МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Natalia Shyriaieva

STATISTICS. BASIC PRINCIPLES

**Lecture notes on Statistics Course
for students of bachelor level in 6.030601 Management
and 6.030508 Finance and credit**

Наталя Ширяєва

СТАТИСТИКА. ОСНОВНІ ПРИНЦИПИ

**Текст лекцій з курсу «Статистика»
для студентів напрямів 6.030601 «Менеджмент»
та 6.030508 «Фінанси і кредит»**

Затверджено
редакційно-видавничою
радою університету,
протокол № 1 від 20.03.2015 р.

Харків
НТУ "ХПІ"
2015

УДК 658: 519.2 = 111(072)
ББК 65.290 – 2 + 60.681.2 Англ – 923
Ш 64

Рецензенти: *М.М. Шевченко*, к.е.н, доц. НТУ "ХПІ"
В.О. Шведун, к.е.н, с.н.с. НУЦЗУ

Текст лекцій містить основи загальної теорії статистики, включаючи групування статистичних даних, абсолютні, відносні і середні величини, статистичні розподіли, вибіркове спостереження, кореляційно-регрессійний аналіз, оцінювання, ряди динаміки, індекси та їх використання в економіко-статистичних дослідженнях. Представлено типові приклади з розв'язаннями за матеріалом, що вивчається.
Призначено для студентів напрямів 6.030601 «Менеджмент» та 6.030508 «Фінанси і кредит»

Ш 64 **Shyriaieva N.** Statistics. Basic principles. Lecture notes on Statistics Course for students of bachelor level in 6.030601 Management and 6.030508 Finance and credit / N.Shyriaieva. – Kharkiv: NTU "KhPI", 2015. – 162 p.

ISBN

The text contains lecture notes of the general theory of statistics, including clustering statistics, absolute, relative and average values, organization and grouping of data, sampling, correlation and regression analysis, estimation, time series, indexes and their use in economics. Typical examples of solving the material are presented in the text.

Designed for students of bachelor level in 6.030601 "Management" and 6.030508 "Finance and Credit"

Іл. 51. Табл. 44. Бібліогр. 14 назв.

УДК 658: 519.2 = 111(072)
ББК 65.290 – 2 + 60.681.2 Англ – 923

ISBN

CONTENTS

	Preface ……..	5
1	The Role of Statistics……	6
	1.1 The Importance of Statistics……..	6
	1.2 Statistical Observation: Forms and Types……	7
	1.3 Basic Definitions……..	8
	1.4 The Importance of Sampling……	11
	1.5 The Functions of Statistics……...	11
	1.6 Levels of Measurement……	12
2	Describing Data Sets. Organization and Grouping……..	14
	2.1 Introduction……	14
	2.2 Methods of Organizing Data……..	14
	2.3. Pictorial Displays……	22
	2.4 Stem – And – Leaf Designs……	28
3	Measures of Central Tendency and Dispersion……..	30
	3.1 Introduction……	30
	3.2 Measures of Central Tendency for Ungrouped Data……	30
	3.3. Measures of Central Tendency for Ungrouped Data……	34
	3.4 Selecting the Appropriate Measure of Central Tendency……	36
	3.5 Measures of Dispersion for Ungrouped Data……..	37
	3.6. Calculating the Variance and Standard Deviation with Grouped Data	41
	3.7 Other Measures of Dispersion……	42
	3.8 Common Uses for the Standard Deviation……..	44
4	Probability Distributions ……	49
	4.1 Principles of Probability……	49
	4.2. Probability Distributions……..	55
5	Sampling Distributions: An Introduction to Inferential Statistics……..	63
	5.1 Introduction……	63
	5.2 The Central Limit Theorem……	67
	5.3 Using the Sample Distribution……	67
	5.4 Types of Samples…….	69
6	Estimation……	71
	6.1 Introduction……	71
	6.2 Confidence Intervals for the Population Mean – Large Samples….	73
	6.3 Confidence Intervals for the Population Mean – Small Samples….	75
	6.4 Confidence Intervals for Population Proportions……	78
	6.5 Controlling the Interval Width……	80
	6.6 Determining the Sample Size……	82
	6.7 Properties of Good Estimators……	85

7 Simple Regression and Correlation Analysis …… 87
7.1 Introduction…….. 87
7.2 The Basic Objective of Regression Analysis…… 89
7.3 Ordinary Least Squares Method (the line of best fit).…….. 91
7.4 The Standard Error of the Estimate: A Measure of Goodness-of-Fit 96
7.5 Correlation Analysis…… 97
7.6 Interval Estimation in Regression Analysis…… 100
8 Time Series Analysis and Forecasting…… 107
8.1 Introduction…… 107
8.2 Time Series and Their Components……. 107
8.3 Time-Series Models…… 111
8.4 Smoothing Techniques…… 112
8.5 Decomposition of a Time Series…… 118
9 Index Numbers…… 128
9.1 Introduction…… 128
9.2 A Simple Price Index…… 129
9.3 Composite Price Indexes…… 131
9.4 Weighted Composited Price Indexes……. 132
9.5 Average of Relatives Method…… 135
9.6 Selection of the Base Period…… 137
9.7 Specific Indexes…….. 139
9.8 Uses for the CPI…….. 141
10 Test your Knowledge…… 145
Literature …… 159
Appendix Distribution Tables …… 160

PREFACE

Our world grows in complexity. People often forced to make decisions on one or another problem in the presence of considerable uncertainty. Yet solutions to these problems are essential to our well-being and even our ultimate survival. We are also continually pressured by distressing economic problems such as raging inflation, a cumbersome tax system, and excessive swings in the business cycle. Our entire social and economic fabric is threatened by environmental pollution, a burdensome public debt, an ever-increasing crime rate, and unpredictable interest rates.

Statistics provides collection, processing and analysis of massive social and economic phenomena that characterize all aspects of life and activities of the population, shows the relationship of various aspects of the economy, studying the dynamics of development and effective management decision making at all levels.

The main goal of this text is to present basic concepts in elementary statistics as usually Statistics is a frightful subject for most students. The text is written for nonstatisticians and can be used by students in all disciplines. The "Chapter checklist" sections at the end of each chapter and the “Test your Knowledge” chapter at the end of text allow students to build confidence by working problems related to the relevant concepts. Examples are taken from a wide variety of disciplines that emphasize the concepts and are to the point.

1 THE ROLE OF STATISTICS

1.1 The Importance of Statistics

From the word “stato” – state or country. Statistics has mentioned first in Ancient Rome and China as some numerical calculations. In 17-18 centuries the so called “countryknolegment” was the subject of statistics.

Virtually every area of serious scientific inquiry can benefit from statistical analysis. For the economic policymaker who must advise the president and other public officials on proper economic procedure? Statistics has proven to be an invaluable tool. Decisions regarding tax rates, social program, defense spending etc. Businessmen and businesswoman, in their eternal quest for profit, find statistics essential in the decision-making process.

For those in marketing research, statistics is of invaluable assistance in determining if a new product is likely to prove successful. Accountants, personal managers, and manufacturers all find unlimited opportunities to utilize statistical analysis. Even the medical researcher, concerned about the effectiveness of a new drug, finds statistics a helpful ally.

Today, most large businesses (as well as many smaller ones) have *quality control* (QC) departments whose function it is to gather performance data and resolve quality problems. Thus, *total quality management* (TQM) represents a growing area of opportunity to those with a statistical background.

One of the most important elements of TQM is a body of statistical tools and methods used to promote *statistical quality control* (SQC). These tools help to organize and analyze data for the purpose of problem solving. One of these tools is the *Pareto chart*. Named after the Italian economist, Vilfredo Pareto, this diagram identifies the quality problems that occur most often or that prove to be the most costly. Figure 1.1 shows a Pareto chart of the defects affecting the production of microwave ovens marketed by JC Penney.

Pareto charts often express the 80/20 rule: 80 percent of all problems are due to 20 percent of the causes. As Figure 1.1 shows, approximately 75 percent of all the problems are caused by the auto defrost feature and the temperature hold feature of the oven.

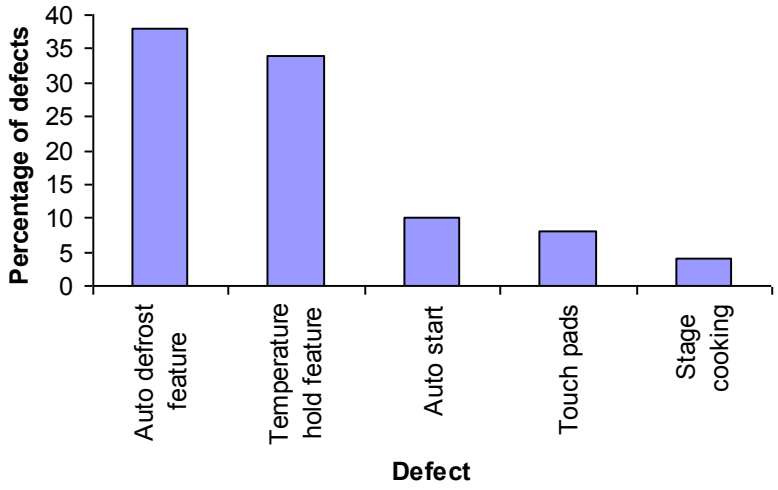


Figure 1.1 – Pareto Chart of the Defects Affecting the Production of Microwave Ovens Marketed By JC Penney

1.2 Statistical Observation: Forms and Types

Forms of Statistical Observation

1. *Reporting* – an organized form, in which data are presented in specially structured manner, obligatory, juridically confirmed by signature of manager. **Example**: financial report, balance, report about financial results, and report about movement of money means.

2. Specially organized observation

a) c*ensus* – periodic or one-time registration of certain sides of social life. **Example**: census of population, census of rare animals, plants.

b) *one-time accounting*. **Example**: registration of stockholders.

c) *special statistical survey* - exploration of mass situations on a certain topic:

- *direct observation***:** 3rd year students of NTU "KhPI", which have advanced level of English;
- *documentary:* accounting report;
- *survey* (oral, poll, interview): internet questionnaire, interview, fillings of certain forms, messages-survey.

Types of Statistical Observation

1. *By Time of Registration*:

a) *current observation* - systematic registration of facts or phenomena as they occur with the aim to study their dynamics. **Example**: registration of birth, marriages and death, registration of accidents and negative events as they occur by insurance companies.

b) *non-current observation*: *periodic observation* – is conducted in certain periods of time using similar tools and programs. **Example**: periodic observation of passengers turnover in public transport, periodic registration of prices of certain products (once a month or once a quarter)

c) *once-only observation* - is conducted one only in order to gather data on quantitative characteristics of phenomena or a process at the moment of its investigation. **Example**. The amount of unemployed people at the period autumn 2013.

2. By Inclusion of Units in the Population

a) *population information* – covers all units of investigated population. **Example**: census of enumeration, census of all national enterprises.

b) *sampling* (statistics) – covers only part of investigated population (a sample). Depending on how this part was chosen there are:

- *basis data array method* – investigates the biggest and the most essential units if the investigated population. **Example.** Investigation of the prices level on markets.
- *selection* – random selection of the sample, all the units investigated were randomly chosen from the population. **Example**. The level of knowledge of foreign languages among students, the amount of people doing sport in Ukraine.
- *monographic* – detailed investigation of units of the investigated population in order to reveal the tendencies began to show. **Example**: investigation of career priorities of International Business Department students.

3. By Type of Registration

a) *expeditionary* – when specially trained people fill in the forms and simultaneously check the veracity of answers

b) *self registration* – the survey through which respondents conduct the answering cards themselves.

c) *correspondences* – is conducted through correspondents who fill in the forms and give them to statistical officials.

d) *questionnaire* – forms are given to respondents or sent by post.

1.3 Basic Definitions

In any statistical study, the researcher is interested in a certain collection or set of observations called the ***population*** (or universe).

Example 1.1. If the chief executive officer (CEO) for a large manufacturing firm wishes to study the output of all plants owned by the firm, then the output of all plants is the population.

A **parameter** is a descriptive measure of the entire population of all observations of interest to the researcher. A parameter *describes* a population.

Example 1.2. The total output of all the manufacturing plants.

A **sample** is a representative portion of the population which is selected for study because the population is too large to examine in its entirely. A sample is a scientifically selected subset of the population. Samples are necessary because studying entire populations is too time-consuming and costly.

Example 1.3. Each month the US Department of Labor calculates the average income of a sample of only several thousand wage earners selected from the entire population of all 121 million workers. The average from this sample is then used as an estimate of the average income for the entire population.

A **statistic** describes a sample and serves as an estimate of the corresponding population parameter.

Example 1.4. The average income of those several thousand workers computed by the Department of Labor is a statistic.

A **variable** is the characteristic of the population that is being examined in the statistical study.

Example 1.5. In a study concerning the income of wage earners in the US, the variable is *income*.

Example 1.6. If the statistical advisor for the mayor of San Francisco is interested in the distance commuters must drive each morning, the variable is *miles driven*.

A variable can be:

1. *A quantitative variable* – if the observations can be expressed numerically. For **example**, the incomes of all the wage earners is an example of a quantitative population.

2. *A qualitative variable* is measured nonnumerically. For **example**, the marital status of credit applicants, the sex of students, hair color, the race etc.

3. *A continuous variable* is one that can take on any value within a given range.

4. *A discrete variable* is restricted to certain values, usually whole numbers. They are often the result of enumeration or counting. The number of students in the class or the number of cars sold by General Motors are the **examples**.

1.3.1 *Absolute and Relative Statistical Values*

Statistics in its resume is based on numerical data. First of all, the results of statistical observation are registered in forms of original *absolute values*.

Absolute statistical values are quantitative indexes, which characterize measurement of public activities in particular time and place conditions. Absolute statistical values do not show the structure of public activity under consideration, its progress in time, and do not give a full idea of it.

In statistics all *absolute statistical values* are

1. Denominate numbers (represent measures of factors).
2. Measured in concrete units.
3. Can be positive and negative (income, lost, decrease, etc.).

Measures of factors can be

1. Natural units (lengths, volume, etc.)
   - simple (meters, liters, tons, etc.);
   - complex – a combination of different units (kilometer per hour, kilowatt per hour, etc.).
2. Pricing units (currency of a country).
3. Labor units (work content of particular operation, etc.).

Absolute statistical values can be divided on groups*:*

1. Individual (characterize the factor size of a particular unit).
2. Resultant (summary).
3. Instant values (a real level of an phenomenon on a certain moment or a date, i.e. floating assets, population size, etc.).
4. Interval values (result over a period of time, i.e. population increase over a period of time, quantum of output over a month or a year, etc.).

Relative value in statistics is a deviation part of two values. The *main rule for calculation of relative values*: numerator is a characteristic under the study. It is called *current (reporting) magnitude*. The magnitude under the comparison called the *base of comparison* or a *base*.

The *comparison result* of two values of the same name can be expressed as:

1. Coefficient (a quantity of times a current magnitude “less” or “more” of the comparison base).
2. Percent (if a base is 100%).
3. Per thousand (the base is 1,000 units).
4. Per ten thousands (the base is 10,000 units).

Per one hundred thousand (the base is 100,000 units).

1.4 The Importance of Sampling

Samples are necessary because populations are often too big to study in their entirety. It is too time-consuming and costly to examine the entire population, that a selection of a sample from the population, calculate the sample statistic, and use it to estimate the corresponding population parameter often is very helpful.

There’re **two branches of statistical analysis**:

1. *Descriptive statistics* is the process of collecting, organizing, and presenting data in some manner that quickly and easily describes these data.

2. *Inferential statistics* involves the use of a sample to draw some inference, or conclusion, about the population from which that sample was taken.

However, all too often, the sample proves not to be very representative of the population, and **sampling error** will result. Because due to the luck of the draw in selecting the sample elements, it’s possible to unknowingly choose atypical elements that misrepresent the population. So, **sampling error** is the difference between the unknown population parameter and the sample statistic used to estimate the parameter.

Sampling bias is the tendency to favor the selection of certain elements over others. It’s a more serious form of sampling error.

If the sampling procedure is incorrectly designed and tends to promote the selection of too many units with a particular characteristic at the expense of units without that characteristic, the sample is said to be biased. For **example**, the sampling process may inherently favor the selection of males to the exclusion of females, or married persons to the exclusion of singles.

1.5 The Functions of Statistics

Statistics is the science concerned with the **(1)** collection, **(2)** organization, **(3)** presentation, **(4)** analysis, and **(5)** interpretation of data.

The first step in any statistical study is the *collection* of the data. Then the researcher has to *organize* and *present* these data in some meaningful and descriptive manner. The data must be put into some logical order that tends to quickly and readily reveal the message they contain. This procedure constitutes the process of **descriptive statistics**. After the data have been *organized* and *presented* for examination, the statistician must analyze and interpret them. These procedures rely on **inferential statistics** and constitute a major benefit of statistical analysis.

Statistical application 1-1 **(to predict the future with some degree of accuracy).** Fortune magazine recently described the problem Nike, a maker of athletic shoes, was having in deciding which type of color design was preferred among its customers. There was some confusion on the part of upper management regarding which of two competing fashions would attract more customer interest.

A sample of over 1000 die-hard runners was carefully selected. The runners were given the opportunity to express their preferences. From this information, management concluded that the population of customers displayed an overwhelming preference for the one design over the other. On this basis, the decision was made as to which design to manufacture and market.

1.6 Levels of Measurements

Variables can be classified on the basis of their level of measurement. Variables can be (1) nominal, (2) ordinal, (3) interval, or (4) ratio.

1. A **nominal** measurement is created when names are used to establish categories into which variables can be exclusively recorded. It carries no indication of order of preference, but merely establishes a categorical arrangement into which each observation can be placed.

Example 1.7. Sex can be classified as “male” and “female”, soft drink could be classified as Coke, Pepsi, or 7-Up. Each drink could be recorded in one of these categories to the exclusion of the others.

2. An **ordinal** scale produces a distinct ordering or arrangement of the data. That is, the observations are ranked on the basis of some criterion.

Example 1.8. Opinion polls often use an ordinal scale such as “strongly agree”, “agree” “no opinion”, “disagree”, and “strongly disagree”. Numbers can be used to order the rankings. The magnitude of the numbers is not important. *The arithmetic differences between the values are meaningless.* A product ranks “2” is not twice as good as one with a ranking of “1”.

3. Variables on an **interval** scale are measured numerically, and, like ordinal data, carry an inherent ranking or ordering. *The differences between the values are important.* Thus, the arithmetic addition and subtraction are meaningful. The value of zero is arbitrarily chosen in an interval scale as reference point.

Example 1.9. The Fahrenheit scale for temperatures. The same difference of 10 degrees exists as between 90 and 100 degrees Fahrenheit as 70 degrees hotter than 60 degrees, or to say it's 10 degrees colder than 10 degrees. Thus, 80 degrees is not twice as hot as 40 degrees and the ratio 80/40 has no meaning.

4. The **ratio** scale is based on a numbering system in which zero is meaningful. The arithmetical operations of multiplication and division also take on a rational interpretation. A **ratio scale** is used to measure many types of data found in business analysis (costs, profits, etc.). Measurements such as weight, time, and distance are also measured on a ratio scale since zero is meaningful.

Example 1.10. A firm with 40 percent market share has twice as much of the market as a firm with 20 percent market share. An item that weight 100 pounds is one-half as heavy as an item weighing 200 pounds.

Chapter Checklist

1. Explain the importance of statistics and how you will use it to solve common business problems?
2. Discuss the manner in which a statistical background can open up career opportunities?
3. Construct and interpret a Pareto chart?
4. Distinguish between populations and their parameters, and samples and their statistics?
5. Define qualitative and quantitative variables, and give examples?
6. Explain sampling error, and cite it causes?
7. Discuss the five functions of statistics?
8. Define and give examples of the four levels of measurement of variables?

2 DESCRIBING DATA SETS. ORGANIZATION AND GROUPING

2.1 Introduction

Almost every statistical effort begins with the process of collecting necessary data and thereby forming the data set to be used in the study. This collection of raw data in itself reveals very little. In order to determine data significance, we must organize them into some form so that, we can get an idea as what the data can tell us.

Statistical tools for organizing data include

- Frequency tables, which place all the data in specific classes.
- Various pictorial displays that can provide a handy visual representation of the data.
- Contingency tables and steam – and – leaf designs, which also allow presentation of a large data set in a concise, discernible form.

2.2 Methods of Organizing Data

Several basic tools can be used to describe and summarize a large data set. The simplest is an **ordered array**.

Example 2.1 (if there's a small data array). The IQ scores of five valedictorians from Podunk University are 75, 73, 91, 83 and 80. An ordered array merely lists these observations in ascending or descending order. The five values might thus appear as 73, 75, 80, 83, 91. The ordered array provides some organization to the data set; it can now readily be seen that the two extreme values are 73 and 91.

2.2.1 *Frequency Distributions*

Example 2.2 As the resident statistician for Pigs & People Airline, you have been asked by the chairman of the board to collect and organize data regarding flight operations. You are primarily interested in the daily values for two variables: (1) number of passengers and (2) number of miles flown, rounded to the nearest tenth. You are able to obtain these data from daily flight records for the past 50 days, and you have recorded this information in Table 2.1 and Table 2.2. However, in this raw form it is unlikely that the chairman could gain any useful knowledge regarding operations.

Table 2.1 – Raw Data on the Number of Passengers for P&P Airlines

68	71	77	83	79
72	74	57	67	69
50	60	70	66	76
70	84	59	75	94
65	72	85	79	71
83	84	74	82	97
77	73	78	93	95
78	81	79	90	83
80	84	91	101	86
93	92	102	80	69

Table 2.2 – Raw Data on Miles Flown for P&P Airlines

569.3	420.4	468.5	443.9	403.7
519.7	518.7	445.3	459.0	373.4
493.7	505.7	453.7	397.1	463.9
618.3	493.3	477.0	380.0	423.7
391.0	553.5	513.7	330.0	419.8
370.7	544.1	470.0	361.9	483.8
405.7	550.6	504.6	343.3	497.9
453.3	604.3	473.3	393.9	478.4
437.9	320.4	473.3	359.3	568.2
450.0	413.4	469.3	383.7	469.1

There are several ways to organize presentation.

1. **Frequency distribution** (frequency table) will provide some order to the data by dividing them into classes and recording the number of observations in each class. In Table 2.3 distribution for the daily number of passengers over the last 50 days is presented.

Each class has a *lower boundary* and *upper boundary*. The class limits of these boundaries are quite important. It is essential that class boundaries do not overlap.

Table 2.3 – Frequency Distribution for Passengers

Class (passengers)	Frequency (days)
50 to 59	3
60 to 69	7
70 to 79	18
80 to 89	12
90 to 99	8
100 to 109	2
Total	50

Boundaries such as

50 to 60
60 to 70
70 to 80
…

are confusing. Since *passengers* is a discrete variable, values such as 59.9 pose no problem since it is impossible to have fractional values. On the other hand, *miles flown* is a continuous variable since it is possible to fly a fraction of a mile. It would be improper to set the boundaries as

300 to 349
350 to 399
400 to 449

since it is unclear in which class observations such as 349.9 or 399.9 should be tallied. The frequency distribution for miles flown might instead appear in Table 2.4. The chairman can now detect a pattern to flight operations not apparent from the raw data in Table 2.2. for example, P&P never flew over 650 miles on any of the 50 days examined. They flew between 450 and 500 miles more often than any other distance. On 26 of the 50 days examined, total mileage was between 400 and 500 miles.

Table 2.4 – Frequency Distribution for Miles Flown

Class (miles)	Frequency (days)
300 and under 350	3
350 and under 400	9
400 and under 450	9

Continuation of Table 2.4

Class (miles)	Frequency (days)
450 and under 500	17
500 and under 550	6
550 and under 600	4
600 and under 650	2
Total	50

2. **The number of classes** in a frequency table is somewhat arbitrary. In general, statistical table should have between 5 and 20 classes. Too few classes would not reveal any details about the data; too many would prove as confusing as the list of raw data itself.

There is a simple rule to approximate the number of classes

$$2^c \geq n, \tag{2.1}$$

or

$$c = 1 + 3.322 * lg\ n, \tag{2.2}$$

where c is the number of classes, n is the number of observations.

In example 2.2 for P&P the number of observations is $n = 50$. thus,

$$2^c \geq 50.$$

Solving for c, it can be found $2^6 = 64$, which exceeds n. This rule suggests that there should be six classes in the frequency table.

This rule should not be taken as the final determining factor. For convenience, more classes or fewer classes may be used.

2.2.2 *Class Intervals and Midpoints*

The **class interval** is the range of values found within a class. It is determined by subtracting the lower (or upper) boundary of one class from the lower (or upper) boundary of the next class. The interval of the first class in Table 2.3.is thus 60 – 50 = 10.

It is desirable to make all class *intervals of equal size* in a frequency distribution. However, equal class sizes may not always be possible. It may

necessary to use *open – ended intervals*, which do not cite a lower boundary for the first class or an upper boundary for the last class. The last class in Table 2.4 might be stated as "600 and up".

The class interval can be determined by the following expression:

$$CI = \frac{\text{Largest value - Smallest value}}{\text{Number of desired classes}}. \qquad (2.3)$$

For the example 2.2 $CI = \frac{102 - 50}{6} = 8.7$. Since 8.7 is an awkward number, the results can be slightly adjusted up or down to facilitate construction of the frequency table. For convenience, the interval of 10 was selected in forming Table 2.3.

It is often necessary to determine a **class midpoint**. This is done by calculating the average of the boundaries of a class by adding the upper and lower boundaries and dividing by 2. thus, the midpoint for the first class in Table 2.5 is (50+59)/2 = 54.5. The remaining midpoints are also shown.

Table 2.5 – Frequency Distribution and Midpoints for Passengers

Class	Frequency	Midpoint
50 to 59	3	54.5
60 to 69	7	64.5
70 to 79	18	74.5
80 to 89	12	84.5
90 to 99	8	94.5
100 to 109	2	104.5

2.2.3 *Cumulative Frequency Distributions*

Determination of the number of observations that are greater than or less than some amount can be done through the use of a *less-than cumulative frequency distribution* or a *more-than cumulative frequency distribution.* Cumulative frequency tables can easily be constructed from their respective frequency tables.

A **less-than cumulative frequency distribution** for a particular class is found by adding the frequency in that class to those in all previous classes. Table 2.6 shows a less-than cumulative frequency distribution for the data set on

passengers. The frequencies from Table 2.3 are repeated in Table 2.6. It can be seen that, on 40 of the 50 days, less than 90 passengers flew the airways of P&P.

Table 2.6 – Less-Than Cumulative Frequency Distribution for the Number of Passengers

Class (passengers)	Frequency (days)	Cumulative Frequency (days)
Less than 50	0	0
Less than 60	3	3
Less than 70	7	10
Less than 80	18	28
Less than 90	12	40
Less than 100	8	48
Less than 110	2	50

For a **more-than cumulative frequency distribution**, the values are found by subtracting frequencies of previous classes. This is reflected in Table 2.7. using Table 2.3 it can be found that on all 50 days at least 50 passengers boarded P&P Airlines. Since less that 60 passengers bought tickets only on three of the 50 days, then on remaining 47 days, 60 or more people flew P&P, on 22 days at least 80 passengers flew the airline.

Table 2.7 – More - Than Cumulative Frequency Distribution for the Number of Passengers

Class (passengers)	Frequency (days)	Cumulative Frequency (days)
50 or more	3	50
60 or more	7	47
70 or more	18	40
80 or more	12	22
90 or more	8	10
100 or more	2	2
110 or more	0	0

2.2.4 *Relative Frequency Distributions*

A *relative frequency distribution* expresses the frequency percentage of the total number of observations in the sample. This allows to make statements regarding the number of observations in a single class relative to the entire sample. Using the data from Table 2.3 on passengers, it can be now computes a relative frequency distribution by dividing each frequency by the sample size 50. See Table 2.8.

Table 2.8 – Relative Frequency Distribution for Passengers

Class (passengers)	Frequency (days)	Relative Frequency
50 to 59	3	3/50 = 6%
60 to 69	7	7/50 = 14%
70 to 79	18	18/50 = 36%
80 to 89	12	12/50 = 24%
90 to 99	8	8/50 = 16%
100 to 109	2	2/50 = 4%
Total	50	100%

In this fashion it can be seen, that 36 percent of the days sampled, between 70 and 79 passengers embarked on a journey using P&P Airlines or 60 percent of the time, P&P served between 70 and 89 travelers.

2.2.5 *Cumulative Relative Frequency Distribution*

A *cumulative relative frequency distribution* expresses the cumulative frequency of each class relative to the entire sample. The cumulative process can be based on a more-than or less-than principle. A less-than cumulative relative frequency distribution is shown in Table 2.9. The last column provides the information in which we are interested. It shows, for example, 80 percent of the time less than 90 passengers boarded P&P airplanes.

Table 2.9 – Less-Than Cumulative Relative Frequency Distribution for the Number of Passengers

Class (passengers)	Frequency (days)	Cumulative Frequency (days)	Cum.Relative Frequency
Less than 50	0	0	0/50 = 0%
Less than 60	3	3	3/50 = 6%
Less than 70	7	10	10/50 = 20%
Less than 80	18	28	28/50 = 56%
Less than 90	12	40	40/50 = 80%
Less than 100	8	48	48/50 = 96%
Less than 110	2	50	50/50 = 100%

2.2.6 *Contingency Tables (Cross-Tabulations or Cross-Tabs)*

The data sets in Example 2.2 involved only one variable: the number of passengers, or the number of miles flown. To examine two variables simultaneously *contingency tables* can be used. These tables indicate the number of observations for both variables that fall jointly in age level category of passengers, for example. If statistician obtained data relating to age and frequency of flying, he could present this information in the form of a contingency table such as Table 2.10.

Table 2.10 – P&P's Contingency Table for Age and Flights

Number of Flights per Year				
Age	1-2	3-5	Over 5	Total
0 to less than 25	1 (0.02)	1 (0.02)	2 (0.04)	4 (0.08)
25 to less than 40	2 (0.04)	8 (0.16)	10 (0.20)	20 (0.40)
40 to less than 65	1 (0.02)	6 (0.12)	15 (0.30)	22 (0.44)
65 and over	1 (0.02)	2 (0.04)	1 (0.02)	4 (0.08)
Total	5 (0.10)	17 (0.34)	28 (0.56)	50 (1.00)

In this table there is four age categories and three flight categories. Each of the 50 people in this sample will fall into one of these 12 joint categories. The percentage within each cell is shown in parentheses. For example, 8 of the 50

passengers, or 16 percent, between ages 25 and 40 fly between three and five times a year; the most of the passengers come from the second and third age categories; people in the 40 to less than 65 age group flew over five times a year more often than any other frequency.

The information contained in each cell could be displayed on the basis of sex. Thus, of the eight people in the second age category who fly between three and five times a year, six might be males and the other two females. The entry in that cell could then appear as 6, 2 (0.16).

2.3. Pictorial Displays

Pictorial displays are another way of describing data sets.

1 A **histogram** is very common method of displaying data. It places the classes of a frequency distribution on the horizontal axis, and the frequencies on the vertical axis. The area in each rectangular bar is proportional to the frequency in that class. See Figure 2.1.

2 A **frequency polygon** expresses the distribution of the data by means of a single line determined by the midpoints of the classes. See Figure.2.2. it is common practice to extend both ends of the frequency polygon to the horizontal axis at those points that would be the midpoints of the adjacent classes at each end. Thus, if a class preceded the first class of 50 to 59, it would have boundaries of 40 to 49 with a midpoint of 44.5. similarly, if the last class of 100 to 109 were followed by another class, it would have boundaries of 110 to 119 and a midpoint 114.5.

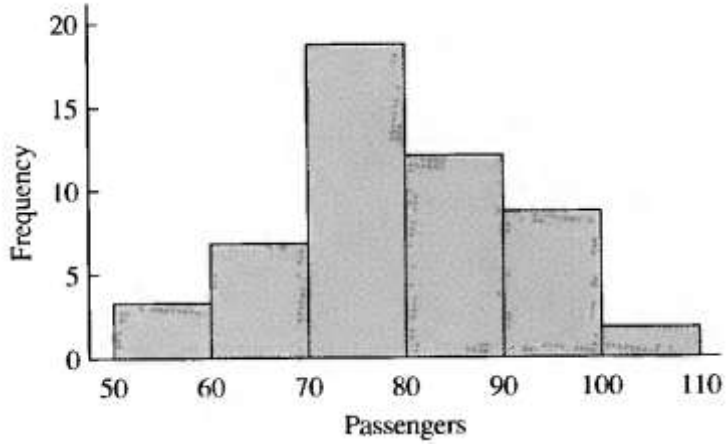


Figure 2.1 – Histograms for P&P's Passengers

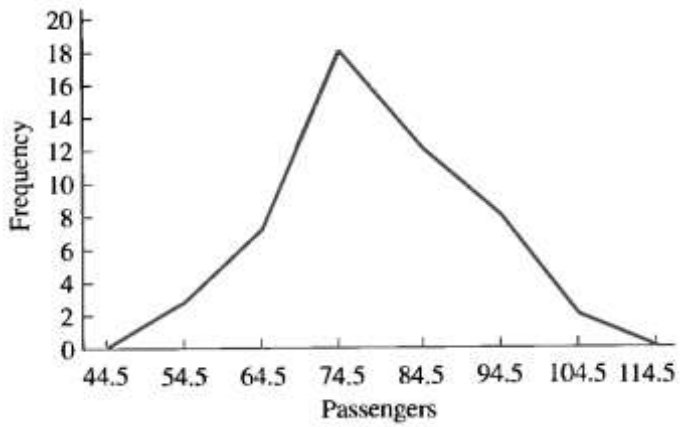


Figure 2.2 – Frequency polygon for P&P's Passengers

3 The information shown in a lees-than cumulative frequency distribution and a more-than cumulative frequency distribution can also be displayed pictorially. Such a graph is called an **ogive** and is shown in Figure 2.3. the less-than ogive shows that on 3 days, less than 60 passengers flew P&P Airlines, and the more-than ogive reveals that on 47 days, more than 59 (60 or more) travelers boarded flights with P&P.

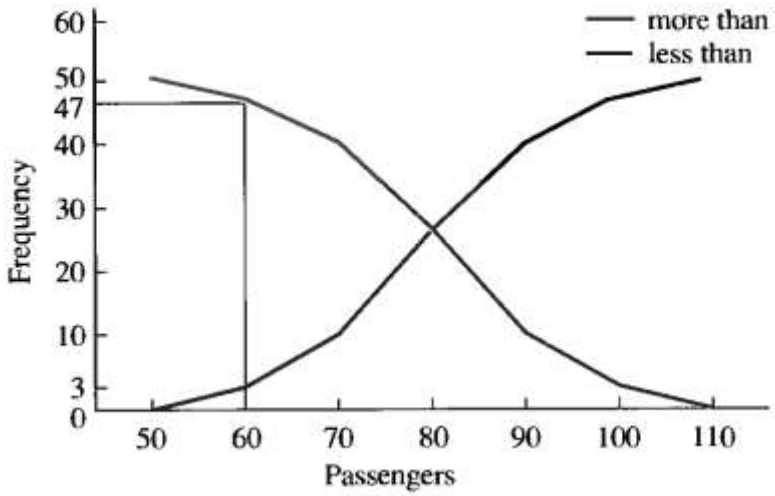


Figure 2.3 – Ogives for Passengers Data

4 A **bar chart** presents data in a manner similar to a histogram. The main difference is that the bar chart need not show frequencies on the axis, but may be presented horizontally or vertically. See Figure 2.4. it is possible to show more that one value at a time on a bar chart. Figure 2.5 displays the revenues and costs for P&P Airlines. Figure 2.6 provides further examples of the use of bar charts.

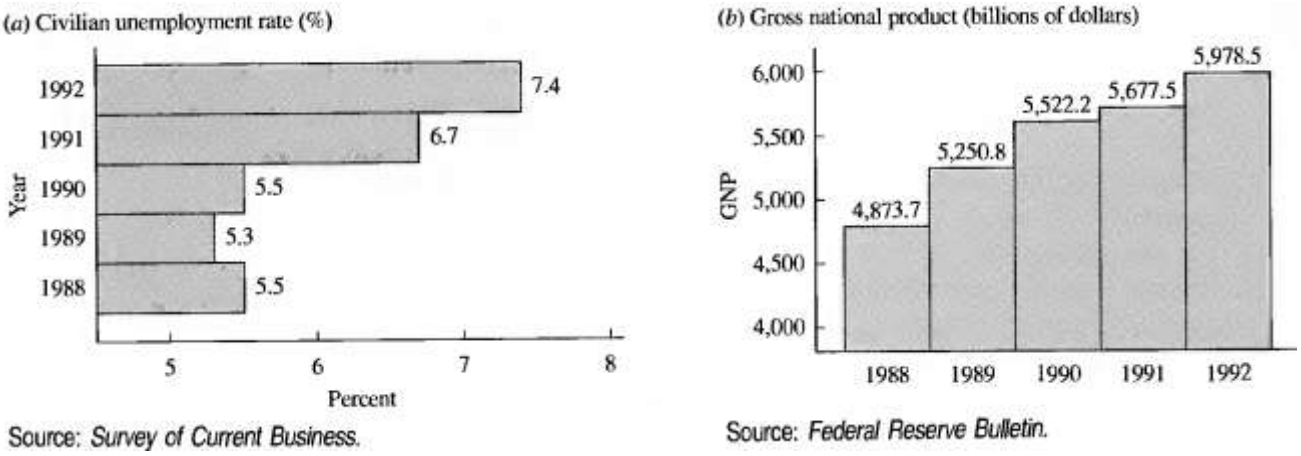


Figure 2.4 – Bar charts for Performance of the US Economy

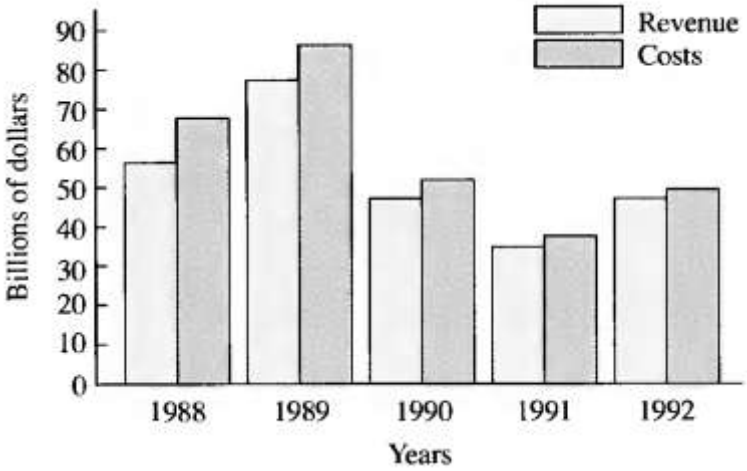
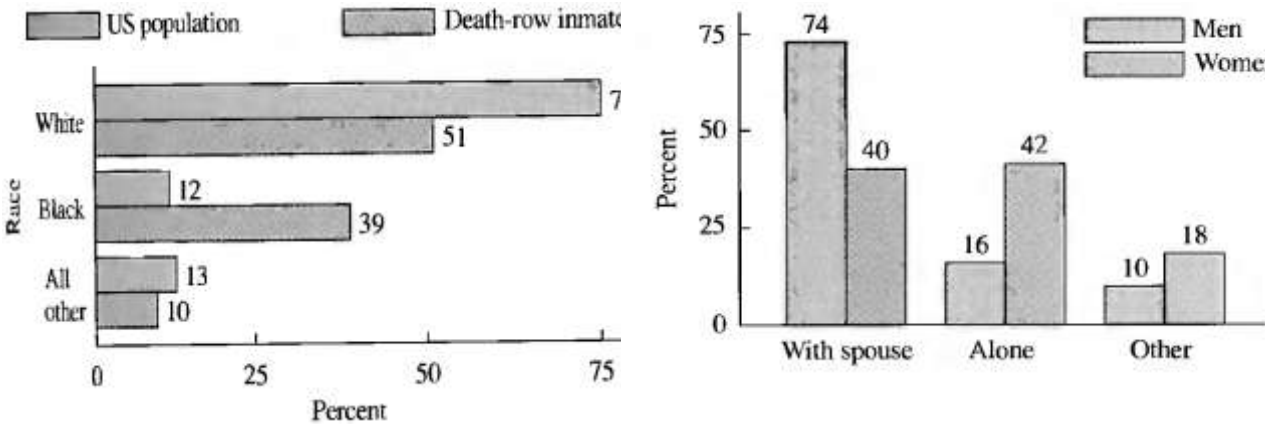
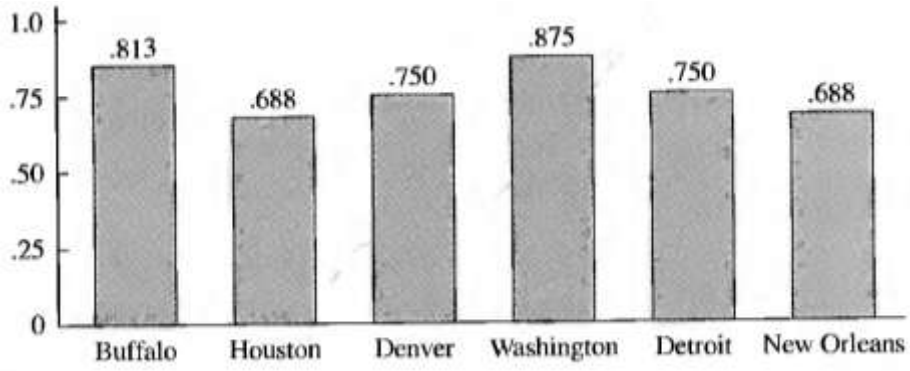


Figure 2.5 – P&P Performance



a *b*

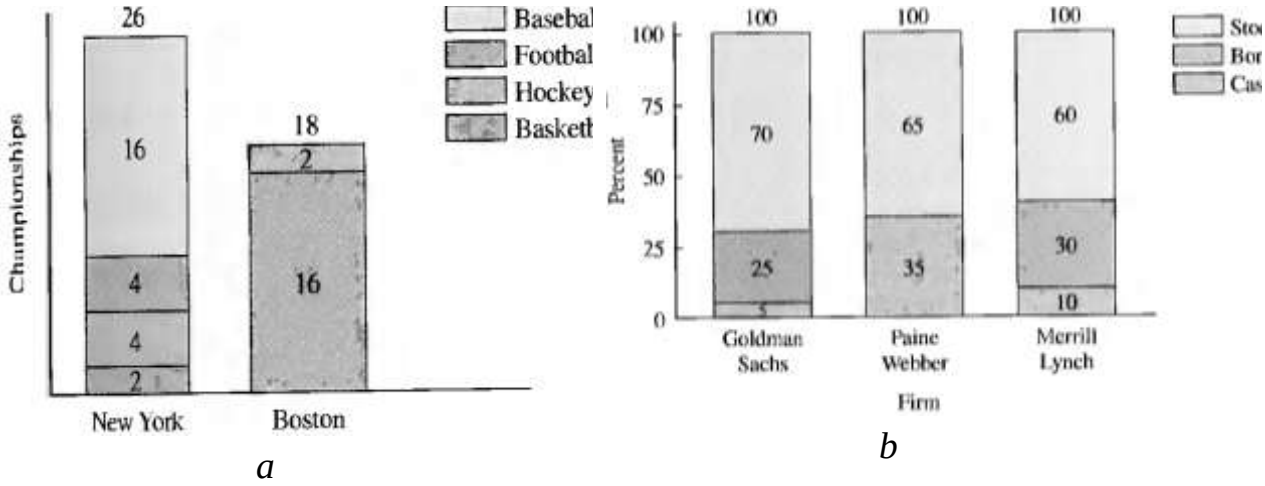
Figure 2.6 – Common Uses of Bar Charts: *a* – Percent of US population and death-row inmates by race; *b* – Living arrangements of people 65 years and older



c

Figure 2.6 – Common Uses of Bar Charts: *c* – Winning percentages of 1991 divisional winners in national football league (continuation)

Figure 2.7 contains **stacked bar charts**, which provide a different look at the use of the displays on Figures 2.4 – 2.6.



a *b*

Figure 2.7 – Stacked Bar Charts: *a* – Championship teams since 1946; *b* – Where to put your money: advice from investment firms

5 A **pie chart** presents the data in the form of circle. The slices represent the absolute or relative (percentage) proportions. Pie charts are quite useful for displaying relative differences among observations, and they are particularly appropriate for illustrating percentage differences. A pie chart is formed by marking off a portion of the pie corresponding to each characteristic being displayed.

To ensure that the pie is properly portioned, each percentage is multiplied by the 360 degrees in a circle. If the values are not already in percentages, it is necessary to convert them to percentages to determine the appropriate number of degrees.

Table 2.11 contains the appropriate breakdown for the characteristics of the nation's work force taken from an edition of The Wall Street Journal. For example, the 48 percent of all workers who never take work home converts to 172.8 degrees. The remaining portions are similarly calculated and the resulting pie chart is shown in Figure 2.8

Table 2.11 – Working Habits of the US Labor Force

How Often Employees Take Work Home	Proportions (%)	Degrees
Never	48	360 * 0.48 = 172.8
Less than once a month	10	360 * 0.10 = 36.0
Once per month	12	360 * 0.12 = 43.2
Twice per week	9	360 * 0.09 = 32.4
3 – 4 times per week	8	360 * 0.08 = 28.8
Every day	13	360 * 0.13 = 46.8
Total	100	360.0

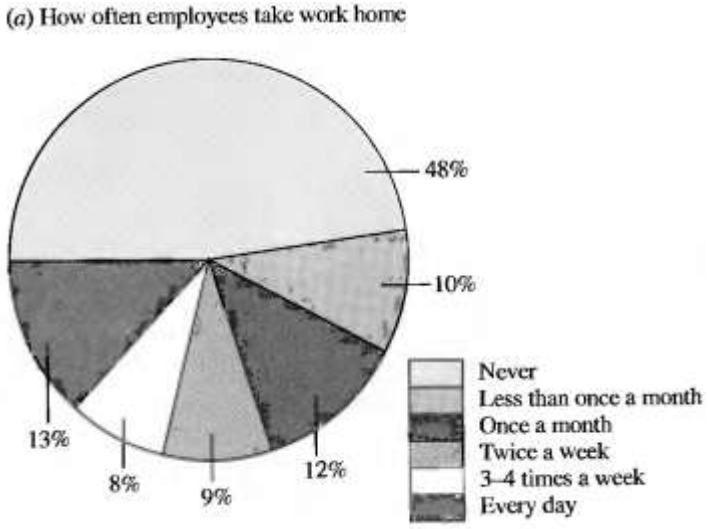


Figure 2.8 – Pie Chart

6 Since much of the business and economic data statisticians work with is measured over time, a **line chart** is useful because it permits us to express units of time on the horizontal axis. A line chart for the unemployment rate in the US is shown in Figure 2.9.

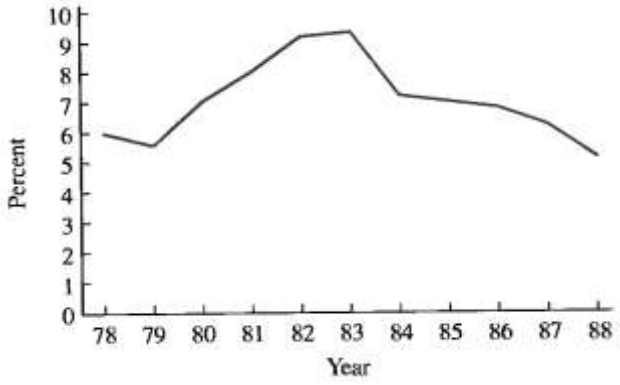


Figure 2.9 – US Unemployment Rate

7 Financial data are often displayed with the aid of a **high – low – close chart (**or *ticks and tabs***)**. It displays the highest value, the lowest value, and the closing value for a selected variable during a given time period. The most well – recognized example is the Down Jones averages found daily in The Wall Street Journal (WSJ). See Figure 2.10.

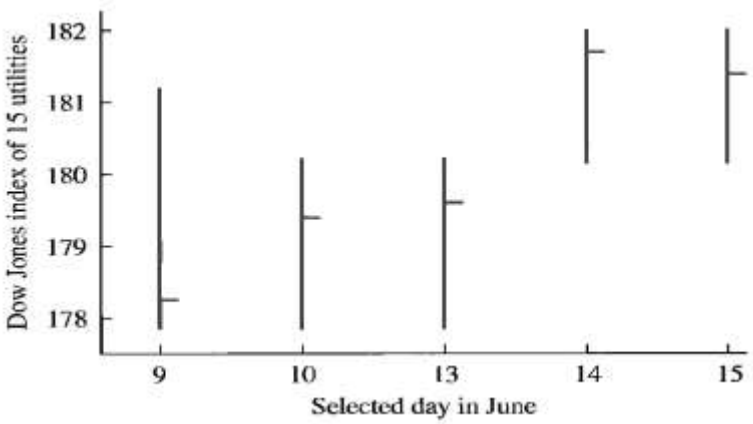


Figure 2.10 – The Dow Jones Average for 15 Utilities on Selected Days in June

The upper end of the vertical line, or *tick*, marks off the highest value that the index reached on that day; the lowest end of a tick indicates the lowest value of the day. The closing value is shown by the little *tab* in between.

2.4 Stem – And – Leaf Designs

The stem-and-leaf design is an alternative to a histogram, which provides a quick visual impression of the number of observations in a class. The stem-and-leaf design was devised by a noted statistician John Tukey.

Each observation in the data set is divided into two parts; a *stem* and a *leaf*. Although there is considerable flexibility in the procedure to be followed, it is often convenient to identify all but the last digit of an observation as the stem. The last digit is then identify as the leaf.

The stems must be placed in an ordered array from the lowest to highest. It is also often desirable to place the values in the leaf in an ordered array. The complete stem-and-leaf design for the data on passengers in Table 2.1 is presented in Table 2.12.

Table 2.12 – Stem-and-Leaf Design for Passenger Data

Stem	Leaf
5	0, 7, 9
6	0, 5, 6, 7, 8, 9, 9
7	0, 0, 1, 1, 2, 2, 3, 4, 4, 5, 6, 7, 7, 8, 8, 9, 9, 9
8	0, 0, 1, 2, 3, 3, 3, 4, 4, 4, 5, 6
9	0, 1, 2, 3, 3, 4, 5, 7
10	1, 2

Now it is apparent that not only are there three observations in the 50s, but their individual values of 50, 57, and 59 are easily seen; the observations range from a low of 50 to a high of 102.

As it can be seen, the *stem-and-leaf design is similar to a histogram*, but offers the advantage of retaining the values of the original observations.

If the data set contains fractional observations, like these 26.0, 28.3, 28.7, 27.8, 29.3, 29.5, it might be advantageous to use as the stem all the digits to the left of the decimal point, while those on the right become the leaf. See Table 2.13

Table 2.13 – Stem-and-Leaf Design for Fractional Observations

Stem	Leaf
26	0
27	8
28	3, 7
29	3, 5

Chapter Checklist

After studying this chapter, as a test of your understanding of the basic concepts, can you

1. Construct a frequency distribution from a raw of data set by determining the proper number of classes and the class boundaries?
2. Determine class midpoints in a frequency table?
3. Construct cumulative, relative, and cumulative relative frequency distributions?
4. Compile contingency tables with the proper number of cells containing the appropriate information?
5. Convey information through the construction of various pictorial displays?
6. Create stem-and-leaf designs from raw data?

3 MEASURES OF CENTRAL TENDENCY AND DISPERSION

3.1 Introduction

A data set of *quantitative variables* can be also called as **variational series** (*series of order statistics*). **Variant** is a parameter point which varies. **Frequency** is the number of each variant. A data set of *qualitative variables* called **attributive series**. The purpose of this chapter is to determine various ways in which the average of a data set can be calculated. These averages are referred to as *measures of central tendency*. It will be also explored ways to judge the extent to which the individual observations in a data set are spread out around their central point. These valuations of spread are called *measures of dispersion*. A measure of central tendency locates the center, or average, of a data set. A measure of dispersion indicates the tendency for the individual observations to deviate from that center point. These important measures can be calculated for raw, ungrouped data, or for data that have already been grouped into classes within a frequency table. These objectives can best be achieved by examining the

- Mean, median, and mode for grouped and ungrouped data.
- Mean absolute deviation.
- Variance and standard deviation for grouped and ungrouped data.
- Quartiles, deciles, and percentiles for grouped and ungrouped data.

It might be easier to first compute the measures of central tendency for ungrouped data.

3.2 Measures of Central Tendency for Ungrouped Data

There are three common methods of identifying the center of a data set around which the data are located. They are the *mean*, the *median*, and the *mode*. The precise determination of these three values can vary. In each case, that is the point around which all the numbers in the data set seem to be grouped.

3.2.1 *Mean*

The **mean** is the measure of central tendency most commonly thought of as the average. The mean of a population, containing N observations, is calculated as

$$\mu = \frac{X_1 + X_2 + X_3 + ... + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}, \quad (3.1)$$

where *N* is the size of the population.

The mean of population is a parameter. The mean is affected by extreme values, or *outliers*.

Example 3.1. Peter wants to compute the mean of his last 10 exams in his statistics course. He adds them up and divide by 10.

If the population as large and it would prove too time-consuming to add up all the observations, the only alternative might be to take a sample and calculate its mean. *The mean for a sample is a statistic*. It is found as follows:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad (3.2)$$

where *n* is the size of a sample.

3.2.2 *Median*

The **median** is sometimes referred to as the *positional average,* because it lies exactly in the middle of the data set after the values have been placed into an ordered array. One-half of the observations will be *above* the median, and one-half will be *below* the median. *Before the median can be calculated, the observations must be put into an ordered array*. The median is not affected by extreme value.

In a data set with an *odd number of observations*, this middle position is found as

$$\text{Median position}^{\text{odd}} = \frac{N+1}{2}. \quad (3.3)$$

If there is an *even number of observations*, the median is the average of the two middle values.

$$\text{Median position}^{\text{even}} = \frac{X_{n/2} + X_{(n/2)+1}}{2}. \quad (3.4)$$

Example 3.2. Denny is rightfully worried about his grade in statistics. He scored the following grades on the five tests given this semester: 63, 59, 71, 41, 32. his professor has warned Denny that any grade below 60 is failing. Calculate and interpret Denny's median grade.

Solution. **(1)** The values have to be put in an ordered array first. There is no matter if it would be ascending or descending order:

32, 41, 59, 63, 71.

The *position* of the median grade is determined as (N+1)/2 = (5+1)/2 = 3. The median is then 59, which is located in the third position after the data have been placed into the ordered array.

(2) Assume a sixth test was given and Denny scored another 63 on it. The ordered array would appear as

32, 41, **59, 63**, 63, 71.

The *median position* is now (N+1)/2 = (6+1)/2 = 3.5; that is, the third and one-half position. The median is the value halfway between the third and fourth observations, and is therefore (59+63)/2 = 61. This means that half of his test grades were below 61 and half were above 61.

3.2.3 *Mode*

The **mode** is the observation that occurs with the greatest frequency.

From the **example 3.2 (2),** the *modal grade* is 63 because it occurred more often than any other grade. If still another test was given and Denny managed another 41 on it, the data set would be *bimodal* because both 41 and 63 occurred with equal frequency.

If all the observations occur with equal frequency, the data set has no mode.

For the **example 3.2 (1)**, with Denny's first five tests of 32, 41, 59, 63, 71, there is no modal grade.

3.2.4 *Weighted Mean*

In the discussion of the mean in 3.2.1, it was assumed that each observation was equally important. In certain cases, some observations could be given greater weight. The proper procedure is given by Formula 3.4.

$$\overline{X}_W = \frac{\sum XW}{\sum W}, \tag{3.5}$$

where $\overline{X}_W$ is the weighted mean; X is the individual observations; W is the weight assigned to each observation.

The weighted mean is higher than the simple arithmetic mean.

Example 3.3. Statistics professor threatens to count the final exam twice as much as the other tests when determining the final grade of each student in the class. Then the score they get on the final must be given twice as much weight. That is, it must be counted twice in figuring the grade. Assume the student scored

89, 92, and 79 on the hour exams, and a 94 on the final exam. These scores and weight can be reflected in Table 3.1. Formula (3.5) yields

$$\overline{X}_W = \frac{\sum XW}{\sum W} = \frac{448}{5} = 89.6.$$

Table 3.1 – Calculation of the weighted mean

Grade (X)	Weight (W)	XW
89	1	89
92	1	92
79	1	79
94	2	188
Total	5	448

This approach is the same as adding the score on the final exam twice in computing the average:

$$\overline{X}_W = \frac{89+92+79+94+94}{5} = 89.6.$$

Example 3.4. Paul the Plumber sells five types of drain cleaner. Each type, along with the profit per can and the number of cans sold, is shown in Table 3.2.

Table 3.2 – Types of Drain Cleaner for Paul the Plumber

Cleaner	Profit per Can (X), \$	Sales Volume in Cans (W)	XW, \$
Glunk Out	2.00	3	6.00
Bubble Up	3.50	7	24.50
Dream Drain	5.00	15	75.00
Clear More	7.50	12	90.00
Main Drain	6.00	15	90.00
	24.00	52	285.50

Solution. The simple arithmetic mean of Paul's profit can be calculated as \$24/5 = \$4.80 per can. This is not a good estimate of Paul's average profit, since he sells more of some types than he does of others. In order to get a financial statement more representative of his true business performance, Paul must give more weight to the more popular types of cleaner. The proper calculation would

therefore be the weighted mean. The proper measure of weight would be the amounts sold. The weighted mean is then

$$\bar{X}_W = \frac{\sum XW}{\sum W} = \frac{\$285.50}{52} = \$5.49 \text{ per can.}$$

Interpretation. In this example the weighted mean is higher than the simple arithmetic mean because Paul sells more of those types of cleaner with a higher profit margin.

3.2.5 *The Geometric Mean*

The **geometric mean** can be used to show percentage changes in a series of positive numbers, and it also represents the average change over time. It has wide application in business and economics (the percentage change in sales, gross national product, etc.).

GM is found by taking the *n*th root of the product of *n* numbers. Thus,

$$GM = \sqrt[n]{X_1 X_2 X_3 \cdots X_n}\,. \tag{3.6}$$

The *geometric mean will always be less than the arithmetic mean* except in the rare case when all the percentage increases are the same. Then the two means are equal.

Example 3.5. The geometric mean of 5, 6, 8, and 12 is

$$GM = \sqrt[4]{5*6*8*12} = \sqrt[4]{2880} = 7.33.$$

3.3. Measures of Central Tendency for Grouped Data

If the data have been grouped into classes in a frequency table, it is impossible to determine measures of central tendency by the methods just discussed since the individual values are not given. Alternative approaches must be found. It should be kept in mind that computations made using grouped data are only approximations.

3.3.1 *Mean*

In calculating the mean from grouped data, the assumption is made that the observations in each class are equal to the class midpoint. Other words, the frequency and midpoint of each class must be taken into consideration when computing the mean using grouped data.

$$\bar{X}_g = \frac{\sum fM}{n} = \frac{\sum fM}{\sum f}, \tag{3.7}$$

where *f* is the frequency or number of observations in each class; *M* is the midpoint of each class; *n* is the sample size and equals the combined frequencies in all classes.

Example 3.6. The frequency table for Pigs&People Airlines was developed in Chapter 2 and repeated in Table 3.3.

Table 3.3 – Frequency Distribution for Passengers

Class (passengers)	Frequency (*f*) (days)	Midpoint (*M*)	*fM*
50 to 59	3	54.5	163.5
60 to 69	7	64.5	451.5
70 to 79	18	74.5	1341.0
80 to 89	12	84.5	1014.0
90 to 99	8	94.5	756.0
100 to 109	2	104.5	208.0
	50		3935.0

Using formula (3.7), it can be seen that P&P flew a daily average of 78.7 passengers.

$$\overline{X}_g = \frac{\sum fM}{n} = \frac{3935}{50} = 78.7.$$

3.3.2 *Median*

If the data have been recorded in a frequency table, they cannot be placed in an ordered array in order to calculate the median. The *median class* of the frequency distribution should be found first. The **median class** is that class whose cumulative frequency is greater than or equal to *n*/2. The **median** can be determined as

$$Median = L_{md} + \left[\frac{n/2 - F}{f_{md}}\right](C), \tag{3.8}$$

where L_{md} is the lower boundary of the median class; *F* is the cumulative frequency of the class preceding the median class; f_{md} is the frequency of the median class; *C* is the class interval of the median class.

As an illustration, the frequency table for P&P Airlines is given in Table 3.4

Table 3.4 – Frequency Distribution for Passengers

Class (passengers)	Frequency (*f*) (days)	Cumulative Frequency (days)
50 to 59	3	3
60 to 69	7	10
70 to 79	18	28
80 to 89	12	40
90 to 99	8	48
100 to 109	2	50

Since *n* is 50, the median class is the third class in the Table 3.4. It has a cumulative frequency of 28. Using formula (3.8), the median is

$$Median = 70 + \left[\frac{50/2 - 10}{18}\right]10 = 78.33.$$

Conclusion: on 25 days – one-half of the 50 days surveyed – less than 78.33 passengers flew on P&P Airlines, and on the other 25 days more than 78.33 passengers flew the friendly skies of P&P.

3.3.3 *Mode*

Since by definition the mode is the observation that occurs more often, it will be found in the class with the highest frequency. This class with the largest frequency is called the modal class. To estimate the mode in the case of grouped data, Formula (3.9) is used.

$$Mode = L_{mo} + \left[\frac{D_a}{D_b + D_a}\right](C), \qquad (3.9)$$

where L_{mo} is the lower boundary of the modal class; D_a is the difference between the frequency of the modal class and the class preceding it; D_b is the difference between the frequency of the modal and the class after it; C is the class interval of the modal class.

3.4 Selecting the Appropriate Measure of Central Tendency

The chosen measure might depend on the nature of the data or the manner in which that data are used.

For **example**, Land's End, a popular retailer of camping equipment, would benefit little from the knowledge that the average size of the hiking boots they sold was 7.3492. More useful in future business decisions would be knowledge of the *modal* size – recognizing that they sold more boots of size 8 than any other.

As another **example**, presume Peter is engaged in a study of consumer incomes. The mean income is $80,000. However, a very large majority of the people has incomes around $35,000, but the presence of a few millionaires (outliers) raises the overall mean level. This mean of $80,000 is somewhat misleading in that it doesn't represent any typical person's income. If businesses were to base decisions on an average of $80,000, or government was to formulate public policy based on this figure, the results could be far from what they anticipated. Perhaps Peter should report the *median* income. He can then rest assured that one-half the people earned incomes above that level and one-half earned incomes below it.

In still another situation, the *mean* would serve as the most useful measure of the average. Assume the Land's End wishes to market a new camping tent. The dimensions of the tent would depend, among other things, on the average height of adults. Experience has shown that the mean serves quite well as measure of central tendency when dealing with products that are built to conform to people's height. The size of doorways, counter tops in homes and retail businesses, and much of the furniture that is manufactured is based on mean heights.

3.5 Measures of Dispersion for Ungrouped Data

In the efforts to describe a set of numbers, it has been seen that it is useful to locate the center of that data set. But identifying a measure of central tendency is not always sufficient. It often proves helpful if statisticians or managers can also cite the extent to which the individual observations are spread out around that center point.

Take the three small data sets shown here:

Data Set 1	Data Set 2	Data Set 3
0, 10	4, 6	5, 5

These data sets are not similar, but all three average exactly five. If not seeing the observations in each data set, and hearing only what the averages were,

somebody might presume a similarity. To provide a more complete description of the data sets, a measure of how spread out the observations are from that mean of 5 is needed. A **measure of dispersion** indicates to what degree the individual observations are dispersed or spread out around their mean.

3.5.1 *Range*

A simple, but not practically useful, measure of dispersion is the range. The **range** is the difference between the highest observation and the lowest observation. Its advantage is that it is easy to calculate and gives at least some impression as to the makeup of the data set. Its disadvantage is that it takes only two of the, perhaps, hundreds of observations in the data set into consideration in its calculation. The rest of the observations are ignored.

3.5.2 *Mean Absolute Deviation*

It might seem that a practical approach to measuring the dispersion in a data set is to simply calculate the average amount by which the observations vary from the mean. This is called the **average deviation (AD).** Formula (3.10) shows how the AD can be calculated.

$$AD = \frac{\sum\left(X_i - \overline{X}\right)}{n}. \tag{3.10}$$

Example 3.7. Professor Willey Doezoff, a long-time resident of the statistics department, gave a quiz to his introductory statistics class last week. Eight of the brightest students scored 73, 82, 64, 61, 68, 52, and 73. The average is $\overline{X}$ = 67, and is used to calculate the AD in the manner shown in Table 3.5.

Table 3.5 – Grades for Professor Doezoff's Stat Class

X_i	$X_i - \overline{X}$
73	73 – 67 = 6
82	82 – 67 = 15
64	64 – 67 = – 3
61	61 – 67 = – 6
63	63 – 67 = – 4
68	68 – 67 = 1
52	52 – 67 = – 15
73	73 – 67 = 6
Total	0 = $\sum\left(X_i - \overline{X}\right)$

The result is an average deviation of 0. This happens because the pluses and minuses cancel each other out. The *average deviation* (AD) of a data set is *always* zero. A solution to this enigma is the **mean absolute deviation (MAD).** MAD takes the absolute value of the differences, so the negatives do not cancel out the positives.

$$MAD = \frac{\sum |X_i - \bar{X}|}{n}. \qquad (3.11)$$

Thus, for the example 3.7, MAD = 7. This value serves as an indication of the amount by which the individual observations are dispersed around their mean of 67: *the higher the MAD the more the dispersion*.

MAD is a "quick and dirty" method of measuring the amount of deviation in a data set.

3.5.3 *Variance and the Standard Deviation for a Population*

The two most important measures of dispersion of the data: the *variance* and the *standard deviation*.

The **variance** is the mean of the squared deviations from the mean. It means (1) finding the amount by which each observation deviates from the mean, (2) squaring those deviations, and (3) finding the average of those squared deviations.

The *variance for a population* is

$$\sigma^2 = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_N - \mu)^2}{N} = \frac{\sum (X_i - \mu)^2}{N}, \qquad (3.12)$$

or

$$\sigma^2 = \frac{\sum X^2 - N\mu^2}{N}, \qquad (3.12a)$$

where X_1, X_2,.., X_N are the individual observations; μ is the population mean; N is the number of observations.

The **standard deviation** is the square root of the variance.

$$\sigma = \sqrt{\sigma^2}. \qquad (3.13)$$

The concept of the standard deviation is often quite important in business and economics. For example, in finance the standard deviation in used as a measure of the risk associated with various investment opportunities. Generally, the higher the standard deviation of the rate of return of a particular investment, the greater the degree of risk.

Example 3.8. Markus Boggs is manager of Nest Egg Investments, a financial planning firm that assists individuals in setting up their personal

portfolios. Recently, Markus was interested in the rates of return over the past five years of two different mutual funds. Megabucks, Inc. showed rates of return over the five-year period of 12, 10, 13, 9 and 11 percent, while Dynamic Corporation yielded 13, 12, 14, 10, and 6 percent, a client approached Boggs and expressed an interest in one of these mutual funds. Which one should Boggs choose for his client?

Solution. The both funds offer an average return of 11 percent. Since that, the safer investment is the one with the smaller degree of risk as measured by the standard deviation. Boggs calculates the variance and takes square root to get the standard deviation for each stock. For Megabucks, it becomes

$$\sigma^2 = \frac{(12-11)^2 + (10-11)^2 + (13-11)^2 + (9-11)^2 + (11-11)^2}{5} = 2.$$

The standard deviation is

$$\sigma = \sqrt{2} = 1.41\%.$$

For Dynamics,

$$\sigma^2 = \frac{(13-11)^2 + (13-11)^2 + (14-11)^2 + (10-11)^2 + (6-11)^2}{5} = 8.$$

The standard deviation is therefore

$$\sigma = \sqrt{8} = 2.83\%.$$

Interpretation. Since Megabucks exhibits less variability in its returns and offers the same rate of return in average as doe Dynamics, Megabucks represents the safer of the two investments and is therefore the preferred investment opportunity.

3.5.4 *Variance and the Standard Deviation for a Sample*

The variance and standard deviation for a sample still represent measures of dispersion around the mean. They are calculated quite similarly to those for a population. The **sample variance** is

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}, \tag{3.14}$$

or

$$s^2 = \frac{\sum X^2 - n\bar{X}^2}{n-1}, \tag{3.14a}$$

where $n - 1$ are degrees of freedom.

The **sample standard deviation** is

$$s = \sqrt{s^2}\ . \tag{3.15}$$

The *number of degrees of freedom* in Formula (3.14) in any statistical operation is equal to the number of observations minus any constraints placed on those observations. A *constraint* is any value that must be computed from the observations.

Another reason for deviation by "n – 1" in (3.14) is that a sample is generally a little less dispersed that the population from which it was taken. There is therefore a tendency for the sample standard deviation s to be a little less that the population standard deviation σ.

Formulas (3.12a) and (3.14a) reduces the required arithmetic, but provide no insight into the nature of a variance.

3.6. Calculating the Variance and Standard Deviation with Grouped Data

If data are grouped into a frequency table, the variance and standard deviation can be calculated as

$$s^2 = \frac{\sum fM^2 - n\overline{X}^2}{n-1}. \tag{3.16}$$

And

$$s = \sqrt{s^2}\ . \tag{3.17}$$

Example 3.9. The flight director for P&P requires information regarding the dispersion of the numbers of passengers. Decisions regarding scheduling and the most efficient size of planes to use depend on the fluctuation in the passenger load. If this variation in number of passengers is large, bigger planes may be needed to avoid overcrowding on those days when the passenger load is extensive. The frequency table for P&P appeared as

Table 3.6 – Frequency Distribution for Passengers

Class (passengers)	Frequency (f) (days)	Midpoint (M)	fM	M^2	$f M^2$
50 to 59	3	54.5	163.5	2916	5,832
60 to 69	7	64.5	451.5	3969	19,845
70 to 79	18	74.5	1341.0	5184	72,576
80 to 89	12	84.5	1014.0	6561	118,098

Continuation of Table 3.6

Class (passengers)	Frequency (*f*) (days)	Midpoint (*M*)	*fM*	M^2	fM^2
90 to 99	8	94.5	756.0	8100	56,700
100 to 109	2	104.5	208.0	9801	39,204
$\sum$	50		3935.0		312,255

Solution. Given that, the mean was calculated in an earlier example 3.6 as

$$\overline{X}_g = \frac{\sum fM}{n} = \frac{3915}{50} = 78.3.$$

Formulas (3.16) and (3.17) give

$$s^2 = \frac{312{,}255 - 50(78.3)^2}{49} = 116.54 \text{ passengers squared,}$$

$$s = \sqrt{116.54} = 10.80 \text{ passengers.}$$

Interpretation. The flight director can now decide if the planes currently in use can accommodate fluctuations in passengers levels as measured by a standard deviation of 10.8. If not, perhaps larger planes will be used to accommodate any overflow that might otherwise occur on those days with heavy traffic.

3.7 Other Measures of Dispersion

There are other ways the dispersion of a data set might be measured. These additional measures of dispersion, which often prove to be quite serviceable, are **quartiles**, **deciles**, and **percentiles**.

Every data set has **three quartiles**, which divide it into four equal parts. As seen in Figure 3.1, if the horizontal line can be thought of as a data set arranged in an ordered array, three quartiles can be identified, which together produce four separate parts or subsets of equal size in the data set.

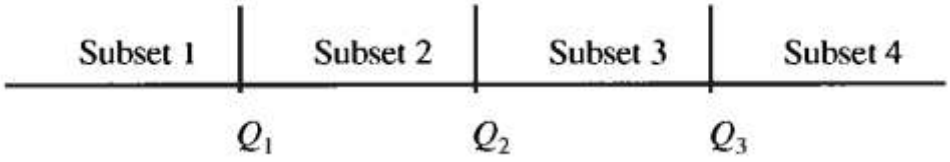


Figure 3.1 - Three Quartiles Produce Four Equal Subsets

The *first quartile* (P_{25}) is that value below which, at most, 25 percent of the observations fall, and above which the remaining 75 percent can be found. The

second quartile (P_{50}) is right in the middle. One-half the observations are below it and one-half the observations are above it; in this sense, it is the same as the median. The *third quartile* (P_{75}) is that value below which, at most, 75 percent of the observations are located, and above which the remaining 25 percent can be found.

The determination of quartiles is often useful. Many graduate schools, for **example**, will admit only those students in the top 25 percent (third quartile) of their applicants, etc.

Deciles separate a data set into 10 equal subsets, and percentiles produce 100 parts. The *first decile* (P_{10}) is that observation below which, at most, 10 percent of the observations are found while the remaining 90 percent are located above it. The *first percentile* (P_1) is that value below which no more than 1 percent of the observations are located, and the rest are above it, etc. Each data set has **9 deciles** and **99 percentiles**.

The **location of the *P*th percentile** is found as

$$Lp = (n+1)\frac{P}{100}\ , \qquad (3.18)$$

where Lp is the location in an ordered array of the desired percentile; n is the number of observations; P is the desired percentile.

Example 3.10. In Table 3.7 the observations for the number of shares for 50 stocks traded on the New York Stock Exchange are shown (in an ordered array). Assume that somebody wishes to calculate the 25^{th} percentile, P_{25}, for the stocks.

Table 3.7 – Numbers of Shares Traded on the NYSE (in 100's)

3	10	19	27	34	38	48	56	67	74
4	12	20	29	34	39	48	59	67	74
7	14	21	31	36	43	52	62	69	76
9	15	25	31	37	45	53	63	72	79
10	17	27	34	38	47	56	64	73	80

First the location of the 25^{th} percentile must be found in an ordered array.

$$L_{25} = (50+1)\frac{25}{100} = 12.75.$$

The resulting value of 12.75 tells that the 25^{th} percentile is located 75 percent of the way between the 12^{th} observation of 20 and the 13^{th} observation of

21, or P_{25} = 20+(0.75)(21-20) = 20.75, or 2,075 shares since the data were originally expressed in hundreds of shares. Thus, 25 percent of the observations are below 20.75, and the remaining 75 percent are above 20.75.

The **interquartile range (IQR)** is a unique measure of dispersion. The IQR is the difference between the first quartile and the third quartile. That is, $P_{75} - P_{25}$. One-half of the observations lie within this range. It consists of the middle 50 percent of the observations in that it cuts off the lower 25 percent and the upper 25 percent of the data points.

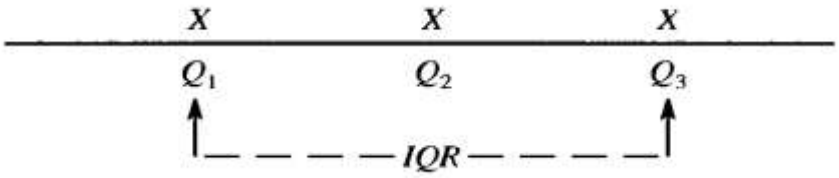


Figure 3.2 - The Interquartile Range

As a result, the IQR provides a measure of dispersion that is not heavily influenced by a few extreme observations.

3.8. Common Uses for the Standard Deviation

There are at least two additional applications for the standard deviation. The first one involves Chebyshev's Theorem and applies to any distribution of observations, while the second is appropriate only if the distribution meets specific conditions of normality.

3.8.1. *Chebyshev's Theorem*

The theorem was formulated by the Russian mathematician P.L.Chebyshev (1821 - 1894). It states that *for **any** data set, at least* $1 - 1/K^2$ *percent of the observations lie within K standard deviations of the mean, where K is any number greater than* 1. Chebyshev's Theorem is expressed as

$$1-\left[\frac{1}{K^2}\right]. \qquad (3.19)$$

Thus, if statistician forms an interval from K = three standard deviations above the mean to three standard deviations below the mean, then at least $1-\frac{1}{3^2}=88.89\%$ of all observations will be within that interval.

Example 3.11. Passengers for P&P averaged 78.3 per day with (see Example 3.9) a standard deviation of 10.8. In order to schedule times for a new

route P&P opened, management wants to know how often the number of passengers is within K = two standard deviations of the mean, and what that interval is.

Solution. Moving two standard deviations (2 x 10.8) = 21.6 passengers above and below the mean of 78.3, an interval of (78.3 – 21.6) = 56.7 to (78.3 + 21.6) = 99.9 passengers will be found. The management can be certain that at least $1-\frac{1}{2^2}=75\%$ of the time, the number of daily passengers was between 56 and 99.

Interpretation. On at least 75 percent of the days, the number of passengers was between 56 and 99. This provides the management of P&P with valuable information regarding how many passengers to prepare for in-flight operations.

3.8.2. *The Normal Distribution and the Empirical Rule*

The concept of a normal distribution is commonly encountered in statistical analysis and is of considerable importance. A **normal distribution** is a distribution of continuous (not discrete) data that produces a bell-shaped, symmetrical curve like that shown in Figure 3.3.

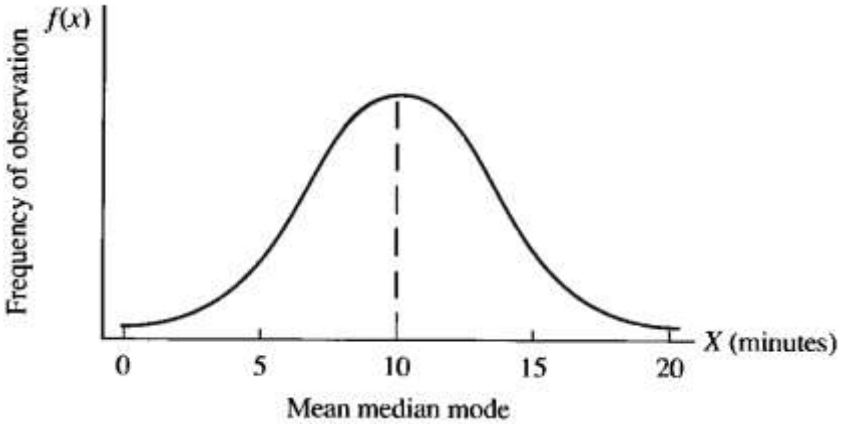


Figure 3.3 – A Normal Distribution

In a normal distribution, the *mean*, *median*, and *mode* are all *equal*. Of importance, one-half of the observations are above the mean and one-half are below it. This means that one-half of the area under the curve is to the left of the mean, and one-half of the area under the curve is to the right of the mean.

The **Empirical Rule** tells that if include all normally distributed observations within one standard deviation of the mean (one standard deviation above the mean and one standard deviation below the mean) then 68.3 percent of all observations will be encompassed. Moving more than one standard deviation

above and below the mean, a larger percentage of observations will be encompassed. The Empirical Rule specifies that

68.3 percent of the observations lie within plus or minus *one standard deviation* of the mean.

95.5 percent of the observations lie within plus or minus *two standard deviations* of the mean.

99.7 percent of the observations lie within plus or minus *three standard deviations* of the mean. Such observations are rarity and happen less that 1 percent of the time if the data are normally distributed.

The Empirical Rule also applies to *sample data*. Thus, for example, $\overline{X} \pm 2s$ produces a range that includes 95.5 percent of all observations in the sample. It is also important to remember that the Empirical Rule describes the total area under the normal curve that is found within a given range.

If the observations are highly dispersed, the bell-shaped curve will be flattened and spread out.

Example 3.12. There are a large number of observations for the time, in minutes, that it takes skiers to complete a particular run. The modal observation (μ = 10 in this case) is the one occurring with the greatest frequency and is therefore at the peak of the distribution. Given the skiers' times, one standard deviation (σ = 2 minutes) above and below the mean of 10 yields a range of 8 to 12 minutes. Two standard deviations (σ = 4 minutes) yields a range of 6 to 14 minutes, and three standard deviations (σ = 6 minutes) – a range of 4 to 16 minutes. This is shown in Figure 3.4.

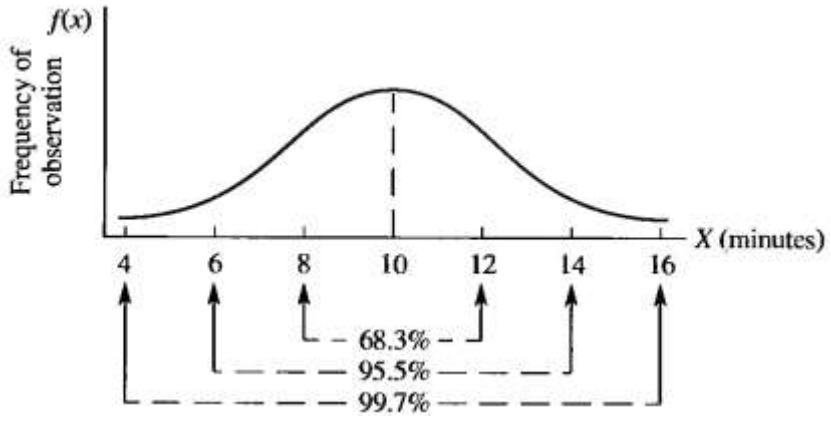
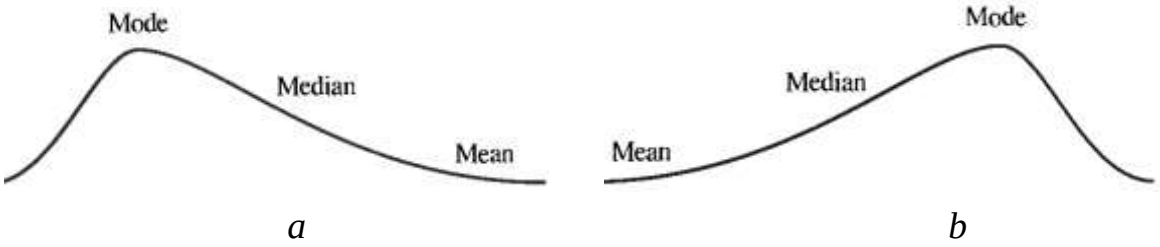


Figure 3.4 – Normally Distributed Times of 1,000 Skiers

According to the Empirical Rule, 997 of the 1,000 skiers took between 4 min and 16 min to complete the run. Thus, only 3 of the 1,000 skiers were either very good skiers and took less that 4 min or were lousy and took more than 16 min.

3.8.3. *Skewness and Coefficient of Variation*

Not all distributions are normal. Some are skewed right (Fig.3.5a) or left (Fig.3.5b). In Figure 3.5, distribution curves for people's weights can be found.



a *b*

Figure 3.5 – Skewed Distribution of People's Weights:
a – Distribution skewed to the right; *b* - Distribution skewed to the left

In both cases in Figure 3.5, the mode is, by definition, that observation occurring with the greatest frequency. It is therefore at the peak of the distribution. The mean is most affected by extreme observations. Therefore it is pulled in the direction of skewness more than is the median, which lies between the mean and the mode. These conditions of skewness are significant and can be measured by the **Pearsonian coefficient of skewness**.

$$P = \frac{3(\overline{X} - median)}{s}. \qquad (3.20)$$

If $P<0$, the data are skewed left; if $P>0$, the are skewed right; if $P = 0$, they are normally distributed.

The **coefficient of variation** serves as a relative measure of dispersion. The coefficient of variation assesses the degree of dispersion of a data set relative to its mean.

$$CV = \frac{s}{\overline{X}}(100). \qquad (3.21)$$

Chapter Checklist

After studying this chapter, can you

1. Distinguish conceptually between measures of central tendency and measures of dispersion?

2. Calculate the mean, median, and mode from both ungrouped and grouped data?

3. Calculate the weighted mean and the geometric mean, and identify when each should be used?

4. Compute percentiles and the variance and standard deviation for ungrouped and grouped data?

5. Explain what the variance and standard deviation are measuring?

6. Apply Chebyshev's theorem and the Empirical Rule to a data set?

7. Explain what is meant by a normal distribution, and draw a normal curve?

4 PROBABILITY DISTRIBUTIONS

4.1 Principles of Probability

Statisticians often take samples for the purpose of gaining knowledge regarding the world around them. On the basis of these samples, they can frequently estimate the probability that specific evens will occur. **Probability** is the numerical likelihood of the occurrence of an uncertain event.

4.1.1. *Experiments, Outcomes and Sets*

The probability of an event is measured by values between 0 and 1. The probability of a certainty is 1. The probability of impossibility is 0, i.e.

$$P \text{ (certain event)} = 1,$$

$$P \text{ (impossible event)} = 0.$$

The process that produces the event is called an **experiment**. An *experiment* is well-defined action leading to a single, well-defined result. That result is called the **outcome**. A **set** is any collection of objects. The objects in a set are its **elements** or **members**. The set of all possible outcomes for an experiment is the **sample space** (SS). For **example**, the sample space for the experiment of flipping a coin is

$$\text{SS} = (\text{heads, tails}).$$

The occurrence of either a head or a tail is a certainty. So, the probability that a head or a tail occurs equals 1. That is,

$$P \text{ (head or tail)} = 1.$$

Properties of probability.

1. The probability that some uncertain event will occur is between 0 and 1. If E_i, is any given event, then it can be said that

$$0 \leq \text{P}\ (E_i) \leq 1.$$

For **example**, (1) the probability that the sun will rise tomorrow is very high – quite close to 1; (2) the probability a student will pass this course without studying it, at the other extreme, close to zero.

2. If E_i is an event representing some element in a sample space, then

$$\Sigma \text{P}(E_i) = 1.$$

There are only three generally accepted ways to approach probability: (1) the relative frequency (or posterior) approach, (2) the subjective approach, and (3) the classical (or a priori) approach.

4.1.2. *Probability Approaches*

The **relative frequency approach** uses past data that have been empirically observed. It notes the frequency with which some event has occurred in the past and estimates the probability of its reoccurrence on the basis of these historic data. The probability of an event based on the relative frequency approach is determined by

$$P(E) = \frac{\text{Number of times the event has occurred in the past}}{\text{Total number of observations}}. \qquad (4.1)$$

For **example**, assume that during the last calendar year there were 50 births at a local hospital. Thirty-two of the little new arrivals were baby girls. The *relative frequency approach* reveals that the probability that the next birth (or any randomly selected birth) is a female is determined as

$$P(\textit{female}) = \frac{\text{Number of females born last year}}{\text{Total number of births}} = \frac{32}{50}.$$

In many instances past data are not available. It is therefore not possible to calculate probability from previous performance. The only alternative is to estimate probability on the basis of our best judgment. The **subjective approach** requires the assignment of the probability of some event on the basis of the best available evidence. *The subjective approach is used when we want to assign probability to an event that has never occurred*. For **example**, the probability, that a woman will be elected president of the United States.

Of the three methods of assessing probability, the classical approach is one most often associated with gambling and games of chance. The classical probability of an event *E* is determined as

$$P(E) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible outcomes}}. \qquad (4.2)$$

Example 4.1. The probability of getting a head in the single flip of a coin is 1/2, i.e. using formula (4.2)

$$P(\textit{head}) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible outcomes}} = \frac{1}{2}.$$

There is only one way that the event can occur (you get a head), and only two possible outcomes (a head or a tail).

Example 4.2. The probability of rolling a "3" with a six-sided die is

$$P(3) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible outcomes}} = \frac{1}{6}.$$

There is only one way that the event can occur (you roll a "3"), and six possible outcomes. Although the probability of rolling a "3" is 1/6, this is not suggest, that of every six rolls, one is a "3". Instead, the implication is that if the die is rolled a large number of times (technically, an infinite number), one – sixth of the rolls produce a "3".

Classical probability involves the determination of the probability of some event in an a priori manner (*before the fact*). Thus, *before* drawing a card from a deck of 52 cards, it can be determined that the probability of drawing an ace is

$$P(ace) = \frac{\text{Number of ways the event can occur}}{\text{Total number of possible outcomes}} = \frac{4}{52}.$$

4.1.3. *Relationships between Events*

Two events are said to be **mutually exclusive** if the occurrence of one event precludes the occurrence of the other: *If one vent happens, the other cannot.*

An **example** of mutually exclusive events is flipping a head or a tail in a single coin flip. If the head occurs, the tail cannot. Other example is rolling a 3 or even number, drawing from a 52-card deck one card that is a queen or an ace, etc.

Collectively exhaustive events are those events that consist of all possible outcomes of an experiment. In a case of rolling a die, the collectively exhaustive events are 1, 2, 3, 4, 5, and 6. *The collectively exhaustive events for an experiment constitute its sample space.*

Independent events are events which independent if the occurrence of one event has no effect on the probability that the second will occur. For **example**, since the result of drawing a card from a deck has no impact on whether it rains tomorrow, the result of the draw is independent of tomorrow's weather.

Complementary events are events such that if one does not occur, the other must. The complement of A is written $\bar{A}$, and is referred to as "not A". For **example**, if A is rolling an even number with a die (2, 4, or 6), the *complement* is rolling an odd number (1, 3, or 5). *Complementary events are also collectively exhaustive*, because if A does not occur, $\bar{A}$ must occur. Thus,

$$P(A) + P(\bar{A}) = 1,$$
$$P(A) = 1 - P(\bar{A}).$$

4.1.4. *Unions, Intersections and Venn Diagrams*

The **intersection** of *A* and *B*, written $A \cap B$ and read as "*A* intersection *B*", consists of those elements that are in common to *both A* and *B*. Given $A \cap B$, the events *A* and *B* are called **joint events**. The probability that two or more events will all occur is called **joint probability**.

A *Venn diagram* is a useful tool to portray the relationship between sets. It was developed by an English mathematician John Venn (1834-1923). It's shown in Figure 4.1.

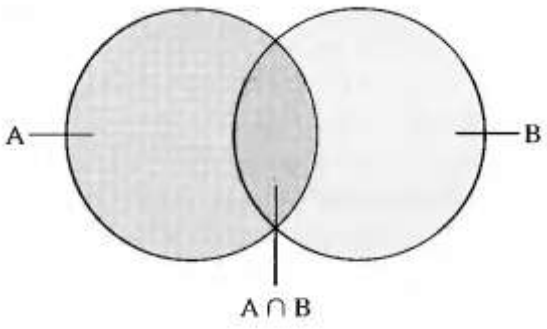


Figure 4.1 – Venn Diagram

The **union** of *A* and *B*, written *A*U*B* and read as "*A* union *B*", consists of those elements that are in either *A* or *B* or both.

Example 4.3. Given a deck of 52 cards, set *A* is all hearts and set *B* is all kings. Identify $A \cap B$ and *A*U*B*.

Solution. The two sets are shown in the Venn diagram. See Figure 4.2.

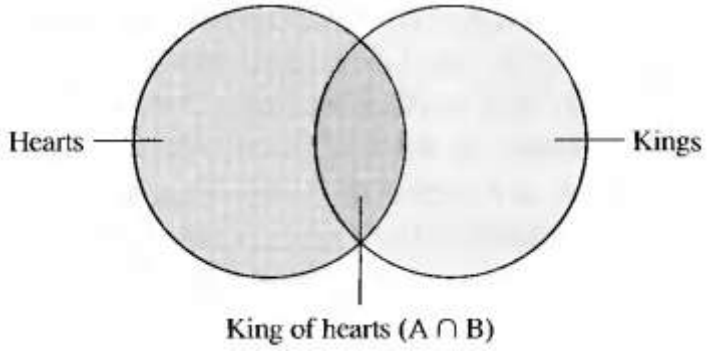


Figure 4.2 – A Venn Diagram for Deck of Cards

*A*U*B* is the set of all cards that are in either set *A* or set *B*, and consists of all hearts (including the king) and all kings (including the heart).

$A \cap B$ contains only those elements that are in both *A* and *B*, i.e. those elements that are both hearts and kings.

Interpretation. The event *AUB* is satisfied if *either a king or a heart* is drawn from the deck. The joint event $A \cap B$ is satisfied *only* if *the king of hearts* is drawn.

4.1.5. *Frequency Tables and Probability Tables*

Frequency tables and probability tables are handy devices for summarizing data that can be used to compute probabilities.

For **example**, consider the data set collected by Sales Director for E-Z-C Eye Care Center. The style and size of frames for eyeglasses were noted for the last 100 sales. The results are shown in Table 4.1. The probability table constructed from the frequency table 4.1 and is shown in Table 4.2.

Table 4.1 – Frequency Table for E-Z-C Eye Care Center

Size	Frame Style			Total
	Plastic	Wire	Composite	
Large	12	8	5	25
Medium	23	31	1	55
Small	6	6	8	20
Total	41	45	14	100

Table 4.2 – A Probability Table for sales for E-Z-C Eye Care Center

Size	Frame Style			Total
	Plastic	Wire	Composite	
Large	12/100 = 0.12	8/100 = 0.08	5/100 = 0.05	25/100 = 0.25
Medium	23/100 = 0.23	31/100 = 0.31	1/100 = 0.01	55/100 = 0.55
Small	6/100 = 0.06	6/100 = 0.06	8/100 = 0.08	20/100 = 0.20
Total	41/100 = 0.41	45/100 = 0.45	14/100 = 0.14	100/100 = 1.00

4.1.6. *Two Rules of Probability*

The **rule of multiplication** is used to determine **joint probability** of "***A* and *B***" ($A \cap B$). It is states that

1. If *A* and *B* are *independent* events, the probability of the event *A* must be multiplied by the probability of event *B*.

$$P(A \text{ and } B) = P(A) \times P(B) , \qquad (4.3)$$

2. If *A* and *B* are *dependent* events, the probability of the event *A* must be multiplied by the probability of event *B*, given event *A* has already occurred. This is based on the principle of *conditional probability* and can be written as $P(B \mid A)$, and read as the "probability of *B* given *A*". in general, conditional probability can be computed as

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \quad . \tag{4.4}$$

Then *the Multiplication Rule for dependent events A* and *B* can be written in (4.5)

$$P(A \text{ and } B) = P(A) \text{ x } P(B \mid A). \tag{4.5}$$

The **rule of addition** is used to find the probability of ***A* or *B*** (*A*U*B*). This rule states that

1. If *A* and *B* are *mutually exclusive* events the probability of event *A* must be added to the probability of event *B*. The joint probability in this case is zero, i.e. $P(A \text{ and } B) = 0$.

$$P(A \text{ or } B) = P(A) + P(B). \tag{4.6}$$

2. If *A* and *B* are *not mutually exclusive* (both can occur at the same time) events the probability of event *A* must be added to the probability of event *B* and subtract the joint probability of events *A* and *B*. See formula (4.7).

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{4.7}$$

4.1.7. *Baye's Rule and Conditional Probability*

The a **priori probability** is the probability estimate before the event occurs, and may be subject to change after further study. **Posterior probabilities** are conditional probabilities based on additional information.

For **example**, the probability of drawing a jack from a deck is 4/52 (a priori probability). After it was learned that the card was a face card, the probability is $\mathbf{P}(J|F) = 4/12$ (posterior probability).

For some events *A* and *B* **Baye's Rule** states that

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(B)} \quad , \tag{4.8}$$

where

$$P(B) = P(A) \times P(B|A) + P(\overline{A}) \times P(B|\overline{A}). \quad (4.9)$$

4.2 Probability Distributions

A **random variable** is a variable whose value is the result of a random event. Number of units sold, daily levels of output, and the height of customers are **examples**.

A **probability distribution** is a display of all possible outcomes of an experiment along with the probabilities of each outcome. Probability distributions are based on the outcomes of random variables. There are several different *types of probability distributions*:

- Binomial distributions.
- Poisson distributions.
- Hypergeometric distributions.
- Uniform distributions.
- Exponential distributions.
- Normal distribution.

The probability that the random variable X can take on some specific value, x_i is written

$$P(X = x_i).$$

It should be noted that

$$0 \leq P(X = x_i) \leq 1 \text{ and } \Sigma P(X = x_i) = 1.$$

The use of a discrete random variable leads to the formation of a **discrete probability distribution**. The number of customers, the number of units sold are **examples**.

The use of continuous random variable leads to formation of a **continuous probability distribution.** A continuous probability distribution is usually the result of measurement. There are no gaps in the observations because no matter how close two observations might be, a third could be found that would fall between the first two.

4.2.1. *The Mean and the Variance of Discrete Random Variables*

The *mean* of probability distribution is called the **expected value** of the random variable. The *expected value* $E(X)$ of a discrete random variable X is the

weighted mean of all possible outcomes, in which the weights are the respective probabilities of those outcomes.

$$\mu = E(X) = \Sigma[(x_i)P(x_i)], \qquad (4.10)$$

where x_i are the individual outcomes, $P(x_i)$ – probability of the proper individual outcome.

Example 4.4. The expected value of the experiment of rolling a die is

$$\mu = E(X) = [1 * 1/6] + [2 * 1/6] + [3 * 1/6] + [4 * 1/6] + [5 * 1/6] + [6 * 1/6] = 3.5$$

In practice, the larger number of rolls, the closer the mean is to 3.5.

The **variance of probability distribution** is the mean of squared deviations from the mean. The variance σ^2 may be written as

$$\sigma^2 = \Sigma[(x_i - \mu)^2 P(x_i)], \qquad (4.11)$$

or as

$$\sigma^2 = \Sigma[(x_i)^2 P(x_i)] - \mu^2 . \qquad (4.12)$$

4.2.2. *The Binomial Distribution*

A **binomial distribution** based on the Bernoulli process and it must fit certain conditions:

1. There must be only two possible outcomes. One is identified as a success, the other as a failure.

2. The probability of success, π, remains constant from one trial to the next, as does probability of a failure, $1 - \pi$.

3. The probability of a success in one trial is totally independent of any other trial.

4. The experiment can be repeated many times.

If the probability that any given trial will result in a success is known, it is possible to estimate how many successes there will be in a given number of trials.

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x} = {}_nC_x(\pi)^x(1-\pi)^{n-x}, \qquad (4.13)$$

where n is the number of trials; π is the probability of a success on any given trial; x is given number.

The *mean* and *variance for binomial distribution* are calculated as

$$\mu = n\pi \quad \text{and} \quad \sigma^2 = n\pi(1 - \pi). \qquad (4.14)$$

It means that on average there are $n\pi$ successes out of n trials.

Figure 4.4 shows the **probability mass function** (PMF), which assigns a probability to each value of *X*, for binomial distributions with different values of π and *n*.

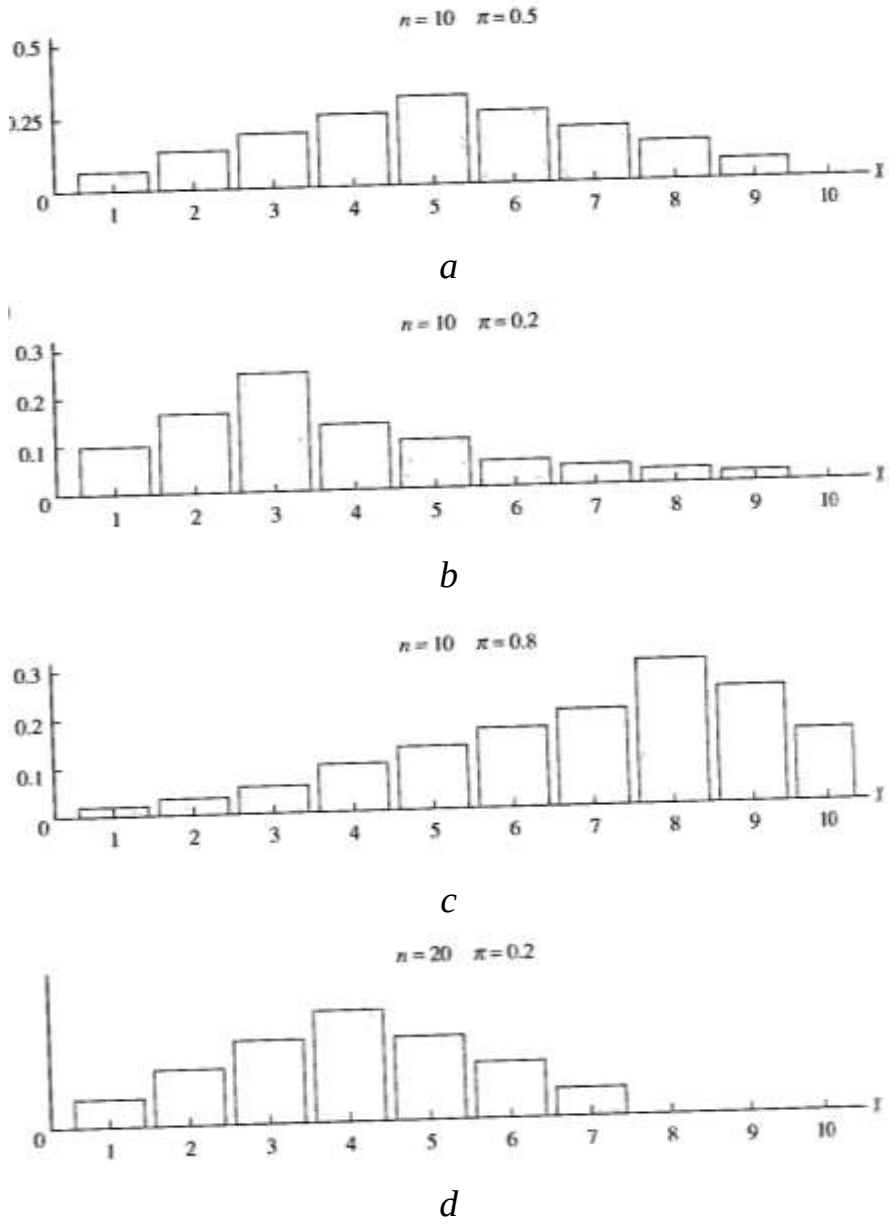


Figure 4.3 – Probability Mass Functions: a – the distribution is symmetrical if π = 0.5; b – the distribution is skewed right if π < 0.5; c – the distribution is skewed left if π > 0.5; d – a distribution close to normality since *n* is bigger

4.2.3. *The Poisson Distribution*

Developed by the French mathematician Simeon Poisson (1781 - 1840), the **Poisson distributions** *is a discrete probability distribution that measures the probability of a random event over some interval of time and space.* The Poisson distribution is also useful as an approximation for binomial probabilities.

It is often used to describe the number of arrivals of customers per hour, the number of industrial accidents each month, or the number of machines that break down and are awaiting repair. In each of these cases, the random variable (customers, accidents, machines) is measured per unit of time and space (distance).

For application of the Poisson distribution two *assumptions* are necessary:

1. The probability of the occurrence of the event is constant for any two intervals of time or space.

2. The occurrence of the event in any interval is independent of the occurrence in any other interval.

The *Poisson probability function* can be expressed as

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}, \tag{4.15}$$

where x is number of times the event occurs; μ is the mean of occurrence per unit of time or space; e = 2.71828 is the base of natural logarithm system.

4.2.4. *The Hypergeometric Distribution*

The binomial distribution is appropriate only if the *probability* of a success remains *constant* for each trial. This will occur if the sampling is done from an infinite (or very large) population. The distinction between the binomial and hypergeometric distributions lies in the *population size*, especially as it relates to the size of the sample.

If the probability of a success is not constant, if a sample is selected without replacement from a known population and contains a relatively large proportion of the population the **hypergometric distribution** is particularly useful.

The *probability function for the hypergeometric distribution* is

$$P(x) = \frac{{}_rC_x \;\; {}_{N-r}C_{n-x}}{{}_NC_n}, \tag{4.16}$$

where x is number in the sample identified as a success; N is the population size; r is the number in the population identified as a success; n is a sample size.

Example 4.5. Assume a racing stable has N = 10 horses, and r = 4 of them have a contagious disease. What is the probability of selecting a sample of n = 3 in which there are x = 2 diseased horses?

$$P(X = 2) = \frac{{}_4C_2 \;\; {}_{10-4}C_{3-2}}{{}_{10}C_3} = \frac{6 \times 6}{120} = 0.30.$$

There is a 30 percent probability of selecting three racehorses, two of which are ill.

The problem concerning diseased racehorses a hypergeometric coz there are only two possible outcomes: the horses are either (1) diseased or (2) not diseased.

4.2.5. *The Uniform Distribution*

A **uniform distribution** is a continuous distribution in which every possible outcome has an equal chance of occurring. The experiment of rolling a die is an **example**. The probabilities of all possible outcomes are all 1/6 (numbers 1 to 6). The number that actually occurs in a single roll is an observation, while the outcome of the roll is a random variable and can take any value from 1 to 6.

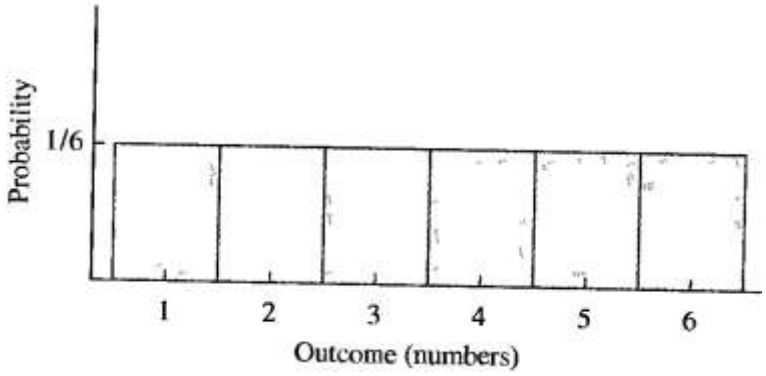


Figure 4.4 – A Uniform Probability Distribution for Rolling a Die

The **mean** of a uniform distribution is halfway between its two end points (*a* – minimum or a lower end point of distribution, *b* – maximum or an upper end point of the distribution)

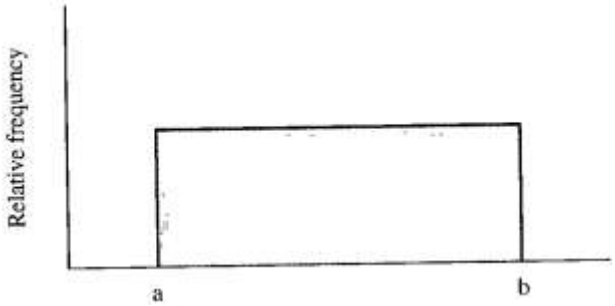


Figure 4.5 – A Uniform Probability Distribution

$$\mu = \frac{a+b}{2}. \tag{4.17}$$

The **standard deviation** of a uniform distribution is

$$\sigma = \frac{b-a}{\sqrt{12}}. \tag{4.18}$$

The *total area* under the curve must equal 1 or 100 percent. Since the area is height times width, the height is

$$Height = \frac{Area}{Width} = \frac{1}{b-a}, \tag{4.19}$$

where $b - a$ is the width or range of the distribution.

The probability of a single observation between two values X_1 and X_2 can be determined as in Formula (4.20)

$$P(X_1 < X < X_2) = \frac{X_2 - X_1}{Range}. \tag{4.20}$$

4.2.6. *The Exponential Distribution*

The **exponential distribution** is a continuous distribution that measures the passage of time between events. It can measure the time lapse as

1. The time that passes between two successive arrivals (people, trucks, telephone calls, etc.);
2. The amount of time that it takes to complete one action, such as serving one customer, loading one truck, handling one call.

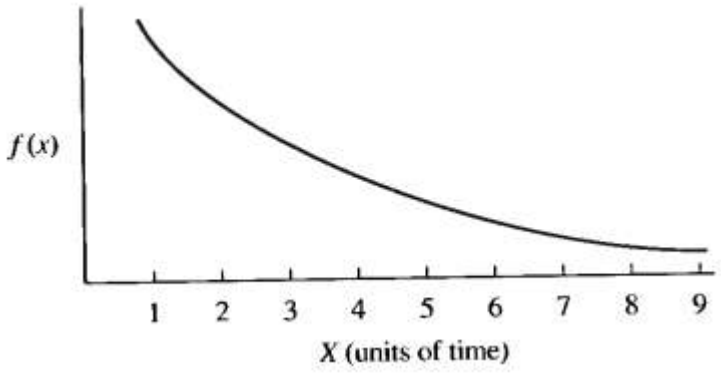


Figure 4.6 – An Exponential Probability Function

The Figure 4.6 shows *that the larger the value of the random variable, as measured in units of elapsed time, the less likely it is to occur*.

If the *arrival process is Poison-distributed*, then the *lapse of time* between arrivals *is exponentially distributed*. Let μ be the mean number of arrivals in a given time period, and let μ* be the mean lapse of time between arrivals. Then,

$$\mu^* = \frac{1}{\mu}. \tag{4.21}$$

For **example**, if an average of four trucks arrive every hour at the loading dock (μ = 4), then, on average, one truck arrives every 0.25 hour. That is,

$$\mu^* = \frac{1}{4} = 0.25 \text{ hour.}$$

The probability that no more than *t* units of time elapse between successive occurrences is

$$P(0 < X < t) = 1 - e^{-\mu t}, \tag{4.22}$$

where μ is the mean rate of occurrence; e = 2.71828 is the base of natural logarithm system.

The exponential distribution has a very common and useful application in business in the evaluation of waiting lines, or queues. Many business operations involve lines. Customers queue up for service, telephone calls come into a switchboard, trucks arrive at the loading dock, and machines break down and must be repaired.

Let Λ be the average rate at which units arrive for service per unit of time, μ the average number of units that can be serviced in that same time unit. Then a *queuing system* can be evaluated as

$$P_0 = 1 - \frac{\Lambda}{\mu},$$

where P_0 is the probability the system is idle; that is, there are no units in the system;

$$P_n = \left[\frac{\Lambda}{\mu}\right]^n (P_0),$$

where P_n is the probability *n* units are in the system;

$$L = \frac{\Lambda}{\mu - \Lambda},$$

where L is the average number of units in the system (those waiting for service plus the one receiving the service);

$$W = \frac{1}{\mu - \Lambda},$$

where W is the average time a unit spends in the system waiting for service and receiving that service (waiting time plus service time);

$$L_q = \frac{\Lambda^2}{\mu(\mu - \Lambda)},$$

where L_q is the average number of units waiting for service;

$$W_q = \frac{\Lambda}{\mu(\mu - \Lambda)},$$

where W_q is the average time spent in the queue waiting for service to begin.

Chapter Checklist

1. Define and give examples of experiments, outcomes, and sets.
2. Describe and provide examples of the three approaches to probability.
3. Define and give examples of events that are mutually exclusive, collectively exhaustive, independent, and complementary.
4. Use Venn diagrams to identify intersections and unions.
5. Construct frequency tables and probability tables
6. Clearly cite when and how the two rules of probability should be used.
7. Explain when and how conditional probability is required and how it relates to Baye's Rule.
8. Define and give an example of a random variable.
9. Describe, define and give an example of a probability distribution.
10. Distinguish between a discrete and a continuous distribution.
11. Calculate the mean and the variance of a probability distribution.
12. Identify and determine when to use each of the distributions discussed in this chapter.
13. Calculate the probabilities associated with each of the distributions discussed in this chapter.

5 SAMPLING DISTRIBUTIONS: AN INTRODUCTION TO INFERENTIAL STATISTICS

5.1 Introduction

Populations are usually too large to study in their entirety. It is necessary to select a representative sample of a more manageable size. This *sample is then used to draw conclusions* about the population in which statisticians are interested.

Inferential statistics involves the use of a statistic to form a conclusion, or inference, about the corresponding parameter.

The inferential process is extremely important in many statistical analyses. *The value of the statistics depends on the sample taken*. From any given population of size *N*, it is possible to get many different samples of size *n*, i.e. ${}_N C_n = \frac{N!}{n!(N-n)!}$. *Each sample may well have a different mean*. That is why some sampling error is likely to occur.

The difference between the population parameter (μ) *and the sample statistic used to estimate the parameter* ($\overline{X}$) is called **sampling error**. If a few extremely large observations are drawn for the sample, the sample mean will *overestimate* μ. If a few extremely small observations are drawn for the sample, the sample mean will *underestimate* μ. The size of the sampling error can never be calculated if the population mean is unknown.

A list of all possible values for a statistic and the probability associated with each value is called a **sampling distribution**.

Example 5.1. There is a population *N* = 4 incomes for four students. These incomes are \$100, \$200, \$300 and \$400. The mean income can is μ = \$250.

1. There are ${}_4 C_2 = \frac{4!}{2!(4-2)!} = 6$ possible samples of size *n* = 2 can be selected from this population (see Table 5.1).

2. Each sample has a different mean with exception of the third and fourth samples. The probability of selecting a sample that yields an $\overline{X}$ of 150 is

$$P(\overline{X} = 150) = \frac{1}{{}_N C_n} = \frac{1}{6}.$$

Table 5.1 – All Possible Samples of Size $n = 2$ from a population of $N = 4$

Sample	Sample Elements X_1	Sample Means $\bar{X}$
1	100, 200	150
2	100, 300	200
3	100, 400	250
4	200, 300	250
5	200, 400	300
6	300, 400	350

3. If a sample 5 was picked (see Table 5.1), the estimate of μ is $\bar{X}$ = 300, which is greater than the actual value for the population mean. So that, samples 1 and 2 in Table 5.1 produce an underestimate of the population mean.

4. Sampling distribution is presented in Table 5.2 and Figure 5.1

Table 5.2 – Sampling Distribution for Samples of Size $n = 2$ from a Population of $N = 4$ Incomes

Sample Mean $\bar{X}$	Number of Samples Yielding $\bar{X}$	Probability $P(\bar{X})$
150	1	1/6
200	1	1/6
250	2	2/6
300	1	1/6
350	1	1/6
Total	-	1

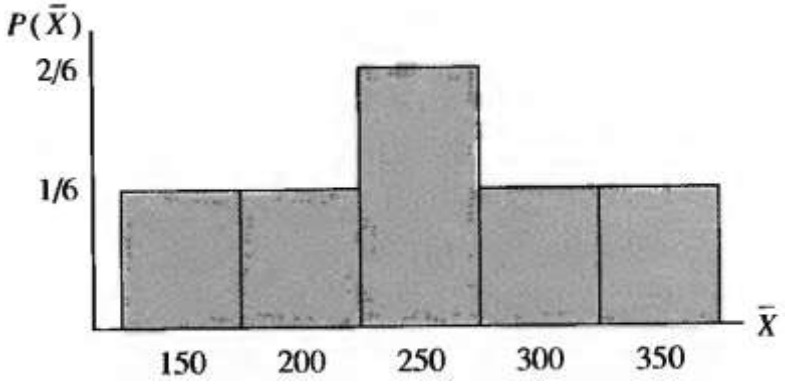


Figure 5.1 - Sampling Distribution for Samples of Size $n = 2$ from a Population of $N = 4$ Incomes

5.1.1. *The Mean of the Sample Means*

The sampling distribution for the sample means is merely a listing of all possible sample means. These sample means have a mean; it is called the "mean of all possible sample means", or the **grand mean**. *The grand mean will always equal the population mean*, i.e. $\overline{\overline{X}} = \mu$. It is calculated in the usual fashion: the individual observations (sample means) are summed, and the result is divided by the number of observations (samples).

$$\overline{\overline{X}} = \frac{\sum \overline{X}}{{}_{N}C_{n}} = \frac{\sum \overline{X}}{K}, \tag{5.1}$$

where N is the population size; n is the number of observations in a single sample; K is the number of possible samples.

5.1.2. *The Standard Error of the Sampling Distributions*

The sampling distribution of the sample means also has a **variance**. It measures the dispersion of the individual observations (sample means) around their mean (the grand mean). It can be found by

1. Determining the amount by which each of observations (sample means) differs from their mean (the grand mean).
2. Squaring those deviations.
3. Averaging the squared deviations by dividing by the number of sample means, K.

$$\sigma^2_{\overline{X}} = \frac{\sum\left(\overline{X} - \overline{\overline{X}}\right)^2}{K}. \tag{5.2}$$

The **standard error of the sampling distribution**, $\sigma_{\overline{X}}$, can be found as

$$\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}}. \tag{5.3}$$

This *standard error of the sampling distribution of sample means* (or just *standard error*) measures the dispersion of a set of a sample means around the grand mean. Thus, the **standard error** is the measure of the variation of the sample means around the grand mean. As such, it measures the tendency to suffer sampling error in the effort to estimate the parameter.

A close approximation of the variance and standard error can be found much more easily with

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}. \tag{5.4}$$

And $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}, \quad (5.5)$$

where σ^2 is the population variance.

The $\bar{X}$'s are less dispersed than original data. That is, $\sigma_{\bar{X}}$, *the standard error of the sampling distribution* of the $\bar{X}$'s, is *smaller than the standard deviation of the original population* σ, i.e. $\sigma_{\bar{X}} < \sigma$. So, as the number of a sample observations ***n*** *increases*, the spread in the sampling distribution, which is measured by the standard error, will *decrease*. Therefore, there is less chance for a larger error.

The **finite population correction** factor (**FPC**) must be used in calculating the standard error. *The FPC is used only if n is more than 10 percent of N*. If drawing is done without replacement from a finite population, the variance is

$$\sigma_{\bar{X}}^2 = \left[\frac{\sigma^2}{n}\right]\left[\frac{N-n}{N-1}\right]. \quad (5.6)$$

And the standard error becomes

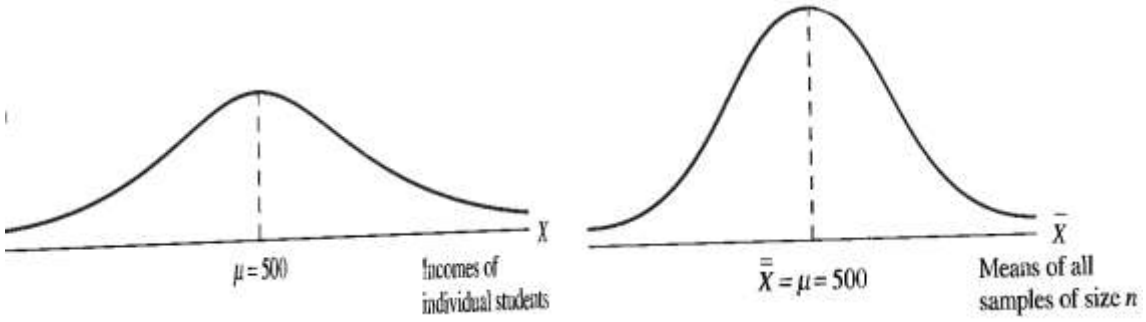
$$\sigma_{\bar{X}} = \left[\frac{\sigma}{\sqrt{n}}\right]\sqrt{\frac{N-n}{N-1}}, \quad (5.7)$$

where $\frac{N-n}{N-1}$ is the FPC. This expression accounts for the fact that *N* is finite, and thereby provides a more accurate statement of the variation in the sampling distribution.

5.1.3. *The Standard Error and Normality*

If the data in a population are *normally distributed*, then the sampling distribution of the sample means will *also be normal*.

For **example**, there are incomes for several thousand students. These incomes average $500 and they are normally distributed. If all samples of size *n* are selected from that normal population of student incomes, then the sampling distribution of the sample means will also be normal. This is displayed in Figure 5.2. In Figure 5.2 b, the $\bar{X}$'s are more closely clustered around their mean than are individual observations in Figure 5.2 a.



a *b*

Figure 5.2 – A Normal Distribution: *a* – Population Distribution; *b* –Sampling Distribution

5.2. The Central Limit Theorem

In many instances the population is not normally distributed. In such cases the **Central Limit Theorem** is used. It is states the following:

As n gets larger, the sampling distribution of sample means will approach a normal distribution with $\overline{\overline{X}} = \mu$ *and* $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

This means that even if the population is not normally distributed, the distribution of sample means will be normal if $\boldsymbol{n \geq 30}$. This is because the standard error gets smaller as ***n*** gets bigger.

5.3 Using the Sample Distribution

Samples have a very direct and consequential impact on decisions that are made. An extremely common and quite useful application of a sampling distribution is to determine the probability that a sample mean will fall within a given range. Given that the sampling distribution will be normally distributed because (1) the sample is taken from a normal population or (2) $n \geq 30$ and the Central Limit Theorem ensures normality in the sampling process.

There is a *conversion formula* for determining the probability of selecting one observation that would fall within a given range

$$Z = \frac{X - \mu}{\sigma},$$

where X is a single observation of interest; σ is the population standard deviation.

However, many business decisions depend on an entire sample – not just one observation. In this case, the conversion formula must be altered to account for

the mean of several observations, $\overline{X}$. Therefore, when sampling is done, the conversion formula becomes

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}, \tag{5.8}$$

where $\overline{X}$ is the mean of several observations; $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ is the standard error of the sampling distribution.

Z – Formula has the **purpose**: *it is used to convert all normal distributions of* $\overline{X}$ *to a standard form.*

Example 5.2. The Telcom recorded telephone messages for its customers. These messages averaged 150 seconds with a standard deviation of 15 seconds. Telcom wished to determine the probability that the mean duration of a sample of *n* = 35 phone calls is between 150 and 155 seconds, that is $P(150 < \overline{X} < 155)$.

Solution. Since *n* > 1 and a sample was taken, Formula (5.8) must be used. Then

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{155-150}{15/\sqrt{35}} = 1.97 \text{ or an area of } 0.4756.$$

Thus, $P(150 < \overline{X} < 155) = P(0 < Z < 1.97) = 0.4756$. See it in Figure 5.3. so the probability that a sample of 35 calls will have duration within the range of 150 and 155 seconds is 47.56 percent. Such a quite big percentage is because the sampling distribution is less dispersed than the original population, i.e. the dispersion of original population is bigger than dispersion of the sampling distribution ($\sigma > \sigma/\sqrt{n}$).

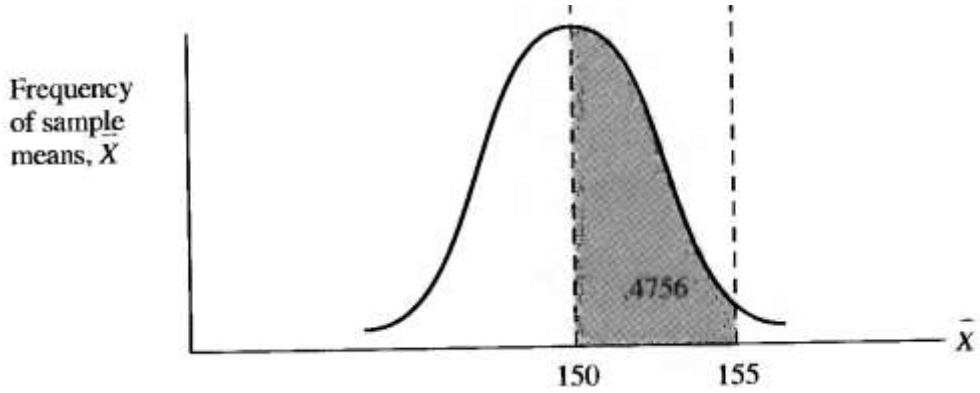


Figure 5.3 – The Mean of a Sample Observations

By being able to predict the likelihood that a certain statistic will fall within a given range, decision making becomes more precise and scientific.

5.4. Types of Samples

1 *Simple Random Sample.* Several difference samples of a given size can be selected from a population. A **simple random sample** involves a method by which each possible sample of some given size has the same likelihood or probability of being selected.

Example 5.3. A national fast-food chain wishes to randomly select 5 of the 50 states to sample consumer taste. **(1)** If the five states are chosen in such a way that any one of the ${}_{50}C_5$ = 2,118,760 possible samples of five states is as likely to be chosen as any other sample of five states, a simple random sample has been taken. **(2)** The simplest technique is merely to list the states on 50 identical pieces of paper, put them in a hat, and draw out five of them.

There is an alternative approach of random sampling. It is a **random number table**. It is often generated by a computer program and formed through the repeated selection of numerical digits. Each time, all 10 digits (0 – 9) have an equal likelihood of being selected.

2 *Systematic Sampling.* A **systematic sampling** is formed by selecting every *i*th item from the population. If *i* is equal to 10, a systematic sample consists of every tenth observation in the population. The population must be ordered or listed in a random fashion. For example, if somebody is needed to select a sample of size 100 from a population of 1,000, *i* must be 10.

3 *Stratified Sampling.* A **stratified sample** is taken by forcing the proportions of the sample from each strata (division of all units of interest into subgroups) to conform to the pattern of the population. It is commonly employed when the population is heterogeneous, or dissimilar, yet certain homogeneous subgroups can be isolated. In this manner the researcher can increase accuracy beyond that obtained by a simple random sample of similar size.

4 *Cluster Sampling.* **Cluster sampling** offers certain advantages over other methods. It consists of dividing the entire population into clusters, or groups, and then selecting a sample of these clusters. All observations in these selected clusters are included in the sample.

Chapter Checklist

1. Distinguish inferential statistics from descriptive statistics.
2. Define and give an example of a sampling distribution.

3. Define and give an example of sampling error.

4. Explain where the grand mean comes from and how it compares to the population mean.

5. Define and give an example of the standard error of the sampling distribution. Tell why the standard error occurs?

6. Discuss the role of the Central Limit Theorem and its importance to statistical analysis.

7. Discuss the role of the Finite Population Correction Factor and its importance to statistical analysis.

8. Use the concept of a sampling distribution to determine the probability that the sample mean falls within a given range.

9. Explain and give examples of different sampling techniques: ()1 simple random sampling, (2) systematic sampling, (3) stratified sampling, (4) cluster sampling.

6 ESTIMATION

6.1 Introduction

Populations are generally too large to study in their entirety. Their size requires that samples should be selected. If a manager of a retail store wished to know tie mean expenditure by her customers last year, she would find it difficult to calculate the average of the hundreds or perhaps thousands of customers who shopped in to store. It would prove much easier to use an estimate of the population mean b; calculating the mean of a representative sample.

There are at least *two types of estimators* commonly used for this purpose:

1) a point estimate,
2) and an interval estimate.

A **point estimate** uses a statistic to estimate tie parameter at a *single value* or point. The store manager may select a sample of n 500 customers and find $\bar{x} = \$37.10$. This value serves as the point estimate for tie population mean.

An **interval estimate** specifies a *range* within which the unknown parameter may lie. The manager may decide the population mean lies somewhere between $35 and $38. Such an interval is often accompanied by a statement as to the level of confidence that can be placed in its accuracy. It is therefore called a *confidence interval*.

Actually there are *three levels of confidence* commonly associated with confidence intervals: **99**, **95**, and **90** percent. There is nothing magical about these three values. It's easy to calculate an 82 percent confidence interval if it's so desirable. These three levels of confidence, called **confidence coefficients**, are simply conventional. The manager referred to above might, for example, be 95 percent confident that tie population mean is between $35 and $38.

Interval estimates enjoy certain advantages over point estimates. Due to sampling error, $\overline{X}$ will likely not equal μ. However, there is no way of knowing how large the sampling error is. Intervals are therefore used to account for this unknown discrepancy.

6.1.1. *The Principle of a Confidence Interval*

A confidence interval has a **lower confidence limit** (LCL) and an **upper confidence limit** (UCL). These two limits are found by calculating a sample mean, $\bar{x}$, as a point estimate, adding a certain amount to it to get the UCL, and

subtracting the same amount from it to get the LCL. The determination of that amount is the subject of this chapter.

The question may arise, "How can we construct an interval and then argue that we can be 95 percent confident that it contains μ if we don't even know what the population mean is?" Recall from the discussion of the Empirical Rule that 95.5 percent of all sample means lie within two standard errors of the population mean. Then, of course, the population mean lies within two standard errors of 95.5 percent of all sample means. Therefore, starting with any sample mean, if we move two standard errors above that mean and two standard errors below that mean, we can be 95.5 percent confident that the resulting interval contains the unknown population mean. Figure 6.1 illustrates.

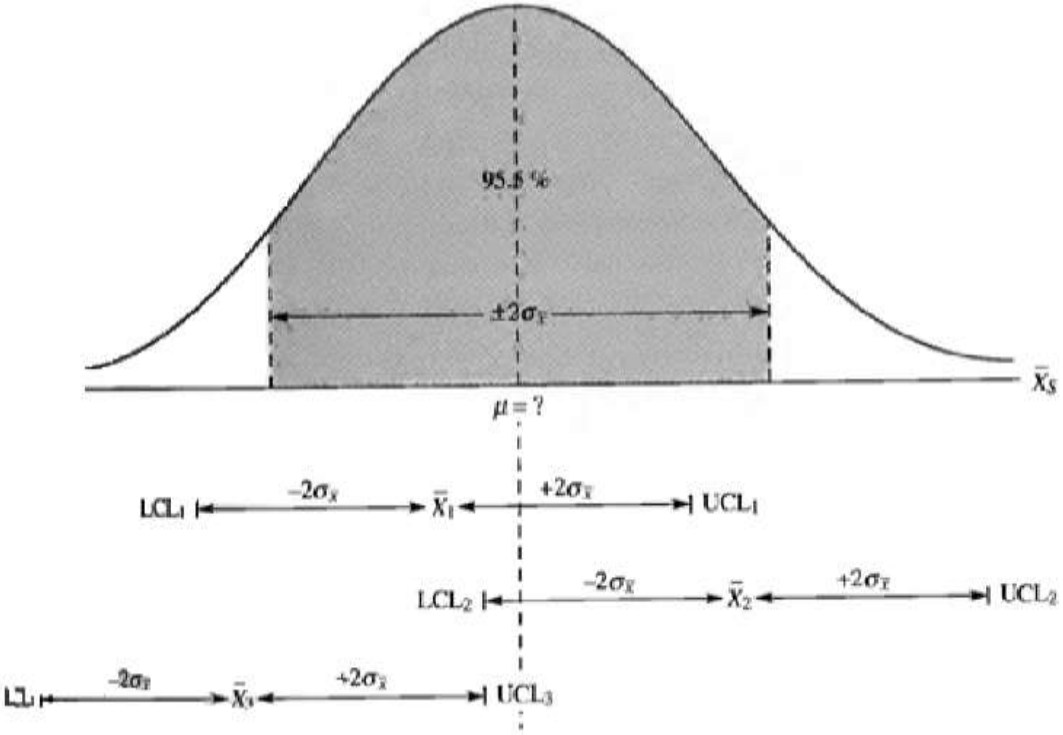


Figure 6.1 – Possible 95.5 Percent Confidence Interval for population mean μ

It should be noticed that if the sample yields $\bar{x}_1$, an interval extending two standard errors above and two standard errors below $\bar{x}_1$ does indeed include the value of the population mean. Similarly, if the sample had yielded a sample mean of $\bar{x}_2$, the resulting interval would also include the population mean. Remember, the discussion of sampling distributions showed that, from any population, we can get many different samples of some given size, each with its own mean. For the purpose of comparison, Figure 6.1 shows three of these possible sample means.

Notice that only $\bar{x}_3$ lies so far from the population mean that an interval ±2 standard errors around either sample mean does not include the population mean. All other sample means will produce an interval that contains the population mean. The key to remember, then, is this: *Since the population mean lies within two standard errors of 95.5 percent of all these sample means, then, given any sample mean, we can be 95.5 percent certain that the interval two standard errors around that sample mean contains the unknown population mean*.

If we wish to construct the more conventional 95 percent interval (rather than the 95.5 percent discussed above), how many standard errors must we move above and below the sample mean? Since the *Z*-table contains only values for the area above or below the mean, we must divide the 95 percent by 2, yielding 47.50 percent, or 0.4750. We then find the *Z*-value corresponding to an area of 0.4750, which is $Z = 1.96$. Thus, to construct a 95 percent confidence interval, we simply specify an interval 1.96 standard errors above and below the sample mean. This value of 95 percent is called the *confidence coefficient*.

In the case of medical research in which lives are at risk, or if our decisions have significant economic consequences, it may be desirable to calculate a higher level of confidence.

6.1.2 *The Probability of Error–The Alpha Value*

In formulating a confidence interval there is, of course, a chance that the interval will be in error and not contain the unknown value of the parameter. If a 95 percent level of confidence is chosen, then 95 percent of all the possible intervals contain the parameter. Thus, the remaining 5 percent do not. This 5 percent is called the **alpha value** and is found as (1 - *confidence coefficient*). A 99 percent confidence interval carries an alpha value of 1 percent and a 90 percent confidence interval carries a 10 percent alpha value.

6.2 Confidence Intervals for the Population Mean – Large Samples

One of the most common uses for confidence intervals is to estimate the population mean. Many typical business situations require an estimate of μ. A producer may wail to estimate the mean monthly level of output for his firm; a marketing representative might be concerned about a drop in mean quarterly sales; a management director might be interested in the mean wage level for hourly

employees. There is an almost infinite number of situations calling for an estimate of the unknown population mean.

Example 6.1. Consider the producer just mentioned, who wishes to construct a confidence interval for his mean monthly output. Recall that a confidence interval consists of an upper confidence limit (UCL) and a lower confidence limit (LCL). Using the sample mean $\bar{x}$ as a point estimator, the producer will add a certain amount to it to get tie UCL, and subtract the same amount to get the LCL. The question is, how much should be added and subtracted? The answer depends on how precise the producer wishes to be. Assume he wants a 95 percent interval. Figure 8-3 illustrates the producer's objective. He must identify an LCL and a UCL that will encompass 95 percent of all the possible values for p, along the axis. If the population standard deviation n is known, the limits on this confidence interval (C.I.) can be found by when past experience and familiarity with the population may reveal its variance but the mean remains a mystery. In any event, in the likely circumstance the population standard deviation is unknown, we simply substitute the sample standard deviation, *s,* and obtain the interval using

$$\text{C.I. for } \mu = \bar{X} \pm Zs_{\bar{x}}. \tag{6.1}$$

Returning to the producer's efforts to construct the 95 percent interval for mean output, assume he takes a sample of $n = 100$ and calculates a sample mean of $\bar{x} = 112$ tons. Past experience has shown that $\sigma = 50$ tons. Then,

$$\text{C.I. for } \mu, = \bar{x} \pm Z\sigma_{\bar{x}} = 112 \pm (Z)\frac{50}{\sqrt{100}}.$$

The producer still needs a value for *Z,* which will be taken from the *Z*-table. Thus, *Z*-value is 1.96. The producer can now complete his answer.

$$\text{C.I. for } \mu, = \bar{x} \pm Z\sigma_{\bar{x}} = 112 \pm (Z)\frac{50}{\sqrt{100}} = 112 \pm (1{,}96)\frac{50}{\sqrt{100}},$$

$$102.2 < \mu < 121.8 \text{ tons.}$$

The producer can draw two inferences about the population from his sample, each based on one of the two interpretations of a confidence interval discussed above:

1. He can be 95 percent confident that the mean daily output lies between 102.2 and 121.8 tons.

2. Ninety-five percent of all confidence intervals formed in this manner will include the true value for μ.

With reference to this second interpretation, consider this question: If the producer would repeat the experiment, would he get the same answer for the interval? No, because he would likely get a different value for $\overline{X}$. However, he can be certain that 95 percent of all the confidence intervals he might construct in this manner would include $\overline{X}$.

6.3 Confidence Intervals for the Population Mean – Small Samples

In the event a small sample (n < 30) must be taken, the normal distribution may not be appropriate. Specifically, when (1) the sample is small and (2) σ is unknown, the *Z*-distribution will not apply. Instead, if the population is known to be normal, an alternative distribution, called Student's t-distribution (or simply the *t*-distribution) must be used. The **Student's t-distribution** was developed in 1908 by William S.Gosset (1876-1937), who worked as a brewmaster for Guinness Breweries in Dublin, Ireland. Guinness would not allow any of its employees to publish their research. When Gosset (who liked to "toy with numbers for pure relaxation") first reported on the t-distribution, he published under the pseudonym "Student" in order to protect his job. Hence, the term Student's *t*-distribution.

Like the *Z*-distribution, the t-distribution has a mean of zero, is symmetrical about the mean, and ranges from – ∞ to +∞. However, while the *Z*-distribution has a variance of $\sigma^2 = 1$, the variance of the *t*-distribution is greater than l. As a result, the *t*-distribution is flatter and more dispersed than the *Z*-distribution (see Figure 6.2).

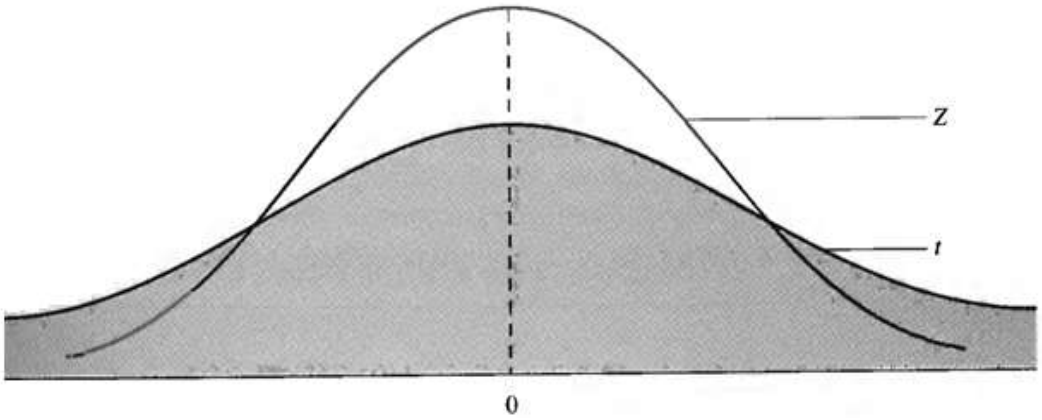


Figure 6.2 – A Comparison of the *t*-Distribution and the *Z*-Distribution

Actually the *t*-distribution is an entire family of distributions with different variances. The variance of the t-distribution depends on the degrees of freedom (d.f.), which is equal to $n - 1$.

The variance for the t-distribution can be written as

$$\sigma^2 = \frac{n-1}{n-3}. \tag{6.2}$$

As n increases, the variance approaches l. When $n > 30$, the distribution will, like the *Z*-distribution, have a variance of 1 (or very close to it). This explains why the *Z*-distribution can be used for large samples.

It is important to remember that the *t*-distribution is used when

(1) the population is assumed to be normal,

(2) a small sample is taken,

(3) σ is unknown.

The *t*-statistic is calculated like the *Z*-statistic.

$$t = \frac{X - \mu}{S_{\bar{x}}}. \tag{6.3}$$

Rewriting (6.3) algebraically a confidence interval for μ, is

$$\text{C.I. for } \mu = \overline{X} \pm (t)(S_{\bar{x}}) = \overline{X} \pm t\frac{s}{\sqrt{n}}. \tag{6.4}$$

The proper *t*-value can be found from Table A. To illustrate, assume you want a 95 percent confidence interval and have a sample of 20 observations. Since $n = 20$, the degrees of freedom are d.f. = n - 1 = 19. Move down the first column Table F under "d.f." to 19. Move across that row to the column headed by a confidence level of 0.95 for two-tailed tests. (Ignore the two rows concerning one-tailed tests.) The resulting entry of 2.093 is the proper *t*-value for a 95 percent confidence interval with a sample size of 20 (d.f. = 19).

Example 6.2 (*from a news story in The Wall Street Journal*)**.** A construction firm was charged with inflating the expense vouchers it files for construction contracts with the federal government. The contract states that a certain type of job should average $1150. In the interest of time, the directors of only 12 government agencies were called on to enter court testimony regarding the firm's vouchers. If a mean of $1275 and a standard deviation of $235 are discovered from testimony, would a 95 percent confidence interval support the firm's legal case? Assume voucher amounts are normal.

A 95 percent level of confidence with d.f. = 12 - 1 = 11 yields from Table A a *t*-value of 2.201. Then

$$\text{C.I. for } \mu = \overline{X} \pm t\frac{s}{\sqrt{n}} = 1275 \pm (2.201)\frac{235}{\sqrt{12}} = 1275 \pm 149.31,$$

$$\$1{,}125.69 \leq \mu \leq \$1{,}424.31.$$

The court can be 95 percent confident that the mean voucher was between \$1125 and \$1424. Since the interval contains the \$1150 amount agreed upon, it would seem to strengthen the firm's defense.

Example 6.3 (*UAW vs FMC*)**.** The labor agreement between the United Auto Workers and Ford Motor Company required that the mean output for a particular production section be held at 112 units per month per employee. Disagreement arose between UAW and FMC as to whether this standard was being maintained. The labor agreement specified that if mean production levels dropped below the stipulated amount of μ = 112, FMC was permitted to take "remedial action". Due to the cost involved, only 20 workers were tested, yielding a mean of 102 units. Assume a standard deviation of 6.5 units was found and that output levels are normally distributed. Does a 90 percent confidence interval tend to suggest a violation of the labor contract, thereby allowing the remedial action?

Solution:

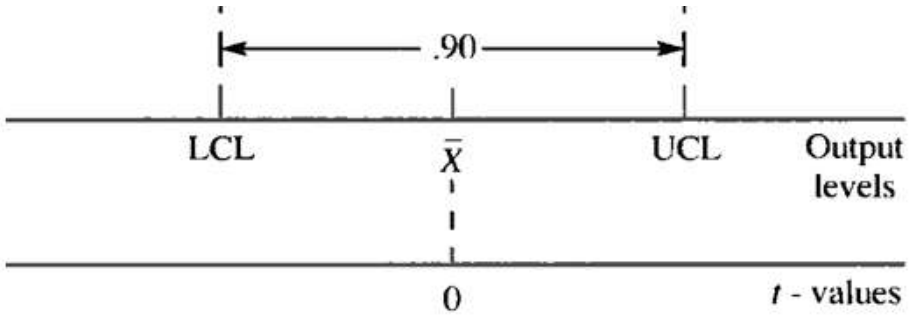


Figure 6.3 – A 90 percent Confidence Interval

With a 90 percent level of confidence and $n - 1 = 19$ d.f., Table A yields a *t*-value of 1.729.

$$\text{C.I. for } \mu = \overline{X} \pm t\frac{s}{\sqrt{n}} = 102 \pm (1.729)\frac{8.5}{\sqrt{20}} = 102 \pm 3.29,$$

$$98.71 \leq \mu \leq 105.29.$$

The mean output level of 112 units specified in the labor contract is not into confidence interval.

Interpretation: There is a 90 percent level of confidence that the contract is being violated. FMC is within its rights to pursue a remedy for lagging productivity.

Obviously, deciding whether to use a *t*-test or a *Z*-test is crucial. Figure 6.4 will aid in selecting the proper test statistic. Remember that the *t*-distribution should be used when all three of these conditions are present:

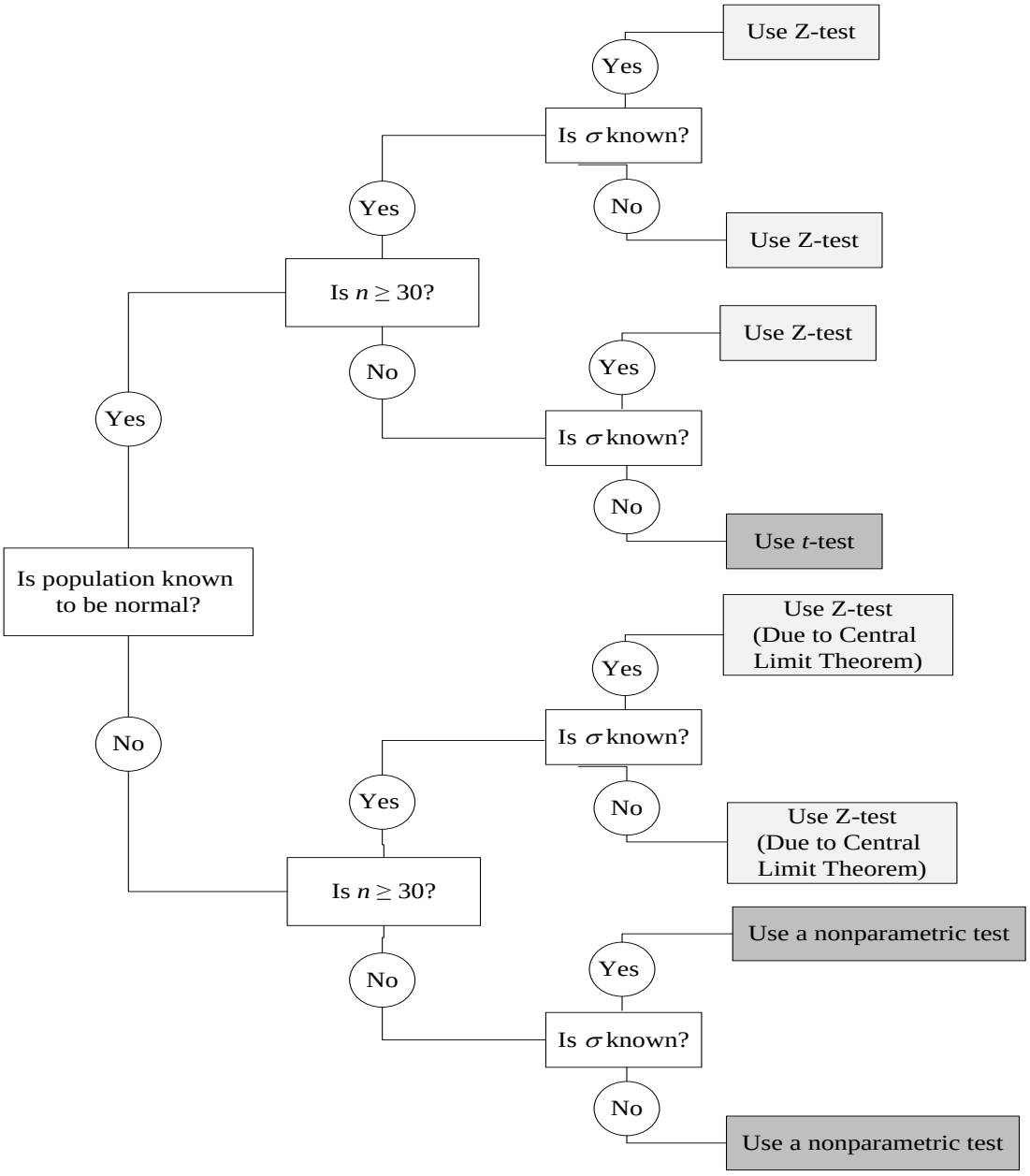


Figure 6.4 - Selecting the Proper Test Statistic for μ

6.4 Confidence Intervals for Population Proportions

Decisions often depend on parameters that are binary – parameters with only two possible categories into which responses may fall. In this event, the parameter of concern is the population proportion.

For example, a firm may want to know what proportion of its customers pay on credit as opposed to those who use cash. Corporations are often interested in what percentage of their products are defective as opposed to the percentage that is not defective, or what proportion of their employees quit after one year in contrast to that proportion who do not quit after one year. In each of these instances, there are only two possible outcomes. Concern is therefore focused on that proportion of responses that fall into one of these two outcomes.

If $n\pi$ and $n(1 - \pi)$ are both greater than 5, the distribution of sample proportions will be normal (n should be greater than 50). Thus, the standard error of the sampling distribution of sample proportions is

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}. \qquad (6.5)$$

However, Formula (6.5) requires π, the parameter is going to be estimated. Therefore, the sample proportion p is used as an estimator for π.

Formula (6.5) becomes

$$S_p = \sqrt{\frac{p(1-p)}{n}}. \qquad (6.6)$$

The confidence interval for population proportion is then

$$\text{C.I. for } \pi = p \pm Zs_p.$$

Example 6.4. The manager of a TV station must determine what percentage of households the city have more than one TV set. A random sample of 500 homes reveals that 273 have two or more sets. What is the 90 percent confidence interval for the proportion all homes with two or more sets? Given these data,

$$p = 275/500 = 0.55,$$

and

$$S_p = \sqrt{\frac{(0.55)(0.45)}{500}} = 0.022.$$

Table B yields a Z of 1.65 for a 90 percent confidence interval.

$$\text{C.I. for } \pi = 0.55 \pm (1{,}65)(0.022) = 0.55 \pm 0.036,$$

$$0.514 \leq \pi \leq 0.586.$$

The manager can be 90 percent confident that between 51.4 percent and 58.6 percent of the homes in the city have more than one TV.

6.5 Controlling the Interval Width

In working through some of the problems encountered thus far in this chapter, you may have become somewhat alarmed by the width of the interval. Some concern might be caused by the fact that an interval was too wide and failed to localize the parameter with sufficient precision. Narrowing the interval will, of course, provide the researcher with a more exact estimation of the parameter's value. *There are two common methods of narrowing the interval*. Both, however, entail a cost of some kind. These procedures to achieve a more precise interval are

(1) Decreasing the level of confidence,

(2) Increasing the sample size.

6.5.1. *Adjusting the Level of Confidence*

By the mere nature of confidence intervals, accepting a lower level of confidence in the interval will generate a more precise, narrower interval.

Example 6.5. **Headhunters in Paradise.** Executive search firms specialize in helping corporations locate and secure top-level management talent. Called "headhunters," these firms are responsible for the placement of many of the nation's top CEOs. Business Week recently reported on the "efforts by headhunters to place executives in a heavenly corporate setting." A source was quoted in the story as saying that "one out of every four CEOs is an outsider – an executive with less than five years at the company he runs."

1) If, in a sample of 350 US corporations, 77 have outsider CEOs, would a 99 percent confidence interval support the quote?

Solution:

$$P = \frac{77}{350} = 0.22,$$

$$S_p = \sqrt{\frac{(0.22)(0.78)}{350}} = 0.022,$$

$$\text{C.I. for } \pi = p \pm Zs_p = 0.22 \pm (2.58)(0.022),$$

$$0.163 \leq \pi \leq 0.277.$$

Interpretation: We are confident at the 99 percent level that between 16.3 percent and 27.7 percent of US corporations have outside CEOs. The quote is supported by these findings, since 25 percent is contained within the interval.

2) If we were willing to accept a 10 percent probability of error, by constructing a 90 percent interval, we would have

$$\text{C.I. for } \mu = 0.22 \pm (1.65)(0.022) = 0.22 \pm 0.036,$$

$$0.184 \leq \mu \leq 0.256.$$

The effect of decreasing the level of confidence from 99 percent to 90 percent is shown in Figure 6.5.

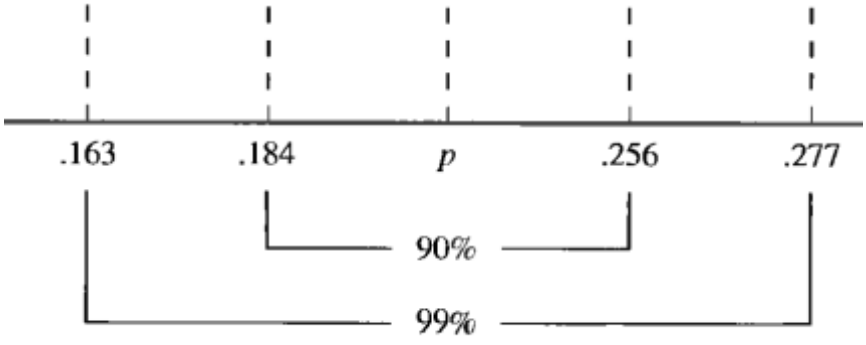


Figure 6.5 – Comparing 99 Percent and 90 Percent Confidence Intervals

The estimate now places π between 18.4 percent and 25.6 percent. This is certainly a narrower interval, but the aforementioned cost is a loss of confidence in the interval and a marked increase in the probability of error. It is up to the researcher, based on how critical his or her work is, to determine what probability of error can be tolerated. A trade-off must be made.

6.5.2. ***Adjusting the Sample Size***

Another common *method of generating a narrower interval is to take a larger sample*. It has been repeatedly emphasized that large samples will reduce the expected error and are more likely to produce an estimate closer to the true value of the parameter. Therefore, the researcher can retain a given level of confidence and still reduce the width of the interval.

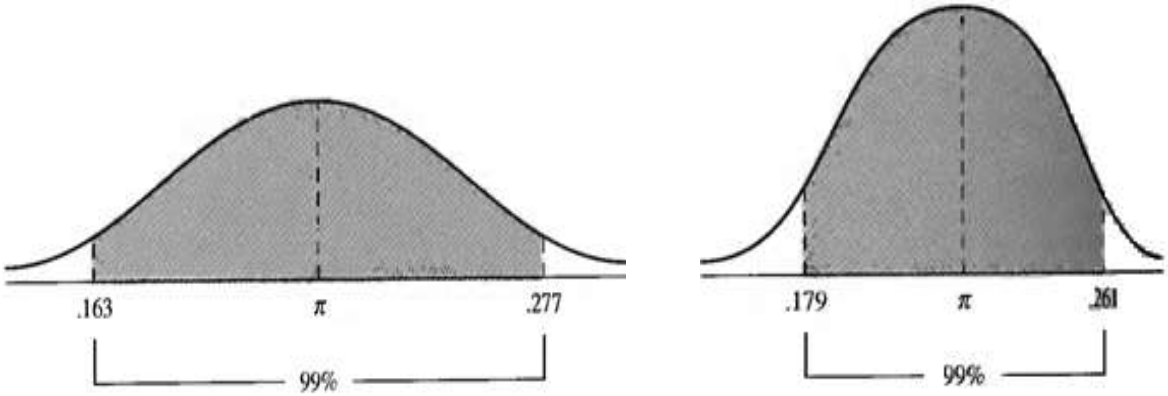
Return to **Example 6.5.** In a sample of $n = 350$, 22 percent of the CEOs were outsiders. The 99 percent confidence interval was 16.3 percent to 27.7 percent In order to narrow the interval and yet retain the 99 percent level of confidence, it is necessary to increase the sample size. Assume that in a sample of $n = 700$, 22 percent are found to be outsiders. The sample proportion of 22 percent is held constant to ensure that the only sample size is changed. The 99 percent confidence interval is then

$$S_p = \sqrt{\frac{(0.22)(0.78)}{700}} = 0.0157,$$

$$\text{C.I. for } \pi = 0.22 \pm (2.58)(0.0157) = .022 \pm 0.041,$$

$$0.179 \leq \pi \leq 0.261.$$

The results of increasing the size of the sample are shown in Figure 6.6.



a *b*

Figure 6.6 – Comparing Sample Sizes: *a* – Sample Size *n* = 350; *b* – Sample Size *n* = 700

Notice that the larger sample size produced a smaller standard error, S_p. The earlier example with *n* = 350 produced a standard error of 0.022, while this sample size of 700 resulted in a standard error of only 0.0157. The 99 percent confidence interval is therefore contained within a narrower range.

There is again, however, a cost associated with producing this more precise interval – in the form of the time and expense required to collect a larger sample. This additional cost must be judged against the higher degree of precision.

6.6 Determining the Sample Size

The size of the sample has an important impact on the probability of error and the precision of the estimate, as well as on other important factors associated with the research effort. Determination of the appropriate sample size is crucial. Given a desired confidence level, two factors are particularly instrumental in influencing the necessary sample size:

(1) The variability of the population, σ^2.

(2) The size of the error that can be tolerated.

While the first factor is beyond the control of the researcher, the size of the tolerable error should be examined at this point.

The extent of error a researcher can tolerate depends on how critical the work must be. Some tasks are extremely delicate and require exacting results: Vital medical procedures upon which lives may depend, or the production of machine

parts that must meet precise measurements, can tolerate only a small error. In other instances, larger errors may be of lesser consequence.

Example 6.6. Presume that in manufacturing a part for compact disk players, an error of 2 centimeters (cm) in diameter will cause no problem. Any error in excess of 2 cm will, however, result in a defective disk player. If the part can vary above and below some desired mean diameter by 2 cm, an interval of 4 cm is allowed. Any given interval is twice the tolerable error. See Figure 6.7 for an illustration.

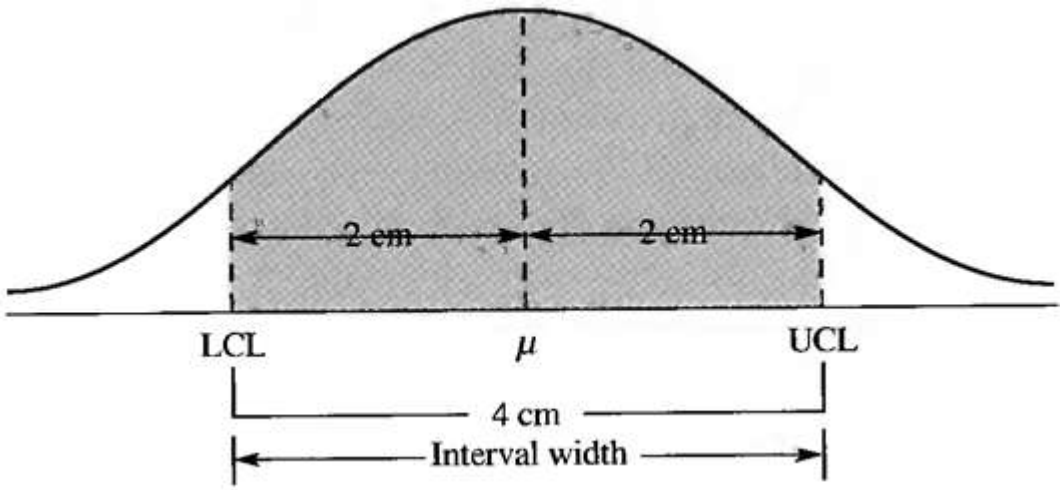


Figure 6.7 – The Tolerable Error is One-Half the Interval

Thus, *the confidence interval can extend above and below the mean by the amount of the tolerable error*.

6.6.1. Sample Size for the population mean μ

Recall that the normal deviate *Z* can be expressed as

$$Z = -\frac{\overline{X} - \mu}{\sigma_{\bar{x}}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}.$$

This can be rewritten algebraically as

$$n = \frac{Z^2 \sigma^2}{(\overline{X} - \mu)^2}, \tag{6.7}$$

where the difference between the sample mean and the population mean $(\overline{X} - \mu)$ is the error. In the **Example 6.6** for the compact disk players with a tolerable error of 2 cm, Formula (6.7) would be written as

$$n = \frac{Z^2 \sigma^2}{(2)^2}.$$

The value of *Z* depends on the level of confidence required. This leaves only σ^2 to be determined in order to calculate the proper sample size. In the likely event σ^2 is unknown, it can be estimated with a *pilot sample* of any reasonable size (n > 30). The variance calculated from this preliminary sample can then be used in Formula (6.7).

Assume, for example, that the manufacturer of the disk players wishes to construct a 95 percent interval for the mean size of the part. A pilot sample has revealed a standard deviation of 6 cm in the part. How large should the sample be? A 95 percent interval calls for a *Z*-value of 1.96. Thus,

$$n = \frac{(1.96)^2(0.6)^2}{(2)^2} = 34.5 \text{ or } 35.$$

The manufacturer should select a sample of 35 parts. From this sample, a 95 percent interval could be constructed for the mean size. This interval would have an error not greater than 2 cm.

6.6.2. ***Sample Size for intervals of population proportion π***

For the population proportion *Z*-value is found by using the following formula

$$Z = \frac{p - \pi}{\sigma_p},$$

where

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

This can be rewritten to produce an expression for sample size.

$$n = \frac{Z^2\pi(1-\pi)}{(p-\pi)^2}, \qquad (6.8)$$

where (p — π) is the difference between the sample proportion and the population proportion, and is therefore *the error*.

Formula (6.8) requires a value for π. However, π is the parameter we wish to estimate, and is unknown. This problem can be handled in one of two ways.

(1) A *pilot sample* can be taken to obtain a preliminary value for π,

(2) π might be set as π = 0.5 for the purpose of determining sample size.

The second approach is often preferred because it is very "safe" or conservative – it will ensure the largest possible sample size given any desired

level of confidence and error. This larger sample results from the fact that the numerator of Formula (6.10), which contains $\pi(1 - \pi)$, will be maximized (thus, n will be maximized) if $\pi = 1 - \pi = 0.5$. There is no value other than 0.5 which you could assign to π that would make $\pi(1 - \pi)$ larger. If $\pi = 0.5$, then $\pi(1 - \pi) = 0.25$. Any value other than 0.5 would result in $\pi(1 - \pi) < 0.25$. Thus, n would be smaller.

Example 6.7. Wally Simpleton is running for governor. He wants to estimate within 1 percentage point the proportion of people who will vote for him. He also wants to be 95 percent confident of his findings. How large should the sample size be?

$$n = \frac{(1.96)^2(0.5)(0.5)}{(0.01)^2} = 9.604 \text{voters.}$$

6.7 Properties of Good Estimators

A distinction should be drawn between an *estimator* and an *estimate*. An **estimator** is the rule or procedure, usually expressed as a formula, that is used to derive the estimate. For example,

$$\overline{X} = \frac{\sum X_i}{n}$$

is the estimator for the population mean.

If the value of the estimator $\overline{X}$ is found to be, say, 10, then 10 is the estimate of the population mean.

To perform reliably, estimators must be

(1) *Unbiased*. An estimator is unbiased if the mean of the sampling distribution equals the corresponding parameter. To cite a specific example, $\overline{X}$ is an unbiased estimator of p because the mean of the sampling distribution of sample means, $\overline{X}$, equals μ. Thus, $E(\overline{X}) = \overline{X} = \mu$.

(2) *Efficient*. Given any unbiased estimators, the most efficient estimator is the one with the smallest variance.

(3) *Consistent*. An estimate is consistent if, as n increases, the value of the statistic approaches the parameter. For an estimate to be consistent, it must be unbiased and its variance must approach zero as n increases. The variance of the sampling distribution of the sample means, $\sigma_{\overline{x}}^2$, is σ^2/n. As n gets larger, $\sigma_{\overline{x}}^2$ will approach zero. Therefore, it can be said that $\overline{X}$ is a consistent estimator of μ.

(4) *Sufficient*. An estimator is sufficient if no other estimator could provide more information about the parameter.

Chapter Checklist

After studying this chapter, can you

1. Calculate a confidence interval for both the population mean and the population proportion?
2. Properly interpret a confidence for both the population mean and the population proportion?
3. Distinguish under what condition a *t*-test should be performed?
4. Explain how to control the width of an interval to affect its precision?
5. Determine the proper sample size necessary to calculate a confidence interval for both μ, and π?
6. Explain the properties of good estimators?

7 SIMPLE REGRESSION AND CORRELATION ANALYSIS

7.1 Introduction

Many empirical studies rely quite heavily on regression and correlation analysis. These tools are perhaps the most commonly used forms of statistical analysis, and the invaluable when making a large number of business and economic decisions.

Regression and correlation analysis recognize that they may be a determinable and quantifiable relationship between two or more variables. That is, one variable depends on another and can be determined by it; or one variable is a function of another. This can be stated as

$$Y = f(X) \tag{7.1}$$

Since Y depends on X, it is the **dependent variable** and X is the **independent variable**.

Example. *Distinction between the dependent and independent variables in the business world*.

A firm's sales depend, at least a part, on the amount of advertising that it does. Sales is seen as the dependent variable and is a function of the independent variable, advertising. In this manner, advertising can be used to predict and forecast sales. The dependent variable Y is also referred to as the *regressand* or the *explained* variable, while the independent variable X is called the *regressor* or the *explanatory* variable.

Regression and correlation are two different but closely related concepts. *Regression* is a quantitative expression of the basic nature of the relationship between the dependent and independent variables. *Correlation* determines the strength of the relationship.

Simple regression holds that the dependent variable Y is a function of only one independent variable, as indicated in formula (7.1). It sometimes called **bivariate** analysis because only two variables are involved - one dependent and one independent.

It is also important to distinguish between *linear* and *curvilinear regression*. ***Linear*** regression attempts to depict the relationship between X and Y by a straight line. This procedure is based on the contention that a change in X is accompanied

by a systematic change in Y, which can be represented by a line. ***Curvilinear*** regression is used if the relationship can better be described by a curve.

Thus, *if X and Y are related in a linear manner, then, as X changes, Y changes by a constant amount. If a curvilinear relationship exists, Y will change by a constant rate as X changes*.

The nature of linear regression, as well as the manner in which it differs from curvilinear regression, can be best illustrated by the *scatter diagrams* (fig. 7.1). Scatter diagrams plot the paired observations of X and Y on a graph. Customarily, the dependent variable is placed on the vertical axes, while the independent variable is on horizontal axis.

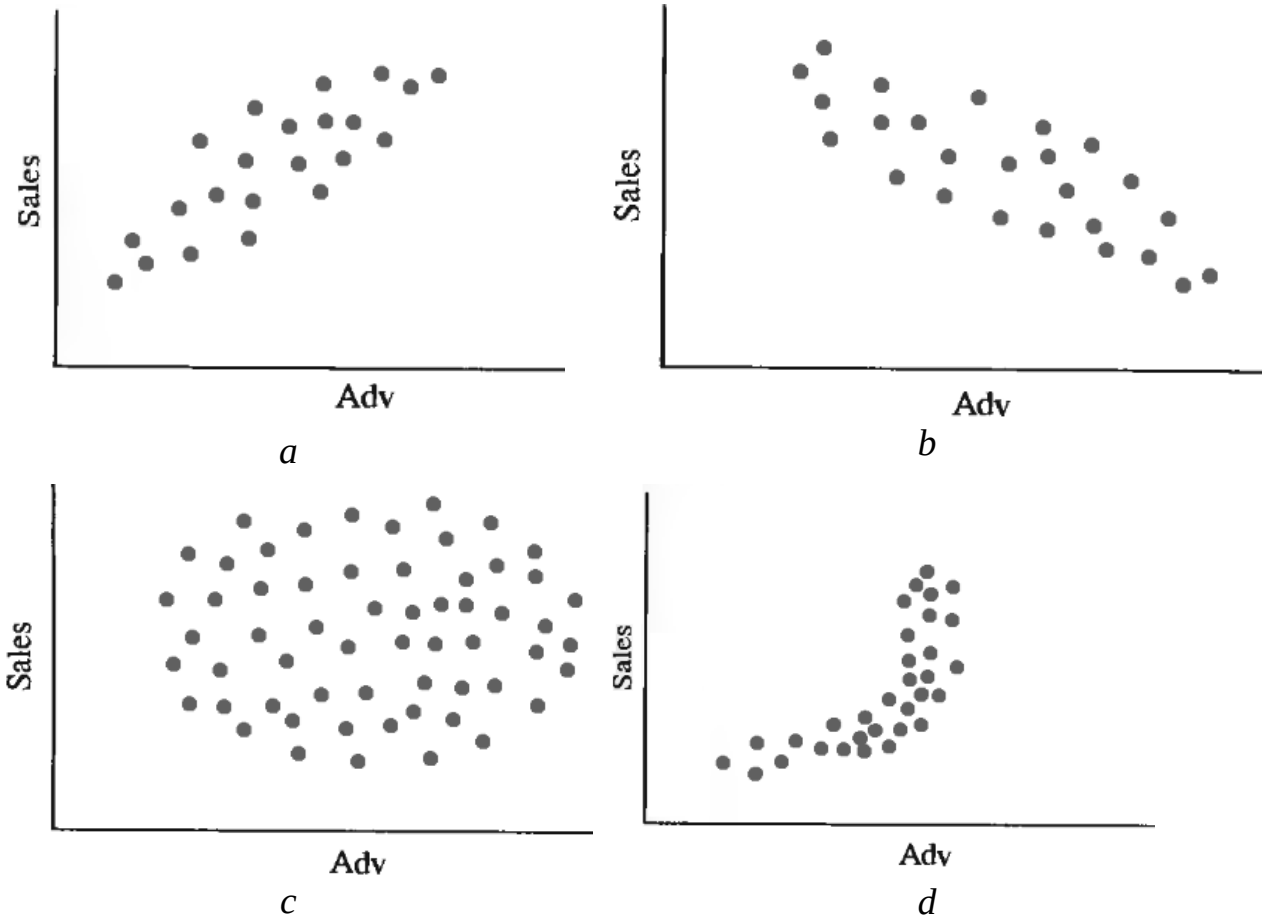


Figure 7.1 – Linear and Curvilinear Relationships:
a – direct or positive linear relationship; *b* – negative or indirect linear relationship; *c* – no relationship exists between two variables; *d* – curvilinear relationship

The objective of regression analysis is to develop a line that passes through the scatter diagram and best represents the data points. Most of this chapter will focus on *simple linear regression*.

7.2 The Basic Objective of Regression Analysis

Relationships between variables are either *deterministic* or *stochastic* (random).

Example 7.2. Deterministic relationship can be expressed by a mathematical model, or formula, that converts speed in miles per hour (mph) into kilometers per hour (kph). Since 1 mile equals approximately 1.6 kilometers, this model is 1 mph = 1.6 kph. Thus, a speed of 5 mph = 5*1.6 = 8.0 kph. This is *deterministic* model because there is no error (except for rounding) in the determination of the rate of speed in kph. Given any value for mph, kph can be determined exactly.

Only few relationships in the business world are also exact or so easily determined. In using advertising to determine sales, for **example**, there is almost always some variation in the relationship. A model of this nature is said to be *stochastic*, due to the presence of random variation. It can be written as

$$Y = \underbrace{\beta_0 + \beta_1(X)}_{\text{(deterministic component)}} + \underbrace{\varepsilon}_{\text{(random component)}}, \qquad (7.2)$$

where β_0 is the vertical interception of the line; β_1 is the slope; ε is a random error term or *disturbance* term designed to capture variation above and below the regression line due to all other factors not included in the model (may be positive or negative, depending on whether a value Y, given any X value, lies above or below the regression line).

Thus,

A deterministic mathematical model is expressed as $Y = \beta_0 + \beta_1 X$. Given any value for X, the value of Y can be determined with precision. A stochastic model contains one or more random components that lead to errors in efforts to predict, and is written as $Y = \beta_0 + \beta_1(X) + \varepsilon$.

Example 7.3. A computer manufacturer wishes to examine the relationship between the number of hard-disk drivers produced and the total cost. The firm's head financial analyst and statistician collects data over a five-period for the number of drivers produced and the corresponding costs. Although a sample of only five observations is most likely insufficient, it will serve the purpose of illustration. The data are displayed in Table 7.1. The data are then plotted in a scatter diagram shown in Figure 7.2. If a line is drawn through the middle of the scatter, some observations fall above it while others fall below it. Therefore, not all

the observations will fall directly on the regression line. There will likely be some variation above and below it. This deviation above and below the line is reflected in Formula (7.2) by ε.

Table 7.1 - Production data for computer hardware

Day	Number of Drivers	Cost, $
1	50	450
2	40	380
3	65	540
4	55	500
5	45	420

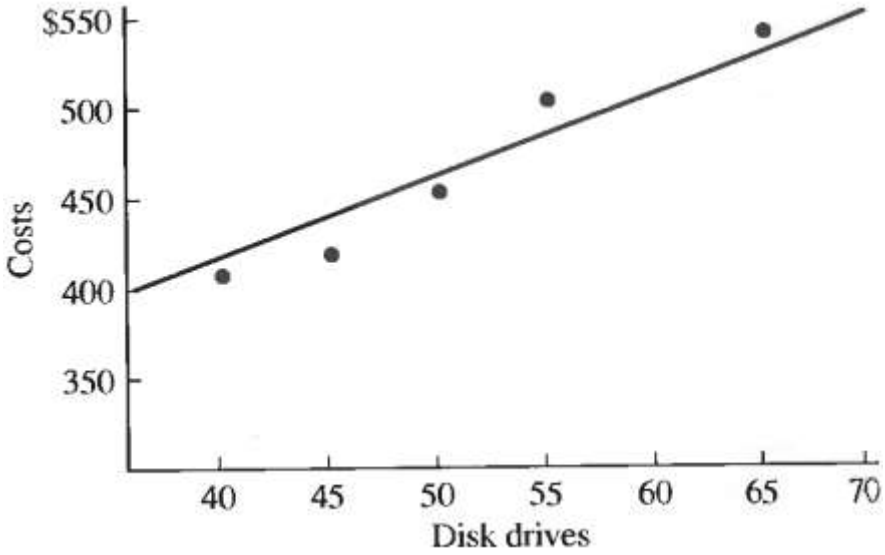


Figure 7.2 - A scatter diagram for Production Data

To estimate the true population regression line the sample model is used

$$Y = b_0 + b_1X + e, \qquad (7.2a)$$

where b_0 - regression constant and b_1 – regression coefficient are estimates for the population parameters β_0 and β_1; e – error component (it usually has a mean value of zero and a variance σ^2 of some amount).

The model (7.2a) is then used to estimate the relationship between X and Y, resulting in the regression line

$$\hat{Y} = b_0 + b_1X \qquad (7.2b)$$

where $\hat{Y}$ (pronounced Y-hat) – is the estimated value for the dependent variable and is represented by a point *on* the regression line.

It should be noticed also that *the regression model can be used to predict or forecast the value for the dependent variable*.

7.3 Ordinary Least Squares Method (the line of best fit).

This method is called OLS because it results in line which is minimizes the squared vertical distances from each observation point to the line itself (Fig.7.3).

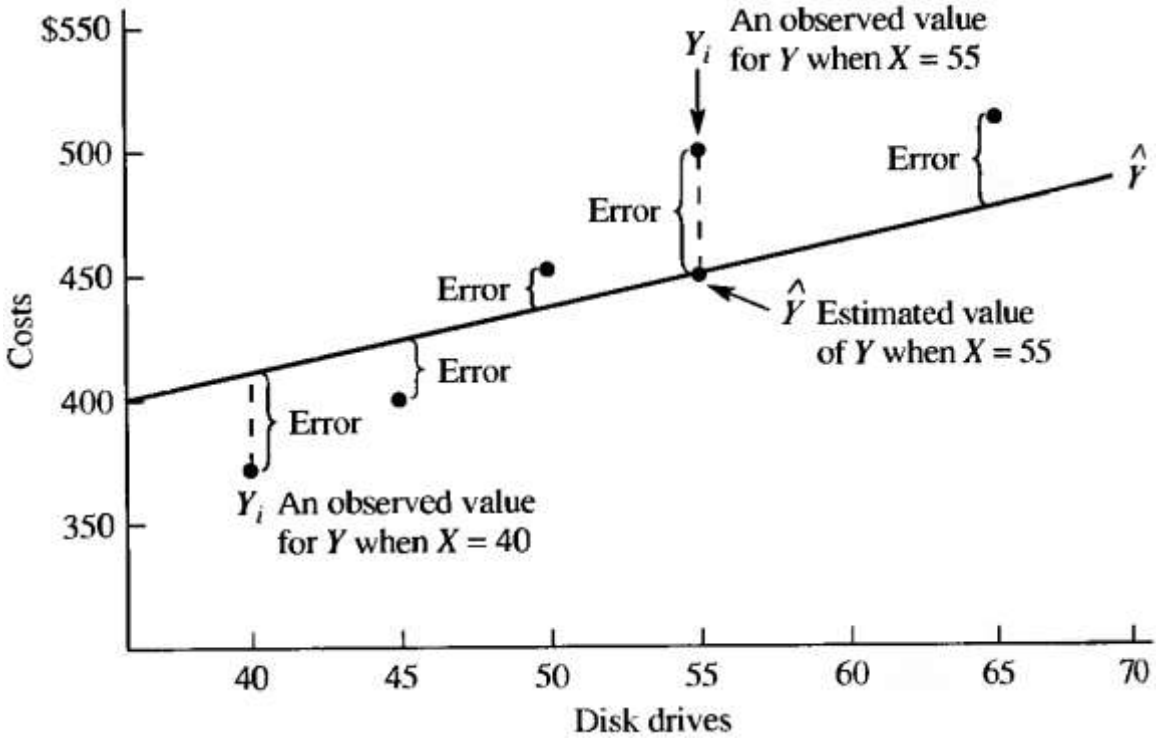


Figure 7.3 – Ordinary Least Squares

In Figure 7.3 Y_i is an actual, observed for the Y variable, $\hat{Y}$ is a value on the line predicted by the equation.

Thus, the vertical difference between all the values Y_i and $\hat{Y}$ should be calculated and then squared to yield $(Y_i - \hat{Y})^2$. All squared differences are summed then and expressed as

$$\Sigma(Y_i - \hat{Y})^2 = \min, \qquad (7.3)$$

where *min* is the number smaller than any summed squared vertical deviations between the actual data points and any other line. Hence, the term *least squares* is used.

The difference $Y_i - \hat{Y}$ is called **residual**, or **error**.

Assumptions of OLS.

1) *The error term is a random variable and is normally distributed*.
2) *Any two errors are independent of each other*, i.e. the error when $X_i = 10$ is totally independent of the error suffered when X_{i+1} is equal to any other value.
3) *All errors have the same variance*. This condition is known as *homoscedastisity* (Fig. 7.4).

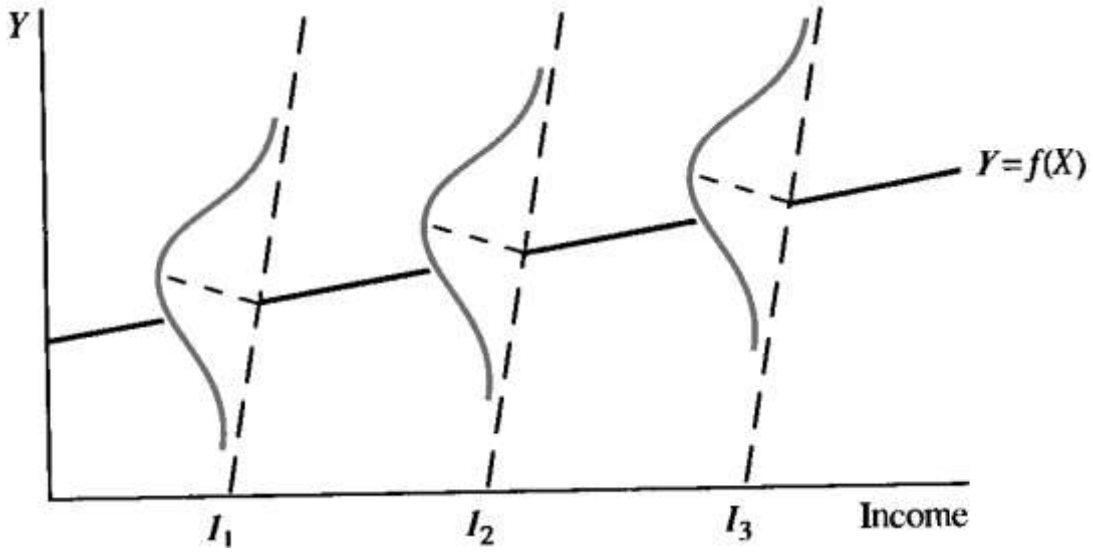


Figure 7.4 – Equality of Variance in the Error Term

4) *The means of the Y – values all lie on a straight line*. Given some value Xi, there will occur a normal distribution of Y-values. This distribution has a mean. The same is true if X is set equal to any other value. OLS assumes that these two means, as well as all others that might be observed, lie on a straight line. This is referred to as the assumption of linearity, and can be expressed as

$$\mu_{y|x} = \beta_0 + \beta_1 X ,$$

where $\mu_{y|x}$ is the mean of the population of Y-values for any given value of X.

To calculate the regression coefficient b_0 and the intercept b_1 in Formula (7.2B) the sums of squares and cross-products are used

$$b_1 = \frac{SS_{xy}}{SS_x}, \qquad b_0 = \bar{Y} - b_1 \bar{X} , \tag{7.4}$$

where

$$SS_{xy} = \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n},$$

$$SS_x = \Sigma(X_i - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}. \quad (7.5)$$

These calculations are extremely sensitive to rounding. In the interest of accuracy it's better to carry out calculations of five or six decimal places.

The meaning of b_1: it indicates by how much Y will change for every one-unit change in the X-variable.

Example 7.4. In order to make decisions regarding allocations for the advertising budget, the accounting department for Hop Scotch Airlines must determine the nature of the relationship between advertising expenditures and the number of passengers. The senior accountant recognizes that regression analysis would be of invaluable assistance.

Table 7.2 – Regression Data for Hop Scotch Airlines

Observation (months)	Advertising (in \$1000's) (X)	Passengers (in \$1000's) (Y)	XY	X^2	Y^2
1	10	15	150	100	225
2	12	17	204	144	289
3	8	13	104	64	169
4	17	23	391	299	529
5	10	16	160	100	256
6	15	21	315	225	441
7	10	14	140	100	196
8	14	20	280	196	400
9	19	24	456	361	575
10	10	17	170	100	589
11	11	16	176	121	256
12	13	18	234	169	324
13	16	23	368	256	529
14	10	15	150	100	225
15	12	16	192	144	256
Total	187	268	3490	2469	4960

Solution: from Formulas (7.4) coefficients of the regression model can be determined as

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{148.93333}{137.73333} = 1.0813166 \text{ or } 1.08,$$

$$b_0 = \overline{Y} - b_1\overline{X} = 17.86667 - (1.08)(12.46667) = 4.3865 \text{ or } 4.4.$$

The regression equation is therefore

$$\hat{Y} = 4.40 + 1.08X.$$

Interpretation:

1) The model tells that if, for example, \$10000 is spent on advertising (X=10), then

$$\hat{Y} = 4.40 + 1.08(10) = 15.2 \text{ passengers (in 1000's)}$$

will choose to fly Hop Scotch.

b) Since b_1 = 1.08, for every additional \$1000 that Hop Scotch spends on advertising, 1080 more passengers will choose this air company, i.e. if advertising is increased by one unit to \$11000, the estimate of total passenger becomes $\hat{Y}$ = 4.40 + 1.08(11) = 16.28 or 16280 passengers.

As there is no evidence of a cause – and – effect relationship the simultaneous increase in X and Y may have been caused by an unknown third variable excluded from the study. It is a common misconception to assume that there exists a cause – and – effect relationship between the two variables.

7.3.1 *The Y-Values Are Assumed to Be Normally Distributed*

The value of the depended variable Y will vary even if the value for X remains fixed. Since Y is different almost all the time, the best regression model can do is estimate the average value for Y given any X - value. Regression analysis is based on the assumption that a linear relationship exists between X and the mean value of Y, E(Y). The regression line can be written

$$E(Y) = \beta_0 + \beta_1(X).$$

A point on the line denotes the average value for Y given any X-value. For this reason, the regression line is often referred to as a ***mean line***.

The point to remember is that for any value of X that may occur several times, a different Y-value can be get each time. An entire distribution of different Y-values would result. Regression analysis assumes that this distribution of Y-values is *normal*. This distribution is centered at the mean of these Y-values, as illustrated in Figure 7.5.

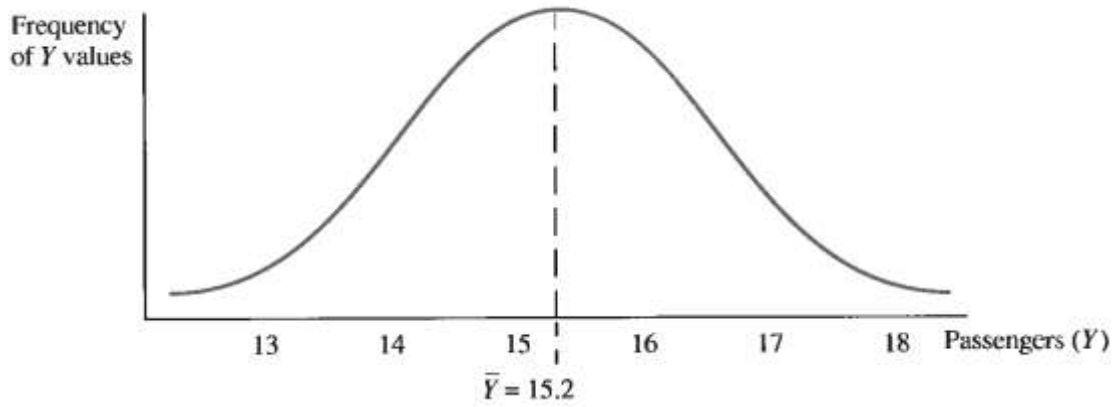


Figure 7.5 – The normal distribution of Y-values for a given single value of X (X=10)

This suggests that if the airlines spends $10000 each month on advertising for several months, the number of passengers, although perhaps different each month, will average 15200.

The normal distribution of Y-values exists for all values of X. thus, if X = 11 on many separate occasions, there would occur an entire distribution of Y-values that would be normally distributed and centered $\hat{Y}$ = 16.28 0r 16280 passengers etc.

When estimating the true, but unknown regression line with a sample regression line, statisticians trying to find that line which passes through the means of the various distributions of Y-values for each X-value. This is illustrated in Figure 7.6.

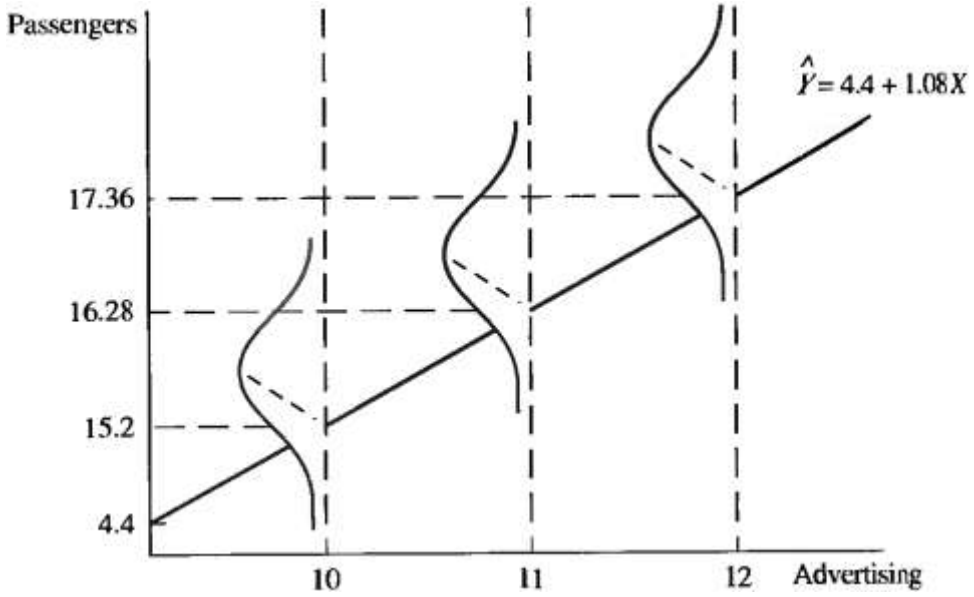


Figure 7.6 – The Normal Distributions of Y-Values for the Various Values of X

Notice that for each value of X there is a distribution of Y-values. The regression line passes through the mean of each of those distributions. Each distribution of Y-values is normal and, like any distribution of numbers, has a variance of σ^2 and a standard deviation of σ. The important point to note here is that this variance is assumed to be the same for each distribution of Y-values regardless of the X-value. That is, the variance of Y-values when X = 10 is the same as the variance Y-values when X = 11 (or anything else).

7.4 The Standard Error of the Estimate: A Measure of Goodness-of-Fit

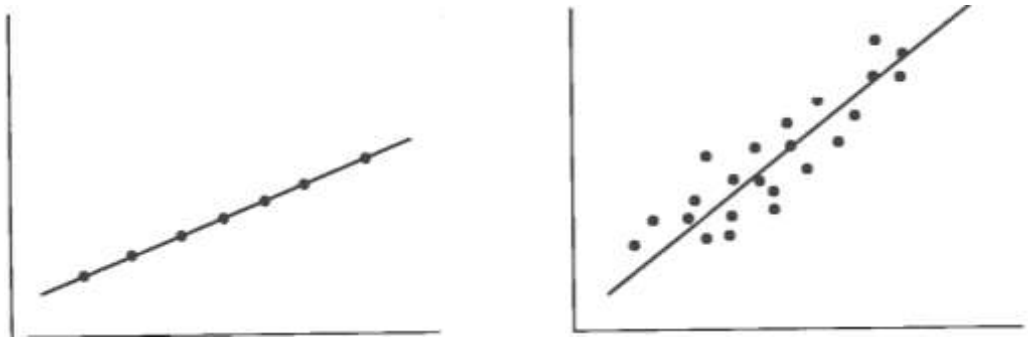
The formula to calculate the standard error of the estimate is the following

$Se = \sqrt{\frac{\Sigma(Y_i - \hat{Y})^2}{n-2}}$, or $Se = \sqrt{MSE}$, where $MSE = \frac{SSE}{n-2}$ - the mean square error;

$SSE = SSy - \frac{(SSxy)^2}{SSx}$ - the error sum of squares;

$SS_y = \Sigma(Y_i - \bar{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$ - the sum of squares of Y.

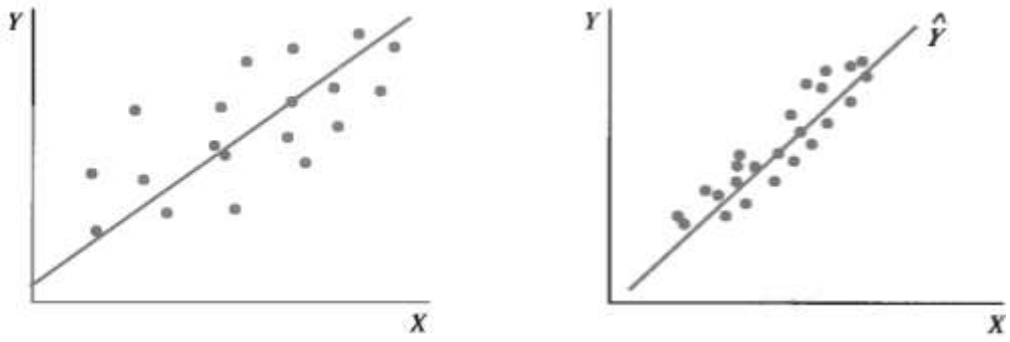
The ***standard error of the estimate***, *Se*, is quite similar to the standard deviation of a single variable, and is a measure of the average amount by which the actual observations for Y vary around the regression line. It gauges the variation of the data points above and below the regression line. As a measure of the dispersion of the Y-values around the regression line, it reflects a tendency to depart from the actual value of Y when using the regression model for prediction purposes. In that sense, it is a measure of the average amount of the error.



a *b*

Figure 7.7 – Possible Scatter Diagrams: *a* – Se = 0; *b* – Se ≠ 0

The more dispersed the original data are, the larger Se will be (Fig.7.8).



a *b*

Figure 7.8 – A Comparison of the Standard Error of the Estimate

As indicated by the scatter diagrams in Figure 7.8, the data for Fig.7.8(a) are much more dispersed than those on Fig.7.8(b). The *Se* for Fig.7.8(a) would therefore be larger.

7.5 Correlation Analysis

7.5.1 *Coefficient of Determination*.

The job of correlation analysis is to measure the strength of the relationship between Y and X in the regression model. This measure of strength is provided by the coefficient of determination R. The ***coefficient of determination*** is one of the measures of goodness-of fit along with the standard error of the estimate S_e.

To understand correlation analysis, the **total deviation of Y (**$Y_i - \bar{Y}$**)** has to be considered. The *total deviation* is the amount by which an actual value of Y, Y_i, differs from $\bar{Y}$, the mean of all the values for the dependent variable.

The total deviation can be broken down into types. The *explained deviation* and *unexplained deviation*. The **explained deviation** is that portion of the total deviation that is explained by the regression model, is the difference between the value predicted by the model and the mean value of Y: ($\hat{Y} - \bar{Y}$). The **unexplained deviation** is the difference between the actual value Y_i and that value predicted by the model: ($Y_i - \hat{Y}$).

As an example the data from Table 7.2 are considered. Taking month 13, it's shown that 23000 people flew on Hop Scotch (Y_i = 23). Since the mean value for the number of passengers is

$$\bar{Y} = 268/15 = 17.87 .$$

The total deviation for the thirteenth month is 23-17.87=5.13.

On the other hand, the regression model forecasts a value for Y of

$$\hat{Y} = 4.4 + 1.08*(16) = 21.68.$$

Using the regression model the error is only Y_i - $\hat{Y}$ =23-21.68 = 1.32. This value is much closer to the actual value for passengers than the average value for Y as a prediction.

This is shown in Figure 7.9.

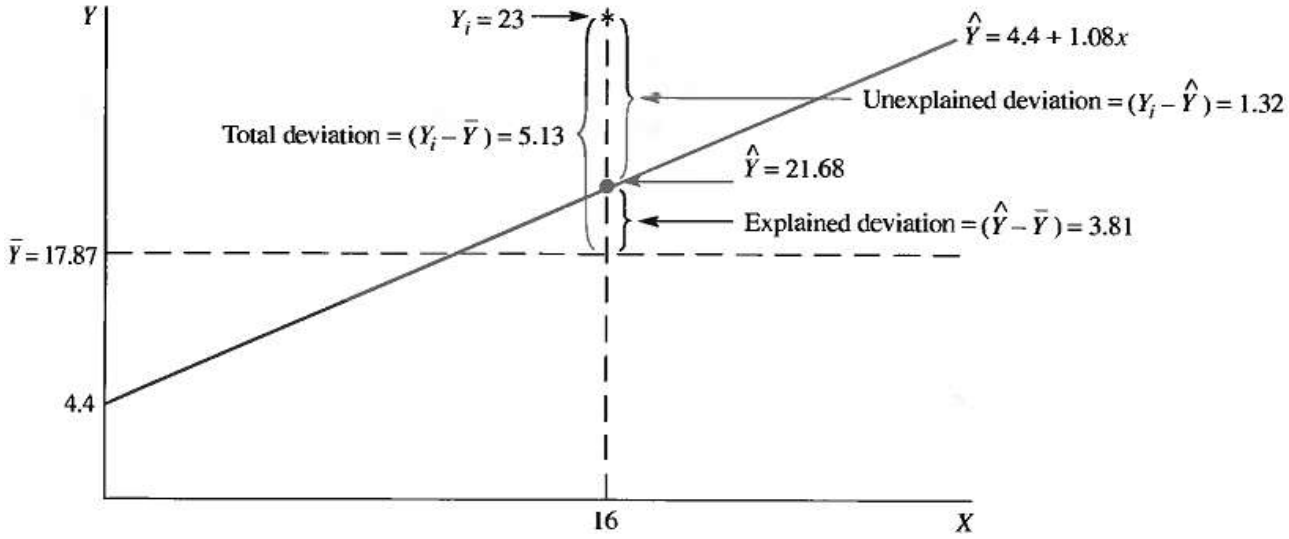


Figure 7.9 – Deviations for Hop Scotch Airlines

Then

Total deviation = Explained deviation + Unexplained deviation

That is,

$$(Y_i - \bar{Y}) = (\hat{Y} - \bar{Y}) + (Y_i - \hat{Y}).$$

To prevent negative errors from offsetting the positive errors, the squaring process is necessary. Thus, the total sum of squares (SST), the regression sum of squares (SSR) and the error sum of squares (SSE) respectively are

$$\text{SST} = \Sigma(Y_i - \bar{Y})^2; \ \text{SSR} = \Sigma(\hat{Y} - \bar{Y})^2; \ \text{SSE} = \Sigma(Y_i - \hat{Y})^2.$$

The **coefficient of determination r^2**,

- it is a ratio of the explained deviation to the total deviation,

- it is a measure of the explanatory power of the regression model by measuring what portion of the change in Y is explained by the change in X,

- measures the strength of the *linear* relationship between X and Y.

$$r^2 = \frac{Explained \text{ var } iation}{Total \text{ var } iation} = \frac{SSR}{SST}.$$

In terms of sums of squares and cross products it can be calculated as

$$r^2 = \frac{(SSxy)^2}{(SSx)(SSy)}.$$

The *value for r^2 must be between 0 and 1* since more that 100 percent of the change in Y cannot be explained. The higher the r^2, the more explanatory power the model has. If r^2 = 70 percent, this means 70 percent of the variation in Y is explained by changes in X.

Example. Given the data for Hop Scotch from Table 7.2, the coefficient of determination is

$$r^2 = \frac{(SSxy)^2}{(SSx)(SSy)} = \frac{(148.9333)^2}{(137.7333)(171.7333)} = 0.93776 \approx 0.94.$$

Interpretation. The coefficient of determination reveals that 94 percent of the change in the number of passengers is explained (not caused) by changes in advertising expenditures.

Since r^2 = 0.94, the model explains 94 percent of the change in Y. the other 6 percent can be explained by some variable(s) other than advertising. This 6 percent sometimes referred to as the ***coefficient of nondetermination***, k^2.

7.5.2 *Coefficient of Correlation*

Designated as *r*, the correlation coefficient is simply the square root of the coefficient of determination. Developed by Karl Pearson around the turn of the century, it is sometimes called the *Pearsonian product-moment correlation coefficient*.

Thus,

$$r = \sqrt{r^2} = \sqrt{0.93776} = 0.96838.$$

The value for r ranges between +1 and -1. The correlation coefficient must be given an algebraic sign after calculating. As it reflects the slope of the regression line, the sign of *r* is the same as the regression coefficient b_1. Thus,

- If $r > 0$, b_1 will be positive and the line will slope up.
- If $r < 0$, b_1 will be negative and the regression line will be negatively sloped.

- If $r = 0$, no linear relationship between X and Y is suggested.
- The absolute value of r ($|r|$) indicates the strength of the relationship between X and Y (Fig. 7.10).

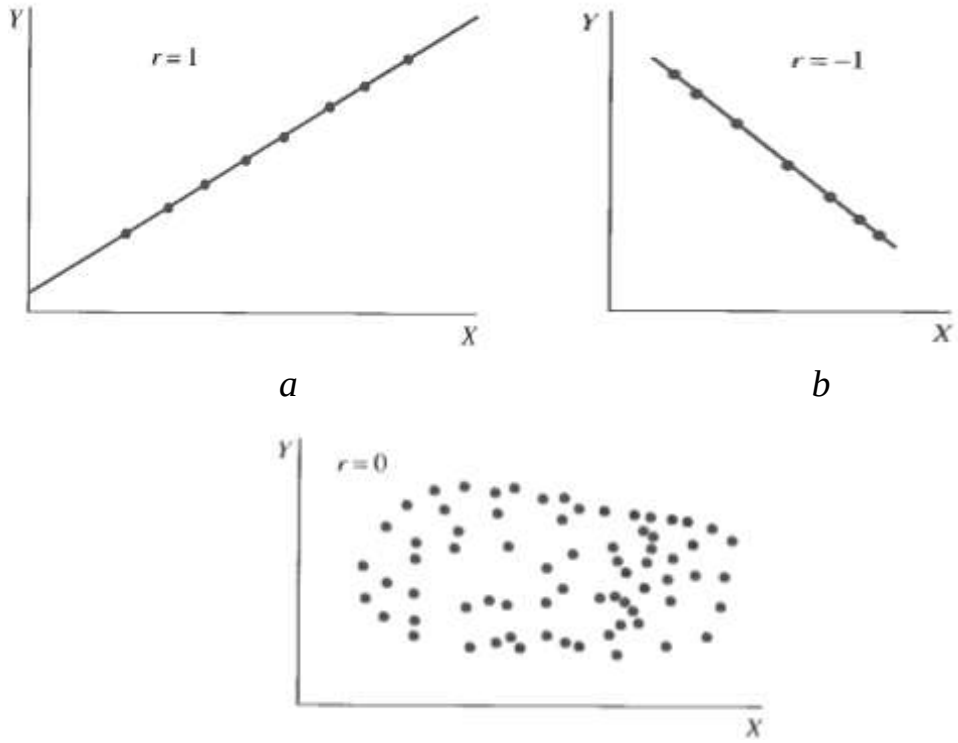


Figure 7.10 – Potential values for *r*: *a* – Perfect positive correlation; *b* – Perfect negative correlation; *c* – No linear correlation

7.6 Interval Estimation in Regression Analysis

One of the basis purposes in conducting regression analysis is to forecast and predict values for the dependent variable. Once the regression equation has be determined, it is a very simple matter to develop a point estimate for the dependent variable by substituting a given value for X into the equation and solving for Y.

But it is more interesting to find out an interval estimates. There are at least *two interval estimates* commonly associated with regression procedures.

1) An interval estimate for the mean value of Y given any X-value, i.e. **the conditional mean** ($\mu_{y|x}$).

2) An estimation of a *single value* of *Y* given that X is set equal to a specific amount. This estimate is referred to as **a predictive interval**.

7.6.1. *The Conditional Mean for Y*

The Conditional Mean for Y is the population mean for all Y-values under the condition that X is equal to a specific value.

To calculate confidence interval for the conditional mean of Y, *the standard error of the conditional mean* S_y must be determined. It recognizes that a sample is used to calculate b_0 and b_1 in the regression equation. It is determined by

$$S_Y = S_e\sqrt{\frac{1}{n}+\frac{(X_i-\overline{X})^2}{SSx}}\,, \qquad (7.6.1)$$

where S_e – the standard error of the estimate;

X_i – the given value for the independent variable.

Thus the C.I. for the conditional mean is

$$\text{C.I. for } \mu_{y|x} = \hat{Y} \pm tS_Y, \qquad (7.6.2)$$

where $\hat{Y}$ – the point estimator found from the original regression equation ;

t-value is based on a selected level of confidence with *n*-2 degrees of freedom. There are *n*-2 d.f. because two values b_0 and b_1 must be calculated from the sample data. Therefore two d.f. are lost.

The C.I. for $\mu_{y|x}$ could be calculated at several X-values. Thus, several C.I. would then form an entire confidence band for $\mu_{y|x}$. The band becomes wider at the extremes. This happens because regression analysis is based on averages, and the father researcher is away from the center point, the less accurate his results (Fig.7.11).

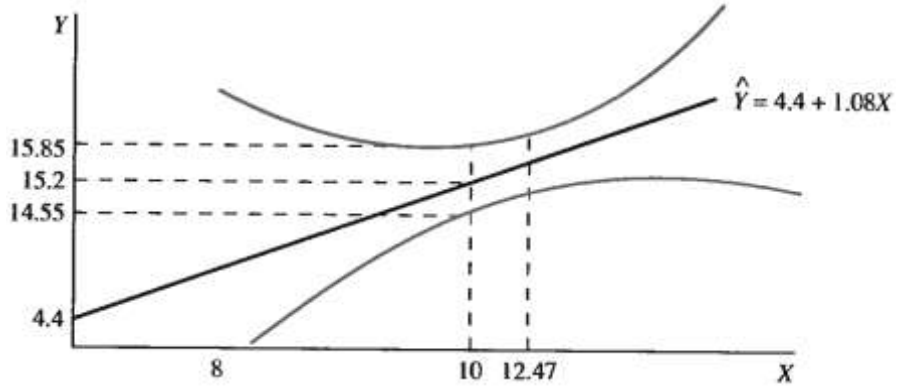


Figure 7.11 – Confidence limits for $\mu_{y|x}$ for Hop Scotch from Table 7.2

7.6.2 *The Predictive Interval for a Single Value of Y*

If X was set to some amount just one time, then the researcher would get one resulting value of Y. So he can be 95 (or other) percent certain that that single value of Y falls within the specific interval.

The standard error of the forecast S_{y_i} (not standard error of the confidence mean, S_Y) must be calculated first. It accounts for the fact that individual values are more dispersed than are means.

$$S_{y_i} = S_e\sqrt{1+\frac{1}{n}+\frac{(X_i-\overline{X})^2}{SSx}}, \qquad (7.6.3)$$

The predictive interval for a single value of Y, Y_x is then

$$\text{C.I. for } Y_x = \hat{Y} \pm tS_{y_i}, \qquad (7.6.4)$$

Example. After receiving the interval estimate for the conditional mean from the marketing division, the CEO now demands to know what the estimate is for passengers the next time they spend X = $ 10000 for advertising. the head of the marketing division realizes that what the CEO is asking for the predictive interval estimate for a single value of X.

Solution.

$$S_{y_i} = S_e\sqrt{1+\frac{1}{15}+\frac{(10-12.47)^2}{137.73333}} = 0.907\sqrt{1.1114} = 0.956.$$

Since

$$\hat{Y} = 4.4 + 1.08(10) = 15.2,$$

$$\text{C.I. for } Y_x = \hat{Y} \pm tS_{y_i} = 15.2 \pm (2.160)(0.956) = 15.2 \pm 2.065,$$

$$13.14 < Y_x < 17.27.$$

Interpretation. We can 95 percent certain that if any single month X = $ 10000, the resulting single value of Y will be between 13140 and 17270 passengers. This interval is wider that the first because researcher is working with less predictable individual values.

7.6.3 *Hypothesis Test about the Population Correlation Coefficient*

In the case the correlation coefficient is not zero ($r \neq 0$) the researcher can conclude on the basis of the sample data that there is a relationship between X and Y. But as always, the interest is with the entire population of all X-values and all Y-values. If the sample data reveal a relationship ($r \neq 0$), there may be no such relationship at the population level ($\rho = 0$).

Therefore it is often desirable to test the hypothesis that the population correlation coefficient is 0. Thus

$$H_0 : \rho = 0;$$

$$H_\alpha : \rho \neq 0.$$

The hypothesis test is done to determine if it is *significantly* different from zero. This test employs the *t*-statistics

$$t = \frac{r}{S_r}\ , \tag{7.6.5}$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}} \tag{7.6.6}$$

has n-2 degrees of freedom, S_r is the standard error of the sampling distribution of *r*. It recognizes that if several samples of size *n* were taken, different values for *r* the researcher would get. If ρ = 0, the r-values would be distributed around ρ, ranging from -1 to +1.

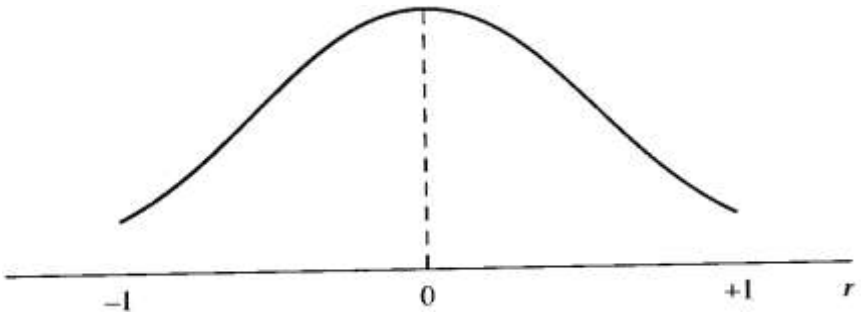


Figure 7.12 – the distribution of Correlation Coefficient *r*

Then the level of confidence is chosen (for example, 95 percent or α = 0.05) at which the null hypothesis $\rho = 0$ is tested. This choice allows to find out a critical value of *t* from the *t*-table. This critical value of *t* is compared with the *t* which was calculated from the Formula (7.6.5) based on the sample data.

In the case the confidence level is 95 percent, and sample size is 15, then it give us 15-2= 13 degrees of freedom and t – values are ±2.160 (Fig.7.13). This means that if $\rho = 0$, there are only 5 percent chance that the sample would yield a t-value below -2.160 or above +2.160. If t-value from Formula (7.6.5.) is outside the range the researcher can be 95 percent certain that $\rho \neq 0$, thus indicating that there is a relationship between X and Y at the population level.

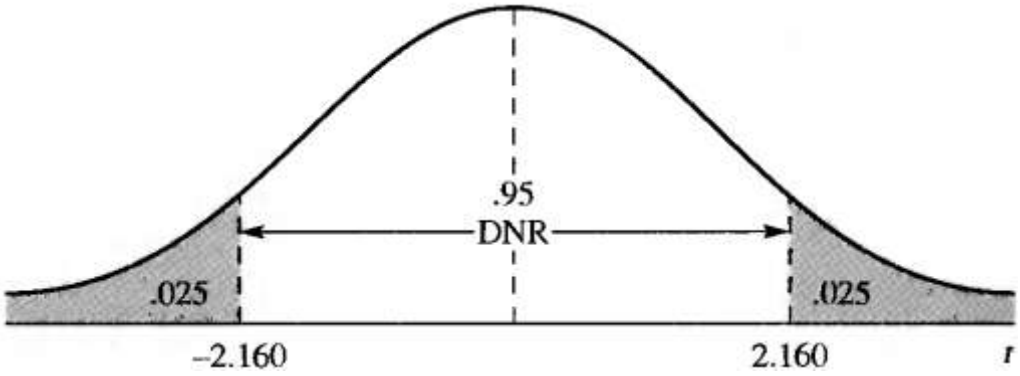


Figure 7.13 – Critical t-Values for Testing the Hypothesis that $\rho = 0$ (confidence level is 95 percent)

If t-value is between -2.160 and +2.160, then the null hypothesis $\rho = 0$ cannot be rejected and despite the sample results the researcher would conclude at the 95 percent level of confidence that there is no relationship between X and Y.

7.6.4 ***Testing Inferences about the Population Regression Coefficient. A Confidence interval for*** β_1

If the slope of the actual but unknown population regression line is zero, there is no relationship between X and Y. but due to the luck of the draw in the sample, the researcher can select sample data that suggest a relationship. It might happen as shown in Figure 7.14

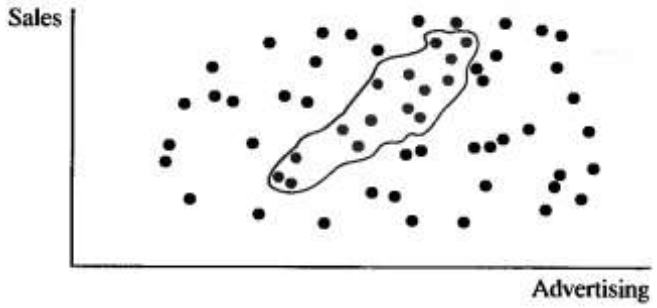


Figure 7.14 – A Possible Pattern of Population Data for Hop Scotch Airlines

As it can be plainly seen, the sample regression would be positively sloped, $b_1 > 0$, and a relationship would be suggested by OLS. It is therefore often a wise practice to test the hypothesis that $\beta_1 = 0$ given $b_1 \neq 0$. The test involves

$$H_0 : \beta_1 = 0;$$

$$H_\alpha : \beta_1 \neq 0;$$

and uses a *t*-statistics defined as

$$t = \frac{b}{S_{b_1}}\ , \tag{7.6.7}$$

$$S_{b_1} = \frac{S_e}{\sqrt{SSx}}, \tag{7.6.8}$$

where S_{b_1} is the standard error of the regression coefficient b_1 and it measures the variation in the regression coefficient;

S_e is the standard error of the estimate.

A critical value for t is obtained from the table and compared with the t-value calculated from the sample by using Formula (7.6.7).

A confidence interval for β_1 is calculated then by using Formula (7.6.9)

$$\text{C.I. for } \beta_1 = b_1 \pm \text{t}S_{b_1}, \tag{7.6.9}$$

where the t-statistics has n-2 degrees of freedom

Chapter Checklist

You Make the Decision

1. The CEO for the Acme Trucking Company is concerned about recent trends in business performance. Profits are falling, and the stockholders are calling for the resignation. He comes to you for some answers regarding this predicament. You collect data on miles driven by the firm's trucks and resulting revenues the firm earned. The CEO asks if there might be a relationship between these two important variables and to what degree miles driven might explain revenues.

a. What do you do? How do you respond?

b. The CEO wants to know if your work will allow him to predict future revenues. How could you use you r statistical results to provide an estimate of revenues in near future?

c. Your results include the regression model

$$\hat{\text{Rev}} = 23.2 = 523.6\ \text{MD},$$

where Rev is earned revenues;

MD is miles driven by the firm's trucks,

with a correlation coefficient of 0.78. How would you interpret these results?

d. Your CEO feels confident in your statistical study given the fact that, he concludes, MD causes 78 percent of the change in Rev. How do you respond?

e. You tell the CEO that the regression line you have computed is a mean line and that it is the line of best fit. Having no knowledge of statistical analysis, he asks you to explain. What do you tell him?

Conceptual Questions

2. What is meant by "minimizing the sum of the errors squared" in your model for the trucking firm in Problem 1?

3. In what way might autocorrelation and heteroscedasticity present a problem in your regression model?

4. What is the difference between regression and correlation?

5. Identify the dependent and independent variables in each case:

   a. Time spent working on a term paper and the grade received.
   b. Height of a son and height of a father.
   c. A woman's age and the cost of her life insurance.
   d. Price of a product and the number of units purchased by an individual.
   e. Demand for a product and the number of consumers in the market.

8 TIME SERIES ANALYSIS AND FORECASTING

8.1 Introduction

Many business and economic studies are based on time-series data. Such data series offer many advantages to statistical analysts who wish o examine the business world in which they live. This is particularly true in their efforts to forecast and predict events. This chapter examines ways in which time-series data can be used to make forecasts, and how those forecasts can be used to make informed decisions, such as

a. The four components of a time series.
b. Two types of time-series models:
    - Additive model;
    - Multiplicative model.
c. Smoothing techniques:
    - Moving average.
    - Exponential smoothing.
d. The decomposition of a time series.

8.2. Time Series and Their Components

Developing a forecast often starts with the collection of past data over several time periods. The resulting data set is called a **time series** because it contains observations over time. The time periods can vary in length. They can be early, quarterly, or even daily. Time periods of only one hour may be used for highly volatile variables such as the price of heavily traded stock on one of the organized stock exchanges. Table 8.1 shows time-series data for the U.S. gross national product. It registers data for some variable (GNP) over time.

Table 8.1 – GNP Time-Series (in billions of dollars)

Year	GNP
1985	4,014.9
1986	4,240.3
1987	4,662.8
1988	4,939.2
1989	5,403.7

Sometimes time-series data are used to forecast future values from past observations. One approach to this effort is to simply estimate the value in the next time period to be equal to that of the last time period. That is,

$$\hat{Y}_{t+1} = Y_t,$$

where $\hat{Y}_{t+1}$ is the estimate of the value of the time series in the next time period, and Y_t is the actual value in the current time period.

Referred to as the **naive method of forecasting**, this approach might be used when the data exhibit a **random walk**.

Random walk movements demonstrate no trend upward or downward and typically shift direction suddenly. They can be expressed as

$$\hat{Y}_{t+1} = Y_t + a_t,$$

where a_t is some random amount, positive or negative, by which Y changes in time period t. because it totally random, it is virtually unpredictable. The best way is to simply use the most recent observation as the prediction for the next value.

That is, the *naive method of forecasting* uses the most recent observation for the forecast of the next observation.

All time series contain at least one of the following four components:

1. Trend.
2. Seasonal variation.
3. Cyclical variation.
4. Irregular or random variation.

8.2.1. *Secular Trend*

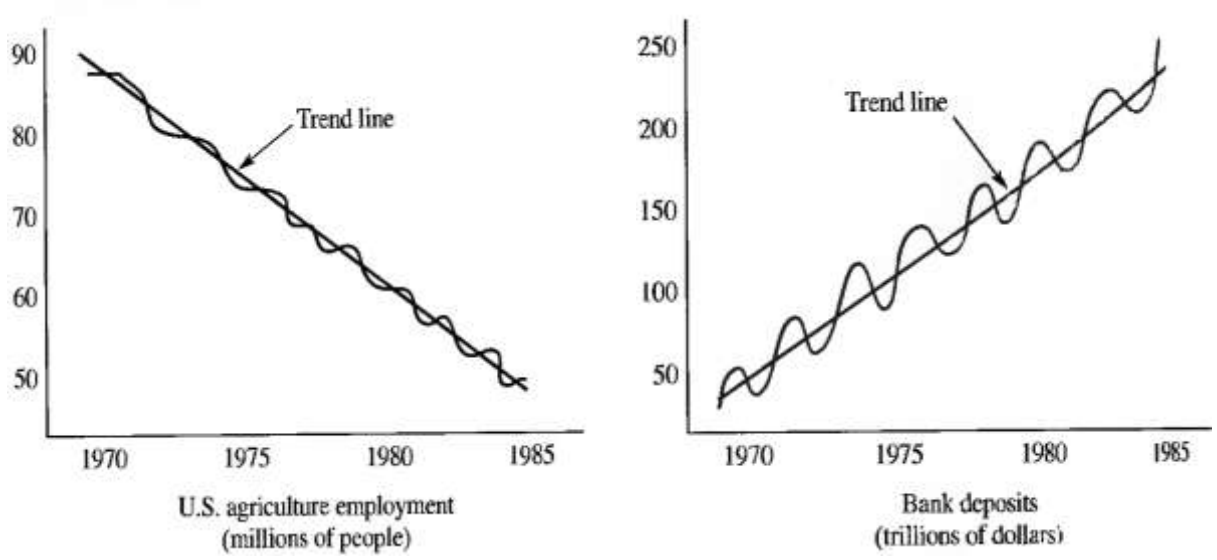
The **secular trend**, or merely trend, is the long-run behavior of the variable over an extended length of time. It reflects the general direction of the time series as upward or downward.

Examples include:

a. The rising number of foreign cars sold in the United States,.
b. The increase in the volume of credit transactions over the past few years.
c. The downward movement in the number of people living in rural areas in the last two decades.

Some variables move smoothly over time, while others show “fits and starts” that produce a rather bumpy ride.

Figure 8.1a shows the downward trend in agricultural employment over the last decade. There is a little variation around the steadily declining trend. Figure 8.1b, on the other hand, shows the upward secular movement in U.S. bank deposits. The data show considerable variation above and below the trend line drawn through the middle of the data.



a *b*

Figure 8.1 – Long-Term Trends in the U.S. Economic Activity: *a* – The Downward Trend; *b* – The Upward Secular Movement

8.2.2. *The Seasonal Component*

A lot of business activity is influenced by changing seasons of the year. For example, sales of certain seasonal goods such as Honda snowmobiles, Jantzen swimwear, and Hallmark Valentine cards would likely display a strong seasonal component.

Seasonal fluctuations are (regular) movements in the time series that reoccur each year about the same time.

Figure 8.2 shows how each year the unemployment rate tends to go up in May when high school students enter the summer job market, and goes down in November when retail stores hire temporary help to handle the Christmas rush. Notice that no apparent trend exists in the unemployment rate.

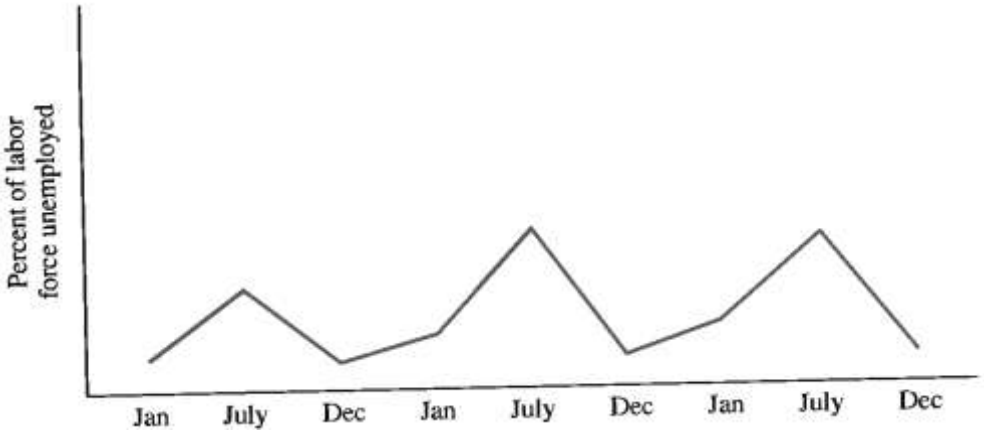


Figure 8.2 – Seasonal Fluctuations in Unemployment

For seasonal fluctuations to be detected, the time period in which the data are reported must be less than one year. Since seasonal fluctuations occur within the year, annual data will not capture or reflect seasonal changes; quarterly, monthly, or weekly data are necessary.

8.2.3. *Cyclical Variations*

Many variables exhibit a tendency to fluctuate above and below the long-term trend over a long period of time. These fluctuations are called **cyclical fluctuations** or **business cycles**. They cover much longer time periods than do seasonal variations, often encompassing three or more years in duration.

A cycle contains four phases:

1. The upswing or expansion, during which the level of business activity is accelerated, unemployment is low, and production is brisk;
2. The peak, at which point the rate of economics activity has “topped out”;
3. The downturn, or contraction, when unemployment rises and activity wanes;
4. The trough, where activity is at the lowest point.

A cycle runs from one phase to the next like phase and, as shown in Figure 8.3, fluctuates above and below the long-term trend in a wavelike manner.

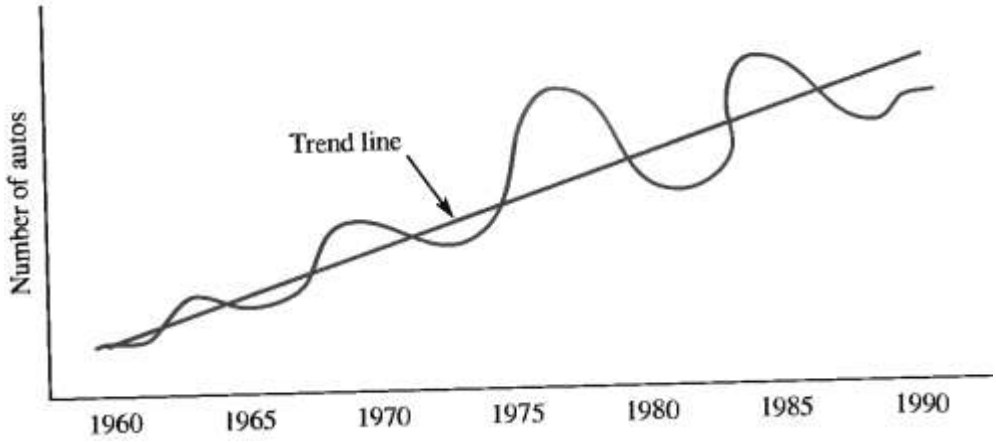


Figure 8.3 – Cyclical Fluctuations of Foreign Auto Imports

8.2.4. *Irregular Fluctuations*

Time series also contain **irregular**, or **random**, **fluctuations** caused by unusual occurrences producing movements that have no discernible pattern. These movements are, like fingerprints and snowflakes, unique, and unlikely to reoccur in similar fashion. They can be caused by events such as wars, floods, earthquakes, political elections, or oil embargoes.

8.3 Time-Series Models

A time-series model can be expressed as some combination of these four components. The model is simply a mathematical statement of the relationship amon g the four components. Two types of models are commonly associated with time series:

1) the additive model;
2) the multiplicative model.

The *additive model* is expressed as

$$Y_t = T_t + S_t + C_t + I_t ,$$

where Y_t is the *value of time series* for time period *t*, and the right-hand side values are the *trend*, the *seasonal variation,* the *cyclical variation*, and the *random* or *irregular variation*, respectively, for the same time period. In the additive model, all values are expressed in original units, and S, C, and I are deviations around T.

Example 8.1. If we were to develop a time-series model for sales in dollars for a local retail store, we might find that T = \$500, S = \$100, C = -\$25, and I = -\$10. Sales would be

$$Y = \$500 + \$100 - \$25 - \$10 = \$565.$$

Notice that the positive value for S indicates that existing seasonal influences have had a positive impact on sales. The negative cyclical value suggests that the business cycle is currently in a downswing. There was apparently some random event that had a negative impact on sales.

The additive y model suffers from the somewhat unrealistic assumption that the components are independent of each other. This is seldom the case in the real world. In most instances, movements in one component will have an impact on other components, thereby negating the assumption of independence. Or, perharps even more commonly, we often find that certain forces at work in the economy simultaneously affect two or more components. Again, the assumption of independence is violated.

As a result, the multiplicative model is often preferred. It assumes that the components interact with each other and do not move independently. The multiplicative model is expressed as

$$Y_t = T_t \times S_t \times C_t \times I_t .$$

In the multiplicative model, only T is expressed in the original units, and S, C, and I are stated in terms of percentages.

Example 8.2. Values for bad debts at a commercial bank might be recorded as T = $ 10 million, S = 1.7, C = 0.91 and I = 0.87. Bad debts could then be computed as

$$Y = (10)(1.7)(0.91)(0.87) = \$13.46 \text{ million.}$$

Since seasonal fluctuations occur within time periods of less than one year, they would not be reflected in annual data. A *time series for annual data* would be expressed as

$$Y_t = T_t \times C_t \times I_t .$$

8.4 Smoothing Techniques

A primary use of time-series analysis is to forecast future values. The general behavior of the variable can often be best discussed by examining its long term trend. However, if the time series contains too many random fluctuations or short-term seasonal changes, the trend may be somewhat obscured and difficult to observe. It is often possible to eliminate many of these confounding factors by averaging the data over several time periods. This is accomplished by the use of

certain smoothing techniques that remove random fluctuations in the series, thereby providing a less obstructed view of the true behavior of the series. Two common methods of smoothing time series data are examined: a moving average and exponential smoothing.

8.4.1. *Moving Averages*

A **moving average** (MA) will have the effect of "smoothing out" the data, producing a movement with fewer peaks and valleys. It is computed by averaging the values in the time series over a set number of time periods. The same number of time periods is retained for each average by dropping the oldest observation and picking up the newest. Assume that the closing prices for a stock on the New York Stock Exchange for Monday through Wednesday were \$20, \$22, and \$18, respectively. We can compute a three-period (day) moving average as

$$(20 + 22 + 18) / 3 = 20.$$

This value of 20 then serves as a forecast or estimate of what the closing price might be at any time in the future. If the closing on Thursday is, say 19, the next moving average is calculated by dropping Monday's value of 20 and using Thursday's closing price of 19. Thus, the forecast becomes

$$(22 + 18 + 19) / 3 = 19.67.$$

The estimate figured in this manner is seen as the long-run averages of the series. It is taken as the forecast for the closing price on any given day in the future.

Thus, **MA**: *a series of arithmetic averages over a given number of time periods, it is the estimate of the long-run average of the variable*.

Example 8.3. The sales for Arthur Momitor's Snowmobiles, Inc., over the past 12 months are shown in Table 8.2. Both a three month MA and a five-month MA are calculated.

Table 8.2 – Snowmobile sales for Arthur Momitor

Month	Sales (\$100)	Three-Month MA	Five-Month MA
January	52		
February	81	60.00	
March	47	64.33	59.00
April	65	54.00	63.20

Continuation of Table 8.2

Month	Sales ($100)	Three-Month MA	Five-Month MA
May	50	62.67	56.00
June	73	56.00	58.60
July	45	59.33	55.60
August	60	51.67	61.40
September	50	63.00	55.80
October	79	58.00	59.20
November	45	62.00	
December	62		

The first entry in the three-month MA is obtained by averaging the sales of snowmobiles in January, February, and March. The resulting value of (52 + 81 + 47)/3 = 60 is centered on the middle time period of February. The next entry is determined by averaging February, March, and April, and centering the value of 64.33 in the middle of those three periods, which is March. The remaining entries are determined similarly.

The first entry in the five-month MA series uses values for months January through May. The average of (52+51+47+65+50)/5 = 59 is centered in the middle of those five time periods at March.

Moving averages have the effect of smoothing out large variations in the data. This smoothing effect occurs because unusually small observations are averaged in with other values, and their impact is thereby restrained. The larger the number of time periods in a moving averageб the more pronounced the smoothing effect will be. Notice that the range of values in the three-month MA is less than that in the original data and greater than the range found in the five-month MA. Figure 8.4 illustrates this tendency for the smoothing effect to increase with the number of time periods in the moving average.

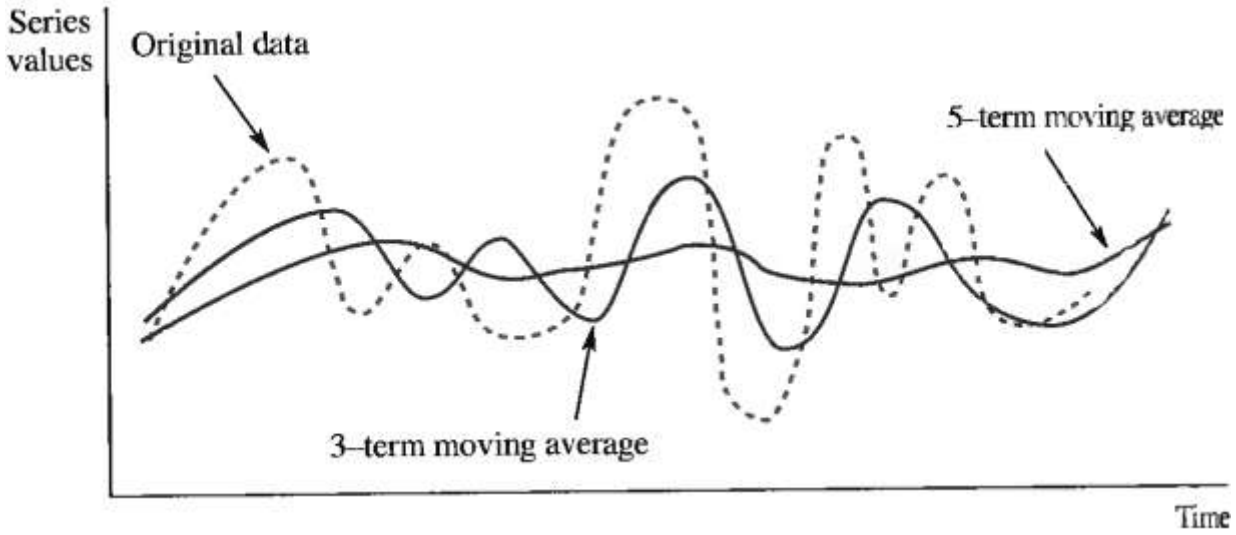


Figure 8.4 – Comparing Moving Averages

Notice that when an odd number of time periods is used in the moving average, the results can be automatically centered at the middle time period. However, if there is an even number of time periods, there is no middle observation at which the value can be automatically centered.

Moving averages can be used to remove irregular and seasonal fluctuations. Each entry in the moving average is derived from four observations of quarterly data – that is, one full year's worth. Thus, the moving average "averages out" any seasonal variations that might occur within the year, effectively eliminating them and leaving only trend and cyclical variations.

In general, if the number of time periods in a moving average is sufficient to encompass a full year (12 if the monthly data are used; 52 if weekly data are used), seasonal variations are averaged out and removed from the series. The data are then said to be **deseasonalized**.

As noted, the use of a larger number of time periods results in a smoother averaged series. Therefore, if the data are quite volatile, a small number of periods should be use in the forecast to avoid placing the forecast too close to the long-run average. If the data do not vary greatly from the long-run mean, a larger number of time periods should be used in forming the moving average.

The moving average method of forecasting is best used when the data show no upward or downward trend. It is a somewhat simplistic approach and finds its most common use in the decomposition of time series.

8.4.2 *Exponential Smoothing*

Exponential smoothing also has the effect of smoothing out a series. It also provides an effective means of prediction. First-order exponential smoothing is used when the data do not exhibit any trend pattern. The model contains a self-correcting mechanism that adjusts forecast on the opposite direction of past errors. The equation is

$$F_{t+1} = \alpha A_t + (1-\alpha)F_t \,, \qquad (8.1)$$

where F_{t+1} is the forecast for the next time period

A_t is the actual, observed value for the current time period

F_t is the forecast previously made for the current time period

α is a "smoothing constant" which is given a value between and 1. Since the data do not trend up or down but fluctuate around some long-term average, the value F_{t+1} is taken as the forecast for any future time period.

Thus, *exponential smoothing is a forecasting tool in which the forecast is based on a weighted average of current and past values*.

Example 8.4. Suppose it is currently the last business day of February. Sales for Uncle Vito's Used Cars for the month have been compiled and total $110 thousand. Uncle Vito has decided to forecast sales for March.

Table 8.3 – Uncle Vito's Auto Sales ($1000)

Month	Forecast	Actual	Error (F_t - A_t)
January	-	105	
February	105	110	-5.0
March	106.5	107	-0.5
April	106.65	112	-5.35

According to (8.1), the March forecast, F_{t+1}, requires

a. February's actual sales, A_t.

b. The forecast for February, F_t.

However, since March is the first month in which Uncle Vito is developing his forecast, there was no forecast made for February and F_t is unknown. The general practice is to simply use the actual value of the previous time period, January in this case, for the first forecast.

Uncle Vito's records show that January sales were $ 105 thousand. Assuming a value of 0.3 for α, the forecast for March is

$$F_{t+1} = \alpha A_t + (1-\alpha)F_t = \alpha A_{Feb} + (1-\alpha)F_{Feb} = 0.3*110 + 0.7*105 =$$

$$= \$\ 106.5 \text{ thousand as the forecast for sales in March.}$$

As table 8.3 reveals, Uncle Vito can plan for sales of $ 106.5 thousand. If actual sales in March are $ 107 thousand, the error is computed as $F_t - A_t = 106.5 - 107 = -0.5$. Also, $F_{Apr} = (0.3)(107) + (0.7)(106.5) = 106.65$.

Assume sales in April prove to be $112 thousand. The error is then - $ 5.35 thousand. Uncle Vito can also predict sales for May:

$$F_{t+1} = \alpha A_t + (1-\alpha)F_t = \alpha A_{Apr} + (1-\alpha)F_{Apr} = 0.3*112 + 0.7*106.65 =$$

$$= \$\ 108.26 \text{ thousand.}$$

The value selected for α is critical. Since it's desirable to produce a forecast with the smallest possible error, the α – value that minimizes the **mean square error** (MSE) is optimal. Trial and error often serves as the best method to determine the proper α – value. Table 8.4 contains Uncle Vito's actual sales data for the first seven months.

Table 8.4 – Sales Data for Uncle Vito

Month	Actual	Forecast (α = 0.3)	Error	Forecast (α = 0.8)	Error
January	105				
February	110	105.00	-5.0	105.00	-5.0
March	107	106.50	-0.5	109.00	2.00
April	112	106.65	-5.35	107.40	-4.60
May	117	108.26	-8.74	111.08	-5.92
June	109	110.88	1.88	115.82	6.82
July	108	110.32	2.32	110.36	2.36
August		109.62		108.47	

Errors are based on forecasts calculated using α – values of 0.3 and 0.8. The MSE is

$$MSE = \frac{\sum (F_t - A_t)^2}{n-1}. \qquad (8.2)$$

For α = 0.3, the MSE is

$$MSE = \frac{(-5)^2 + (-0.5)^2 + (-5.35)^2 + (-8.74)^2 + (1.88)^2 + (2.32)^2}{7-1} = 23.20.$$

An α of 0.8 yields

$$MSE = \frac{(-5)^2 + (2)^2 + (-4.6)^2 + (-5.92)^2 + (6.82)^2 + (2.36)^2}{7-1} = 22.88.$$

An α of 0.8 produces better forecasting results since it generates a smaller error factor. Other values of α may be tried to determine their impact on MSE and the accuracy of the resulting forecasts. Generally speaking, if the data are rather volatile, a lower α-value is called for. This is because smaller values for α assign less weight to more recent observations. If the data show considerable movement, the last observation may not be representative of the long-run average.

Remember, first-order exponential smoothing in the manner described here is appropriate if the data show no trend, but move around some average value over the long run. If a downward or upward trend can be detected by plotting the data, second-order exponential smoothing, the mechanics of which will not be examined here, should be used.

Unlike moving averages, which use only a set number of time periods of data, exponential smoothing uses all past values of the time series. This is because F_{t+1} depends on A_t and F_t. Yet, F_t used A_{t-1} and F_{t-1} in its calculation, and F_{t-1} used A_{t-2} and F_{t-2}. Thus, each forecast depends on previous actual values of A_{t-n} all the way back to where the forecasts first began. The farther back in time you go, the less emphasis a value of A has on the current forecast.

8.5 Decomposition of a Time Series

In this section the techniques used to isolate the four components of a time series are examined. This procedure is known as **decomposition.** Decomposition can be used to measure the degree of impact each component has on the direction of the time series itself. We begin by measuring the trend in a time series.

8.5.1. *Secular Trend*

By estimating the trend in a time series, it can be removed from the actual value and thereby the size of the remaining components is determined. If the change in a time series is relatively constant, a linear trend model will likely provide a fairly accurate forecast. Since many business and economic variables

display a tendency to change by a constant amount over time, a linear model is appropriate.

The most widely used method to fit a linear trend is the method of ordinary least squares discussed in Chapter 9. The main difference between the procedure discussed earlier and that presented here is that, in trend analysis, the independent (right- hand side) variable is time. The linear relationship can be written as

$$\hat{Y}_t = b_0 + b_1 t. \qquad (8.3)$$

where $\hat{Y}_t$ is the estimated value of the dependent variable; b_0 is the intercept of the trend line; b_1 is the slope of the trend line; *t* is the independent variable, *X*.

Table 8.5 contains data for the number of housing starts in Happy Valley, California, over a 16-year period. Mayfield Construction wants to fit these time-series observations, using OLS, to develop a model that can predict future housing starts.

Table 8.5 - Housing starts in Happy Valley (in 100's)

Year	t(X)	Housing Starts (Y)	XY	X^2
1977	1	7.0	7.0	1
1978	2	7.1	14.2	4
1979	3	7.9	23.7	9
1980	4	7.3	29.2	16
1981	5	8.2	41.0	25
1982	6	8.3	49.8	36
1983	7	8.1	56.7	49
1984	8	8.6	68.8	64
1985	9	8.8	79.2	81
1986	10	8.9	89.0	100
1987	11	8.7	95.7	121
1988	12	9.1	109.2	144
1989	13	9.4	122.2	169
1990	14	9.1	127.4	196
1991	15	9.5	142.5	225
1992	16	9.9	158.4	256
	136	135.9	1,214.0	1,496.0

The values for t are obtained by coding each time period starting with 1 for the first time period, 2 for the second, and so on. The sums of squares and cross-products, used to calculate the regression line, are

$$SSx = \sum X^2 - \frac{(\sum X)^2}{n} = 1.496 - \frac{136^2}{16} = 360,$$

$$SSxy = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 1.214 - \frac{136*135.9}{16} = 58.85.$$

The formulas for b_1 and b_0 are

$$b_1 = \frac{SSxy}{SSx} = \frac{58.85}{340} = 0.173, \qquad b_0 = \overline{Y} - b_1 \overline{X} = 7.02.$$

The equation for the trend line is

$$\hat{Y}_t = 7.02 + 0.173t.$$

Figure 8.5 displays the raw data and the trend line they produce.

Given this equation, it is possible to predict the number of housing starts for future time periods merely by substituting the appropriate value for *t*. Suppose Mayfield Construction wants to forecast the housing starts for 1993. Since the value of *t* would be 17 in 1993, the forecast becomes

$$\hat{Y} = 7.02 + 0.173(17) = 9.96.$$

or 9,960 starts, since the data were expressed in units of 1,000.

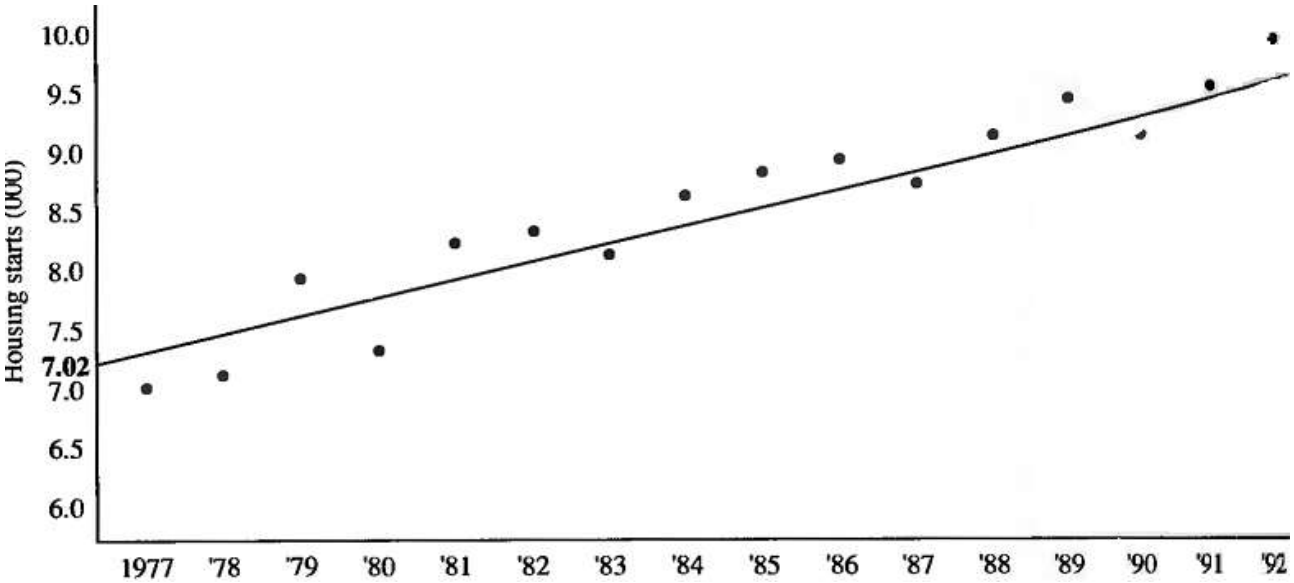


Figure 8.5 – The number of housing starts in Happy Valley and the trend line they produce

Similarly, since 1995 would carry a *t* value of 19, the forecast for 1995 would be

$$Y = 7.02 + 0.173(19) = 10.31.$$

It is estimated that there will be 1,031 housing starts in 1995.

Of course, the farther into the future a forecast is made, the less confidence you can place in its precision. Additionally, its accuracy is based on the condition that the past provides a representative picture of future trends.

8.5.2. *Seasonal Variation*

Many businesses experience seasonal variations in the level of their activity. Changes in weather and climate affect business conditions in agriculture and construction, as well as related industries such as farm implements and timber. Many commodities, such as swimwear and snow skis are influenced by changes in the season. Artificial seasons based on social customs, including Christmas, June weddings, and May graduations, have an impact on business activity. Thanksgiving and Easter affect the poultry and egg industries.

Seasons do not have to be as long as those implied above. Organized stock exchanges find that trading is heavier on Fridays and Mondays than it is on other weekdays. Here, the "season" is a single day.

The study of seasonal fluctuations lends much to our ability to evaluate and understand business behavior. The ultimate aim is to determine a seasonal index, which can be used to analyze and predict business activity.

The data in Table 8.7 show monthly profits for Vinnie's Video Village. A superficial examination reveals that profits seem to be higher during the summer months when school is out, and lower at other times of the year. This suggests the presence of seasonal factors.

Table 8.7 - Seasonal Fluctuations in Vinnie`s Profits

Time Period	(Y) Profits ($100's)	Month MA (T-C)	Centered MA	Ratio to MA V/CMA = S*I
1991				
January	10			
February	9			
March	11			
April	12			

Continuation of Table 8.7

Time Period	(Y) Profits ($100's)	Month MA (T-C)	Centered MA	Ratio to MA V/CMA = S*I
May	18			
June	23			
July	27	15.5833	15.5417	1.7373
August	26	15.5000	15.5833	1 6685
September	18	15.6667	15.6250	1.1520
October	13	15.5833	15.5833	0.8342
November	10	15.5833	15.6250	0.6400
December	10	15.6667	15.7500	0.6349
1992				
January	9	15.9167	15.8750	0.5669
February	11	16.3333	16.1250	0.6822
March	10	16.6667	16.5000	0.6061
April	12	16.8333	16.7500	0.7164
May	19	16.9167	16.8750	1.1259
June	25	17.0833	17.0000	1.4706
July	28	17.1667	17.1250	1.6350
August	31	16.9167	17.0417	1.8191
September	22	16.9167	16.9167	1.3005
October	15	16.9167	16.9167	0.8867
November	11	16.9167	16.9167	0.6502
December	12	16.9167	16.9167	0.7094
1993				
January	10	17.0000	16.9583	0.5897
February	8	17.0000	17.0000	0.4706
March	10	16.9167	16.9583	0.5897
April	12	17.0000	16.9583	0.7076
May	19	17.5833	17.2916	1.0988
June	25	18.1667	17.8750	1.3986
July	29			
August	31			
September	21			
October	16			
November	18			
December	19			

The first step in developing a seasonal index is to calculate a centered moving average. Since Vinnie's profits tend to fluctuate over the course of the year, and monthly data are used, a 12-period (month) moving average should be calculated. If activity on organized stock exchanges has to be analyzed, the daily data should be used and a five-period (for the five business days) moving average has to be constructed.

Table 8.7 shows the 12-month moving average and the centered moving average (CMA). As noted, the moving average eliminates recurring seasonal movements as well as any random effects over the course of the year. Thus, given a multiplicative model Y = T * C * S * I**,** the moving average eliminates S and I and contains only T and C. That is, MA = T * C.

It is now possible to calculate the *ratio to moving average*. To do this, divide the original series value Y by the moving average. The result produces the S and I components of the time series.

$$\frac{Y}{MA} = \frac{T \times C \times S \times I}{T \times C} = S \times I.$$

By dividing the time-series values by the moving average the I component will be removed shortly.

Ratio to Moving Average. *By dividing the original time-series data by the moving average, the ratio to moving average is obtained, which contains the S and I components.*

A **mean ratio to moving average** is calculated for each month as shown in Table 8.8. These mean ratios are then normalized to produce the seasonal indexes. The purpose of this normalization procedure is to ensure that the seasonal indexes will sum to 12, since a 12-period moving average is used. This is accomplished by multiplying each mean by a *normalization ratio***,** which is the ratio of 12 to the sum of the means. Notice in Table 8.8 that the sum of the means of the ratio to moving averages is 11.8454. Thus, the normalization ratio is

$$\frac{12}{11.8454} = 1.01305.$$

Then the seasonal indexes are found by multiplying each mean by 1.01305. This normalization removes the irregular component, leaving only the seasonal factor.

Uses of the Seasonal Index

1. After going to all the trouble of calculating these seasonal indexes, you will be glad to know that they are put to vital use. For example, the seasonal index for a particular month indicates *how that month performs relative to the year as a whole.*

Table 8.8 - Seasonal Indexes for Vinnie`s Profits

Month	1991	1992	1993	Mean Ratio to MA	Seasonal Index
January		0.5669	0.5897	0.5783	0.5858
February		0.6822	0.4706	0.5764	0.5839
March		0.6061	0.5897	0.5979	0.6057
April		0.7164	0.7076	0.7120	0.7213
May		1.1259	1.0988	1.1124	1.1269
June		1.4706	1.3986	1.4346	1.4533
July	1.7373	1.6350		1.6861	1.7082
August	1.6685	1.8191		1.7438	1.7665
September	1.1520	1.3005		1.2262	1 2422
October	0.8342	0.8867		0.8605	0.8717
November	0.6400	0.6502		0.6451	0.6535
December	0.6349	0.7094		0.6721	0.6809
				11.8454	11.9999 ≈12

The index of 0.5858 for January tells Vinnie that profits in January are only 58.58 percent of the average for the full year. Profits are 41.42 percent (1.000 — 0.5858) below the year's monthly average.

2. Perhaps more importantly, the indexes can be used to *deseasonalize data.* This has the effect of removing seasonal variation from a series to determine what the values would be in the absence of seasonal variations. It yields the average value per month that would occur if there were no seasonal changes. The deseasonalized value is found by dividing the actual value during the month by the seasonal index in that month. For example, in January 1991, the deseasonalized value is

$$\frac{10}{0.5858} = 17.07.$$

In other words, if Vinnie's business was not subject to seasonal variation, profits in January 1991 would have been $1,707.

Deseasonalized values are also called *seasonally adjusted* because they tell us what the values would be if we adjust for seasonal influences. The classic example involves unemployment rates. Since unemployment is usually higher in May than most other months due to school dismissals and the influx of many teenagers into the job market, the seasonal index for May will be greater than 1. If actual unemployment in May is 7.2 percent and the index is, say, 1.103, the deseasonalized, or seasonally adjusted, rate of unemployment is 7.2/1.103 = 6.53 percent. This is not to say that unemployment was 6.53 percent. (It was actually 7.2 percent.) But when we adjust for seasonal forces, which typically inflate the rate of unemployment in May, the deseasonalized rate is lower. In this manner a measure or index of seasonal variation can be used to determine if the change in some series is more or less than what might be expected given the typical seasonal behavior.

Deseasonalized Values *Deseasonalized values are obtained by dividing the actual values by their respective seasonal indexes. They reflect what the variable would be if we adjusted for seasonal influence*.

3. The reverse is possible, in that the seasonal index can be used to *seasonalize data* to get a better picture of what any one month might generate in profits. Assume Vinnie felt that profits might total 190 during the year. Without any seasonalization it might be argued that each month would generate 190/12 = 15.83, or $1,583 in profits. However, Vinnie knows that monthly variations will occur. He could seasonalize the data to determine the extent of that monthly variation by multiplying 15.83 by the seasonal index. He knows that in January profits tend to be 58.58 percent of the yearly total. His estimate of profits for January is (15.83X0.5858) = 9.27, or $927.

Or perhaps Vinnie is working with the trend equation which, given the data, is

$$Y_t = 13.85 + 0.167t.$$

The forecast for January 1994, the 37^{th} time period, is

$$Y_t = 13.85 + 0.167(37) = 20.31.$$

This doesn't account for the seasonal lows that occur in January. The value can be seasonalized by multiplying by the seasonal index for January, yielding (20)*(0.5858) = 11.73, which is probably more accurately reflects profits during that month.

8.5.3. *Cyclical Variation*

Many businesses are affected by swings in the business cycle. When the economy in general turns up, their business activity may accelerate, while an economic downturn brings on a drop in business. Some industries often exhibit movements in the opposite direction of the cycle. The entertainment industry, for example, has been known to experience counter cyclical movements. Presumably, when economic conditions worsen, many people seek relief from harsh reality by escaping to the movies.

The cyclical component can be identified by first obtaining the trend and seasonal components as described earlier.

The statistical norm is then calculated by multiplying the trend projection by the seasonal index. This is called the *norm* because it represents the values that would occur if only the trend and seasonal variations were present.

The cyclic and irregular components are obtained next by dividing the original data by the statistical norm, which contains *T* and *S***.** That is, since T=T×S×C×L

$$\frac{Y}{T \times S} = \frac{T \times S \times C \times I}{T \times S} = C \times I.$$

The results are then multiplied by 100 to express the answer in percentage form. Note that if annual data are used, they will, by definition, contain no seasonal variations. The seasonal index would be unnecessary. The values of the time series would consist only of

$$Y=T\times C\times \mathrm{I}.$$

The components *CI* could be found directly by dividing only by the trend values:

$$\frac{Y}{T} = \frac{T \times C \times I}{T} = C \times I.$$

8.5.4. *Irregular Variation*

Having isolated the other three components of a time series, little more need be said about irregular variations. Suffice it to say that it is often possible to smooth out and effectively eliminate them by using a moving average.

Figure 8.6 portrays the relationship between the various forecasting tools. The methods in this chapter are by no means exhaustive. There are many other techniques, some considerably more sophisticated than those examined here, that can be used to provide accurate forecasts.

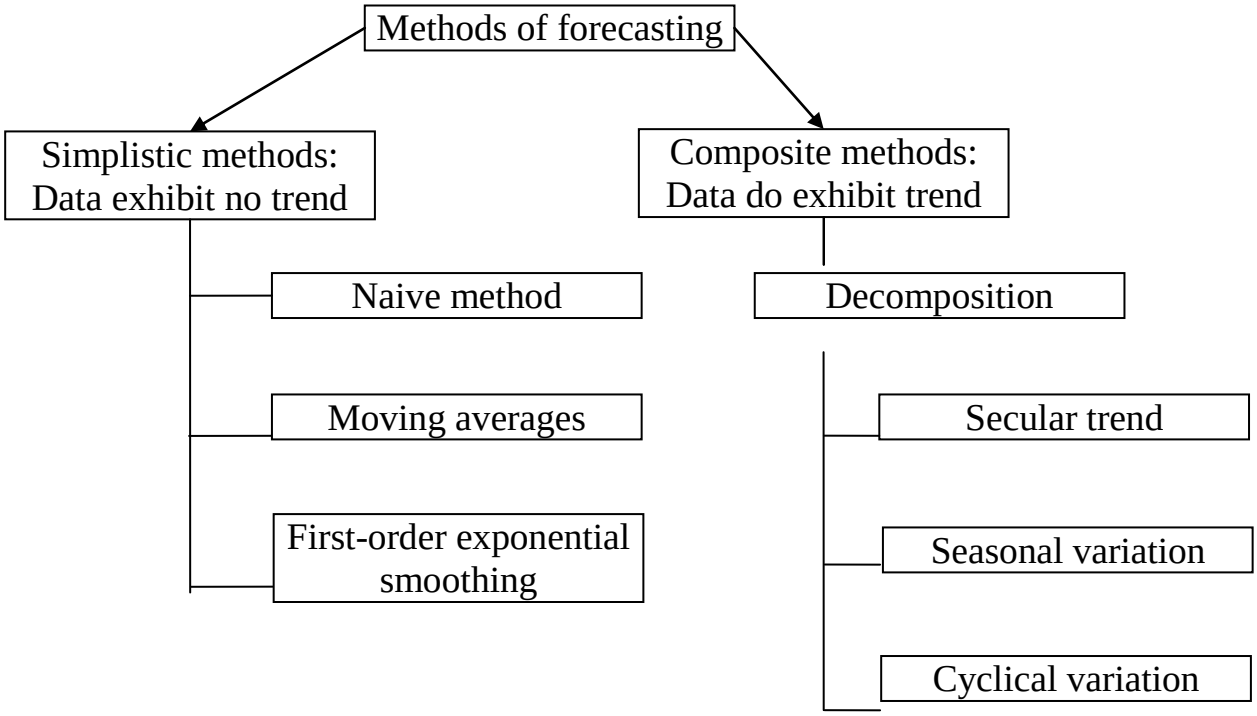


Figure 8.6 - Methods of forecasting

Chapter Checklist

1. After studying this chapter, can you
2. Define a time series and its four components?
3. Decompose a time series?
4. Calculate a moving average and forecast with the results?
5. Explain exponential smoothing and forecast with it?
6. Explain the effect of the smoothing constant?
7. Use the ordinary least squares method to forecast?
8. Explain seasonal variation?
9. Calculate and explain seasonal indexes?
10. Calculate and explain the nature of seasonally adjusted values?

9 INDEX NUMBERS

9.1 Introduction

Analysts of business and economic data find many uses for price indexes. Many facts and conditions can be discovered only through the application of index numbers to data bases containing economic variables. This chapter examines how index numbers are used in the analysis of business problems (Figure 9.1).

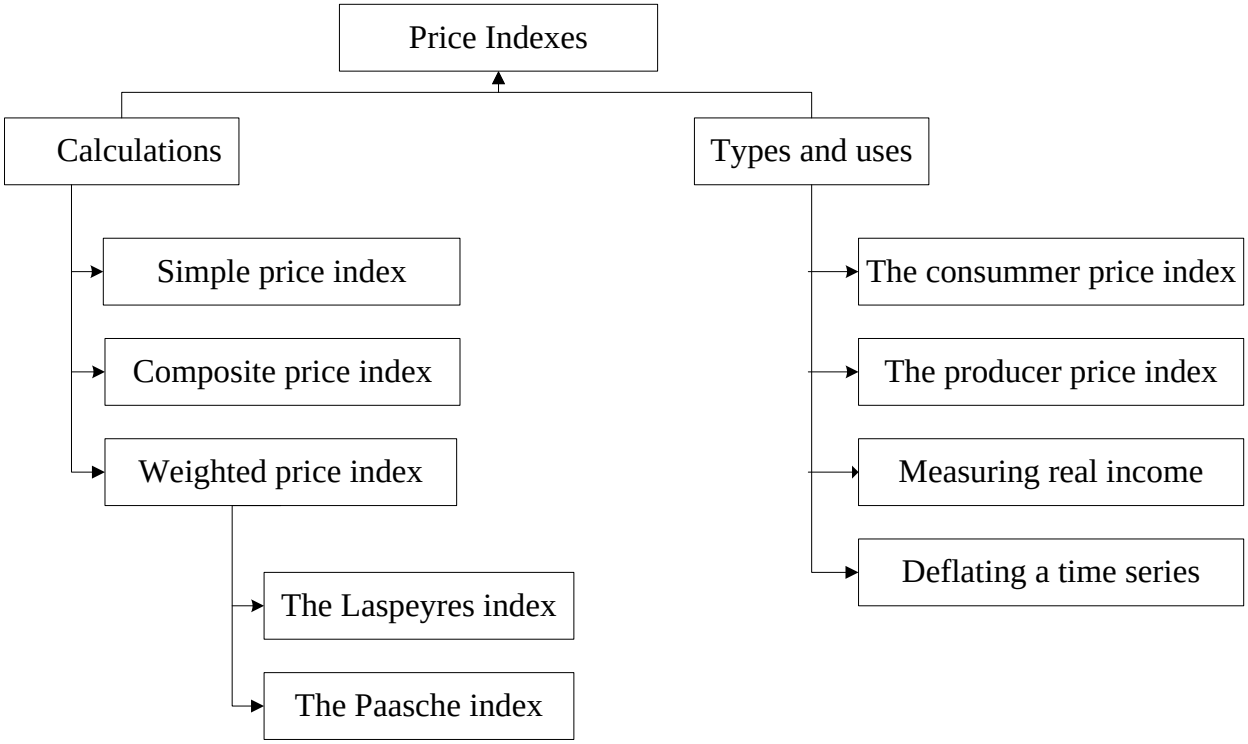


Figure 9.1 – Chapter 11 Structure

Business and economic conditions tend to fluctuate widely over time. These vacilations make it difficult to analyze essential business data or to properly interpret economic variables. Comparisons from one time period to the next often become misleading.

The use of index numbers can alleviate many of these problems. Decision makers obtain a more accurate picture of the behavior of economic variables and the relationships that exist among these variables. An index number relates a value in one time period, called the **base period**, to another value in a different time period, called the **reference** (or **current**) **period**. The use of index numbers has become prevalent within our business and economic communities over the last several years.

Index numbers are important to many private business decisions, and are crucial in measuring the impact of many government socioeconomic programs. The description of the role of index numbers covers

3. Simple price index.
4. Composite price index.
5. Weighted price indexes:
   – The Laspeyres index.
   – The Paasche index.
6. Relative indexes.
7. The base period and how to change it.
8. Specific indexes such as the consumer price index (CPI).

9.2 A Simple Price Index

A simple price index characterizes the relationship between the price of a good or service at one point in time, called the *base period,* to the price of that same good or service at a different point in time, called the *reference period.*

To calculate a simple index, you merely divide the price of the commodity in the reference period by its price in the base period and multiply by 100.

Example 9.1. Determine the price index for the reference period 2003 and chose 2008 as the base period. Thus,

$$PI_R = \frac{P_R}{P_B} \times 100, \tag{9.1}$$

$$PI_{2003} = \frac{P_{2003}}{P_{2008}} \times 100.$$

where *PI* is the price index and *P* is the price in the respective years.

Example 9.2. Jack Nipp and his partner, Harry Tuck, own a meat packing plant in Duluth. Data for their three most popular items are shown in Table 9.1.

Table 9.1 – Data for Nip and Tuck, Inc.

Item	Unit	Price/Unit		
		2003	2004	2005
Beef	1pound	3	3.3	4.5
Pork	1pound	2	2.2	2.1
Veal	1pound	4	4.5	3.64

Nipp tells Tuck to compute a simple price index for each product with 2003 as the base period. Using Formula (9.1), Tuck finds that the price indexes for beef in each of the three years are

$$PI_{2003} = \frac{P_{2003}}{P_{2003}} \times 100 = \frac{3.00}{3.00} \times 100 = 100,$$

$$PI_{2004} = \frac{P_{2004}}{P_{2003}} \times 100 = \frac{3.30}{3.00} \times 100 = 110,$$

$$PI_{2005} = \frac{P_{2005}}{P_{2003}} \times 100 = \frac{4.50}{3.00} \times 100 = 150.$$

From the base year of 2003 to 2004, the price index rose from 100 to 110. Tuck can therefore conclude that the price of beef increased by 10 percent. This is calculated as the difference between the two index numbers divided by the base number. That is,

$$\frac{PI_{2004} - PI_{2003}}{PI_{2003}} = \frac{110 - 100}{100} = 10\%.$$

Similarly, it can be concluded that a 50 percent increase occurred from 2003 to 2005:

$$\frac{PI_{2005} - PI_{2003}}{PI_{2003}} = \frac{150 - 100}{100} = 50\%.$$

You might want to conclude that a 40 percent increase in price occurred from 2004 to 2005 since the price index increased by 40. However, this is not the case. The percentage increase from 2004 to 2005 is

$$\frac{PI_{2005} - PI_{2004}}{PI_{2004}} = \frac{150 - 100}{110} = 36.4\%.$$

The 40 percent difference between the index numbers in 2004 and 2005 is called the ***percentage point*** increase, not the percentage increase.

Notice that the price index in the base year is always 100. This will always be the case since the price in the base year is, of course, 100 percent of itself.

The indexes for pork and veal are calculated in similar fashion and are shown in Table 9.2. Notice that the 2005 index for veal is less than 100. This reflects the fact that veal prices in 2005 were lower than they were in the base year of 2003. Specifically, prices for veal went down by (100 — 91)/100 = 9 percent from 2003 to 2005.

Table 9.2 – Price Indexes for Nipp and Tuck, Inc. (2003=100)

Item	2003	2004	2005
Beef	$\frac{3.00}{3.00}\times 100 = 100$	$\frac{3.30}{3.00}\times 100 = 110$	$\frac{4.50}{3.00}\times 100 = 150$
Pork	$\frac{2.00}{2.00}\times 100 = 100$	$\frac{2.20}{2.00}\times 100 = 110$	$\frac{2.10}{2.00}\times 100 = 105$
Veal	$\frac{4.00}{4.00}\times 100 = 100$	$\frac{4.50}{4.00}\times 100 = 112.5$	$\frac{3.64}{4.00}\times 100 = 91$

9.3 Composite Price Indexes

Often we want to calculate a price index for several goods. This is called a composite price index. Firms that produce two or more products are usually interested in a composite index. So are many government agencies who chart consumer behavior. The U.S. Department of Labor compiles the consumer price index, which measures relative prices for a typical "market basket'' of goods and services consumed by the general public.

The composite index is computed by adding the price of the individual commodities in the reference year and dividing by the summation of those prices in the base year. The result is then multiplied by 100.

$$PI_R = \frac{\sum P_R}{\sum P_B} \times 100. \qquad (9.2)$$

Using the data for Nipp and Tuck, the 2003 composite index for all three products, retaining 2003 as the base period, is

$$PI_{2003} = \frac{3.00 + 2.00 + 4.00}{3.00 + 2.00 + 4.00}(100) = 100.0.$$

The index for 2004 is

$$PI_{2004} = \frac{3.30 + 2.20 + 4.50}{3.00 + 2.00 + 4.00}(100) = 111.11.$$

And 2005 produces

$$PI_{2005} = \frac{4.50 + 2.10 + 3.64}{3.00 + 2.00 + 4.00}(100) = 113.8.$$

9.4 Weighted Composited Price Indexes

At least two problems arise with the use of composite price indexes. The first concerns the arbitrary nature in which the units are expressed. Had Nipp and Tuck priced beef at $1.50 per half pound instead of $3.00 per pound, the price index would have been entirely different. Second, the composite indexes as computed do not take into account the fact that some goods sell in larger quantities than do other, less popular, products. No consideration is given to the respective amounts of each product that are sold.

For example, the composite index calculated for Nipp and Tuck gives the same importance, or weight, to beef as to pork, even though twice as much of the former may have been purchased by the consumers. In this case it's better to compute a **weighted price index**. Such a calculation assigns different weights to individual prices. These weights are established so as to measure the amounts sold of each product. This provides a more accurate reflection of the true cost of the consumer's market basket of goods.

The quantities selected as weights can be taken from the number of units sold in (1) the base period or (2) the reference period.

Two common indexes are the *Laspeyres index* and the *Paasche index*. The Laspeyres index uses quantities sold in the base year as weights; the Paasche index relies on quantities sold in the reference year as weights. Each procedure has its own advantages and disadvantages.

The **Laspeyres index** uses **base period** weights (quantities) in its calculation. The rationale is that these quantities will not change from one calculation to the next thereby permitting more meaningful comparisons over time.

To illustrate, consider the data for Nipp and Tuck in Table 9.3, which also includes the amounts sold for each product.

Table 9.3 – Nipp and Yuck, Inc.

Item	Unit	Price/Unit			Quantity Sold (100's lb)		
		2003	2004	2005	2003	2004	2005
Beef	1 pound	3.00	3.30	4.50	250	320	350
Pork	1 pound	2.00	2.20	2.10	150	200	225
Veal	1 pound	4.00	4.50	3.64	80	90	70

The Laspeyres index is

$$L = \frac{\sum(P_R \times Q_B)}{\sum(P_B \times Q_B)} \times 100, \tag{9.3}$$

where P_R is the price in the reference period, and P_B and Q_B are the price and quantities sold in the period selected as the base period.

Table 9.4 shows computations necessary for the Laspeyres index using 2003 as the base year.

Table 9.4 – The Laspeyers Index for Nipp and Tuck (2003 = 100)

Item	Prices			Quantities in 2003	P_R x Q_B		
	2003	2004	2005		$P_{2003} \times Q_{2003}$	$P_{2004} \times Q_{2003}$	$P_{2005} \times Q_{2003}$
Beef	3.00	3.30	4.50	250	750	825	1125
Pork	2.00	2.20	2.10	150	300	330	315
Veal	4.00	4.50	3.64	80	320	360	291.2
-	-	-	-	-	1370	1515	1731.2

The numerator for L is figured by first multiplying each price by the quantities sold in the base period of 2003. The denominator is then determined by multiplying the price in the base year by the quantity in the base year. The index for 2003 is

$$L_{2003} = \frac{\sum(P_{2003} \times Q_{2003})}{\sum(P_{2003} \times Q_{2003})} \times 100 = \frac{1{,}370}{1{,}370}(100) = 100.$$

The index for 2004 uses the prices in the reference year (2004) and the quantities in the base year (2003) for the numerator:

$$L_{2004} = \frac{\sum(P_{2004} \times Q_{2003})}{\sum(P_{2003} \times Q_{2003})} \times 100 = \frac{1515}{1370}(100) = 110.58.$$

The numerator for 2005 uses prices in 2005 and quantities in 2003:

$$L_{2005} = \frac{\sum(P_{2005} \times Q_{2003})}{\sum(P_{2003} \times Q_{2003})} \times 100 = \frac{1731.2}{1370}(100) = 126.36.$$

The interpretation of the Laspeyres index is like that for the earlier indexes. From 2003 to 2005, the price of the market basket for these three meat items increased by 26.36 percent. It would take \$126.36 in 2005 to buy what \$100 did in 2003. Or, alternatively, it would require \$1.26 in 2005 to buy what \$1.00 did in 2003.

Notice that *the denominator is the same* in all three years: the Laspeyres index always uses quantities from the base period.

The **Paasche index,** on the other hand, uses as weights the quantities sold in each of the various reference years. This has the advantage of basing the index on current consumer behavior patterns. As consumers change their buying habits, these changes in consumer tastes are reflected by the index. Commodities that no longer attract consumers' interest, such as buggy whips and top hats, do not receive as much consideration. However, using of different quantity measures makes it impossible to attribute any differences in the index to changes in prices alone.

Thus, *Paasche Index uses quantities sold in the reference period as the weight factor*.

Its calculation is a bit more involved than the Laspeyres:

$$P = \frac{\sum(P_R \times Q_B)}{\sum(P_B \times Q_B)} \times 100. \qquad (9.4)$$

The quantities for the reference years appear in both the numerator and denominator. Table 9.5 provides the computation necessary for the Paasche, using the Nippand Tuck data with 2003 as the base.

Table 9.5 – Paache Index for Nipp and Yuck (2003 = 100)

Item	P	Q	P	Q	P	Q
Beef	3.00	250	3.30	320	4.50	350
Pork	2.00	150	2.20	200	2.10	225
Veal	4.00	80	4.50	90	3.64	70
		$P_{03} \times Q_{03}$	$P_{04} \times Q_{04}$	$P_{05} \times Q_{05}$	$P_{03} \times Q_{04}$	$P_{03} \times Q_{05}$
		750	1056	1575	960	1050
		300	440	472.5	400	450
		320	405	254.8	360	280
		1370	1901	2302.3	1720	1780

We must first multiply prices and quantities for all three years to get $P_R \times Q_R$, which is used in the numerator. We also need the value for price in the base year, 2003, times the quantity for each reference year to get $P_B \times Q_R$, which is used in the denominator. The Paasche index for 2003 is

$$P_{2003} = \frac{\sum(P_{03} \times Q_{03})}{\sum(P_{03} \times Q_{03})} \times 100 = \frac{1370}{1370}(100) = 100.$$

For 2004, it is

$$P_{2004} = \frac{\sum(P_{04} \times Q_{04})}{\sum(P_{03} \times Q_{04})} \times 100 = \frac{1901}{1720}(100) = 110.5.$$

For 2005, it is

$$P_{2005} = \frac{\sum(P_{05} \times Q_{05})}{\sum(P_{03} \times Q_{05})} \times 100 = \frac{2302.3}{1780}(100) = 129.3.$$

The usual interpretation applies.

The Laspeyres index requires quantity data for only one year and is easier to compute. Therefore, it is used more frequently than the Paasche. Since the base period quantities are always used, more meaningful comparisons over time are permitted.

However, the Laspeyres tends to overweigh goods whose prices increase. This occurs because the increase in price will decrease quantities sold, but the lower quantity will not be reflected by the Laspeyres index because it uses quantities from the base year. The Paasche, on the other hand, tends to overweigh goods whose prices go down. In an effort to offset these shortcomings. **Fisher's ideal index** is sometimes suggested. This index combines the Laspeyres and the Paasche by finding the square root of their product:

$$F = \sqrt{L \times P}.$$

The interpretation of the Fisher index is subject to some dispute. For this reason, it is not widely used.

9.5 Average of Relatives Method

Another type of composite index is the **average of relatives index.** As the name suggests, this index finds the average of several relative indexes. Formula (9.5) illustrates.

$$AR = \frac{\sum\left[\frac{P_R}{P_B}(100)\right]}{N}. \qquad (9.5)$$

The price of each good in a given reference period is divided by its price in the base period and multiplied by 100. The results are summed and averaged over the number of goods ***N***

Table 9.6 – Relative Advantages and Disadvantages of Laspeyers and Paasche Indexes

Index	Advantages	Disadvantages
Laspeyers	Requires quantity data for only one time period. Thus: (1) data are obtained easier, and (2) a more meaningful comparison over time can be made since any changes can be attributed to price movements.	Overweighs goods whose prices increase. Does not reflect changes in buying patterns over time.
Paasche	Reflects changes in buying habits since it uses quantity data for each reference period.	Requires quantity data for each year; these data are often difficult to obtain. Since different quantities are used, it is impossible to attribute differences in the index to price changes alone. Overweighs goods whose prices decrease.

Example 9.3 Consider again the data for Nipp and Tuck. With 2003 as the base period, the average of relatives index for 2004 is found by dividing the price of each good in the reference year, 2004, by its price in the base year, and multiplying by 100. This is done for all three goods. The results are added and divided by 3 to get the average Thus, the average of relatives index for 2004 is

For beef: $\frac{P_{04}}{P_{03}}(100) = \frac{3.30}{3.00}(100) = 110.$

For pork: $\frac{P_{04}}{P_{03}}(100) = \frac{2.20}{2.00}(100) = 110.$

For veal: $\frac{P_{04}}{P_{03}}(100) = \frac{4.50}{4.00}(100) = 112.5.$

Then

$$AR_{04} = \frac{110 + 110 + 112.5}{3} = 110.83.$$

Table 9.7 shows the full set of computations.

Table 9.7 – The Average of Relatives Index for Nipp and Ruck (2003=100)

Item	2003	2004	2005
Beef	$\frac{3.00}{3.00}\times 100 = 100$	$\frac{3.30}{3.00}\times 100 = 110$	$\frac{4.50}{3.00}\times 100 = 150$
Pork	$\frac{2.00}{2.00}\times 100 = 100$	$\frac{2.20}{2.00}\times 100 = 110$	$\frac{2.10}{2.00}\times 100 = 105$
Veal	$\frac{4.00}{4.00}\times 100 = 100$	$\frac{4.50}{4.00}\times 100 = 112.5$	$\frac{3.64}{4.00}\times 100 = 91$
	/ 300	/ 332.5	/ 346
AR	100	110.8	115.3

On the average, prices rose by 10.8 percent from 2003 to 2004 and by 15.3 percent from 2003 to 2005. What $100 would buy in 2003 would cost $115.30 in 2005.

An obvious drawback of this method is its failure to account for weights. Although in 2003, consumers bought more beef than either of the other two products, the price of beef is given the same importance as the prices of pork and veal in calculating the index.

This deficiency is corrected by calculating a **weighted average of relatives index,** which accounts for different quantities. This is precisely what the Laspeyres and Paasche indexes do. That is, *the Laspeyres and Paasche indexes are weighted average of relatives indexes*.

9.6 Selection of the Base Period

Choosing an appropriate base period is critical. You must use a time period that is "representative" of the prevailing economic life-style. However, determining what is representative is quite subjective and often difficult. *Time periods containing wars or major depressions should obviously be avoided*.

It is often necessary to change the base period to modernize the index and better reflect current trends and conditions. To shift the base to a different time

period, divide the existing index numbers by the index number in the new base period and multiply by 100. Table 9.8 shifts the base period from 1983 to 1985. Each of the index numbers based on 1983 are divided by 118, the index number of 1985. This produces the index numbers with 1985 as the base period.

Table 9.8 – Shifting the Base from 1983 to 1985

Year	Index (1983 = 100)	Index (1985 =100)
1981	89	(89/118)(100) = 75
1982	95	(95/118)(100) = 81
1983	100	(100/118)(100) = 85
1984	118	(110/118)(100) = 93
1985	121	(118/118)(100) = 100
1986	125	(121/118)(100) = 103
1987	131	(125/118)(100) = 106
1988	132	(131/118)(100) = 111
1989	138	(132/118)(100) = 112

If two indexes with different base periods are to be compared, it is advisable to shift the base period of one index to that of the other index. Comparisons are thereby more meaningful. Furthermore, it frequently becomes necessary to combine two indexes with different base periods. This is particularly true if, in conducting a study, we find that our time frame includes the period in which the series was rebased. We must then splice the two series together.

Notice from Table 9.9 that the index for producer prices is given for years 1984 to 1987 with 1975 = 100, and for 1987 to 1990 with the base of 1979 = 100.

Table 9.9 – Splicing Indexes for Producer Prices

Year	Index (1975 = 100)	Index (1979 =100)	Spliced Index (1979 =100)
1984	142	-	98.19
1985	147	-	101.65
1986	153	-	105.80
1987	162	112	112.00
1988	-	114	114.00
1989	-	121	121.00
1990	-	130	130.00

In order to splice, you must have at least one time period (1987 in this case) in which the index number is given for both series. Start at 1986 and work backwards (1985, 1984, etc.). Find the spliced value for 1986 by dividing the 1986 value by the 1987 value in the 1975 series, and multiplying by the 1987 value in the 1979 series.

That is,

$$\frac{153}{162}(112) = 105.8.$$

The 105.8 is the first value we find for the spliced series. The spliced value for 1985 is found similarly to be

$$\frac{147}{153}(105.8) = 101.65.$$

The value for 1984 is

$$\frac{142}{147}(101.65) = 98.19.$$

9.7 Specific Indexes

Numerous governmental agencies as well as the Federal Reserve System (which is not part of the federal government) and private businesses compile different indexes for a variety of purposes. The use for a specific index depends on who is compiling it and what factors go into its formulation. Perhaps the best-known index series is the **consumer price index**.

1. **Consumer Price Index**

The consumer price index (CPI) is reported monthly by the Bureau of Labor Statistics (BLS) of the U.S. Department of Labor (by State Committee of Statistics in Ukraine). It was first reported in 1914 as a means to determine if the wages of industrial workers were keeping pace with the inflation pressures brought on by World War 1. Prior to 1978, there was only one CPI. This traditional measure reflected changes in prices of a fixed market basket of about 400 goods and services commonly purchased by "typical" urban and clerical workers. It encompassed about 40 percent of the nation's total population.

In January 1978, the BLS began reporting a more comprehensive index, the *consumer price index for all urban consumers*. It is called CPI-U, while the older index is CPI-W. The newer CPI-U covers about 80 percent of the population and

includes around 3,000 consumer products ranging from basic necessities, such as food, clothing, and housing, to allowances for educational and entertainment expenses. In 1988 both CPI series were rebased from 1967 to 1982-1984.

Both the CPI-W and the CPI-U employ a weighting system for the types of goods and services purchased by consumers. Food, for example, is assigned a weight, or measure of relative importance, of about 18, while housing is given a weight of about 43. Medical care and entertainment each receive a weight of 5. The total weights for all commodities sum to 100. The weights on these products are adjusted about every 10 years. In this manner, the CPI-W and the CPI-U are similar to the Laspeyres index. Technically, the CPI differs slightly from a true Laspeyres because the weighting system used by the CPI is not revised at the same time that the index is rebased. The CPI is therefore sometimes referred to as a *fixed-weight aggregate price index.*

In Ukraine the CPI is calculated since 1991. The index is updated monthly, usually in the beginning of the month following the subject one and 100 corresponds to the beginning of the reported month.

The CPI is highly useful in gauging inflation, measuring "real" changes in monetary values by removing the impact of price changes, and, to a limited extent, serving as a cost-of-living index. It is even instrumental in determining raises in Social Security benefits and negotiated wage settlements in labor contracts. Its many uses will be more fully examined in the next section.

2. **Other Indexes**

The **producer price index** (formerly, the wholesale or industrial price index) is also published monthly by the BLS. It measures changes in the prices of goods in primary markets for raw materials used in manufacturing. It, too, is similar to the Laspeyres index and covers almost 3,000 producer goods. It is calculated differently for Ukrainian market.

The industrial production index is reported by the Federal Reserve System. It is not a monetary measurement, but tracks changes in the volume of industrial output in the nation. The base period is currently 1977.

There are numerous stock market indexes. Perhaps the most well-known is the **Dow Jones industrial average**. This index covers 30 selected industrial stocks to represent the almost 1,800 stocs traded in the New York Stock Exchange.

Standard & Poor's composite index of 500 industrial stocks is also highly watched.

9.8 Uses for the CPI

Movements in the CPI have a major impact on many business conditions and economic considerations. As noted, the CPI is often viewed as a measure of inflation in the economy. Annual rates of inflation are measured by the percentage change in the CPI from one year to the next. The inflation rate from year to year is

$$\frac{CPI_t - CPI_{t-1}}{CPI_{t-1}} \times 100,$$

where CPI_t, is the *CPI* in time period *t*, and CPI_{t-1} is the *CPI* in the previous time period. Table 9.10 shows the CPI for 1986 to 1989 using 1982-1984 as the base.

Table 9.10 - *CPI* and Inflation Rate for Selected Years

Year	CPI	Inflation Rate (%)
1986	109.6	-
1987	113.6	3.6
1988	118.3	4.1
1989	124.3	5.1
1990	127.2	2.3

The inflation rate for 1987, for example, is

$$\frac{113.6 - 109.6}{109.6}(100) = 3.6\%.$$

Changes in the CPI are also often taken as a *measure of the cost of living*. It is argued, however, that such a practice is questionable. The CPI does not reflect certain costs or expenditures such as taxes, nor does it account for changes in the quality of goods available. Further, the CPI fails to measure other valued items in our economic structure, such as increased leisure time by the average worker or improvements in the variety of commodities from which consumers can choose. Nevertheless, the CPI is often cited in the popular press as a measure of the cost of living.

The CPI is often the basis for adjustments in wage rates, Social Security payments, and even rental and lease agreements. Many labor contracts contain

cost-of-living adjustments (COLAs) which stipulate that an increase in the CPI of an agreed- upon amount will automatically trigger a rise in the workers' wage levels.

The CPI can also be used to **deflate** a time series. Deflating a series removes the effect of price changes and expresses the series in *constant* dollars. Economists often distinguish between nominal (or current) dollars and real (or constant) dollars. If a time series, such as your annual income over several years, is expressed in terms of 1982 dollars, that income is said to be real income.

Example 9.3. Assume your money (nominal) income was as shown in Table 9.11.

Table 9.11 – Money and Real Incomes for Selected Years

Year	Money Income, $	CPI (1982-84 = 100)	Real Income, $	Purchasing Power of a Dollar
1986	42,110	109.6	38,421	0.91
1987	46,000	113.6	40,493	0.88
1988	49,800	118.3	42,096	0.85
1989	53,500	124.3	43,041	0.80

In 1986, you actually earned $42,110. It would seem that you are doing quite well financially. Your income increased from $42,110 to $53,500 over that time period. However, prices have been going up also. To obtain a measure of how much your income has really increased, in real terms, you must deflate your income stream. This is done by dividing your money income by the CPI and multiplying by 100. The result is your real income expressed in constant (real) dollars of a given base year.

Thus, *real income is the purchasing power of your money income*.

$$\text{Real income} = \frac{\text{Money income}}{CPI} \times 100.$$

You earned $42,110 in 1986, but it was worth only $38,421 in 1982-84 prices. That is, keeping prices constant at the 1982-84 level, you are earning an equivalent of only $38,421. Your constant (real) income based on 1982-84 price levels is $38,421. The difference between $42,110 and $38,421 was consumed by rising prices from 1982-84 to 1986. Your money income rose by $53,500 — $42,110 = $11,390, but your real income went up by only $43,041 — $38,421 =

$4,620. If prices had gone up faster than your money income, your real income would have actually decreased.

The purchasing power of your dollar is found to be 100/CPI. For 1986 we have 100/109.6 = 0.91. This means that $1.00 in 1986 would buy what $0.91 would purchase in 1982-84.

Economists commonly deflate gross national product (GNP) to obtain a measurement of the increase in our nation's real output. Gross national product is the monetary value of all final goods and services produced in our economy. By deflating GNP over time, economists eliminate any increase due to price inflation, and arrive at a measure of the actual increase in the production of goods and services available for consumption. Table 9.12 illustrates that real GNP is found by dividing nominal (or current) GNP by the CPI and multiplying by 100.

Table 9.12 – Nominal and Real GNP (in $ billions)

Year	Nominal GNP	CPI	Real GNP
1986	4,140.3	109.6	3,777.6
1987	4 526 7	113.6	3,984.8
1988	4,864.3	118.3	4,111.8
1989	5,116.8	124.3	4,116.5

Thus, *Real GNP measures the value of our nation's output in constant dollars in some base period*. It omits any fluctuation due to changing prices.

$$\text{Real GNP}=\frac{National GNP}{CPI}\times 100.$$

Example 9.4 describes how indexes can have international implications.

According to *Business Week***,** Eastman Kodak Company was experiencing a problem maintaining accurate records on the amount of exports they sold to European and Eastern markets. Persistent fluctuations in prices and exchange rates made comparisons over time almost meaningless. It was impossible to determine if movements in export levels were due to a change in the actual volume of business or simply a result of instability in the financial world. The company's inability to gain a full appreciation of its international market made planning and decision making quite tenuous. The company finally decided to index export levels, using

prices and exchange rates in a selected base period. In this manner, it was possible for Kodak to get a more realistic picture of their true market position.

Chapter Checklist

After studying this chapter, as a test of your knowledge of the essential points, can you

1. Cite uses for index numbers?
2. Calculate and interpret a simple price index?
3. Calculate and interpret a composite price index?
4. Explain the principle of a weighted price index?
5. Calculate and interpret the Laspeyres and Paasche indexes?
6. Explain the advantages and disadvantages of the Laspeyres and Paasche indexes?
7. Calculate an index based on the average of relatives method, and explain its nature?
8. Identify the issues important in selecting a base period?
9. Shift the base period of an index series?
10. Splice index numbers?
11. Discuss the nature and application of the specific indexes presented in the chapter?
12. Show how the consumer price index can be used to distinguish between money income and real income?

10 TEST YOUR KNOWLEDGE

10.1 Graphical Displays of Data.

1. The section of statistics which involves the collection, organization, summarizing, and presentation of data relating to some population or sample is

(a) inferential statistics.

(b) descriptive statistics.

(c) an example of a frequency distribution.

(d) the study of statistics.

2. A subset of the population selected to help make inferences on a population is called

(a) a population.

(b) inferential statistics.

(c) a census.

(d) a sample.

3. A set of all possible data values for a subject under consideration is called

(a) descriptive statistics.

(b) a sample.

(c) a population.

(d) statistics.

4. The number of occurrences of a data value is called

(a) the class limits.

(b) the frequency.

(c) the cumulative frequency.

(d) the relative frequency.

5. A large collection of data may be condensed by constructing

(a) classes.

(b) a frequency polygon.

(c) class limits.

(d) a frequency distribution.

6. When constructing a frequency distribution for a small data set, it is wise to use

(a) 5 to 20 classes.

(b) 5 to 15 classes.

(c) 5 to 10 classes.

(d) less than 10 classes.

7. When constructing a frequency distribution for a large data set, it is wise to use

(a) 5 to 20 classes.

(b) 5 to 15 classes.

(c) 5 to 10 classes.

(d) less than 10 classes.

8. When straight-line segments are connected through the midpoints at the top of the rectangles of a histogram with the two ends tied down to the horizontal axis, the resulting graph is called

(a) a bar chart.

(b) a pie chart.

(c) a frequency polygon.

(d) a frequency distribution.

10.2 Numerical Measures of Central Tendency for Ungrouped Data

1. A student has seven statistics books open in front of him. The page numbers are as follows: 231, 423, 521, 139, 347, 400, and 345. The median for this set of numbers is

(a) 139.

(b) 347.

(c) 346.

(d) 373.5.

2. A cyclist recorded the number of miles per day that she cycled for 5 days. The recordings were as follows: 13, 10, 12, 10, and 11. The mean number of miles she cycled per day is

(a) 13.

(b) 11.

(c) 10.

(d) 11.2.

3. An instructor recorded the following quiz scores (out of a possible 10 points) for the 12 students present: 7, 4, 4, 7, 2, 9, 10, 6, 7, 3, 8, 5. The mode for this set of scores is

(a) 9.5.

(b) 7.

(c) 6.

(d) 3.

4. It is stated that more students are purchasing graphing calculators than any other type of calculator. Which measure is being used here?

(a) Mean

(b) Median

(c) Mode

(d) None of the above

5. Which of the following is not a measure of central tendency?

(a) Mode

(b) Variability

(c) Median

(d) Mean

Use the following frequency distribution for **Problems 6 to 8.**

X	**Frequency**
20	2
29	4
30	4
39	3
44	2

6. The mean of the distribution is

(a) 32.4.

(b) 30.

(c) 39.

(d) 32.07.

7. The median of the distribution is

(a) 4.

(b) 30.

(c) 29.5.

(d) 34.5.

8. The mode of the distribution is

(a) 29.

(b) 30.

(c) 29 and 30.

(d) none of the above.

9. Given the following data set: 12, 32, 45, 14, 24, and 31. The total deviation from the mean for the data values is

(a) 0.

(b) 26.3333.

(c) 29.5.

(d) 12.

10. The most frequently occurring value in a data set is called the

(a) spread.

(b) mode.

(c) skewness.

(d) maximum value.

11. A single numerical value used to describe a characteristic of a sample data set, such as the sample median, is referred to as a

(a) sample parameter.

(b) sample median.

(c) population parameter.

(d) sample statistic.

12. Which of the following is true for a positively skewed (right-skewed) distribution?

(a) Mode = Median = Mean

(b) Mean < Median < Mode

(c) Mode < Median < Mean

(d) Median < Mode < Mean

13. Which of the following would be affected the most if there is an extremely large value in the data set?

(a) The mode

(b) The median

(c) The frequency

(d) The mean

14. If the number of values in a data set is even, and the numbers are ordered, then

(a) the median cannot be found.

(b) the median is the average of the two middle numbers.

(c) the median, mode, and mean are equal.

(d) none of the above answers are correct.

10.3 Numerical Measures of Variability for Ungrouped Data

1. A sample of 10 students was asked by the instructor to record the number of hours each spent studying for a given exam from the time the exam was announced in class. The following data values were the recorded numbers of hours: 12, 15, 8, 9, 14, 8, 17, 14, 8, and 15.

The variance for the number of hours spent studying for this sample is

(a) 10.0000.

(b) 9.0000.

(c) 3.4641.

(d) approximately 12.

2. The numbers of minutes spent in the computer lab by 20 students working on a project are given below:

Numbers of Minutes

30 | 0, 2, 5, 5, 6, 6, 6, 8

40 | 0, 2, 2, 5, 7, 9

50 | 0,1, 3, 5

60 | 1, 3

The range for this data set is

(a) 300.

(b) 303.

(c) 603.

(d) 600.

3. The numbers of minutes spent in the computer lab by 20 students working on a project are given below:

Numbers of Minutes

30 | 0, 2, 5, 5, 6, 6, 6, 8

40 | 0, 2, 2, 5, 7, 9

50 | 0, 1, 3, 5

60 | 1, 3

The standard deviation for this data set is

(a) 101.6.

(b) 403.8.

(c) 306.0.

(d) 500.5.

4. The price increases on 5 stocks were $7, $1, $8, $4, and $5. The standard deviation for these price increases is

(a) 2.3.

(b) 2.7.

(c) 3.2.

(d) 4.1.

5. Which of the following is not affected by an extreme value in a data set?

The mean absolute deviation

The median

The range

The standard deviation

6. Given the following set of numbers 15, 20, 40, 25, 35. What is the variance?

(a) 9.27

(b) 86.0

(c) 10.37

(d) 107.5

7. Which of the following is the crudest measure of dispersion?

(a) The mean absolute deviation

(b) The variance

(c) The mode

(d) The range

8. Which of the following is not a measure of central tendency?

(a) Mean

(b) Median

(c) Q3

(d) Mode

9. Given the following data set: 12, 32, 45, 14, 24, and 31. The total deviation from the mean for the data values is

(a) 0.

(b) 26.3333.

(c) 29.5.

(d) 12.

10. Given that a sample is approximately bell-shaped with a mean of 60 and a standard deviation of 3, the approximate percentage of data values that is expected to fall between 54 and 66 is

(a) 75 percent.

(b) 95 percent.

(c) 68 percent.

(d) 99.7 percent.

10.4 Simple Regression

1. In simple linear regression analysis with *X* representing the independent variable and *Y* representing the dependent variable, if the *Y* intercept is negative, then

(a) the correlation between *X* and *Y* is negative.

(b) the correlation between *X* and *Y* is positive.

(c) the correlation between *X* and *Y* could be either negative, positive, or zero.

(d) the value of the predicted *Y* value is always negative.

2. In regression analysis, the input variable that is used to get a predicted value is

(a) the dependent variable.

(b) the independent variable.

(c) the least-squares variable.

(d) the random variable.

3. In the simple linear regression model with *X* representing the independent variable and *Y* representing the dependent variable, correlation analysis is used to

(a) find the least-squares regression line.

(b) find the slope of the regression line.

(c) measure the strength of the linear relationship between *x* and *y*.

(d) draw a scatter plot.

4. If the correlation coefficient is zero, the slope of a linear regression line will be

(a) positive.

(b) negative.

(c) positive or negative.

(d) none of the above.

5. In the simple linear regression model, if there is a very strong correlation between the independent and dependent variables, then the correlation coefficient should be

(a) close to –1.

(b) close to +l.

(c) close to either –1 or +l.

(d) close to zero.

6. For the simple linear regression model, if all the points on a scatter plot lie on a straight line with correlation coefficient r = -1, then the slope of the regression line is

(a) –1.

(b) +l.

(c) positive.

(d) negative.

7. The least-squares equation for the line of best fit

(a) minimizes the error sum of squares.

(b) maximizes the error sum of squares.

(c) does not change the error sum of squares.

(d) does none of the above.

8. If through some analysis, one can conclude that the slope of the line of best fit is not equal to zero, then the simple linear regression model indicates that there is

(a) a positive relationship between the independent and dependent variables.

(b) a negative relationship between the independent and dependent variables.

(c) a positive or negative relationship between the independent and dependent variables.

(d) no relationship between the independent and dependent variables.

9. Which of the following is not a possible value of the correlation coefficient?

(a) +1

(b) -1

(c) 0.011

(d) 1.11

10. A negative correlation coefficient between the dependent variable *Y* and the independent variable *X* indicates that

(a) large values of *X* are associated with small values of *Y*.

(b) large values of *X* are associated with large values of *Y*.

(c) small values of *X* are associated with small values of *Y*.

(d) none of the above answers are correct.

11. For the simple linear regression model, if the unit for the dependent variable is square feet, then the unit for the independent variable

(a) must be square feet.

(b) can be some unit of square measurement.

(c) can be any unit.

(d) cannot be a unit of square measurement.

12. In simple linear regression analysis, there

(a) is only one independent variable in the model.

(b) could be several linear independent variables in the model.

(c) is only one nonlinear term in the model.

(d) is at least one nonlinear term in the model.

10.5 Sampling Distributions and the Central Limit Theorem

1. As the sample size increases,

(a) the population mean decreases.

(b) the population standard deviation decreases.

(c) the standard deviation for the distribution of the sample means increases.

(d) the standard deviation for the distribution of the sample means decreases.

2. The concept of sampling distribution applies to

(a) only discrete probability distributions from which random samples are obtained.

(b) only continuous probability distributions from which random samples are obtained.

(c) only the normal probability distribution.

(d) any probability distribution from which random samples are obtained.

3. When we consider sampling distributions, if the sampling population is normally distributed, then the distribution of the sample means

(a) will be exactly normally distributed.

(b) will be approximately normally distributed.

(c) will have a discrete distribution.

(d) will be none of the above.

4. The expected value of the sampling distribution of the sample mean is equal to

(a) the standard deviation of the sampling population.

(b) the mean of the sampling population.

(c) the mean of the sample.

(d) the population size.

5. The sample statistic $\bar{x}$ is the point estimate of

(a) the population standard deviation σ.

(b) the population median.

(c) the population mean μ.

(d) the population mode.

6. If repeated random samples of size 40 are taken from an infinite population, the distribution of sample means

(a) will always be normal because we do not know the distribution of the population.

(b) will always be normal because the sample mean is always normal.

(c) will always be normal because the population is infinite.

(d) will be approximately normal because of the Central Limit Theorem.

7. The mean TOEFL score of international students at a certain university is normally distributed with a mean of 490 and a standard deviation of 80. Suppose groups of 30 students are studied. The mean and the standard deviation for the distribution of sample means will respectively be

(a) 490, 8/3.

(b) 16.33, 80.

(c) 490, 14.61.

(d) 490, 213.33.

8. A certain brand of light bulb has a mean lifetime of 1500 hours with a standard deviation of 100 hours. If the bulbs are sold in boxes of 25, the parameters of the distribution of sample means are

(a) 1,500, 100.

(b) 1,500, 4.

(c) 1,500, 2.

(d) 1,500, 20.

9. Samples of size 49 are drawn from a population with a mean of 36 and a standard deviation of 15. Then $P(\bar{x} < 33)$ is

(a) 0.5808.

(b) 0.4192.

(c) 0.1608.

(d) 0.0808.

10. A tire manufacturer claims that its tires will last an average of 40,000 miles with a standard deviation of 3,000 miles. Forty-nine tires were placed on test and the average failure miles was recorded. The probability that the average failure miles was less than 39,500 is

(a) 0.3790.

(b) 0.8790.

(c) 0.1210.

(d) 0.6210.

10.6 Confidence Intervals Large Samples

1. If we are constructing a 98 percent confidence interval for the population mean, the confidence level will be

(a) 2 percent.

(b) 2.29.

(c) 98 percent.

(d) 2.39.

2. The *z* value corresponding to a 97 percent confidence interval is

(a) 1.88.

(b) 2.17

(c) 1.96.

(d) 3 percent.

3. As the sample size increases, the confidence interval for the population mean will

(a) decrease.

(b) increase.

(c) stay the same.

(d) decrease and then increase.

4. If we change the confidence level from 98 % to 95 % when constructing a confidence interval for the population mean, we can expect the size of the interval to

(a) increase.

(b) decrease.

(c) stay the same.

(d) do none of the above.

5. Generally, lower confidence levels will yield

(a) smaller standard deviations for the sampling distribution

(b) larger margins of error.

(c) broader confidence intervals.

(d) narrower confidence intervals.

6. If the 98 percent confidence limits for the population mean μ are 73 and 80, which of the following could be the 95 percent confidence limits?

(a) 73 and 81

(b) 72 and 79

(c) 72 and 81

(d) 74 and 79

7. A 90 percent confidence interval for a population mean indicates that

(a) we are 90 percent confident that the interval will contain all possible sample means with the same sample size taken from the given population.

(b) we are 90 percent confident that the population mean will be the same as the sample mean used in constructing the interval.

(c) we are 90 percent confident that the population mean will fall within the interval.

(d) none of the above is true.

8. Interval estimates of a parameter provide information on

(a) how close an estimate of the parameter is to the parameter.

(b) what proportion of the estimates of the parameter are contained in the interval.

(c) exactly what values the parameter can assume.

(d) the *z* score.

9. Which of the following confidence intervals will be the widest?

(a) 90 percent

(b) 95 percent

(c) 80 percent

(d) 98 percent

10. The best point estimate for the population variance is

(a) a statistic.

(b) the sample standard deviation.

(c) the sample mean.

(d) the sample variance.

11. When determining the sample size in constructing confidence intervals for the population mean μ, for a fixed maximum error of estimate and level of confidence, the sample size will

(a) increase when the population standard deviation is decreased.

(b) increase when the population standard deviation is increased.

(c) decrease when the population standard deviation is increased.

(d) decrease and then increase when the population standard deviation is increased.

12. When computing the sample size to help construct confidence intervals for the population proportion, for a fixed margin of error of estimate and level of confidence, the sample size will be maximum when

(a) $p = 0.25$.

(b) $(1 - p) = 0.25$.

(c) $p(1 - p) = 0.5$.

(d) $p = 0.5$.

13. What value of the population proportion p will maximize $p(1 - p)$*?*

(a) 0.50

(b) 0.25

(c) 0.75

(d) 0.05

14. Suppose that a sample of size 100 is selected from a population with unknown variance. If this information is used in constructing a confidence interval for the population mean, which of the following statements is true?

(a) The sample must have a normal distribution.

(b) The population is assumed to have a normal distribution.

(c) Only 95 percent confidence intervals may be computed.

(d) The sample standard deviation cannot be used to estimate the population standard deviation because the sample size is large.

LITERATURE

1. Лугінін О. Є. Статистика : підручник / О. Є. Лугінін, С. В. Білоусова. – К.: Центр навчальної літератури, 2005. – 580 с.

2. Бек В. Л. Теорія статистики : навч. посіб. / В. Л. Бек. – К.: ЦУЛ, 2003. – 288 с.

3. Захожай В.Б. Статистика: підручник / В. Б. Захожай, І. І. Попов. – К. : МАУП, 2006. –536с.

4. Попов І. І. Теорія статистики. Практикум : навч. посіб. / І. І. Попов. – К. : КНТЕУ, 2006. –290 с.

5. Практикум по теории статистики : учеб. пособие / под ред. Р. А. Шмойловой. – М.: Финансы и статистика, 2002. –416с.

6. Сигел Э. Практическая бизнес-статистика : пер. с анг / Э. Сигел. – М. : Вильямс, 2002. –1021 с.

7. Статистика : структурно-логічні схеми та задачі : навч. посіб. – КНЕУ, 2007. – 304 с.

8. Галицька, Е. В. Фінансова статистика [Текст] : навч. посібник / Е. В. Галицька, Н. В. Ковтун ; дар. Е. М. Затуловська. – К. : Кондор, 2008. – 434 с.

9. Фінансово-банківська статистика [Текст] : навч. посібник / П. Г. Вашків [та ін.]. – К. : Либідь, 2007. – 512 с.

10. Лутчин, Наталія Павлівна. Статистика фінансів [Текст] : навчальний посібник / Н. П. Лутчин, А. К. Миронюк. – Львів : Новий світ–2000, 2011. – 324 с.

11. Webster Allen L. Applied Statistics for Business and Economics. Second Edition / Allen L. Webster. – The USA: Richard D.Irwin, Inc., 1995. – 1068 p.

12. Feller W. An Introduction to probability Theory and its Applications. Volume 1. Third Edition / W. Feller. – The USA: John Wiley&Sons, Inc. 1968. – 527 p.

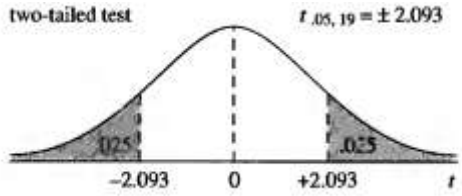
13. Fernandes M. Statistics for Business and Economics / M. Fernandes. Marcelo Fernandes & Ventus Publishing ApS. 2009. – 150 p.

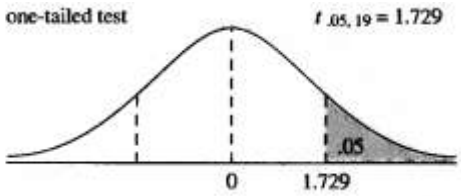
14. Jaisingh Lloyd R. Statistics for Utterly Confused / Lloyd R. Jaisingh. McGraw–Hill. 2000. – 337 p.

APPENDIX

DISTRIBUTION TABLES

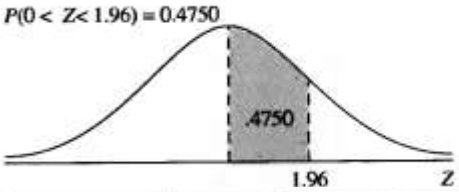
Table A – The *t* Distribution





	0.900	0.700	0.500	0.300	0.200	0.100	0.050	0.020	0.010	α value (two tailed-test)
	0.100	0.300	0.500	0.700	0.800	0.900	0.950	0.980	0.990	CL (two tailed-test)
	0.450	0.350	0.250	0.150	0.100	0.050	0.025	0.010	0.005	α value (one tailed-test)
	0.550	0.650	0.750	0.850	0.900	0.950	0.975	0.990	0.995	CL (one tailed-test)
d.f.	Values of *t*									
1	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657	
2	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925	
3	0.137	0.424	0.765	1.250	1.638	2.353	3.182	4.541	5.841	
4	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.747	4.604	
5	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032	
6	0.131	0.404	0.718	1.134	1.440	1.943	2.447	3.143	3.707	
7	0.130	0.402	0.711	1.119	1.415	1.895	2.365	2.998	3.499	
8	0.130	0.399	0.706	1.108	1.397	1.860	2.306	2.896	3.355	
9	0.129	0.398	0.703	1.100	1.383	1.833	2.262	2.821	3.250	
10	0.129	0.397	0.700	1.093	1.372	1.812	2.228	2.764	3.169	
11	0.129	0.396	0.697	1.088	1.363	1.796	2.201	2.718	3.106	
12	0.128	0.395	0.695	1.083	1.356	1.782	2.179	2.681	3.055	
13	0.128	0.394	0.694	1.079	1.350	1.771	2.160	2.650	3.012	
14	0.128	0.393	0.692	1.076	1.345	1.761	2.145	2.624	2.977	
15	0.128	0.393	0.691	1.074	1.341	1.753	2.131	2.602	2.947	
16	0.128	0.392	0.690	1.071	1.337	1.746	2.120	2.583	2.921	
17	0.128	0.392	0.689	1.069	1.333	1.740	2.110	2.567	2.898	
18	0.127	0.392	0.688	1.067	1.330	1.734	2.101	2.552	2.878	
19	0.127	0.391	0.688	1.066	1.328	**1.729**	**2.093**	2.539	2.861	
20	0.127	0.391	0.687	1.064	1.325	1.725	2.086	2.528	2.845	
21	0.127	0.391	0.686	1.063	1.323	1.721	2.080	2.518	2.831	
22	0.127	0.390	0.686	1.061	1.321	1.717	2.074	2.508	2.819	
23	0.127	0.390	0.685	1.060	1.319	1.714	2.069	2.500	2.807	
24	0.127	0.390	0.685	1.059	1.318	1.711	2.064	2.492	2.797	
25	0.127	0.390	0.684	1.058	1.316	1.708	2.060	2.485	2.787	
26	0.127	0.390	0.684	1.058	1.315	1.706	2.056	2.479	2.779	
27	0.127	0.389	0.684	1.057	1.314	1.703	2.052	2.473	2.771	
28	0.127	0.389	0.683	1.056	1.313	1.701	2.048	2.467	2.763	
29	0.127	0.389	0.683	1.055	1.311	1.699	2.045	2.462	2.756	
30	0.127	0.389	0.683	1.055	1.310	1.697	2.042	2.457	2.750	
40	0.126	0.388	0.681	1.050	1.303	1.684	2.021	2.423	2.704	
60	0.126	0.387	0.679	1.045	1.296	1.671	2.000	2.390	2.660	
120	0.126	0.386	0.677	1.041	1.289	1.658	1.980	2.358	2.617	
∞	0.126	0.385	0.674	1.036	1.282	1.645	1.960	2.326	2.576	

Table B – Z-test (The 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	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	**0.4750**	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Навчальне видання

SHYRIAIEVA Natalia

STATISTICS. BASIC PRINCIPLES

Lecture notes on Statistics Course
for students of bachelor level in 6.030601 «Management»
and 6.030508 «Finance and credit»

ШИРЯЄВА Наталя

СТАТИСТИКА. ОСНОВНІ ПРИНЦИПИ

Текст лекцій з курсу «Статистика»
для студентів напрямів 6.030601 «Менеджмент»
та 6.030508 «Фінанси і кредит»

Роботу до видання рекомендував проф. В. А. Міщенко

В авторській редакції

План 2015 р. п. 124
Підписано до друку 20.03.15. Формат 60× 84 1/16 . Папір офсет.
Друк - ризографія. Гарнітура Таймс. Ум. друк. арк. ___.
Наклад 100 прим. Зам № . Ціна договірна.

Видавничий центр НТУ "ХПІ".
Свідоцтво про державну реєстрацію ДК № 9657 від 24.12.2009р.
61002, Харків 2, вул. Фрунзе, 21

Друкарня _____.