there are no other signals indicating a lack of control. Thus, there are no indications of special sources of variation on the R-chart. In other words, only common-cause variation appears to exist. In this case, we can proceed further to calculate the upper control limit, center line, and lower control limit for the $\overline{X}$-chart

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{r} = 78.926 \\ CL &= \bar{\bar{x}} = 58.680 \\ LCL &= \bar{\bar{x}} - A_2\bar{r} = 38.434 \end{aligned}$$

where the value of A_2 for a sample of size $n = 5$ is $A_2 = 0.577$.

The $\overline{X}$-chart is shown in Figure 5.3(b), where the sample number 2 appears to be out-of-control. Further investigation is warranted to determine the source(s) of this special variation. Thus, the $\overline{X}$-chart can be interpreted without concern that the observed variability in the sample means could be associated with a lack of control of process variability. Therefore, the sample number 2 is discarded when remedial actions have been taken to remove special causes, and then new limits are calculated using the remaining 24 samples (i.e. samples 1 and 3 to 25). These limits are referred to as **revised control limits**.



(a) R-chart

(b) $\overline{X}$-chart

FIGURE 5.3: R- and $\overline{X}$-charts for assembly data.

With sample 2 deleted, the revised mean range and grand mean are:

$$\bar{\bar{x}} = 57.56 \qquad \bar{r} = 34.69$$

The revised control limits for the new R-chart are

$$\begin{aligned} UCL &= D_4\bar{r} = 73.340 \\ CL &= \bar{r} = 34.684 \\ LCL &= D_3\bar{r} = 0 \end{aligned}$$

and the revised control limits for the new $\overline{X}$-chart are

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{r} = 77.564 \\ CL &= \bar{\bar{x}} = 57.558 \\ LCL &= \bar{\bar{x}} - A_2\bar{r} = 37.551 \end{aligned}$$

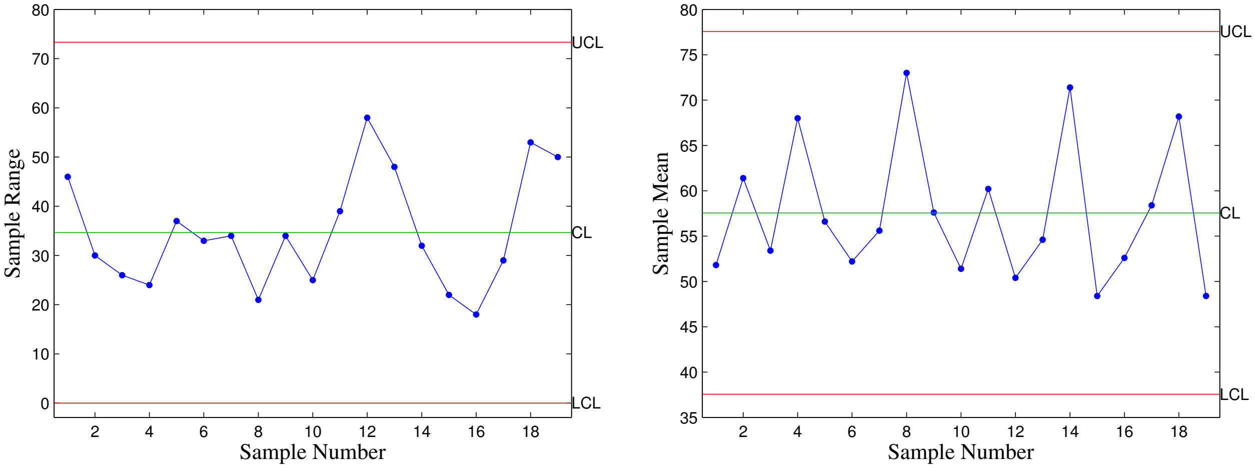
The new $\overline{X}$- and R-charts are shown in Figure 5.4. Notice now that all the points fall within the limits, indicating that the process may be stable. ■

5.2.2 Control Charts for the Mean and Standard Deviation

It is customary to examine the control chart for standard deviation first and verify that the all the plotted points fall within the control limits, and then proceed to constructing the X-chart. In fact, the concept of bringing the process variability under control first and then proceeding to control the mean does make a lot of sense. This is due to the fact that without controlling the process variability, it is almost impossible to bring the process mean under control.

Control Chart for the Standard Deviation (S-chart):

As the sample size n increases, the range becomes increasingly less efficient as a measure of variability. This is the case because the range ignores all information between the two most extreme values (minimum and maximum sample



(a) R-chart

(b) $\overline{X}$-chart

FIGURE 5.4: Revised R- and $\overline{X}$-charts for assembly data.

values). A standard deviation control chart (or simply S-chart) is sensitive to changes in variation in the measurement process, and it preferable for larger sample sizes ($n \geq 10$). The S-chart plots the sample standard deviations vs the sample number. Suppose m preliminary samples are available, each of size n. Denote by S_i the sample standard deviation of the i-th sample. An unbiased estimator of μ_S is the mean $\overline{S}$ of the sample standard deviations of m samples

$$CL = \overline{S} = \frac{1}{m}\sum_{i=1}^{m} S_i \tag{5.2.9}$$

The upper and lower control limits are located at a distance of $3\sigma_S$ above and below the center line. It can be shown that $E(\overline{S}) = \mu_{\overline{S}} = c_4\sigma$ and $\sigma_{\overline{S}} = \sigma\sqrt{1-c_4^2}$, where c_4 is a constant that depends on the sample size n (see Appendix Table V). Thus, $\overline{S}/c_4$ is an unbiased estimator of σ, which in turn implies that $(\overline{S}/c_4)\sqrt{1-c_4^2}$ is an estimator of $\sigma_{\overline{S}}$. Therefore, the upper and lower control limits may be expressed as

$$UCL = \overline{S} + 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2} = \left(1 + \frac{3}{c_4}\sqrt{1-c_4^2}\right)\overline{S} = B_4\overline{S} \tag{5.2.10}$$

and

$$LCL = \overline{S} - 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2} = \left(1 - \frac{3}{c_4}\sqrt{1-c_4^2}\right)\overline{S} = B_3\overline{S} \tag{5.2.11}$$

where B_3 and B_4 are constants that depend on the sample size n (see Appendix Table V).

Control chart for standard deviation (S-chart):
The upper control limit, center line, and lower control limit of the S-chart are given by

$$\begin{aligned} UCL &= B_4\bar{s} \\ CL &= \bar{s} \\ LCL &= B_3\bar{s} \end{aligned} \tag{5.2.12}$$

Example 5.4 *Containers are produced by a process where the volume of the containers is subject to quality control. Twenty-five samples of size 5 each were used to establish the quality control parameters, and the sum of the sample standard deviations is*

$$\sum_{i=1}^{25} s_i = 0.903$$

(i) Find the control limits for the S-chart

(ii) Assuming the process is in control, estimate the process standard deviation.

Solution: From the given information, the number of samples is $m = 25$ and the sample size is $n = 5$. Thus, Appendix Table V gives $B_3 = 0$, $B_4 = 2.089$, and $c_4 = 0.940$.

(i) The average of the sample standard deviations is

$$\bar{s} = \frac{1}{m}\sum_{i=1}^{m} s_i = \frac{0.903}{25} = 0.0361$$

Therefore, the control limits for the S-chart are

$$\begin{aligned} UCL &= B_4\bar{s} = (2.089)(0.0361) = 0.0754 \\ CL &= \bar{s} = 0.0361 \\ LCL &= B_3\bar{s} = 0 \end{aligned}$$

(ii) The process standard deviation can be estimated as follows

$$\hat{\sigma} = \frac{\bar{s}}{c_4} = \frac{0.0361}{0.940} = 0.0384.$$

Control Chart for the Mean (from $\bar{s}$):

We can now write the parameters of the corresponding $\overline{X}$-chart involving the use of the sample standard deviation. Let us assume that the sample standard deviation S and the sample mean $\overline{X}$ are available from the base preliminary sample. Since $\overline{S}/c_4$ as an unbiased estimator of σ, the upper and control limits of the $\overline{X}$-chart may also be written as

$$UCL = \overline{\overline{X}} + 3\frac{\overline{S}}{c_4\sqrt{n}} = \overline{\overline{X}} + A_3\overline{S} \tag{5.2.13}$$

and

$$LCL = \overline{\overline{X}} - 3\frac{\overline{S}}{c_4\sqrt{n}} = \overline{\overline{X}} - A_3\overline{S} \tag{5.2.14}$$

where A_3 is a constant that depends on the sample size n (see Appendix Table V).

Control chart for mean ($\overline{X}$-chart from $\bar{s}$):
The upper control limit, center line, and lower control limit of the $\overline{X}$-chart (from $\bar{s}$) are given by

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_3\bar{s} \\ CL &= \bar{\bar{x}} \\ LCL &= \bar{\bar{x}} - A_3\bar{s} \end{aligned} \tag{5.2.15}$$

Example 5.5 *Using the data of Table 5.1, construct the S- and the $\overline{X}$-charts.*

Solution: The upper control limit, center line, and lower control limit for the S-chart are given by

$$\begin{aligned} UCL &= B_4\bar{s} = 29.502 \\ CL &= \bar{s} = 14.122 \\ LCL &= B_3\bar{s} = 0 \end{aligned}$$

where $B_3 = 0$ and $B_4 = 2.089$ for samples of size $n = 5$.
Similar to the analysis that we performed in the case of R-chart, we first examine the S-chart for signs of special variation. The S-chart is given in Figure 5.5(a), which shows that all plotted points fall within the control limits and there is no evidence of any special pattern. Thus, we may conclude that the only variation present in the process is

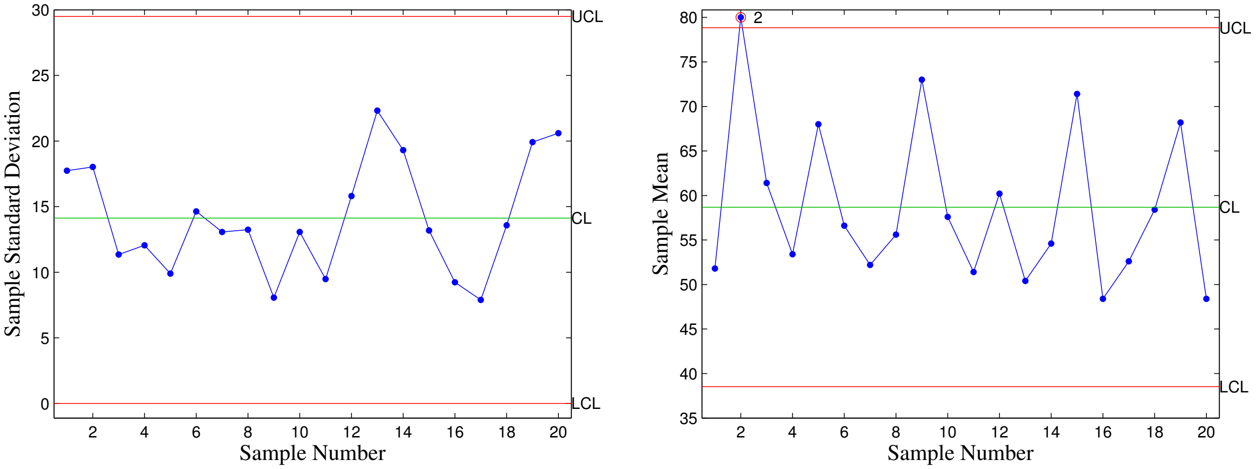
due to common causes. In this case, we can proceed further to calculate the upper control limit, center line, and lower control limit for the $\overline{X}$-chart:

$$\begin{aligned}
UCL &= \bar{\bar{x}} + A_3\bar{s} = 78.837\\
CL &= \bar{\bar{x}} = 58.680\\
LCL &= \bar{\bar{x}} - A_3\bar{s} = 38.523
\end{aligned}$$

where $A_3 = 1.427$ for samples of size $n = 5$.

```
load assembly.mat %load assembly data
[stats ,plotdata]=controlchart(X,'chart',{'s','xbar'},'sigma','std'); %S- and Xbar-charts
```

The $\overline{X}$-chart is shown in Figure 5.5(b), where the sample number 2 is out-of-control. This indicate that the process may not be under control and there are some special causes present that are affecting the process mean. Thus, a thorough investigation should be launched to find the special causes, and appropriate action should be taken to eliminate these special causes before we proceed to recalculate the control limits for the ongoing process.



(a) s-chart

(b) $\overline{X}$-chart

FIGURE 5.5: S- and $\overline{X}$-charts for assembly data.

With sample 2 deleted, the new upper control limit, center line, and lower control limit for the S-chart are given by

$$\begin{aligned}
UCL &= B_4\bar{s} = 29.072\\
CL &= \bar{s} = 13.917\\
LCL &= B_3\bar{s} = 0
\end{aligned}$$

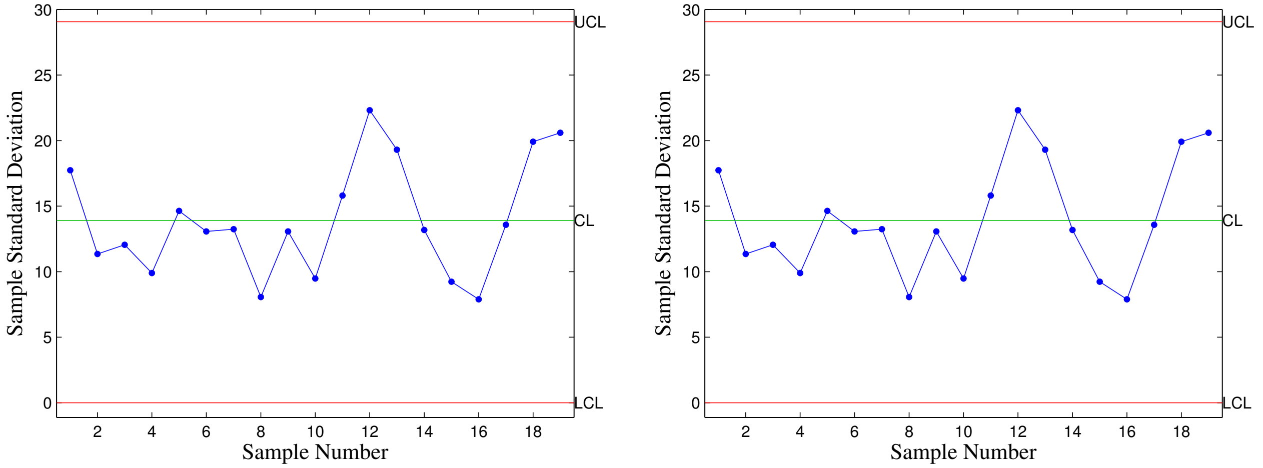
The new upper control limit, center line, and lower control limit for the $\overline{X}$-chart are given by

$$\begin{aligned}
UCL &= \bar{\bar{x}} + A_3\bar{s} = 77.421\\
CL &= \bar{\bar{x}} = 57.558\\
LCL &= \bar{\bar{x}} - A_3\bar{s} = 37.694
\end{aligned}$$

The new $\overline{X}$- and S-charts are shown in Figure 5.6. Notice now that all the points fall within the limits, indicating that the process may be stable.

Example 5.6 *A component part for a jet aircraft engine is manufactured by an investment casting process. The vane opening on this casting is an important functional parameter of the part. Table 5.2 presents 20 samples of five parts each. The values given in the table have been coded by using the last three digits of the dimension; that is, 31.6 should be 0.50316 inch.*

(i) Estimate the process mean and standard deviation.



(a) S-chart

(b) $\overline{X}$-chart

FIGURE 5.6: Revised S- and $\overline{X}$-charts for assembly data.

TABLE 5.2: Vane-Opening Data.

Sample	Data				
Number	X_1	X_2	X_3	X_4	X_5
1	33	29	31	32	33
2	33	31	35	37	31
3	35	37	33	34	36
4	30	31	33	34	33
5	33	34	35	33	34
6	38	37	39	40	38
7	30	31	32	34	31
8	29	39	38	39	39
9	28	33	35	36	43
10	38	33	32	35	32
11	28	30	28	32	31
12	31	35	35	35	34
13	27	32	34	35	37
14	33	33	35	37	36
15	35	37	32	35	39
16	33	33	27	31	30
17	35	34	34	30	32
18	32	33	30	30	33
19	25	27	34	27	28
20	35	35	36	33	30

(ii) Construct the R- and the $\overline{X}$-charts.

(iii) Construct the S- and the $\overline{X}$-charts.

Solution: Table 5.3 shows the vane-opening data with extra columns displaying the sample means, sample ranges, and sample standard deviations. The grand mean, average range, and average standard deviation are also listed at the bottom of the table. For a sample size of $n = 5$, we have $D_3 = 0$, $D_4 = 2.114$, $A_2 = 0.577$, $B_3 = 0$, $B4 = 2.089$, and $A_3 = 1.427$.

(i) Using Table 5.3, the process mean and standard deviation can be estimated as follows

$$\overline{\overline{X}} = 33.32 \quad \text{and} \quad \hat{\sigma} = \frac{\overline{R}}{d_2} = \frac{5.8}{2.326} = 2.4936$$

where $d_2 = 2.326$ for samples of size $n = 5$.

TABLE 5.3: Vane-Opening Data with Sample Means, Ranges, and Standard Deviations.

Sample	Data							
Number	X_1	X_2	X_3	X_4	X_5	$\overline{X}_i$	R_i	S_i
1	33	29	31	32	33	31.60	4	1.67
2	33	31	35	37	31	33.40	6	2.61
3	35	37	33	34	36	35.00	4	1.58
4	30	31	33	34	33	32.20	4	1.64
5	33	34	35	33	34	33.80	2	0.84
6	30	31	32	34	31	31.60	4	1.52
7	38	33	32	35	32	34.00	6	2.55
8	31	35	35	35	34	34.00	4	1.73
9	27	32	34	35	37	33.00	10	3.81
10	33	33	35	37	36	34.80	4	1.79
11	35	37	32	35	39	35.60	7	2.61
12	33	33	27	31	30	30.80	6	2.49
13	35	34	34	30	32	33.00	5	2.00
14	32	33	30	30	33	31.60	3	1.52
15	35	35	36	33	30	33.80	6	2.39
16	33	33	27	31	30	30.80	6	2.49
17	35	34	34	30	32	33.00	5	2.00
18	32	33	30	30	33	31.60	3	1.52
19	25	27	34	27	28	28.20	9	3.42
20	35	35	36	33	30	33.80	6	2.39
						$\overline{\overline{X}} = 33.32$	$\overline{R} = 5.8$	$\overline{S} = 2.345$

(ii) The upper control limit, center line, and lower control limit of the R-chart are

$$\begin{aligned} UCL &= D_4\bar{r} = 12.27 \\ CL &= \bar{r} = 5.8 \\ LCL &= D_3\bar{r} = 0 \end{aligned}$$

The R-chart is analyzed first to determine if it is stable. Figure 5.7(a) shows that there is an out-of-control point on the R-chart at sample (subgroup) 9. Assuming that the out-of-control point at sample 9 has an assignable cause, it can be discarded from the data.

The $\overline{X}$-chart can now be analyzed. The control limits of the $\overline{X}$-chart are

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{r} = 36.67 \\ CL &= \bar{\bar{x}} = 33.32 \\ LCL &= \bar{\bar{x}} - A_2\bar{r} = 29.97 \end{aligned}$$

Figure 5.3(b) shows that there are out-of-control points at samples 6, 8, 11, and 19. Assuming assignable cause, we can discard these samples from the data.

MATLAB code

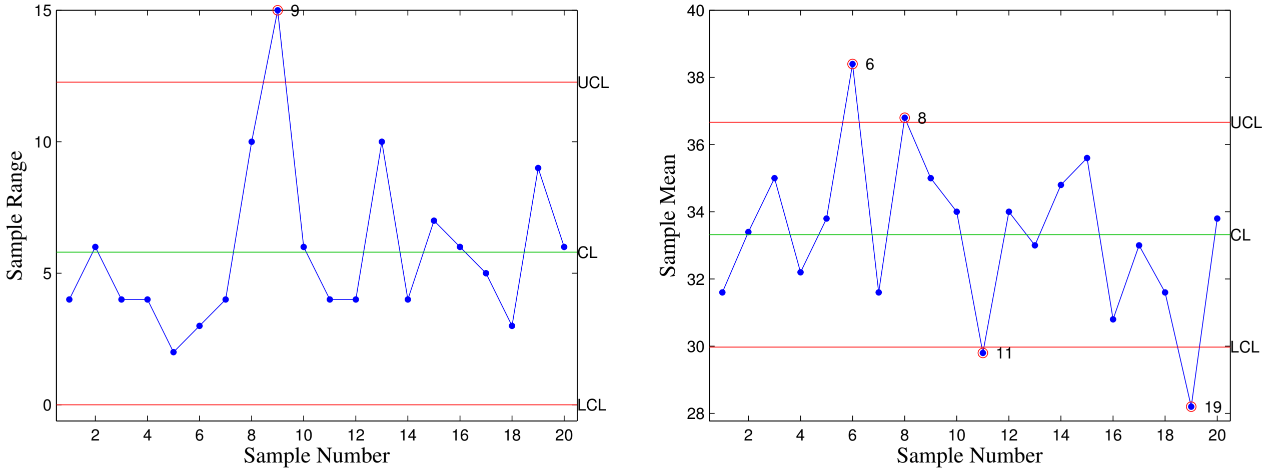
```
>> load vaneopening.mat; %load vane-opening data
>> [stats,plotdata]=controlchart(X,'chart',{'r','xbar'},'sigma','range'); %r & Xbar charts
```

Thus, if the out-of-control points at samples 6, 8, 9, 11, and 19 are discarded, then the new control limits are calculated using the remaining 15 samples. The revised mean range and grand mean are the given by

$$\bar{\bar{x}} = 33.21 \qquad \bar{r} = 5$$

The revised control limits for the new R-chart are

$$\begin{aligned} UCL &= D_4\bar{r} = 10.57 \\ CL &= \bar{r} = 5 \\ LCL &= D_3\bar{r} = 0 \end{aligned}$$



(a) R-chart

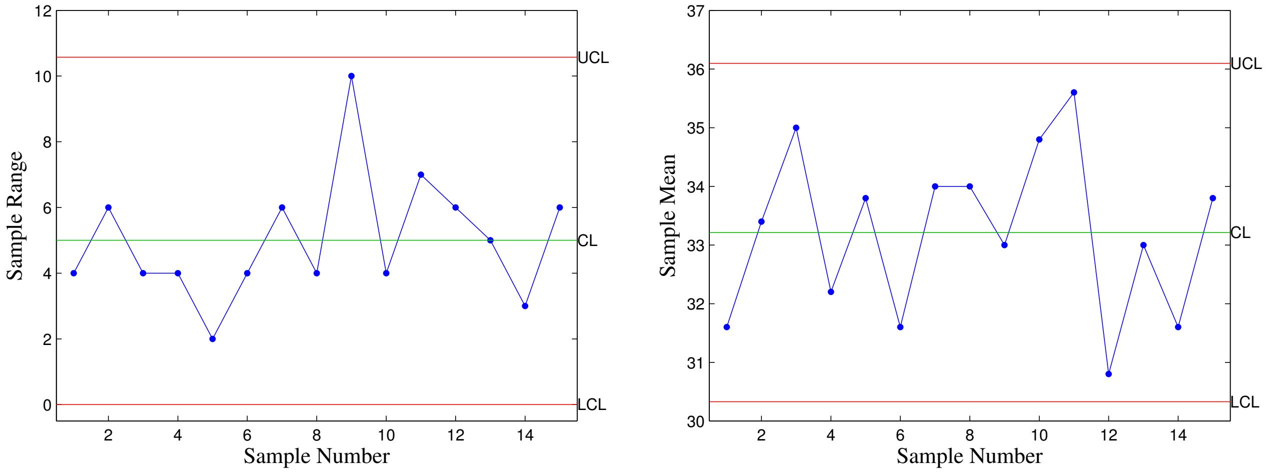
(b) $\overline{X}$-chart

FIGURE 5.7: R- and $\overline{X}$-charts for the vane-opening data.

and the revised control limits for the new $\overline{X}$-chart are

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{r} = 36.10 \\ CL &= \bar{\bar{x}} = 33.21 \\ LCL &= \bar{\bar{x}} - A_2\bar{r} = 30.33 \end{aligned}$$

The new $\overline{X}$- and R-charts are shown in Figure 5.8. Notice now that all the points fall within the limits, indicating that the process may be stable.



(a) R-chart

(b) $\overline{X}$-chart

FIGURE 5.8: Revised R- and $\overline{X}$-charts for the vane-opening data.

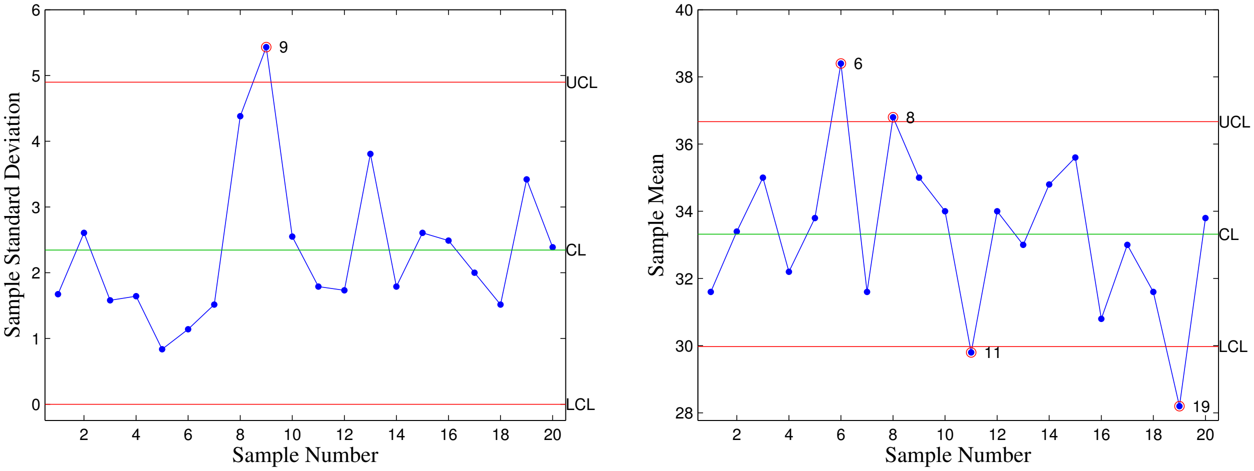
(iii) The upper control limit, center line, and lower control limit of the S-chart are given by

$$\begin{aligned} UCL &= B_4\bar{s} = 4.899 \\ CL &= \bar{s} = 2.345 \\ LCL &= B_3\bar{s} = 0 \end{aligned}$$

The upper control limit, center line, and lower control limit of the $\overline{X}$-chart are given by

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_3\bar{s} = 36.67 \\ CL &= \bar{\bar{x}} = 33.32 \\ LCL &= \bar{\bar{x}} - A_3\bar{s} = 29.97 \end{aligned}$$

We first examine the S-chart for signs of special variation. The S-chart are shown in Figure 5.9(a), where samples 6, 8, 11, and 19 are out-of-control. The $\overline{X}$-chart is shown in Figure 5.9(b), where the sample number 9 is out-of-control.



(a) S-chart

(b) $\overline{X}$-chart

FIGURE 5.9: S- and $\overline{X}$-charts for vane-opening data.

MATLAB code

```
>> load vaneopening.mat; %load vane-opening data
>> [stats,plotdata]=controlchart(X,'chart',{'s','xbar'},'sigma','std'); %s & Xbar charts
```

With samples 6, 8, 9, 11, and 19 deleted, the new upper control limit, center line, and lower control limit of the S-chart are given by

$$\begin{aligned} UCL &= B_4\bar{s} = 4.281 \\ CL &= \bar{s} = 2.049 \\ LCL &= B_3\bar{s} = 0 \end{aligned}$$

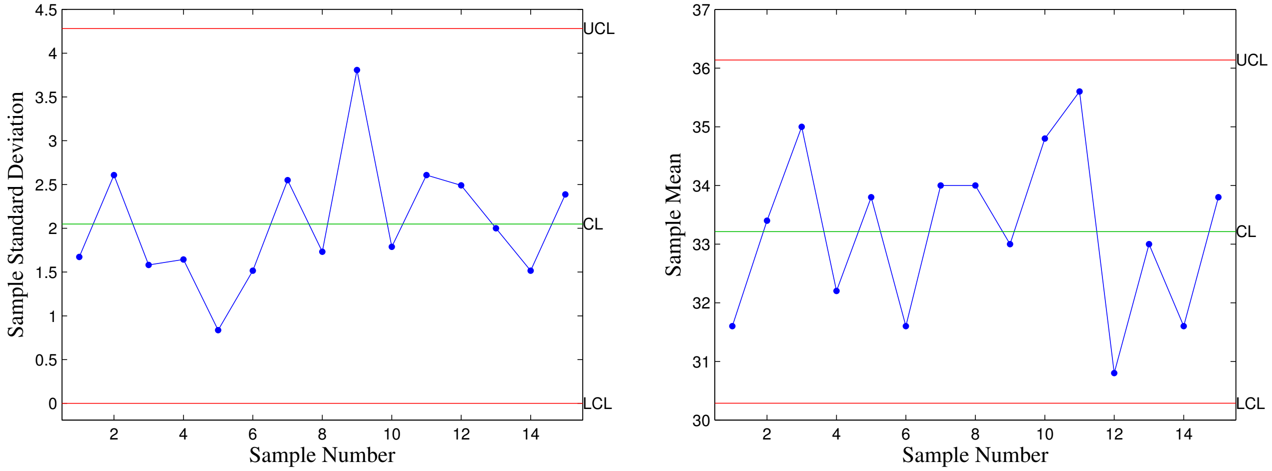
and the new upper control limit, center line, and lower control limit of the $\overline{X}$-chart are given by

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_3\bar{s} = 36.14 \\ CL &= \bar{\bar{x}} = 33.21 \\ LCL &= \bar{\bar{x}} - A_3\bar{s} = 30.29 \end{aligned}$$

As shown in Figure 5.10, the remaining plotted points on the S- and $\overline{X}$-charts indicate a stable process. ■

5.3 PROCESS CAPABILITY ANALYSIS

Process capability is the long-term performance level of the process after it has been brought under statistical control. In other words, process capability is the range over which the natural variation of the process occurs as determined by the system of common causes. A process may be in statistical control, but due to a high level of variation may not be



(a) s-chart

(b) $\overline{X}$-chart

FIGURE 5.10: Revised S- and $\overline{X}$-charts for vane-opening data.

capable of producing output that is acceptable to customers. A process capability study assumes the system is stable and that the data are normally distributed. A process is stable when only normal random variation is present.

A process capability analysis is simply the comparison of the distribution of a process output with the product tolerances. Control charts limits can be compared by specification limits to determine the process capability. These specification limits are often set by customers, management, and/or product designers. Moreover, specification limits are usually two-sided, with upper specification limit (USL) and lower specification limit (LSL); or can be one-sided, with either USL or LSL. Knowing the capability of your processes, we can specify better the quality performance requirements for new machines, parts and processes. The capability of a process centered on the desired mean can be measured using the **process capability potential** C_p, which is defined as the ratio between the specification spread ($USL - LSL$) and the process spread (6σ):

$$C_p = \frac{\text{Specification spread}}{\text{Process spread}} = \frac{USL - LSL}{6\sigma} \tag{5.3.1}$$

where σ is the process standard deviation. The idea is illustrated graphically in Figure 5.11. Note that the specification spread is the performance spread acceptable to customers, management, and/or product designers.
Let X be the process quality characteristic that we want to monitor. The performance of the process with respect to the specification limits USL and LSL are defined as follows:

$$P(X > USL) = \text{Percentage of nonconforming produced by the process at the upper end.}$$

$$P(X < LSL) = \text{Percentage of nonconforming produced by the process at the lower end.}$$

Thus, the total percentage of nonconforming produced by the process is defined as

$$P(X < USL \text{ or } X > USL) = 1 - P(LSL < X < USL) \tag{5.3.2}$$

Other capability indices that are frequently used in process capability analysis include:

- **Upper capability index:** $C_{pU} = (USL - \mu)/(3\sigma)$
- **Lower capability index:** $C_{pL} = (\mu - LSL)/(3\sigma)$
- **Process capability index:** $C_{pk} = \min(C_{pU}, C_{pL})$

These four measures of process capability quantify the degree to which your process produces output that meets the customer's specification, and can be used effectively to summarize process capability information in a convenient unitless system. Calculating the process capability measures requires knowledge of the process mean and standard deviation, μ and σ, which are usually estimated from data collected from the process. Assume m preliminary samples

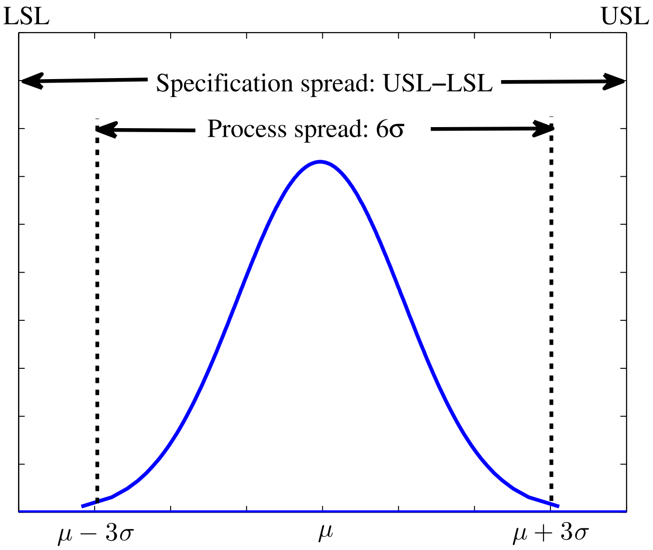


FIGURE 5.11: Specification spread vs. process spread.

TABLE 5.4: Process Capability Indices with $\hat{\sigma} = \overline{R}/d_2$.

Index	Estimated Equation
C_p	$(USL - LSL)/(6\hat{\sigma})$
C_{pU}	$(USL - \overline{\overline{X}})/(3\hat{\sigma})$
C_{pL}	$(\overline{\overline{X}} - LSL)/(3\hat{\sigma})$
C_{pk}	$\min(C_{pU}, C_{pL})$

(with equal sample size n) are available, then $\hat{\mu} = \overline{\overline{X}}$ is the grand mean and $\hat{\sigma} = \overline{R}/d_2$, where $\overline{R}$ is the mean of sample ranges. Table 5.4 shows a summary of the process capability measures. Note that $C_{pk} \leq C_p$, and are equal when $\overline{\overline{X}}$ is at target.

The lower and upper capability indices C_{pL} and C_{pU} are used when only one direction from the mean is important. The process capability index, C_{pk}, measures the distance of the process average $\overline{\overline{X}}$ from the closest specification. In other words, unlike C_p, the index C_{pk} takes process location into account. Moreover, C_{pk} can be calculated in situations where there is only one specification limit. Thus, C_{pk} can be used in place of the other three capability measures. Three possible cases can be considered:

Case 1: If $C_{pk} < 1$, the process in not capable of consistently producing product within the specifications. The process produces more than 2700 non-conforming units per million. It is impossible for the current process to meet specifications even when it is in statistical control. If the specifications are realistic, an effort must be immediately made to improve the process (i.e. reduce variation) to the point where it is capable of producing consistently within specifications.

Case 2: If $C_{pk} \geq 1.33$, the process in highly capable and produces less than 64 non-conforming units per million. C_{pk} values of 1.33 or greater are considered to be industry benchmarks.

Case 3: If $1 \leq C_{pk} < 1.33$, the process in barely capable and produces more than 64 but less than 2700 non-conforming units per million. This process has a spread just about equal to specification width. It should be noted that if the process mean moves to the left or the right, a significant portion of product will start falling outside one of the specification limits. This process must be closely monitored.

Example 5.7 *A pharmaceutical company carried out a process capability study on the weight of tablets produced and showed that the process was in-control with a process mean $\overline{\overline{X}} = 2504$ mg and a mean range $\overline{R} = 91$ mg from samples of size $n = 4$. Compute the process capability indices for the specifications limits 2800 mg and 2200 mg, and interpret your result.*

Solution: From the given information, we have $\overline{\overline{X}} = 2504$, $\overline{R} = 91$, $LSL = 2200$, $USL = 2800$, and $n = 4$. For a sample of size $n = 4$, Appendix Table V gives $d_2 = 2.059$. Thus,

$$\begin{aligned}
\hat{\sigma} &= \frac{\overline{R}}{d_2} = \frac{91}{2.059} = 44.1962 \\
C_p &= \frac{USL - LSL}{6\hat{\sigma}} = 2.263 \\
C_{pU} &= \frac{USL - \overline{\overline{X}}}{3\hat{\sigma}} = 2.232 \\
C_{pL} &= \frac{\overline{\overline{X}} - LSL}{3\hat{\sigma}} = 2.293 \\
C_{pk} &= \min(C_{pU}, C_{pL}) = 2.232
\end{aligned}$$

Since $C_{pk} = 2.232 \geq 1.33$, the process is highly capable. Moreover, C_{pU} is smaller that C_{pL}, which indicates that the process is skewed more to the high side. Thus, the process is highly capable of meetings the requirements but not centered. Some corrective action may have to be taken to centralize the process.

MATLAB code

```
n = 4; d2 = 2.059; %sample size equal 4
LSL = 2200; USL = 2800;
Xbarbar  = 2504; Rbar = 91;
sigma = Rbar/d2;
Cp = (USL - LSL)/(6*sigma);
Cpl = (Xbarbar-LSL)/(3*sigma);
Cpu = (USL - Xbarbar)/(3*sigma);
Cpk = min(Cpl,Cpu);
fprintf('Process Capability Indices:  Cp=%.3f, Cpl=%.3f, Cpu=%.3f, ...
         Cpk=%.3f\n',Cp,Cpl,Cpu,Cpk);
```

Example 5.8 *Assume that the vane-opening data (Example 5.6) are normally distributed and that the specifications are* 34 ± 6.5.

(i) *Determine the process capability index.*

(ii) *What proportion of the product will not meet specifications?*

Solution: From the given information, we have $LSL = 27.5$ and $USL = 40.5$. From the revised R- and $\overline{\overline{X}}$-charts in Example 5.6, we have $\overline{\overline{X}} = 33.21$ and $\overline{R} = 5$. Thus, the estimated process standard deviation is $\hat{\sigma} = \overline{R}/d_2 = 5/2.326 = 2.1496$.

(i) The process capability indices are given by

$$\begin{aligned}
C_p &= \frac{USL - LSL}{6\hat{\sigma}} = 1.0079 \\
C_{pU} &= \frac{USL - \overline{\overline{X}}}{3\hat{\sigma}} = 1.1299 \\
C_{pL} &= \frac{\overline{\overline{X}} - LSL}{3\hat{\sigma}} = 0.8859 \\
C_{pk} &= \min(C_{pU}, C_{pL}) = 0.8859
\end{aligned}$$

The C_{pk} value is quite low ($C_{pk} < 1$), which indicates that the process in not capable of consistently producing product within the specification. Since C_{pL} is smaller than C_{pU}, the process is skewed more to the low side and reflects a relatively poor capability in meeting the low side of the design specification.

(ii) The total percentage of nonconforming produced by the process is

$$\begin{aligned}\text{Percentage of Nonconforming} &= 1 - P(LSL < X < USL)\\ &= 1 - P\left(\frac{LSL - \overline{\overline{X}}}{\hat{\sigma}} < \frac{X - \overline{\overline{X}}}{\hat{\sigma}} < \frac{USL - \overline{\overline{X}}}{\hat{\sigma}}\right)\\ &= 1 - P(-2.6578 < Z < 3.3898)\\ &= 1 - (\Phi(3.3898) - \Phi(-2.6578)) = 0.0043\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution $N(0,1)$. Thus, the proportion of product not meeting specifications is 0.43%, which is quite low. Although we found that the process is in control, it is however not capable of meeting the stated specifications. In this case, common causes must be found for process improvement.■

Example 5.9 *A dimension has specifications of* 2.125 ± 0.005*. Data from the process indicate that the distribution is normally distributed, and the* $\overline{X}$*- and* R*-charts indicate that the process is stable. The control chart used a sample of size five and it is found that* $\overline{\overline{X}} = 2.1261$ *and* $\overline{R} = 0.0055$*. Determine the fraction of the manufactured product that will have this particular dimension outside the specification limits.*

Solution: From the given information, we have $LSL = 2.120$ and $USL = 2.130$, $\overline{\overline{X}} = 2.1261$, and $\overline{R} = 0.0055$. For a sample of size $n = 5$, Appendix Table V gives $d_2 = 2.326$. Thus, the estimated process standard deviation is $\hat{\sigma} = \overline{R}/d_2 = 0.0055/2.326 = 0.00236$.
The total percentage of nonconforming produced by the process is

$$\begin{aligned}\text{Percentage of Nonconforming} &= 1 - P(LSL < X < USL)\\ &= 1 - P\left(\frac{LSL - \overline{\overline{X}}}{\hat{\sigma}} < \frac{X - \overline{\overline{X}}}{\hat{\sigma}} < \frac{USL - \overline{\overline{X}}}{\hat{\sigma}}\right)\\ &= 1 - P(-2.58 < Z < 1.65)\\ &= 1 - (\Phi(1.65) - \Phi(-2.58)) = 0.0544\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution $N(0,1)$. Thus, approximately 5.44% of the products will fall outside specification for this quality characteristic. ■

5.4 Individual Control Charts

Some situations exist in which the sample consists of a single observation, that is, the sample size is equal to 1. A sample of size one can occur when production is very slow or costly, and it is impractical to allow the sample size to be greater than one. The individual control charts, I- and MR-charts, for variable data are appropriate for this type of situation. The I-chart (also called X-chart) serves the same function as the $\overline{X}$-chart except that now X is the value of the individual measurement. Assuming that $X \sim N(\mu_x, \sigma_x^2)$, the control limits of the I-chart are then given by

$$\begin{aligned}UCL &= \mu_x + 3\sigma_x\\ CL &= \mu_x\\ LCL &= \mu_x - 3\sigma_x\end{aligned} \tag{5.4.1}$$

Since μ_x and σ_x are unknown, they need to be estimated. Suppose m preliminary observations (samples) are available, each of size one. Then the process mean μ_x can be estimated as the average of the individual measurements $\hat{\mu}_x = \bar{x} = (1/m)\sum_{i=1}^{m} x_i$.

Since only individual measurements are available, the moving ranges $MR_i = |X_i - X_{i-1}|$, $i = 2, \ldots, m$ between two successive samples X_{i-1} and X_i need to be calculated to estimate the process variability (standard deviation) σ as follows:

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} \tag{5.4.2}$$

where

$$\overline{MR} = \frac{1}{m-1}\sum_{i=2}^{m} MR_i \tag{5.4.3}$$

Note that since the data are taken as pairs $\{X_{i-1}, X_i\}$ to calculate the moving ranges $MR_i = |X_i - X_{i-1}|$, the value of d_2 is then equal to 1.128. Notice that division is done by $m-1$ since only $m-1$ moving range values are calculated (there is no moving range for subgroup 1), where m is the number of observations.

Control chart for individuals (I-chart):
The upper control limit, center line, and lower control limit for the I-chart are given by

$$\begin{aligned} UCL &= \bar{x} + 3\frac{\overline{mr}}{d_2} = \bar{x} + 3\frac{\overline{mr}}{1.128} \\ CL &= \bar{x} \\ LCL &= \bar{x} - 3\frac{\overline{mr}}{d_2} = \bar{x} - 3\frac{\overline{mr}}{1.128} \end{aligned} \tag{5.4.4}$$

Example 5.10 *A company is manufacturing high precision tubes. Quality Control department is interested to determine if the production process in under control. For simplicity, we assume that from each batch is measured the outer diameter of a random selected tube. Measured data are given in Table 5.5. Construct the I-chart.*

TABLE 5.5: Outer diameter of tubes.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Diameter	99.82	99.63	99.89	99.45	100.03	99.76	100.23	99.81	99.91	100.12	100.05	99.78	100.01	100.04	99.95

Solution: The upper control limit, center line, and lower control limit for the I-chart are given by

$$\begin{aligned} UCL &= \bar{x} + 3\frac{\overline{mr}}{d_2} = 99.8987 + 3\frac{0.2593}{1.128} = 100.588 \\ CL &= \bar{x} = 99.8987 \\ LCL &= \bar{x} - 3\frac{\overline{mr}}{d_2} = 99.8987 - 3\frac{0.2593}{1.128} = 99.2093 \end{aligned}$$

```
X=[99.82 99.63 99.89 99.45 100.03 99.76 100.23 99.81 ...
   99.91 100.12 100.05 99.78 100.01 100.04 99.95]; %outer diameter data
[stats,plotdata]=controlchart(X,'chart','i','width',2); %plot I-chart
fprintf('Control limits for i-chart: UCL=%g, CL=%g, LCL=%g\n',...
         plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
```

The I-chart is shown in Figure 5.12, where all samples appear to be in-control. Thus, the process of producing the tubes is considered to be in statistical control. ■

On the other hand, the MR-chart is used to monitor the process variability. it can be shown that μ_{MR} and σ_{MR} can be estimated as

$$\hat{\mu}_{MR} = \overline{MR} \quad \text{and} \quad \hat{\sigma}_{MR} = \frac{d_3}{d_2}\overline{MR} \tag{5.4.5}$$

Control chart for moving ranges (MR-chart):
The upper control limit, center line, and lower control limit for the MR-chart are given by

$$\begin{aligned} UCL &= D_4\overline{mr} = 3.267\,\overline{mr} \\ CL &= \overline{mr} \\ LCL &= D_3\overline{mr} = 0 \end{aligned} \tag{5.4.6}$$

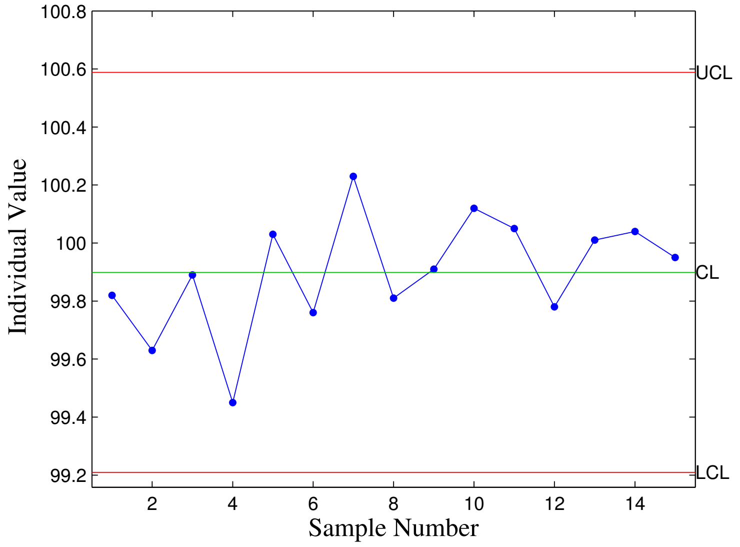


FIGURE 5.12: *I*-chart for outer diameter data.

It is important to note that the moving range control chart can not be interpreted in the same way as the R chart presented earlier, with respect to patterns or trends. Patterns or trends identified on the moving range chart do not necessarily indicate that the process is out of control. The moving ranges $MR_i = |X_i - X_{i1}|$ are correlated. There is a natural dependency between successive MR_i values.

Example 5.11 *Using the data of Table 5.5, construct the MR-chart.*

Solution: The upper control limit, center line, and lower control limit for the MR-chart are given by

$$\begin{aligned} UCL &= D_4\overline{mr} = (3.267)(0.2593) = 0.8471 \\ CL &= \overline{mr} = 0.2593 \\ LCL &= D_3\overline{mr} = 0 \end{aligned}$$

```
[stats,plotdata]=controlchart(X,'chart','mr','width',2); %plot MR-chart
fprintf('Control limits for mr-chart:  UCL=%g, CL=%g, LCL=%g\n',...
         plotdata.ucl(2),plotdata.cl(2),plotdata.lcl(2));
```

The *mr*-chart is shown in Figure 5.13, where all samples appear to be in-control. ■

Example 5.12 *Packages of a particular instant dry food are filled by a machine and weighed. The weights (in ounces) for 15 successive packages have been collected and are displayed in Table 5.6. Construct the I- and MR-charts.*

TABLE 5.6: Weights for dry food packages.

Bottle	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Weight	19.85	19.92	19.93	19.26	20.36	19.96	19.87	19.80	20.40	19.98	20.17	19.81	20.21	19.64	20.15

Solution: The moving ranges are calculated using $MR_i = |X_i - X_{i-1}|$. To illustrate, consider the first moving range at subgroup 2:

$$MR_2 = |X_2 - X_1| = 19.92 - 19.85 = 0.07$$

The remaining moving ranges are calculated accordingly and are given in Table 5.7.

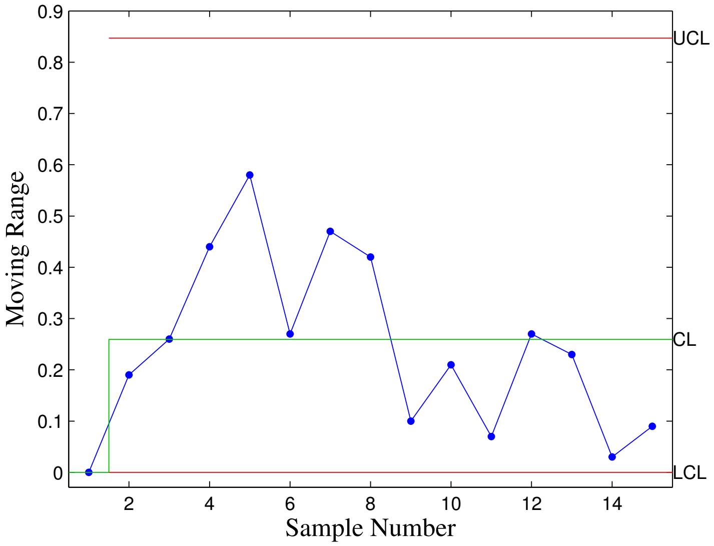


FIGURE 5.13: *MR*-chart for outer diameter data.

TABLE 5.7: Weights for dry food packages with moving ranges.

Bottle	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Weight	19.85	19.92	19.93	19.26	20.36	19.96	19.87	19.80	20.40	19.98	20.17	19.81	20.21	19.64	20.15
Moving Range	-	0.07	0.01	0.67	1.10	0.40	0.09	0.07	0.60	0.42	0.19	0.36	0.40	0.57	0.51

The upper control limit, center line, and lower control limit for the *I*-chart are given by

$$\begin{aligned} UCL &= \bar{x} + 3\frac{\overline{mr}}{d_2} = 19.954 + 3\frac{0.39}{1.128} = 20.9909 \\ CL &= \bar{x} = 19.954 \\ LCL &= \bar{x} - 3\frac{\overline{mr}}{d_2} = 19.954 - 3\frac{0.39}{1.128} = 18.9171 \end{aligned}$$

The upper control limit, center line, and lower control limit for the *MR*-chart are given by

$$\begin{aligned} UCL &= D_4\overline{mr} = (3.267)(0.39) = 1.27395 \\ CL &= \overline{mr} = 0.39 \\ LCL &= D_3\overline{mr} = 0 \end{aligned}$$

The *I*- and *MR*-charts are displayed in Figure 5.14, which shows that all samples fall within the control limits. Thus, the process appears to be in statistical control. ■

5.5 CUMULATIVE SUM CONTROL CHART (CUSUM-CHART)

Although Shewhart charts with 3σ limits can quickly detect large process changes, they are ineffective for small, sustained process changes (for example, changes smaller than 1.5σ). An alterative control chart has been developed to detect small shifts in the process mean is the so-called cumulative sum (CUSUM) control chart (CUSUM-chart), which is more sensitive to small shifts in the process because it is based on not only the current observation, but also the most recent past observations. Moreover, a CUSUM-chart is especially effective with samples of size $n = 1$.

Suppose m preliminary samples of size $n \geq 1$ are available. Then, the CUSUM-chart plots the cumulative sums C_i of deviations of the observations from some target mean μ_0

$$C_i = \sum_{j=1}^{i}(\bar{x}_j - \mu_0) = \left[\sum_{j=1}^{i-1}(\bar{x}_j - \mu_0)\right] + (\bar{x}_i - \mu_0) = C_{i-1} + (\bar{x}_i - \mu_0), \quad i = 1, \ldots, m \tag{5.5.1}$$

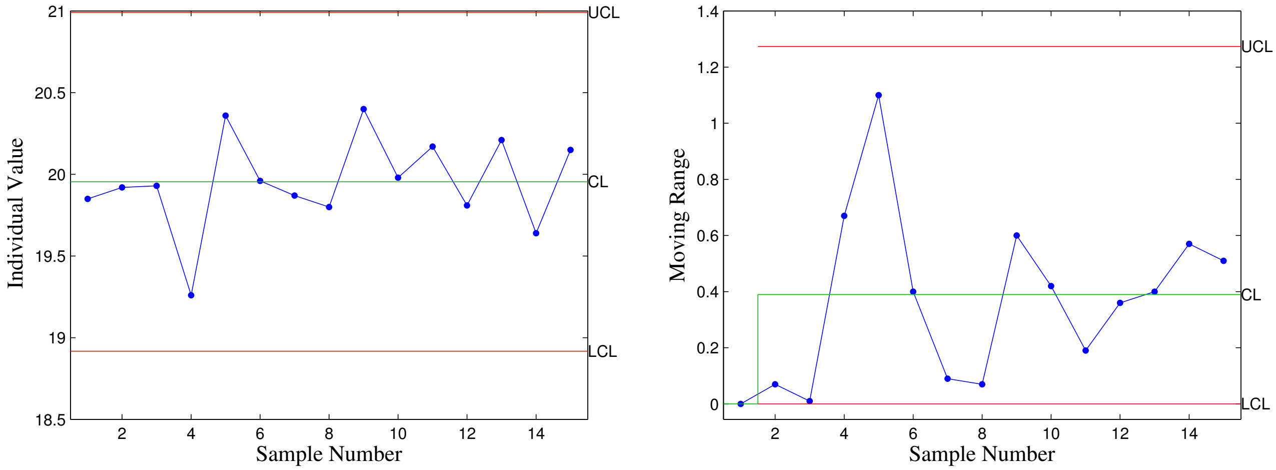


FIGURE 5.14: I- and MR-charts for package weights.

against the subgroup (sample i), where $\bar{x}_j$ is the mean of the j-th sample ($j = 1, \ldots, i$).

As long as the process remains "in control" at the target mean μ_0, the cumulative sums, C_i, will be approximately zero. Otherwise, if the process shifts away from the target mean μ_0, then C_i become increasingly large in absolute value. The tabular CUSUM for monitoring the process mean involves two statistics, C_i^+ and C_i^-, defined as

$$\begin{aligned} C_i^+ &= \max[0, \bar{x}_i - (\mu_0 + K) + C_{i-1}^+] \\ C_i^- &= \max[0, (\mu_0 - K) - \bar{x}_i + C_{i-1}^-] \end{aligned} \tag{5.5.2}$$

where

- C_i^+ is the accumulation of deviations above the target mean, with initial value $C_0^+ = 0$
- C_i^- is the accumulation of deviations below the target mean, with initial value $C_0^- = 0$
- K is called the reference value given by $K = |\mu_1 - \mu_0|/2$, where μ_1 is the out-of-control mean that we are interested in detecting.

A deviation from the target that is larger than K increases either the one-sided upper CUSUM C_i^+ or the one-sided lower CUSUM C_i^-. If the out-of-control mean μ_1 is unknown, there are methods for determining the value K. In this situation we can let $K = k\sigma_{\bar{x}}$, where k is some constant chosen so that a particular shift is detected, and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ with σ denoting the process standard deviation. For example, say a shift from target of 1 standard deviation is important to detect (i.e., detect whether the target has shifted to $\mu_0 + 1\sigma_{\bar{x}}$ or $\mu_0 - 1\sigma_{\bar{x}}$, then $k = 1$, and $K = 1\sigma_{\bar{x}}$). If the process standard deviation is not known, it must be estimated from the data provided. The two-sided CUSUM-chart plots the values of C_i^+ and C_i^i for each sample i. A control limit violation (out-of-control) occurs when either C_i^+ or C_i^i exceeds a specified control limit (or threshold) $H = h\sigma_{\bar{x}}$, where h is typically equal to 5.

For simplicity of CUSUM limits calculation, it is preferable to use the standardized value $y_i = (\bar{x}_i - \mu_0)/\sigma_{\bar{x}} = (\bar{x}_i - \mu_0)/(\sigma/\sqrt{n})$ of the variable x_i. Since the value of σ is unknown, it is usually estimated as $\hat{\sigma} = \overline{MR}/d_2$ for individual observations, and as $\hat{\sigma} = \overline{R}/d_2$ for samples of size $n > 1$.

Control chart for standardized cumulative sums (CUSUM-chart):
The one-sided upper and lower CUSUMs of the standardized CUSUM-chart are given by

$$\begin{aligned} C_i^+ &= \max[0, y_i - k + C_{i-1}^+] \\ C_i^- &= \max[0, -k - y_i + C_{i-1}^-] \end{aligned} \tag{5.5.3}$$

The upper control limit, center line, and lower control limit of the CUSUM-chart are given by

$$\begin{aligned} UCL &= h \\ CL &= 0 \\ LCL &= -h \end{aligned} \tag{5.5.4}$$

Example 5.13 *The data given in Table 5.8 are average readings from a process, taken every hour. (Read the observations down, from left). The target value for the mean is* $\mu_0 = 160$.

TABLE 5.8: Process readings.

159.0480	160.2766	160.3368	162.0092	162.2135
156.8969	159.2432	157.7590	159.2716	164.2690
159.1034	160.2861	160.8015	160.2469	159.7530
156.8976	163.2101	162.4781	162.5760	158.2998
161.8597	162.6982	162.8001	161.5964	158.6755
161.8039	159.1031	157.9376	160.6725	159.1114

1. *Estimate the process standard deviation*
2. *Set up and apply a tabular CUSUM for this process, using standardized values* $h = 5$ *and* $k = 0.5$. *Interpret this chart.*

Solution:

1. Since the data are individual observations, an estimate the process standard deviation is then given by

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{2.0112}{1.128} = 1.7830$$

```
X=[159.0480 156.8969 159.1034 156.8976 161.8597 161.8039 160.2766 159.2432 ...
   160.2861 163.2101 162.6982 159.1031 160.3368 157.7590 160.8015 162.4781 ...
   162.8001 157.9376 162.0092 159.2716 160.2469 162.5760 161.5964 160.6725 ...
   162.2135 164.2690 159.7530 158.2998 158.6755 159.1114];
mu0 = 160; %target value for the mean
MR = slidefun(@range,2,X); %moving range
MR2bar = mean(MR(2:end));
d2 = 1.128; %from Table in the Appendix
sigmahat = MR2bar/d2; %estimate of the process standard deviation
h = 5;
k = 0.5;
[Cplus,Cminus] = cusumchart(X,mu0,sigmahat,h,k); %plot CUSUM-chart
```

2. The CUSUM-chart is shown in Figure 5.15, where the process appears to be in-control. So there does not appear to have been a shift of 0.5σ from the target value of 160. ■

5.6 Exponentially Weighted Moving Average Control Chart (EWMA-chart)

Another alterative control chart that is generally used for detecting small shifts in the process mean is the so-called exponentially weighted moving average (EWMA) chart, and it plots weighted moving average values. Like the CUSUM-chart, the EWMA-chart is also preferred when the samples are of size $n = 1$ (i.e. individual measurements). Suppose

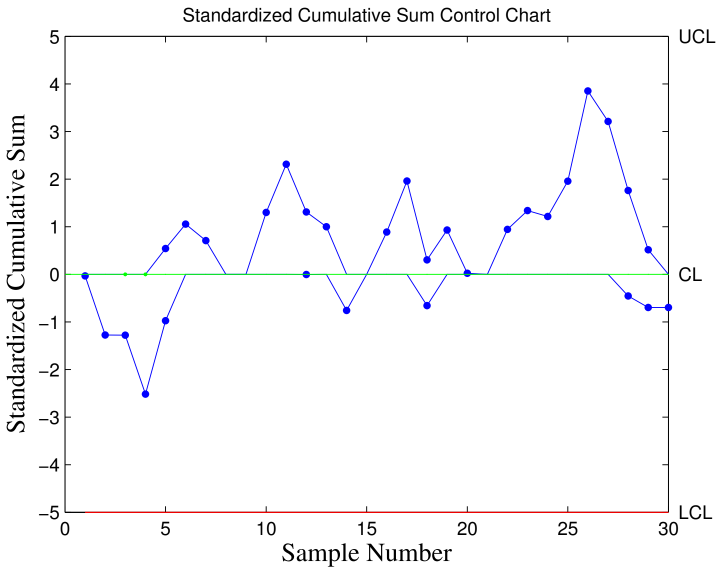


FIGURE 5.15: CUSUM-chart for process readings.

m preliminary samples of size $n \geq 1$ are available. The EWMA statistic is defined by

$$z_i = \lambda \bar{x}_i + (1-\lambda) z_{i-1}, \quad i = 1, \ldots, m \tag{5.6.1}$$

where

- the weighting factor, $0 < \lambda \leq 1$, determines the depth of memory for the EWMA
- $\bar{x}_i$ is the most current observation (i.e. average of the sample at time i)
- z_{i-1} is the previous EWMA statistic in which the initial value z_0 is equal to the process target mean μ_0. If μ_0 is unknown, then $z_0 = \bar{\bar{x}}$.

The EWMA-chart plots z_i against the sample i. Let σ be the process standard deviation (i.e. $\sigma = \sigma_{\bar{x}}$). It can be shown that for large i, the standard deviation of z_i is given by

$$\sigma_{z_i} = \sigma_{\bar{x}} \sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]} \approx \frac{\sigma}{\sqrt{n}} \sqrt{\frac{\lambda}{2-\lambda}}, \quad \text{as} \quad i \to \infty \tag{5.6.2}$$

Note that large i means that EWMA control chart has been running for several time periods.

Control chart for the exponentially weighted moving average (EWMA-chart):
The upper control limit, center line, and lower control limit of the EWMA-chart are given by

$$\begin{aligned} UCL &= \mu_0 + L \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{\lambda}{2-\lambda}} \\ CL &= \mu_0 \\ LCL &= \mu_0 - L \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{\lambda}{2-\lambda}} \end{aligned} \tag{5.6.3}$$

where L is the width of the control limits. For individual observations, we use $\hat{\sigma} = \overline{MR}/d_2$; and for samples of size $n > 1$, we use $\hat{\sigma} = \overline{R}/d_2$.

The values of the parameters L and λ can have a considerable impact on the performance of the chart. The parameter λ determines the rate at which past (historical) data enter into the calculation of the EWMA statistic. A value of $\lambda = 1$

implies that only the most recent observation influences the EWMA. Thus, a large value of λ gives more weight to recent data and less weight to historical data, while a small value of λ gives more weight to historical data. Although the choice of λ and L is arbitrary. In practice, the values $0.05 \leq \lambda \leq 0.25$ and $2.6 \leq L \leq 3$ work well. Note that $L = 3$ matches other control charts, but it may be necessary to reduce L slightly for small values of λ.

Example 5.14 *Using the data of Table 5.8:*

1. *Apply an EWMA control chart to these data using $\lambda = 0.15$ and $L = 2.7$*

Solution: The EWMA-chart is shown in Figure 5.16, where the process is out-of-control at sample number 26. ■

```
lambda = 0.15;
L = 2.7;
ewmachart(X,lambda,L,mu0,sigmahat); %plot EWMA-chart
```

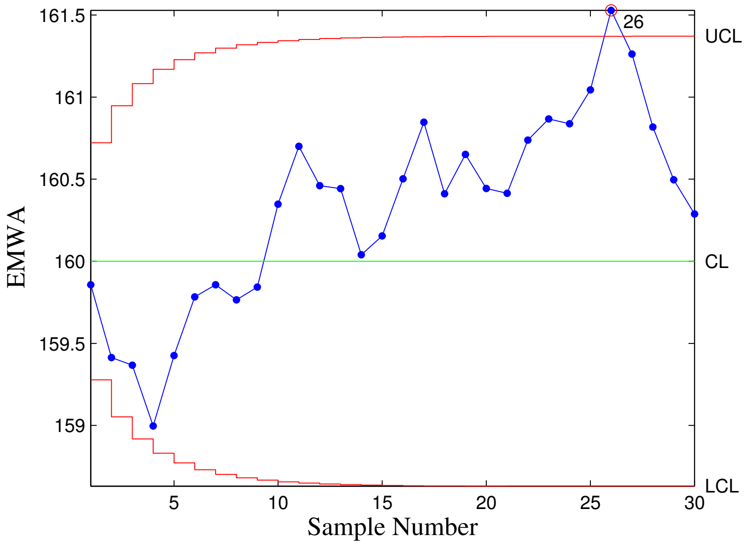


FIGURE 5.16: EWMA-chart for process readings.

Example 5.15 *The concentration of a chemical product is measured by taking four samples from each batch of material. The average concentration of these measurements is shown for the last 20 batches in in Table 5.9. Assume the target value of concentration for this process is 100.*

TABLE 5.9: Concentration Measurements.

Batch	Concentration	Batch	Concentration
1	104.50	11	95.40
2	99.90	12	94.50
3	106.70	13	104.50
4	105.20	14	99.70
5	94.80	15	97.70
6	94.60	16	97.00
7	104.40	17	95.80
8	99.40	18	97.40
9	100.30	19	99.00
10	100.30	20	102.60

(i) *Estimate the process standard deviation*

(ii) *Set up and apply a tabular CUSUM for this process, using standardized values $h = 5$ and $k = 0.5$. Does the process appear to be in control at the target?*

(iii) *Apply an EWMA control chart to these data using $\lambda = 0.1$ and $L = 2.7$. Interpret this chart.*

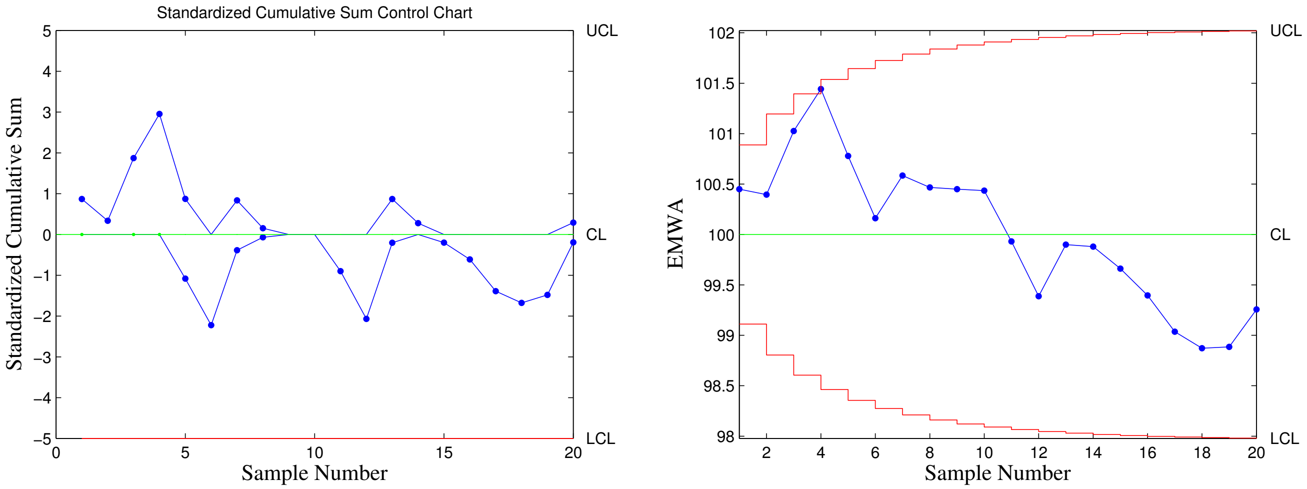
Solution:

(i) Since the data are individual observations, an estimate the process standard deviation is then given by

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{3.71}{1.128} = 2.29$$

(ii) The CUSUM-chart is given in Figure 5.17(a), which shows that the process is in-control.

(iii) The EWMA-chart is given in Figure 5.17(b), which shows that the process is in-control.



(a) (b)

FIGURE 5.17: CUSUM- and EWMA-charts for concentration measurements.

Example 5.16 *Packages of a particular instant dry food are filled by a machine and weighed. The weights (in ounces) for 24 successive packages have been collected and are displayed in Table 5.10. Assume the target mean weight for this process is 20 ounces.*

TABLE 5.10: Weights for dry food packages.

Package	Weight	Package	Weight
1	20.26	13	20.30
2	19.97	14	19.77
3	19.76	15	20.40
4	19.72	16	19.98
5	19.69	17	19.91
6	19.85	18	20.18
7	19.96	19	20.08
8	20.03	20	20.05
9	20.06	21	20.20
10	19.71	22	19.90
11	19.68	23	19.95
12	19.94	24	20.12

(i) *Estimate the process standard deviation* *(ii) Set up and apply a tabular CUSUM for this process, using standardized values $h = 5$ and $k = 0.5$. Does the process appear to be in control at the target?*

(iii) Apply an EWMA control chart to these data using $\lambda = 0.1$ and $L = 2.7$. Interpret this chart.

Solution:

(i) Since the data are individual observations, an estimate the process standard deviation is then given by

$$\hat{\sigma} = \frac{\overline{MR}}{d_2} = \frac{0.20}{1.128} = 0.18$$

(ii) The CUSUM-chart is given in Figure 5.18(a), which shows that the process is in-control. Thus, there does not appear to have been a shift of 0.5σ from the target value of 20 ounces.

(iii) The EWMA-chart is given in Figure 5.18(b), which shows that the process is in-control since all EWMA statistics fall within the control limits.

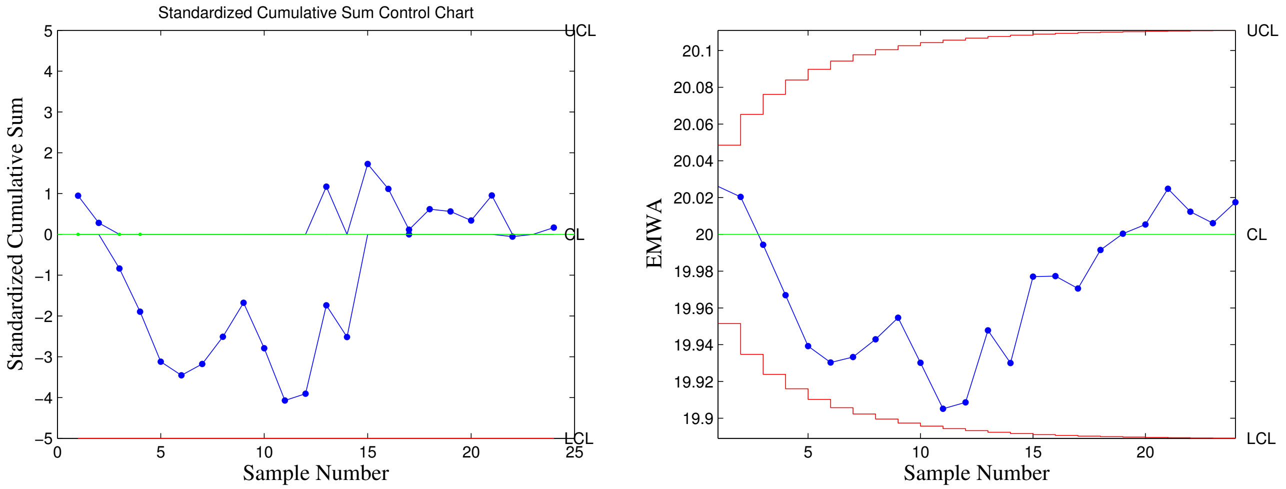


FIGURE 5.18: CUSUM- and EWMA-charts for dry food packages.

5.7 Control Charts for Attributes

In quality control, a defective quality characteristic is called a defect or non-conformance, whereas a unit that has at least one defect is called a defective or nonconforming unit. In other words, a defective unit may have more than one defect.

5.7.1 Control Chart for Proportion Defective: p-chart

The proportion or fraction defective in a population is defined as the ratio of the number of defective items in the population to the total number of items in that population. The goal of the fraction nonconforming control chart is to monitor the proportion of defective units (fraction defective) for a process of interest using the data collected over m samples (subgroups) each of size n. The p-chart plots the sample proportions (fraction defectives) vs the sample number. If D is the number of units that are defective in a random sample of size n, then $D \sim bino(n, p)$. The sample proportion defective $\hat{P} = D/n$ is the ratio of the number of defective units in the sample, D, to the sample size n. The mean and standard deviation of $\hat{P}$ are $\mu_{\hat{P}} = p$ and $\sigma_{\hat{P}} = \sqrt{p(1-p)/n}$, respectively.

Since p is unknown, it must be estimated from the available data. Suppose m preliminary samples are available, each of size n. If D_i is the number of defectives in the i-th sample, then the fraction defective of the i-th sample is $\hat{P}_i = D_i/n$, $i = 1, \ldots, m$. An unbiased estimator of $\mu_{\hat{P}}$ is the mean $\overline{P}$ of the fraction defectives of m samples,

$$CL = \overline{P} = \frac{1}{m}\sum_{i=1}^{m}\hat{P}_i = \frac{1}{mn}\sum_{i=1}^{m}D_i$$

Since $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, the upper and control limits may be expressed as

$$UCL = \overline{P} + 3\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}$$

and

$$LCL = \overline{P} + 3\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}$$

Control chart for proportion defective (p-chart):
The upper control limit, center line, and lower control limit of the p-chart are given by

$$\begin{aligned} UCL &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ CL &= \bar{p} \\ LCL &= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \end{aligned} \tag{5.7.1}$$

Example 5.17 *A quality control inspector wishes to construct a fraction-defective control chart for a light bulb production line. Packages containing 1000 light bulbs are randomly selected, and all 1000 bulbs are light-tested. The results of the tests are given in Table 5.11. Construct and plot a p-chart.*

TABLE 5.11: Number of defectives observed in samples of 1000 light bulbs.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13
Number of Defectives	9	12	13	12	11	9	7	0	12	8	9	7	11

Solution: The center line CL is given by

$$CL = \bar{p} = \frac{\sum(\text{Number of Defectives})}{(13)(1000)} = \frac{120}{13000} = 0.0092$$

The upper and lower control limits are given by

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.0183$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.0001$$

```
X = [9 12 13 12 11 9 7 0 12 8 9 7 11]; % Number of defectives
n = 1000; %total number of inspected items
[stats,plotdata]=controlchart(X,'chart','p','unit',n); %plot p-chart
fprintf('Control limits for p-chart: UCL=%g, CL=%g, LCL=%g\n', ...
          plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
```

The p-chart, shown in Figure 5.19, indicates that the sample number 8 is outside the control limits. ■

Example 5.18 *When a coupon redemption process is in control, then a maximum of 3% of the rebates are done incorrectly, for a maximum acceptable proportion of errors of 0.03. For 20 sequential samples of 100 coupon redemptions each, an audit reveals that the number of errors found in the rational subgroup samples are given in Table 5.12. Construct and plot a p-chart.*

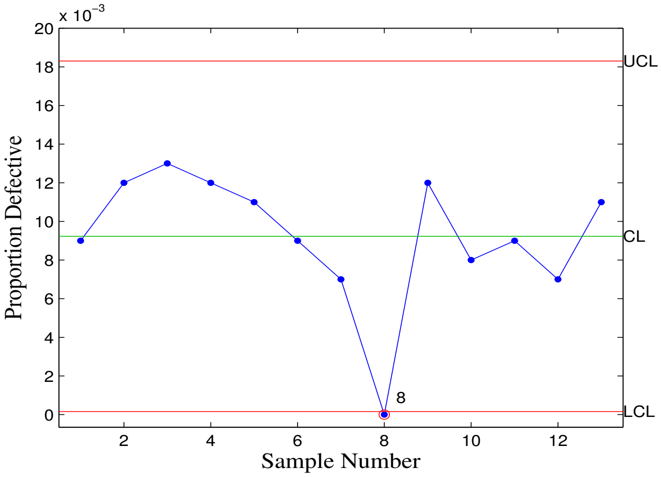


FIGURE 5.19: p-chart for lightbulb data.

TABLE 5.12: Number of errors observed in samples of 100 coupon redemptions.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number of Errors	2	2	3	6	1	3	6	4	7	2	5	6	3	2	4	5	3	8	1	4

Solution: The center line CL is given by

$$CL = \bar{p} = \frac{\sum(\text{Number of Errors})}{(20)(100)} = \frac{77}{2000} = 0.0385$$

The upper and lower control limits are given by

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.09622$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0$$

The p-chart is shown in Figure 5.20. As can be observed in the figure, only common-cause variation is included in the chart, and the process appears to be stable.

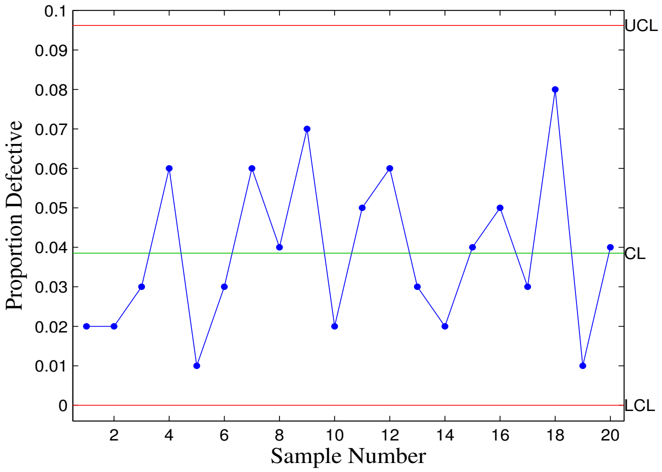


FIGURE 5.20: p-chart for the coupon data.

5.7.2 Control Chart for Number of Defectives: np-chart

The np control chart is a slight variation of the p-chart, except now the actual number of defective items D_i are plotted on the control chart against the sample number, where $i = 1, \ldots, m$. The control limits are based on the number of defective units instead of the fraction defective. The np-chart and the p-chart will give the same resulting information. That is, if the p-chart indicates an out-of-control situation for a process, then the np-chart for the same data will also signal out-of-control. For the np-chart the average fraction defective is estimated as

$$\overline{P} = \frac{1}{mn} \sum_{i=1}^{m} D_i$$

Control chart for number of defectives (np-chart):
The upper control limit, center line, and lower control limit of the np-chart are given by

$$\begin{aligned} UCL &= n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} \\ CL &= n\bar{p} \\ LCL &= n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} \end{aligned} \tag{5.7.2}$$

Example 5.19 *Using the data in Table 5.11, construct and plot an np-chart.*

Solution: The center line CL is given by

$$CL = n\bar{p} = 9.23077$$

The upper and lower control limits are given by

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 18.3033$$

$$LCL = \bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 0.15828$$

```
X = [9 12 13 12 11 9 7 0 12 8 9 7 11]; % Number of defectives
n = 1000; %total number of inspected items
[stats,plotdata]=controlchart(X,'chart','np','unit',n); %plot np-chart
fprintf('Control limits for np-chart: UCL=%g, CL=%g, LCL=%g\n', ...
         plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
```

The np-chart, shown in Figure 5.21, indicates that the sample number 8 is outside the control limits. ■

Example 5.20 *For each of the 15 days, a number of magnets used in electric relays are inspected and the number of defectives is recorded. The total number of magnets tested is 15,000. The results are given in Table 5.20. Construct and plot the p-and np-charts.*

Table 5.13: Number of defectives observed in samples of 1000 magnets.

Day Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Defectives	201	152	190	214	186	193	183	225	168	182	174	182	182	169	187

Solution: Since the total number of defectives in 15 days is 15000, the average sample size is $n = 15000/15 = 1000$.

- Control limits for the p-chart: $UCL = 0.22277$, $CL = 0.185867$, $LCL = 0.148963$
- Control limits for the np-chart: $UCL = 222.77$, $CL = 185.867$, $LCL = 148.963$

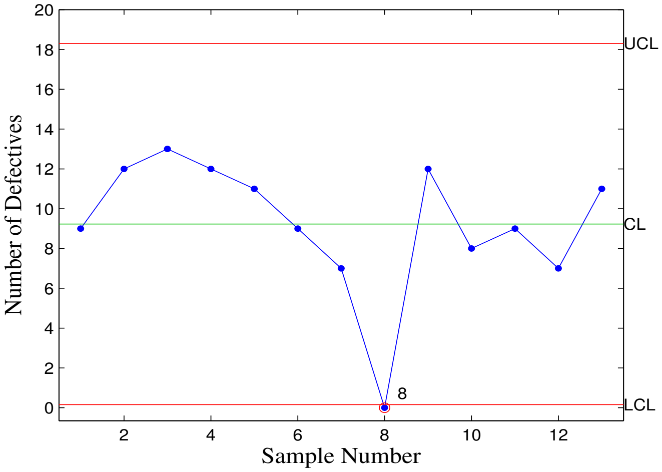
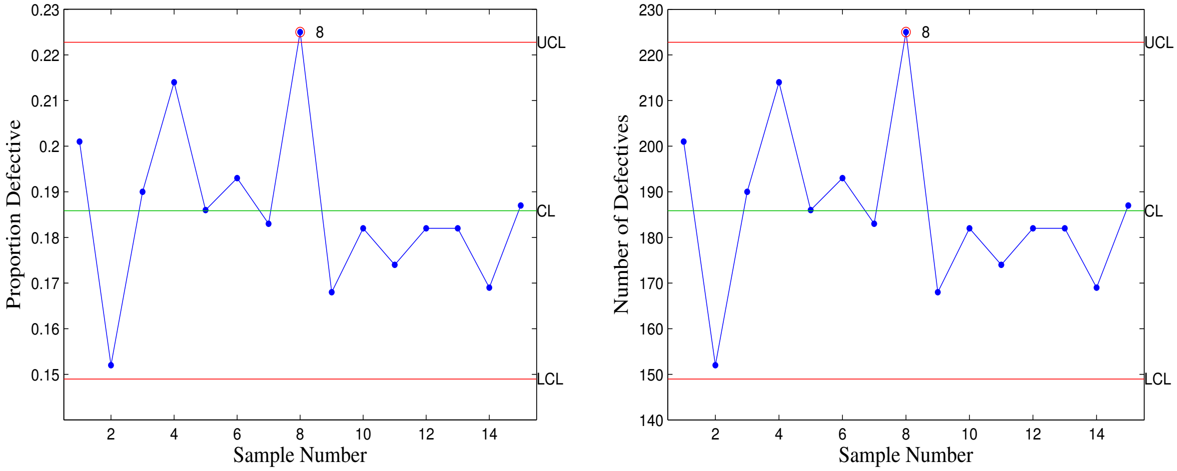


FIGURE 5.21: p-chart for lightbulb data.

A close examination of the control charts, shown in Figure 5.22, reveals that for the 8th week, the sample is above the upper control limit. This indicates a significantly high percentage defective, which implies that there is an assignable cause on the manufacturing process. Such cases may also result due to a lapse from the inspection department or the sample size being quite different from the average used to calculate the control limits. ■

```
X = [201 152 190 214 186 193 183 225 168 182 174 182 182 169 187]; % Number of defectives
n = 1000; %total number of inspected items
[stats,plotdata]=controlchart(X,'chart','p','unit',n); %plot p-chart
fprintf('Control limits for p-chart: UCL=%g, CL=%g, LCL=%g\n', ...
         plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
figure;
[stats,plotdata]=controlchart(X,'chart','np','unit',n); %plot np-chart
fprintf('Control limits for np-chart: UCL=%g, CL=%g, LCL=%g\n', ...
         plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
```



(a) (b)

FIGURE 5.22: (a) p-chart, and (b) np-chart for magnet data.

5.7.3 Control Chart for Count of Defects: c-chart

The c-chart plots the numbers of defects per item. Assume that the number of defects X in a given inspection unit follows a Poisson distribution $poiss(c)$, with parameter c. Then, the mean and standard deviation of X are $\mu_X = c$ and $\sigma_X = \sqrt{c}$, respectively. Since c is unknown, it must be estimated from the available data. Suppose m preliminary samples are available, each of size n. If X_i is the number of defects per sample, then an estimator of the number of defects over the entire data set is given by

$$\overline{C} = \frac{\sum_{i=1}^{m} X_i}{m}$$

Control chart for count of defects (c-chart):
The upper control limit, center line, and lower control limit of the c-chart are given by

$$\begin{aligned} UCL &= \bar{c} + 3\sqrt{\bar{c}} \\ CL &= \bar{c} \\ LCL &= \bar{c} - 3\sqrt{\bar{c}} \end{aligned} \qquad (5.7.3)$$

Example 5.21 *The number of noticeable defects found by quality control inspectors in a randomly selected 1-square-meter specimen of woolen fabric from a certain loom is recorded each hour for a period of 20 hours. The results are shown in Table 5.14. Construct and plot a c-chart to monitor the textile production process.*

TABLE 5.14: Number of defects observed in specimens of woolen fabric over 20 consecutive hours.

Hour	1	2	3	4	5	6	7	8	9	10
Number of Defects	11	14	10	8	3	9	10	2	5	6
Hour	11	12	13	14	15	16	17	18	19	20
Number of Defects	12	3	4	5	6	8	11	8	7	9

Solution: The center line CL is the mean number of defects per square meter of woolen fabric given by

$$CL = \bar{c} = \frac{\sum(\text{Number of Defects})}{20} = \frac{151}{20} = 7.55$$

The upper and lower control limits are given by

$$UCL = \bar{c} + 3\sqrt{\bar{c}} = 15.7932$$

$$LCL = \bar{c} - 3\sqrt{\bar{c}} = -0.69$$

```
X = [11 14 10 8 3 9 10 2 5 6 12 3 4 5 6 8 11 8 7 9]; % Number of defects
n = 20; %size of inspected units (every 20 hours): sample size
[stats,plotdata]=controlchart(X,'chart','c','unit',n); %plot c-chart
fprintf('Control limits for c-chart: UCL=%g, CL=%g, LCL=%g\n', ...
         plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
```

Since a negative number of defects cannot be observed, the LCL value is set to 0. The c-chart, shown in Figure 5.23, indicate that the process is in control. ∎

Example 5.22 *Samples of fabric from a textile mill, each 100 m^2, are selected, and the number of occurrences of foreign matter are recorded. Data for 25 samples are shown in Table 5.15. Construct and plot a c-chart for the number of nonconformities.*

Solution: The center line CL is the mean number of nonconformities

$$CL = \bar{c} = \frac{\sum(\text{Number of Nonconformities})}{100} = \frac{151}{25} = 7.56$$

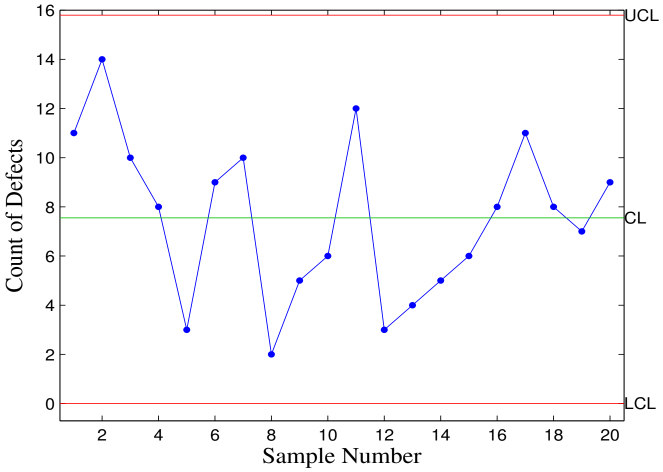


FIGURE 5.23: *c*-chart for woolen fabric data.

TABLE 5.15: Foreign Matter Data.

Sample Number	1	2	3	4	5	6	7	8	9	10	11	12	13
Nonconformities	5	4	7	6	8	5	6	5	16	10	9	7	8
Sample Number	14	15	16	17	18	19	20	21	22	23	24	25	
Nonconformities	11	9	5	7	6	10	8	9	9	7	5	7	

The upper and lower control limits for the *c*-chart are given by

$$UCL = \bar{c} + 3\sqrt{\bar{c}} = 15.8086$$

$$LCL = \bar{c} - 3\sqrt{\bar{c}} = 0$$

The *c*-chart is displayed in Figure 5.24, which indicates that the sample number 9 is out-of-control.

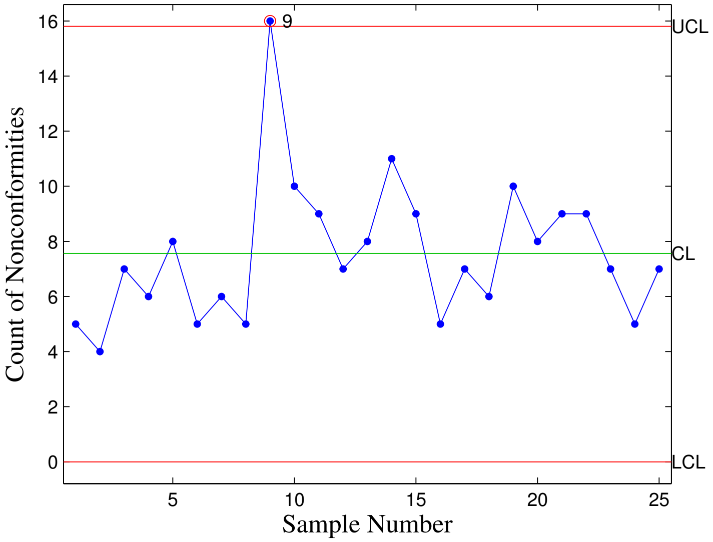


FIGURE 5.24: *c*-chart for foreign matter data.

Assuming special causes for the out-of-control point (sample 9 is deleted), the revised centerline and control limits for the c-chart are

$$\begin{aligned} CL &= \bar{c} = 7.20833 \\ UCL &= \bar{c} + 3\sqrt{\bar{c}} = 15.2628 \\ LCL &= \bar{c} - 3\sqrt{\bar{c}} = 0 \end{aligned}$$

The revised c-chart is given in Figure 5.25, which shows that all the remaining points fall within the control limits.

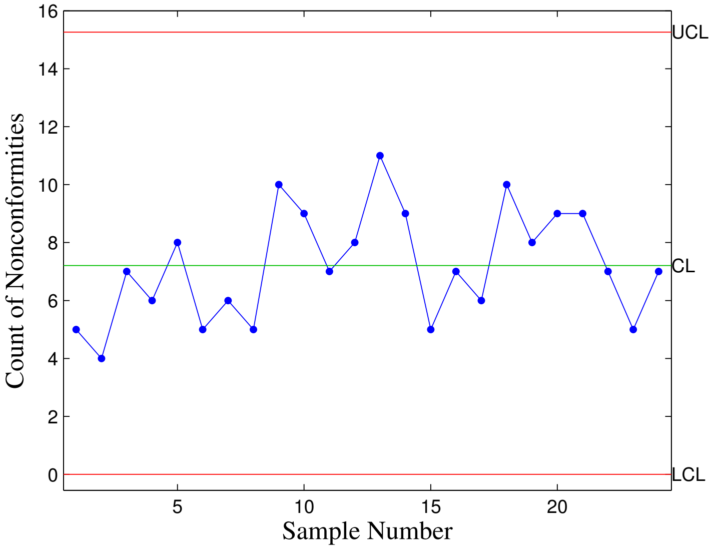


FIGURE 5.25: Revised c-chart for foreign matter data.

5.7.4 CONTROL CHART FOR DEFECTS PER UNIT (u-CHART)

The u-chart monitors the average number of defects. Like the c-chart, the control limits in the u-chart are computed based on the Poisson distribution. Suppose m preliminary samples are available, each of size n. In this chart we plot the rate of defects $U_i = X_i/n$, which is the number of defects X_i per sample divided by the number n (sample size) of units inspected. Thus, an estimator of the average number of defects $U_i = X_i/n$ over m samples is given by

$$\overline{U} = \frac{\sum_{i=1}^{m} U_i}{m}$$

Unlike the c-chart, the u-chart does not require a constant number of units, and it can be used, for example, when the samples are of different sizes.

Control chart for defects per unit (u-chart):
The upper control limit, center line, and lower control limit of the u-chart are given by

$$\begin{aligned} UCL &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} \\ CL &= \bar{u} \\ LCL &= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} \end{aligned} \tag{5.7.4}$$

Example 5.23 *Using the data of Table 5.14, construct and plot the u-chart.*

Solution: The upper control limit, center line, and lower control limit of the u-chart are given by

$$\begin{aligned} UCL &= \bar{u}+3\sqrt{\frac{\bar{u}}{n}} = 0.3775+3\sqrt{\frac{0.3775}{20}} = 0.7897 \\ CL &= \bar{u} = 0.3775 \\ LCL &= \bar{u}-3\sqrt{\frac{\bar{u}}{n}} = 0.3775-3\sqrt{\frac{0.3775}{20}} = -0.0347 \end{aligned}$$

Since a negative number of defects per unit cannot be observed, the LCL value is set to 0. The u-chart, shown in Figure 5.26, indicate that the process is in control. ■

```
X = [11 14 10 8 3 9 10 2 5 6 12 3 4 5 6 8 11 8 7 9]; % Number of defects
n = 20; %size of inspected units (every 20 hours): sample size
[stats,plotdata]=controlchart(X,'chart','u','unit',n); %plot u-chart
fprintf('Control limits for u-chart:  UCL=%g, CL=%g, LCL=%g\n', ...
         plotdata.ucl(1),plotdata.cl(1),plotdata.lcl(1));
```

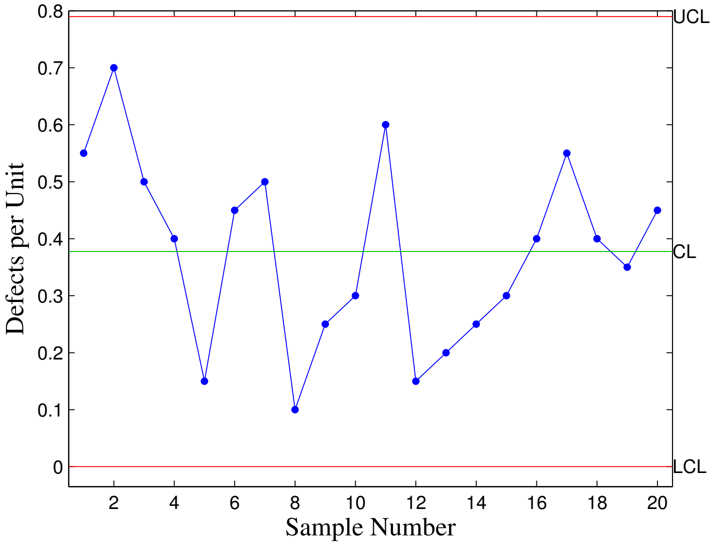


FIGURE 5.26: u-chart for woolen fabric data.

5.8 Acceptance Sampling

Acceptance sampling is a methodology commonly used in quality control and improvement to determine whether to accept or reject a particular lot or batch of products before shipped to customers. This is done by divising a sampling plan that sets the product acceptability criteria. A 100% inspection does not guarantee 100% compliance and is too time consuming and costly. Rather than evaluating all items, a specified sample is taken, inspected or tested, and a decision is made about accepting or rejecting the entire production lot. There are two major classifications of acceptance plans: by attributes and by variables. When the decision to accept or reject a lot based on classification of the items as either defective (conforming) or nondefective (nonconforming), the sampling plan is called **inspection by attributes**. The lot is accepted if no more than an allowable number of defective items are found. The attribute case is the most common for acceptance sampling, and will be assumed for the rest of this section. A sampling plan based on one sample is known as a **single sampling plan**, while sampling plans based on two or more successively drawn samples are known as **double** or **multiple sampling plans**. Selection of an acceptance sampling plan will depend on the nature of the inspection test and the data produced.

An acceptance sampling plan works in the following way. A fixed number n of items is sampled from each lot of size N, carefully inspected, and each item is judged to be either defective or nondefective. If the number d of defectives in the sample is less than or equal to prespecified acceptance number c, the lot is accepted. Otherwise, the lot is rejected. Thus, to design a sampling plan we need to know how many items n to sample and how many defectives items c in that sample are enough to convince us that the lot is unacceptable.

In quality control, there is some relationship between the lot size N and the sample size n because the probability distribution for the number d of defectives in a sample of n items from a lot will depend on the lot size N. For example, a good sampling plan will provide for effective decisions with a sample of 10% or less of the lot size. If N is large and n is small relative to N, then the probability distribution the number d of defectives follows a binomial distribution

$$P(d=k)=\binom{n}{k}p^k(1-p)^{n-k}, \qquad k=0,1,\ldots,n$$

where p denotes the **lot fraction defective** (true proportion nonconforming).
Thus, for a sampling plan with sample size n and acceptance number c, the **probability of accepting a lot** with lot fraction defective p is given by

$$P_a(p)=P(\text{Accept lot})=P(d\leq c)=\sum_{k=0}^{c}P(d=k)=\sum_{k=0}^{c}\binom{n}{k}p^k(1-p)^{n-k} \tag{5.8.1}$$

5.8.1 Sampling Plan Criteria

In order to design a sampling plan, four values must be known and are typically determined from past experience, engineering estimate, and/or management decisions:

1. *Acceptable Quality Level (AQL)*. This is the maximum fraction defective, p_1, considered acceptable off the producer's line, and it is generally in the order of 1-2%. That is, $P_a(p_1)=P_a(AQL)$ should be large, typically near 0.95.

2. *Producer's Risk*, α. This is the probability of rejecting a lot that is within acceptable quality level (AQL). That is, the producer's risk is the probability of rejecting $H_0\,:\,p=p_1$ when H_0 is true. Thus, $\alpha=1-P_a(AQL)$ is the probability of a Type I error. That is, the producer's risk is the probability that a lot containing an acceptable quality level is rejected.

3. *Lot Tolerance Percent Defective (LTPD*. This is the largest lot fraction defective, p_2, that can be tolerated by a consumer. The LTPD has a low probability of acceptance.

4. *Consumer's Risk*, β. This is the probability that a "bad" lot containing a greater number of defects than the LTPD limit will be accepted. Thus, $\beta=P_a(LTPD)$ is the probability of making a Type II error.

Example 5.24 *A manufacturer of USB drives ships a particular USB in lots of 500 each. The acceptance sampling plan used prior to shipment is based on a sample size $n=10$ and acceptance number $c=1$.*

1. *Find the producer's risk if $AQL=0.05$*

2. *Find the consumer's risk if $LTPD=0.2$*

Solution: The lot size $N=500$ is much larger than the sample size $n=10$. Thus, the probability of lot acceptance is given by

$$P_a(p)=\sum_{k=0}^{1}\binom{10}{k}p^k(1-p)^{10-k}=\binom{10}{0}(1-p)^{10}+\binom{10}{1}p(1-p)^9=(1-p)^{10}+10p(1-p)^9$$

1. For $p=0.05$, we have $P_a(0.05)=0.914$. Thus, the producer's risk is $\alpha=1-P_a(AQL)=1-P_a(0.05)=1-0.914=0.086$. That is, the producer will reject 8.6% of the lots, even if the lot fraction defective is as small as 0.05.

2. The consumer's risk is $\beta=P_a(LTPD)=P_a(0.2)=0.376$. ■

```
c=1; % acceptance number
n=10; % sample size
AQL = 0.05; % acceptable quality level
alpha = 1-cdf('bino',c,n,AQL); % producer's risk
fprintf('Producer Risk = %.3f\n',alpha)
LTPD = 0.2; % lot tolerance percent defective
beta = cdf('bino',c,n,LTPD); % consumer's risk
fprintf('Consumer Risk = %.3f\n',beta)
```

5.8.2 Operating Characteristic (OC) Curves

The operating characteristic (OC) curve is an excellent graphical tool for evaluating quality sampling plans. The OC curve depicts the plot of the probability of lot acceptance P_a versus lot fraction defective p. The OC curve for Example 5.24 is shown in Figure 5.27. Note that as the lot fraction defective increases, the probability of lot acceptance decreases until it reaches 0.

```
c=1; % acceptance number
n=10; % sample size
p=0:0.01:1; % lot fraction defective
Pa=cdf('bino',c,n,p); % probability of lot acceptance
plot(p,Pa); % plot OC curve
```

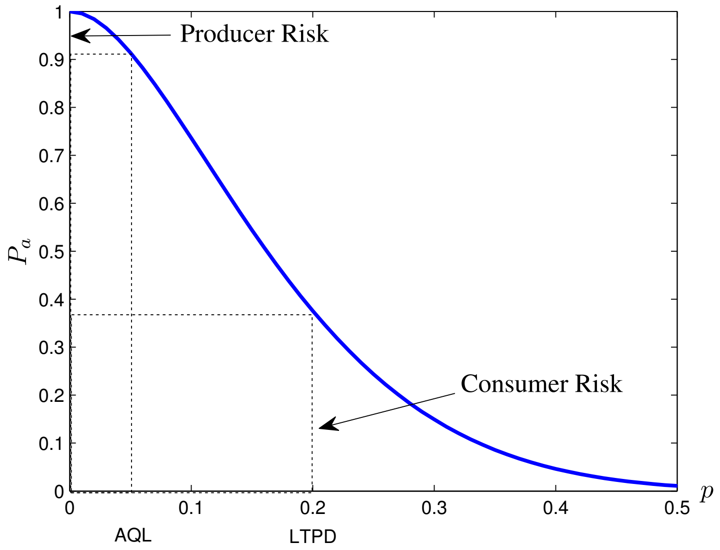


FIGURE 5.27: OC curve for Example 5.24.

In general, the steeper the OC curve the better the protection for both consumer and producer. In fact, the ideal OC curve would be a parallel line to the y-axis at the AQL value as shown in Figure 5.28.

When sample sizes are increased, the OC curve becomes steeper as shown in Figure 5.29(b). In the same vein, when the acceptance number is decreased, the curve gets steeper. Moreover, changing the acceptance number does not significantly change the OC curve as shown in Figure 5.29(c).

5.8.3 Average Outgoing Quality

The average outgoing quality (AOQ) is the expected average quality of outgoing products for a given value of incoming product quality. The AOQ curve can be used to evaluate a sampling plan by showing the average quality accepted by

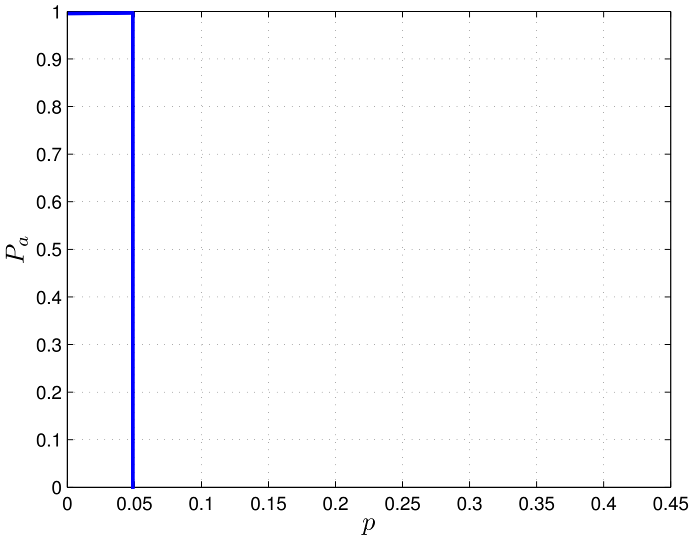
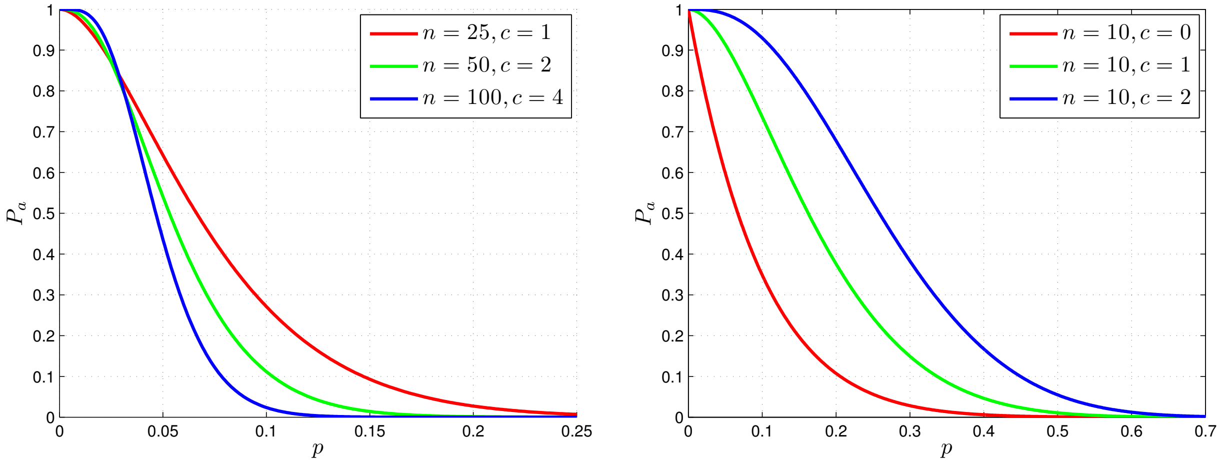


FIGURE 5.28: Ideal OC curve (100% inspection).



(a)

(b)

FIGURE 5.29: (a) OC curves for different sample sizes; (b) OC curves for the same sample sizes.

the consumer for a given fraction defective.

$$AOQ = \left(1 - \frac{n}{N}\right) p\, P_a \tag{5.8.2}$$

where N is the lot size and n is the sample size. Note that when $N \gg n$, then $AOQ \approx p\, P_a$. The AOQ curve for Example 5.24 is shown in Figure 5.30.

```
c=1; % acceptance number
n=10; % sample size
N=500; % lot size
p=0:0.01:1; % lot fraction defective
Pa=cdf('bino',c,n,p); % probability of lot acceptance
AOQ = (1-n/N)*p.*Pa; % average outgoing quality
plot(p,AOQ); % plot AOQ curve
```

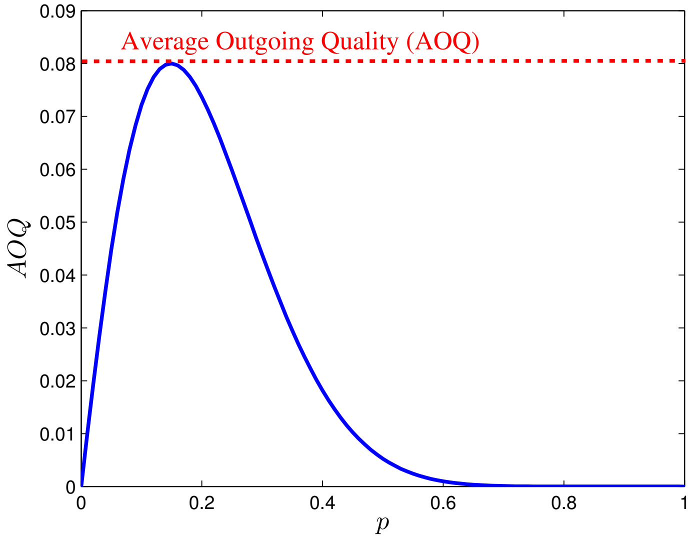


FIGURE 5.30: AOQ curve for Example 5.24.

The AOQ curve initially increases as more defectives are produced, more are released. As more and more lots are rejected, 100% inspections become more common and the AOQ curve starts to decrease as a result. The average outgoing quality limit (AOQL) is simply the maximum value on the AOQ curve, and represents the maximum possible fraction defective for the sampling plan.

5.8.4 AVERAGE TOTAL INSPECTION

In practice, a rejected lot usually means that the lot is to follow a particular routine, to be screened, repaired, corrected, or even rejected, or perhaps accepted after argument or waiver of specifications. In screening, the rest of the lot is 100% inspection. Defective items are supposedly removed, and what is left of the lot is supposedly perfect. But the question that rises is: What is the total amount of inspection when rejected lots are screened? If all lots contain zero defectives, no lot will be rejected. If all items are defective, all lots will be inspected, and the amount to be inspected is N. Finally, if the lot quality is $0 < p < 1$, the average amount of inspection per lot will vary between the sample size n, and the lot size N.

The average total inspection (ATI) is the average total number of items inspected per lot of size N. Let the quality of the lot be p and the probability of lot acceptance be P_a, then the ATI per lot is

$$ATI = n + (1 - P_a)(N - n) \tag{5.8.3}$$

The ATI curve for Example 5.24 is shown in Figure 5.31.

```
c=1; % acceptance number
n=10; % sample size
N=500; % lot size
p=0:0.01:1; % lot fraction defective
Pa=cdf('bino',c,n,p); % probability of lot acceptance
ATI = n+(N-n)*(1-Pa); % average total inspection
plot(p,ATI); % plot ATI curve
```

5.9 MULTIVARIATE CONTROL CHARTS

Many manufacturing and service businesses use univariate statistical control charts to monitor the performance of their processes [1]. However, in most processes, there are more than one measurement process to monitor [2], and it is increasingly difficult to determine the root cause of defects if multiple process variables exhibit faults or process deviations at the same moment in time. Moreover, most processes are highly correlated, particularly for assembly operations and chemical processes [2, 3]. Univariate control charts not only lead to frequent adjustments of the process but also

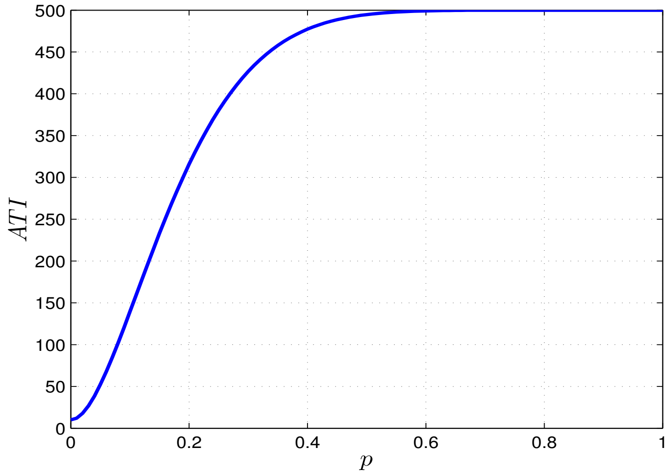


FIGURE 5.31: AOQ curve for Example 5.24.

do not account for the correlation information between the measurement processes [3]. Multivariate quality control methods overcome these limitations by monitoring the interactions of several process variables simultaneously and also by determining hidden factors using dimensionality-reduction techniques [2]. The use of multivariate statistical process control is also facilitated by the proliferation of sensor data that is typically complex, high-dimensional and generally correlated. Multivariate charts are used to detect shifts in the mean or the relationship (covariance) between several related parameters.

In recent years, several multivariate statistical process control techniques have been proposed to analyze and monitor multivariate data [5, 4, 6]. With multivariate quality control charts, it is possible to have well-defined control limits, while taking in consideration the cross-correlation between the variables. In addition, these multivariate charts may be used to analyze the stability of the processes without the complication of simultaneously monitoring several univariate control charts [2].

Mapping a multivariate situation as a univariate may lead to results where processes might seem to be in control when in fact they are not and vice-versa, as illustrated in Fig. 5.32 which depicts the result of modelling two highly-correlated variables as independent. The ellipse defines a region where the process is operating under normal operating conditions. Any observation falling outside the ellipse is identified as a fault. If the variables were, however, modeled as independent, then the control region would be defined between the rectangle. As can be seen in Fig. 5.32, some out-of-control observations would be misidentified, indicating that the correlation structure between the variables should be taken into account in order to accurately characterize the behavior of multivariate industrial environments [2].

5.9.1 χ^2 and Hotelling T^2 Control Chart

In many industrial applications, the output of a process is characterized by p variables that are measured simultaneously. Independent variables can be charted individually, but if the variables are correlated, a multivariate chart is needed to determine whether the process is in control. Generally, the univariate process variables make a random vector, $\boldsymbol{x} = (X_1, X_2, \ldots, X_p)'$. That is, the process has p quality characteristics $X_1, X_2, \ldots, X_p$ that we are interested to monitor. Suppose the random vector $\boldsymbol{x}$ follows a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ and a covariance matrix Σ. The statistic given by

$$\chi^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \tag{5.9.1}$$

follows a $\chi^2(p)$ distribution with p degrees of freedom, where $\bar{\boldsymbol{x}}$ is the sample mean for each of the p quality characteristics from a sample of size n, $\boldsymbol{\mu}$ is the vector of in-control means for each quality characteristic, and Σ^{-1} is the inverse of the covariance matrix. The upper control limit of the χ^2 control chart is given by $UCL = \chi^2_{\alpha,p}$, where α is a given significance level.

Since $\boldsymbol{\mu}$ and Σ are unknown in practice, we usually estimate them on the basis of preliminary samples (subgroups), taken when the process is thought to be in control. Suppose m preliminary samples are available, each of size n. The Hotelling's T^2 control chart is the most common monitoring technique for multivariate data, and it can be thought of

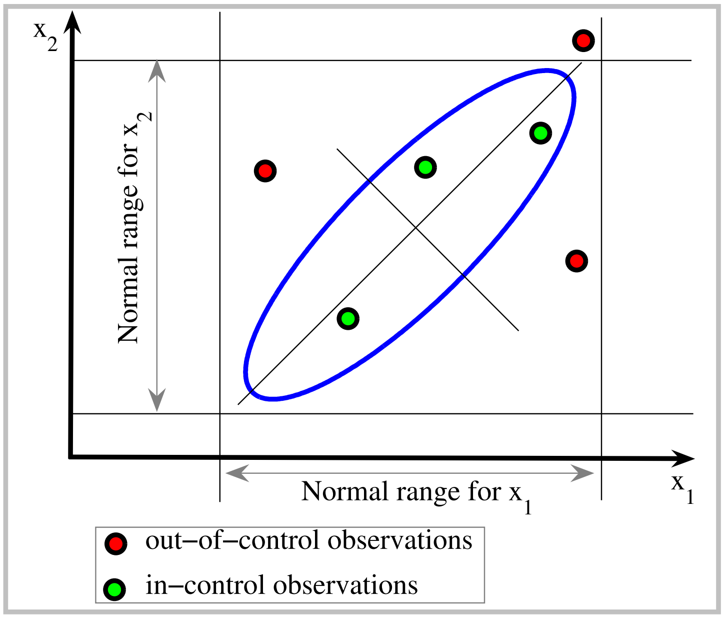


FIGURE 5.32: Motivation behind using multivariate control charts.

as the multivariate counterpart of the Shewhart $\bar{x}$-chart. The T^2 statistic is given by

$$T^2 = n(\bar{\mathbf{x}} - \hat{\boldsymbol{\mu}})'\widehat{\Sigma}^{-1}(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}) \tag{5.9.2}$$

where $\bar{\mathbf{x}}$ is the vector of sample means, $\hat{\boldsymbol{\mu}}$ is the estimated vector of in-control means, and $\widehat{\Sigma}$ is the estimated covariance matrix for the quality characteristics when the process is in control.

To find $\hat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$, consider the vector of sample means given by

$$\bar{\mathbf{x}}_k = \begin{pmatrix} \bar{x}_{1k} \\ \bar{x}_{2k} \\ \vdots \\ \bar{x}_{pk} \end{pmatrix} \qquad k = 1, 2, \ldots, m \tag{5.9.3}$$

where

$$\bar{x}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ijk} \qquad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases} \tag{5.9.4}$$

is the sample mean of the jth quality characteristic for the kth sample, and x_{ijk} is the ith observation on the jth quality characteristic in the kth sample.

The sample variances for the jth quality characteristic in the kth sample are given by

$$s^2_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ijk} - \bar{x}_{jk}\right)^2 \qquad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases} \tag{5.9.5}$$

The sample covariance between the jth and hth quality characteristics in the kth sample is given by

$$s^2_{jhk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ijk} - \bar{x}_{jk}\right)\left(x_{ihk} - \bar{x}_{hk}\right), \qquad \begin{cases} k = 1, 2, \ldots, m \\ j \neq h \end{cases} \tag{5.9.6}$$

The target mean of each quality characteristic for m samples is given by

$$\bar{\bar{x}}_j = \frac{1}{m}\sum_{k=1}^{m} \bar{x}_{jk}, \quad j = 1, 2, \ldots, p \tag{5.9.7}$$

and the averaged sample variance and covariance are given by

$$\bar{s}_j^2 = \frac{1}{m}\sum_{k=1}^{m} s_{jk}^2 \qquad j = 1, 2, \ldots, p \tag{5.9.8}$$

and

$$\bar{s}_{jh} = \frac{1}{m}\sum_{k=1}^{m} s_{jhk} \qquad j \neq h \tag{5.9.9}$$

Thus, $\hat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$ can be estimated by the mean vector $\bar{\bar{\boldsymbol{x}}}$ and the average covariance matrix C, respectively, as follows:

$$\bar{\bar{\boldsymbol{x}}} = \begin{pmatrix} \bar{\bar{x}}_1 \\ \bar{\bar{x}}_2 \\ \vdots \\ \bar{\bar{x}}_p \end{pmatrix} \qquad \text{and} \qquad C = \begin{pmatrix} \bar{s}_1^2 & \bar{s}_{12} & \cdots & \bar{s}_{1p} \\ \bar{s}_{12} & \bar{s}_2^2 & \cdots & \bar{s}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{s}_{1p} & \bar{s}_{2p} & \cdots & \bar{s}_p^2 \end{pmatrix} \tag{5.9.10}$$

The Hotelling's T^2 statistic for the ith subgroup is given by

$$T_k^2 = n(\bar{\boldsymbol{x}}_k - \bar{\bar{\boldsymbol{x}}})' \, C^{-1} \, (\bar{\boldsymbol{x}}_k - \bar{\bar{\boldsymbol{x}}}), \quad k = 1, 2, \ldots, m \tag{5.9.11}$$

- For phase I, the upper control limit is

$$UCL = \frac{p(m-1)(n-1)}{mn - m - p + 1} F_{\alpha,p,mn-m-p+1} \tag{5.9.12}$$

- For phase II, the upper control limit is

$$UCL = \frac{p(m+1)(n-1)}{mn - m - p + 1} F_{\alpha,p,mn-m-p+1} \tag{5.9.13}$$

Example 5.25 *Table 5.16 shows bivariate data for two quality characteristics X_1 and X_2 for 25 samples, each of size 5. Plot the Hotelling's T^2 control chart. Assume $\alpha = 0.05$.*

Solution: We have $m = 25$, $n = 5$, and $p = 2$. The summary statistics are shown in Table 5.17.
For phase I, the upper control limit is

$$UCL = \frac{p(m-1)(n-1)}{mn - m - p + 1} F_{\alpha,p,mn-m-p+1} = 1.9394 \, F_{0.05,2,99} = (1.9394)(3.0882) = 5.9893$$

where $F_{0.05,2,99}$ can be computed using MATLAB as follows: $\gg$ icdf('F',1-0.05,2,99). The Hotelling's T^2 control chart is shown in Figure 5.33, where the samples 5 and 18 appears to be out-of-control. However, all the samples in the $\overline{X}$-charts for each quality characteristic appear to be in-control as shown in Figure 5.34. Thus, the process is out-of-control. ■

5.10 Principal Components Analysis

Principal components analysis (PCA) is an explanatory technique to learn about data sets. The objective of PCA to reduce the dimensionality of the data set while retaining as much as possible the variation in the data set. Principal components (PCs) are linear transformations of the original set of variables, and are uncorrelated and ordered so that the first few components carry most of the variation in the original data set. The first PC has the geometric interpretation that it is a new coordinate axis that maximizes the variation of the projections of the data points on the new coordinate axis. The general idea of PCA is as follows: if we have a set of moderately or strongly correlated variables (i.e. the variables share much common information), it may then be possible to construct new variables that are combinations of these variables that account for much of the original information contained in the data. The output

Sample Number k	Quality Characteristic									
	X_1					X_2				
1	65	69	79	66	62	33	41	36	36	42
2	64	71	72	73	72	35	37	37	37	39
3	78	75	73	59	60	40	34	39	32	38
4	63	66	74	69	76	34	34	39	34	38
5	87	81	71	62	67	37	33	39	36	38
6	76	72	80	67	75	34	38	33	36	38
7	64	63	60	75	61	34	38	32	34	37
8	66	65	68	85	75	34	35	41	33	42
9	73	81	78	65	67	39	39	39	39	36
10	75	64	69	73	68	37	34	38	37	34
11	78	61	74	65	67	38	41	42	41	36
12	72	78	56	70	74	35	31	43	39	41
13	61	72	72	91	83	37	36	35	34	39
14	70	87	78	70	76	40	39	37	30	41
15	57	67	51	64	69	33	37	38	37	36
16	74	74	67	77	63	35	37	37	36	37
17	74	69	59	76	70	40	35	38	39	32
18	70	72	61	57	61	42	43	38	40	40
19	62	56	74	64	68	34	36	40	36	37
20	79	71	78	65	77	38	39	29	37	37
21	77	75	67	67	68	33	41	37	35	40
22	63	69	65	70	79	39	36	36	41	36
23	58	66	83	65	65	37	41	36	37	38
24	77	69	68	72	73	34	38	37	35	35
25	74	52	72	73	72	39	33	39	36	35

TABLE 5.16: Bivariate data for two quality characteristics.

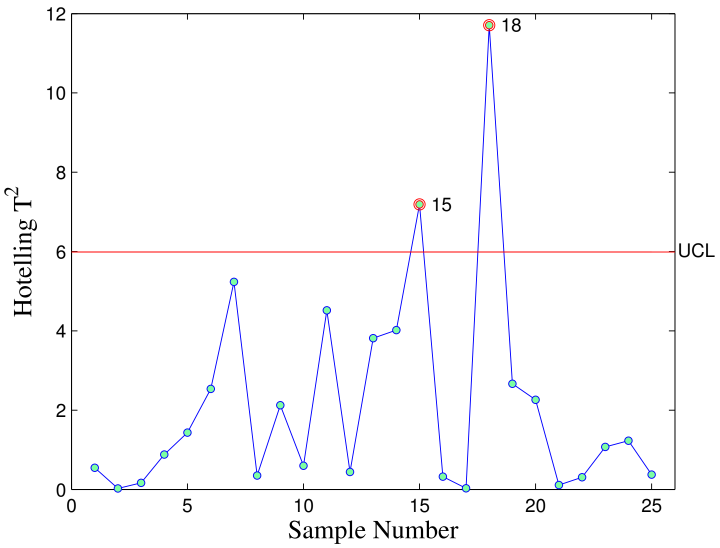


FIGURE 5.33: T^2 control chart for bivariate data.

of PCA consists of coefficients that define the linear combinations used to obtain the new variables (PC loadings) and the new variables (PCs) themselves. Examining the PC loadings and plotting the PCs can aid in data interpretation, particularly with higher dimensional data.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ be n observation vectors on a random vector $\boldsymbol{x} = (X_1, X_2, \dots, X_p)'$ with p quality characteristics,

Sample	Summary Statistics					
Number k	$\bar{x}_{1k}$	$\bar{x}_{2k}$	s^2_{1k}	s^2_{2k}	s_{12k}	T^2_k
1	68.20	37.60	42.70	14.30	-5.90	0.55
2	70.40	37.00	13.30	2.00	4.00	0.02
3	69.00	36.60	78.50	11.80	14.50	0.17
4	69.60	35.80	29.30	6.20	11.90	0.88
5	73.60	36.60	104.80	5.30	-7.45	1.43
6	74.00	35.80	23.50	5.20	-6.00	2.54
7	64.60	35.00	36.30	6.00	-2.00	5.24
8	71.80	37.00	69.70	17.50	-5.25	0.35
9	72.80	38.40	47.20	1.80	4.35	2.12
10	69.80	36.00	18.70	3.50	5.50	0.60
11	69.00	39.60	47.50	6.30	-3.00	4.52
12	70.00	37.80	70.00	23.20	-30.00	0.44
13	75.80	36.20	132.70	3.70	-4.95	3.81
14	76.20	37.40	49.20	19.30	11.40	4.02
15	61.60	36.20	55.80	3.70	0.10	7.19
16	71.00	36.40	33.50	0.80	-3.00	0.32
17	69.60	36.80	43.30	10.70	3.65	0.03
18	64.20	40.60	41.70	3.80	10.35	11.71
19	64.80	36.60	45.20	4.80	11.40	2.66
20	74.00	36.00	35.00	16.00	-8.25	2.26
21	70.80	37.20	23.20	11.20	-2.20	0.11
22	69.20	37.60	38.20	5.30	-3.65	0.31
23	67.40	37.80	86.30	3.70	-5.90	1.07
24	71.80	35.80	12.70	2.70	-5.30	1.23
25	68.60	36.40	86.80	6.80	18.20	0.37

TABLE 5.17: Summary Statistics for the bivariate data.

where

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

is a p-dimensional column vector, which represents the vector of quality characteristic means. All n observation vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ on p process variables can be transposed to row vectors and placed in a matrix X, of dimension $n \times p$, called data matrix or data set, as follows:

$$X = \begin{pmatrix} \boldsymbol{x}'_1 \\ \boldsymbol{x}'_2 \\ \vdots \\ \boldsymbol{x}'_i \\ \vdots \\ \boldsymbol{x}'_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} \tag{5.10.1}$$

An individual column of X corresponds to that data collected on a particular variable, while an individual row (observation) refers to the data collected on a particular individual or object for all variables. The value of the jth variable for the ith observation $\boldsymbol{x}'_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ is x_{ij}, which is the element of the ith row and jth column of X.

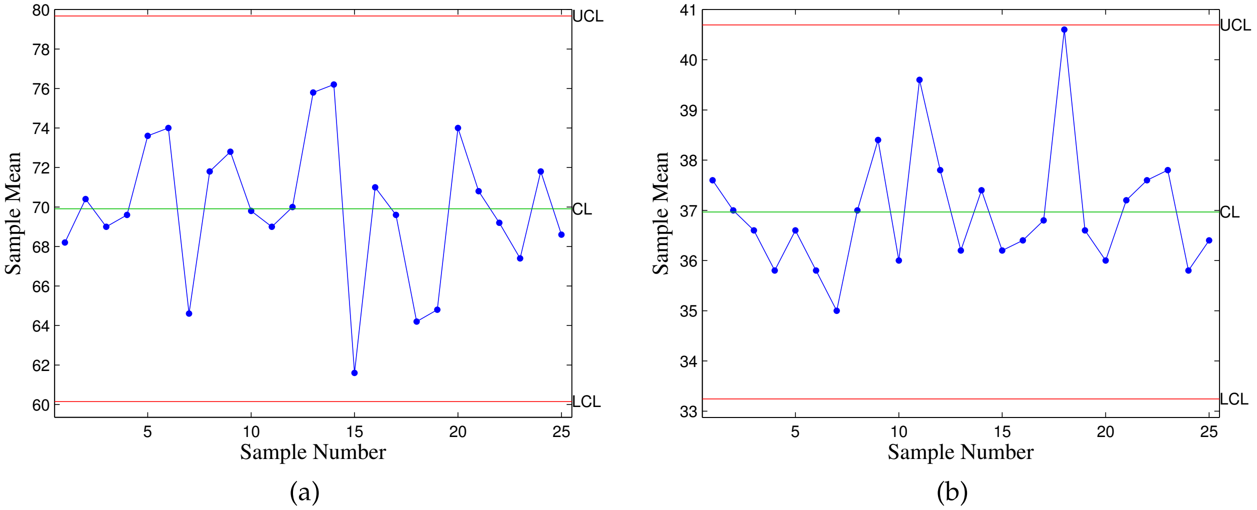


FIGURE 5.34: $\overline{X}$-chart for each quality characteristic of the bivariate data: (a) $\overline{X}$-chart for X_1. (b) $\overline{X}$-chart for X_2.

- The sample mean vector $\bar{\boldsymbol{x}}$ is a p-dimensional vector given by

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \frac{1}{n}X'\mathbf{1} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{pmatrix} \tag{5.10.2}$$

where $\mathbf{1} = (1, 1, \ldots, 1)'$ is a n-dimensional column vector of all 1s, X' denotes the transpose of X, and $\bar{x}_j$ is the mean of the jth variable. That is, $\bar{x}_j$ is the mean of the jth column of the data matrix X.

- The sample covariance matrix $S = (s_{ij})$ is a $p \times p$ symmetric matrix given by

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1i} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2i} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{1i} & s_{2i} & \cdots & s_{ii} & \cdots & s_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{ip} & \cdots & s_{pp} \end{pmatrix} \tag{5.10.3}$$

where $s_{ii} = s_i^2$ is the sample variance of the ith variable

$$s_{ii} = s_i^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \tag{5.10.4}$$

and s_{ij} is the sample covariance of the ith and jth variables

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j). \tag{5.10.5}$$

The sample covariance matrix $S = (s_{ij})$ can be written as

$$S = \frac{1}{n-1}X'HX \tag{5.10.6}$$

where $H = I - J/n$ is a called centering matrix, and $J = \mathbf{11}'$ is an $n \times n$ matrix of all 1s. Note that a centered data matrix Y is obtained by multiplying the centering matrix H with the data matrix X, that is $Y = HX$.

- The sample correlation matrix $R = (r_{ij})$ is a symmetric $p \times p$ matrix given by

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{pmatrix} \tag{5.10.7}$$

where r_{ij} is the sample correlation between ith and jth variables

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} = \frac{s_{ij}}{s_i s_j}. \tag{5.10.8}$$

Note that if $D = \operatorname{diag}(s_1, s_2, \ldots, s_p)$ is a $p \times p$ diagonal matrix of sample standard deviations, then the covariance and correlation matrices can be written as $S = DRD$ and $R = D^{-1}SD^{-1}$, where $D^{-1} = \operatorname{diag}(1/s_1, 1/s_2, \ldots, 1/s_p)$ is the inverse matrix of D.

5.10.1 PCA Algorithm

Given a data matrix X, the PCA algorithm consists of four main steps:

Step 1: Compute the centered data matrix $Y = HX$ by subtracting off-column means.

Step 2: Compute the $p \times p$ covariance matrix S of the centered data matrix as follows

$$S = \frac{1}{n-1} Y'Y \tag{5.10.9}$$

Step 3: Compute the eigenvectors and eigenvalues of S using eigen-decomposition

$$S = A\Lambda A' = \sum_{j=1}^{p} \lambda_j \boldsymbol{a}_j \boldsymbol{a}_j' \tag{5.10.10}$$

where

- $A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_p)$ is a $p \times p$ orthogonal matrix ($A'A = I$) whose columns $\boldsymbol{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jp})'$ are the eigenvectors of S such that $\boldsymbol{a}_j'\boldsymbol{a}_j = 1$, $j = 1, \ldots, p$. The eigenvectors tell us a direction or linear combination of existing data. The first eigenvector $\boldsymbol{a}_1$ defines a variable with the most variation in the data matrix. The eigenvectors are the *principal component (PC)* coefficients, also known as *loadings*. That is, each eigenvector $\boldsymbol{a}_j$ contains coefficients for the jth PC. These eigenvectors are in order of decreasing component variance.
- $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a $p \times p$ diagonal matrix whose elements are the eigenvalues of S arranged in decreasing order, i.e. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. The eigenvalues tell us about the amount of variation in the data matrix, that is the variance in a particular direction. The eigenvalues measure the importance of the PCs.

Step 4: Compute the transformed data matrix $Z = YA$ of size $n \times p$

$$Z = (\boldsymbol{z}_1', \boldsymbol{z}_2', \ldots, \boldsymbol{z}_i', \ldots, \boldsymbol{z}_p') = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ z_{i1} & z_{i2} & \cdots & z_{ij} & \cdots & z_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nj} & \cdots & z_{np} \end{pmatrix} \tag{5.10.11}$$

which contains the coordinates of the original data in the new coordinate system defined by the PCs. The rows of Z correspond to observations $\boldsymbol{z}_i = A'(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, while its columns correspond to *PC scores*. The eigenvalues are the variances of the columns of the PCs. Thus, the first PC score accounts for as much of the variability in the data as possible, and each succeeding component score accounts for as much of the remaining variability as possible.

To apply PCA to the data matrix X, we use the MATLAB function pca as follows:

```
>> [A,Z,lambda,Tsquare] = pca(X)
```

where A is the eigenvector matrix, Z is the transformed data matrix, lambda is a p-dimensional vector of eigenvalues i.e. $\Lambda = \text{diag}(\text{lambda}) = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, and Tsquare is Hotelling's T^2, a statistical measure of the multivariate distance of each observation from the center of the data set. Tsquare $= (T_1^2, T_2^2, \ldots, T_p^2)$ is p-dimensional vector whose values are given by

$$T_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' S^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \tag{5.10.12}$$

Using the fact that $S^{-1} = A\Lambda^{-1}A'$ and $\boldsymbol{z}_i = A'(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, it follows that

$$T_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' A\Lambda^{-1}A' (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \boldsymbol{z}_i' \Lambda^{-1} \boldsymbol{z}_i \tag{5.10.13}$$

Standardizing the data is often preferable when the variables are in different units or when the variance of the different columns is substantial. In this case, we use pca(zscore(X)) instead of pca(X). That is, PCA is performed on the correlation matrix instead of the covariance.

Hotelling Control Chart:

The Hotelling control chart is a multivariate extension of the $\overline{X}$-chart that does take the correlation into account. The plotted points on this chart are given by the Hotelling's statistic T_i^2 for individual observations:

$$T_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' S^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$$

- For phase I, the upper control limit is

$$UCL = \frac{(n-1)^2}{n} B_{\alpha, p/2, (n-p-1)/2} \tag{5.10.14}$$

  where $B(\cdot)$ the inverse cdf of the Beta distribution and α is the significance level (typically set to 0.05 or 0.01).

- For phase II, the upper control limit is

$$UCL = \frac{p(n+1)(n-1)}{n(n-p)} F_{\alpha, p, n-p} \tag{5.10.15}$$

Example 5.26 *The phosphorus content data set contains 18 observations, where each observation has 3 variables:*

- *X1: Inorganic phosphorus*
- *X2: Organic phosphorus*
- *X3: Plant phosphorus*

This data set studies the effect of organic and inorganic phosphorus in the soil in comparison with the phosphorus content of the corn grown. Start by loading the data in phosphorus.mat*:*

```
>> load phosphorus.mat;
>> whos
  Name                Size              Bytes  Class     Attributes
  Description         7x72               1008  char
  X                   18x3                432  double
  observations        18x2                 72  char
  variables            3x2                 12  char
```

1. *Display the box plot for the phosphorus data matrix.*

2. *Compute the mean vector $\mathbf{x}$, covariance S, and correlation R.*
3. *Plot the second PC score vs. the first PC score.*
4. *Plot the Hotelling's T^2 control chart.*

Solution:

```
load phosphorus.mat;
[n,p]=size(X); %size of data matrix
xbar = mean(X); %mean vector
S = cov(X); %covariance matrix
R = corr(X); %correlation matrix
boxplot(X);
[A,Z,lambda,Tsquare]=pca(X); %perform PCA on data matrix using covariance
% PC2 score vs. PC1 score
scatter(Z(:,1),Z(:,2),3,'o','MarkerFaceColor',[.49 1 .63],'LineWidth',1);
```

1. The side-by-side box plots for the data is shown in Figure 5.35. We can see that the third column of the data (i.e. plant phosphorus) contains an outlier.

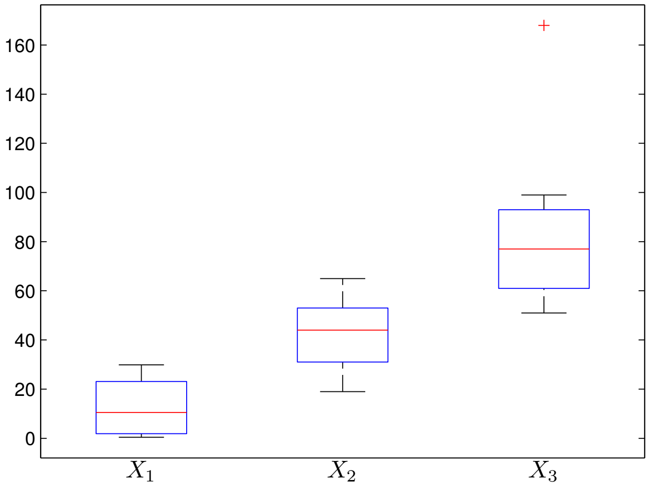


FIGURE 5.35: Boxplot for the phosphorus content data.

2. The mean vector $\bar{\mathbf{x}}$, covariance S, and correlation R are

$$\bar{\mathbf{x}} = \begin{pmatrix} 11.94 \\ 42.11 \\ 81.28 \end{pmatrix} \qquad S = \begin{pmatrix} 103.12 & 63.86 & 190.09 \\ 63.86 & 185.63 & 130.38 \\ 190.09 & 130.38 & 728.80 \end{pmatrix} \qquad R = \begin{pmatrix} 1.00 & 0.46 & 0.69 \\ 0.46 & 1.00 & 0.35 \\ 0.69 & 0.35 & 1.00 \end{pmatrix}$$

Note that R can also be obtained from S using the MATLAB command corrcov as follows:

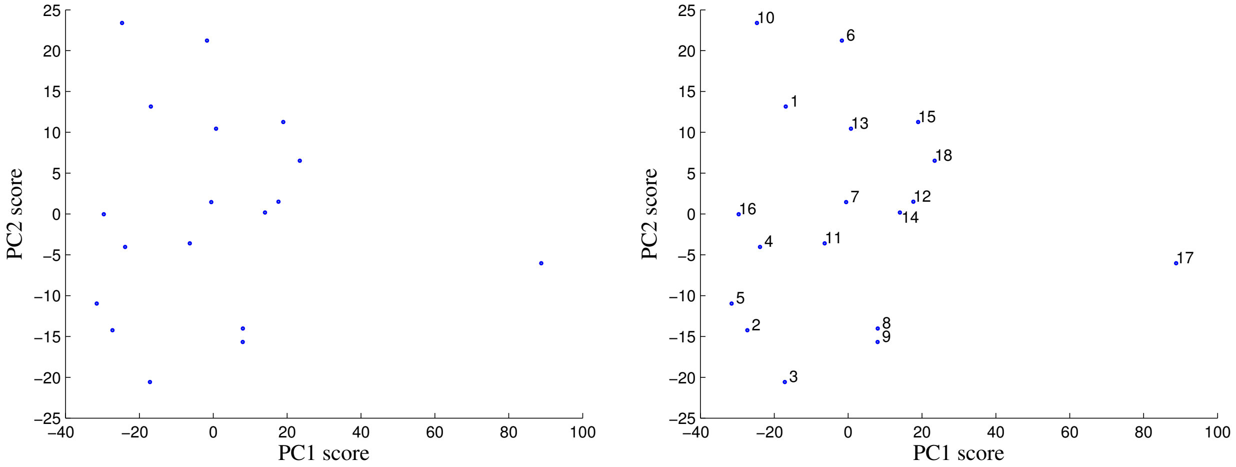
```
>> [R,sigma]=corrcov(S);
```

where sigma is the vector of sample standard deviations.

3. The plot the second PC score vs. the first PC score is shown in Figure 5.36. The labels displayed in Figure 5.36(b) represent the observation numbers. Notice that the sample number 17 is an outlier.

4. The plot of the Hotelling's T^2 control chart is displayed in Figure 5.37, which shows that the sample numbers 6 and 17 appear to be out-of-control.

```
%Plot Hotelling's T^2 control chart
alpha=0.05;
UCL = ((n-1)^2/n)*icdf('beta',1-alpha,p/2,(n-p-1)/2);
plot(Tsquare,'bo-','MarkerFaceColor',[.49 1 .63],'MarkerSize',2);
```



(a) unlabeled plot (b) labeled plot

FIGURE 5.36: PC2 vs. PC1 score for the phosphorus content data.

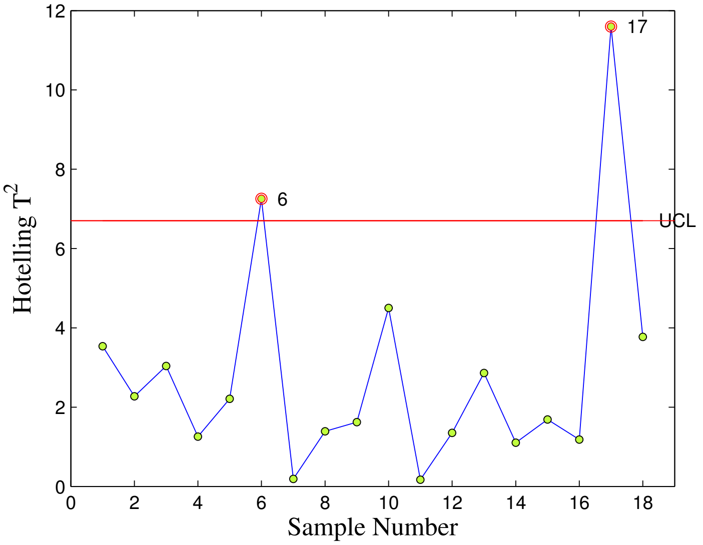


FIGURE 5.37: Hotelling's T^2 control chart for the phosphorus content data.

5.10.2 PCA THEORY

Let $\boldsymbol{x} = (X_1, X_2, \ldots, X_p)'$ denote the original variables measured on the objects or individuals. In principal component analysis, the p original variables are transformed into linear combinations of uncorrelated variables (PCs) $Z_1, Z_2, \ldots, Z_p$ such that the jth PC Z_j is given by $Z_j = \boldsymbol{a}_j'\boldsymbol{x}$, where $\boldsymbol{a}_j = (a_{1j}, \ldots, a_{pj})'$ is the jth eigenvector of S. That is,

$$Z_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p \tag{5.10.16}$$

Since $S\boldsymbol{a}_j = \lambda_j \boldsymbol{a}_j$, it follows that $var(Z_j) = \lambda_j$. Thus, the first PC Z_1 has the largest variance, while the last PC Z_p has the smallest. Moreover, the PCs are uncorrelated, that is $corr(Z_j, Z_k) = 0$ for $j \neq k$, which results from the orthogonality of the eigenvectors $\boldsymbol{a}_j'\boldsymbol{a}_k = 0$.

If $\boldsymbol{x}$ is standardized to have zero mean and unit variance for each X_i, then the PCs Z_j are correlated with the original variables X_i such that

$$corr(Z_j, X_i) = a_{ij}\sqrt{\lambda_j} \tag{5.10.17}$$

Denote by $\boldsymbol{z} = (Z_1, Z_2, \ldots, Z_p)'$ the p-dimensional vector of PCs. Then, the correlation matrix between the PCs and the

the original variables is given by

$$C = corr(\boldsymbol{z}, \boldsymbol{x}) = A\Lambda^{1/2} \tag{5.10.18}$$

where A and $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p})$ are obtained by performing PCA on the correlation matrix, that is by applying PCA on the standardized data matrix using pca(zscore(X)).

Example 5.27 *Using the phosphorus content data*

1. *Determine the first and second PCs*
2. *Display the scatter plot of PC2 coefficients vs. PC1 coefficients, and label the points*
3. *Compute the correlation matrix between the PCs and the original variables*

Solution

1. Using the columns of the eigenvector matrix

$$A = \begin{pmatrix} 0.27 & 0.16 & 0.95 \\ 0.22 & 0.95 & -0.22 \\ 0.94 & -0.27 & -0.22 \end{pmatrix},$$

   the first PC is then given by

$$Z_1 = 0.27\,X_1 + 0.22\,X_2 + 0.94\,X_3$$

   and the second PC is given by

$$Z_2 = 0.16\,X_1 + 0.95\,X_2 - 0.27\,X_3$$

2. The scatter plot of PC2 coefficients vs. PC1 coefficients is shown in Figure 5.38. This plot helps understand which variables have a similar involvement within PCs. As can be seen in Figure 5.38, the variables X_1 and X_2 are located on the left of the plot, while the variable X_3 is located on the bottom right. This is consistent with the values of the coefficients of PC1 and PC2.

```
scatter(A(:,1),A(:,2),3,'o','MarkerFaceColor',[.49 1 .63],'LineWidth',1);
gname(variables); %press the Enter or Escape key to stop labeling.
```

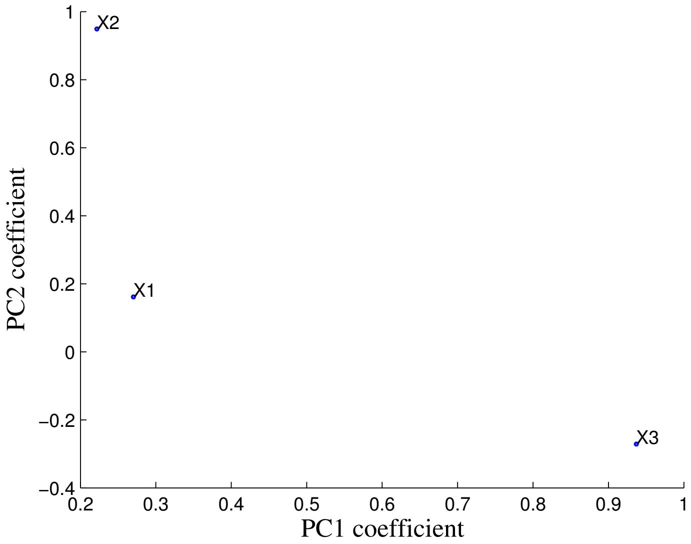


FIGURE 5.38: Coefficients of PC2 vs. PC1 for the phosphorus content data.

3. The correlation matrix between the PCs and the original variables is given by

```
[Ac,Zc,lambdac,Tsquarec]=pca(zscore(X)); %PCA on standardized data
C = Ac*sqrt(diag(lambdac)); %component correlation matrix
```

$$C = \begin{pmatrix} 0.90 & -0.19 & 0.40 \\ 0.70 & 0.71 & -0.09 \\ 0.85 & -0.39 & -0.35 \end{pmatrix}$$

Note that all the three variables are highly correlated with the first PC. We can also see that the organic phosphorus (X_2) is highly correlated with PC2. ■

5.10.3 Scree Plot

The percentage of variance accounted for by the jth PC is called *explained variance*, and it is given by

$$\ell_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\%, \quad j = 1, \ldots, p. \tag{5.10.19}$$

After computing the eigenvectors, we sort them by their corresponding eigenvalues and then we pick the d principal components with the largest eigenvalues. The other components will be discarded. The natural question that arises is: How do we select the value of d? Well, the proportion of variance retained by mapping down from p to d dimensions can be found as the normalized sum of the d largest eigenvalues

$$v = \sum_{j=1}^{d} \ell_j = \frac{\sum_{j=1}^{d} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\%. \tag{5.10.20}$$

In many applications, d is chosen such that a relatively high percentage, say $70 - 95\%$ of explained variance is retained. The remaining variance is assumed to be due to noise. The number of principal components d ($d << p$) can also be determined using the scree graph, which is the plot of the explained variance against the component number k. The optimal number k is usually selected as the one where the kink (elbow) in the curve appears.

Example 5.28 *Using the phosphorus content data*

1. *Compute the explained variance*

2. *Plot the explained variance vs. the number of PCs.*

3. *What would be the lowest-dimensional space to represent the phosphorus content data?*

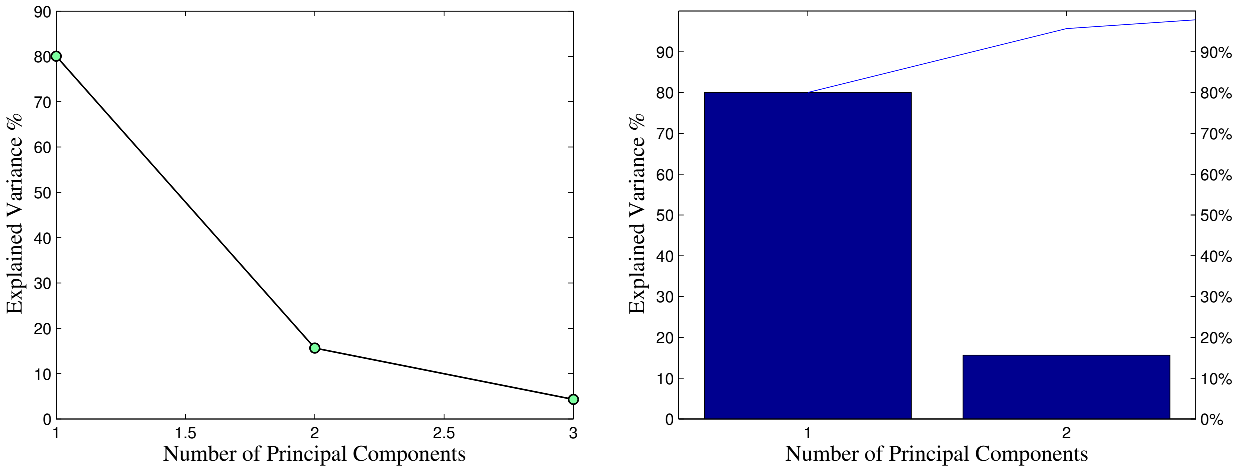
Solution

```
expvar=100*lambda/sum(lambda);%percent of the total variability explained by each PC.
plot(expvar,'ko-','MarkerFaceColor',[.49 1 .63],'LineWidth',1); %Scree plot
pareto(expvar); %Pareto plot
```

1. The explained variance by the three PCs are: $\ell_1 = 80.04\%$, $\ell_2 = 15.65\%$, and $\ell_3 = 4.31\%$. Notice that PC1 and PC2 combined account for 95.69% of the variance in the data.

2. The scree and pareto plots of the explained variance vs. the number of PCs are shown in Figure 5.39.

3. Based on the explained variance by both PC1 and PC2 and also from the scree and Pareto plots, it can be deduced that the lowest-dimensional space to represent the phosphorus content data corresponds to $d = 2$. ■



(a) Scree plot

(b) Pareto plot

FIGURE 5.39: Scree and Pareto plots for the phosphorus content data.

5.10.4 BIPLOT

The plot of the principal component coefficients is a graphical display of the variables, while the plot of the principal component scores is a graphical display of the observations. The biplot was originally proposed by Gabriel (1971) as a graphical tool that allows information on both observations and variables of a data matrix to be displayed graphically, and hence the "bi" in the name. Observations are displayed as points while variables are displayed as vectors. The biplot helps visualize both the principal component coefficients for each variable and the principal component scores for each observation in a single plot. Each of the p variables is represented in the biplot by a vector, and the direction and length of the vector indicates how each variable contributes to the two principal components in the biplot, as shown in Figure 5.40(a). The axes in the biplot represent the principal components (columns of eigenvector matrix A), and the observed variables (rows of A) are represented as vectors. A biplot allows us to visualize the magnitude and sign of each variable's contribution to the first two or three principal components, and how each observation is represented in terms of those components. Each of the n observations is represented in the biplot by a point, and their locations indicate the score of each observation for the two principal components in the plot. For example, points near the left edge of this plot have the lowest scores for the first principal component.

A 2D biplot displays the first two PCs, i.e. PC2 vs. PC1, while a 3D biplot displays the first 3 PCs, i.e., PC1, PC2, and PC3 as shown in Figure 5.40(b). It is usually difficult to visualize a 3D biplot on a 2D plane, but rotating 3D biplot can be very useful when the first two PCs do not explain most of the variance in the data. The axes in the biplot represent the PCs, and the observed variables are represented as vectors.

Example 5.29 *Using the phosphorus content data*

1. *Display the 2D biplot of PC2 vs. PC1*

2. *Display the 3D biplot of PC1, PC2, and PC3*

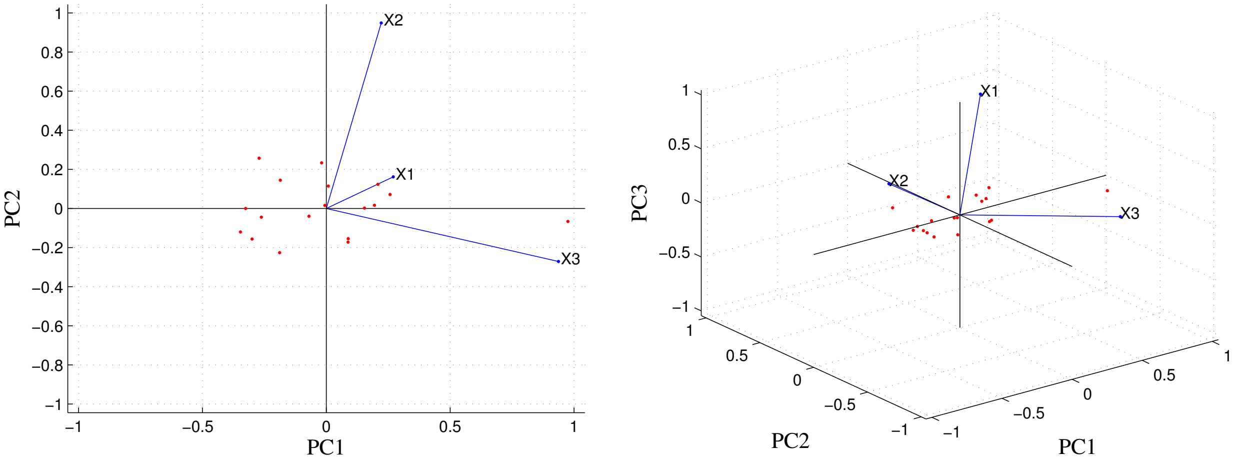
Solution

```
biplot(A(:,1:2),'Scores',Z(:,1:2),'VarLabels',variables); %2D biplot
biplot(A(:,1:3),'Scores',Z(:,1:3),'VarLabels',variables) %3D biplot
```

1. The 2D biplot of PC2 vs. PC1 is shown in Figure 5.40(a). The first principal component, represented in this biplot by the horizontal axis, has positive coefficients for all 3 variables. That corresponds to the 3 vectors directed into the right half of the plot. The second principal component, represented by the vertical axis, has 2 positive coefficients for the variables X_1 and X_2, and 1 negative coefficient for the variable X_3. That corresponds to vectors directed into the top and bottom halves of the plot, respectively. This indicates that this component distinguishes between observations that have high values for the first set of variables and low for the second, and observations

that have the opposite. Each of the 18 observations (rows of scores) is represented in this plot by a point, and their locations indicate the score of each observation for the two principal components in the plot. For example, points near the left edge of this plot have the lowest scores for the first principal component. The angles of between the vectors representing the variables and the PCs indicate the contribution of the variable to the PCs. A narrow angle indicates that the variable plays a major role in the PC. For example, plant phosphorus (X_3) is important in the first PC, while organic phosphorus (X_2) is important in the second PC.

2. The 3D biplot of PC1, PC2, and PC3 is shown in Figure 5.40(b). ■



(a) 2D biplot

(b) 3D biplot

FIGURE 5.40: 2D and 3D biplots for the phosphorus content data.

5.10.5 PCA Control Chart

Recall that the mean and variance of the jth principal component Z_j are $\mu_{Z_j} = 0$ and $var(Z_j) = \lambda_j$. That is, $\sigma_{Z_j} = \sqrt{\lambda_j}$.

PCA Control chart for jth PC:
The upper control limit, center line, and lower control limit of the jth PC are given by

$$\begin{aligned} UCL &= 3\sqrt{\lambda_j} \\ CL &= 0 \\ LCL &= -3\sqrt{\lambda_j} \end{aligned} \tag{5.10.21}$$

The control chart for the first PC of the phosphorus content data is depicted in Figure 5.41, which shows that the sample number 17 is out-of-control.

Example 5.30 *The European Jobs data are the percentage employed in different industries in European countries during 1979. The job categories are agriculture, mining, manufacturing, power supplies, construction, service industries, finance, social and personal services, and transportation and communications. It is important to note that these data were collected during the Cold War. The European Jobs data set contains 26 observations (countries), where each observation has 9 variables:*

- X_1: *agriculture (Agr)*
- X_2: *mining (Min)*
- X_3: *manufacturing (Man)*

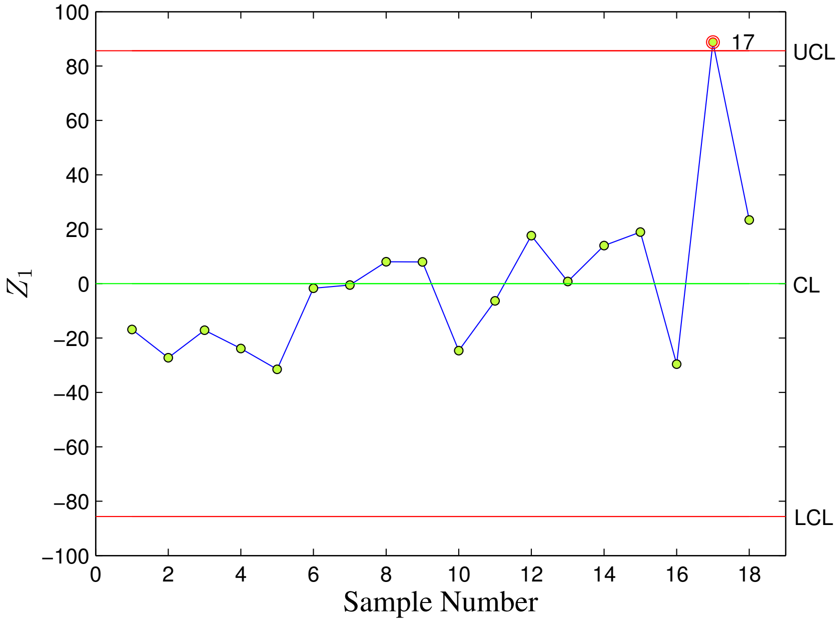


FIGURE 5.41: PC1 control chart for the phosphorus content data.

- *X_4: power supply industries (PS)*
- *X_5: construction (Con)*
- *X_6: service industries (SI)*
- *X_7: finance (Fin)*
- *X_8: social and personal services (SPS)*
- *X_9: transport and communications (TC)*

To load the data set into the MATLAB workspace, type:

```
>> load europeanjobs.mat
>> whos
   Name                Size              Bytes   Class     Attributes
    X                  26x9               1872   double
   countries           26x14               728   char
   description         15x96              2880   char
   variables            9x3                 54   char
```

1. *Display the correlation matrix of the data.*
2. *Display the scatterplot matrix of the data.*
3. *Plot the second PC score vs. the first PC score.*
4. *Determine the first and second PCs*
5. *Display the scatter plot of PC2 coefficients vs. PC1 coefficients, and label the points*
6. *Compute the explained variance, and plot it against the number of PCs. What would be the lowest-dimensional space to represent the phosphorus content data?*
7. *Display the 2D biplot of PC2 vs. PC1. Then, display the 3D biplot of PC1, PC2, and PC3*
8. *Plot the Hotelling and first PC control charts. Identify the out-of-control points*

Solution:

MATLAB code

```
>> load europeanjobs.mat
>> whos
   Name               Size             Bytes  Class     Attributes
   X                  26x9              1872  double
   countries          26x1              1970  cell
   description        15x96             2880  char
   variables          1x9               588   cell
>> [n,p]=size(X); %size of data matrix
>> R = corrcoef(X); %correlation matrix
>> plotmatrix(X);
>> boxplot(X);
>> [A,Z,lambda,Tsquare]=pca(X); %perform PCA on data matrix using covariance
>> scatter(Z(:,1),Z(:,2),3,'o','MarkerFaceColor',[.49 1 .63],'LineWidth',1);
```

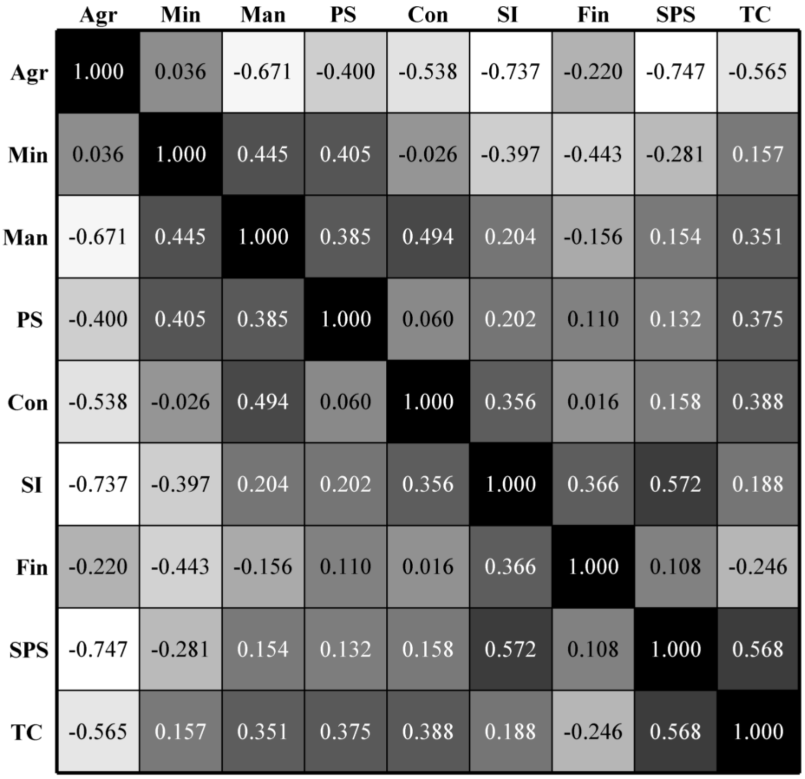
The data matrix X for the European jobs data is a 26 × 9 matrix given by

$$\boldsymbol{X} = \begin{array}{l} \\ \text{Belgium} \\ \text{Denmark} \\ \text{France} \\ \text{W. Germany} \\ \text{Ireland} \\ \text{Italy} \\ \text{Luxembourg} \\ \text{Netherlands} \\ \text{United Kingdom} \\ \text{Austria} \\ \text{Finland} \\ \text{Greece} \\ \text{Norway} \\ \text{Portugal} \\ \text{Spain} \\ \text{Sweden} \\ \text{Switzerland} \\ \text{Turkey} \\ \text{Bulgaria} \\ \text{Czechoslovakia} \\ \text{E. Germany} \\ \text{Hungary} \\ \text{Poland} \\ \text{Rumania} \\ \text{USSR} \\ \text{Yugoslavia} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \text{Agr} & \text{Min} & \text{Man} & \text{PS} & \text{Con} & \text{SI} & \text{Fin} & \text{SPS} & \text{TC} \end{array} \\ \left(\begin{array}{ccccccccc}
3.3 & 0.9 & 27.6 & 0.9 & 8.2 & 19.1 & 6.2 & 26.6 & 7.2 \\
9.2 & 0.1 & 21.8 & 0.6 & 8.3 & 14.6 & 6.5 & 32.2 & 7.1 \\
10.8 & 0.8 & 27.5 & 0.9 & 8.9 & 16.8 & 6.0 & 22.6 & 5.7 \\
6.7 & 1.3 & 35.8 & 0.9 & 7.3 & 14.4 & 5.0 & 22.3 & 6.1 \\
23.2 & 1.0 & 20.7 & 1.3 & 7.5 & 16.8 & 2.8 & 20.8 & 6.1 \\
15.9 & 0.6 & 27.6 & 0.5 & 10.0 & 18.1 & 1.6 & 20.1 & 5.7 \\
7.7 & 3.1 & 30.8 & 0.8 & 9.2 & 18.5 & 4.6 & 19.2 & 6.2 \\
6.3 & 0.1 & 22.5 & 1.0 & 9.9 & 18.0 & 6.8 & 28.5 & 6.8 \\
2.7 & 1.4 & 30.2 & 1.4 & 6.9 & 16.9 & 5.7 & 28.3 & 6.4 \\
12.7 & 1.1 & 30.2 & 1.4 & 9.0 & 16.8 & 4.9 & 16.8 & 7.0 \\
13.0 & 0.4 & 25.9 & 1.3 & 7.4 & 14.7 & 5.5 & 24.3 & 7.6 \\
41.4 & 0.6 & 17.6 & 0.6 & 8.1 & 11.5 & 2.4 & 11.0 & 6.7 \\
9.0 & 0.5 & 22.4 & 0.8 & 8.6 & 16.9 & 4.7 & 27.6 & 9.4 \\
27.8 & 0.3 & 24.5 & 0.6 & 8.4 & 13.3 & 2.7 & 16.7 & 5.7 \\
22.9 & 0.8 & 28.5 & 0.7 & 11.5 & 9.7 & 8.5 & 11.8 & 5.5 \\
6.1 & 0.4 & 25.9 & 0.8 & 7.2 & 14.4 & 6.0 & 32.4 & 6.8 \\
7.7 & 0.2 & 37.8 & 0.8 & 9.5 & 17.5 & 5.3 & 15.4 & 5.7 \\
66.8 & 0.7 & 7.9 & 0.1 & 2.8 & 5.2 & 1.1 & 11.9 & 3.2 \\
23.6 & 1.9 & 32.3 & 0.6 & 7.9 & 8.0 & 0.7 & 18.2 & 6.7 \\
16.5 & 2.9 & 35.5 & 1.2 & 8.7 & 9.2 & 0.9 & 17.9 & 7.0 \\
4.2 & 2.9 & 41.2 & 1.3 & 7.6 & 11.2 & 1.2 & 22.1 & 8.4 \\
21.7 & 3.1 & 29.6 & 1.9 & 8.2 & 9.4 & 0.9 & 17.2 & 8.0 \\
31.1 & 2.5 & 25.7 & 0.9 & 8.4 & 7.5 & 0.9 & 16.1 & 6.9 \\
34.7 & 2.1 & 30.1 & 0.6 & 8.7 & 5.9 & 1.3 & 11.7 & 5.0 \\
23.7 & 1.4 & 25.8 & 0.6 & 9.2 & 6.1 & 0.5 & 23.6 & 9.3 \\
48.7 & 1.5 & 16.8 & 1.1 & 4.9 & 6.4 & 11.3 & 5.3 & 4.0
\end{array}\right) \end{array}.$$

1. The correlation matrix of the data is shown in Figure 5.42. From this correlation matrix plot we can see that the percentage of people employed in agriculture is negatively correlated with virtually of the other employment areas indicating a contrast between industrial and agricultural economies. We also see that the percent of people employed in manufacturing is positively correlated with employment areas which are required support manufacturing such as power supply, mining, construction and transportation. Other interesting relationships between these variables are also evident.

MATLAB code

```
>> R = corrcoef(X); %correlation matrix
>> imagesc(R);
>> set(gca,'XTick',1:p); set(gca,'YTick',1:p);
```



	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Agr	1.000	0.036	-0.671	-0.400	-0.538	-0.737	-0.220	-0.747	-0.565
Min	0.036	1.000	0.445	0.405	-0.026	-0.397	-0.443	-0.281	0.157
Man	-0.671	0.445	1.000	0.385	0.494	0.204	-0.156	0.154	0.351
PS	-0.400	0.405	0.385	1.000	0.060	0.202	0.110	0.132	0.375
Con	-0.538	-0.026	0.494	0.060	1.000	0.356	0.016	0.158	0.388
SI	-0.737	-0.397	0.204	0.202	0.356	1.000	0.366	0.572	0.188
Fin	-0.220	-0.443	-0.156	0.110	0.016	0.366	1.000	0.108	-0.246
SPS	-0.747	-0.281	0.154	0.132	0.158	0.572	0.108	1.000	0.568
TC	-0.565	0.157	0.351	0.375	0.388	0.188	-0.246	0.568	1.000

FIGURE 5.42: Correlation matrix of the European jobs data.

```
>> set(gca,'XTickLabel',variables); set(gca,'YTickLabel',variables);
>> axis([0 p+1 0 p+1]); grid; colorbar;
```

2. When interpreting correlations it is important to visualize the bivariate relationships between all pairs of variables. This can be achieved by looking at a scatterplot matrix, which is shown in Figure 5.43

3. The plot the second PC score vs. the first PC score is shown in Figure 5.44. The labels displayed in Figure 5.44(b) represent the names of the countries.

MATLAB code

```
>> scatter(Z(:,1),Z(:,2),15,'ko','MarkerFaceColor',[.49 1 .63],'LineWidth',1);
>> xlabel('PC1 score','fontsize',14,'fontname','times');
>> ylabel('PC2 score','fontsize',14,'fontname','times');
>> gname(countries); %press Enter or Escape key to stop labeling.
```

4. Using the columns of the eigenvector matrix

$$A = \begin{pmatrix} 0.8918 & -0.0068 & 0.1185 & -0.0968 & -0.1800 & 0.1526 & -0.0916 & -0.0687 & 0.3354 \\ 0.0019 & 0.0923 & 0.0794 & -0.0102 & 0.0011 & -0.4564 & 0.7665 & -0.2905 & 0.3240 \\ -0.2713 & 0.7703 & 0.1847 & -0.0104 & -0.3360 & 0.2009 & -0.1620 & -0.0741 & 0.3375 \\ -0.0084 & 0.0120 & -0.0068 & 0.0181 & 0.0025 & -0.2309 & 0.0629 & 0.9092 & 0.3399 \\ -0.0496 & 0.0690 & -0.0773 & -0.0829 & 0.7243 & 0.5584 & 0.1943 & 0.0045 & 0.3253 \\ -0.1918 & -0.2344 & -0.5796 & -0.6076 & -0.2659 & 0.0216 & -0.0879 & -0.1044 & 0.3367 \\ -0.0311 & -0.1301 & -0.4700 & 0.7812 & -0.1211 & 0.0553 & -0.0800 & -0.1228 & 0.3344 \\ -0.2980 & -0.5668 & 0.5977 & 0.0483 & -0.2359 & 0.2479 & -0.0045 & -0.0521 & 0.3324 \\ -0.0454 & -0.0099 & 0.1594 & -0.0378 & 0.4349 & -0.5459 & -0.5675 & -0.2238 & 0.3342 \end{pmatrix}$$

the first PC is then given by

$$Z_1 = 0.89\,X_1 - 0.27\,X_3 - 0.192\,X_6 - 0.298\,X_8 = 0.89\,\texttt{Agr} - 0.27\,\texttt{Man} - 0.192\,\texttt{SI} - 0.298\,\texttt{SPS}$$

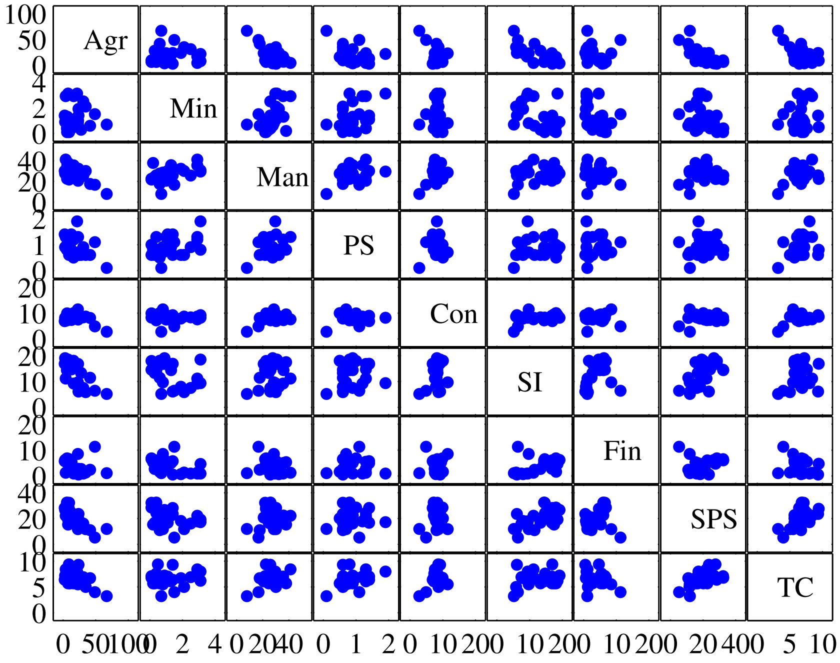
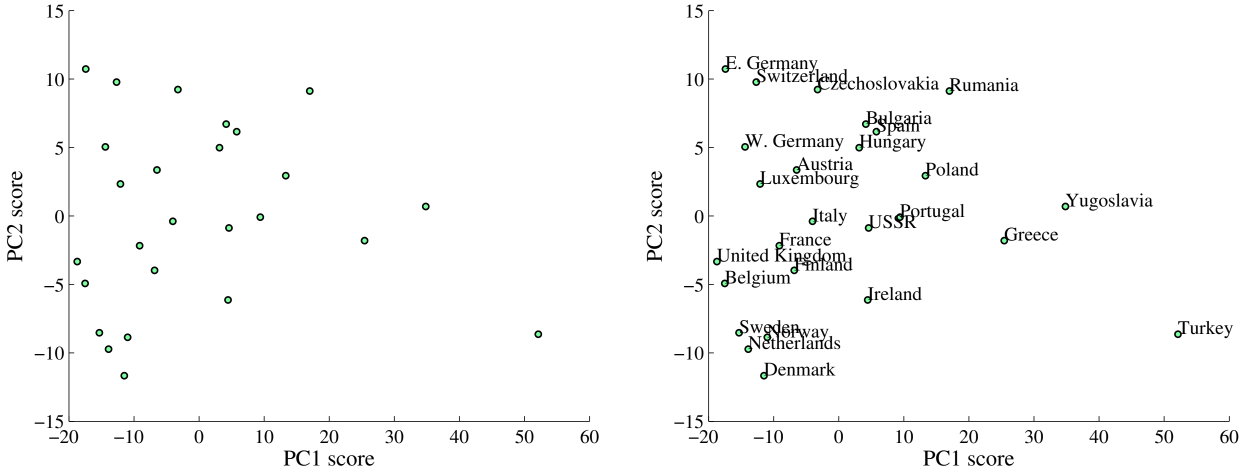


FIGURE 5.43: Scatterplot matrix of the European jobs data.



(a) unlabeled plot (b) labeled plot

FIGURE 5.44: PC2 vs. PC1 score for the European jobs data.

We can see that the first PC is essentially a contrast between agriculture and industrial/urban employment areas. This is evidenced by the positive coefficient for agriculture and the negative coefficients for manufacturing, service industries, and social and personal services.

The second PC is given by

$$Z_2 = 0.77\,X_3 - 0.234\,X_6 - 0.13\,X_7 - 0.567\,X_8 = 0.77\,\texttt{Man} - 0.234\,\texttt{SI} - 0.13\,\texttt{Fin} - 0.567\,\texttt{SPS}$$

We can see that the second PC appears to be a contrast between manufacturing and non-industrial areas such as service industries and finance. This is evidenced by the positive coefficient for manufacturing and the negative coefficients for service industries, finance, and social and personal services.

5. The scatter plot of PC2 coefficients vs. PC1 coefficients is shown in Figure 5.45. This plot helps understand which variables have a similar involvement within PCs. As can be seen in Figure 5.45, the variables Agr is located on

the right of the plot, while the other variables are located on the left of the plot. This is consistent with the values of the coefficients of PC1 and PC2.

MATLAB code

```
>> scatter(A(:,1),A(:,2),3,'o','MarkerFaceColor',[.49 1 .63],'LineWidth',1);
>> gname(variables); %press Enter or Escape key to stop labeling.
```

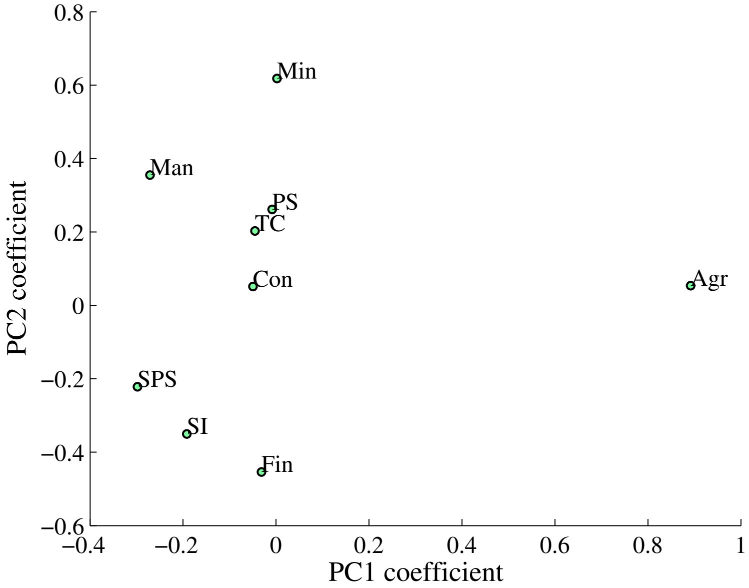
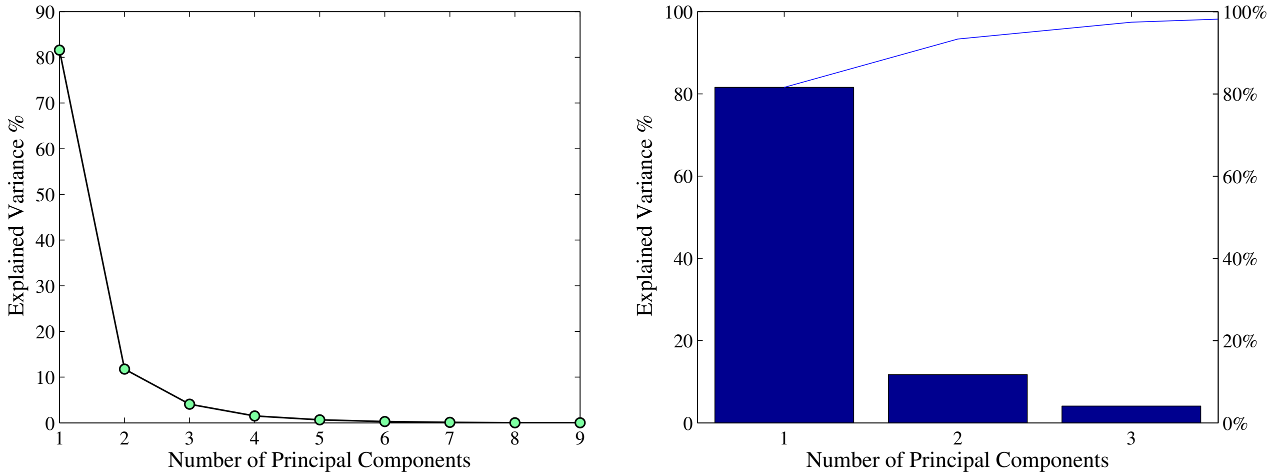


FIGURE 5.45: Coefficients of PC2 vs. PC1 for the European jobs data.

6. The explained variance by the three PCs are: $\ell_1 = 81.58\%$, $\ell_2 = 11.75\%$, and $\ell_3 = 4.09\%$. Notice that PC1 and PC2 combined account for 93.33% of the variance in the data. The scree and Pareto plots of the explained variance vs. the number of PCs are shown in Figure 5.46. Based on the explained variance by both PC1 and PC2 and also from the scree and Pareto plots, it can be deduced that the lowest-dimensional space to represent the European jobs data corresponds to $d = 2$.

7. The 2D biplot of PC2 vs. PC1 is shown in Figure 5.47(a). The axes in the biplot represent the principal components (columns of A), and the observed variables (rows of A) are represented as vectors. Each observation (row of Z) is represented as a point in the biplot. From Figure 5.47(a), we can see that the first principal component has 1 positive coefficient for the first variable `Agr` and 3 negative coefficients for the variables `Man`, `SI`, and `SPS`. That corresponds to 1 vector directed into the right half of the plot, and 3 vectors directed into the left half of the plot, respectively. The second principal component, represented by the vertical axis, has 1 positive coefficient for the variable `Man`, and 3 negative coefficients for the variables `SI`, `FIN`, and `SPS`. That corresponds to vectors directed into the top and bottom halves of the plot, respectively. This indicates that this component distinguishes between observations that have high values for the first set of variables and low for the second, and observations that have the opposite. Each of the 26 countries is represented in this plot by a red point, and their locations indicate the score of each observation for the two principal components in the plot. For example, points near the left edge of this plot have the lowest scores for the first principal component. The variables are represented by rays extending out from the plot origin. Rays that tend to point in the same direction represent variables that are positively correlated. For example we can see that the rays for manufacturing, mining, and power supply all point in the same direction, indicating the positive correlation of these employment areas with one another. The ray for agriculture is fairly isolated indicating its weak positive correlation and more often times negative correlation with the other employment areas. The cases are plotted in accordance with their scores on the first three PCs. We can see that Turkey and Yugoslavia both extend far to the right in the direction of ray for the agriculture variable. This indicates that these countries have a larger percentage of the workforce employed in agriculture in comparison to the other countries in this data set. We can also see that Norway has a relatively large percentage of its workforce employed in the social and service areas. The 3D biplot of PC1, PC2, and PC3 is shown in Figure 5.47(b)



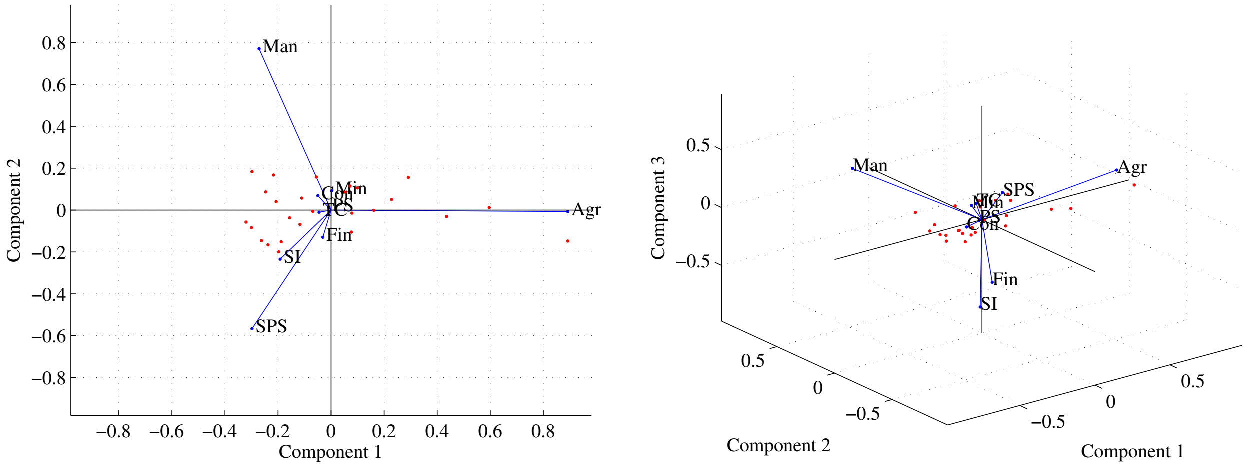
(a) Scree plot

(b) Pareto plot

FIGURE 5.46: Scree and Pareto plots for the European jobs data.

MATLAB code

```
>> expvar=100*variance/sum(variance);%percent of the total variability explained by each PC.
>> plot(expvar,'ko-','MarkerFaceColor',[.49 1 .63],'LineWidth',1);
>> figure;
>> pareto(expvar);
```



(a) 2D biplot

(b) 3D biplot

FIGURE 5.47: 2D and 3D biplots for the European jobs data.

MATLAB code

```
biplot(A(:,1:2),'Scores',Z(:,1:2),'VarLabels',variables)
figure; biplot(A(:,1:3),'Scores',Z(:,1:3),'VarLabels',variables)
```

8. The Hotelling and first PC charts are displayed in Figure 5.48. The Hotelling chart indicates that the samples 7 (Luxembourg), 18 (Yugoslavia), and 26 (Turkey) are out-of-control. All the plotted points on the first PC chart are within the control limits.

MATLAB code

```
>> alpha = 0.05;
>> [outliers, h] = tsquarechart(X,alpha); %T^2 chart
>> figure;
>> k=1;
>> [outliers, h] = pcachart(X,k); %1st PC control chart
```

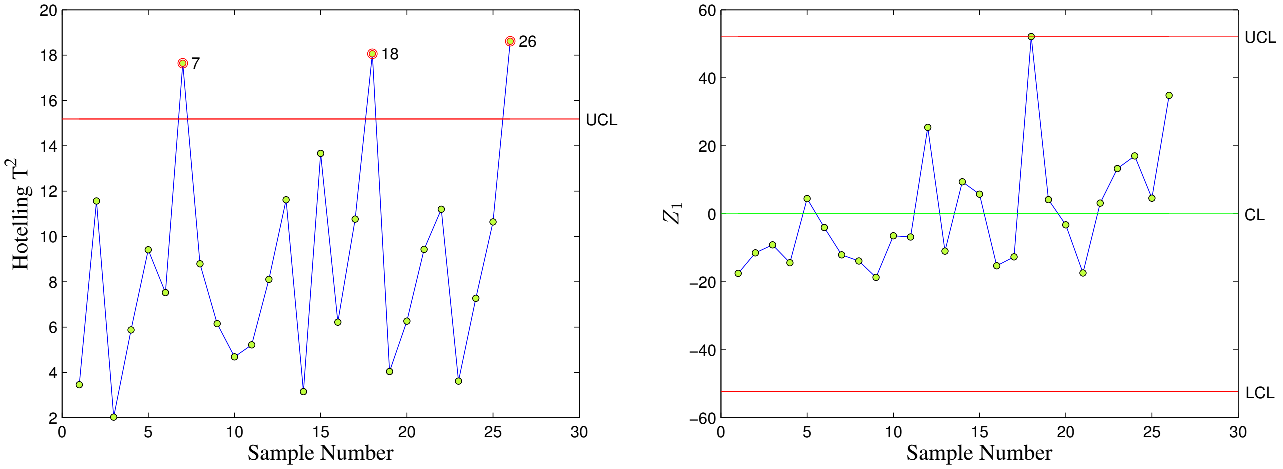


FIGURE 5.48: Hotelling and PCA charts for the European jobs data.

5.11 PROBLEMS

❶ The management of a bank has embarked on a program of statistical process control and has decided to use variable control charts to study the waiting time of customers during the peak noon to 1 p.m. lunch hour to detect special causes of variation. Four customers are selected during the one-hour period; the first customer to enter the bank every 15 minutes. Each set of four measurements makes up a subgroup (sample). Table 5.18 lists the waiting time (operationally defined as the time from when the customer enters the line until he or she begins to be served by the teller) for 20 days.

TABLE 5.18: Waiting Time for Customers at a Bank.

Sample	Data			
Number	X_1	X_2	X_3	X_4
1	7.2	8.4	7.9	4.9
2	5.6	8.7	3.3	4.2
3	5.5	7.3	3.2	6.0
4	4.4	8.0	5.4	7.4
5	9.7	4.6	4.8	5.8
6	8.3	8.9	9.1	6.2
7	4.7	6.6	5.3	5.8
8	8.8	5.5	8.4	6.9
9	5.7	4.7	4.1	4.6
10	3.7	4.0	3.0	5.2
11	2.6	3.9	5.2	4.8
12	4.6	2.7	6.3	3.4
13	4.9	6.2	7.8	8.7
14	7.1	6.3	8.2	5.5
15	7.1	5.8	6.9	7.0
16	6.7	6.9	7.0	9.4
17	5.5	6.3	3.2	4.9
18	4.9	5.1	3.2	7.6
19	7.2	8.0	4.1	5.9
20	6.1	3.4	7.2	5.9

a) Construct a table that shows the waiting time data with extra columns displaying the sample means and sample ranges.

b) Estimate the process mean and standard deviation.

c) Construct the R- and the $\overline{X}$-charts. Identify the out-of-control points using all Western Electric rules. If necessary, revise your control limits, assuming that any samples that violate Western Electric rules can de discarded.

❷ A sample data set called `parts.mat` in the MATLAB Statistics Toolbox contains measurements on newly machined parts, taken at one hour intervals for 36 hours. Each row of the runout matrix contains the measurements for 4 parts chosen at random. The values indicate, in thousandths of an inch, the amount the part radius differs from the target radius. To load the data set into the MATLAB workspace, type:

```
>> load parts
>> whos
  Name          Size             Bytes  Class     Attributes
  runout        36x4              1152  double
```

(i) Construct the R- and the $\overline{X}$-charts. Identify the out-of-control points. If necessary, revise your control limits, assuming that any samples that plot outside the control limits can de discarded.

(ii) Assuming the process is in control, estimate the process mean and standard deviation.

(iii) Construct the s- and the $\overline{X}$-charts. Identify the out-of-control points. If necessary, revise your control limits, assuming that any samples that plot outside the control limits can de discarded.

❸ Table 5.19 presents the weights, in ounces, for a sequence of 15 rational subgroup samples of potato chips, with $n = 4$ for each sample. Assume that the specifications are 14 ± 1.37.

TABLE 5.19: Potato chip Data.

Sample Number	Package Weights (oz)			
	X_1	X_2	X_3	X_4
1	15.01	14.98	15.16	14.80
2	15.09	15.14	15.08	15.03
3	15.04	15.10	14.93	15.13
4	14.90	15.03	14.94	14.92
5	15.04	15.05	15.08	14.98
6	14.96	14.81	14.96	14.91
7	15.01	15.10	14.90	15.03
8	14.71	14.92	14.77	14.95
9	14.81	14.80	14.64	14.95
10	15.03	14.89	14.99	15.03
11	15.16	14.91	14.95	14.83
12	14.92	15.05	15.01	15.02
13	15.06	15.03	14.95	15.02
14	14.99	15.14	15.04	15.11
15	14.94	15.08	14.90	15.17

(i) Construct the R- and the $\overline{X}$-charts. Is the process under statistical control? Explain.

(ii) Construct the s- and the $\overline{X}$-charts. Is the process under statistical control? Explain.

(iii) Assuming that the package weights are normally distributed, calculate the process capability index and the proportion of the product that will not meet specifications.

(iv) Comment on the ability of the process to produce items that meet specifications?

❹ The diameter of holes is measured in consecutive order by an automatic sensor. The results of measuring 25 holes are given in in Table 5.20. Assume the target diameter is 10 millimeters.

(i) Estimate the process standard deviation

TABLE 5.20: Diameter measurements.

Sample	Diameter	Sample	Diameter
1	9.94	14	9.99
2	9.93	15	10.12
3	10.09	16	9.81
4	9.98	17	9.73
5	10.11	18	10.14
6	9.99	19	9.96
7	10.11	20	10.06
8	9.84	21	10.11
9	9.82	22	9.95
10	10.38	23	9.92
11	9.99	24	10.09
12	10.41	25	9.85
13	10.36		

(ii) Set up and apply a tabular cusum for this process, using standardized values $h = 5$ and $k = 0.5$. Does the process appear to be operating in a state of statistical control at the desired target level?

(iii) Apply an EWMA control chart to these data using $\lambda = 0.4$ and $L = 3$. Interpret this chart.

❺ The wafer dataset (`wafer.mat`) is based on one found in "Statistical case studies for industrial process improvement" by Czitrom and Spagon. It consists of oxide layer thickness measured in 9 locations on each of 116 semiconductor wafers. The measurements were taken by position on the wafer as shown in Figure 5.49. Note that the first is in the center, the next 4 halfway out, and the last 4 on the edge. To load the data set into the MATLAB workspace, type:

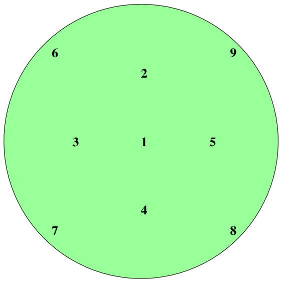


FIGURE 5.49: Layout of wafer data.

```
>> load wafer.mat
>> whos
   Name               Size              Bytes   Class     Attributes
   X                 116x9               8352   double
   description        5x89               890    char
   variables          9x2                 36    char
```

a) Display the correlation matrix of the data.

b) Display the side-by-side boxplots of the data. Comment on the plots.

c) Plot the second PC score vs. the first PC score. Comment on the plot.

d) Determine the first and second PCs.

e) Display the scatter plot of PC2 coefficients vs. PC1 coefficients, and label the points. Comment on the plot.

f) Compute the explained variance, and plot it against the number of PCs. What would be the lowest-dimensional space to represent the phosphorus content data?

g) Display the 2D biplot of PC2 vs. PC1. Then, display the 3D biplot of PC1, PC2, and PC3. Comment on the plots.

h) Plot the Hotelling and first PC control charts. Identify the out-of-control points.

5.12 REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, 6th edition, 2009.

[2] K. Yang and J. Trewn, *Multivariate Statistical Process Control in Quality Management*, Mc-Graw Hill Professional, 2004.

[3] K.H. Chen, D.S. Boning, and R.E. Welch, “Multivariate statistical process control and signature analysis using eigenfactor detection methods,” *Proc. Symposium on the Interface of Computer Science and Statistics*, Costa Mesa, CA, June 2001.

[4] N.D. Tracy, J.C. Young, and R.L. Mason, “Multivariate quality control charts for individual observations,” *Journal of Quality Technology*, vol. 24, no. 22, pp. 88-95, 1992.

[5] J.A. Vargas, “Robust estimation in multivariate control charts for individual observations,” *Journal of Quality Technology*, vol. 35, no. 4, pp. 367-376, 2003.

[6] J.H. Sullivan and W.H. Woodall, “A comparison of multivariate control charts for individual observations,” *Journal of Quality Technology*, vol. 28, no. 24, pp. 398-408, 1996.

[7] I.T. Jolliffe, *Principal Component Analysis*, New York: Springer, 1986.

[8] T.F. Cox, *An Introduction to Multivariate Data Analysis*, Hodder Arnold, 2005.

Chapter 6

Regression and Analysis of Variance

> The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic.
>
> *Ronald Fisher*

Regression analysis is a statistical tool for the investigation of relationships between variables. The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a given set of data observations. Regression analysis is very useful in quality control because you can determine how much to change if you have a quality issue. The simplest type of regression analysis is linear regression, which is a method for predicting the value of a dependent variable, based on the value of one or more independent variables. For example, in the relationship between age and weight of a cow during a specific phase of production, age is the independent variable and weight is the dependent variable. As the cows age increases, its weight will also increase.

On the other hand, Analysis of Variance (ANOVA) is a method for testing differences among two or more means by analyzing variance. An ANOVA conducted on an experimental design in which there is only one factor is called a *one-way ANOVA*. If an experiment has two factors, then the ANOVA is called a *two-way ANOVA*. ANOVA is used in the hope of showing that there is a difference between distribution means. For example, several groups of patients, all suffering from high blood pressure, are submitted to several new treatments (one treatment for each group). These treatments are expected to have different efficacies, and it is hoped that some of them will turn out to be particularly effective. After the treatments, ANOVA will be used on blood pressure measurements, and it is hoped that it will reject the hypothesis that all treatments are equally effective (or ineffective).

6.1 Linear Regression

Regression analysis concerns the study of relationships between quantitative variables with the object of identifying, estimating, and validating the relationship. The estimated relationship can then be used to predict one variable from the value of the other variable(s). Suppose we have n pairs of observations (x_i, y_i). Plotting a scatter plot is an important preliminary step prior to undertaking a formal statistical analysis of the relationship between the variables x and y. The variable x is called **independent** or **predictor** variable, while the variable y is called **dependent** or **response** variable. The scatter plot shows whether or not there is a linear (straight-line) relation between the dependent and independent variables. For example, the scatter plot of the car plant data given in Table 6.1 is shown in Figure 6.1 and it reveals that the relationship between x and y is approximately linear; that is, the points seem to cluster around a straight line.

Example 6.1 *The manager of a car plant wishes to investigate how the plant's electricity usage depends upon the plant's production. A sample of 12-month data recorded from January to December of the previous year is selected. The production x (in million dollars) and the electricity usage y (in million kWh) for each month are listed in Table 6.1. Construct a scatter plot of the data.*

Solution:

MATLAB code

```
>> x = [4.51 3.58 4.31 5.06 5.64 4.99 5.29 5.83 4.70 5.61 4.90 4.20]’;
>> y = [2.48 2.26 2.47 2.77 2.99 3.05 3.18 3.46 3.03 3.26 2.67 2.53]’;
>> scatter(x,y,’ko’,’MarkerFaceColor’,[.49 1 .63]);
```

TABLE 6.1: Car plant data.

Observation	Production x	Electricity usage y
January	4.51	2.48
February	3.58	2.26
March	4.31	2.47
April	5.06	2.77
May	5.64	2.99
June	4.99	3.05
July	5.29	3.18
August	5.83	3.46
September	4.70	3.03
October	5.61	3.26
November	4.90	2.67
December	4.20	2.53

The output scatter plot is displayed in Figure 6.1. ■

Since a linear relation is the simplest relationship to handle mathematically, we present the details of the statistical regression analysis for this case. In simple linear regression, we assume that each observed value y_i of the response variable Y_i can be described as follows:

Linear Regression Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{6.1.1}$$

where

- y_i is the dependent variable that we wish to predict or explain
- x_i is the independent variable used to predict or explain y_i
- β_0 and β_1 are unknown parameters, intercept and slope of the line, respectively, also called *regression coefficients*.
- $\varepsilon_i \sim N(0, \sigma^2)$ is a random error with unknown variance

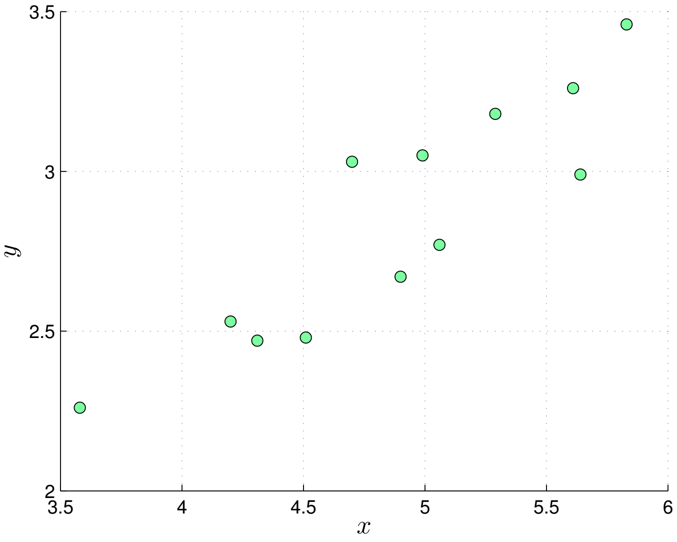


FIGURE 6.1: Scatter plot of the car plant data.

The term *linear* is used because Eq. (6.1.1) is a linear function of the unknown parameters β_0 and β_1. The simple linear regression model given by Eq. (6.1.1) may be written in matrix form as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.1.2}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{6.1.3}$$

Since the random error $\varepsilon_i \sim N(0, \sigma^2)$, it follows that the values y_i are observations from the independent random variables $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Using the least-square method to fit the simple linear regression model (6.1.1), the unknown parameters β_0 and β_1 can be found by minimizing the sum of squared deviations

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2 \tag{6.1.4}$$

Thus, the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize L must satisfy

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{6.1.5}$$

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{6.1.6}$$

Simplifying these two equations yields to the *least squares normal equations* given by

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \displaystyle\sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} \Longrightarrow \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{6.1.7}$$

Using the identity $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, we can also write S_{xx} and S_{xy} as

$$S_{xx} = \sum_{i=1}^{n} x_i(x_i - \bar{x}) \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x}). \tag{6.1.8}$$

Therefore, the estimated or *fitted value* of y_i is given by the prediction or regression equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1, respectively.

Least-squares estimates:

Slope: $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

y-intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Example 6.2 *Using the data of Table 6.1,*

1. *Compute the least-squares estimate $\hat{\beta}_0$ and $\hat{\beta}_1$*

2. *Find the regression line and plot it on the scatter plot.*

Solution: We treat the production as the predictor variable, x, and electricity usage as the response variable, y.

```
n = 12; % Sample size
x = [4.51 3.58 4.31 5.06 5.64 4.99 5.29 5.83 4.70 5.61 4.90 4.20]';
y = [2.48 2.26 2.47 2.77 2.99 3.05 3.18 3.46 3.03 3.26 2.67 2.53]';
scatter(x,y,'ko','MarkerFaceColor',[.49 1 .63]);
X = [ones(size(x)) x];
b = regress(y,X);
beta0hat = b(1); % Estimate of intercept
beta1hat = b(2); % Estimate of slope

% Sums of squares
Sxx = (n-1)*var(x);
Syy = (n-1)*var(y);
Sxy = beta1hat*Sxx;
SSr = beta1hat*Sxy; % Residual sum of squares
SSt = Syy; % Total sum of squares

yhat = beta0hat + beta1hat*x; % Regression line
hold on; plot(x,yhat,'r');
```

1. The returned values of the estimated parameters intercept and slope are $\hat{\beta}_0 = 0.409$ and $\hat{\beta}_1 = 0.499$, respectively.

2. The estimated regression line is $\hat{y} = 0.409 + 0.499x$. The value of $\hat{\beta}_1$ means that for each increase of 1 unit in x, the mean value of y is estimated to increase by 0.499. In other words, the positive slope value implies that the estimated mean electricity usage increases by 0.499 million kWh for each additional million of production. Figure 6.2 shows a scatter plot of the data and the graph of the estimated regression line (in red color). ■

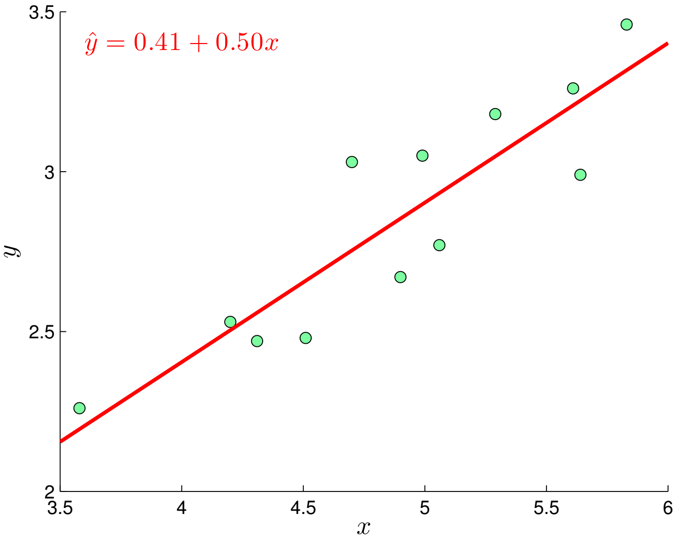


FIGURE 6.2: Scatter plot and regression line of the car plant data.

Example 6.3 *Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1*

Solution: Since $E(y_i) = \beta_0 + \beta_1 x_i$ and $E(S_{xy}) = \sum_{i=1}^{n} E(y_i)(x_i - \bar{x})$, it follows that $E(S_{xy}) = \beta_1 S_{xx}$. Thus, $E(\hat{\beta}_1) = \beta_1$. Also, $E(\bar{y}) = \beta_0 + \beta_1 \bar{x}$ implies that $E(\hat{\beta}_0) = \beta_0$. Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. ■

Example 6.4 *Find the variances of β_0 and β_1*

Solution: Since $var(y_i) = \sigma^2$ and $var(S_{xy}) = \sum_{i=1}^{n} var(y_i)(x_i - \bar{x})^2$, it follows that $var(S_{xy}) = \sigma^2 S_{xx}$. Thus, $var(\hat{\beta}_1) = \sigma^2/S_{xx}$. Similarly, it can be shown that $var(\hat{\beta}_0) = \sigma^2(1/n + \bar{x}^2/S_{xx})$. Therefore, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$ and $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(1/n + \bar{x}^2/S_{xx}))$. ■

6.1.1 Residual Analysis

In order to check the regression model assumptions, we need to compute and examine the normality of the residuals $e_i = y_i - \hat{y}_i$, where y_i is the observed value and $\hat{y}_i$ is the predicted (fitted) value. The residuals are used to evaluate the goodness of fit of the regression (also called least square) line. The calculated predicted values and residuals for the car plant data are shown in Table 6.2.

Table 6.2: Predicted values and residuals for car plant data.

Production x	Electricity usage y	Predicted value $\hat{y}$	Residual $e = y - \hat{y}$
4.51	2.48	2.6588	-0.1788
3.58	2.26	2.1949	0.0651
4.31	2.47	2.5590	-0.0890
5.06	2.77	2.9331	-0.1631
5.64	2.99	3.2225	-0.2325
4.99	3.05	2.8982	0.1518
5.29	3.18	3.0479	0.1321
5.83	3.46	3.3172	0.1428
4.70	3.03	2.7535	0.2765
5.61	3.26	3.2075	0.0525
4.90	2.67	2.8533	-0.1833
4.20	2.53	2.5041	0.0259

The first assumption we consider is that the random errors ε_i are normally distributed. A **normal probability plot** of the residuals (or standardized residuals $z_i = (y_i - \hat{y}_i)/\hat{\sigma}$) will give an indication of whether or not the assumption of normality of the random errors is appropriate. Recall that a normal probability plot is found by plotting the quantiles of the observed sample against the corresponding quantiles of a standard normal distribution $N(0,1)$. If the normal probability plot shows a straight line, it is reasonable to assume that the observed sample comes from a normal distribution. If, on the other hand, the points deviate from a straight line, there is statistical evidence against the assumption that the random errors are an independent sample from a normal distribution.

The normal probability plot for the car plant data, displayed in Figure 6.3, shows a relatively linear pattern. The fact that the points in the lower and upper extremes of the normal probability plot do not deviate significantly from the straight-line pattern indicates the normality assumption of the errors is not violated. That is, we can quite reasonably conclude that the normal distribution provides a good model for the residuals.

```
% Normal probability plot of residuals
res = y-yhat; % Residuals
normplot(res); % Display normal probability plot
```

The second assumption we consider is that the random errors ε_i have zero mean and same variance σ^2. A **residual plot**, which is a scatter plot of the residuals e_i against the predicted (fitted) values $\hat{y}_i$ or against the independent variable x, can be used to check this assumption. More precisely, if the second assumption is satisfied then we would expect the residuals to vary randomly around zero and we would also expect the spread of the residuals to be about the same throughout the residual plot. That is, if there is no violation in assumptions, this scatter plot should look like a band around zero with randomly distributed points and no discernible pattern. There should be no relation between the residuals and fitted values (or independent variables).

The residual plot against the predicted values for the car plant data, shown in Figure 6.4, indicates that the points seem to be fluctuating randomly around zero. The residual plot against the independent variable is also displayed in Figure 6.4. Although there is widespread scatter in the residual plots, there is no apparent pattern or relationship between the residuals and $\hat{y}$ (or x). The residuals appear to be evenly spread above and below zero. Thus, the residual

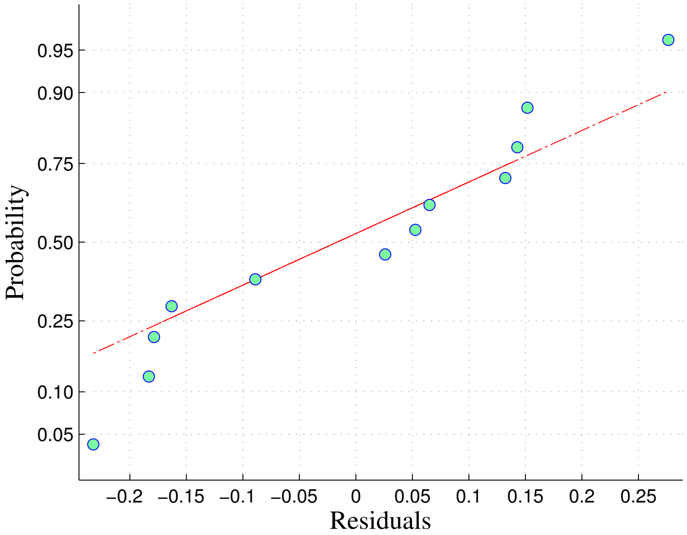


FIGURE 6.3: Normal probability plot of the residuals for the car plant data.

plots do not suggest violations of the assumptions of zero means and same variance of the random errors. Therefore, the linear regression fits the data adequately.

```
% Residual plot (residuals vs. fitted values)
res = y-yhat; % Residuals
scatter(yhat,y-yhat,'ko'); % Display residual plot
h = refline(0); set(h,'linestyle', '-', 'color', 'r'); % Horizontal zero-line
```

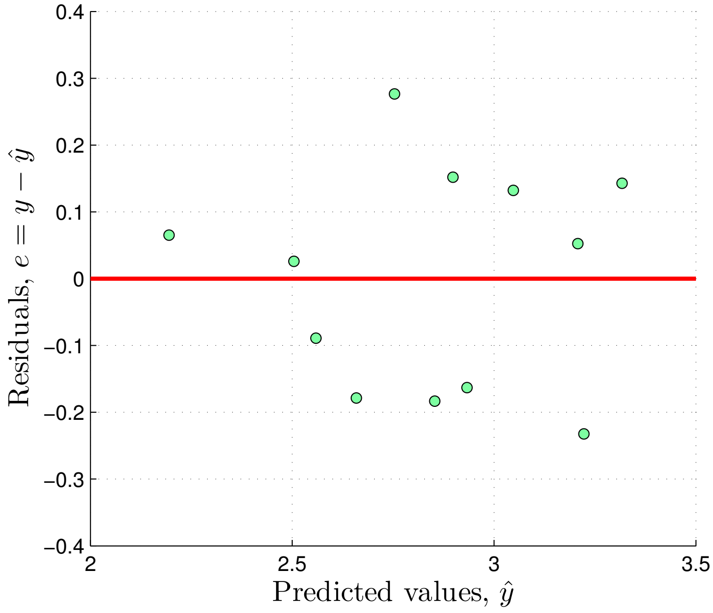


FIGURE 6.4: Residual plot against the predicted values for the car plant data.

```
% Residual plot (residuals vs. fitted values)
```

```
res = y-yhat; % Residuals
scatter(x,y-yhat,'ko'); % Display residual plot
h = refline(0); set(h,'linestyle', '-', 'color', 'r'); % Horizontal zero-line
```

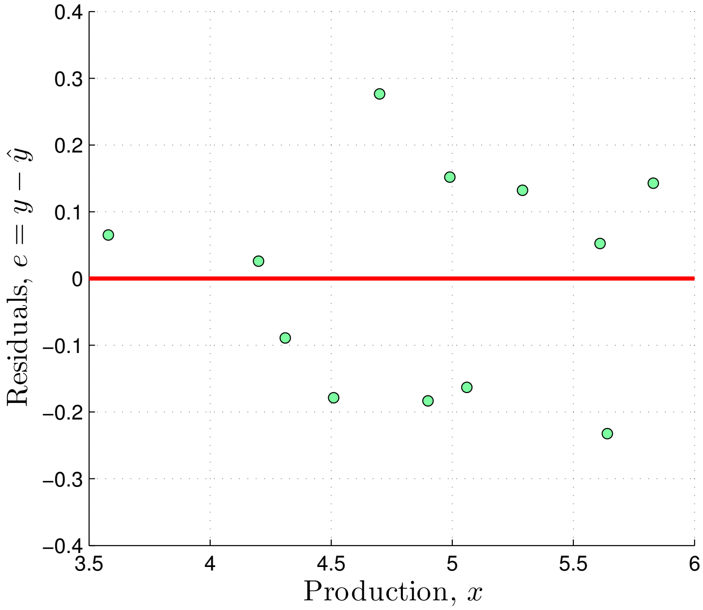


FIGURE 6.5: Residual plot against the independent variable for the car plant data.

Therefore, in order to perform statistical inference on the regression, it must be the case that the residuals are distributed with constant error variance. This requirement is easily verified through normal probability and residual plots.

Example 6.5 *Let $SS_E = \sum_{i=1}^{n} e_i^2$ be the sum of residuals. Show that $SS_E/(n-2)$ is an unbiased estimator of σ^2.*

Solution: Denote by $\mathcal{S}^2 = SS_E/(n-2)$. Since $(n-2)\mathcal{S}^2/\sigma^2 \sim \chi^2(n-2)$, it follows that $E((n-2)\mathcal{S}^2/\sigma^2) = n-2$. Thus, $E(\mathcal{S}^2) = \sigma^2$. That is, $\hat{\sigma}^2 = SS_E/(n-2)$. This estimate of σ^2 has $n-2$ degrees of freedom (2 is subtracted from n because two unknown parameters, β_0 and β_1, are estimated). ■

Standard errors of β_0 and β_1:

$$s.e.(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \tag{6.1.9}$$

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{6.1.10}$$

Example 6.6 *Using the data of Table 6.1,*

1. *Compute the estimate of σ^2*

2. *Compute the standard errors $s.e.(\hat{\beta}_0)$ and $s.e.(\hat{\beta}_1)$*

Solution:

```
% SSe and variance estimate
SSe = SSt - SSr;
sigma2hat = SSe/(n-2);
```

```

% Standard errors of beta0hat and beta1hat
se_beta0hat = sqrt(sigma2hat*(1/n + mean(x)^2/Sxx));
se_beta1hat = sqrt(sigma2hat/Sxx);
```

1. The returned value is $\hat{\sigma}^2 = 0.030$. The estimated value of standard deviation $\hat{\sigma} = 0.1732$ implies that most of the observed electricity usage (y) values will fall within approximately $2\hat{\sigma} = 0.3464$ million kWh of their respective predicted values.

2. The returned standard errors are $s.e.(\hat{\beta}_0) = 0.386$ and $s.e.(\hat{\beta}_1) = 0.078352$. ■

6.1.2 Confidence Intervals for the Intercept β_0 and Slope β_1

It can be shown that the statistics

$$\frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \sim t(n-2) \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)} \sim t(n-2) \tag{6.1.11}$$

follow a t distribution with $n-2$ degrees of freedom, where n is the number of observations.

Confidence Intervals for the Slope β_1:

- A $100(1-\alpha)\%$ confidence interval on β_1 is given by

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\, s.e.(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\, s.e.(\hat{\beta}_1)$$

- A $100(1-\alpha)\%$ upper-confidence interval on β_1 is given by

$$\beta_1 \leq \hat{\beta}_1 + t_{\alpha,n-2}\, s.e.(\hat{\beta}_1)$$

- A $100(1-\alpha)\%$ lower-confidence interval on β_1 is given by

$$\hat{\beta}_1 - t_{\alpha,n-2}\, s.e.(\hat{\beta}_1) \leq \beta_1$$

Similarly, the confidence intervals for the intercept can be obtained by replacing β_1 and $\hat{\beta}_1$ with β_0 and $\hat{\beta}_0$, respectively, in the above box.

Example 6.7 *Using the data of Table 6.1, find a 95% confidence interval for the true slope of the line. Interpret your result.*

Solution:

```
% 95% confidence interval
alpha = 0.05; % Significance level
t_alpha = icdf('t',1-alpha/2,n-2); %t_{alpha/2,n-2}
%Confidence Intervals Limits
lcl = beta1hat-t_alpha*se_beta1hat; % Lower-confidence limit
ucl = beta1hat+t_alpha*se_beta1hat; % upper-confidence limit
```

$100(1-\alpha)\% = 95\%$ implies that $\alpha = 0.05$. Thus, the 95% confidence interval for the slope β_1 is

$$(\hat{\beta}_1 \pm t_{\alpha/2,n-2}\, s.e.(\hat{\beta}_1)) = (0.499 \pm (2.228139)(0.078352)) = (0.324252, 0.673409)$$

Thus, $0.324252 \leq \beta_1 \leq 0.673409$. That is, on average, the monthly electricity usage increases by an amount between 0.32 and 0.67 for every extra monthly production. ■

6.1.3 Hypothesis Testing about the Slope β_1

When the residuals are normally distributed with constant error variance, we can perform inference on the linear regression equation. We will now test the hypothesis of no linear relation between the predictor and the response variable. In testing the population mean μ, there are three ways to structure the hypothesis test:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$	$H_1 : \beta_1 > 0$	$H_1 : \beta_1 < 0$

These hypotheses relate to the *significance of regression*. Failure to reject the null hypothesis is equivalent to concluding that there is no linear relationship between x and y. On the other hand, rejection of the null hypothesis implies that x is of value in explaining the variability in y. In the two-tailed test, we are testing the claim that a linear relation exists between two variables without regard to the sign of the slope. In the upper-tailed test, we are testing that the claim that the slope of the true regression line is positive. In the lower-tailed test, we are testing that the claim that the slope of the true regression line is negative.
Under the assumption that the null hypothesis is true (i.e. $H_0 : \beta_1 = 0$), the test statistic

$$T_0 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} \sim t(n-2) \tag{6.1.12}$$

follows a t distribution with $n-2$ degrees of freedom, where n is the number of observations.

Recall that a critical region or rejection region is the set of all values such that the null hypothesis is rejected. Let t_0 be the numerical value, calculated from the sample data, of the test statistic T_0. Then, for a selected significance level α, the critical regions are

Two-tailed	Upper-Tailed	Lower-Tailed
If $t_0 < -t_{\alpha/2,n-2}$ or $t_0 > t_{\alpha/2,n-2}$	If $t_0 > t_{\alpha,n-2}$	If $t_0 < -t_{\alpha,n-2}$
reject H_0	reject H_0	reject H_0

Denote by $F(\cdot)$ the cdf of the $t(n-2)$ distribution. Then, the t-test about the slope β_1 may be summarized as follows:

Hypotheses:

Two-tailed	Upper-Tailed	Lower-Tailed
$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$	$H_1 : \beta_1 > 0$	$H_1 : \beta_1 < 0$

Test Statistic (t-test): $T_0 = \dfrac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$

Critical Regions:

Two-tailed	Upper-Tailed	Lower-Tailed
$\lvert t_0 \rvert > t_{\alpha/2,n-2}$	$t_0 > t_{\alpha,n-2}$	$t_0 < -t_{\alpha,n-2}$

P-values:

Two-tailed	Upper-Tailed	Lower-Tailed
$2[1 - F(\lvert t_0 \rvert)]$	$1 - F(t_0)$	$F(t_0)$

Example 6.8 *Using the data of Table 6.1,*

1. *Conduct a test to determine if the true slope of the line differs from 0. Use $\alpha = 0.05$*
2. *Conduct a test to determine whether y is positively related to x. Use $\alpha = 0.05$*

Solution: Since the residuals are normally distributed with constant error variance, we can perform statistical inference on the regression.

```
alpha = 0.05; % Significance level
t_alpha = icdf('t',1-alpha/2,n-2); %t_{alpha/2,n-2}
t0 = beta1hat/se_beta1hat; % Value of T0 statistic
pvalue = 2*(1-cdf('t',t0,n-2)); % p-value
```

1. We are testing the claim that there is no linear relation between production and electricity usage.

   Step 1: We test the hypotheses (two-tailed)

   $$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

   Step 2: The value of the t-test is given by

   $$t_0 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = 6.366551$$

   Step 3: The rejection region is $|t_0| > t_{\alpha/2,n-2}$, where $t_{\alpha/2,n-2} = t_{0.025,10} = 2.228139$.

   Step 4: Since $|t_0| = 6.366551 > t_{\alpha/2,n-2} = 2.228139$, we reject the null hypothesis H_0.

   Thus, the null hypothesis $H_0 \ : \ \beta_1 = 0$ should be rejected. The same conclusion can be drawn by computing the p-value $= 2[1 - F(|t_0|)] = 0.000082$, which is much smaller than $\alpha = 0.05$.

2. We are testing the claim that the slope of the true regression is positive.

   Step 1: We test the hypotheses (upper-tailed)

   $$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 > 0 \end{aligned}$$

   Step 2: The value of the t-test is given by

   $$t_0 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = 6.366551$$

   Step 3: The rejection region is $t_0 > t_{\alpha,n-2}$, where $t_{\alpha,n-2} = t_{0.05,10} = 1.8125$.

   Step 4: Since $t_0 = 6.366551 > t_{\alpha,n-2} = 1.8125$, we reject the null hypothesis H_0.

   Thus, the null hypothesis $H_0 \ : \ \beta_1 = 0$ should be rejected. ■

6.1.4 Analysis of Variance Procedure for Testing Significance of Regression

The total variability $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (called *total sum of squares*) in the dependent variable y can be partitioned into the variability explained by the regression line SS_R and the variability SS_E (*error sum of squares*) about the regression line as follows:

$$SS_T = SS_R + SS_E \tag{6.1.13}$$

where $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy} = S_{xy}^2/S_{xx}$ is referred to as the *regression sum of squares*, and it represents the variation in y that is accounted for by regression on x.

Example 6.9 *Show that the expected value of the regression sum of squares is* $E(SS_R) = \sigma^2 + \beta_1^2 S_{xx}$

Solution: Since $E(S_{xy}^2) = var(S_{xy}) + (E(S_{xy}))^2 = \sigma^2 S_{xx} + \beta_1^2 S_{xx}^2$, it follows that $E(SS_R) = E(S_{xy}^2)/S_{xx} = \sigma^2 + \beta_1^2 S_{xx}$.■

Note that SS_T, SS_R and SS_E have $n-1$, 1, and $n-2$ degrees of freedom, respectively. Since $SS_E/\sigma^2 \sim \chi^2(n-2)$ and $SS_R/\sigma^2 \sim \chi^2(1)$, it follow that if the null hypothesis $H_0 \ : \ \beta_1 = 0$ is true then $E(SS_R) = \sigma^2$, and the statistic

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \sim F(1, n-2) \tag{6.1.14}$$

where $MS_R = SS_R/1$ and $MS_E = SS_E/(n-2)$ are referred to as mean square regression and mean square error, respectively.

Let f_0 is the numerical value, calculated from the data, of the test statistic F_0 and $f_{\alpha,1,n-2}$ is the upper-percentage points of the F-distribution $F(1, n-2)$. If $f_0 > f_{\alpha,1,n-2}$ then $H_0 : \beta_1 = 0$ would be rejected.

This analysis of variance (ANOVA) for testing significance of regression is usually represented in tabular form, called **ANOVA table**, as shown in Table 6.3. The ANOVA table for a linear regression gives the sum of squares, degrees of freedom, mean squares for regression and error, and the value of the ANOVA F statistic.

TABLE 6.3: Analysis of Variance for Testing Significance of Regression.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	1	MS_R	$\frac{MS_R}{MS_E}$
Error	SS_E	$n-2$	MS_E	
Total	SS_T	$n-1$		

Example 6.10 *Using the data of Table 6.1, test the hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ using the analysis of variance procedure with $\alpha = 0.05$.*

Solution: The ANOVA for testing the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative hypothesis $H_1 : \beta_1 \neq 0$ is summarized in Table 6.4. We want to test the significance of regression with $\alpha = 0.05$.

Step 1: We test the hypotheses

$$\begin{aligned} H_0 &: \text{There is no significant linear relationship} \\ H_1 &: \text{There is a significant linear relationship} \end{aligned}$$

or equivalently

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_R}{MS_E} = 40.533$$

Step 3: The rejection region is $f_0 > f_{\alpha,1,n-2}$, where $f_{\alpha,1,n-2} = f_{0.05,1,10} = 4.9646$.

Step 4: Since $f_0 = 40.533 > f_{\alpha,1,n-2} = 4.9646$, we reject the null hypothesis H_0.

Thus, we conclude that the monthly electricity usage of the plant is linearly related to its monthly production. ■

TABLE 6.4: ANOVA table for testing significance of regression on car plant data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	1.212	1	1.212	40.533
Error	0.299	10	0.030	
Total	1.511	11		

6.1.5 Correlation and Determination Coefficients

The *correlation coefficient*, r, is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{6.1.15}$$

and it measures the strength of the linear relationship between the variables y and x. The value of r is always between -1 and 1. The sign of r is the same as the sign of $\hat{\beta}_1$. A value of $r \approx 0$ implies little or no linear relationship between y and x, while a value of $r \approx 1$ or $r \approx -1$ indicates a strong positive or negative linear relationship, respectively. A positive correlation coefficient ($r > 0$) indicates a positive slope, while a negative correlation coefficient ($r < 0$) indicates a negative slope for the regression line.

The proportion of the total variability accounted for by the regression line is the *coefficient of determination*, r^2, given by

$$r^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{6.1.16}$$

which takes a value between 0 and 1.

For example, the coefficient of determination for the car plant data is $r^2 = 0.802$, meaning that 80.2% of the sample variation in electricity usage values can be explained by using the production to predict the electricity usage in the straight-line model. The correlation coefficient is $r = 0.8955$ because the sign of $\hat{\beta}_1$ is positive. Thus, the relationship between production and electricity usage is linear with a positive slope.

6.1.6 Regression Analysis: A Complete Example

Example 6.11 *A random sample of eight drivers insured with a company and having similar auto insurance policies was selected. Table 6.5 lists their driving experiences (in years) and monthly auto insurance premiums (in dollars).*

Table 6.5: Auto insurance data.

Driving Experience	Monthly Auto Insurance Premium
5	64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

1. *Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium? Do you expect a positive or negative relationship between these two variables.*

2. *Find the least-squares regression line by choosing appropriate dependent and independent variables based on your answer in part 1.*

3. *Interpret the meaning of the values $\hat{\beta}_0$ and $\hat{\beta}_1$ calculated in part 2.*

4. *Display the scatter plot and the regression line.*

5. *Calculate the correlation and determination coefficients, and explain what they mean.*

6. *Predict the monthly auto insurance premium for a driver with 10 years of driving experience.*

7. *Calculate the standard errors*

8. *Display the normal probability plot. Does the normality assumption seem reasonable?*

9. *Display the residual plot. Does the constant variance assumption seem reasonable?*

10. *Construct a 90% confidence interval for β_1*

11. *Test at the 5% significance level whether β_1 is negative.*

12. *Conduct a t test (concerning β_1) for a significant regression. Use a significance level of 0.05.*

13. *Construct the ANOVA table and conduct an F-test for a significant regression. Use a significance level of 0.05.*

Solution:

1. Intuitively, we expect the insurance premium to depend on driving experience. Thus, the insurance premium is a dependent variable and driving experience is an independent variable in the regression model. A new driver is considered a high risk by insurance companies, and has to pay a higher premium for auto insurance. Therefore, the insurance premium is expected to decrease with an increase in the years of driving experience. Thats is, we expect a negative relationship between these two variables.

2. The least-squares regression line is $\hat{y} = 76.66 - 1.55x$, where $\hat{\beta}_0 = 76.66$ and $\hat{\beta}_1 = -1.55$.

3. The value $\hat{\beta}_0 = 76.66$ gives the value of $\hat{y}$ for $x = 0$; that is, it gives the monthly auto insurance premium for a driver with no driving experience. The value $\hat{\beta}_1 = -1.55$ gives the change in $\hat{y}$ due to a change of one unit in x, indicating that, on average, for every extra year of driving experience, the monthly auto insurance premium decreases by \$1.55.

4. The scatter plot and the regression line are shown in Figure 6.6. As expected, we can see a negative relationship between the dependent and independent variables.

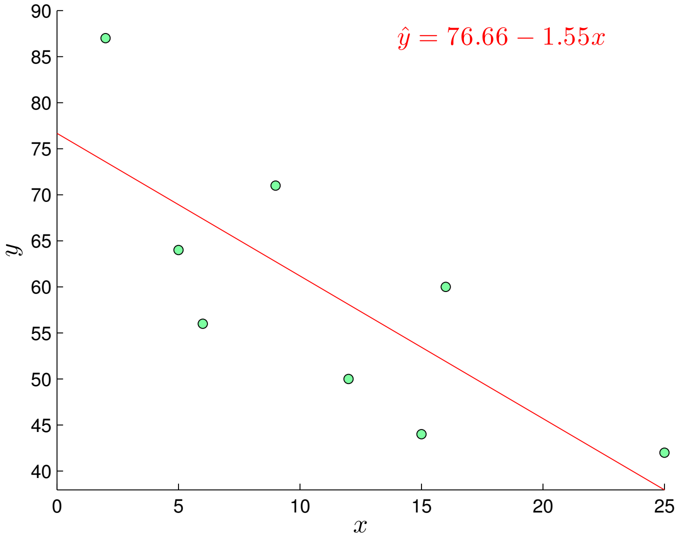


FIGURE 6.6: Scatter plot and regression line for the auto insurance data.

5. The value of $r = -0.77$ indicates that the driving experience and the monthly auto insurance premium are negatively related. The value of $r^2 = 0.59$ states that 59% of the total variation in insurance premiums is explained by years of driving experience and 41% is not.

6. Using the regression line, we can predict the monthly auto insurance premium for a driver with $x = 10$ years of driving experience as

$$\hat{y} = 76.66 - 1.55x = 76.66 - (1.55)(10) = \mathbf{\$61.18}$$

7. The estimated value of standard deviation is $\hat{\sigma} = 10.3199$. Thus, the standard errors are $s.e.(\hat{\beta}_0) = 6.961$ and $s.e.(\hat{\beta}_1) = 0.52698$.

8. A normal probability plot of the auto insurance data is provided in Figure 6.7. The plot is roughly linear and all the data values lie within the bounds of the normal probability plot, indicating that the data are roughly normal.

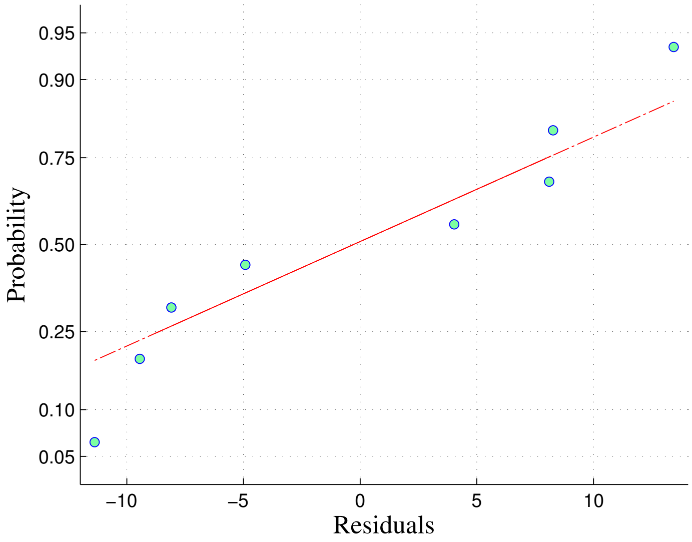


FIGURE 6.7: Normal probability plot of the residuals for the auto insurance data.

9. The residual plot for the auto insurance data, shown in Figure 6.8, indicate that the points seem to be randomly dispersed around zero. That is, a linear regression model is appropriate for the data. Thus, the residual plot does not suggest violations of the assumptions of zero means and same variance of the random errors.

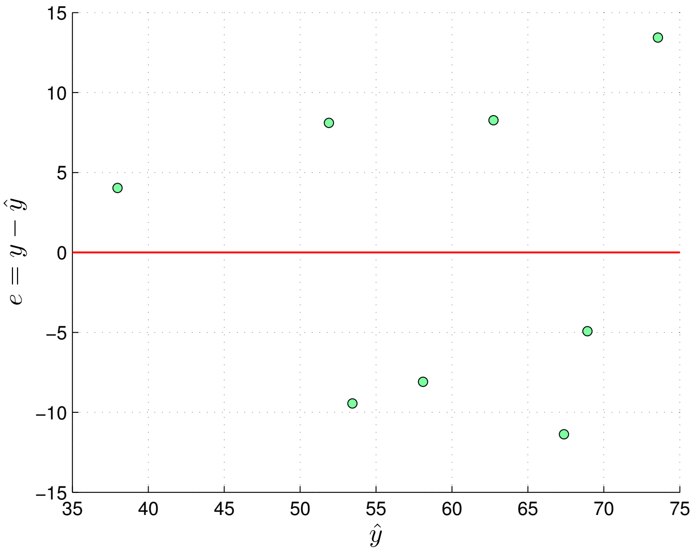


FIGURE 6.8: Residual plot for the auto insurance data.

10. $100(1-\alpha)\% = 90\%$ implies that $\alpha = 0.10$. Thus, the 90% confidence interval for the slope β_1 is

$$(\hat{\beta}_1 \pm t_{\alpha/2,n-2}\, s.e.(\hat{\beta}_1) = (-2.57, -0.52)$$

Thus, $-2.571 \leq \beta_1 \leq -0.523$. That is, on average, the monthly auto insurance premium of a driver decreases by an amount between \$0.52 and \$2.57 for every extra year of driving experience.

11. Since the residuals are normally distributed with constant error variance, we can perform statistical inference on the regression. We are testing whether β_1 is negative at a significance level $\alpha = 0.05$.

    Step 1: We test the hypotheses (lower-tailed)

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 < 0 \end{aligned}$$

    Step 2: The value of the t-test is given by

$$t_0 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = -2.93671$$

    Step 3: The rejection region is $t_0 < -t_{\alpha,n-2}$, where $t_{\alpha,n-2} = t_{0.05,6} = 1.94318$.

    Step 4: Since $t_0 = -2.93671 < -t_{\alpha,n-2} = -1.94318$, we reject the null hypothesis H_0.

    Thus, we reject the null hypothesis $H_0 : \beta_1 = 0$ and conclude that $\beta_1 < 0$. That is, the monthly auto insurance premium decreases with an increase in years of driving experience.

12. The t test (concerning β_1) for a significant regression is two-tailed. We are testing for significance of regression at a significance level $\alpha = 0.05$.

    Step 1: We test the hypotheses (two-tailed)

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

    Step 2: The value of the t-test is given by

$$t_0 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = -2.93671$$

    Step 3: The rejection region is $|t_0| > t_{\alpha/2,n-2}$, where $t_{\alpha/2,n-2} = t_{0.025,6} = 2.446912$.

    Step 4: Since $|t_0| = 2.93671 > t_{\alpha/2,n-2} = 2.446912$, we reject the null hypothesis H_0.

    We conclude that there is evidence to suggest that $\beta_1 \neq 0$, and the regression is therefore significant.

13. The ANOVA for testing significance of linear regression for the auto insurance data is summarized in Table 6.6.

TABLE 6.6: ANOVA table for simple linear regression using auto insurance data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	918.493	1	918.493	8.624
Error	639.007	6	106.501	
Total	1557.500	7		

    We want to test the significance of regression using ANOVA with $\alpha = 0.05$.

    Step 1: We test the hypotheses

$$\begin{aligned} H_0 &: \text{There is no significant linear relationship} \\ H_1 &: \text{There is a significant linear relationship} \end{aligned}$$

    or equivalently

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_R}{MS_E} = 8.624$$

Step 3: The rejection region is $f_0 > f_{\alpha,1,n-2}$, where $f_{\alpha,1,n-2} = f_{0.05,1,6} = 5.9874$.

Step 4: Since $f_0 = 8.624 > f_{\alpha,1,n-2} = 5.9874$, we reject the null hypothesis H_0.

Thus, we conclude that the monthly auto insurance premium is linearly related to driving experience. ■

6.2 Multiple Regression

A multiple regression model is a probabilistic model that includes more than one independent variable. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. This section provides a brief introduction to multiple regression analysis. In multiple regression analysis, we assume that each observed value y_i of the response variable Y_i can be described as follows:

Multiple Linear Regression Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n \tag{6.2.1}$$

where

- y_i is the dependent variable
- $x_1, x_2, \ldots, x_k$ are the independent variables
- $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters
- $\varepsilon_i \sim N(0, \sigma^2)$ is a random error with unknown variance

Thus, the multiple regression model uses two or more independent variables (x_i) to predict the value of a dependent variable (y_i). The observed data for the multiple regression model can be represented by the format shown in Table 6.7, where x_{ij} is the j-th observation or level of variable x_j.

Table 6.7: Data for Multiple Linear Regression.

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

The multiple regression model can be written in matrix form as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{6.2.2}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Denote by X' the transpose of X, and $(X'X)^{-1}$ the inverse of $X'X$. Then, the vector $\boldsymbol{\beta}$ of unknown parameters can be estimated using the least-square estimation:

Least-squares estimates:

Regression parameters: $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$

Regression line: $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$

Note that $X'X$ is a matrix of size $(k+1) \times (k+1)$.

Example 6.12 *A study was performed on the systolic blood pressure (SBP) y and its relationship to x_1 = weight in pounds and x_2 = age in a class of males of approximately the same height. From 13 subjects preselected according to weight and age, the data in Table 6.8 were obtained.*

TABLE 6.8: Systolic blood pressure data.

Systolic blood pressure y	Weight x_1	Age x_2
120	152	50
141	183	20
124	171	20
126	165	30
117	158	30
129	161	50
123	149	60
125	158	50
132	170	40
123	153	55
132	164	40
155	190	40
147	185	20

1. *Compute the least-squares estimate $\hat{\boldsymbol{\beta}}$*
2. *Find the regression equation and plot it on the scatter plot*
3. *Predict the systolic blood pressure of a 37-year old male who weighs 180 pounds.*

Solution: We treat the weight and age as the predictor variables, x_1 and x_2, and systolic blood pressure as the response variable, y. The number of observations is $n = 13$. Thus, the multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, \ldots, 13$$

MATLAB code

```
>> y = [120   141   124   126   117   129   123   125   132   123   132   155   147]';
>> x1 = [152   183   171   165   158   161   149   158   170   153   164   190   185]';
>> x2 = [50    20    20    30    30    50    60    50    40    55    40    40    20]';
>> X = [ones(size(x1)) x1 x2];
>> b = regress(y,X)
b =
  -65.0997
    1.0771
    0.4254
>> beta0hat = b(1); beta1hat = b(2); beta2hat = b(3);
>> x1fit = min(x1):3:max(x1);
>> x2fit = min(x2):3:max(x2);
>> [X1FIT,X2FIT] = meshgrid(x1fit,x2fit);
>> YFIT = beta0hat + beta1hat*X1FIT + beta2hat*X2FIT;
>> surf(X1FIT,X2FIT,YFIT);
>> hold on; scatter3(x1,x2,y,'filled')
```

1. The returned values are $\hat{\beta}_0 = -65.0997$ and $\hat{\beta}_1 = 1.0771$, and $\hat{\beta}_2 = 0.4254$. That is, the least-squares estimate $\hat{\boldsymbol{\beta}}$ is given by

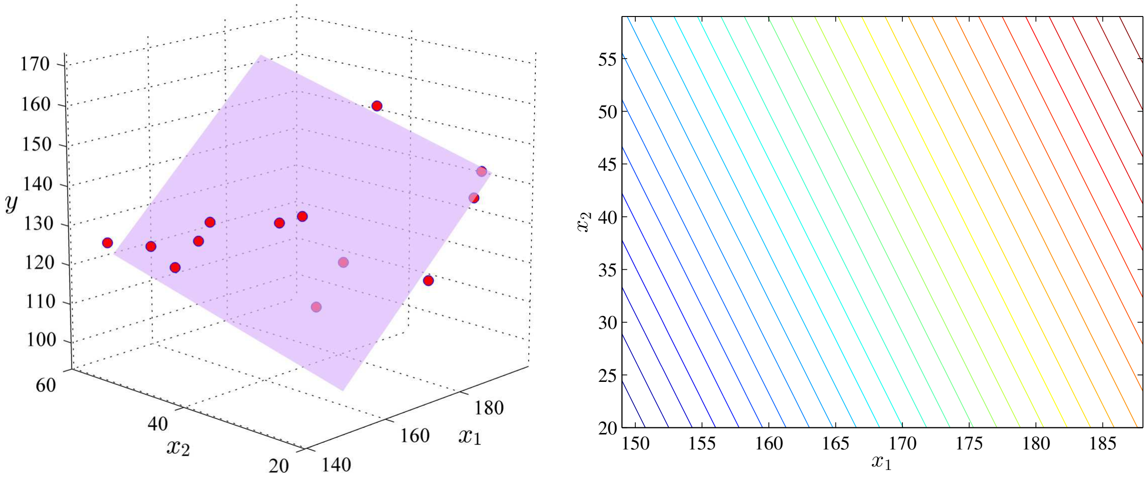
$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -65.0997 \\ 1.0771 \\ 0.4254 \end{pmatrix}$$

   The value of the estimated intercept, $\hat{\beta}_0$, represents the mean value of y when both x_1 and x_2 equal 0. Because the systolic blood pressure cannot be less than 0, the intercept has no practical interpretation. The value of $\hat{\beta}_1$ means that for each increase of 1 unit in x_1, the mean value of y is estimated to increase by 1.0771, holding constant the effect of x_2. In other words, holding constant the age of a person, for each increase of 1 pound in the weight, the fitted model predicts that the expected systolic blood pressure is estimated to increase by 1.0771. Similarly, the value of $\hat{\beta}_2$ means that for each increase of 1 unit in x_2, the mean value of y is estimated to increase by 0.4254, holding constant the effect of x_1. In other words, holding constant the weight of a person, for each increase of 1 year in the age, the fitted model predicts that the expected systolic blood pressure is estimated to increase by 0.4254.

2. The fitted regression equation is given by $\hat{y} = -65.0997 + 1.0771x_1 + 0.4254x_2$, and it describes a plane in the 3D space of y, x_1, and x_2 as shown in Figure 6.9(a). Note that the predicted values lie on the plane. The contour plot of the regression model is also shown in Figure 6.9(b). Notice that the contour lines in this plot are straight lines. This fitted regression equation can be used to predict systolic blood pressure for pairs of values of the regressor variables weight (x_1) and age (x_2).

3. For $x_1 = 180$ and $x_2 = 37$, the fitted value is

$$\begin{aligned} \hat{y} &= -65.0997 + 1.0771x_1 + 0.4254x_2 \\ &= -65.0997 + 1.0771(180) + 0.4254(37) \\ &\approx 145 \end{aligned}$$



(a) (b)

FIGURE 6.9: Systolic blood pressure data: (a) Scatter plot and regression plane; (b) Contour plot.

6.2.1 SCATTERPLOT MATRIX

A scatterplot matrix is a graphical representation for visualizing the multivariable data in Table 6.8 as a matrix of two-dimensional scatter plots. The scatterplot matrix displays scatterplots of all possible pairs of variables or columns of the matrix data. Histograms of each variable or column are shown along the diagonal of the scatterplot matrix. Figure 6.10 shows the scatterplot matrix of the systolic blood pressure data, where each of the three variables taking the role of x-variable and y-variable against all the others. For example, the top panel (row) displays from left to right the individuals scatterplots of y against x_1 and y against x_2, respectively.

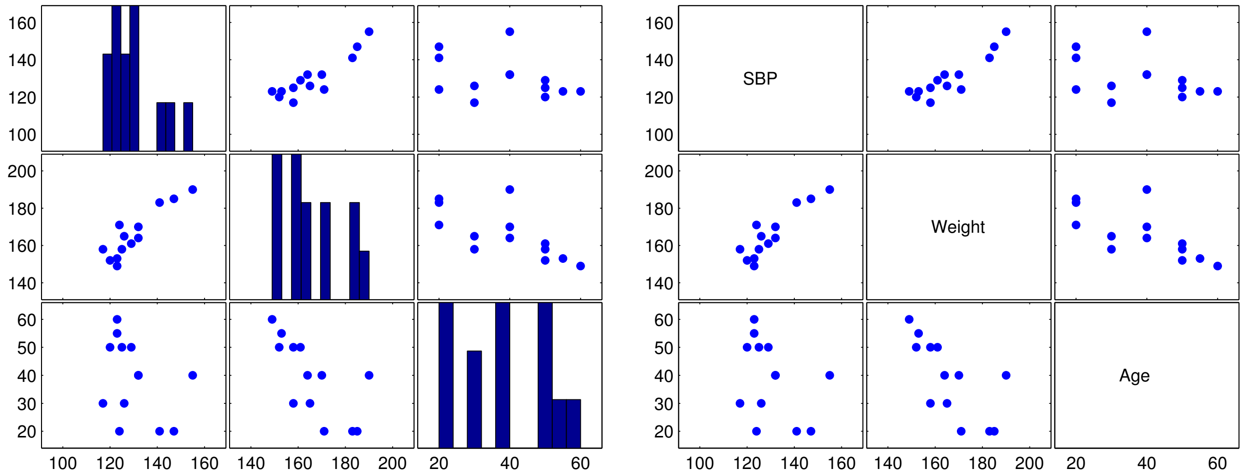


FIGURE 6.10: Scatterplot matrix of the systolic blood pressure data.

MATLAB code

```
>> y = [120   141   124   126   117   129   123   125   132   123   132   155   147]';
>> x1 = [152   183   171   165   158   161   149   158   170   153   164   190   185]';
>> x2 = [50    20    20    30    30    50    60    50    40    55    40    40    20]';
>> data = [y x1 x2];
>> plotmatrix(data);
```

6.2.2 Residual Analysis

In order to check the regression model assumptions, we need to compute and examine the normality of the residuals $e_i = y_i - \hat{y}_i$, where y_i is the observed value and $\hat{y}_i$ is the predicted (fitted) value. Thus, the $n \times 1$ vector of residuals is given by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}} \tag{6.2.3}$$

The normal probability and residual plots for the the systolic blood pressure data are shown in Figure 6.11, which indicate that the residuals are normally distributed with constant error variance. Thus, the regression model assumptions are valid. Notice the relatively linear appearance of the normal probability plot and the relative scatter of the residuals versus fitted values.

6.2.3 Hypothesis Tests and Analysis of Variance in Multiple Linear Regression

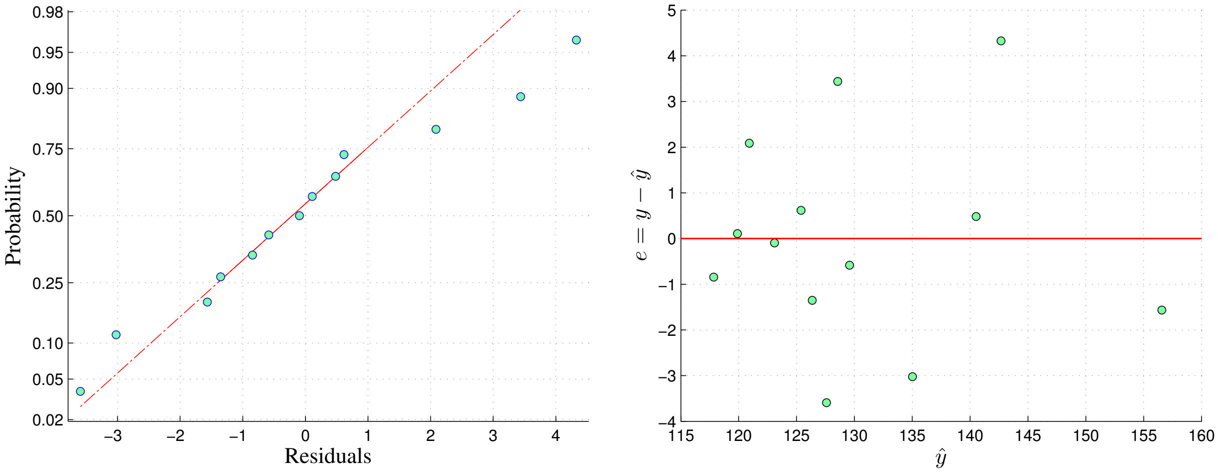
The test for significance of regression is a test to determine whether a linear relationship exists between the response variable y and a subset of the regressor variables $x_1, x_2, \ldots, x_k$. The appropriate hypotheses are

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_1 : \text{At least one } \beta_i \neq 0 \end{aligned} \tag{6.2.4}$$

Rejection of $H_0 : \beta_1 = \beta_1 = \cdots = \beta_k = 0$ implies that at least one of the regressor variables $x_1, x_2, \ldots, x_k$ contributes significantly to the model.

The test for significance of regression is a generalization of the procedure used in simple linear regression. The total variability (i.e. total sum of squares) $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$ in the dependent variable y can be partitioned into the variability explained by the regression line SS_R (sum of squares due to regression) and the variability SS_E (sum of squares due to error) about the regression line as follows:

$$SS_T = SS_R + SS_E \tag{6.2.5}$$



(a) (b)

FIGURE 6.11: Systolic blood pressure data: (a) Normal probability plot; (b) Residual plot.

where

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \boldsymbol{y}'\boldsymbol{y} - n\bar{\boldsymbol{y}}^2, \quad SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}'X'\boldsymbol{y} - n\bar{\boldsymbol{y}}^2, \quad SS_E = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e}$$

If the null hypothesis $H_0 : \beta_1 = \beta_1 = \cdots = \beta_k = 0$ is true, then $SS_R/\sigma^2 \sim \chi^2(k)$ and $SS_E/\sigma^2 \sim \chi^2(n-k-1)$. Moreover, the test statistic is given by

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E} \sim F(k, n-k-1) \tag{6.2.6}$$

where $MS_R = SS_R/k$ and $MS_E = SS_E/(n-k-1)$ are the mean square regression and mean square error, respectively.

Let f_0 is the numerical value, calculated from the data, of the test statistic F_0 and $f_{\alpha,k,n-k-1}$ is the upper-percentage points of the F-distribution $F(k, n-k-1)$. If $f_0 > f_{\alpha,k,n-k-1}$ then $H_0 : \beta_1 = \beta_1 = \cdots = \beta_k = 0$ would be rejected.

It can be shown that an unbiased estimator of the variance σ^2 of the error term in the multiple linear regression model is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n-k-1} = MS_E \tag{6.2.7}$$

Notice that the estimate of σ^2 has $n-k-1 = n-(k+1)$ degrees of freedom ($k+1$ is subtracted from n because $k+1$ unknown parameters, $\beta_0, \beta_1, \ldots, \beta_k$, are estimated).

The analysis of variance (ANOVA) for testing significance of multiple regression is usually represented in tabular form (ANOVA table), as shown in Table 6.9.

TABLE 6.9: Analysis of Variance for Testing Significance of Regression in Multiple Linear Regression.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	$\dfrac{MS_R}{MS_E}$
Error	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Example 6.13 *Using the data of Table 6.8, test for significance of regression using the analysis of variance procedure with the level of significance $\alpha = 0.05$.*

Solution: The ANOVA for testing significance of multiple regression for the systolic blood pressure data is summarized in Table 6.10. We want to test the significance of regression with $\alpha = 0.05$.

TABLE 6.10: ANOVA table for the systolic blood pressure data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	1423.838	2	711.919	113.126
Error	62.931	10	6.293	
Total	1486.769	12		

Step 1: We test the hypotheses

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \\ H_1 &: \text{At least one } \beta_i \neq 0 \end{aligned}$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_R}{MS_E} = 113.126$$

Step 3: The rejection region is $f_0 > f_{\alpha,k,n-k-1}$, where $f_{\alpha,k,n-k-1} = f_{0.05,2,10} = 4.1028$.

Step 4: Since $f_0 = 113.126 > f_{\alpha,k,n-k-1} = 4.1028$, we reject the null hypothesis H_0.

Thus, we conclude that systolic blood pressure is linearly related to either weight or age, or both. ■

Table 6.10 shows that the estimate of σ^2 for the systolic blood pressure regression model is $\hat{\sigma}^2 = MS_E = 6.293$. Also, the coefficient of determination R^2 for the systolic blood pressure data is given by

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = 0.958$$

Thus, 95.8% of the total variation has been explained by the regression model.

It is worth pointing out that the F-statistic is related to the coefficient of determination R^2 by the formula:

$$F_0 = \frac{R^2/k}{(1-R^2)/(n-k-1)} \tag{6.2.8}$$

Therefore, R^2 is large implies F_0 is large, and vice versa.

6.3 Analysis of Variance

In Chapter 4 we used sampling theory to test the significance of differences $\mu_1 - \mu_2$ between two sampling means μ_1 and μ_2. In many situations there is a need to test the significance of differences between more than two sampling means or, equivalently, to test the null hypothesis that more that two sample means are equal. Analysis of Variance (ANOVA) is a statistical method that allows us to analyze and interpret observations from several populations by partitioning the total variation in a data set according to the sources of variation that are present.

One-way analysis of variance (one-way ANOVA) is a technique used to compare means of two or more groups of data using the F distribution. This technique can be used only for numerical data. The one-way ANOVA tests the null hypothesis that samples in two or more groups are drawn from the same population, and it is conducted on an experimental design in which there is only one factor (independent variable). The purpose of one-way ANOVA is to find out whether data from several groups have a common mean. That is, to determine whether the groups are actually different in the measured characteristic.

6.3.1 One-Way ANOVA: Completely Randomized Design

In this section we consider a one-factor completely randomized design (experiment), where observations or measurements are obtained for a independent random samples (called treatments or groups) of n measurements. By "completely randomized", we mean that the participants have been randomly assigned to one of the unique levels of the factor. The observed data for the response measurements can be represented by the format shown in Table 6.11, where y_{ij} is the j-th observation on treatment i. The summary statistics

$$\begin{aligned}
y_{i.} &= \sum_{j=1}^{n} y_{ij} \quad i = 1,\ldots,a \quad \text{(Total of the observations under the } i\text{-th treatment)}\\
\bar{y}_{i.} &= y_{i.}/n \quad \text{(Mean of the observations under the } i\text{-th treatment.)}\\
y_{..} &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \quad \text{(Grand total of all observations)}\\
\bar{y}_{..} &= y_{..}/an \quad \text{(Grand mean of all observations)}
\end{aligned}$$

also appear in the last two columns of the table.

Table 6.11: Typical Data for a Single-Factor Experiment.

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$
					$y_{..}$	$\bar{y}_{..}$

6.3.2 Model for One-Way ANOVA

To implement a formal statistical test for no difference among treatment effects, we need to have a population model for the experiment. To this end, we assume that the response measurements with the ith treatment constitute a random sample from a normal population $N(\mu_i, \sigma^2)$ with a mean μ_i and a common variance σ^2. The samples are assumed to be mutually independent. Thus, comparing a treatments can be described by the statistical model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases} \tag{6.3.1}$$

or, equivalently,

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases} \tag{6.3.2}$$

where $\mu_i = \mu + \tau_i$ is the mean of the i-th treatment, μ is the overall mean common to all treatments, and τ_i is the i-th treatment effect such that $\sum_{i=1}^{a} \tau_i = 0$. The errors $\varepsilon_{ij} \sim N(0, \sigma^2)$ are i.i.d., and Y_{ij} is a random variable denoting the (i,j)-th observation y_{ij}. It can be shown that the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_{i.}$, which implies that the estimated or fitted value of y_{ij} is $\hat{y}_{ij} = \bar{y}_{i.}$. Thus, the estimate of the i-th treatment effect is given by

$$\hat{\tau}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i.} - \bar{y}_{..} \qquad \text{for } i = 1,2,\ldots,a \tag{6.3.3}$$

In other words, $\hat{\tau}_i$ is the difference between the mean of the observations under the i-th treatment and the grand mean of all observations.

6.3.3 Residual Analysis

In order to check the regression model assumptions, we need to compute and examine the residuals

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.} \tag{6.3.4}$$

where y_{ij} is the observed value and $\bar{y}_{i.}$ is the predicted (fitted) value. The normal probability and residual plots are used to examine the assumptions underlying the ANOVA, and will also show whether there are any serious outliers. The normality assumption of the errors ε_{ij} can be checked by constructing a normal probability plot of the residuals, while the assumption of equal variances of the errors ε_{ij} at each factor level can be checked by plotting the residuals e_{ij} against the factor levels and compare the spread in the residuals. It is also useful to plot the residuals e_{ij} against the fitted values $\hat{y}_{ij} = \bar{y}_{i.}$.

- *Normal probability plot*: if the observations come from a normal distribution, then the points in the plot fall along a straight line. If the scatter plot is nonlinear, then there is evidence to suggest that the data did not come from a normal distribution.

- *Residual plot*: if the fitted model is appropriate for the data then there will be no apparent pattern in the plot; that is there will only be random variation around 0. However, if the fitted model is not appropriate, there will be a clear relationship between the factor levels or the fitted values and the residuals.

6.3.4 Hypotheses in One-Way ANOVA

The one-way ANOVA evaluates the hypothesis that the samples all have the same mean against the alternative that the means are not all the same. It is assumed that the a populations are independent and normally distributed with means $\mu_1, \mu_2, \ldots, \mu_a$ and common variance σ^2. The null hypothesis that no difference exists among the a means μ_i and the alternative hypothesis is that not all the μ_i's are equal can be formulated as a one-sided hypothesis test:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \ldots = \mu_a = \mu \\ &H_1 : \text{At least one } \mu_i \neq \mu \end{aligned} \tag{6.3.5}$$

or, equivalently,

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0 \\ &H_1 : \text{At least one } \tau_i \neq 0 \end{aligned} \tag{6.3.6}$$

If H_0 is true, then the observations will all have the same normal distribution $N(\mu, \sigma^2)$, implying that there is no significant difference between the treatments.

6.3.5 Total Variability Decomposition

The total variability in the data is described by the total sum of squares

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{6.3.7}$$

and can be partitioned, using ANOVA, into the within-treatments variability SS_E and the between-treatments variability $SS_{\text{Treatments}}$, as follows:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2}_{\text{Total } (SS_T)} = \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2}_{\text{Within } (SS_E)} + \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}_{\text{Between } (SS_{\text{Treatments}})} \tag{6.3.8}$$

Note that SS_T has $an - 1$ degrees of freedom and $SS_{\text{Treatments}}$ has $a - 1$ degrees of freedom. Thus, the error sum of squares SS_E has $a(n-1)$. That is,

$$\underbrace{SS_T}_{an-1 \text{ degrees of freedom}} = \underbrace{SS_E}_{a(n-1) \text{ degrees of freedom}} + \underbrace{SS_{\text{Treatments}}}_{a-1 \text{ degrees of freedom}} \tag{6.3.9}$$

The sums of squares SS_T, $SS_{\text{Treatments}}$, and SS_E can be computed efficiently using the following simplified formulas:

$$
\begin{aligned}
SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{an} \\
SS_{\text{Treatments}} &= \sum_{i=1}^{a} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{an} \\
SS_E &= SS_T - SS_{\text{Treatments}}
\end{aligned}
\tag{6.3.10}
$$

Note that in practice, the sums of squares SS_T and $SS_{\text{Treatments}}$ should be calculated or computed before the error sum of squares SS_E.

6.3.6 Expected Values of the Variabilities and Testing Equality of Means

It can be shown that the expected values of $SS_{\text{Treatments}}$ and SS_E are given by:

- $E(SS_{\text{Treatments}}) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$
- $E(SS_E) = a(n-1)\sigma^2$

Moreover,

- The error mean square $MS_E = \dfrac{SS_E}{a(n-1)}$ is an unbiased estimator of σ^2
- If H_0 is true, then the mean square for treatments $MS_{\text{Treatments}} = SS_{\text{Treatments}}/(a-1)$ is an unbiased estimator of σ^2
- If H_0 is true, then $F_0 = \dfrac{SS_{\text{Treatments}}/(a-1)}{SS_E/a(n-1)} = \dfrac{MS_{\text{Treatments}}}{MS_E} \sim F(a-1, a(n-1))$

It is customary to present the decomposition of the sum of squares and the degrees of freedom in a tabular form (ANOVA table), as shown in Table 6.12. It is important to point out that at an α-significance level, we would reject H_0 if $f_0 > f_{\alpha,a-1,a(n-1)}$, where f_0 is the numerical value, calculated from the data, of the test statistic F_0 and $f_{\alpha,a-1,a(n-1)}$ is the upper-percentage points of the distribution $F(a-1, a(n-1))$. The computed value of f_0 is usually presented in the last column of the ANOVA table. We can also reject H_0 if p-value $= 1 - F(f_0) < \alpha$, where $F(\cdot)$ is the cdf of the F-distribution $F(a-1, a(n-1))$.

Table 6.12: ANOVA table for one-factor experiment.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

Example 6.14 *A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. They decide to make up six test specimens at each concentration level, using a pilot plant. All 24 specimens are tested on a laboratory tensile tester, in random order. The data from this experiment are shown in Table 6.13.*

(i) Calculate the totals and averages for each treatment (i.e. each hardwood concentration level).

(ii) Estimate the treatment effects for the four hardwood concentration levels.

(iii) Construct the ANOVA table.

(iv) Is there sufficient evidence to indicate a difference in hardwood concentrations affect the mean tensile strength of the paper? Test using $\alpha = 0.01$.

(v) Analyze the residuals and comment on model adequacy.

TABLE 6.13: Tensile Strength of Paper (psi).

Hardwood	Observations					
Concentration	1	2	3	4	5	6
5%	7	8	15	11	9	10
10%	12	17	13	18	19	15
15%	14	18	19	17	16	18
20%	19	25	22	23	18	20

Solution: The experiment involves a single factor, hardwood concentration, at four levels. Thus, we have a completely randomized design with $a = 4$ treatments.

(i) The totals and averages for each treatment (i.e. hardwood concentration) are shown in Table 6.14, which presents the data for the tensile strength of paper (psi) with totals and averages. From the four sample means, we observe that the mean for the 20% hardwood concentration level is considerably larger than the means for the other hardwood concentration levels. This can also be observed in the side-by-side boxplots in Figure 6.12. Thus, the team of engineers need to determine whether the four population means differ.

TABLE 6.14: Tensile Strength of Paper (psi) with totals and averages.

Hardwood	Observations							
Concentration	1	2	3	4	5	6	Totals ($y_{i.}$)	Averages ($\bar{y}_{i.}$)
5%	7	8	15	11	9	10	60	10.00
10%	12	17	13	18	19	15	94	15.67
15%	14	18	19	17	16	18	102	17.00
20%	19	25	22	23	18	20	127	21.17
							$y_{..} = 383$	$\bar{y}_{..} = 15.96$

(ii) The estimate of the i-th treatment effect is $\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$ for $i = 1, 2, \ldots, a$. Thus,

$$\hat{\tau}_1 = 10 - 15.96 = -5.96;\ \hat{\tau}_2 = 15.67 - 15.96 = -0.29;\ \hat{\tau}_3 = 17 - 15.96 = 1.04;\ \hat{\tau}_4 = 21.17 - 15.96 = 5.21.$$

(iii) The output results of ANOVA for the tensile strength data are exhibited in Table 6.15.

```
data = [7   8 15 11   9 10
        12 17 13 18 19 15
        14 18 19 17 16 18
        19 25 22 23 18 20];
[p,table,stats] = anova1(data'); % Performs a one–way ANOVA
```

(iv) Let μ_1, μ_2, μ_3, and μ_4 represent the mean tensile strengths for hardwood concentration levels 5%, 10%, 15%, and 20%, respectively. We want to test the equality of the 4 treatment means $\mu_1, \mu_2, \mu_3, \mu_4$ with $\alpha = 0.01$.

Step 1: We test the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu \\ H_1 &: \text{At least one } \mu_i \neq \mu \end{aligned}$$

where μ is the overall mean common to all treatments.

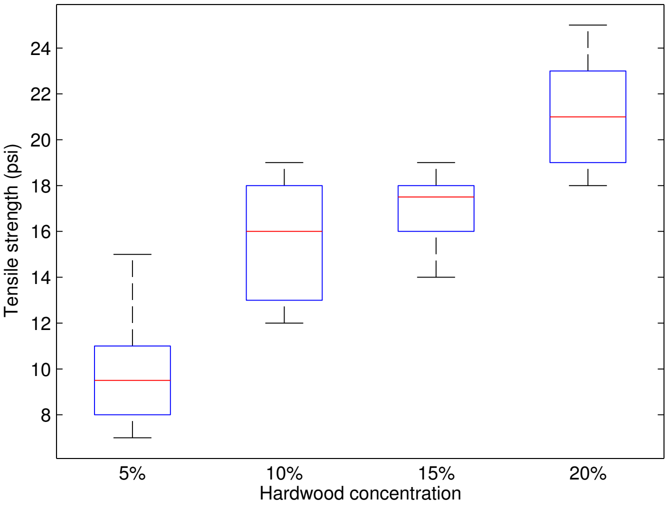
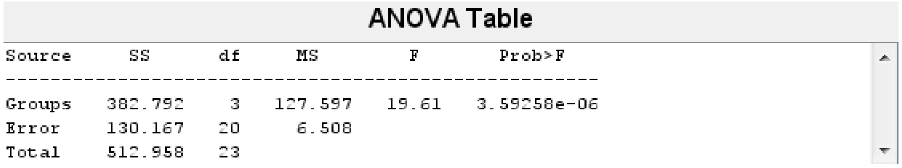


FIGURE 6.12: Boxplots for the tensile strength data.

TABLE 6.15: ANOVA table for the tensile strength data.



ANOVA Table

```
Source       SS       df      MS         F        Prob>F
---------------------------------------------------------
Groups    382.792     3    127.597    19.61    3.59258e-06
Error     130.167    20      6.508
Total     512.958    23
```

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_{\text{Treatments}}}{MS_E} = 19.61$$

Step 3: The rejection region is $f_0 > f_{\alpha,a-1,a(n-1)}$, where $f_{\alpha,a-1,a(n-1)} = f_{0.01,3,20} = 4.9382$.

Step 4: Since $f_0 = 19.61 > f_{\alpha,a-1,a(n-1)} = 4.9382$, we reject the null hypothesis H_0.

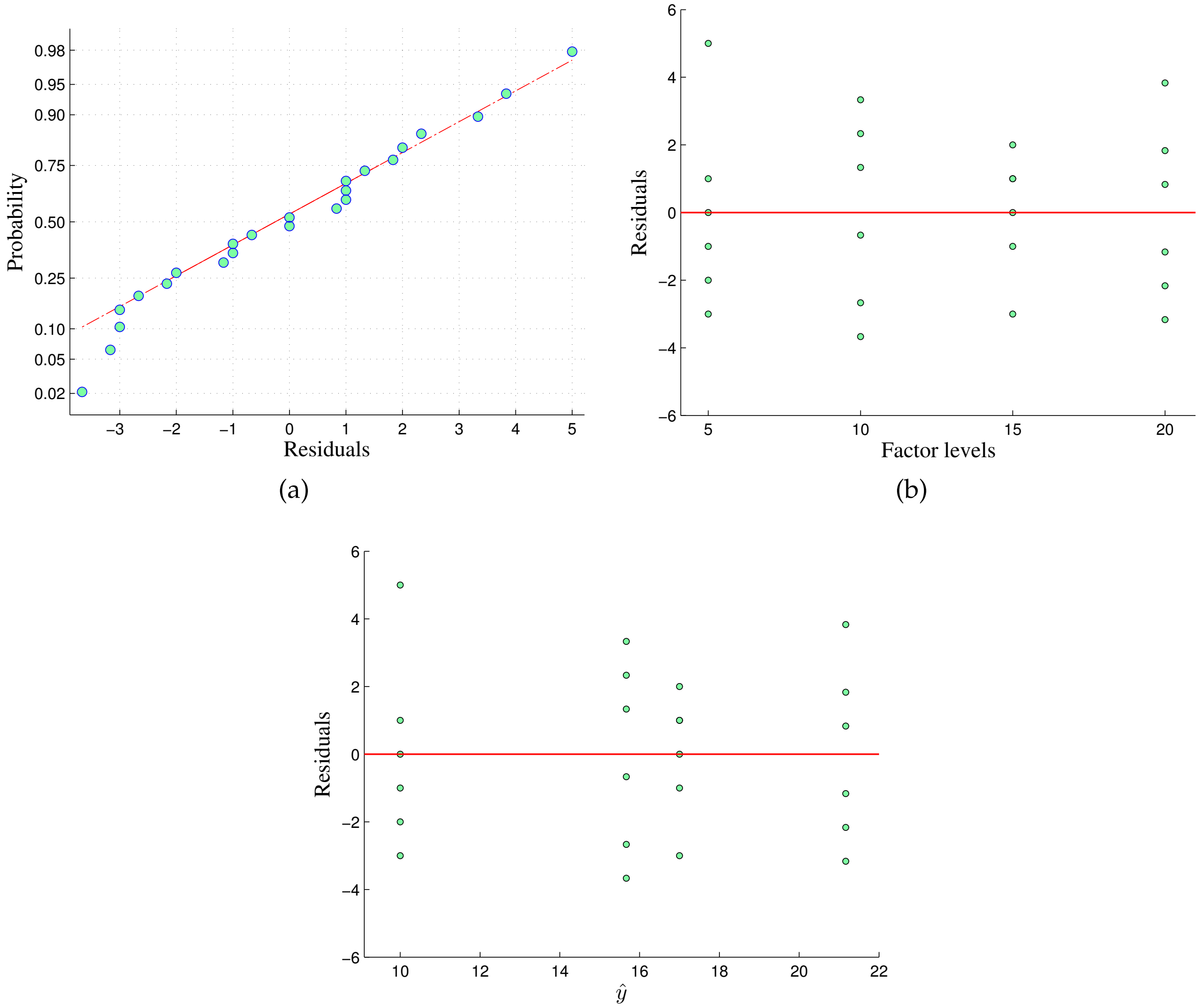
Thus, we reject the null hypothesis of equality of the mean tensile strengths for the four hardwood concentration levels of treatment, and we conclude that the hardwood concentration in the pulp significantly affects the mean strength of the paper. We can arrive at the same conclusion by noting that p-value $= 1 - F(f_0) = 1 - F(19.61) = 3.59258 \times 10^{-6} < \alpha = 0.01$, where $F(\cdot)$ is the cdf of the F-distribution $F(3,20)$.

(v) The calculated residual values for the tensile strength experiment are shown in Table 6.16, and the residual plots are depicted in Figure 6.13. The data of normal probability plot appear fairly linear, suggesting that no reason to doubt the normality assumption. The residual plots against the factor levels and fitted values exhibit random scatter around 0.

```
>> data = [7  8 15 11  9 10
          12 17 13 18 19 15
          14 18 19 17 16 18
          19 25 22 23 18 20];
>> res = bsxfun(@minus,data,mean(data,2)); %residuals
>> normplot(res(:));
>> figure;
>> levels = [5 10 15 20]';
>> x1 = repmat(levels,n,1); y = res(:);
>> scatter(x1,y,3,'ko','MarkerFaceColor',[.49 1 .63]);
>> hold on; h = refline(0); set(h,'linestyle', '-', 'color', 'r','LineWidth',1);
>> hold off; grid off;
>> figure;
>> yibar = mean(data,2); x2 = repmat(yibar,n,1);
>> scatter(x2,y,3,'ko','MarkerFaceColor',[.49 1 .63]);
>> hold on; h = refline(0); set(h,'linestyle', '--', 'color', 'r','LineWidth',1);
>> hold off; grid off;
```

TABLE 6.16: Residual values for the tensile strength experiment.

Hardwood Concentration	Residuals					
5%	-3.00	-2.00	5.00	1.00	-1.00	0.00
10%	-3.67	1.33	-2.67	2.33	3.33	-0.67
15%	-3.00	1.00	2.00	0.00	-1.00	1.00
20%	-2.17	3.83	0.83	1.83	-3.17	-1.17



(a)

(b)

(c)

FIGURE 6.13: Residual analysis of the tensile strength experiment.

Example 6.15 *An experiment was conducted to compare the wearing qualities of three types of paints when subjected to the abrasive action of a slowly rotating cloth-surfaced wheel. Ten paint specimens were tested for each paint type, and the number of hours until visible abrasion was apparent was recorder for each specimen. The data are shown in Table 6.17.*

(i) Is there sufficient evidence to indicate a difference in mean time until abrasion is visibly evident for the three paint types? Test using $\alpha = 0.05$.

(ii) *Analyze the residuals and comment on model adequacy.*

TABLE 6.17: Data for the paint experiment.

Paint	Observations									
Type	1	2	3	4	5	6	7	8	9	10
1	148	76	393	520	236	134	55	166	415	153
2	513	264	433	94	535	327	214	135	280	304
3	335	643	216	536	128	723	258	380	594	465

Solution: The experiment involves a single factor, paint type, at three levels. Thus, we have a completely randomized design with $a = 3$ treatments. Table 6.18 displays the data for the paint experiment with totals and averages.

TABLE 6.18: Data for the paint experiment with totals and averages.

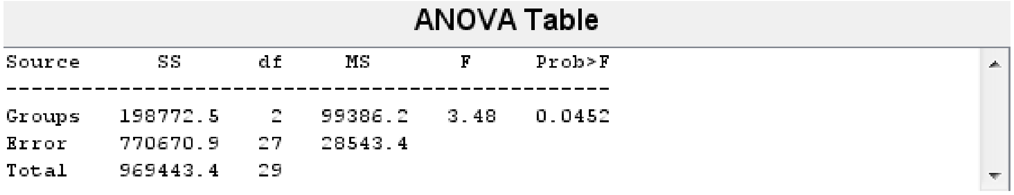
Paint	Observations											
Type	1	2	3	4	5	6	7	8	9	10	Totals ($y_{i.}$)	Averages ($\bar{y}_{i.}$)
1	148	76	393	520	236	134	55	166	415	153	2296	229.60
2	513	264	433	94	535	327	214	135	280	304	3099	309.90
3	335	643	216	536	128	723	258	380	594	465	4278	427.80
											$y_{..} = 9673$	$\bar{y}_{..} = 322.43$

The output results of ANOVA for the paint data are presented in Table 6.19.

MATLAB code

```
data = [148  76 393 520 236 134  55 166 415 153
        513 264 433  94 535 327 214 135 280 304
        335 643 216 536 128 723 258 380 594 465];
[p,table,stats] = anova1(data'); % Performs a one-way ANOVA
```

TABLE 6.19: ANOVA for the paint type experiment.



ANOVA Table

```
Source      SS       df      MS        F      Prob>F
-------------------------------------------------------
Groups    198772.5    2    99386.2   3.48    0.0452
Error     770670.9   27    28543.4
Total     969443.4   29
```

(i) Let μ_1, μ_2, and μ_3 represent the mean abrasion times for paint types 1, 2, and 3, respectively. We want to test the equality of the 3 treatment means μ_1, μ_2, μ_3 with $\alpha = 0.05$.

Step 1: We test the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu \\ H_1 &: \text{At least one } \mu_i \neq \mu \end{aligned}$$

where μ is the overall mean common to all treatments.

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_{\text{Treatments}}}{MS_E} = 3.48$$

Step 3: The rejection region is $f_0 > f_{\alpha,a-1,a(n-1)}$, where $f_{\alpha,a-1,a(n-1)} = f_{0.05,2,27} = 3.35413$.

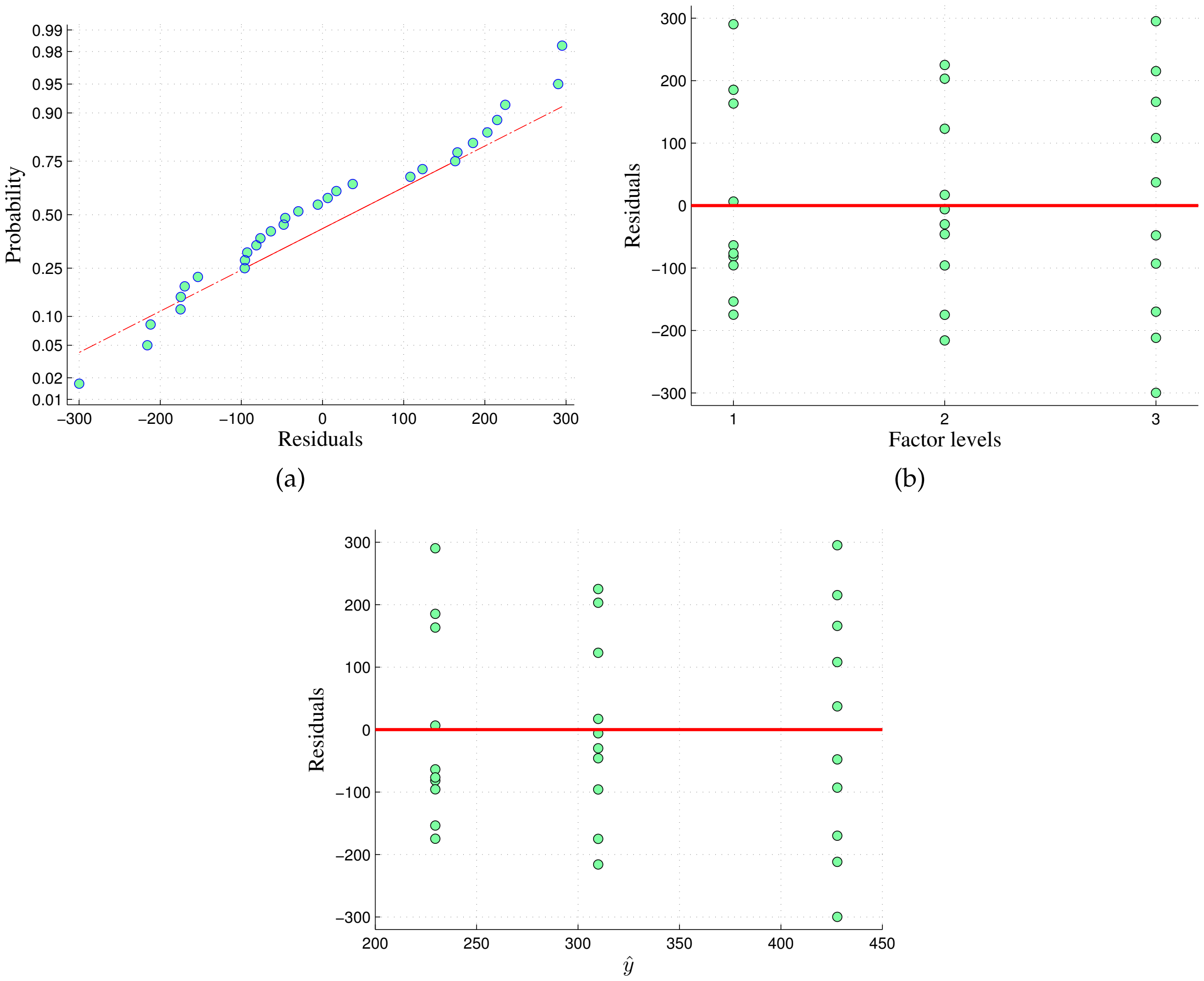
Step 4: Since $f_0 = 3.48 > f_{\alpha,a-1,a(n-1)} = 3.35413$, we reject the null hypothesis H_0.

Thus, we conclude that the mean time to visible abrasion differs for at least two of the three paint types. We can arrive at the same conclusion by noting that p-value $= 1 - F(3.48) = 0.0452 < \alpha = 0.05$.

(ii) The calculated residual values for the paint type experiment are shown in Table 6.20. Figure 6.14 displays the residual plots for the paint type experiment. These plots do not reveal any model inadequacy or unusual problem with the assumptions.

TABLE 6.20: Residual values for the paint experiment.

Paint Type	Residuals									
1	-81.60	-153.60	163.40	290.40	6.40	-95.60	-174.60	-63.60	185.40	-76.60
2	203.10	-45.90	123.10	-215.90	225.10	17.10	-95.90	-174.90	-29.90	-5.90
3	-92.80	215.20	-211.80	108.20	-299.80	295.20	-169.80	-47.80	166.20	37.20



(a)

(b)

(c)

FIGURE 6.14: Residual analysis of the paint type experiment.

Example 6.16 *Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6* *samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data from this experiment are presented in Table 6.21.*

(i) Is there sufficient evidence to indicate a difference in concrete aggregate affect the mean absorption of moisture in concrete? Draw comparative boxplots and perform an analysis of variance. Test using $\alpha = 0.05$.

(ii) Analyze the residuals and comment on model adequacy.

TABLE 6.21: Absorption of Moisture in Concrete Aggregates.

Aggregate	Absorption of Moisture					
1	551	457	450	731	499	632
2	595	580	508	583	633	517
3	639	615	511	573	648	677
4	417	449	517	438	415	555
5	563	631	522	613	656	679

Solution: The experiment involves a single factor, aggregate, at five levels. Thus, we have a completely randomized design with $a = 5$ treatments. A box plot for each aggregate is shown in Figure 6.15. From these side-by-side box plots it is evident that the absorption is not the same for all aggregates. In fact, it appears as if aggregate 4 stands out from the rest. The output results of ANOVA for the absorption of moisture data are presented in Table 6.22.

MATLAB code

```
>> data = [551   457   450   731   499   632
           595   580   508   583   633   517
           639   615   511   573   648   677
           417   449   517   438   415   555
           563   631   522   613   656   679];
>> [a,n] = size(data);
>> boxplot(data');
>> [p,table,stats] = anova1(data');
```

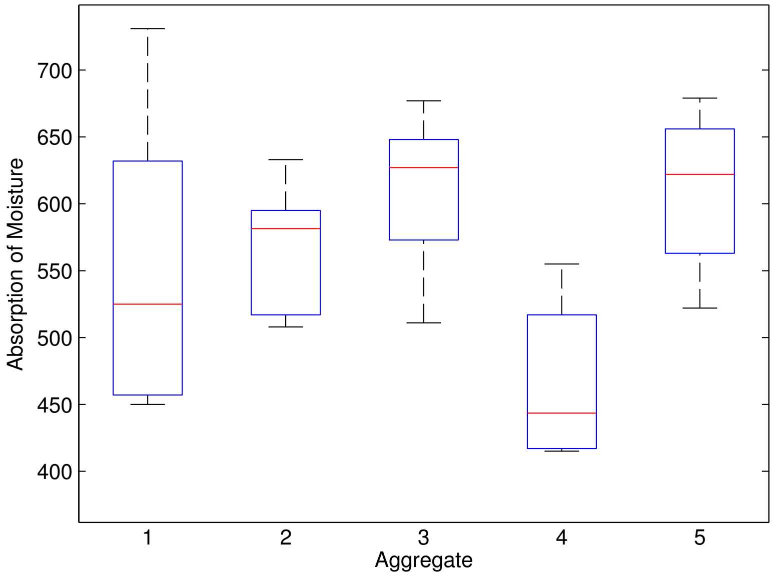
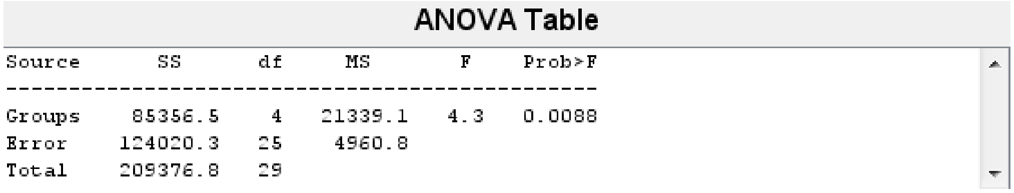


FIGURE 6.15: Box plots for the absorption of moisture in concrete aggregates.

(i) Let μ_1, μ_2, μ_3, and μ_4 represent the mean absorption of moisture in concrete for aggregate 1, 2, 3, 4, 5, respectively.

TABLE 6.22: ANOVA for the absorption of moisture data.



ANOVA Table

```
Source        SS       df      MS         F     Prob>F
----------------------------------------------
Groups     85356.5      4    21339.1    4.3    0.0088
Error     124020.3     25     4960.8
Total     209376.8     29
```

We want to test the equality of the 5 treatment means $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ with $\alpha = 0.05$. From the given information, we have $a = 5$ and $n = 6$.

Step 1: We test the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu \\ H_1 &: \text{At least one } \mu_i \neq \mu \end{aligned}$$

where μ is the overall mean common to all treatments.

Step 2: The value of the test statistic is given by

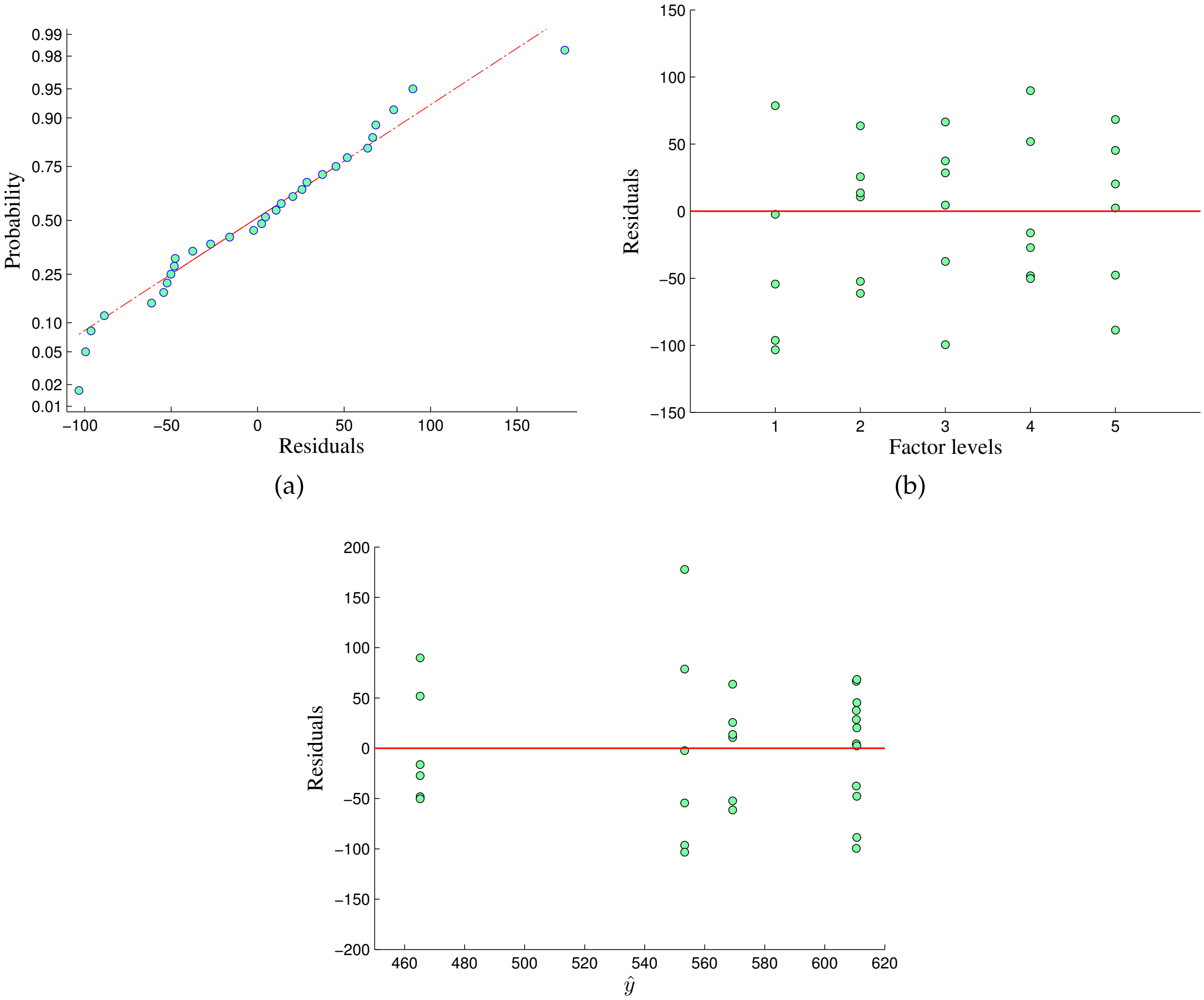
$$f_0 = \frac{MS_{\text{Treatments}}}{MS_E} = 4.3$$

Step 3: The rejection region is $f_0 > f_{\alpha,a-1,a(n-1)}$, where $f_{\alpha,a-1,a(n-1)} = f_{0.01,4,25} = 4.17742$.

Step 4: Since $f_0 = 4.3 > f_{\alpha,a-1,a(n-1)} = 4.17742$, we reject the null hypothesis H_0.

Thus, we conclude that the aggregare in concrete significantly affects the absorption of moisture in concrete. We can arrive at the same conclusion by noting that p-value $= 1 - F(4.3) = 0.0088 < \alpha = 0.05$.

(ii) The residual plots are depicted in Figure 6.16. The data of normal probability plot appear fairly linear, suggesting that no reason to doubt the normality assumption. The residual plots against the factor levels and fitted values exhibit random scatter around 0.



(a)

(b)

(c)

FIGURE 6.16: Residual analysis of the absorption of moisture experiment.

6.4 Problems

❶ Every Saturday, Montreal Gazette publishes in its HomeFront section a random sampler of recent real-estate transactions for houses and condominiums in Montreal. Data collected from a sample of 15 transactions are shown in Table 6.23, where x is the asking price (in thousand dollars) and y is the selling price (in thousand dollars) of the property.

Table 6.23: Montreal-area real-estate transactions data.

Asking Price	419	239	279	430	669	399	598	339	395	294	295	575	228	170	349
Selling Price	400	200	265	399	671	393	580	315	363	290	277	520	223	165	342

a) Does the selling price depend on the asking price or does the asking price depend on the selling price? Do you expect a positive or negative relationship between these two variables?

b) Find the least-squares regression line by choosing appropriate dependent and independent variables based on your answer in part a).

c) Interpret the meaning of the values $\hat{\beta}_0$ and $\hat{\beta}_1$ calculated in part b).

d) Display the scatter plot and the regression line.

e) Calculate the correlation and determination coefficients, and explain what they mean.

f) Predict the selling price for a house with an asking price of $350K.

g) Calculate the standard errors

h) Display the normal probability plot. Does the normality assumption seem reasonable?

i) Display the residual plot. Does the constant variance assumption seem reasonable?

j) Find a 95% confidence interval estimate on the slope.

k) Conduct a test to determine if the true slope of the line differs from 0 using $\alpha = 0.05$. What is the P-value for this test?

l) Test at the 5% significance level whether β_1 is positive. What is the P-value for this test?

❷ How do real estate agents decide on the asking price for a newly listed condominium? A computer database in a small community contains the listed selling price y (in thousands of dollars), the amount of living area x_1 (in hundreds of square feet), and the numbers of floors x_2, bedrooms x_3, and bathrooms x_4, for $n = 15$ randomly selected condos currently on the market. The data are shown in Table 6.24.

Table 6.24: Data on 15 condominiums.

List Price, y	Living Area, x_1	Floors, x_2	Bedrooms, x_3	Bathrooms, x_4
169.0	6	1	2	1
218.5	10	1	2	2
216.5	10	1	3	2
225.0	11	1	3	2
229.9	13	1	3	1.7
235.0	13	2	3	2.5
239.9	13	1	3	2
247.9	17	2	3	2.5
260.0	19	2	3	2
269.9	18	1	3	2
234.9	13	1	4	2
255.0	18	1	4	2
269.9	17	2	4	3
294.5	20	2	4	3
309.9	21	2	4	3

a) Fit the multiple linear regression model to these data.

b) Display the normal probability.

c) Display the residual plot against the fitted values.

d) Display the residual plots against each of the independent variables.

e) Predict the selling price of a 900 square feet condo with one floor, two bedrooms and one bath.

f) Test the significance of regression using the ANOVA procedure with $\alpha = 0.05$.

g) Calculate the coefficient of determination, and explain what is means.

❸ The compressive strength of concrete is being studied, and four different mixing techniques are being investigated. Table 6.25 presents the collected data.

a) Calculate the totals and averages for each mixing technique.

b) Estimate the treatment effects for the four mixing techniques.

c) Does mixing technique affect compressive strength of the concrete? Draw comparative box plots and perform an analysis of variance. Use $\alpha = 0.05$.

d) Analyze the residuals and comment on model adequacy.

TABLE 6.25: Data on compressive strength of concrete.

Mixing Technique	Compressive Strength (psi)			
1	3129	3000	2865	2890
2	3200	3300	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

❹ A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are randomly assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the results presented in Table 6.26.

a) Calculate the totals and averages for each program.

b) Estimate the treatment effects for the four programs.

c) Using a 5% significance level, determine whether the treatments differ in their effectiveness.

d) Draw comparative box plots and perform an analysis of variance. Use $\alpha = 0.05$.

e) Analyze the residuals and comment on model adequacy.

TABLE 6.26: Number of times the task is performed per minute.

	Observations				
Program	1	2	3	4	5
1	9	12	14	11	13
2	10	6	9	9	10
3	12	14	11	13	11
4	9	8	11	7	8

❺ An economist wants to test whether mean housing prices are the same regardless of which of three air-pollution levels typically prevails. A random sample of house purchases in three areas yields the price data presented in Table 6.27.

a) Calculate the totals and averages for each pollution level.

b) Estimate the treatment effects for the four pollution levels.

c) Using a 2.5% significance level, test whether housing prices differ by level of pollution.

d) Draw comparative box plots and perform an analysis of variance. Use $\alpha = 0.05$.

e) Analyze the residuals and comment on model adequacy.

f) Draw the box plot of all the residuals and comment on your plot.

TABLE 6.27: Mean housing prices (in thousands of dollars).

Pollution	Observations				
Level	1	2	3	4	5
Low	120	68	40	95	83
Medium	61	59	110	75	80
High	40	55	73	45	64

6.5 REFERENCES

[1] D.C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, 6th Edition, 2009.

[2] I. Bass and B. Lawton, *Lean Six Sigma using SigmaXL and Minitab*, McGraw-Hill Professional, 1st Edition, 2009.

Chapter 7

Design of Experiments

> All life is an experiment. The more experiments you make, the better.
>
> *Ralph Waldo Emerson*

Experiments with two or more factors are encountered frequently. A factorial experiment is one that investigates the effects of two or more independent variables (factors) on a single dependent variable (response). By a factor, we mean a discrete variable used to classify experimental units, such as temperature, time, or pressure that may be varied from trial to trial. Traditionally, experiments are designed to determine the effect of one variable (factor) upon one response. Factorial design involves two or more factors in a single experiment, and can reduce the number of experiments one has to perform by studying multiple factors simultaneously. Moreover, it can be used to find both main effects (from each independent factor) and interaction effects (when both factors must be used to explain the outcome).

7.1 Factorial Experiments

Modeling real world phenomena often requires more than just one factor. The analysis of variance (ANOVA) can be extended to handle the two-factor factorial experiment. Unlike one-way analysis of variance (one-way ANOVA) tests which measure significant effects of one factor only, two-way analysis of variance (**two-way ANOVA**) tests (also called two-factor ANOVA) measure the effects of two factors simultaneously. For example, an experiment might be defined by two parameters, such as treatment and time point. One-way ANOVA tests would be able to assess only the treatment effect or the time effect. Two-way ANOVA, on the other hand, would not only be able to assess both time and treatment in the same test, but also whether there is an interaction between the parameters.

We consider a two-factor completely randomized factorial design, where there are two factors that each has two or more levels. By "completely randomized", we mean that the participants have been randomly assigned to one of the unique levels of the factors. Let A and B be two fixed factors, with a levels of factor A and b levels of factor B. The two factors are crossed so that there are ab treatment combinations, and a total of abn experimental units are randomly allocated to the ab treatments with n per experiment. The n observations obtained under each treatment combination are called **replicates**; that is, the experiment has n replicates, and each replicate contains all ab treatment (or factor) combinations. This two-factor factorial design that has equal number of replicates is called **balanced two-way layout**, as shown in Table 7.1, where y_{ijk} is the observation in the (i,j)-th cell for the k-th replicate, $i = 1,2,\ldots,a$, $j = 1,2\ldots,b$, and $k = 1,2,\ldots,n$.

Let Y_{ijk} be a random variable denoting the observation in the (i,j)-th cell for the k-th replicate. Thus, the observations in a completely randomized design with a two-factor factorial treatment structure and $n > 1$ replicates may be described by the linear statistical model

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,b \\ k = 1,2,\ldots,n \end{cases} \tag{7.1.1}$$

TABLE 7.1: Data layout for a two-factor factorial experiment.

Factor A	Factor B levels					
Levels	1	2	$\ldots$	j	$\ldots$	b
1	$y_{111}, \ldots, y_{11n}$	$y_{121}, \ldots, y_{12n}$	$\ldots$	$y_{1j1}, \ldots, y_{1jn}$	$\ldots$	$y_{1b1}, \ldots, y_{1bn}$
2	$y_{211}, \ldots, y_{21n}$	$y_{221}, \ldots, y_{22n}$	$\ldots$	$y_{2j1}, \ldots, y_{2jn}$	$\ldots$	$y_{2b1}, \ldots, y_{2bn}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$y_{i11}, \ldots, y_{i1n}$	$y_{i21}, \ldots, y_{i2n}$	$\ldots$	$y_{ij1}, \ldots, y_{ijn}$	$\ldots$	$y_{ib1}, \ldots, y_{ibn}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	$y_{a11}, \ldots, y_{a1n}$	$y_{a21}, \ldots, y_{a2n}$	$\ldots$	$y_{aj1}, \ldots, y_{ajn}$	$\ldots$	$y_{ab1}, \ldots, y_{abn}$

where the terms of the model are defined as follows:

- y_{ijk} is the observed (response variable) value in the (i, j)-th cell for the k-th replicate. That is, y_{ijk} is the response from the k-th experimental unit receiving the i-th level of factor A and the j-th level of factor B
- μ is the overall mean effect (gran mean)
- τ_i is the effect of the i-th level of factor A
- β_j is the effect of the j-th level of factor B
- $(\tau\beta)_{ij}$ is the effect of the interaction between factor A and factor B
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$ is a random error.

Assuming a fixed-effects model, the τ_i, β_j, and $(\tau\beta)_{ij}$ satisfy the following conditions

$$\sum_{i=1}^{a} \tau_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} (\tau\beta)_{ij} = 0, \quad \sum_{j=1}^{b} (\tau\beta)_{ij} = 0.$$

Using the dot notation, let $y_{...}$ denote the grand total of all the observations, $\bar{y}_{i..}$ denote the mean of the observations taken at the i-th level of factor A; $\bar{y}_{.j.}$ denote the mean of the observations taken at the j-th level of factor B; $\bar{y}_{ij.}$ denote the mean of the observations in the (i, j)-th cell of Table 7.1; and $\bar{y}_{...}$ denote the grand mean of all the observations.

$$\begin{aligned}
y_{...} &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk} \\
\bar{y}_{i..} &= \frac{1}{bn} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk} \\
\bar{y}_{.j.} &= \frac{1}{an} \sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk} \\
\bar{y}_{ij.} &= \frac{1}{n} \sum_{k=1}^{n} y_{ijk} \\
\bar{y}_{...} &= \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}
\end{aligned} \tag{7.1.2}$$

It can be shown that the least squares estimates of parameters of the statistical model for a two-factor factorial experiment with interaction given by Eq. (7.1.1) are

$$\hat{\mu} = \bar{y}_{...}; \quad \hat{\tau}_i = \bar{y}_{i..} - \bar{y}_{...}; \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}; \quad \widehat{(\tau\beta)}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...} \tag{7.1.3}$$

Thus, the fitted model for a balanced two-factor factorial experiment is given by

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \widehat{(\tau\beta)}_{ij} = \bar{y}_{ij.} \tag{7.1.4}$$

That is, the (i,j)-th cell sample mean $\bar{y}_{ij.}$ is the fitted value for all observations in the (i,j)-th cell. A plot of the mean response $\bar{y}_{ij.}$ versus the level of factor A for different levels of B is called **interaction plot**, and it is used to check the presence of interactions between the factors A and B. The interaction plot gives a pictorial view of the tendency in the data to show the effect of changing one factor as one moves from one level to another of a second factor.

7.1.1 Hypotheses in Two-Way ANOVA and Total Variability Decomposition

We are interested in testing the hypotheses of no main effect for factor A, no main effect for B, and no AB interaction effect:

1. Hypotheses of no main effect for factor A

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0 \\ &H_1 : \text{At least one } \tau_i \neq 0 \end{aligned} \tag{7.1.5}$$

2. Hypotheses of no main effect for factor B

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = \ldots = \beta_b = 0 \\ &H_1 : \text{At least one } \beta_j \neq 0 \end{aligned} \tag{7.1.6}$$

3. Hypotheses of no AB interaction

$$\begin{aligned} &H_0 : (\tau\beta)_{11} = (\tau\beta)_{12} = \ldots = (\tau\beta)_{ab} = 0 \\ &H_1 : \text{At least one } (\tau\beta)_{ij} \neq 0 \end{aligned} \tag{7.1.7}$$

As in the previous Chapter, the ANOVA tests these hypotheses by partitioning the total variability (total sum of squares) SS_T in the data into the sum of squares SS_A of the factor A, the sum of squares SS_B of the factor B, the sum of squares SS_{AB} of the interaction AB, and the error sum of squares SS_E as follows:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E \tag{7.1.8}$$

where

$$\begin{aligned} SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{...}^2}{abn} \\ SS_A &= bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 \\ SS_B &= an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2 \\ SS_{AB} &= n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\ SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2 = SS_T - SS_A - SS_B - SS_{AB} \end{aligned} \tag{7.1.9}$$

Note that SS_T has $abn-1$ degrees of freedom (d.f.), SS_A has $a-1$ d.f, SS_B has $b-1$ d.f, and SS_{AB} has $(a-1)(b-1)$ d.f.. Thus, the error sum of squares SS_E has $ab(n-1)$ d.f. That is,

$$\underbrace{SS_T}_{abn-1\text{ d.f.}} = \underbrace{SS_A}_{a-1\text{ d.f.}} + \underbrace{SS_B}_{b-1\text{ d.f.}} + \underbrace{SS_{AB}}_{(a-1)(b-1)\text{ d.f}} + \underbrace{SS_E}_{ab(n-1)\text{ d.f.}} \tag{7.1.10}$$

It is worth pointing out that SS_A measures the main effect of factor A, SS_B measures the main effect of factor B, SS_{AB} measures the interaction effect of factors A and B, and SS_E represents the variability in the y_{ijk}'s not accounted for by the main effects and interaction effects.

The mean squares are defined as the sum of squares divided by their degrees of freedom:

$$MS_A = \frac{SS_A}{a-1}, \quad MS_B = \frac{SS_B}{b-1}, \quad MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}, \quad MS_E = \frac{SS_E}{ab(n-1)}.$$

Thus, in the case of a 2-way ANOVA, the total variability is divided up into four components: variability among the levels of each of the two factors, variability due to interaction of the two factors, and variability within cells (error variability). Three separate statistical tests are performed (based on the F-statistic), comparing each of the first three sources of variability (variability due to first factor, variability due to second factor, and variability due to interaction) to the error variability. In each test, the resulting F-statistic value or p-value allows us to determine whether that specific effect is significant.

The test statistics for testing the hypotheses of no main effect for factor A, no main effect for B, and no AB interaction effect are computed by dividing the mean square and the mean error sum of square:

1. Hypotheses of no main effect for factor A

$$\begin{aligned} H_0 &: \tau_1 = \tau_2 = \ldots = \tau_a = 0 \\ H_1 &: \text{At least one } \tau_i \neq 0 \end{aligned}$$

   If H_0 is true, then $F_0 = MS_A/MS_E \sim F(a-1, ab(n-1))$. At an α significance level, the null hypothesis would be rejected if the numerical value f_0 of F_0 is greater than $f_{\alpha,a-1,ab(n-1)}$.

2. Hypotheses of no main effect for factor B

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_b = 0 \\ H_1 &: \text{At least one } \beta_j \neq 0 \end{aligned}$$

   If H_0 is true, then $F_0 = MS_B/MS_E \sim F(b-1, ab(n-1))$. At an α significance level, the null hypothesis would be rejected if the numerical value f_0 of F_0 is greater than $f_{\alpha,b-1,ab(n-1)}$.

3. Hypotheses of no AB interaction

$$\begin{aligned} H_0 &: (\tau\beta)_{11} = (\tau\beta)_{12} = \ldots = (\tau\beta)_{ab} = 0 \\ H_1 &: \text{At least one } (\tau\beta)_{ij} \neq 0 \end{aligned}$$

   If H_0 is true, then $F_0 = MS_{AB}/MS_E \sim F((a-1)(b-1), ab(n-1))$. At an α significance level, the null hypothesis would be rejected if the numerical value f_0 of F_0 is greater than $f_{\alpha,(a-1)(b-1),ab(n-1)}$.

It is customary to present the decomposition of the sum of squares and the degrees of freedom in a tabular form (ANOVA table), as shown in Table 7.2

Example 7.1 *An experiment was conducted to determine the effects of four different pesticides on the yield of fruit from different varieties (B_1, B_2, B_3) of a citrus tree. Eight trees from each variety were randomly selected from an orchard. The four pesticides were then randomly assigned to two trees of each variety and applications were made according recommended levels. Yields of fruit (in bushels per tree) were obtained after the test period. The data for this experiment are given in Table 7.3.*

(i) Write an appropriate model for this experiment

(ii) Construct a plot of the treatment means (i.e. interaction plot).

(iii) Set up an analysis of variance table and conduct the appropriate F-tests on main effects and interactions using a significance level $\alpha = 0.05$

TABLE 7.2: ANOVA table for a two-factor experiment.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	$a-1$	$MS_A = \dfrac{SS_A}{a-1}$	$\dfrac{MS_A}{MS_E}$
B	SS_B	$b-1$	$MS_B = \dfrac{SS_B}{b-1}$	$\dfrac{MS_B}{MS_E}$
Interaction	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \dfrac{SS_{AB}}{(a-1)(b-1)}$	$\dfrac{MS_{AB}}{MS_E}$
Error	SS_E	$ab(n-1)$	$MS_E = \dfrac{SS_E}{ab(n-1)}$	
Total	SS_T	$abn-1$		

TABLE 7.3: Data for the two-factor factorial experiment of fruit tree yield.

Pesticide	Variety (Factor B)		
(Factor A)	1	2	3
1	49 39	55 41	66 68
2	50 55	67 58	85 92
3	43 38	53 42	69 62
4	53 48	85 73	85 99

Solution: From the given information, the experiment is a completely randomized factorial experiment with factor, A, pesticides, having $a = 4$ levels and factor, B, variety, having $b = 3$ levels. There are $n = 2$ replicates of the 12 factor-level combination of the two factors.

(i) The model for this experiment is given by

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1,2,3,4 \\ j = 1,2,3 \\ k = 1,2 \end{cases} \tag{7.1.11}$$

where μ is the overall mean yield per tree, τ_i is the effect of the i-th level of pesticide, β_j is the effect of the j-th level of variety of citrus tree, and $(\tau\beta)_{ij}$ is the interaction effect of the i-th level of pesticide with the j-th level of variety of citrus tree.

(ii) Figure 7.1 depicts the interaction plot for fruit yield experiment. From this plot, we can observe a lack of interaction. We also observe that the differences in mean yields among the three varieties of citrus trees remain nearly constant across the four pesticide levels. That is, the three lines for the three varieties are nearly parallel lines and hence the interaction between the levels of variety and pesticide is not significant.

(iii) The analysis of variance is shown in the ANOVA Table 7.4.

MATLAB code

```
>> n=2; %number of replicates
>> a = 4; %factor A has a levels
>> b = 3; %factor B has b levels
>> data = [49 39 55 41 66 68
           50 55 67 58 85 92
           43 38 53 42 69 62
           53 48 85 73 85 99];
>> y = data'; %transpose the data
>> [p,table,stats] = anova2(y,n);
```

Because the interaction is not significant as shown in the interaction plot, we can next test the main effects of the two factors. These tests separately examine the differences among the levels of variety and the levels of pesticides. Then, we test the interaction between the different pesticides and the different varieties. The appropriate F tests on main effects and interactions using a significance level $\alpha = 0.05$ are as follows:

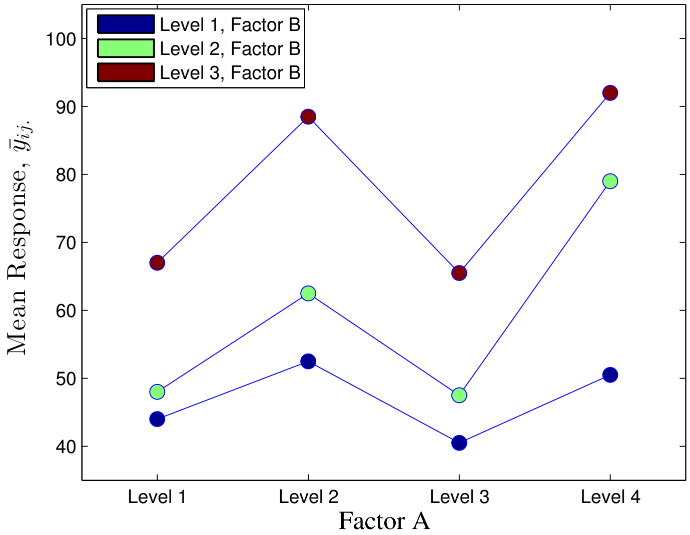
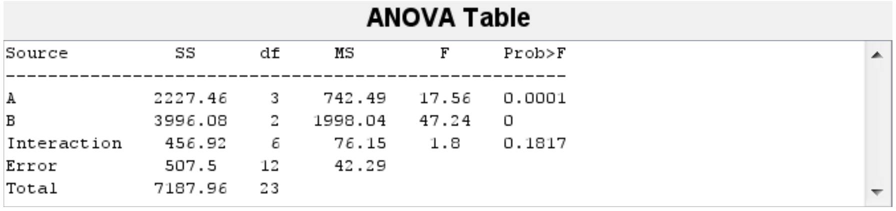


FIGURE 7.1: Interaction plot for fruit yield experiment.

TABLE 7.4: ANOVA table for fruit yield experiment.



ANOVA Table

```
Source            SS       df     MS         F      Prob>F
-------------------------------------------------------
A              2227.46     3    742.49    17.56    0.0001
B              3996.08     2   1998.04    47.24    0
Interaction     456.92     6     76.15     1.8     0.1817
Error           507.5     12     42.29
Total          7187.96    23
```

- F-test for factor A

  Step 1: We test the hypotheses of no main effect for factor A (pesticides):

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$
$$H_1 : \text{At least one } \tau_i \neq 0$$

  Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_A}{MS_E} = 17.56$$

  Step 3: The rejection region is $f_0 > f_{\alpha,a-1,ab(n-1)}$, where $f_{\alpha,a-1,ab(n-1)} = f_{0.05,3,12} = 3.4903$.

  Step 4: Since $f_0 = 17.56 > f_{\alpha,a-1,ab(n-1)} = 3.4903$, we reject the null hypothesis H_0. Thus, we have sufficient evidence to indicate a significant difference in the mean yields among the four pesticide levels.

- F-test for factor B

  Step 1: We test the hypotheses of no main effect for factor B (varieties):

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1 : \text{At least one } \beta_j \neq 0$$

  Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_B}{MS_E} = 47.24$$

  Step 3: The rejection region is $f_0 > f_{\alpha,b-1,ab(n-1)}$, where $f_{\alpha,b-1,ab(n-1)} = f_{0.05,2,12} = 3.8853$.

Step 4: Since $f_0 = 47.24 > f_{\alpha,b-1,ab(n-1)} = 3.8853$, we reject the null hypothesis H_0. Thus, we have sufficient evidence to indicate a significant difference in the mean yields among the three varieties of citrus trees.

- F-test for interaction AB

Step 1: We test the hypotheses

$$H_0 : (\tau\beta)_{11} = (\tau\beta)_{12} = \ldots = (\tau\beta)_{43} = 0$$
$$H_1 : \text{At least one } (\tau\beta)_{ij} \neq 0$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_{AB}}{MS_E} = 1.8$$

Step 3: The rejection region is $f_0 > f_{\alpha,(a-1)(b-1),ab(n-1)}$, where $f_{\alpha,(a-1)(b-1),ab(n-1)} = f_{0.05,6,12} = 2.9961$.

Step 4: Since $f_0 = 1.8 \ngtr f_{\alpha,(a-1)(b-1),ab(n-1)} = 2.9961$, we do not reject the null hypothesis H_0. Thus, we have insufficient evidence to indicate and interaction between pesticide levels and variety of trees levels.

7.1.2 Residual Analysis

The residuals are simply the differences between the data values and the corresponding cell sample means:

$$e_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \bar{y}_{ij.} \tag{7.1.12}$$

The normal probability plot is used to check the normality assumption of the errors $\varepsilon_{ijk} \sim N(0, \sigma^2)$, while the residual plot is used to check the constant variance assumption. In terms of residuals, the error sum of squares can be written as

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} e_{ijk}^2 \tag{7.1.13}$$

Figure 7.2 shows the residual plots for fruit yield experiment. The data of normal probability plot appear fairly linear, suggesting that no reason to doubt the normality assumption. Also, the residual plots against the fitted values as well as against the factor levels of A and B exhibit random scatter around 0. Thus, the model assumptions are valid.

Example 7.2 *Aircraft primer paints are applied to aluminum surfaces by two methods: dipping and spraying. The purpose of the primer is to improve paint adhesion, and some parts can be primed using either application method. The process engineering group responsible for this operation is interested in learning whether three different primers differ in their adhesion properties. A factorial experiment was performed to investigate the effect of paint primer type and application method on paint adhesion. For each combination of primer type and application method, three specimens were painted, then a finish paint was applied, and the adhesion force was measured. The data from this experiment are given in Table 7.5.*

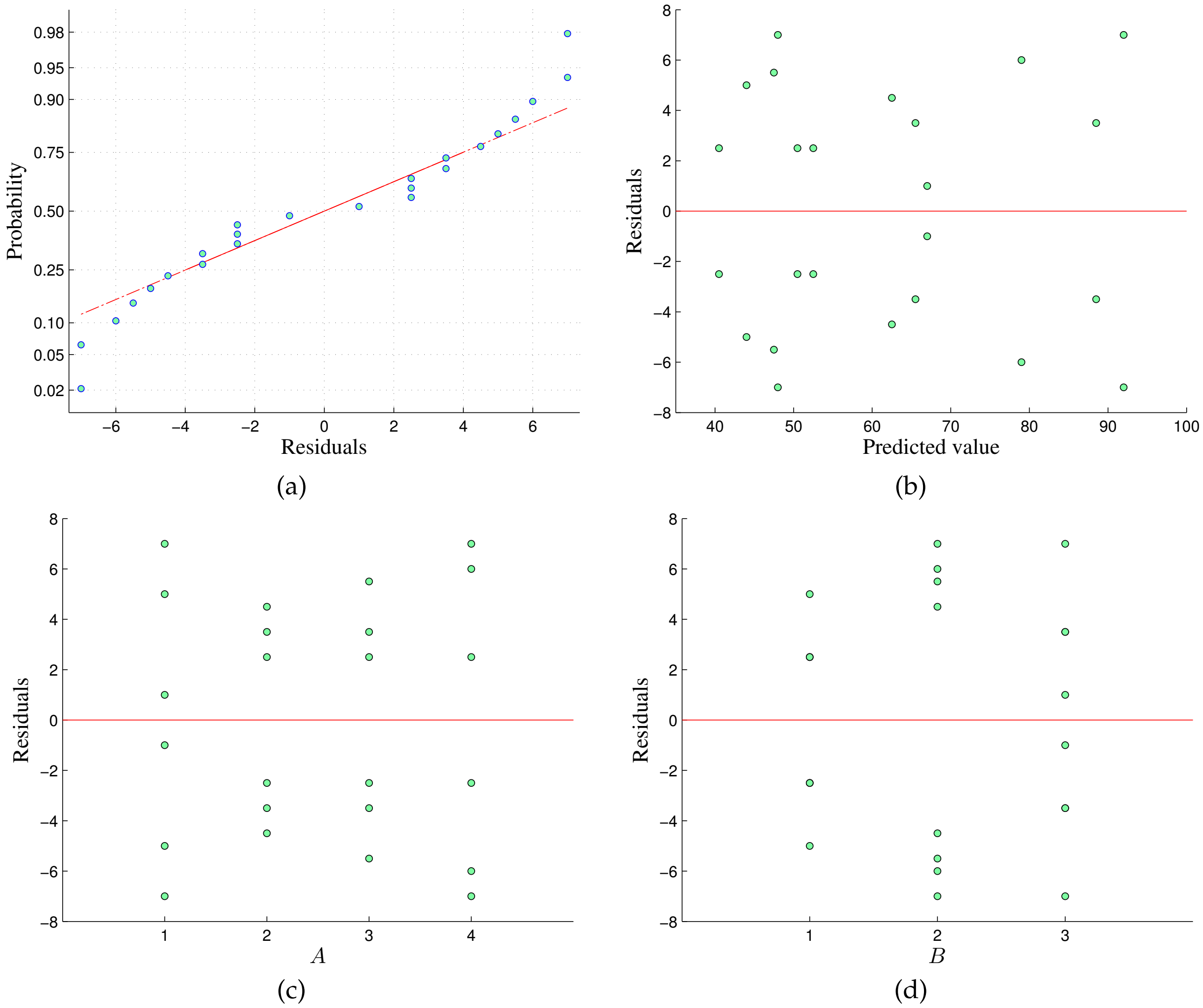
(i) Write an appropriate statistical model for this experiment

(ii) Set up an analysis of variance table and conduct the appropriate F-tests on main effects and interactions using a significance level $\alpha = 0.05$

(iii) Construct the interaction plot.

(iv) Perform the residual analysis.

Solution: From the given information, the experiment is a completely randomized factorial experiment with factor, A, primer type, having $a = 3$ levels and factor, B, application method, having $b = 2$ levels. There are $n = 3$ replicates of the 6 factor-level combination of the two factors.



(a) (b)

(c) (d)

FIGURE 7.2: Residual analysis for fruit yield experiment: (a) Normal probability plot; (b) Residual plot against fitted values; (c) Residual plot against factor A; (d) Residual plot against factor B.

TABLE 7.5: Data for the two-factor factorial experiment of adhesion force.

Primer Type	Application Method	
	Dipping	Spraying
1	4.0, 4.5, 4.3	5.4, 4.9, 5.6
2	5.6, 4.9, 5.4	5.8, 6.1, 6.3
3	3.8, 3.7, 4.0	5.5, 5.0, 5.0

(i) The model for this experiment is given by

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1,2,3 \\ j = 1,2 \\ k = 1,2,3 \end{cases} \tag{7.1.14}$$

where μ is the overall mean adhesion force, τ_i is the effect of the i-th level of primer type, β_j is the effect of the j-th level of application method, and $(\tau\beta)_{ij}$ is the interaction effect of the i-th level of primer type with the j-th level of application method.

(ii) The MATLAB output for the analysis of variance table is given in the Table 7.4. The appropriate F-tests on main effects and interactions using a significance level $\alpha = 0.05$ are as follows:

- F-test for factor A

  Step 1: We test the hypotheses of no main effect for factor A (primer type):

  $$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \tau_3 = 0 \\ &H_1 : \text{At least one } \tau_i \neq 0 \end{aligned}$$

  Step 2: The value of the test statistic is given by

  $$f_0 = \frac{MS_A}{MS_E} = 27.86$$

  Step 3: The rejection region is $f_0 > f_{\alpha,a-1,ab(n-1)}$, where $f_{\alpha,a-1,ab(n-1)} = f_{0.05,2,12} = 3.8853$.

  Step 4: Since $f_0 = 27.86 > f_{\alpha,a-1,ab(n-1)} = 3.8853$, we reject the null hypothesis H_0. Thus, we have sufficient evidence to indicate a significant difference in the mean adhesion force among the three levels of primer type.

- F-test for factor B

  Step 1: We test the hypotheses of no main effect for factor B (application method):

  $$\begin{aligned} &H_0 : \beta_1 = \beta_2 = 0 \\ &H_1 : \text{At least one } \beta_j \neq 0 \end{aligned}$$

  Step 2: The value of the test statistic is given by

  $$f_0 = \frac{MS_B}{MS_E} = 59.7$$

  Step 3: The rejection region is $f_0 > f_{\alpha,b-1,ab(n-1)}$, where $f_{\alpha,b-1,ab(n-1)} = f_{0.05,1,12} = 4.7472$.

  Step 4: Since $f_0 = 59.7 > f_{\alpha,b-1,ab(n-1)} = 4.7472$, we reject the null hypothesis H_0. Thus, we have sufficient evidence to indicate a significant difference in the mean adhesion force among the two application methods.

- F-test for interaction AB

Step 1: We test the hypotheses

$$\begin{aligned} &H_0 : (\tau\beta)_{11} = (\tau\beta)_{12} = \ldots = (\tau\beta)_{32} = 0 \\ &H_1 : \text{At least one } (\tau\beta)_{ij} \neq 0 \end{aligned}$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_{AB}}{MS_E} = 1.47$$

Step 3: The rejection region is $f_0 > f_{\alpha,(a-1)(b-1),ab(n-1)}$, where $f_{\alpha,(a-1)(b-1),ab(n-1)} = f_{0.05,2,12} = 3.8853$.

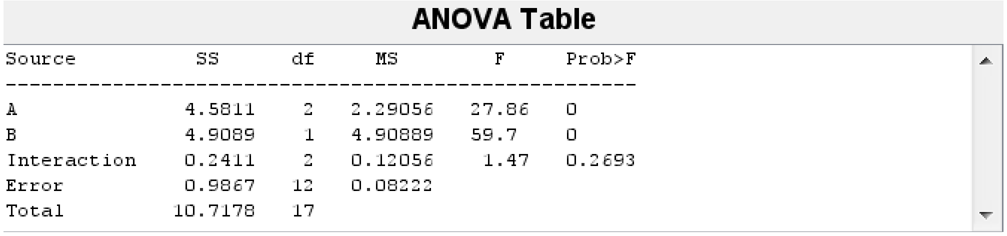
Step 4: Since $f_0 = 1.47 \ngtr f_{\alpha,(a-1)(b-1),ab(n-1)} = 3.8853$, we do not reject the null hypothesis H_0. Thus, we have insufficient evidence to indicate and interaction between levels primer type and levels of application method.

These conclusions can also be drawn from the last column of the ANOVA table by noticing that the p-values for the two test statistics on the main effects are considerably smaller than $\alpha = 0.05$, while the p-value for the test statistic on the interaction is greater than $\alpha = 0.05$.

(iii) Figure 7.3 depicts the interaction plot for adhesion force experiment. From this plot, we can observe a lack of interaction because the lines are nearly parallel. Thus, the interaction between the levels of primer type and application method is not significant. Furthermore, since a large response indicates greater adhesion force, we conclude that spraying is the best application method and that primer type 2 is most effective.

(iv) The residuals are presented in Table 7.7. Figure 7.4 shows the residual plots for adhesion force experiment. The data of normal probability plot appear fairly linear, suggesting that no reason to doubt the normality assumption. Also, the residual plots against the fitted values as well as against the factor levels of A and B exhibit random scatter around 0. Thus, the model assumptions are valid.

TABLE 7.6: ANOVA table for for adhesion force experiment.



ANOVA Table

```
Source            SS       df      MS          F      Prob>F
---------------------------------------------------------
A               4.5811     2    2.29056    27.86    0
B               4.9089     1    4.90889    59.7     0
Interaction     0.2411     2    0.12056     1.47    0.2693
Error           0.9867    12    0.08222
Total          10.7178    17
```

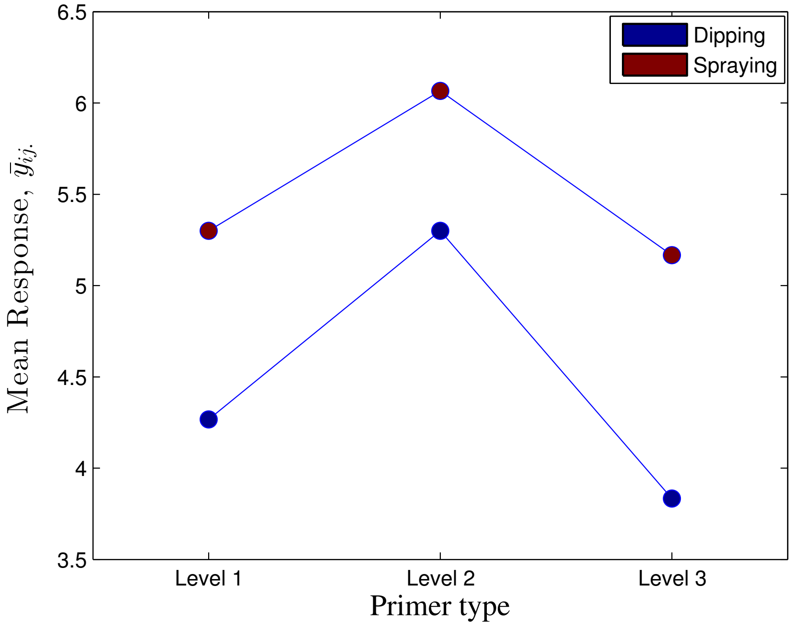


FIGURE 7.3: Interaction plot for adhesion force experiment.

TABLE 7.7: Residuals for the two-factor factorial experiment of adhesion force.

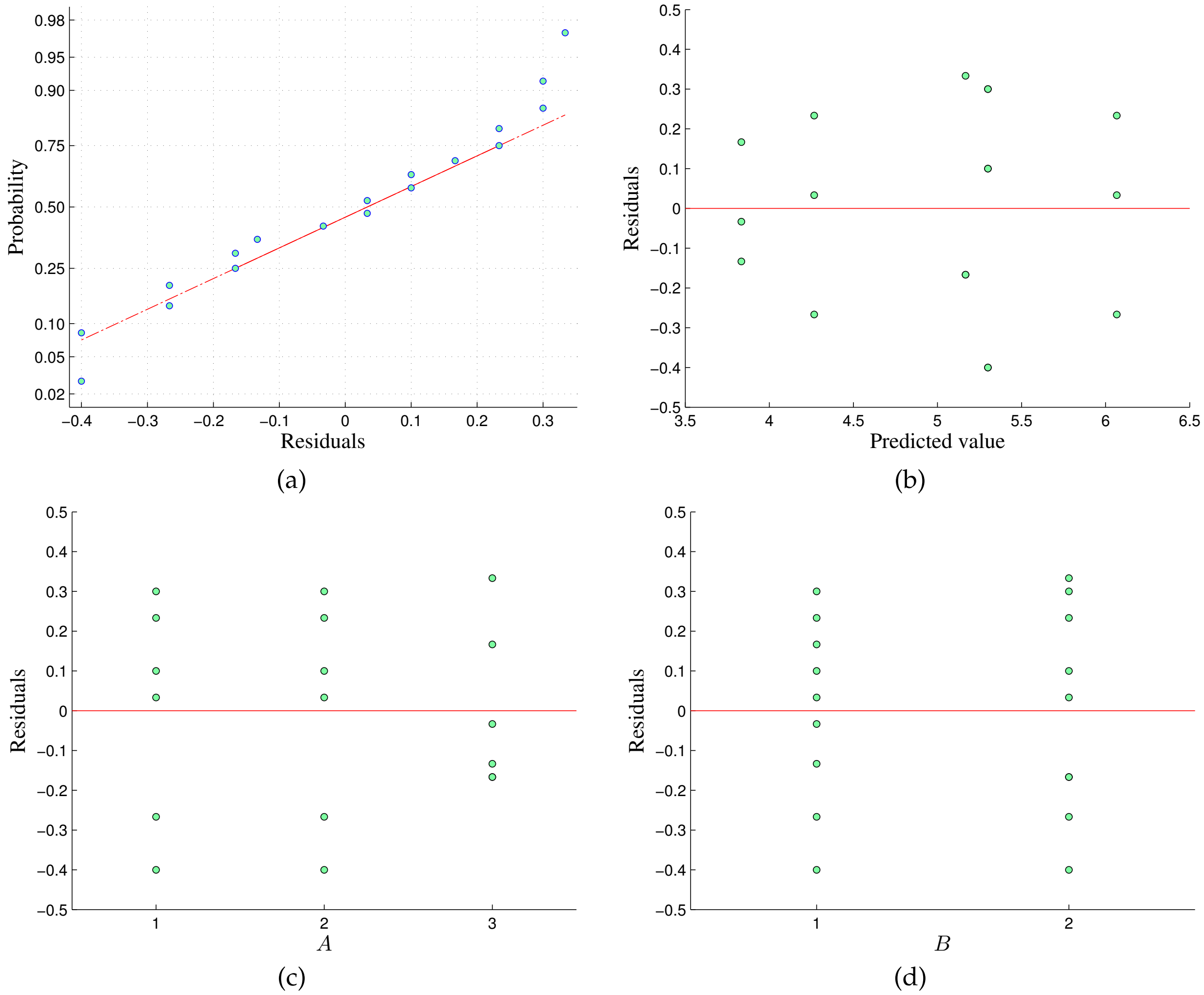
Primer Type	Application Method	
	Dipping	Spraying
1	-0.2667, 0.2333, 0.0333	0.1000, -0.4000, 0.3000
2	0.3000, -0.4000, 0.1000	-0.2667, 0.0333, 0.2333
3	-0.0333, -0.1333, 0.1667	0.3333, -0.1667, -0.1667

7.1.3 TWO-WAY ANOVA WITH ONE REPLICATE

Two-way tests can also be analyzed on data with only one replicate ($n = 1$) per treatment combination. However, the interaction between the factors cannot be tested. In other words, it is not possible to test the null hypothesis of no interaction, but we can test the null hypotheses regarding main effects as long as we assume no interaction. Two-way ANOVA table without replication is shown in Table 7.8.

Example 7.3 *A building contractor employs three construction engineers, E_1, E_2, and E_3, to estimate and bid on jobs. To determine whether one tends to be a more conservative (or liberal) estimator than the others, the contractor selects four projected construction jobs and has each estimator independently estimate the cost (in dollars per square foot) of each job. The data for this experiment are given in Table 7.9. Set up an analysis of variance table and conduct the appropriate F-tests on main effects using a significance level $\alpha = 0.05$*

Solution: From the given information, the experiment is a completely randomized factorial experiment with factor, A, estimator, having $a = 3$ levels and factor, B, construction job, having $b = 4$ levels. There is $n = 1$ replicate of the 12



(a)

(b)

(c)

(d)

FIGURE 7.4: Residual analysis for adhesion force experiment: (a) Normal probability plot; (b) Residual plot against fitted values; (c) Residual plot against factor A; (d) Residual plot against factor B.

TABLE 7.8: ANOVA table for a two-factor experiment with one replicate.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	$a-1$	$MS_A = \dfrac{SS_A}{a-1}$	$\dfrac{MS_A}{MS_E}$
B	SS_B	$b-1$	$MS_B = \dfrac{SS_B}{b-1}$	$\dfrac{MS_B}{MS_E}$
Error	SS_E	$(a-1)(b-1)$	$MS_E = \dfrac{SS_E}{ab-1}$	
Total	SS_T	$ab-1$		

factor-level combination of the two factors. The analysis of variance is shown in the ANOVA Table 7.10.

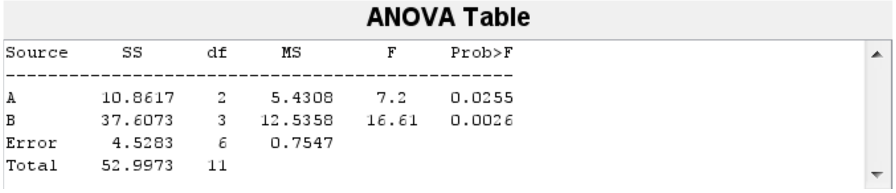
We test the main effects of the two factors. These tests separately examine the differences among the levels of estimator and the levels of construction job. The appropriate F tests on main effects and interactions using a significance level $\alpha = 0.05$ are as follows:

- F-test for factor A

TABLE 7.9: Data for the two-factor factorial experiment of the building contractor.

Estimator	Construction Job (Factor B)			
(Factor A)	1	2	3	4
E_1	35.10	34.50	29.25	31.60
E_2	37.45	34.60	33.10	34.40
E_3	36.30	35.10	32.45	32.90

TABLE 7.10: ANOVA table for the building contractor experiment.



ANOVA Table

```
Source       SS        df       MS          F       Prob>F
----------------------------------------------------------
A          10.8617     2      5.4308      7.2     0.0255
B          37.6073     3     12.5358     16.61    0.0026
Error       4.5283     6      0.7547
Total      52.9973    11
```

Step 1: We test the hypotheses of no main effect for factor A (pesticides):

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$$
$$H_1 : \text{At least one } \tau_i \neq 0$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_A}{MS_E} = 7.2$$

Step 3: The rejection region is $f_0 > f_{\alpha,a-1,(a-1)(b-1)}$, where $f_{\alpha,a-1,(a-1)(b-1)} = f_{0.05,2,6} = 5.1433$.

Step 4: Since $f_0 = 7.2 > f_{\alpha,a-1,(a-1)(b-1)} = 5.1433$, we reject the null hypothesis H_0. Thus, we have sufficient evidence to indicate a significant difference in the mean among the three estimators.

- F-test for factor B

Step 1: We test the hypotheses of no main effect for factor B (varieties):

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 : \text{At least one } \beta_j \neq 0$$

Step 2: The value of the test statistic is given by

$$f_0 = \frac{MS_B}{MS_E} = 16.61$$

Step 3: The rejection region is $f_0 > f_{\alpha,b-1,(a-1)(b-1)}$, where $f_{\alpha,b-1,(a-1)(b-1)} = f_{0.05,3,6} = 4.7571$.

Step 4: Since $f_0 = 47.24 > f_{\alpha,b-1,(a-1)(b-1)} = 4.7571$, we reject the null hypothesis H_0. Thus, we have sufficient evidence to indicate a significant difference in the mean among the four construction jobs.

7.2 HIGH-FACTOR EXPERIMENTS

The two-factor experiments can be generalized to more than 2 factors. However, each additional factor adds a layer of complexity to the analysis. For example, in the case of three-factor experiment there are a levels of factor A, b levels of factor B, and c levels of factor C. The statistical model for a three-factor experiment can be written as follows:

$$y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl} \quad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,b \\ k = 1,2,\ldots,c \\ l = 1,2,\ldots,n \end{cases} \tag{7.2.1}$$

where τ_i, β_j, and γ_k are the main effects; $(\tau\beta)_{ij}$, $(\tau\gamma)_{ik}$, and $(\beta\gamma)_{jk}$ are the two-interaction effects that have the same interpretation as in the case of the two-factor experiment; and $(\tau\beta\gamma)_{ijk}$ is the three-factor interaction effect.

7.3 2^k Factorial Designs

In this section, we focus on experimental designs in which the experimental plan calls for the study of the effect on a response of k factors, each at two levels. These are commonly known as 2^k **factorial experiments**. For example, 2^2 (two-level) factorial experiments are factorial experiments in which each factor is investigated at only two levels.

The factorial experiments, where all combination of the levels of the factors are run, are usually referred to as full factorial experiments. Full factorial two-level experiments are also referred to as 2^k designs where k denotes the number of factors being investigated in the experiment. These designs are referred to as two-level full factorial designs. A full factorial two level design with k factors requires 2^k runs for a single replicate. For example, a two-level experiment with three factors will require $2^3 = 8$ runs. The choice of the two levels of factors used in two level experiments depends on the factor - some factors naturally have two levels. For example, if gender is a factor, then male and female are the two levels.

Assume we have n observations $y_{ij}, i = 1, \ldots, 2^k; j = 1, \ldots, n$ that are made at each of each of the 2^k runs, and denote by $\bar{y}_i = (\sum_j^n y_{ij})/n$ the sample mean of the data of each run, as shown in Table 7.11. The grand mean, $\bar{\bar{y}}$, is given by

$$\bar{\bar{y}} = \frac{1}{2^k} \sum_{i=1}^{2^k} \bar{y}_i = \frac{1}{2^k n} \sum_{i=1}^{2^k} \sum_{j=1}^{n} y_{ij} \tag{7.3.1}$$

Table 7.11: Data Layout for a 2^k Design.

	Data			
Run	Rep 1	…	Rep n	Averages
1	y_{11}	$\cdots$	y_{1n}	$\bar{y}_1$
2	y_{21}	$\cdots$	y_{2n}	$\bar{y}_2$
3	y_{31}	$\cdots$	y_{3n}	$\bar{y}_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2^k	$y_{2^k 1}$	$\cdots$	$y_{2^k n}$	$\bar{y}_{2^k}$

7.3.1 2^2 Factorial Design

The simplest of the two-level factorial experiments is the 2^2 design, where two factors A and factor B are investigated at two levels. A single replicate of this design will require four runs ($2^2 = 4$). The geometry of the 2^2 design can be represented using a square with the four treatment combinations lying at the four corners, as shown in Figure 7.5 (right). Note that the design can be represented geometrically as a square with the runs forming the corners of the square. Figure 7.5 shows the 4 runs in a tabular format often called the **test matrix**. The signs "-" and "+" denote the low and high levels of a factor, respectively. Figure 7.5 (left) shows the treatments for this design, where letters are used to represent the treatments. For example, (1) represents the treatment combination where all factors involved are at the low level or the level represented by -1; a represents the treatment combination where factor A is at the high level, while the remaining factors (in this case, factor B) are at the low level. Similarly, b represents the treatment combination where factor B is at the high level, while factor A is at the low level and ab represents the treatment combination where factors A and B are at the high level. The letters (1), a, b, and ab represent the totals of all n observations taken at these design points; that is n observations are made at each of the $2^2 = 4$ treatment combinations (runs). Denote these n observations by $y_{ij}, i = 1, \ldots, 4; j = 1, \ldots, n$, as shown in Table 7.12. The averages, $\bar{y}_i = (\sum_{j=1}^n y_{ij})/n$, $i = 1, \ldots 4$, are also shown in the last column of the table.

The design matrix, displayed in Figure 7.5 (center), shows that the sum of the terms resulting from the product of any two columns of the design matrix is zero. That is, the 2^2 design is an orthogonal design. In fact all 2^k designs are orthogonal designs. Column I represents the intercept term. Columns A and B represent the respective factor settings. Column AB represents the interaction and is the product of columns A and B. Note that with the exception of identity column I, every column has an equal number of + and signs.

The effects investigated by the 2^2 design are the main effects A and B and the two-factor interaction AB. The main effect of A (resp. B) is defined as the difference in the mean response between the high level of A (resp. B), averaged

TABLE 7.12: Data Layout for a 2^2 Design.

Treatment	Factor		Data			
Combination	A	B	Rep 1	$\dots$	Rep n	Averages
(1)	-	-	y_{11}	$\dots$	y_{1n}	$\bar{y}_1$
a	+	-	y_{21}	$\dots$	y_{2n}	$\bar{y}_2$
b	-	+	y_{31}	$\dots$	y_{3n}	$\bar{y}_3$
ab	+	+	y_{41}	$\dots$	y_{4n}	$\bar{y}_4$

Test Matrix

Run	A	B
(1)	-	-
a	+	-
b	-	+
ab	+	+

Design Matrix

$$\begin{array}{c} \begin{array}{cccc} \text{I} & \text{A} & \text{B} & \text{AB} \end{array} \\ \begin{pmatrix} + & - & - & + \\ + & + & - & - \\ + & - & + & - \\ + & + & + & + \end{pmatrix} \end{array}$$

Design Geometry

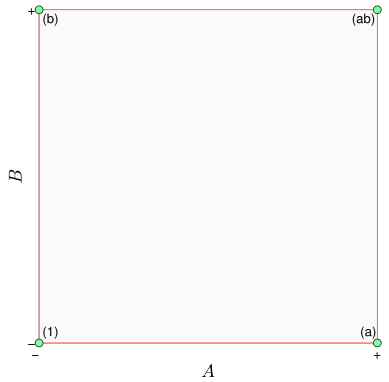


FIGURE 7.5: 2^2 design.

over the levels of B (resp. A), while the main effect of the interaction AB is defined as the difference between the mean effect of A at the high level of B and at the low level of B. Thus, Effect$_A$, Effect$_B$, and Effect$_{AB}$ are estimated as follows:

$$\begin{aligned} \text{Effect}_A &= A = \bar{y}_{A+} - \bar{y}_{A-} = \frac{a+ab}{2n} - \frac{b+(1)}{2n} = \frac{1}{2n}\left(a+ab-b-(1)\right) = \frac{\text{Contrast}_A}{2n} \\ \text{Effect}_B &= B = \bar{y}_{B+} - \bar{y}_{B-} = \frac{b+ab}{2n} - \frac{a+(1)}{2n} = \frac{1}{2n}\left(b+ab-a-(1)\right) = \frac{\text{Contrast}_B}{2n} \\ \text{Effect}_{AB} &= AB = \frac{ab+(1)}{2n} - \frac{a+b}{2n} = \frac{1}{2n}\left(ab+(1)-a-b\right) = \frac{\text{Contrast}_{AB}}{2n} \end{aligned} \tag{7.3.2}$$

where the contrast of a factor or interaction is obtained by point-wise multiplication of the treatment combinations in the first column of Table 7.13 by the signs in the corresponding factor or interaction column.
Similarly, the sums of squares can be estimated as follows:

$$\begin{aligned} SS_A &= \frac{\left(a+ab-b-(1)\right)^2}{4n} = \frac{(\text{Contrast}_A)^2}{4n} \\ SS_B &= \frac{\left(b+ab-a-(1)\right)^2}{4n} = \frac{(\text{Contrast}_B)^2}{4n} \\ SS_{AB} &= \frac{\left(ab+(1)-a-b\right)^2}{4n} = \frac{(\text{Contrast}_{AB})^2}{4n} \end{aligned} \tag{7.3.3}$$

In general, for any 2^k design with n replicates, the effect estimates and the sum of squares for any effect or interaction are computed as follows:

TABLE 7.13: Signs for Contrasts in the 2^2 Design.

Treatment Combination	Factorial Effect		
	A	B	AB
(1)	-	-	+
a	+	-	-
b	-	+	-
ab	+	+	+

$$\begin{aligned} \text{Effect} &= \frac{\text{Contrast}}{n2^{k-1}} \\ SS &= \frac{(\text{Contrast})^2}{n2^k} = n(\text{Effect})^2\, 2^{k-2} \end{aligned} \tag{7.3.4}$$

Example 7.4 *Consider the data in Table 7.14 for a two-factor factorial experiment, where the data numbers denote the averages of the cells.*

TABLE 7.14: Factorial experiment without interaction.

Factor A	Factor B	
	B_{low}	B_{high}
A_{low}	15	25
A_{high}	40	50

1. *Draw the design geometry*

2. *Find the main effects of A, B, and AB*

3. *Draw the interaction plot.*

Solution:

1. The design geometry is shown in Figure 7.6.

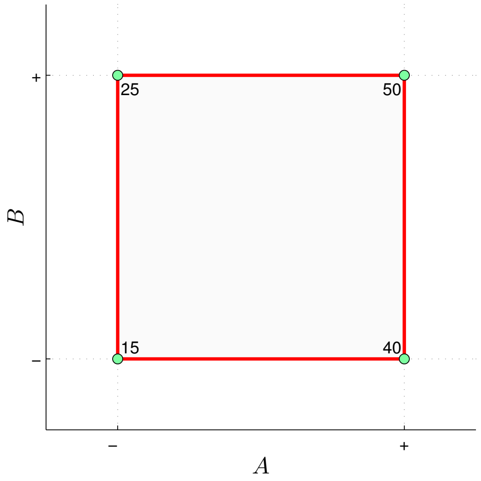


FIGURE 7.6: Design geometry.

2. The main effects of A, B, and AB are

$$A = \frac{40+50}{2} - \frac{15+25}{2} = 25$$

$$B = \frac{25+50}{2} - \frac{15+40}{2} = 10$$

$$AB = \frac{15+50}{2} - \frac{25+40}{2} = 0$$

3. The interaction plot, shown in Figure 7.7, reveals that there is no interaction (parallel lines) between the factors A and B.

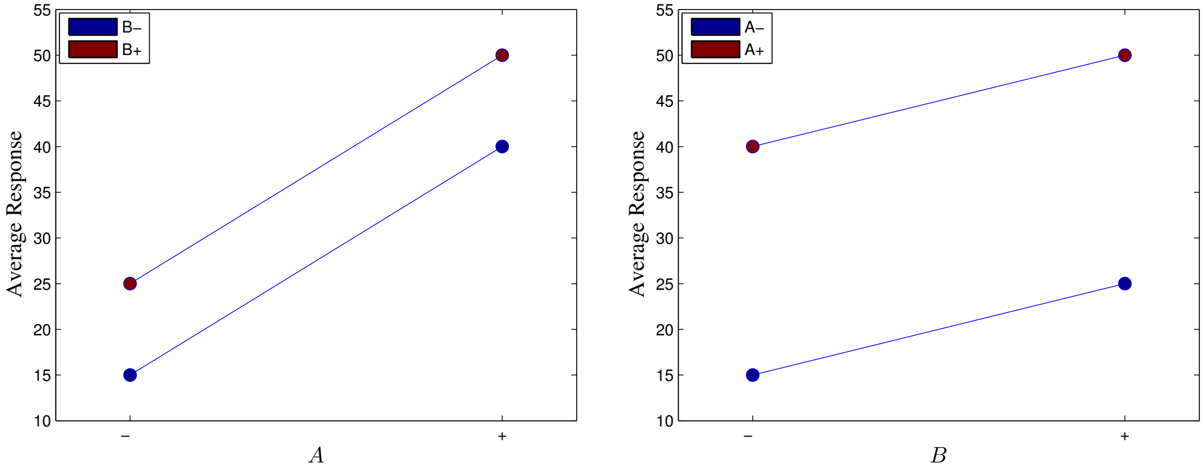


FIGURE 7.7: Interaction plots.

MATLAB code

```
>> y = [15 40 25 50]; %average response
>> M = fracfact('b a ba'); % order: A B AB
>> interactionplot(y,{M(:,1) M(:,2)},'varnames',{'A','B'})
```

Example 7.5 *Consider the data in Table 7.15 for a two-factor factorial experiment, where the data numbers denote the averages of the cells.*

TABLE 7.15: Factorial experiment without interaction.

Factor A	Factor B	
	B_{low}	B_{high}
A_{low}	15	25
A_{high}	40	0

1. Draw the design geometry

2. Find the main effects of A, B, and AB

3. Draw the interaction plot.

Solution:

1. The design geometry is shown in Figure 7.8.

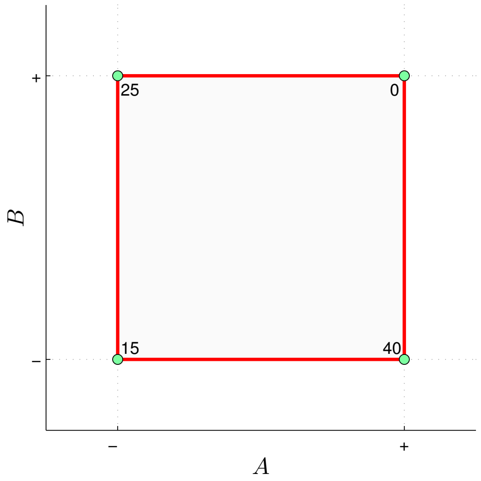


FIGURE 7.8: Design geometry.

2. The main effects of A, B, and AB are

$$A = \frac{40+0}{2} - \frac{15+25}{2} = 0$$

$$B = \frac{25+0}{2} - \frac{15+40}{2} = -15$$

$$AB = \frac{15+0}{2} - \frac{25+40}{2} = -25$$

3. The interaction plot, shown in Figure 7.9, reveals that there is interaction between the factors A and B.

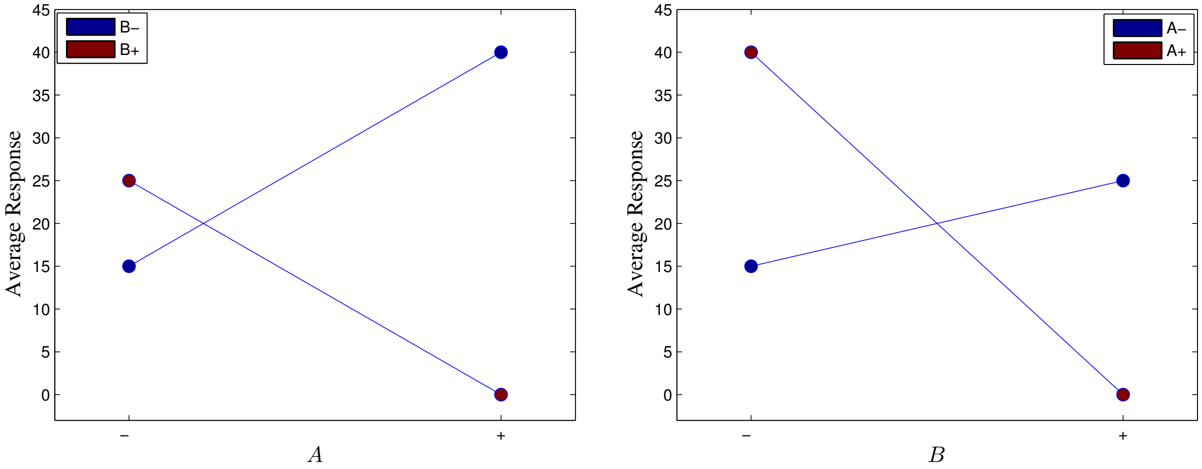


FIGURE 7.9: Interaction plots.

MATLAB code

```
>> y = [15 40 25 0]; %average response
>> M = fracfact('b a ba'); % order: A B AB
>> interactionplot(y,{M(:,1) M(:,2)},'varnames',{'A','B'})
```

7.3.2 Analysis of Variance for 2^k Design

The total sum of squares, which has $2^k n - 1$ degrees of freedom, is given by

$$SS_T = \sum_{i=1}^{2^k} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2 \tag{7.3.5}$$

where $\bar{y}$ is the grand mean of all observations. Thus, the error sum of squares, which has $2^k n - 1$ degrees of freedom, is obtained by subtraction

$$SS_E = SS_T - SS_A - SS_B - SS_{AB}. \tag{7.3.6}$$

The ANOVA table for the two-factor (2^k with $k = 2$) factorial design is shown in Table 7.16.

TABLE 7.16: ANOVA table for a two-factor ($k = 2$) factorial experiment.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
A	SS_A	1	$MS_A = SS_A$	$\frac{MS_A}{MS_E}$
B	SS_B	1	$MS_B = SS_B$	$\frac{MS_B}{MS_E}$
Interaction	SS_{AB}	1	$MS_{AB} = SS_{AB}$	$\frac{MS_{AB}}{MS_E}$
Error	SS_E	$2^k(n-1)$	$MS_E = \frac{SS_E}{2^k(n-1)}$	
Total	SS_T	$2^k n - 1$		

7.3.3 Statistical Inference

It can be shown that variance of the effect estimate for a 2^k design is given by

$$var\,(\text{Effect}) = var\left(\frac{\text{Contrast}}{n2^{k-1}}\right) = \frac{\sigma^2}{n2^{k-2}} \tag{7.3.7}$$

Since the variance σ^2 can be estimated by $\hat{\sigma}^2 = MS_E$, the standard error of the effect is then given by

$$s.e.\,(\text{Effect}) = \frac{\hat{\sigma}}{\sqrt{n2^{k-2}}} = \sqrt{\frac{MS_E}{n2^{k-2}}} \tag{7.3.8}$$

Thus, a $100(1-\alpha)\%$ confidence interval for an effect is given by

$$\text{Effect} \pm t_{\alpha/2,2^k(n-1)}\, s.e.\,(\text{Effect}) = \text{Effect} \pm t_{\alpha/2,2^k(n-1)} \frac{\hat{\sigma}}{\sqrt{n2^{k-2}}} = \text{Effect} \pm t_{\alpha/2,2^k(n-1)} \sqrt{\frac{MS_E}{n2^{k-2}}} \tag{7.3.9}$$

The significance of an effect can be tested using the t-statistic

$$t_{\text{ratio}} = \frac{\text{Effect}}{s.e.\,(\text{Effect})} = \frac{(\text{Effect})\sqrt{n2^{k-2}}}{\hat{\sigma}} \tag{7.3.10}$$

or equivalently the F-statistic

$$f_{\text{ratio}} = t^2_{\text{ratio}} = \frac{(\text{Effect})^2\,(n2^{k-2})}{\hat{\sigma}^2} = \frac{MS_{\text{Effect}}}{MS_E} \tag{7.3.11}$$

An effect is considered **significant** at level α if

$$|t_{\text{ratio}}| > t_{\alpha/2,2^k(n-1)} \quad \text{or equivalently if} \quad f_{\text{ratio}} > f_{\alpha,1,2^k(n-1)} \tag{7.3.12}$$

7.3.4 Residual Analysis for for 2^2 design

The regression model used to obtain the predicted values is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{7.3.13}$$

where x_1 represents factor A, x_2 represents factor B, and x_{12} represents the interaction AB. The quantities x_1 and x_2 are sometimes referred to as contrast coefficients and are given by

$$x_1 = \begin{cases} -1 & \text{if } A \text{ is low} \\ +1 & \text{if } A \text{ is high} \end{cases} \quad \text{and} \quad x_2 = \begin{cases} -1 & \text{if } B \text{ is low} \\ +1 & \text{if } B \text{ is high.} \end{cases}$$

The regression coefficients β_0, β_1, β_2, and β_{12} are estimated as follows:

$$\hat{\beta}_0 = \bar{\bar{y}} \qquad \hat{\beta}_1 = \frac{A}{2} \qquad \hat{\beta}_2 = \frac{B}{2} \qquad \hat{\beta}_{12} = \frac{AB}{2}$$

Thus, the residuals are defined as $e_{ij} = y_{ij} - \hat{y}_{ij}$, where the fitted values $\hat{y}_{ij}$ are obtained using the prediction (regression) equation

Regression Equation for 2^2 Design:

$$\hat{y} = \bar{\bar{y}} + \frac{A}{2} x_1 + \frac{B}{2} x_2 + \frac{AB}{2} x_1 x_2 \tag{7.3.14}$$

From Eq. (7.3.12), an effect (or factor) in a 2^2 design is considered **significant** at a level α if $f_{\text{ratio}} > f_{\alpha,1,4(n-1)}$. Thus, any factor that is not significant is excluded from the regression equation. For example, if the interaction AB is not significant, i.e. $f_{\text{ratio}} = MS_{AB}/MS_E \ngtr f_{\alpha,1,4(n-1)}$, then the regression equation becomes $\hat{y} = \bar{\bar{y}} + (A/2)x_1 + (B/2)x_2$.

Similar to the design matrix, we define the contrast matrix X for a 2^2 design as:

$$X = \begin{pmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{pmatrix}$$

Using multiple regression analysis, the regression line may be estimated as $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{y}} = (\bar{y}_1\ \bar{y}_2\ \bar{y}_3\ \bar{y}_4)'$ is the column vector of averages.

```
k = 2; % 2^2 design
M = fracfact('b a ab'); % note the correct order
ybar = mean(data,2); % column vector of averages (data of size 4xn)
X = [ones(2^k,1) M]; % contrast matrix for full model
betahat = regress(ybar,X); % regression coefficients
yhat = X*betahat; % fitted values
```

The residuals can be used to check the normality assumption of the error by displaying the normal probability plot, and also to check the constant error variance by displaying the residual plot against the fitted values or factors.

Example 7.6 *A router is used to cut locating notches on a printed circuit board. The vibration level at the surface of the board as it is cut is considered to be a major source of dimensional variation in the notches. Two factors are thought to influence vibration: bit size (A) and cutting speed (B). Two bit sizes (1/16 and 1/8 inch) and two speeds (40 and 90 rpm) are selected, and four boards are cut at each set of conditions shown below. The response variable is vibration measured as a resultant vector of three accelerometers (x, y, and z) on each test circuit board. The resulting data are given in Table 7.17.*

(i) Draw the design geometry and the interaction plot. Comment on the interaction. What levels of bit size and speed would you recommend for routine operation?

(ii) Calculate the estimate of all factorial effects and the sums of squares by the contrast method.

(iii) Perform an analysis of variance.

TABLE 7.17: Data for the router experiment.

Treatment Combination	Factor A	B	Data Rep 1	Rep 2	Rep 3	Rep 4
(1)	-	-	18.2	18.9	12.9	14.4
a	+	-	27.2	24.0	22.4	22.5
b	-	+	15.9	14.5	15.1	14.2
ab	+	+	41.0	43.9	36.3	39.9

(iv) *Find the regression equation at a significance level* $\alpha = 0.05$.

(v) *Construct a normal probability plot of the residuals, and plot the residuals versus the predicted vibration level. Interpret these plots.*

(vi) *Find a 95% confidence interval for the interaction.*

Solution: This is a 2^k factorial experiment with $k = 2$ and $n = 4$. The data for the router experiment with averages in shown in Table 7.18.

TABLE 7.18: Data for the router experiment with averages.

Treatment Combination	Factor A	B	Data Rep 1	Rep 2	Rep 3	Rep 4	Averages
(1)	-	-	18.2	18.9	12.9	14.4	16.10
a	+	-	27.2	24.0	22.4	22.5	24.02
b	-	+	15.9	14.5	15.1	14.2	14.93
ab	+	+	41.0	43.9	36.3	39.9	40.27

(i) Figure 7.10 depicts the design geometry and the interaction plot for the router experiment. From the interaction plot, we can observe that there is interaction. Note that the large positive effect of speed occurs primarily when bit size is at the high level.

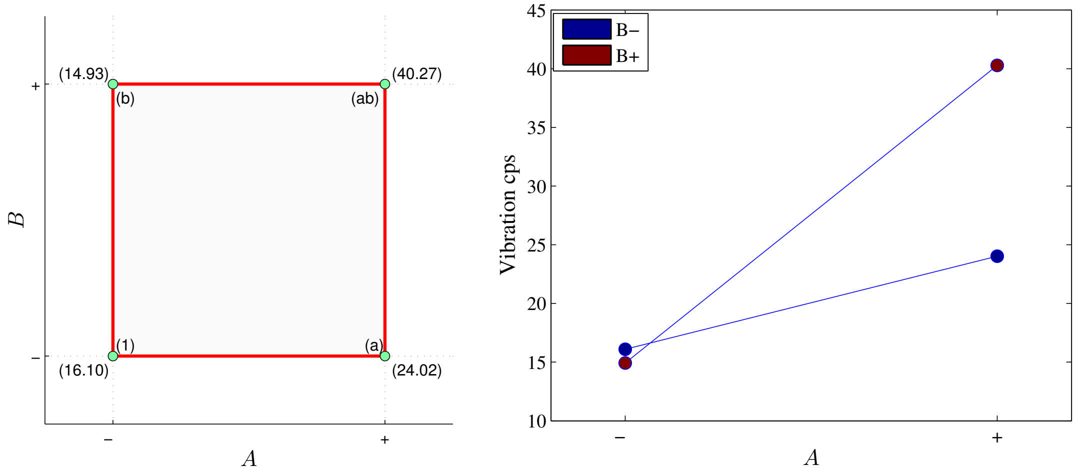


FIGURE 7.10: Router experiment. Left: design geometry; Right: interaction plot.

To reduce the vibration, use the smaller bit. Once the small bit is specified, either speed will work equally well, because the slope of the curve relating vibration to speed for the small tip is approximately zero. The process is robust to speed changes if the small bit is used.

(ii) The effect estimates are given by

$$A = \frac{\text{Contrast}_A}{n2^{k-1}} = \frac{1}{2n}[-(1) + a - b + ab] = 16.64$$
$$B = \frac{\text{Contrast}_B}{n2^{k-1}} = \frac{1}{2n}[-(1) - a + b + ab] = 7.54$$
$$AB = \frac{\text{Contrast}_{AB}}{n2^{k-1}} = \frac{1}{2n}[(1) - a - b + ab] = 8.71$$

Using the formula $SS = n(\text{Effect})^2 2^{k-2}$ for the sum of squares, we have

$$SS_A = nA^2 2^{k-2} = (4)(16.64)^2 = 1107.23$$
$$SS_B = nB^2 2^{k-2} = (4)(7.54)^2 = 227.26$$
$$SS_{AB} = n(AB)^2 2^{k-2} = (4)(8.71)^2 = 303.63$$
$$SS_T = \sum_{i=1}^{2^k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{4}\sum_{j=1}^{4}(y_{ij} - \bar{y})^2 = 1709.83$$

where $\bar{y}$ is the grand mean of all observations y_{ij}, that is

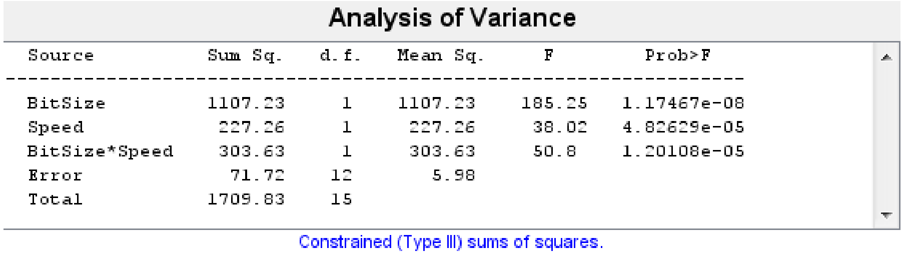
$$\bar{y} = \frac{1}{2^k n}\sum_{i=1}^{2^k}\sum_{j=1}^{n} y_{ij} = \frac{1}{16}\sum_{i=1}^{4}\sum_{j=1}^{4} y_{ij} = 23.83$$

Thus, the error sum of squares is

$$SS_E = SS_{\text{Total}} - SS_A - SS_B - SS_{AB} = 71.72$$

(ii) The MATLAB output for the analysis of variance table is given in Table 7.19.

TABLE 7.19: ANOVA table for the router experiment.



```
Analysis of Variance

  Source            Sum Sq.   d.f.   Mean Sq.     F        Prob>F
-------------------------------------------------------------------
  BitSize           1107.23     1    1107.23    185.25   1.17467e-08
  Speed              227.26     1     227.26     38.02   4.82629e-05
  BitSize*Speed      303.63     1     303.63     50.8    1.20108e-05
  Error               71.72    12       5.98
  Total             1709.83    15

Constrained (Type III) sums of squares.
```

```
k = 2; n = 4;  %two−level experiment with 4 replicates
%data from the router experiment
y = [18.2 18.9 12.9 14.4
     27.2 24.0 22.4 22.5
     15.9 14.5 15.1 14.2
     41.0 43.9 36.3 39.9];
%Use fracfact to find the design matrix (please note the reverse order)
M = fracfact('b a ba'); % order: A B AB
R = repmat(M,n,1); %replicate of design matrix
g1 = R(:,1); g2 = R(:,2);
[p, table, stats] = anovan(y(:), {g1 g2}, 'varnames',{'BitSize' 'Speed'},'model', 'full');
```

(iv) At a significance level $\alpha = 0.05$, all the values of the F-test statistics for factor A, factor B, and interaction AB are greater than $f_{\alpha,1,4(n-1)} = f_{0.05,1,12} = 4.7472$ as shown in ANOVA Table 7.19; that is, $f_{\text{ratio}} > f_{\alpha,1,4(n-1)} = 4.7472$. Thus, the regression equation is given by

$$\hat{y} = \bar{y} + \frac{A}{2}x_1 + \frac{B}{2}x_2 + \frac{AB}{2}x_1x_2 = 23.83 + 8.32x_1 + 3.77x_2 + 4.36x_1x_2$$

For example, for $x_1 = -1$ and $x_2 = -1$ the value of $\hat{y}$ is equal to 16.10.

(v) Figure 7.11 shows the residual plots for the router experiment. The data of normal probability plot appear fairly linear, suggesting that no reason to doubt the normality assumption. Also, the residual plots against the fitted values as well as against the factor levels of A and B exhibit random scatter around 0. Thus, the model assumptions are valid. There is nothing unusual about the residual plots.

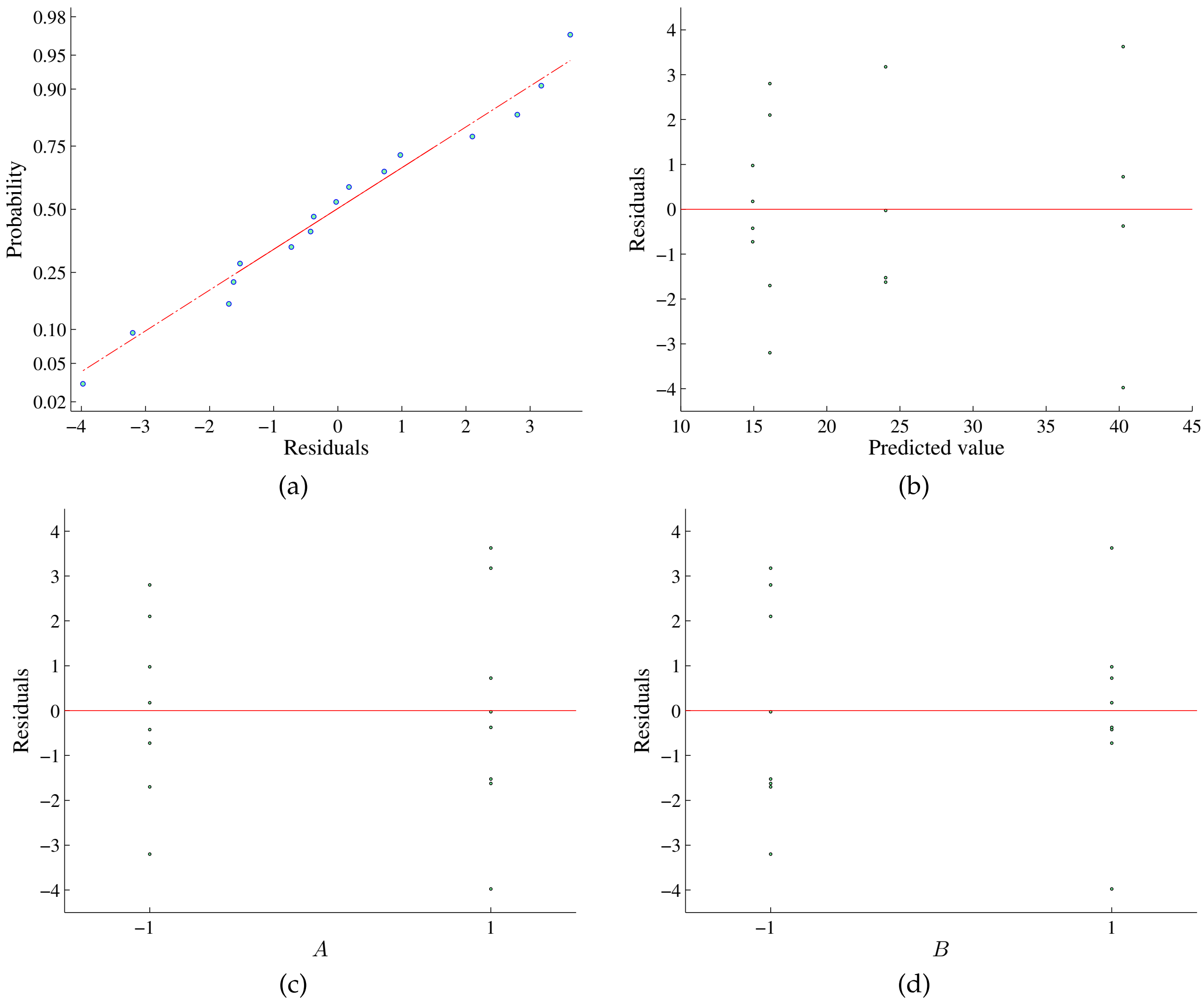


FIGURE 7.11: Residual analysis for the router experiment: (a) Normal probability plot; (b) Residual plot against fitted values; (c) Residual plot against factor A (bit size); (d) Residual plot against factor B (speed).

(vi) The 95% confidence interval for the interaction AB is given by

$$AB \pm t_{\alpha/2, 2^k(n-1)} \sqrt{\frac{MS_E}{n2^{k-2}}} = 8.71 \pm t_{0.025,12} \sqrt{\frac{5.98}{4}} = 8.71 \pm 2.66$$

Thus, the interaction effect is between 6.05 and 11.37.

7.3.5 2^3 Factorial Design

The 2^3 design is a two-level factorial experiment design with three factors A, B and C. This design tests three ($k = 3$) main effects, A, B and C; three ($\binom{k}{2} = \binom{3}{2} = 3$) two-factor interaction effects, AB, BC, AC; and one ($\binom{k}{3} = \binom{3}{3} = 1$) three-factor interaction effect, ABC. The design requires $2^3 = 8$ runs per replicate. The eight treatment combinations corresponding to these runs are (1), a, b, ab, c, ac, bc and abc, which represent the totals of all n observations taken at these design points; that is n observations are made at each of the $2^3 = 8$ treatment combinations. The data layout for a 2^3 design is shown in Table 7.20, where n observations by $y_{ij}, i = 1, \ldots, 8; j = 1, \ldots, n$ are made for each of the $2^3 = 8$ runs.

TABLE 7.20: Data Layout for a 2^3 Design.

Treatment	Factor			Data			
Combination	A	B	C	Rep 1	$\ldots$	Rep n	Averages
(1)	-	-	-	y_{11}	$\ldots$	y_{1n}	$\bar{y}_1$
a	+	-	-	y_{21}	$\ldots$	y_{2n}	$\bar{y}_2$
b	-	+	-	y_{31}	$\ldots$	y_{3n}	$\bar{y}_3$
ab	+	+	-	y_{41}	$\ldots$	y_{4n}	$\bar{y}_4$
c	+	-	+	y_{51}	$\ldots$	y_{5n}	$\bar{y}_5$
ac	+	-	+	y_{61}	$\ldots$	y_{6n}	$\bar{y}_6$
bc	+	+	+	y_{71}	$\ldots$	y_{7n}	$\bar{y}_7$
abc	+	+	+	y_{81}	$\ldots$	y_{8n}	$\bar{y}_8$

The geometry of the 2^3 design can be represented using a using a cube with the eight treatment combinations lying at the eight corners, as shown in Figure 7.12 (right). The contrasts of the factors A, B, C and the interactions AB, AC, BC, and ABC can be calculated using Table 7.13. Similar to the 2^2 design, the contrast of a factor or interaction is obtained by point-wise multiplication of the treatment combinations in the first column of Table 7.13 by the signs in the corresponding factor or interaction column. That is,

Test Matrix

Run	A	B	C
(1)	-	-	-
a	+	-	-
b	-	+	-
ab	+	+	-
c	-	-	+
ac	+	-	+
bc	-	+	+
abc	+	+	+

Design Matrix

$$\begin{array}{cccccccc}
\mathrm{I} & \mathrm{A} & \mathrm{B} & \mathrm{AB} & \mathrm{C} & \mathrm{AC} & \mathrm{BC} & \mathrm{ABC} \\
\left(\begin{array}{c} + \\ + \\ + \\ + \\ + \\ + \\ + \\ + \end{array}\right.
& \begin{array}{c} - \\ + \\ - \\ + \\ - \\ + \\ - \\ + \end{array}
& \begin{array}{c} - \\ - \\ + \\ + \\ - \\ - \\ + \\ + \end{array}
& \begin{array}{c} + \\ - \\ - \\ + \\ + \\ - \\ - \\ + \end{array}
& \begin{array}{c} - \\ - \\ - \\ - \\ + \\ + \\ + \\ + \end{array}
& \begin{array}{c} + \\ - \\ + \\ - \\ - \\ + \\ - \\ + \end{array}
& \begin{array}{c} + \\ + \\ - \\ - \\ - \\ - \\ + \\ + \end{array}
& \left.\begin{array}{c} - \\ + \\ + \\ - \\ + \\ - \\ - \\ + \end{array}\right)
\end{array}$$

Design Geometry

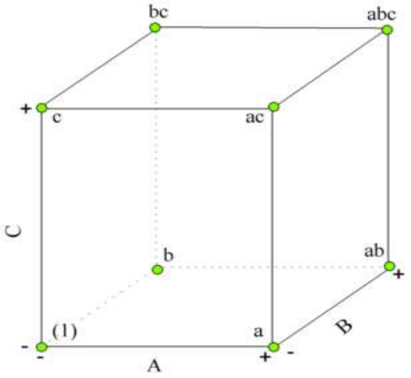


FIGURE 7.12: 2^3 design.

$$\begin{aligned}
\text{Contrast}_A &= -(1) + a - b + ab - c + ac - bc + abc \\
\text{Contrast}_B &= -(1) - a + b + ab - c - ac + bc + abc \\
\text{Contrast}_{AB} &= (1) - a - b + ab + c - ac - bc + abc \\
\text{Contrast}_C &= -(1) - a - b - ab + c + ac + bc + abc \\
\text{Contrast}_{AC} &= (1) - a + b - ab - c + ac - bc + abc \\
\text{Contrast}_{BC} &= (1) + a - b - ab - c - ac + bc + abc \\
\text{Contrast}_{ABC} &= -(1) + a + b - ab + c - ac - bc + abc
\end{aligned}$$

Using Eq. (7.3.4), the effect estimates and the sum of squares for any effect or interaction are computed as follows:

TABLE 7.21: Signs for Contrasts in the 2^3 Design.

Treatment	Factorial Effect						
Combination	A	B	AB	C	AC	BC	ABC
(1)	-	-	+	-	+	+	-
a	+	-	-	-	-	+	+
b	-	+	-	-	+	-	+
ab	+	+	+	-	-	-	-
c	-	-	+	+	-	-	+
ac	+	-	-	+	+	-	-
bc	-	+	-	+	-	+	-
abc	+	+	+	+	+	+	+

Effect and Sum of Squares for 2^3 Design:

$$\begin{aligned} \text{Effect} &= \frac{\text{Contrast}}{4n} \\ SS &= \frac{(\text{Contrast})^2}{8n} = 2n(\text{Effect})^2 \end{aligned} \tag{7.3.15}$$

7.3.6 Residual Analysis for 2^3 Design

The residuals are defined as $e_{ij} = y_{ij} - \hat{y}_{ij}$, where the fitted values $\hat{y}_{ij}$ are obtained using the prediction (regression) equation

Regression Equation for 2^3 Design:

$$\hat{y} = \bar{\bar{y}} + \frac{A}{2}x_1 + \frac{B}{2}x_2 + \frac{AB}{2}x_1x_2 + \frac{C}{2}x_3 + \frac{AC}{2}x_1x_3 + \frac{BC}{2}x_2x_3 + \frac{ABC}{2}x_1x_2x_3 \tag{7.3.16}$$

with contrast coefficients

$$x_1 = \begin{cases} -1 & \text{if } A \text{ is low} \\ +1 & \text{if } A \text{ is high} \end{cases} \quad x_2 = \begin{cases} -1 & \text{if } B \text{ is low} \\ +1 & \text{if } B \text{ is high} \end{cases} \quad \text{and} \quad x_3 = \begin{cases} -1 & \text{if } C \text{ is low} \\ +1 & \text{if } C \text{ is high.} \end{cases}$$

Similar to the design matrix, we define the contrast matrix X for a 2^3 design as follows:

$$X = \begin{pmatrix} +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \end{pmatrix}$$

Using multiple regression analysis, the regression line may be estimated as $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{y}} = (\bar{y}_1\ \bar{y}_2\ \bar{y}_3\ \bar{y}_4\ \bar{y}_5\ \bar{y}_6\ \bar{y}_7\ \bar{y}_8)'$ is the column vector of averages.

```
k = 3; % 2^3 design
M = fracfact('c b cb a ac ab abc'); % note the correct order
ybar = sum(data,2); % column vector of averages (data of size 8xn)
X = [ones(2^k,1) M]; % contrast matrix for full model
betahat = regress(ybar,X); % regression coefficients
yhat = X*betahat; % fitted values
```

The residuals can be used to check the normality assumption of the error by displaying the normal probability plot, and also to check the constant error variance by displaying the residual plot against the fitted values or factors.

Example 7.7 *A quality engineer is studying the surface roughness of a part produced in a metal-cutting operation. Three factors, feed rate (A), depth of cut (B), and tool angle (C), are of interest. All three factors have been assigned two levels, and two replicates of a factorial design are run. The coded surface roughness data are given in Table 7.22.*

TABLE 7.22: Data for the surface roughness experiment.

Treatment	Factor			Data	
Combination	A	B	C	Rep 1	Rep 2
(1)	-	-	-	9	7
a	+	-	-	10	12
b	-	+	-	9	11
ab	+	+	-	12	15
c	-	-	+	11	10
ac	+	-	+	10	13
bc	-	+	+	10	8
abc	+	+	+	16	14

(i) Calculate the estimate of all factorial effects and the sums of squares by the contrast method.

(ii) Perform an analysis of variance.

(iii) Find the regression equation at a significance level $\alpha = 0.10$.

(iv) Construct a normal probability plot of the residuals, and display the residuals plots. Comment on these plots.

Solution: This is a 2^3 factorial design in the factors feed rate (A), depth of cut (B), and tool angle (C), with $n = 2$ replicates. The data for the surface roughness experiment with averages are shown in Table 7.23.

TABLE 7.23: Data for the surface roughness experiment with averages.

Treatment	Factor			Data		
Combination	A	B	C	Rep 1	Rep 2	Averages
(1)	-	-	-	9	7	8.00
a	+	-	-	10	12	11.00
b	-	+	-	9	11	10.00
ab	+	+	-	12	15	13.50
c	-	-	+	11	10	10.50
ac	+	-	+	10	13	11.50
bc	-	+	+	10	8	9.00
abc	+	+	+	16	14	15.00

(i) The effect estimates are given by

$$\begin{aligned}
A &= \frac{\text{Contrast}_A}{n2^{k-1}} = \frac{1}{4n}\left(-(1)+a-b+ab-c+ac-bc+abc\right) = 3.38\\
B &= \frac{\text{Contrast}_B}{n2^{k-1}} = \frac{1}{4n}\left(-(1)-a+b+ab-c-ac+bc+abc\right) = 1.63\\
C &= \frac{\text{Contrast}_C}{n2^{k-1}} = \frac{1}{4n}\left(-(1)-a-b-ab+c+ac+bc+abc\right) = 0.88\\
AB &= \frac{\text{Contrast}_{AB}}{n2^{k-1}} = \frac{1}{4n}\left((1)-a-b+ab+c-ac-bc+abc\right) = 1.37\\
AC &= \frac{\text{Contrast}_{AC}}{n2^{k-1}} = \frac{1}{4n}\left((1)-a+b-ab-c+ac-bc+abc\right) = 0.12\\
BC &= \frac{\text{Contrast}_{BC}}{n2^{k-1}} = \frac{1}{4n}\left((1)+a-b-ab-c-ac+bc+abc\right) = -0.63\\
ABC &= \frac{\text{Contrast}_{ABC}}{n2^{k-1}} = \frac{1}{4n}\left(-(1)+a+b-ab+c-ac-bc+abc\right) = 1.13
\end{aligned}$$

Using the formula $SS = n(\text{Effect})^2 2^{k-2}$ for the sum of squares, we have

$$\begin{aligned}
SS_A &= nA^2 2^{k-2} = 45.56\\
SS_B &= nB^2 2^{k-2} = 10.56\\
SS_C &= nC^2 2^{k-2} = 3.06\\
SS_{AB} &= n(AB)^2 2^{k-2} = 7.56\\
SS_{AC} &= n(AC)^2 2^{k-2} = 0.06\\
SS_{BC} &= n(BC)^2 2^{k-2} = 1.56\\
SS_{ABC} &= n(ABC)^2 2^{k-2} = 5.06\\
SS_T &= \sum_{i=1}^{2^k}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2 = \sum_{i=1}^{8}\sum_{j=1}^{2}(y_{ij}-\bar{\bar{y}})^2 = 92.94
\end{aligned}$$

where $\bar{\bar{y}}$ is the grand mean of all observations y_{ij}, that is

$$\bar{\bar{y}} = \frac{1}{2^k n}\sum_{i=1}^{2^k}\sum_{j=1}^{n} y_{ij} = \frac{1}{16}\sum_{i=1}^{8}\sum_{j=1}^{4} y_{ij} = 11.06$$

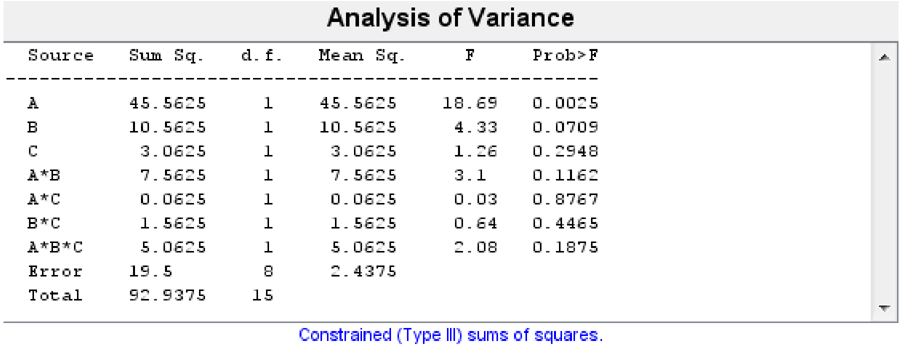
Thus, the error sum of squares is

$$SS_E = SS_{\text{Total}} - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC} - SS_{ABC} = 19.50$$

(ii) The MATLAB output for the analysis of variance table is given in Table 7.24.

```
k = 3; n = 2; %three-level experiment with 2 replicates
%data from the roughness experiment
y = [9 7
      10 12
      9 11
      12 15
      11 10
      10 13
      10 8
      16 14];
%Use fracfact to find the design matrix (Note the reverse order)
M = fracfact('c b cb a ac ab abc'); %order: A B AB C AC BC ABC
R = repmat(M,n,1); %replicate of design matrix
g1 = R(:,1); g2 = R(:,2); g3 = R(:,4); %factors A, B, and C
[p, table, stats] = anovan(y(:), {g1 g2 g3}, 'varnames',{'A' 'B' 'C'},'model', 'full');
```

TABLE 7.24: ANOVA table for the surface roughness 2^3 factorial experiment.



Analysis of Variance

```
 Source      Sum Sq.    d.f.   Mean Sq.      F      Prob>F
----------------------------------------------------------
 A          45.5625      1    45.5625    18.69    0.0025
 B          10.5625      1    10.5625     4.33    0.0709
 C           3.0625      1     3.0625     1.26    0.2948
 A*B         7.5625      1     7.5625     3.1     0.1162
 A*C         0.0625      1     0.0625     0.03    0.8767
 B*C         1.5625      1     1.5625     0.64    0.4465
 A*B*C       5.0625      1     5.0625     2.08    0.1875
 Error      19.5         8     2.4375
 Total      92.9375     15
```

Constrained (Type III) sums of squares.

(iv) The F-ratios for all three main effects and the interactions are shown in ANOVA Table 7.24. At a significance level $\alpha = 0.10$, the critical value for each of these F-ratios is $f_{\alpha,1,2^k(n-1)} = f_{0.10,1,8} = 3.46$. Notice that there is a strong main effect of feed rate (A), since its F-ratio is well into the critical region. However, there is some indication of an effect due to the depth of cut (B). That is, for both factors A and B, we have $f_{\text{ratio}} > f_{\alpha,1,2^k(n-1)} = 3.46$. Thus, the regression equation is given by

$$\hat{y} = \bar{\bar{y}} + \frac{A}{2}x_1 + \frac{B}{2}x_2 = 11.06 + 1.69x_1 + 0.81x_2$$

For example, for $x_1 = -1$ and $x_2 = -1$ the value of $\hat{y}$ is equal to 9.25.

(v) Figure 7.13 shows the normal probability plot for the surface roughness experiment. The data appears to lie approximately along a straight line, suggesting that there is no reason to doubt the normality assumption. Figure 7.14 shows the the residual plots against the fitted values as well as against the factor levels of A, B, and C exhibit random scatter around 0. Thus, the model assumptions are valid. There is nothing unusual about the residual plots.

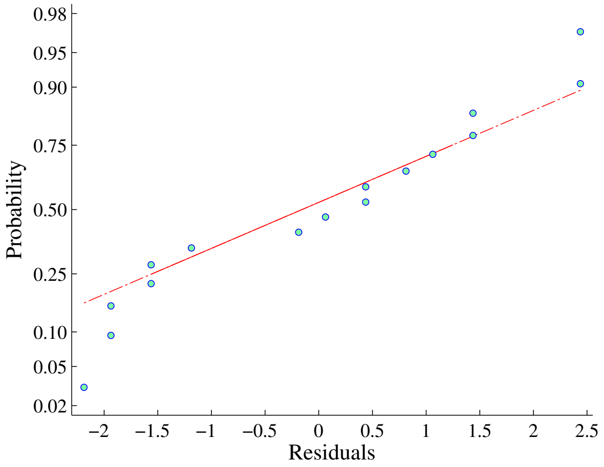
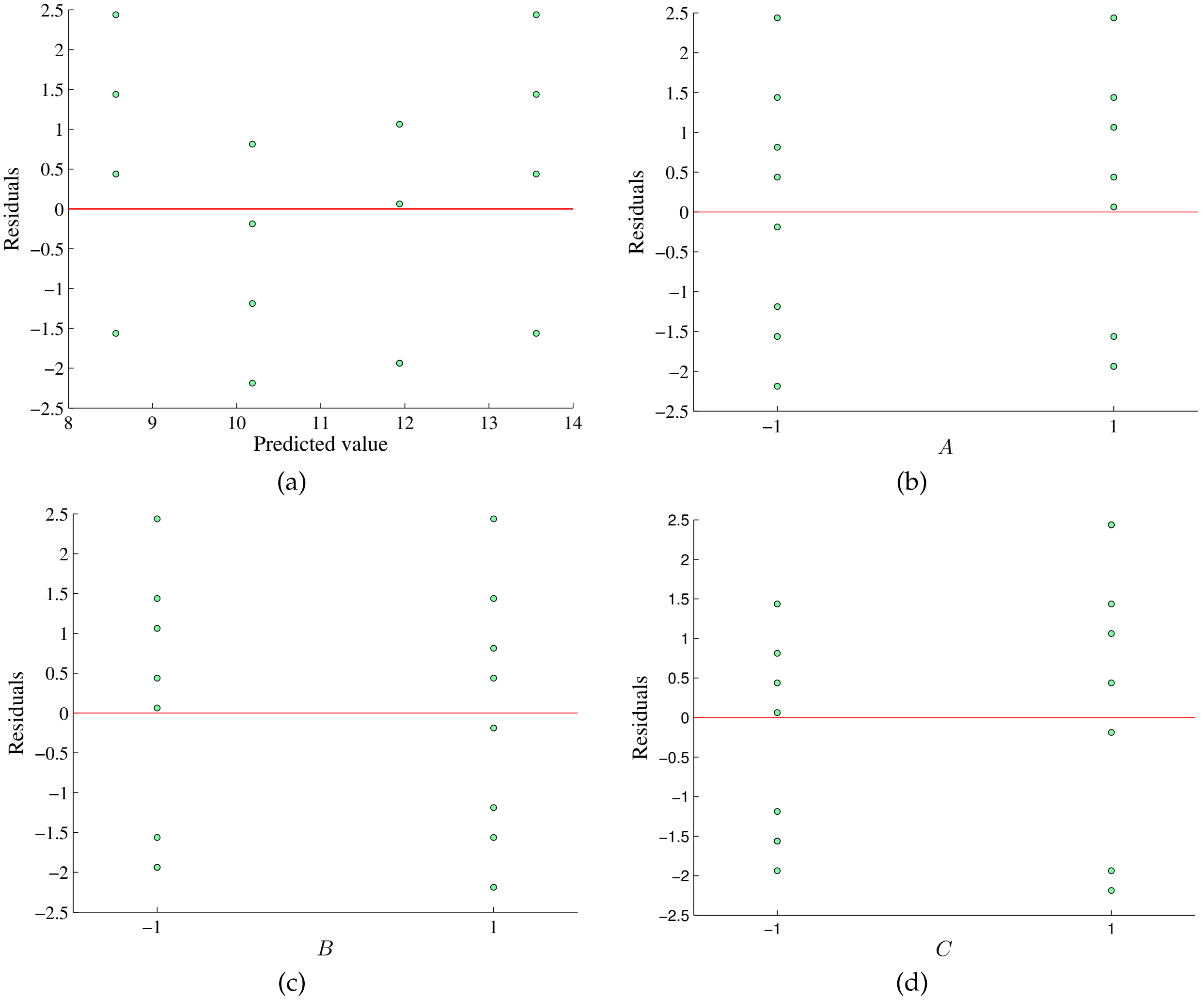


FIGURE 7.13: Normal probability plot for the surface roughness experiment.

Example 7.8 *In the manufacturing industry, the quality of surface finish of graded lumber is an important characteristic. Three factors are to be tested, each at two levels, for their impact on the surface finish. Factor A is the type of wood: oak (level -1) or pine (level 1). Factor B is the rate of feed: 2 m/min (level -1) or 4 m/min (level 1). Factor C is the depth of cut: 1 mm (level -1) and 3 mm (level 1). For each treatment combination, three replications are carried out using a completely randomized design. The resulting surface finish data are given in Table 7.25.*

(i) Calculate the main effects and the interaction effects



(a) (b) (c) (d)

FIGURE 7.14: Residual analysis for the surface roughness experiment: (a) Residual plot against fitted values; (b) Residual plot against factor A; (c) Residual plot against factor B; (d) Residual plot against factor C.

(ii) *Calculate the sums of squares of each of the main effects and interaction effects.*

(iii) *Display the interaction plot, and perform an analysis of variance.*

(iv) *Find the regression equation at a significance level* $\alpha = 0.05$.

(v) *Construct a normal probability plot of the residuals, and display the residuals plots. Comment on these plots.*

Solution: This is a 2^3 factorial design in the factors feed rate (A), depth of cut (B), and tool angle (C), with $n = 2$ replicates. The data for the surface finish experiment with averages are shown in Table 7.26.

(i) Using the formula, Effect = Contrast$/(n2^{k-1})$, the main effects and interaction effects are:

$$A = -2.667, B = 3.333, C = 18.833, AB = -8.833, AC = -3.333, BC = 1.333, ABC = -3.500$$

(ii) Using the formula, $SS = n(\text{Effect})^2 2^{k-2}$ for the sum of squares, we have

$$SS_A = 42.667, SS_B = 66.667, SS_C = 2128.167, SS_{AB} = 468.167, SS_{AC} = 66.667, SS_{BC} = 10.667, SS_{ABC} = 73.5$$

(iii) The interaction plot for the surface finish experiment in displayed in Figure 7.15. Notice that the interaction between A and B is the most significant. The MATLAB output for the analysis of variance table is given in Table 7.27.

TABLE 7.25: Data for the surface finish experiment.

Treatment Combination	Factor			Data		
	A	*B*	*C*	Rep 1	Rep 2	Rep 3
(1)	-	-	-	6	8	9
a	+	-	-	10	16	15
b	-	+	-	18	12	15
ab	+	+	-	12	9	10
c	-	-	+	20	26	29
ac	+	-	+	34	28	32
bc	-	+	+	36	44	46
abc	+	+	+	25	22	24

TABLE 7.26: Data for the surface finish experiment with averages.

Treatment Combination	Factor			Data			Averages
	A	*B*	*C*	Rep 1	Rep 2	Rep 3	
(1)	-	-	-	6	8	9	7.67
a	+	-	-	10	16	15	13.67
b	-	+	-	18	12	15	15.00
ab	+	+	-	12	9	10	10.33
c	-	-	+	20	26	29	25.00
ac	+	-	+	34	28	32	31.33
bc	-	+	+	36	44	46	42.00
abc	+	+	+	25	22	24	23.67

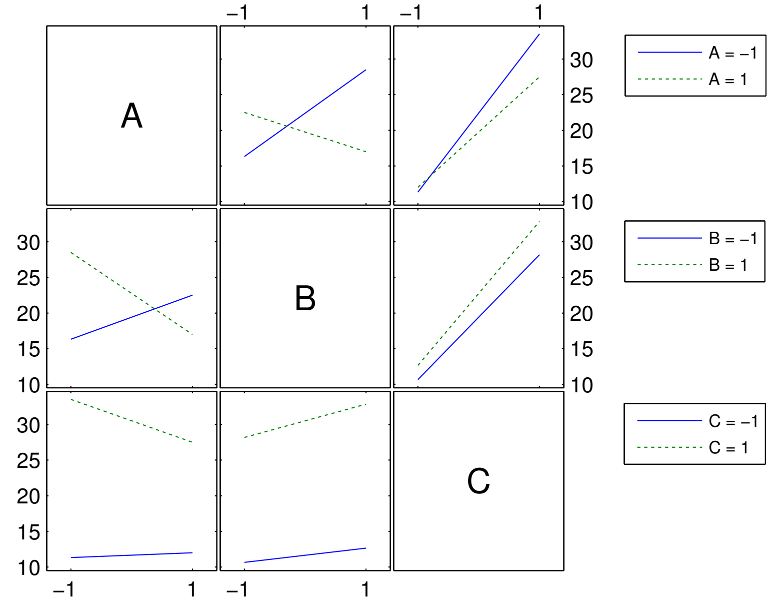
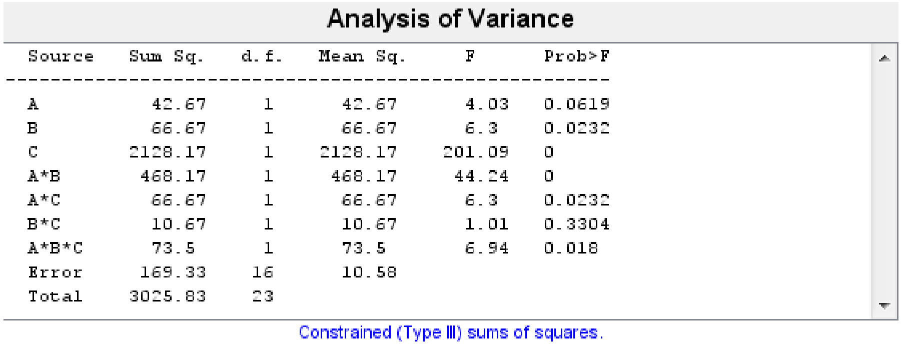


FIGURE 7.15: Interaction plot for the surface finish experiment.

```
k = 3; n = 3; %three−level experiment with 3 replicates
%data from the roughness experiment
y = [6 8 9
      10 16 15
      18 12 15
      12 9 10
```

TABLE 7.27: ANOVA table for the surface finish 2^3 factorial experiment.



Analysis of Variance

```
 Source    Sum Sq.   d.f.   Mean Sq.     F       Prob>F
---------------------------------------------------------
 A           42.67     1      42.67      4.03    0.0619
 B           66.67     1      66.67      6.3     0.0232
 C         2128.17     1    2128.17    201.09    0
 A*B        468.17     1     468.17     44.24    0
 A*C         66.67     1      66.67      6.3     0.0232
 B*C         10.67     1      10.67      1.01    0.3304
 A*B*C       73.5      1      73.5       6.94    0.018
 Error      169.33    16      10.58
 Total     3025.83    23
```

Constrained (Type III) sums of squares.

```
      20 26 29
      34 28 32
      36 44 46
      25 22 24];
%Use fracfact to find the design matrix (Note the reverse order)
M = fracfact('c b cb a ac ab abc'); %order: A B AB C AC BC ABC
R = repmat(M,n,1); %replicate of design matrix
g1 = R(:,1); g2 = R(:,2); g3 = R(:,4); %factors A, B, and C
 [p, table, stats]  = anovan(y(:), {g1 g2 g3}, 'varnames',{'A' 'B' 'C'},'model', 'full');
 figure; interactionplot(y(:),{g1 g2 g3},'varnames',{'A' 'B' 'C'}); %Interaction plot
```

(iv) The F-ratios for all main effects and the interactions are shown in ANOVA Table 7.27. At a significance level $\alpha = 0.05$, the critical value for each of these F-ratios is $f_{\alpha,1,2^k(n-1)} = f_{0.05,1,16} = 4.494$. From the ANOVA table, it can be seen that the F-ratio of B, C, AB, AC, and ABC are greater than the critical value, i.e. $f_{\text{ratio}} > f_{\alpha,1,2^k(n-1)} = 4.494$. Thus, the regression equation is given by

$$\begin{aligned}\hat{y} &= \bar{\bar{y}} + \frac{A}{2}x_1 + \frac{B}{2}x_2 + \frac{AB}{2}x_1x_2 + \frac{C}{2}x_3 + \frac{AC}{2}x_1x_3 + \frac{BC}{2}x_2x_3 + \frac{ABC}{2}x_1x_2x_3 \\ &= 21.088 - 1.333x_1 + 1.667x_2 - 4.417x_1x_2 + 9.417x_3 - 1.667x_1x_3 + 0.667x_2x_3 - 1.750x_1x_2x_3\end{aligned}$$

Notice that the factors A and C are also included in the regression model to preserve the hierarchy of the model. For example, for $x_1 = -1$ and $x_2 = -1$ the value of $\hat{y}$ is equal to 7.667.

(v) Figure 7.16 shows the normal probability plot for the surface finish experiment. The data appears to lie approximately along a straight line, suggesting that there is no reason to doubt the normality assumption. Figure 7.17 shows the the residual plots against the fitted values as well as against the factor levels of A, B, and C exhibit random scatter around 0. Thus, the model assumptions are valid. There is nothing unusual about the residual plots.

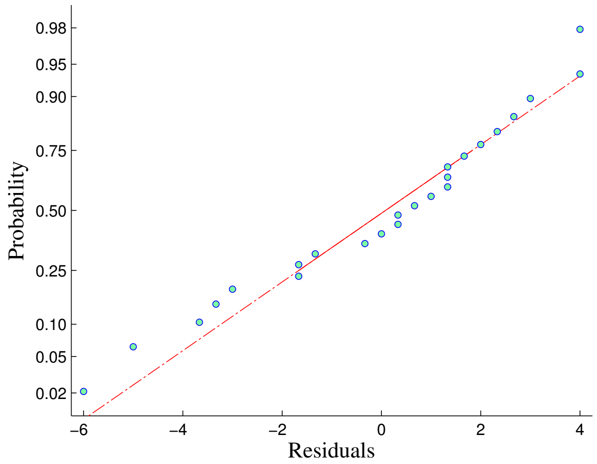
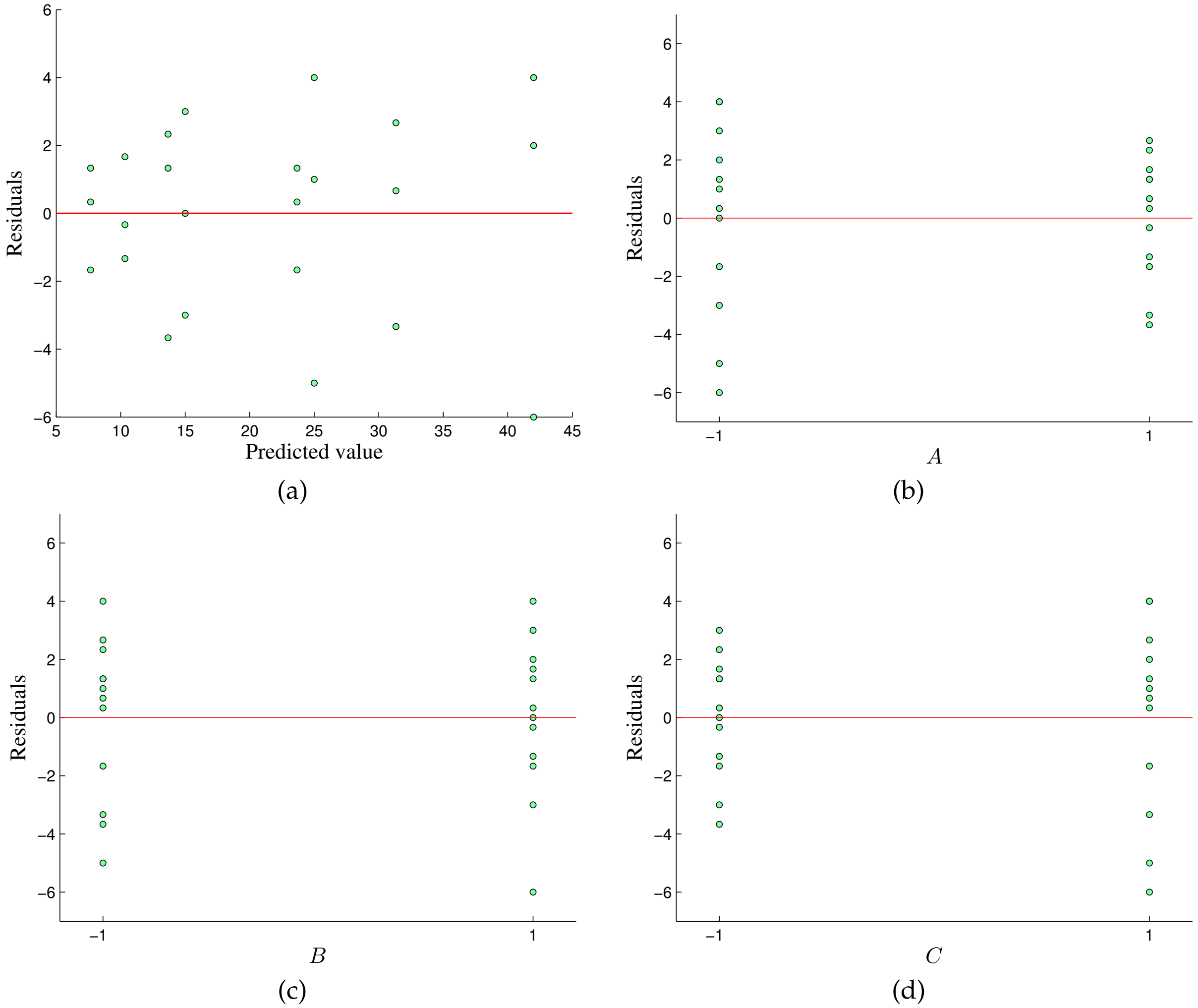


FIGURE 7.16: Normal probability plot for the surface finish experiment.



(a)

(b)

(c)

(d)

FIGURE 7.17: Residual analysis for the surface finish experiment: (a) Residual plot against fitted values; (b) Residual plot against factor A; (c) Residual plot against factor B; (d) Residual plot against factor C.

7.4 Problems

❶ In an experiment conducted to determine which of 3 different missile systems is preferable, the propellant burning rate for 24 static firings was measured. Four different propellant types were used. The experiment yielded duplicate observations of burning rates at each combination of the treatments. The data from this experiment are given in Table 7.28.

a) Write an appropriate statistical model for this experiment

b) Set up an analysis of variance table and conduct the appropriate F-tests on main effects and interactions using a significance level $\alpha = 0.05$

c) Construct the interaction plot and draw conclusions.

d) Perform the residual analysis and draw conclusions.

Table 7.28: Data for the two-factor factorial experiment of propellant burning rates.

Missile System	Propellant Type			
	b_1	b_2	b_3	b_4
a_1	34.0, 32.7	30.1, 32.8	29.8, 26.7	29.0, 28.9
a_2	32.0, 33.2	30.2, 29.8	28.7, 28.1	27.6, 27.8
a_3	28.4, 29.3	27.3, 28.9	29.7, 27.3	28.8, 29.1

❷ An electrical engineer is investigating a plasma etching process used in semiconductor manufacturing. It is of interest to study the effects of two factors, the C_2F_6 gas flow rate (A) and the power applied to the cathode (B). The response is the etch rate. Each factor is run at 3 levels, and 2 experimental runs on etch rate are made for each of the 9 combinations. The setup is that of a completely randomized design. The data from this experiment are given in Table 7.29.

a) Write an appropriate statistical model for this experiment

b) Set up an analysis of variance table and conduct the appropriate F-tests on main effects and interactions using a significance level $\alpha = 0.05$

c) Construct the interaction plot and draw conclusions.

d) Perform the residual analysis and draw conclusions.

Table 7.29: Data for the two-factor factorial experiment of plasma etching process.

C_2F_6 Flow Rate	Power Supplied		
	1	2	3
1	288, 360	488, 465	670, 720
2	385, 411	482, 521	692, 724
3	488, 462	595, 612	761, 801

❸ A manufacturer wishes to determine the effectiveness of four types of machines (A, B, C, and D) in the production of bolts. To accomplish this, the numbers of defective bolts produced by each machine in the days of a given week are obtained for each of two shifts; the results are shown in Table 7.30.

a) Write an appropriate statistical model for this experiment

b) Set up an analysis of variance table and conduct the appropriate F-tests on main effects and interactions using a significance level $\alpha = 0.05$

c) Construct the interaction plot and draw conclusions.

d) Perform the residual analysis and draw conclusions.

TABLE 7.30: Data for the two-factor factorial experiment of number of defective bolts.

Machine	First Shift					Second Shift				
Type	Mon.	Tue.	Wed.	Thu.	Fri.	Mon.	Tue.	Wed.	Thu.	Fri.
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

❹ A basic processing step in the integrated circuit manufacturing industry is to grow an epitaxial layer on polished silicon wafers. The wafers are mounted on a susceptor and positioned inside a bell jar. Chemical vapors are introduced through nozzles near the top of the jar. The susceptor is rotated, and heat is applied. These conditions are maintained until the epitaxial layer is thick enough. Table 7.31 presents the results of a 2^2 factorial design with $n = 4$ replicates using the factors A = deposition time and B = arsenic flow rate. The two levels of deposition time are $-$ = short and $+$ = long, and the two levels of arsenic flow rate are $-$ = 55% and $+$ = 59%. The response variable is epitaxial layer thickness (μm).

TABLE 7.31: Data for the epitaxial process experiment.

Treatment	Factor		Data			
Combination	A	B	Rep 1	Rep 2	Rep 3	Rep 4
(1)	-	-	14.037	14.165	13.972	13.907
a	+	-	14.821	14.757	14.843	14.878
b	-	+	13.880	13.860	14.032	13.914
ab	+	+	14.888	14.921	14.415	14.932

a) Draw the design geometry and the interaction plot. Comment on the interaction.

b) Calculate the estimate of all factorial effects and the sums of squares by the contrast method.

c) Perform an analysis of variance.

d) Find the regression equation at a significance level $\alpha = 0.05$.

e) Construct a normal probability plot of the residuals, and plot the residuals versus the predicted value as well as versus the main factors. Interpret these plots.

f) Find a 95% confidence interval for the interaction.

❺ Consider an investigation into the effect of the concentration of the reactant and the amount of the catalyst on the conversion (yield) in a chemical process. The objective of the experiment was to determine if adjustments to either of these two factors would increase the yield. The factor reactant concentration, denoted by A, has two levels, 15 and 25 percent. The other factor catalyst, denoted by B, has also two levels, 2 and 1 pounds. The experiment is replicated three times. The order in which the runs are made is random. The response variable data are given in Table 7.31.

TABLE 7.32: Data for the catalyst experiment.

Treatment	Factor		Data		
Combination	A	B	Rep 1	Rep 2	Rep 3
(1)	-	-	25	28	27
a	+	-	36	32	32
b	-	+	18	19	23
ab	+	+	29	30	33

a) Draw the design geometry and the interaction plot. Comment on the interaction.

b) Calculate the main effects and the interaction effects

c) Calculate the sums of squares of each of the main effects and interaction effects.

d) Perform an analysis of variance.

e) Find the regression equation at a significance level $\alpha = 0.05$.

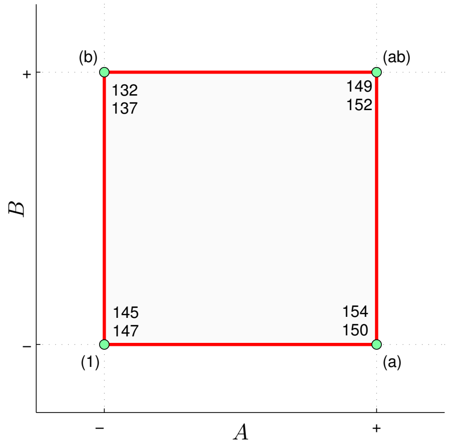
f) Calculate the residuals using the regression model at $\alpha = 0.05$. Then, display all the residuals in one single boxplot. Are there any outliers?

g) Construct a normal probability plot of the residuals, and plot the residuals versus the predicted value as well as versus the main factors. Interpret these plots.

h) Find a 95% confidence interval for each main factor.

i) Find a 95% confidence interval for the interaction.

❻ It is important to study the effect of the concentration of the reactant and the feed rate on the viscosity of the product from a chemical process. Let the reactant concentration be factor A, at levels 15% and 25%. Let the feed rate be factor B, with levels 20 lb/hr and 30 lb/hr. The experiment involves two experimental runs at each of the four combinations ($-$ = low and $+$ = high). The viscosity readings are shown in Figure 7.18, i.e. the data given inside the square or in the table.



Treatment	Factor		Data	
Combination	A	B	Rep 1	Rep 2
(1)	-	-	145	147
a	+	-	154	150
b	-	+	132	137
ab	+	+	149	152

FIGURE 7.18: Data for the viscosity experiment.

a) Draw the design geometry and the interaction plot. Comment on the interaction.

b) Calculate the main effects and the interaction effects

c) Calculate the sums of squares of each of the main effects and interaction effects.

d) Perform an analysis of variance.

e) Find the regression equation at a significance level $\alpha = 0.05$.

f) Construct a normal probability plot of the residuals, and plot the residuals versus the predicted value as well as versus the main factors. Interpret these plots.

g) Find a 95% confidence interval for each main factor.

h) Find a 95% confidence interval for the interaction.

❼ In an experiment conducted to study a particular filtering system for coal, a coagulant was added to a solution in a tank containing coal and sludge, which was then placed in a recirculation system in order that the coal could be washed. Three factors were varied in the experimental process:

Factor A:	percent solids circulated initially in the overflow
Factor B:	flow rate of the polymer
Factor C:	pH of the tank

The amount of solids in the underflow of the cleansing system determines how clean the coal has become. Two levels of each factor were used and two experimental runs were made for each of the $2^3 = 8$ treatment combinations. The response measurements in percent solids by weight in the underflow of the circulation system are given in Table 7.33.

TABLE 7.33: Data for the coal experiment.

Treatment	Factor			Data	
Combination	A	B	C	Rep 1	Rep 2
(1)	-	-	-	4.65	5.81
a	+	-	-	21.42	21.35
b	-	+	-	12.66	12.56
ab	+	+	-	18.27	16.62
c	-	-	+	7.93	7.88
ac	+	-	+	13.18	12.87
bc	-	+	+	6.51	6.26
abc	+	+	+	18.23	17.83

(i) Calculate the estimate of all factorial effects and the sums of squares by the contrast method.

(ii) Perform an analysis of variance.

(iii) Find the regression equation at a significance level $\alpha = 0.10$.

(iv) Construct a normal probability plot of the residuals, and plot the residuals versus the predicted value as well as versus the main factors. Interpret these plots.

❽ In the search for lower-pollution synthetic fuel, researchers are experimenting with three different factors, each controlled at two levels, for the processing of such a fuel. Factor A is the concentration of corn extract at 5% and 10%, factor B is the concentration of an ethylene-based compound at 15% and 25%, and factor C is the distillation temperature at 120°C and 150°C . The levels of undesirable emission of the fuel are are given in Table 7.34 for three replications of each treatment; each level is randomly assigned to a treatment. The larger the level of emission, the worse the impact on the environment.

TABLE 7.34: Data for the emission level experiment.

Treatment	Factor			Data		
Combination	A	B	C	Rep 1	Rep 2	Rep 3
(1)	-	-	-	30	24	26
a	+	-	-	18	22	24
b	-	+	-	30	32	25
ab	+	+	-	43	47	41
c	-	-	+	28	24	22
ac	+	-	+	54	49	46
bc	-	+	+	58	48	50
abc	+	+	+	24	20	22

a) Calculate the main effects and the interaction effects

b) Calculate the sums of squares of each of the main effects and interaction effects.

c) Display the interaction plot, and perform an analysis of variance.

d) Find the regression equation at 5% level of significance. Which factors are significant?

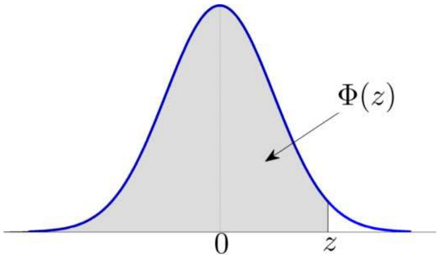
e) Construct a normal probability plot of the residuals, and plot the residuals versus the predicted value as well as versus the main factors. Interpret these plots.

7.5 REFERENCES

[1] D.C. Montgomery, *Introduction to Statistical Quality Control*, Wiley, 6th Edition, 2009.

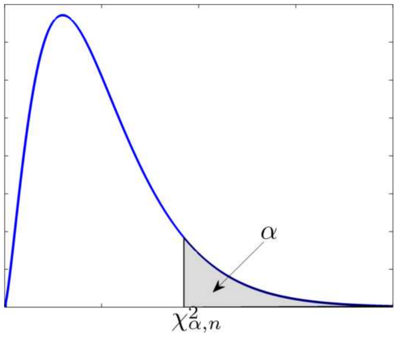
[2] I. Bass and B. Lawton, *Lean Six Sigma using SigmaXL and Minitab*, McGraw-Hill Professional, 1st Edition, 2009.

Table I: Cumulative Standard Normal Distribution



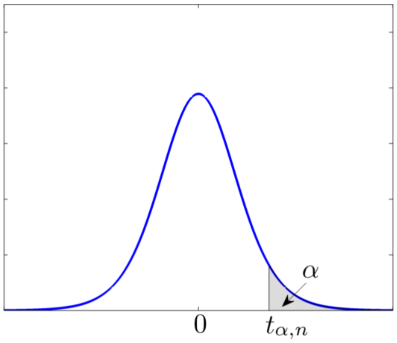
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.53292	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55576	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67365	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76731	0.77035	0.77337	0.77637	0.77935	0.78231	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84135	0.84375	0.84614	0.84850	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89616	0.89796	0.89973	0.90148
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95544	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96637	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97671
2.0	0.97725	0.97778	0.97831	0.97882	0.97933	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99491	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99586	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99745	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99860
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99967	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99982	0.99982	0.99983	0.99984
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Table II: Percentage Points $\chi^2_{\alpha,n}$



	α								
n	0.99	0.975	0.95	0.90	0.50	0.10	0.05	0.025	0.01
1	0.0002	0.001	0.004	0.02	0.45	2.71	3.84	5.02	6.63
2	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21
3	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34
4	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28
5	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09
6	0.87	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81
7	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48
8	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09
9	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67
10	2.56	3.24	3.94	4.87	9.34	15.99	18.31	20.48	23.21
11	3.05	3.81	4.57	5.58	10.34	17.28	19.68	21.92	24.72
12	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22
13	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69
14	4.66	5.62	6.57	7.79	13.34	21.06	23.68	26.12	29.14
15	5.23	6.26	7.26	8.55	14.34	22.31	25.00	27.49	30.58
16	5.81	6.90	7.96	9.31	15.34	23.54	26.30	28.85	32.00
17	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41
18	7.01	8.23	9.39	10.86	17.34	25.99	28.87	31.53	34.81
19	7.63	8.90	10.12	11.65	18.34	27.20	30.14	32.85	36.19
20	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57
21	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93
22	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29
23	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64
24	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98
25	11.52	13.11	14.61	16.47	24.34	34.38	37.65	40.65	44.31
26	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64
27	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96
28	13.56	15.30	16.93	18.94	27.34	37.92	41.34	44.46	48.28
29	14.26	16.04	17.71	19.77	28.34	39.09	42.56	45.72	49.59
30	14.95	16.78	18.49	20.60	29.34	40.26	43.77	46.98	50.89
40	22.16	24.42	26.51	29.05	39.34	51.81	55.76	59.34	63.69
50	29.71	32.35	34.76	37.69	49.33	63.17	67.50	71.42	76.15
60	37.48	40.47	43.19	46.46	59.33	74.40	79.08	83.30	88.38
70	45.44	48.75	51.74	55.33	69.33	85.53	90.53	95.02	100.43
80	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33
90	61.75	65.64	69.13	73.29	89.33	107.57	113.15	118.14	124.12
100	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81

TABLE III: PERCENTAGE POINTS $t_{\alpha,n}$

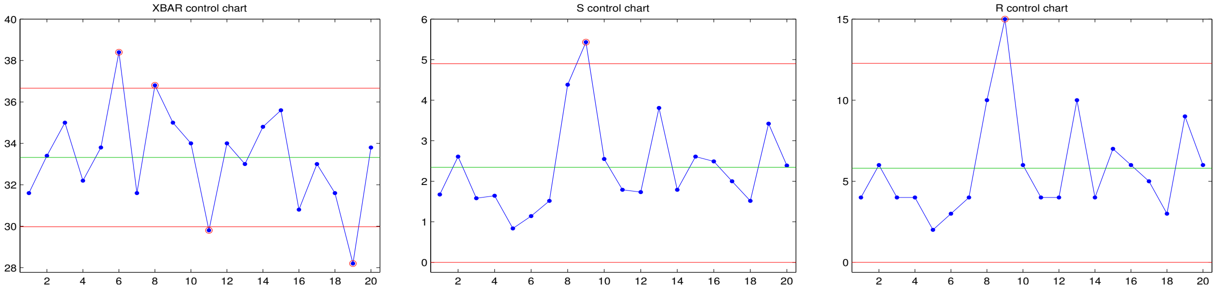


	α							
n	0.25	0.10	0.05	0.025	0.01	0.0833	0.00625	0.005
1	1.000	3.078	6.314	12.706	31.821	38.204	50.923	63.657
2	0.816	1.886	2.920	4.303	6.965	7.649	8.860	9.925
3	0.765	1.638	2.353	3.182	4.541	4.857	5.392	5.841
4	0.741	1.533	2.132	2.776	3.747	3.961	4.315	4.604
5	0.727	1.476	2.015	2.571	3.365	3.534	3.810	4.032
6	0.718	1.440	1.943	2.447	3.143	3.287	3.521	3.707
7	0.711	1.415	1.895	2.365	2.998	3.128	3.335	3.499
8	0.706	1.397	1.860	2.306	2.896	3.016	3.206	3.355
9	0.703	1.383	1.833	2.262	2.821	2.933	3.111	3.250
10	0.700	1.372	1.812	2.228	2.764	2.870	3.038	3.169
11	0.697	1.363	1.796	2.201	2.718	2.820	2.981	3.106
12	0.695	1.356	1.782	2.179	2.681	2.779	2.934	3.055
13	0.694	1.350	1.771	2.160	2.650	2.746	2.896	3.012
14	0.692	1.345	1.761	2.145	2.624	2.718	2.864	2.977
15	0.691	1.341	1.753	2.131	2.602	2.694	2.837	2.947
16	0.690	1.337	1.746	2.120	2.583	2.673	2.813	2.921
17	0.689	1.333	1.740	2.110	2.567	2.655	2.793	2.898
18	0.688	1.330	1.734	2.101	2.552	2.639	2.775	2.878
19	0.688	1.328	1.729	2.093	2.539	2.625	2.759	2.861
20	0.687	1.325	1.725	2.086	2.528	2.613	2.744	2.845
21	0.686	1.323	1.721	2.080	2.518	2.601	2.732	2.831
22	0.686	1.321	1.717	2.074	2.508	2.591	2.720	2.819
23	0.685	1.319	1.714	2.069	2.500	2.582	2.710	2.807
24	0.685	1.318	1.711	2.064	2.492	2.574	2.700	2.797
25	0.684	1.316	1.708	2.060	2.485	2.566	2.692	2.787
26	0.684	1.315	1.706	2.056	2.479	2.559	2.684	2.779
27	0.684	1.314	1.703	2.052	2.473	2.552	2.676	2.771
28	0.683	1.313	1.701	2.048	2.467	2.546	2.669	2.763
29	0.683	1.311	1.699	2.045	2.462	2.541	2.663	2.756
30	0.683	1.310	1.697	2.042	2.457	2.536	2.657	2.750
40	0.681	1.303	1.684	2.021	2.423	2.499	2.616	2.704
60	0.679	1.296	1.671	2.000	2.390	2.463	2.575	2.660
120	0.677	1.289	1.658	1.980	2.358	2.428	2.536	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.394	2.498	2.576

Table IV: Percentage Points f_{α,n_1,n_2} for $\alpha = 0.05$

	n_1																
n_2	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60
1	161.50	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	243.90	246.00	248.00	249.30	250.10	251.10	252.20
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.57
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.43
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.74
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.30
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.01
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.79
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.62
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.49
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.38
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.22
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.16
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.11
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.06
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.02
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.95
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.89
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.86
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.80
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.79
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.75
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.74
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.64
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.53
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.50	1.43
∞	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.51	1.46	1.39	1.32

Table V: Control Chart Factors



Sample	Factors for Control Limits								
	$\overline{X}$-chart		S-chart			R-chart			
Number (n)	A_2	A_3	c_4	B_3	B_4	d_2	d_3	D_3	D_4
2	1.880	2.659	0.798	0.000	3.267	1.128	0.853	0.000	3.267
3	1.023	1.954	0.886	0.000	2.568	1.693	0.888	0.000	2.575
4	0.729	1.628	0.921	0.000	2.266	2.059	0.880	0.000	2.282
5	0.577	1.427	0.940	0.000	2.089	2.326	0.864	0.000	2.114
6	0.483	1.287	0.952	0.030	1.970	2.534	0.848	0.000	2.004
7	0.419	1.182	0.959	0.118	1.882	2.704	0.833	0.076	1.924
8	0.373	1.099	0.965	0.185	1.815	2.847	0.820	0.136	1.864
9	0.337	1.032	0.969	0.239	1.761	2.970	0.808	0.184	1.816
10	0.308	0.975	0.973	0.284	1.716	3.078	0.797	0.223	1.777
11	0.285	0.927	0.975	0.321	1.679	3.173	0.787	0.256	1.744
12	0.266	0.886	0.978	0.354	1.646	3.258	0.778	0.283	1.717
13	0.249	0.850	0.979	0.382	1.618	3.336	0.770	0.307	1.693
14	0.235	0.817	0.981	0.406	1.594	3.407	0.763	0.328	1.672
15	0.223	0.789	0.982	0.428	1.572	3.472	0.756	0.347	1.653
16	0.212	0.763	0.984	0.448	1.552	3.532	0.750	0.363	1.637
17	0.203	0.739	0.985	0.466	1.534	3.588	0.744	0.378	1.622
18	0.194	0.718	0.985	0.482	1.518	3.640	0.739	0.391	1.609
19	0.187	0.698	0.986	0.497	1.503	3.689	0.733	0.404	1.596
20	0.180	0.680	0.987	0.510	1.490	3.735	0.729	0.415	1.585
21	0.173	0.663	0.988	0.523	1.477	3.778	0.724	0.425	1.575
22	0.167	0.647	0.988	0.534	1.466	3.819	0.720	0.435	1.565
23	0.162	0.633	0.989	0.545	1.455	3.858	0.716	0.443	1.557
24	0.157	0.619	0.989	0.555	1.445	3.895	0.712	0.452	1.548
25	0.153	0.606	0.990	0.565	1.435	3.931	0.708	0.459	1.541