University of Baghdad
College of Engineering
Civil Engineering Department

ENGINEERING STATISTICS

FIRST YEAR COURSE

(FRESHMAN COURSE)

PREPARED BY

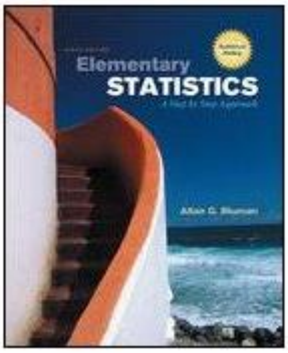
ASSISTANT PROF. DR. SALAH R. AL-ZAIDEE
E-Mail: Salah.R.Al.Zaidee@coeng.uobaghdad.edu.iq
CIVIL ENGINEERING DEPARTMENT
COLLEGE OF ENGINEERING
UNIVERSITY OF BAGHDAD

ASSISTANT PROF. DR. AHMED M. RAOOF
E-Mail: Ahmed.mahjoob@coeng.uobaghdad.edu.iq
CIVIL ENGINEERING DEPARTMENT
COLLEGE OF ENGINEERING
UNIVERSITY OF BAGHDAD

Syllabus

Text Book

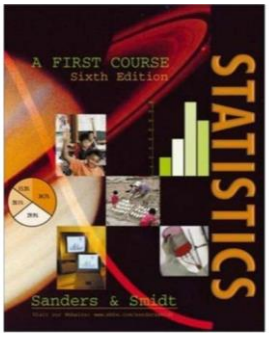
Elementary Statistics: A Step by Step Approach, by Allan G. Bluman, 6th edition.



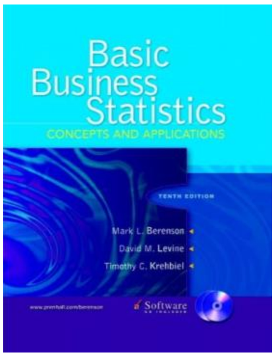
References

General References

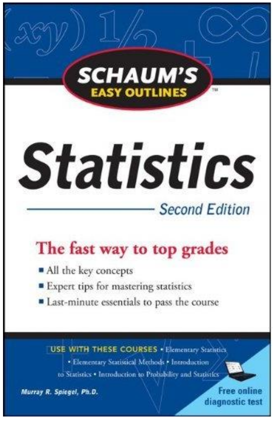
1. Statistics: A First Course by Donald Sanders and Robert Smidt, 6th Edition.



2. Basic Business Statistics: Concepts and Applications, by Mark L. Berenson, David M. Levine, and Timothy C. Krehbiel, 10th Edition.

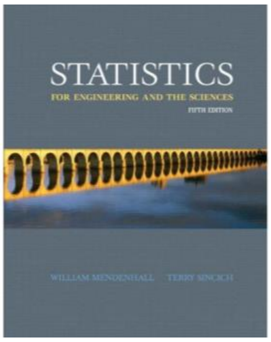


3. Schaum's Easy Outline of Statistics, by Murray Spiegel

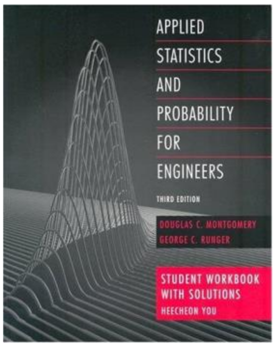

SCHAUM'S
EASY OUTLINES
Statistics
Second Edition
The fast way to top grades
All the key concepts
Expert tips for mastering statistics
Last-minute essentials to pass the course
USE WITH THESE COURSES • Elementary Statistics • Elementary Statistical Methods • Introduction to Statistics • Introduction to Probability and Statistics
Murray R. Spiegel, Ph.D.
Free online diagnostic test

ENGINEERING REFERENCES

1. Statistics for Engineering and the Sciences, by William Mendenhall and William Mendenhall, 5th Edition.

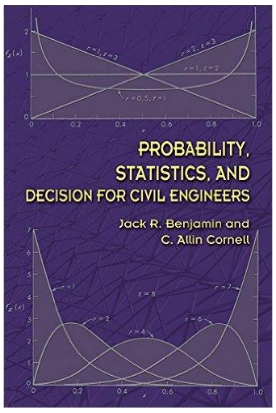

STATISTICS
FOR ENGINEERING AND THE SCIENCES
FIFTH EDITION
WILLIAM MENDENHALL TERRY SINCICH

2. Applied Statistics and Probability for Engineers, 3rd Edition, by Douglas C. Montgomery and George C. Runger.

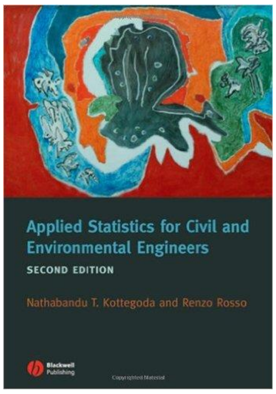


CIVIL ENGINEERING REFERENCE

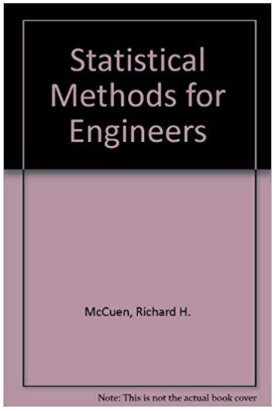
1. Probability, Statistics, and Decision for Civil Engineering, by Jack R. Benjamin and C. Allin Cornell.



2. Applied Statistics for Civil and Environmental Engineers, 2nd Edition, by N. T. Kottegoda, and R. Rosso


Applied Statistics for Civil and
Environmental Engineers
SECOND EDITION
Nathabandu T. Kottegoda and Renzo Rosso
Blackwell Publishing

3. Statistical Methods for Engineers, 1st Edition by Richard H. McCuen 1985.

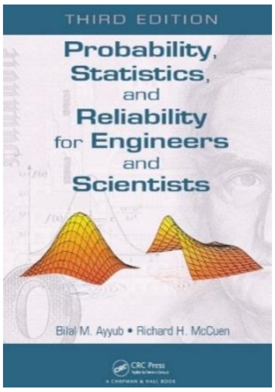


Statistical Methods for Engineers

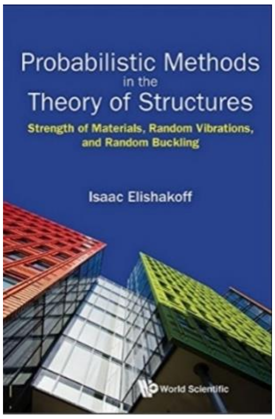
McCuen, Richard H.

Note: This is not the actual book cover

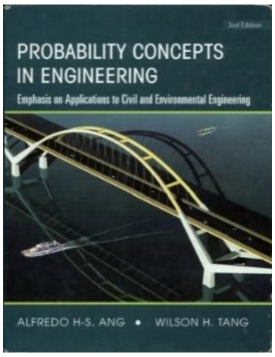
4. Probability, Statistics, and Reliability for Engineers and Scientists, 3rd Edition by Bilal M. Ayyub and Richard H. McCuen.


THIRD EDITION
Probability, Statistics, and Reliability for Engineers and Scientists
Bilal M. Ayyub • Richard H. McCuen
CRC Press

5. Probabilistic Methods in The Theory of Structures. Strength of Materials, Random Vibrations, and Random Buckling. By Isaac Elishakoff.


Probabilistic Methods in the Theory of Structures
Strength of Materials, Random Vibrations, and Random Buckling
Isaac Elishakoff
World Scientific

6. Probability Concepts in Engineering Emphasis on Applications in Civil & Environmental Engineering. By Alfredo H-S. Ang and Wilson H. Tang.


2nd Edition
PROBABILITY CONCEPTS IN ENGINEERING
Emphasis on Applications to Civil and Environmental Engineering
ALFREDO H-S. ANG • WILSON H. TANG

7. Probabilistic solutions in geotechnics. By Laszlo Rethati.

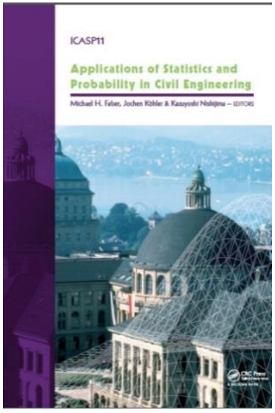
DEVELOPMENTS IN GEOTECHNICAL ENGINEERING 46

Probabilistic solutions in geotechnics

by
László Rétháti
Institute for Geodesy
and Geotechnics, Budapest

Elsevier
Amsterdam–Oxford–New York–Tokyo 1988

8. Applications of Statistics and Probability in Civil Engineering. Edited by Michael H. Faber,



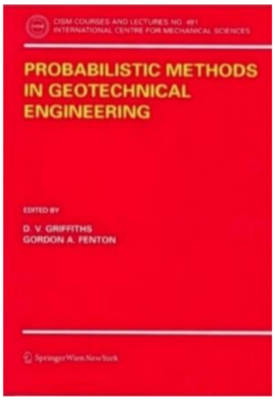
ICASP11

Applications of Statistics and Probability in Civil Engineering

Michael H. Faber, Jochen Köhler & Kazuyoshi Nishijima – EDITORS

CRC Press

9. Probabilistic Methods in Geotechnical Engineering. Edited by D. V. Griffiths.



CISM COURSES AND LECTURES NO. 491
INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES

PROBABILISTIC METHODS IN GEOTECHNICAL ENGINEERING

EDITED BY
D. V. GRIFFITHS
GORDON A. FENTON

SpringerWienNewYork

10. Probabilistic Methods in the Mechanics of Solids and Structures. Edited by S. Eggwertz and N. C. Lind

International Union of Theoretical and Applied Mechanics

Probabilistic Methods in the Mechanics of Solids and Structures

Symposium Stockholm, Sweden
June 19–21, 1984
To the Memory of Waloddi Weibull

Editors
S. Eggwertz and N. C. Lind

With 149 Figures

Springer-Verlag
Berlin Heidelberg New York Tokyo

OPEN COURSEWARE PROGRAM AND VIDEO COURSES

- In 2000, ***Massachusetts Institute of Technology***, ***MIT***, proposed the ***Open Courseware Program***. Within this program, about 1800 full videos in 33 academic disciplines have been offered on internet for free wide use.
- This program is a Non-Degree Program.
- Ninety percent of MIT staff have voluntarily shared their educational materials at the ***Open Courseware Program***.
- This program can be used to full any possible shortages in labs, and study materials in other counties.
- About 35,000,000 individuals from more than 220 countries and territories have adopted the program.
- Today, about 150 universities around the world share their courses and study materials for ***Open Courseware Program.***
- Regarding to statistics and probability theory, the following courses have been offered freely by ***MIT*** and ***University of California***, UC, ***at Berkeley*** on YouTube as apart from their open courseware programs.

Course Code	Course Title	No. of Lec.	Notes
Statistics 21	Introductory Probability and Statistics for Business. Fall 2009 Professor Philip Stark UC Berkeley Link for the first lecture is: https://youtu.be/mnbbRtFxWHA	25	• This course is complete in nature. It had been prepared to be adopted later in summer online courses. • The course is dedicated to business and economy. • Textbook adopted is an online book. https://www.stat.berkeley.edu/~stark/SticiGui/index.htm prepared by the instructor to emphasis some items that not covered adequately in his opinion. • Instructure accent is relatively clear to be understood. • Syllabus is generally similar to our course. • Course web site where students download study materials and upload their home works and assignments have been widely adopted with ***GIS*** technique. • The instructor encouraged his students to prepare a google group where they can post their discussion about home works.
MIT Course Number 6.041 / 6.431	Probabilistic Systems Analysis and Applied Probability. Fall 2010 Prof. John Tsitsiklis Link for the first lecture is: https://www.youtube.com/watch?v=j9WZyLZCBzs Course link of MIT web site: http://ocw.mit.edu/6-041F10	25	• The course aims to introduce the probability models, skills, and tools, by combining mathematics with conceptual understanding and intuition. • Textbook: Introduction to Probability. 2nd edition. Athena Scientific, 2008. ISBN: 9781886529236. By Bertsekas, Dimitri, and John Tsitsiklis.

- A student that has some listening problem to the aforementioned courses can used the icon

  on YouTube to show English subtitles. Some of these subtitles are auto generated by YouTube and may contain some errors. Questionable words are indicated with a gray color in the auto generated subtitle.

GRADING SYSTEM

- Thirty percent, 30%, of the student grade is determined based on short quizzes at end of each chapter. Usually, the student has about 8 to 10 quizzes until the end of academic year. Two or three quizzes with minimum grades will be dropped and the average will be determined accordingly.
- There is no makeup time for the quizzes. Any possible problem in a specific quiz would be implicitly accounted through the dropped quizzes.
- Up to 3 degrees may be added to some students based on their contribution and interactive in the lecture.
- Finally, the remaining 70 degrees are determined based on the final exam.

Chapter 1

The Nature

of

Probability and Statistics

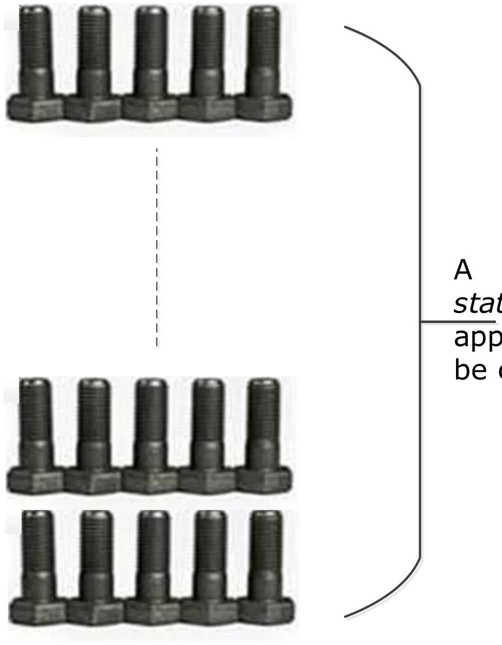
1.1 Overview

1.1.1 The Basic Idea

- The basic idea behind all the statistical methods of data analysis is to make inferences about a population by studying a relatively small sample chosen from it.
- Overview of statistics, probability, and basic pertained terminology can be illustrated using the following example.

1.1.2 Population

- Assume a batch of approximately 10,000 bolts to be tested for strength in tension. The whole batch is called ***population*** according to statistics terminology, Figure 1.1-1.



A batch, ***Population*** *in statistics terminology*, of approximately 10,000 bolts to be checked for strength.

Figure 1.1-1: Population of high strength bolts.

- It is almost impossible to test each member within the population for the following reasons:
  - The test may be destructive in nature.
  - Some items may be physically unobtainable, e.g. bolts that are not produced yet.
  - The test would be too costly.
  - The test would be time-consuming.
- Therefore, a statistical approach should be used. The current statistical approach is based on the following steps.

1.1.3 Step 1: Select a Sample

Unbiased sample to be selected based on basic principle of probability theory and sampling techniques. A brief introduction to the sampling techniques has been presented in section 1.6.

1.1.4 Step 2: Sampling Testing

- The selected sample is then tested by measuring the pertained property, tensile strength in this example. In general physical, chemical, geological, economical, psychological, and other aspects of the tests are out the scope of the statistics and they should be sought in the specialized branches.

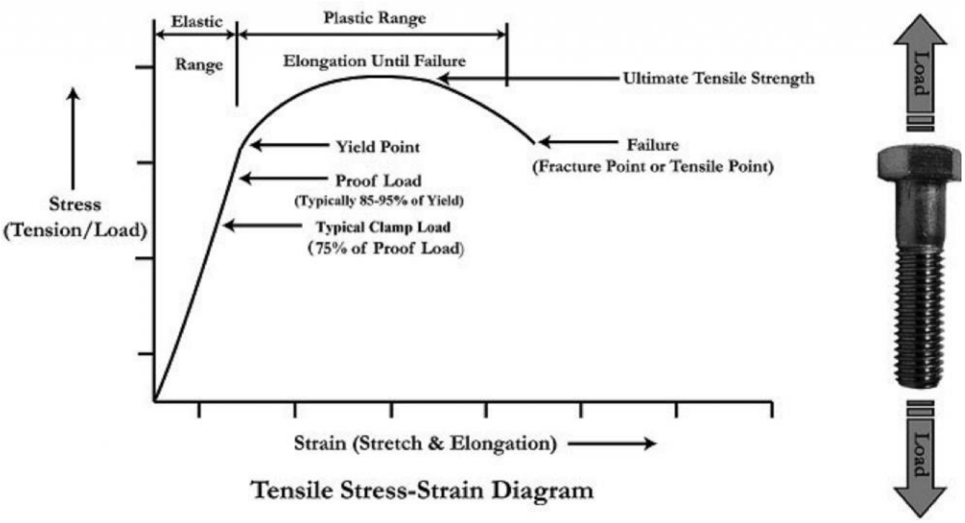

Elastic
Range
Plastic Range
Elongation Until Failure
Ultimate Tensile Strength
Yield Point
Failure
(Fracture Point or Tensile Point)
Proof Load
(Typically 85-95% of Yield)
Typical Clamp Load
(75% of Proof Load)
Stress
(Tension/Load)
Strain (Stretch & Elongation)
Tensile Stress-Strain Diagram
Load
Load

Figure 1.1-2: Bolt testing for tensile strength.

- The sample either noted from
  - An observational point of view:
    Where the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
  - An experimental point of view:
    Where the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.
- In this example, if we only intend to check bolts strength, the statistical process is classified as an observational, one otherwise it is classified as experimental one.

1.1.5 Step 3: Description and Summarization of Data

Observation or experimental data to be described and summarized based on tools and principles of ***Descriptive Statistics***. This can be in one of the following forms:

1.1.5.1 Graphically

This may be in terms of histograms as would be discussed in ***Chapter 02***, see Figure 1.1-3 below.

1.1.5.2 Numerically

Data may also be presented in a numerical form as will be discussed in, ***Chapter 03***:

$$Average\ Tensile\ Strength = 120\ kN$$

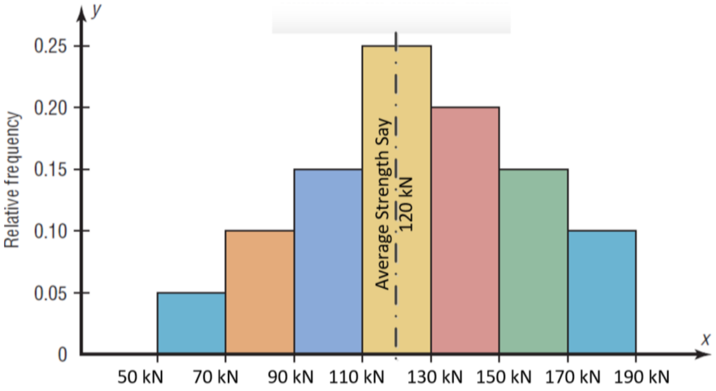


Figure 1.1-3: Data presentation in terms of histogram.

1.1.6 Step 4: Mathematical Framework

- The probability theory represents the mathematical framework to transfer from the descriptive statistical part to the inferential part.

- The probability theory is used also in selection of a random sample and in the design of the experiments. These aspects represent the stating points for any statistical analysis.

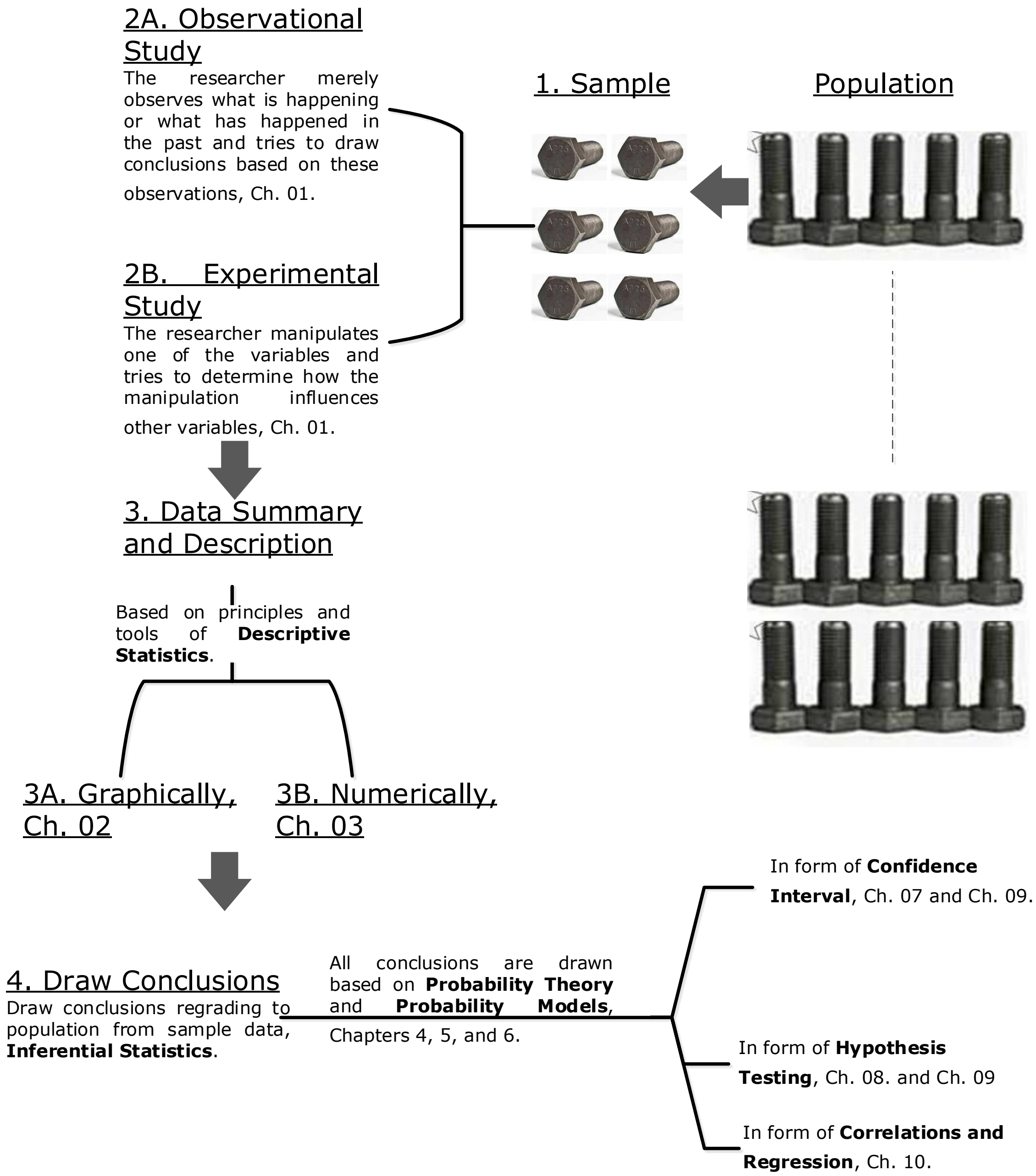
1.1.7 STEP 5: DRAW CONCLUSIONS REGARDING TO POPULATION FROM SAMPLE DATA

- Finally, conclusions regarding to population can be drawn from sample data, ***Inferential Statistics***.
- Single variable conclusions, e.g. variation of tensile strength, may be written in form of
  - ***Confidence interval***, Chapter 7

    $120\ kN - \epsilon \leq Avg.Tensile\ Strength\ for\ Bolts\ in\ the\ Whole\ Population \leq 120\ kN + \epsilon$

    In general, the variation term, ϵ, is inversely proportional to sample size. Then the larger sample size the smaller variation, ϵ, and the more accurate predation for tensile strength.
  - ***Hypothesis testing***. Chapter 8.

How to select a suitable probability density model to fit a data has been discussed in Chapter 9.



The researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations, Ch. 01.

The researcher manipulates one of the variables and tries to determine how the manipulation influences other variables, Ch. 01.

Based on principles and tools of **Descriptive Statistics**.

Draw conclusions regrading to population from sample data, **Inferential Statistics**.

All conclusions are drawn based on **Probability Theory** and **Probability Models**, Chapters 4, 5, and 6.

In form of **Confidence Interval**, Ch. 07 and Ch. 09.

In form of **Hypothesis Testing**, Ch. 08. and Ch. 09

In form of **Correlations and Regression**, Ch. 10.

Figure 1.1-4: An overview of statistical process.

- Bivariate and multivariate descriptions and conclusions, e.g. relation between bolt strength and a chemical agent, are usually presented in terms of ***correlation and regression***, Chapter 10.
- An overview of the statistical process with a cross reference to the corresponding chapters of syllabus has been summarized in Figure 1.1-4.

1.1.8 OTHER EXAMPLES WHERE SAMPLING IS NECESSARY

In civil engineering, there are many situations like the above illustration example where samples should be taken to make decision about the population. See Figures below:

1.1.8.1 Cement Batch

Figure 1.1-5: **Cement batch.**

1.1.8.2 Course and Fine Aggregates

Figure 1.1-6: **Course and fine aggregates.**

1.1.8.3 Reinforcing Bars and Steel Section

Figure 1.1-7: **Reinforcing bars.**

Figure 1.1-8: Steel beams.

1.2 STATISTICS

Statistics is the science that deals with conducting studies to:

- collect,
- organize,
- summarize,
- analyze,
- and draw conclusions from data[1].

<u>Linguistic Aspects:</u>

From linguistic point of view, the word ***statistics*** was developed from the word of ***state***. And it can be:

- Uncountable singular noun where the 's' is a part of the word. In this case it refers to the science of data gathering and analysis.

- Countable noun where 's' is for plural and it refers to measure(s) in the description and inferential processes.

1.3 VARIABLE

- A variable ***is a characteristic or attribute that can be assumed as different values***.
- Usually an object has many variables depending on which aspects we are interested in, for example
  - Human being may be considered for different aspects, variables, e.g. height, weight, race, color, gender, etc.
  - A Bolt, an object, may also has many variables to be considered, e.g. its strength, diameter, length, etc.

1.4 DESCRIPTIVE AND INFERENTIAL STATISTICS

Statistics can be divided into two main areas, depending on the use of the data. The two areas are:

- Descriptive statistics
- Inferential statistics

1.4.1 DESCRIPTIVE STATISTICS

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

1.4.2 EXAMPLES ON DESCRIPTIVE STATISTICS

- Surveying, and summarization of previous measurements for earthquake motions, see Figure 1.4-1 below.
- Surveying, and summarization of previous measurements for water levels in a dam upstream, see Figure 1.4-2 below.

1.4.3 INFERENTIAL STATISTICS

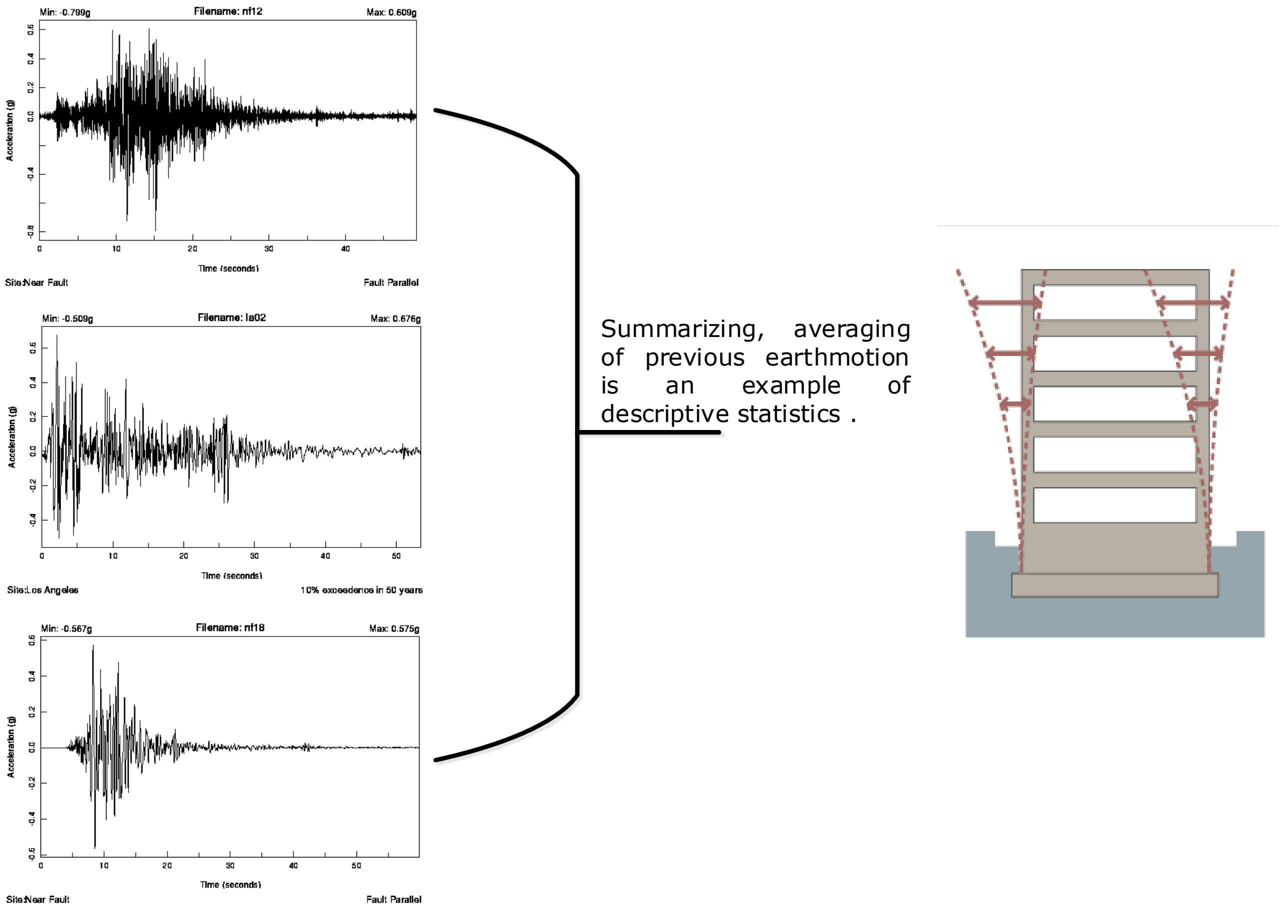
Inferential statistics consists of:

- Generalizing from samples to populations,
- Performing estimations and hypothesis tests,
- Determining relationships among variables and making predictions.

1.4.4 EXAMPLES OF INFERENTIAL STATISTICS

- Expecting an earthquake that may a building be subjected to during the design period, see Figure 1.4-1.
- Expecting maximum water lever acting on a dam upstream, Figure 1.4-2.

[1] "Data" is a plural noun; the singular form is "datum."



Summarizing, averaging of previous earthmotion is an example of descriptive statistics .

Figure 1.4-1: Earthquake motions.

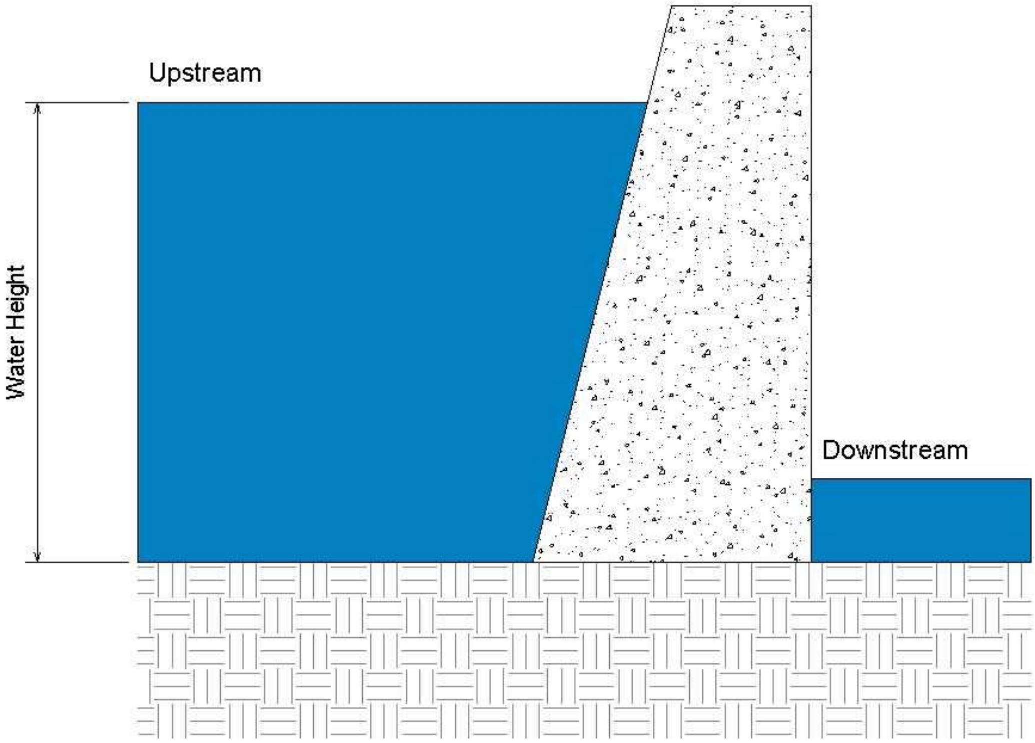


Figure 1.4-2: Water level in dam upstream.

1.5 Variables and Types of Data

Variables can be classified as ***qualitative*** or ***quantitative***.

1.5.1 Qualitative Variables

- Qualitative variables are variables that can be placed into distinct categories, according to some characteristics or attributes.
- For example,
  - Gender (male, female, or other),
  - Religious preference,
  - Geographical locations.
- Qualitative and categorical variables typically ***do not have units***.
- Qualitative and categorical variables ***have neither a "size" nor, typically, a natural ordering to their values***.
- They answer questions such as "***which kind***?"
- ***Arithmetic*** with qualitative variables usually ***does not make sense, even if the variables take numerical values***.

1.5.2 Quantitative Variables

- Quantitative variables are numerical variables that can be ordered or ranked.
- For example,
  - The variable (age) is numerical, and people can be ranked in accordance to the value of their ages.
  - Heights,
  - Weights,
  - Body temperatures.

Example 1.5-1

In a statistical survey, number "1" have been assigned to males while number "2" have been assigned to females. With this assignment, are the males and females classified as qualitative or quantitative data?

Solution

Even with this assignment, classifications of human beings into males and females still a qualitative one for following reasons:

- This assignment is arbitrary. For example, another survey may assign number "1" for females and number "2" for males.
- Algebraic operations on assigned numbers have no meaning. For example, there is no meaning for the difference between number "2" that assigned for females and number "1" that assigned to males.

Example 1.5-2

The Journal of Performance of Constructed Facilities (Feb., 1990) reported on the performance of water distribution networks dimensions in the Philadelphia area. For a part of the study, the following variables were measured for each sampled water pipe section. Identify the data produced by each as quantitative or qualitative.

- Pipe diameter (measured in inches)
- Pipe material (steel or PVC)
- Pipe location (Center City or suburbs)
- Pipe length (measured in feet)

Solution

Both pipe diameter (in inches) and pipe length (in feet) are measured on a meaningful numerical scale: hence, these two variables produce quantitative data.

Both of pipe material and pipe location can only classify the material as either steel or PVC and the location as either Center City or the suburbs. Consequently, pipe material and pipe location are both qualitative variables.

Home Work 1.5-1

Classify each variable as qualitative or quantitative.

a. Marital status of nurses in a hospital.

Qualitative

b. Time it takes to run a marathon.

Quantitative

c. Weights of lobsters in a tank in a restaurant.

Quantitative

d. Colors of automobiles in a shopping center parking lot.

Qualitative

e. Ounces of ice cream in a large milkshake.

Quantitative

f. Capacity of the NFL football stadiums.

Quantitative

g. Ages of people living in a personal care home.

Quantitative

Home Work 1.5-2

Classify each variable as qualitative or quantitative.

a. Marital status of nurses in a hospital.

Qualitative

b. Time it takes to run a marathon.

Quantitative

c. Weights of lobsters in a tank in a restaurant.

Quantitative

d. Colors of automobiles in a shopping center parking lot.

Qualitative

e. Ounces of ice cream in a large milkshake.

Quantitative

f. Capacity of the NFL football stadiums.

Quantitative

g. Ages of people living in a personal care home.

Quantitative

1.5.3 SUB CLASSIFICATION OF QUANTITATIVE VARIABLES

Quantitative variables can be further classified into two groups:

1.5.3.1 Discrete Variables

- Discrete variables assume values that can be counted.
- Examples of discrete variables are:
  - The number of children in a family,
  - The number of students in a classroom,
  - The number of calls received by a switchboard operator each day for a month.

1.5.3.2 Continuous Variables

- Continuous variables can assume an infinite number of values between any two specific values.
- They are obtained by measuring.
- They often include fractions and decimals.
- Examples of continuous variables are:
  - Height
  - Weight

1.5.4 SUMMARY OF DATA CLASSIFICATION

Above data classifications have been summarized in ***Figure 1.5-1***.

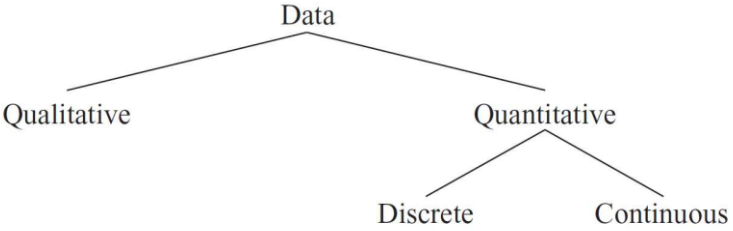


Figure 1.5-1: Summary of data classification.

Example 1.5-3

Classify each variable as discrete or continuous.

a. Number of pizzas sold by Pizza Express each day.

Discrete

b. Relative humidity levels in operating rooms at local hospitals.

Continuous

c. Number of bananas in a bunch at several local supermarkets.

Discrete

d. Lifetimes (in hours) of 15 iPod batteries.

Continuous

e. Weights of the backpacks of first graders on a school bus.

Continuous

f. Number of students each day who make appointments with a math tutor at a local college.

Discrete

g. Blood pressures of runners in a marathon.

Continuous

1.5.5 Data Classifications Based on Measurement Scales

In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured into:

- Nominal,
- Ordinal,
- Interval,
- Ratio.

1.5.5.1 Nominal Level of Measurement

- The nominal level of measurement classifies data into mutually exclusive (non-overlapping) categories in which no order or ranking can be imposed on the data.
- Examples
  - Classification of college instructors according to the subject they are teaching (e.g., English, history, psychology, or mathematics).
  - Classification according to religion (Islam, Christianity, Judaism, etc.),

1.5.5.2 Ordinal Level of Measurement

- The ordinal level of measurement classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.
- Examples
  - Based on students' evaluations, guest speakers might be ranked as superior, average, or poor.
  - People may be classified according to their build (small, medium, or large).
- Distance in the Ordinal Scale

  In the ordinal scale, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as:
  - 0=less than high school;
  - 1=some high school;
  - 2=high school degree;
  - 3=some college;
  - 4=college degree;
  - 5=post college.

  In this scale, higher numbers mean more education. But, is distance from 0 to 1 the same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure.

1.5.5.3 Interval Level of Measurement

- The interval level of measurement ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.
- Examples
  - IQ, Intelligence Quotient, is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110.
  - Temperature is another example of interval measurement, since there is a meaningful difference of 1°F between each unit, such as 72 and 73°F.

- Distance in the Interval Scale:
  In interval measurement, the distance between attributes does have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80.
- Ratio in the Interval Scale:
  In interval measurement, ratios do not make any sense. For example, 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

1.5.5.4 Ratio Level of Measurement

- The ratio level of measurement possesses all the characteristics of interval measurement, and ***a true zero does exist***.
- In addition, true ratios exist when the same variable is measured on two different members of the population. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1.
- Ratio scales have differences between units (1 inch, 1 pound, etc.),
- Examples of ratio scales are those used to measure:
  - Height,
  - Weight,
  - Area,
  - Number of phone calls received.

1.5.5.5 Summary of the four measurement scales:

The above four measurement scales have been summarized in ***Figure 1.5-2*** below.

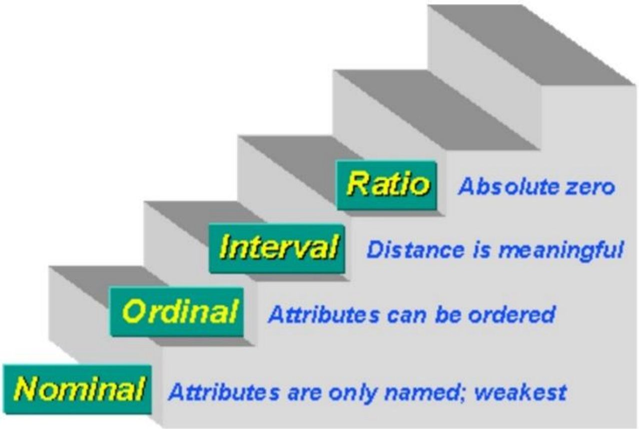
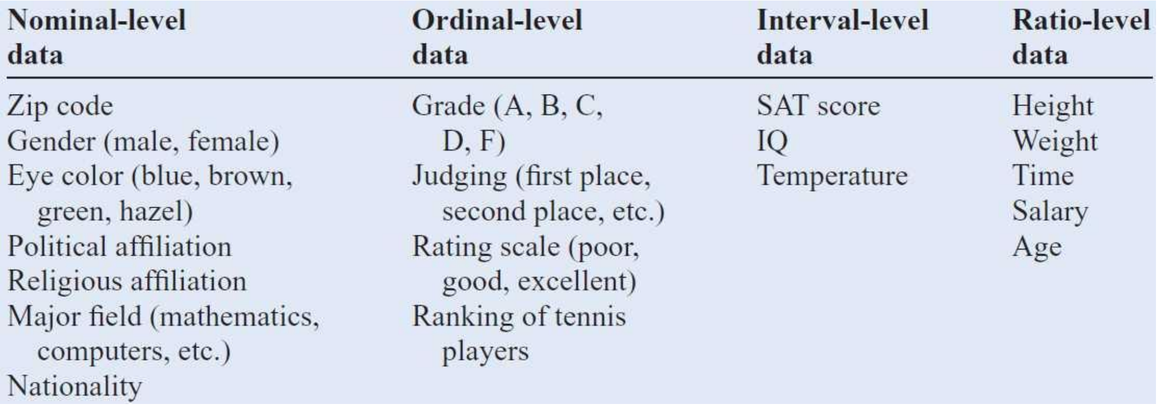


Figure 1.5-2: Summary of the four basic measurement scale.

1.5.5.6 Other Examples of Measurement Scales

Other examples of measurement scales are presented in Table 1.5-1.

Table 1.5-1: Examples of measurement scales

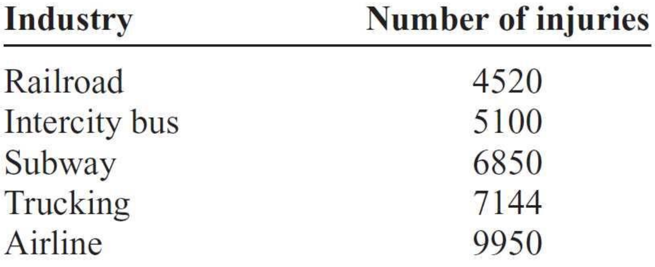


Nominal-level data	Ordinal-level data	Interval-level data	Ratio-level data
Zip code	Grade (A, B, C, D, F)	SAT score	Height
Gender (male, female)	Judging (first place, second place, etc.)	IQ	Weight
Eye color (blue, brown, green, hazel)	Rating scale (poor, good, excellent)	Temperature	Time
Political affiliation	Ranking of tennis players		Salary
Religious affiliation			Age
Major field (mathematics, computers, etc.)			
Nationality			

Example 1.5-4

Read the following information about the transportation industry and answer the questions.

Transportation Safety



Industry	Number of injuries
Railroad	4520
Intercity bus	5100
Subway	6850
Trucking	7144
Airline	9950

The chart shows the number of job-related injuries for each of the transportation industries for 1998.

1. What are the variables under study?
2. Categorize each variable as quantitative or qualitative.
3. Categorize each quantitative variable as discrete or continuous.
4. Identify the level of measurement for each variable.
5. The railroad is shown as the safest transportation industry. Does that mean railroads have fewer accidents than the other industries? Explain.

Solution

1. The variables are industry and number of job-related injuries.
2. The type of industry is a qualitative variable, while the number of job-related injuries is quantitative.
3. As there is no meaning for a fraction of injury, then the number of job-related injuries is discrete.
4. Type of industry is nominal, and the number of job-related injuries is ratio.
5. The railroads do show fewer job-related injuries; however, there may be other things to consider. For example, railroads employ fewer people than the other transportation industries in the study.

Example 1.5-5

Data in Table 1.5-2 represent live loads act on a building floor. Classify data into nominal, ordinal, interval, or ratio.

Table 1.5-2 Floor live loads, in psf.

79	76	68	70	90

Solution

As data have precise rank with meaningful zero value, therefore data are classified as Ratio Level.

Example 1.5-6

Data in

Table 1.5-3 represent angle of internal friction, in degrees, gathering during a soil investigation process.

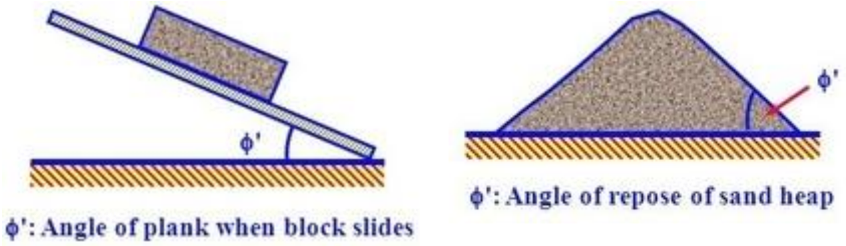

ϕ'
ϕ': Angle of plank when block slides
ϕ'
ϕ': Angle of repose of sand heap

Figure 1.5-3: Intuitive meaning for the angle of internal fiction.

Table 1.5-3: Angle of internal friction of soil, in degrees.

30	29	30	29

Solution

As data have a precise rank with meaningful zero value, therefore data are classified as ***Ratio Level***.

1.5.6 PROBLEMS

Problem 1.5-1

Drinking-water quality study Disasters (Vol. 28, 2004) published a study of the effects of a tropical cyclone on the quality of drinking water on a remote Pacific island. Water samples (size 500 milliliters) were collected for approximately 4 weeks after Cyclone Anti hit the island. The following variables were recorded for each water sample. Identify each variable as quantitative or qualitative.

1. Town where sample was collected
2. Typed water supply (river intake, stream, or borehole)
3. Acidic level (pH scale, 1 to 14)
4. Temperature (degrees Centigrade)

Answers

1. Qualitative
2. Qualitative
3. Quantitative
4. Quantitative

Problem 1.5-2

According to the National Bridge Inventory, All highway bridges in the United States are inspected periodically for structural deficiency by the Federal Highway Administration (FHWA). Data from the FHWA inspections are compiled into the National Bridge Inventory (NBI). Several of the nearly 100 variables maintained by the NBI are listed below. Classify each variable as quantitative or qualitative.

1. Length of maximum span (feet)
2. Number of vehicle lanes
3. Toll bridge (yes or no)
4. Average daily traffic
5. Condition of deck (good, fair, or poor)
6. Bypass or detour length (miles)
7. Route type (interstate, U.S., state, county, or city)

Answers

1. Quantitative
2. Quantitative
3. Qualitative
4. Quantitative
5. Qualitative
6. Quantitative
7. Qualitative

1.6 Accuracy, Precision, and Significant Figures

- Engineers must be aware of three principles that govern the proper use of numbers: ***accuracy***, ***precision***, and ***significant digits***.
- For engineering measurements, they are defined in subsequent articles.

1.6.1 Accuracy

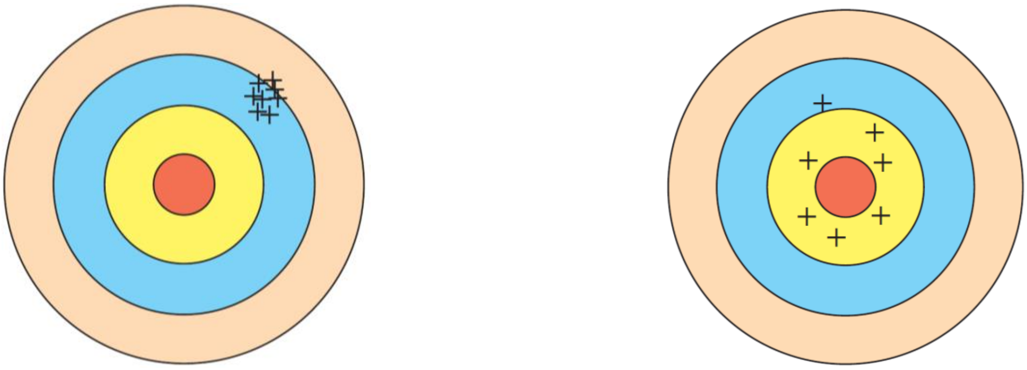
- The ***accuracy error*** is ***the value of one reading minus the true value***.
- In general, ***accuracy of a set of measurements refers to the closeness of the average reading to the true value***.
- Accuracy is generally associated with repeatable, fixed errors.

1.6.2 Precision

- The ***precision error*** is ***the value of one reading minus the average of readings***.
- In general, ***precision of a set of measurements refers to the fineness of the resolution and the repeatability of the instrument***.
- Precision is generally associated with unrepeatable, random errors.
- Accuracy versus precision:
  - A measurement or calculation can be very precise without being very accurate, and vice versa.
  - For example, suppose the ***true value of wind speed is 25.00 m/s***. Two anemometers A and B take five wind speed readings each:
    - Anemometer A:
      25.50, 25.69, 25.52, 25.58, and 25.61 m/s.
      Average of all readings = 25.58 m/s.
    - Anemometer B:
      26.3, 24.5, 23.9, 26.8, and 23.6 m/s.
      Average of all readings = 25.02 m/s.

    Clearly, ***anemometer A is more precise***, ***since none of the readings differs by more than 0.11 m/s from the average***. However, the average is 25.58 m/s, 0.58 m/s greater than the true wind speed; this indicates ***significant bias error***, also called ***constant error*** or ***systematic error***.

    On the other hand, ***anemometer B is not very precise***, since its readings swing wildly from the average; but its overall average is much closer to the true value. Hence, ***anemometer B is more accurate*** than anemometer A, at least for this set of readings, even though it is less precise.
  - Shooting arrows analogy for accuracy versus precision:
    - The difference between accuracy and precision can be illustrated effectively by analogy to shooting arrows at a target, as sketched in ***Figure 1.6-1***.
    - Shooter A is very precise, but not very accurate, while shooter B has better overall accuracy, but less precision.



Shooter A

Shooter B

Figure 1.6-1: Illustration of accuracy versus precision. Shooter A is more precise, but less accurate, while shooter B is more accurate, but less precise.

1.6.3 SIGNIFICANT FIGURES

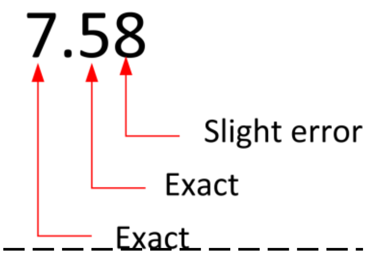
1.6.3.1 Basic Definitions

- Statistics deals extensively with approximations. Consequently, before discussing the errors associated with numerical values, it is useful to review basic concepts related to approximate representation of the numbers themselves.
- The meaning of significant figures is simply presented with referring to ***Table 1.6-1*** below.

Table 1.6-1: Significate figures.

If you know the number to:	Then report this many significant figures:
1 part of 10	1
1 part of 100	2
1 part of 1000	3
1 part of 10,000	4
1 part of 100,000	5
1 part of 1,000,000	6
Etc.	Etc.

- A significate figure is an ***accurate figure digit***, ***although the last digit is accepted to have some error***. For example, if 7.58 is a three significant figures number, then:



Example 1.6-1

With refereeing to the speedometer of Figure 1.6-2, one person estimates the speed 48.8, whereas another 48.9 km/h. What are the significant figures of the measured speed?

Solution

Because of the limits of the instrument, only the first two digits can be used with confidence. Estimates of the third digit (or higher) must be viewed as approximations. Therefore, the two estimations have ***three significant numbers***.

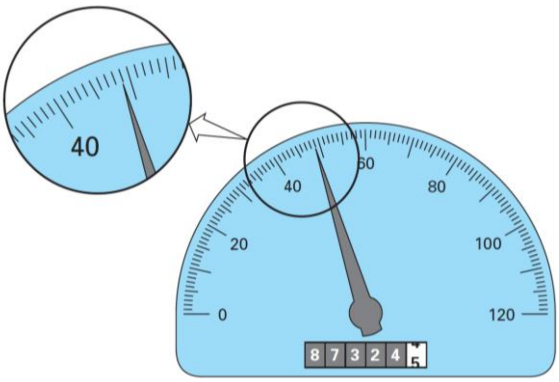


Figure 1.6-2: An automobile speedometer and odometer.

Example 1.6-2

With refereeing to Figure 1.6-2, one person estimates that the car has traveled slightly less than 87324.45 km during its lifetime. What are the significant figures for the estimated number?

Solution

Based on the above definition of the significant figures, the estimated value has ***seven significant figures***.

1.6.3.2 Basic Rules for Significant Figures

Although it is usually a straightforward procedure to ascertain the significant figures of a number, some cases can lead to confusion and the following rules would be useful.

- Rule 1:
  Non-zero numbers are ***always*** significant.
- Rule 2:
  In between zeros are ***always*** significant.
- Rule 3:
  Leading zeros are ***never*** significant.
- Rule 4:
  Trailing zeros are ***sometimes*** significant.

Example 1.6-3

Determine the significant figures for the following numbers.

2.34 56.6 7264

Solution

According to rule 1, as all digits are non-zero, therefore the first and second number has three significant figures while the third number has four significant figures.

Example 1.6-4

Determine the significant figures for the following numbers.

2.0004 90306 4000607

Solution

According to rule 2, as in between zeros are always significant, then

2.0004	90306	4000607
(5 significant figures)	(5 significant figures)	(7 significant figures)

Example 1.6-5

Determine the significant figures for the following numbers.

0.000067 0.000302

Solution

According to rule 3 as leading zeros are never significant, the indicated numbers have the following significant figures:

0.000067	0.000302
Two significant digits	Three significant digits.

Example 1.6-6

Determine the significant figures for the following numbers.

80000 80000. 2040 2040.0000

Hint

This example relates to rule 4 that is the trickiest one as the trailing zeros are ***sometimes*** significant. For further explanation, it is useful to say that ***if there is no decimal point the trailing zeros would be insignificant*** and vice versa.

Solution

Based on rule 4 and above further explanation, the significant figures for the indicated numbers would be:

80000	80000.	2040	2040.0000
One significant figure	Five significant figures	Three significant figures	Eight significant figures

1.6.3.3 Number of Significant Figure in Civil Engineering Applications

Generally, in civil engineering, measurements are accurate to 1 part of 1000. So, three significant figures are appropriate.

1.6.3.4 Adding and Subtracting of the Significant Figures

- To show how adding and/or subtracting significant figures, consider the adding of the following two number:

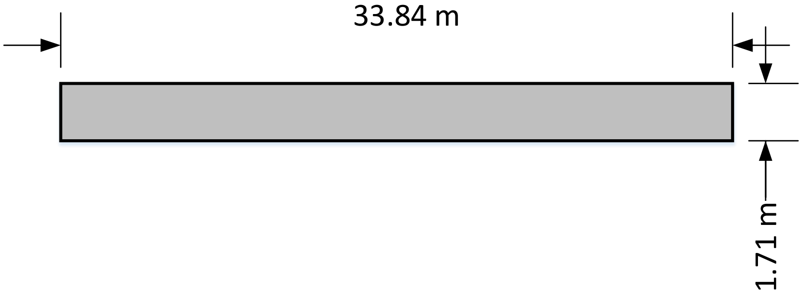
1.6.3.5 Significate Figures: Multiplication/Divisions

To determine the number of significant figures for a multiplication or division operation for two number, say AxB, the following steps should be followed:

- Indicate the number of significant figures for each number.
- Calculate the answer.
- Round the answer to have same number of significant figures as the least precise number.

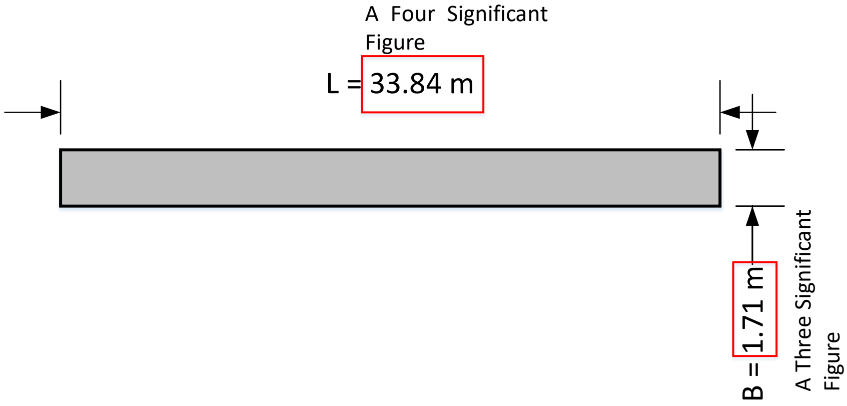
Example 1.6-7

Compute area for the rectangular shape shown below, your results should contain no more significant figures than the least accurate figure used in obtaining it.



Solution

Indicate the number of significant figures for each number.



Calculate the answer.

$$A = 33.84 \times 1.71$$
$$= 57.8664\, m^2$$

Round the answer to have same number of significant figures as the least precise number.

$$A = 57.9\ \ m^2$$

1.7 DATA COLLECTION AND SAMPLING TECHNIQUES*

- As stated previously, researchers use samples to collect data and information about a particular variable from a large population.
- Using samples saves time and money and in some cases enables the researcher to get more detailed information about a particular subject.

1.7.1 DATA COLLECTION

- Data can be collected in various ways such as surveys.
- Surveys can be done by using different methods. Three of the most common methods are:
  - Telephone survey,
  - Mailed questionnaire,
  - Personal interview.

1.7.2 SAMPLING TECHNIQUES

- Samples cannot be selected in haphazard ways because the information obtained might be biased.
- To obtain samples that are unbiased—i.e., that gives each subject in the population an equally likely chance of being selected—statisticians use four basic methods of sampling:
  - Random,
  - Systematic,
  - Stratified,
  - Cluster sampling.

1.7.2.1 Random Sample

Random samples are selected by using chance methods or random numbers.

1.7.2.1.1 Cards Method

- In the cards method:
  - Number each subject in the population.
  - Then place numbered cards in a bowl, mix them thoroughly,
  - Finally, select as many cards as needed.
- The subjects whose numbers are selected constitute the sample.

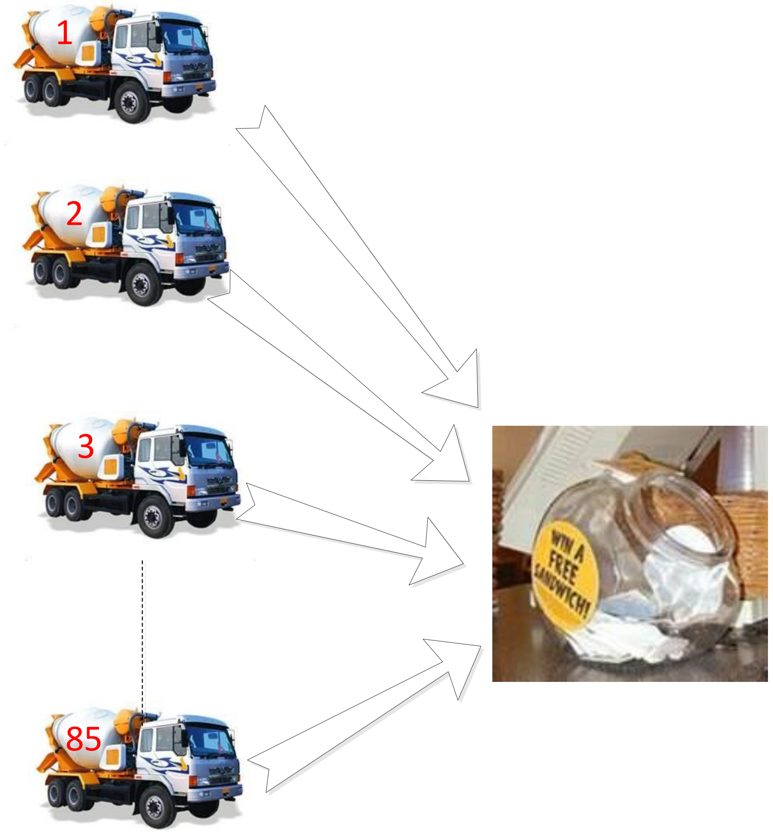
Example 1.7-1

Suppose that a site engineer has a concrete casting job that required a concrete volume of 85 truck. And assume that the site engineer intends to select 15 trucks from which concrete samples will be taken. How can site engineer select a random sample with a size of 15 from the 85 trucks.

Solution

One method is to prepare 85 numbered chards to simulate the number of trucks (a chard for each truck).

Then place numbered cards in a bowl, mix them thoroughly, and select the required 15 chards.



1.7.2.1.2 Tables for Random Numbers

- Since it is difficult to mix the cards thoroughly, there is a chance of obtaining a biased sample.
- For this reason, statisticians use another method of obtaining numbers. They generate random numbers with a computer or calculator.
- Before the invention of computers, random numbers were obtained from tables. Some two-digit random numbers are shown in Table 1.7-1 below.

Table 1.7-1: Random Numbers

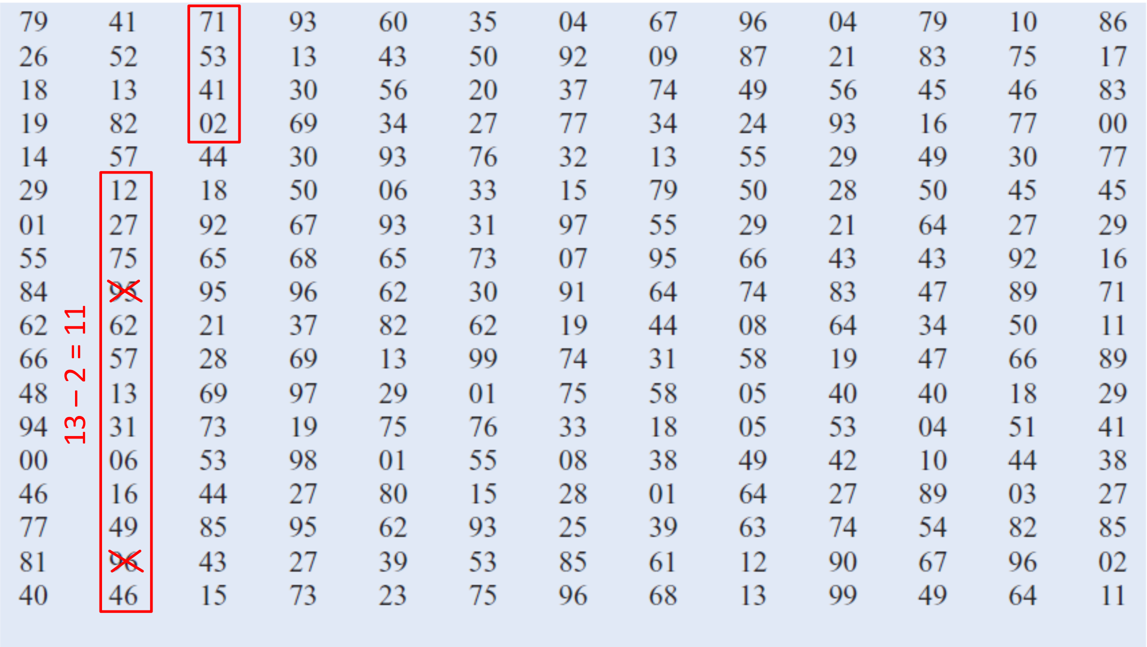
79	41	71	93	60	35	04	67	96	04	79	10	86
26	52	53	13	43	50	92	09	87	21	83	75	17
18	13	41	30	56	20	37	74	49	56	45	46	83
19	82	02	69	34	27	77	34	24	93	16	77	00
14	57	44	30	93	76	32	13	55	29	49	30	77
29	12	18	50	06	33	15	79	50	28	50	45	45
01	27	92	67	93	31	97	55	29	21	64	27	29
55	75	65	68	65	73	07	95	66	43	43	92	16
84	95	95	96	62	30	91	64	74	83	47	89	71
62	62	21	37	82	62	19	44	08	64	34	50	11
66	57	28	69	13	99	74	31	58	19	47	66	89
48	13	69	97	29	01	75	58	05	40	40	18	29
94	31	73	19	75	76	33	18	05	53	04	51	41
00	06	53	98	01	55	08	38	49	42	10	44	38
46	16	44	27	80	15	28	01	64	27	89	03	27
77	49	85	95	62	93	25	39	63	74	54	82	85
81	96	43	27	39	53	85	61	12	90	67	96	02
40	46	15	73	23	75	96	68	13	99	49	64	11

Example 1.7-2

Resolve ***Example 1.7-1*** above based on ***Table 1.6 1*** for Random Numbers.

Solution

- To select a random sample of 15 subjects out of 85 trucks:
  - It is necessary to number each trucks from 01 to 85.
  - Then select a starting number by closing your eyes and placing your finger on a number in the table. (Although this may sound somewhat unusual, it enables us to find a starting number randomly.) In this case ***suppose your finger landed on the number 12 in the second column***. (It is the sixth number down from the top.)
  - Then proceed downward until you have selected 15 different numbers between 01 and 85. When you reach the bottom of the column, go to the top of the next column.
  - If you select a number greater than 85 or the number 00 or a duplicate number, just omit it.
  - In our example, we will use the trucks numbered 12, 27, 75, 62, 57, 13, 31, 06, 16, 49, 46, 71, 53, 41, and 02.



79	41	71	93	60	35	04	67	96	04	79	10	86
26	52	53	13	43	50	92	09	87	21	83	75	17
18	13	41	30	56	20	37	74	49	56	45	46	83
19	82	02	69	34	27	77	34	24	93	16	77	00
14	57	44	30	93	76	32	13	55	29	49	30	77
29	12	18	50	06	33	15	79	50	28	50	45	45
01	27	92	67	93	31	97	55	29	21	64	27	29
55	75	65	68	65	73	07	95	66	43	43	92	16
84	~~95~~	95	96	62	30	91	64	74	83	47	89	71
62	62	21	37	82	62	19	44	08	64	34	50	11
66	57	28	69	13	99	74	31	58	19	47	66	89
48	13	69	97	29	01	75	58	05	40	40	18	29
94	31	73	19	75	76	33	18	05	53	04	51	41
00	06	53	98	01	55	08	38	49	42	10	44	38
46	16	44	27	80	15	28	01	64	27	89	03	27
77	49	85	95	62	93	25	39	63	74	54	82	85
81	~~96~~	43	27	39	53	85	61	12	90	67	96	02
40	46	15	73	23	75	96	68	13	99	49	64	11

13 – 2 = 11

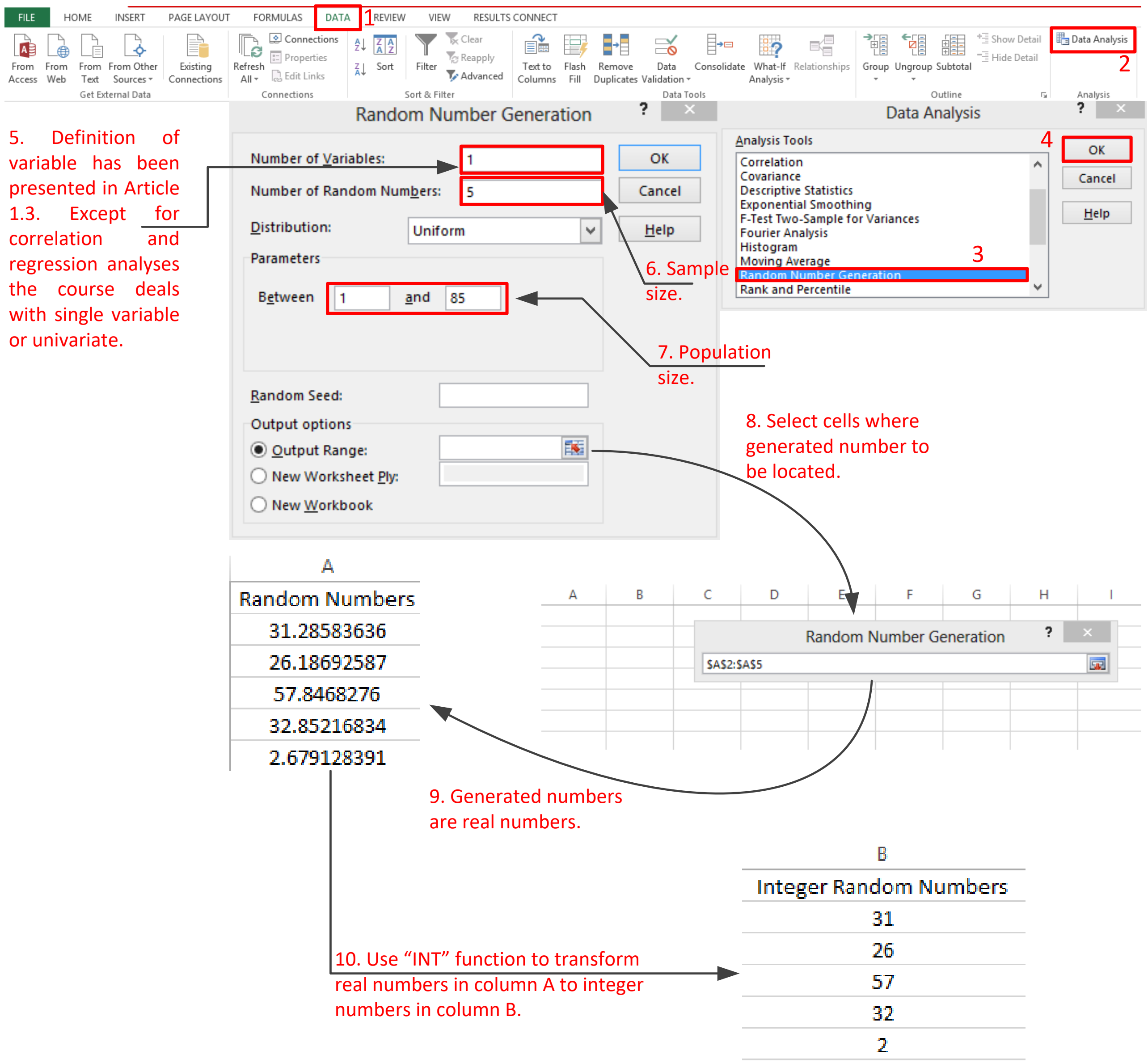
Example 1.7-3

Resolve ***Example 1.7-1*** above but with using Microsoft Excel to generate the required random number for the 15 trucks that should be selected from completely 85 trucks.

Solution

Microsoft Excel is so powerful and easy to use software that can be adopted in most of statistical analyses. To generate random number using Excel, steps indicated below can be adopted. As will be discussed in ***Chapters 4*** and ***5***, random variables should follow a specific probability model.

In the indicated steps, the uniform model is adopted to generate the random numbers. With this model, each truck has the same likely chance to be selected in the sample.



Data Analysis

Analysis Tools: Correlation, Covariance, Descriptive Statistics, Exponential Smoothing, F-Test Two-Sample for Variances, Fourier Analysis, Histogram, Moving Average, Random Number Generation, Rank and Percentile

OK | Cancel | Help

Random Number Generation

Number of Variables: 1

Number of Random Numbers: 5

Distribution: Uniform

Parameters

Between 1 and 85

Random Seed:

Output options

● Output Range:

○ New Worksheet Ply:

○ New Workbook

OK | Cancel | Help

Random Number Generation: $A$2:$A$5

5. Definition of variable has been presented in Article 1.3. Except for correlation and regression analyses the course deals with single variable or univariate.

6. Sample size.

7. Population size.

8. Select cells where generated number to be located.

9. Generated numbers are real numbers.

10. Use “INT” function to transform real numbers in column A to integer numbers in column B.

A
Random Numbers
31.28583636
26.18692587
57.8468276
32.85216834
2.679128391

B
Integer Random Numbers
31
26
57
32
2

1.7.2.2 Systematic Sampling:

Researchers obtain ***systematic samples*** by numbering each subject of the population and then selecting every kth subject.

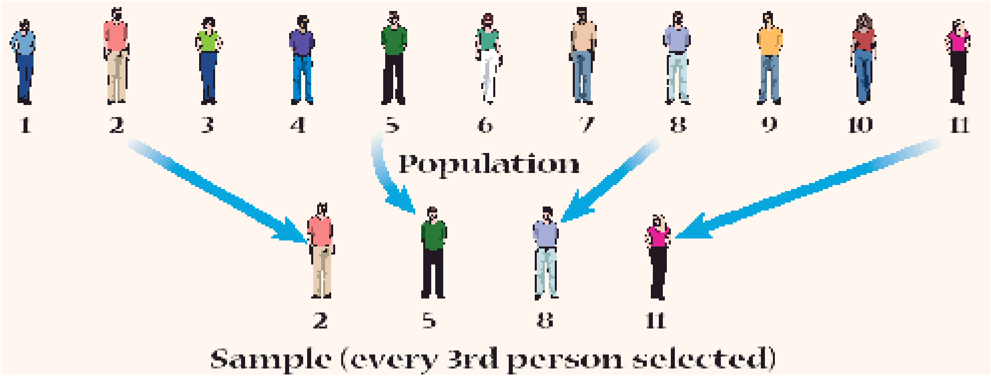


Figure 1.7-1: Sketch for systematic sample.

Example 1.7-4

Resolve Example 1.7-1 above based on a systematic sampling technique.

Solution

- As there were 85 trucks in the population and a sample of 15 trucks were needed. Since $85 \div 15 = 5.67$, then $k = 5$, and every 5th truck would be selected;

- However, the first truck (numbered between 1 and 5) would be selected at random. Suppose truck 3 were the first subject selected; then the sample would consist of the truck whose numbers were

3 8 13 18 23 28 33 38 43 48 53 58 63 68 73

1.7.3 Stratified Sampling

- Researchers obtain ***stratified samples*** by
  - Dividing the population into groups (called strata) according to some characteristics that are important to the study.
  - Then sampling from each group. ***Samples within the strata should be randomly selected***.
- For example:
  Suppose that a site manager aims to collect opinions of his staff about live support level in the site. Furthermore, the site manager wishes to see if the opinions of the engineering staff differ from those of the non-engineering staff. Then, he will classified the staff into engineering and non-engineering and select randomly from each stratum.

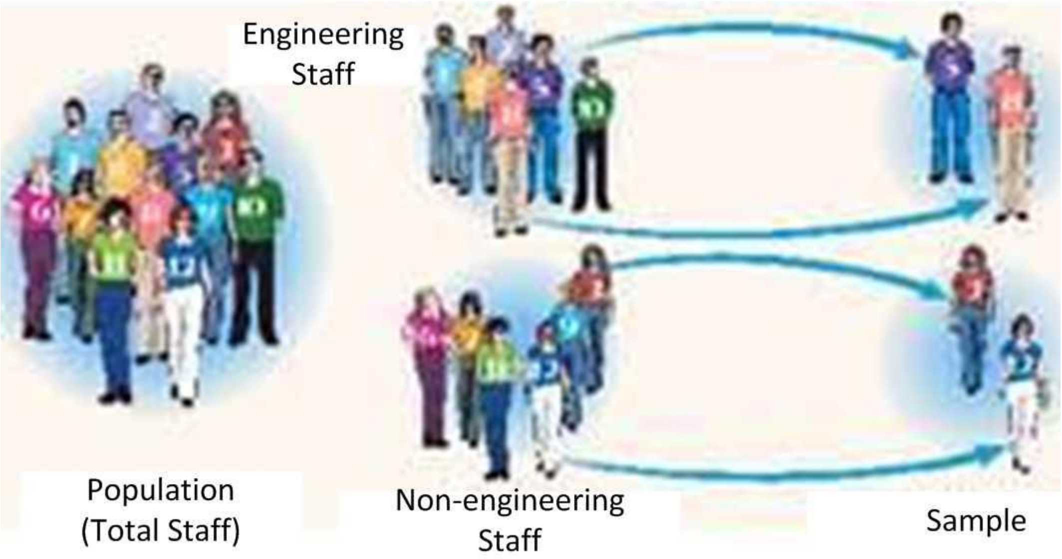


Figure 1.7-2: An example on stratification sampling.

1.7.4 Cluster Sampling

- Here the population is divided into groups called clusters by some means such as geographical area.
- Then the researcher randomly selects some of these clusters and uses all members of the selected clusters as the subjects of the samples.
- Cluster sampling is used when:
  - The population is large,
  - Or when it involves

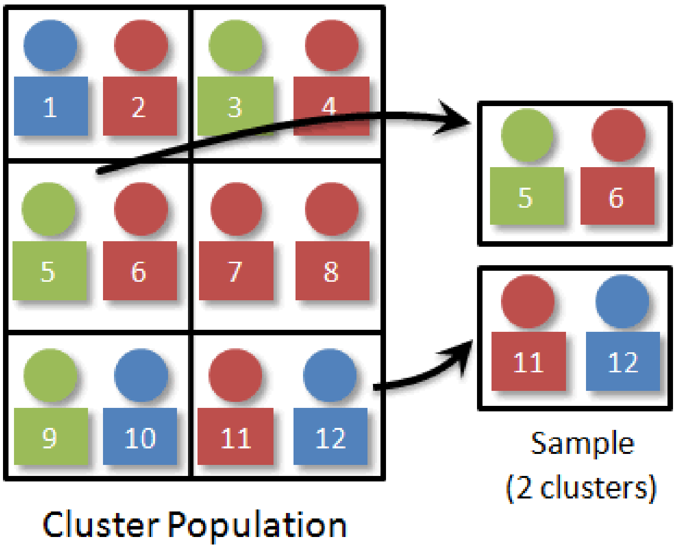


Figure 1.7-3: A sketch for cluster sampling.

subjects residing in a large geographical area.

- For example:
  If one wanted to do a study involving the patients in the hospitals of Baghdad, it would be very costly and time-consuming to try obtaining a random sample of patients since they would be spread over a large area. Instead, a few hospitals could be selected randomly, and the patients from these hospitals would be interviewed in a cluster.

1.8 UNCERTAINTY IN ENGINEERING*

- As the engineer in general, and civil engineer in particular, has insufficient information, therefore he/she should make a decision with uncertainty.
- This section aims to show that there are two uncertainty types:
  - First Type: ***Uncertainty Associated with Randomness*** (The ***aleatory uncertainty***)

    It is the uncertainty associated with the randomness of the underlying phenomenon that is exhibited as variability in the observed information.
  - Second Type: ***Uncertainty Associated with Imperfect Knowledge*** (The ***Epistemic Uncertainty***)

    It is the uncertainty associated with imperfect models of the real world because of insufficient or imperfect knowledge of reality.
- The two types of uncertainty may ***be combined*** and ***analyzed as a total uncertainty*** or ***treated separately***. In either case, the principles of ***probability*** and ***statistics*** apply equally.

1.8.1 UNCERTAINTY ASSOCIATED WITH RANDOMNESS (ALEATORY UNCERTAINTY)

- The aleatory uncertainty (sometimes called ***databased***, ***natural***, ***intrinsic***, ***irreducible*** or ***fundamental*** uncertainty) ***is associated with the inherent variability of basic information***, which is part of the real world (within our ability to observe and describe).
- For the bolt example, even we have a perfect condition with full information about the strength of each bolt in the population, we still have uncertain variation in strength of the bolt. This type of uncertain or random variation even with full population survey represents the aleatory uncertainty.
- It results when observed measurements are different from one experiment (or one observation) to another, even if conducted or measured under ***apparently identical conditions***, at least as the research though.
- In the bolt example, they ***appear identical*** and to be ***tested under identical conditions***. Therefore, result differences are understood as a random or uncertain in nature.
- It is ***probabilistic in nature*** and ***studied based on the probability theory*** of ***Chapters 4, 5, and 6 of this course***.

1.8.2 UNCERTAINTY ASSOCIATED WITH IMPERFECT KNOWLEDGE (THE EPISTEMIC UNCERTAINTY)

- The epistemic (or knowledge-based) uncertainty is ***associated with imperfect knowledge of the real world*** and ***may be reduced through application of better prediction models and/or improved experiments***.
- To understand this type of uncertainty it useful to return to the bolt example the further explanation of the analysis process.
  - After the data describing, the analyst uses the sample measurements or statistics to estimate the population parameters. Hence, there is an epistemic uncertainty about how accurately the sample can represent the population.
  - Then he/she should select a suitable probability model to replace the actual histogram for the subsequent analysis. See Figure 1.8-1. Each probability model has its own equation and parameters.

  - In summary, in the epistemic uncertainty, we deal with the selection of a suitable model with suitable parameters based on the incomplete sample data.
- Based on the definition of the statistics science, the epistemic uncertainty is ***statistical in nature***. It is not fundamental and ***can be virtually eliminated at the expense of a very large sample size***.

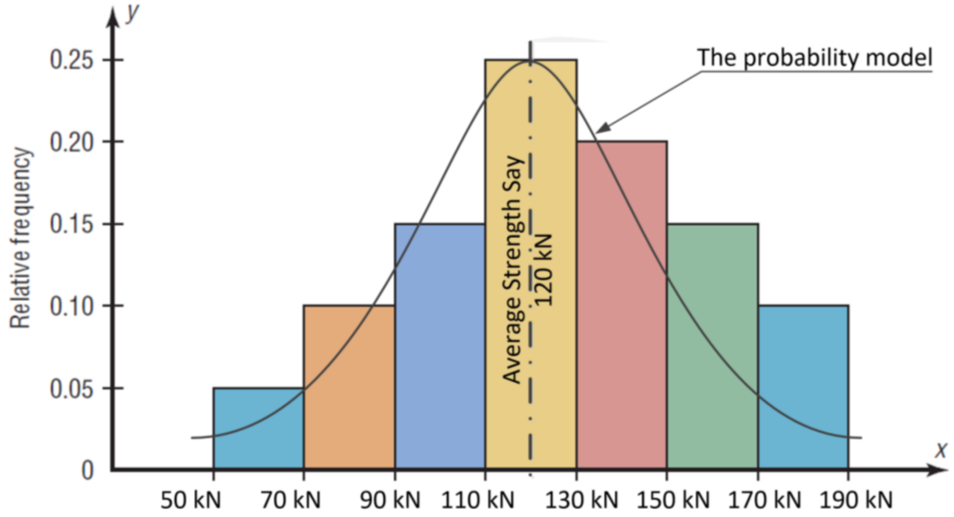


Figure 1.8-1: Proposing of a probability model to simulate the data histogram.

Example 1.8-1

With refereeing to bolt strength data of Figure 1.8-1:

- What type of uncertainty is there if the analyst intends to analyze the variation in the sample itself?
- What type of uncertainty is there if the analyst intends to use the sample to infer the population features?

Solution

- In the first case, the analyst implicitly considers the sample as a population by itself. Therefore, he/she deals with an aleatory uncertainty that is probabilistic in nature.
- In the second case, the analyst has an epistemic uncertainty that is statistical in nature and can be reduced or eliminated with a large sample.

Example 1.8-2

Consider the calculation of the deflection of a prismatic cantilever beam under the concentrated load P as shown in Figure 1.8-2. For engineering purposes, the deflection at the end of the beam B is usually calcu1ated based on the simple beam theory, which gives

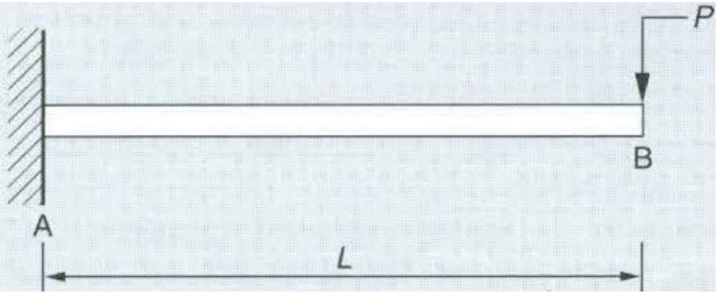


Figure 1.8-2: A cantilever beam subjected to concentrated load P.

$$\Delta_{\mathrm{B}} = \frac{PL^3}{3EI} \qquad \text{Eq. 1.8-1}$$

in which

E = the modulus of elasticity of the material, and

I = the moment of inertia of the beam cross section.

The ***Eq. 1.8-1*** is based on several idealized assumptions, which are a. follows:

1. The material is linearly elastic.
2. Under the load P, plane sections of the beam remain plane.
3. The support of the beam at A is perfectly rigid.

In reality, each of the above assumptions may not be totally valid. For example, depending on the material of the beam and the magnitude of the load, the behavior may not be linearly elastic. Also, for high loads. plane sections of the beam will not remain plane; finally, it is seldom that the support at A can be perfectly rigid. Therefore, the calculation of the deflection, Δ_B, with Eq. 1.8-1 will involve some undetermined error and thus some uncertainty. Unless there is reason to believe otherwise, it is reasonable to assume that the calculated deflection is the mean deflection of the cantilever beam. implying that the error of Eq. 1.8-1 is symmetric about the average deflection. However, if necessary, the result may be connected for any bias in the calculation.

One way to evaluate or assess the error of ***Eq. 1.8-1*** is to test a number of similar cantilever beams of the same material under the same conditions, and with precisely measured values of P, L, E, and I. The results of the tests should provide us with a basis for evaluating the error underlying the equation. For illustration, suppose (hypothetically) that 10 wood beams were tested and the results (in terms of the ratios of the measured to the calculated deflections) were as follows:

$\frac{\Delta_{B\ Measured}}{\Delta_{B\ theoritical}}$	1.05	0.95	1.10	0.98	1.15	0.97	1.20	1.00	1.12	1.08

These test results would yield the following sample mean and sample standard deviation (see Chapter 3) for the ratio of the measured to the calculated deflection Δ_B:

$Mean\ (or\ average)\ ratio\ =\ 1.06$

$Standard\ deviation\ of\ ratio\ =\ 0.084$

According to the above test results, the deflection of the beam calculated with Eq. 1.8-1 tends to ***underestimate*** the correct deflection and, therefore, needs to be corrected by the ***bias factor*** of 1.06. whereas its coefficient of variation (defined as ***the standard deviation divided by the mean***) representing the epistemic uncertainty is 0.084/1.06 = 0.0793.

Depending on the material, there may also be significant variability in the EI of the beam. This variability in El will also lead to ***aleatory uncertainty*** in the calculated end deflection Δ. Finally. if the El of ***the beam is estimated from sampled observations***, ***there will also be sampling error in the estimated mean El***, and this will add ***further epistemic uncertainty*** to the calculated deflection of the beam.

1.9 Design and Decision Making under Uncertainty*

- Imperfect models is indispensable
- As we indicated earlier, in engineering information no single observation or measurement is representative; and any evaluation or prediction must be performed based on ***imperfect models*** of the real world. That is, uncertainty (aleatory and/or epistemic) is unavoidable in engineering.
- Under the preceding situation, how should engineering designs be formulated or decisions affecting a design and planning be determined? Presumably, we may ***assume consistently worst*** conditions (e.g., specify the highest possible flood, smallest observed fatigue life of materials) and develop conservative designs on this basis. From ***the standpoint of system performance and safety, this approach may be suitable***, and indeed has been the basis for much of engineering designs and planning in the past and can be expected to continue in the future. However, this approach eschews any information on risk and lacks a systematic basis for evaluating the degree of conservativeness; ***a resulting design that is overly conservative may be excessively costly***, whereas one with insufficient conservatism may be inexpensive but will sacrifice performance or safety. ***The optimal decision ought to be based on a trade-off between cost and benefit***, in order to achieve a balance between cost and system performance. As the available information and evaluative models are invariably imperfect or insufficient, and thus contain uncertainties, the required trade-off analysis ought to be performed with in the context of probability and risk.
- The situations described above are common to many problems in engineering; in the following we describe several examples illustrating some of these problems. The examples are idealized to simplify the presentations; nevertheless, they serve to illustrate the essence of the decision-making aspects of engineering under conditions of uncertainty.

1.9.1 Design of Structures and Machines

- In general, the randomness in the structural behavior is classified into the randomness in the strength of the section, member, or system and the randomness in the applied loads.
- Randomness in martial strength:
  - The strengths of structural material and components are random and thus contain variabilities as illustrated in ***Figure 1.9-1*** and ***Figure 1.9-2*** for concrete and steel;
  - Therefore, the calculation of the capacity of a structure (invariably based on an idealized model) will contain both aleatory and epistemic uncertainties.

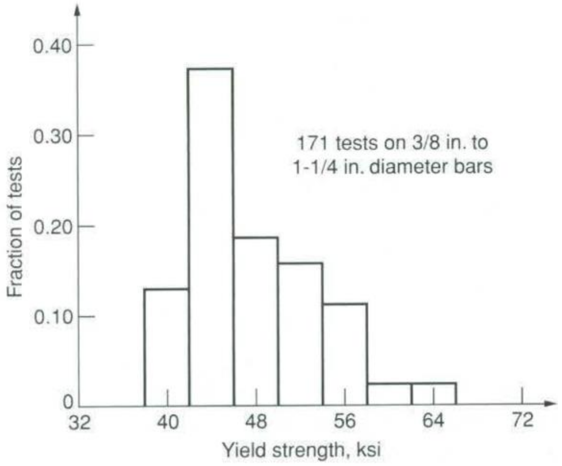


Figure 1.9-1: Yield strength of reinforcing bars.

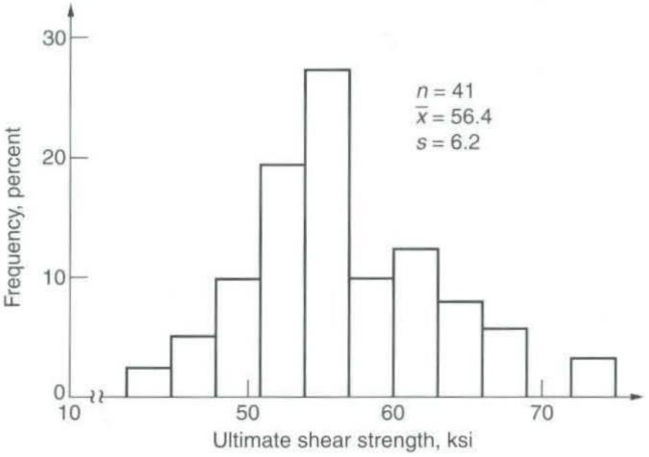
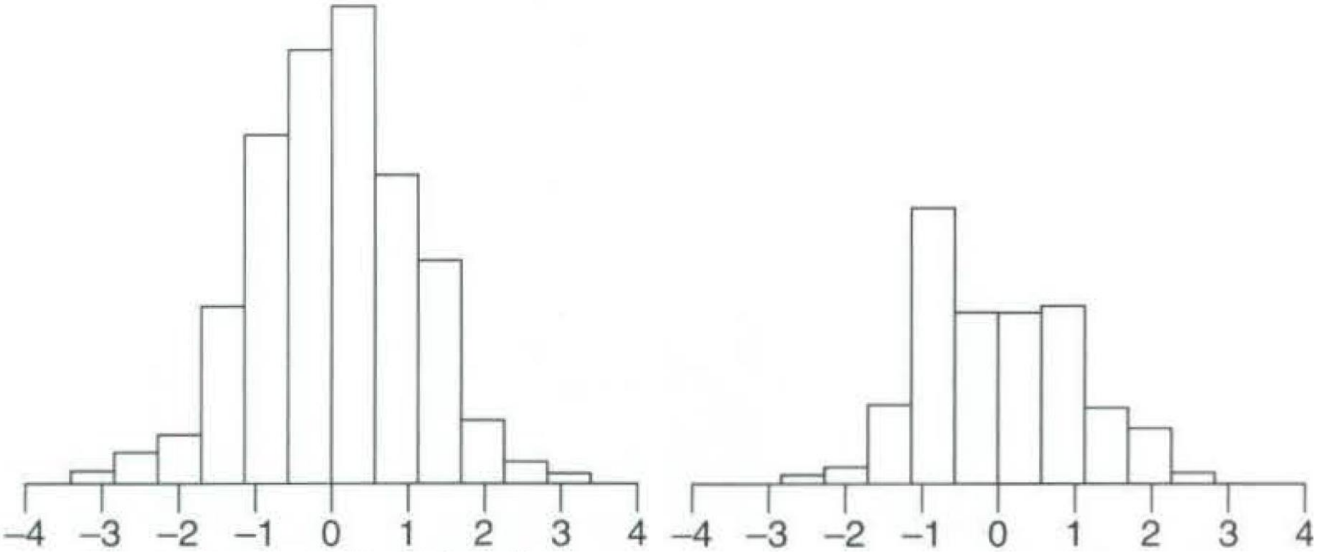


Figure 1.9-2: Ultimate shear strength of fillet welds.

- Randomness in the applied loads:
  On the other hand. the applied loads on a structure are also invariably random, as illustrated in ***Figure 1.9-3*** for wind pressures on tall buildings.



(a) Typhoon Ruby 16/7/1970 7SW-3 **(b) Typhoon Georgia 13/9/1970 7NE-4**

Figure 1.9-3: Pressure fluctuations on tall buildings during typhoons.

- Therefore, the design of a structure, i.e., involving the determination of the load carrying capacity of the structure, must consider the question of "***how safe is safe enough***? "a question that realistically requires the consideration of risk and the probability of nonperformance or failure.

1.9.2 PLANNING AND DESIGN OF TRANSPORTATION INFRASTRUCTURES

This article is not complete yet.

1.10 REASONING AND FALLACIES*

- This article reviews logical rules that produce ***valid arguments*** and common rule violations that lead to ***fallacies***.
- Understanding fallacies helps us to avoid committing them and to recognize fallacious arguments made by others.

1.10.1 TYPES OF REASONING

Reasoning can be ***inductive*** or ***deductive***.

1.10.2 DEDUCTIVE REASONING

- It is what we call "***logic***" informally.
- It is a way of ***thinking mathematically about all kinds of things***: ***Given*** a set of assumptions (***premises***), ***what must then be true***?
- Deductive reasoning—if logical—is ***as certain as mathematics can be***.

1.10.3 INDUCTIVE REASONING

- Inductive reasoning attempts to generalize from experience (data) to new situations.
- It is ***inherently uncertain***.

1.10.4 WHICH TYPE OF REASONING IS THE STATISTICS

- Much of Statistics concerns ***inductive reasoning***.
- Exceptional care is needed to draw reliable conclusions by inductive reasoning.
- ***Good inductive reasoning requires correct deductive reasoning, the subject of this article***.

1.10.5 VALID, FALLACIOUS, AND SOUND DEDUCTIVE REASONING

- Deductive reasoning that is ***mathematically correct (logical) is valid***.
- Deductive reasoning that is ***incorrect (logically faulty, illogical) is fallacious***.
- Reasoning ***can be valid even if the assumptions on which it is based are false***. ***If reasoning is valid and based on true premises, it is sound***.
- ***Many deductive and inductive arguments rely on statistical evidence***. Even ***the best statistical evidence can lead to wrong conclusions if it is used in a fallacious argument***.

1.10.6 ARGUMENTS

- An argument is a ***sequence of statements***, ***one of which is called the conclusion***. The other statements are ***premises*** (***assumptions***).
- The argument presents the premises—collectively— as evidence that the conclusion is true.

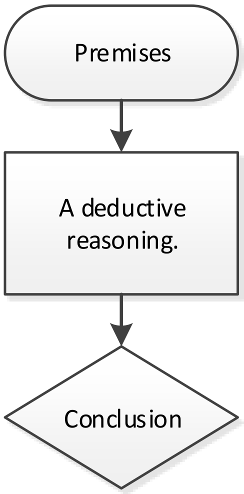


Figure 1.10-1: The structure of argument.

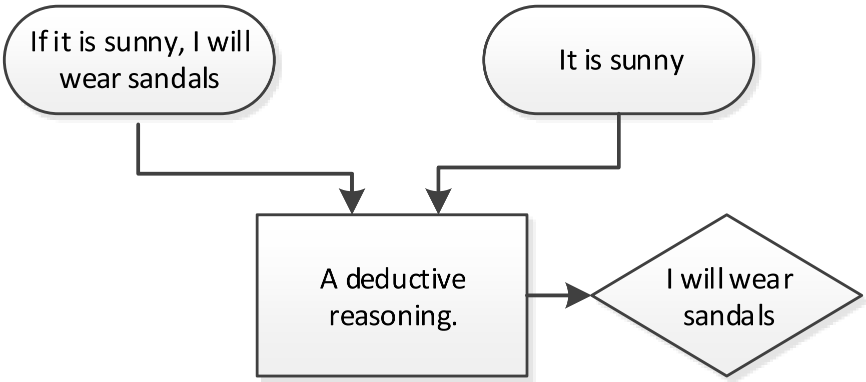
- For instance, the following is an argument:

  $If\ A\ is\ true\ then\ B\ is\ true. A\ is\ true. Therefore, B\ is\ true.$

  The conclusion is that $B\ is\ true$.
  - The ***premises*** are ***If A is true then B is true and A is true***.
  - The premises support ***the conclusion*** that $B\ is\ true$.
  - The word "$therefore$" ***is not part of the conclusion***. ***It is a signal that the statement after it is the conclusion***.

Example 1.10-1

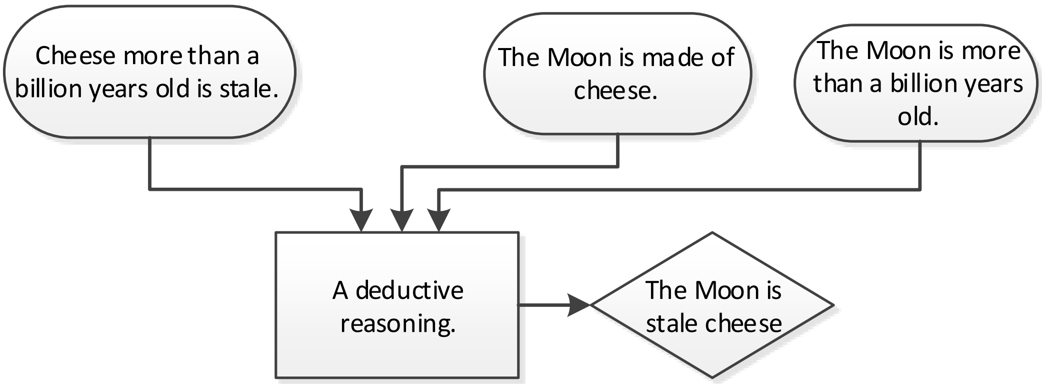
Example for a ***valid argument*** that ***may be sound or not***:



As the premise $I\ will\ wear\ sandals$ is questionable, therefore the argument may be sound or not.

Example 1.10-2

Example for a ***valid*** but ***unsound argument*** that may be sound or not:


Cheese more than a billion years old is stale.
The Moon is made of cheese.
The Moon is more than a billion years old.
A deductive reasoning.
The Moon is stale cheese

As one is well convinced that the moon is not made from cheese, therefore the argument is ***valid*** but ***unsound***.

Example 1.10-3

Formulate the argument of ***Example 1.10-2*** in a logical form.

Solution

The logical form of the argument of ***Example 1.10-2*** is (roughly):

$For\ any\ x, if\ x\ is\ A\ and\ x\ is\ B\ then\ x\ is\ C. y\ is\ A. y\ is\ B. Therefore, y\ is\ C.$

Here

- $A\ is$ "$made\ of\ cheese$", $B\ is$ "$more\ than\ a\ billion\ years\ old$" and $C\ is$ "$stale$".
- The symbol x is a free variable that can stand for anything;
- The symbol y stands for the Moon.

1.10.7 Some Logical Terms

- In the conditional $If\ A\ then\ B$, ***A is called the antecedent*** and ***B is called the consequent***.
- Affirming: To affirm something is to assert that it is true.
- Denying: To deny something is to assert that it is false.

1.10.8 Some Valid Rules of Reasoning

Some valid rules of reasoning are presented in ***Table 1.10-1*** below.

Table 1.10-1: Valid rules of reasoning.

A or not A. (***Law of the excluded middle***).

Not (A and not A).

A. Therefore, A or B.

A. B. Therefore, A and B.

A and B. Therefore, A.

Not A. Therefore, not (A and B).

A or B. Not A. Therefore, B. (***Denying the disjunction***).

Not (A and B). Therefore, (not A) or (not B). (***de Morgan***).

Not (A or B). Therefore, (not A) and (not B). (***de Morgan***).

If A then B. A. Therefore, B. ***(Affirming the precedent***, "***affirming by affirming***").

If A then B. Not B. Therefore, not A. (***Denying the consequent***, "***denying by denying***").

1.10.9 Formal Fallacies

They are errors that result from misapplying or not following the rules of ***Table 1.10-1***.

Example 1.10-4

Show if the following argument is valid or not:

$If\ A\ then\ B. B. Therefore, A.$

Solution

It is very useful for the student to "plug in" values in ***abstract expressions*** to get plain-language examples. Therefore, if premises of ***Example 1.10-1*** above in to the argument:

$If\ it\ is\ sunny, I\ will\ wear\ sandals. I\ will\ wear\ sandals. Therefore, it\ is\ sunny.$

This is a fallacy known as ***affirming the consequent***. The premises say that $if\ A\ is\ true, B\ must\ also\ be\ true$. It does $not\ follow\ that\ if\ B\ is\ true, A\ must\ also\ be\ true$.

To draw the conclusion that A is true, we need ***an additional premise***: $If\ B\ then\ A$. That premise, together with the other two premises, would allow us to conclude that A is true.

1.11 GENERAL EXAMPLES

Example 1.11-1

In each of these statements, tell whether descriptive or inferential statistics have been used.

a. By 2040 at least 3.5 billion people will run short of water (World Future Society). ***Inferential***

b. Nine out of ten on-the-job fatalities are men (Source: USA TODAY Weekend). ***Descriptive***

c. Expenditures for the cable industry were $5.66 billion in 1996 (Source: USA TODAY). ***Descriptive***

d. The median household income for people aged 25–34 is $35,888 (Source: USA TODAY). ***Descriptive***

e. Allergy therapy makes bees go away (Source: Prevention). ***Inferential***

f. Drinking decaffeinated coffee can raise cholesterol levels by 7% (Source: American Heart Association).

g. The national average annual medicine expenditure per person is $1052 (Source: The Greensburg Tribune Review). ***Descriptive***

h. Experts say that mortgage rates may soon hit bottom (Source: USA TODAY). ***Inferential.***

Example 1.11-2

Classify each as nominal-level, ordinal-level, interval-level, or ratio-level measurement.

a. Pages in the 25 best-selling mystery novels. ***Ratio***

b. Rankings of golfers in a tournament. ***Ordinal***

c. Temperatures inside 10 pizza ovens. ***Interval***

d. Weights of selected cell phones. ***Ratio***

e. Salaries of the coaches in the NFL. ***Ratio***

f. Times required to complete a chess game. ***Ratio***

g. Ratings of textbooks (poor, fair, good, excellent). ***Ordinal***

h. Number of amps delivered by battery chargers. ***Ratio***

i. Ages of children in a day care center. ***Ratio***

j. Categories of magazines in a physician’s office (Sports, women’s, health, men’s, news). ***Nominal***

Example 1.11-3

Classify each sample as random, systematic, stratified, or cluster.

a. In a large school district, all teachers from two buildings are interviewed to determine whether they believe the students have less homework to do now than in previous years. ***Cluster***

b. Every seventh customer entering a shopping mall is asked to select her or his favorite store. ***Systematic***

c. Nursing supervisors are selected using random numbers to determine annual salaries. ***Random***

d. Every 100th hamburger manufactured is checked to determine its fat content. ***Systematic***

e. Mail carriers of a large city are divided into four groups according to gender (male or female) and according to whether they walk or ride on their routes. Then 10 are selected from each group and interviewed to determine whether they have been bitten by a dog in the last year. ***Stratified***

Example 1.11-4

Identify each study as being either observational or experimental.

a. Subjects were randomly assigned to two groups, and one group was given an herb and the other group a placebo. After 6 months, the numbers of respiratory tract infections each group had were compared. ***Experimental***

b. A researcher stood at a busy intersection to see if the color of the automobile that a person drives is related to running red lights. ***Observational***

c. A researcher finds that people who are more hostile have higher total cholesterol levels than those who are less hostile. ***Observational***

d. Subjects are randomly assigned to four groups. Each group is placed on one of four special diets—a low-fat diet, a high-fish diet, a combination of low-fat diet and high-fish diet, and a regular diet. After 6 months, the blood pressures of the groups are compared to see if diet has any effect on blood pressure. ***Experimental***

1.12 CONTENTS

CHAPTER 2 FREQUENCY DISTRIBUTIONS AND GRAPHS

2.1 INTRODUCTION

- Data Organization
    - To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way.
    - The most convenient method of organizing data is to construct a ***frequency distribution***.
- Charts and Graphs
    - After organizing the data, the researcher presents them so they can be understood by those who will benefit from reading the study.
    - The most useful method of presenting the data is by constructing ***statistical charts and graphs***. There are many different types of charts and graphs, and each one has a specific purpose.
- Chapter Overview

  This chapter explains
    - How to organize data by constructing frequency distributions is presented ***Article 2.2***.
    - How to present the data by constructing charts and graphs is presented in ***Article 2.3***. The charts and graphs illustrated in this include:
        - Histograms,
        - Frequency polygons,
        - Ogives.
    - Finally, some Matlab codes to prepare a frequency table and to draw a histogram are presented in ***Article 2.4***. This article is out the scope of this undergraduate course.

2.2 Organizing Data

2.2.1 Basic Concepts Related to Frequency Distribution

- Suppose a researcher wishes to study the working ages (in months) of 50 portable concrete mixers that are produced by a specific manufacturer.
- In general, he/she shall follow steps indicated in below:

2.2.1.1 Gathering of Raw Data

- The researcher first would have to get the data on the ages of the mixers.
- In this case, these data are collected from different sites that used this type of concrete mixer. When the data are in original form, they are called ***raw data*** and are listed next.

Table 2.2-1: Raw data for working ages of concrete mixer.

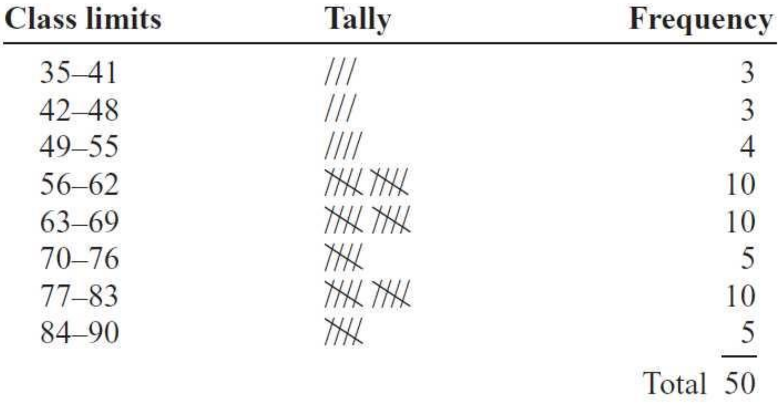
49	57	38	73	81
74	59	76	65	69
54	56	69	68	78
65	85	49	69	61
48	81	68	37	43
78	82	43	64	67
52	56	81	77	79
85	40	85	59	80
60	71	57	61	69
61	83	90	87	74

Figure 2.2-1: A Concrete mixer.

2.2.1.2 Frequency Distribution

- Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a ***frequency distribution***.
- A frequency distribution consists of classes and their corresponding tallies and frequencies.
- For concrete mixer example, a frequency distribution is presented in ***Table 2.2-2***.
- Left-hand Side Convention
    - It is important to note that the left-hand side convention has been adopted in the class definition and in the counting of tallies and frequencies.
    - According to this convention, the class would be closed on the left side but open on the right. In mathematical terms, the class would be:
      [ )
    - Then, if a value located exactly on the boundary on the intervals, it would be included with the upper interval.

Table 2.2-2: Frequency distribution for working age concrete mixers



Class limits	Tally	Frequency
35–41	///	3
42–48	///	3
49–55	////	4
56–62	𝍸 𝍸	10
63–69	𝍸 𝍸	10
70–76	𝍸	5
77–83	𝍸 𝍸	10
84–90	𝍸	5
	Total	50

2.2.1.3 Definition of Frequency Distribution

Frequency distribution can be defined as the organization of raw data in table form, using classes and frequencies.

2.2.1.4 Benefits of Frequency Distribution

- Some general observations can be made from looking at the frequency distribution.
- For example, it can be stated that the majority of the portable mixer in the study are over 55-month-old.

2.2.1.5 Types of Frequency Distributions

Two types of frequency distributions that are most often used are:

- The categorical frequency distribution:
- It is usually related to qualitative data.
- The grouped frequency distribution:
- It is usually related to quantitative data.

2.2.2 CATEGORICAL FREQUENCY DISTRIBUTIONS

The ***categorical frequency distribution*** is used for ***qualitative data*** that can be placed in specific categories, such as political affiliation, or major field of study.

Example 2.2-1

To select the best source for structural steel, a construction company has asked 80 of its engineers to indicate their preference from among the following four steel companies: ***Company 1***, ***Company 2***, ***Company 3***, and ***Company 4***. ***Table 2.2-3*** below shows the data, consisting of the 80 choices made. Reduce data in terms of a frequency distribution.

Table 2.2-3: Eighty choices for preference between four companies of Example 2.2-1.

Company 3	Company 4	Company 2	Company 3	Company 2
Company 2	Company 4	Company 4	Company 1	Company 2
Company 4	Company 1	Company 4	Company 4	Company 4
Company 1	Company 3	Company 3	Company 4	Company 2
Company 4	Company 2	Company 4	Company 1	Company 4
Company 1	Company 1	Company 2	Company 2	Company 4
Company 2	Company 2	Company 4	Company 2	Company 2
Company 4	Company 4	Company 3	Company 4	Company 3
Company 2	Company 1	Company 3	Company 4	Company 2
Company 3	Company 2	Company 4	Company 4	Company 3
Company 4	Company 4	Company 4	Company 1	Company 4
Company 1	Company 1	Company 2	Company 2	Company 2
Company 3	Company 4	Company 3	Company 2	Company 2
Company 4	Company 4	Company 1	Company 3	Company 4
Company 1	Company 1	Company 2	Company 1	Company 4
Company 2	Company 1	Company 1	Company 2	Company 2

Solution

Since the data are categorical, therefore discrete classes can be used. The class width will be predefined based on the categories of the gathered data, the four companies in this example. The procedure for constructing a frequency distribution for categorical data is given next.

Step 1: Make a table as shown.

A	B	C	D
Class	Tally	Frequency	Percent
Company 1			
Company 2			
Company 3			
Company 4			

Step2: Tally the data and place the results in column B.

A	B	C	D
Class	Tally	Frequency	Percent
Company 1	𝍸 𝍸 𝍸 |		
Company 2	𝍸 𝍸 𝍸 𝍸 ||||		
Company 3	𝍸 𝍸 ||		
Company 4	𝍸 𝍸 𝍸 𝍸 𝍸 |||		

Step 3: Count the tallies and place the results in column C.

A	B	C	D
Class	Tally	Frequency	Percent
Company 1	𝍸 𝍸 𝍸 |	16	
Company 2	𝍸 𝍸 𝍸 𝍸 ||||	24	
Company 3	𝍸 𝍸 ||	12	
Company 4	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28	

Step 4 Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \times 100\%$$

where

f is frequency of the class and n is total number of values.

For example, in the class of Company A, the percentage is

$$\% = \frac{16}{16_{Number\ of\ rows\ in\ raw\ data} \times 5_{Number\ of\ columns\ in\ raw\ data}} \times 100\%$$

$$\% = \frac{16}{80} \times 100\% = 20\%$$

A	B	C	D
Class	Tally	Frequency	Percent
Company 1	𝍸 𝍸 𝍸 |	16	20
Company 2	𝍸 𝍸 𝍸 𝍸 ||||	24	30
Company 3	𝍸 𝍸 ||	12	15
Company 4	𝍸 𝍸 𝍸 𝍸 𝍸 |||	28	35

Percentages are not normally part of a frequency distribution, but they can be added since they are used in certain types of graphs. Also, the decimal equivalent of a percent is called a ***relative frequency***.

Step 5: Find the totals for columns C (frequency) and D (percent). The completed table is shown.

A	B	C	D
Class	Tally	Frequency	Percent
Company 1	卌 卌 卌 |	16	20
Company 2	卌 卌 卌 卌 ||||	24	30
Company 3	卌 卌 ||	12	15
Company 4	卌 卌 卌 卌 卌 |||	28	35
	Total	80 $= 16 \times 5\ Ok$	100 Ok.

2.2.3 Grouped Frequency Distributions

- When the range of the data is large, the quantitative data must be grouped into classes that are more than one unit in width, in what is called a ***grouped frequency distribution*.**
- Therefore, a criterion is needed to determine a suitable class number. One of the most common criteria has been presented in Example 2.3-2 of Article 2.3.

Example 2.2-2

Data in ***Table 2.2-4*** represent the record of high temperatures in degrees Fahrenheit (^{0}F) for each of the 50 states. Construct a grouped frequency distribution for the data.

Table 2.2-4: High temperatures in degrees Fahrenheit (^{0}F) for each of the 50 states.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Solution

The procedure for constructing a grouped frequency distribution for numerical data follows.

- Step 1: Determine the classes.
  - Find the highest value and lowest value: H = 134 and L = 100.
  - Find the range:
    $Range\ =\ R\ =\ highest\ value\ -\ lowest\ value\ =\ H\ -\ L$
    So
    $Range\ =\ R\ =\ 134\ -\ 100\ = 34$
  - Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.
  - Find the class width by dividing the range by the number of classes.
    $$Class\ Width\ = \frac{Range}{Number\ of\ Class} = \frac{34}{7} =\ 4.9$$
    Round the answer up to the nearest whole number if there is a remainder: $4.9\ \approx 5$.

- Notes on Class Width:
  - Rounding up is different from rounding off. A number is rounded up if there is any decimal remainder when dividing. For example,
    $$\frac{85}{6} = 14.167$$
    and is rounded up to 15
  - After dividing, if there is no remainder, you will need to add an extra class to accommodate all the data.
- Select a starting point for the lowest class limit.
  This can be the smallest data value or any convenient number less than the smallest data value. In this case, 100 is used.
- Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 7 classes, as shown, 100, 105, 110, etc.

  Class Limits
  100-105
  105-110
  110-115
  115-120
  120-125
  125-130
  130-135

- Step 2: Tally the data.

Class Limits	**Tally**
100-105	//
105-110	𝍸 ///
110-115	𝍸 𝍸 𝍸 ///
115-120	𝍸 𝍸 ///
120-125	𝍸 //
125-130	/
130-135	/

- Step 3 Find the numerical frequencies from the tallies.

Class Limits	**Tally**	Frequency
100-105	//	2
105-110	𝍸 ///	8
110-115	𝍸 𝍸 𝍸 ///	18
115-120	𝍸 𝍸 ///	13
120-125	𝍸 //	7
125-130	/	1
130-135	/	1

2.2.4 Cumulative Frequency Distribution

- Sometimes it is necessary to use a cumulative frequency distribution.
- A ***cumulative frequency distribution*** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary). The values are found by adding the frequencies of the classes less than or equal to the upper-class boundary of a specific class. This gives an ascending cumulative frequency.
- In this example, the cumulative frequency for:
  - The first class is
    $0 + 2 = 2$;
  - For the second class it is
    $0 + 2 + 8 = 10$;
  - For the third class it is
    $0 + 2 + 8 + 18 = 28$.
- Naturally, a shorter way to do this would be to just add the cumulative frequency of the class below to the frequency of the given class.

	Cumulative Frequency
Less than 105	2
Less than 110	8+2=10
Less than 115	18+10=28
Less than 120	13+28=41
Less than 125	7+41=48
Less than 130	1+48=49
Less than 135	1+49=50

Example 2.2-3

Data in ***Table 2.2-5*** below represent wind speed in *mph* at a specific site. Summarized these data in terms of a frequency distribution.

Table 2.2-5: Wind speed in *mph.*

86	133	110	151	132
137	147	165	152	156
128	154	169	162	116
137	168	147	139	119
118	126	143	135	151

In your solution, use the formula below to estimate number of intervals.

$k = 1 + 3.3 \log n$

This empirical relation usually used to determine the number of interval would be discussed in some details in ***Example 2.3-3*** below.

Solution

$n = 5 \times 5 = 25$

$k = 1 + 3.3 \times \log 25 = 5.6$

Try 6 classes

$Maximum\ wind\ speed = 169\ mph$

$Minimum\ wind\ speed = 86\ mph$

$Range\ of\ wind\ speed = 169 - 86 = 83\ mph$

$Class\ width = \frac{83}{6} = 13.8$

Use 6 classes with width of 14 mph for each class.

Class Limits		Frequency
86	100	1
100	114	1
114	128	4
128	142	7
142	156	7
156	170	5

Example 2.2-4

Data in ***Table 2.2-6*** below represent live loads act on a building floor. Summarize data in term of a frequency distribution and cumulative frequency distribution with number of intervals, k, computed based on following relation:

$k = 1 + 3.3 \log n$

where n is number of data.

Table 2.2-6 Floor live loads, in psf, for Example 2.2-4.

79	76	68	70	90
83	88	56	75	89
80	80	87	103	68
84	80	90	73	97
81	83	84	75	84

Solution

Data Summarization:

Data size, n, is:

$n = 5 \times 5 = 25$

Number of classes would be:

$k = 1 + 3.3 \log n = 1 + 3.3 \times \log(25) = 5.61$

Try 6 classes.

$Maximum\ Value = 103\ psf$

$Minimum\ Value = 56\ psf$

$Range = 103 - 56 = 47$

Therefore, class width would be:

$$Class\ width = \frac{47}{6} = 7.833$$

***Use 6 classes each is 8 psf in width*.**

Frequency distribution is:

Classes		**Frequency**
56	64	1
64	72	3
72	80	5
80	88	10
88	96	4
96	104	2
Summation		25 Okay

While cumulative frequency distribution is presented in table below.

		Cumulative Frequency
Less than	64	1
Less than	72	4
Less than	80	9
Less than	88	19
Less than	96	23
Less than	104	25

2.2.5 UNGROUPED FREQUENCY DISTRIBUTION

- When the range of the data values is relatively small, less than 25 according to (Kottegoda & Rosso, 2008), a frequency distribution can be constructed using single data values for each class.
- This type of distribution is called an **ungrouped frequency distribution** and is shown next.
- In this course, data are generally treated as grouped data to practice students. In their future works when they deal with huge data, they can apply the recommendation of (Kottegoda & Rosso, 2008).

Example 2.2-5

The data shown here represent the compressive strength of concrete cylinder samples in MPa (N/mm^2). Construct a frequency distribution and analyze the distribution.

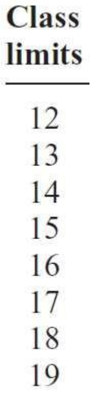
Table 2.2-7: Compressive strength of concrete cylinder samples in MPa.

12	17	12	14	16	18
16	18	12	16	17	15
15	16	12	15	16	16
12	14	15	12	15	15
19	13	16	18	16	14

Figure 2.2-2: Compression test of concrete cylinder.

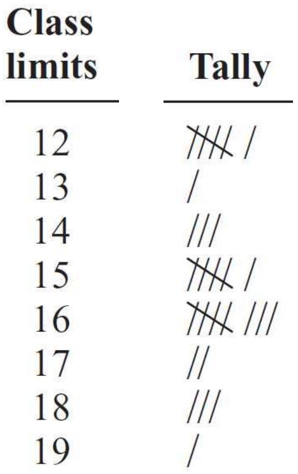
Solution

- Step 1: Determine the classes.
  Since the range of the data set is small
  19 - 12 = 7
  Then classes consisting of a single data value can be used. They are:



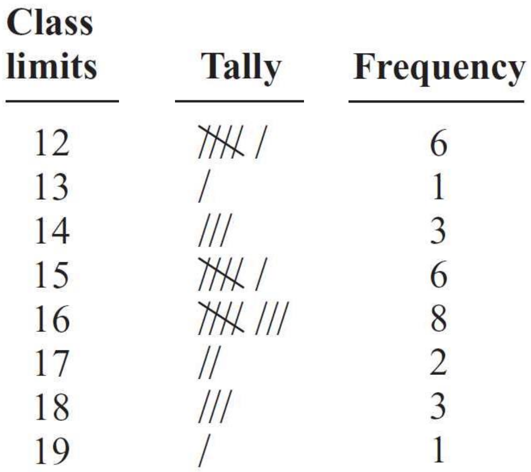
Class limits
12
13
14
15
16
17
18
19

- Step 2 Tally the data.



Class limits	Tally
12	𝍸 /
13	/
14	///
15	𝍸 /
16	𝍸 ///
17	//
18	///
19	/

- Step 3: Find the numerical frequencies from the tallies, and find the cumulative frequencies.



Class limits	Tally	Frequency
12	𝍸 /	6
13	/	1
14	///	3
15	𝍸 /	6
16	𝍸 ///	8
17	//	2
18	///	3
19	/	1

	Cumulative Frequency
Less than or equal to 12	6
Less than or equal to 13	1+6=7
Less than or equal to 14	3+7=10
Less than or equal to 15	6+10=16
Less than or equal to 16	8+16=24
Less than or equal to 17	2+24=26
Less than or equal to 18	3+26=29
Less than or equal to 19	1+29=30

2.2.6 Problem

Problem 2.2-1

Name the three types of frequency distributions and explain when each should be used.

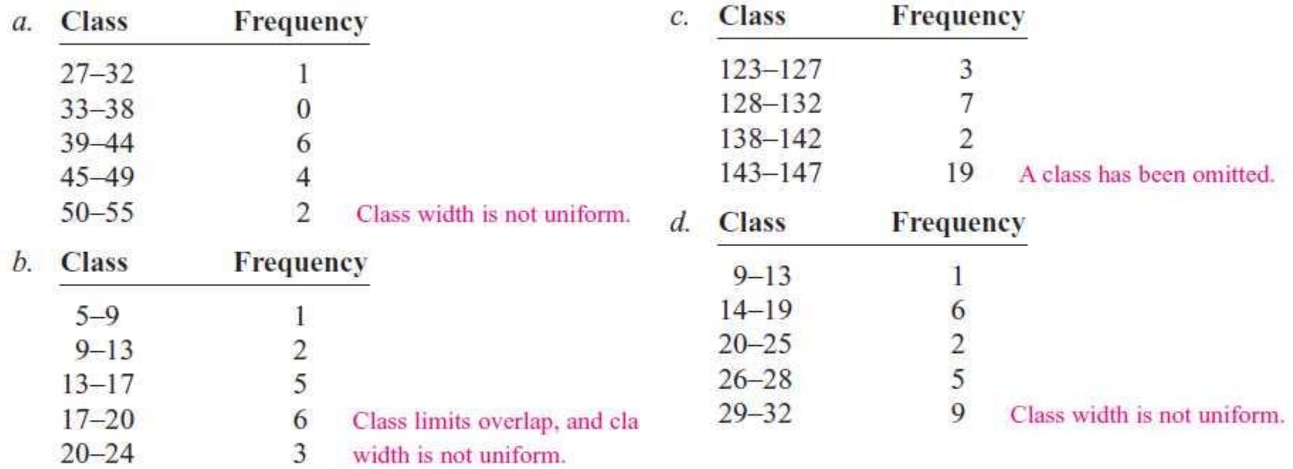
Solution

In general, frequency distributions classified into:

- Categorical, usually adopted for qualitative data.
- Grouped, usually for large number, greater than 25, of quantitative data.
- Ungrouped, usually for small number, less than 25, of quantitative data.

Problem 2.2-2

Shown here are four frequency distributions. Each is incorrectly constructed. State the reason why. (Answers are indicated in the red font).



a.

Class	Frequency
27–32	1
33–38	0
39–44	6
45–49	4
50–55	2

Class width is not uniform.

b.

Class	Frequency
5–9	1
9–13	2
13–17	5
17–20	6
20–24	3

Class limits overlap, and cla width is not uniform.

c.

Class	Frequency
123–127	3
128–132	7
138–142	2
143–147	19

A class has been omitted.

d.

Class	Frequency
9–13	1
14–19	6
20–25	2
26–28	5
29–32	9

Class width is not uniform.

Problem 2.2-3

What are open-ended frequency distributions? Why are they necessary?

Solution

- An open-ended frequency distribution has either a first class with no lower limit or a last class with no upper limit.
- They are necessary to accommodate all the data.

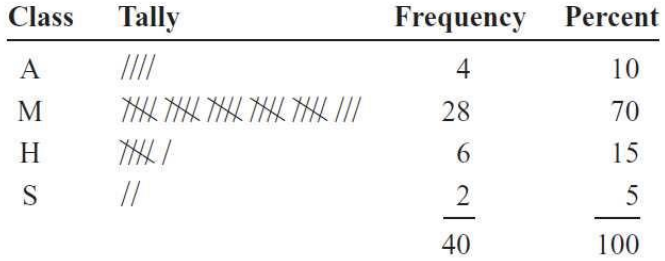
Problem 2.2-4

A survey was taken on how much trust people place in the information they read on the Internet. Construct a categorical frequency distribution for the data.

A = trust in everything they read,
M = trust in most of what they read,
H = trust in about one-half of what they read,
S = trust in a small portion of what they read.

M	M	M	A	H	M	S	M	H	M
S	M	M	M	M	A	M	M	A	M
M	M	H	M	M	M	H	M	H	M
A	M	M	M	H	M	M	M	M	M

Answer



Class	Tally	Frequency	Percent
A	////	4	10
M	𝍸 𝍸 𝍸 𝍸 𝍸 ///	28	70
H	𝍸 /	6	15
S	//	2	5
		40	100

2.3 HISTOGRAMS, FREQUENCY POLYGONS, AND OGIVES

- After you have organized the data into a frequency distribution, you can present them in graphical form.
- The purpose of graphs in descriptive statistics is to:
  - Convey the data to the viewers in pictorial form.
  - Summarize a data.
  - Ease data presentation as most people are used to comprehend the meaning of data presented graphically better than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.
  - Get the audience's attention in a publication or a speaking presentation.
- The purpose of graphs in inferential statistics is to:
  - Discover a trend or pattern in the data.
  - Analyze the data.
- The three most commonly used graphs in research are
  - The histogram.
  - The frequency polygon.
  - The cumulative frequency graph, or ogive (pronounced o-jive).

2.3.1 THE HISTOGRAM

Example 2.3-1

Present frequency distribution of ***Example 2.2-2***, ***Temperature Example,*** in terms of a ***Histogram***.

Solution

For convenience, the frequency distribution of ***Example 2.2-2*** is represented here.

Classes	Frequency
100-105	2
105-110	8
110-115	18
115-120	13
120-125	7
125-130	1
130-135	1

As indicated in ***Figure 2.3-1*** below, in a histogram the abscissa, horizontal axis, represents classes with their units while ordinate, vertical axis, represents frequency.

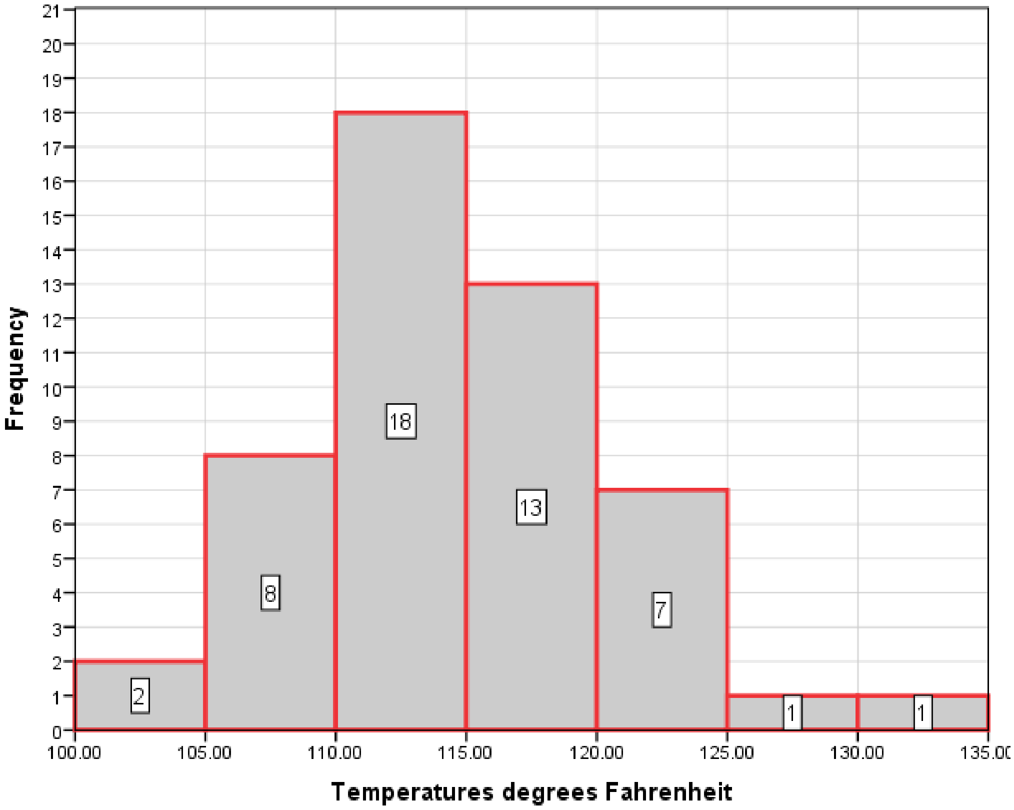
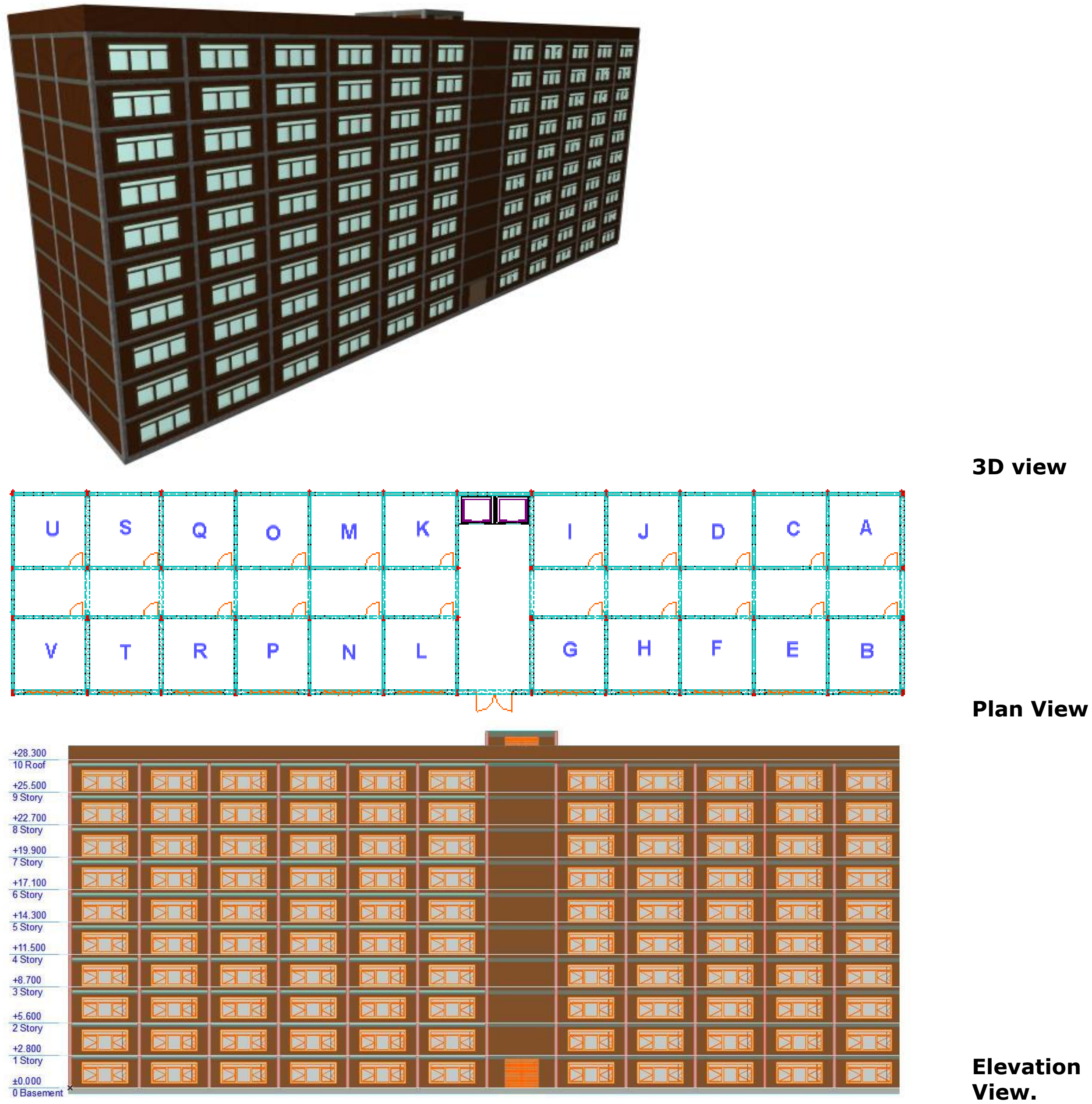


Figure 2.3-1: Histogram for Example 2.3-1.

Example 2.3-2

Numbers in ***Table 2.3-1*** below represent the live loads observed in a warehouse building indicated in ***Figure 2.3-2***. In his work, the engineer should understand the nature of load distribution. Put these data in a histogram form.



3D view

Plan View

Elevation View.

Figure 2.3-2: Warehouse building for Example 2.3-2.

Table 2.3-1: Live load for Example 2.3-2.

Bay	***Basement***	***1st***	***2d***	***3d***	***4th***	***5th***	***6th***	***7th***	***8th***	***9th***
A	0	7.8	36.2	60.6	64.0	64.2	79.2	88.4	38.0	72.7
B	72.2	72.6	74.4	21.8	17.1	48.5	16.8	105.9	57.2	75.7
C	225.7	42.5	59.8	41.7	39.9	55.5	67.2	122.8	45.2	62.9
D	55.1	55.9	87.7	59.2	63.1	58.8	67.7	90.4	43.3	55.2
E	36.6	26.0	90.5	23.0	43.5	52.1	102.1	71.7	4.1	37.3
F	129.4	66.4	138.7	127.9	90.9	46.9	197.5	151.1	157.3	197.0
G	134.6	73.4	80.9	53.3	80.1	62.9	150.8	102.2	6.4	45.4
H	121.0	106.2	94.4	139.6	152.5	70.2	111.8	174.1	85.4	83.0
I	178.8	30.2	44.1	157.0	105.3	87.0	50.1	198.0	86.7	64.6
J	78.6	37.0	70.7	83.0	179.7	180.2	60.6	212.4	72.2	86.0
K	94.5	24.1	87.3	80.6	74.8	72.4	131.1	116.1	53.6	99.1
L	40.2	23.4	8.4	42.6	43.4	27.4	63.8	18.4	16.2	58.7
M	92.2	49.8	50.9	116.4	122.9	132.3	105.2	160.3	28.7	46.8
N	99.5	106.9	55.9	136.8	110.4	123.5	92.4	160.9	45.4	96.3
0	88.5	48.4	62.3	71.3	133.2	92.1	111.7	67.9	53.1	39.7
P	93.2	55.0	80.8	143.5	122.3	184.2	150.0	57.6	6.8	53.3
Q	96.1	54.8	63.0	228.3	139.3	59.1	112.1	50.9	158.6	139.1

Bay	***Basement***	***1st***	***2d***	***3d***	***4th***	***5th***	***6th***	***7th***	***8th***	***9th***
R	213.7	65.7	90.3	198.4	97.5	155.1	163.4	155.3	229.5	75.0
S	137.6	62.5	156.5	154.1	134.3	81.6	194.4	155.1	89.3	73.4
T	79.8	68.7	85.6	141.6	100.7	106.0	131.1	157.4	80.2	65.0
U	78.5	118.2	126.4	33.8	124.6	78.9	146.0	100.3	97.8	75.3
V	24.8	55.6	135.6	56.3	66.9	72.2	105.4	98.9	101.7	58.2

Solution

- Let us divide this range into 20-psf intervals, 0 to 20, 20.0 to 40, etc., and tally the number of occurrences in each interval.

Classes		Frequency
0	20	10
20	40	17
40	60	41
60	80	42
80	100	35
100	120	19
120	140	22
140	160	16
160	180	6
180	200	7
200	220	2
220	240	3

- Plotting the frequency of occurrences in each interval as a bar yields a histogram, as shown in ***Figure 2.3-3***.

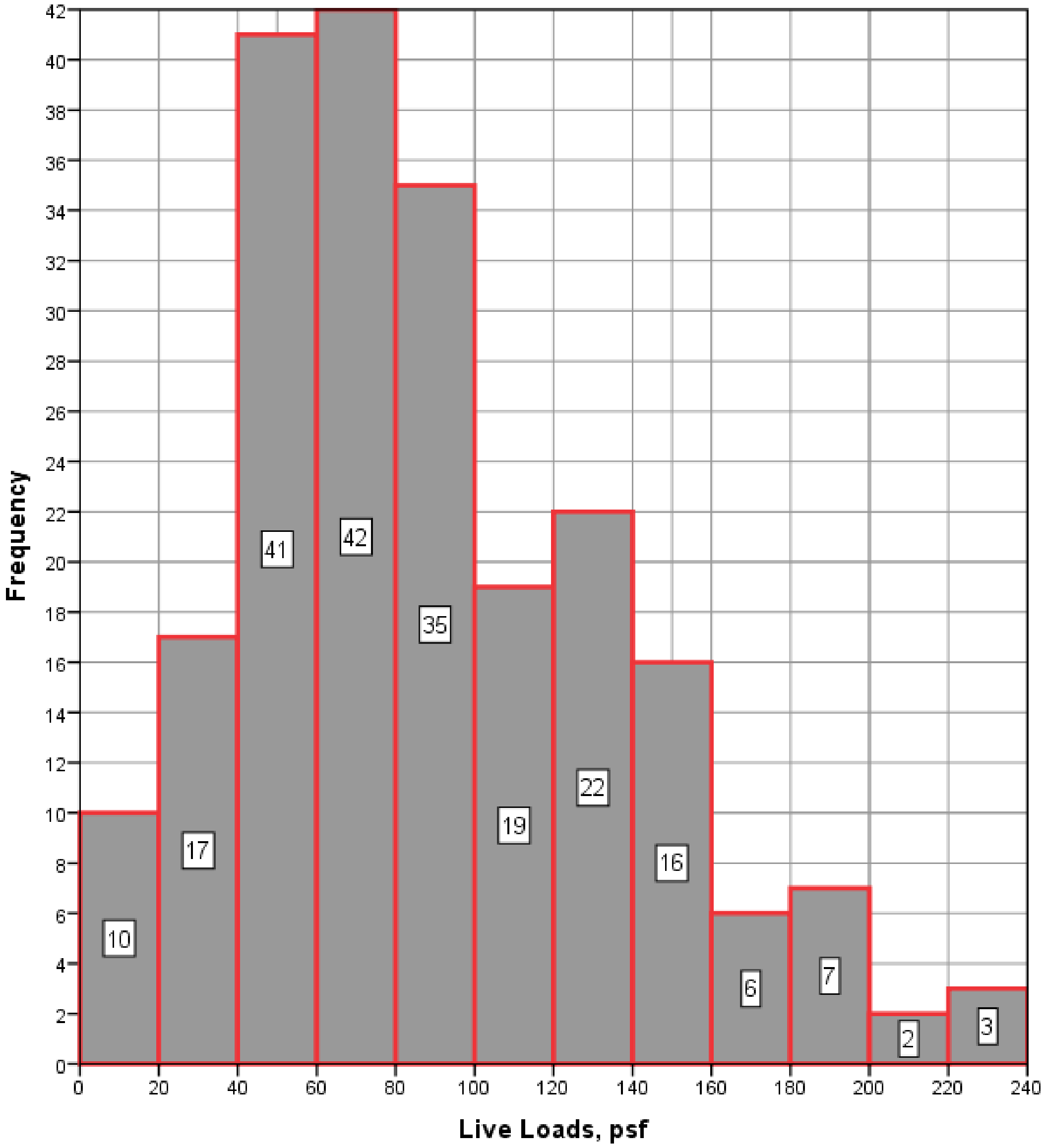


Figure 2.3-3: Histogram for live load of Example 2.3-2 with 20-psf intervals.

Example 2.3-3

Re-draw the histogram for live loads of the previous example with using of 10 psf interval firstly and 50 psf secondly. Compare between the two graphs.

Solution

Adopting the same procedures of the examples above, the histogram with an interval of 10 psf is shown in ***Figure 2.3-4*** below. While with an interval of 50 psf, the histogram is presented in ***Figure 2.3-5*** below.

Comparison between the two histograms:

- Unfortunately, if the number of values is small, the choice of the precise point at which the interval divisions are to occur also may alter significantly the appearance of the histogram.
- Such variations in shape are indicative of a failure of the set of data to display any sharply defined features, a piece of information which is valuable to the engineer.
- An ***empirical practical guide*** has been suggested. If the number of data values is n, then the number of intervals, k, between the minimum and maximum value observed should be about:

  $k = 1 + 3.3 \log n$ **Eq. 2.3-1**
- This relation has been drawn in ***Figure 2.3-6*** below where it can be noted that the k asymptotes a constant value, in the range of 12, for large sample sizes.

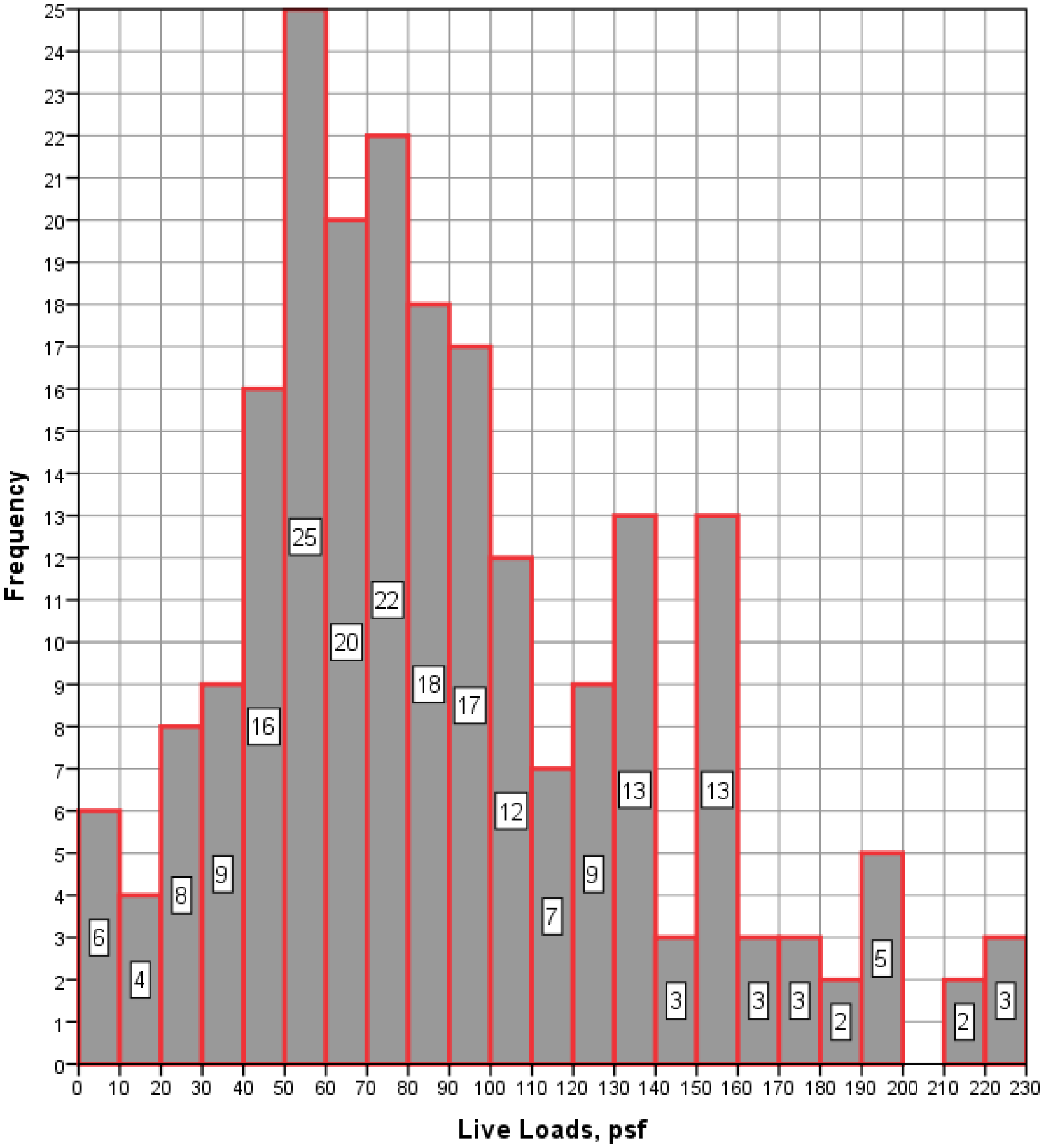


Figure 2.3-4: Histogram for live load of Example 2.3-2 with 10-psf intervals.

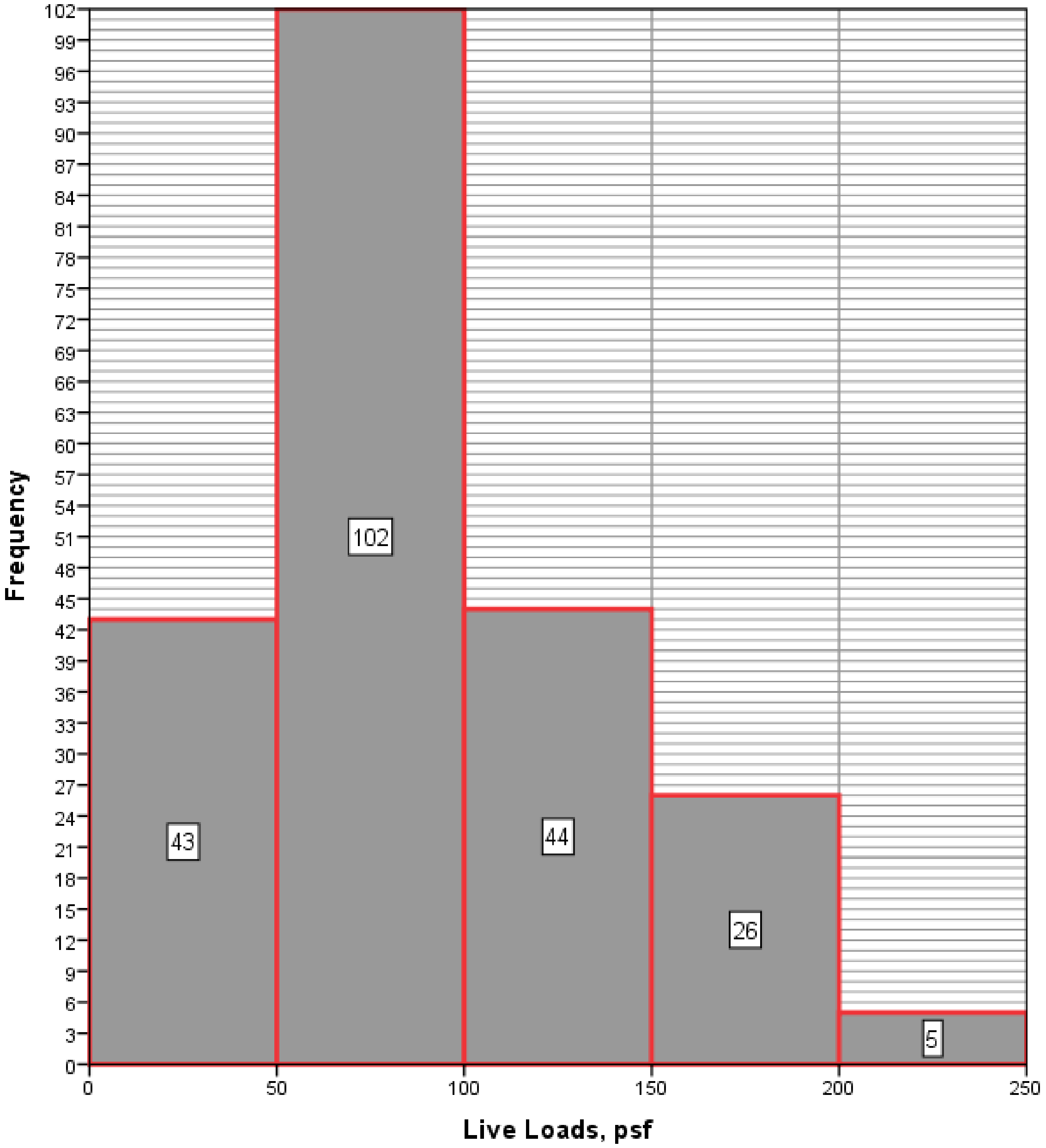


Figure 2.3-5: Histogram for live load of Example 2.3-2 with 50-psf intervals.

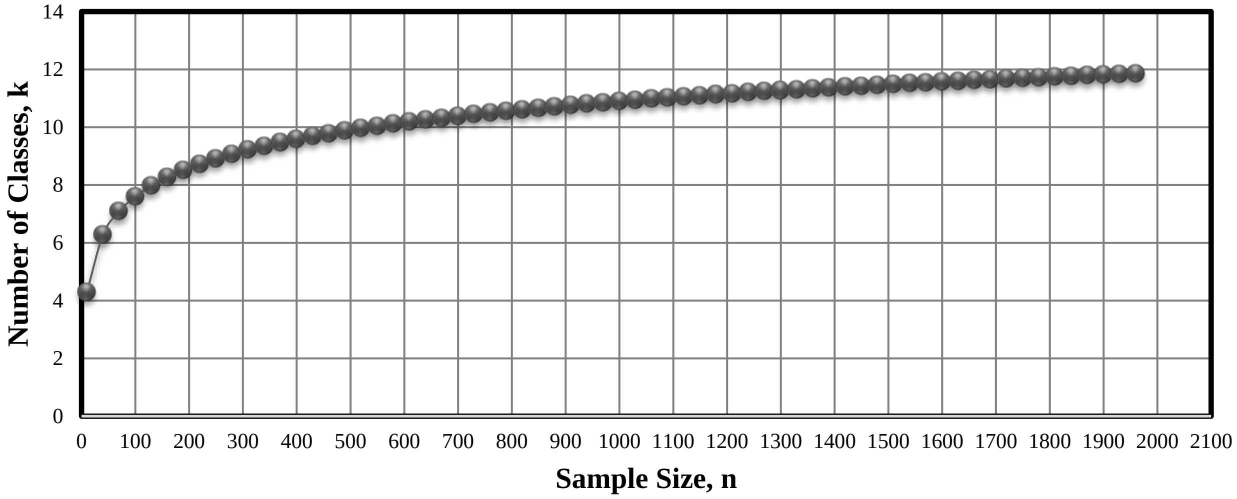


Figure 2.3-6: Number of classes versus sample size according to Eq. 2.3-1.

- If Eq. 2.3-1 is adopted, the most suitable class number would be:
  $k = 1 + 3.3 \times \log(10 \times 22) = 8.7$
  Then about 9 intervals should be adopted:
  $$Class\ width = \frac{Maximum - Minimum}{9} = \frac{229.50 - 0.00}{9} = 25.5$$

Using 10 classes with width of 25lb, histogram would be as indicated in ***Figure 2.3-7***.

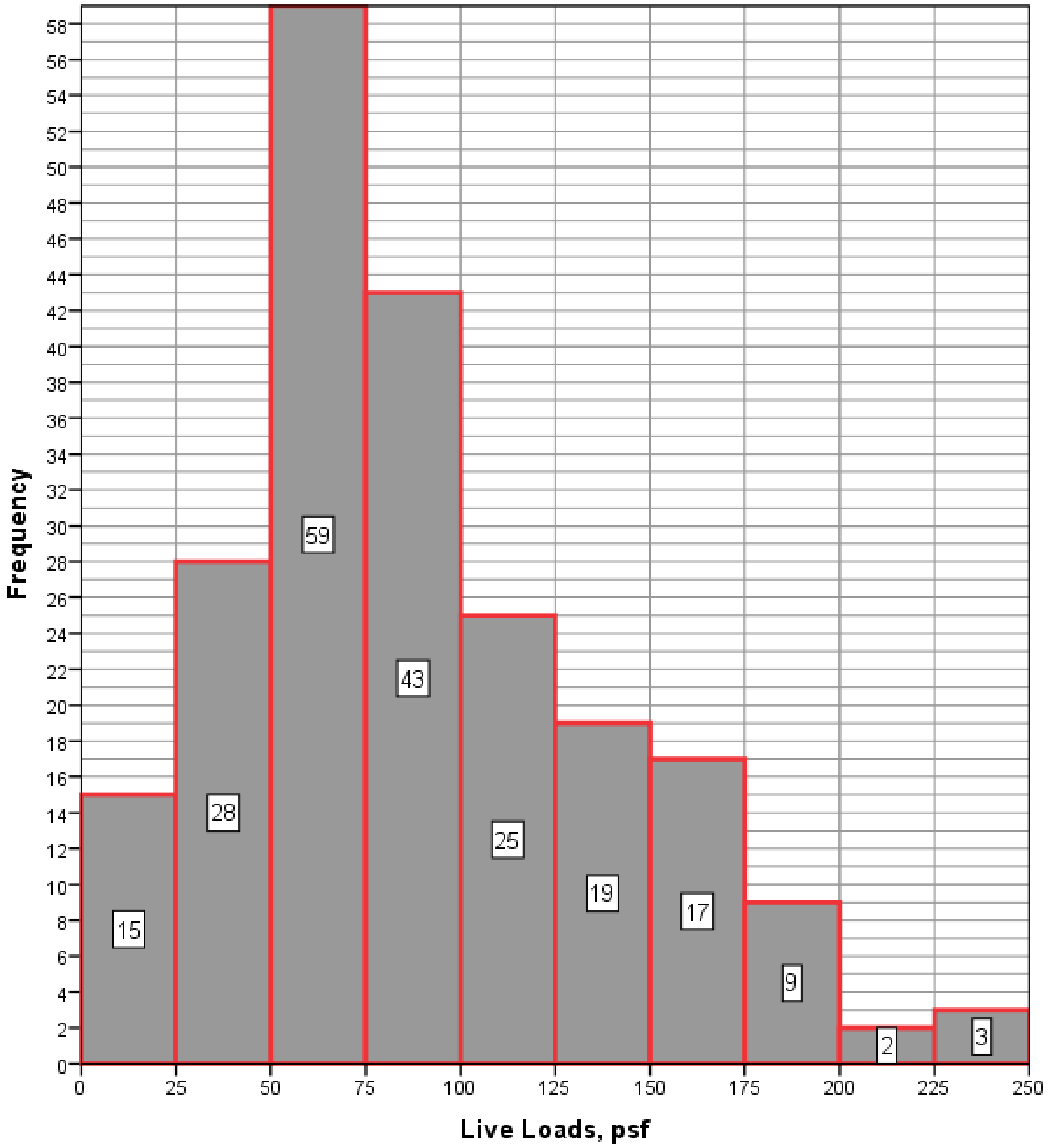


Figure 2.3-7: Histogram for live load of Example 2.3-2 with 25-psf intervals.

Example 2.3-4

Draw a histogram for data shown in ***Table 2.3-2*** below that represents timber strength expressed in terms of modulus of rupture in MPa. The physical or mechanical meaning of modulus of rupture is presented in ***Figure 2.3-8*** below.

Table 2.3-2: Timber strength, modulus of rapture in MPa.

0.00	28.00	31.60	34.44	36.84	39.21	41.75	44.30	47.25	53.99
17.98	28.13	32.02	34.49	36.85	39.33	41.78	44.36	47.42	54.04
22.67	28.46	32.03	34.56	36.88	39.34	41.85	44.36	47.61	54.71
22.74	28.69	32.40	34.63	36.92	39.60	42.31	44.51	47.74	55.23
22.75	28.71	32.48	35.03	37.51	39.62	42.47	44.54	47.83	56.60
23.14	28.76	32.68	35.17	37.65	39.77	43.07	44.59	48.37	56.80
23.16	28.83	32.76	35.30	37.69	39.93	43.12	44.78	48.39	57.99
23.19	28.97	33.06	35.43	37.78	39.97	43.26	44.78	48.78	58.34
24.09	28.98	33.14	35.58	38.00	40.20	43.33	45.19	49.57	65.35
24.25	29.11	33.18	35.67	38.05	40.27	43.33	45.54	49.59	65.61
24.84	29.90	33.19	35.88	38.16	40.39	43.41	45.92	49.65	69.07
25.39	29.93	33.47	35.89	38.64	40.53	43.48	45.97	50.91	70.22
25.98	30.02	33.61	36.00	38.71	40.71	43.48	46.01	50.98	
26.63	30.05	33.71	36.38	38.81	40.85	43.64	46.33	51.39	
27.31	30.33	33.92	36.47	39.05	40.85	43.99	46.50	51.90	
27.90	30.53	34.12	36.53	39.15	41.64	44.00	46.86	53.00	
27.93	31.33	34.40	36.81	39.20	41.72	44.07	46.99	53.63	
n = 165									

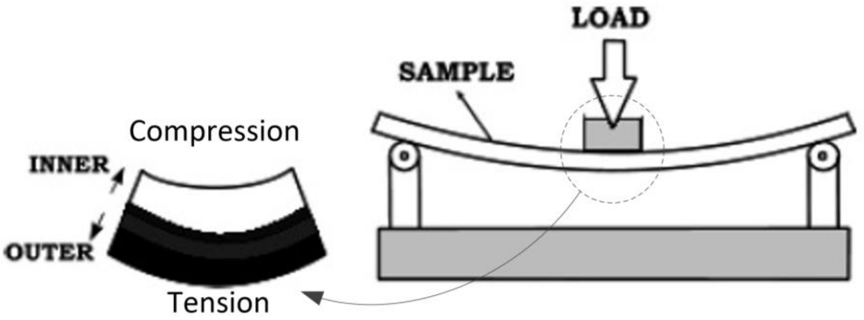


Figure 2.3-8: Test for modulus of rapture for Example 2.3-4.

Solution

Based on Eq. 2.3-1

$k = 1 + 3.3\log n$

The number of classes that should be adopted would be:

$k = 1 + 3.3\log 165 = 8.31$

Say 9 classes. Class width would be:

$$Class\ width = \frac{Maximum - Minimum}{9} = \frac{70.22 - 0.00}{9} = 7.80$$

Then adopt class with width of 8 MPa. Prepare frequency distribution and draw histogram as indicated in ***Table 2.3-3*** and ***Figure 2.3-9*** below.

Table 2.3-3: Frequency distribution table for timber modulus of rupture.

Classes		Frequency
0	8	1
8	16	0
16	24	7
24	32	27
32	40	58
40	48	48
48	56	16
56	64	4
64	72	4

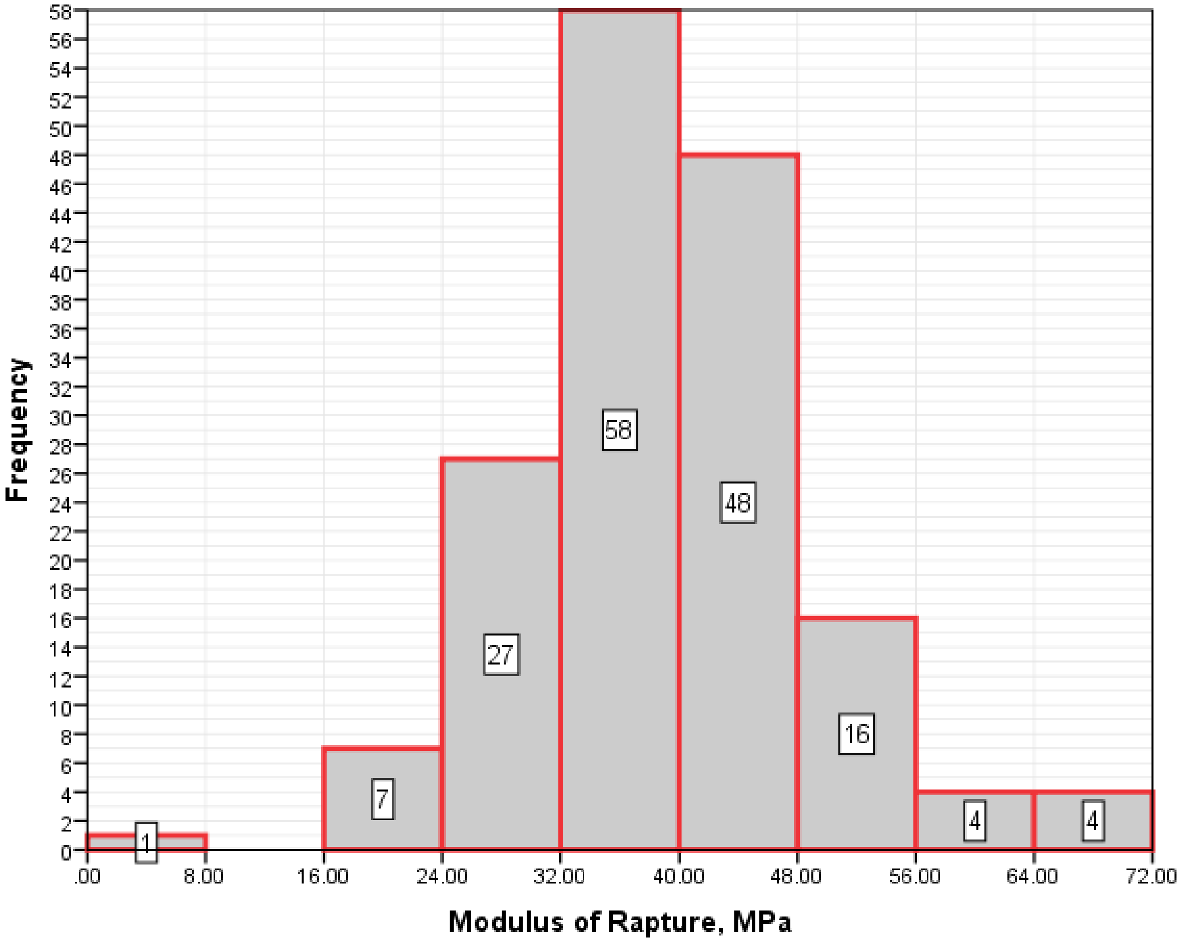


Figure 2.3-9: Histogram for the modulus of rupture.

2.3.2 The Frequency Polygon

The **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

Example 2.3-5

Using the frequency distribution given in ***Example 2.3-1***, ***Temperatures Example***, construct a frequency polygon.

Solution

Find the midpoints of each class with the corresponding frequency:

Table 2.3-4: Frequency distribution table for Example 2.3-5.

Classes		Midpoints	Frequency
100	105	102.5	2
105	110	107.5	8
110	115	112.5	18
115	120	117.5	13
120	125	122.5	7
125	130	127.5	1
130	135	132.5	1

Draw the x and y-axes. Label the x-axis with the midpoint of each class, and then use a suitable scale on the y-axis to draw the frequencies.

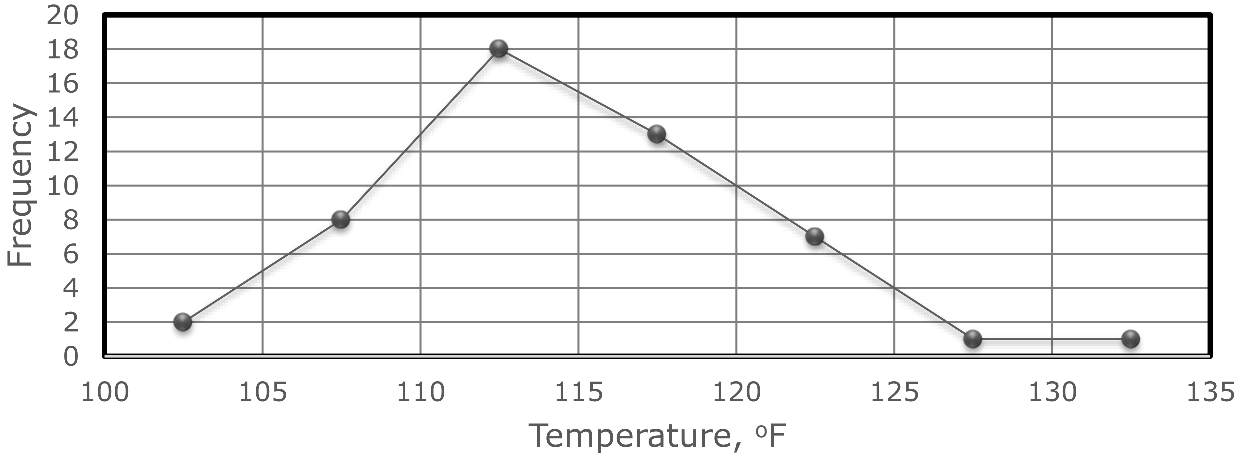


Figure 2.3-10: Frequency polygon for Example 2.3-5.

Example 2.3-6

Using the frequency distribution given in ***Example 2.3-2***, ***Live Loads Example***, to construct a frequency polygon.

Solution

Find the midpoints of each class with the corresponding frequency.

Table 2.3-5: Frequency distribution table for Example 2.3-6.

Classes		Midpoints	Frequency
0	20	10	10
20	40	30	17
40	60	50	41
60	80	70	42
80	100	90	35
100	120	110	19
120	140	130	22
140	160	150	16
160	180	170	6
180	200	190	7
200	220	210	2
220	240	230	3

Draw the x and y-axes. Label the x-axis with the midpoint of each class, and then use a suitable scale on the y-axis to draw the frequencies.

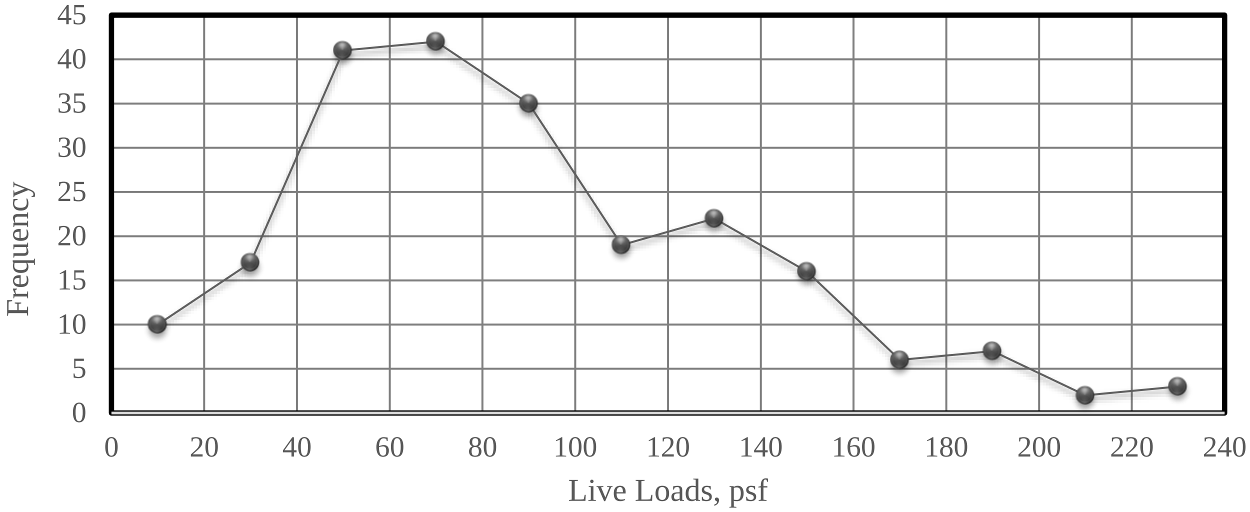


Figure 2.3-11: Frequency polygon for Example 2.3-6.

Example 2.3-7

Using the frequency distribution given in Example 2.3-4 ***Timber Strength Example***, to construct a frequency polygon.

Solution

Find the midpoints of each class with the corresponding frequency:

Table 2.3-6: Frequency distribution table for Example 2.3-7.

Classes		Midpoints	Frequency
0	8	4	1
8	16	12	0
16	24	20	7
24	32	28	27
32	40	36	58
40	48	44	48
48	56	52	16
56	64	60	4
64	72	68	4

Draw the x and y-axes. Label the x-axis with the midpoint of each class, and then use a suitable scale on the y-axis to draw the frequencies.

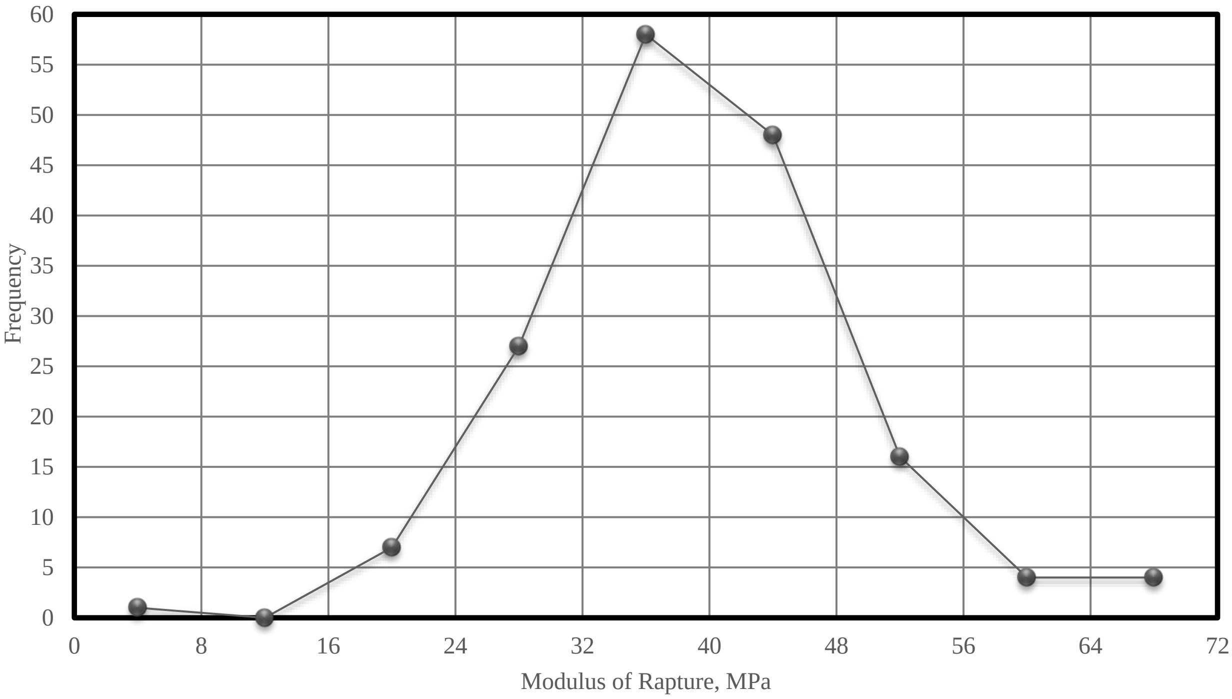


Figure 2.3-12: Frequency polygon for Example 2.3-7.

2.3.3 THE OGIVE

The **ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

Example 2.3-8

Construct an ogive for the frequency distribution described in ***Example 2.3-1***, ***Temperature Example***.

Solution

From the frequency distribution of ***Example 2.3-1*** shown in ***Table 2.3-7***. Find the cumulative frequency for each class as indicated in ***Table 2.3-8***. Plot the cumulative frequency indicated in ***Figure 2.3-13*** where the ***horizontal axis represents the upper-class limit***, while the vertical axis represents cumulative frequency.

Table 2.3-7: Frequency distribution table for Example 2.3-8.

Classes	Frequency
100-105	2
105-110	8
110-115	18
115-120	13
120-125	7
125-130	1
130-135	1

Table 2.3-8: Cumulative Frequency distribution table for Example 2.3-8.

		Cumulative Frequency
Less than	105	2
Less than	110	10
Less than	115	28
Less than	120	41
Less than	125	48
Less than	130	49
Less than	135	50

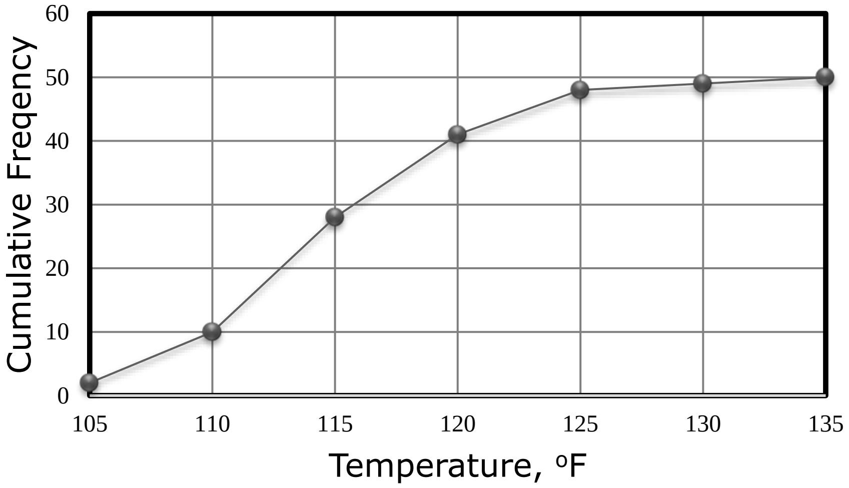


Figure 2.3-13: Cumulative frequency diagram for Example 2.3-8.

Example 2.3-9

Construct an ogive for the frequency distribution described in ***Example 2.3-2***, ***Live Loads Example***.

Solution

From the frequency distribution of ***Example 2.3-2*** shown in ***Table 2.3-9***. Find the cumulative frequency for each class as indicated in ***Table 2.3-10***. Finally, plot the cumulative frequency of ***Figure 2.3-14*** where the horizontal axis represents the upper-class limit, while the vertical axis represents cumulative frequency.

Table 2.3-9: Frequency distribution table for Example 2.3-9.

Classes		Frequency
0	20	10
20	40	17
40	60	41
60	80	42
80	100	35
100	120	19
120	140	22
140	160	16
160	180	6
180	200	7
200	220	2
220	240	3

Table 2.3-10: Cumulative Frequency distribution table for Example 2.3-9.

		Cumulative Frequency
Less than	20	10
Less than	40	27
Less than	60	68
Less than	80	110
Less than	100	145
Less than	120	164
Less than	140	186
Less than	160	202
Less than	180	208
Less than	200	215
Less than	220	217
Less than	240	220

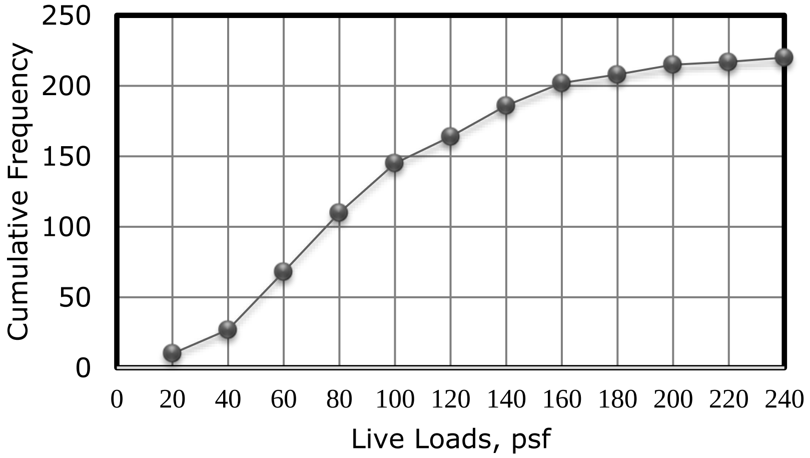


Figure 2.3-14: Cumulative frequency diagram for Example 2.3-9.

Example 2.3-10

Construct an ogive for the frequency distribution described in ***Example 2.3-4 timber strength*** expressed in terms of the modulus of rupture.

Solution

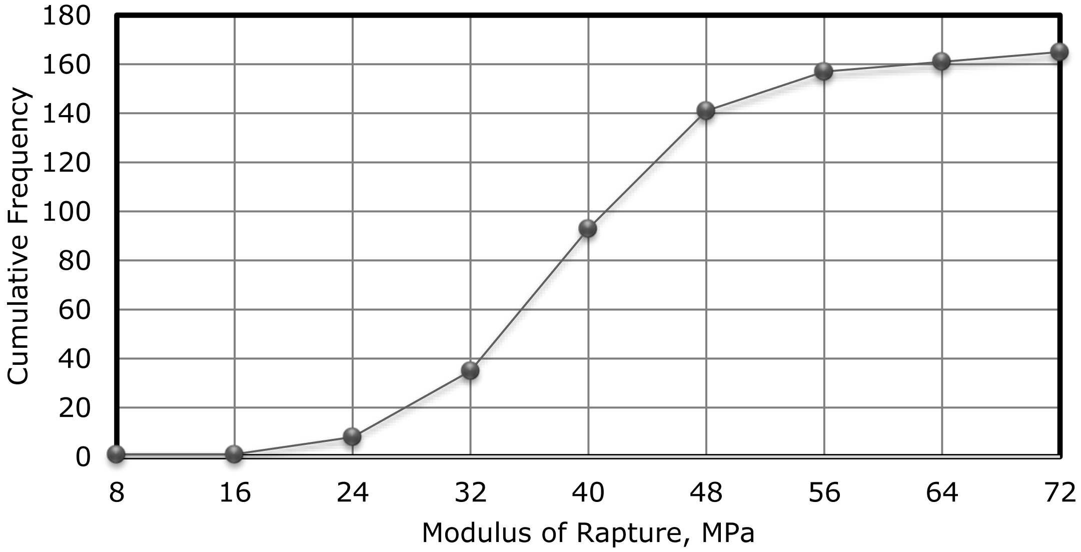
From the frequency distribution of is shown below:

Classes		Frequency
0	8	1
8	16	0
16	24	7
24	32	27
32	40	58
40	48	48
48	56	16
56	64	4
64	72	4

Find the cumulative frequency for each class.

		Cumulative Frequency
Less than	8	1
Less than	16	1
Less than	24	8
Less than	32	35
Less than	40	93
Less than	48	141
Less than	56	157
Less than	64	161
Less than	72	165

Plot the cumulative frequency where the horizontal axis represents the upper-class limit, while the vertical axis represents cumulative frequency.



2.3.4 RELATIVE FREQUENCY GRAPHS

- The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using **proportions** instead of raw data as frequencies. These types of graphs are called **relative frequency graphs**.
- Graphs of relative frequencies instead of frequencies are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class.
- To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of the frequencies.
- The sum of the relative frequencies will always be 1. These graphs are similar to the ones that use raw data as frequencies, but the values on the *y-axis* are in terms of proportions.

Example 2.3-11

Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution (shown here) of the compressive strength for 20 randomly selected concrete cylinder samples.

Table 2.3-11: Frequency table for compressive strength of Example 2.3-11.

Classes	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
Summation	20

Solution

Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.

Classes		Midpoint	Frequency	Relative Frequency
5.5	10.5	8	1	0.05
10.5	15.5	13	2	0.10
15.5	20.5	18	3	0.15
20.5	25.5	23	5	0.25
25.5	30.5	28	4	0.20
30.5	35.5	33	3	0.15
35.5	40.5	38	2	0.10
		Summation	20	1.00, Okay

Find the cumulative relative frequencies.

		Cumulative Frequency	Cumulative Relative Frequency
Less than	10.5	1	0.05
Less than	15.5	3	0.15
Less than	20.5	6	0.30
Less than	25.5	11	0.55
Less than	30.5	15	0.75
Less than	35.5	18	0.90
Less than	40.5	20	1.00

Draw each graph as shown in Figures below. For the histogram and ogive, use the class along the x-axis. For the frequency polygon, use the midpoints on the x-axis. The scale on the y-axis uses proportions.

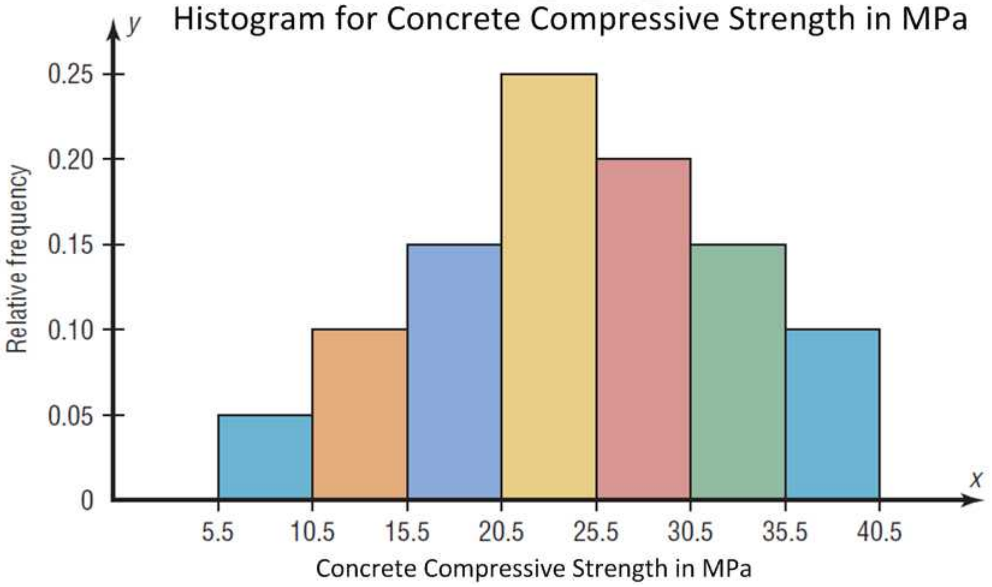


Figure 2.3-15: Histogram for concrete compressive strength of Example 2.3-11.

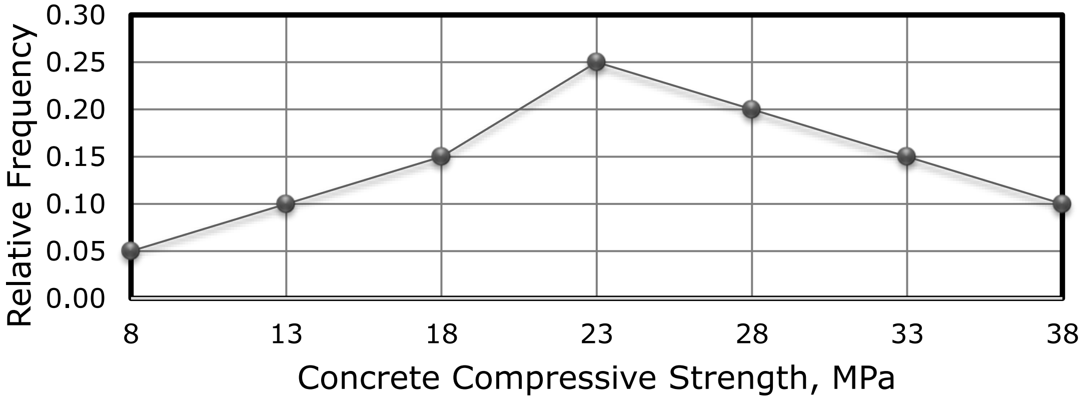


Figure 2.3-16: Frequency polygon for concrete compressive strength of Example 2.3-11.

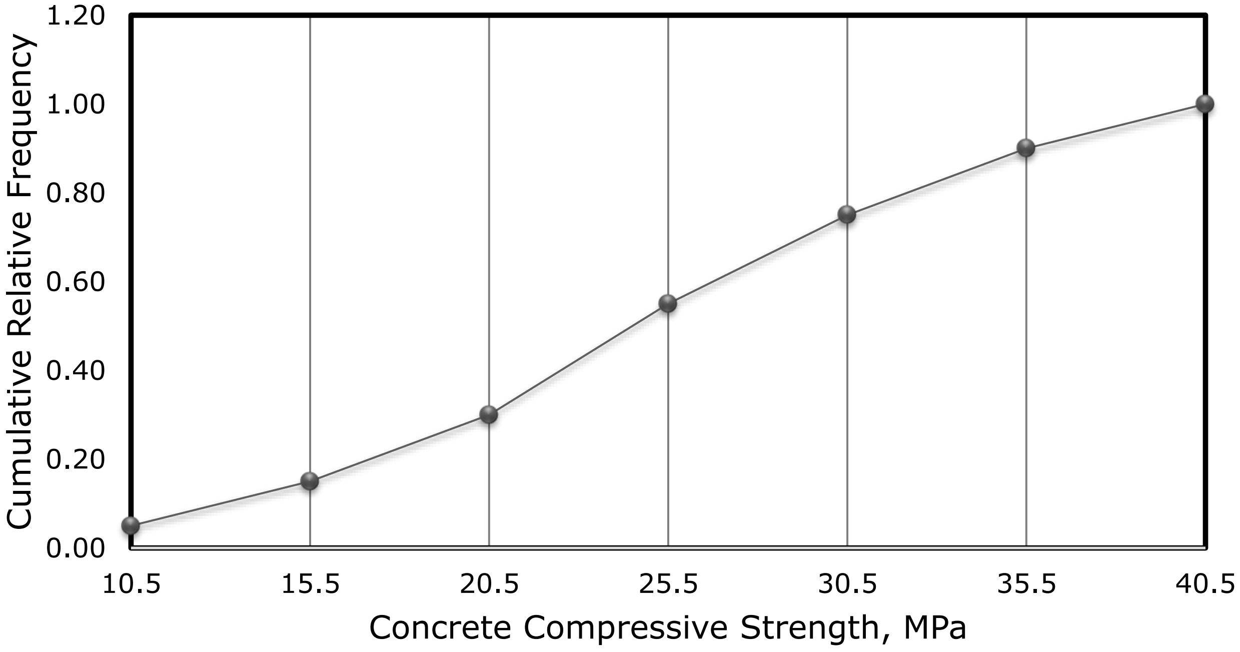


Figure 2.3-17: Cumulative frequency polygon for concrete compressive strength of Example 2.3-11.

2.3.5 Different-width Classes

This article presents some practical aspects when different class widths are predefined by codes and specifications should be adopted in data presentation.

Example 2.3-12

Soil resistivity is used in studies of corrosion of pipes buried in the soil. For example, a resistivity of ohms/cm represents:

- 0 to 400 ***extremely severe corrosion conditions***;
- 400 to 900, ***very severe***;
- 900 to 1500, ***severe***;
- 1500 to 3500, ***moderate***;
- 3500 to 8000, ***mild***;
- 8000 to 20,000, ***slight risk***.

As indicated in ***Table 2.3-12*** below, there were 32 measurements of soil resistivity made at a construction site. Construct a ***frequency distribution***, and ***histogram*** using the classification bands above as intervals.

Table 2.3-12: Soil resistivity in ohms/cm of Example 2.3-12.

1750	1240	1490	2500
960	510	1610	2300
740	910	1110	1240
1030	840	1340	3060
530	1340	2180	1880
1170	1240	1340	6550
5770	1370	1680	1180
2300	1260	1550	2760

Solution

The frequency distribution and the histogram are presented in ***Table 2.3-13*** and ***Figure 2.3-21*** respectively.

Table 2.3-13: Frequency distribution for Example 2.3-12.

Classes		Frequency
0	400	0
400	900	4
900	1500	15
1500	3500	11
3500	8000	2
8000	20000	0

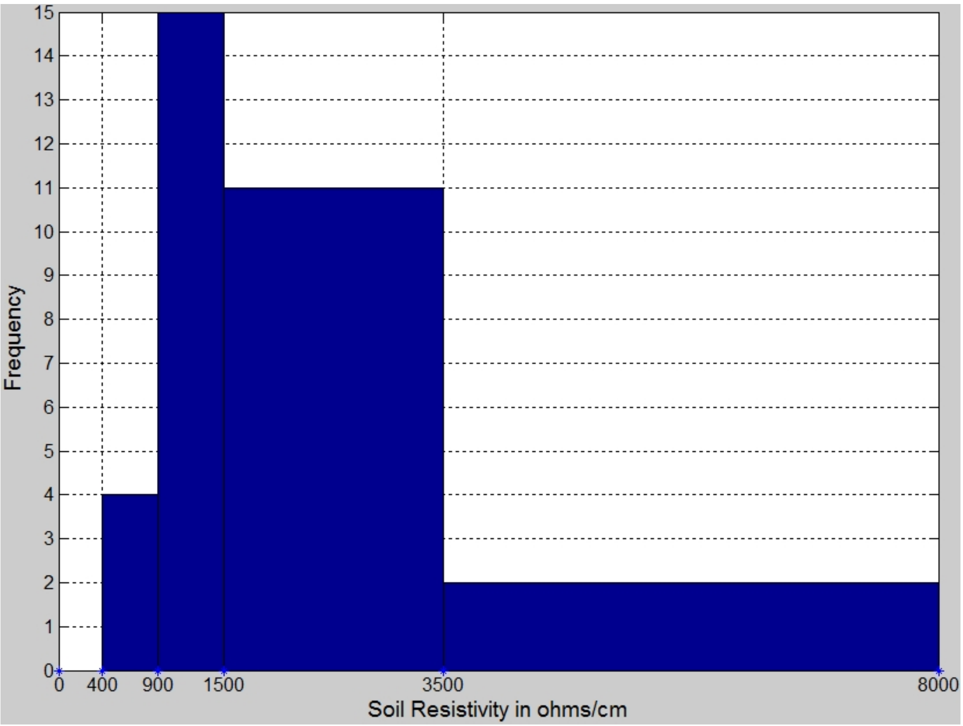


Figure 2.3-18: Histogram for data of Example 2.3-12.

2.3.6 Distribution Shapes

- When one is describing data, it is important to be able to recognize the shapes of the distribution values.
- In later chapters, you will see that the shape of a distribution also determines the appropriate statistical methods used to analyze the data.
- A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution.
- Several of the most common shapes are shown in ***Figure 2.3-19*** below.

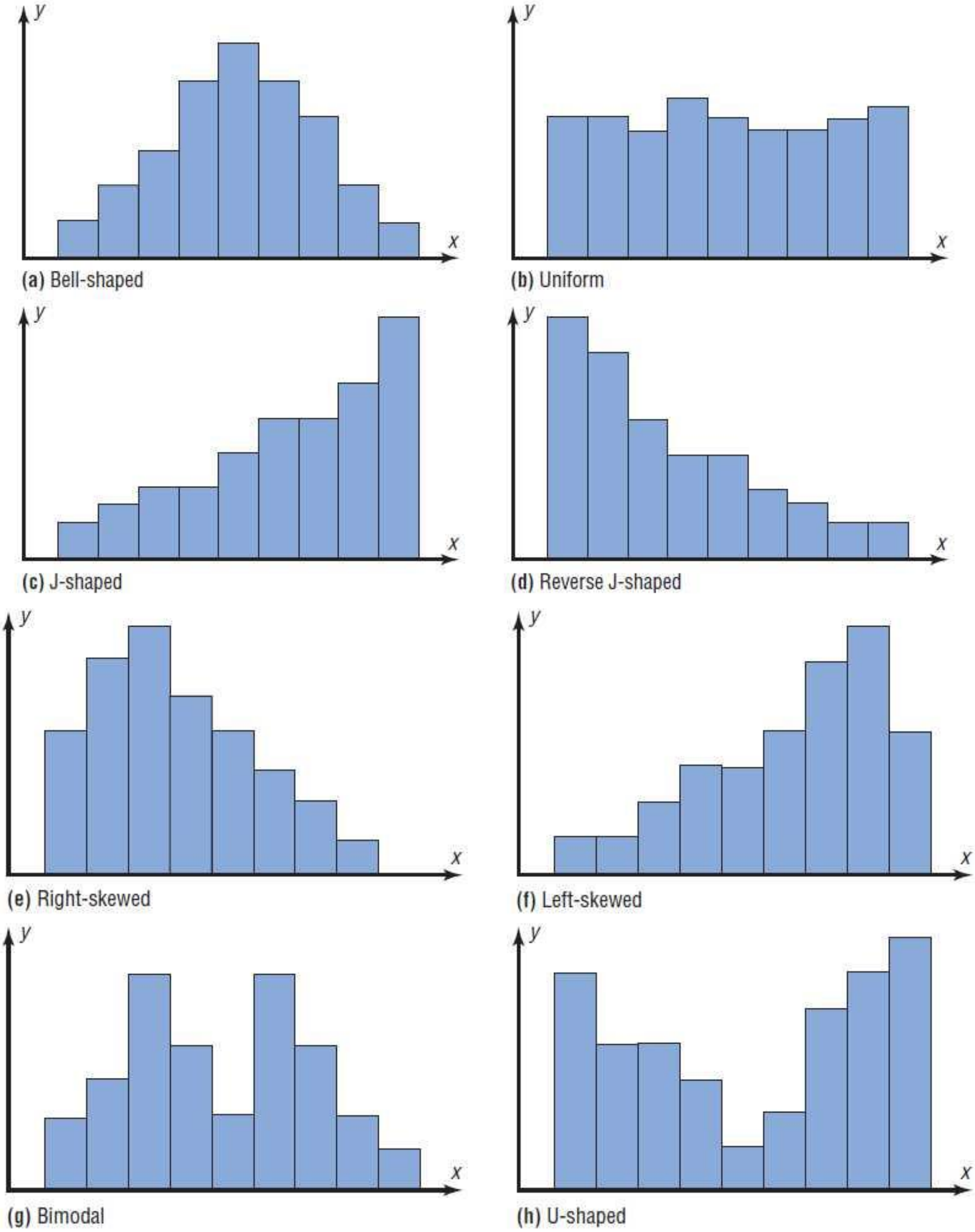


Figure 2.3-19: Most common distribution shapes.

2.3.7 Problems

Problem 2.3-1

Data for yield stresses of reinforcing bars in MPa have been presented in terms of frequency distribution below. Represent these data graphically in terms of histograms, frequency polygon, and ogives.

Table 2.3-14: Frequency table for yield stresses of reinforcing bars in MPa.

Classes		Frequency
350	365	1
365	380	1
380	395	2
395	410	9
410	425	8
425	440	8
440	455	9
455	470	1
470	485	1

Answers

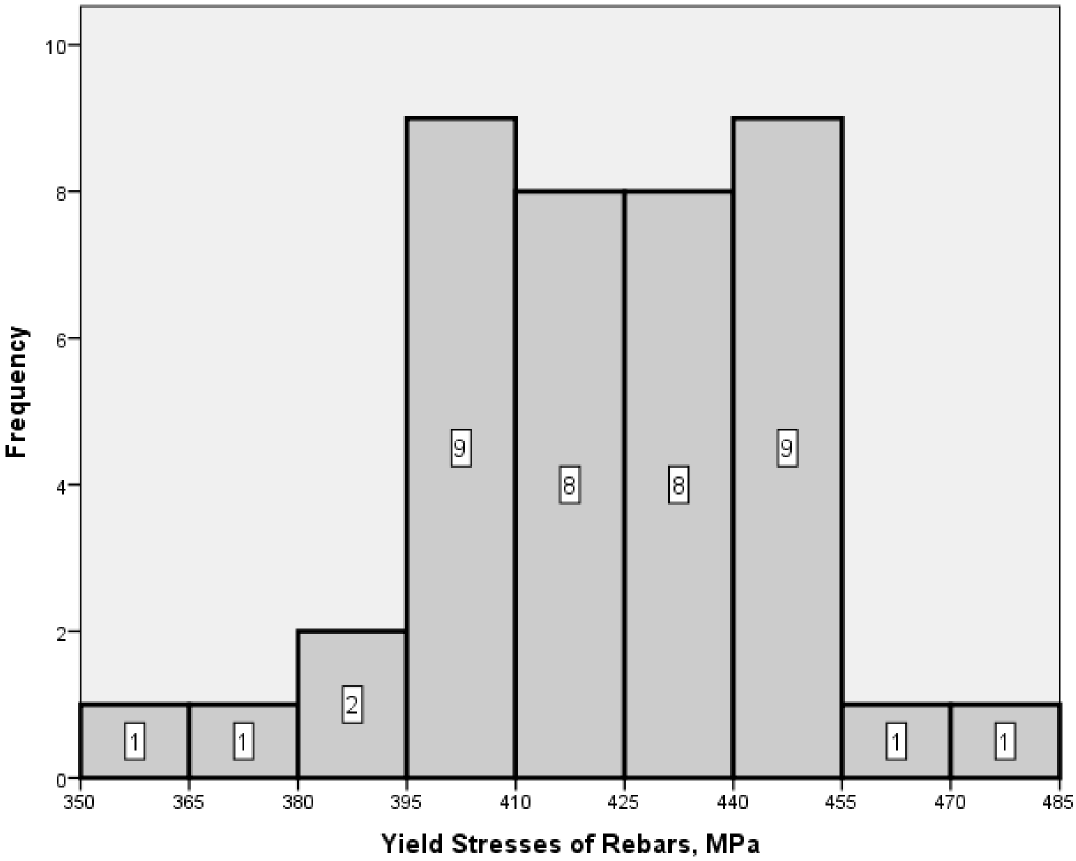


Figure 2.3-20: Histogram for yield stresses of reinforcing bars in MPa of Problem 2.3-1

Problem 2.3-2

Data in ***Table 2.3-15*** below represent angle of friction between concrete and sand. As indicated in ***Table 2.3-15***, data have been summarized in form of frequency distribution. For these data, draw **histogram**, **frequency polygon**, and **ogives**.

Table 2.3-15: Frequency distribution for angle of friction between concrete and sand, in degrees, for Problem 2.3-2.

Classes		Frequency
26	29	3
29	32	3
32	35	9
35	38	8
38	41	2
41	44	4

Answers

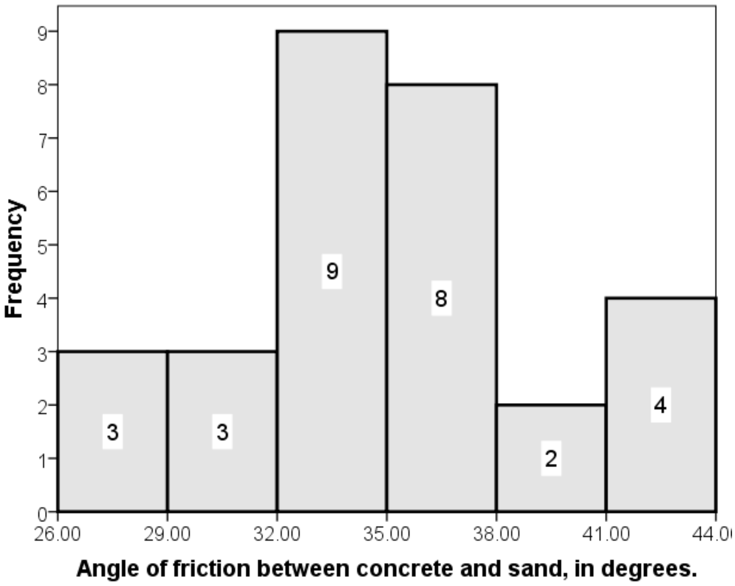


Figure 2.3-21: Histogram for data of Problem 2.3-2.

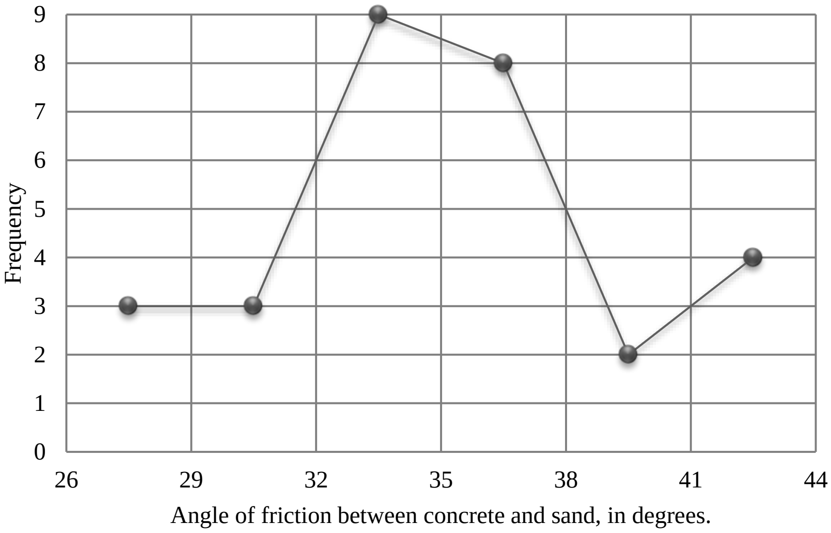


Figure 2.3-22: Frequency polygon for data of Problem 2.3-2.

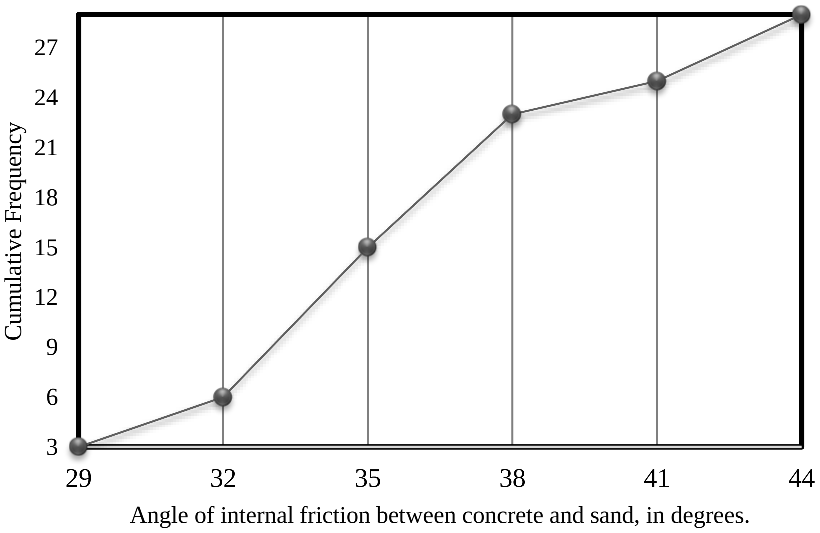


Figure 2.3-23: Ogives for data of Problem 2.3-2.

Problem 2.3-3

Maximum flow rates in m^3/sec, from 1921 to 1965 for a dam in the Mexico city were recorded and presented in ***Table 2.3-16*** below. For these data, construct a ***frequency distribution***, ***cumulative frequency distribution***, ***histogram***, ***frequency polygon***, and ***ogives***.

Table 2.3-16: Maximum flow rates, $m^3/second$.

1340	1380	1450	618	523	508	1220	1780
1090	944	397	282	353	597	995	611
985	1430	778	1280	1020	1300	1060	412
184	1480	876	113	516	1000	1890	611
409	780	674	969	870	329	458	1556
1217	819	576	1324				

Answers

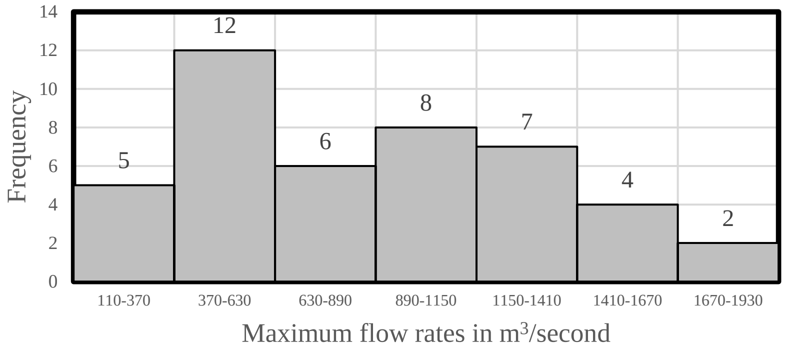


Figure 2.3-24: Histogram for data of Problem 2.3-3.

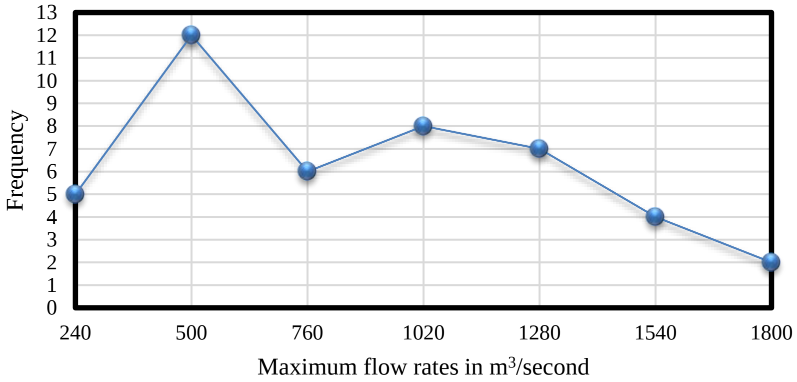


Figure 2.3-25: Frequency polygon for data of Problem 2.3-3.

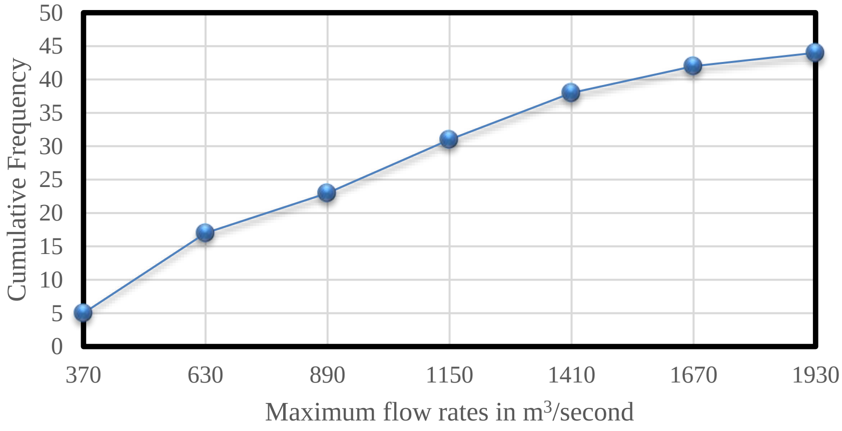


Figure 2.3-26: Ogives for data of Problem 2.3-3.

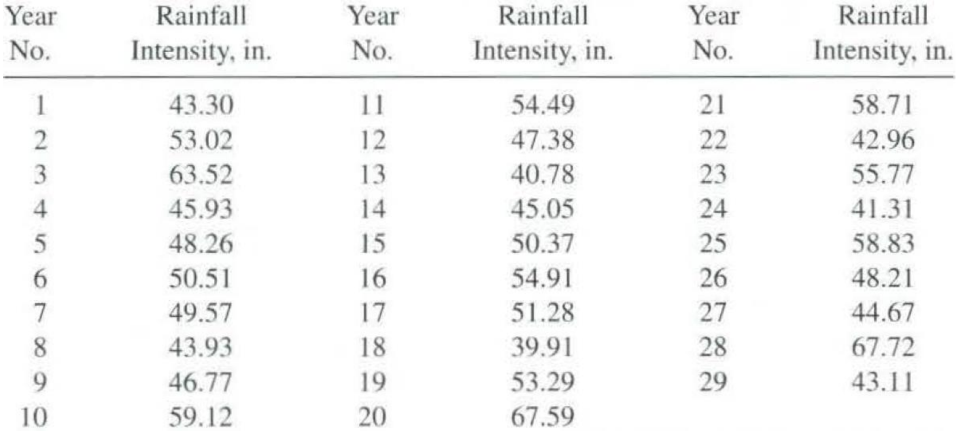
2.4 DENSITY SCALE*

- Oftentimes, there may be reasons to compare an empirical frequency diagram, such as a histogram, with a theoretical frequency distribution (such as a probability density function, PDF, discussed later in Chapter 6).
- For this purpose, the area under the empirical frequency diagram must be equal to unity. This can be obtained by dividing each of the ordinates in a histogram by its total area to have a ***density scale***.
- When classes have different widths, the histogram may be misleading, as a wider class may seem erroneous to have a larger amount of the data. To avoid this misleading presentation, the density scale is usually adopted with different width classes.

Example 2.4-1

The 29 years of annual cumulative rainfall intensity in a watershed area recorded over a period of 29 years as presented in ***Table 2.4-1***. These data have been put in a frequency table and presented in a histogram form as indicated in ***Table 2.4-2*** and ***Figure 2.4-1*** respectively. Reformulate these data in terms of a density scale.

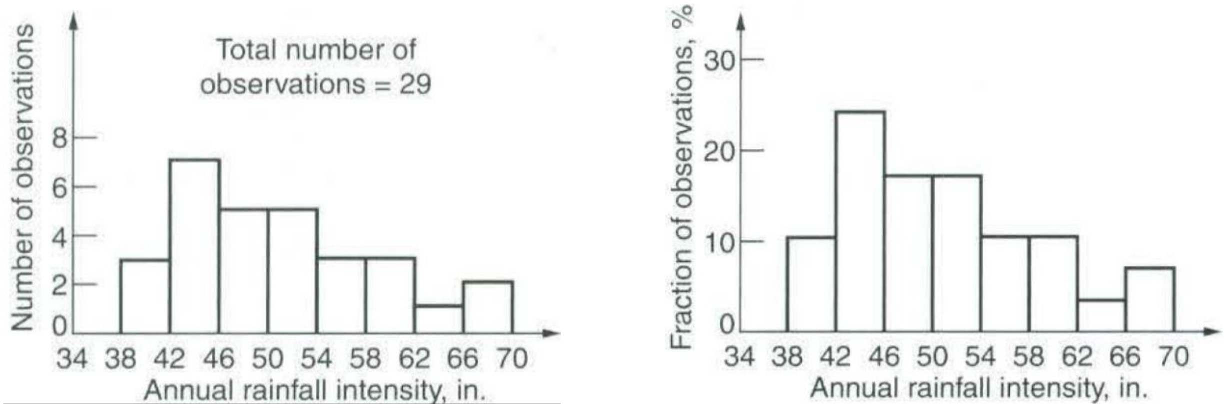
Table 2.4-1: Rainfall intensity data recorded over a period of 29 years.



Year No.	Rainfall Intensity, in.	Year No.	Rainfall Intensity, in.	Year No.	Rainfall Intensity, in.
1	43.30	11	54.49	21	58.71
2	53.02	12	47.38	22	42.96
3	63.52	13	40.78	23	55.77
4	45.93	14	45.05	24	41.31
5	48.26	15	50.37	25	58.83
6	50.51	16	54.91	26	48.21
7	49.57	17	51.28	27	44.67
8	43.93	18	39.91	28	67.72
9	46.77	19	53.29	29	43.11
10	59.12	20	67.59		

Table 2.4-2: Number and fraction of total observations in each interval.

Col.1		Col. 2	Col. 3
Intervals		Frequency	Relative Frequency
38	42	3	10.34
42	46	7	24.14
46	50	5	17.24
50	54	5	17.24
54	58	3	10.34
58	62	3	10.34
62	66	1	3.45
66	70	2	6.90
	Total	29	100.0



(a) In number of observations. **(b) In fraction of total observations.**

Figure 2.4-1: Histograms of annual rainfall intensity.

Solution

The histogram in terms of a density scale can be formulated starting the fraction of total observation indicated in the third column of ***Table 2.4-2*** through divided these values by the class width. The results are indicated in the fourth column of the ***Table 2.4-3***. The units for the fourth column are % per in. The % in the numerator is taken from the unit of the relative frequency of the third column while the inch unit in the denominator is taken from the class width of the rainfall intensity.

Table 2.4-3: Rainfall frequency distribution in terms of the density scale.

Col.1		Col. 2	Col. 3	Col. 4
Intervals		Frequency	Relative Frequency	Relative Frequency/Class Width
38	42	3	10.34	2.59
42	46	7	24.14	6.03
46	50	5	17.24	4.31
50	54	5	17.24	4.31
54	58	3	10.34	2.59
58	62	3	10.34	2.59
62	66	1	3.45	0.86
66	70	2	6.90	1.72
	Total	29	100.0	25.0

In terms of the density scale, the histogram would be as indicated in ***Figure 2.4-2***. Through multiplication the sum of the fourth column of Table 2.4-3 (25.0 % per inch) by the common class width of 4 inches, one can conclude that the area under the histogram of Figure 2.4-2 is 100% or 1 unit as required.

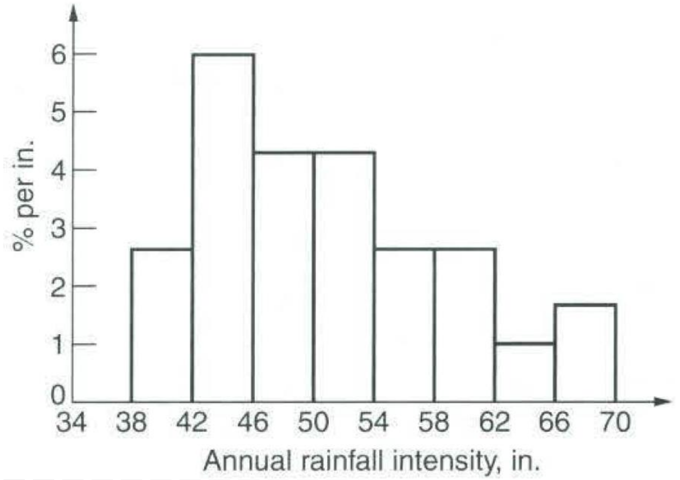


Figure 2.4-2: Histograms of annual rainfall intensity in terms of the density scale.

Example 2.4-2

Redraw the histogram for ***Example 2.3-12*** above but with using a density scale.

Solution

As discussed previously, the density scale is obtained when a relative frequency is divided by class width. Classes, frequency, relative frequency, and density scale for this example are presented in ***Table 2.4-4***. As an example, for these calculations, consider the density scale for second class:

$$Density\ Scale = \frac{\frac{Frequency}{Sample\ Size}}{Class\ width} = \frac{4/32}{900-400} = 0.00025\,1\ \frac{1}{\text{ohms/cm}}$$

Since relative frequency is dimensionless, hence the density scale will has units of reciprocal for those of measured quantities. In terms of density scale, the histogram would be as presented in ***Figure 2.4-3*** below.

With simple calculations, one will note that the area under histograms, that drawn in terms of density scale, is equal to 1 unit. Due to this feature, these diagrams is related to ***Probability Density Function***, PDF, of ***Chapter 5***.

Table 2.4-4: Classes, frequency, relative frequency, and density scale for Example 2.4-2.

Classes		Frequency	Class Width	Relative Frequency	Density Scale 1/(ohms/cm)
0	400	0	400	0	0
400	900	4	500	0.125	0.00025
900	1500	15	600	0.46875	0.00078125
1500	3500	11	2000	0.34375	0.000171875
3500	8000	2	4500	0.0625	1.38889E-05
8000	20000	0	12000	0	0

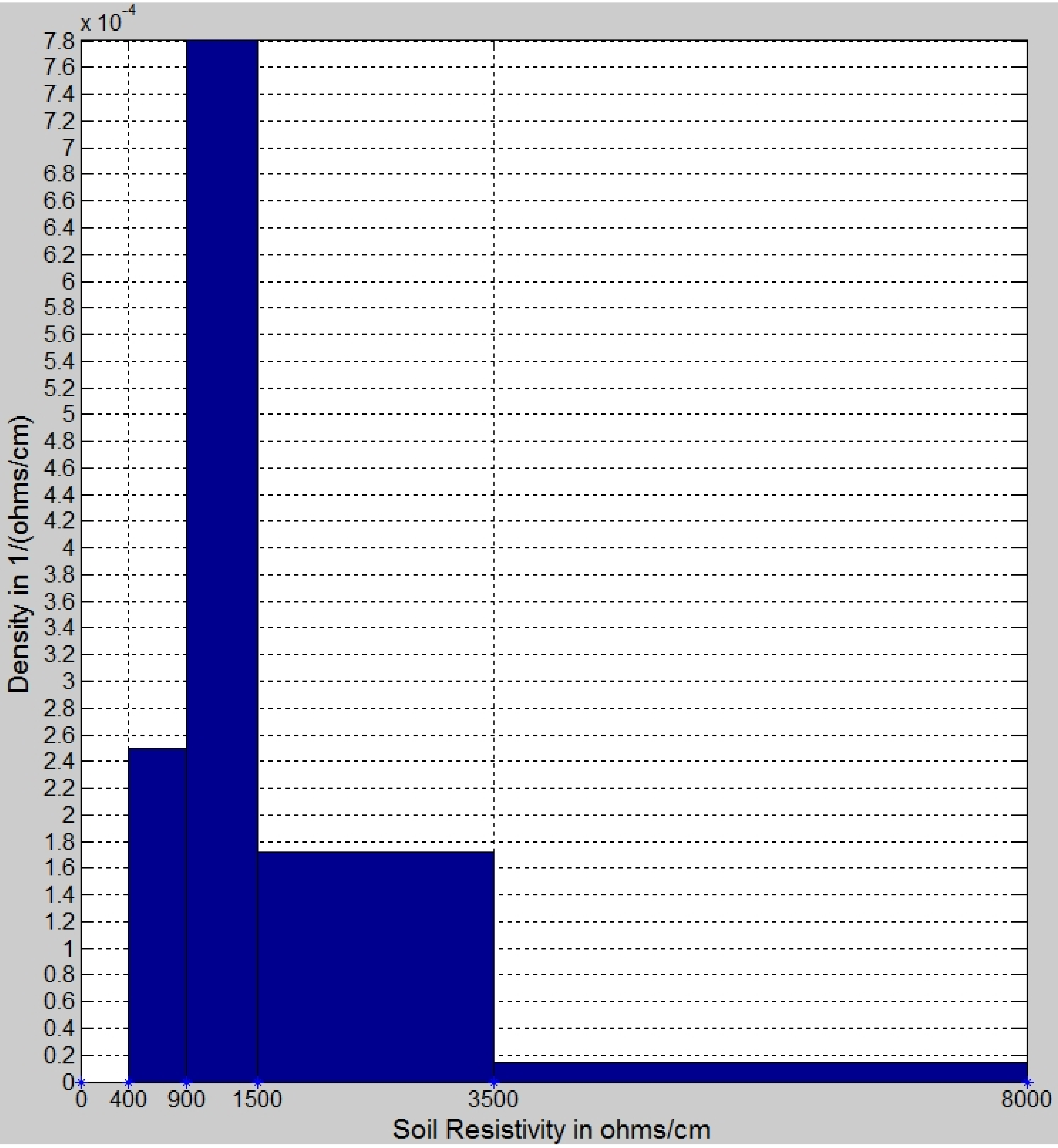


Figure 2.4-3: Histogram in terms of density scale for Example 2.4-2.

2.5 MATLAB COMPUTER APPLICATIONS*

Different Matlab functions can be adopted in graphical presentations of data. Some of them are presented in this article.

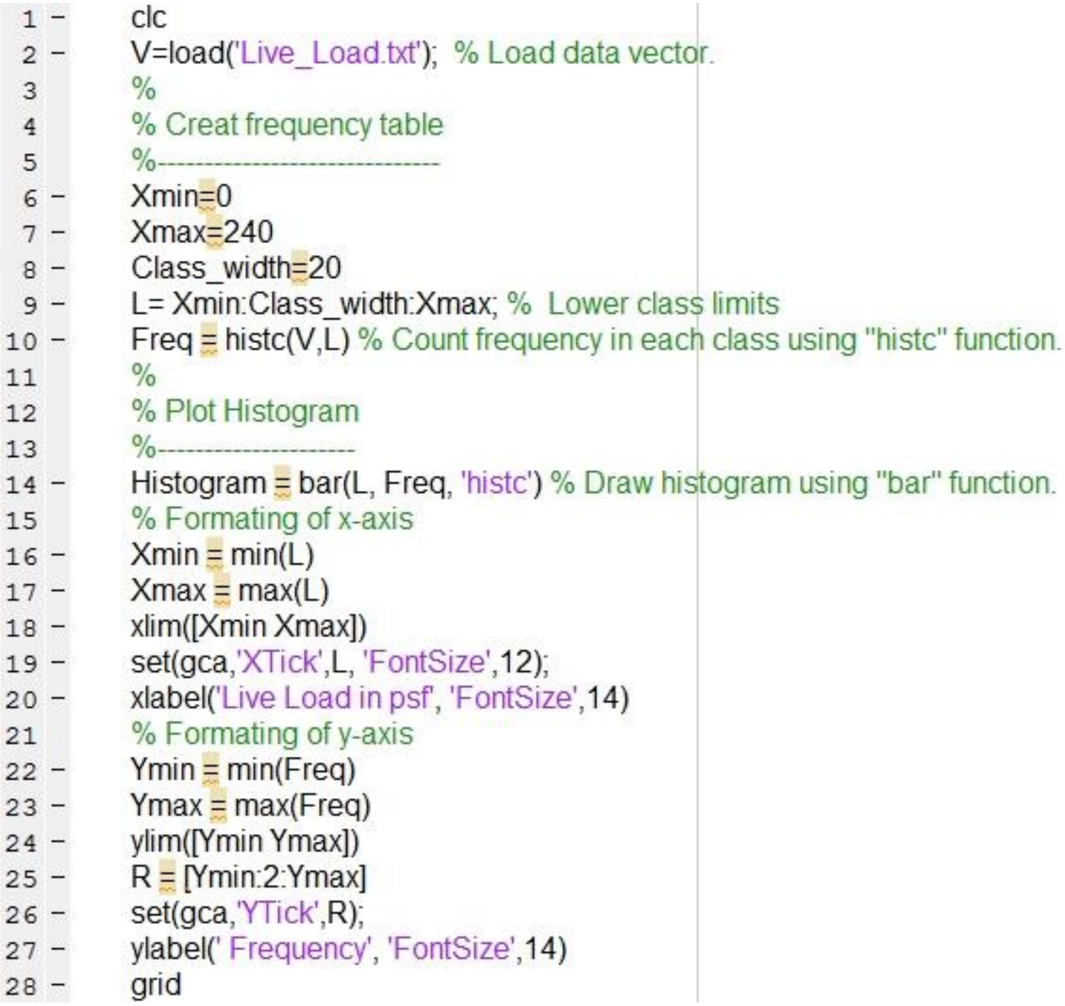
Example 2.5-1

Write a Matlab code to prepare the frequency distribution and to draw the histogram for ***Example 2.3-2***, ***Live Load Example***, using a class width of 20 psf.

Solution

As indicated in ***Table 2.5-1*** below, Matlab functions ***histc*** and ***bar*** have been adopted respectively to count frequencies and draw the histogram.

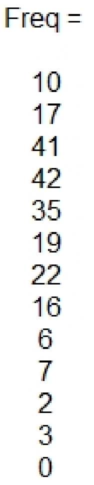
Table 2.5-1: Matlab code to count frequency and draw histogram for Example 2.5-1.



```
clc
V=load('Live_Load.txt');  % Load data vector.
%
% Creat frequency table
%-----------------------------
Xmin=0
Xmax=240
Class_width=20
L= Xmin:Class_width:Xmax; %  Lower class limits
Freq = histc(V,L) % Count frequency in each class using "histc" function.
%
% Plot Histogram
%---------------------
Histogram = bar(L, Freq, 'histc') % Draw histogram using "bar" function.
% Formating of x-axis
Xmin = min(L)
Xmax = max(L)
xlim([Xmin Xmax])
set(gca,'XTick',L, 'FontSize',12);
xlabel('Live Load in psf', 'FontSize',14)
% Formating of y-axis
Ymin = min(Freq)
Ymax = max(Freq)
ylim([Ymin Ymax])
R = [Ymin:2:Ymax]
set(gca,'YTick',R);
ylabel(' Frequency', 'FontSize',14)
grid
```

When the code is run, the frequency would be as indicated in ***Table 2.5-2*** while the histogram would be as indicated in ***Figure 2.5-1*** below. As a verification, one can note that these results coincide with those of ***Figure 2.3-3***.

Table 2.5-2: Frequency for Example 2.5-1 determined from Matlab code.



```
Freq =

   10
   17
   41
   42
   35
   19
   22
   16
    6
    7
    2
    3
    0
```

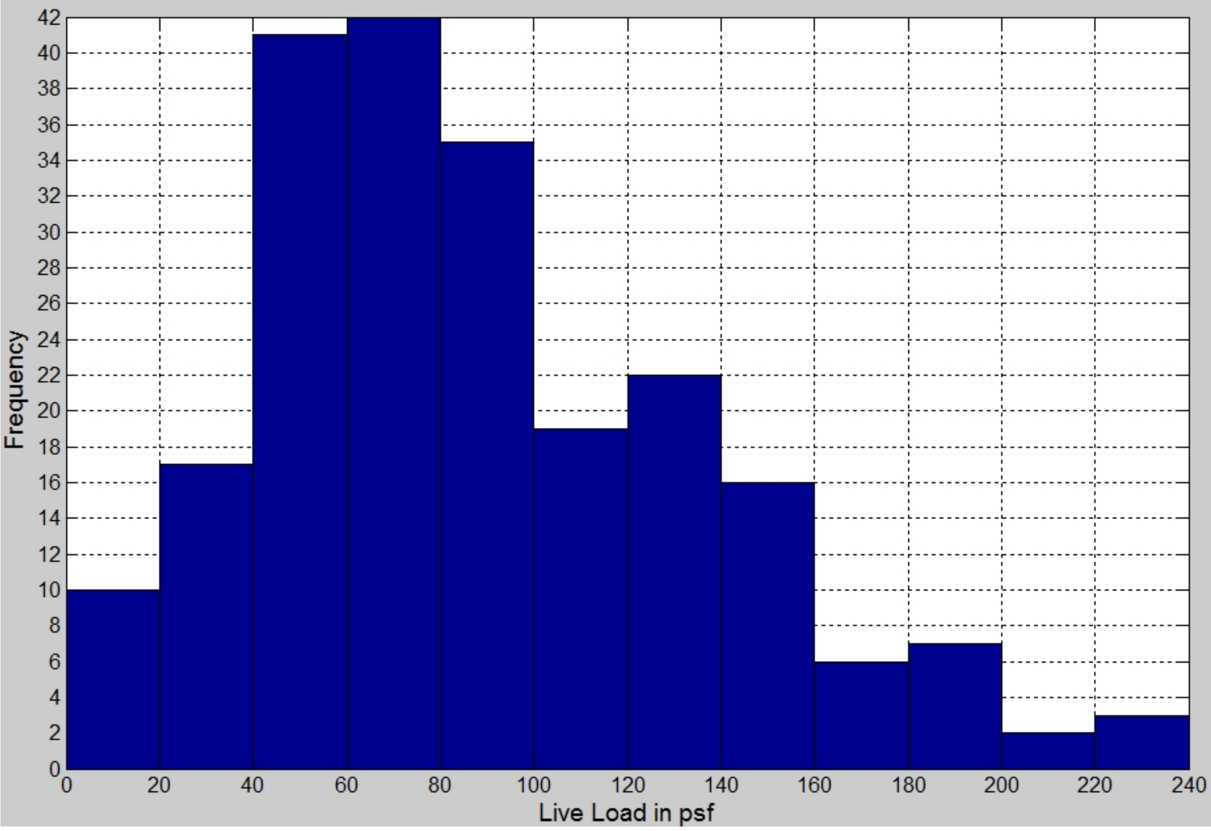


Figure 2.5-1: Histogram for Example 2.5-1 determined from Matlab code.

Example 2.5-2

Resolve ***Example 2.5-1*** but with using a class width of 10 psf.

Solution

Only the command line 8 of Matlab code, 8 - Class_width=20, should be modified, to change class width into 10 psf. The histogram would be as indicated in ***Figure 2.5-2*** below. As a verification, one can note that this histogram coincides with ***Figure 2.3-4***.

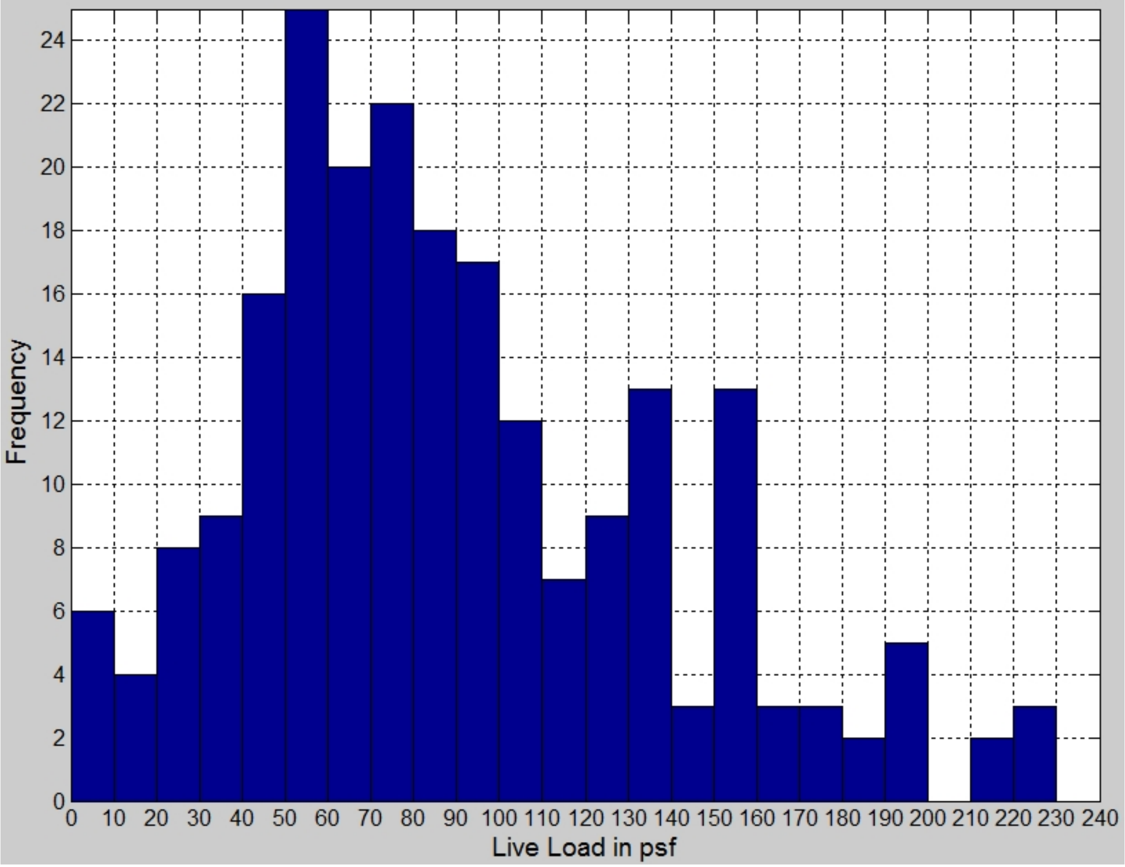


Figure 2.5-2: Histogram for Example 2.5-2 determined from Matlab code.

Contents

CHAPTER 3

DATA DESCRIPTION

3.1 INTRODUCTION

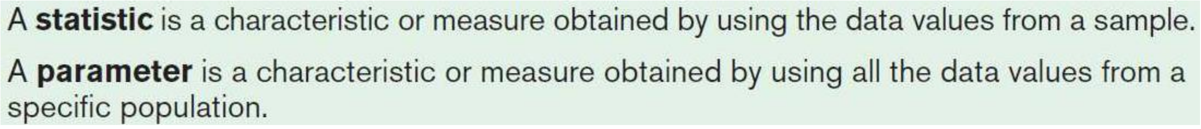
- Chapter 2 showed how one can gain useful information from raw data by:
  - Organizing them into a frequency distribution,
  - Presenting the data by using various graphs,
- ***Section 3.2*** defines the statistic and the parameter and shows the current practice of using the roman letters to describe the statistic while the Greek letters for the parameter.
- ***Section 3.3*** shows the central tendency measures that usually used to describe the data congestion about a central value. The measures of central tendency include:
  - Mean,
  - Median,
  - Mode.
- In addition to knowing the average, one must know how the data values are dispersed. That is, do the data values cluster around the mean, or are they spread more evenly throughout the distribution? The measures that determine the spread of the data values are called ***measures of variation,*** or ***measures of dispersion***. These measures include the:
  - Range,
  - Variance,
  - Standard deviation.

  and they are discussed in ***Section 3.4***.
- ***Section 3.5*** shows how a linear transformation affects the measures of ***Section 3.3*** and ***Section 3.4***.
- Other measures to indicate asymmetry and peakedness of data are presented in ***Section 3.6***.
- Another set of measures is necessary to describe data. These measures are called ***measures of position***. They tell where a specific data value falls within the data set or its relative position in comparison with other data values. The most common position measures are:
  - Percentiles,
  - Deciles,
  - Quartiles.

  These measures are discussed in ***Section 3.7***.
- Sections ***3.6*** and ***3.7*** are advanced in nature and they are beyond the scope of the undergraduate courses.

3.2 BASIC DEFINITIONS AND NOTATIONS

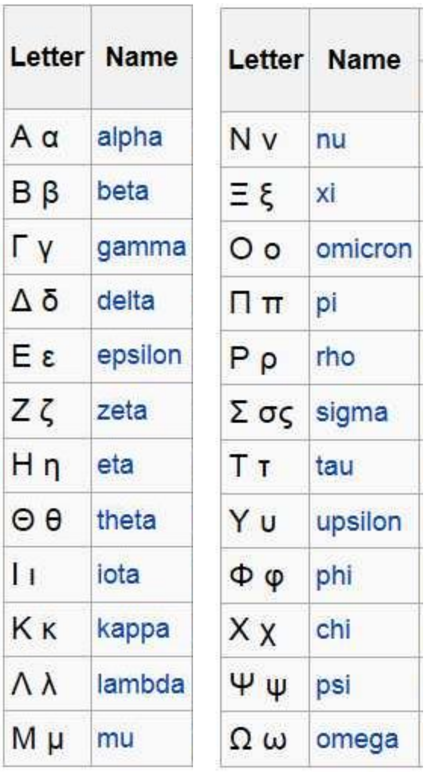
- Measures found by using all the data values in the population are called ***parameters***.
- Measures obtained by using the data values from samples are called ***statistics***.



A **statistic** is a characteristic or measure obtained by using the data values from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

- In statistics, ***Greek letters are used to denote parameters***, and ***Roman letters are used to denote statistics***. Greek letters have been summarized in below:



Letter	Name	Letter	Name
Α α	alpha	Ν ν	nu
Β β	beta	Ξ ξ	xi
Γ γ	gamma	Ο ο	omicron
Δ δ	delta	Π π	pi
Ε ε	epsilon	Ρ ρ	rho
Ζ ζ	zeta	Σ σς	sigma
Η η	eta	Τ τ	tau
Θ θ	theta	Υ υ	upsilon
Ι ι	iota	Φ φ	phi
Κ κ	kappa	Χ χ	chi
Λ λ	lambda	Ψ ψ	psi
Μ μ	mu	Ω ω	omega

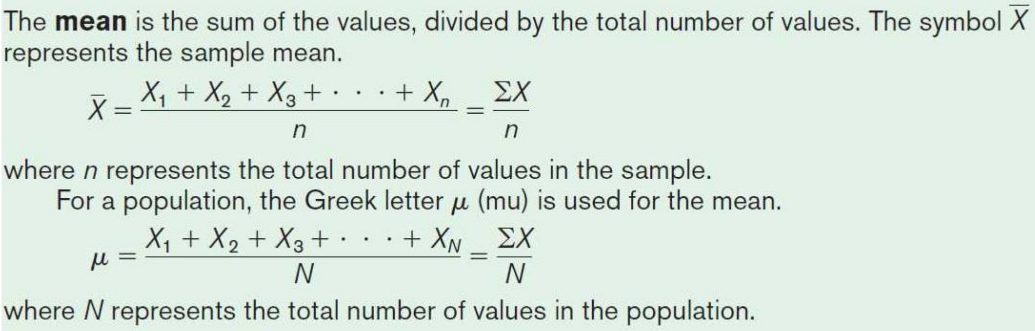
- Assume that the data are obtained from samples unless otherwise specified.

3.3 MEASURES OF CENTRAL TENDENCY

3.3.1 THE FIRST MEASURE "THE MEAN" OR "THE ARITHMETIC MEAN"

3.3.1.1 For Ungrouped Data

The ***mean***, also known as the ***arithmetic average*** can be defined as follows:



The **mean** is the sum of the values, divided by the total number of values. The symbol $\overline{X}$ represents the sample mean.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdot\cdot\cdot + X_n}{n} = \frac{\Sigma X}{n}$$

where n represents the total number of values in the sample.

For a population, the Greek letter μ (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdot\cdot\cdot + X_N}{N} = \frac{\Sigma X}{N}$$

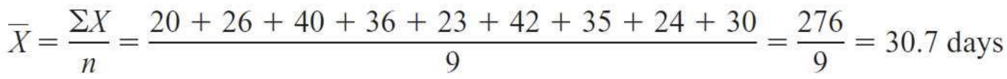
where N represents the total number of values in the population.

Example 3.3-1

The data represent the number of compressive strengths for a sample of concrete. Find the mean.

20, 26, 40, 36, 23, 42, 35, 24, 30.

Solution



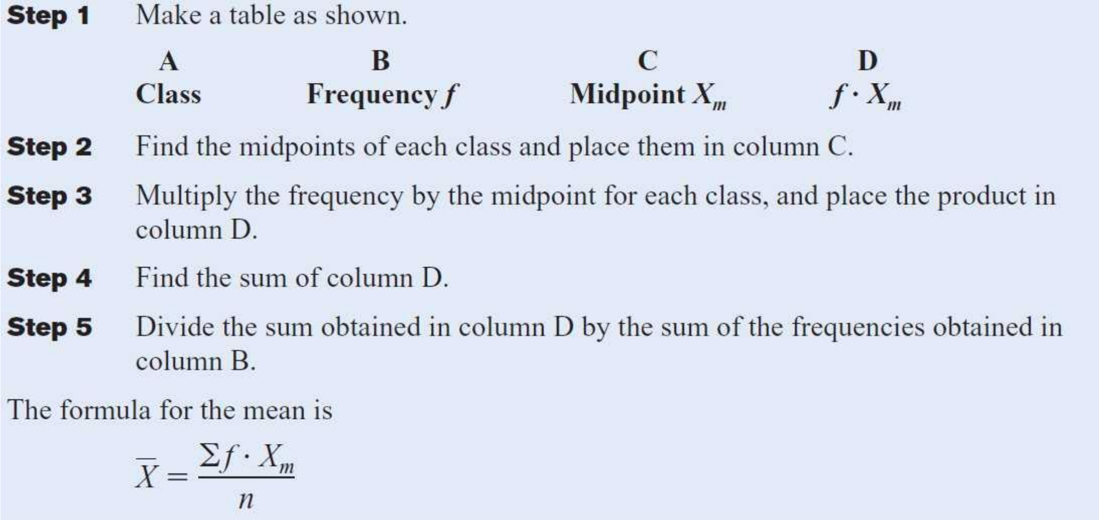
$$\overline{X} = \frac{\Sigma X}{n} = \frac{20 + 26 + 40 + 36 + 23 + 42 + 35 + 24 + 30}{9} = \frac{276}{9} = 30.7 \text{ days}$$

3.3.1.2 Rounding Rule for the Mean

The mean should be rounded to one more decimal place than occurs in the raw data. For example, if the raw data are given in whole numbers, the mean should be rounded to the nearest tenth. If the data are given in tenths, the mean should be rounded to the nearest hundredth, and so on.

3.3.1.3 For Grouped Data

For grouped data, mean can be computed based on following procedure:



Step 1 Make a table as shown.

A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

Example 3.3-2

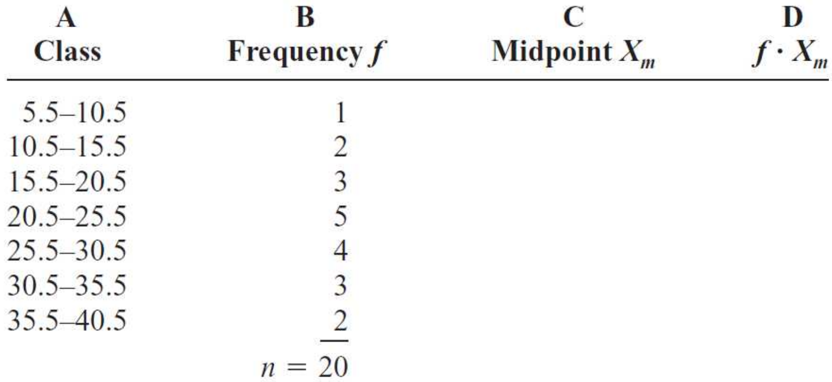
Find the mean for following grouped data. This data represents the compressive strength for 20 randomly selected concrete cylinder samples.

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
Summation	20

Solution

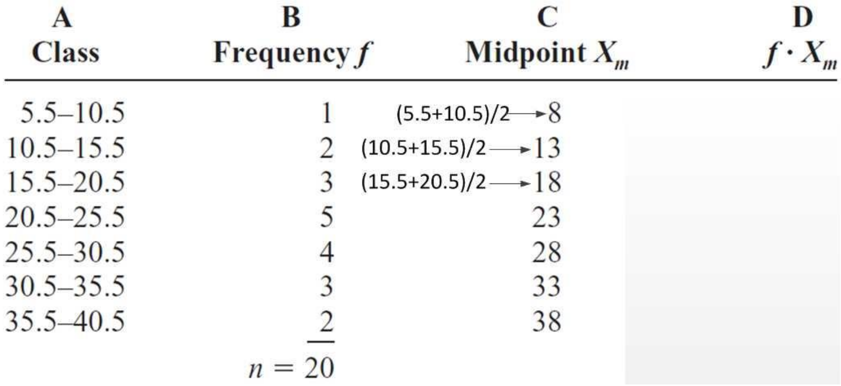
The procedure for finding the mean for grouped data is given here.

Step 1: Make a table as shown.



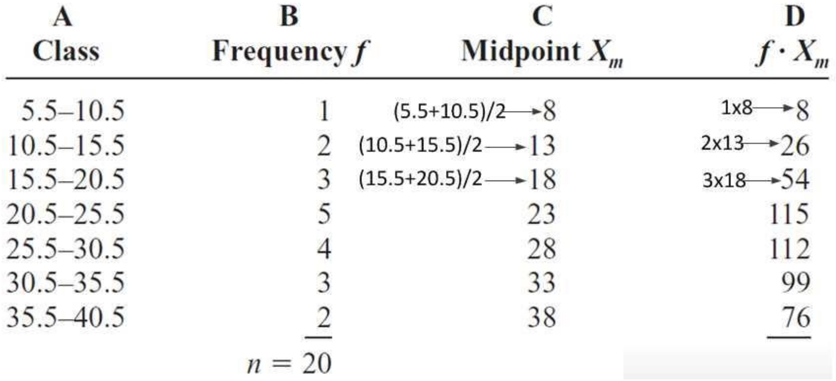
A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$
5.5–10.5	1		
10.5–15.5	2		
15.5–20.5	3		
20.5–25.5	5		
25.5–30.5	4		
30.5–35.5	3		
35.5–40.5	2		
	$n = 20$		

Step 2: Find the midpoints of each class and enter them in column C.



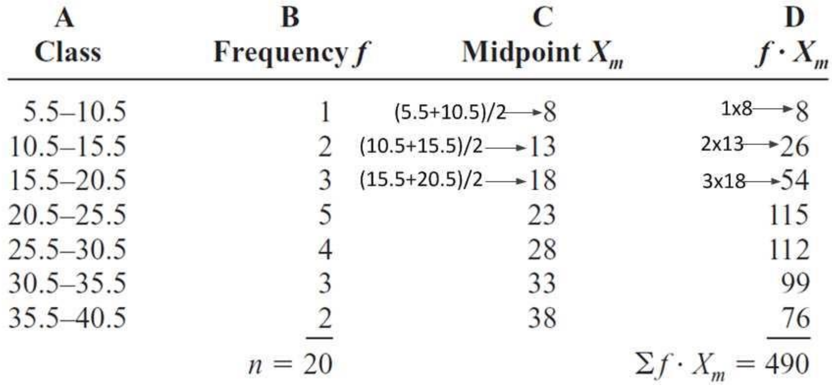
A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$
5.5–10.5	1	(5.5+10.5)/2 → 8	
10.5–15.5	2	(10.5+15.5)/2 → 13	
15.5–20.5	3	(15.5+20.5)/2 → 18	
20.5–25.5	5	23	
25.5–30.5	4	28	
30.5–35.5	3	33	
35.5–40.5	2	38	
	$n = 20$		

Step 3 For each class, multiply the frequency by the midpoint, as shown, and place the product in column D.



A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$
5.5–10.5	1	(5.5+10.5)/2 → 8	1x8 → 8
10.5–15.5	2	(10.5+15.5)/2 → 13	2x13 → 26
15.5–20.5	3	(15.5+20.5)/2 → 18	3x18 → 54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		

Step 4: Find the sum of column D.



A	B	C	D
Class	Frequency f	Midpoint X_m	$f \cdot X_m$
5.5–10.5	1	(5.5+10.5)/2 → 8	1x8 → 8
10.5–15.5	2	(10.5+15.5)/2 → 13	2x13 → 26
15.5–20.5	3	(15.5+20.5)/2 → 18	3x18 → 54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

Step 5: Divide the sum by n to get the mean.

$$X = \frac{\Sigma f.X_m}{n} = \frac{490}{20} = 24.5\ MPa$$

3.3.1.4 Important Note Related to Mean of Grouped Data:

- The procedure for finding the mean for grouped data assumes that the mean of all the raw data values in each class is equal to the midpoint of the class.
- This is not true, since the average of the raw data values in each class usually will not be exactly equal to the midpoint.
- However, using this procedure will give an acceptable approximation of the mean, since some values fall above the midpoint and other values fall below the midpoint for each class, and the midpoint represents an estimate of all values in the class.

3.3.2 OTHER MEAN VALUES

3.3.2.1 Harmonic Mean

There are two approaches to compute the mean for ratios indicated in below: $x_1, x_2, \ldots.., x_n$

- When ratios x_i have ***common denominator***, their mean is an arithmetic mean:
  $$\bar{x} = \frac{x_1 + x_2 + \cdots.. + x_n}{n}$$
- On the other hand, when ratios have ***common numerator***, the mean is called the ***harmonic mean*** and computed as follows:
  $$\bar{x}_h = \frac{1}{\frac{1}{n\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots. + \frac{1}{x_n}\right)}}$$ **Eq. 3.3-1**

Example 3.3-3

A tourist purchases gasoline at three filling stations, where the prices are 33 1/3, 25, and 20 cent per gallon. What is the average price?

Solution

If gallon is taken as a fixed unit, then the ratios have common denominator and the average would be the arithmetic mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}(33.333 + 25 + 20) = 26.1\ cent\ per\ gallon$$

While when cent is considered as fixed unit, therefore the ratios would have common numerator and the mean would be a harmonic one:

$$\bar{x}_h = \frac{1}{\frac{1}{n\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots. + \frac{1}{x_n}\right)}} = \frac{1}{\frac{1}{3\left(\frac{1}{33.33} + \frac{1}{25} + \frac{1}{20}\right)}} \approx 25\ cents\ per\ galoon$$

The former would be the correct estimator when the tourist intends to buy same number of gallons from each station. While the latter would be the correct estimator when the tourist intends to spend same cents, say 50 cents, at each station.

Example 3.3-4

Three velocities of $0.20\frac{m}{s}, 0.24\frac{m}{s}$, and $0.16\frac{m}{s}$ have been determined for stream flow through using of a floating device. Select which mean value should be adopted and determine mean speed accordingly.

Solution

From physics, velocity v is defined as:

$$v = \frac{L}{t}$$

where L is the distance between two points and t is the time required to travel between them. With a floating device, the distance L between two point is assumed fixed and required time t is computed accordingly, therefore the ratios, i.e. velocities, have common numerator and harmonic mean should be adopted:

$$\bar{x}_h = \frac{1}{\frac{1}{n\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots.+\frac{1}{x_n}\right)}} = \frac{1}{\frac{1}{3\left(\frac{1}{0.2}+\frac{1}{0.24}+\frac{1}{0.16}\right)}} \approx 0.194\frac{m}{s}$$

Example 3.3-5

On a certain country road that runs from a coastal town to a village in the mountains, the average speed of motorcars is 80 km/h uphill and 100 km/h downhill. What is the average speed for a journey from the town to the village and back?

Solution

From physics, velocity v is defined as:

$$v = \frac{L}{t}$$

where L is the distance between two points and t is the time required to travel between them.

Assume that car speedometer measures speed based on different times for a constant distance, therefore we have common numerator and different denominators and mean value should be determined in terms of harmonic mean:

$$\bar{v} = \frac{1}{\frac{1}{n\left(\frac{1}{v_1}+\frac{1}{v_2}+\cdots.+\frac{1}{v_n}\right)}} = \frac{1}{\frac{1}{2\left(\frac{1}{80}+\frac{1}{100}\right)}} \approx 88.9\frac{km}{h}$$

3.3.2.2 Geometric Mean

- The geometric mean is used in ***averaging values that represent a rate of change***.
- Here the variable follows an exponential, that is, a logarithmic law.
- For a sample of observations,

  $x_1, x_2, \ldots\ldots, x_n$

  the geometric mean is defined as:

  $$\bar{x}_g = (x_1 \times x_2 \times x_3 \times \ldots\ldots \times x_n)^{\frac{1}{n}}$$ **Eq. 3.3-2**

  This average is used when dealing with observations each of bears an approximately constant ratio to the preceding one, for example, in average rates of growth (increase or decrease).
- Geometric mean is applicable only for positive data.
- The use of the geometric mean can be avoided by transforming the original variables x into $\log x$. The antilog of the arithmetic mean of the new variable will the give the right answer.

  $$\log \bar{x}_g = \frac{\Sigma(\log x_i)}{n}$$ **Eq. 3.3-3**

Example 3.3-6

A value of 100 falls to 50 and subsequently raises to 100. The ratio of change is ½ and 2. Use the arithmetic mean and the geometric mean to determine the average rate of change. Comment on the results.

Solution

This example aims to show intuitively how the geometric mean is more representative than the arithmetic mean when the data is related to the rate of change.

$$\bar{x} = \frac{\frac{1}{2}+2}{2} = 1.25$$

While the geometric mean is:

$$\bar{x}_g = (x_1 \times x_2 \times x_3 \times \dots \dots \times x_n)^{\frac{1}{n}} = \left(\frac{1}{2} \times 2\right)^{\frac{1}{2}} = 1$$

The answer according to the geometric mean is intuitively correct as the overall change is zero.

Example 3.3-7

Consider the case of populations of towns and cities that increase geometrically, which means that a future increase is expected that is proportional to the current population.

Such information is invaluable for planning and designing urban water supplies and sewerage systems.

Suppose, for example, that according to a survey conducted in 1970 and again in 1990 the population of a city had increased from 230,000 to 310,000. An engineer needs to verify, for purposes of design, the per capita consumption of water in the intermediate period and hence tries to estimate the population in 1980.

Solution

The central value to use in this situation is the geometric mean of the two numbers which is:

$$\bar{x}_g = (x_1 \times x_2 \times x_3 \times \dots \dots \times x_n)^{\frac{1}{n}} = (230000 \times 310000)^{\frac{1}{2}} = 267021\ capita\ in\ 1980.$$

Example 3.3-8

The number of degrees *cum laude* awarded at a university during six consecutive years is given in the table below. What is the average percentage of increase in the number of such degrees per annum?

Year	Number of Degrees	Ratio to Previous Year's Value
1959	5	-
1960	6	1+(6-5)/5 = 1.20
1961	9	1+(9-6)/6 = 1.50
1962	15	1+(15-9)/9 = 1.67
1963	30	1+(30-15)/15 = 2.00
1964	50	1+(50-30)/30 = 1.67

Solution

To find the average we calculate the geometric mean of the ratios given in the last column of the table:

$$\bar{x}_g = (x_1 \times x_2 \times x_3 \times \dots \dots \times x_n)^{\frac{1}{n}} = (1.2 \times 1.5 \times 1.67 \times 2.00 \times 1.67)^{\frac{1}{5}} = 1.586$$

That is an average increase per year of 58.6 percent.

Example 3.3-9

A company's year to year changes in fuel consumption expenditures were -5, 10, 20, 40, and 60 percent. Using the geometric mean of growth factor, determine the average yearly percent change in expenditures.

Solution

Converting the percent changes to growth factors:

0.95, 1.10, 1.20, 1.40, and 1.60

Then, compute the geometric mean for these factors:

$$\bar{x}_g = (x_1 \times x_2 \times x_3 \times \dots \dots \times x_n)^{\frac{1}{n}} = (0.95 \times 1.10 \times 1.20 \times 1.40 \times 1.60)^{\frac{1}{5}} = 1.229$$

Subtracting 1 gives 0.229 or 22.9% average increase per year.

3.3.3 Relations between the Three Different Mean Values

- For a set of positive numbers,
  - The geometric mean is less than or equal to the arithmetic mean. They are equal in the rare case where all the numbers in the data are the same.

  - The arithmetic mean is greater than or equal to the harmonic mean.
- The relations between the three means can be summarized according to the following relation:

$\bar{x}_h \leq \bar{x} \leq \bar{x}_g$ **Eq. 3.3-4**

3.3.4 THE SECOND MEASURE "THE MEDIAN"

3.3.4.1 Basic Definition

The **median** is the midpoint of the data array. The symbol for the median is MD.

3.3.4.2 Steps in computing the median of a data array

Step 1 Arrange the data in order.

Step 2 Select the middle point.

Example 3.3-10

Find the median for following data.

713, 300, 618, 595, 311, 401, and 292

Solution

Step 1 Arrange the data in order.

292, 300, 311, 401, 595, 618, 713

Step 2 Select the middle value.

292, 300, 311, 401, 595, 618, 713
↑
Median

Hence, the median is 401 rooms.

Example 3.3-11

Find the median for the following data

684, 764, 656, 702, 856, 1133, 1132, 1303

Solution

656, 684, 702, 764, 856, 1132, 1133, 1303
↑
Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

3.3.5 THE THIRD MEASURE "THE MODE"

3.3.5.1 For Ungrouped Data

- Basic Definition:
  The value that occurs most often in a data set is called the **mode.**
- A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**.
- If a data set has two values that occur with the same greatest frequency, both values are the mode and the data set is said to be **bimodal**.
- If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal**.
- When no data value occurs more than once, the data set is said to have no mode.

Example 3.3-12: (Unimodal Data)

Find the mode for the following data:

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

Solution

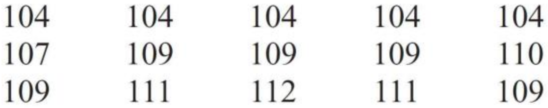
It is helpful to arrange the data in order although it is not necessary.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5

Since 10 occurred 3 times—a frequency larger than any other number—the mode is 10.

Example 3.3-13: (Bimodal Data)

Find the mode for following data:



104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

Solution

Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

Example 3.3-14: (Data that has no Mode)

Find the mode for following data

401, 344, 209, 201, 227, 353

Solution

Since each value occurs only once, there is no mode.

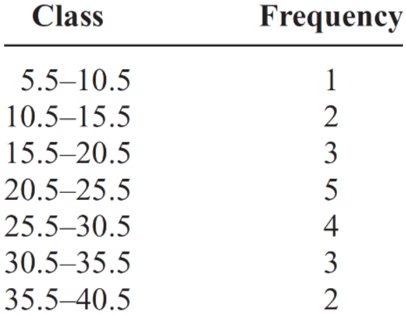
Note: Do not say that the mode is zero. That would be incorrect, because in some data, such as temperature, zero can be an actual value.

3.3.5.2 For Grouped Data

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

Example 3.3-15

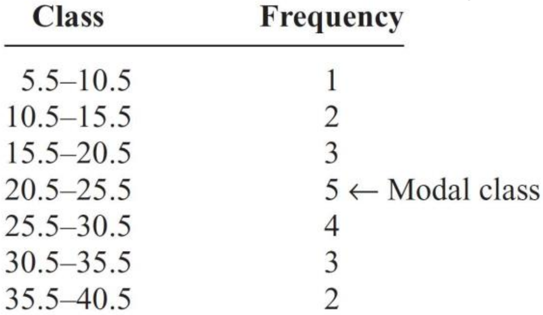
Find the modal class for the frequency distribution of Example 3.3-2.



Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

Solution

The modal class is 20.5–25.5, since it has the largest frequency.



Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5 ← Modal class
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

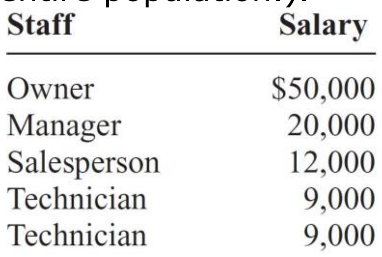
Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 23 MPa.

3.3.6 Which Central Measure should be Adopted

- An extremely high or extremely low data value in a data set can have a striking effect on the mean of the data set. These extreme values are called **outliers**.
- This is one reason why when analyzing a frequency distribution, you should be aware of any of these values.
- For the data set shown in Example below, the mean, median, and mode can be quite different because of extreme values.

Example 3.3-16

A small company consists of the owner, the manager, the salesperson, and two technicians, all whose annual salaries are listed here. (Assume that this is the entire population.).



Staff	Salary
Owner	$50,000
Manager	20,000
Salesperson	12,000
Technician	9,000
Technician	9,000

Find the mean, median, and mode.

Solution

$$\mu = \frac{\Sigma X}{N} = \frac{50{,}000 + 20{,}000 + 12{,}000 + 9000 + 9000}{5} = \$20{,}000$$

Hence, the mean is $20,000, the median is $12,000, and the mode is $9,000. In this Example, the mean is much higher than the median or the mode. This is so because the extremely high salary of the owner tends to raise the value of the mean.

In this and similar situations, the median should be used as the measure of central tendency.

3.3.7 Distribution Shapes

Frequency distributions can assume many shapes. The three most important shapes are:

- Positively skewed,
- Symmetric,
- Negatively skewed.

3.3.7.1 Positively skewed or right-skewed distribution

Most of the data values fall to the left of the mean and cluster at the lower end of the distribution; the "tail" is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median.

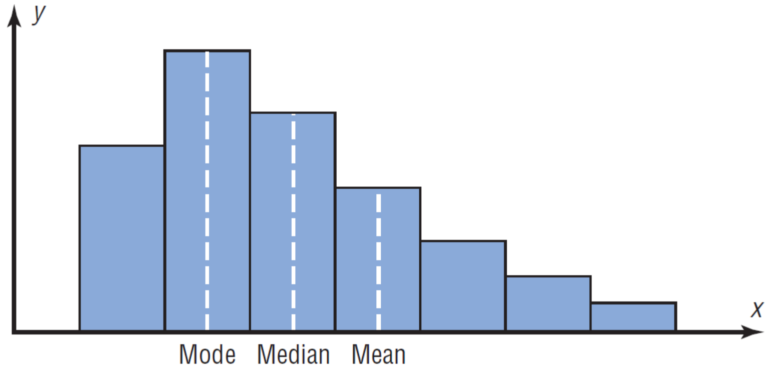


Figure 3.3-1: A positively skewed or right-skewed.

3.3.7.2 Symmetric Distribution

In a **symmetric distribution**, the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution.

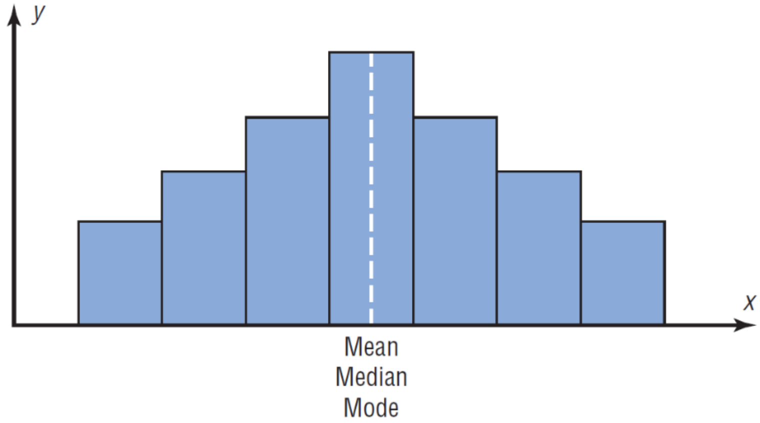


Figure 3.3-2: A symmertic distribution.

3.3.7.3 Negatively Skewed or Left-Skewed

When most of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left, the distribution is said to be **negatively skewed or left-skewed**. Also, the mean is to the left of the median, and the mode is to the right of the median.

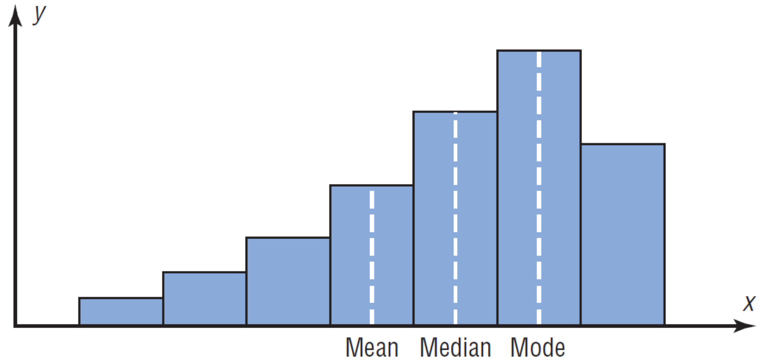


Figure 3.3-3: A negatively skewed distribution.

3.4 Measures of Variation

In statistics, to describe the data set accurately, ***statisticians must know more than the measures of central tendency***. Consider the example below.

Example 3.4-1

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Solution

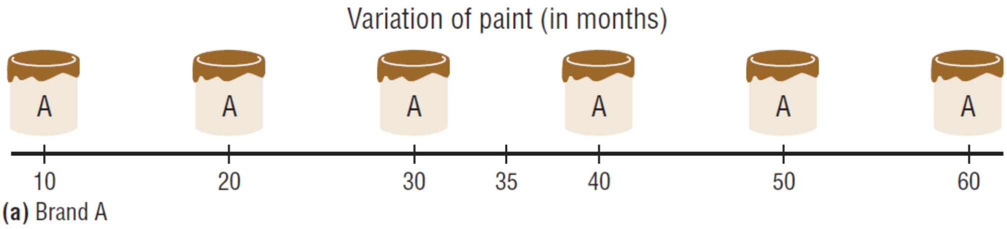
The mean for brand A is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35\, month$$

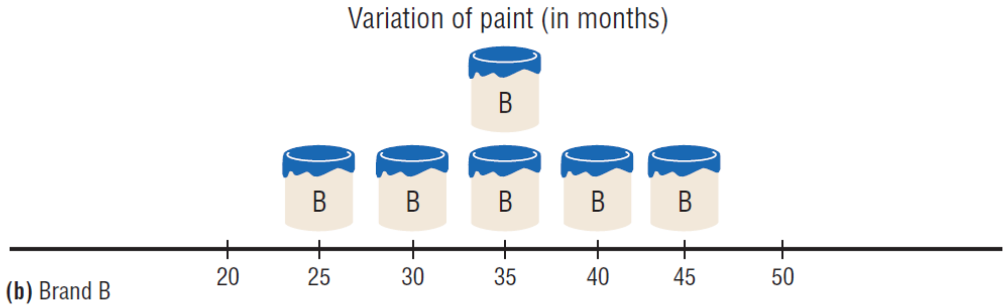
The mean for brand B is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35\, months$$

Since the means are equal, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure below.



(a) Brand A



(b) Brand B

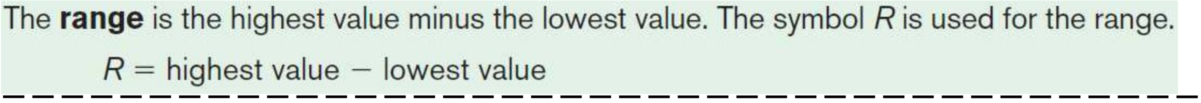
As Figure above shows, even though the means are the same for both brands, the spread, or variation, is quite different. Brand B performs more consistently; it is less variable.

For the spread or variability of a data set, three measures are commonly used:

- Range,
- Variance and Standard deviation.

Each measure will be discussed in this section.

3.4.1 THE FIRST MEASURE "RANGE"



The **range** is the highest value minus the lowest value. The symbol R is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

Example 3.4-2

Find the ranges for the paints in previous Example.

Solution

For brand A, the range is

$R = 60 - 10 = 50$ months

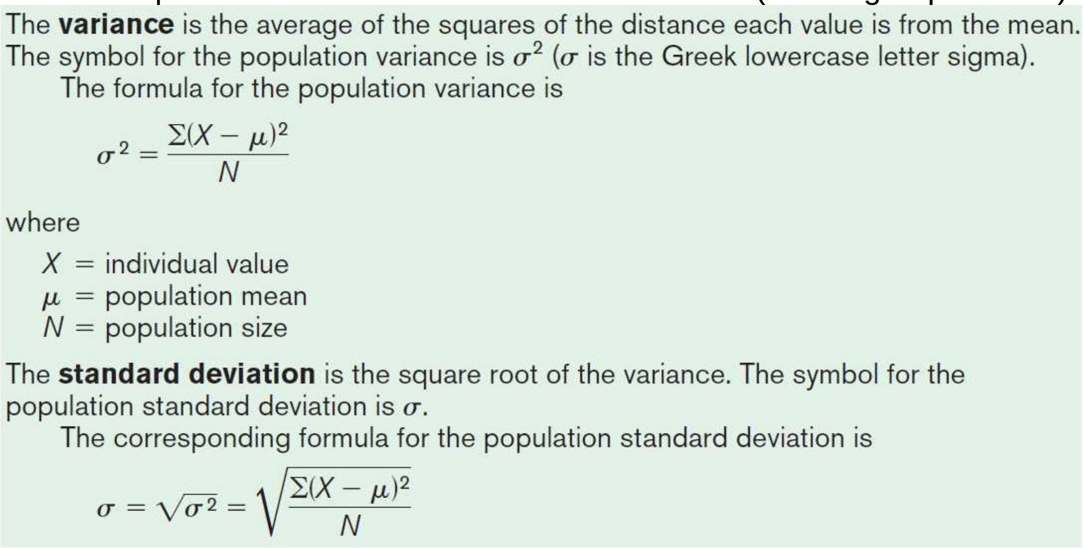
For brand B, the range is

$R = 45 - 25 = 20$ months

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

3.4.2 THE SECOND AND THIRD MEASURES "VARIANCE AND STANDARD DEVIATION"

3.4.2.1 Population Variance and Standard Deviation (For Ungrouped Data)



The **variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

where

X = individual value
μ = population mean
N = population size

The **standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

3.4.2.2 Rounding Rule for the Standard Deviation

The rounding rule for the standard deviation is the same as that for the mean. The final answer should be rounded to one more decimal place than that of the original data.

Example 3.4-3

Find the variance and standard deviation for the data set for brand A paint in Example 10.

10, 60, 50, 30, 40, 20

Solution

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each value, and place the result in column B of the table.

Step 3 Square each result and place the squares in column C of the table.

A X	B $X-\mu$	C $(X-\mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Step 4 Find the sum of the squares in column C.

$$\Sigma(X-\mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

Step 5 Divide the sum by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} = \frac{250}{6} = 41.7$$

Step 6 Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} = \sqrt{41.7} = 6.5$$

Hence, the standard deviation is 6.5.

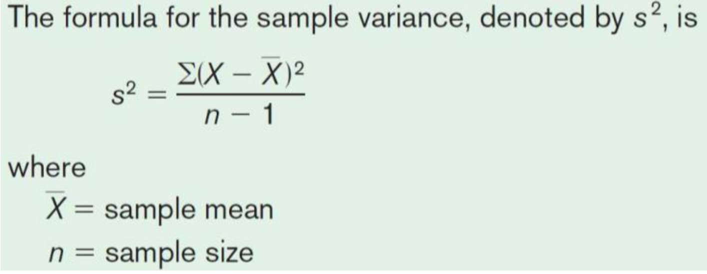
- Since the standard deviation of brand A is 17.1 and the standard deviation of brand B is 6.5, the data are more variable for brand A.
- ***In summary, when the means are equal, the larger the variance or standard deviation is, the more variable the data are.***

3.4.2.3 Sample Variance and Standard Deviation (For Ungrouped Data)

- When computing the variance for a sample, one might expect the following expression to be used:

  $$\frac{\Sigma(X-\overline{X})^2}{n}$$

  where $\bar{X}$ is the sample mean and n is the sample size.
- Above expression does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance.
- Therefore, instead of dividing by n, find the variance of the sample by dividing by $n-1$, giving a slightly larger value estimate of the population variance.



The formula for the sample variance, denoted by s^2, is

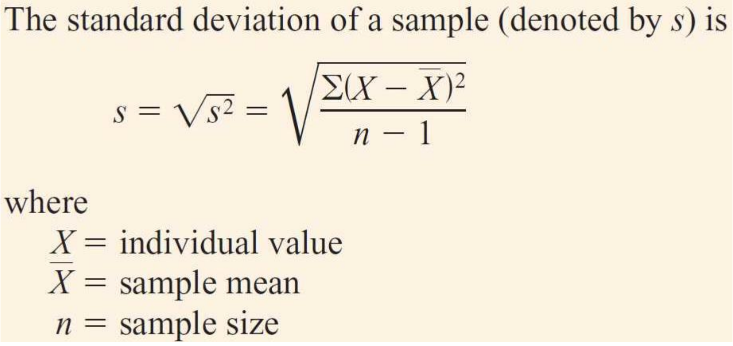
$$s^2 = \frac{\Sigma(X-\overline{X})^2}{n-1}$$

where

$\overline{X}$ = sample mean

n = sample size

- To find the standard deviation of a sample, you must take the square root of the sample variance, which was found by using the preceding formula.



The standard deviation of a sample (denoted by s) is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$$

where

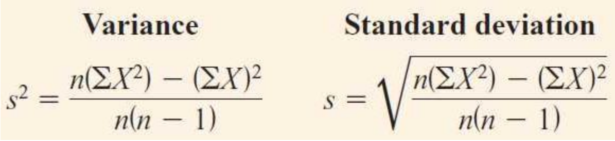
X = individual value

$\overline{X}$ = sample mean

n = sample size

- Shortcut Formulas:

  Shortcut formulas for computing the variance and standard deviation are presented below and will generally be used later.



Variance **Standard deviation**

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)} \qquad s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$$

Example 3.4-4

Find the sample variance and standard deviation for the following sample data:

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

Solution

Step 1 Find the sum of the values.

$\Sigma X = 11.2 + 11.9 + 12.0 + 12.8 + 13.4 + 14.3 = 75.6$

Step 2 Square each value and find the sum.

$\Sigma X^2 = 11.2^2 + 11.9^2 + 12.0^2 + 12.8^2 + 13.4^2 + 14.3^2 = 958.94$

Step 3 Substitute in the formulas and solve.

$$\begin{aligned} s^2 &= \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)} \\ &= \frac{6(958.94) - 75.6^2}{6(6-1)} \\ &= \frac{5753.64 - 5715.36}{6(5)} \\ &= \frac{38.28}{30} \\ &= 1.276 \end{aligned}$$

The variance is 1.28 rounded.

$s = \sqrt{1.28} = 1.13$

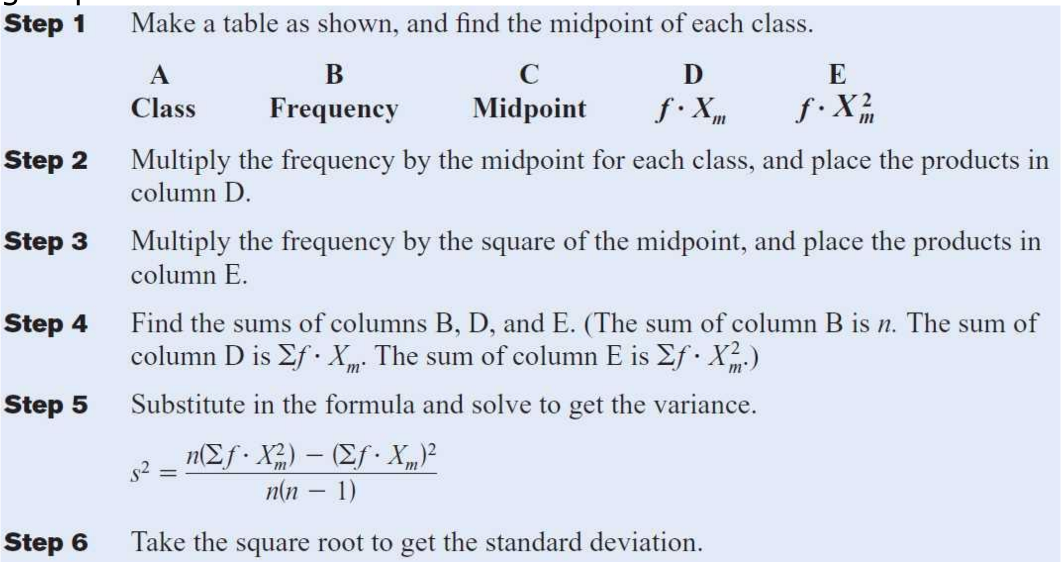
Hence, the sample standard deviation is 1.13.

Note that ΣX^2 is not the same as $(\Sigma X)^2$.

- The notation ΣX^2 means to square the values first, then sum;
- $(\Sigma X)^2$ means to sum the values first, then square the sum.

3.4.2.4 Variance and Standard Deviation for Grouped Data

- The procedure for finding the variance and standard deviation for grouped data is like that for finding the mean for grouped data, and it uses the midpoints of each class.
- The steps for finding the variance and standard deviation for grouped data are summarized in this Procedure Table.



Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is n. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$

Step 6 Take the square root to get the standard deviation.

Example 3.4-5

Find the variance and the standard deviation for the frequency distribution of the data below.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Solution

Step 1 Make a table as shown, and find the midpoint of each class.

A	B	C	D	E
	Frequency	Midpoint		
Class	f	X_m	$f \cdot X_m$	$f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$

Step 4 Find the sums of columns B, D, and E. The sum of column B is n, the sum of column D is $\Sigma f.X_m$, and the sum of column E is $\Sigma f.X_m^2$. The completed table is shown.

A	B	C	D	E
Class	Frequency	Midpoint	$f \cdot X_m$	$f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13{,}310$

Step 5 Substitute in the formula and solve for s^2 to get the variance.

$$s^2 = \frac{n(\Sigma f.X_m^2) - (\Sigma f.X_m)^2}{n(n-1)} = \frac{20 \times 13310 - 490^2}{20 \times (20-1)} = 68.7$$

Step 6 Take the square root to get the standard deviation.

$s = \sqrt{68.7} = 8.3$

Example 3.4-6

Sample data represent strength in ton for wooden piles have been gathered and presented in the histogram indicated in ***Figure 3.4-1***. For these grouped data, determine the mean value and the stranded deviation. Are the computed values exact or approximate? Explain your answer.

Solution

For the ground data indicated in the histogram, the mean value can be determined based on the following relation:

$$X = \frac{\Sigma f . X_m}{n}$$

The mid-point and frequency for each class can be deduced from the indicated diagram:

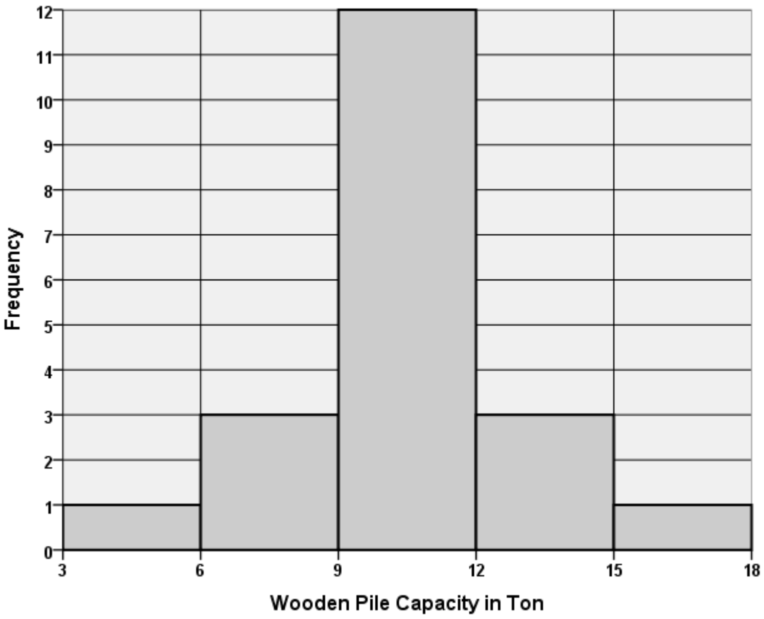


Figure 3.4-1: Histogram for compressive strength of wooden piles.

Classes		Mid-point	Frequency	$f . X_m$
3	6	4.5	1	4.5
6	9	7.5	3	22.5
9	12	10.5	12	126
12	15	13.5	3	40.5
15	18	16.5	1	16.5
		Summation	20	210

$$X = \frac{\Sigma f . X_m}{n} = \frac{210}{20} = 10.5 \, ton$$

The variance, s^2, for the grouped data can be determined based on the following relation:

$$s^2 = \frac{n(\Sigma f . X_m^2) - (\Sigma f . X_m)^2}{n(n-1)}$$

The details for the calculations have been prepared with the referring in below:

Classes		Mid-point, X_m	Frequency, f	$f . X_m$	X_m^2	$f . X_m^2$
3	6	4.5	1	4.5	20.25	20.25
6	9	7.5	3	22.5	56.25	168.75
9	12	10.5	12	126	110.25	1323
12	15	13.5	3	40.5	182.25	546.75
15	18	16.5	1	16.5	272.25	272.25
		Summation	20	210		2331

$$s^2 = \frac{n(\Sigma f . X_m^2) - (\Sigma f . X_m)^2}{n(n-1)} = \frac{20 \times 2331 - (210)^2}{20 \times (20-1)} = 6.63 \, ton$$

$$s = \sqrt{s^2} = \sqrt{6.63} = 2.57 \, ton$$

These statistics are approximated in nature as they have been determined in terms of the grouped data.

Home Work 3.4-1

Data for crane cable strength in ton have been gathered and presented in the histogram of ***Figure 3.4-1***. Use the histogram to determine the mean and the standard deviation for the indicated grouped data. Are the determined values approximate or exact? Example your answer.

Answer

$\bar{X} = 58.4\ ton$

$s = 5.07\ ton$

There are approximated.

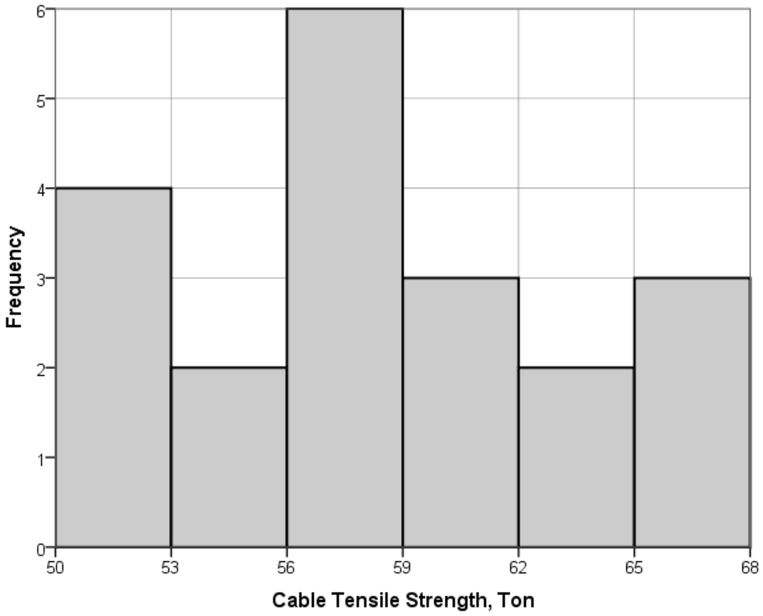


Figure 3.4-2: Histogram for the cable tensile strength, ton.

3.4.2.5 Uses of the Variance and Standard Deviation:

- Variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
- The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or the parts will not fit together.
- Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later in this course.

3.4.3 Coefficient of Variation

- When comparing the relative dispersion of more than one kind of data, it is convenient to have a dimensionless description such as the commonly quoted sample ***coefficient of variation***.
- This quantity ν, read nu, is defined as the ratio of the sample standard deviation, s, to the sample mean, $\bar{x}$.

$$\nu = \frac{s}{\bar{x}}$$

Example 3.4-7

Fifteen reinforced-concrete beams built by engineering students to the same specifications and fabricated from the same batch of concrete were tested in flexure. The observed results of first-crack and ultimate loads, recorded to the nearest 50 lb, are presented in Table 3.4-1 below. Using the coefficient of variation, ν, which one of the loads for first crack or failure load seems more scattered?

Table 3.4-1: Tests of identical reinforced concrete beams.

Beam No.	Load at which the first crack was observed, lb	Failure load, lb
1	5100	9300
2	5800	9300
3	6000	9400
4	6000	9500
5	6000	9550
6	6000	9550
7	6500	9600
8	6500	9900
9	6500	10100
10	7200	10200
11	8450	10200
12	9300	10300
13	9500	10350
14	10350	10500
15	10600	10600

Solution

Compute mean, $\bar{x}$, and standard deviation, s, for each of load for first crack and failure load:

$$\bar{x}_{Load\ for\ first\ crack} = 7320\ lb$$

$$s_{Load\ for\ first\ crack} = 1815\ lb$$

$$\bar{x}_{Failure\ load} = 9890\ lb$$

$s_{Failure\ load} = 455\ lb$

The coefficients of variation would be:

$$v_{Load\ for\ first\ crack} = \frac{1815}{7320} = 0.248$$

$$v_{Load\ for\ first\ crack} = \frac{455}{9890} = 0.0460 \ll v_{Load\ for\ first\ crack}$$

Therefore, first crack loads are ***more variable*** or more difficult to predict closely than failure loads. Such information is important when appearance as well as strength is a design criterion.

Example 3.4-8

For a residential building indicated in ***Figure*** 3.4-3 below, and based on survey of similar buildings, floor live loads, W_{Live}, in kPa, have been gathered and presented in ***Table*** 3.4-2 below. From these data determined ***mean*** value, $\bar{W}_{Live}$, and ***standard deviation***, $s_{W_{Live}}$ for the floor live loads.

Due to building dimensions and the number of its floors, the axial force that is supported by an interior column, $P_{Interior\ column}$, can be estimated based on following relation:

$$P_{Interior\ column} = W_{Live} \times A_{Supported\ by\ an\ interior\ column} \times No.of\ Floors = W_{Live} \times (5 \times 6) \times 3$$

Adopting the relation above and based on principles of ***linear transformation***, what are the mean value of axial force supported by an interior column, $\bar{P}_{Interior\ column}$ and its standard deviation, $s_{P_{Interior\ column}}$?

In term of the ***coefficient of variation***, v, which one of the floor live load, W_{Live}, and the column axial force, $P_{Interior\ column}$, has more scatter? Explain your answer.

Table 3.4-2: Floor live loads of residential buildings, in kPa.

1.77	1.56	1.91	1.55	2.00
1.99	1.92	1.71	1.62	1.78
1.93	2.41	2.36	2.25	2.38
1.91	1.95	1.78	2.18	1.93
1.54	1.92	1.89	2.15	2.00

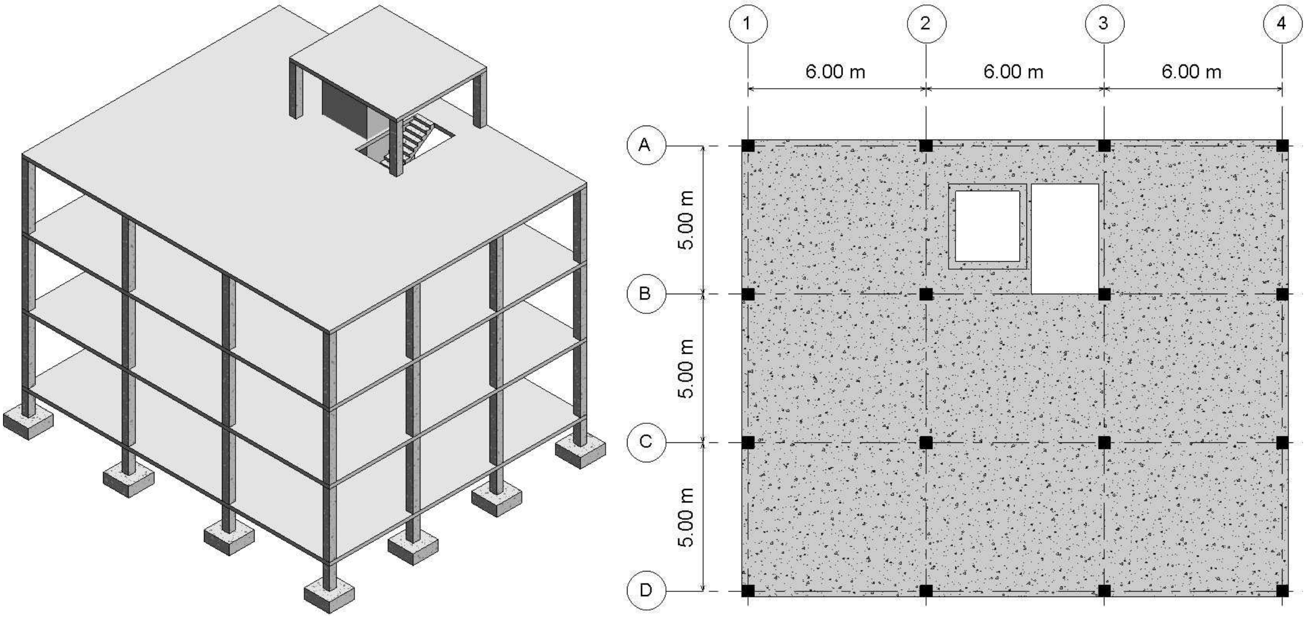


Figure 3.4-3: A residential building.

Solution

Mean can be computed based on following relation:

$$\bar{W}_{Live\ loads} = \frac{\Sigma W_{(Live\ load)_i}}{n} = \frac{48.39}{25} = 1.936\ kPa\ \blacksquare$$

The standard deviation for floor live load would be determined with referring to table below:

No.	*Live Load in kPa*	$W_{Live} - \overline{W}_{Live}$	$(W_{Live} - \overline{W}_{Live})^2$
1	1.77	-0.17	0.02742336
2	1.99	0.05	0.00295936
3	1.93	-0.01	0.00003136
4	1.91	-0.03	0.00065536
5	1.54	-0.40	0.15649936
6	1.56	-0.38	0.14107536
7	1.92	-0.02	0.00024336
8	2.41	0.47	0.22505536
9	1.95	0.01	0.00020736
10	1.92	-0.02	0.00024336
11	1.91	-0.03	0.00065536
12	1.71	-0.23	0.05089536
13	2.36	0.42	0.18011536
14	1.78	-0.16	0.02421136
15	1.89	-0.05	0.00207936
16	1.55	-0.39	0.14868736
17	1.62	-0.32	0.09960336
18	2.25	0.31	0.09884736
19	2.18	0.24	0.05973136
20	2.15	0.21	0.04596736
21	2.00	0.06	0.00414736
22	1.78	-0.16	0.02421136
23	2.38	0.44	0.19749136
24	1.93	-0.01	0.00003136
25	2.00	0.06	0.00414736
Summation	48.39	0.00	1.50

$$s^2_{Live\ load} = \frac{\Sigma(W_{Live} - \overline{W}_{Live})^2}{N-1} = \frac{1.50}{25-1} = 0.0625$$

$$s_{Live\ load} = \sqrt{0.0625} = 0.25\ kPa\ ■$$

With referring to standard form of linear equation:

$$y = k_1 + k_2 x$$

and with adopting the $P_{Interior\ column}$ as the dependent variable, y, and W_{Live} as the independent variable, x, the linear relation would be:

$$P_{Interior\ column} = k_1 + k_2 W_{Live}$$

Comparing this relation with that given in the problem statement:

$$P_{Interior\ column} = W_{Live} \times A_{Supported\ by\ an\ interior\ column} \times No.of\ Floors = W_{Live} \times (5 \times 6) \times 3$$

one concludes that:

$$k_1 = 0,\ k_2 = (5 \times 6) \times 3 = 90\ m^2$$

$$\therefore\ P_{Interior\ column} = 0 + 90\ W_{Live}$$

From relations for linear transformation:

$$\therefore\ \overline{P}_{Interior\ column} = 0 + 90\ \overline{W}_{Live} = 90 \times 1.936 = 174.24\ kN\ ■$$

$$s_P = |k_2| s_{W_{Live}} = 90 \times 0.25 = 22.5\ kN\ ■$$

Finally, the scatter for floor live load, $W_{Live\ load}$, and axial force in an interior column, $P_{Interior\ column}$, in terms of the ***coefficient of variation***, v, would be:

$$v_{W_{Live\ load}} = \frac{s_{Live\ load}}{\overline{W}_{Live\ loads}} = \frac{0.25}{1.936} = 0.129$$

$$\nu_{W_{Live\ load}} = \frac{s_P}{\bar{P}_{Interior\ column}} = \frac{22.5}{174.24} = 0.129$$

As expected, as both phenomena are related linearly, therefore they should have same scatter.

Example 3.4-9

For a sandy soil, ten specimens have been collected and tested. Angle of internal friction, ϕ, in degrees, and elastic modulus,, E_s, in kN/m^2, have been presented in ***Table*** 3.4-2 below.

For these data determine the mean value and standard deviation for each of the angle of internal friction, ϕ, and the elastic modulus, E_s.

In term of the ***coefficient of variation***, ν, which one of the angle of internal friction, ϕ, and the elastic modulus, E_s, has more scatter? Explain your answer.

Table 3.4-3: Angle of friction and elastic modulus for sandy soil.

Angle of Internal Friction, ϕ, in Degrees	**Elastic Modulus, E_s, in kN/m^2**
34.0	24000
33.7	23000
33.7	23000
32.6	19000
36.1	32000
33.2	21000
33.7	23000
31.8	16000
34.0	24000
33.7	23000

Solution

Angle of Internal Friction

For sample of angle of friction would be determined from relation below:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{336.6}{10} = 33.7^o \blacksquare$$

While the sample variation, s^2, would be determined with referring the formula and table below:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Angle of Internal Friction	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
34	0.35	0.1225
33.7	0.05	0.0025
33.7	0.05	0.0025
32.6	-1.05	1.1025
36.1	2.45	6.0025
33.2	-0.45	0.2025
33.7	0.05	0.0025
31.8	-1.85	3.4225
34	0.35	0.1225
33.7	0.05	0.0025
	$\sum_{i=1}^{n}(X_i - \bar{X})^2$	10.985

Therefore

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{10.985}{10-1} = 1.22$$

Finally, the standard deviation would be:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{1.22} = 1.10^o\blacksquare$$

Elastic Modulus

For sample of elastic modulus would be determined from relation below:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{228000}{10} = 22800\,MPa\blacksquare$$

While the sample variation, s^2, would be determined with referring the formula and table below:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Angle of Internal Friction	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
24000	1200	1440000
23000	200	40000
23000	200	40000
19000	-3800	14440000
32000	9200	84640000
21000	-1800	3240000
23000	200	40000
16000	-6800	46240000
24000	1200	1440000
23000	200	40000
	$\sum_{i=1}^{n}(X_i - \bar{X})^2$	151600000

Therefore

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{151600000}{10-1} = 16844444\ \text{MPa}$$

Finally, the standard deviation would be:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{16844444} = 4104\,MPa\blacksquare$$

Assessment of Data Scatter in Terms of Coefficient of Variation

$$v_{Angle\ of\ friction} = \frac{s}{\bar{X}} = \frac{1.10}{33.7} = 0.0326$$

$$v_{Elastic\ modulus} = \frac{s}{\bar{X}} = \frac{4104}{22800} = 0.18$$

Therefore, the ***Coefficient of Variation*** indicates that ***the elastic modulus is more scatter that the angle of variation***.

3.4.4 Additional Example

- This article presents general examples that show a detailed data analysis including the measures of central tendency and the measures of variation.
- These examples have been prepared based on quizzes and exams of the previous years.

Example 3.4-10

Data below represent wind speed in *mph* at a specific site. For these data compute: mean, median, mode, range, and standard deviation.

Table 3.4-4: Wind speed in *mph* at a specific site.

86	133	110	151	132
137	147	165	152	156
128	154	169	162	116
137	168	147	139	119
118	126	143	135	151

Solution

Mean can be computed based on following:

$\bar{x} = \frac{\Sigma x_i}{n} = 139\, mph$ ■

For median value, put data in ascending or descending order. As data number is odd, median value would be:

$Median = 139$ ■

86
110
116
118
119
126
128
132
133
135
137
137
139
143
147
147
151
151
152
154
156
162
165
168
169

The data is multimodal where each of 137, 147, and 151 occurs two times in data.

$Range\ =\ Maximum\, Value - \ Minimum\, Value\ = 169 - 86 = 83\, mph$ ■

Wind Speed, mph	$x - \bar{x}$	$(x - \bar{x})^2$
86	-53	2786.6
110	-29	859.4

Wind Speed, mph	$x - \bar{x}$	$(x - \bar{x})^2$
116	-23	530.4
118	-21	432.4
119	-21	422.5
126	-13	170.3
128	-11	120.2
132	-7	47.5
133	-7	42.8
135	-4	19.7
137	-3	6.3
137	-2	5.7
139	0	0.1
143	4	14.2
147	7	55.9
147	8	57.9
151	12	134.4
151	12	147.3
152	13	177.2
154	14	209.0
156	17	274.9
162	22	502.1
165	26	650.5
168	28	809.8
169	30	882.6
	Σ	**9360**

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{N-1} = \frac{9360}{25-1} = 390$$

$$s = \sqrt{390} = 19.75\ mph \ ■$$

Example 3.4-11

Data presented in ***Table*** 3.4-5 below represent culvert discharge in $\boldsymbol{m^3}$ per second. For these data, compute ***mean***, ***median***, ***range***, ***variance***, and ***standard deviation***.

Table 3.4-5: Discharge through a culvert, in $m^3\ per\ second$. Data for Example 3.4-11.

1.77	1.56	1.91	1.55	2.00
1.99	1.92	1.71	1.62	1.78
1.93	2.41	2.36	2.25	2.38
1.91	1.95	1.78	2.18	1.93
1.54	1.92	1.89	2.15	2.00

Solution

Mean for ungrouped data is:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{48.39}{25} = 1.936\ m^3\ per\ sec.$$

To determine median, put data in order. The median would be mid-point of the array as indicated in Table 3.4-6 below.

Table 3.4-6: Data for Example 3.4-11 presented in order.

1.54
1.55
1.56
1.62
1.71
1.77

1.78
1.78
1.89
1.91
1.91
1.92
1.92
1.93
1.93
1.95
1.99
2.00
2.00
2.15
2.18
2.25
2.36
2.38
2.41

Range for data:

$Maximum\ value = 2.41\ m^3\ per\ sec.$

$Minimum\ value = 1.54\ m^3\ per\ sec.$

$Range = Maximum\ value - Minimum\ value = 2.41 - 1.54 = 0.87\ m^3\ per\ sec.$

Variance and Standard Deviation:

For ungrouped data, variance, s^2, can be computed based on following relation:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

In tabulated form, calculations are presented in Table 3.4-7 below.

$$s^2 = \frac{1.495}{25 - 1} = 0.0623$$

$$s = \sqrt{0.0623} \approx 0.25\ m^3\ per\ sec.$$

Table 3.4-7: Calculations of variance for data of Example 3.4-11.

Discharge through Culvert, in $m^3\ per\ sec.$	$(x - \bar{x})$	$(x - \bar{x})^2$
1.77	-0.17	0.02742336
1.99	0.05	0.00295936
1.93	-0.01	0.00003136
1.91	-0.03	0.00065536
1.54	-0.40	0.15649936
1.56	-0.38	0.14107536
1.92	-0.02	0.00024336
2.41	0.47	0.22505536
1.95	0.01	0.00020736
1.92	-0.02	0.00024336
1.91	-0.03	0.00065536
1.71	-0.23	0.05089536
2.36	0.42	0.18011536
1.78	-0.16	0.02421136
1.89	-0.05	0.00207936
1.55	-0.39	0.14868736
1.62	-0.32	0.09960336
2.25	0.31	0.09884736
2.18	0.24	0.05973136

Discharge through Culvert, in m^3 $per\ sec.$	$(x-\bar{x})$	$(x-\bar{x})^2$
2.15	0.21	0.04596736
2.00	0.06	0.00414736
1.78	-0.16	0.02421136
2.38	0.44	0.19749136
1.93	-0.01	0.00003136
2.00	0.06	0.00414736
		$\Sigma = 1.495$

Example 3.4-12

During a survey for output of a bulldozer when working in a sandy soil, following data have been collected:

Output Rate in yd^3 per hr.			
23	275	82	33
148	38	405	43
108	74	39	254
75	145	77	52
108	14	71	10

For above data:

- Compute mean, median, and mode.
- Compute range, variance, and standard deviation.

Solution

Compute mean, median, and mode:

$Sum = 2074$

$$mean = \frac{2074}{20} = 103.7$$

Median:

Put data in an order form:

Data in Order Form
10
14
23
33
38
39
43
52
71
74
75
77
82
108
108
145
148
254
275
405

$$Median = \frac{74 + 75}{2} = 74.5$$

Mode:

Mode = 108

Compute range, variance, and standard deviation:

Maximum = 405

Minimum = 10

Range = 395

$$Variance \ = \frac{(x-\bar{x})^2}{n-1}$$

Output Rate	$x-\bar{x}$	$(x-\bar{x})^2$
108	4.3	18.49
10	-93.7	8779.69
14	-89.7	8046.09
23	-80.7	6512.49
33	-70.7	4998.49
38	-65.7	4316.49
39	-64.7	4186.09
43	-60.7	3684.49
52	-51.7	2672.89
71	-32.7	1069.29
74	-29.7	882.09
75	-28.7	823.69
77	-26.7	712.89
82	-21.7	470.89
108	4.3	18.49
145	41.3	1705.69
148	44.3	1962.49
254	150.3	22590.09
275	171.3	29343.69
405	301.3	90781.69
	Sum	193576.2
	Variance	10188.22105

$$Variance \ = \frac{193576.2}{19} = 10188$$

$$Standard\ Deviation \ = s = \sqrt{Variance}$$

$$s = \sqrt{10188} = 100.9$$

Example 3.4-13

A tire manufacture wants to determine the inter diameter of a certain grade of tire. Ideally, the diameter would be 570mm. The data are as follows:

572	572	573	568	569	575	565	570

For above data:

- Compute mean, median, and mode.
- Compute range, variance, and standard deviation.

Solution

Mean:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{4564}{8} = 570.5$$

Median:

Put data in order form:

Data in Order Form
565
568
569
570
572
572
573
575

$$Median\ = \frac{570 + 572}{2} = 571$$

Mode:

$Mode\ =\ 572$

Range:

$Range\ =\ Maximum - \ Minimum$

$Range = \ \ 575 - 565 = 10$

Variance:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{\mathrm{n} - 1}$$

x	$(x - \bar{x})^2$
572	2.25
572	2.25
573	6.25
568	6.25
569	2.25
575	20.25
565	30.25
570	0.25
	Sum = 70

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{\mathrm{n} - 1} = \frac{70}{8 - 1} = 10$$

Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{10} = 3.16$$

Example 3.4-14

Table 3.4-8 presents the piezometer gauge pressures at a critical location (ordered data in inches of mercury). Compute ***sample mean***, $\bar{X}$, ***median***, ***mode, range***, and ***standard deviation***. Use relation between ***mean***, ***mode***, and ***median*** to ***estimate shape of the histogram***.

Table 3.4-8: Piezometer gauge pressures, inch of mercury.

12.01	12.08	12.18	12.23	12.27
12.37	12.49	12.53	12.58	12.69
12.76	12.83	12.84	12.88	12.88
12.90	12.92	13.00	13.08	13.35

Solution

Mean for ungrouped data is:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{252.87}{20} = 12.64\ \text{inches of mercury}$$

The Mode:

As only ***12.88*** value occurs twice, therefore the mode of data is ***12.88 inches of mercury***.

The Median:

As data have been already put data in order, therefore, the median, that is the mid-point of the array, would be:

$$Median = \frac{12.69 + 12.76}{2} = 12.73 \text{ inches of mercury}$$

The Range of data:

$$Maximum\ value = 13.35\ inches\ of\ mercury$$
$$Minimum\ value = 12.01\ inches\ of\ mercury$$
$$Range = Maximum\ value - Minimum\ value = 13.35 - 12.01 = 1.34 \text{ inches of mercury}$$

Variance and Standard Deviation:

For ungrouped data, variance, s^2, can be computed based on following relation:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

In tabulated form, calculations are presented in Table 3.4-9.

$$s^2 = \frac{2.519}{20 - 1} = 0.1325$$

$$s = \sqrt{0.1325} \approx 0.364 \text{ inches of mercury}$$

Table 3.4-9: Calculations of variance for data of Example 3.4-14.

X	$(X - \bar{X})$	$(X - \bar{X})^2$
12.01	-0.63	0.401
12.37	-0.27	0.075
12.76	0.12	0.014
12.90	0.26	0.066
12.08	-0.56	0.318
12.49	-0.15	0.024
12.83	0.19	0.035
12.92	0.28	0.076
12.18	-0.46	0.215
12.53	-0.11	0.013
12.84	0.20	0.039
13.00	0.36	0.127
12.23	-0.41	0.171
12.58	-0.06	0.004
12.88	0.24	0.056
13.08	0.44	0.191
12.27	-0.37	0.140
12.69	0.05	0.002
12.88	0.24	0.056
13.35	0.71	0.499
	Summation	2.519

Estimate the shape of the histogram:

As

$$\bar{X} < Median < Mode$$

therefore, the histogram shape would be as indicated in ***Figure*** 3.4-4 below.

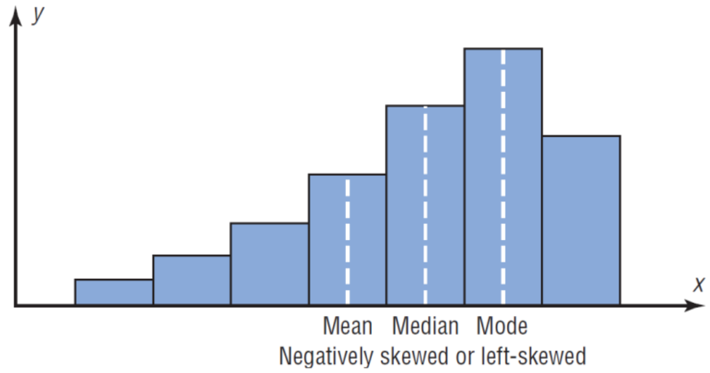


Figure 3.4-4: Estimate histogram of Example 3.4-14.

Example 3.4-15

Time in hours for two steps, namely, load and haul, of a heavy construction process have been presented in Table 3.4-10. Find the sample means, median, mode, standard deviations, and coefficients of variation for each of the two steps in a cycle. Which step is "more variable?"

Table 3.4-10: Time in hours for load and haul for a heavy construction process.

Load					
1.58	1.84	1.80	2.06	1.85	2.00
1.72	1.74	1.67	1.92	2.23	1.88
2.52	1.71	2.38	2.03	1.94	1.93

Haul					
2.08	2.08	2.17	2.07	1.96	2.13
1.95	2.15	2.25	2.18	2.16	2.03
2.18	2.20	1.99	2.17	1.99	1.99

Solution

Mean for ungrouped data is:

$$\bar{X}_{Load} = \frac{\Sigma X}{n} = \frac{34.80}{18} = 1.93 \text{ hours}$$

$$\bar{X}_{Haul} = \frac{\Sigma X}{n} = \frac{37.73}{18} = 2.10 \text{ hours}$$

The Mode:

The load data has no mode. While the model for haul data is 1.99 hours.

The Median:

After sorting, the data would be:

Load	Haul
1.58	1.95
1.67	1.96
1.71	1.99
1.72	1.99
1.74	1.99
1.80	2.03
1.84	2.07
1.85	2.08
1.88	*2.08*
1.92	*2.13*
1.93	2.15
1.94	2.16
2.00	2.17
2.03	2.17
2.06	2.18
2.23	2.18
2.38	2.20
2.52	2.25

$$Median_{Load} = \frac{1.88 + 1.92}{2} = 1.90 \text{ hour} \quad Median_{Haul} = \frac{2.08 + 2.13}{2} = 2.11 \text{ hour}$$

Standard Deviation:

For ungrouped data, variance, s^2, can be computed based on following relation:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

to be:

$$s_{Load} = 0.246 \text{ hour}$$

$$s_{Haul} = 0.093 \text{ hour}$$

Coefficients of variation

$$\nu_{Load} = \frac{s_{Load}}{\bar{X}_{Load}} = \frac{0.246}{1.93} = 0.127$$

$$\nu_{Haul} = \frac{s_{Haul}}{\bar{X}_{Haul}} = \frac{0.093}{2.10} = 0.044$$

Based on the coefficients of variation, the load process is more variable than the haul process.

3.5 LINEAR TRANSFORMATION

This articles shows how mean, standard deviation, median, mode, and range are affected when data x are linearly transformed into y according to ***Eq. 3.5-1*** below.

$$y = k_1 + k_2 x \qquad \text{Eq. 3.5-1}$$

The coefficient k_1 is ***function intercept*** and it represents a ***shifting scale***. While the coefficient k_2 is ***function slope*** and it represents a ***scaling factor***.

The mean value of "y" in terms of mean value of "x" would be,

$$\bar{y} = \frac{\sum_1^n y_i}{n} = \frac{\sum_1^n (k_1 + k_2 x_i)}{n} = \frac{nk_1}{n} + \frac{k_2 \sum_i^n x_i}{n}$$

$$\bar{y} = k_1 + k_2 \bar{x} \qquad \text{Eq. 3.5-2}$$

Regarding to standard deviation,

$$s_y = \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum_1^n \left((\cancel{k_1} + k_2 x_i) - (\cancel{k_1} + k_2 \bar{x})\right)^2}{n-1}} = k_2 \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}}$$

$$s_y = k_2 s_x$$

As standard deviation is a positive value, then the final form of transformation would be:

$$s_y = |k_2| s_x \qquad \text{Eq. 3.5-3}$$

In the same approach, one can prove the following relations:

$$Mode_y = \mathrm{k}_1 + k_2 Mode_x \qquad \text{Eq. 3.5-4}$$

$$Median_y = \mathrm{k}_1 + k_2 Mode_x, \qquad k_2 \; is \; positive \qquad \text{Eq. 3.5-5}$$

$$Range_y = |k_2| Range_x \qquad \text{Eq. 3.5-6}$$

Example 3.5-1

Data below represents a sample of person's weight with mean of 60.9 kg and standard deviation of 5.7 kg. Transform mean and standard deviation into pound value.

	Weight, kg
	55
	67
	66
	64
	61
	53
Mean	60.9
Standard Deviation	5.7

Solution

Pound is related to kilogram based on following linear relation:

$lb = kg \times 2.2$

then

$k_1 = 0, k_2 = 2.2$

$\bar{x}_{in\,pound} = 2.2 \times 60.9 = 134\; lb, s_{in\,pound} = 2.2 \times 5.7 = 12.5\; lb$

Example 3.5-2

Data below represents temperature in Celsius for six different sites. Transform mean and standard deviation into Fahrenheit. Fahrenheit and Celsius scales are related through following linear relation:

$$T_F = 32 + \frac{9}{5} T_C$$

	Temperature, °C
	26
	35
	35
	33
	30
	25
Mean	30.7
Standard Deviation	4.6

Solution

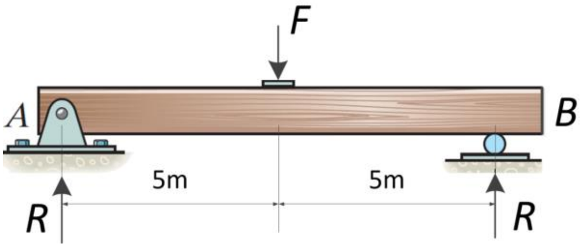
$$k_1 = 32, k_2 = \frac{9}{5}$$

Then

$$\bar{x}_F = 32 + \frac{9}{5} \times 30.7 = 87.3\; {}^oF, s_F = \frac{9}{5} \times 4.6 = 8.3\; {}^oF$$

Example 3.5-3

For the simply supported beam shown, the applied force "F" has a mean value of 50 kN and a standard deviation of 6 kN. What are the mean value and standard deviation for the reaction "R"?



Solution

Based on equilibrium conditions, reaction "R" relates to applied force "F" by the following linear relation:

$R = \frac{1}{2}F$

Therefore

$k_1 = 0$ and $k_2 = \frac{1}{2}$

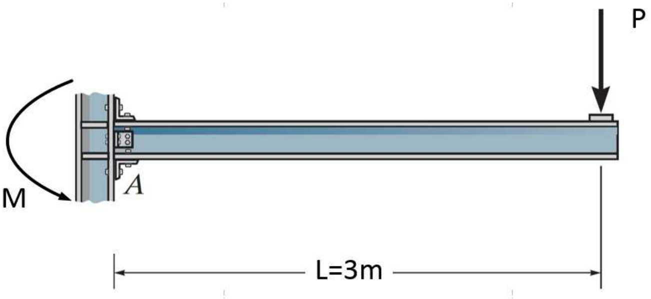
$\bar{x}_{for\,R} = \frac{1}{2}\bar{x}_{for\,F} = \frac{1}{2} \times 50 = 25\,kN,\ \ s_{for\,R} = \frac{1}{2}s_{for\,F} = \frac{1}{2} \times 6 = 3\,kN$

Example 3.5-4

Applied force, $\boldsymbol{P}$, and fixed end moment, $\boldsymbol{M}$, for cantilever beam shown are related to each other based on following relation:

$M = PL = P \times 3 = 3P$

The applied force, P, has a mean value of 12 kN and a standard deviation, s, of 5 kN. Based on linear transformation, what are the mean value and standard deviation of moment, M?



Solution

Based on linear transformation,

$\bar{M} = k_1 + k_2\bar{P}$

Based on linear relation between P and M:

$k_1 = 0, k_2 = 3\ \Rightarrow \bar{M} = 0 + 3\bar{P} = 3 \times 12 = 36\,kN.m$ ■

Standard deviations for bending moment and applied force are related to each other based on following relation:

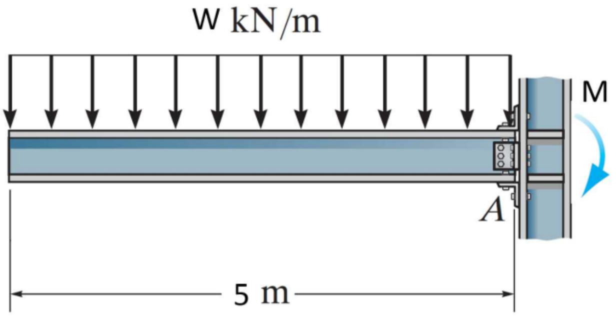
$s_M = |k_2|s_P \Rightarrow s_M = 3 \times 5 = 15\,kN.m$ ■

Example 3.5-5

Uniformly distributed load, $\boldsymbol{W}$, and fixed end moment, $\boldsymbol{M}$, for cantilever beam shown are related based on following relation:

$M = \frac{WL^2}{2} = \frac{W \times 5^2}{2} = 12.5W$

The uniform load, W, has a mean value of 40 kN/m and a standard deviation, s, of 5 kN/m. Based on linear transformation, what are the mean value and standard deviation of moment, M?.



Solution

Based on linear transformation,

$\bar{M} = k_1 + k_2\bar{W}$

Based on linear relation between P and M:

$k_1 = 0, k_2 = 12.5\ \Rightarrow \bar{M} = 0 + 12.5\bar{W} = 12.5 \times 40 = 500\,kN.m$ ■

Standard deviations for bending moment and uniformly distributed load are related based on following relation:

$s_M = |k_2|s_W \Rightarrow s_M = 12.5 \times 5 = 62.5\,kN.m$ ■

3.6 OTHER MEASURES*

In addition the measures of central tendency and the measures of variation that have been discussed in Section 3.3 and Section 3.4, other two measures, namely the measure of asymmetry and the measure of peakdness, are necessary for full description of the data.

3.6.1 MEASURE OF ASYMMETRY

- The ***sample coefficient of skewness***, g_1, determined from relation below:

$$g_1 = \frac{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \qquad \text{Eq. 3.6-1}$$

  is nondimensional used to provides a measure of the degree of asymmetry about the mean of the data.
- The coefficient is positive when more data are larger than the mean while it is negative when more data is lower than the mean.
- Zero skewness results from symmetry but does not necessarily imply it.

Example 3.6-1

Compute the coefficient of skewness, g_1, for beam failure load presented in Table 3.4-1 above.

Solution

After computing of mean and standard deviation for the failure load:

$\bar{x}_{Failure\ load} = 9890\ lb$

$s_{Failure\ load} = 455\ lb$

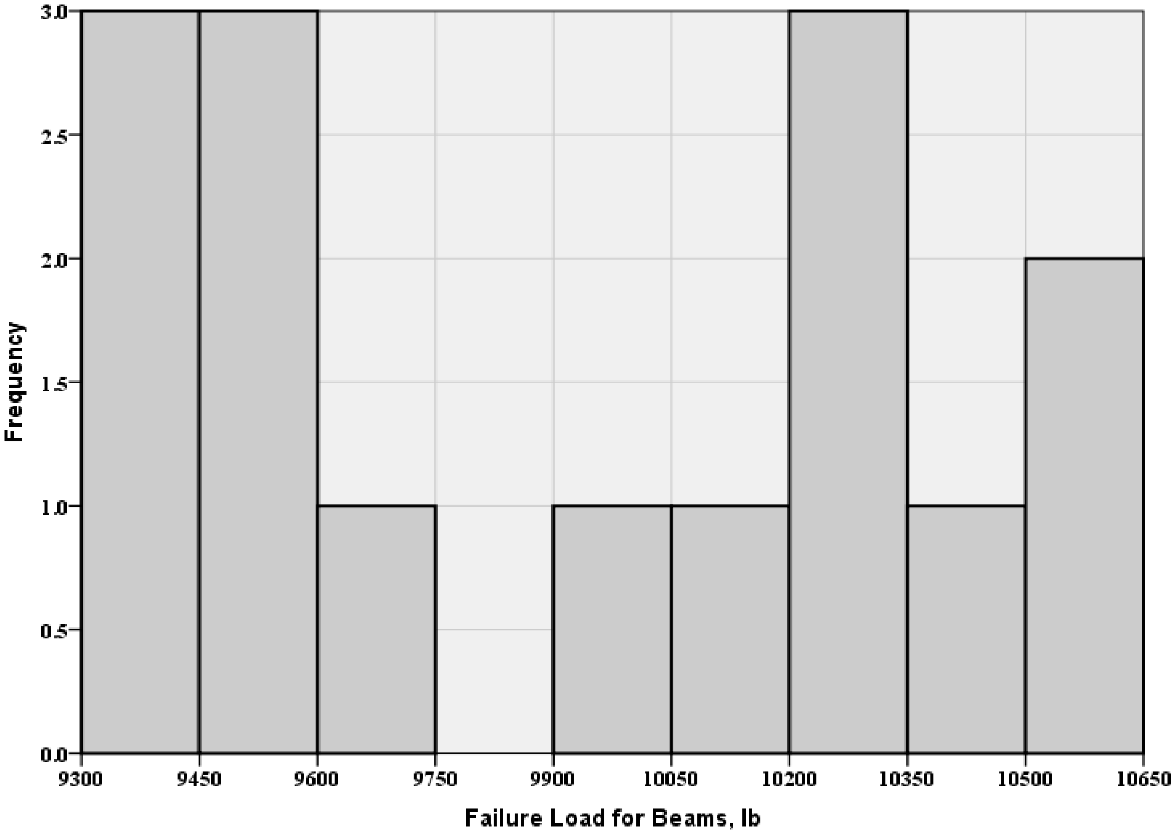
The $(x_i - \bar{x})^3$ for each test and their summation $\sum_{i=1}^{15}(x_i - \bar{x})^3$ can be determined with referring to table below:

Beam No.	Failure load, lb	$x_i - \bar{x}$	$(x_i - \bar{x})^3$
1	9300	-590	-205379000
2	9300	-590	-205379000
3	9400	-490	-117649000
4	9500	-390	-59319000
5	9550	-340	-39304000
6	9550	-340	-39304000
7	9600	-290	-24389000
8	9900	10	1000
9	10100	210	9261000
10	10200	310	29791000
11	10200	310	29791000
12	10300	410	68921000
13	10350	460	97336000
14	10500	610	226981000
15	10600	710	357911000
		$\sum_{i=1}^{15}(x_i - \bar{x})^3$	129270000

Then the coefficient of skewness would be:

$$g_1 = \frac{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} = \frac{\left(\left(\frac{1}{15}\right) \times 129270000\right)}{455^3} = 0.0915$$

Therefore, data for beam failure load is skewed to the right of the mean value. Same conclusion can be drawn with referring to data histogram presented in below.



3.6.2 Measures of Peakedness

3.6.2.1 Basic Definition

- The extent of the relative steepness of ascent in the vicinity and on either side of the mode in a histogram or frequency polygon is said to be a measure of its ***peakedness*** or ***tail weight***.
- This is quantified by the dimensionless sample ***coefficient of kurtosis*** (its is a Greek word with meaning "humped"), which is defined for a sample of observations, x_1, x_2, . . . , x_n by:

$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} \qquad \text{Eq. 3.6-2}$$

Example 3.6-2

Two samples have been selected randomly from two different concrete batches. Cylindrical compressive strength has been tested and presented in MPa as indicated below. For each sample, draw histogram and compute coefficient of kurtosis and comment on peakedness of data based on graphical aspects and numerical aspects. In your drawing of histogram, adopt a class width of 3.0 MPa for each sample data.

Specimen No.	Sample 1	Sample 2
1	26	23
2	25	25
3	30	32
4	26	30
5	18	30
6	29	26
7	28	30
8	28	26
9	26	27
10	32	29
11	24	25
12	32	25

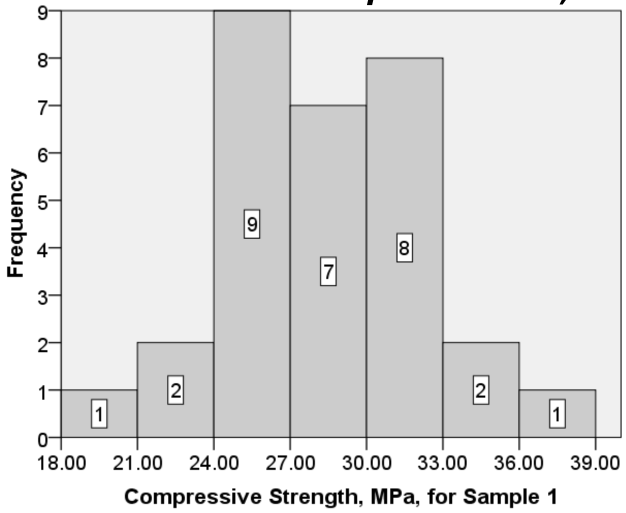
Specimen No.	Sample 1	Sample 2
13	29	28
14	28	27
15	32	27
16	21	29
17	34	26
18	26	29
19	23	30
20	26	25
21	32	33
22	37	27
23	26	28
24	30	26
25	24	26
26	29	27
27	33	24
28	29	29
29	31	32
30	30	29

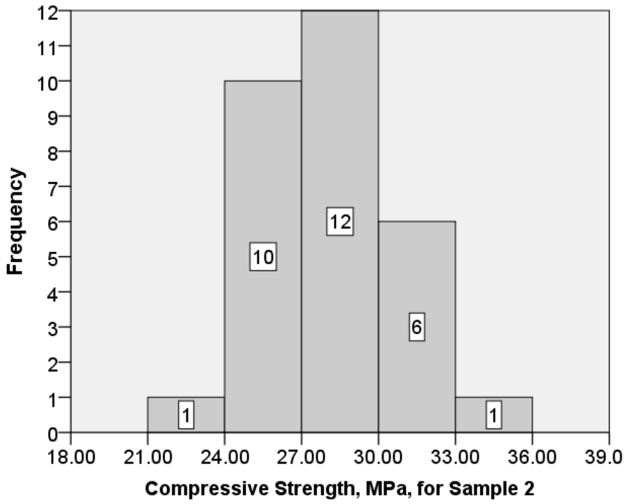
Solution

Graphical Aspects

Histogram for each sample data can be drawn as discussed in ***Chapter 2***.

As indicated in below, for comparison purposes both histograms have been drawn starting from same point. The histograms indicate that *Sample 1* is flatter than *Sample 2*. The comparison gives ***an indication that Sample 1 has larger standard deviation and smaller peakedness, i.e. smaller coefficient of kurtosis***.





Numerical Aspects

Regarding to the coefficient of kurtosis, for *Sample 1*, the mean and standard deviation can be computed as discussed previously:

$\bar{x}_{Sample\ 1} = 28\ MPa$

$s_{Sample\ 1} = 4.0\ MPa$

then, the coefficient of kurtosis can be determined with referring to table below:

$$g_{2\ for\ Sample\ 1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} = \frac{24879}{30 \times 4^4} = 3.23 \blacksquare$$

No.	Sample 1	$(x_i - \bar{x})^4$
1	26	10
2	25	47
3	30	28
4	26	6

No.	Sample 1	$(x_i - \bar{x})^4$
5	18	11124
6	29	4
7	28	0
8	28	0
9	26	13
10	32	210
11	24	197
12	32	267
13	29	2
14	28	0
15	32	194
16	21	2783
17	34	1687
18	26	7
19	23	435
20	26	16
21	32	293
22	37	6165
23	26	14
24	30	36
25	24	400
26	29	2
27	33	773
28	29	0
29	31	134
30	30	33
	$\sum_{i=1}^{n}(x_i - \bar{x})^4$	24879

In the same approach, after computing mean and standard deviation for the *Sample 2*,

$\bar{x}_{Sample\ 2} = 28\ MPa$

$s_{Sample\ 2} = 2.0\ MPa$

it’s coefficient of kurtosis can be determined with referring to table below:

$$g_{2\ for\ Sample\ 2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} = \frac{2104}{30 \times 2^4} = 4.38\ ■$$

No.	Sample 2	$(x_i - \bar{x}')^4$
1	23	495
2	25	69
3	32	199
4	30	9
5	30	22
6	26	15
7	30	32
8	26	10
9	27	0
10	29	2
11	25	81
12	25	63
13	28	0
14	27	0
15	27	4
16	29	1
17	26	35
18	29	4

No.	Sample 2	$(x_i - \overline{x'})^4$
19	30	6
20	25	104
21	33	441
22	27	1
23	28	0
24	26	23
25	26	5
26	27	1
27	24	268
28	29	5
29	32	202
30	29	5
	$\sum_{i=1}^{n}(x_i - \bar{x})^4$	2104

Larger standard deviation and smaller kurtosis for *Sample 1* comparing with those of *Sample 2*, indicate that *Sample 1* has larger variation and smaller peakedness comparing with *Sample 2*.

Engineering Important for Conclusions

Irrespective they have been concluded from graphs or from numerical measures, larger standard deviation and smaller coefficient of kurtosis for *Sample 1* comparing with *Sample 2* indicate that it has been produced under a process with a smaller quality control comparing with that process adopted to produce *Sample 2*.

3.6.2.2 Data Modeling with the Normal Curve

- According to their coefficient of kurtosis, the curves can be classified into three types indicated ***Figure 3.6-1*** below.

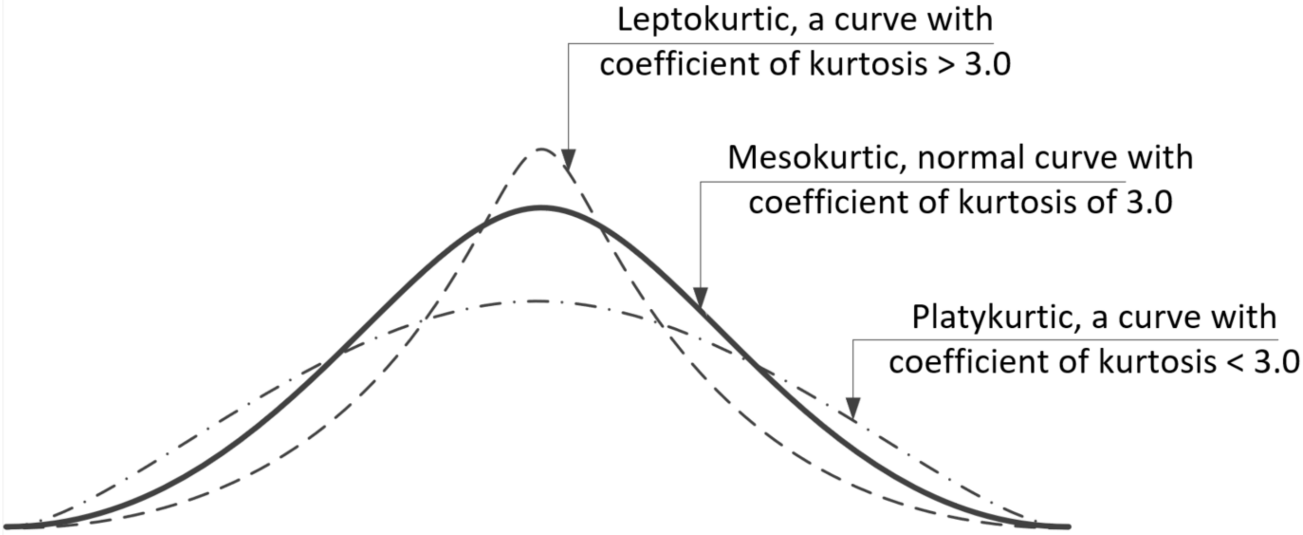
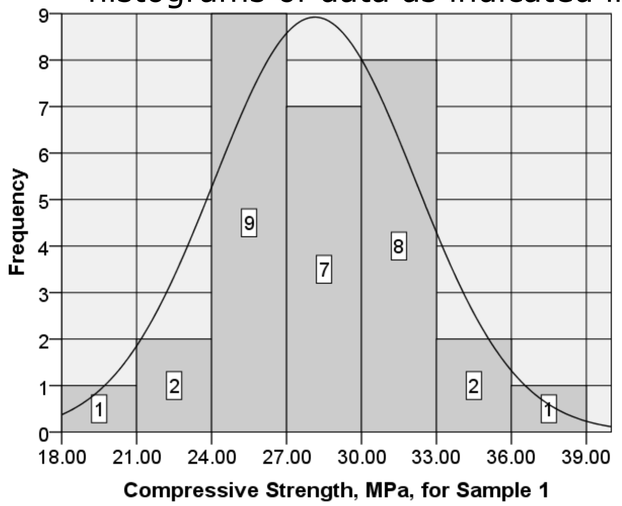
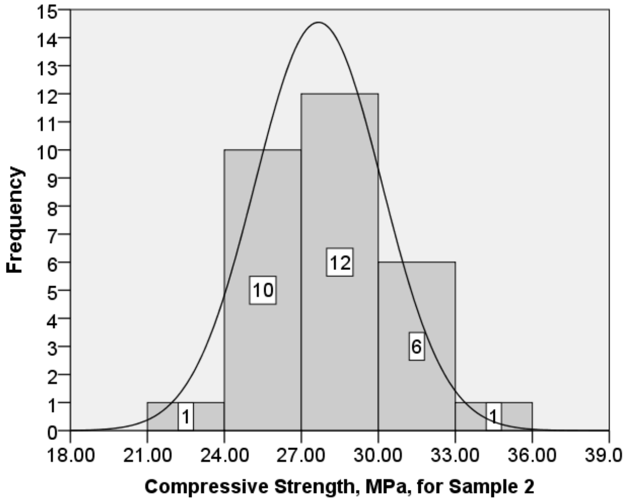

Leptokurtic, a curve with coefficient of kurtosis > 3.0
Mesokurtic, normal curve with coefficient of kurtosis of 3.0
Platykurtic, a curve with coefficient of kurtosis < 3.0

Figure 3.6-1: Three types of curves according their coefficient of kurtosis.

- The ***normal curve***, sometimes called ***bell shape*** or ***Gaussian curve***, is one of the most common probability models. It is usually used to model different natural phenomena and industrial processes.
- The normal curve has a coefficient of kurtosis of ***three***. This basic feature is usually used in statistical analysis to show how data is closed to normal curve through comparing its coefficient of kurtosis with standard value of three.
- Referring to ***Example 3.6-2*** above, one concludes that *Sample 1* that has a coefficient of kurtosis of 3.23 is closer to the normal curve than *Sample 2* that has a coefficient of kurtosis of 4.38.

Therefore, the normal model is more suitable to simulate data of Sample 1 than data of Sample 2. This fact can be concluded from graphical interpretation when the normal curve is added to the histograms of data as indicated in below.





3.7 MEASURES OF POSITION*

- These measures aim to locate a specific data is located relative to the other data in a sample or population. They include:
  - Standard scores,
  - Percentiles,
  - Deciles,
  - Quartiles.
- For example,
  - If a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it.
  - The ***median*** is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one-half of the values fall above it.

3.7.1 STANDARD SCORES

- There is an old saying, "***You can't compare apples and oranges***". However, with the use of statistics, it can be done to some extent.
- Suppose that a student scored 90 on a music test and 45 on an English exam.
  - Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on.
  - However, a comparison of a relative standard similar to both can be made. This comparison uses the mean and standard deviation and is called a ***standard score*** or ***z score***.
- A standard score or z score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. In general, it is computed according to ***Eq. 3.7-1*** below:

$$z = \frac{value - mean}{standard\ deviation} \qquad \text{Eq. 3.7-1}$$

- For samples, the formula Eq. 3.7-1 would be:

$$z = \frac{x - \bar{x}}{s} \qquad \text{Eq. 3.7-2}$$

- While for population, the formula would be:

$$z = \frac{x - \mu}{\sigma} \qquad \text{Eq. 3.7-3}$$

- Note:
  - If the z score is positive, the score is above the mean.
  - If the z score is 0, the score is the same as the mean.
  - If the z score is negative, the score is below the mean.
  - As indicated in formulation ***Eq. 3.7-1***, standard score, z, is a nondimensional number.
  - It is insensitive for scaling nor for units system adopted.

Example 3.7-1: (Positive z Score)

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

First, find the z scores. For calculus the z score is

$$z = \frac{X - \bar{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

The calculus score of 65 was actually 1.5 standard deviations above the mean of 50.

Example 3.7-2 (Negative z Score)

Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\bar{X} = 40$	$s = 5$
Test B	$X = 94$	$\bar{X} = 100$	$s = 10$

Solution

For test A,

$$z = \frac{X - \bar{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.

3.7.2 PERCENTILES*

Percentiles divide the data set into 100 equal groups.

3.7.2.1 Difference between Percentages and Percentiles:

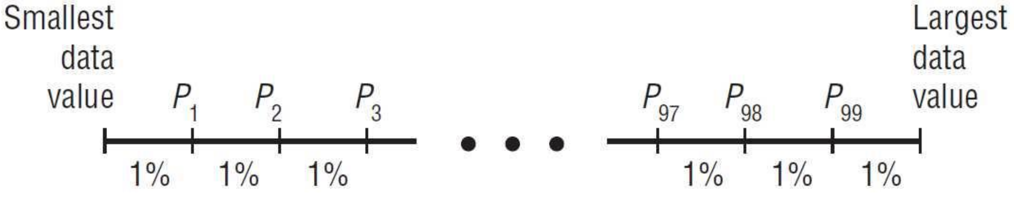
- Percentiles are not the same as percentages.
- That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72.
- There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between.
- On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

3.7.2.2 Percentiles Symbols

Percentiles are symbolized by:

$$P_1, P_2, P_3, \ldots, P_{99}$$

and divide the distribution into 100 groups.



3.7.2.3 Construction of Percentiles Graphs

Percentile graphs use the same values as the cumulative relative frequency graphs described previously, except that the proportions have been converted to percent.

Example 3.7-3 (Construction of Percentiles Graph)

The frequency distribution for the blood pressure readings (in millimeters of mercury, mm Hg) of 200 randomly selected college students is shown here. Construct a percentile graph.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24		
104.5–119.5	62		
119.5–134.5	72		
134.5–149.5	26		
149.5–164.5	12		
164.5–179.5	4		
	200		

Solution

Step 1

Find the cumulative frequencies and place them in column C.

A Class boundaries	B Frequency	C Cumulative frequency
89.5–104.5	24	24
104.5–119.5	62	86
119.5–134.5	72	158
134.5–149.5	26	184
149.5–164.5	12	196
164.5–179.5	4	200
	200	

Step 2

Find the cumulative percentages and place them in column D. To do this step, use the formula

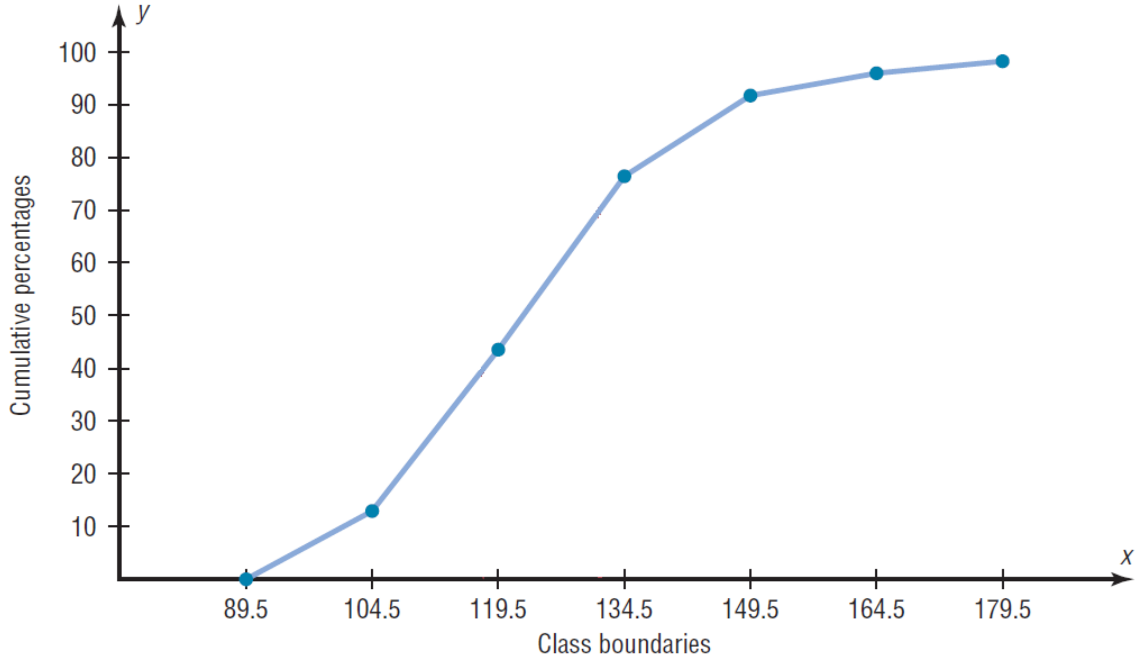
$$Cumulative\ \% = \frac{Cumulative\ freq.}{n} \times 100$$

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24	24	12
104.5–119.5	62	86	43
119.5–134.5	72	158	79
134.5–149.5	26	184	92
149.5–164.5	12	196	98
164.5–179.5	4	200	100
	200		

$$\text{Cumulative \%} = \frac{24}{200} \cdot 100 = 12\%$$

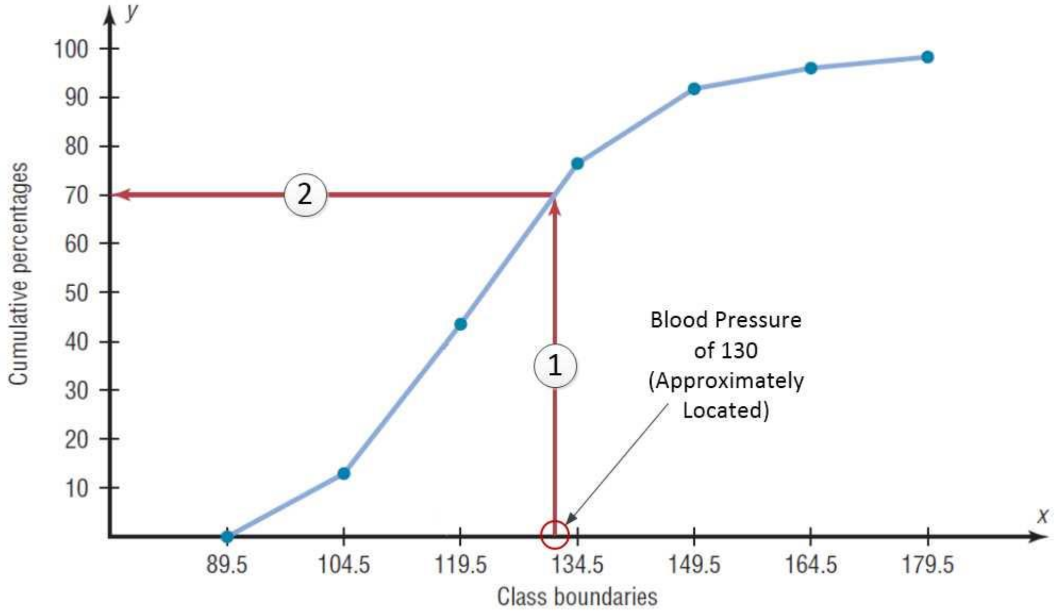
Step 3

Graph the data, using class boundaries for the x axis and the percentages for the y axis, as shown in Figure below.

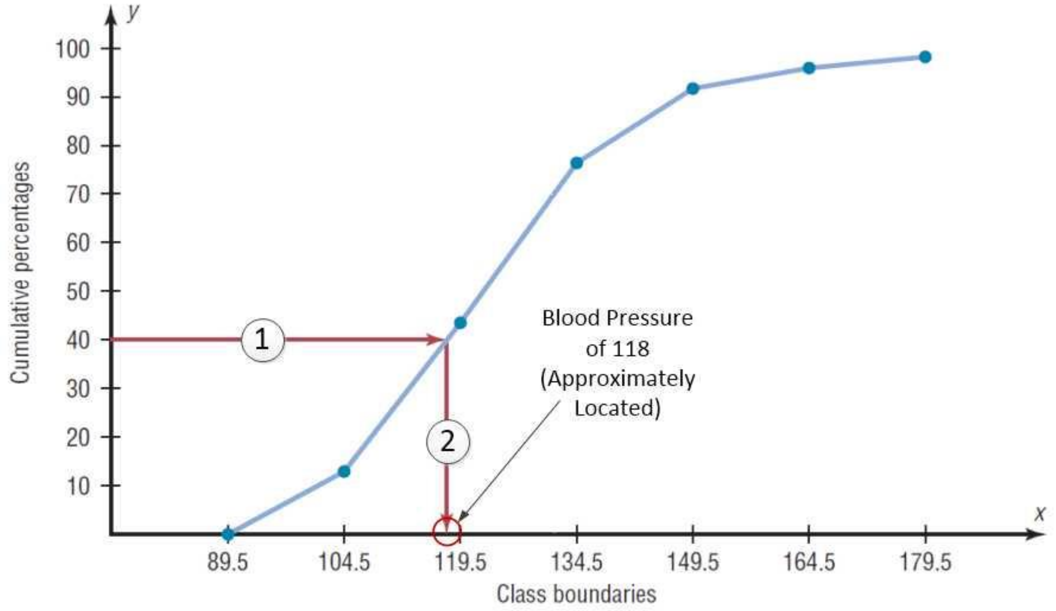


Use of Percentiles Graph:

- Once a percentile graph has been constructed, one can:
  - Find the approximate corresponding percentile ranks for given blood pressure values.
  - Or find approximate blood pressure values for given percentile ranks.
- For example, to find the percentile rank of a blood pressure reading of 130, find 130 on the x axis of Figure below, and draw a vertical line to the graph. Then move horizontally to the value on the y axis. Note that a blood pressure of 130 corresponds to approximately the 70th percentile.



- If the value that corresponds to the 40th percentile is desired, start on the y axis at 40 and draw a horizontal line to the graph. Then draw a vertical line to the x axis and read the value. In below, the 40th percentile corresponds to a value of approximately 118.



3.7.2.4 Percentile Formula

- Finding values and the corresponding percentile ranks by using a graph yields only approximate answers. Several mathematical methods exist for computing percentiles for data.
- These methods can be used to:
  - Find the approximate percentile rank of a data value
  - Or to find a data value corresponding to a given percentile.

3.7.2.5 Formula Approximate Percentile Rank of a Data Value

The percentile corresponding to a given value X is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

Example 3.7-4

A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

Since there are six values below a score of 12, the solution is

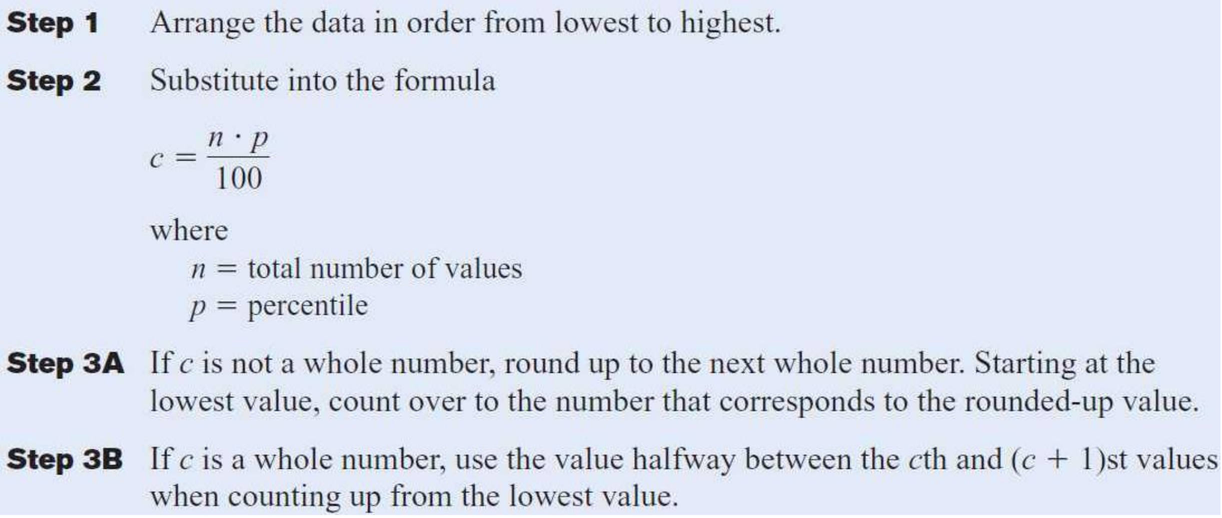
$$\boxed{2, 3, 5, 6, 8, 10,}\ 12, 15, 18, 20$$

$$\text{Percentile} = \frac{\boxed{6} + 0.5}{10} \cdot 100 = 65\text{th percentile}$$

Thus, a student whose score was 12 did better than 65% of the class.

3.7.2.6 Finding a Data Value Corresponding to a Given Percentile:

The steps for finding a value corresponding to a given percentile are summarized below:



Step 1 Arrange the data in order from lowest to highest.

Step 2 Substitute into the formula

$$c = \frac{n \cdot p}{100}$$

where

n = total number of values

p = percentile

Step 3A If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

Step 3B If c is a whole number, use the value halfway between the cth and $(c + 1)$st values when counting up from the lowest value.

Example 3.7-5

Using the scores in previous Example, find the value corresponding to the 25th percentile.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Solution

Step 1

Arrange the data in order from lowest to highest.

$2, 3, 5, 6, 8, 10, 12, 15, 18, 20$

Step 2

Compute

$$c = \frac{n \cdot p}{100}$$

where

$n =$ total number of values

$p =$ percentile

Thus,

$$c = \frac{10 \cdot 25}{100} = 2.5$$

Step 3

If c is not a whole number, round it up to the next whole number; in this case,

$$c = 3$$

Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

Third Data

$2, 3, \boxed{5}, 6, 8, 10, 12, 15, 18, 20$

Example 3.7-6

Using the scores in previous Example, find the value corresponding to the 60th percentile.

$18, 15, 12, 6, 8, 2, 3, 5, 20, 10$

Solution

Step 1

Arrange the data in order from smallest to largest.

$2, 3, 5, 6, 8, 10, 12, 15, 18, 20$

Step 2

Substitute in the formula.

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6$$

Step 3

If c is a whole number, use the value halfway between the c and $c + 1$ values when counting up from the lowest value in this case, the 6th and 7th values.

$$2, 3, 5, 6, 8, \underset{\substack{\nearrow \\ \text{6th value}}}{10}, \underset{\substack{\nwarrow \\ \text{7th value}}}{12}, 15, 18, 20$$

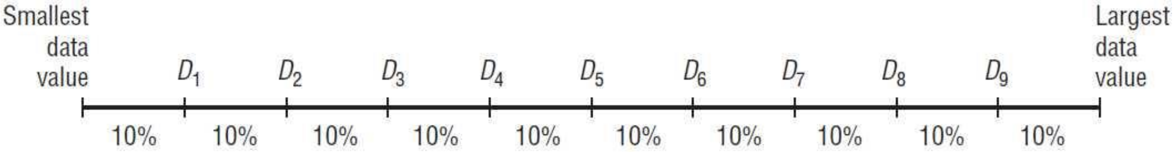
The value halfway between 10 and 12 is 11. Find it by adding the two values and dividing by 2.

$$\frac{10 + 12}{2} = 11$$

Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

3.7.3 Deciles*

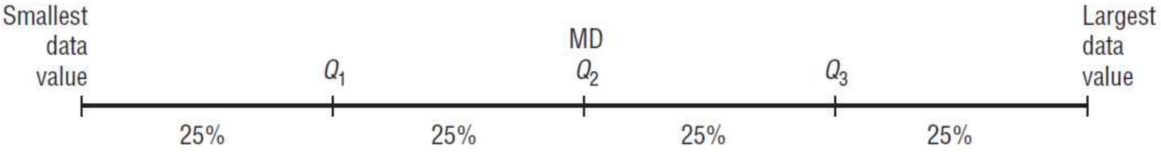
Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2, etc.



3.7.4 Quartiles

3.7.4.1 Definition

Quartiles divide the distribution into four groups, separated by Q_1, Q_2, Q_3.



3.7.4.2 Finding Procedures

Method for finding quartiles is found in this Procedure Table.

Step 1	Arrange the data in order from lowest to highest.
Step 2	Find the median of the data values. This is the value for Q_2.
Step 3	Find the median of the data values that fall below Q_2. This is the value for Q_1.
Step 4	Find the median of the data values that fall above Q_2. This is the value for Q_3.

Example 3.7-7

Find Q_1, Q_2, and Q_3 for the data set

15, 13, 6, 5, 12, 50, 22, 18

Solution

Step 1

Arrange the data in order.

5, 6, 12, 13, 15, 18, 22, 50

Step 2

Find the median (Q_2).

5, 6, 12, 13, 15, 18, 22, 50
↑
MD

$$\text{MD} = \frac{13 + 15}{2} = 14$$

Step 3

Find the median of the data values less than 14.

5, 6, 12, 13
↑
Q_1

$$Q_1 = \frac{6 + 12}{2} = 9$$

Step 4

Find the median of the data values greater than 14.

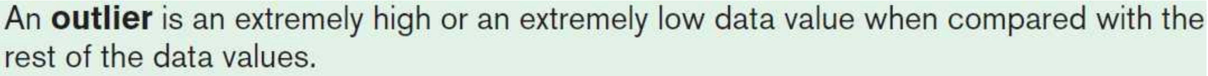
15, 18, 22, 50
↑
Q_3

$$Q_3 = \frac{18 + 22}{2} = 20$$

3.7.5 DETERMINATION OF OUTLIERS

3.7.5.1 Definition

A data set should be checked for extremely high or extremely low values. These values are called ***outliers.***



An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

3.7.5.2 Outliers Importance

- An outlier can strongly affect the mean and standard deviation of a variable.
- For example, suppose a researcher mistakenly recorded an extremely high data value. This value would then make the mean and standard deviation of the variable much larger than they really were.
- Outliers can have an effect on other statistics as well.

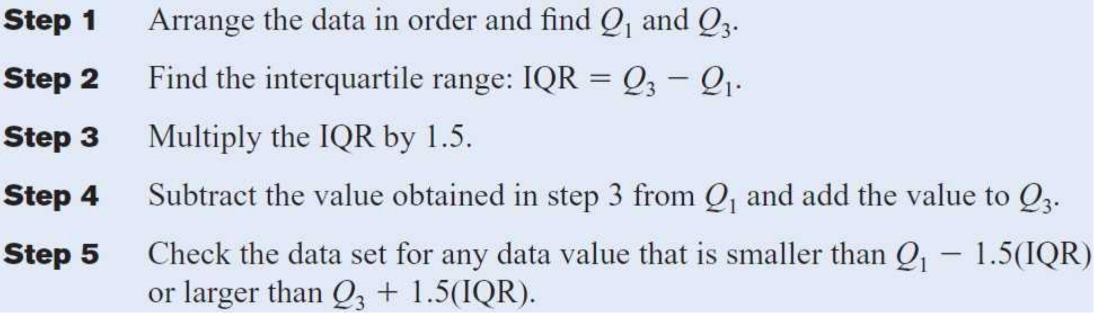
3.7.5.3 Causes for Outliers

There are several reasons why outliers may occur.

- First, the data value may have resulted from a measurement or observational error. Perhaps the researcher measured the variable incorrectly.
- Second, the data value may have resulted from a recording error. That is, it may have been written or typed incorrectly.
- Third, the data value may have been obtained from a subject that is not in the defined population. For example, suppose test scores were obtained from a seventh-grade class, but a student in that class was actually in the sixth grade and had special permission to attend the class. This student might have scored extremely low on that particular exam on that day.
- Fourth, the data value might be a legitimate value that occurred by chance (although the probability is extremely small).

3.7.5.4 Check a Data Set for Outliers

There are several ways to check a data set for outliers. One method is shown in this Procedure Table.



Step 1	Arrange the data in order and find Q_1 and Q_3.
Step 2	Find the interquartile range: IQR $= Q_3 - Q_1$.
Step 3	Multiply the IQR by 1.5.
Step 4	Subtract the value obtained in step 3 from Q_1 and add the value to Q_3.
Step 5	Check the data set for any data value that is smaller than $Q_1 - 1.5(\text{IQR})$ or larger than $Q_3 + 1.5(\text{IQR})$.

Example 3.7-8

Check the data set of ***Example 3.7-7*** above for outliers.

5, 6, 12, 13, 15, 18, 22, 50

Solution

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1

Find Q_1 and Q_3. This was done in Example above; Q_1 is 9 and Q_3 is 20.

Step 2

Find the interquartile range (IQR), which is $Q_3 \_ Q_1$.

$\text{IQR} = Q_3 - Q_1 = 20 - 9 = 11$

Step 3

Multiply this value by 1.5.

$1.5(11) = 16.5$

Step 4

Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q3.

$9 - 16.5 = -7.5$ and $20 + 16.5 = 36.5$

Step 5

Check the data set for any data values that fall outside the interval from -7.5 to 36.5.

The value 50 is outside this interval; hence, it can be considered an outlier.

Home Work 3.7-1

Data indicated in ***Table 3.7-1*** below, represents sulfate content, in form of SO_3, for a site in Baghdad. Does the data contain outlier values?

Table 3.7-1: Sulfate content in form of SO_3 for Home Work 3.7-1.

So3	
.06	.05
.05	.08
.08	.06
.07	.09
.07	.07
.08	.22
.08	.07
.06	.06
.09	.05
.07	.08
.07	.07
.07	.07
.06	.05

Answers

Yes, the value of 0.22 represents an outlier reading.

Home Work 3.7-2

Data in ***Table 3.7-2*** represents cubical compressive strength at age of one day for concrete adopted in a sliding form construction process similar to that indicated in ***Figure 3.7-1*** below. Does this data include any outlier values?

Figure 3.7-1: Sliding form construction.

Table 3.7-2: Cubical compressive strength for concrete at age of one day in MPa.

Specimen No.	$f_{cu}\ MPa$
1	35.58
2	34.48

Specimen No.	$f_{cu}\ MPa$
3	27.14
4	25.45
5	27.04
6	26.3
7	45.56
8	46.15
9	26.18
10	39.8
11	39.3
12	19.52
13	21.66
14	29.89
15	30.91

Answer

$Q_1 = 26.18\ MPa, Q_2 = 29.89\ MPa$, and $Q_3 = 39.3\ MPa$

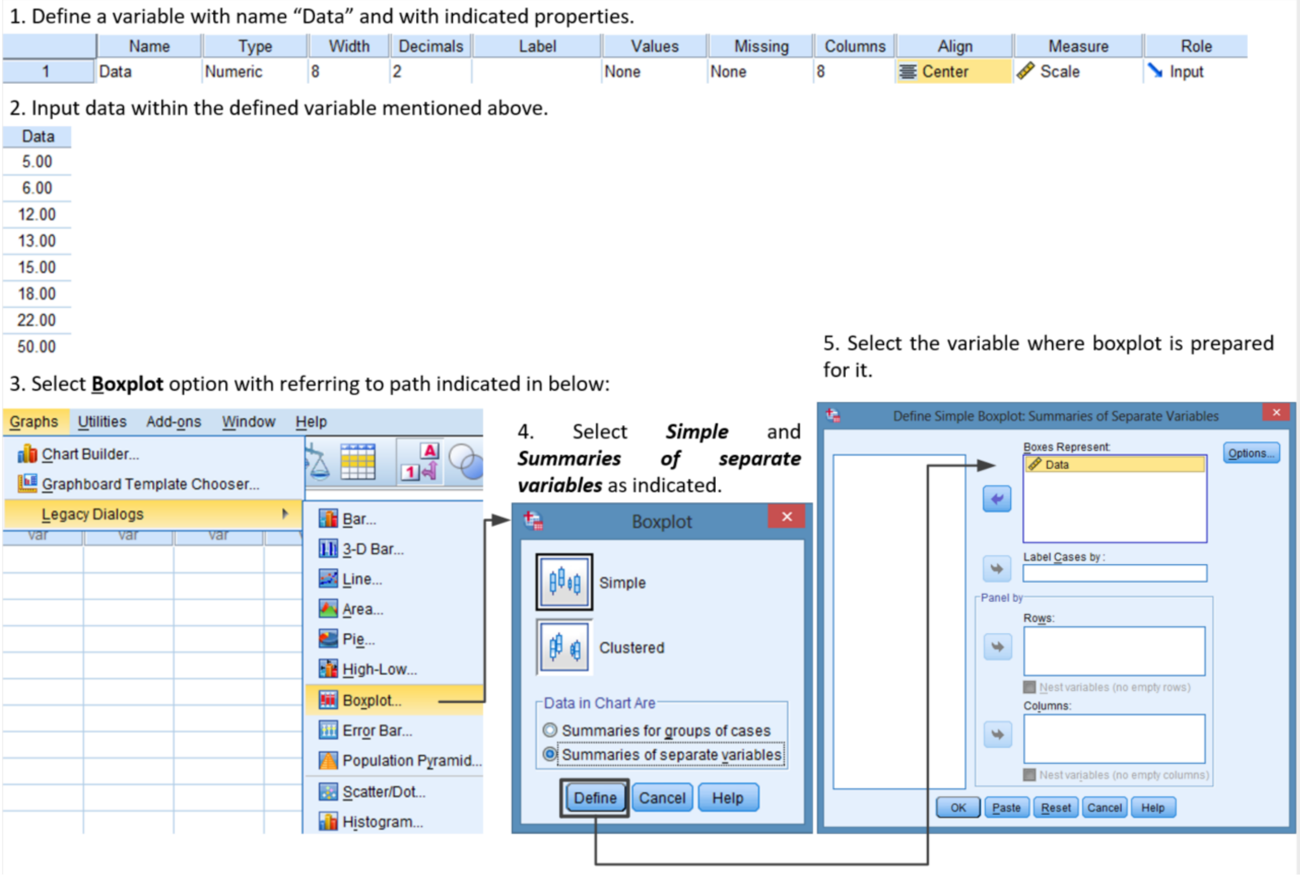
$IQR = Q_3 - Q_1 = 39.3 - 26.18 = 13.12\ MPa$

$Maximum\ Value = 46.15\ MPa < 39.3 + 1.5 \times 13.12 \approx 59\ MPa, \therefore No\ outlier$

$Minimum\ Value = 19.52\ MPa > 26.18 - 1.5 \times 13.12 \approx 6.50\ MPa, \therefore No\ outlier$

3.7.5.5 Box Plot*

- ***Box plot*** is a useful technique adopted by most of statistical software to described distribution of data and diagnose outliers if any.
- As box plot is usually drawn with aid of a software, therefore its details have been presented in this article with referring to SPSS software.
- With referring to data of ***Example 3.7-8*** above, steps for drawing Box plot in SPSS have been presented in ***Figure 3.7-2*** below.



1. Define a variable with name "Data" and with indicated properties.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Data	Numeric	8	2		None	None	8	Center	Scale	Input

2. Input data within the defined variable mentioned above.

Data
5.00
6.00
12.00
13.00
15.00
18.00
22.00
50.00

3. Select **Boxplot** option with referring to path indicated in below:

4. Select ***Simple*** and ***Summaries of separate variables*** as indicated.

5. Select the variable where boxplot is prepared for it.

Figure 3.7-2: Steps to draw a Box plot in SPSS environment.

- When pressing ok button, the SPSS box plot would be as indicated in ***Figure 3.7-3*** below. Some interest values are indicated also.
- In SPSS, a value far by more than ***1.5IQR*** from ends of box is called an ***outlier*** and ***it is indicated with a circle "o"*** while a value far by more than ***3IQR*** from ends of box is called ***extreme outlier*** and ***it is indicated with an asterisk "*"***.

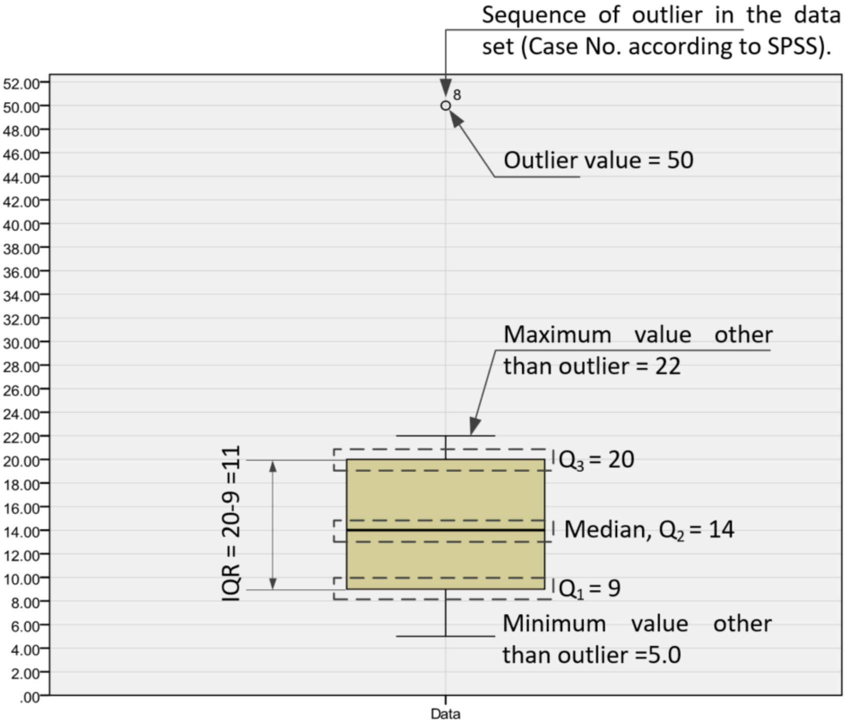


Figure 3.7-3: Box plot in SPSS environment for data of Example 3.7-8.

Example 3.7-9

Use SPSS software to resolve ***Example 3.7-8*** above when the 50 value is change into 100.

Solution

When 50 value is change into 100, the box plot in SPSS environment would be as indicated in ***Figure 3.7-4*** below. As the value is far from box upper end by more that ***3IQR***

$100 > 20 + 3 \times 11 = 53$

therefore it has been classified as an extreme outlier value and indicated with an asterisk "*" instead of "o" as done in ***Example 3.7-8*** above.

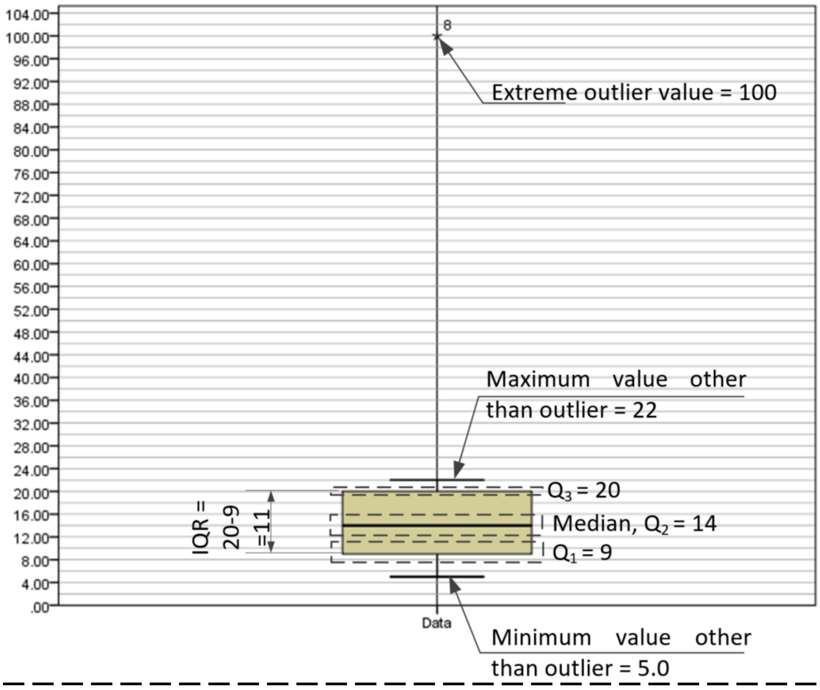


Figure 3.7-4: Box plot in SPSS environment for data of Example 3.7-9.

Home Work 3.7-3

Resolve ***Home Work 3.7-1*** above with using box plot. According to SPSS notation, does the value of 0.22 represent an outlier or extreme outlier?

Answer

The box plot is indicated in ***Figure 3.7-5*** below. According to notations of SPSS, the value of 0.22 is classified as an extreme outlier.

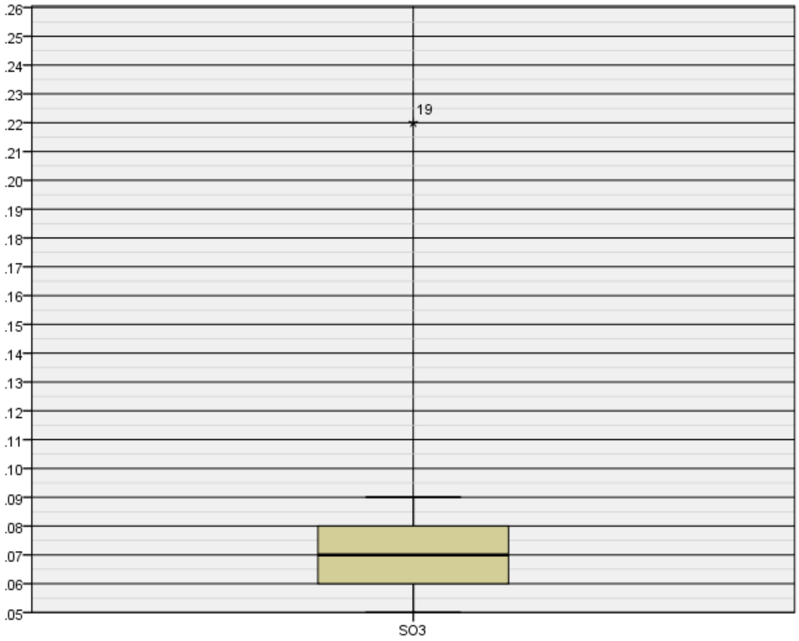


Figure 3.7-5: Box plot for Home Work 3.7-3.

Home Work 3.7-4

Using box plot, resolve ***Home Work 3.7-2*** above for sliding form construction to show if there is any outlier.

Answer

The box plot is indicated in ***Figure 3.7-6*** below. According to SPSS box plot, there is no outlier in the data.

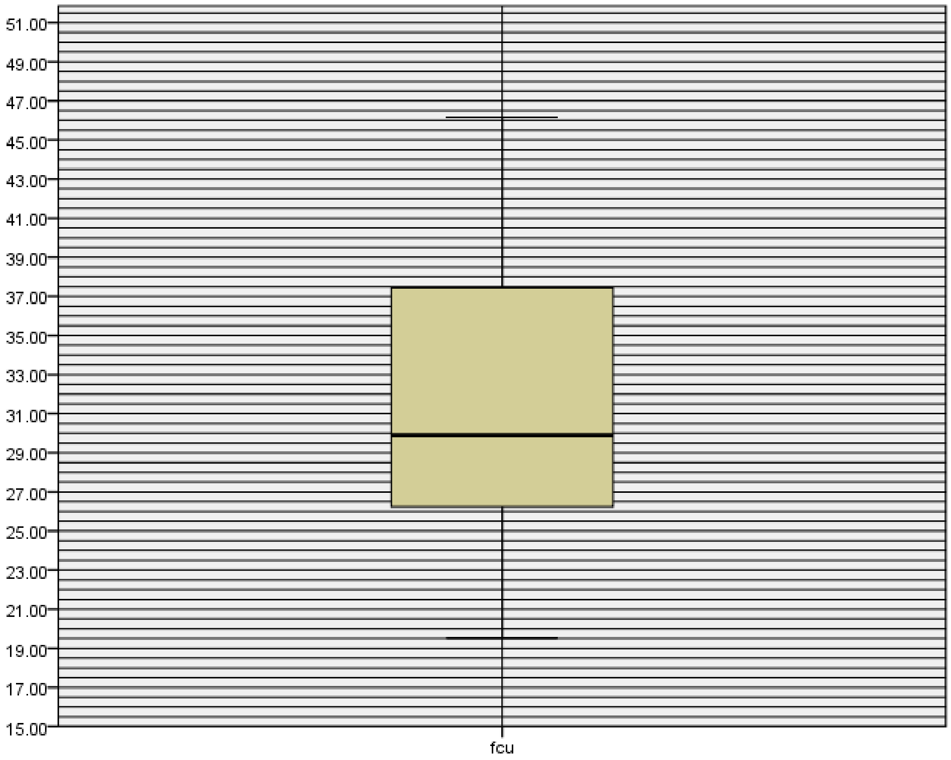


Figure 3.7-6: Box plot for Home Work 3.7-4.

Contents

CHAPTER 4

PROBABILITY AND COUNTING RULES

4.1 INTRODUCTION

4.1.1 CHAPTER LAYOUT

- Many people are familiar with probability from observing or playing games of chance.
- In addition to being used in games of chance, probability is the basis of inferential statistics.
- The basic concepts of probability are explained in this chapter. These concepts include:
  - Probability experiments and sample spaces. These include rules for counting.
  - Deferent interpretations for probability.
  - Addition rule, multiplication rule, probabilities of complementary events, and conditional probability.

4.1.2 PROBABILITY THEORY IN ENGINEERING

- Every engineering problem involves phenomena that exhibit scatter of the type illustrated in the previous chapters.
- To deal with such situations in a manner in which incorporates this variability in his analyses, the engineer makes use of the ***theory of probability***, a branch of mathematics dealing with uncertainty.

4.1.3 SOURCES FOR UNCERTAINTY IN ENGINEERING

- Uncertainty is introduced into engineering problems through the:
  - Through lack of sufficient data:
    Lacking a full-depth hole, the depth of soil to rock at a building site can only be estimated. This uncertainty is the result of incomplete information.
  - Variation inherent in nature:
    This type of uncertainty occurs even one surveys whole population.
    For example, even with a long history of data, one cannot predict the maximum flood that will occur in the next 10 years in a given area. This uncertainty is a product of natural variation.
    Both the depth to rock and the maximum flood are uncertain, and both can be dealt with using the same theory, namely ***the probability theory***.
  - Through man's lack of understanding of all the causes and effects in physical systems
- As a result of uncertainties like those mentioned above, ***the future can never be entirely predicted*** by the engineer.
- The engineer must, rather, ***consider the possibility of the occurrence of particular events and then determine the likelihood of their occurrence***.

4.1.4 MODELING IN ENGINEERING

- A fundamental step in any engineering investigation is the formulation of a set of mathematical models that is, descriptions of real situations in ***a simplified***, ***idealized form suitable for computation***.

- In civil engineering, one frequently:
  - Ignores friction,
  - Assumes rigid bodies,
  - Adopts an ideal fluid

  to arrive at relatively simple mathematical models, which are amenable to analysis by arithmetic or calculus.

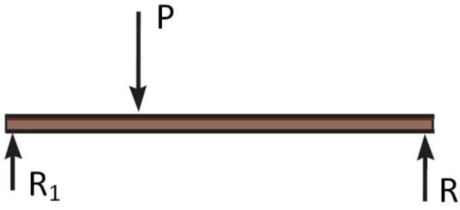
4.1.5 Deterministic versus Probabilistic Models

- Frequently the models in engineering are ***deterministic***, where a single number describes each independent variable and a formula (a model) predicts a specific value for the dependent variable.
- When the element of uncertainty, owing to:
  - Natural variation,
  - Or incomplete professional knowledge,

  is to be considered explicitly, the models derived are ***probabilistic*** and subject to analysis by the rules of probability theory.
- In the ***probabilistic model***, the values of the independent variables are not known with certainty, and thus the variable related to them through the physical model cannot be precisely predicted.
- In addition, the physical model may itself contain elements of uncertainty. Many examples of both situations will follow.

Example 4.1-1

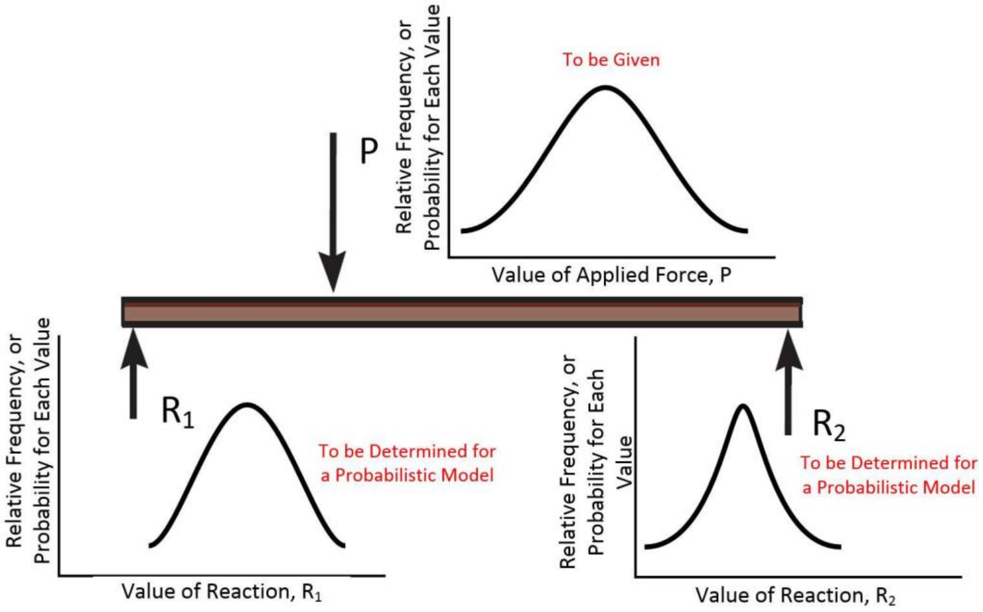
For the simply supported beam shown below, where a deterministic or probabilistic model is adopted?



Solution

When deterministic model is adopting, as in ***Engineering Mechanic***, the magnitude, position, and direction for applied force, P, are assumed completely defined.

While when we are uncertain about, for example, the magnitude of applied force, P, we should adopt a probabilistic model where frequency distributions for reactions, R_1 and R_2 , are determined from frequency distribution for the applied force, P.


Relative Frequency, or Probability for Each Value
To be Given
P
Value of Applied Force, P
Relative Frequency, or Probability for Each Value
R1
To be Determined for a Probabilistic Model
Value of Reaction, R1
Relative Frequency, or Probability for Each Value
R2
To be Determined for a Probabilistic Model
Value of Reaction, R2

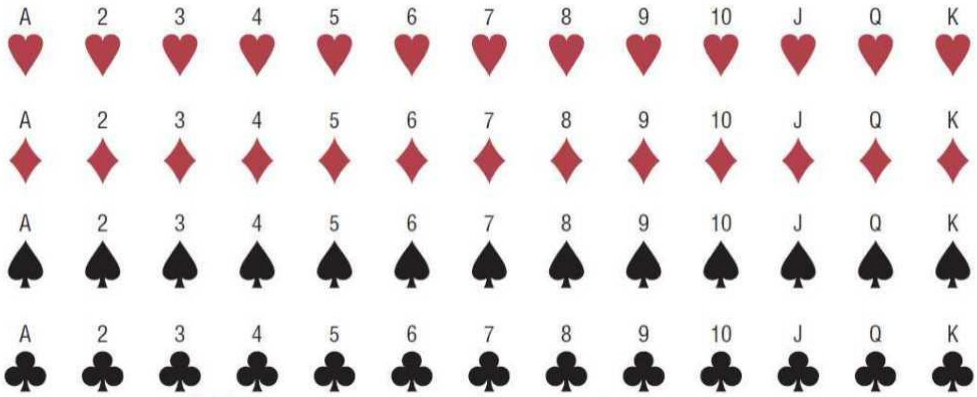
4.1.6 Common Examples in Probability Theory

As theory of probability grew out of the study of various games of chance, then following common examples are usually used to explain of probability basic concepts.

- Coins Examples:

- Dice or Die (old name) Examples:

- Cards Examples:


A 2 3 4 5 6 7 8 9 10 J Q K
A 2 3 4 5 6 7 8 9 10 J Q K
A 2 3 4 5 6 7 8 9 10 J Q K
A 2 3 4 5 6 7 8 9 10 J Q K

Hearts ♥, Clubs ♣, Diamonds ♦, and Spades ♠

- Since these devices lend themselves well to the application of concepts of probability, they will be used in this chapter as examples.

4.2 Sample Space

4.2.1 Definition

- Processes such as flipping a coin, rolling a die, or drawing a card from a deck are called probability experiments.
- An outcome is the result of a single trial of a probability experiment.
- A sample space is the set of all possible outcomes of a probability experiment. Some sample spaces for various probability experiments are shown here.

Experiment	Sample space
Toss one coin	Head, tail
Roll a die	1, 2, 3, 4, 5, 6
Answer a true/false question	True, false
Toss two coins	Head-head, tail-tail, head-tail, tail-head

Example 4.2-1

Find the sample space for rolling two dice.

Solution

Since each die can land in six different ways, and two dice are rolled, the sample space can be presented by a rectangular array, as shown in Figure below. The sample space is the list of pairs of numbers in the chart.

	Die 2					
Die 1	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Example 4.2-2

Find the sample space for drawing one card from an ordinary deck of cards.

Solution

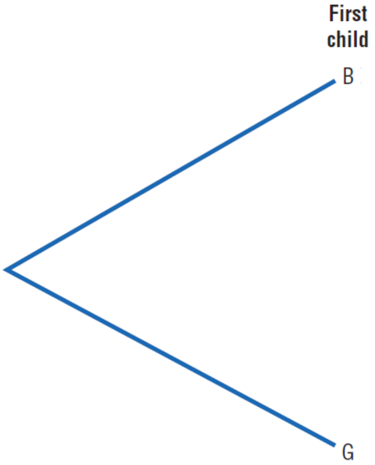
Since there are 4 suits and 13 cards for each suit (ace through king), there are 52 outcomes in the sample space.

Example 4.2-3

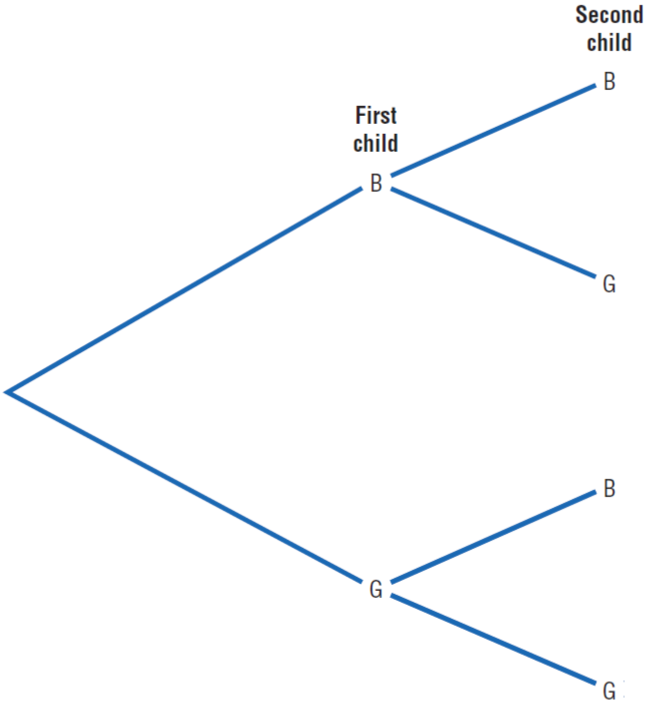
Find the sample space for the gender of the children if a family has three children. Use B for boy and G for girl.

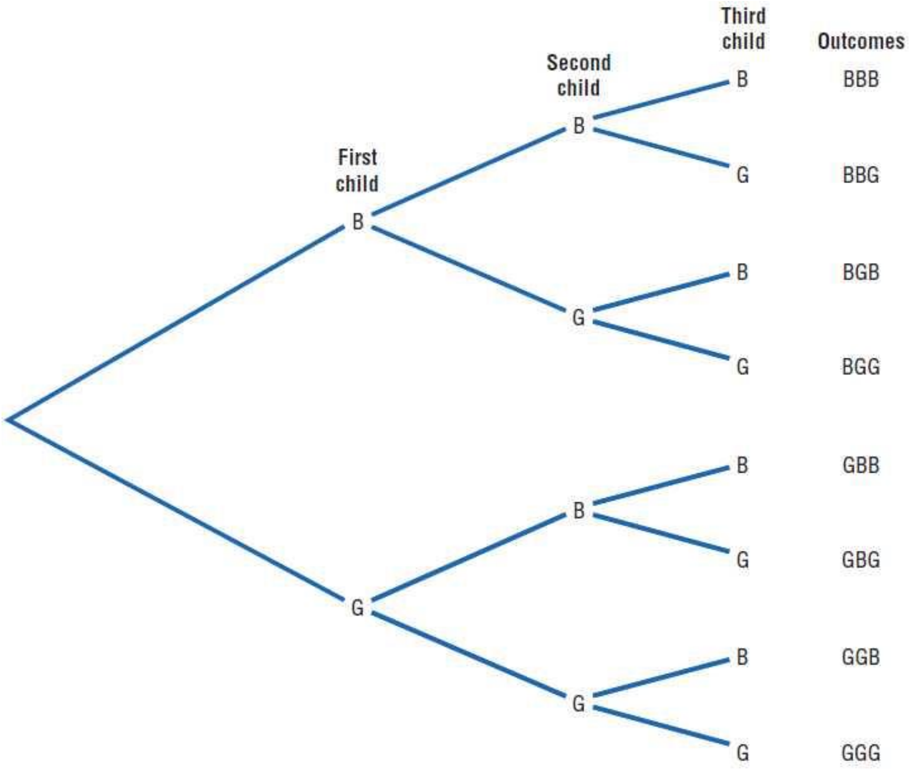
Solution

In previous examples, the sample spaces were found by observation and reasoning; however, another way to find all possible outcomes of a probability experiment is to use a ***tree diagram***.



1. Since there are two possibilities (boy or girl) for the first child, draw two branches from a starting point and label one B and the other G.
2. Then if the first child is a boy, there are two possibilities for the second child (boy or girl), so draw two branches from B and label one B and the other G. Do the same if the first child is a girl.
3. Follow the same procedure for the third child. The completed tree diagram is shown in in Figure below:





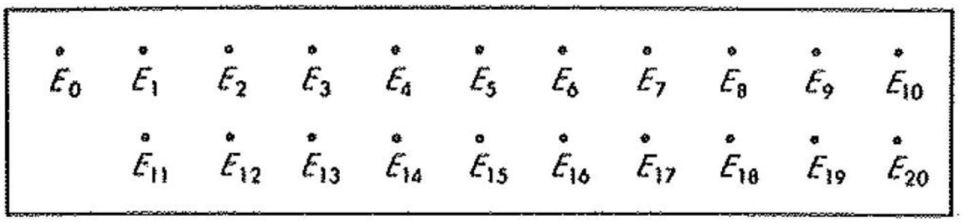
Example 4.2-4, Traffic Example

Suppose that a traffic engineer goes to a particular street intersection exactly at noon each weekday and waits until the traffic signal there has gone through one cycle. The engineer records the number of southbound vehicles which had to come to a complete stop before their light turned green. If a minimum vehicle length is 15 ft and the block is 300 ft long, the maximum possible number of cars in the queue is 20.

If only the total number of vehicles is of interest, what is the sample space for this experiment?

Solution

The sample space is a set of 21 points labeled, say, $E_0, E_1, \dots\dots\dots\dots.., E_{20}$ each associated with a particular number of observed vehicles. These might be represented as in Figure below.



Example 4.2-5

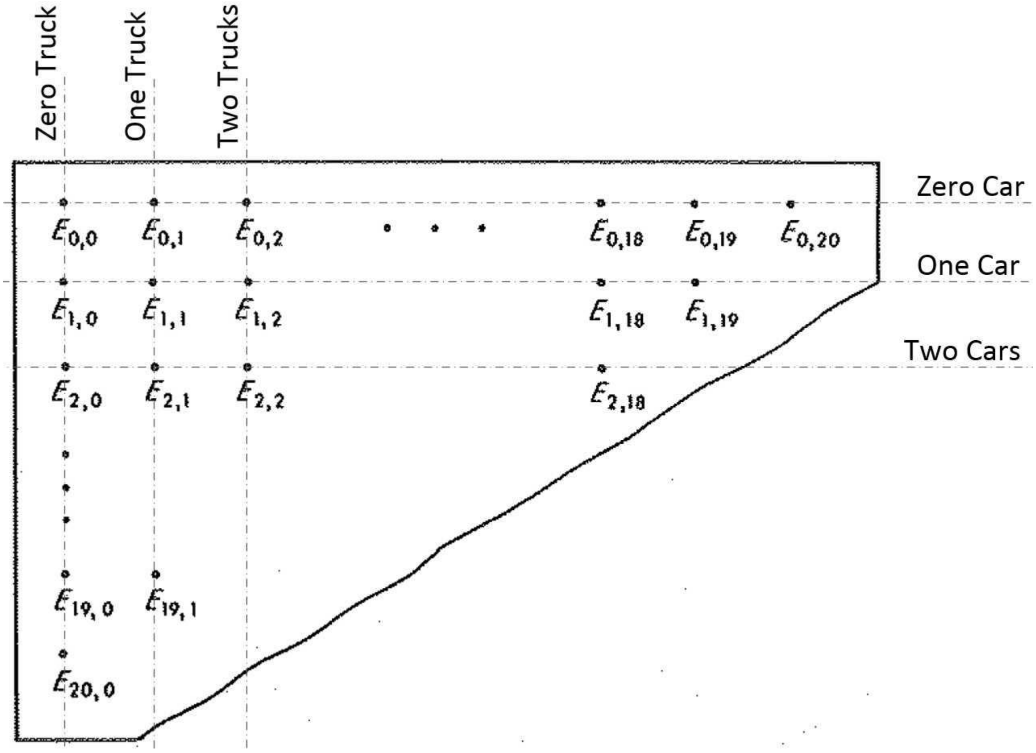
What is the sample space for experiment of previous example when the engineer needed other information regarding for differentiating between trucks and automobiles?

Solution

As more information are required, then the sample space for the experiment would then be larger, containing an individual sample point $E_{i,j}$ for each possible combination of i cars and j trucks such that the maximum value of

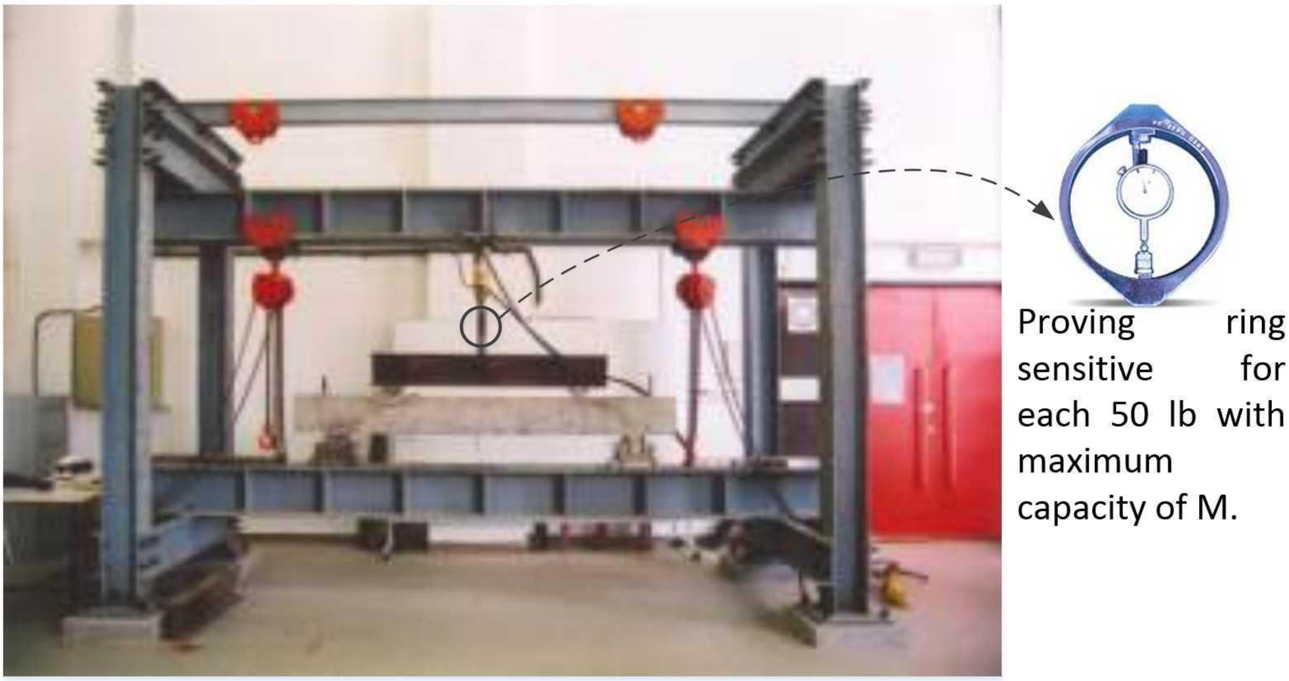
$i + j = 20$

Therefore, sample space would be as in shown in Figure below.



Example 4.2-6

In testing the ultimate strength of reinforced-concrete beams, what is sample space when testing machine is only sensitive to the nearest 50 lb.



Proving ring sensitive for each 50 lb with maximum capacity of M.

Solution

The sample space for the experiment consists of a set of points, each associated with an outcome,

$0, 50, 100, \ldots, or\ M\ lb$

where M maximum capacity of testing machine, a multiple of 50 lb.

Example 4.2-7

Resolve previous example of testing machine with assuming the machine is so precise that can match any load value between zero and its maximum capacity, $M\ lb$.

Solution

Instead of discrete sample space of previous example, the sample space is a continuous one from zero to $M\ lb$. In set notation, sample space can be written as,

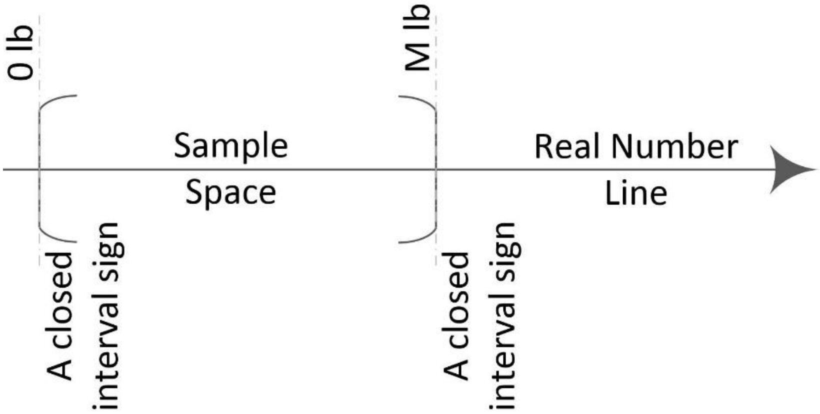
$\Omega \equiv \{S: 0 \leq S \leq M\}$

where

Ω is the sample space,

S is strength of a reinforced concrete beam measured by proving ring, in lb.

M is the maximum strength of proving ring, in lb.



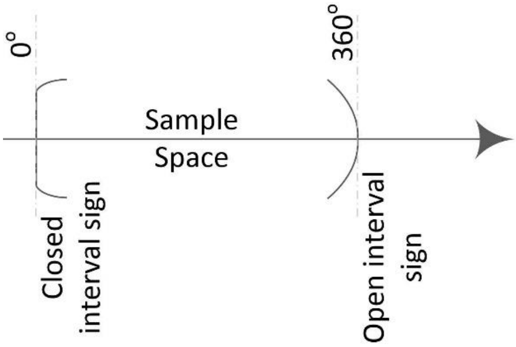
Example 4.2-8, Wind Direction in an Airport Site

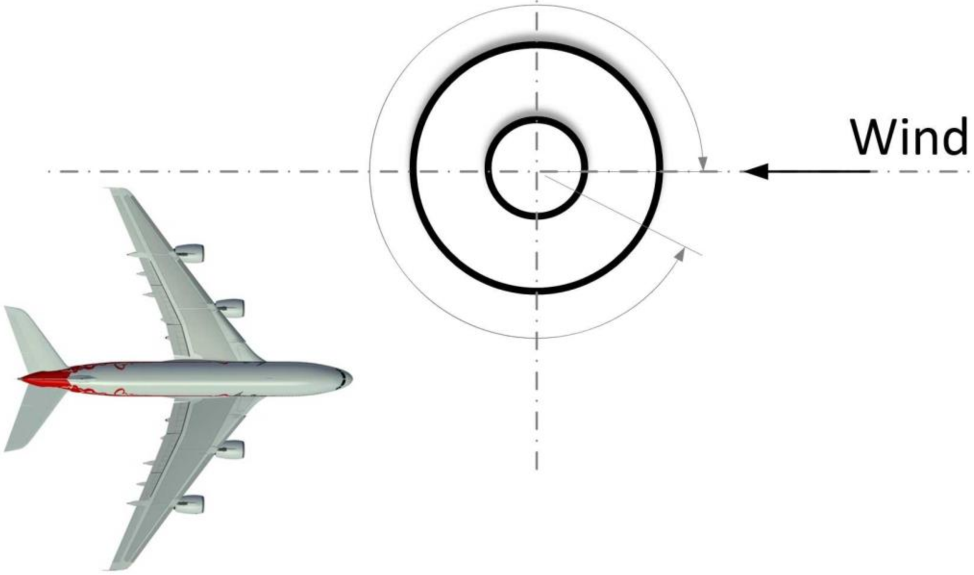
In an airport site, what is the sample space of wind direction for an observing tower located in an open region?

Solution

Assuming completely open region and neglecting seasonal effect, wind can blow from any side and sample space would be:

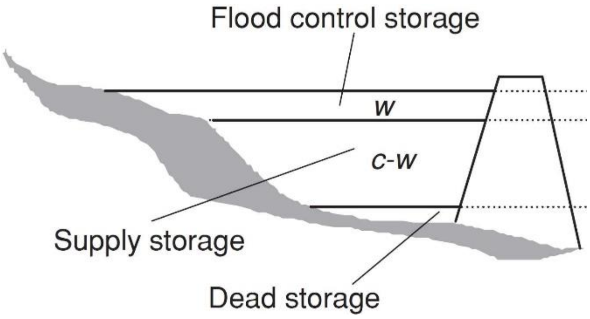
$\Omega \equiv \{W_D: 0^o \leq W_D < 360^o\}$





Example 4.2-9

The amount of water, S, stored in a reservoir varies with time from 0 to c, the active reservoir capacity, owing to the combined effect of inflows and outflows (see Figure below). What is the sample space for water in this reservoir?

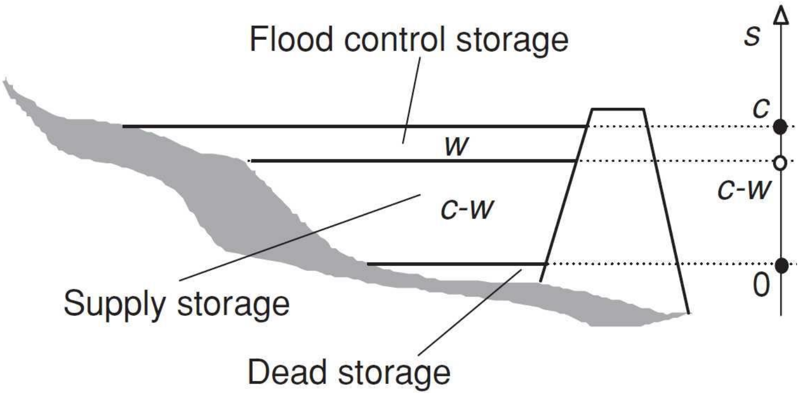


Solution

As flood control storage, for example a spillway, works just when water reaches maximum level, C, then the sample space will be:

$\Omega \equiv \{S: 0 \leq S < c\}$

Graphically, the sample space will be:



4.2.2 Event

4.2.2.1 Basic Definition

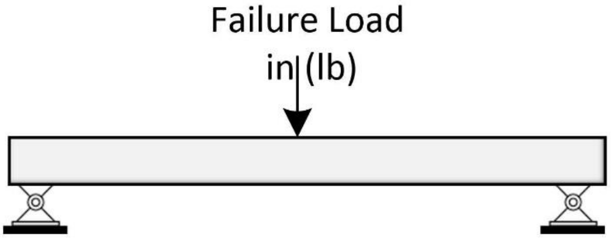
- An outcome was defined previously as the result of a single trial of a probability experiment. In many problems, one must find the probability of two or more outcomes.
- For this reason, it is necessary to distinguish between an **outcome** and an **event**.

An **event** consists of a set of outcomes of a probability experiment.

- An event can be one outcome or more than one outcome, then events could be classified into:
  - Simple Event
    It is an event that consists from single outcome. For example getting 6 in die rolling.
  - Compound Event
    It is an event that consists from more than one outcome. For example, getting odd number in die rolling.

Example 4.2-10

In an experiment to measure beam strength expressed in terms of maximum point load, in pound, applied at mid-span of beam, sketch following events:



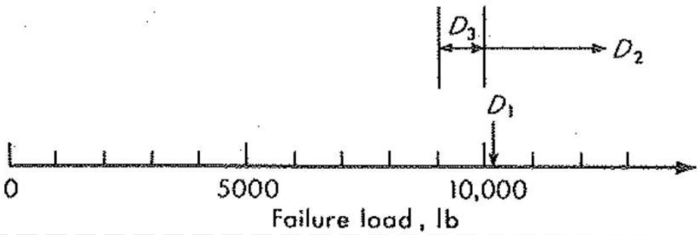
$D_1 = [the\ observation\ was\ 10101]$

$D_2 = [the\ outcome\ was\ greater\ than\ 10{,}000, that\ is, > \ 10{,}000]$

$D_3 = [the\ result\ was\ greater\ than\ 9000\ but\ less\ than\ or\ equal\ to$
$10{,}000, that\ is, > \ 9000\ and\ \leq;\ 10{,}000]$

Solution

The sketch for aforementioned event is presented in below:



4.2.2.2 Relationships Among Events

Events in a sample space may be related in a number of ways:

4.2.2.2.1 Mutually Exclusive Events

The events are said to be ***mutually exclusive*** or ***disjoint*** when ***events contain no sample point in common***.

Example 4.2-11

For the Example 4.2-4, Traffic Example, if events A and B are defined as:

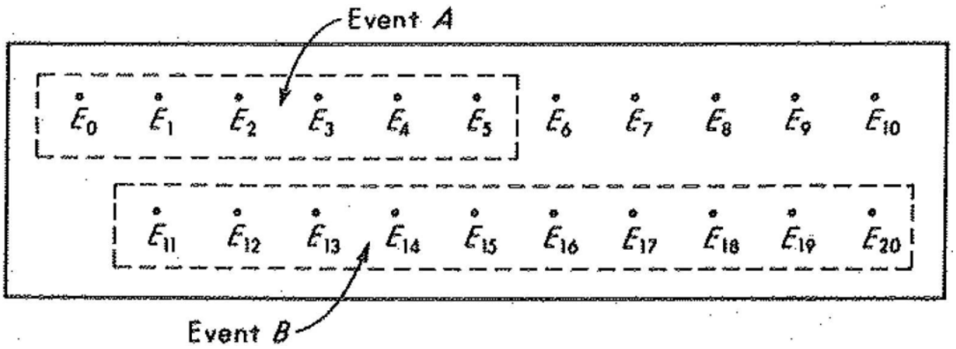
A, "fewer than 6 stopped vehicles were observed"

B, "more than 10 stopped vehicles were observed"

then, are the two events mutually exclusive?

Solution

If two events are shaded on sample space as shown in the indicated, one concludes that events are ***mutually exclusive,*** as they have no common sample point.

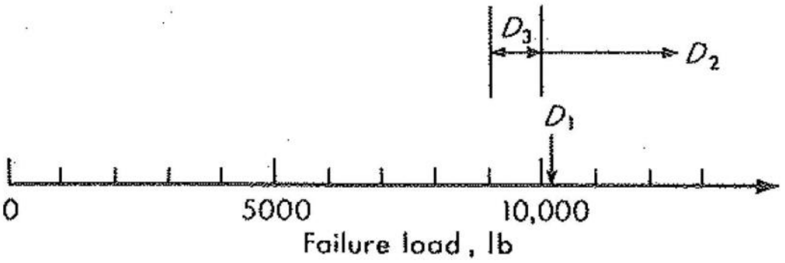


Example 4.2-12

Are events D_1, and D_3 or D_2 and D_3 of Example 4.2-10 above mutually exclusive?

Solution

For convenience, definitions and graphical description of events are represented in below:



$D_1 = [\textit{the observation was } 10{,}101]$

$D_2 = [\textit{the outcome was greater than } 10{,}000, \textit{that is}, > \ 10{,}000]$

$D_3 = [\textit{the result was greater than } 9000 \textit{ but less than or equal to}$
$10{,}000, \textit{that is}, > \ 9000 \textit{ and } \ \leq 10{,}000]$

The events D_1 and D_3, defined above are mutually exclusive as they have no common point.

The D_2 and D_3 are also mutually exclusive, owing to the care with which the ***inequality*** $(\leq)$ and ***strict inequality*** $(>)$ have been written at 10,000.

4.2.2.2.2 Intersection Events

If a pair of events A and B are ***not mutually exclusive***, the ***set of points which they have in common is called their intersection***, denoted $A \cap B$.

4.2.2.2.3 Collectively Exhaustive Events

Example 4.2-13

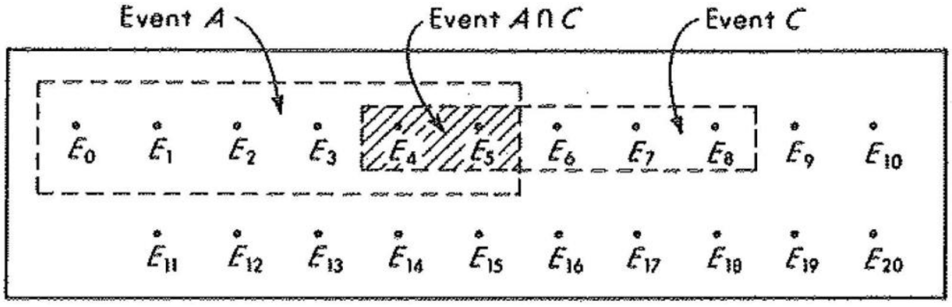
For the Example 4.2-4, Traffic Example, if events A and C are defined as:

A, "fewer than 6 stopped vehicles were observed"

C, "from four to eight stopped vehicles were observed"

then, are the two events mutually exclusive?

Solution


Event A
Event $A \cap C$
Event C
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
E_{11} E_{12} E_{13} E_{14} E_{15} E_{16} E_{17} E_{18} E_{19} E_{20}

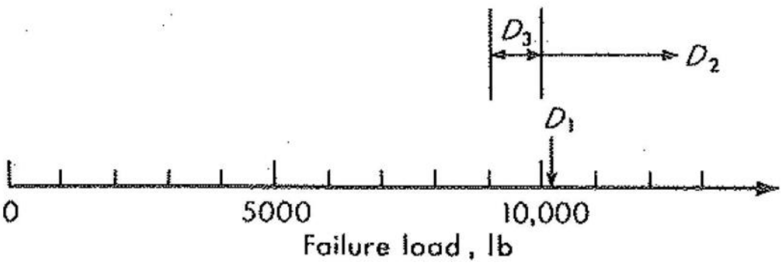
If two events are shaded on sample space as shown in the figure, one concludes that events are ***not mutually exclusive*** as they have common sample points.

Example 4.2-14

Are events D_1, and D_2 of Example 4.2-10 above mutually exclusive?

Solution

For convenience, definitions and graphical description of events are represented in below:



$D_1 =\ [the\ observation\ was\ 10101]$

$D_2 =\ [the\ outcome\ was\ greater\ than\ 10{,}000, that\ is, >\ 10{,}000]$

$D_3 =\ [the\ result\ was\ greater\ than\ 9000\ but\ less\ than\ or\ equal\ to$
$10{,}000, that\ is, >\ 9000\ and\ \leq;\ 10{,}000]$

The intersection of the events D_1 and D_2, in Figure above is simply the event D_1, itself.

If the intersection of two events is equivalent to one of the events, that event is said to be contained in the other. This is written

$D_1 \subset D_3$

where operator $\subset$ read as a subset.

4.2.2.2.4 The Union of Two Events

The union of two events A and C is the event which is the collection of all sample points which occur at least once in either A or C, written $A \cup C$.

Example 4.2-15

For the Example 4.2-4, Traffic Example, if events A and C are defined as:

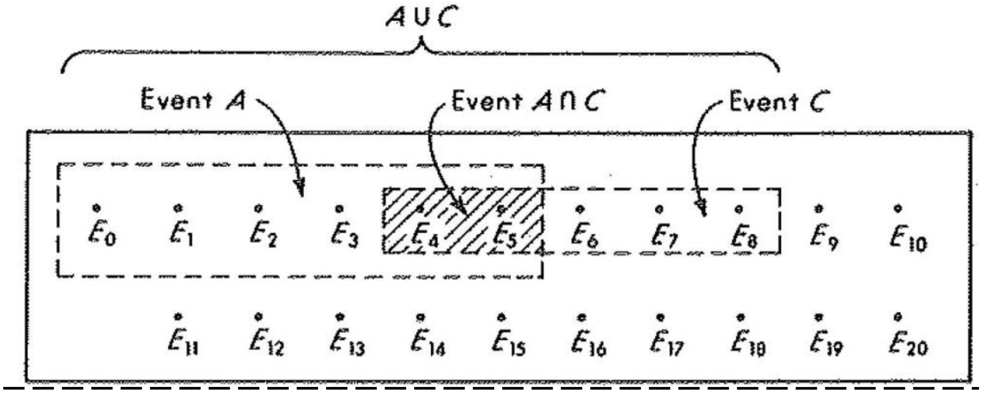
A, "fewer than 6 stopped vehicles were observed"

C, "from four to eight stopped vehicles were observed"

What is the union of the events A and C?

Solution

The union of the events A and C is presented graphically in the indicated figure.


$A \cup C$
Event A
Event $A \cap C$
Event C
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
E_{11} E_{12} E_{13} E_{14} E_{15} E_{16} E_{17} E_{18} E_{19} E_{20}

4.2.2.2.5 Complementary Events

The ***complement of an event*** E is the set of outcomes in the sample space that are not included in the outcomes of event E. The complement of E is denoted by E^c.

Above definition is valid irrespective the type of interpretation that adopted.

Example 4.2-16

Find the complement of each event.

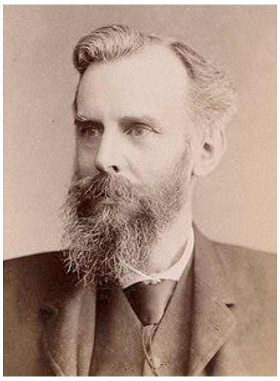
a. Rolling a die and getting a 4
b. Selecting a month and getting a month that begins with a J
c. Selecting a day of the week and getting a weekday

Solution

a. Getting a 1, 2, 3, 5, or 6
b. Getting February, March, April, May, August, September, October, November or December
c. Getting Saturday or Sunday

4.2.2.3 VENN DIAGRAM

- Venn diagrams were developed by mathematician John Venn and are used in set theory and symbolic logic.
- Such diagrams provide a very useful visual representation of sets and set operations such as the complement, union, intersection, and other combinations.
- Because sample points, sample spaces, and events are sets, one can use this type of illustration to show the events in a sample space and important relationships among events.
- Because the algebra of events is analogous to areas on a plane, Venn diagrams such as those of Figure 4.2-1 are most helpful for interpretation.



John Venn, (4 August 1834–4 April 1923) was an English mathematician, logician and philosopher noted for introducing the Venn diagram, used in the fields of set theory, probability, logic, statistics, competition math, and computer science.

In 1866, Venn published The Logic of Chance, a ground-breaking book which adopted the frequency theory of probability, offering that probability should be determined by how often something is forecast to occur as opposed to "educated" assumptions.

Venn then further developed George Boole's theories in the 1881 work Symbolic Logic, where he highlighted what would become known as Venn diagrams.

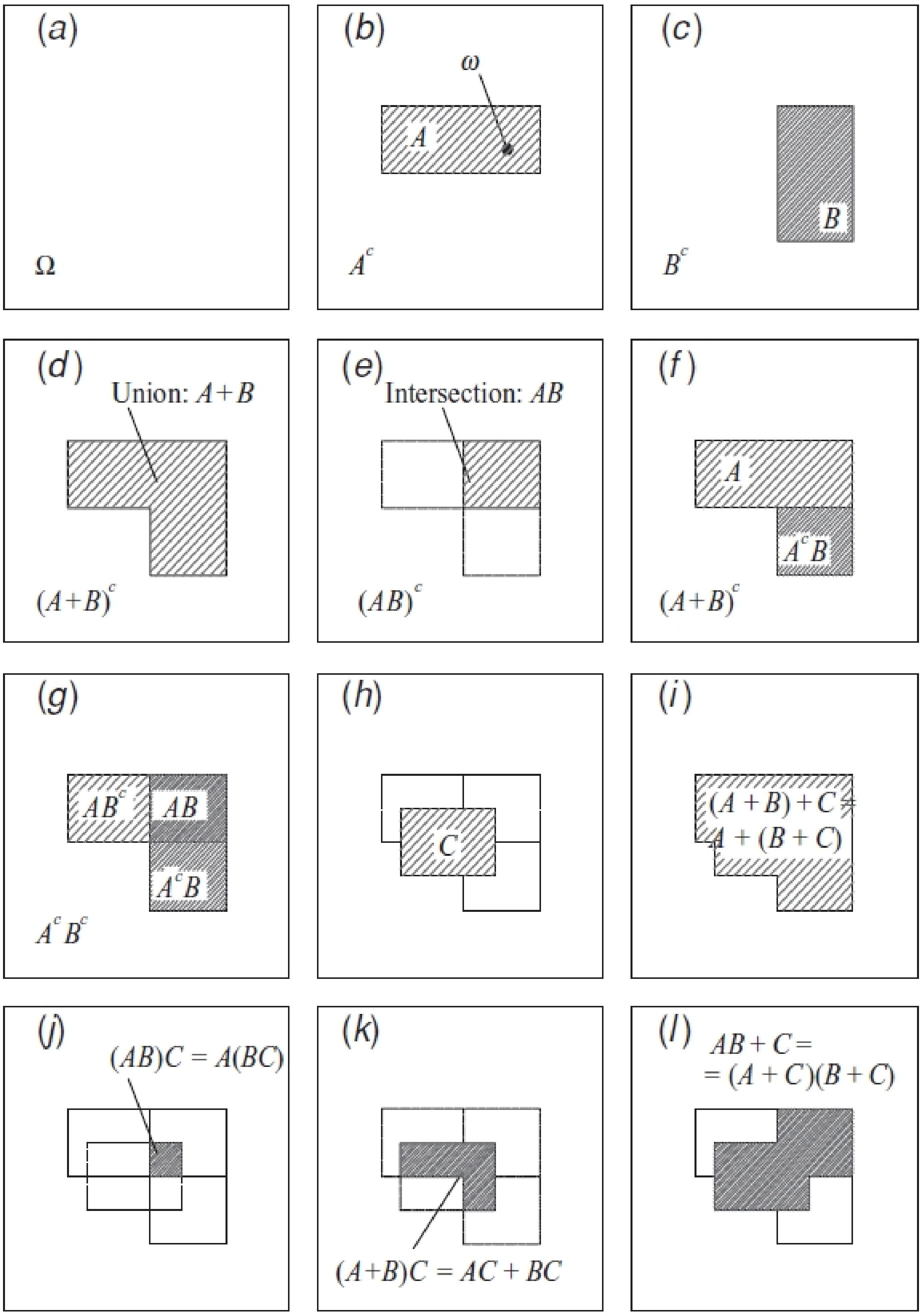


Figure 4.2-1: Venn diagrams representing the sample space and different random events.

4.2.3 Two-dimensional Sample Space

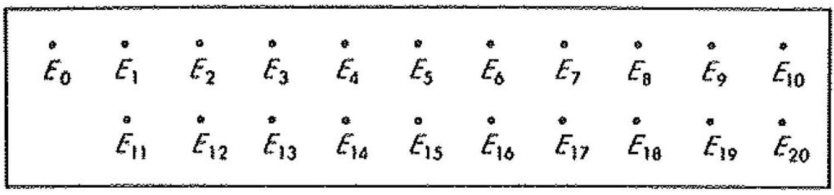
Two-dimensional sample space is useful when where the experiment involves observing two numbers, for example the number of cars and the number of trucks.

Example 4.2-17

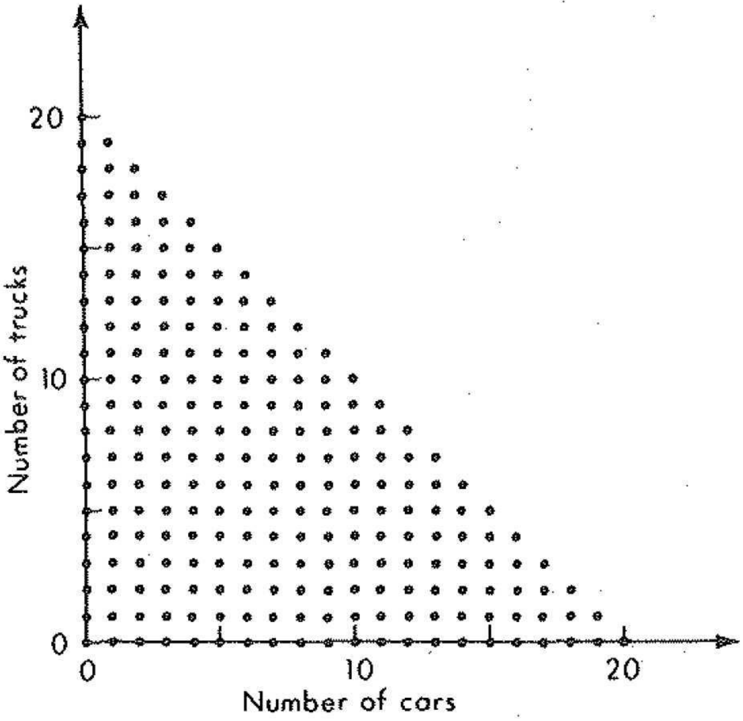
Represent sample space for Example 4.2-4, Traffic Example in terms of two-dimensional sample space.

Solution

For convenience, the original sample space for Example 4.2-4, Traffic Example is represented in below,



In terms of two-dimensional, the sample space is shown below. Where each point on the grid represents a possible outcome, a sample point.

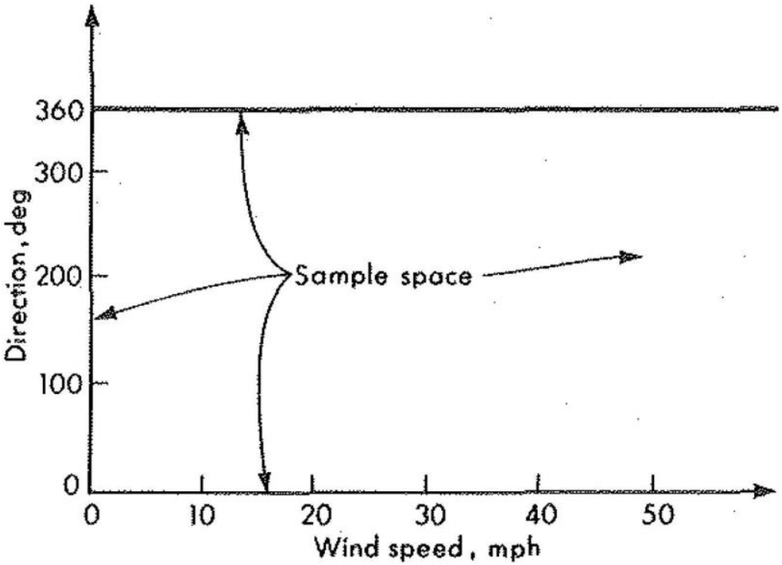


Example 4.2-18

Use two-dimensional sample space to describe the sample space for an experiment concerns with direction and speed of wind in an airport site.

Solution

The two-dimensional sample space for wind speed and direction is presented in below. Wind is implicitly assumed to blow from any direction and has no upper bound velocity limit.



4.2.4 CONDITIONAL SAMPLE SPACE

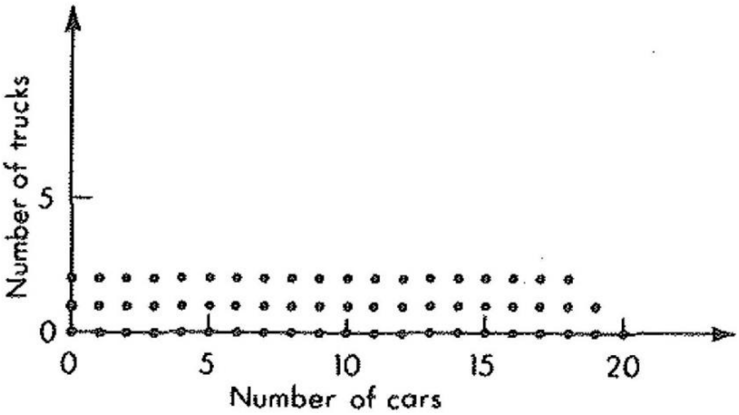
If the engineer is interested in the possible outcomes of an experiment given that some event A has occurred, the set of events associated with event A can be considered a new, reduced sample space called as ***conditional sample space***.

Example 4.2-19

For the Example 4.2-4, Traffic Example, what is conditional sample space if we only concern with observation of two or fewer trucks?

Solution

After excluding more than two trucks from two-dimensional sample space for Example 4.2-4, Traffic Example, its conditional sample space would be as shown below:

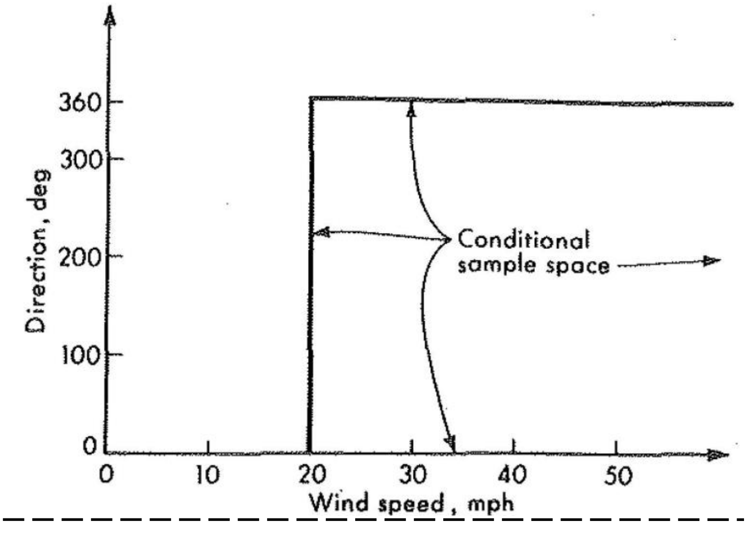


Example 4.2-20

For the experiment concerns with wind direction and speed in an airport site, what is the conditional sample space if the engineer interests only with wind velocity greater than 20 mph?

Solution

Assuming wind to blow from any direction and has no upper bound velocity limit, the conditional sample space would be:



Example 4.2-21

A civil engineer is asked to assess the reliability of a balcony overlooking a football stadium. The maximum number of people who can be accommodated in the balcony is 20. The weight of an individual can be approximately 50, 75, or 100 kg.

(a) Sketch the sample space.

(b) Show the following events involving numbers of people and their weights at any time:

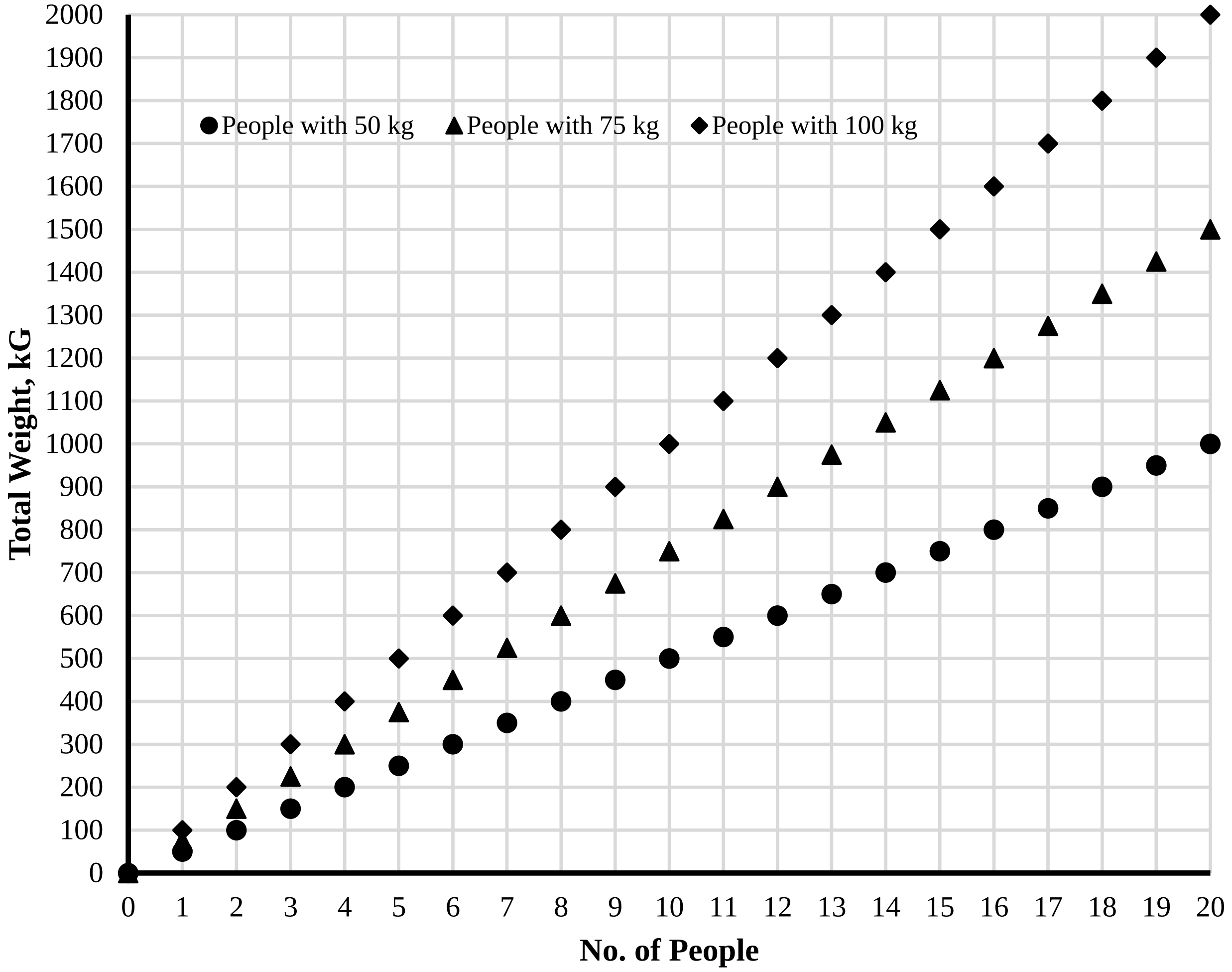
A ≡ {there are more than 16 people in the balcony},

B ≡ {the total weight of people in the balcony is 1500 kg},

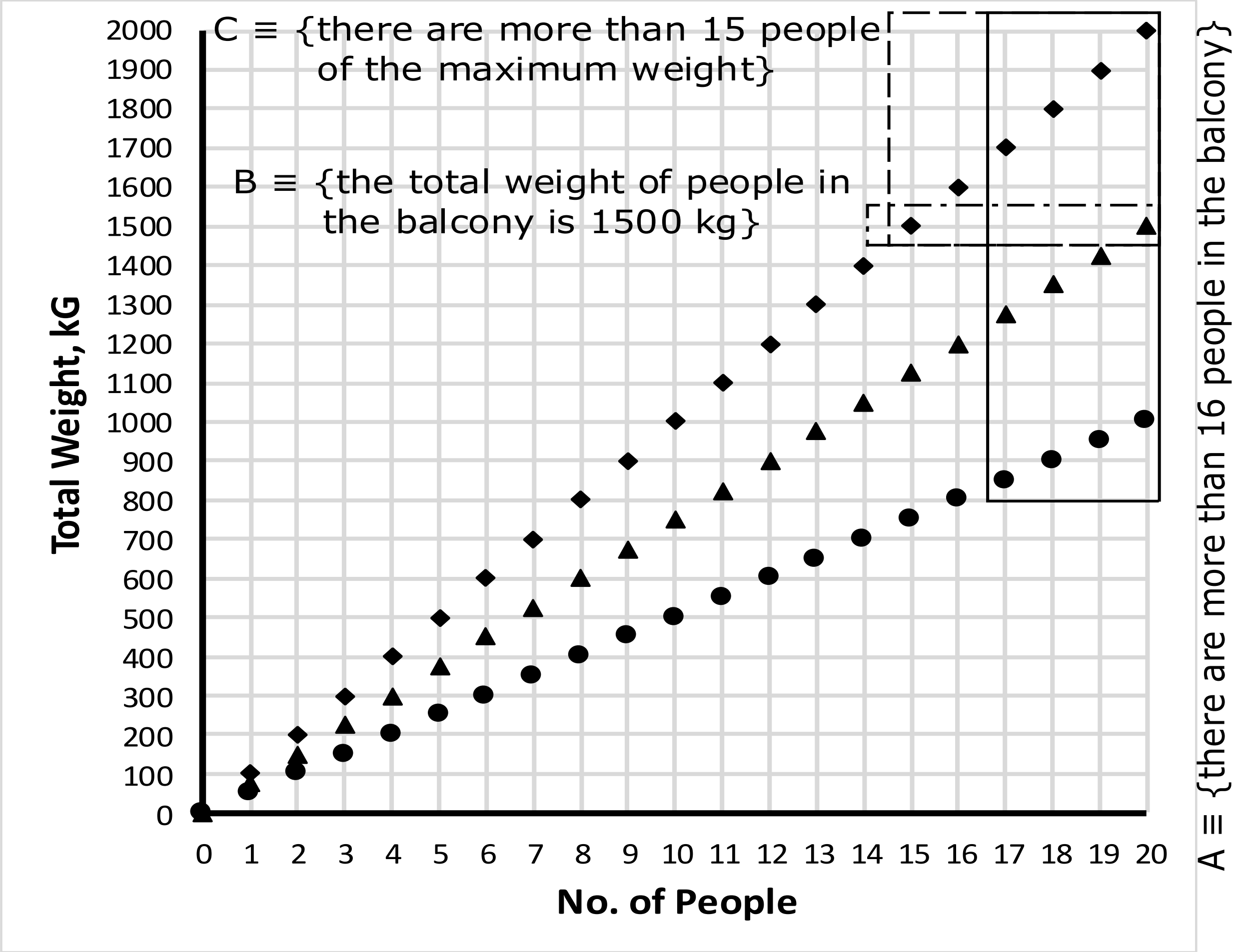
C ≡ {there are more than 15 people of the maximum weight}.

Solution

The sample space is shown below:



Events A, B, and C are shown in below:



Example 4.2-22

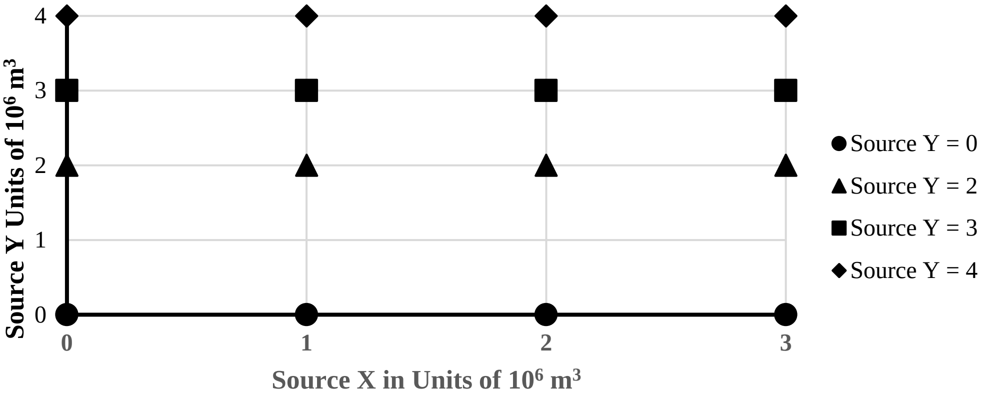
A reservoir impounds water from a stream X and receives water Y deviated via a tunnel from an adjoining catchment. The annual inflow from source X can be approximated to 1 or 2 or 3 units of 10^6 m^3, and that from source Y is 2 or 3 or 4 units of 10^6 m^3. On an appropriate sample space, show the following events:

A ≡ {source X is less than 3 units}.

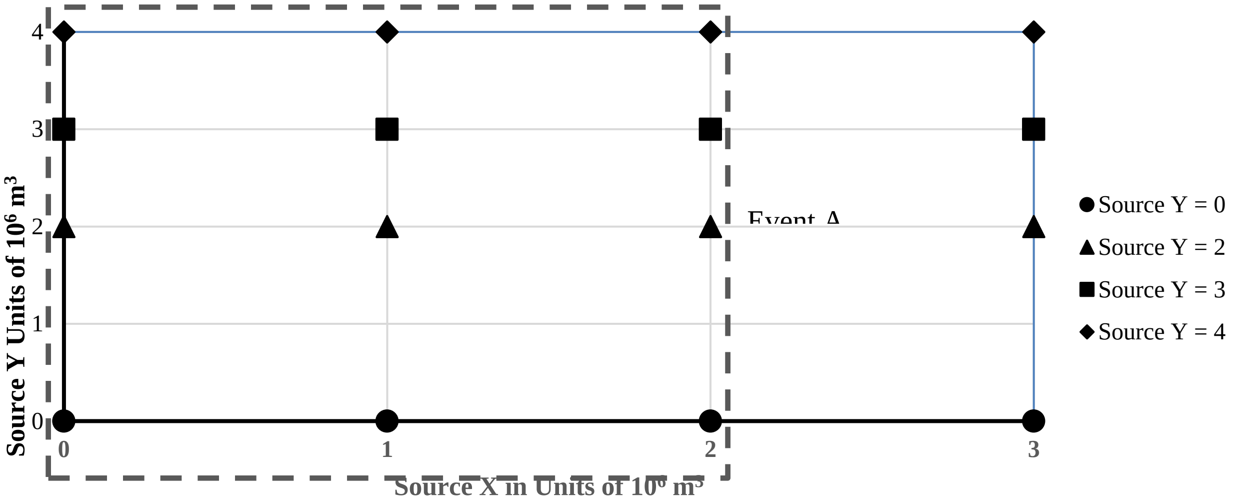
B ≡ {source Y is more than 2 units}.

Solution

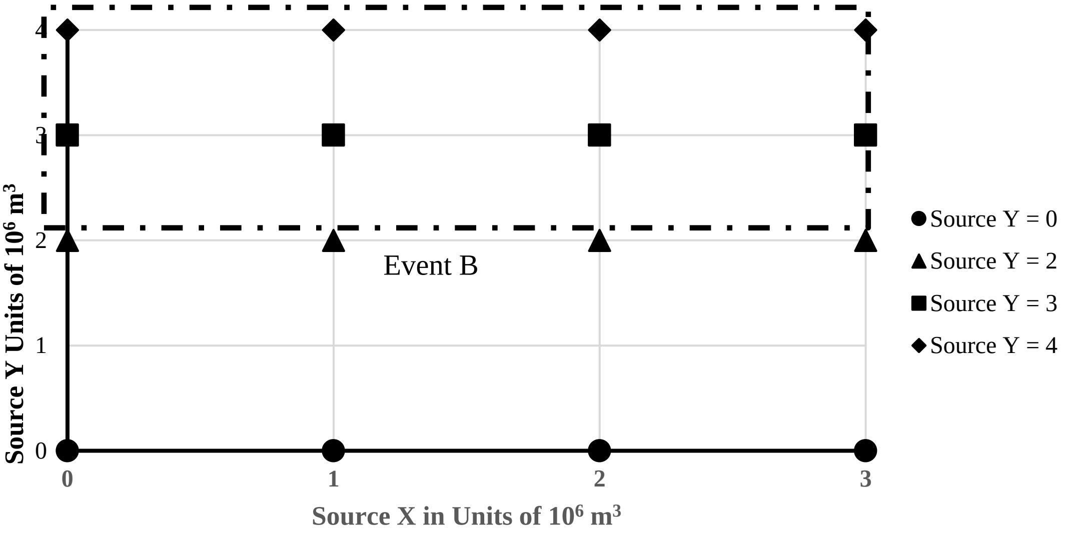
Sample space is shown below:



Event A ≡ {source X is less than 3 units} is shown below:



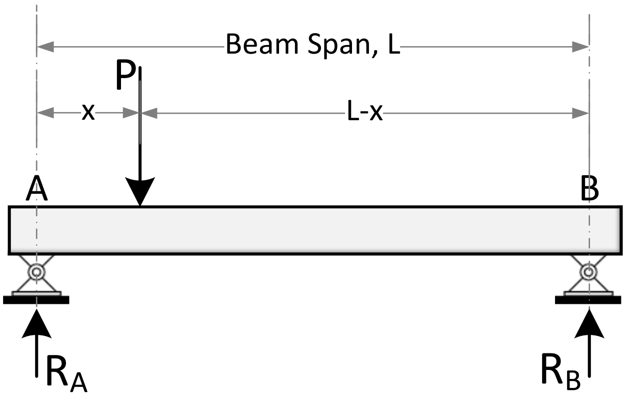
While event B ≡ {source Y is more than 2 units} will be:



Example 4.2-23

Use two-dimensional sample space to describe sample space for reactions R_A and R_B of a simply supported beam shown below. The concentered force, P, can move anywhere between points A and B.

Then indicate the events B_1 corresponding to location $0 \le x \le L/2$, and B_2, with $L/2 < x \le L$, are mutually exclusive?



Solution

From engineering mechanics, one can show that:

$$R_A = P \times \frac{L-x}{L}, \quad R_B = P \times \frac{x}{L}$$

Therefore, the relation is linear and one needs two points to draw sample space.

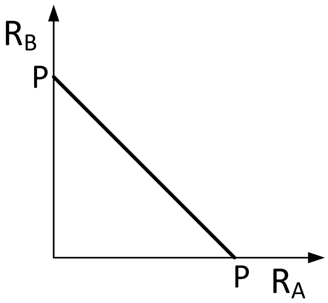
When force P acts at point A,

$$R_A = P, \qquad R_B = 0$$

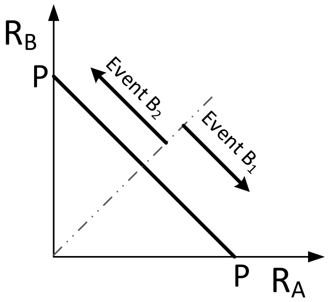
while when force P acts at point B,

$$R_A = 0, \qquad R_B = P$$

The sample space is shown in Figure below:



With events B_1 and B_2 indicated on it, sample space would be



The events B_1 and B_2 are mutually exclusive.

4.3 COUNTING RULES FOR SAMPLE SPACE

Instead of computing number of all possible outcomes for an experiment based on drawing of a related tr-e or sample space (***these methods may be time-consuming*** and ***inapplicable in sometimes***) following three mathematical methods could be used:

- Fundamental counting rule.
- Permutation rule.
- Combination rule.

These rules are explained below.

4.3.1 THE FUNDAMENTAL COUNTING RULE

4.3.1.1 BASIC CONCEPTS

In a sequence of n events in which the first one has k_1 possibilities and the second has k_2 and the third has k_3, and so forth, the total number of possibilities of the sequence will be:

$$\text{Total number of possibilities} = \mathrm{k}_1 . k_2 . k_3 . \dots \dots k_n \qquad \text{Eq. 4.3-1}$$

Note: In this case ***and*** means to multiply.

Example 4.3-1

A coin is tossed, and a die is rolled. Find the number of outcomes for the sequence of events.

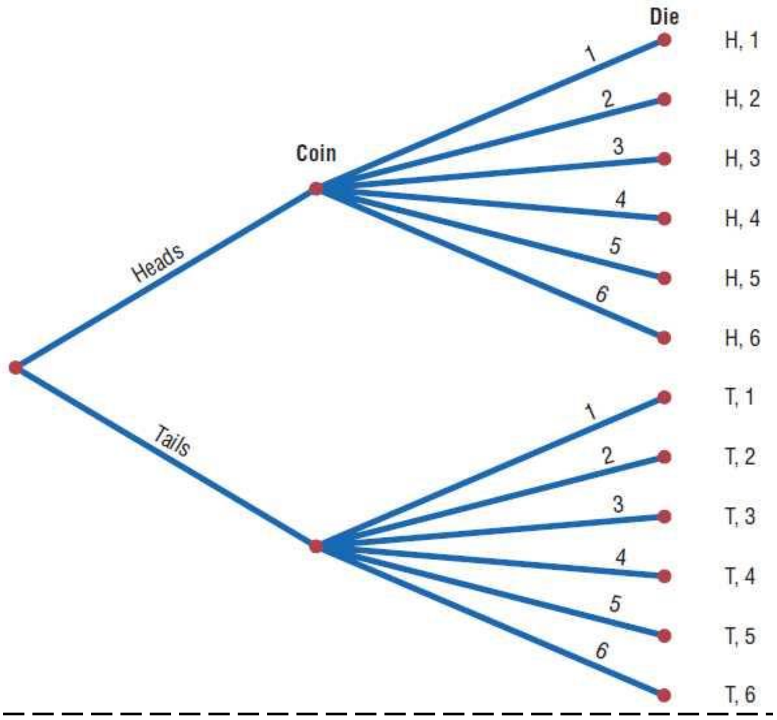
Solution

This example could be solved based on tree method and as shown below:

Or could be solved based on fundamental counting rule that based on following reasoning:

Since the coin can land either heads up or tails up and since the die can land with any one of six numbers showing face up, then there are

$$\begin{aligned} Possible\ outcomes &= 2_{Coin} \times 6_{Die} \\ &= 12 \end{aligned}$$



Example 4.3-2

A paint manufacturer wishes to manufacture several different paints. The categories include:

Color	Red, blue, white, black, green, brown, yellow
Type	Latex, oil
Texture	Flat, semigloss, high gloss
Use	Outdoor, indoor

How many different kinds of paint can be made if you can select one color, one type, one texture, and one use?

Solution

You can choose one color and one type and one texture and one use. Since there are 7 color choices, 2 type choices, 3 texture choices, and 2 use choices, the total number of possible different paints is:

$$Possible\ outcomes = 7_{Color} \times 2_{Type} \times 3_{Texture} \times 2_{Use} = 84$$

Example 4.3-3

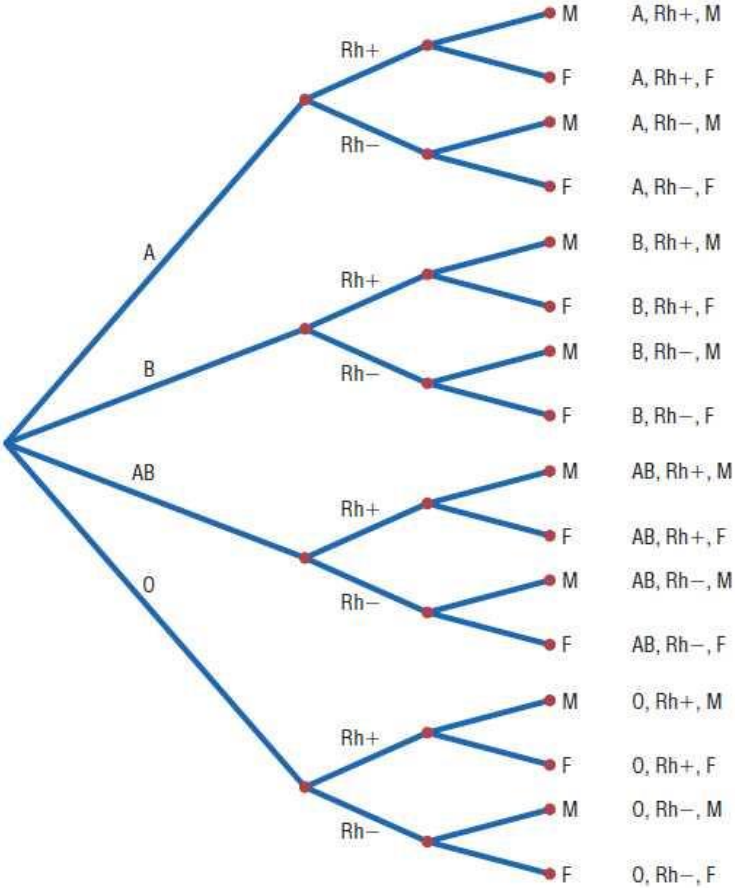
There are four blood types, A, B, AB, and O. Blood can also be Rh+ and Rh-. Finally, a blood donor can be classified as either male or female. How many different ways can a donor have his or her blood labeled?

Solution

Based on tree diagram, the possible outputs are presented in the indicated tree.

Based on fundamental computing rule: Since there are 4 possibilities for blood type, 2 possibilities for Rh factor, and 2 possibilities for the gender of the donor, there are:

$$Possible\ outcomes = 4_{Blood\ type} \times 2_{RH} \times 2_{Gender} = 16$$



4.3.1.2 Notes about Repetitions

When determining the number of different possibilities of a sequence of events, you must know whether repetitions are permissible.

Example 4.3-4

The manager of a department store chain wishes to make four-digit identification cards for her employees. How many different cards can be made if she uses the digits 1, 2, 3, 4, 5, and 6 and the following two cases:

- Repetitions are permitted.
- Repetitions are not permitted.

Solution

- Repetitions are permitted:
  Since there are 4 spaces to fill on each card and there are 6 choices for each space, the total number of cards that can be made is
  $$\text{Possible number of cards} = 6 \times 6 \times 6 \times 6 = 1296$$

- Repetitions are not permitted:
  Here the first digit can be chosen in 6 ways. But the second digit can be chosen in only 5 ways, since there are only five digits left, etc. Thus, the solution is
  $6\times\ 5\times\ 4\times\ 3=\ 360$

4.3.1.3 Factorial Notation

In order to put the counting rules in a compact form, factorial notation is usually used. For any counting n:

$n! = n(n-1)(n-2)\ \dots\dots\dots.1$

To use the formulas in the permutation and combination rules, a special definition of $0!$ is needed. $0!$ = 1.

For example:

$5!\ = 5\times\ 4\times\ 3\times\ 2\times\ 1$

$9! =\ 9\times\ 8\times\ 7\times\ 6\times\ 5\times 4\times\ 3\times\ 2\times\ 1$

4.3.2 Permutations

A **permutation** is an arrangement of *n* objects in a specific order.

4.3.2.1 Permutation Based on Fundamental Counting Rule

Permutation could be computed based on fundamental counting rule and as shown in following examples.

Example 4.3-5

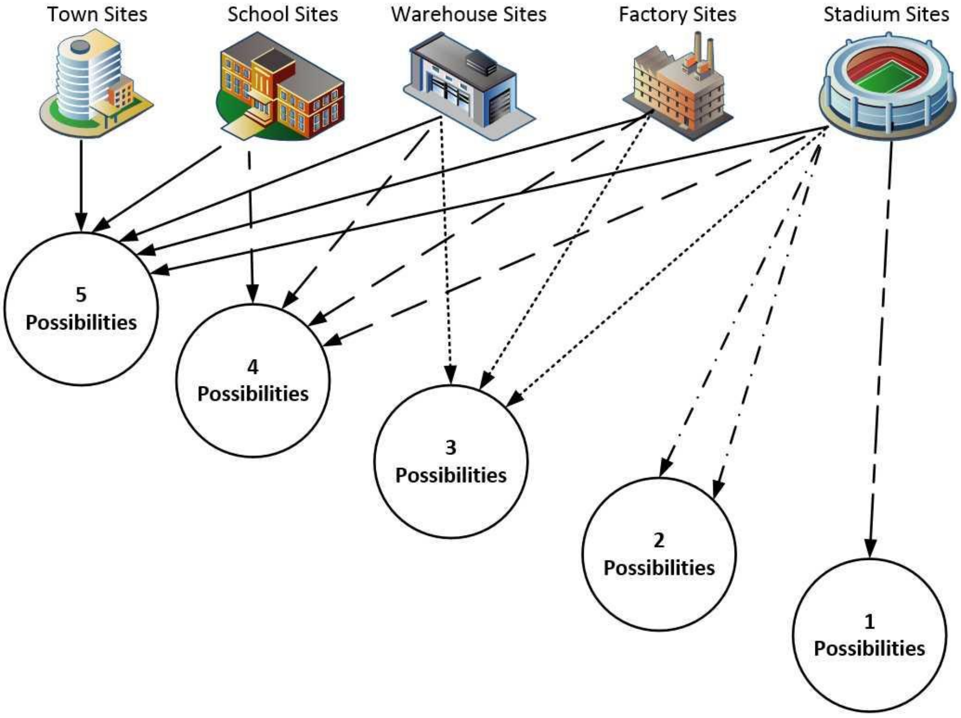
Suppose a business owner has a choice of 5 locations in which to establish her business. She decides to rank each location according to certain criteria, such as price of the store and parking facilities. How many different ways can she rank the 5 locations?

Solution

She has 5 choices for the first location, 4 choices for the second location, 3 choices for the third location, etc. There are:

$5! = 5\times 4\times 3\times 2\times 1$
$= 120$

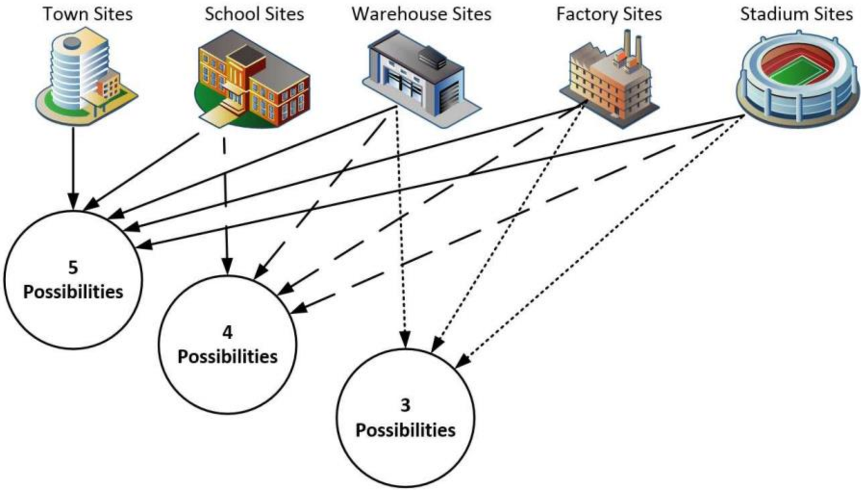
different possible rankings.



Example 4.3-6

Suppose the business owner in previous example wishes to rank only the top 3 of the 5 locations. How many different ways can she rank them?

Solution



Using the fundamental counting rule, she can select any one of the 5 for first choice, then any one of the remaining 4 locations for her second choice, and finally, any one of the remaining locations for her third choice, as shown.

$$Possible\ choices = 5 \times 4 \times 3$$
$$= 60$$

4.3.2.2 Permutation Based on Permutation Rule

The arrangement of n objects in a specific order using r objects at a time is called ***permutation of* n *objects taking* r *objects at a time***. It is written as ${}_nP_r$, and the formula is:

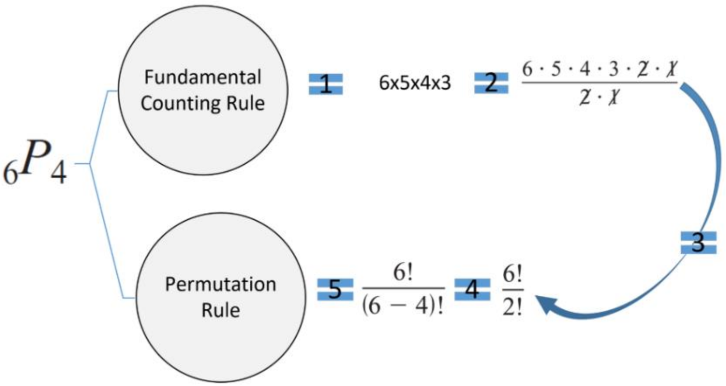
$${}_nP_r = \frac{n!}{(n-r)!} \qquad \text{Eq. 4.3-2}$$

Example 4.3-7

Based on computing ${}_6P_4$, show the relation between fundamental computing rule and permutation rule.

Solution

Relation between fundamental computing rule and permutation rule have been summarized below:


${}_6P_4$
Fundamental Counting Rule
1
6x5x4x3
2
$\frac{6\cdot5\cdot4\cdot3\cdot\cancel{2}\cdot\cancel{1}}{\cancel{2}\cdot\cancel{1}}$
3
4
$\frac{6!}{2!}$
5
$\frac{6!}{(6-4)!}$
Permutation Rule

Example 4.3-8

Resolve Example 4.3-5 above with using the permutation relation of Eq. 4.3-2.

Solution

This example aims to show that the permutation rule, Eq. 4.3-2, is applicable for $r = n$.

$${}_nP_r = \frac{n!}{(n-r)!} \Longrightarrow \quad {}_5P_5 = \frac{5!}{(5-5)!} = \frac{5!}{0!} = \frac{5\times4\times3\times2\times1}{1} = 120$$

Home Work 4.3-1

In how many ways can a party of 7 persons arrange themselves in a row of 7 chairs? Answer ${}_7P_7 = 7!$

Home Work 4.3-2

A contractor wishes to build 9 houses, each different in design. In how many ways can he place these houses on a street if 6 lots are on one side of the street and 3 lots are on the opposite side? Answer $_9P_9 = 362880$

Example 4.3-9

The advertising director for a television show has 7 ads to use on the program. If she selects 1 of them for the opening of the show, 1 for the middle of the show, and 1 for the ending of the show, how many possible ways can this be accomplished?

Solution

Since order is important, the solution is:

$$_nP_r = \frac{n!}{(n-r)!} \Longrightarrow \quad _7P_3 = \frac{7!}{(7-3)!} = \frac{7!}{4!} = 210$$

Hence, there would be 210 ways to show 3 ads.

Example 4.3-10

A school musical director can select 2 musical plays to present next year. One will be presented in the fall, and one will be presented in the spring. If she has 9 to pick from, how many different possibilities are there?

Solution

Order is important since one play can be presented in the fall and the other play in the spring.

$$_nP_r = \frac{n!}{(n-r)!} \Longrightarrow \quad _9P_2 = \frac{9!}{(9-2)!} = \frac{9!}{7!} = \frac{9 \times 8 \times \cancel{7!}}{\cancel{7!}} = 72$$

There are 72 different possibilities.

4.3.2.3 Circular Permutation

In general, n objects can be arranged in a circular in (n-1)! Ways.

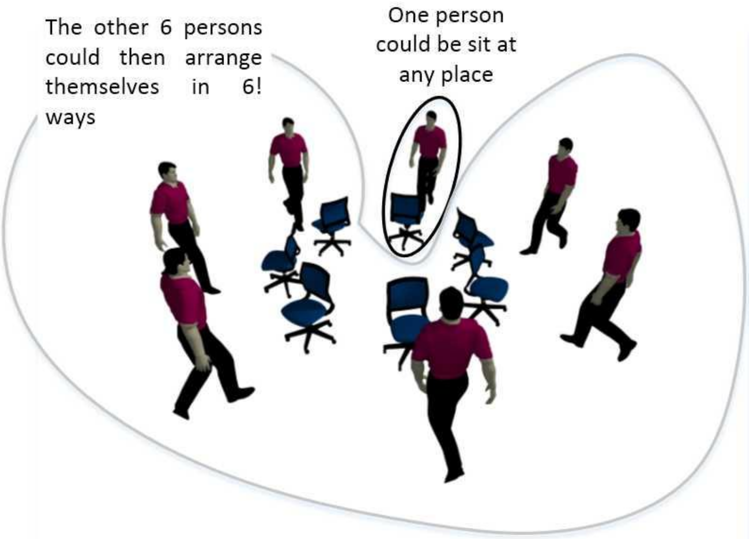
Example 4.3-11

In how many ways can a party of 7 persons arrange themselves around a circular table?

Solution

Referring to the indicated figure,

$$_7P_{7\ in\ circular\ manner} = (7-1)!$$


The other 6 persons could then arrange themselves in 6! ways
One person could be sit at any place

4.3.3 COMBINATIONS

4.3.3.1 BASIC CONCEPTS

- Suppose a dress designer wishes to select two colors of material to design a new dress, and she has on hand four colors. How many different possibilities can there be in this situation?
- This type of problem differs from previous ones in that the ***order of selection is not important***. That is, if the designer selects yellow and red, this selection is the same as the selection red and yellow.
- This type of selection is called a combination. ***The difference between a permutation and a combination is that in a combination, the order or arrangement of the objects is not important; by contrast, order is important in a permutation***. Example below illustrates this difference.
  A selection of distinct objects without regard to order is called a **combination.**

Example 4.3-12

Given the letters A, B, C, and D, list the permutations and combinations for selecting two letters.

Solution

The permutations are:

AB	BA	CA	DA
AC	BC	CB	DB
AD	BD	CD	DC

In ***permutations***, AB is different from BA.

AB	~~BA~~	~~CA~~	~~DA~~
AC	BC	~~CB~~	~~DB~~
AD	BD	CD	~~DC~~

But in ***combinations***, AB is the same as BA since the order of the objects does not matter in combinations. Therefore, if duplicates are removed from a list of permutations, what is left is a list of combinations, as shown. Hence the combinations of A, B, C, and D are AB, AC, AD, BC, BD, and CD. (Alternatively, BA could be listed, and AB crossed out, etc.). The combinations have been listed alphabetically for convenience, but this is not a requirement.

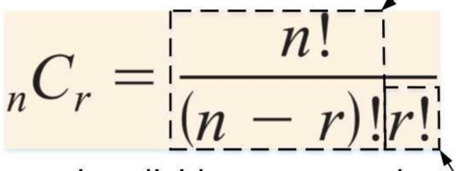
4.3.3.2 COMBINATION RULE

- Instead of working based on direct exclusion of one of outcomes that differs in order only from related tree diagram, combinations could be computed based on following rule:

$$ {}_nC_r = \frac{n!}{(n-r)!\,r!} \qquad \text{Eq. 4.3-3} $$

- Essen of the combination rule might be understood based on the indicated reasoning:

This part of relation is equivalent to permutation rule, then it considers outcomes as different even they differ in order only.



$$ {}_nC_r = \frac{n!}{(n-r)!\,r!} $$

This r! divides out the duplicates from the number of permutations.

Example 4.3-13

A newspaper editor has received 8 books to review. He decides that he can use 3 reviews in his newspaper. How many different ways can these 3 reviews be selected?

Solution

$$_nC_r = \frac{n!}{(n-r)!\,r!} \Longrightarrow \quad {}_8C_3 = \frac{8!}{(8-3)!\,3!} = \frac{8!}{5!\,3!} = \frac{8\times 7\times 6\times \cancel{5!}}{\cancel{5!}3!} = \frac{8\times 7\times 6}{3\times 2\times 1} = 56$$

There are 56 possibilities.

Example 4.3-14

In a staff, there are 7 women and 5 men. A committee of 3 women and 2 men is to be chosen. How many different possibilities are there?

Solution

Number of required possibilities could be computed based on following reasoning:

- First, you must select 3 women from 7 women, which can be done in $_7C_3$, or 35, ways.
- Next, 2 men must be selected from 5 men, which can be done in $_5C_2$, or 10, ways.
- Finally, by the fundamental counting rule, the total number of different ways is
  $35\times\ 10 = 350$
  since you are choosing both men and women.
- Above reasoning could be put in a formula form:
  $$_7C_3 \cdot {}_5C_2 = \frac{7!}{(7-3)!3!} \cdot \frac{5!}{(5-2)!2!} = 350$$

4.4 PROBABILITY AND ITS INTERPRETATIONS

- In our discussion of probability theory, we shall distinguish between the ***interpretation of probability*** and the ***calculus of the probability***.
- Interpretation of Probability:
  - The interpretation of probability deals with the question:
    What do we mean by probability and how should we interpret it?
  - For example, the probability of a head occurring when a coin is tossed is 1/2. How should this statement be interpreted?
  - There are three basic interpretations of probability:
    1. ***Classical Probability***.
    2. ***Objective***, ***Empirical***[1] or ***Relative Frequency Probability***.
    3. ***Subjective***[2] Probability.
- Calculus of Probability:
  The calculus of probability deals with the mathematical deductions obtained from the basic postulates (probability rules that should be satisfied regardless interpretation type) of probability.

4.4.1 INTERPRETATION OF PROBABILITY

Probability can be interpreted based on one of following approaches:

- ***Classical*** or ***prior*** probability,
- ***Objective*** or ***empirical*** or ***posterior*** probability,
- ***Subjective*** probability.

These interpretations have been discussed briefly in below.

4.4.2 CLASSICAL OR PRIOR PROBABILITY

- Classical probability ***assumes that all outcomes in the sample space are equally likely to occur***.
- For example, when a single die is rolled, each outcome has the same probability of occurring. Since there are six outcomes, each outcome has a probability of 1/6.
- On the other hand, when a card is selected from an ordinary deck of 52 cards, you assume that the deck has been shuffled, and each card has the same probability of being selected. In this case, it is 1/52.

4.4.2.1 ORIGIN OF CLASSICAL PROBABILITY INTERPRETATION

- Classical interpretation is based on the principle of ***insufficient reason*** (or ***principle of indifference***) was used by the famous Swiss mathematician ***Jacob Bernoulli*** (1655-1705) to define probabilities.

[1] According to Merriam-Webster Dictionary, ***empirical*** can be defined as originating in or based on observation or experience. it is from the Greek ***empeirikos***, meaning experienced.

[2] Subjective versus Objective Knowledge:
Subjective Knowledge:
That which has been acquired by direct experience and interpreted by the experiencer. Subjective knowledge is not about objective reality though the two are frequently confused, it is about subjective reality or as some call it, the soul. The two can be very similar if the owner is mature enough to have created an accurate model of the world.
Objective Knowledge:
That which has been acquired by indirect means primarily observation and analysis when one is not personally involved in the activity being observed.

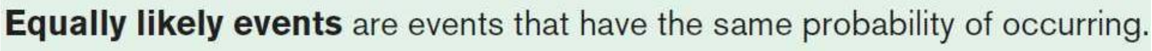
- Suppose a fair die is tossed and a student is asked the probability that a 2 will occur. He will probably give the answer 1/6. If a fair coin is tossed and he is asked the probability that a head will occur, he will probably give the answer 1/2. However, if he is asked why he answered 1/6 or 1/2, he may have trouble giving a precise reason.
- The ***principle of insufficient reason*** proposes that when there is no basis for preferring any one of the possible events (outcomes) to any other; then all should be treated as if they were equally likely to occur.
- The famous French mathematician ***P. S. Laplace*** (1749-1827) stated this principle as follows: "***The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible that is to say, to such as we may be equally undecided about in regard to their existence***".
- This principle of insufficient reason has several characteristics,
  - One of which is that it assumes ***symmetry of events***. Thus, we have a ***fair die***, or a ***fair coin***, or a ***fair deck of cards***.
  - A second characteristic is that it ***is based-on abstract reasoning and does not depend on experience***.

Jacob Bernoulli[a] (6 January 1655 August 1705) was one of the many prominent mathematicians in the ***Bernoulli family***.

He was an early proponent of ***Leibnizian calculus*** and had sided with Gottfried Wilhelm Leibniz during the Leibniz–Newton calculus controversy.

He is known for his numerous contributions to ***calculus***, and along with his brother Johann, was one of the founders of the ***calculus of variations***. He also discovered the ***fundamental mathematical constant*** e . However, ***his most important contribution*** was in ***the field of probability***, where ***he derived the first version of the law of large numbers***.

4.4.2.2 Equally Likely Events



Equally likely events are events that have the same probability of occurring.

4.4.2.3 Formula for Classical Probability

The probability of any event E is:

$$\mathrm{P(E)} = \frac{\text{Number of oucomes in E}}{Total\ number\ of\ oucomes\ in\ the\ sampe\ spave} = \frac{\mathrm{n(E)}}{\mathrm{n(S)}} \qquad \text{Eq. 4.4-1}$$

This probability is called ***classical probability***.

4.4.3 How to Express Probability?

- Probabilities can be expressed as
  - Fractions,
  - Decimals,
  - Or — where appropriate—percentages.
- If you ask, “What is the probability of getting a head when a coin is tossed?” typical responses can be any of the following three.
  - “One-half.”
  - “Point five.”
  - “Fifty percent.”

These answers are all equivalent and correct irrespective type of interpretation.

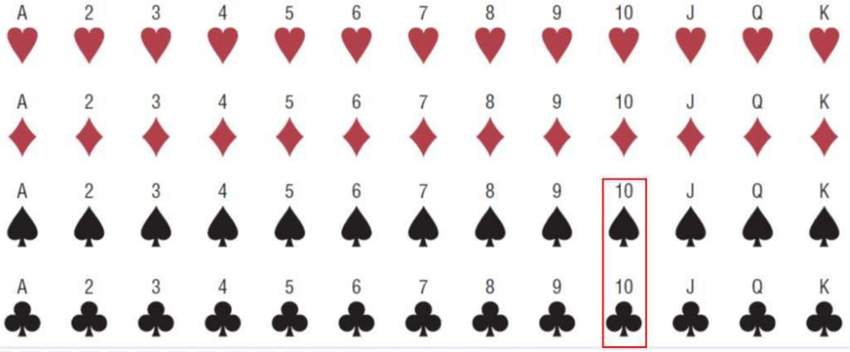
Example 4.4-1

Find the probability of getting a black 10 when drawing a card from a deck.

Solution

There are 52 cards in a deck, and there are two black 10s—the 10 of spades and the 10 of clubs. Hence the probability of getting a black 10 is:

$$P_{black\ 10} = \frac{2}{52} = \frac{1}{26}$$



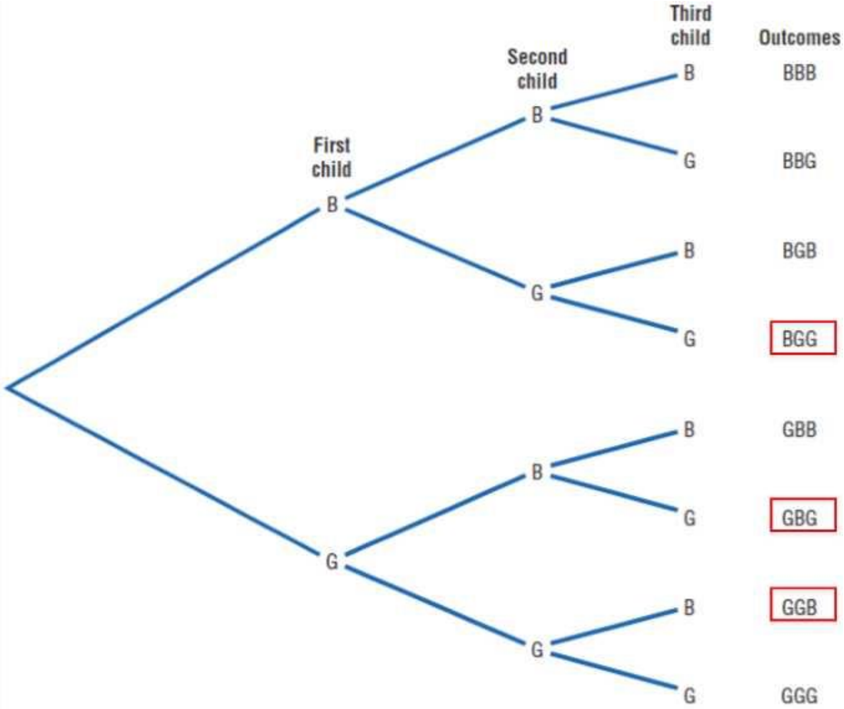
Example 4.4-2

If a family has three children, find the probability that two of the three children are girls.

Solution

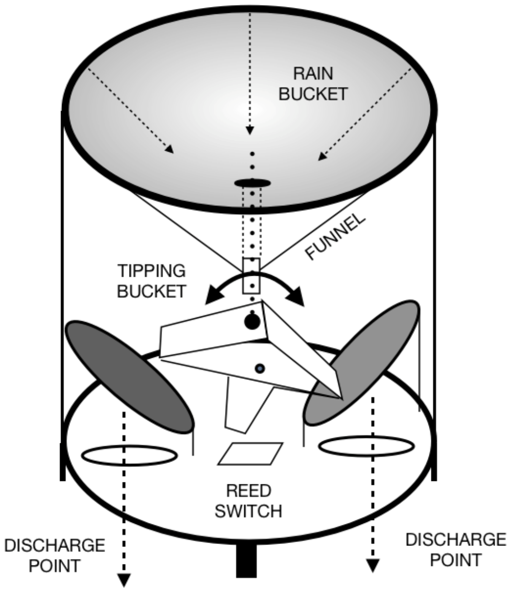
The sample space for the gender of the children for a family that has three children has eight outcomes, that is, BBB, BBG, BGB, GBB, GGG, GGB, GBG, and BGG. Since there are three ways to have two girls, namely, GGB, GBG, and BGG,

$$P_{two\ girls} = \frac{3}{8}$$



Example 4.4-3

A tipping bucket rain gauge type operates by means of a pair of buckets (see Figure below). After the rain has been collected through a funnel at the top, it fills the first bucket, which then overbalances and empties, thereby directing the flow of water into the second bucket. The alternating motion of the tipping buckets is transmitted to a recording time device, which provides a measure of the rainfall intensity by counting the rate of tilting. What is the probability that the water flows into either the left or the right bucket?



Solution

Because the two buckets are equal in volume and based on ***principle of insufficient reason*** (or ***principle of indifference***) 1/2 probability can be assigned to flow into either the left or the right bucket.

Example 4.4-4 Flood occurrence.

Consider the floods that exceed the previously established design flood in the outlet reach of the Bisagno River at Genoa, Italy, observed from 1931 to 1995. Records indicate that six floods occurred in the period, namely, in 1945, 1951 (twice), 1953, 1970, and 1992. The engineer is interested in evaluating the likelihood of the occurrence of such a flood in any year.

Solution

Let, A event representing the occurrence of at least one flood in a year**.** Based on ***principle of insufficient reason***

$$\mathrm{P(A)} = \frac{\mathrm{n(E)}}{\mathrm{n(S)}} = \frac{5}{65} = 0.077$$

This number measures the hydrological risk affecting the river site.

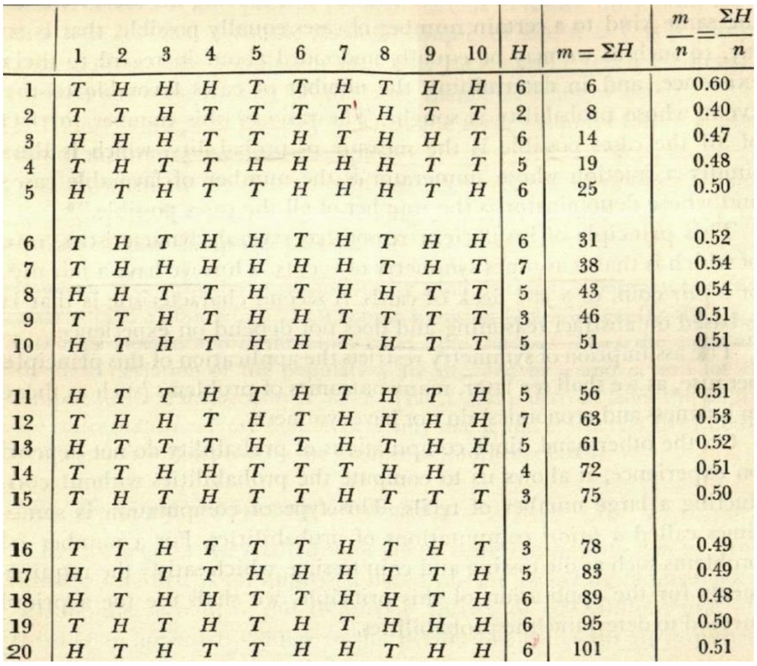
4.4.4 Objective or Empirical or Posterior Probability

- The difference between classical and empirical probability is that classical probability assumes that certain outcomes are equally likely (such as the outcomes when a die is rolled), while empirical probability relies on actual experience to determine the likelihood of outcomes.
- The basic reference to this approach is the Russian mathematician ***A. N. Kolmogorov's*** book, ***Foundations of the Theory of Probability (1933)***.

Andrey Nikolaevich Kolmogorov (25 April 1903 – 20 October 1987) was a 20th-century Soviet mathematician who made significant contributions to the ***mathematics of probability*** theory, ***topology***, ***intuitionistic logic***, ***turbulence***, ***classical mechanics***, ***algorithmic information theory*** and ***computational complexity***.

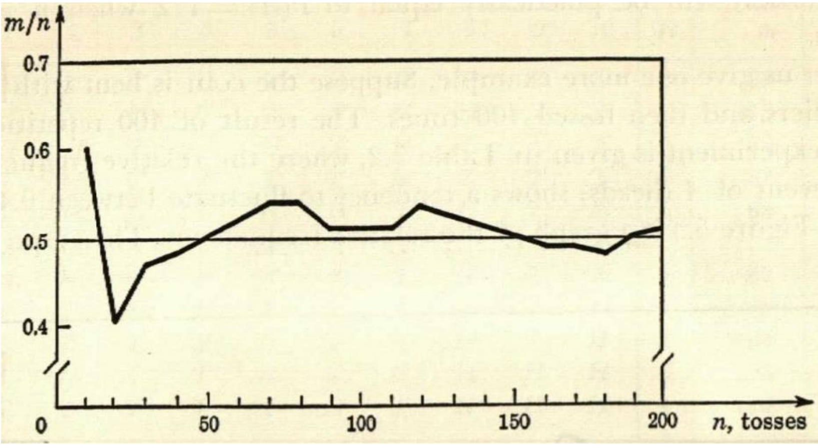
4.4.4.1 Basic Illustration

- Let us explain this approach with an illustration. Consider an experiment of tossing a fair coin. There are 2 possible outcomes (events), *E1* (heads) and *E2* (tails). Let us repeat this experiment 200 times under uniform conditions. The results are given in indicated tabke
  - The column labeled *H* shows the number of heads per 10 tosses.



	1	2	3	4	5	6	7	8	9	10	H	$m = \Sigma H$	$\frac{m}{n} = \frac{\Sigma H}{n}$
1	T	H	H	H	T	T	H	T	H	H	6	6	0.60
2	T	T	H	T	T	T	T	H	T	T	2	8	0.40
3	H	H	H	T	T	H	T	H	H	T	6	14	0.47
4	T	H	T	T	H	H	H	H	T	T	5	19	0.48
5	H	T	H	T	T	H	H	H	T	H	6	25	0.50
6	T	H	T	H	H	T	H	T	H	H	6	31	0.52
7	T	H	H	H	H	H	H	T	H	T	7	38	0.54
8	H	H	T	T	H	T	H	H	T	T	5	43	0.54
9	T	T	H	T	H	H	T	T	T	T	3	46	0.51
10	H	T	H	T	H	H	T	H	T	T	5	51	0.51
11	H	T	T	H	T	H	T	H	T	H	5	56	0.51
12	T	H	H	T	H	T	H	H	H	H	7	63	0.53
13	H	T	T	T	H	T	H	T	H	H	5	61	0.52
14	T	T	H	H	T	T	T	H	H	T	4	72	0.51
15	T	H	T	H	T	T	H	T	T	T	3	75	0.50
16	T	T	H	T	T	T	H	T	H	T	3	78	0.49
17	H	T	T	T	H	H	H	T	T	H	5	83	0.49
18	H	T	H	H	T	T	H	H	T	H	6	89	0.48
19	T	H	T	H	T	H	T	H	H	H	6	95	0.50
20	H	T	H	T	T	H	H	T	H	H	6	101	0.51

  - The column $m = \Sigma H$ is the cumulative sum of heads.
  - The column m/n is the relative frequency of heads for n tosses.
  - For example, in the third row, m/n. = 14/30=0.47 is the relative frequency of heads in 30 tosses.
- The fluctuations of the relative frequencies of heads, m/ n, fluctuate considerably when *n* is small, but as *n* becomes large, the amplitude of the fluctuations decreases. This phenomenon is expressed by saying: ***the relative frequency becomes stable, or the relative frequency shows statistical regularity, as n becomes large.***
- Let us show about statistical regularity, in terms of a graph. In Fig. below, we have relative frequencies on the vertical axis and the number of tosses n on the horizontal axis. We see from the figure that the amplitude of the fluctuations gradually decreases as n becomes larger and, in our present case, tends to fluctuate around the value 0.5.



- In mathematical form, statistical regularity can be written as follows:

$$Probability\ of\ Event\ A\ =\ P(A) = \lim_{n\to\infty}\frac{m}{n}$$

4.4.4.2 Basic Definition

Given a frequency distribution, the probability of an even being in a given class is:

$$\mathrm{P(E)} = \frac{\text{Frequency of the class}}{Total\ frequencies\ in\ the\ distribution} = \frac{\mathrm{f}}{\mathrm{n}} \qquad \text{Eq. 4.4-2}$$

The probability is called ***empirical probability*** and ***is based on observation***.

Example 4.4-5

After observing the vehicles at the intersection every weekday for a year, the traffic engineer in the ***Example 4.2-4***, Traffic Example, of the previous section might assign the observed relative frequencies of the simple events, "no cars," "one stopped car," etc., to the sample points Eo, E1, ... , E2o. Which type of probability he had assigned?

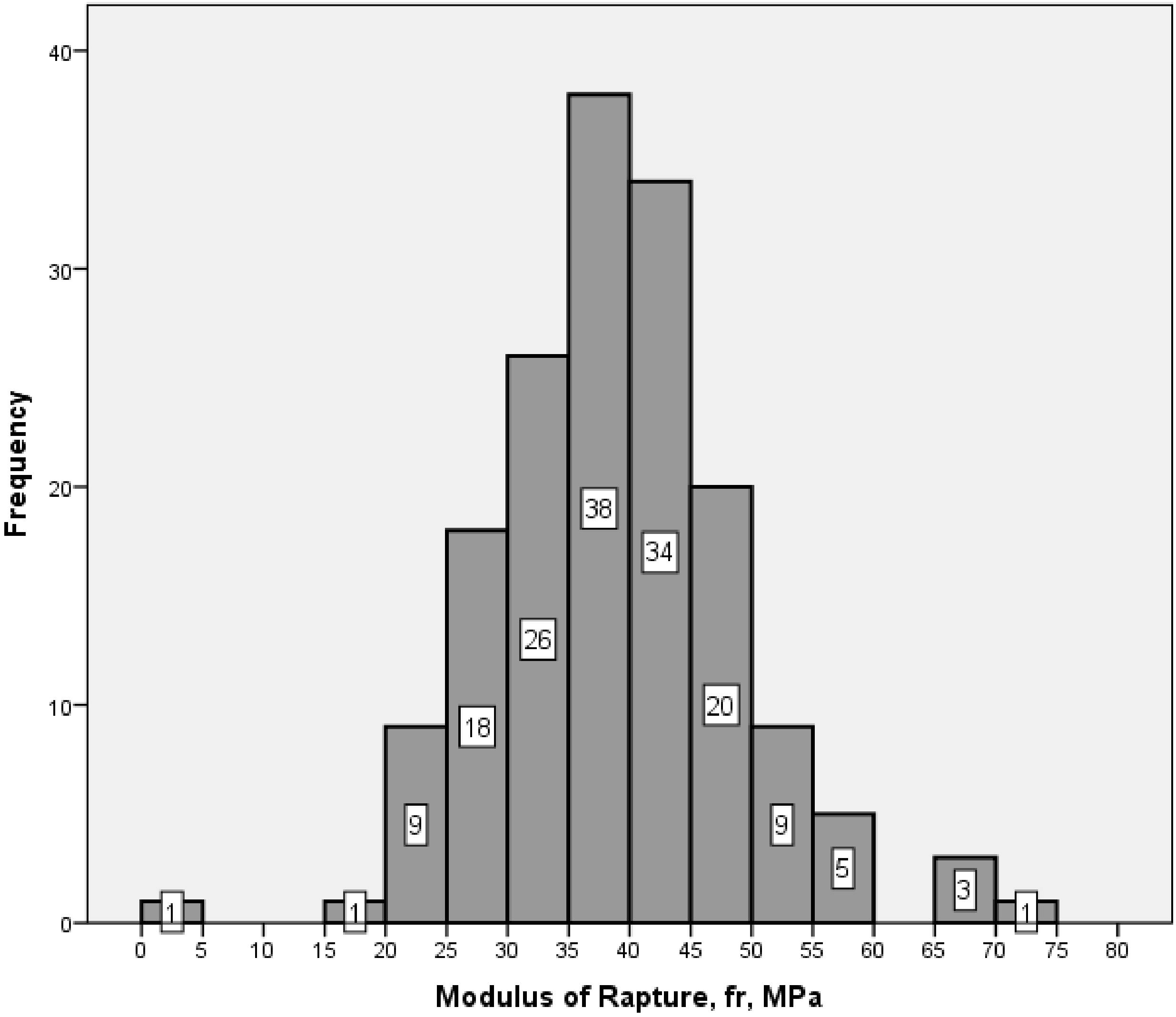
Solution

As they based on relatively a long period of observation and data gathering, a year, therefore assigned relative frequency can be considered as an ***objective probability***.

Example 4.4-6

Data of 165 tests to measure tensile strength of timber beams, expressed in terms of modulus of rapture, f_r, (MPa), have been summarized in terms of the histogram shown below. In terms of ***objective probability interpretation***:

- What is the probability that modulus of rapture in the range from 20 to 25 MPa?
- If one is to use this material under a maximum design strength of 25 N/mm^2, what is the reliability of a structure using this timber?



Solution

According to objective interpretation of probability, relative frequency can be to measure probability.

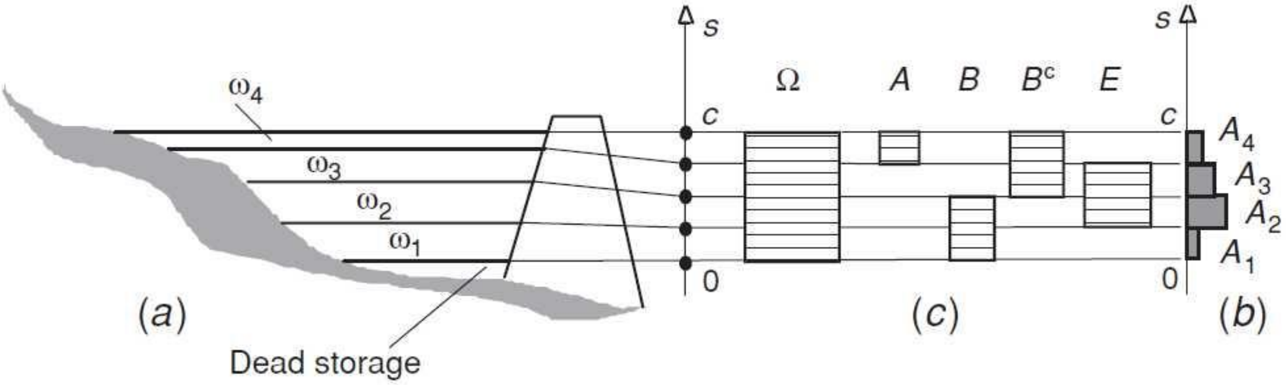
$$P(20 \leq f_r \leq 25) \approx \frac{9}{165} = 0.0545 = 5.45\ \%$$

With design strength of 25 MPa, the structure reliability will equal to the probability of:

$$P \geq 25 = \frac{18 + 26 + 38 + 34 + 20 + 9 + 5 + 3 + 1}{165} = \frac{154}{165} = 0.933 = 93.3\%$$

Example 4.4-7

A storage S in a reservoir was discretized into four states, ω1, ω2, ω3, and ω4 shown in Figure below. Observations of reservoir storage are made at the end of each period of operation, say, one year. Suppose that after 36 years of reservoir operation, the following frequencies have been observed: 5, 15, 10, and 6, for simple events A1, A2, A3, and A4, respectively. Use objective interpretation to estimates the probability for these events.



Solution

According to objective probability interpretation, relative frequency can be used to estimate events probability.

$P(A_1) = \frac{5}{36}, P(A_2) = \frac{15}{36}, P(A_3) = \frac{10}{36}, P(A_4) = \frac{6}{36}$

Example 4.4-8

A quality-control engineer samples 100 items manufactured by a certain process and finds that 15 of them are defective. Answer by true or false for the following statements with explanations for your answers.

a. The probability that an item produced by this process is defective is 0.15.
b. The probability that an item produced by this process is defective is likely to be close to 0.15, but not exactly equal to 0.15.

Solution

Statement "a" is ***false*** as with sample data one cannot exactly determine the probability for a possible outcome.

Statement "b" is ***true*** as with sample data one can conclude that the probability of a possible outcome would be close to its relative frequency.

4.4.5 SUBJECTIVE PROBABILITY

- In subjective probability, a person or group makes an ***educated guess*** at the chance that an event will occur.
- This guess is based on the ***person's experience*** and evaluation of a solution.
- For example,
  - A physician might say that, based on her diagnosis, there is a 30% chance the patient will need an operation.
  - A seismologist might say there is an 80% probability that an earthquake will occur in a certain area.
- The Subjective Approach has following application:
  - The subjective approach may be applied to events that have not yet occurred.
  - The subjective approach may be applied to events that occur only once and does not required an experiment with a large number of trails or the assumption of statistical regularity.
  - Meaning of $n \to \infty$ in objective interpretation could be interpret in terms of a subjective approach.
- ***All three types of probability (classical, empirical, and subjective) are used to solve a variety of problems in business, engineering, and other fields.***

Example 4.4-9

Calling upon past experience in similar situations, a knowledge of local geology, and the taste of a handful of the material on the surface, a soils engineer might state the probabilities that each of several types of soil might be found below a particular footing. Which types of probabilities has assigned?

Solution

As he has assigned probability based on his experience without any data gathering, therefore a subjective probability has been assigned.

Home Work 4.4-1

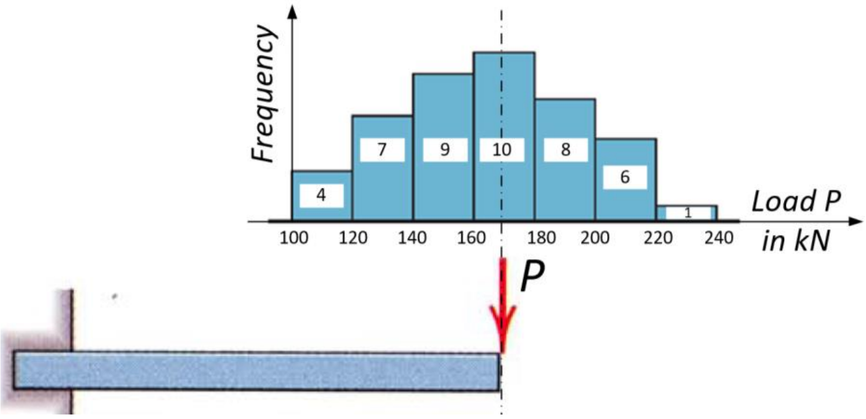


Figure 4.4-1: Cantilever beam for Home Work 4.4-1.

For the cantilever beam indicated in ***Figure 4.4-1***:

- What is the failure probability if the beam has a strength of 200 kN?
  What probability interpretation you use in your answer.
- What would the failure probability if the beam has a strength of 210 kN and based on which interpretation it is determined?

Home Work 4.4-2

For the simple beam indicated in ***Figure 4.4-2***:

- What is the failure probability if the beam has a strength of 300 kN? What probability interpretation you use in your answer.
- What would the failure probability if the beam has a strength of 310 kN and based on which interpretation it is determined?

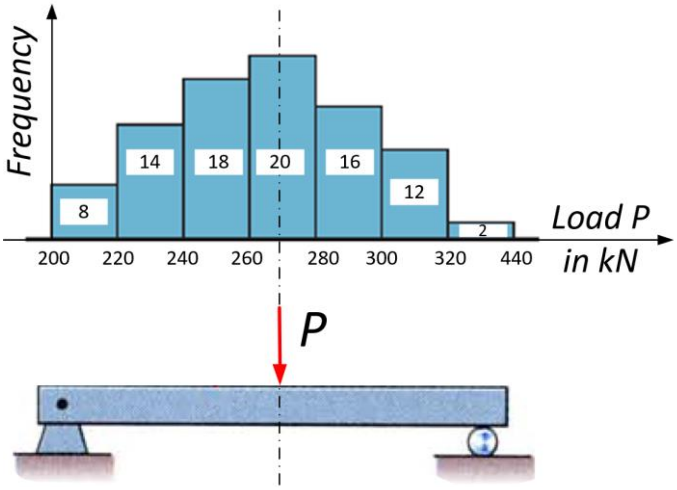


Figure 4.4-2: Simple beam for Home Work 4.4-2.

4.5 Axioms of Probability

- No matter how the engineer chooses to interpret the meaning of the probability measure and no matter what its source, as long as the assignment of these weights is consistent with three simple axioms, the mathematical validity of any results derived through the correct application of the axiomatic theory of probability is assured.
- We use the notation P(A) to denote the probability of an event A, which in the context of probability is frequently called a random event.
- The following conditions must hold on the probabilities assigned to the events in the sample space:

4.5.1 Axiom I

The probability of an event is a number greater than or equal to zero but less than or equal to unity:

$$0 \leq P(A) \leq 1 \qquad \text{Eq. 4.5-1}$$

4.5.2 Axiom II

The probability of the certain event S is unity:

$$\mathrm{P(S)} = 1 \qquad \text{Eq. 4.5-2}$$

where S is the event associated with all the sample points in the sample space.

4.5.3 Axiom III

The probability of an event which is the union of two mutually exclusive events is the sum of the probabilities of these two events:

$$P(A \cup B) = P(A) + P(B) \qquad \text{Eq. 4.5-3}$$

Since S is the union of all simple events, the third axiom implies that Axiom II could be written:

$$\sum_{all\, i} P(E_i) = 1 \qquad \text{Eq. 4.5-4}$$

in which the E, are simple events associated with individual sample points.

Example 4.5-1

When a single die is rolled, find the probability of getting a 9.

Solution

Since the sample space is 1, 2, 3, 4, 5, and 6, it is impossible to get a 9. Hence, the probability is:

$$P(9) = \frac{0}{6} = 0$$

Example 4.5-2

When a single die is rolled, what is the probability of getting a number less than 7?

Solution

Since all outcomes 1, 2, 3, 4, 5, and 6 are less than 7, the probability is

$$P(\text{number less than 7}) = \frac{6}{6} = 1$$

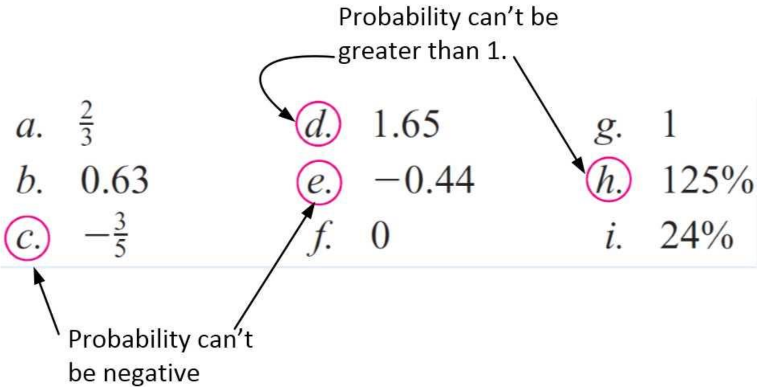
The event of getting a number less than 7 is ***certain***.

Example 4.5-3

A probability experiment is conducted. Which of these cannot be considered a probability outcome?

a. $\frac{2}{3}$
b. 0.63
c. $-\frac{3}{5}$
d. 1.65
e. -0.44
f. 0
g. 1
h. 125%
i. 24%

Solution



Probability can't be greater than 1.

$a.\ \frac{2}{3}$ $\quad d.\ 1.65$ $\quad g.\ 1$

$b.\ 0.63$ $\quad e.\ -0.44$ $\quad h.\ 125\%$

$c.\ -\frac{3}{5}$ $\quad f.\ 0$ $\quad i.\ 24\%$

Probability can't be negative

Example 4.5-4

Based on his personal experience, a geotechnics engineer has assigned following possibility for soil profile at a specific site:

$P(C) = \frac{1}{4}$, $P(S) = \frac{1}{4}$

where C and S are clay and sand respectively. What is the probability of that the profile be a rock?

Solution

As profile is classified as either sand, or clay, or rock, therefore these classifications are mutually exclusive events that represent whole sample space. According to Axiom III

$$P(S) + P(C) + P(R) = 1$$

$$P(R) = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$$

4.6 PROBABILITY OF COMPLEMENTARY EVENTS

Probability of a complementary event could be computed based on following relation (this relation has been derived based on definition of a complementary event and based on probability rules):

$P(\bar{E}) = 1 - P(E)$

$P(E) = 1 - P(\bar{E})$

$P(E) + P(\bar{E}) = 1$ Eq. 4.6-1

Example 4.6-1

If the probability that a person lives in an industrialized country of the world is 1/5, find the probability that a person does not live in an industrialized country.

Solution

$P_{not\ living\ in\ an\ industrialized\ country} = 1 - P_{living\ in\ an\ industrialized\ country}$

$P_{not\ living\ in\ an\ industrialized\ country} = 1 - \frac{1}{5} = \frac{4}{5}$

Example 4.6-2

If the probability that it will rain tomorrow is 0.20, what is the probability that it will not rain tomorrow? Would you recommend taking an umbrella?

Solution

$P_{no\ rain\ tomorrow} = 1 - 0.2 = 0.80$

Since the probability that it will not rain is 80%, you could leave your umbrella at home and be fairly safe.

Example 4.6-3

Consider the floods that exceed the previously established design flood in the outlet reach of the Bisagno River at Genoa, Italy, observed from 1931 to 1995. Records indicate that six floods occurred in the period, namely, in 1945, 1951 (twice), 1953, 1970, and 1992. What is the probability that no flood occurs in any year?

Solution

Let, A event representing the occurrence of at least one flood in a year. Therefore $\bar{A}$ represents complementary event where no flood occurs in any year.

Based on ***principle of insufficient reason***

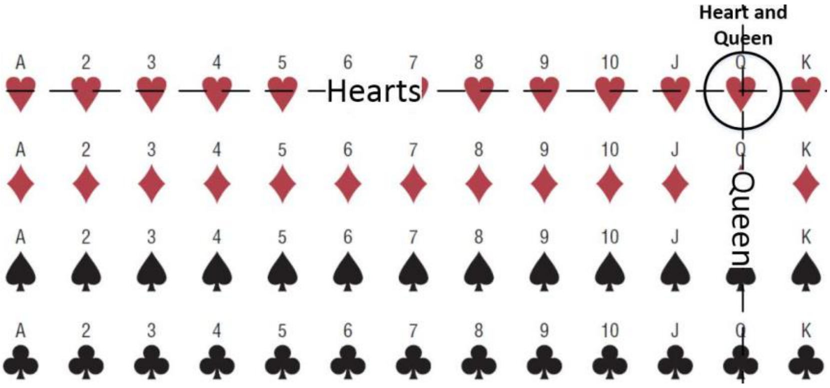
$$\mathrm{P(A)} = \frac{\mathrm{n(E)}}{\mathrm{n(S)}} = \frac{5}{65} = 0.077$$

Then, the probability that no flood occurs in any year would be:

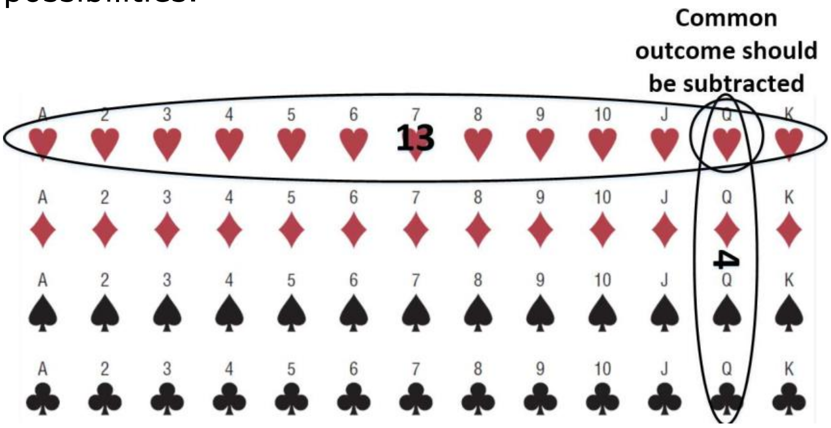
$P(\bar{A}) = 1 - 0.077 = 0.923$

4.7 MEANING OF "AND" AND "OR"

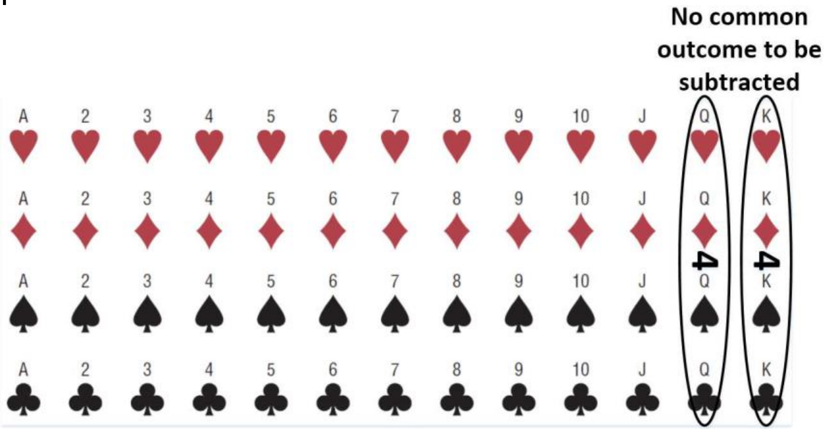
- In probability theory (**irrespective what interpretation is adopted**), it is important to understand the meaning of the words "*and*" and "*or*".
- For example, if you were asked to find the probability of getting ***a queen "and" a heart*** when you were drawing a single card from a deck, you would be looking for the queen of hearts. ***Here the word and means "at the same time."***



- The word *or* has two meanings.
  - For example, if you were asked to find the probability of selecting ***a queen or a heart*** when one card is selected from a deck, you would be looking for one of the 4 queens or one of the 13 hearts. In this case, the queen of hearts would be included in both cases and counted twice. So there would be
    4 + 13 -1 = 16
    possibilities.



  - On the other hand, if you were asked to find the probability of getting a queen *or* a king, you would be looking for one of the 4 queens or one of the 4 kings. In this case, there would be
    4 + 4 = 8
    possibilities.



- Definition of “Inclusive Or” and “Exclusive Or”
  - Inclusive or:
    In the first case, both events can occur at the same time; we say that this is an example of the ***inclusive or***.
  - Exclusive or:
    In the second case, both events cannot occur at the same time, and we say that this is an example of the ***exclusive or***.

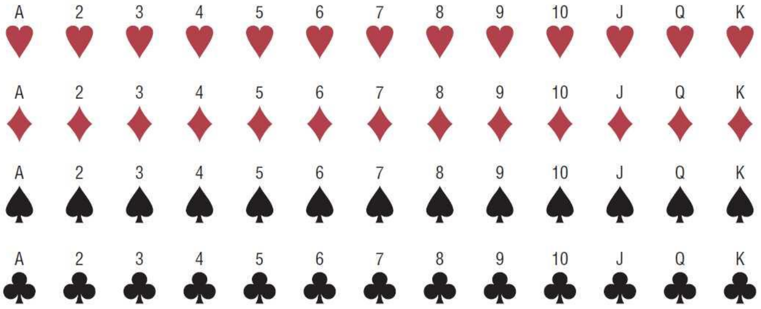
Example 4.7-1

A card is drawn from an ordinary deck. Find these probabilities.

a. Of getting a jack
b. Of getting the 6 of clubs (i.e., a 6 and a club)
c. Of getting a 3 or a diamond
d. Of getting a 3 or a 6

Solution

Refer to the sample space in the indicated figure



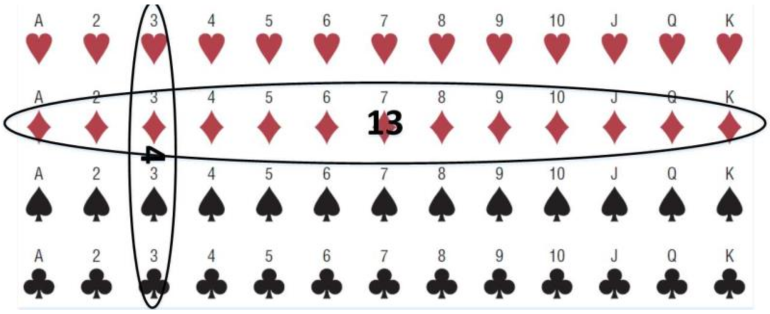
a. Of getting a jack:
There are 4 jacks so there are 4 outcomes in event E and 52 possible outcomes in the sample space. Hence,

$$P_{jack} = \frac{4}{52} = \frac{1}{13}$$

a. Of getting the 6 of clubs (i.e., a 6 and a club):
Since there is only one 6 of clubs in event E, the probability of getting a 6 of clubs is:

$$P(6\,of\;clubs) = \frac{1}{52}$$

b. Of getting a 3 or a diamond:
There are four 3s and 13 diamonds, but the 3 of diamonds (i.e. 3 and diamonds) is counted twice in this listing. Hence, there are 16 possibilities of drawing a 3 or a diamond, so



$$P(3\;or\;diamond) = \frac{13+4-1}{52} = \frac{16}{52} = \frac{4}{13}$$

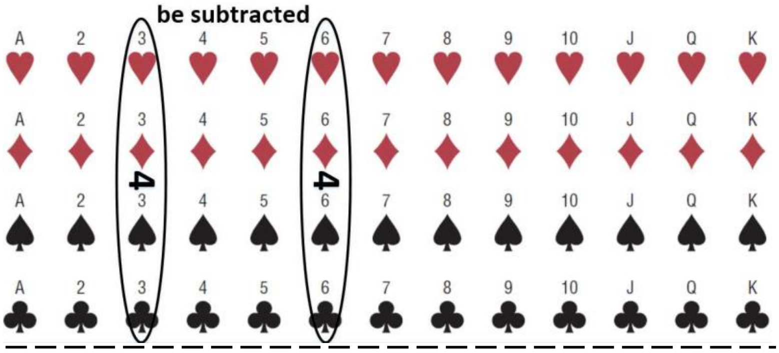
This is an example of the ***inclusive or***.

c. Of getting a 3 or a 6:

$$P(3\;or\;6) = \frac{8}{52} = \frac{2}{13}$$

This is an example of the ***exclusive or***.

No common outcome to be subtracted



4.8 THE ADDITION RULES FOR PROBABILITY

4.8.1 FIRST ADDITION RULE (FOR EXCLUSIVE MUTUALLY EVENTS)

The probability of two or more events can be determined by the ***addition rules.*** The first addition rule is used when the events are ***mutually exclusive***.

$$\mathrm{P(Aor\ B) = P(A) + P(B)} \qquad \text{Eq. 4.8-1}$$

Example 4.8-1

A city has 9 coffee shops: 3 Starbuck's, 2 Caribou Coffees, and 4 Crazy Mocho Coffees. If a person selects one shop at random to buy a cup of coffee, find the probability that it is either a Starbuck's or Crazy Mocho Coffees.

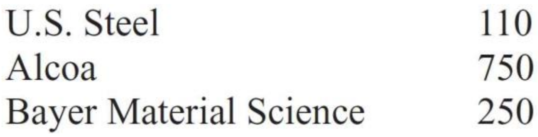
Solution

Since there are 3 Starbuck's and 4 Crazy Mochos, and a total of 9 coffee shops, and since the events are mutually exclusive, then:

$$P(\text{Starbuck's or Crazy Mochos}) = P(\text{Starbuck's}) + \mathrm{P}(\text{Crazy Mochos}) = \frac{3}{9} + \frac{4}{9} = \frac{7}{9}$$

Example 4.8-2

The corporate research and development centers for three local companies have the following number of employees:



U.S. Steel	110
Alcoa	750
Bayer Material Science	250

If a research employee is selected at random, find the probability that the employee is employed by U.S. Steel or Alcoa.

Solution

Since the events are mutually exclusive, then:

$$P(U.S.Stee\ or\ Alcoa) = P(U.S.Steel) + P(Alcoa) = \frac{110}{1110} + \frac{750}{1110} = \frac{860}{1110} = \frac{86}{111}$$

Example 4.8-3

A day of the week is selected at random. Find the probability that it is a weekend day.

Solution

Since the events are mutually exclusive, then:

$$P(Saturday\ or\ Sunday) = P(Saturady) + P(Sunday) = \frac{1}{7} + \frac{1}{7} = \frac{2}{7}$$

4.8.2 SECOND ADDITION RULE (FOR EXCLUSIVE MUTUALLY EVENTS)

When two events are ***not mutually exclusive***, we must subtract one of the two probabilities of the outcomes that are common to both events, since they have been counted twice.

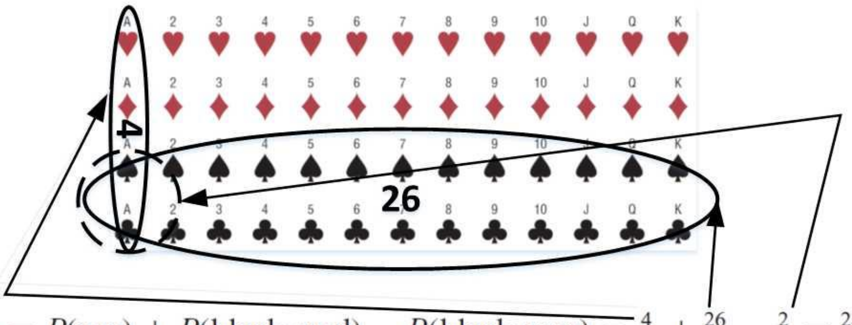
$$\mathrm{P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)} \qquad \text{Eq. 4.8-2}$$

Example 4.8-4

A single card is drawn at random from an ordinary deck of cards. Find the probability that it is either an ace or a black card.

Solution

Since there are 4 aces and 26 black cards (13 spades and 13 clubs), 2 of the aces are black cards, namely, the ace of spades and the ace of clubs. Hence the probabilities of the two outcomes must be subtracted since they have been counted twice.



$$P(\text{ace or black card}) = P(\text{ace}) + P(\text{black card}) - P(\text{black aces}) = \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13}$$

Example 4.8-5

In a hospital unit, there are 8 nurses and 5 physicians; 7 nurses and 3 physicians are females. If a staff person is selected, find the probability that the subject is a nurse or a male.

Solution

The sample space is shown here.

Staff	Females	Males	Total
Nurses	7	1	8
Physicians	3	2	5
Total	10	3	13

The probability is:

$$P(nurse\ or\ male) = P(nurse) + P(male) - P(male\ nurse)$$

$$P(nurse\ or\ male) = \frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13}$$

4.8.3 Which One of Two Addition Rule is More Fundamental?

- Rule 2 can also be used when the events are mutually exclusive, since
- P (A and B) will always equal 0, then it is more fundamental than rule 2.
- However, it is important to make a distinction between the two situations.

4.8.4 Addition Rules for More Than Two Events

- The probability rules can be extended to three or more events. For three ***mutually exclusive events*** A, B, and C,

  $$\mathrm{P(A\ or\ B\ or\ C) = P(A) + P(B) + P(C)} \qquad \text{Eq. 4.8-3}$$

- For three events that are ***not mutually exclusive***,

  $$\mathrm{P(A\ or\ B\ or\ C) = P(A) + P(B) + P(C) - P(A\ and\ B) - P(A\ and\ C) - P(B\ and\ C)} + P(A\ and\ B\ and\ C) \qquad \text{Eq. 4.8-4}$$

Example 4.8-6

If one card is drawn from an ordinary deck of cards, find the probability of getting the following.

- A king or a queen or a jack
- A club or a heart or a spade
- A king or a queen or a diamond
- An ace or a diamond or a heart

Solution

Refereeing to the deck of cards:

- A king or a queen or a jack:
  As events are mutually exclusive, then:

  $$P(king\ or\ queen\ or\ jack) = \frac{4}{52} + \frac{4}{52} + \frac{4}{52} = \frac{12}{52} = \frac{3}{13}$$

- A club or a heart or a spade:
  As events are mutually exclusive, then:
  $P(\text{club or a heart or a spade}) = \frac{13}{52} + \frac{13}{52} + \frac{13}{52} = \frac{39}{52} = \frac{3}{4}$
- A king or a queen or a diamond:
  As events are not mutually exclusive, then:
  $P(\text{king or a queen or a diamond}) = \frac{4}{52} + \frac{4}{52} + \frac{13}{52} - \frac{2}{52} = \frac{19}{52}$
- An ace or a diamond or a heart:
  As events are not mutually exclusive, then:
  $P(\text{ace or a diamond or a heart}) = \frac{4}{52} + \frac{13}{52} + \frac{13}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13}$

Example 4.8-7

For data of timber strength with size of 165 and that has histogram shown below. Define following events:

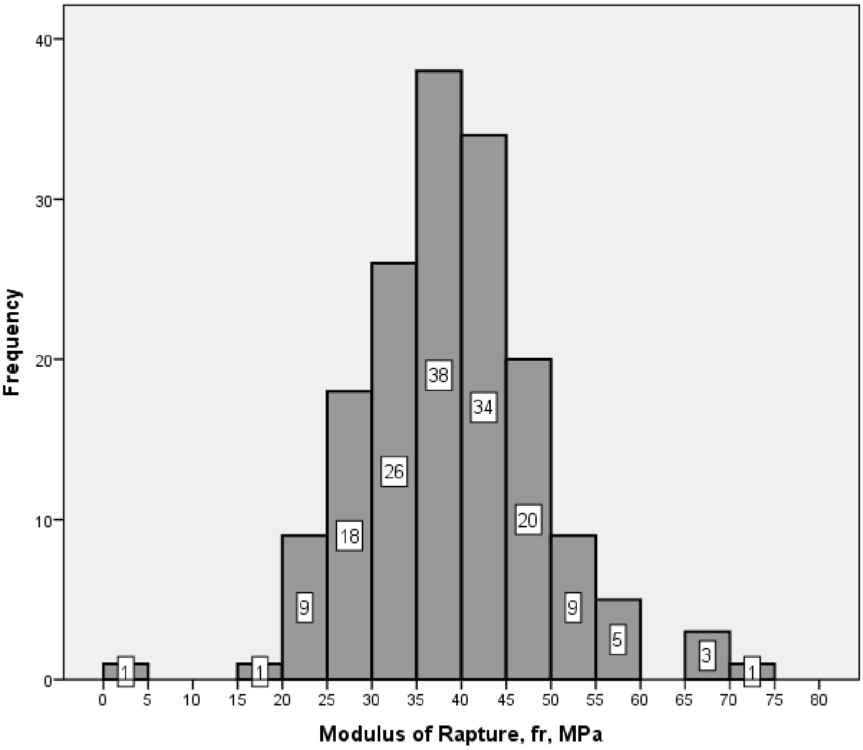
$A \equiv \{25 < S < 50\,MPa\}$

and

$B \equiv \{35 < S < 60\,MPa\}$

where S denotes the modulus of rupture of a test sample in MPa. Compute following probabilities:

$P(A)$, $P(B)$, $P(AB)$, and $P(A+B)$



Solution

Using the relative frequencies to estimate probabilities:

$$P(A) = \frac{18 + 26 + 38 + 34 + 20}{165} = 0.824, P(B) = \frac{38 + 34 + 20 + 9 + 5}{165} = 0.642$$

The intersection event, AB, is given by the common outcomes of A and B, so that:

$AB \equiv \{35 < S < 50\,MPa\}$

Accordingly,

$$P(AB) = P(A\ and\ B) = \frac{38 + 34 + 20}{165} = 0.557$$

Finally, the probability of the union event A + B can be computed by:

$P(A + B) = P(A) + P(B) - P(AB) = 0.824 + 0.642 - 0.557 = 0.909$

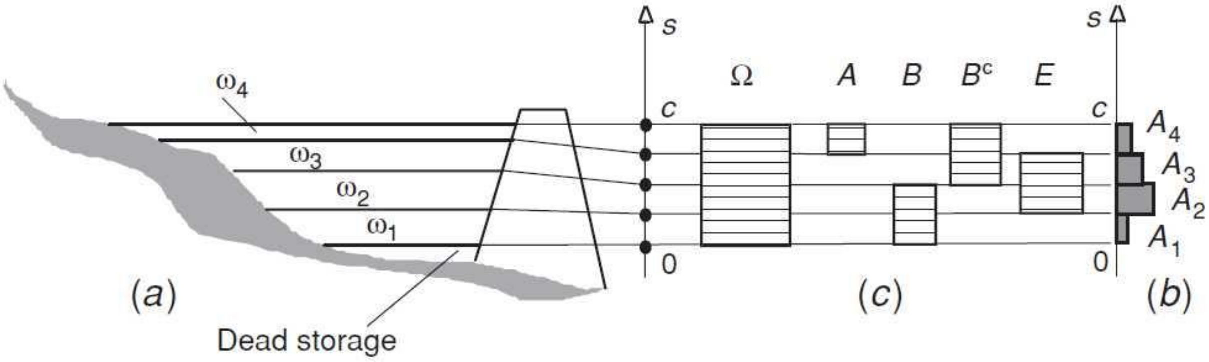
As verification, probability of event A+B can be determined directly from definition of the event and histogram based on objective interpretations of probability.

$A + B \equiv \{25 < S < 60\,MPa\}$

$$P(A + B) = \frac{18 + 26 + 38 + 34 + 20 + 9 + 5}{165} = \frac{150}{165} = 0.909$$

Example 4.8-8

A storage S in a reservoir was discretized into four states, ω1, ω2, ω3, and ω4 shown in Figure below. Observations of reservoir storage are made at the end of each period of operation, say, one year. Suppose that after 36 years of reservoir operation, the following frequencies have been observed: 5, 15, 10, and 6, for simple events A1, A2, A3, and A4, respectively.



Find the probability for following event:

$D = A_3 + A_4 \equiv \{S: c/2 \leq S < c\}$

Solution

According to objective probability interpretation, relative frequency can be used to estimate events probability.

$P(A_3) = \dfrac{10}{36}$

$P(A_4) = \dfrac{6}{36}$

As two events are mutually exclusive, therefore, $P(D)$ would be:

$P(D) = P(A_3) + P(A_4) = \dfrac{10}{36} + \dfrac{6}{36} = \dfrac{16}{36}$

4.8.5 Further Properties of Probability Functions

Some other useful properties can be demonstrated by using the three preceding axioms with basic aspects of set theory. These properties hold for any probability space.

4.8.5.1 Probability of a Contained Event

The probability of an event A that is contained in another event B does not exceed the probability of B; that is,

$P(A) \leq P(B), if\ A \subset B$ Eq. 4.8-5

4.8.5.2 Boole's Inequality

According to ***Boole's Inequality***, the probability of the union of n events does not exceed the sum of their probabilities; that is:

$P(A_1 + A_2 + \ldots + A_n) \leq P(A_1) + P(A_2) + \ldots + P(A_n)$ Eq. 4.8-6

4.8.5.3 Usefulness of the Further Properties

In certain cases, the above inequalities of a ***Contained Event*** and ***Boole's Inequality*** can provide conservative approximations of the required design probability in the absence of sufficient knowledge to determine the probability of a design event.

Example 4.8-9 Dam Failure

Two natural events can result in the failure of a dam in an earthquake-prone area.

- Firstly, a very high flood, exceeding the design capability of its spillway, say, event A, may destroy it.
- Secondly, a destructive earthquake can cause a structural collapse, say, event B.

Hydrological and seismological consultants estimate that the probability measures characterizing flood exceedance and earthquake occurrence on a yearly basis are

$P(A) = 0.02 \, and \, P(B) = 0.01$

respectively. The occurrence of one or both events can result in the failure of the dam. What is probability of dam failure?

Solution

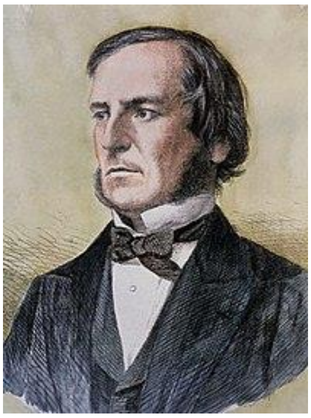
According to addition rule, probability of dam failure is:

$P(A + B) = P(A) + P(B) - P(AB)$

As nothing is known about $P(AB)$, therefore $P(A + B)$ can be approximately and conservatively estimated based on ***Boole's Inequality***:

$P(A + B) \approx P(A) + P(B) = 0.02 + 0.01 = 0.03$

This indicates a probability of not more than 3% that the dam will collapse in a given year.



George Boole (2 November 1815 – 8 December 1864) was a largely ***self-taught*** English ***mathematician***, ***philosopher*** and ***logician***, most of whose short career was spent as the first professor of mathematics at Queen's College, Cork in Ireland. He worked in the fields of ***differential equations*** and ***algebraic logic*** and is best known as the author of ***The Laws of Thought*** (1854) which contains ***Boolean algebra***. ***Boolean logic*** is credited with ***laying the foundations for the information age***.

4.9 THE MULTIPLICATION RULES AND CONDITIONAL PROBABILITY

- Previous section showed that the addition rules are used to compute probabilities for mutually exclusive and non-mutually exclusive events.
- This section introduces the multiplication rules.

4.9.1 THE MULTIPLICATION RULES FOR INDEPENDENT EVENTS

The ***multiplication rules*** can be used to find the probability of two or more events that occur ***in sequence***.

4.9.1.1 INDEPENDENT EVENTS

- To derive a multiplication rule, one should define ***independent events*** as:

Two events *A* and *B* are **independent events** if the fact that *A* occurs does not affect the probability of *B* occurring.

- Examples of Independent Events:
  - Rolling a die and getting a 6, and then rolling a second die and getting a 3.
  - Drawing a card from a deck and getting a queen, replacing it, and drawing a second card and getting a queen.

4.9.1.2 FORMULA OF MULTIPLICATION RULE

- To find the probability of ***two independent events*** that occur in sequence, you must find the probability of each event occurring separately and then multiply the answers.

$$\mathrm{P(A\ and\ B) = P(A).P(B)} \qquad \text{Eq. 4.9-1}$$

- For example, if a coin is tossed twice, the probability of getting two heads is:

$$\frac{1}{2}.\frac{1}{2}=\frac{1}{4}$$

- This result can be verified by looking at the sample space HH, HT, TH, TT. Then:

$$P(HH)=\frac{1}{4}$$

Example 4.9-1

A coin is flipped, and a die is rolled. Find the probability of getting a head on the coin and a 4 on the die.

Solution

As the two evets are intuitively independent, then:

$$P(head\ and\ 4)=P(head).P(4)=\frac{1}{2}.\frac{1}{6}=\frac{1}{12}$$

Example 4.9-2

Resolve previous example but based on analysis of sample space.

Solution

The sample space is:

H1 H2 H3 H4 H5 H6 T1 T2 T3 T4 T5 T6

Size of sample space is 12. As there is only one outcome in the specified event (H4), then the probability is:

$$P(H4)=\frac{1}{12}$$

Example 4.9-3

A card is drawn from a deck and ***replaced***; then a second card is drawn. Find the probability of getting a queen and then an ace.

Solution

The probability of getting a queen is 4/52, and since the card is replaced, the probability of getting an ace is 4/52 and two events are independent. Hence, the probability of getting a queen and an ace is:

$$P(queen\ and\ ace) = P(queen).P(ace) = \frac{4}{52}.\frac{4}{52} = \frac{16}{2704} = \frac{1}{169}$$

Example 4.9-4

An urn contains 3 red balls, 2 blue balls, and 5 white balls. A ball is selected, and its color noted. Then it is ***replaced***. A second ball is selected, and its color noted. Find the probability of each of these.

- Selecting 2 blue balls.
- Selecting 1 blue ball and then 1 white ball.
- Selecting 1 red ball and then 1 blue ball.

Solution

The replacement indicates two independent events.

- Selecting 2 blue balls:

  $$P(blue\ and\ blue) = P(blue).P(blue) = \frac{2}{10}.\frac{2}{10} = \frac{4}{100} = \frac{1}{25}$$

- Selecting 1 blue ball and then 1 white ball:

  $$P(blue\ and\ white) = P(blue).P(white) = \frac{2}{10}.\frac{5}{10} = \frac{10}{100} = \frac{1}{10}$$

- Selecting 1 red ball and then 1 blue ball:

  $$P(red\ and\ blue) = P(red).P(blue) = \frac{3}{10}.\frac{2}{10} = \frac{6}{10} = \frac{3}{50}$$

4.9.1.3 Multiplication Rule for Three or More Independent Events

Multiplication rule 1 can be extended to three or more independent events by using the formula:

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots \text{ and } K) = P(A) \cdot P(B) \cdot P(C) \cdots P(K)$$

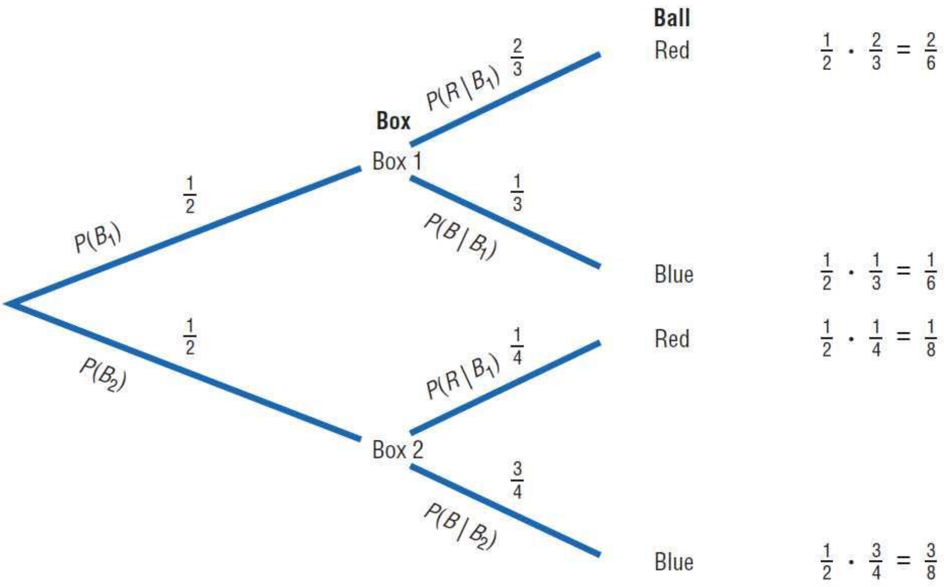
4.9.1.4 Using of Tree Diagram to Solve Probability Problems when Events are Sequential

Aforementioned multiplication formula can be intuitively grasped if one use tree diagram in his solution.

Example 4.9-5

Box 1 contains 2 red balls and 1 blue ball. Box 2 contains 3 blue balls and 1 red ball.

A coin is tossed. If it falls heads up, box 1 is selected and a ball is drawn. If it falls tails up, box 2 is selected and a ball is drawn. Find the probability of selecting a red ball.


Ball
Box
Box 1
Box 2
$P(B_1)$
$\frac{1}{2}$
$P(B_2)$
$\frac{1}{2}$
$P(R|B_1)$ $\frac{2}{3}$
$P(B|B_1)$ $\frac{1}{3}$
$P(R|B_1)$ $\frac{1}{4}$
$P(B|B_2)$ $\frac{3}{4}$
Red $\frac{1}{2} \cdot \frac{2}{3} = \frac{2}{6}$
Blue $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$
Red $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$
Blue $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$

Solution

Tree diagram for this example is:

$P(red\ ball) = \frac{2}{6} + \frac{1}{8} = \frac{11}{24}$

4.9.1.5 NON-REPLACEMENT COULD BE NEGLECTED FOR LARGE POPULATIONS

When a small sample is selected from a large population and the subjects are not replaced, the probability of the event occurring changes so slightly that for the most part, it is considered to remain the same. Two Examples below illustrate this concept.

Example 4.9-6

A Harris poll found that 46% of Americans say they suffer great stress at least once a week. If three people are selected at random, find the probability that all three will say that they suffer great stress at least once a week.

Solution

As the population is so large (all Americans) compared to sample, then the probability of simple event (48%) could be assumed constant even without replacement.

Let S denote stress. Then

$P(S\ and\ S\ and\ S) = P(S).P(S).P(S) = 0.46 \times 0.46 \times 0.46 \approx 0.097$

Example 4.9-7

Approximately 9% of men have a type of color blindness that prevents them from distinguishing between red and green. If 3 men are selected at random, find the probability that all of them will have this type of red-green color blindness.

Solution

Let C denote red-green color blindness. Then

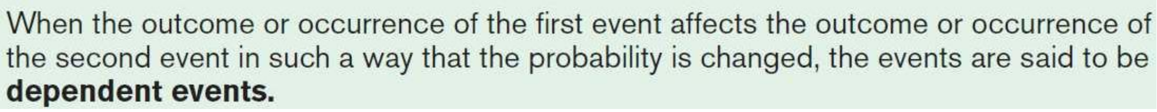
$P(C\ and\ C\ and\ C) = P(C).P(C).P(C) = 0.09 \times 0.09 \times 0.09 = 0.000729$

Hence, the rounded probability is 0.0007.

4.9.2 MULTIPLICATION RULE FOR DEPENDENT EVENT

4.9.2.1 DEFINITION OF DEPENDENT EVENT

- In all previous examples, the events were independent of one another, since the occurrence of the first event in no way affected the outcome of the second event.
- On the other hand, when the occurrence of the first event changes the probability of the occurrence of the second event, the two events are said to be ***dependent***.
- Dependent events are formally defined now.



When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be **dependent events.**

4.9.2.2 EXAMPLES ON DEPENDENT EVENTS

Here are some examples of dependent events:

- Drawing a card from a deck, not replacing it, and then drawing a second card.
- Selecting a ball from an urn, not replacing it, and then selecting a second ball.
- Having high grades and getting a scholarship.
- Parking in a no-parking zone and getting a parking ticket.

4.9.2.3 Formula for Multiplication Rule for Dependent Events

To find probabilities when events are dependent, use the multiplication rule with a modification in notation.

Example 4.9-8

Suppose a card is drawn from a deck and ***not replaced***, and then a second card is drawn. What is the probability of selecting an ace on the first card and a king on the second card?

Solution

The probability of getting an ace on the first draw is 4/52, and the probability of getting a king on the second draw is 4/51. By the multiplication rule, the probability of both events occurring is:

$$P(ace\ and\ king)=\frac{4}{52}.\frac{4}{51}=\frac{16}{2652}=\frac{4}{663}$$

4.9.2.4 Conditional Probability

4.9.2.4.1 Basic Definition

- In above example, the event of getting a king on the second draw *given* that an ace was drawn the first time is called a *conditional probability.*
- The ***conditional probability*** of an event ***B*** in relationship to an event ***A*** is the probability that event ***B*** occurs after event ***A*** has already occurred.
- The notation for conditional probability is:
  $P(B/A)$
- This notation does not mean that B is divided by A; rather, it means the probability that event B occurs given that event A has already occurred.
- With notation of ***conditional probability***, multiplication rule for dependent events can be written as:
  $$P(A\ and\ B)=P(A).P(B/A) \qquad \text{Eq. 4.9-2}$$
  Or
  $$P(B/A)=\frac{P(A\ and\ B)}{P(A)} \qquad \text{Eq. 4.9-3}$$

4.9.2.4.2 Venn Diagram for Conditional Probability

- The Venn diagram for conditional probability is shown in ***Figure 4.9-1*** above.
- In this case,
  $$P(B/A)=\frac{P(A\ and\ B)}{P(A)}$$
  which is represented by the area in the ***intersection or overlapping part of the circles A and B***, divided by ***the area of circle A***.
- *The reasoning here is that if you assume A has occurred, then A becomes the sample space for the next calculation and is the denominator of the probability fraction:*
  $$\frac{P(A\ and\ B)}{P(A)}$$

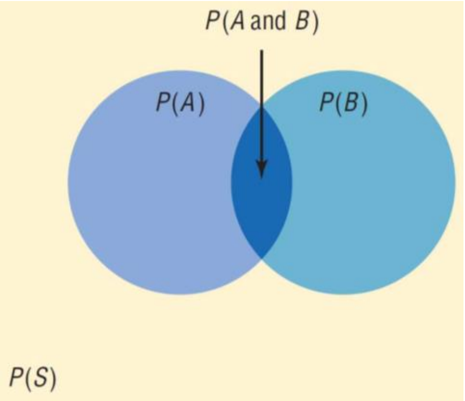


Figure 4.9-1: Venn diagrams representing the conditional probability.

- *The numerator* P(A and B) *represents the probability of the part of B that is contained in A.*
- Imposing a condition reduces the sample space.

4.9.2.4.3 Examples on Conditional Probability

Example 4.9-9

At a university in western Pennsylvania, there were 5 burglaries سطو reported in 2003, 16 in 2004, and 32 in 2005. If a researcher wishes to select at random two burglaries to further investigate, find the probability that both will have occurred in 2004.

Solution

In this case, the events are dependent since the researcher wishes to investigate two distinct cases. Hence the first case is selected and not replaced.

$$P(C_1 \, and \, C_2) = P(C_1).P(C_2/C_1) = \frac{16}{53}.\frac{15}{52} = \frac{60}{689}$$

Example 4.9-10

World Wide Insurance Company found that 53% of the residents of a city had homeowner's insurance (H) with the company. Of these clients, 27% also had automobile insurance (A) with the company. If a resident is selected at random, find the probability that the resident has both homeowner's and automobile insurance with World Wide Insurance Company.

Solution

As 27% has already given regarding the clients instead of the residents, therefore it represents the conditional probability $P(A/H)$ and can be used directly in the multiplication rule.

$$P(H \, and \, A) = P(H).P(A/H) = 0.53 \times 0.27 = 0.1431$$

Example 4.9-11

A box contains black chips and white chips. A person selects two chips ***without replacement***. If the probability of selecting a black chip and a white chip is 15/56, and the probability of selecting a black chip on the first draw is 3/8, find the probability of selecting the white chip on the second draw, given that the first chip selected was a black chip.

Solution

Let:

B is selecting a black chip and W is selecting a white chip. Then

$P(B \, and \, W) = P(B).P(W/B)$

As $P(B \, and \, W)$ and $P(B)$ are given and $P(W/B)$ is required, therefore the relation can be rewritten as:

$$P(W/B) = \frac{P(B \, and \, W)}{P(W/B)} = \frac{\frac{15}{56}}{\frac{3}{8}} = \frac{15}{56} \times \frac{8}{3} = \frac{5}{7}$$

Hence, the probability of selecting a white chip on the second draw given that the first chip selected was black is 5/7.

Example 4.9-12

The probability that Sam parks in a no-parking zone and gets a parking ticket is 0.06, and the probability that Sam cannot find a legal parking space and has to park in the no-parking zone is 0.20.

On Tuesday, Sam arrives at school and has to park in a no-parking zone. Find the probability that he will get a parking ticket.

Solution

Let N is parking in no-parking zone and T is getting a ticket. The example statement information can be written as:

$Sam\ parks\ in\ a\ no-parking\ zone\ and\ gets\ a\ parking\ ticket = P(N\ and\ T) = 0.06$

$Sam\ has\ to\ park\ in\ the\ no-parking\ zone\ = P(N) = 0.20$

Then

$$P(N\ and\ T) = P(N).P(T/N) \Longrightarrow P(T/N) = \frac{P(N\ and\ T)}{P(N)} = \frac{0.06}{0.20} = 0.30$$

Hence, Sam has a 0.30 probability of getting a parking ticket, given that he parked in a no-parking zone.

Example 4.9-13

A recent survey asked 100 people if they thought women in the armed forces should be permitted to participate in combat تشـارك في القتال. The results of the survey are shown.

Gender	Yes	No	Total
Male	32	18	50
Female	8	42	50
Total	40	60	100

Find these probabilities.

a. The respondent answered yes, given that the respondent was a female.

b. The respondent was a male, given that the respondent answered no.

Solution

Let

M respondent was male Y respondent answered yes

F respondent was a female N respondent answered no

a. The respondent answered yes, given that the respondent was a female.

$$P(Y/F) = \frac{P(Y\ and\ F)}{P(F)} = \frac{\frac{8}{100}}{\frac{50}{100}} = \frac{8}{50} = \frac{4}{25}$$

b. The respondent was a male, given that the respondent answered no.

$$P(M/N) = \frac{P(M\ and\ N)}{P(N)} = \frac{\frac{18}{100}}{\frac{60}{100}} = \frac{18}{60} = \frac{3}{10}$$

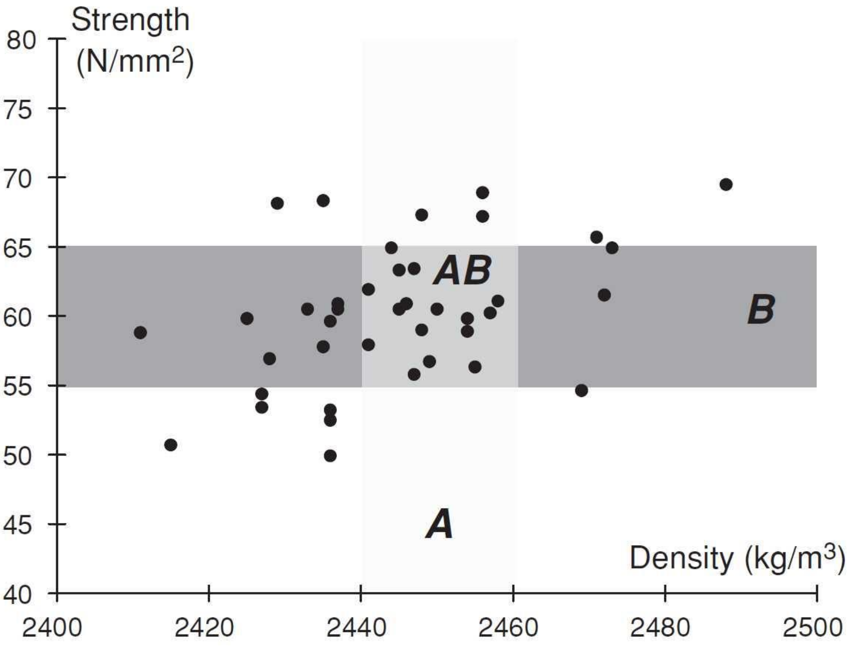
Example 4.9-14

Consider the n = 40 paired data of densities and compressive strengths of concrete given have been measured and presented in scatter plot below with following events:

$$A \equiv \left\{2440 < \rho_c < 2460 \frac{kg}{m3}\right\}$$

and

$$B \equiv \{55 < S_c < 65 \frac{N}{mm^2}\}$$



where ρ_c denotes the density of a concrete cube under test, measured in

kg/m^3, and S_c denotes the compressive strength of that cube, measured in N/mm^2. Based on above data, compute $P(A)$, $P(B)$, $P(AB)$ and $P(A/B)$.

Solution

Based on classical interpretation of probability:

$$P(A) = \frac{n_A}{n} = \frac{19}{40}, \quad P(B) = \frac{n_b}{n} = \frac{26}{40}$$

$$P(AB) = \frac{n_{AB}}{n} = \frac{16}{40}, P(A/B) = \frac{n_{AB}}{n_B} = \frac{16}{26}$$

Example 4.9-15

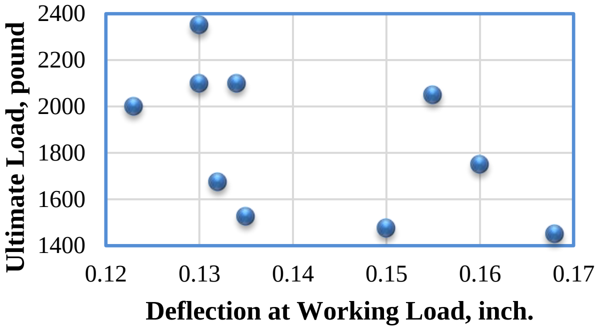
Ten timber beams were tested on simply supported span by a single concentrated load at mid-span. Test results are presented in scatter diagram of figure below. If two events, A and B have been defined as follows:

$A \equiv \{1800 < Ultimate\ Load < 2200\ lb\}$

and

$B \equiv \{0.14 < Deflection < 0.17\ inches\}$

Based on classical interpretation of probability compute the following probabilities: $P(A)$, $P(B)$, $P(AB)$ and $P(A/B)$.



Solution

Based on classical interpretation of probability:

$$P(A) = \frac{n_A}{n} = \frac{4}{10} = 0.4$$

$$P(B) = \frac{n_b}{n} = \frac{4}{10} = 0.4$$

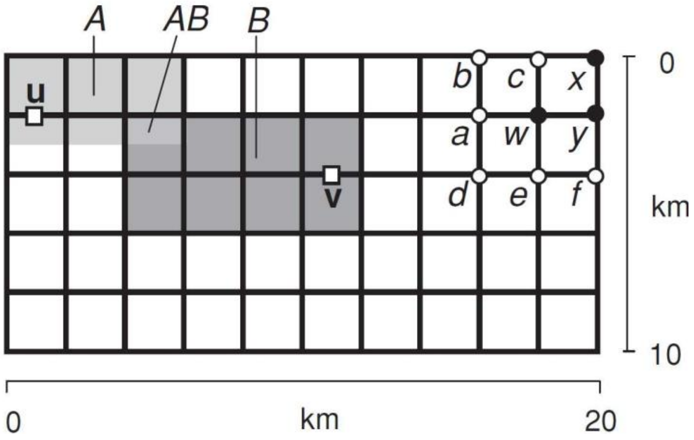
$$P(AB) = \frac{n_{AB}}{n} = \frac{1}{10} = 0.1$$

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{0.1}{0.4} = 0.25$$

Example 4.9-16

Consider a pipeline for the distribution of a water supply of an urban area of 200 km². The city plan is approximately rectangular with dimensions of 10 by 20 km, and it is uniformly covered by the network shown in below.

Pressures and flow rates are uniform throughout the whole network, so that losses are equally likely to occur within it. Define the events A and B as shown below, compute $P(A)$, $P(B)$, $P(AB)$, and $P(A/B)$.



Solution

As uniform pressures and uniform losses are mentioned in example statement, therefore classical probability interpretation has been implicitly assumed. Therefore, probabilities for above events will be:

$$P(A) = \frac{n_A}{n} = \frac{3 + 3 \times \frac{1}{2}}{10 \times 5} = 0.09, \qquad P(B) = \frac{n_B}{n} = \frac{8}{10 \times 5} = 0.16$$

$$P(AB) = \frac{\frac{1}{2}}{10 \times 5} = 0.01, \quad P(A/B) = \frac{n_{AB}}{n_B} = \frac{\frac{1}{2}}{8} = 0.0625$$

Example 4.9-17

The foundation of a wall can fail either by excessive settlement or from bearing capacity. The respective failures are represented by events A and B, with probabilities $P(A) = 0.005$, and $P(B) = 0.002$. The probability of failure in bearing capacity given that the foundation displays excessive settlement is $P(B|A) = 0.2$. Based on these information, find the probabilities for the following compound events:

- The probability of failure of the wall foundation, $P(A + B)$.
- The probability that there is excessive settlement in the foundation but there is no failure in bearing capacity, $P(A\bar{B})$.
- The conditional probability that the foundation has excessive settlement given that it fails, $P(A/B)$.

Solution

The probability of failure of the wall foundation, $P(A + B)$:

$$P(A + B) = P(A) + P(B) - P(AB) = P(A) + P(B) - P(B/A)\,P(A)$$

$$P(A + B) = 0.005 + 0.002 - 0.2 \times 0.005 = 0.006$$

The probability that there is excessive settlement in the foundation but there is no failure in bearing capacity, $P(A\bar{B})$:

$$P(A\bar{B}) = P(\bar{B}/A)\,P(A) = (1 - P(B/A)P(A) = (1 - 0.2) \times 0.005 = 0.004$$

The conditional probability that the foundation has excessive settlement given that it fails, $P(A/B)$:

$$P(A/B) = \frac{n_{AB}}{n_B} = \frac{n_{AB}}{n_B} \times \frac{n_A}{n_A} = \frac{n_{AB}}{n_A}\frac{n_A}{n_B} = P(B/A)\frac{P(A)}{P(B)} = 0.2 \times \frac{0.005}{0.002} = 0.5$$

Example 4.9-18

A question of the acceptability of an existing concrete culvert to carry an anticipated flow has arisen. Records are sketchy, and the engineer assigns estimates of annual maximum flow rates and their likelihoods of occurrence as follows:

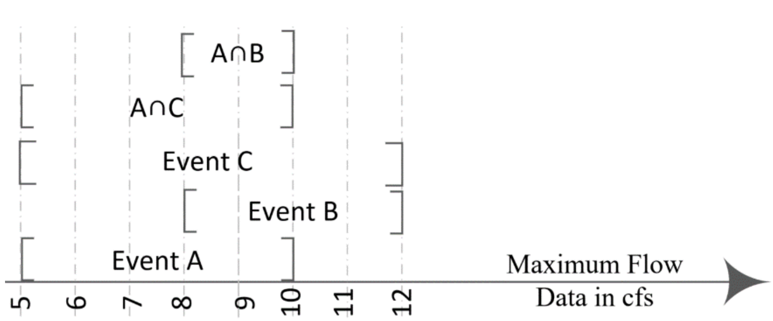
$Event\ A = [5\ to\ 10\ cfs]$	$P(A) = 0.6$
$Event\ B = [8\ to\ 12\ cfs]$	$P(B) = 0.6$
$Event\ C = A \cup B$	$P(C) = 0.7$

- Construct the sample space. Indicate events A, B, C, $A \cap C$, and $A \cap B$ on the sample space.
- Find $P(A \cap B)$, $P(\bar{A})$, $P(A|B)$, $P(B|A)$, and $P(B \cup \bar{A})$

Solution

- Sample space:

The sample space would be as indicated:



- Probabilities:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$0.7 = 0.6 + 0.6 - P(A \cap B)$

$P(A \cap B) = P(A\&B) = 0.6 + 0.6 - 0.7 = 0.5$■

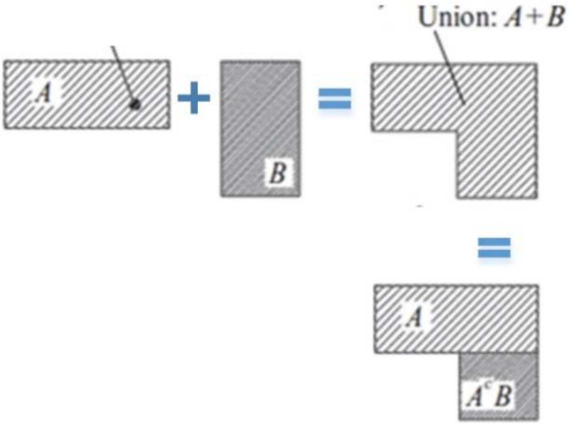
$P(\bar{A}) = 1 - P(A) = 1 - 0.6 = 0.4$■

$$P(A|B) = \frac{P(A\&B)}{P(B)} = \frac{P(A \cap B)}{P(B)} = \frac{0.5}{0.6} = 0.833■$$

$$P(B|A) = \frac{P(A\&B)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{0.5}{0.6} = 0.833■$$

$P(B \cup \bar{A}) = P(B) + P(\bar{A}) - P(B\bar{A})$

The probability of $P(B\bar{A})$ can be determined with referring to following ***Venn diagrams***:



$P(B\bar{A}) = P(A \cup B) - P(A) = 0.7 - 0.6 = 0.1$

$P(B \cup \bar{A}) = 0.6 + 0.4 - 0.1 = 0.9$ ■

4.9.2.5 Multiplication Rule for Three or More Dependent Events

The multiplication rule can be extended to three or more events, as shown in example below:

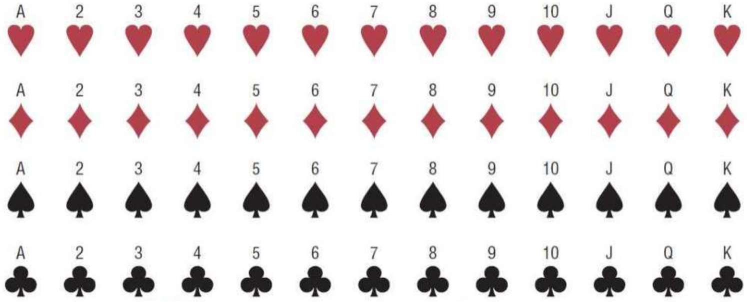
Example 4.9-19

Three cards are drawn from an ordinary deck and ***not replaced***. Find the probability of these events.

- Getting 3 jacks.
- Getting an ace, a king, and a queen in order.
- Getting a club, a spade, and a heart in order.
- Getting 3 clubs.

Solution

With referring to the indicated standard deck of cards:


A 2 3 4 5 6 7 8 9 10 J Q K
A 2 3 4 5 6 7 8 9 10 J Q K
A 2 3 4 5 6 7 8 9 10 J Q K
A 2 3 4 5 6 7 8 9 10 J Q K

- Getting 3 jacks.

$$P(3\,jacks) = \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{24}{132600} = \frac{1}{5525}$$

- Getting an ace, a king, and a queen in order.

$$P(ace\ and\ king\ and\ queen) = \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50} = \frac{64}{132600} = \frac{8}{16575}$$

- Getting a club, a spade, and a heart in order.

$$P(club\ and\ spade\ and\ heart) = \frac{13}{52} \cdot \frac{13}{51} \cdot \frac{13}{50} = \frac{2197}{132600} = \frac{169}{10200}$$

- Getting 3 clubs.

$$P(club\ and\ club\ and\ club) = \left(\frac{13}{52}\right) \times \left(\frac{12}{51}\right) \times \left(\frac{11}{50}\right) = \frac{11}{850}$$

4.10 THE THEOREM OF TOTAL PROBABILITY*

4.10.1 BASIC CONCEPTS THROUGH AN EXAMPLE

- On occasion, the probability of an event, say A, cannot be determined directly: its occurrence will depend on the occurrence or nonoccurrence of other events such as E_i, $i = 1, 2, .. n$, and the probability of A will depend on which of the $E_i's$ has occurred. On such an occasion, the probability of A would be composed of the conditional probabilities (conditioned on each of the $E_i's$) and weighted by the respective probabilities of the $E_i's$. Such problems require the ***theorem of total probability***.
- Before formally presenting the mathematical theorem. let us examine the following example to illustrate the essential elements of the theorem.

Example 4.10-1

The flooding of a river in the spring season will depend on the accumulation of snow in the mountains during the past winter season. The accumulation of snow may be described as ***heavy***, ***normal***, and ***light***. Clearly. if the snow accumulation in the mountains is heavy. the probability of flooding in the following spring will be high whereas, if the snow accumulation is light, this probability will be low.

Flooding, of course, may also be caused by rainfalls in the spring. With the following notations.

F is occurrence of flooding in the river,

H is heavy accumulation of snow,

N is normal accumulation of snow,

L is light (including no) accumulation of snow

Assume that:

$P(F/H) = 0.90, P(F/N) = 0.4, P(F/L) = 0.1$

whereas in a given winter season,

$P(H) = 0.20;\ P(N) = 0.50;\ P(L) = 0.30$

Based on these assumption and information, find the probability of flooding in the river during the following spring.

Solution

Based on the definition of the conditional probability, the probability of flooding in the river during the following spring season will be:

$P(F) = P(F/H)P(H) + P(F/N)P(N) + P(F/L)P(L)$

$P(F) = 0.9 \times 0.20 + 0.4 \times 0.50 + 0.10 \times 0.30 = 0.41$ ■

In Example 4.10-1, we can observe the following:

- The accumulations of snow in the past winter, namely. heavy, normal, and light. Are mutually exclusive.
- The probabilities of the three levels of snow accumulation, namely, H, N. and L add up to 1.0.

Therefore, the three events H, N, and L are ***mutually exclusive*** and also ***collectively exhaustive***.

4.10.2 FORMAL FORMULATION

- Formally. consider ***n events*** that are ***mutually exclusive*** and ***collectively exhaustive***, namely $E_1, E_2, \dots \dots, E_n$. That is,
  $E_1 \cup E_2 \cup E_3 \cup \dots \dots . . \cup E_n = S$

- Then, if A is an event in the same sample space S as shown in Figure 4.10-1, we derive the theorem of total probability as follows:
  $A = AS = A(E_1 \cup E_2 \cup E_3 \cup \dots\dots \cup E_n)$
  $A = (AE_1 \cup AE_2 \cup AE_3 \cup \dots\dots \cup AE_n)$
  where $AE_1 \cup AE_2 \cup AE_3 \cup \dots\dots \cup AE_n$ are also mutually exclusive as can be seen in the Venn diagram of Figure 4.10-1.

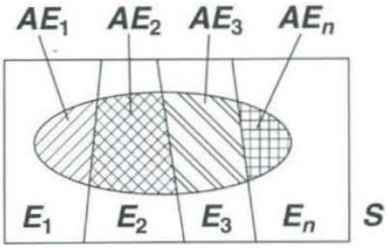


Figure 4.10-1: Intersection of A and $E_1, E_2, \dots\dots, E_n$, in sample space S.

- Therefore,
  $P(A) = P(AE_1) + P(AE_2) + P(AE_3) + \cdots \dots P(AE_n)$
- Finally, by virtue of the multiplication rule, we obtain the theorem of total probability as:
  $P(A) = P(A|E_1) + P(A|E_2) + P(A|E_3) + \cdots \dots P(A|E_n)$ Eq. 4.10-1
- Again, in applying the above total probability theorem, it is important to observe that the conditioning events $E_1, E_2, \dots\dots, E_n$, must be ***mutually exclusive*** and ***collectively exhaustive***.

Example 4.10-2

Hurricanes along the Gulf of Mexico and the eastern seaboard of the United States occur every year, mostly in the summer and fall. These hurricanes are classified into five categories from $C1$ through C5; to be classified as a hurricane, the wind speed must be at least 75 miles per hour (120 km/hr). The frequencies of hurricanes, of course, would decrease with the categories; for example, a Category $C5$ hurricane, with sustained wind > 150 mph (242 km/hr), would very seldom occur.

Assume that annually there can be at most one hurricane striking a particular area in the southern coast of Louisiana along the Gulf of Mexico, and the annual occurrence probabilities of the different hurricane categories are as follows:

$P(C1) = 0.35, P(C2) = 0.25, P(C3) = 0.14, P(C4) = 0.05, P(C5) = 0.01$

Structural damage to an engineered building in the reference area can be expected to occur depending on the category of hurricane that the building will be subjected to. Suppose the conditional probabilities of damage to the building are as follows:

$$P(D|C1) = 0.05; P(D|C2) = 0.10; P(D|C3) = 0.25; P(D|C4) = 0.60; P(D|C5) = 1.00; P(D|C0) = 0$$

Find the annual probability of wind damage to the building.

Solution

We might observe first that the five categories of hurricanes, C1, C2, C3, C4, C5, are ***mutually exclusive***, as it is reasonable to assume that no two categories can occur at the same time and cover all possible hurricanes; thus, these five categories ***plus non-hurricane winds (denoted C0) are also collectively exhaustive***. Therefore, Eq. 4.10-1 is applicable and the total probability of the damage would be:

$$P(D) = P(D|C1) \times P(C1) + P(D|C2) \times P(C2) + P(D|C3) \times P(C3) + P(D|C4) \times P(C4) + P(D|C5) \times P(C5)$$

$$P(D) = 0.05 \times 0.35 + 0.10 \times 0.25 + 0.25 \times 0.14 + 0.60 \times 0.05 + 1.00 \times 0.01 = 0.1175$$

Therefore, annually the probability of hurricane wind damage to the building is about 12%.

We might observe from the above calculations that the greatest contributions to the annual damage probability are from hurricanes of Categories 3 and 4, $P(D|C3) \times P(C3) = 0.25 \times 0.14 = 0.035\ and\ P(D|C4) \times P(C4) = 0.60 \times 0.05 = 0.03$

Observe also that even though damage to the building will certainly occur under a Category 5 hurricane $P(D|C5) = 1.00$ the occurrence of such hurricanes is very rare, annually $P(C5) = 0.01$, which means that it might occur (on the average) only about a once in every 100 years.

Example 4.10-3

Figure 4.10-2 shows the eastbound directions of two interstate highways I1 and I2 merging into another highway I3. Interstates I1 and I2 have the same traffic capacities; however, the rush-hour traffic volume on I2 is about twice that on I1 so that during rush hour, the probabilities of traffic congestion, denoted, respectively, as E_1 and E_2, are as follows:

$P(E_1) = 0.10; P(E_2) = 0.20$

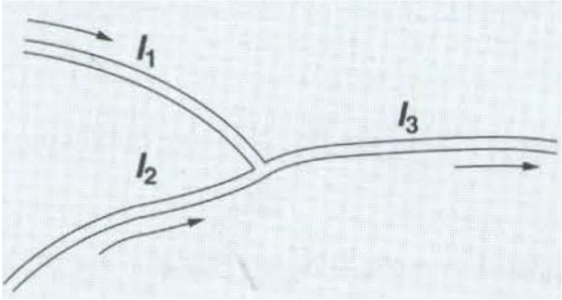


Figure 4.10-2: Eastbound interstate highways.

Also, when one route has excessive traffic, the chance of excessive traffic on the other route can be expected to increase; assume that these conditional probabilities are:

$P(E_1|E_2) = 0.40$ whereas, from Bayes' theorem (see Section 4.11), we must have:

$P(E_2|E_1) = 0.80$

Determine the probability of traffic congestion on the third route I3, $P(E_3)$ with the following two cases:

- First: Let us assume that the capacity of I3 is the same as that of I1 or I2 and that when I1 and I2 are both carrying less than their respective traffic capacities, there is a 20% probability that I3 will experience excessive traffic; i.e., $P(E3|\bar{E}_1\bar{E}_2) = 0.20$.
-

Solution

We would expect that the relevant probability will depend on the traffic conditions on I1 and I2, which may be E_1E_2, $\bar{E}_1E_2$, or $\bar{E}_1\bar{E}_2$; observe that these four joint events are ***mutually exclusive*** and ***collectively exhaustive***. Their respective probabilities are then:

The solution is not complete yet.

4.12 CONTENTS

CHAPTER 5

DISCRETE PROBABILITY DISTRIBUTIONS

5.1 INTRODUCTION

- This chapter explains the concepts and applications of what is called a ***probability distribution***.
- In addition, special probability distributions, such as the ***binomial***, and ***Poisson distributions***, are explained.

5.2 MODELING CONCEPT

- In mathematics, curves can be represented by equations. For example, the equation of the circle is:

  $x^2 + y^2 = r^2$

  where r is the radius.
- A circle can be used to represent many physical objects, such as a wheel or a gear. Even though it is not possible to manufacture a wheel that is perfectly round, the equation and the properties of a circle can be used to study many aspects of the wheel, such as area, velocity, and acceleration.

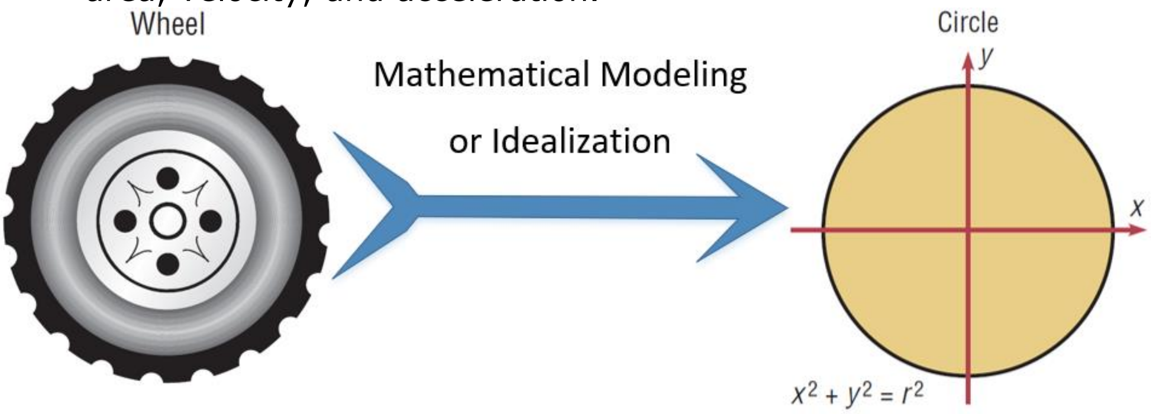


Figure 5.2-1: Using circle to model of a wheel.

- In a similar manner, many theoretical curves, called ***probability distribution curves***, can be used to simulate many variables that are not perfectly coincide with these distributions.

5.3 Random Variable and Probability Distributions*

5.3.1 Random Variable

A ***random variable*** is a ***variable whose values are determined by chance***.

Example 5.3-1

Should strength of concrete be modeled as variable or as random variable?

Solution

If strength of concrete can be related to other parameters such as constituent materials, sample shape (cylinder or cube), test age (7 days, 28 days, etc.), then it can be simulated as a ***variable*** (***mathematical variable***).

Experimentally, it has found that different concrete sample that have been produced from same constituent materials with same sample shape and tested at same age and under same conditions will have differences in their strength. In this sense, concrete strength should be simulated as a ***random variable*** as it depends on parameters that are out of controlled (***determined by chance***).

Example 5.3-2

Consider ***Figure 5.3-1*** above that represents temperature in °F measured at a Columbia on May 2010.

Should temperature at this location be simulated as a mathematical variable or as a random variable?

Solution

When temperature at a specific date, for example 10th of May, at 2010 is compared to temperature records on same time but at different years, one notes a significant different and concludes that the temperature should be simulated as a random variable.

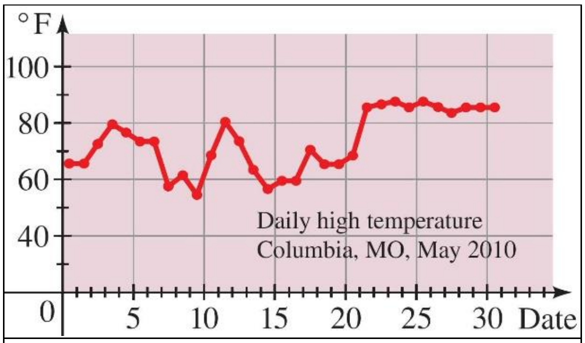
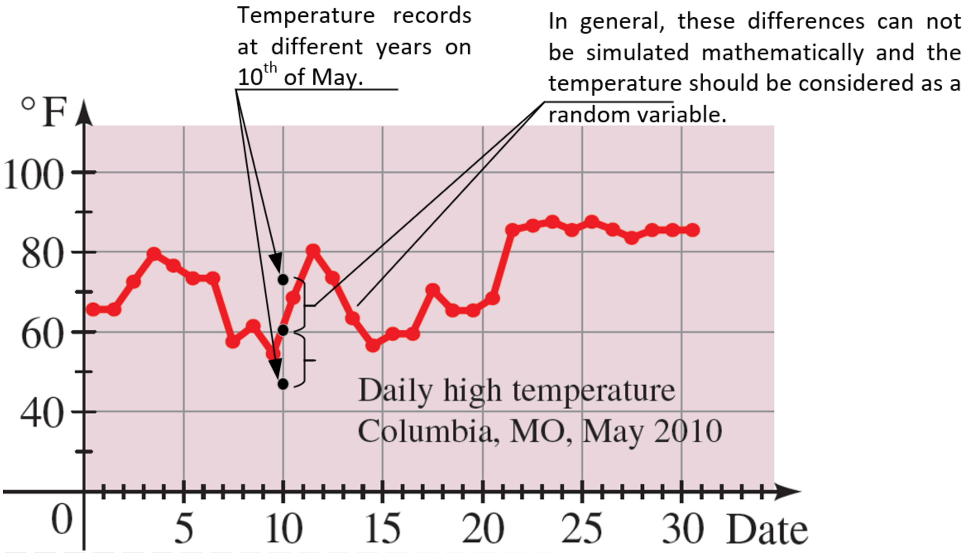


Figure 5.3-1: Temperature versus date, data for Example 5.3-2



Temperature records at different years on 10th of May.

In general, these differences can not be simulated mathematically and the temperature should be considered as a random variable.

5.3.2 Probability Mass Function

- For the case of discrete variable, the Probability Mass Function, abbreviated ***pmf***, is defined as:

  $p_X(x) = P_r(X = x)$

  where the capital letter X, refers to the random variable, while the small letter x refers to a specific value for the random variable.
- Based on probability axioms:

  $0 \leq p_X(x) \leq 1$, for all possible x,

  $p_X(x) = 0$ for all unrealizable x,

  $\Sigma p_X(x) = 1$ which is summed over all possible x,

$P_X(x_1 + x_2 + \cdots + x_n) = p_X(x_1) + p_X(x_2) + \cdots + p_X(x_n)$ for mutually exclusive outcomes $x_1, x_2, \ldots, x_n$.

Example 5.3-3

Construct a probability distribution for rolling a single die.

Solution

Since the sample space is 1, 2, 3, 4, 5, 6 and each outcome has a probability of $\frac{1}{6}$, ***according to classical or prior probability interpretation***, the distribution is as shown.

Outcome *X*	1	2	3	4	5	6
Probability *P*(*X*)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Probability distributions can be shown graphically by representing the values of X on the x-axis and the probabilities P(X) on the y-axis.

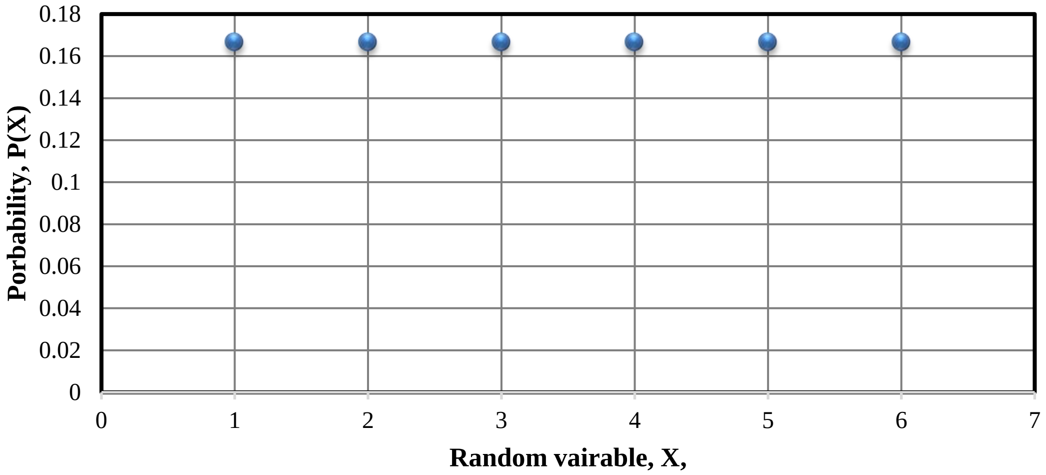


Figure 5.3-2: Probability distribution for Example 5.3-3.

Example 5.3-4

Represent graphically the probability distribution for the sample space for tossing three coins and getting a head.

Solution

Using ***classical*** or ***prior probability interpretation***, when three coins are tossed, the sample space is represented as

TTT, TTH, THT, HTT, HHT, HTH, THH, HHH;

and if X is the random variable for the number of heads, then X assumes the value 0, 1, 2, or 3. Probabilities for the values of X can be determined as follows:

No heads	**One head**			**Two heads**			**Three heads**
TTT	TTH	THT	HTT	HHT	HTH	THH	HHH
$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
$\frac{1}{8}$	$\frac{3}{8}$			$\frac{3}{8}$			$\frac{1}{8}$

Hence, the probability of getting no heads is, one head is $\frac{3}{8}$, two heads is $\frac{3}{8}$, and three heads is $\frac{1}{8}$. From these values, a probability distribution can be constructed by listing the outcomes and assigning the probability of each outcome, as shown here.

Number of heads *X*	0	1	2	3
Probability *P*(*X*)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The values that X assumes are located on the x axis, and the values for P(X) are located on the y axis. The graph is shown in Figure 5–1.

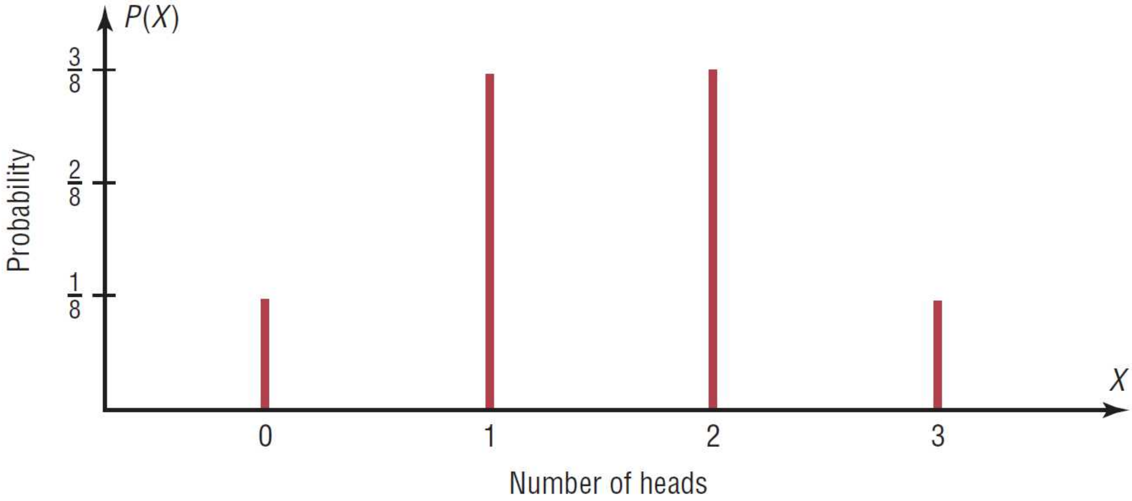


Figure 5.3-3: Probability distribution for Example 5.3-4.

Example 5.3-5

Plot a probability distribution for the data indicated in ***Table 5.3-1*** below that represents the number of floods recorded per year at a gauging station for a specific river.

Table 5.3-1

$p_X(0) = .00;$ $p_X(1) = 2/34 = .06;$ $p_X(2) = 6/34 = .18;$

$p_X(3) = 7/34 = .20;$ $p_X(4) = 9/34 = .26;$ $p_X(5) = 4/34 = .12;$

$p_X(6) = 1/34 = .03;$ $p_X(7) = 4/34 = .12;$ $p_X(8) = 1/34 = .03;$

$p_X(x) = .00 \text{ for } x > 8.$

Solution

As data have been gathered from a recording and measuring process, therefore presented probabilities are objective or empirical probabilities.

Consider number of floods per year, X, as a random variable, the probability distribution would be as indicated in Figure 5.3-4 below.

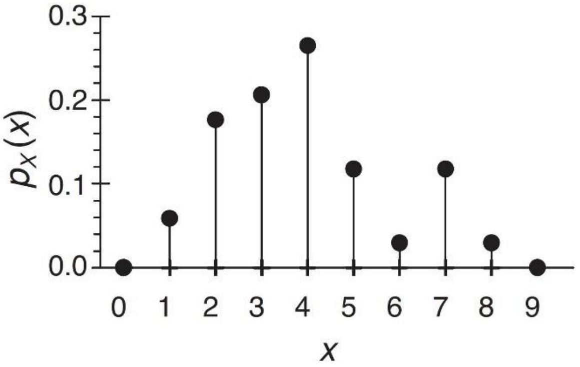


Figure 5.3-4: Probability distribution for number of floods per year recorded at a specific station.

5.4 Mean, Variance, Standard Deviation, and Expectation*

- The ***mean***, ***variance***, and ***standard deviation*** for a ***probability distribution*** are ***computed differently*** from the mean, variance, and standard deviation for samples.
- This section explains how these measures as well as a ***new measure*** called ***the expectation*** are calculated for probability distributions.

5.4.1 Mean

- In Chapter 3, the mean for a sample or population was computed by adding the values and dividing by the total number of values, as shown in these formulas:

$$\mu = \frac{\sum X}{N} \qquad \text{Eq. 5.4-1}$$

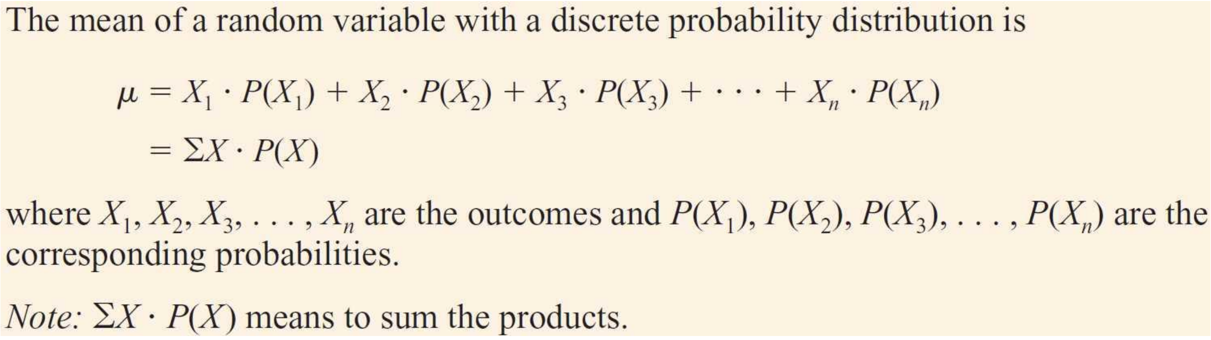
- As discussed in ***Chapter 4***, according to objective (empirical) interpretation of probability, exact parameter, for example the mean, μ, can be obtained from sample only when sample size is infinity larger, i.e. $n \rightarrow \infty$. Since this task is impossible, the previous formula cannot be used because the denominator would be infinity.
- Hence, a new method of computing the mean is necessary. To derive the new formula, the above equation has been re-written in the following form:

$$\mu = \frac{f_1X_1 + f_2X_2 + \cdots + f_NX_N}{N} \qquad \text{Eq. 5.4-2}$$

- Based on objective (empirical) interpretation of probability, each term of above relation can be re-written in the following form:

$$\lim_{N\to\infty} \frac{Frequency\ of\ X_i}{N} = P(X_i) \qquad \text{Eq. 5.4-3}$$

- Therefore, in its final form the relation for the mean, μ, would be as indicated in below:



The mean of a random variable with a discrete probability distribution is

$$\mu = X_1 \cdot P(X_1) + X_2 \cdot P(X_2) + X_3 \cdot P(X_3) + \cdots + X_n \cdot P(X_n)$$
$$= \Sigma X \cdot P(X)$$

where $X_1, X_2, X_3, \ldots, X_n$ are the outcomes and $P(X_1), P(X_2), P(X_3), \ldots, P(X_n)$ are the corresponding probabilities.

Note: $\Sigma X \cdot P(X)$ means to sum the products.

- ***This method gives the exact theoretical value of the mean as if it were possible to roll the die an infinite number of times***.

5.4.2 Variance and Standard Deviation

- Recall from ***Chapter 3*** that to measure this spread or variability, statisticians use the ***variance*** and ***standard deviation***. These formulas were used:

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \qquad \text{Eq. 5.4-4}$$

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} \qquad \text{Eq. 5.4-5}$$

- As for mean, to be suitable for cases where $N \rightarrow \infty$, above relations are rewritten as indicated in below in terms of objective interpretations of probability.

Find the variance of a probability distribution by multiplying the square of each outcome by its corresponding probability, summing those products, and subtracting the square of the mean. The formula for the variance of a probability distribution is

$$\sigma^2 = \Sigma[X^2 \cdot P(X)] - \mu^2$$

The standard deviation of a probability distribution is

$$\sigma = \sqrt{\sigma^2} \qquad \text{or} \qquad \sqrt{\Sigma[X^2 \cdot P(X)] - \mu^2}$$

Example 5.4-1

Find the mean, variance, and standard deviations of the number of spots that appear when a die is tossed.

Solution

In the toss of a die, the mean can be computed thus.

Outcome *X*	1	2	3	4	5	6
Probability *P*(*X*)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\mu = \Sigma X \cdot P(X) = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6}$$

$$= \tfrac{21}{6} = 3\tfrac{1}{2} \text{ or } 3.5$$

That is, when a die is tossed many times, the theoretical mean will be 3.5. ***Note that even though the die cannot show a 3.5, the theoretical average is 3.5***.

Square each outcome and multiply by the corresponding probability, sum those products, and then subtract the square of the mean.

$$\sigma^2 = (1^2 \cdot \tfrac{1}{6} + 2^2 \cdot \tfrac{1}{6} + 3^2 \cdot \tfrac{1}{6} + 4^2 \cdot \tfrac{1}{6} + 5^2 \cdot \tfrac{1}{6} + 6^2 \cdot \tfrac{1}{6}) - (3.5)^2 = 2.9$$

To get the standard deviation, find the square root of the variance.

$$\sigma = \sqrt{2.9} = 1.7$$

Example 5.4-2

The data indicated in below represents the observed number of days per month of high-amplitude waves acting on a sea pier is given below. Determine the mean and variance of X.

X	P(X)
0	0.38
1	0.22
2	0.18
3	0.13
4	0.09
5	0.06
6	0.03
≥7	0.01

Solution

Based on relation for the mean value, μ,

$\mu = \Sigma X.P(X)$

data table has been updated as indicated in below:

X	P(X)	X.P(X)
0	0.38	0
1	0.22	0.22
2	0.18	0.36
3	0.13	0.39
4	0.09	0.36
5	0.06	0.3
6	0.03	0.18
≥7	0.01	0.07
	μ = ∑X.P(X)	1.88

To determine variance based on the relation of

$\sigma^2 = \Sigma\big(X^2.P(X)\big) - \mu^2$

Data table has been updated to follows

X	P(X)	X.P(X)	X^2	X^2.P(X)
0	0.38	0	0	0
1	0.22	0.22	1	0.22
2	0.18	0.36	4	0.72
3	0.13	0.39	9	1.17
4	0.09	0.36	16	1.44
5	0.06	0.3	25	1.5
6	0.03	0.18	36	1.08
≥7	0.01	0.07	49	0.49
	μ = ∑X.P(X)	1.88		6.62

Then:

$\sigma^2 = \Sigma\big(X^2.P(X)\big) - \mu^2 = 6.62 - 1.88^2 = 3.09$

Finally, the standard deviation would be:

$\sigma = \sqrt{\sigma^2} = \sqrt{3.09} = 1.78$

Example 5.4-3

In a building project, the construction of the foundations takes time T1 and the construction of the superstructure takes time T2. Because of inclement weather, labor problems, and other factors, T1 and T2 behave like random variables with empirical probabilities indicated in below. Calculate the mean times taken for the foundations and the superstructure.

Time in Weeks, T	$P(T_1)$	$P(T_2)$
1	0.1	0
2	0.3	0
3	0.4	0
4	0.2	0.1
5	0	0.5
6	0	0.4
7	0	0

Solution

Based on relation for the mean value, μ,

$\mu = \Sigma T.P(T)$

data table has been updated as indicated in below:

Time in Weeks, T	$P(T_1)$	$P(T_2)$	$T.P(T_1)$	$T.P(T_2)$
1	0.1	0	0.1	0
2	0.3	0	0.6	0
3	0.4	0	1.2	0
4	0.2	0.1	0.8	0.4
5	0	0.5	0	2.5
6	0	0.4	0	2.4
7	0	0	0	0
		Summation	2.7	5.3

5.5 The Binomial Distribution

5.5.1 Experiments with Two Possible Outcomes

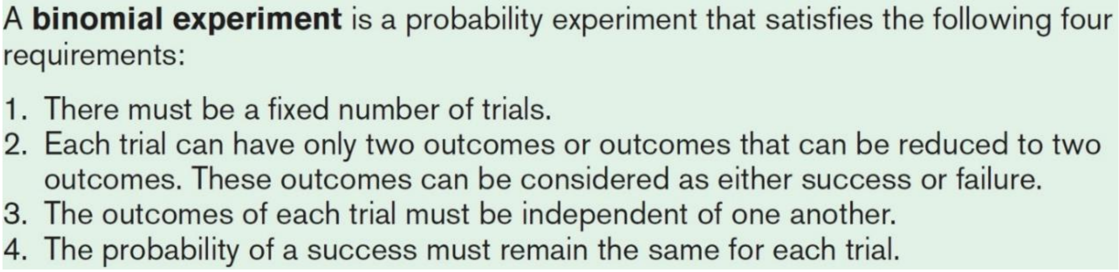
- Many types of probability problems have only two outcomes or can be reduced to two outcomes.
- For example,
  - When a coin is tossed, it can land heads or tails.
  - In a basketball game, a team either wins or loses.
  - A true/false item can be answered in only two ways, true or false.

5.5.2 Situations that can be Reduced to Two Outcomes

- Other ***situations can be reduced to two outcomes***. For example,
  - A medical treatment can be classified as effective or ineffective, depending on the results.
  - A person can be classified as having normal or abnormal blood pressure, depending on the measure of the blood pressure gauge.
  - A multiple-choice question, even though there are four or five answer choices, can be classified as correct or incorrect.

5.5.3 Binominal Experiment

Situations like these are called binomial experiments.

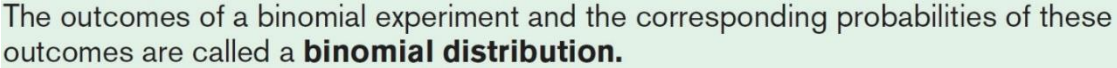


A **binomial experiment** is a probability experiment that satisfies the following four requirements:

1. There must be a fixed number of trials.
2. Each trial can have only two outcomes or outcomes that can be reduced to two outcomes. These outcomes can be considered as either success or failure.
3. The outcomes of each trial must be independent of one another.
4. The probability of a success must remain the same for each trial.

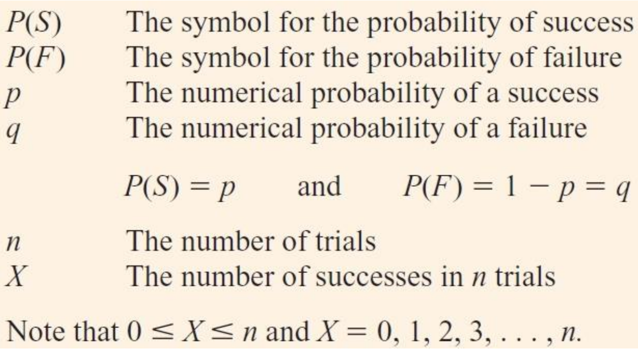
5.5.4 Binominal Distribution and Its Formula

- A ***binomial experiment*** and its results give rise to a special probability distribution called the ***binomial distribution***.



The outcomes of a binomial experiment and the corresponding probabilities of these outcomes are called a **binomial distribution.**

- In binomial experiments, ***the outcomes are usually classified as successes or failures***. For example, the ***correct answer*** to a multiple-choice item can be ***classified as a success***, but any of the ***other choices would be incorrect*** and hence ***classified as a failure***.
- The ***notation*** that is commonly used for binomial experiments and the binomial distribution is defined now.



$P(S)$	The symbol for the probability of success
$P(F)$	The symbol for the probability of failure
p	The numerical probability of a success
q	The numerical probability of a failure

$$P(S) = p \quad \text{and} \quad P(F) = 1 - p = q$$

n	The number of trials
X	The number of successes in n trials

Note that $0 \leq X \leq n$ and $X = 0, 1, 2, 3, \ldots, n$.

- In a binominal experiment, the probability of exactly X successes in n trials is:

$$P(X) = \frac{n!}{(n-X)!\,X!} \times p^{X} \times q^{n-X} \qquad \text{Eq. 5.5-1}$$

Example 5.5-1

A coin is tossed 3 times. Based on the sample space, find the probability of getting exactly two heads.

Solution

This problem can be solved by looking at the sample space. There are three ways to get two heads.

HHH, HHT, HTH, THH, TTH, THT, HTT, TTT

The answer is $\frac{3}{8}$, or 0.375.

Example 5.5-2

Resolve

Example 5.5-1 above based on binominal probability formula if applicable.

Solution

Looking at the

Example 5.5-1 from the standpoint of a binomial experiment, one can show that it meets the four requirements.

1. There are a fixed number of trials (three).
2. There are only two outcomes for each trial, heads or tails.
3. The outcomes are independent of one another (the outcome of one toss in no way affects the outcome of another toss).
4. The probability of a success (heads) is ½ in each case.

Therefore, the binominal probability formula is applicable. In this case, n = 3, X = 2, p = 1/2 , and q = 1/2 . Hence, substituting in the formula gives:

$$P(X) = \frac{n!}{(n-X)!\,X!} \times p^{X} \times q^{n-X} = \frac{3!}{(3-2)! \times 2!} \times \frac{1^{2}}{2} \times \frac{1^{(3-2)}}{2} = 0.375$$

which is the same answer obtained by using the sample space.

Explaining Binominal Formula with Referring to

5.5.5 Example 5.5-1 and Example 5.5-2

- ***Example 5.5-1*** and ***Example 5.5-2*** above can be used to explain the formula.
- First, note that there are ***three ways*** to get ***exactly two heads and one tail*** from a possible ***eight ways***. They are HHT, HTH, and THH.
- In this case, then, the number of ways of obtaining two heads from three-coin tosses is $_{3}C_{2}$, or 3, as shown in ***Chapter 4***.
- In general, the number of ways to get X successes from n trials without regard to order is:

  $$_{n}C_{X} = \frac{n!}{(n-X)!X!}$$

  This is the first part of the binomial formula.
- Next, each success has a probability of p and can occur twice.
- Likewise, each failure has a probability of q and can occur once, giving the second part of the formula.
- To generalize, then, each success has a probability of p and can occur X times, and each failure has a probability of q and can occur $n - X$ times. Putting it all together yields the binomial probability formula.

Example 5.5-3

A survey found that one out of five Americans say he or she has visited a doctor in any given month. If 10 people are selected at random, find the probability that exactly 3 will have visited a doctor last month.

Solution

$$P(X) = \frac{n!}{(n-X)!\,X!} \times p^X \times q^{n-X} \Rightarrow P(3) = \frac{10!}{(10-3)! \times 3!} \times \left(\frac{1}{5}\right)^3 \times \left(\frac{4}{5}\right)^{(10-3)} = 0.201$$

Example 5.5-4

A survey from Teenage Research Unlimited (Northbrook, Illinois) found that 30% of teenage consumers receive their spending money from part-time jobs. If 5 teenagers are selected at random, find the probability that at least 3 of them will have part-time jobs.

Solution

To find the probability that at least 3 have part-time jobs, it is necessary to find the individual probabilities for 3, or 4, or 5 and then add them to get the total probability.

$$\because P(X) = \frac{n!}{(n-X)!\,X!} \times p^X \times q^{n-X}$$

$$P(3) = \frac{5!}{(5-3)! \times 3!} \times \left(\frac{30}{100}\right)^3 \times \left(\frac{70}{100}\right)^{(5-3)} = 0.132$$

$$P(4) = \frac{5!}{(5-4)! \times 4!} \times \left(\frac{30}{100}\right)^4 \times \left(\frac{70}{100}\right)^{(5-4)} = 0.028$$

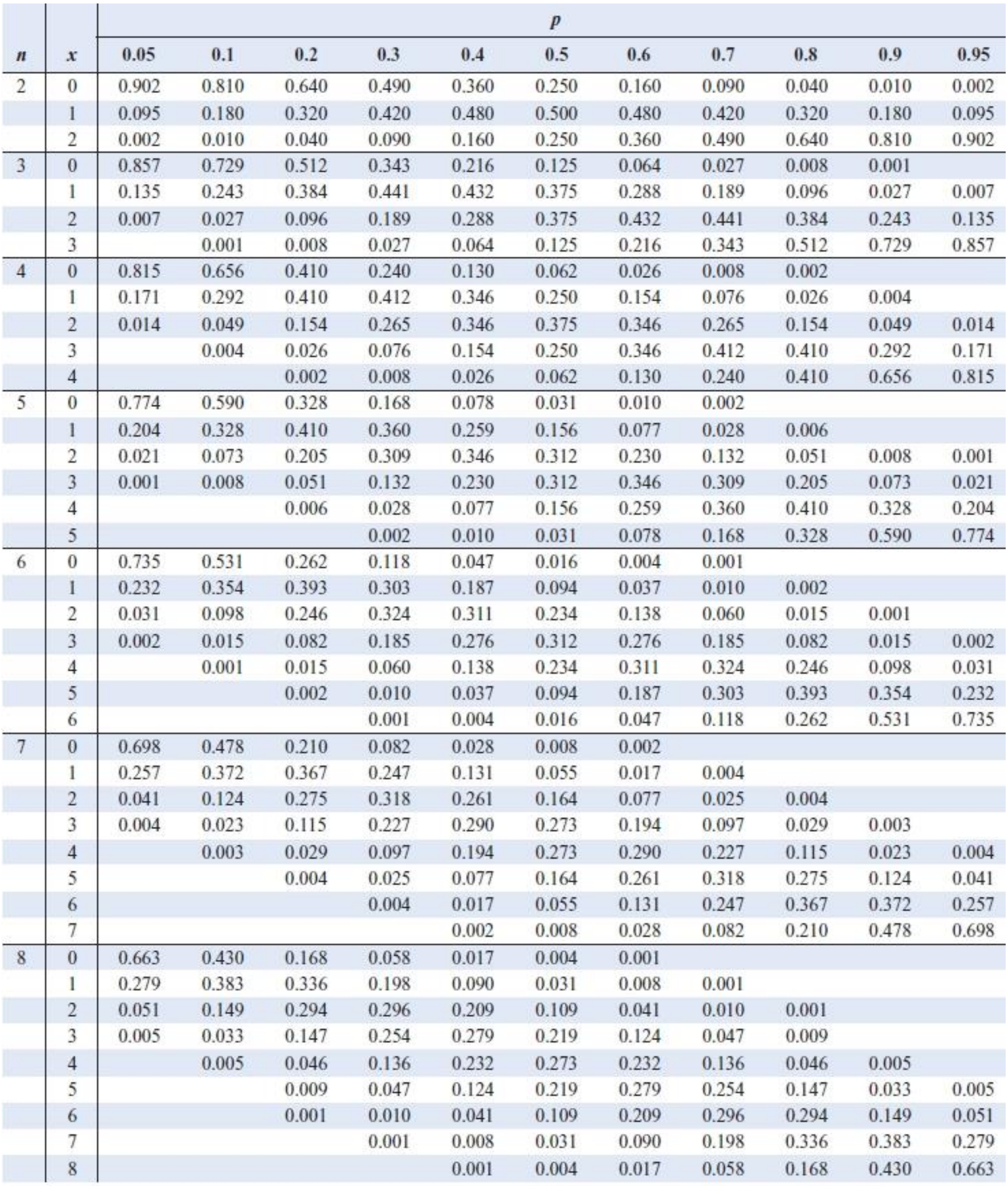
$$P(5) = \frac{5!}{(5-5)! \times 5!} \times \left(\frac{30}{100}\right)^5 \times \left(\frac{70}{100}\right)^{(5-5)} = 0.002$$

$$P(at\ least\ three\ teenagers\ have\ part-time\ jobs) = 0.132 + 0.028 + 0.002 = 0.162$$

5.5.6 TABLE TO COMPUTE PROBABILITIES USING THE BINOMIAL PROBABILITY FORMULA

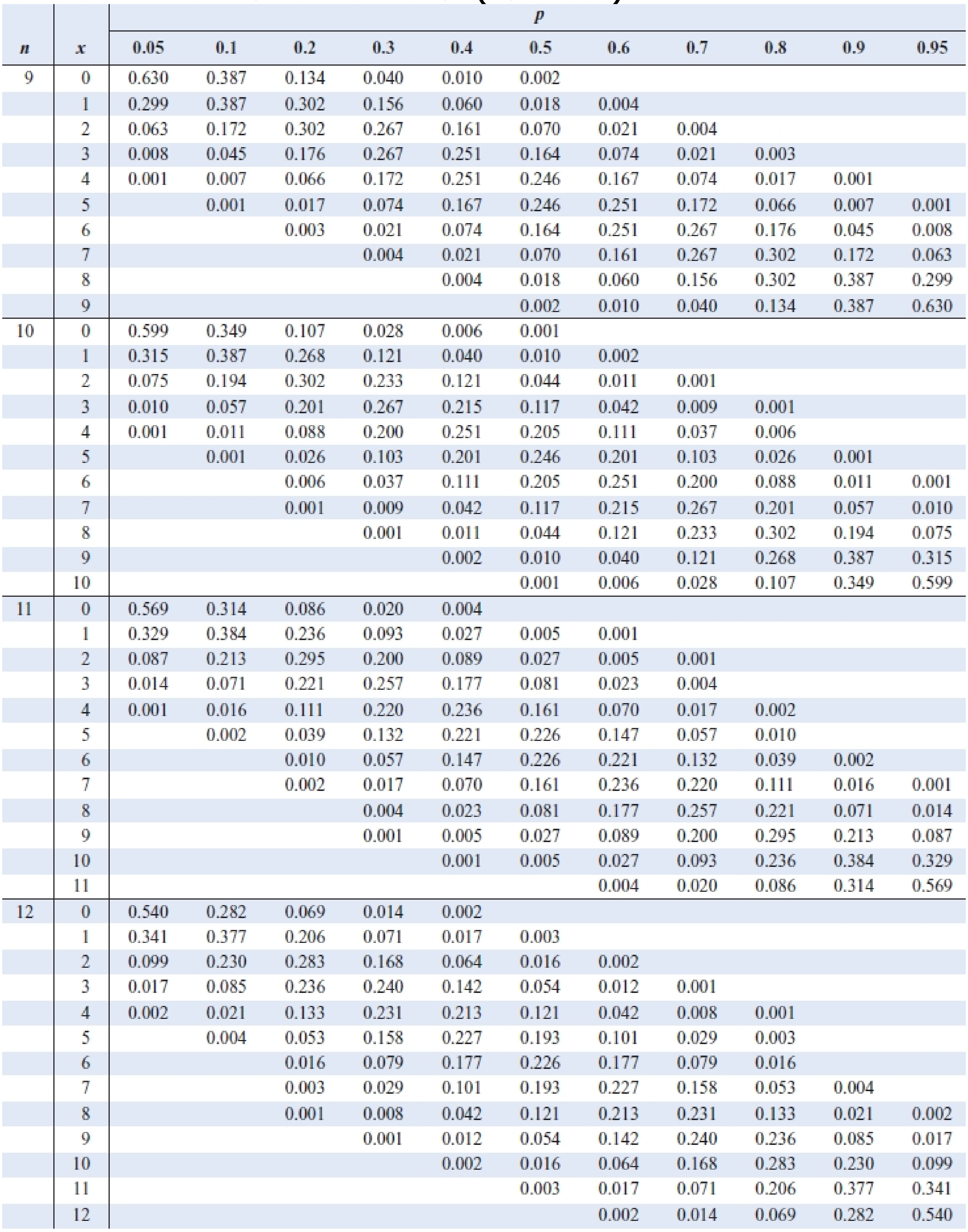
- Computing probabilities by using the binomial probability formula can be ***quite tedious at times***, so tables have been developed for selected values of n and p.
- Table 5.5-1 below gives the probabilities for individual events.

Table 5.5-1: The Binomial Distribution



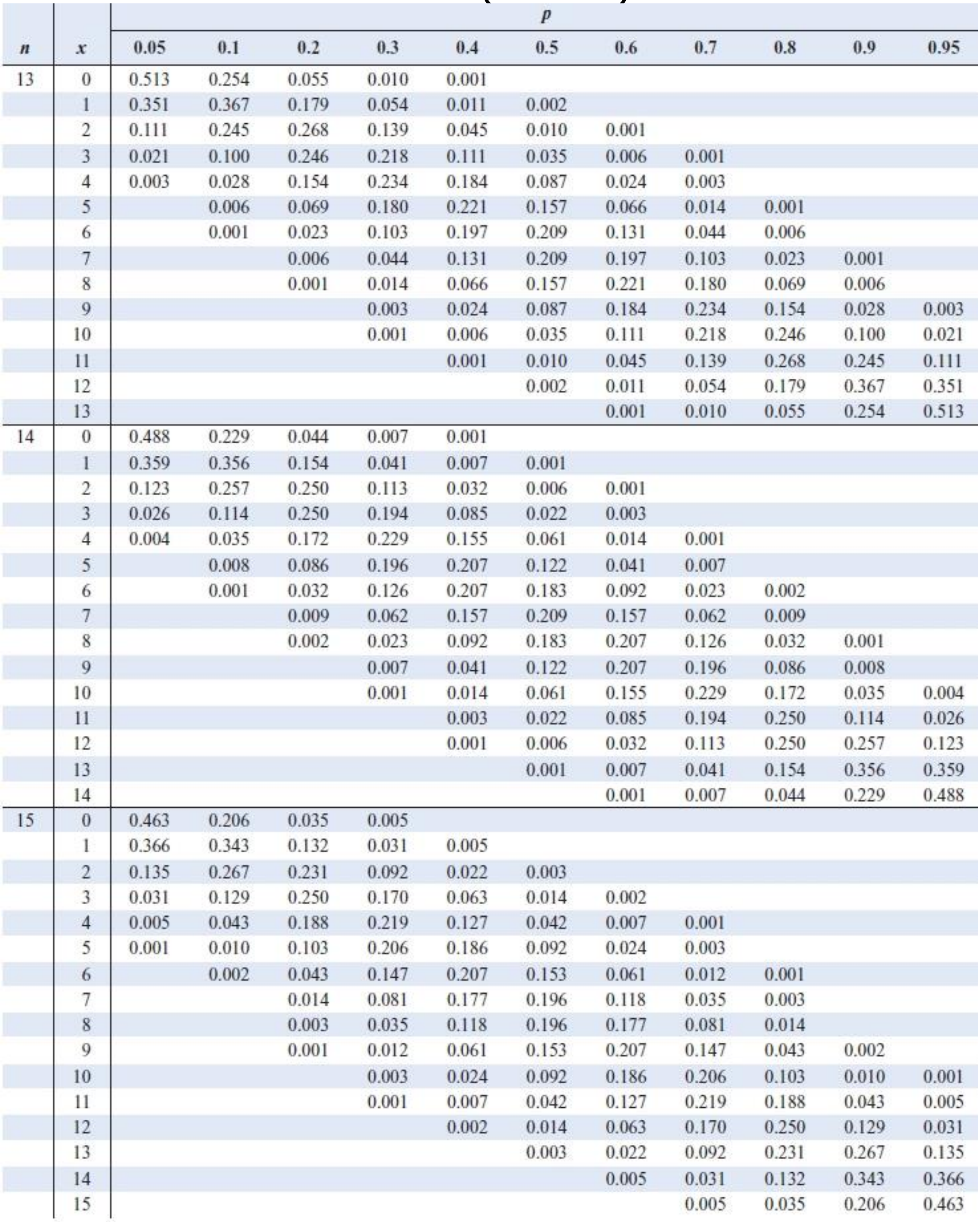
		p										
n	*x*	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
2	0	0.902	0.810	0.640	0.490	0.360	0.250	0.160	0.090	0.040	0.010	0.002
	1	0.095	0.180	0.320	0.420	0.480	0.500	0.480	0.420	0.320	0.180	0.095
	2	0.002	0.010	0.040	0.090	0.160	0.250	0.360	0.490	0.640	0.810	0.902
3	0	0.857	0.729	0.512	0.343	0.216	0.125	0.064	0.027	0.008	0.001	
	1	0.135	0.243	0.384	0.441	0.432	0.375	0.288	0.189	0.096	0.027	0.007
	2	0.007	0.027	0.096	0.189	0.288	0.375	0.432	0.441	0.384	0.243	0.135
	3		0.001	0.008	0.027	0.064	0.125	0.216	0.343	0.512	0.729	0.857
4	0	0.815	0.656	0.410	0.240	0.130	0.062	0.026	0.008	0.002		
	1	0.171	0.292	0.410	0.412	0.346	0.250	0.154	0.076	0.026	0.004	
	2	0.014	0.049	0.154	0.265	0.346	0.375	0.346	0.265	0.154	0.049	0.014
	3		0.004	0.026	0.076	0.154	0.250	0.346	0.412	0.410	0.292	0.171
	4			0.002	0.008	0.026	0.062	0.130	0.240	0.410	0.656	0.815
5	0	0.774	0.590	0.328	0.168	0.078	0.031	0.010	0.002			
	1	0.204	0.328	0.410	0.360	0.259	0.156	0.077	0.028	0.006		
	2	0.021	0.073	0.205	0.309	0.346	0.312	0.230	0.132	0.051	0.008	0.001
	3	0.001	0.008	0.051	0.132	0.230	0.312	0.346	0.309	0.205	0.073	0.021
	4			0.006	0.028	0.077	0.156	0.259	0.360	0.410	0.328	0.204
	5				0.002	0.010	0.031	0.078	0.168	0.328	0.590	0.774
6	0	0.735	0.531	0.262	0.118	0.047	0.016	0.004	0.001			
	1	0.232	0.354	0.393	0.303	0.187	0.094	0.037	0.010	0.002		
	2	0.031	0.098	0.246	0.324	0.311	0.234	0.138	0.060	0.015	0.001	
	3	0.002	0.015	0.082	0.185	0.276	0.312	0.276	0.185	0.082	0.015	0.002
	4		0.001	0.015	0.060	0.138	0.234	0.311	0.324	0.246	0.098	0.031
	5			0.002	0.010	0.037	0.094	0.187	0.303	0.393	0.354	0.232
	6				0.001	0.004	0.016	0.047	0.118	0.262	0.531	0.735
7	0	0.698	0.478	0.210	0.082	0.028	0.008	0.002				
	1	0.257	0.372	0.367	0.247	0.131	0.055	0.017	0.004			
	2	0.041	0.124	0.275	0.318	0.261	0.164	0.077	0.025	0.004		
	3	0.004	0.023	0.115	0.227	0.290	0.273	0.194	0.097	0.029	0.003	
	4		0.003	0.029	0.097	0.194	0.273	0.290	0.227	0.115	0.023	0.004
	5			0.004	0.025	0.077	0.164	0.261	0.318	0.275	0.124	0.041
	6				0.004	0.017	0.055	0.131	0.247	0.367	0.372	0.257
	7					0.002	0.008	0.028	0.082	0.210	0.478	0.698
8	0	0.663	0.430	0.168	0.058	0.017	0.004	0.001				
	1	0.279	0.383	0.336	0.198	0.090	0.031	0.008	0.001			
	2	0.051	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001		
	3	0.005	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009		
	4		0.005	0.046	0.136	0.232	0.273	0.232	0.136	0.046	0.005	
	5			0.009	0.047	0.124	0.219	0.279	0.254	0.147	0.033	0.005
	6			0.001	0.010	0.041	0.109	0.209	0.296	0.294	0.149	0.051
	7				0.001	0.008	0.031	0.090	0.198	0.336	0.383	0.279
	8					0.001	0.004	0.017	0.058	0.168	0.430	0.663

Table 5.5-1: The Binomial Distribution (Continued)



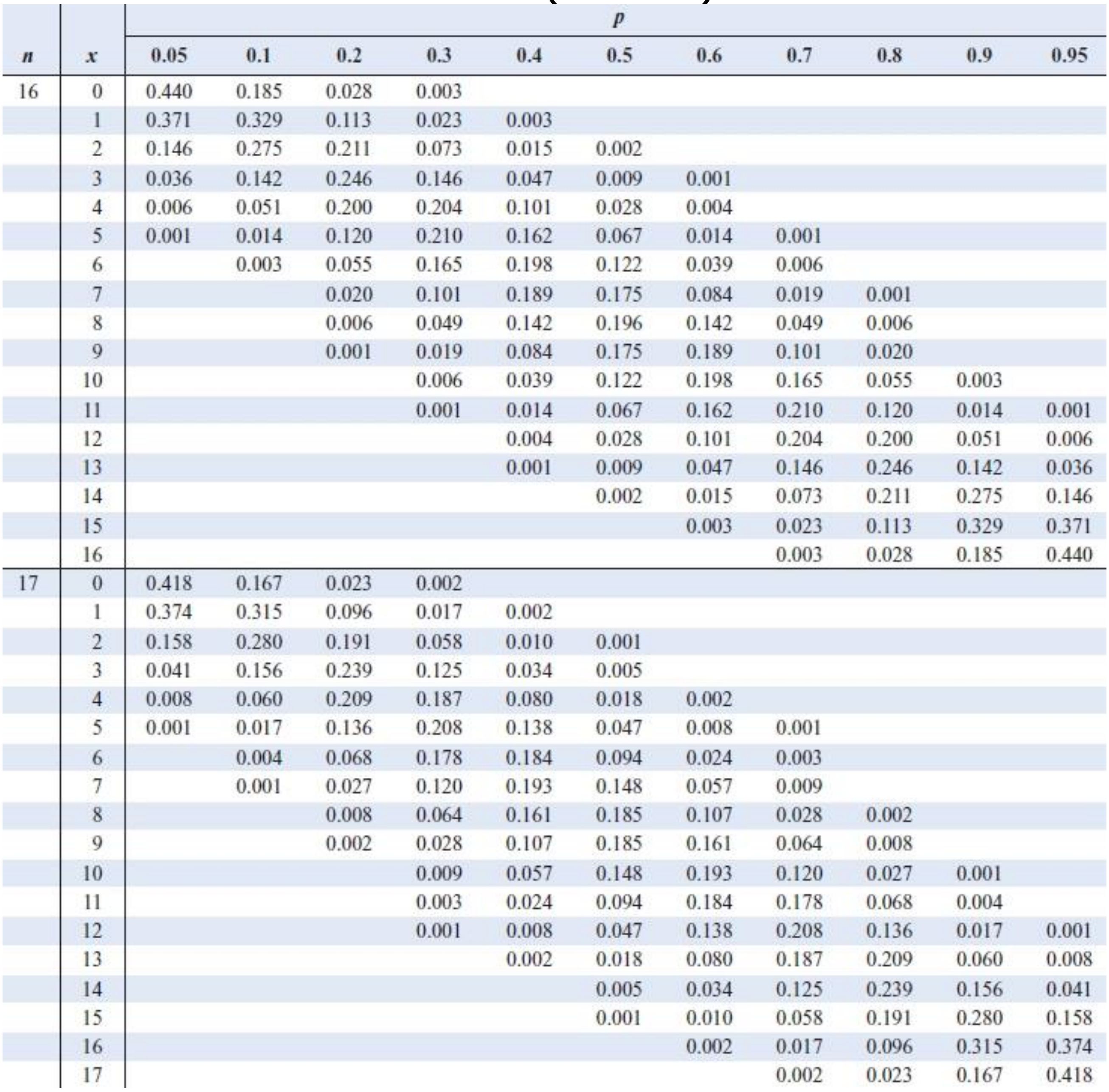
		p										
n	x	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
9	0	0.630	0.387	0.134	0.040	0.010	0.002					
	1	0.299	0.387	0.302	0.156	0.060	0.018	0.004				
	2	0.063	0.172	0.302	0.267	0.161	0.070	0.021	0.004			
	3	0.008	0.045	0.176	0.267	0.251	0.164	0.074	0.021	0.003		
	4	0.001	0.007	0.066	0.172	0.251	0.246	0.167	0.074	0.017	0.001	
	5		0.001	0.017	0.074	0.167	0.246	0.251	0.172	0.066	0.007	0.001
	6			0.003	0.021	0.074	0.164	0.251	0.267	0.176	0.045	0.008
	7				0.004	0.021	0.070	0.161	0.267	0.302	0.172	0.063
	8					0.004	0.018	0.060	0.156	0.302	0.387	0.299
	9						0.002	0.010	0.040	0.134	0.387	0.630
10	0	0.599	0.349	0.107	0.028	0.006	0.001					
	1	0.315	0.387	0.268	0.121	0.040	0.010	0.002				
	2	0.075	0.194	0.302	0.233	0.121	0.044	0.011	0.001			
	3	0.010	0.057	0.201	0.267	0.215	0.117	0.042	0.009	0.001		
	4	0.001	0.011	0.088	0.200	0.251	0.205	0.111	0.037	0.006		
	5		0.001	0.026	0.103	0.201	0.246	0.201	0.103	0.026	0.001	
	6			0.006	0.037	0.111	0.205	0.251	0.200	0.088	0.011	0.001
	7			0.001	0.009	0.042	0.117	0.215	0.267	0.201	0.057	0.010
	8				0.001	0.011	0.044	0.121	0.233	0.302	0.194	0.075
	9					0.002	0.010	0.040	0.121	0.268	0.387	0.315
	10						0.001	0.006	0.028	0.107	0.349	0.599
11	0	0.569	0.314	0.086	0.020	0.004						
	1	0.329	0.384	0.236	0.093	0.027	0.005	0.001				
	2	0.087	0.213	0.295	0.200	0.089	0.027	0.005	0.001			
	3	0.014	0.071	0.221	0.257	0.177	0.081	0.023	0.004			
	4	0.001	0.016	0.111	0.220	0.236	0.161	0.070	0.017	0.002		
	5		0.002	0.039	0.132	0.221	0.226	0.147	0.057	0.010		
	6			0.010	0.057	0.147	0.226	0.221	0.132	0.039	0.002	
	7			0.002	0.017	0.070	0.161	0.236	0.220	0.111	0.016	0.001
	8				0.004	0.023	0.081	0.177	0.257	0.221	0.071	0.014
	9				0.001	0.005	0.027	0.089	0.200	0.295	0.213	0.087
	10					0.001	0.005	0.027	0.093	0.236	0.384	0.329
	11							0.004	0.020	0.086	0.314	0.569
12	0	0.540	0.282	0.069	0.014	0.002						
	1	0.341	0.377	0.206	0.071	0.017	0.003					
	2	0.099	0.230	0.283	0.168	0.064	0.016	0.002				
	3	0.017	0.085	0.236	0.240	0.142	0.054	0.012	0.001			
	4	0.002	0.021	0.133	0.231	0.213	0.121	0.042	0.008	0.001		
	5		0.004	0.053	0.158	0.227	0.193	0.101	0.029	0.003		
	6			0.016	0.079	0.177	0.226	0.177	0.079	0.016		
	7			0.003	0.029	0.101	0.193	0.227	0.158	0.053	0.004	
	8			0.001	0.008	0.042	0.121	0.213	0.231	0.133	0.021	0.002
	9				0.001	0.012	0.054	0.142	0.240	0.236	0.085	0.017
	10					0.002	0.016	0.064	0.168	0.283	0.230	0.099
	11						0.003	0.017	0.071	0.206	0.377	0.341
	12							0.002	0.014	0.069	0.282	0.540

Table 5.5-1: The Binomial Distribution (Continued)



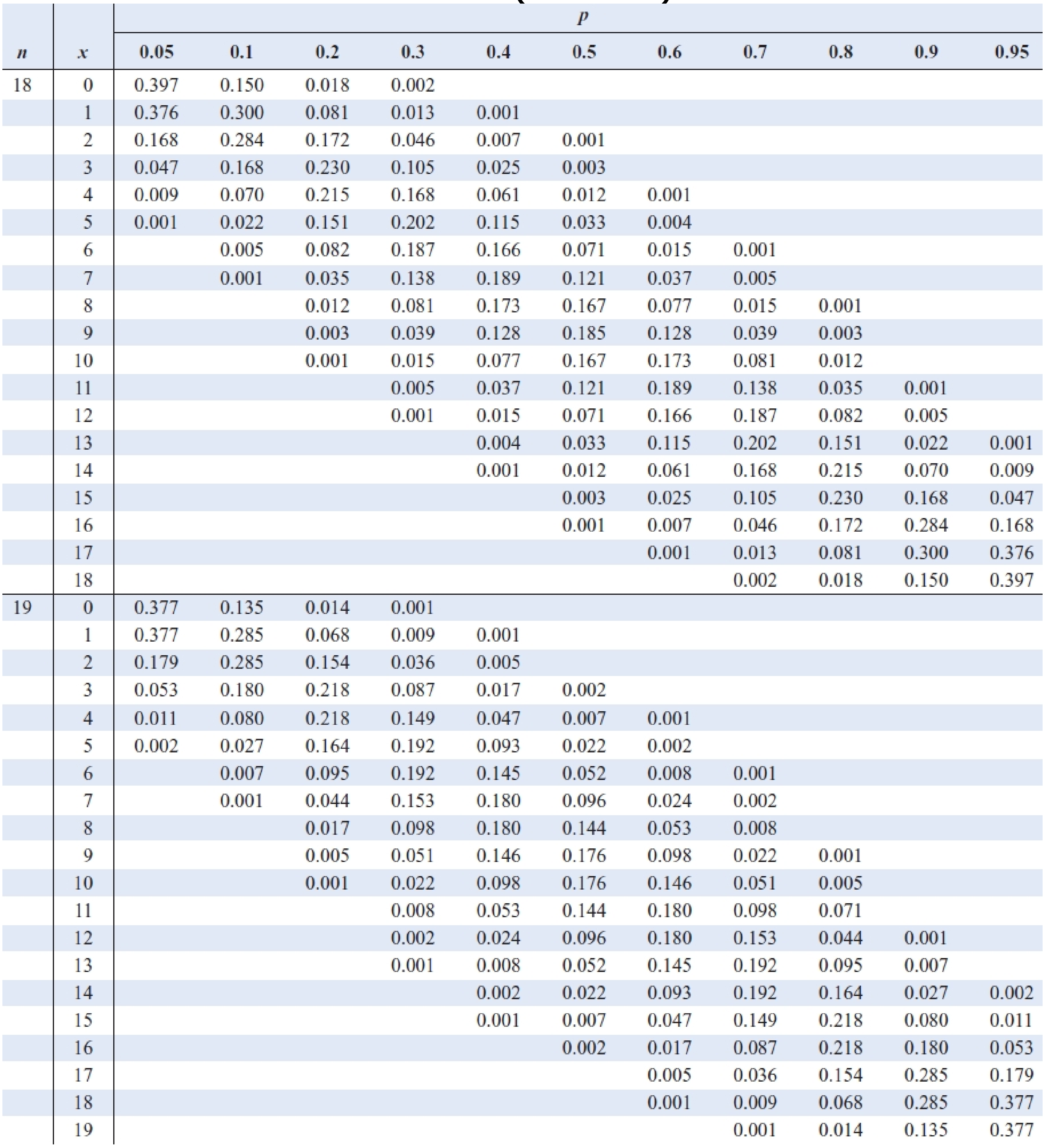
		p										
n	*x*	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
13	0	0.513	0.254	0.055	0.010	0.001						
	1	0.351	0.367	0.179	0.054	0.011	0.002					
	2	0.111	0.245	0.268	0.139	0.045	0.010	0.001				
	3	0.021	0.100	0.246	0.218	0.111	0.035	0.006	0.001			
	4	0.003	0.028	0.154	0.234	0.184	0.087	0.024	0.003			
	5		0.006	0.069	0.180	0.221	0.157	0.066	0.014	0.001		
	6		0.001	0.023	0.103	0.197	0.209	0.131	0.044	0.006		
	7			0.006	0.044	0.131	0.209	0.197	0.103	0.023	0.001	
	8			0.001	0.014	0.066	0.157	0.221	0.180	0.069	0.006	
	9				0.003	0.024	0.087	0.184	0.234	0.154	0.028	0.003
	10				0.001	0.006	0.035	0.111	0.218	0.246	0.100	0.021
	11					0.001	0.010	0.045	0.139	0.268	0.245	0.111
	12						0.002	0.011	0.054	0.179	0.367	0.351
	13							0.001	0.010	0.055	0.254	0.513
14	0	0.488	0.229	0.044	0.007	0.001						
	1	0.359	0.356	0.154	0.041	0.007	0.001					
	2	0.123	0.257	0.250	0.113	0.032	0.006	0.001				
	3	0.026	0.114	0.250	0.194	0.085	0.022	0.003				
	4	0.004	0.035	0.172	0.229	0.155	0.061	0.014	0.001			
	5		0.008	0.086	0.196	0.207	0.122	0.041	0.007			
	6		0.001	0.032	0.126	0.207	0.183	0.092	0.023	0.002		
	7			0.009	0.062	0.157	0.209	0.157	0.062	0.009		
	8			0.002	0.023	0.092	0.183	0.207	0.126	0.032	0.001	
	9				0.007	0.041	0.122	0.207	0.196	0.086	0.008	
	10				0.001	0.014	0.061	0.155	0.229	0.172	0.035	0.004
	11					0.003	0.022	0.085	0.194	0.250	0.114	0.026
	12					0.001	0.006	0.032	0.113	0.250	0.257	0.123
	13						0.001	0.007	0.041	0.154	0.356	0.359
	14							0.001	0.007	0.044	0.229	0.488
15	0	0.463	0.206	0.035	0.005							
	1	0.366	0.343	0.132	0.031	0.005						
	2	0.135	0.267	0.231	0.092	0.022	0.003					
	3	0.031	0.129	0.250	0.170	0.063	0.014	0.002				
	4	0.005	0.043	0.188	0.219	0.127	0.042	0.007	0.001			
	5	0.001	0.010	0.103	0.206	0.186	0.092	0.024	0.003			
	6		0.002	0.043	0.147	0.207	0.153	0.061	0.012	0.001		
	7			0.014	0.081	0.177	0.196	0.118	0.035	0.003		
	8			0.003	0.035	0.118	0.196	0.177	0.081	0.014		
	9			0.001	0.012	0.061	0.153	0.207	0.147	0.043	0.002	
	10				0.003	0.024	0.092	0.186	0.206	0.103	0.010	0.001
	11				0.001	0.007	0.042	0.127	0.219	0.188	0.043	0.005
	12					0.002	0.014	0.063	0.170	0.250	0.129	0.031
	13						0.003	0.022	0.092	0.231	0.267	0.135
	14							0.005	0.031	0.132	0.343	0.366
	15								0.005	0.035	0.206	0.463

Table 5.5-1: The Binomial Distribution (Continued)



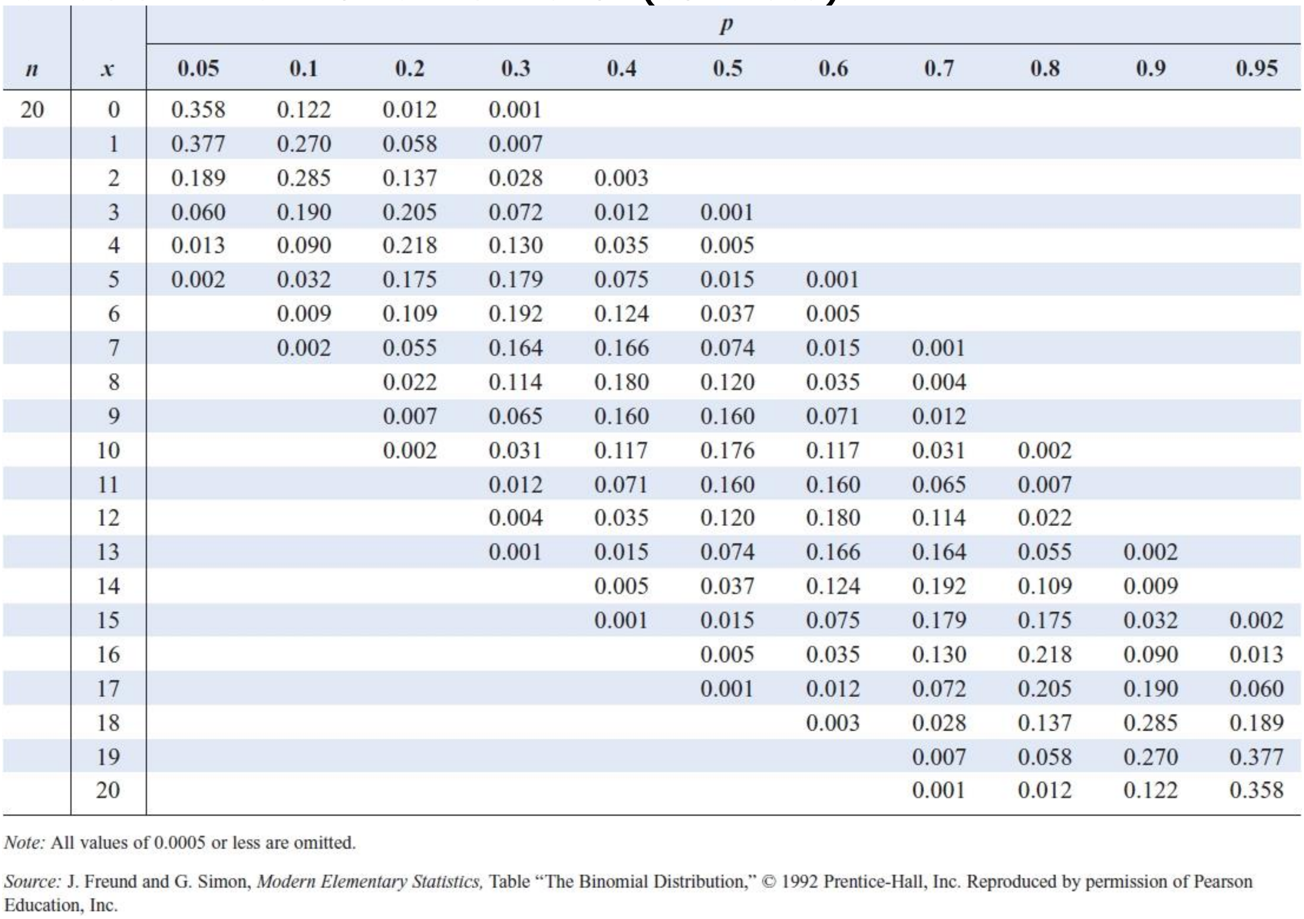
		p										
n	*x*	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
16	0	0.440	0.185	0.028	0.003							
	1	0.371	0.329	0.113	0.023	0.003						
	2	0.146	0.275	0.211	0.073	0.015	0.002					
	3	0.036	0.142	0.246	0.146	0.047	0.009	0.001				
	4	0.006	0.051	0.200	0.204	0.101	0.028	0.004				
	5	0.001	0.014	0.120	0.210	0.162	0.067	0.014	0.001			
	6		0.003	0.055	0.165	0.198	0.122	0.039	0.006			
	7			0.020	0.101	0.189	0.175	0.084	0.019	0.001		
	8			0.006	0.049	0.142	0.196	0.142	0.049	0.006		
	9			0.001	0.019	0.084	0.175	0.189	0.101	0.020		
	10				0.006	0.039	0.122	0.198	0.165	0.055	0.003	
	11				0.001	0.014	0.067	0.162	0.210	0.120	0.014	0.001
	12					0.004	0.028	0.101	0.204	0.200	0.051	0.006
	13					0.001	0.009	0.047	0.146	0.246	0.142	0.036
	14						0.002	0.015	0.073	0.211	0.275	0.146
	15							0.003	0.023	0.113	0.329	0.371
	16								0.003	0.028	0.185	0.440
17	0	0.418	0.167	0.023	0.002							
	1	0.374	0.315	0.096	0.017	0.002						
	2	0.158	0.280	0.191	0.058	0.010	0.001					
	3	0.041	0.156	0.239	0.125	0.034	0.005					
	4	0.008	0.060	0.209	0.187	0.080	0.018	0.002				
	5	0.001	0.017	0.136	0.208	0.138	0.047	0.008	0.001			
	6		0.004	0.068	0.178	0.184	0.094	0.024	0.003			
	7		0.001	0.027	0.120	0.193	0.148	0.057	0.009			
	8			0.008	0.064	0.161	0.185	0.107	0.028	0.002		
	9			0.002	0.028	0.107	0.185	0.161	0.064	0.008		
	10				0.009	0.057	0.148	0.193	0.120	0.027	0.001	
	11				0.003	0.024	0.094	0.184	0.178	0.068	0.004	
	12				0.001	0.008	0.047	0.138	0.208	0.136	0.017	0.001
	13					0.002	0.018	0.080	0.187	0.209	0.060	0.008
	14						0.005	0.034	0.125	0.239	0.156	0.041
	15						0.001	0.010	0.058	0.191	0.280	0.158
	16							0.002	0.017	0.096	0.315	0.374
	17								0.002	0.023	0.167	0.418

Table 5.5-1: The Binomial Distribution (Continued)



		p										
n	*x*	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
18	0	0.397	0.150	0.018	0.002							
	1	0.376	0.300	0.081	0.013	0.001						
	2	0.168	0.284	0.172	0.046	0.007	0.001					
	3	0.047	0.168	0.230	0.105	0.025	0.003					
	4	0.009	0.070	0.215	0.168	0.061	0.012	0.001				
	5	0.001	0.022	0.151	0.202	0.115	0.033	0.004				
	6		0.005	0.082	0.187	0.166	0.071	0.015	0.001			
	7		0.001	0.035	0.138	0.189	0.121	0.037	0.005			
	8			0.012	0.081	0.173	0.167	0.077	0.015	0.001		
	9			0.003	0.039	0.128	0.185	0.128	0.039	0.003		
	10			0.001	0.015	0.077	0.167	0.173	0.081	0.012		
	11				0.005	0.037	0.121	0.189	0.138	0.035	0.001	
	12				0.001	0.015	0.071	0.166	0.187	0.082	0.005	
	13					0.004	0.033	0.115	0.202	0.151	0.022	0.001
	14					0.001	0.012	0.061	0.168	0.215	0.070	0.009
	15						0.003	0.025	0.105	0.230	0.168	0.047
	16						0.001	0.007	0.046	0.172	0.284	0.168
	17							0.001	0.013	0.081	0.300	0.376
	18								0.002	0.018	0.150	0.397
19	0	0.377	0.135	0.014	0.001							
	1	0.377	0.285	0.068	0.009	0.001						
	2	0.179	0.285	0.154	0.036	0.005						
	3	0.053	0.180	0.218	0.087	0.017	0.002					
	4	0.011	0.080	0.218	0.149	0.047	0.007	0.001				
	5	0.002	0.027	0.164	0.192	0.093	0.022	0.002				
	6		0.007	0.095	0.192	0.145	0.052	0.008	0.001			
	7		0.001	0.044	0.153	0.180	0.096	0.024	0.002			
	8			0.017	0.098	0.180	0.144	0.053	0.008			
	9			0.005	0.051	0.146	0.176	0.098	0.022	0.001		
	10			0.001	0.022	0.098	0.176	0.146	0.051	0.005		
	11				0.008	0.053	0.144	0.180	0.098	0.071		
	12				0.002	0.024	0.096	0.180	0.153	0.044	0.001	
	13				0.001	0.008	0.052	0.145	0.192	0.095	0.007	
	14					0.002	0.022	0.093	0.192	0.164	0.027	0.002
	15					0.001	0.007	0.047	0.149	0.218	0.080	0.011
	16						0.002	0.017	0.087	0.218	0.180	0.053
	17							0.005	0.036	0.154	0.285	0.179
	18							0.001	0.009	0.068	0.285	0.377
	19								0.001	0.014	0.135	0.377

Table 5.5-1: The Binomial Distribution (Continued)



		p										
n	*x*	**0.05**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**
20	0	0.358	0.122	0.012	0.001							
	1	0.377	0.270	0.058	0.007							
	2	0.189	0.285	0.137	0.028	0.003						
	3	0.060	0.190	0.205	0.072	0.012	0.001					
	4	0.013	0.090	0.218	0.130	0.035	0.005					
	5	0.002	0.032	0.175	0.179	0.075	0.015	0.001				
	6		0.009	0.109	0.192	0.124	0.037	0.005				
	7		0.002	0.055	0.164	0.166	0.074	0.015	0.001			
	8			0.022	0.114	0.180	0.120	0.035	0.004			
	9			0.007	0.065	0.160	0.160	0.071	0.012			
	10			0.002	0.031	0.117	0.176	0.117	0.031	0.002		
	11				0.012	0.071	0.160	0.160	0.065	0.007		
	12				0.004	0.035	0.120	0.180	0.114	0.022		
	13				0.001	0.015	0.074	0.166	0.164	0.055	0.002	
	14					0.005	0.037	0.124	0.192	0.109	0.009	
	15					0.001	0.015	0.075	0.179	0.175	0.032	0.002
	16						0.005	0.035	0.130	0.218	0.090	0.013
	17						0.001	0.012	0.072	0.205	0.190	0.060
	18							0.003	0.028	0.137	0.285	0.189
	19								0.007	0.058	0.270	0.377
	20								0.001	0.012	0.122	0.358

Note: All values of 0.0005 or less are omitted.

Source: J. Freund and G. Simon, *Modern Elementary Statistics,* Table "The Binomial Distribution," © 1992 Prentice-Hall, Inc. Reproduced by permission of Pearson Education, Inc.

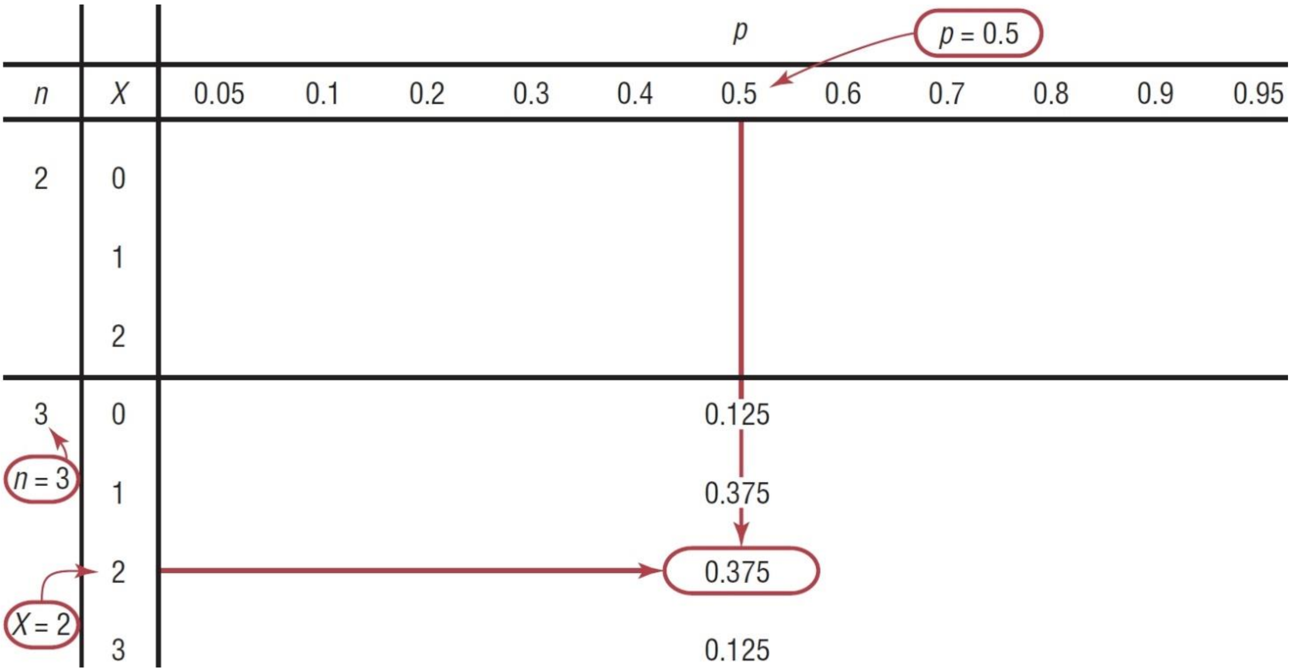
Example 5.5-5

Resolve

Example 5.5-1 using ***Table 5.5-1*** above.

Solution

Since n = 3, X = 2, and p = 0.5, the value 0.375 is found as shown in ***Figure 5.5-1***.



		p										
n	X	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
2	0											
	1											
	2											
3	0						0.125					
	1						0.375					
	2						0.375					
	3						0.125					

Figure 5.5-1: Using Table 5.5-1 to solve Example 5.5-1.

Example 5.5-6

Public Opinion reported that 5% of Americans are afraid of being alone in a house at night. If a random sample of 20 Americans is selected, find these probabilities by using the binomial table.

a. There are exactly 5 people in the sample who are afraid of being alone at night.

b. There are at most 3 people in the sample who are afraid of being alone at night.

c. There are at least 3 people in the sample who are afraid of being alone at night.

Solution

a. n = 20, p = 0.05, and X = 5. From the table, we get 0.002.

b. n = 20 and p = 0.05. “At most 3 people” means 0, or 1, or 2, or 3.

Hence, the solution is

$P(0) + P(1) + P(2) + P(3) = 0.358 + 0.377 + 0.189 + 0.060 = 0.984$

c. n = 20 and p = 0.05. “At least 3 people” means 3, 4, 5, . . . , 20.

This problem can best be solved by finding P(0) + P(1) + P(2) and subtracting from 1.

$P(0) + P(1) + P(2) = 0.358 + 0.377 + 0.189 = 0.924$

$1 - 0.924 = 0.076$

Example 5.5-7

A report from the Secretary of Health and Human Services stated that 70% of single-vehicle traffic fatalities that occur at night on weekends involve an intoxicated driver. If a sample of 15 single-vehicle traffic fatalities that occur at night on a weekend is selected, find the probability that exactly 12 involve a driver who is intoxicated.

Solution

Now, n = 15, p = 0.70, and X = 12. From Table 5.5-1,

$P\ (12) = 0.170$

Hence, the probability is 0.170.

5.5.7 Mean, Variance, and Standard Deviation for the Binomial Distribution

The mean, variance, and standard deviation of a variable that has the binomial distribution can be found by using the following formulas.

$Mean = \mu = n.p$ Eq. 5.5-2

$Variance, \sigma^2 = n.p.q$ Eq. 5.5-3

$Standard\ deviation, \sigma = \sqrt{n.p.q}$ Eq. 5.5-4

Example 5.5-8

A coin is tossed 4 times. Find the mean, and standard deviation of the number of heads that will be obtained.

Solution

With the formulas for the binomial distribution and n = 4, p = 1/2, and q = 1/2 , the results are:

$$\mu = 4 \times \frac{1}{2} = 2$$

$$\sigma = \sqrt{n.p.q} = \sqrt{4 \times \frac{1}{2} \times \frac{1}{2}} = 1$$

From this example, when four coins are tossed many, many times***, the average of the number of heads that appear is 2***, and ***the standard deviation of the number of heads is 1. Note that these are theoretical values***.

Example 5.5-9

A die is rolled 480 times. Find the mean, and standard deviation of the number of 3s that will be rolled.

Solution

This is a binomial experiment since getting a 3 is a success and not getting a 3 is considered a failure.

Hence n = 480, $p = 1/6$, and $q = 5/6$.

$$\mu = n.p = 480 \times \frac{1}{6} = 80$$

$$\sigma = \sqrt{n.p.q} = \sqrt{480 \times \frac{1}{6} \times \frac{5}{6}} \approx 8$$

Example 5.5-10

The Statistical Bulletin published by Metropolitan Life Insurance Co. reported that 2% of all-American births result in twins. If a random sample of 8000 births is taken, find the mean, and standard deviation of the number of births that would result in twins.

Solution

$$\mu = n.p = 8000 \times \frac{2}{100} = 160$$

$$\sigma = \sqrt{n.p.q} = \sqrt{8000 \times \frac{2}{100} \times \frac{98}{100}} \approx 13$$

5.5.8 ENGINEERING EXAMPLES

Example 5.5-11

The quality assurance department in a structural-steel factory inspects every product coming off its production line. The product either fails (F) or passes (S) the inspection. Past experience indicates that the probability of failure (having a defective product) is 5%. When production line manufactures (on the average) 1000 units of the product daily, what are the mean, standard deviation, and coefficient of variation for the nondefective units.

Solution

The mean, standard deviation, and coefficient of variation for the nondefective units:

$$\mu = n.p = 1000 \times \frac{95}{100} = 950 \text{ nondefective units per day}$$

$$\sigma = \sqrt{n.p.q} = \sqrt{1000 \times \frac{95}{100} \times \frac{5}{100}} \approx 7 \text{ nondefective units per day}$$

$$coefficient\ of\ variation = \nu = \frac{\sigma}{\mu} = \frac{7}{950} = 0.00737$$

Example 5.5-12

Suppose a road is flooded with probability p = 0.1 during a year and not more than one flood occurs during a year. What is the probability that it will be flooded at least once during a 5-year period?

Solution

One needs to determine the probability of having no floods and subtracting this from unity, which is the sum of the probabilities of having 0, 1, 2, 3, 4, or 5 floods during the 5-year period. This procedure is followed when it is easier to compute the probability of the complementary event than the probability of the stated event. Thus, the probability that the road will be flooded at least once is:

$$P(X) = \frac{n!}{(n-X)!\,X!} \times p^X \times q^{n-X}$$

$$n = 5, X = 0, and\ p = 0.1$$

$$P(0) = \frac{5!}{(5-0)! \times 0!} \times (0.1)^0 \times (0.9)^{(5-0)} = 0.590$$

$$P(X \geq 1) = 1 - P(0) = 1 - 0.590 = 0.41$$

5.5.9 Homework Problems

Home Work 5.5-1

The probability of a flood in any one year is 0.1. In a 10 years period, what is the probability of: (a) No floods, (b) Two or fewer floods?

Ans. $P(\text{No floods}) = 0.349$ ■

Home Work 5.5-2

Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Find:

The probability that in the next 18 samples, exactly 2 contain the pollutant.

The probability that at least four samples contain the pollutant.

Ans. $P(X \geq 4) = 0.098$ ■

Home Work 5.5-3

A particularly long traffic light on your morning commute is green 20% of the time that you approach it. Assume that each morning represents an independent trial.

(a) Over five mornings, what is the probability that the light is green on exactly one day? Ans $P(1) = 0.4096$ ■

(b) Over 20 mornings, what is the probability that the light is green on exactly four days? Ans. $P(4) = 0.2182$ ■

(c) Over 20 mornings, what is the probability that the light is green on more than four days? Ans. $P(X > 4) = 0.370$ ■

Home Work 5.5-4

Suppose a construction company has an experience record showing that 60% of its projects were completed on schedule. If this record prevails, the probability of the number of on-schedule completions in the next six projects can be described by the binomial distribution. Determine the probability that at least two projects complete on the schedule.

Ans. $P(X \geq 2) = 0.959$■

5.6 The Poisson Distribution

5.6.1 Definition, Applications, and Formula

- It is a discrete probability distribution that is useful:
  - When n is large and p is small,
  - When the independent variables occur over a period of time is called the Poisson distribution.
  - when a density of items is distributed over a given area or volume, such as the number of plants growing per acre or the number of defects in a given length of videotape.
- Formula for the Poisson Distribution:

  The probability of X occurrences in an interval of time, volume, area, etc., for a variable where λ (Greek letter lambda) is the mean number of occurrences per unit (time, volume, area, etc.) is:

  $$P(X;\lambda)=\frac{e^{-\lambda}\lambda^{X}}{X!} \qquad \text{Eq. 5.6-1}$$

  where X = 0, 1, 2, . . .

  The letter e is a constant approximately equal to 2.7183.

5.6.2 Mean and Variance of the Poisson Distribution

For a Poisson distinction with a mean number of occurrences per unit, λ, the mean and variance of X are given by:

$$\mu_X=\lambda \qquad \text{Eq. 5.6-2}$$

$$\sigma_X^2=\lambda \qquad \text{Eq. 5.6-3}$$

Example 5.6-1

If there are 200 typographical errors randomly distributed in a 500-page manuscript, find the probability that a given page contains exactly 3 errors.

Solution

First, find the mean number λ of errors. Since there are 200 errors distributed over 500 pages, each page has an average of:

$$\lambda=\frac{200}{500}=0.4\, error\ per\ page$$

Since X = 3, substituting into the formula yields:

$$P(X;\lambda)=\frac{e^{-\lambda}\lambda^{X}}{X!}$$

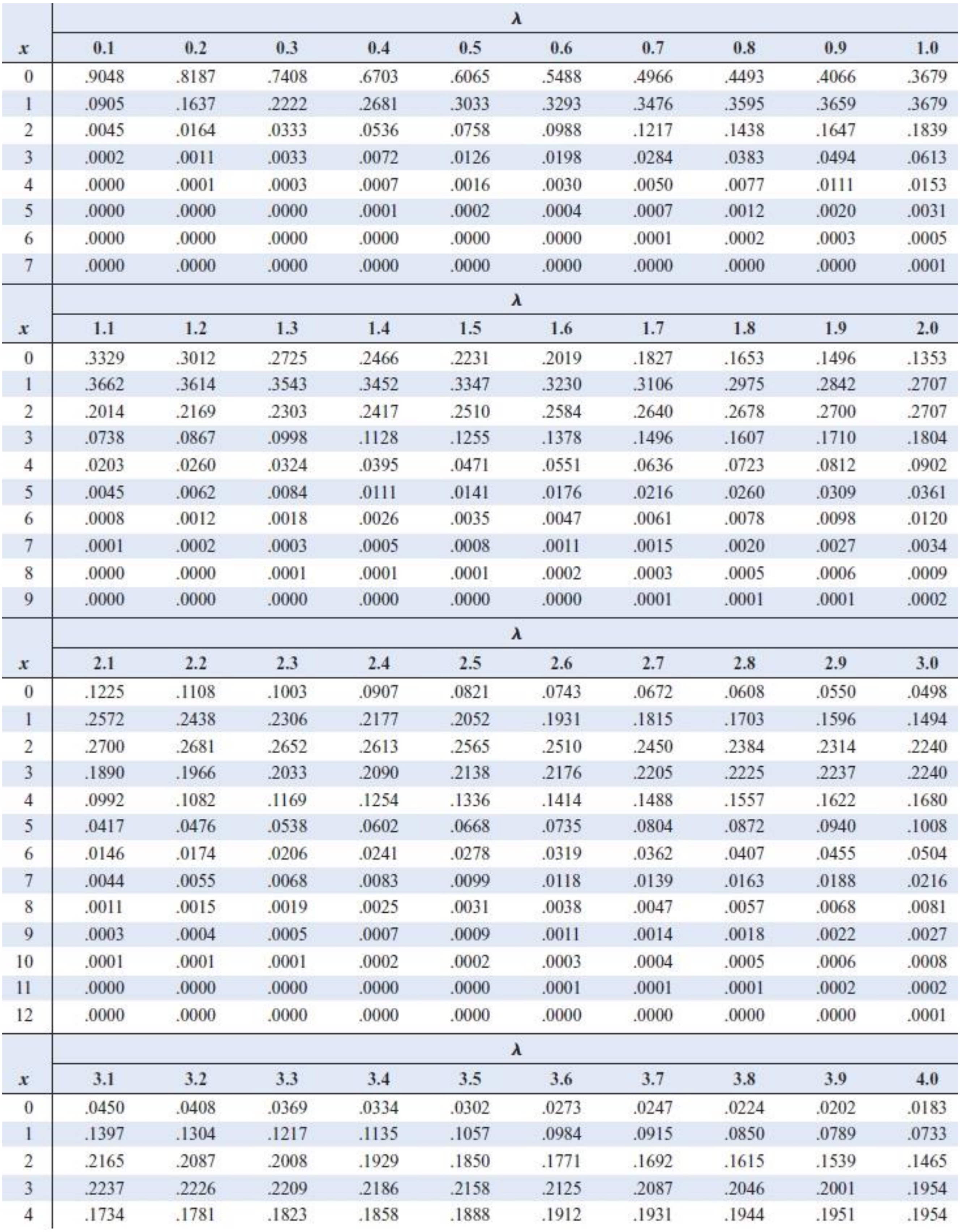
$$P(3;0.4)=\frac{e^{-0.4}\times 0.4^{3}}{3!}=0.00715$$

Thus, there is less than a 1% chance that any given page will contain exactly 3 errors.

5.6.3 TABLE FOR THE POISSON DISTRIBUTION

- Since the mathematics involved in computing Poisson probabilities is somewhat complicated, tables have been compiled for these probabilities.
- ***Table 5.6-1*** below gives P for various values for λ and X.

Table 5.6-1: The Poisson Distribution.



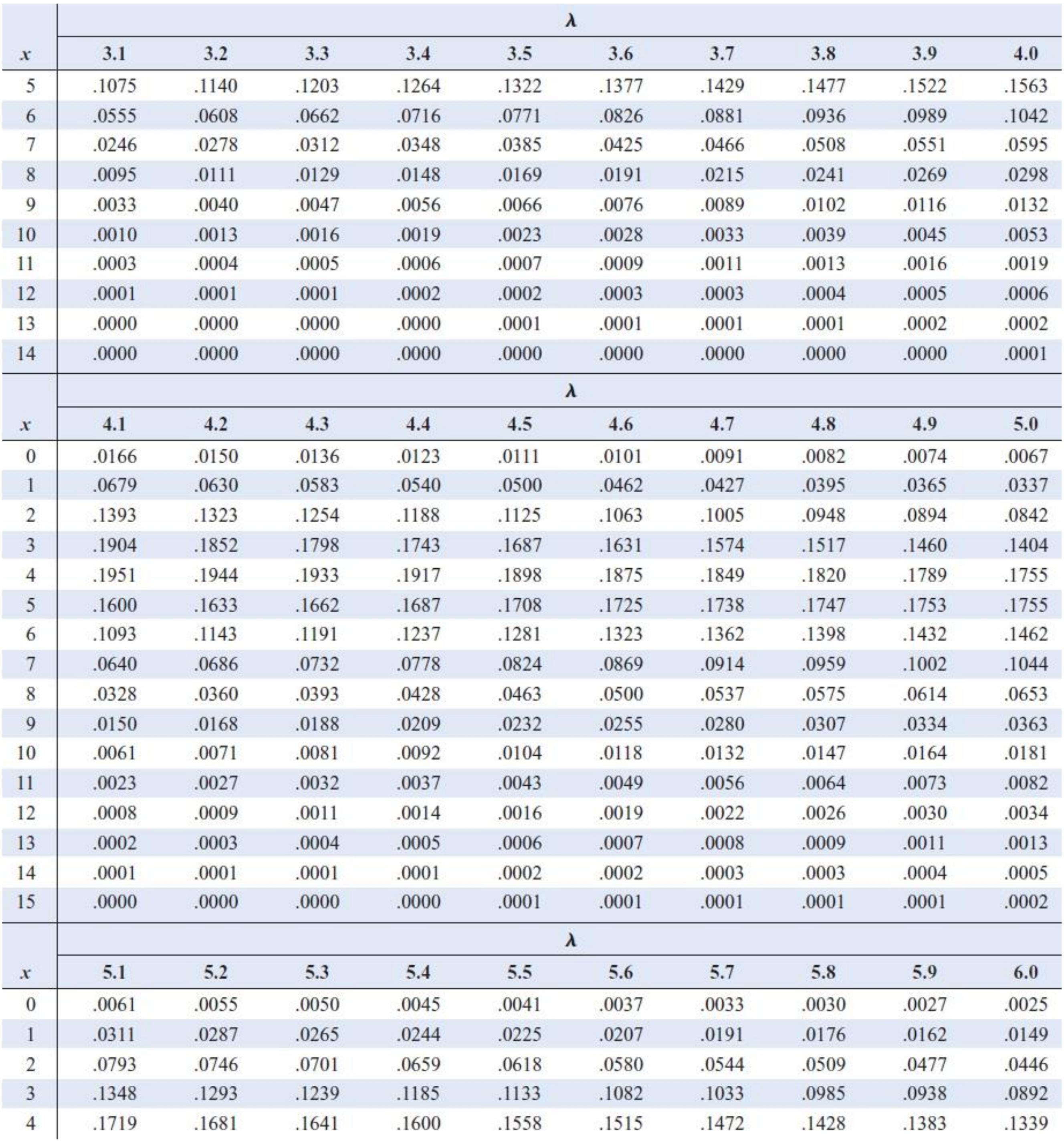
	λ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

	λ									
x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1734	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954

Table 5.6-1: The Poisson Distribution. Continued

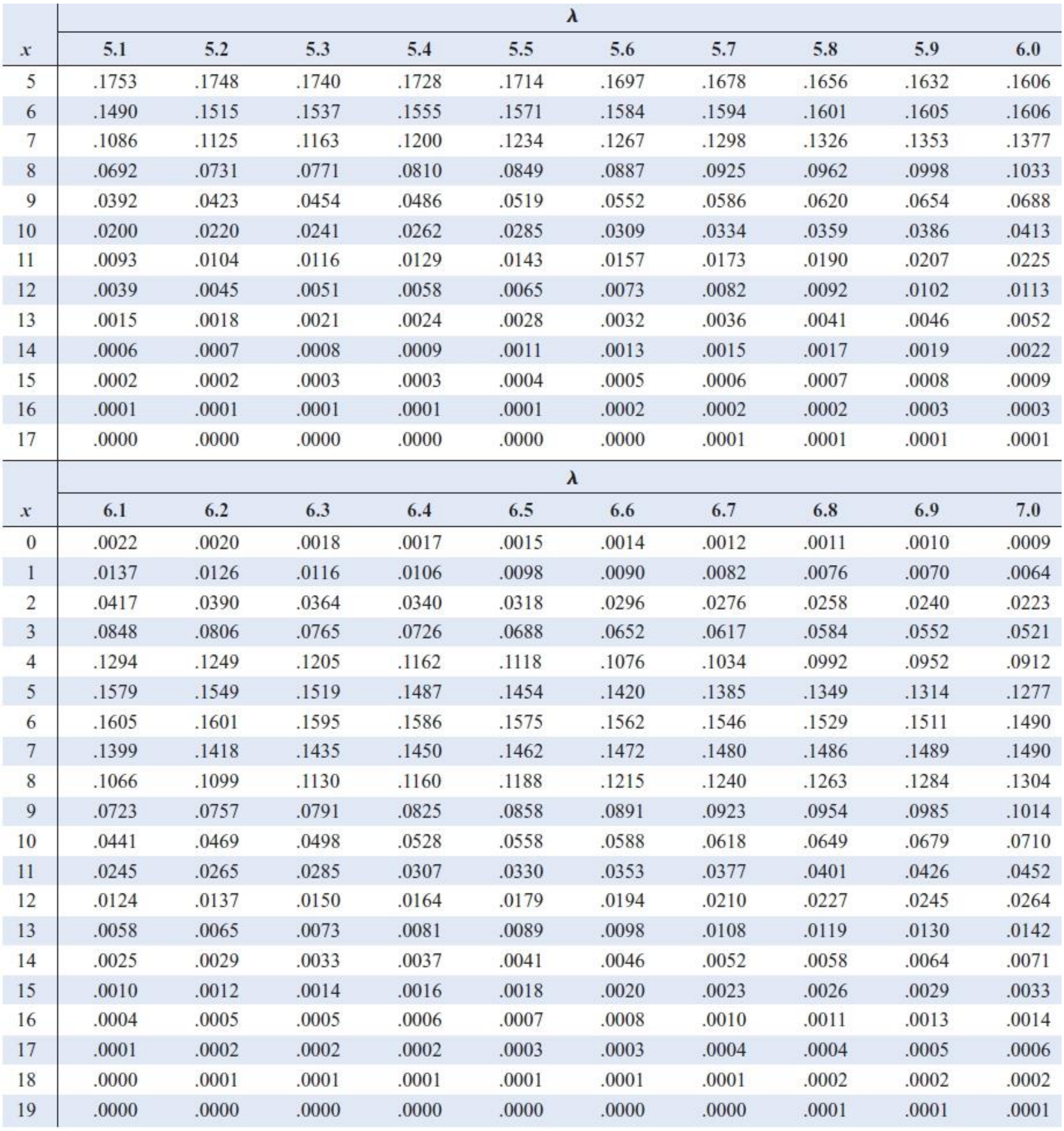


	λ									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0280	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

	λ									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339

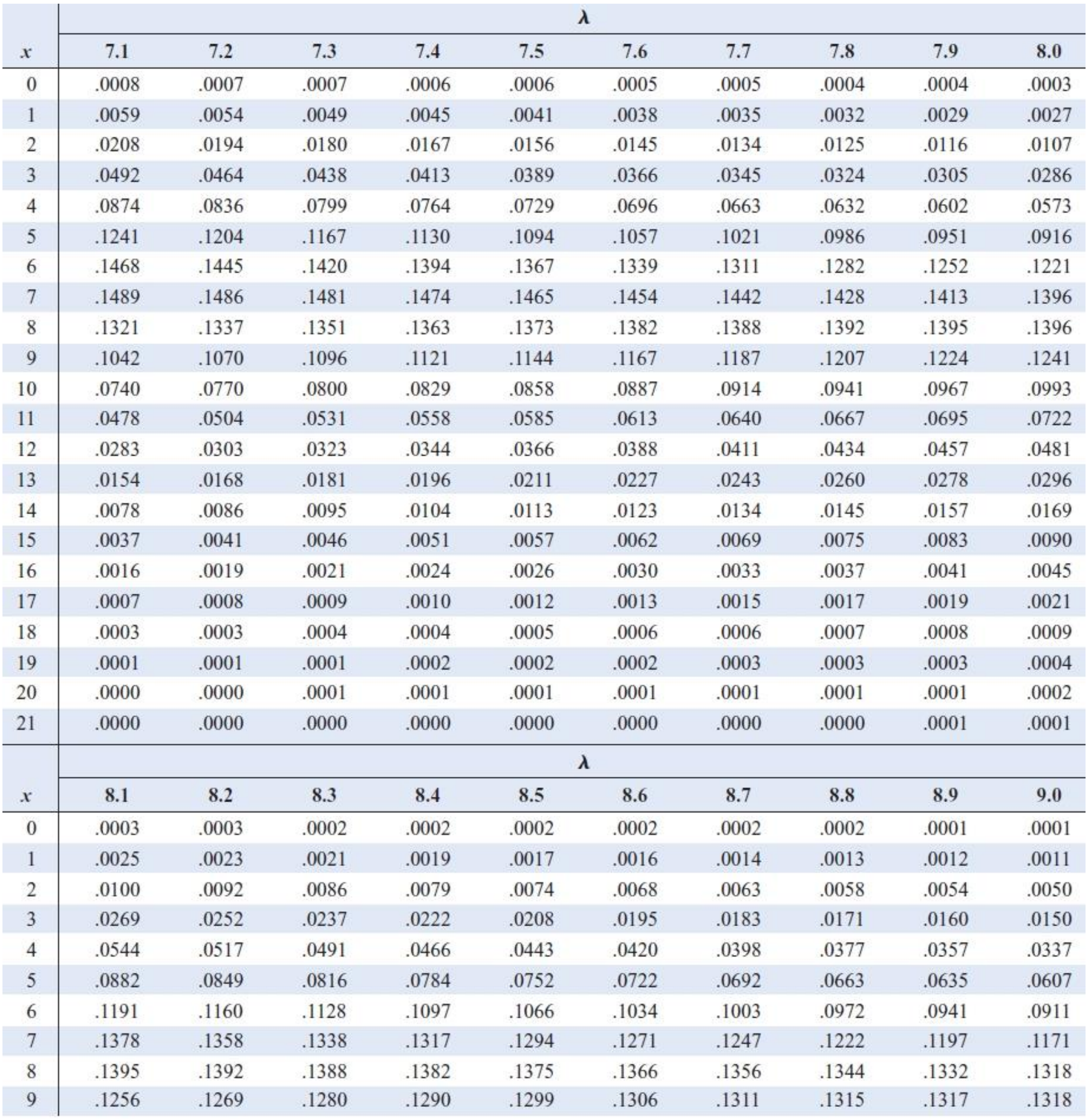
Table 5.6-1: The Poisson Distribution. Continued



	λ									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001

	λ									
x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0245	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0264
13	.0058	.0065	.0073	.0081	.0089	.0098	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

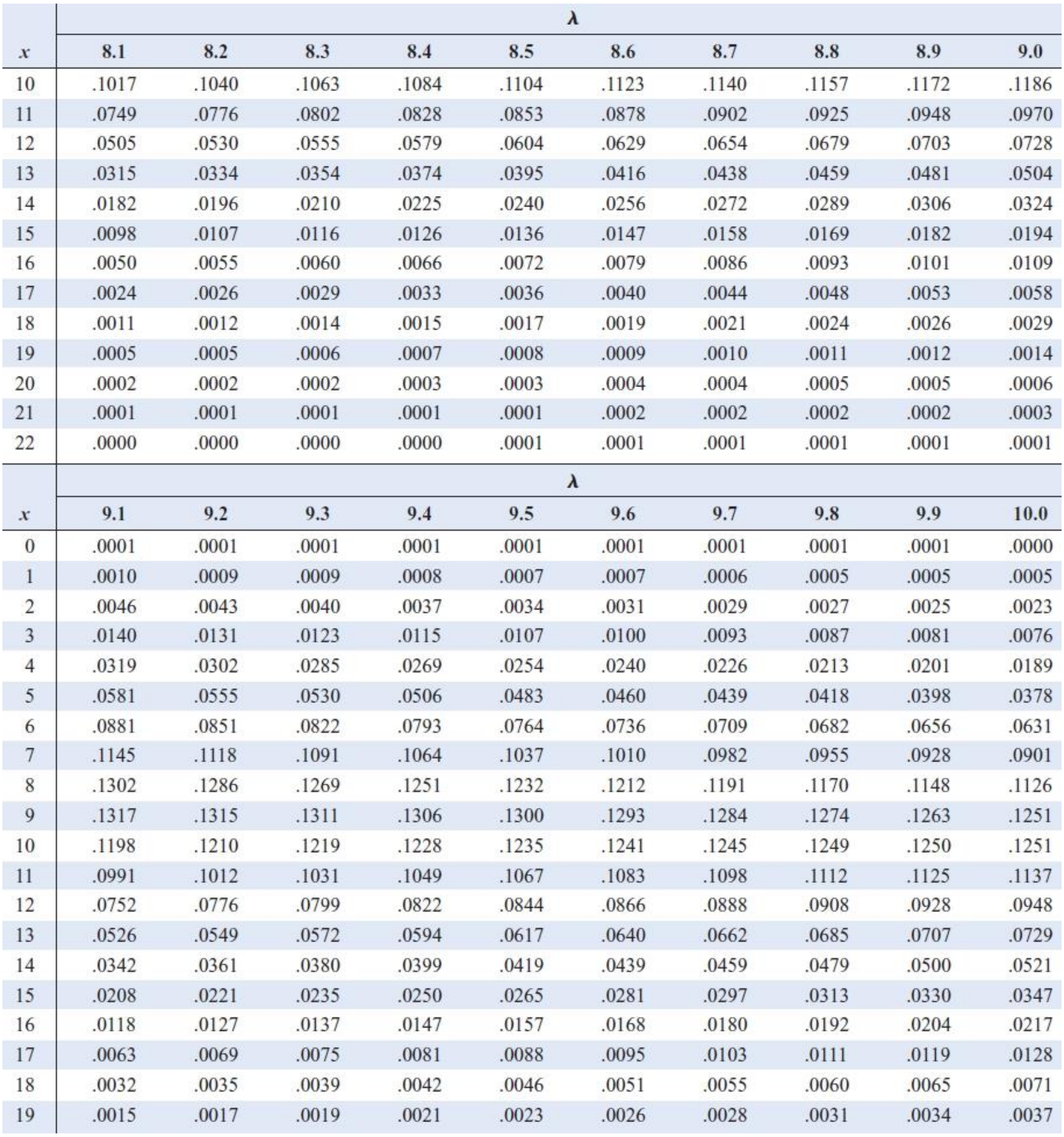
Table 5.6-1: The Poisson Distribution. Continued



	λ									
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1382	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

	λ									
x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1256	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318

Table 5.6-1: The Poisson Distribution. Continued



	λ									
x	**8.1**	**8.2**	**8.3**	**8.4**	**8.5**	**8.6**	**8.7**	**8.8**	**8.9**	**9.0**
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0776	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001

	λ									
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.9**	**10.0**
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1210	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037

Table 5.6-1: The Poisson Distribution. Continued

x	λ									
	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

x	λ									
	11	12	13	14	15	16	17	18	19	20
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0018	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516
16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0256	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125

x	λ									
	11	12	13	14	15	16	17	18	19	20
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0010	.0020
34	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0012
35	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0007
36	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

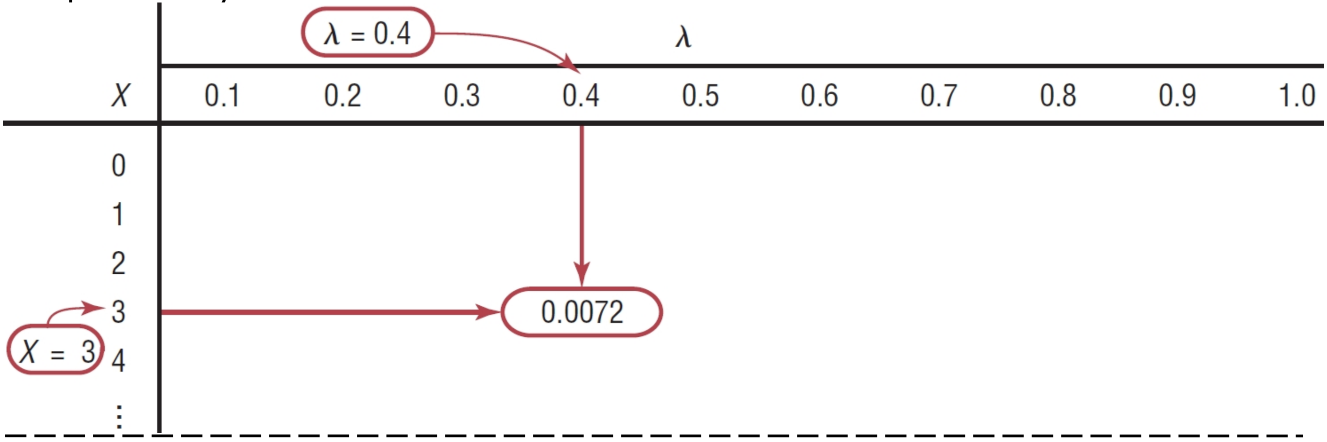
Reprinted with permission from W. H. Beyer, *Handbook of Tables for Probability and Statistics*, 2nd ed. Copyright CRC Press, Boca Raton, Fla., 1986.

Example 5.6-2

Resolve ***Example 5.6-1*** above with using Table 5.6-1.

Solution

In ***Example 5.6-1***, where X is 3 and λ is 0.4, the table gives the value 0.0072 for the probability.



Example 5.6-3

A sales firm receives, on average, 3 calls per hour on its toll-free number. For any given hour, find the probability that it will receive the following.

a. At most 3 calls,
b. At least 3 calls,
c. 5 or more calls

Solution

a. "At most 3 calls" means 0, 1, 2, or 3 calls. Hence,

$$P(0;\,3) + P(1;\,3) + P(2;\,3) + P(3;\,3) = 0.0498 + 0.1494 + 0.2240 + 0.2240$$
$$= 0.6472$$

b. "At least 3 calls" means 3 or more calls. It is easier to find the probability of 0, 1, and 2 calls and then subtract this answer from 1 to get the probability of at least 3 calls.

$$P(0;\,3) + P(1;\,3) + P(2;\,3) = 0.0498 + 0.1494 + 0.2240 = 0.4232$$

and

$$1 - 0.4232 = 0.5768$$

c. For the probability of 5 or more calls, it is easier to find the probability of getting 0, 1, 2, 3, or 4 calls and subtract this answer from 1. Hence,

$$P(0;\,3) + P(1;\,3) + P(2;\,3) + P(3;\,3) + P(4;\,3)$$
$$= 0.0498 + 0.1494 + 0.2240 + 0.2240 + 0.1680$$
$$= 0.8152$$

and

$$1 - 0.8152 = 0.1848$$

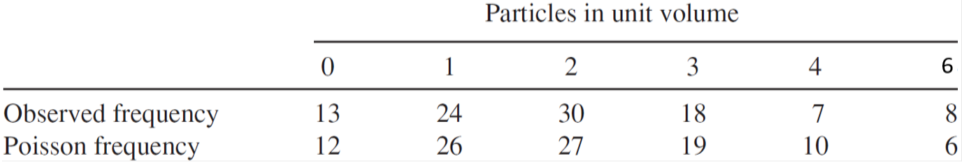
Thus, for the events described, the part "a" event is most likely to occur, and the part "c" event is least likely to occur.

5.6.4 Engineering Examples

Example 5.6-4

Atmospheric dust particles at a particular location cause an environmental problem. The number of particles within a unit volume is observed by focusing a powerful microscope on the particles and making counts. The results of tests on 100 such volumes are shown in Table 5.6-2 below.

Table 5.6-2: Poisson distribution of dust particles in the atmosphere.



	Particles in unit volume					
	0	1	2	3	4	6
Observed frequency	13	24	30	18	7	8
Poisson frequency	12	26	27	19	10	6

Use these data to estimate the mean value for dust particle in unit volume and then use Poisson distribution to generate the corresponding theoretical frequencies.

Solution

As discussed in ***Chapter 3***, mean value for grouped data can be estimated based on following relation:

$$\bar{x} = \frac{\Sigma x_i f_i}{n} = \frac{0 \times 13 + 1 \times 24 + 2 \times 30 + 3 \times 18 + 4 \times 7 + 6 \times 8}{100} = 2.14 \Longrightarrow \lambda \approx \bar{x} = 2.14$$

The theoretical frequencies would be:

$$P(X; \lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$$

$$Frequency(X) = P(X; \lambda) \times 100$$

$$Frequency(0) = P(0; 2.14) \times 100 = \frac{e^{-2.14} \times 2.14^0}{0!} \times 100 \approx 12$$

$$Frequency(1) = P(1; 2.14) \times 100 = \frac{e^{-2.14} \times 2.14^1}{1!} \times 100 \approx 26$$

$$Frequency(2) = P(2; 2.14) \times 100 = \frac{e^{-2.14} \times 2.14^2}{2!} \times 100 \approx 27$$

$$Frequency(3) = P(3; 2.14) \times 100 = \frac{e^{-2.14} \times 2.14^3}{3!} \times 100 \approx 19$$

$$Frequency(4) = P(4; 2.14) \times 100 = \frac{e^{-2.14} \times 2.14^4}{4!} \times 100 \approx 10$$

$$Frequency(6) = P(6; 2.14) \times 100 = \frac{e^{-2.14} \times 2.14^6}{6!} \times 100 \approx 2$$

Example 5.6-5

Particles suspended in a liquid medium at a concentration of 10 particles per mL. A large volume of the suspension is thoroughly agitated, i.e. disturbed, and then 1 mL is withdrawn. What is the probability that exactly eight particles are withdrawn?

Solution

$$P(X; \lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$$

$X = 8$

$\lambda = 10$ because the mean number of particles in 1 mL of suspension (the volume withdrawn) was 10.

$$P(8; 10) = \frac{e^{-10} \times 10^8}{8!} = 0.1126$$

Example 5.6-6

Particles are suspended in a liquid medium at a concentration of 6 particles per mL. A large volume of the suspension is thoroughly agitated, and then 3 mL are withdrawn. What is the probability that exactly 15 particles are withdrawn?

Solution

$\lambda_{for\ volume\ of\ 3ml} = 6 \times 3 = 18\ particle$

$X = 15\ Particles$

$\because P(X;\lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$

$\therefore P(15;18) = \frac{e^{-18} \times 18^{15}}{15!} = 0.0786$

Example 5.6-7

Assume that the number of hits, i.e. visits, on a certain website during a fixed time interval follows a Poisson distribution. Assume that the mean rate of hits is 5 per minute. Find the probability that there will be exactly 17 hits in the next three minutes.

Solution

$\lambda_{for\ 3\ minutes} = 5 \times 3 = 15\ hits$

$X = 17\ Hits\ per\ 3\ minutes$

$\because P(X;\lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$

$\therefore P(17;15) = \frac{e^{-15} \times 15^{17}}{17!} = 0.0847$

Example 5.6-8

The annual rate is assumed to equal the annual probability of fatal accidents per 1000 miles (i.e., $\lambda = 1.8 \times 10^{-5}$). The variable t is considered to be the travel distance in thousands of miles. Find the probability of having ***one fatal accident*** in 10,000 miles of travel.

Solution

$\lambda_{for\ 10000} = (1.8 \times 10^{-5}) \times 10 = 1.8 \times 10^{-4}\ fatal\ accident\ per$ thousand of miles

$\because P(X;\lambda) = \frac{e^{-\lambda}\lambda^X}{X!}$

$\therefore P(1; 1.8 \times 10^{-4}) = \frac{e^{-1.8\times10^{-4}} \times (1.8 \times 10^{-4})^1}{1!} = 1.7997 \times 10^{-4}$

Example 5.6-9

In a certain city, the number of potholes on a major street follows a Poisson distribution with a rate of 3 per mile. Let X represent the number of potholes in a two-mile stretch of road. Find

a. P(X = 4)
b. P(X ≤ 1)
c. μ_x
d. σ_x

Solution

a. P(X = 4):

$\lambda_{for\ 2\ mile} = 3 \times 2 = 6$

$P(X;\lambda) = \frac{e^{-\lambda}\lambda^X}{X!} \Rightarrow P(4;6) = \frac{e^{-6}6^4}{4!} = 0.1339$■

b. P(X ≤ 1)

$P(X \le 1; 6) = \frac{e^{-6}6^1}{1!} + \frac{e^{-6}6^0}{0!} = 0.0174$■

c. μ_x

$\mu_x = \lambda = 6$■

d. σ_x

$\sigma_x^2 = \lambda \Rightarrow \sigma_x = \sqrt{\lambda} = \sqrt{6} = 2.45$ ■

5.6.5 Homework Problems

Home Work 5.6-1

All the pumps at a water treatment plant have been made to the same specifications by a single manufacturer. From tests made over ***4-week period***, it has been determined that ***there are on average two breakdowns during each period***. A new plant manager assumes that ***the problem is not serious if there are no more than four breakdowns over a period of 4 weeks***. ***What is the probability of such an occurrence***? It is assumed that the failures occur randomly in time, the occurrences are independent, and the rate of failure is constant. Thus, ***the Poisson model is applicable***.

Ans. $P(No\ more\ than\ 4\ breakdown; 2) = 0.9473$ ■

Home Work 5.6-2

The number of cracks in a section of interstate highway that are significant enough to require repair is assumed to follow a ***Poisson distribution*** with a mean of ***two cracks per mile***.

(a) What is the probability that there are no cracks that require repair in 5 miles of highway? Ans. $P(0; 10) = 4.54 \times 10^{-5}$

(b) What is the probability that at least one crack requires repair in half mile of highway? Ans. $P(X \geq 1,1) = \ 0.6321$

Home Work 5.6-3

A flow of magnitude 40 m^3/s is exceeded at a particular site on a river once in 3 months on average. What is the probability of having at least one such flood in a year? State assumptions made. Ans. $P(@\ least\ one\ such\ flood\ in\ a\ year) = 0.982$

Home Work 5.6-4

Historical records of severe rainstorms in a town over the last 20 years indicated that there had been an average number of four rainstorms per year. Assuming that the occurrences of rainstorms may be modeled with the Poisson process, what is the probability of two or more rainstorms in the next year?

Ans. $P(X \geq 2; 4) = 0.908$ ■

Contents

Chapter 6
The Normal Distribution

6.1 INTRODUCTION

For our subsequent discussion of statistical inference, we need two preliminaries:

- Probability, which was discussed in ***Chapter 4***, and
- The ***normal curve*** and ***normal area table***, which are the topics of this chapter.

As subsequent discussion will reveal, the ***normal curve holds central importance in statistics because a vast number of phenomena may be explained in terms of the normal distribution, which we shall discuss later***.

6.1.1 THE NORMAL CURVE

The normal curve is the graphical expression of the normal distribution, which is a frequency distribution that has:

- Many frequencies near the center of the distribution and
- Then gradually tapers off symmetrically.

6.1.2 BASIC EXAMPLE

- For example, if a researcher selects a random sample of 100 women, measures their heights, and constructs a histogram, the researcher gets a graph similar to the one shown in ***Figure 6.1-1*** below.

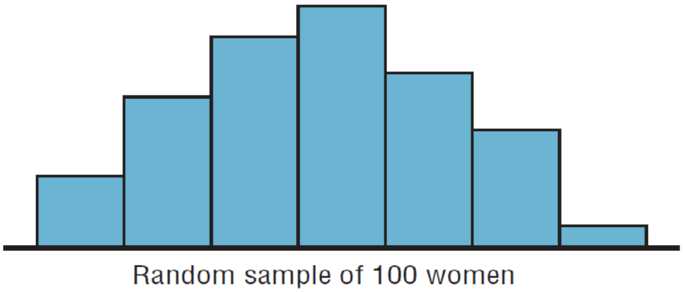


Figure 6.1-1: Histogram for data of the basic example, with sample size =100.

- Now, if the researcher increases the sample size and decreases the width of the classes, the histograms will look like the ones shown in ***Figure 6.1-2*** below.

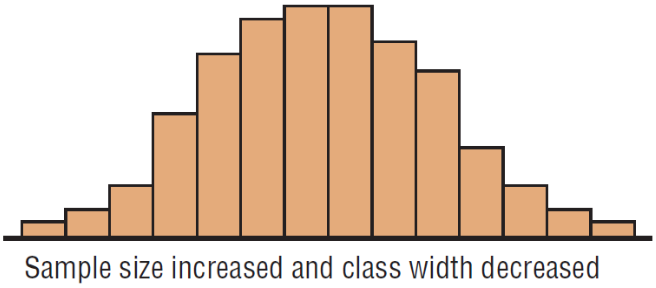


Figure 6.1-2: Histogram for data of the basic example, with sample size >100.

- Finally, if it were possible to measure exactly the heights of all adult females in the United States and plot them, the histogram would approach what is called a normal distribution, shown in ***Figure 6.1-3*** below.

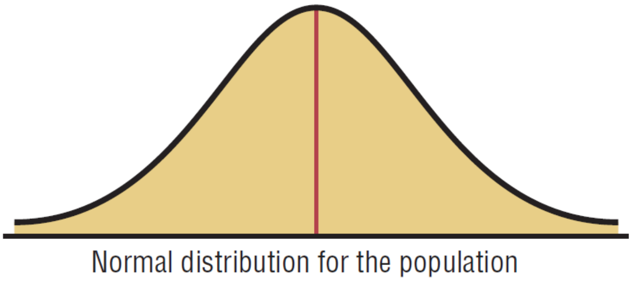


Figure 6.1-3: Histogram for data of the basic example, with sample size >>100.

- This distribution is also known as a ***bell curve*** or a ***Gaussian distribution***, named for the German mathematician Carl Friedrich Gauss (1777–1855), who derived its equation.
- ***No variable fits a normal distribution perfectly***, ***since a normal distribution is a theoretical distribution***. However, a normal distribution can be used to describe many variables, because the deviations from a normal distribution are very small.

6.1.3 Chapter Preview

This chapter will present:

- The properties of a normal distribution and discuss its applications.
- Then a very important fact about a normal distribution called the ***central limit theorem*** will be explained.

Johann Carl Friedrich Gauss (30 April 1777– 23 February 1855) was a German ***mathematician*** and ***physicist*** who made significant contributions to many fields in mathematics and sciences. Sometimes referred to as the "***the foremost of mathematicians***" and "***the greatest mathematician since antiquity***".

Gauss had an exceptional influence in many fields of mathematics and science and is ranked among history's most influential mathematicians.

6.2 Normal Distributions

6.2.1 Mathematical Equation for a Normal Distribution

- The mathematical equation for a normal distribution is:

$$y = \frac{e^{\frac{-(X-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Eq. 6.2-1

  where
  $e \approx 2.718$
  $\pi \approx 3.14$
  μ is the population mean,
  σ is the population standard deviation.
- The shape and position of a normal distribution curve depend on two parameters, the
  - Mean
  - Standard deviation.
- ***Figure 6.2-1*** below shows two normal distributions with the same mean values but different standard deviations. ***The larger the standard deviation, the more dispersed, or spread out, the distribution is***.

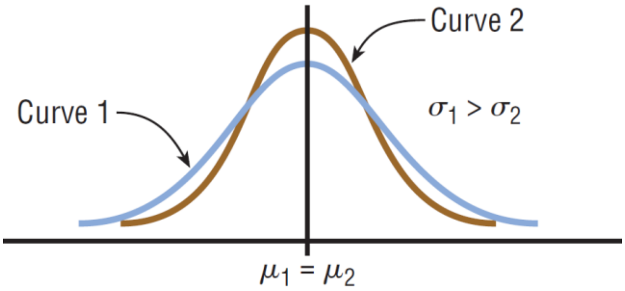


Figure 6.2-1: Two normal curves with same means but with different standard deviations.

- ***Figure 6.2-2*** below shows two normal distributions with the same standard deviation but with different means. These curves have the same shapes but are located at different positions on the x-axis.

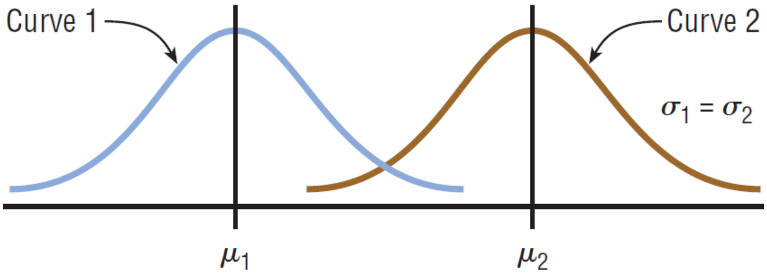


Figure 6.2-2: Two normal curves with different means but with same standard deviations.

- ***Figure 6.2-3*** shows two normal distributions with different means and different standard deviations.

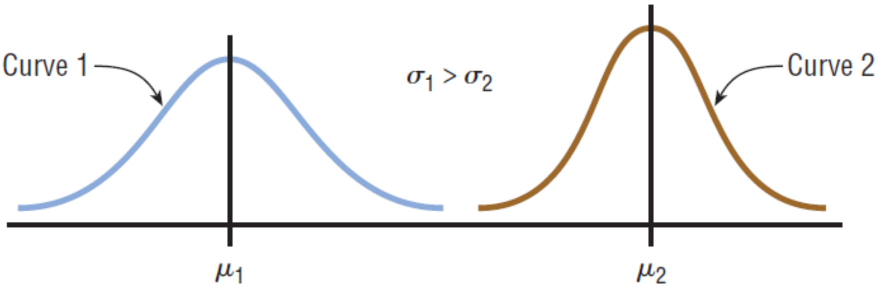


Figure 6.2-3: Two normal curves with different means and different standard deviations.

6.2.2 PROPERTIES OF THE THEORETICAL NORMAL DISTRIBUTION

- A normal distribution curve is bell-shaped.
- The mean, median, and mode are equal and are located at the center of the distribution.
- A normal distribution curve is unimodal (i.e., it has only one mode).
- The curve is symmetric about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.
- The curve is continuous; that is, there are no gaps or holes. For each value of X, there is a corresponding value of Y.
- The curve never touches the x-axis. Theoretically, no matter how far in either direction the curve extends, it never meets the x-axis but it gets increasingly closer.
- The total area under a normal distribution curve is equal to 1.00, or 100%. This fact may seem unusual, since the curve never touches the x-axis, but one can prove it mathematically by using calculus.
- The area under the part of a normal curve that lies:
  - Within 1 standard deviation of the mean is approximately 0.68, or 68%;
  - Within 2 standard deviations, about 0.95, or 95%;
  - Within 3 standard deviations, about 0.997, or 99.7%.

  Figure 6.2-4 shows the area in each region.

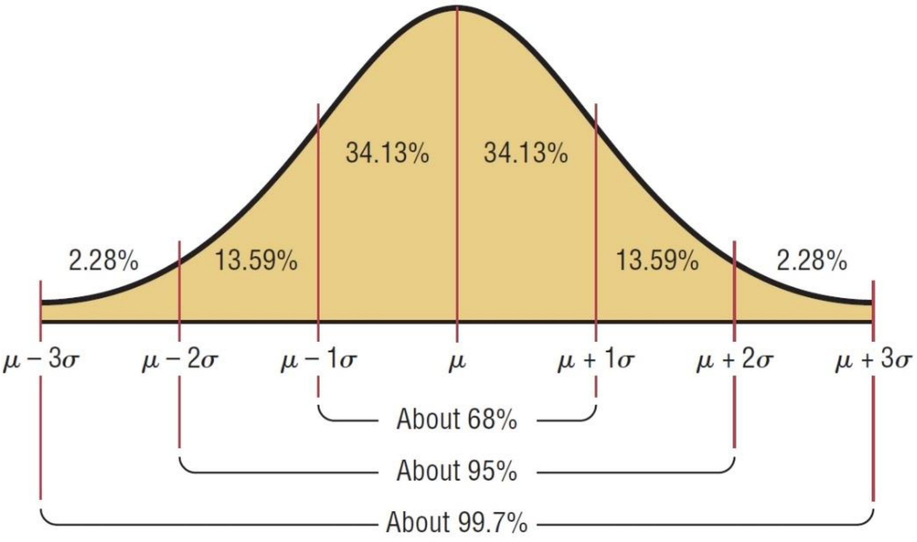


Figure 6.2-4: Areas under a normal distribution curve.

6.3 THE STANDARD NORMAL DISTRIBUTION

6.3.1 BASIC DEFINITION

- Since each normally distributed variable has its own mean and standard deviation, as stated earlier, the shape and location of these curves will vary.
- In practical applications, then, you would have to have a table of areas under the curve for each variable.
- To simplify this situation, statisticians use what is called the ***standard normal distribution.***

  The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1.

- The standard normal distribution is shown in ***Figure 6.3-1*** below.

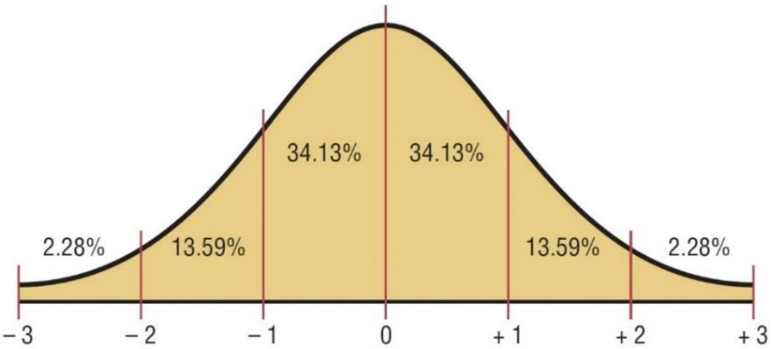


Figure 6.3-1: Standard normal distribution.

- All normally distributed variables can be transformed into the standard normally distributed variable by using the formula for the ***standard score***:

$$z = \frac{X - \mu}{\sigma}$$ **Eq. 6.3-1**

6.3.2 HIGHLIGHT ON MEANING OF TRANSFORMATION

- For example, suppose that the scores for a standardized test are normally distributed, have a mean of 100, and have a standard deviation of 15.

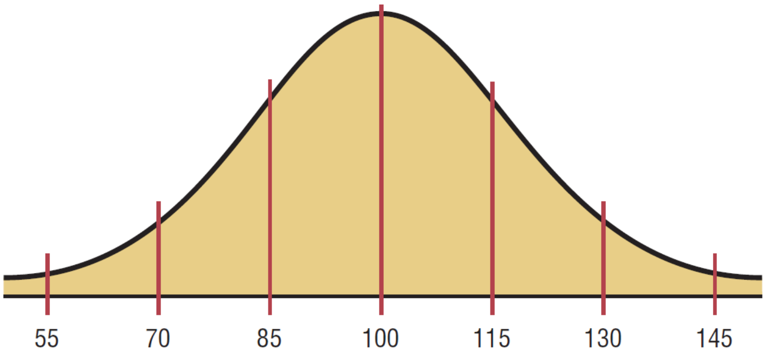


Figure 6.3-2: Normal distribution curve in terms of original data.

- When the scores are transformed to z values, the two distributions coincide, as shown in ***Figure 6.3-3*** (Recall that the z distribution has a mean of 0 and a standard deviation of 1.)

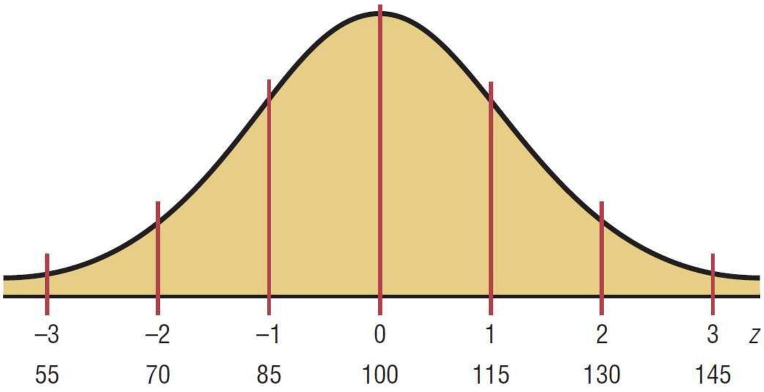
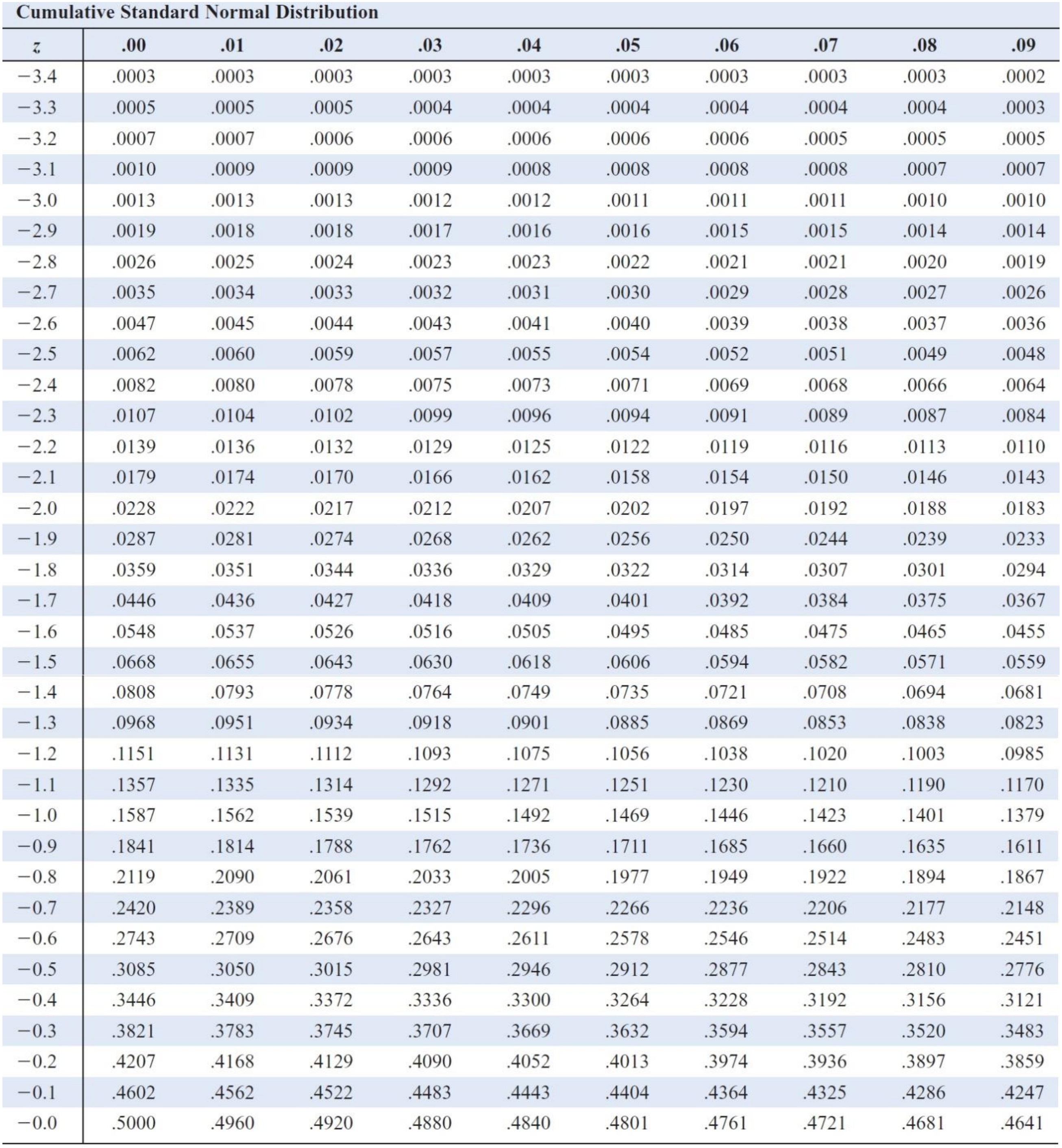


Figure 6.3-3: Normal distribution curve in terms of original data and corresponding standard scores.

6.3.3 TABLES FOR STANDARD NORMAL CURVES

- The major emphasis of this section will be to show the procedure for finding the area under the standard normal distribution curve for any z value.
- Once the X values are transformed by using the preceding formula, they are called ***z values***. The ***z value*** or ***z score*** is actually the number of standard deviations that a particular X value is away from the mean.
- ***Table 6.3-1*** below gives the area (to ***four decimal places***) under the standard normal curve for any z value from -3.49 to 3.49.

Table 6.3-1: The standard normal distribution.



Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

For z values less than −3.49, use 0.0001.

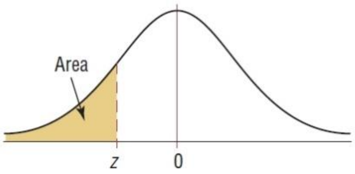
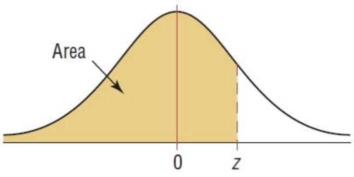


Table 6.3-1: The standard normal distribution. Continued.

Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

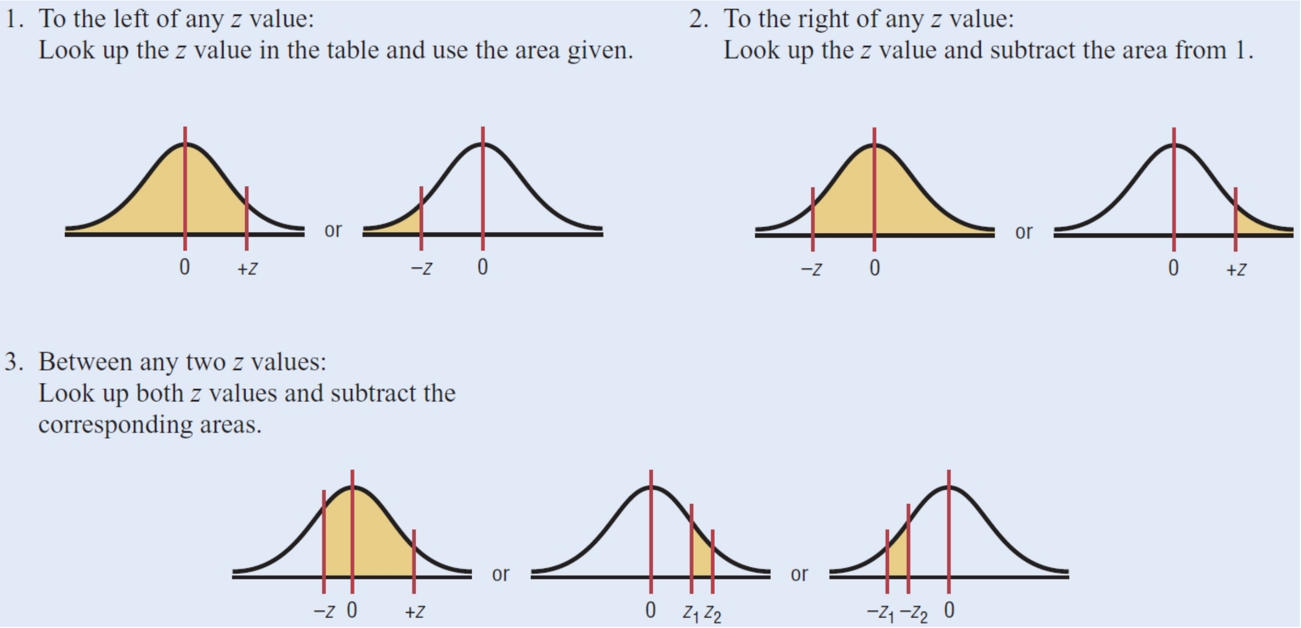
For *z* values greater than 3.49, use 0.9999.



6.3.4 FINDING AREAS UNDER THE STANDARD NORMAL DISTRIBUTION CURVE

- The solution of problems using the standard normal distribution is related to area under the normal curve for a specific z value.
- With regarding to area finding, there are three basic types of problems as indicated in ***Table 6.3-2*** below.

Table 6.3-2: Finding the area under the standard normal distribution curve.



1. To the left of any z value:
   Look up the z value in the table and use the area given.
2. To the right of any z value:
   Look up the z value and subtract the area from 1.
3. Between any two z values:
   Look up both z values and subtract the corresponding areas.

0
+z
or
−z
0
−z
0
or
0
+z
−z 0
+z
or
0
z1 z2
or
−z1 −z2 0

6.3.5 General Examples for Determining Area under the Normal Curve

Example 6.3-1

Find the area to the left of z = 2.06.

Solution

Draw the figure. The desired area is shown in ***Figure 6.3-4*** below.

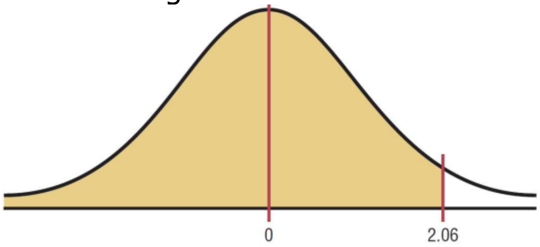
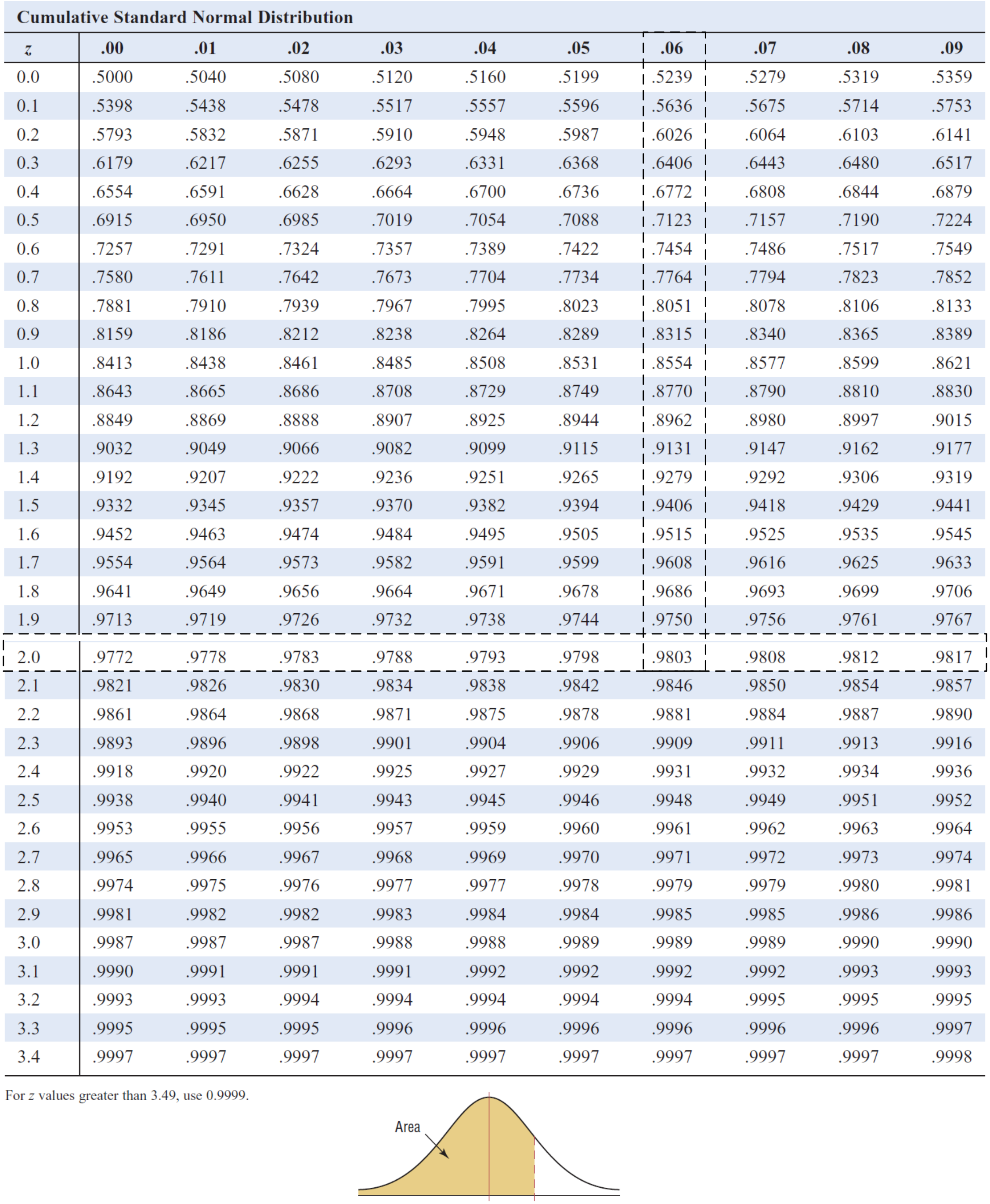


Figure 6.3-4: Standard z value for Example 6.3-1.

We are looking for the area under the standard normal distribution to the left of z = 2.06. It is 0.9803. Hence, 98.03% of the area is less than z = 2.06.



Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

For *z* values greater than 3.49, use 0.9999.

Figure 6.3-5: Standard normal curve Table applied to Example 6.3-1.

Example 6.3-2

Find the area to the right of z = -1.19.

Solution

Draw the figure. The desired area is shown in ***Figure 6.3-6*** below.

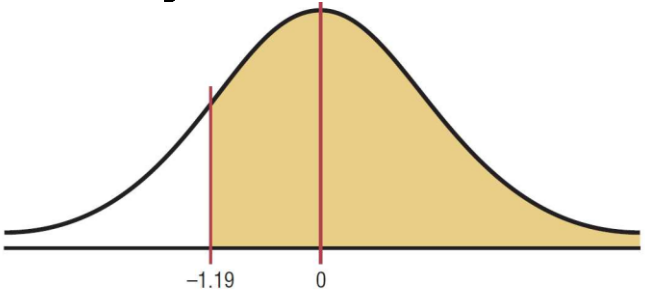
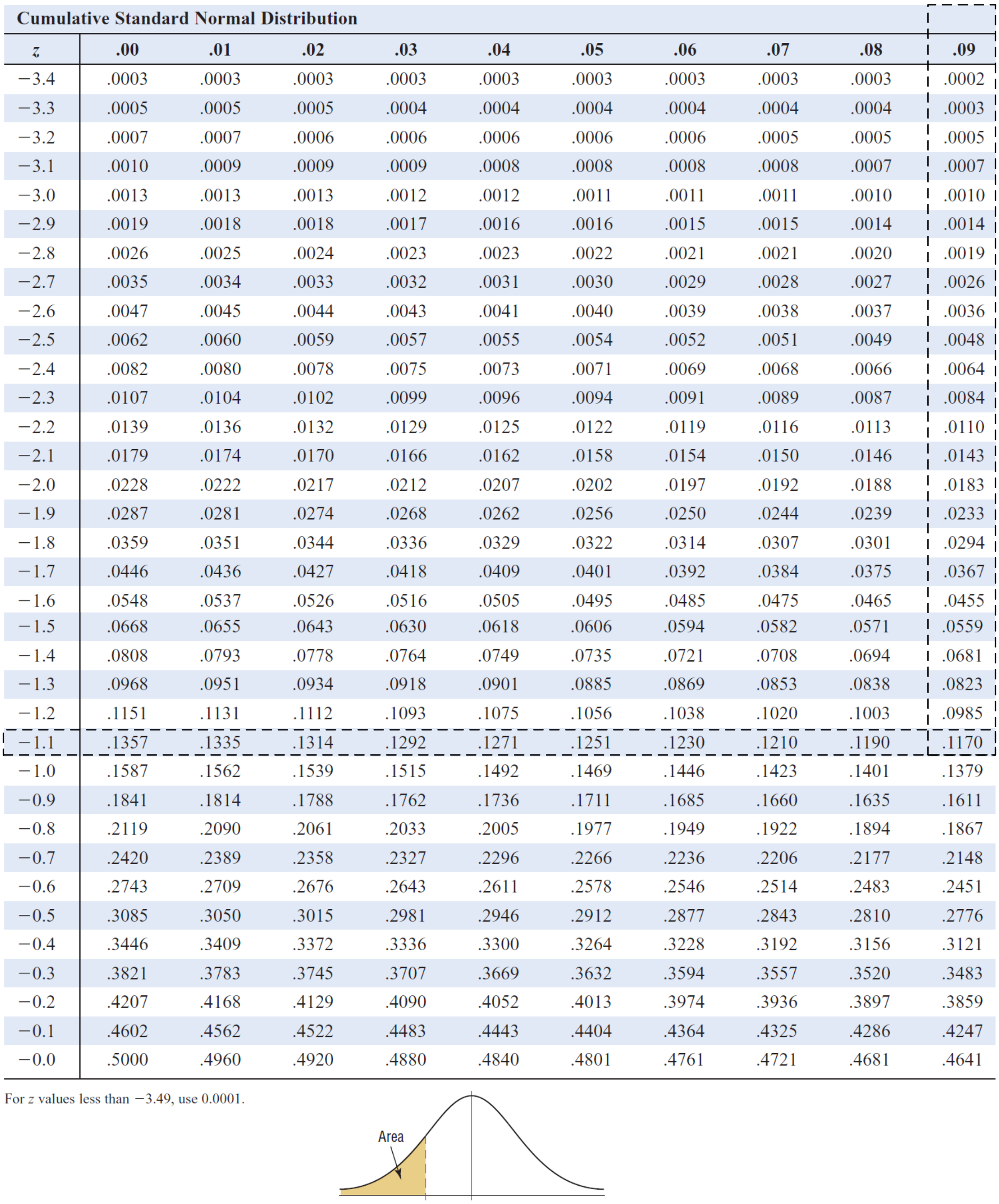


Figure 6.3-6: Standard z value for Example 6.3-2.

We are looking for the area to the right of z=-1.19. This is an example of the second case. It is 0.1170. Subtract it from 1.0000:

$1.000 - 0.1170 = 0.8830$

Hence, 88.30% of the area under the standard normal distribution curve is to the left of z = 1.19.



Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

For *z* values less than −3.49, use 0.0001.

Figure 6.3-7: Standard normal curve Table applied to Example 6.3-2.

Example 6.3-3

Find the area between z = 1.68 and z = -1.37.

Solution

Draw the figure as shown. The desired area is shown in ***Figure 6.3-8*** below:

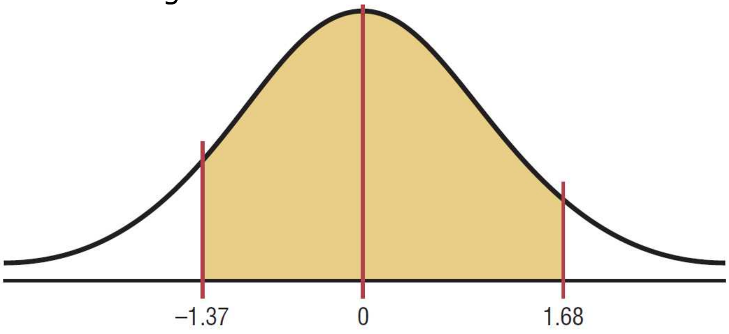


Figure 6.3-8: Standard z values for Example 6.3-3.

Since the area desired is between two given z values, look up the areas corresponding to the two z-values and subtract the smaller area from the larger area.

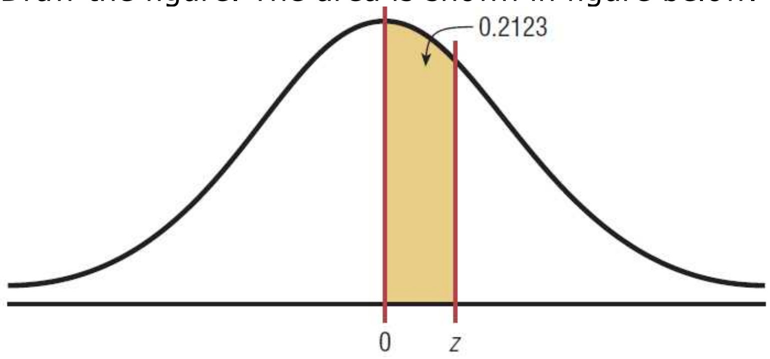
- The area for z = 1.68 is 0.9535.
- The area for z = -1.37 is 0.0853.
- Then, the area between the two z values is
  $0.9535 - 0.0853 = 0.8682$
  or 86.82%.

Example 6.3-4

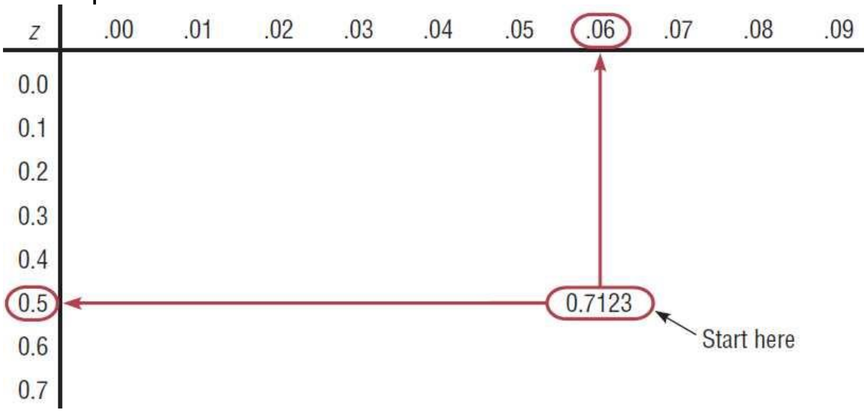
Find the z value such that the area under the standard normal distribution curve between 0 and the z value is 0.2123.

Solution

- Draw the figure. The area is shown in figure below.



- In this case it is necessary to add 0.5000 to the given area of 0.2123 to get the cumulative area of 0.7123.
- Look up the area in ***Table 6.3-1***. The value in the left column is 0.5, and the top value is 0.06.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0										
0.1										
0.2										
0.3										
0.4										
0.5							0.7123			
0.6										
0.7										

- Add these two values (0.5 + 0.06) to get z = 0.56.

6.3.6 NOTES FOR REQUIRED AREA IS NOT EXACTLY FOUND IN THE TABLE

If the exact area cannot be found, use the closest value. For example, if you wanted to find the z value for an area 0.9241, the closest area is 0.9236, which gives a z value of 1.43.

6.4 THE NORMAL DISTRIBUTION CURVE AS A PROBABILITY DISTRIBUTION CURVE

6.4.1 BASIC CONCEPTS

- The area under the standard normal distribution curve can be thought of as a probability.
- That is,
  - If it were possible to select any z value at random, the probability of choosing one, say, between 0 and 2.00 would be the same as the area under the curve between 0 and 2.00. In this case, the area is 0.4772.
  - Therefore, the probability of randomly selecting any z value between 0 and 2.00 is 0.4772.
- The problems involving probability are solved in the same manner as the previous examples involving areas in this section.
- For probabilities, ***a special notation is used***. For example, if the problem is to find the probability of any z value between 0 and 2.32, this probability is written as
  $P(0 \leq z \leq 2.32)$
- Inequality, $\lesssim \geq$, versus strict inequality:
  In a continuous distribution, ***the probability of any exact z value is 0*** since ***the area would be represented by a vertical line above the value***. But vertical lines in theory have no area. So
  $P(a \leq z \leq b) = P(a < z < b)$

6.4.2 THE PROCEDURE

- The standard normal distribution curve can be used to solve a wide variety of practical problems. The only requirement is that:
  - The variable be normally
  - Or approximately normally distributed.
- To solve problems by using the standard normal distribution:
  - Transform the original variable to a standard normal distribution variable by using the formula:
    $z = \frac{X - \mu}{\sigma}$
  - Once the variable is transformed, then the ***Table 6.3-1*** can be used to solve problems.

6.4.3 DETERMINING PROBABILITY FOR A GIVEN X OR Z VALUES

Example 6.4-1

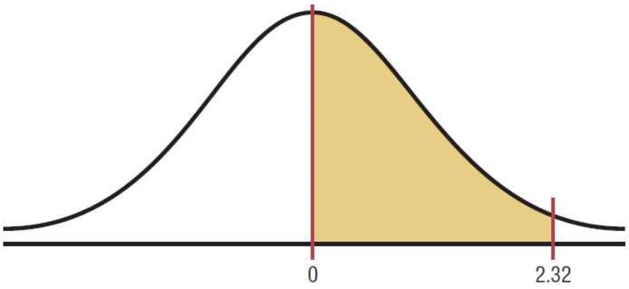
Find the probability for each.

- $P(0 < z < 2.32)$
- $P(z < 1.65)$
- $P(z > 1.91)$

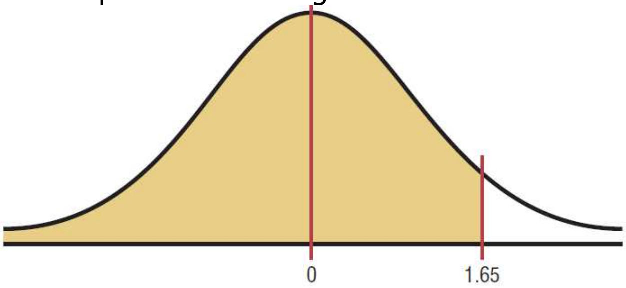
Solution

- $P(0 < z < 2.32)$:
  - Means to find the area under the standard normal distribution curve between 0 and 2.32.
  - First, look up the area corresponding to 2.32. It is 0.9898.
  - Then look up the area corresponding to z = 0. It is 0.500. Subtract the two areas:
    $0.9898 - 0.5000 = 0.4898$
    Hence, the probability is:
    $P(0 < z < 2.32) = 0.4898 \text{ or } 48.98\%$

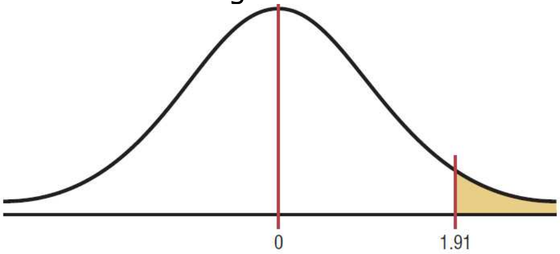
  This is shown in figure below.



- $P(z < 1.65)$:
  - It is represented in figure below.



  - Look up the area corresponding to z = 1.65 in ***Table 6.3-1***. It is 0.9505. Hence,
    $P(z < 1.65) = 0.9505 \text{ or } 95.05\%$
- $P(z > 1.91)$
  - It is shown in figure below:



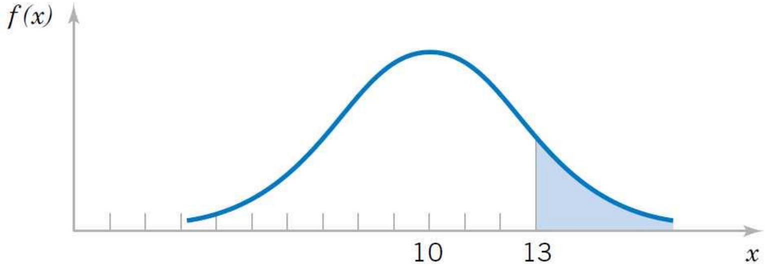
  - Look up the area that corresponds to z = 1.91. It is 0.9719.
  - Then subtract this area from 1.0000.
    $P(z > 1.91) = 1.0000 - 0.9719 = 0.0281 \text{ or } 2.81\%$

Example 6.4-2

Assume that the current measurements in a strip of wire follow a normal distribution with a mean of 10 milliamperes and a variance of 4 (milliamperes)^2. What is the probability that a measurement exceeds 13 milliamperes?

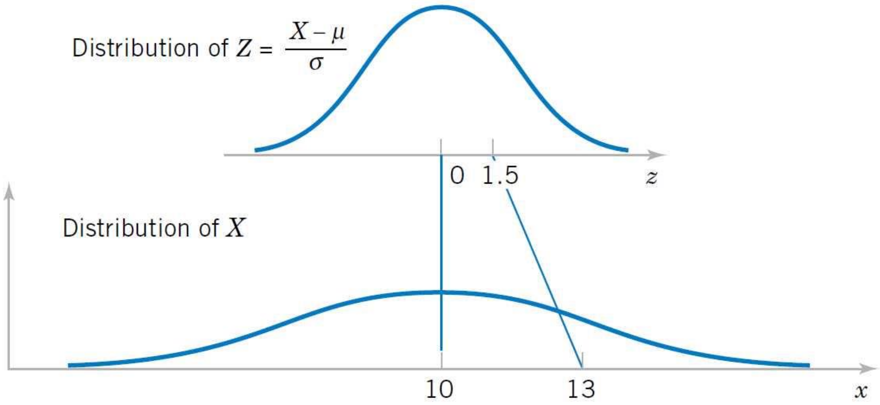
Solution

Step 1 Draw the figure and represent the area as shown in the figure.



Step 2 Find the z value corresponding to 13:

$$z = \frac{X-\mu}{\sigma} = \frac{13-10}{2} = 1.5$$



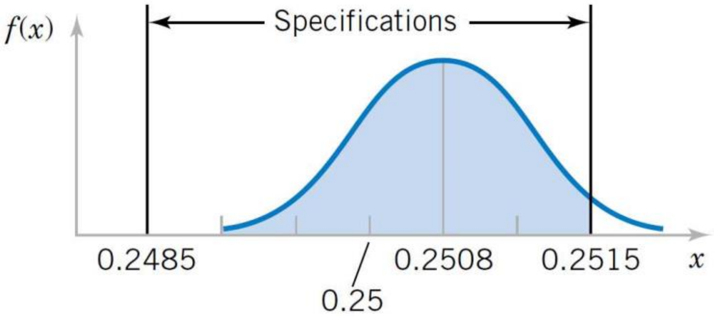
Step 3 Find the area, using ***Table 6.3-1***.

$P(X > 13) = P(z > 1.5) = 1 - P(z < 1.5) = 1 - 0.9332 = 0.0668$

Example 6.4-3

The diameter of a shaft in an optical storage drive is normally distributed with mean 0.2508 inch and standard deviation 0.0005 inch. The specifications on the shaft are 0.2500 ± 0.0015 inch. What proportion of shafts conforms to specifications?

Solution



Let X denote the shaft diameter in inches. The requested probability is shown in the figure.

$$P(0.2485 < X < 0.2515) = P\left(\frac{0.2485 - 0.2508}{0.0005} < Z < \frac{0.2515 - 0.2508}{0.0005}\right)$$

$$P(-4.6 < Z < 1.4) = P(Z < 1.4) - P(Z < -4.6) = 0.91924 - 0.0000 = 0.91924$$

Most of the nonconforming shafts are too large, because the process mean is located very near to the upper specification limit. If the process is centered so that the process mean is equal to the target value of 0.2500:

$$P(0.2485 < X < 0.2515) = P\left(\frac{0.2485 - 0.2500}{0.0005} < Z < \frac{0.2515 - 0.2500}{0.0005}\right)$$

$$P(-3 < Z < 3) = P(Z < 3) - P(Z < -3) = 0.99865 - 0.00135 = 0.9973$$

By re-centering the process, the yield is increased to approximately 99.73%.

6.4.4 Finding z and x Values for a Given Probability

- A normal distribution can also be used to find specific data values for given percentages.
- Use Table 6.3-1, to find z value for a specific probability.
- Then use the following relation to determine the corresponding X value:
  $X = \mu + z.\sigma$

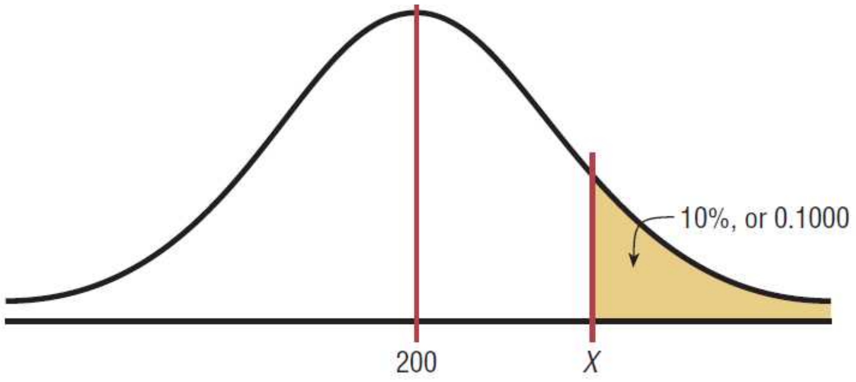
Example 6.4-4

To qualify for a police academy, candidates must score in the top 10% on a general abilities test. The test has a mean of 200 and a standard deviation of 20. Find the lowest possible score to qualify.

Assume the test scores are normally distributed.

Solution

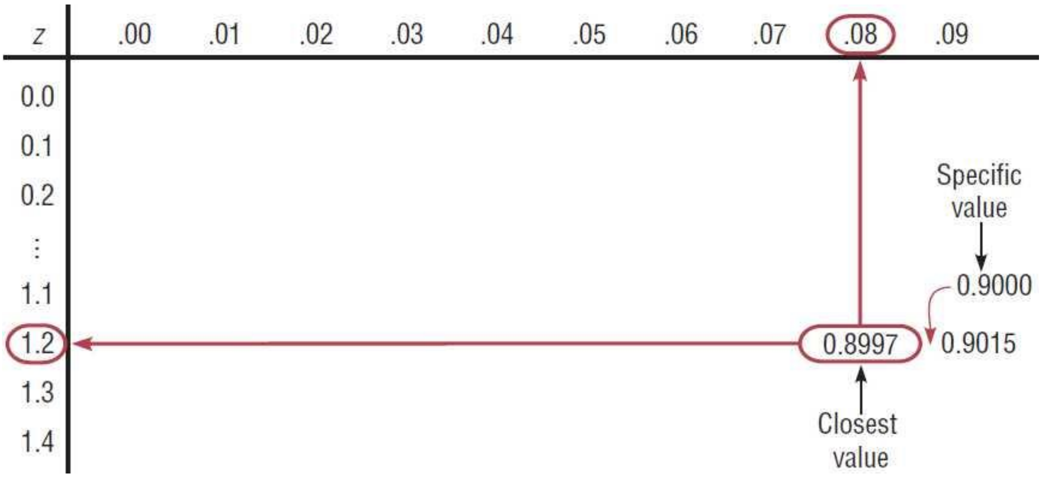
Since the test scores are normally distributed, the test value X that cuts off the upper 10% of the area under a normal distribution curve is desired. This area is shown in Figure below:



Work backward to solve this problem.

Step 1 Subtract 0.1000 from 1.000 to get the area under the normal distribution to the left of X: 1.0000 - 0.10000 = 0.9000.

Step 2 Find the z value that corresponds to an area of 0.9000 by looking up 0.9000 in the area portion of ***Table 6.3-1***. If the specific value cannot be found, use the closest value—in this case 0.8997, as shown in figure below.


z
.00
.01
.02
.03
.04
.05
.06
.07
.08
.09
0.0
0.1
0.2
⋮
1.1
1.2
1.3
1.4
Specific value
0.9000
0.8997
0.9015
Closest value

The corresponding z value is 1.28.

Step 3 Substitute in the formula:

$X = \mu + z.\sigma \Longrightarrow X = 200 + 1.28 \times 20 = 226$

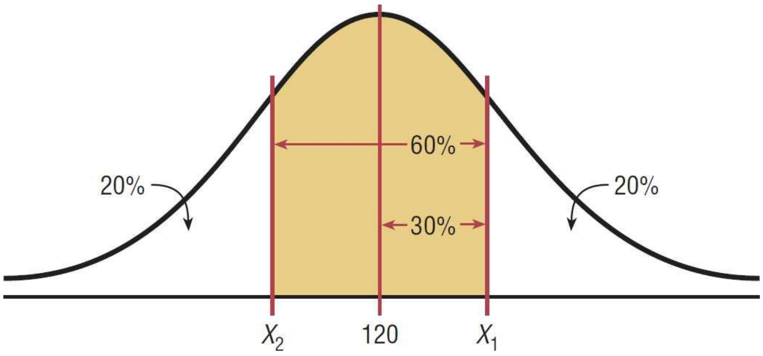
A score of 226 should be used as a cutoff. Anybody scoring 226 or higher qualifies.

Example 6.4-5

For a medical study, a researcher wishes to select people in the middle 60% of the population based on blood pressure. If the mean systolic الانقباضي blood pressure is 120 and the standard deviation is 8, find the upper and lower readings that would qualify people to participate in the study.

Solution

Assume that blood pressure readings are normally distributed; then cutoff points are as shown in the figure.



This figure shows that two values are needed, one above the mean and one below the mean. To get the area to the left of the positive z value, add

$A_1 = 0.5000 + 0.3000 = 0.8000$

The z value with area to the left closest to 0.8000 is:

$z_1 = 0.84$

Substituting in the formula

$X_1 = \mu + z\sigma = 120 + 0.84 \times 8 = 126.72$

The area to the left of the negative z value is 20%, or 0.2000.

$A_2 = 0.2000$

The area closest to 0.2000 is

$z_2 = -0.84$

Substituting in the formula

$X_2 = \mu + z\sigma = 120 - 0.84 \times 8 = 113.28$

Therefore, the middle 60% will have blood pressure readings of:

$113.28 < X < 126.72$

Example 6.4-6

Dead load acting on the beam shown in ***Figure 6.4-1*** below follows a ***normal probability curve*** with mean value of $\mu = 20\ kN/m$ and a standard deviation of $\sigma = 3\ kN/m$. What design value, $W_{D\ Design}$, should be adopted if the probability of over load in the future not greater than 0.02?

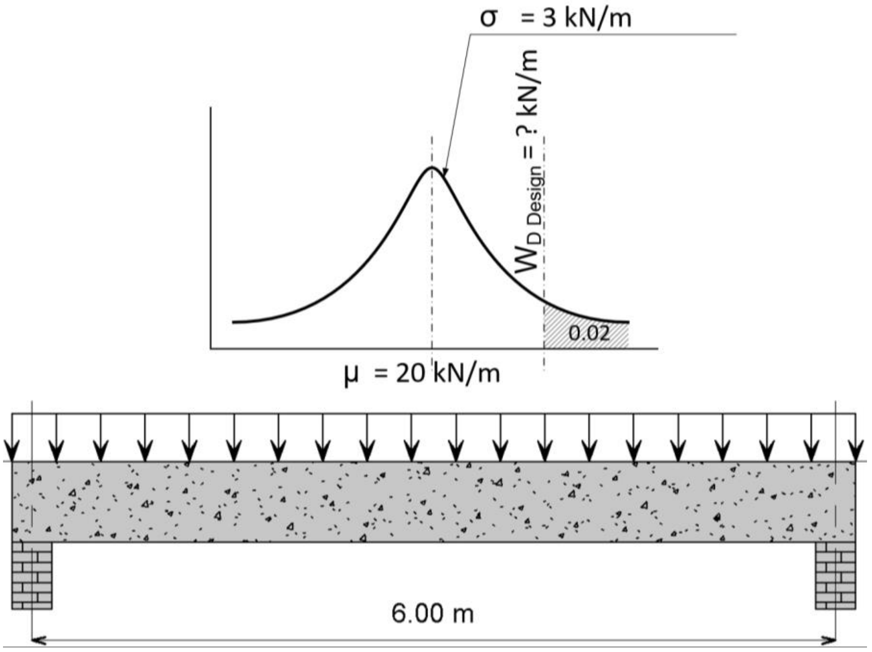


Figure 6.4-1: Simply supported beam with random uniformly distributed load for Example 6.4-6.

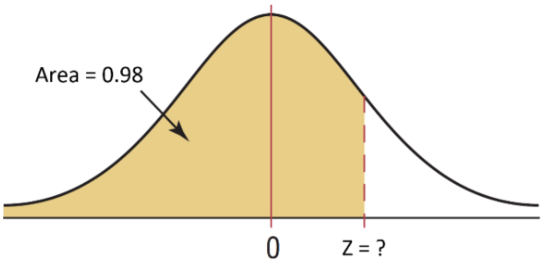
Solution

In terms of ***Table 6.3-1*** above, we are searching for z_{+ve} that making area under the curve equal to:

$Area = 1 - 0.02 = 0.98$

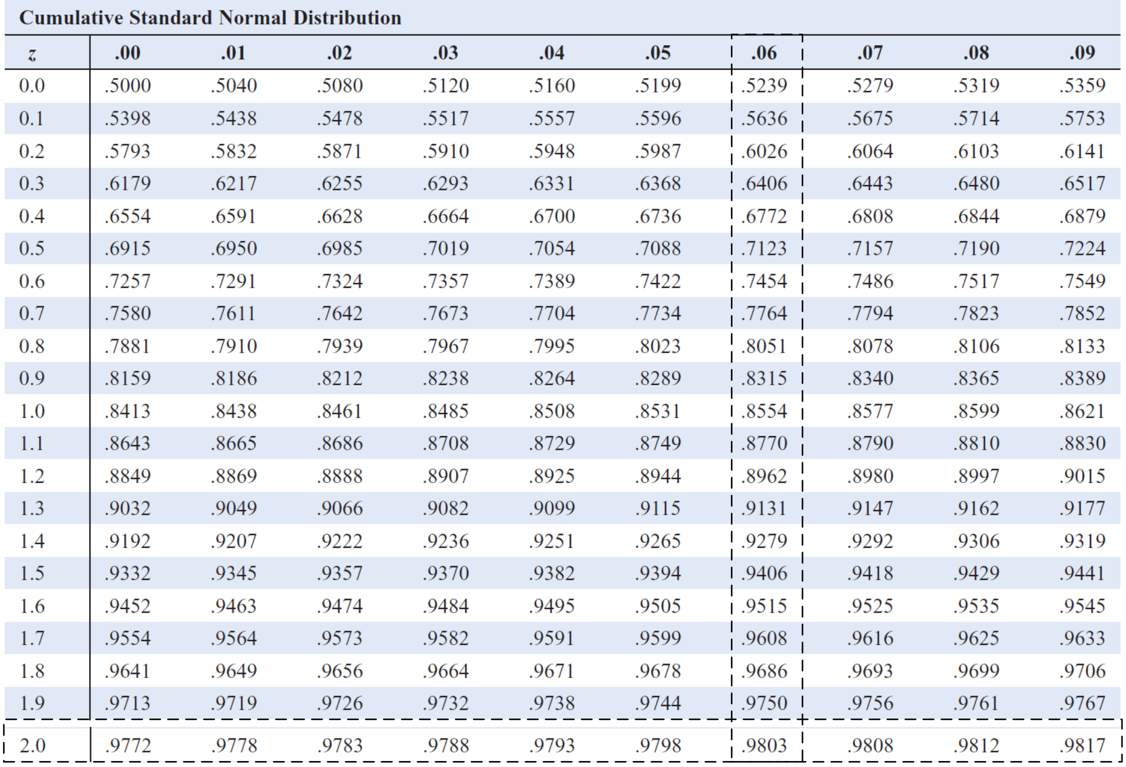
Therefore, the Z value would be, see ***Figure 6.4-2*** below:

$z = \ 2.06$



Therefore, the required design dead load would be:

$W_{D\ Design} = \mu_{D\ Load} + z.\sigma_{D\ Load} = 20 + 2.06 \times 3 = 26.18\frac{kN}{m}$ ■



Cumulative Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

Figure 6.4-2: Standard normal curve Table applied to Example 6.4-6.

Example 6.4-7

An advertising company plans to market a product to low-income families. A study states that for a particular area, the average income per family is μ=$24,596 and the standard deviation is σ=$6256. If the company plans to target the bottom 18% of the families based on income, find the cutoff income. Assume the variable is normally distributed.

Solution

Based on example statement, we are searching on Z value indicated in ***Figure 6.4-3***. From ***Table 6.3-1***, the required Z value would be:

$Z = -0.915$

The corresponding X value would be:

$X = \mu + z\sigma$

$X = 24{,}596 + (-0.915) \times 6256 = 18872$

Then, the cutoff income is $ 18872.

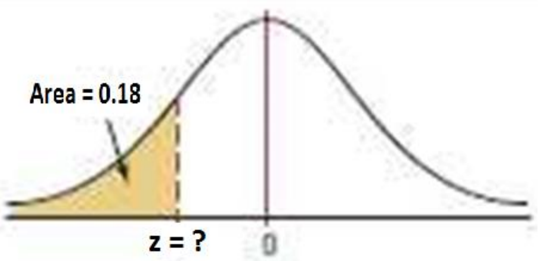


Figure 6.4-3: Required Z value for Example 6.4-7.

Example 6.4-8

Pipe supporting structure presented in ***Figure 6.4-4*** will be rejected when Post "A" ***or*** Post "B" ***settle by more than 2.5 cm***. Based on investigation of load nature and properties of underneath soil, a geotechnical engineer has proposed a ***normal model for settlement*** with a ***mean value of*** $\mu = 2.0\ \text{cm}$ and a ***standard deviation of*** $\sigma = 0.5\ \text{cm}$. What is the probability to reject the structure due to excessive settlement?

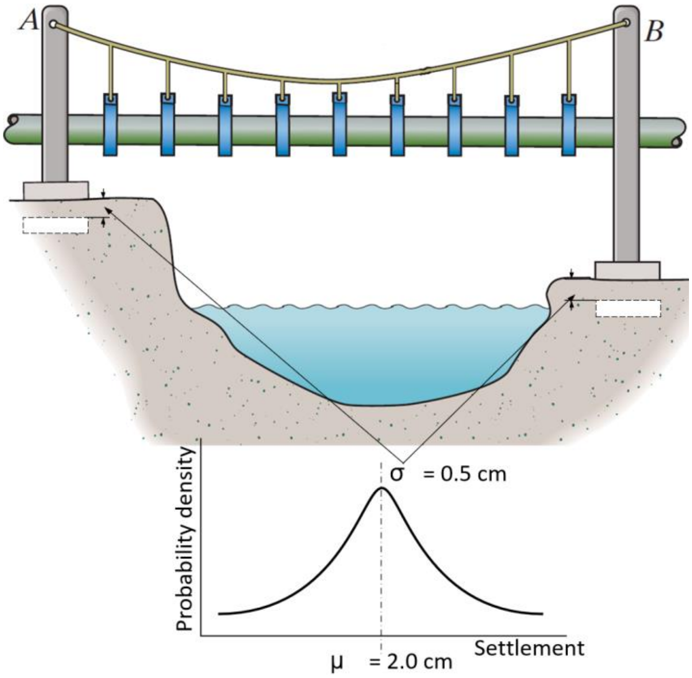
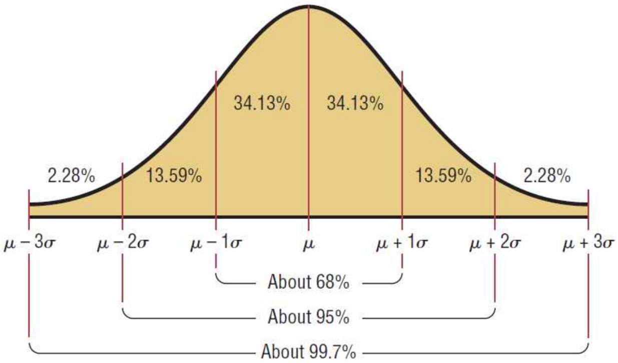


Figure 6.4-4: Pipe supporting structure of Example 6.4-8.

Solution

As the limiting settlement of 2.5 cm is equal to $\mu + \sigma$, therefore the problem can be solved directly based on properties of normal curve that presented in below:



$$P_{Settlement\ of\ Post\ \mathrm{A}\ by\ more\ than\ 2.5cm} = \frac{1-0.68}{2} = 0.16$$

In the same way,

$$P_{Settlement\ of\ Post\ \mathrm{B}\ by\ more\ than\ 2.5cm} = \frac{1-0.68}{2} = 0.16$$

Therefore, the probability of structure rejection due to excessive settlement would be:

$P \;=\; 0.16 + 0.16 = 0.32$ ■

6.4.5 HOMEWORK PROBLEMS

Home Work 6.4-1

Concentrated force "F" that acting on bracket shown ***Figure 6.4-5*** above has a mean value of $\mu = 100\ kN$ and a standard deviation of $\sigma = 10\ kN$. Select a design value of "F" with a 0.25 overload probability.

Ans. $F_{Design} = 106\ kN$■

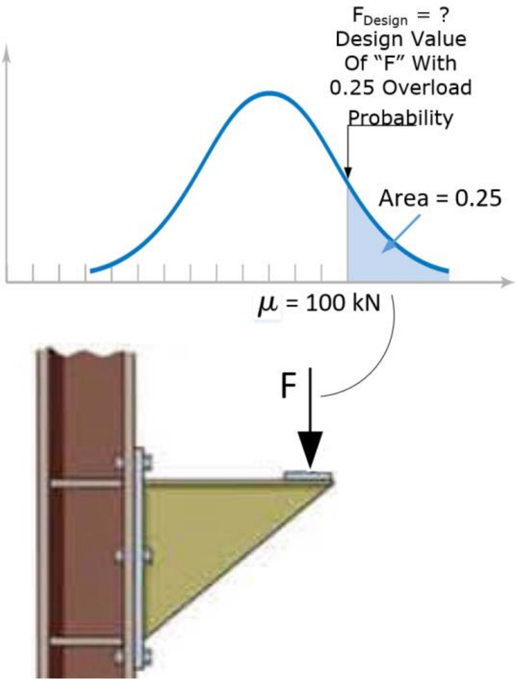


Figure 6.4-5: Bracket connection for Home Work 6.4-1.

Home Work 6.4-2

The compressive strength of samples of cement can be modeled by a ***normal distribution*** with a mean, μ, of $6000\ kg/cm^2$ and a standard deviation, σ, of $100\ kg/cm^2$.

(a) What is the probability that a sample's strength is less than $6250\ kg/cm^2$?

Ans. 0.9938.

(b) What is the probability that a sample's strength is between 5800 and $5900\ kg/cm^2$?

Ans. 0.1359

(c) What strength is exceeded by 95% of the samples?

Ans. 5836.5

Home Work 6.4-3

The water content of soil from a borrow site is normally distributed with a mean of $\mu = 14.2\%$ and standard deviation of $\sigma = 2.3\%$. What is the probability that a sample taken from the site will have a water content above 16% or below 12%?

Ans. $38.7\ \%$

6.5 Determining Normality

6.5.1 Basic Concepts

- A normally shaped or bell-shaped distribution is only one of many shapes that a distribution can assume; however, it is very important since many statistical methods require that the distribution of values (shown in subsequent chapters) be normally or approximately normally shaped.
- There are several ways statisticians check for normality:
  - The easiest way is to draw a histogram for the data and check its shape. If the histogram is not approximately bell shaped, then the data are not normally distributed.
  - Skewness can be checked by using the ***Pearson coefficient of skewness*** (PC) also called ***Pearson's index of skewness***. The formula is:

    $$PC = \frac{\left(3(\bar{X} - median)\right)}{s} \qquad \textbf{Eq. 6.5-1}$$

    If the index is greater than or equal to 1 or less than or equal to -1, it can be concluded that the data are significantly skewed.
  - In addition, the data should be ***checked for outliers*** by using the method shown in ***Chapter 3***. Even one or two outliers can have a big effect on normality.

6.5.2 Numerical Examples

Example 6.5-1

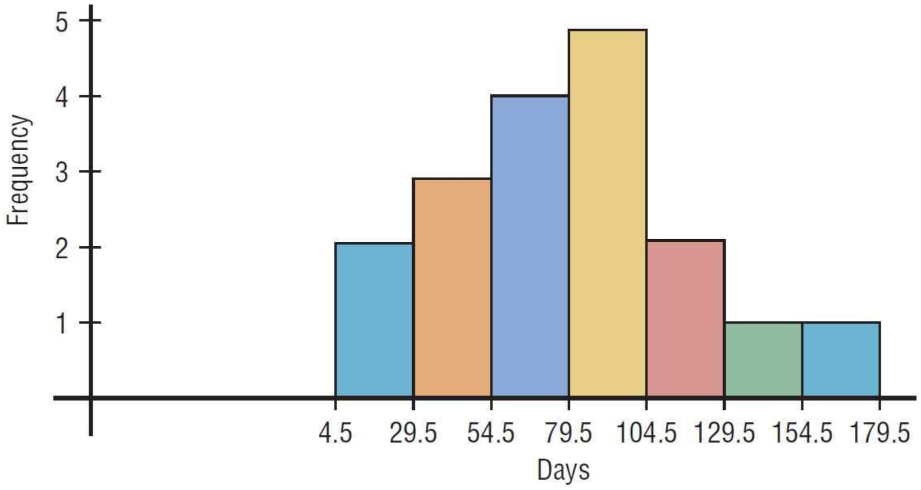
A survey of 18 high-technology firms showed the number of days' inventory they had on hand. Determine if the data are approximately normally distributed.

5	29	34	44	45	63	68	74	74
81	88	91	97	98	113	118	151	158

Solution

Step 1 Construct a frequency distribution and draw a histogram for the data, as shown in figure below:

Class	Frequency
5–29	2
30–54	3
55–79	4
80–104	5
105–129	2
130–154	1
155–179	1



Since the histogram is approximately bell-shaped, we can say that the distribution is approximately normal.

Step 2 Check for skewness. For these data, $\bar{X} = 79.5$, median = 77.5, and s = 40.5. Using the Pearson coefficient of skewness gives:

$$PC = \frac{3 \times (79.5 - 77.5)}{40.5} = 0.148$$

In this case, ***the PC is not greater than +1 or less than -1, so it can be concluded that the distribution is not significantly skewed***.

Step 3[1] Check for outliers. Recall that an outlier is a data value that lies more than 1.5(IQR) units below Q_1 or 1.5(IQR) units above Q_3.

In this case,

$Q_1 = 45$ and $Q_3 = 98$

[1] In exams and quizzes, this step can be skipped to save time.

Hence,
IQR $= Q_3 - Q_1 =$ 98 - 45 $=$ 53
An outlier would be:
A data value less than
45 - 1.5(53) =34.5
Or a data value larger than
98 + 1.5(53) = 177.5
In this case, there are no outliers.
Conclusion:
Since:

- The histogram is approximately bell-shaped,
- The data are not significantly skewed,
- And there are no outliers,

It can be concluded that the distribution is approximately normally distributed.

6.5.3 Homework Problems

Home Work 6.5-1

The compressive strength of concrete is being tested by a civil engineer. He tests ***12 specimens*** and obtains the following data:

Mean, $\bar{X}$	2260psi
Median	2259psi
Standard deviation, s	35.6 psi

Based on Person Coefficient,

$$PC = \frac{\left(3(\bar{X} - median)\right)}{s}$$

is there evidence to support the assumption that compressive strength is normally distributed?

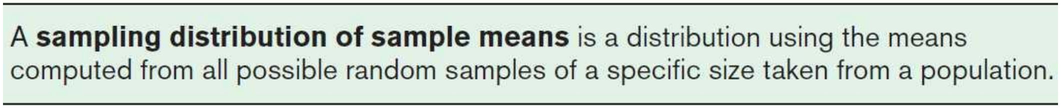
Ans. The data almost obeys normal distribution model.

6.6 The Central Limit Theorem*

In addition to knowing how individual data values vary about the mean for a population, statisticians are interested in knowing how the means of samples of the same size taken from the same population vary about the population mean.

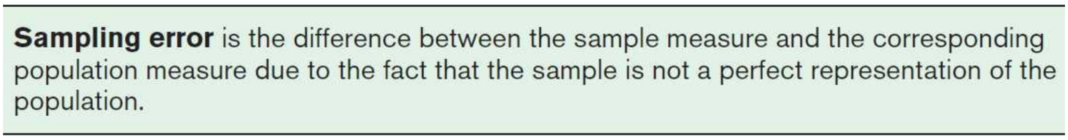
6.6.1 Distribution of Sample Means

- Suppose a researcher selects a sample of 30 adult males and finds the mean of the measure of the triglyceride levels for the sample subjects to be 187 milligrams/deciliter.
- Then suppose a second sample is selected, and the mean of that sample is found to be 192 milligrams/deciliter.
- Continue the process for 100 samples. What happens then is that the mean becomes a random variable, and the sample means 187, 192, 184, . . . , 196 constitute a ***sampling distribution of sample means***.



A **sampling distribution of sample means** is a distribution using the means computed from all possible random samples of a specific size taken from a population.

- If the samples are randomly selected with replacement, the sample means, for the most part, will be somewhat different from the population mean μ. These differences are caused by sampling error.



Sampling error is the difference between the sample measure and the corresponding population measure due to the fact that the sample is not a perfect representation of the population.

6.6.2 Properties of the Distribution of Sample Means

When all possible samples of a specific size are selected with replacement from a population, the distribution of the sample means for a variable has Three important properties:

- 1st Property that Related to Mean:
  The mean of the sample means will be the same as the population mean.
- 2nd Property that Related to Standard Devotion:
  The standard deviation of the sample means will be smaller than the standard deviation of the population, and it will be equal to the population standard deviation divided by the square root of the sample size.
- 3rd Property that Related to Distribution:
  According to ***Central Limit Theorem***: ***As the sample size n increases without limit, the shape of the distribution of the sample means taken with replacement from a population with mean*** μ ***and standard deviation*** σ ***will approach a normal distribution***.

6.6.3 Illustration of Properties of Distribution of Sample Means

- The following example illustrates above two properties:
- Suppose a professor gave an 8-point quiz to a small class of four students. The results of the quiz were 2, 6, 4, and 8. For the sake of discussion, assume that the four students constitute the population.
  - The mean of the population is:
    $$\mu = \frac{2+6+4+8}{4} = 5$$
  - The standard deviation of the population is:
    $$\sigma = \sqrt{\frac{(2-5)^2 + (6-5)^2 + (4-5)^2 + (8-5)^2}{4}} = 2.236$$
  - The graph of the original distribution is shown in ***Figure 6.6-1*** below. This is called a uniform distribution.

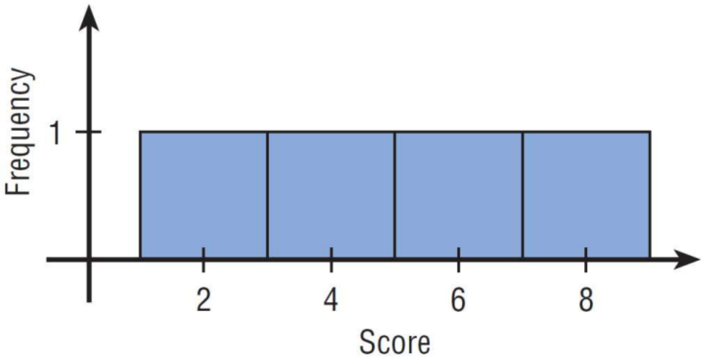


Figure 6.6-1: Original distribution for an eight-point quiz for a class of four student.

- Now, if all samples of size 2 are taken with replacement and the mean of each sample is found, the distribution is as shown.

Table 6.6-1: Sixteen possible samples with size 2 to be taken from a class of four student.

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

- A frequency distribution of sample means is as follows.

Table 6.6-2: Frequency distribution for sixteen possible samples with size 2 to be taken from a class of four student.

$\overline{X}$	f
2	1
3	2
4	3
5	4
6	3
7	2
8	1

- For the data from the example just discussed, ***Figure 6.6-2*** below shows the graph of the sample means. The histogram appears to be approximately normal. The mean of the sample means, denoted by $\mu_{\bar{x}}$ is:

$$\mu_{\overline{X}} = \frac{2 + 3 + \cdots + 8}{16} = \frac{80}{16} = 5$$

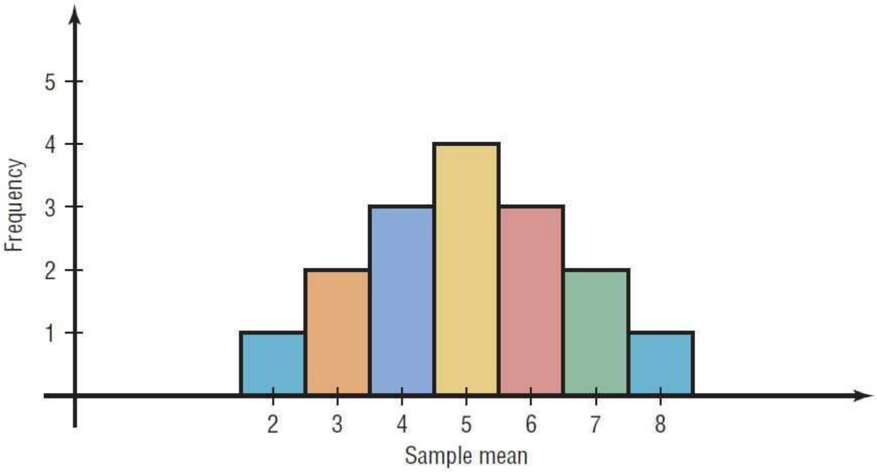


Figure 6.6-2: Histogram for sixteen possible samples with size 2 to be taken from a class of four student.

- Based on first property "***The mean of the sample means will be the same as the population mean***":
  $\mu_{\bar{X}} = \mu$ ■

- Based on second property "***The standard deviation of the sample means is equal to the population standard deviation divided by the square root of the sample size***":

  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ ■

  $$\sigma_{\bar{X}} = \frac{2.236}{\sqrt{2}} = 1.581$$

6.6.4 Numerical Examples Related to Central Limit Theorem

Example 6.6-1

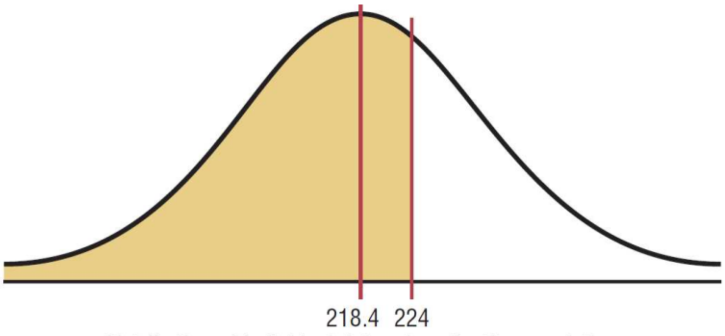
The average number of pounds of meat that a person consumes per year is 218.4 pounds.

Assume that the standard deviation is 25 pounds and the distribution is approximately normal.

- Find the probability that a person selected at random consumes less than 224 pounds per year.
- If a sample of 40 individuals is selected, find the probability that the mean of the sample will be less than 224 pounds per year.

Solution

- The probability that a person selected at random consumes less than 224 pounds per year:
  - Since the question asks about an individual person and the question not related to a sample, then we will use the distribution original random variable that shown.



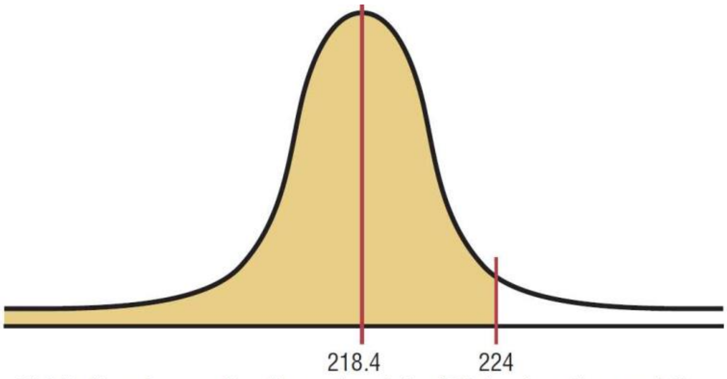
Distribution of individual data values for the population

  - The specified value of variable (consuming of 224 lb of meat) should be transformed to corresponding z value based on following relation:

    $$z = \frac{X-\mu}{\sigma} = \frac{224-218.4}{25} = 0.224$$

  - The area to the left of z < 0.224 is 0.5871.
  - Hence, the probability of selecting an individual who consumes less than 224 pounds of meat per year is 0.5871, or 58.71% or

    $P(X < 224) = 0.5871$ ■



Distribution of means for all samples of size 40 taken from the population

- The probability that the mean of the sample will be less than 224 pounds per year:
  - Since the question concerns the mean of a sample with a size of 40, then one should use distribution for sample means that is normal (based on central limit theorem) and has mean of:

    $\mu_{\bar{x}} = \mu = 218.4\ lb$

    and a standard deviation of:

    $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{40}} = 3.95$$

  - Then, the distribution of sample means will be:
  - It is useful to note that, above distribution is narrower that distribution of meat consuming as it has lower standard deviation.

- The specified value of sample mean should be transformed into related z value based on following relation:

  $$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{224 - 218.4}{25/\sqrt{40}} = 1.42$$

- The area to the left of z = 1.42 is 0.9222.
- Hence, the probability that the mean of a sample of 40 individuals is less than 224 pounds per year is 0.9222, or 92.22%. That is,

  $$P(\bar{X} < 224) = 0.9222$$

- Comparing the two probabilities:
  - You can see that the probability of selecting an individual who consumes less than 224 pounds of meat per year is 58.71%.
  - However, the probability of selecting a sample of 40 people with a mean consumption of meat that is less than 224 pounds per year is 92.22%.
  - ***This rather large difference is due to the fact that the distribution of sample means is much less variable than the distribution of individual data values.***

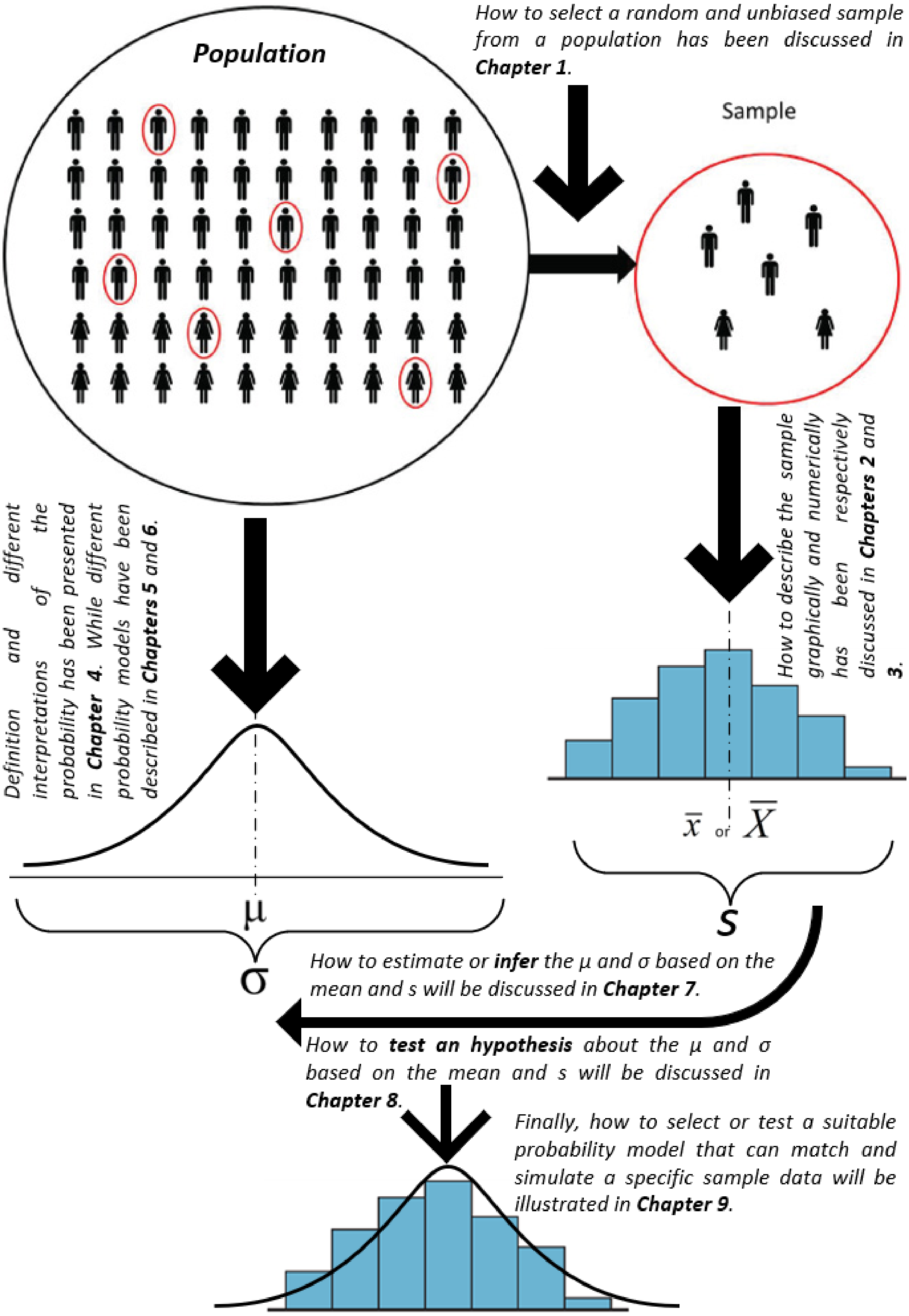
Contents

CHAPTER 7

CONFIDENCE INTERVALS

7.1 INTRODUCTION

- As this chapter is the first one in the inferential statistical aspects, therefore it is useful to summarize what has been achieved in the previous chapters and what is remaining to this and the next chapters of the univariate analysis.
- This summary has been presented in ***Figure 7.1-1***.



*How to select a random and unbiased sample from a population has been discussed in **Chapter 1**.*

*How to describe the sample graphically and numerically has been respectively discussed in **Chapters 2** and **3**.*

*Definition and different interpretations of the probability has been presented in **Chapter 4**. While different probability models have been described in **Chapters 5** and **6**.*

*How to estimate or **infer** the μ and σ based on the mean and s will be discussed in **Chapter 7**.*

*How to **test an hypothesis** about the μ and σ based on the mean and s will be discussed in **Chapter 8**.*

*Finally, how to select or test a suitable probability model that can match and simulate a specific sample data will be illustrated in **Chapter 9**.*

Figure 7.1-1: Summary of what has been achieved and what is remaining in the univariate analysis.

- One aspect of inferential statistics is the ***estimation***, ***which is the process of guessing the value of a parameter from information obtained from a sample***.
- Since the ***populations from which these values were obtained are large***, ***these values are only estimates*** of the true parameters and are derived from data collected from samples.
- The statistical procedures for estimating
  - Population mean,
  - Proportion,
  - Variance, and Standard deviation,
  - Goodness of fit

  will be explained in this chapter.
- An important question in estimation is that of ***sample size***. ***How large should the sample be in order to make an accurate estimate***? The question of sample size will be explained in this chapter also.
- Basic assumption for the inferential techniques:

  Inferential statistical techniques have various ***assumptions that must be met before valid conclusions can be obtained***:
  - One common assumption is that the samples must be ***randomly selected***. ***Chapter 1*** explains how to obtain a random sample.
  - The other common assumption is that ***either the sample size must be greater than or equal to 30*** or ***the population must be normally or approximately normally distributed if the sample size is less than 30***. To check this assumption, you can use the methods explained in ***Chapter 6***.

7.2 Mean Estimation

7.2.1 Basic Estimation Concepts

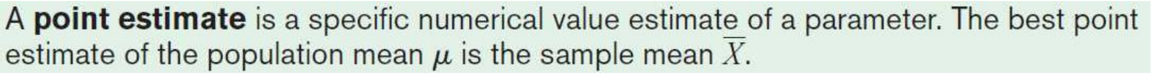
7.2.1.1 Essence of Problem through an Example

Suppose a college president wishes to estimate the average age of students attending classes this semester.

- The president could select a random sample of 100 students and find the average age of these students, say, 22.3 years (he has measured a statistic $\bar{x}$).
- From the sample mean, the president could infer the average age of all the students (he intends to transfer for the statistic $\bar{x}$ to the parameter μ). This could be done based on one of following two approaches of ***Sections 7.2.1.2*** and ***7.2.1.3***.

7.2.1.2 Point Estimate

- Point estimate could be defined as:

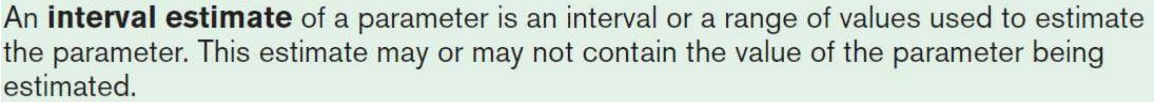


A **point estimate** is a specific numerical value estimate of a parameter. The best point estimate of the population mean μ is the sample mean $\overline{X}$.

- For above example, point estimate of μ could be written as:
  $Average\ age\ of\ all\ students\ (\mu) \approx\ Average\ age\ of\ students\ sample\ (\bar{x})$
- Why mean is used as an estimator and what about other measures of central tendency?
  You might ask why other measures of central tendency, such as the median and mode, are not used to estimate the population mean.
  - The reason is that the means of samples ***vary less than other statistics*** (such as medians and modes) ***when many samples are selected from the same population***.
  - Therefore, the sample mean is the best estimate of the population mean.

7.2.1.3 Interval Estimate

- The sample mean will be, for the most part, somewhat different from the population mean due to sampling error.
- Therefore, you might ask a second question:
  How good is a point estimate?
  - The answer is that there is no way of knowing how close a particular point estimate is to the population mean.
  - This answer ***places some doubt*** on the accuracy of point estimates.
- For this reason, ***statisticians prefer*** another type of estimate, called an ***interval estimate***, which could be defined as follows:



An **interval estimate** of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.

- For above example:
  In an interval estimate, the parameter is specified as being between two values. For example, an interval estimate for the average age of all students might be
  $21.9\ \leq\ \mu\ \ \leq\ 22.7$ or $22.3\ \pm\ 0.4\ years$

7.2.1.4 Confidence Interval

- Either the interval contains the parameter, or it does not. A degree of confidence (usually a percent) can be assigned before an interval estimate is made.

- For instance, you may wish to be 95% confident that the interval contains the true population mean.

7.2.1.5 How to Determine the Confidence Interval

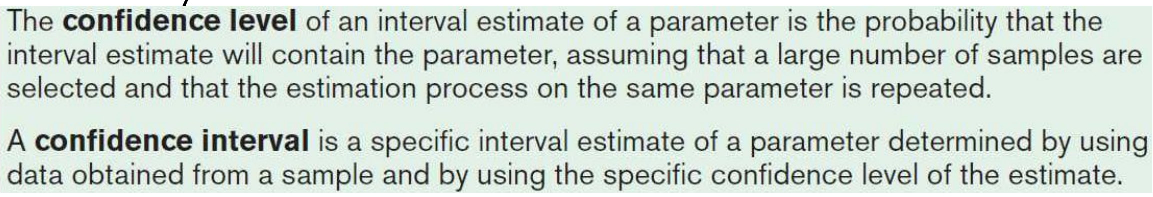
- Another question then arises. Why 95%? Why not 99 or 99.5%?
- If you desire to be ***more confident***, such as 99 or 99.5% confident, then you must make the ***interval larger***.
- For example, a 99% confidence interval for the mean age of college students might be:

  $21.7 \leq \mu \leq 22.9$

  or

  22.3 ± 0.6

- In summary:



The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.

A **confidence interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.

7.2.2 Confidence Intervals for the Mean when σ is Known

- The central limit theorem states that when the sample size is large ($n \geq 30$), interval estimation for mean will be:

$$\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \qquad \textbf{Eq. 7.2-1}$$

  where:
  - In the symbol $z_{\frac{\alpha}{2}}$ (read "zee sub alpha over two"), the Greek letter α (alpha) represents the total area in both tails of the standard normal distribution curve, and $\frac{\alpha}{2}$ represents the area in each one of the tails.
  - The term $z_{\alpha/2}(\sigma/\sqrt{n})$ is called the margin of error (also called the maximum error of the estimate).
- For a specific value, say, $\alpha = 0.05$, 95% of the sample means will fall within this error value on either side of the population mean. See ***Figure 7.2-1*** below.
- Common confidence intervals:
  - Confidence interval of 90%, $z_{\frac{\alpha}{2}} = 1.65$
  - Confidence interval of 95%, $z_{\frac{\alpha}{2}} = 1.96$
  - Confidence interval of 99%, $z_{\frac{\alpha}{2}} = 2.58$

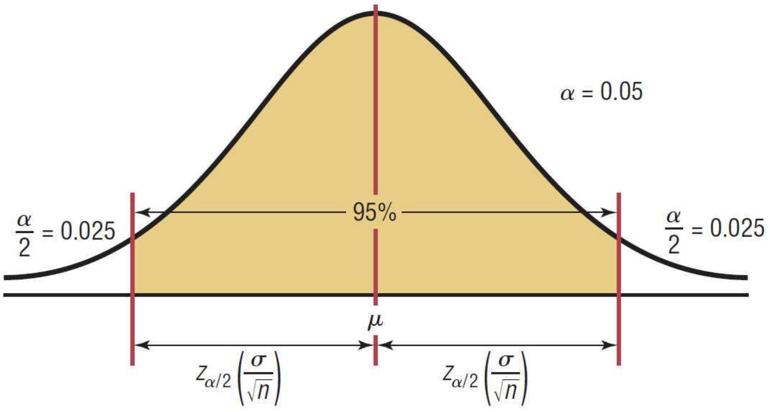


Figure 7.2-1Confidence interval of 95%.

Example 7.2-1

A researcher wishes to estimate the number of days it takes an automobile dealer to sell a Chevrolet Aveo. A sample of 50 cars had a mean time on the dealer's lot of 54 days. Assume the population standard deviation to be 6.0 days. Find the best point estimate of the population mean and the 95% confidence interval of the population mean.

Solution

The best point estimate of the mean is:

$\mu = \bar{X} = 54\ days$

For the 95% confidence interval use z = 1.96:

$$\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Rightarrow 54 - 1.96\frac{6.0}{\sqrt{50}} \le \mu \le 54 + 1.96\frac{6.0}{\sqrt{50}} \Rightarrow 52.3 \le \mu \le 55.7$$

Hence, one can say with 95% confidence that the interval between 52.3 and 55.7 days does contain the population mean, based on a sample of 50 automobiles.

Example 7.2-2

A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes. Assuming that the population standard deviation is 46.5 minutes, find the best point estimate of the population mean and the 99% confidence of the population mean.

Solution

The best point estimate is:

$\mu = \bar{X} = 174.3\ minutes$

The 99% confidence is interval is:

$$\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Rightarrow 174.3 - 2.58 \times \frac{46.5}{\sqrt{30}} \le \mu \le 174.3 + 2.58 \times \frac{46.5}{\sqrt{30}} \Rightarrow 152.4 \le \mu \le 196.2$$

Hence, one can be 99% confident that the mean waiting time for emergency room treatment is between 152.4 and 196.2 minutes.

Important Notes

- When the ***original variable is normally distributed*** and σ ***is known***, the ***standard normal distribution*** can be used to find confidence intervals ***regardless of the size of the sample***.
- When $n \ge 30$, ***the distribution of means will be approximately normal even if the original distribution of the variable departs from normality***.
- When σ ***is unknown***, s ***can be used as an estimate of*** σ, and t ***distribution should be used*** as discussed in ***Section 7.2.4***.

Example 7.2-3

ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at 60ºC are as follows:

64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3.

Assume that impact energy is normally distributed with $\sigma = 1J$. We want to find a 95% confidence interval for μ, the mean impact energy.

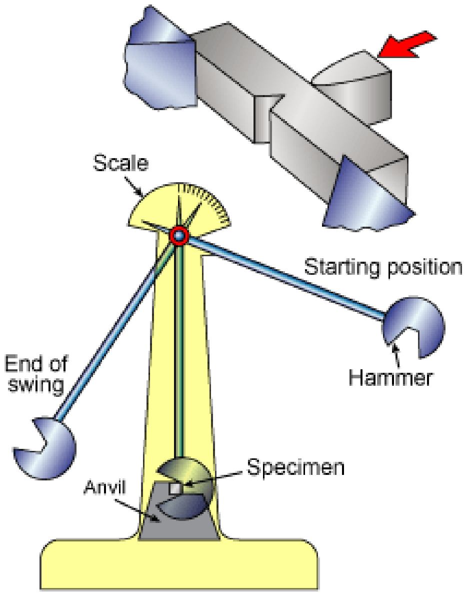


Figure 7.2-2: Charpy test for Example 7.2-3.

Solution

As the original variable is normally distributed and is σ known, the standard normal distribution can be used to find confidence intervals regardless of the size of the sample.

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$64.46 - 1.96 \times \frac{1}{\sqrt{10}} \leq \mu \leq 64.46 + 1.96 \times \frac{1}{\sqrt{10}}$$

$$63.84 \leq \mu \leq 65.08$$

That is, based on the sample data, a range of highly reasonable values for mean impact energy for A238 steel at 60°C is:

$$63.84J \leq \mu \leq 65.08J$$

Example 7.2-4

An article in the 1993 volume of the Transactions of the American Fisheries Society reports the results of a study to investigate the mercury contamination in largemouth bass.

A sample of fish was selected from 53 Florida lakes and mercury concentration in the muscle tissue was measured (ppm). The mercury concentration values are:

1.230	0.490	0.490	1.080	0.590	0.280	0.180	0.100	0.940
1.330	0.190	1.160	0.980	0.340	0.340	0.190	0.210	0.400
0.040	0.830	0.050	0.630	0.340	0.750	0.040	0.860	0.430
0.044	0.810	0.150	0.560	0.840	0.870	0.490	0.520	0.250
1.200	0.710	0.190	0.410	0.500	0.560	1.100	0.650	0.270
0.270	0.500	0.770	0.730	0.340	0.170	0.160	0.270	

Above data has been summarized in numerical form below:

$n = 53, \bar{X} = 0.5250, s = 0.3486$

and in graphical form as shown in ***Figure 7.2-3*** below.

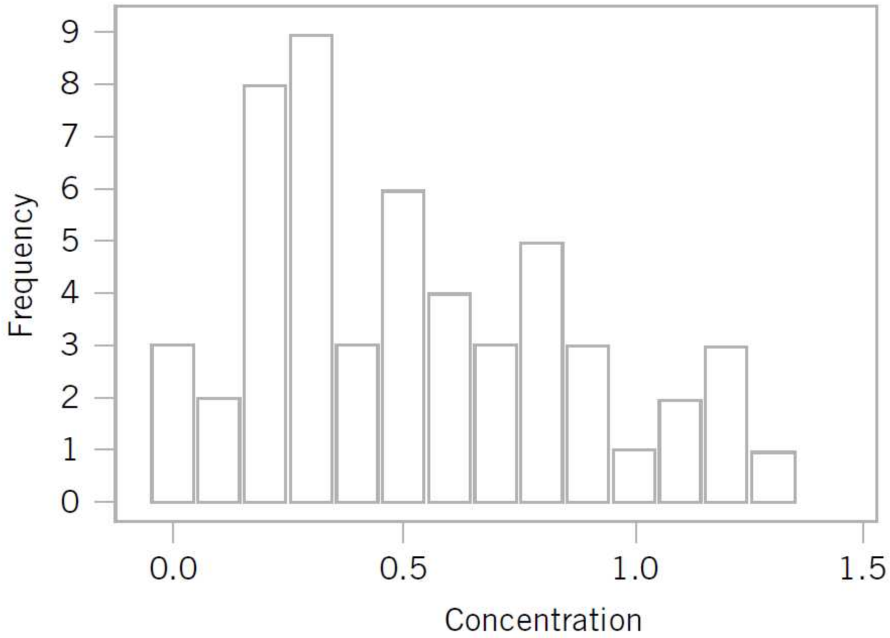


Figure 7.2-3: Histogram for mercury contamination concentration of Example 7.2-4.

Find an approximate 95% confidence interval on μ.

Solution

Histogram plot indicates that the distribution of mercury concentration is not normal and is positively skewed.

Because $n \geq 30$, the assumption of normality is not. The required quantities are: $n = 53, \bar{X} = 0.5250, s = 0.3486$ and $z_{0.025} = 1.96$

The approximate 95% confidence interval on μ is:

$$\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Rightarrow 0.5250 - 1.96\frac{0.3486}{\sqrt{53}} \leq \mu \leq 0.5250 + 1.96\frac{0.3486}{\sqrt{53}}$$

$$0.4311 \leq \mu \leq 0.6189$$

This interval is fairly wide because there is a lot of variability in the mercury concentration measurements.

Example 7.2-5

In an attempt to estimate a friction coefficient to be fixed in his catalog, a ladder-manufacturer has executed 40 related experiments. Results are presented in ***Figure 7.2-4*** above. Based on these results, and with adopted a ***confidence level of 0.95***, ***what is the estimated value for friction coefficient***?

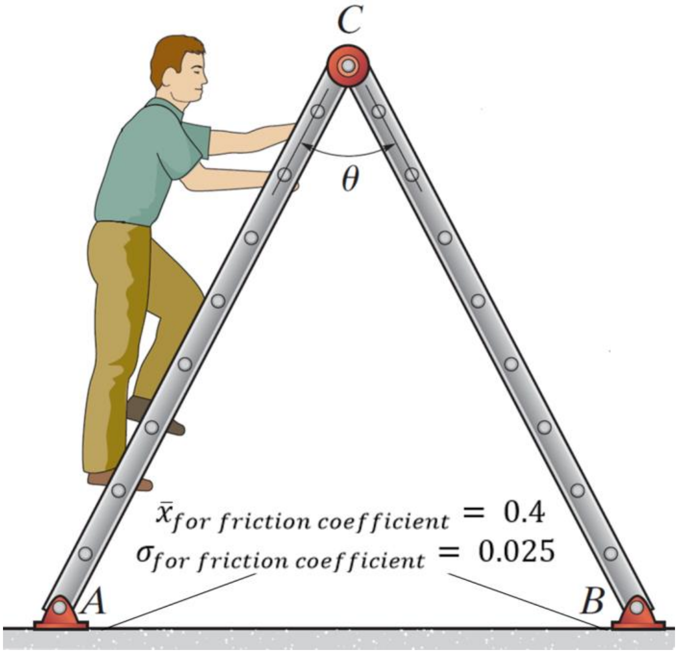

C
θ
$\bar{x}_{for\ friction\ coefficient} = 0.4$
$\sigma_{for\ friction\ coefficient} = 0.025$
A
B

Figure 7.2-4: Ladder for Example 7.2-5.

Solution

As

$n = 40 > 30$

Therefore, normal distribution shall be adopted for sample distribution even when population is not normally distributed.

With confidence level of 95%, interval estimation would be:

$$\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}$$

$$\because z_{0.025} = 1.96$$

$$0.4 - 1.96\frac{0.025}{\sqrt{40}} \leq \mu \leq 0.4 + 1.96\frac{0.025}{\sqrt{40}}$$

$$0.39 \leq \mu \leq 0.41$$

To be conservative regarding to sliding, the manufacturer should recommend a friction coefficient in the range of 0.39 in his catalog.

7.2.3 Sample Size

- Sample size determination is closely related to statistical estimation.
- Quite often one may ask how large a sample is necessary to make an accurate estimate?
- The answer is not simple, since it depends on three things:
  - The margin of error,
  - The population standard deviation,
  - The degree of confidence.
- Derivation for Sample Size Formula:
  - Basic Assumption:
    For the purpose of this chapter, it will be assumed that the population standard deviation of the variable is known or has been estimated from a previous study.
  - The formula for sample size is derived from the margin of error formula:
    $$E = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$
    and this formula is solved for n as follows:
    $$E\sqrt{n} = z_{\frac{\alpha}{2}}\,\sigma$$
    $$\sqrt{n} = \frac{z_{\frac{\alpha}{2}}\,\sigma}{E}$$
    Hence,
    $$n = \left(\frac{z_{\frac{\alpha}{2}}\,\sigma}{E}\right)^2 \quad \blacksquare \qquad \textbf{Eq. 7.2-2}$$
    where E is the margin of error.
  - If necessary, round the answer up to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size n.

Example 7.2-6

A scientist wishes to estimate the average depth of a river. He wants to be 99% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.33 feet.

Solution

Since

$\because \alpha = 1 - 0.99 = 0.01, \therefore z_{\frac{\alpha}{2}} = 2.58$ and $E = 2$

$$n = \left(\frac{z_{\frac{\alpha}{2}}\,\sigma}{E}\right)^2 = \left(\frac{2.58 \times 4.33}{2}\right)^2 = 31.2$$

Round the value 31.2 up to 32.

Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs at least a sample of 32 measurements.

In most cases in statistics, we round off. However, when determining sample size, we always round up to the next whole number.

7.2.4 Confidence Intervals for the Mean when σ is Unknown

7.2.4.1 Basic Concepts

- ***Most of the time***, the value of σ ***is not known***, so it must be estimated by using s, namely, the standard deviation of the sample.
- When s is used, ***especially when the sample size is small***, ***critical values greater than the values for are used in confidence intervals in order to keep the interval at a given level***, such as the 95%. These values are taken from the ***Student t distribution***, most often called the ***t distribution.***
- To use this method:
  - The samples must be simple random samples,
  - The ***population from which the samples were taken must be normally or approximately normally distributed***, or ***the sample size must be 30 or more***.
- The t distribution ***shares some characteristics of the normal distribution*** and ***differs from it in others***.
  - The t distribution is ***similar to the standard normal distribution*** in these ways:
    - It is bell-shaped.
    - It is symmetric about the mean.
    - The mean, median, and mode are equal to 0 and are located at the center of the distribution.
    - The curve never touches the x axis.
  - The t distribution ***differs from the standard normal distribution*** in the following ways:
    - The variance is greater than 1.
    - The t distribution is actually a family of curves based on the concept of degrees of freedom, which is related to sample size.
    - As the sample size increases, the t distribution approaches the standard normal distribution. See ***Figure 7.2-5*** below.

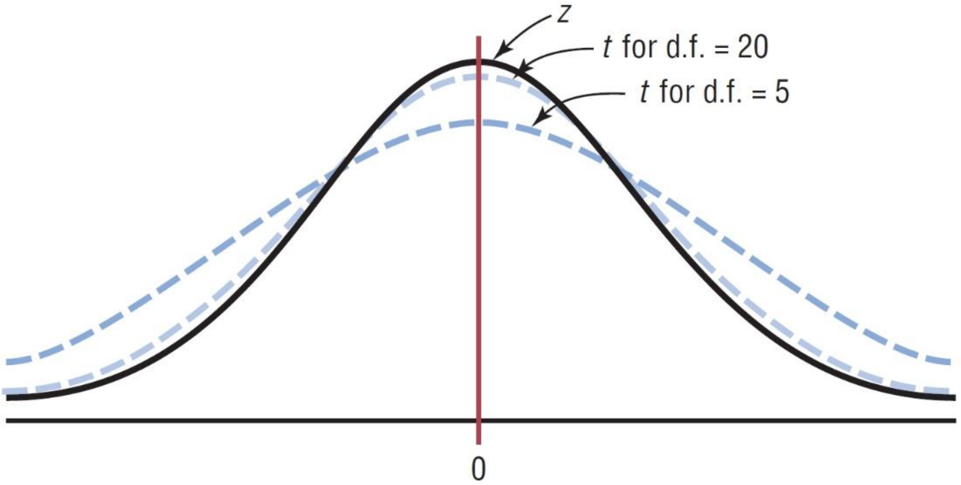


Figure 7.2-5: The t Family of Curves.

- ***Degrees of Freedom Concept:***
  - The ***degrees of freedom*** are the number of values that are free to vary after a sample statistic has been computed, and they tell the researcher which specific curve to use when a distribution consists of a family of curves.
  - For example, if the mean of 5 values is 10, then 4 of the 5 values are free to vary. But once 4 values are selected, the fifth value must be a specific number to get a sum of 50, since $50 \div$

$5 = 10$. Hence, the degrees of freedom are $5 - 1 = 4$, and this value tells the researcher which t curve to use.

- The symbol d.f. will be used for degrees of freedom. The degrees of freedom for a confidence interval for the mean are found by subtracting 1 from the sample size. That is,
  $d.f. = n - 1$
- For some statistical tests used later in this book, the degrees of freedom are not equal to n – 1.

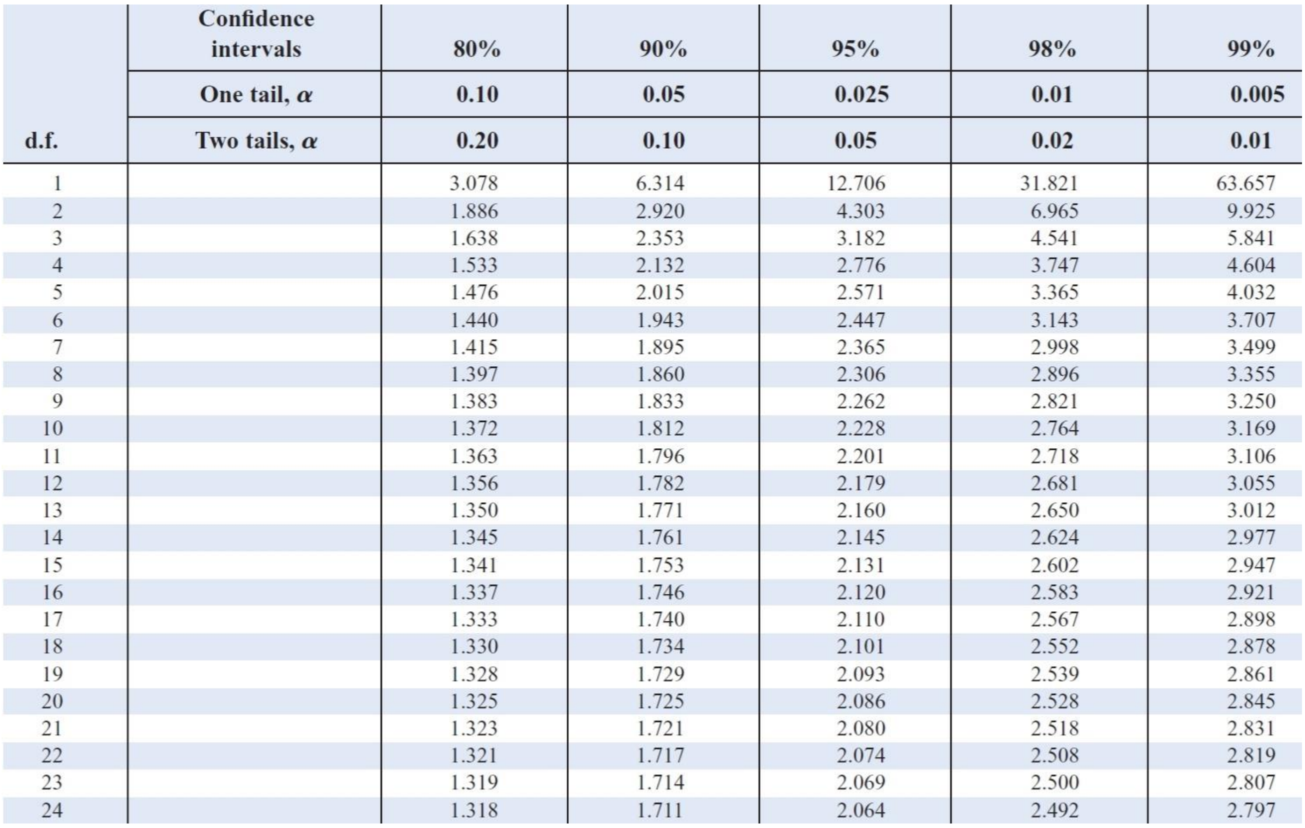
7.2.4.2 FORMULA USING t TEST

- The formula for finding a confidence interval about the mean by using the t distribution is given now:

$$\bar{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \qquad \textbf{Eq. 7.2-3}$$

- The values for $t_{\frac{\alpha}{2}}$ are found in ***Table 7.2-1*** below.

Table 7.2-1: The t Distribution.



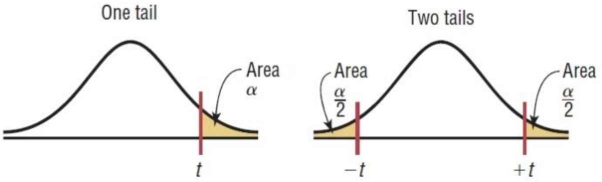
	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797

Table 7.2-1: The t Distribution. Continued

25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
34		1.307	1.691	2.032	2.441	2.728
36		1.306	1.688	2.028	2.434	2.719
38		1.304	1.686	2.024	2.429	2.712
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
55		1.297	1.673	2.004	2.396	2.668
60		1.296	1.671	2.000	2.390	2.660
65		1.295	1.669	1.997	2.385	2.654
70		1.294	1.667	1.994	2.381	2.648
75		1.293	1.665	1.992	2.377	2.643
80		1.292	1.664	1.990	2.374	2.639
90		1.291	1.662	1.987	2.368	2.632
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
(z) ∞		1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

[a] This value has been rounded to 1.28 in the textbook.
[b] This value has been rounded to 1.65 in the textbook.
[c] This value has been rounded to 2.33 in the textbook.
[d] This value has been rounded to 2.58 in the textbook.

Source: Adapted from W. H. Beyer, *Handbook of Tables for Probability and Statistics,* 2nd ed., CRC Press, Boca Raton, Fla., 1986. Reprinted with permission.

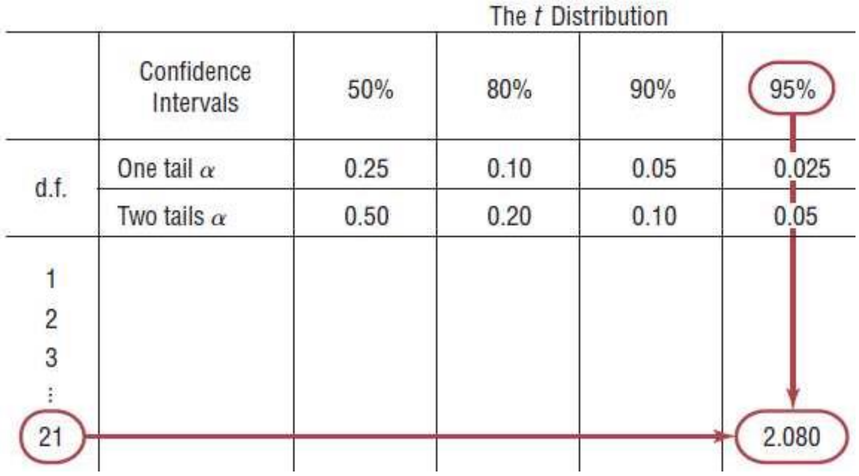


Example 7.2-7

Find the $t_{\frac{\alpha}{2}}$ value for a 95% confidence interval when the sample size is 22.

Solution

The d.f. = 22 - 1, or 21. Find 21 in the left column and 95% in the row labeled Confidence Intervals. The intersection where the two meet gives the value for $t_{\frac{\alpha}{2}}$, which is 2.080.



The *t* Distribution

	Confidence Intervals	50%	80%	90%	95%
d.f.	One tail α	0.25	0.10	0.05	0.025
	Two tails α	0.50	0.20	0.10	0.05
1					
2					
3					
⋮					
21					2.080

Note:

When d.f. is greater than 30, it may fall between two table values. For example, if d.f. it falls between 65 and 70. A conservative approach could be used. In this case, always round down to the nearest table value. In this case, 68 rounds down to 65.

Example 7.2-8

Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was $s = 0.78\, hour$. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

Solution

Since σ is unknown and s must replace it, the t distribution (***Table 7.2-1***) must be used for the confidence interval. Hence, with 9 degrees of freedom $t_{\frac{\alpha}{2}} = 2.262$.

The 95% confidence interval can be found by substituting in the formula.

$$\bar{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \Rightarrow 7.1 - \ 2.262\frac{0.78}{\sqrt{10}} \le \mu \le 7.1 + \ 2.262\frac{0.78}{\sqrt{10}}$$

$6.54 \le \mu \le 7.66$

Therefore, one can be 95% confident that the population mean is between 6.54 and 7.66 hours.

Example 7.2-9

The data represent a sample of the number of home fires started by candles for the past several years. (Data are from the National Fire Protection Association.) Find the 99% confidence interval for the mean number of home fires started by candles each year.

5460	5900	6090	6310	7160	8440	9930

Assume that the phenomenon has a normal probability density function, PDF.

Solution

Step 1 Find the mean and standard deviation for the data. Use the formulas in ***Chapter 3***.

The mean:

$$\bar{X} = \frac{\Sigma X_i}{n} = 7041.4$$

The standard deviation:

$s \ = \ 1610.3$

Step 2 Find $t_{\frac{\alpha}{2}}$ in ***Table 7.2-1***. Use the 99% confidence interval with d.f. = 6. It is 3.707.

Step 3 Substitute in the formula and solve.

$$\bar{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \Rightarrow 7041.4 - 3.707\frac{1610.3}{\sqrt{7}} \le \mu \le 7041.4 + 3.707\frac{1610.3}{\sqrt{7}}$$

$4785.2 \le \mu \le 9297.6$

One can be 99% confident that the population mean number of home fires started by candles each year is between 4785.2 and 9297.6, based on a sample of home fires occurring over a period of 7 years.

Example 7.2-10

An article in the journal Materials Engineering (1989, Vol. II, No. 4, pp. 275–281) describes the results of tensile adhesion tests on 22 U-700 alloy specimens. The load at specimen failure is as follows (in MPa):

10.1	14.9	7.5	15.4	15.4	18.5	7.9	12.7	11.9	19.8	15.4
11.4	14.1	17.6	16.7	15.8	8.8	13.6	11.9	11.4	11.4	19.5

The sample mean is $\bar{X} = 13.71$, and the sample standard deviation is $s \ = \ 3.55$. Assuming a normal probability find a 95% confidence interval on μ.

Solution

Since $n = 22$, we have n - 1 = 21 degrees of freedom for $t_{\frac{\alpha}{2}}$,

$t_{\frac{\alpha}{2}} = 2.080$

The resulting confidence interval is:

$$\bar{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \Longrightarrow 13.71 - 2.080\frac{3.55}{\sqrt{22}} \leq \mu \leq 13.71 + 2.080\frac{3.55}{\sqrt{22}}$$

$12.14 \leq \mu \leq 15.28$

7.2.5 Summary of Using z and t Distributions

Using of z and t distribution have been summarized in ***Figure 7.2-6*** below.

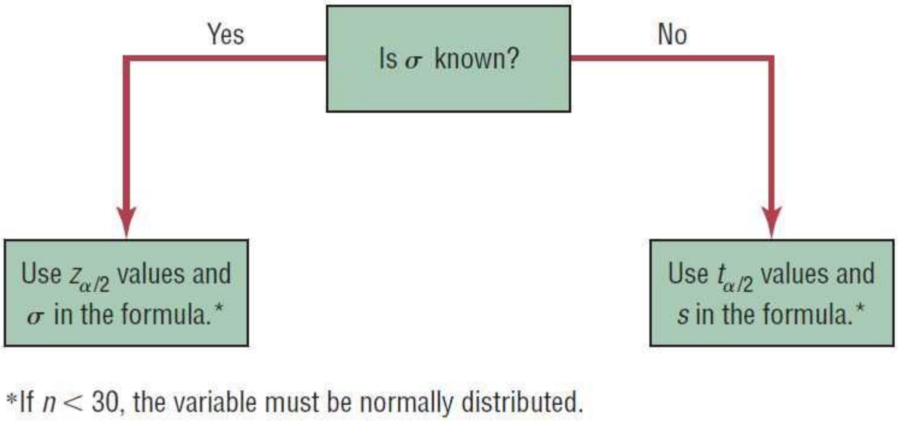

Is σ known?
Yes
No
Use $z_{\alpha/2}$ values and σ in the formula.*
Use $t_{\alpha/2}$ values and s in the formula.*
*If n < 30, the variable must be normally distributed.

Figure 7.2-6: When to use the z or t distribution.

7.2.6 Homework Problems

Home Work 7.2-1

Suppose the annual maximum stream flow of a given river has been observed for 10 years yielding the following statistics:

$x = 10000\,cfs \;\; s^2 = 9 \times 10^6\,(cfs)^2$

(a) Establish the two-sided 90% confidence interval on the mean annual maximum stream flow. Assume a normal population. **Ans.** (8261, 11739)

(b) If it is desired to estimate the mean annual maximum stream flow to within $\pm 1000\,cfs$ with 90% confidence, how many additional years of observation will be required? Assume the variance, σ^2, based on the new set of data will be approximately $9 \times\ 10^6\,(cfs)^2$. (**Ans.** 14)

Home Work 7.2-2

Test No.	Actual Capacity A	Predicted Capacity P	N =A/P
1	20.5	13.6	1.51
2	18.5	20.4	0.91
3	10.0	8.8	1.14
4	15.3	14.3	1.07
5	26.2	22.8	1.15
		$\bar{x}\ of\ N$	1.154
		$s\ of\ N$	0.22

As indicated in the table, five piles have been tested until failure; the load measured at failure denotes the actual capacity of the given pile. Observe that the capacity of each pile has also been predicted by a theoretical model as indicated in the table. The factor N is simply the ratio of the actual pile capacity to the predicted pile capacity; i.e., N = A/P.

(a) Determine the 95% confidence interval of the mean value of N. **Ans.** (0.881, 1.427).

(b) In order to estimate the mean value of N to ± 0.02 with 90% confidence, how many additional piles should be tested? Assume that the variance of N, σ^2, is known and equal to 0.045 for this part. (**Ans.** 300).

Home Work 7.2-3

Concrete placed on a structure was subsequently cored after 28 days, and the following results were obtained of the compressive strengths from five test specimens:

4142 3405 3402 4039 3373 psi.

Determine the 90% two-sided confidence interval of the mean concrete strength. (**Ans.** $x = 3672\,psi\ \ s = 383.8\,psi$ and $3306 \le \mu \le 4038$.)

Home Work 7.2-4

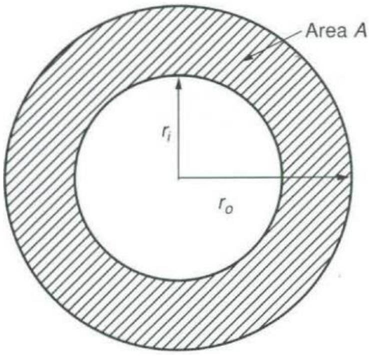
At a weigh station, the weights of trailer trucks were observed before crossing a highway bridge.

(a) Suppose observations on 30 trucks yielded a sample mean of 12.5 tons. Assume that the standard deviation of truck weights is known to be 3 tons. Determine the two-sided 99% confidence intervals of the mean weight of trailer trucks on the particular highway. **Ans.** $11.1 \le \mu \le 13.9\ tons$

(b) In part (a), how many additional trucks should be observed such that the mean truck weight can be estimated to within ± 1.0 ton with 99% confidence? (**Ans.** 30 additional trucks.)

Home Work 7.2-5

Each of the inner and outer radii of a circular ring, as shown in the following figure was measured five times with the following readings in cm:



						$\bar{x}$	s
Outer radius, r_o, cm	2.5	2.4	2.6	2.6	2.4	2.5	0.10
Inner radius, r_i, cm	1.6	1.5	1.6	1.4	1.4	1.5	0.10
Area, cm^2	11.59	11.02	13.19	15.07	11.93	12.56	1.61

With 90% confidence, determine the standard error of the estimated area. (**Ans.** $1.53\ cm^2$)

7.3 Confidence Intervals for Proportions*

7.3.1 Basic Definition and Notations

- A ***proportion represents a part of a whole***. It can be expressed as a fraction, decimal, or percentage. In this case, $12\% = 0.12 = 12/100$ or $3/25$.
- Proportions can be obtained from samples or populations. The following symbols will be used.
  p is population proportion,
  $\hat{p}$ is sample proportion,
  For a sample proportion,
  $$\hat{p} = \frac{X}{n}$$
  $$\hat{q} = \frac{n - X}{n} = 1 - \hat{p}$$
  where X is number of sample units that possess the characteristics of interest,
  n is sample size.

Example 7.3-1

In a recent survey of 150 households, 54 had central air conditioning. Find $\hat{p}$ and $\hat{q}$, where $\hat{p}$ is the proportion of households that have central air conditioning.

Solution

$$\hat{p} = \frac{X}{n} = \frac{54}{150} = 0.36$$
$$\hat{q} = \frac{n - X}{n} = \frac{150 - 54}{150} = 0.64$$

7.3.2 Estimation of Proportion

- As with means, the statistician, ***given the sample proportion, tries to estimate the population proportion***.
- ***Point*** and ***interval estimates*** for a population proportion can be made by using the sample proportion.

7.3.2.1 Point Estimation

- For a point estimate of p (the population proportion), (the sample proportion) $\hat{p}$ is used.
  $p \approx \hat{p}$
- On the basis of the three properties of a good estimator, is $\hat{p}$ ***unbiased***, ***consistent***, and ***relatively efficient***.
- But as with means, one is not able to decide how good the point estimate of p is. Therefore, statisticians ***also use an interval estimate for a proportion***, and ***they can assign a probability that the interval will contain the population proportion***.

7.3.2.2 Confidence Intervals

- To construct a confidence interval about a proportion, you must use the margin of error, which is:
  $$E = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$
- Confidence intervals about proportions must meet the criteria that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.
- Formula for a Specific Confidence Interval for a Proportion:

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

- Assumptions for Finding a Confidence Interval for a Population Proportion:
  - The sample is a random sample.
  - The conditions for a binomial experiment are satisfied (See ***Chapter 5***).

Example 7.3-2

A survey of leaks from water pipes in a city water distribution system, conducted over a representative area, shows that substantial loss occurs in 7 out of 37 pipes tested.

Find population proportion of leaking pipes in the city with 95% confidence limits.

Solution

Sample proportions are:

$$\hat{p} = \frac{X}{n} = \frac{7}{37} = 0.189$$

$$\hat{q} = \frac{n - X}{n} = \frac{37 - 7}{37} = 0.811$$

Estimating of population proportion:

$\because n\hat{p} = 30 \times 0.189 = 5.67 > 5$

$\because n\hat{q} = 30 \times 0.811 = 24.33 > 5$

Therefore, the relation below can be adopted to estimate the population proportion:

$$\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence interval of 95%, $z_{\frac{\alpha}{2}} = 1.96$,

$$0.189 - 1.96\sqrt{\frac{0.189 \times 0.811}{37}} \leq p \leq 0.189 + 1.96\sqrt{\frac{0.189 \times 0.811}{37}}$$

$$0.063 \leq p \leq 0.315$$

That is, we can say with 95% confidence that the interval (0.063, 0.315) includes the true proportion of pipes in the city from which there is a substantial waste.

7.4 Confidence Intervals for Variances and Standard Deviations*

7.4.1 Basic Concepts and Chi-square Distribution

- This section will explain how to find ***confidence intervals for variances and standard deviations***.
- To calculate these confidence intervals, a new statistical distribution is needed. It is called the ***chi-square distribution***.
- The ***chi-square variable is similar to the t variable in that its distribution is a family of curves based on the number of degrees of freedom***.
- The symbol for chi-square is χ^2 (Greek letter chi, pronounced "ki").
- Several of the distributions are shown in ***Figure 7.4-1*** below, along with the corresponding degrees of freedom.
- The chi-square distribution is obtained from the values of

  $$\frac{(n-1)s^2}{\sigma^2}$$

  when random samples are selected from a normally distributed population whose variance is σ^2.
- Notes on chi-square variable:
  - A chi-square variable cannot be negative, and the distributions are skewed to the right.
  - At about 100 degrees of freedom, the chi-square distribution becomes somewhat symmetric.
  - The area under each chi-square distribution is equal to 1.00, or 100%.

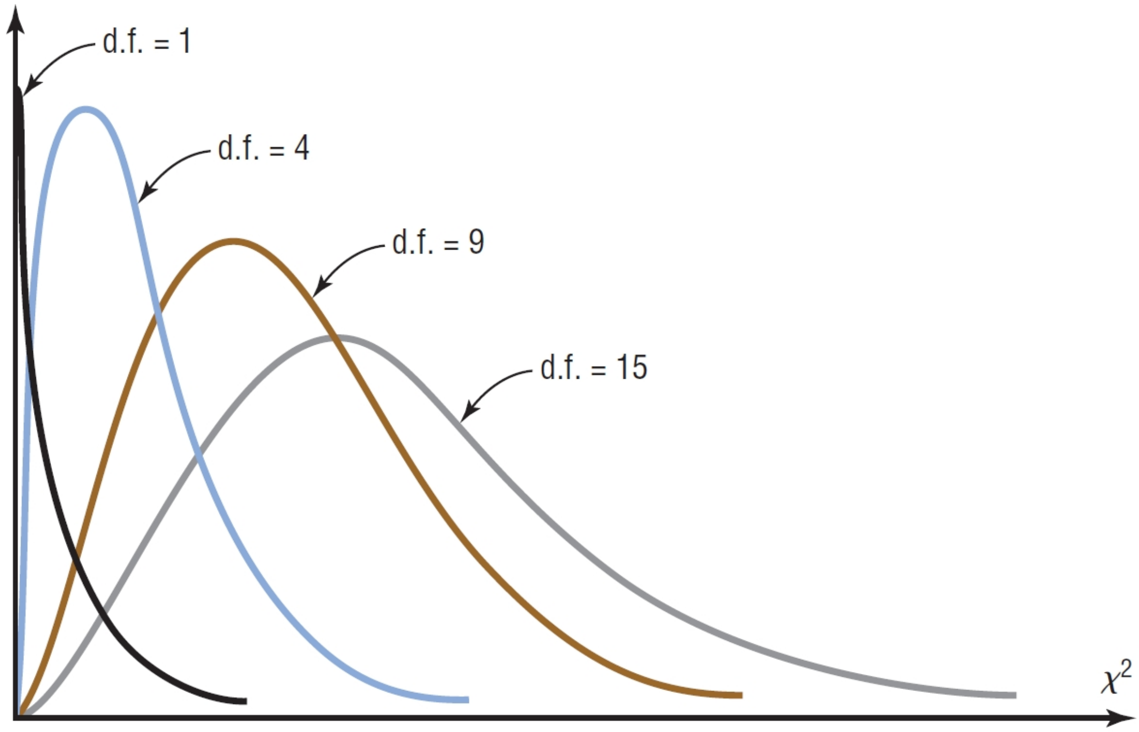


Figure 7.4-1: The Chi-Square Family of Curves.

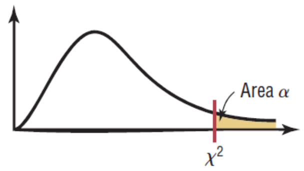
7.4.2 Table for the Chi-square Distribution

- ***Table 7.4-1*** below gives the values for the chi-square distribution. These values are used in the denominators of the formulas for confidence intervals.

Table 7.4-1: The Chi-Square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Source: Owen, *Handbook of Statistical Tables,* Table A–4 "Chi-Square Distribution Table," © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.



- Two different values are used in the formula because the distribution is not symmetric. One value is found on the left side of the table, and the other is on the right, see ***Figure 7.4-2*** below.

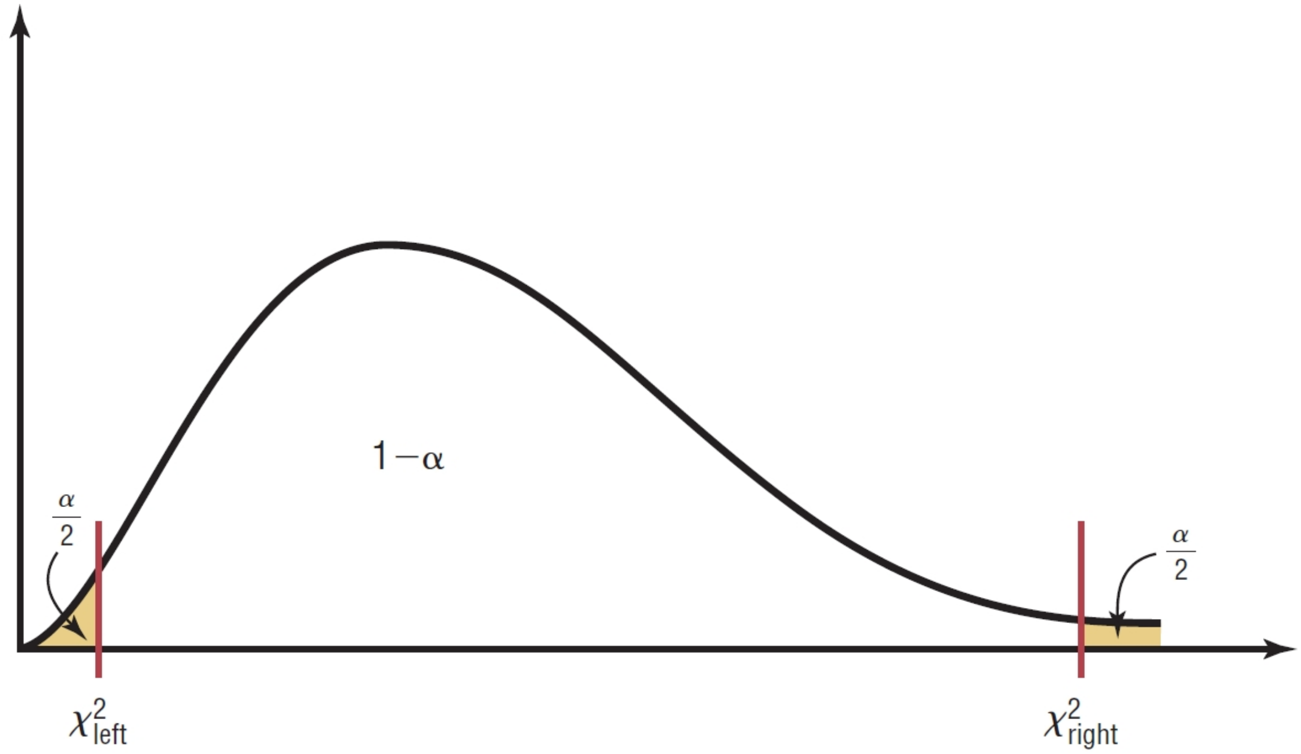


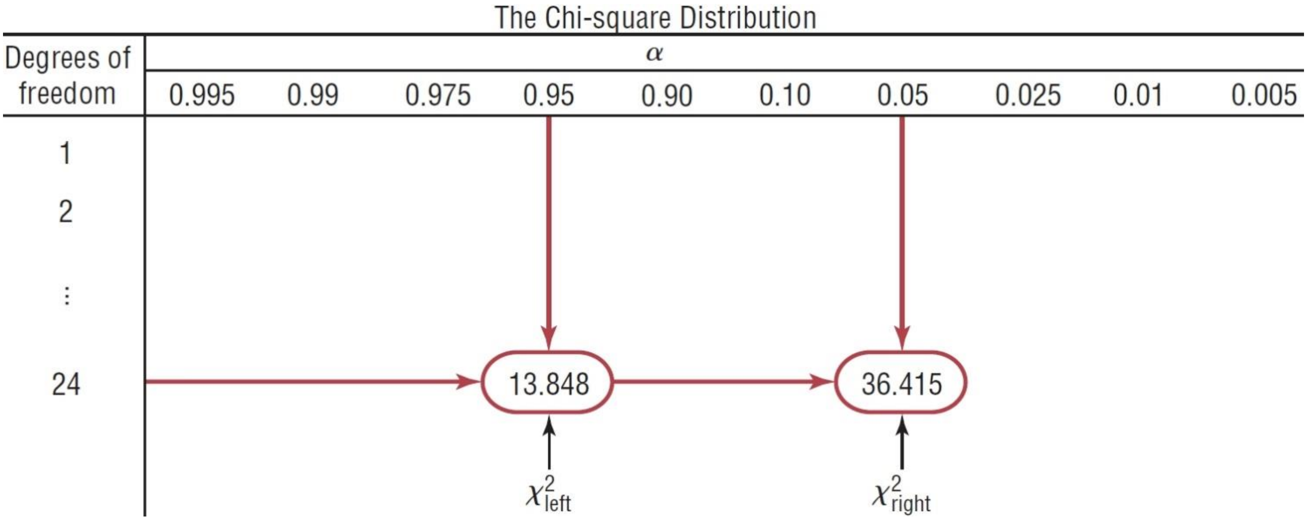
Figure 7.4-2: Chi-Square Distribution for d.f. = n – 1.

Example 7.4-1

Find the values for χ^2_{right} and χ^2_{left} =or a 90% confidence interval when n = 25.

Solution

- To find χ^2_{right}, subtract 1 - 0.90 = 0.10 and divide by 2 to get 0.05.
- To find χ^2_{left}, subtract 1 - 0.05 to get 0.95.
- Hence, use the 0.95 and 0.05 columns and the row corresponding to 24 d.f. See ***Figure 7.4-3***.


The Chi-square Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1										
2										
⋮										
24				13.848			36.415			

χ^2_{left}
χ^2_{right}

Figure 7.4-3: χ^2 Table for Example 7.4-1.

- The answers are:
  $\chi^2_{left} = 13.848$ and $\chi^2_{right} = 36.415$
  See ***Figure 7.4-4***.

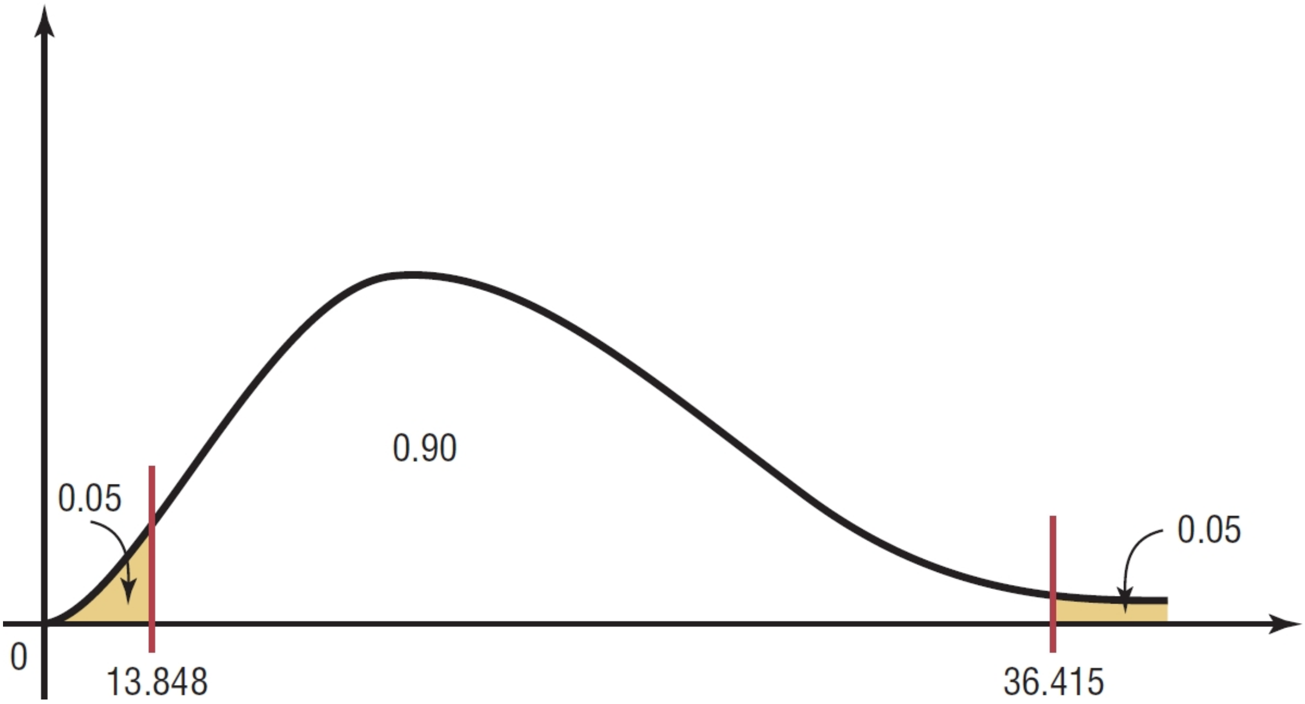


Figure 7.4-4: χ^2 Distribution for Example 7.4-1.

7.4.3 Formula for the Confidence Interval for a Variance and Standard Deviation

- To find ***confidence intervals for variances and standard deviations, you must assume that the variable is normally distributed***.
- The formulas for the confidence intervals are shown here.

  $$\frac{(n-1)s^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{left}}$$

  $d.f. = n-1$
- Formula for the confidence interval for a standard deviation would be:

  $$\sqrt{\frac{(n-1)s^2}{\chi^2_{right}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{left}}}$$

Example 7.4-2

Considering the variance of water depths resulting from irrigation. Twenty-five measurements produced a variance of

$s^2 = 0.003969 \left(\frac{cm}{hr}\right)^2$, or $S = 0.063\frac{cm}{hr}$

In this case, the greater the variance, the poorer the equipment, so an upper limit would be of interest. Thus, estimate the standard deviation, σ, with a 95% level of confidence (α = 0.05) and 24 degrees of freedom.

Solution

As discussed in ***Example 7.4-1*** above, with a 95% level of confidence (α = 0.05) and 24 degrees of freedom,

$\chi^2_{left} = 13.848$ and $\chi^2_{right} = 36.415$

And the standard deviation, σ, would be:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{right}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{left}}}$$

$$\sqrt{\frac{(25-1)\times 0.003969}{13.848}} < \sigma < \sqrt{\frac{(25-1)\times 0.003969}{36.415}}$$

$$0.0829\frac{cm}{hr} < \sigma < 0.0511\frac{cm}{hr}$$

Example 7.4-3

Fourth test cubes, $n = 40$, indicates that the compressive strengths of have an estimated standard deviation of

$$s = 5.02\frac{N}{mm^2}$$

Assuming normal population estimate σ with a one-sided upper 99% confidence limit.

Solution

As the example concerns with upper limit only, therefore, we concern with:

$$\sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{left}}}$$

and whole confidence limit of 0.99 should be adopted to determine χ^2_{left}. With $\because d.f. = 40 - 1 \approx 40$, χ^2_{left} would be, see ***Figure 7.4-5***:

$$\chi^2_{left} \approx 22.164$$

$$\sigma < \sqrt{\frac{(40-1)\times 5.02^2}{22.164}}$$

$\sigma < 6.66\ MPa$ ■

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Source: Owen, *Handbook of Statistical Tables,* Table A–4 "Chi-Square Distribution Table," © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.

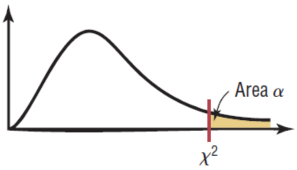


Figure 7.4-5: χ^2 Table for Example 7.4-3.

Home Work 7.4-1

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153\ (fluid\ ounces)^2$.

If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. Assume that the fill volume is approximately normally distributed, with adopting a 95% upper-confidence interval determine upper limit for the standard deviation, σ.

Answers

$\sigma < 0.17\ MPa$ ■

Solution

Home Work 7.4-2

For truss connection indicated in ***Figure 7.4-6***, a rivet is to be inserted into a hole. A random sample of n = 15 parts is selected, and the hole diameter is measured. The sample standard deviation of the hole diameter measurements is s = 0.008 millimeters. Construct a 99% lower confidence bound for σ^2. Ans.

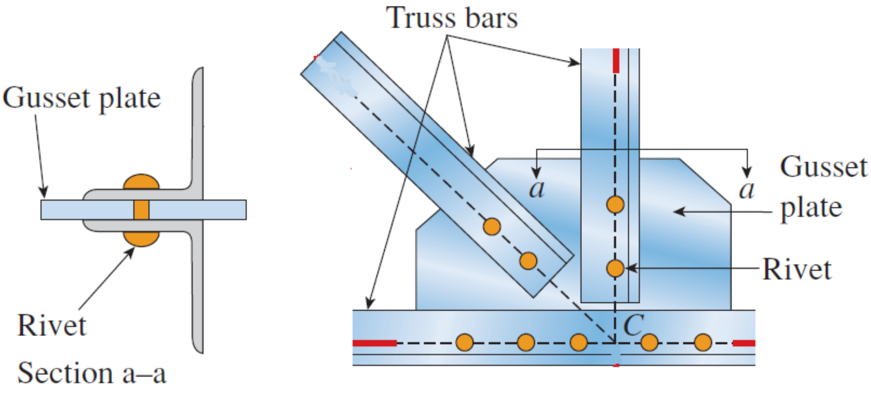


Figure 7.4-6: Truss connection for Home Work 7.4-2.

Ans. $\sigma^2 > 3.075 \times 10^{-5}\ mm$ ■

Solution

Contents

CHAPTER 8
HYPOTHESIS TESTING

8.1 INTRODUCTION

- Researchers are interested in answering many types of questions. For example:
  - A scientist might want to know whether the earth is warming up.
  - A physician might want to know whether a new medication will lower a person's blood pressure.
  - An educator might wish to see whether a new teaching technique is better than a traditional one.
  - A retail merchant might want to know whether the public prefers a certain color in a new line of fashion.
  - Automobile manufacturers are interested in determining whether seat belts will reduce the severity of injuries caused by accidents.
- In ***civil*** and ***water-resources engineering***, similar questions include:
  - A product-development engineer who wants to decide whether or not a ***new concrete additive has a significant influence on curing time***.
  - A soils engineer who wants to report whether ***a simple, in-the-field "vane test" provides an unbiased estimate of the lab-measured unconfined compressive strength of soil***.
  - A water-resources engineer who wants to verify an assumption used in his system model, namely, that ***the correlation coefficient between the monthly rainfalls on two particular watersheds is zero***.
- These types of questions can be addressed through ***statistical hypothesis testing***, which is a ***decision-making process for evaluating claims about a population***.
- In hypothesis testing, the researcher must:
  - Define the population under study,
  - State the particular hypotheses that will be investigated,
  - Give the significance level,
  - Select a sample from the population, collect the data, perform the calculations required for the statistical test,
  - Reach a conclusion.
- Hypotheses concerning parameters such as ***means*** and ***proportions*** can be investigated.
- There are two specific statistical tests used for hypotheses concerning means:
  - The z test,
  - The t test.
- This chapter:
  - We shall explain in detail the hypothesis-testing procedure along with the ***z test*** and the ***t test***.
  - In addition, a hypothesis-testing procedure for testing a ***single variance*** or ***standard deviation*** using the ***chi-square distribution*** is explained also.

- The three methods used to test hypotheses are:
    - The traditional method:
      It has been used since the hypothesis testing method was originally formulated.
    - The P-value method:
      It is a new method that has become popular with the advent of ***modern computers*** and ***high-powered statistical calculators***.
    - The confidence interval method:
      With this method, one can illustrate the relationship between hypothesis testing and confidence intervals.

8.2 Hypothesis Testing—Traditional Method

8.2.1 Null and Alternative Hypothesis

- Every hypothesis-testing situation begins with the statement of a hypothesis.
- A ***statistical hypothesis*** is ***a claim about a population parameter***. This claim may or may not be true.
- There are two types of statistical hypotheses for each situation: the ***null hypothesis*** and the ***alternative hypothesis***.
  - The ***null hypothesis***, symbolized by H_0, is a ***statistical hypothesis that states that there is no difference between a parameter and a specific value***, or ***that there is no difference between two parameters***.
  - The ***alternative hypothesis***, symbolized by H_1, is ***a statistical hypothesis that states the existence of a difference between a parameter and a specific value***, or ***states that there is a difference between two parameters***.

Example 8.2-1

State the null and alternative hypotheses when a medical researcher is interested in finding out whether a new medication will have any undesirable side effects. The researcher is particularly concerned with whether the pulse rate will increase, decrease, or remain unchanged after a patient takes the medication. He knows that the mean pulse rate for the population under study is 82 beats per minute.

Solution

Since the researcher knows that the mean pulse rate for the population under study is 82 beats per minute, the hypotheses for this situation are:

$H_0: \mu = 82 \qquad H_1: \mu \neq 82$

where

- The null hypothesis specifies that the mean will remain unchanged,
- The alternative hypothesis states that it will be different.

This test is called a ***two-tailed test*** (a term that will be formally defined later), since the possible side effects of the medicine could be to ***raise*** or ***lower*** the pulse rate.

Example 8.2-2

State the null and alternative hypotheses when a chemist invents an additive to increase the life of an automobile battery. The mean lifetime of the automobile battery without the additive is 36 months.

Solution

As the mean lifetime of the automobile battery without the additive is 36 months, and as the chemist only interests with increasing of battery life, therefore the null and alternative hypotheses would be:

$H_0: \mu = 36 \qquad H_1: \mu > 36$

This test is called ***right-tailed***, since the interest is in an increase only.

Example 8.2-3

State the null and alternative hypotheses when a contractor wishes to lower heating bills by using a special type of insulation in houses. The average of the monthly heating bills is $78.

Solution

As the average of the monthly heating bills is \$78 and the contactor wishes to lower heating bills, therefore the null and alternative hypotheses would be:

$H_0: \mu = \$78 \quad H_1: \mu < \$78$

This test is a ***left-tailed test***, since the contractor is interested only in lowering heating costs.

8.2.2 TRANSLATE CLAIM FORM A VERBAL FORM INTO A MATHEMATICAL FORM

- To state hypotheses correctly, researchers must ***translate the claim from words into mathematical symbols***.
- The basic symbols used are as follows:

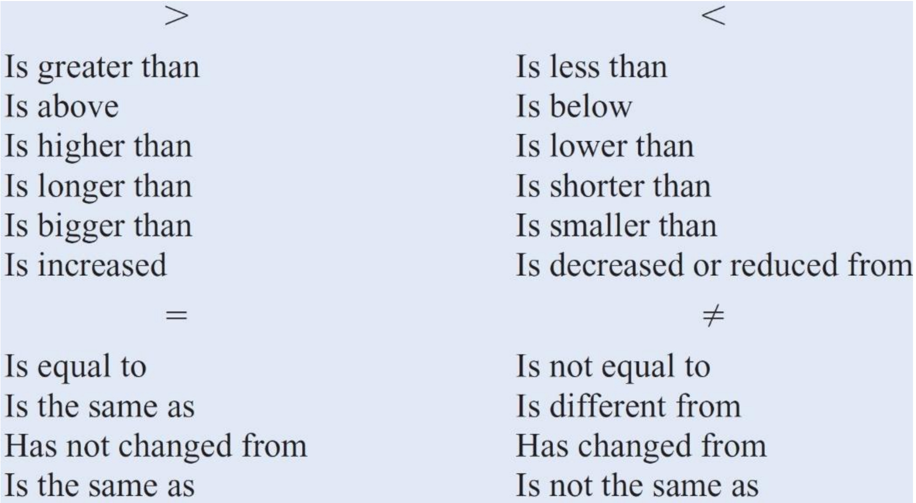
Equal to	$=$	Greater than	$>$
Not equal to	$\neq$	Less than	$<$

- The null and alternative hypotheses are stated together, and ***the null hypothesis contains the equal sign***, as shown (where k represents a specified number).

Two-tailed test	Right-tailed test	Left-tailed test
H_0: $\mu = k$ H_1: $\mu \neq k$	H_0: $\mu = k$ H_1: $\mu > k$	H_0: $\mu = k$ H_1: $\mu < k$

- ***Table 8.2-1*** shows some common phrases that are used in hypotheses and claims, and the corresponding symbols.
- This table should be helpful in translating verbal claims into mathematical symbols.

TABLE 8.2-1: HYPOTHESIS-TESTING COMMON PHRASES.



$>$	$<$
Is greater than	Is less than
Is above	Is below
Is higher than	Is lower than
Is longer than	Is shorter than
Is bigger than	Is smaller than
Is increased	Is decreased or reduced from
$=$	$\neq$
Is equal to	Is not equal to
Is the same as	Is different from
Has not changed from	Has changed from
Is the same as	Is not the same as

8.2.3 HOW TO ASSIGN NULL AND ALTERNATIVE HYPOTHESES IN A STUDY OR RESEARCH

- In most ***professional journals***, ***the assumption is that the mean, proportion, or standard deviation is equal to a given specific value***. Therefore, the ***null hypothesis is always stated using the equals sign***.
- Also, when a researcher conducts a study, ***he or she is generally looking for evidence to support a claim***. ***Therefore, the claim should be stated as the alternative hypothesis***, i.e., using $<$ or $>$ or $\neq$.
- Because of this, the ***alternative hypothesis is sometimes called the research hypothesis***.

8.2.4 SIGNIFICANT LEVEL

After stating the hypothesis, the researcher designs the study adopting the following steps:

- Selects the correct statistical test,
- Chooses an appropriate level of significance,
- Formulates a plan for conducting the study.

EXAMPLE 8.2-4

State the steps that have been adopted in the medical researcher of ***Example 8.2-1*** to find out whether a new medication will have any undesirable side effects.

Solution

In this situation, the researcher will:

- Select a sample of patients who will be given the drug. Recall that:
    - When samples of a specific size are selected from a population, ***the means of these samples will vary about the population mean***,
    - The distribution of the sample means will be approximately normal when the sample size is 30 or more
- After allowing a suitable time for the drug to be absorbed, the researcher will measure each person's pulse rate.

EXAMPLE 8.2-5

Assume that the medical researcher of ***Example 8.2-5*** has taken a sample of patients who use the drug for a sufficient time. The sample mean was unequal to the population mean of 82 beats per minute. Discuss how the sample results may be interpreted in the context of hypothesis testing.

Solution

- Even if the null hypothesis, $H_0: \mu = 82$, is true, the mean of the pulse rates of the sample of patients ***will not***, in most cases, ***be exactly equal to the population mean of 82 beats per minute***.
- There are two possibilities.
    - Either the null hypothesis is true, and the difference between the sample mean and the population mean is due to chance;
    - Or the null hypothesis is false, and the sample came from a population whose mean is not 82 beats per minute but is some other value that is not known.

  These situations are shown in Figure 8.2-1.
- The difference between sample and population means:
    - The ***farther away the sample mean is from the population mean***, ***the more evidence there would be for rejecting the null hypothesis***.
    - If the mean pulse rate of the sample were, say, 83, the researcher would probably ***conclude that this difference was due to chance*** and ***would not reject the null hypothesis***.
    - But if the sample mean were, say, 90, then in all likelihood the researcher would ***conclude that the medication increased the pulse rate of the users*** and ***would reject the null hypothesis***.
- Significant versus Insignificant Differences:
    - The question is, ***where does the researcher draw the line***?
    - This decision is ***not made on feelings or intuition***; ***it is made statistically***. Here is where ***the concepts of statistical test and level of significance are used***.

A **statistical test** uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected.

The numerical value obtained from a statistical test is called the **test value.**

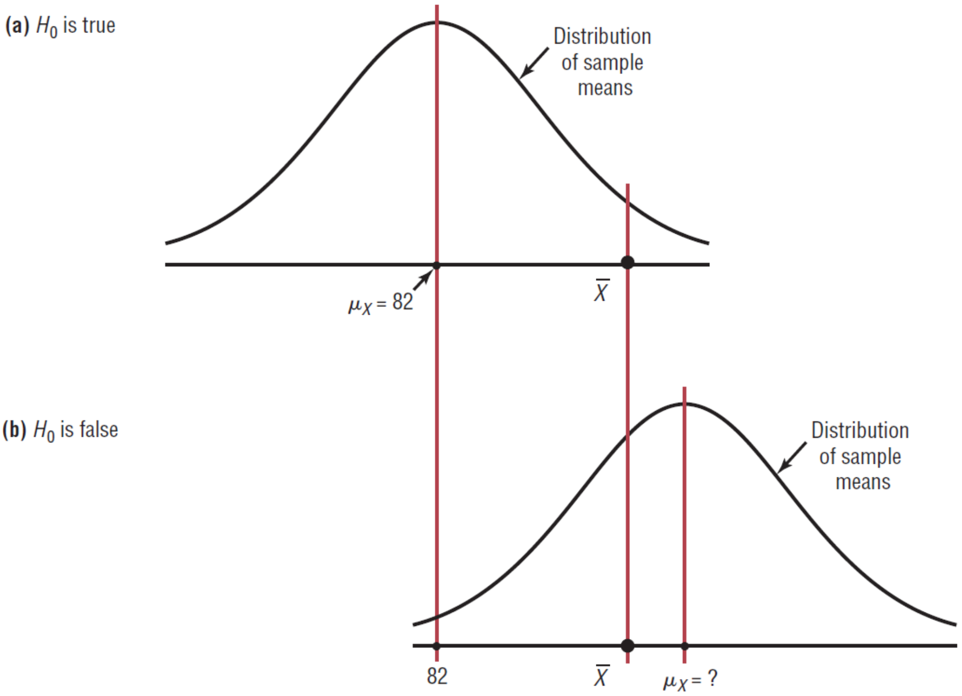
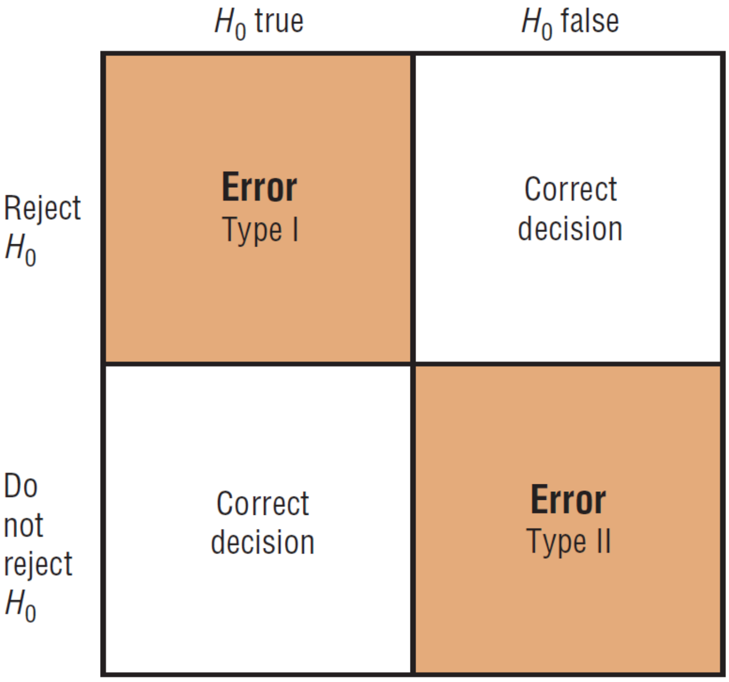


FIGURE 8.2-1: SITUATIONS IN HYPOTHESIS TESTING FOR EXAMPLE 8.2-5

8.2.5 POSSIBLE OUTCOMES

- In the hypothesis-testing situation, ***there are four possible outcomes***.
- In reality, the null hypothesis may or may not be true, and a decision is made to reject or not reject it on the basis of the data obtained from a sample.
- The four possible outcomes are shown in ***Figure 8.2-2***. Notice that there are ***two possibilities for a correct decision*** and ***two possibilities for an incorrect decision***.



	H_0 true	H_0 false
Reject H_0	**Error** Type I	Correct decision
Do not reject H_0	Correct decision	**Error** Type II

FIGURE 8.2-2: POSSIBLE OUTCOMES OF A HYPOTHESIS TEST.

- Type I error:
  A ***type I error*** occurs ***if you reject the null hypothesis when it is true***.

- Type II error:
  A ***type II error*** occurs ***if you do not reject the null hypothesis when it is false***.

EXAMPLE 8.2-6

State the type I error and type II error for the medical researcher of ***Example 8.2-5***.

Solution

- Type I error:
  - The medication ***might not significantly change the pulse rate*** of all the users in the population; but ***it might change the rate***, by ***<u>chance</u>***, of ***the subjects in the sample***.
  - In this case, the researcher will ***reject the null hypothesis when it is really true***, thus committing a ***type I error***.
- Type II error:
  - On the other hand, ***the medication might not change the pulse rate*** of the ***subjects in the sample***, but when it is given to the general population, ***it might cause a significant increase or decrease in the pulse rate of users***.
  - The researcher, on the basis of the data obtained from the sample, ***will not reject the null hypothesis***, thus committing a ***type II error***.

EXAMPLE 8.2-7

State the null and alternative hypotheses when a materials-testing engineer must report, on the basis of three batches, whether the mean asphalt content of all batches of material leaving a particular plant is equal to the desired value of 5 percent. The concern is that unknown changes in the mixing procedure or raw materials supply might have increased or decreased this mean. Comment on the theme of the proposed hypotheses.

Solution

As the material-testing engineer aims to know whether the mean asphalt is equal to the desired value or differ from it and he has no concern if different value is larger or smaller than the indicated mean value, therefore the null and alternative hypotheses would be:

$H_0: \mu = 5\% \qquad H_1: \mu \neq 5\%$

and the test would be a ***two-tailed test***.

As the hypothesis testing is based on a sample statistic, therefore the observed value of this sample statistic may:

- ***Lie some distance from the expected value of 5 even if this is the true value of the mean***, ***type I error***.
- On the other hand, if the true value of ***the mean asphalt content is no longer 5 but has slipped to 4.5 or increased to 5.6, the observed sample average may nonetheless lie close to the hypothesized value of 5***, ***type II error***.

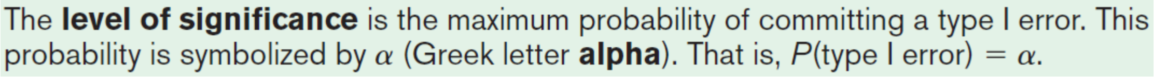
8.2.6 The Critical Value

- From the previous discussion, the ***hypothesis testing decision is made on the basis of probabilities***.
- That is:

  $|X - \mu|$ is large $\Longrightarrow$ The null hypothesis is probably not true.

  $|X - \mu|$ is small $\Longrightarrow$ The null hypothesis is probably true.

- The question is, ***How large a difference is necessary to reject the null hypothesis***?
  - Here is where the level of significance is used.

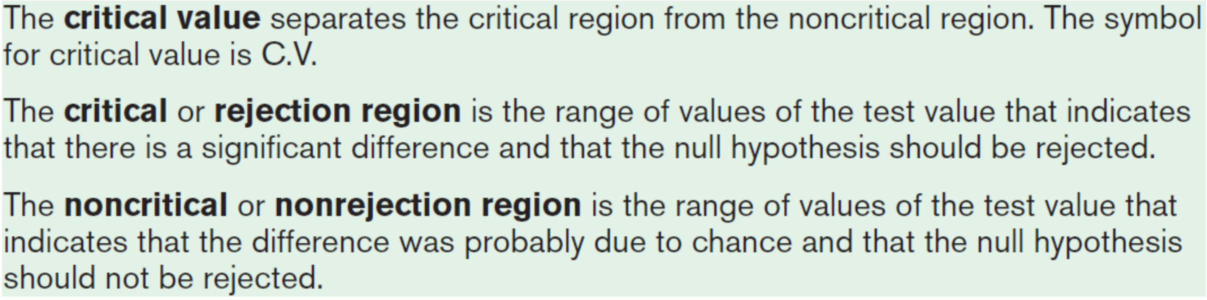


The **level of significance** is the maximum probability of committing a type I error. This probability is symbolized by α (Greek letter **alpha**). That is, $P(\text{type I error}) = \alpha$.

  - The probability of a type II error is symbolized by β, the Greek letter beta. That is, $P(type\ II\ error)\ = \beta$
  - In most hypothesis-testing situations, β cannot be easily computed; however, α and β are related in that decreasing one increases the other.
- In a hypothesis-testing situation, the researcher decides what level of significance to use. After a significance level is chosen, a critical value is selected from a table for the appropriate test. If a z test is used, for example, ***the z table of Chapter 6 is consulted to find the critical value***. The $z_{critical}$ for the most common levels of significant, α, are presented in below for convenience:

Confidence Interval	**Significance Level, α**	**Critical z for Two-tails Test, i.e. $z_{\frac{\alpha}{2}}$**
90%	10%	1.65
95%	5%	1.96
99%	1%	2.58

- The critical value determines the ***critical*** and ***noncritical regions***.



The **critical value** separates the critical region from the noncritical region. The symbol for critical value is C.V.

The **critical** or **rejection region** is the range of values of the test value that indicates that there is a significant difference and that the null hypothesis should be rejected.

The **noncritical** or **nonrejection region** is the range of values of the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected.

8.3 Hypothesis Testing for the Mean

In this section, two statistical tests will be explained:

- The z test is used when σ is known,
- The t test is used when σ is unknown.

8.3.1 The z Test for a Mean

8.3.1.1 Basic Concepts and Test Procedures

- The z test is defined formally as follows:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \textbf{Eq. 8.3-1}$$

  where
  $\bar{X}$ sample mean
  μ hypothesized population mean
  σ population standard deviation
  n is sample size
  The denominator $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean.
- Assumptions for the z test for a mean when σ is known:
  The z test implicitly based on the following two assumptions:
  - The sample is a random sample.
  - Either $n \geq 30$ or the ***population is normally distributed*** if $n < 30$.
- A step by step procedure for the test:
  From the previous discussion, there are five steps for solving hypothesis-testing problems:

  Step 1: State the hypotheses and identify the claim, see ***Sections 8.2.1*** through ***8.2.3***.
  Step 2: Find the critical value(s) as discussed in ***Section 8.2.6***.
  Step 3: Compute the test value based on ***Eq. 8.3-1***.
  Step 4: Make the decision to reject or not reject the null hypothesis.
  Step 5: Summarize the results.

8.3.1.2 Examples

Example 8.3-1

A researcher wishes to see if the mean number of days that a basic, low-price, small automobile sits on a dealer's lot is 29. A sample of 30 automobile dealers has a mean of 30.1 days for basic, low-price, small automobiles. At $\alpha = 0.05$, test the claim that the mean time is greater than 29 days. The standard deviation of the population is 3.8 days.

Solution

As $n \geq 30$, therefore Eq. 8.3-1 can be applied even when the population is not normally distributed.

Step 1: State the hypotheses and identify the claim.

$H_0\!: \mu = 29\ day$ and $H_1\!: \mu > 29\ day$

Step 2: Find the critical value(s).

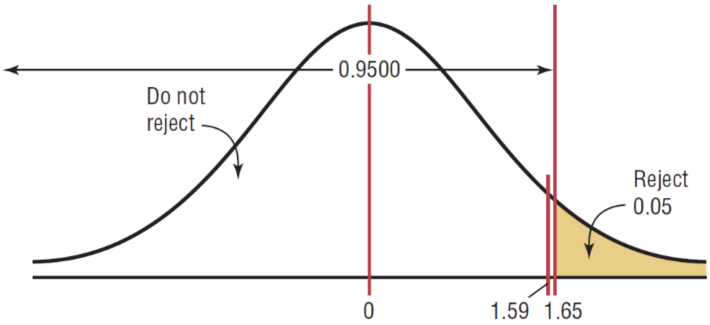
$z_{\frac{\alpha}{2}} = 1.65$

Step 3: Compute the test value based on ***Eq. 8.3-1***.

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{30.1 - 29}{\frac{3.8}{\sqrt{30}}} = 1.59$$

Step 4: Make the decision to reject or not reject the null hypothesis.

$\because z = 1.59 < z_{\frac{\alpha}{2}}$



The null hypotheses cannot be rejected.

Step 5: Summarize the results. There is not enough evidence to support the claim that the mean time is greater than 29 days.

Comment:

Even though in the sample mean of 30.1 is higher than the hypothesized population mean of 29, ***it is not significantly higher. Hence, the difference may be due to chance***. When ***the null hypothesis is not rejected***, ***there is still a probability of a type II error***, i.e., of not rejecting the null hypothesis when it is false.

Example 8.3-2

Suppose the specification for the yield strength of rebars required a mean value of 38 ksi. It is, therefore, essential that the population of rebars to be used in the construction of a reinforced concrete structure has the required mean strength. From the rebars delivered to the construction site by the supplier, the engineer ordered that a sample of 25 rebars be randomly selected and tested for yield strengths. The sample mean from the 25 tests yielded a value of 37.5 ksi. It is known that the standard deviation of rebar strength from the supplier is $\sigma = 3.0\ ksi$. In your solution assumes a normally distributed population and adopts $\alpha = 10\%$.

Solution

Step 1: State the hypotheses and identify the claim.

$H_0: \mu = 38\ ksi$ and $H_1: \mu < 38\ ksi$

Step 2: Find the critical value(s).

$z_{\frac{\alpha}{2}} = 1.65$

Step 3: Compute the test value based on ***Eq. 8.3-1***.

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{37.5 - 38}{\frac{3.0}{\sqrt{25}}} = -0.833$$

Step 4: Make the decision to reject or not reject the null hypothesis.

$\because |z| = 0.833 < z_{\frac{\alpha}{2}} = 1.65$

The null hypotheses cannot be rejected.

Step 5: Summarize the results.

There is not enough evidence to support the claim that the mean of the yield strength is smaller than $38\ ksi$. Therefore, the rebars from the supplier satisfy the required yield strength of the specification and are acceptable.

Example 8.3-3

Based on the experience of a pile company, it has been found that a specific pile type has a capacity of $\mu = 180\ ton$ with a standard deviation of $\sigma = 24\ ton$. In a specific site, 200 piles of the same type have been derived and tested and the mean value of their capacity is $\bar{x} = 184\ ton$. Test the hypotheses with $\alpha = 0.05$ to see if there is a significant difference between $\bar{x}$ and μ.

Solution

As $n \geq 30$, therefore ***Eq. 8.3-1*** can be applied even when the population is not normally distributed.

Step 1: State the hypotheses and identify the claim.

$H_0: \mu = 180\ ton$ and $H_1: \mu > 180\ ton$

Step 2: Find the critical value(s).

$z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96$

Step 3: Compute the test value based on ***Eq. 8.3-1***.

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{184 - 180}{\frac{24}{\sqrt{200}}} = 2.36$$

Step 4: Make the decision to reject or not reject the null hypothesis.

$\because z = 2.36 > z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = 1.96$

The null hypotheses can be rejected.

Step 5: Summarize the results.

There is enough evidence to support the claim that the mean of the pile capacity is larger than $180\ ton$.

EXAMPLE 8.3-4

Resolve Example 8.3-3 but with adopting $\alpha = 0.01$.

Solution

Step 1: State the hypotheses and identify the claim.

$H_0 = 180\ ton$ and $H_1 > 180\ ton$

Step 2: Find the critical value(s).

$z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = 2.58$

Step 3: Compute the test value based on ***Eq. 8.3-1***.

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{184 - 180}{\frac{24}{\sqrt{200}}} = 2.36$$

Step 4: Make the decision to reject or not reject the null hypothesis.

$\because z = 2.36 < z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = 2.58$

The null hypotheses cannot be rejected.

Step 5: Summarize the results.

There is not enough evidence to support the claim that the mean of the pile capacity is larger than $180\ ton$.

This example shows how the selected significant level, α, can affect the final decision regarding accepting or rejecting a hypothesis.

8.3.2 The t Test for a Mean

8.3.2.1 Basic Concepts and Test Procedures

- When the population standard deviation is unknown, ***the z test is not normally used for testing hypotheses involving means***.
- A different test, ***the t test, is used***. The ***distribution of the variable should be approximately normal***.
- As stated in Chapter 7, the t distribution is similar to the standard normal distribution and as the sample size increases, the it approaches the normal distribution.
- The t test is defined as follows:
  The t test is a statistical test for the mean of a population and is used when the population is normally or approximately normally distributed, and σ is unknown.
- The formula for the t test is:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$ **Eq. 8.3-2**

  The degrees of freedom are $d.f. = n - 1$.
- The critical values for the ***t test*** are given in ***Table 8.3-1***. Notice that the degrees of freedom are given for values from 1 through 30, then at intervals above 30 you should always round down to the nearest table value. For example, if d.f. = 59, use d.f. = 55 to find the critical value or values. ***This is a conservative approach***.
- Assumptions for the t test for a mean when σ is unknown:
  - The sample is a random sample.
  - Either $n \geq 30$ or the ***population is normally distributed*** if $n < 30$.
- When you test hypotheses by using the t test (traditional method), follow the same procedure as for the z test with minor modifications as in below:

  Step 1 State the hypotheses and identify the claim.
  Step 2 Find the critical value(s) from ***Table 8.3-1***.
  Step 3 Compute the test value based on ***Eq. 8.3-2***.
  Step 4 Make the decision to reject or not reject the null hypothesis.
  Step 5 Summarize the results.

  Remember that the t test should be used when the population is approximately normally distributed, and the population standard deviation is unknown.

TABLE 8.3-1: THE t DISTRIBUTION.

	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
34		1.307	1.691	2.032	2.441	2.728
36		1.306	1.688	2.028	2.434	2.719
38		1.304	1.686	2.024	2.429	2.712
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
55		1.297	1.673	2.004	2.396	2.668
60		1.296	1.671	2.000	2.390	2.660
65		1.295	1.669	1.997	2.385	2.654
70		1.294	1.667	1.994	2.381	2.648
75		1.293	1.665	1.992	2.377	2.643
80		1.292	1.664	1.990	2.374	2.639
90		1.291	1.662	1.987	2.368	2.632
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
$(z)\ \infty$		1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]

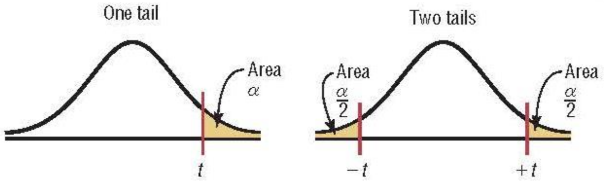
[a] This value has been rounded to 1.28 in the textbook.

[b] This value has been rounded to 1.65 in the textbook.

[c] This value has been rounded to 2.33 in the textbook.

[d] This value has been rounded to 2.58 in the textbook.

Source: Adapted from W. H. Beyer, *Handbook of Tables for Probability and Statistics,* 2nd ed., CRC Press, Boca Raton, Fla., 1986. Reprinted with permission.



8.3.2.2 Examples

Example 8.3-5

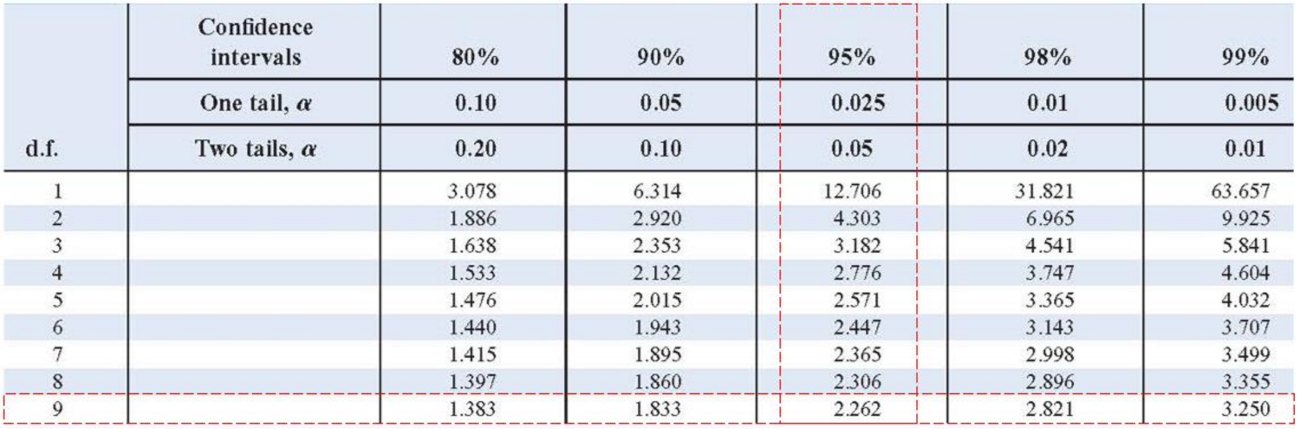
A medical investigation claims that the average number of infections per week at a hospital in southwestern Pennsylvania is 16.3. A random sample of 10 weeks had a mean number of 17.7 infections. The sample standard deviation is 1.8. Is there enough evidence to reject the investigator's claim at $\alpha = 0.05$?

Solution

Step 1 $H0: \mu = 16.3\ (claim)$ and $H1: \mu \neq 16.3$.

Step 2 For $\alpha = 0.05$ and $d.f. = 10 - 1 = 9$, the critical value is:

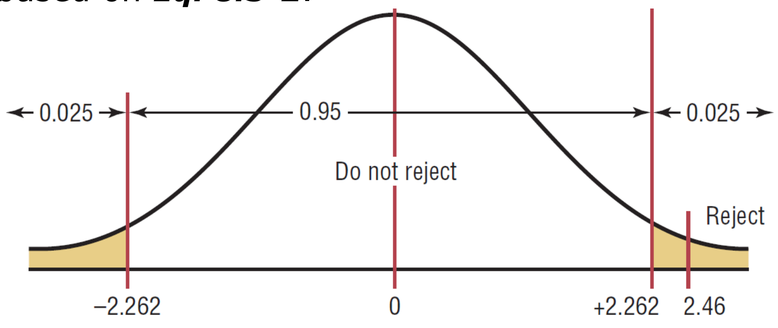
$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = 2.262$



	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250

Step 3 Compute the test value based on ***Eq. 8.3-2***.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{17.7 - 16.3}{\frac{1.8}{\sqrt{10}}} = 2.46$$



Step 4 Reject the null hypothesis since:

$t = 2.46 > t_{\frac{\alpha}{2}} = 2.262$

Step 5 There is enough evidence to reject the claim that the average number of infections is 16.3.

Example 8.3-6

In ***Example 8.3-2*** for the yield strength of rebars, the standard deviation, σ, of the population of rebars is known to be $3.0\ ksi$. In many cases, the information on the population standard deviation may not be known and must be estimated also from the available sampled data. Resolve ***Example 8.3-2*** suppose the 25 tests showed the following results: $\bar{x} = 37.5\ ksi$ and $s = 3.50\ ksi$. As for ***Example 8.3-2***, in your solution assumes a normally distributed population and adopts $\alpha = 10\%$.

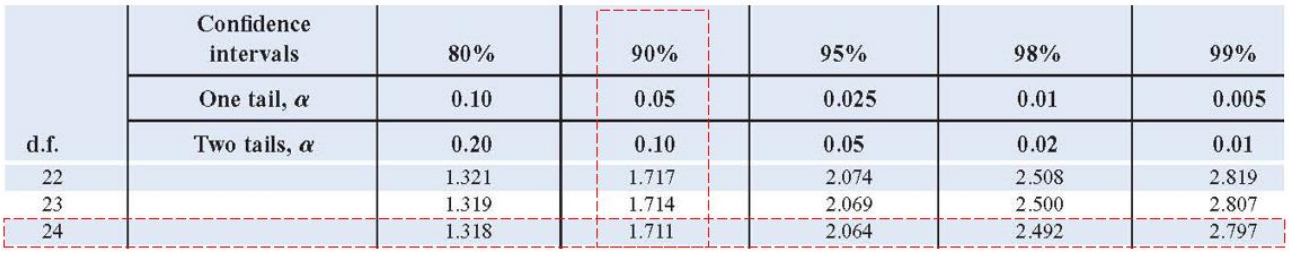
Solution

Step 1: State the hypotheses and identify the claim.

$H_0: \mu = 38\ ksi$ and $H_1: \mu < 38\ ksi$

Step 2 For $\alpha = 0.10$ and $d.f. = 25 - 1 = 24$, the critical value is:

$t_{\frac{\alpha}{2}} = t_{\frac{0.1}{2}} = 1.711$



	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797

Step 3 Compute the test value based on ***Eq. 8.3-2***.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{37.5 - 38}{\frac{3.50}{\sqrt{25}}} = -0.714$$

Step 4 Not reject the null hypothesis since:

$\because |t| = -0.714 < t_{\frac{\alpha}{2}} 2.262$

Step 5: Summarize the results.

There is not enough evidence to support the claim that the mean of the yield strength is smaller than $38\, ksi$. Therefore, the rebars from the supplier satisfy the required yield strength of the specification and are acceptable.

8.3.3 Homework Problems

Home Work 8.3-1

The foundation for a building is designed to rest on 100 piles based on the individual pile capacity of 80 tons. Nine test piles were driven at random locations into the supporting soil stratum and loaded until failure of each pile occurred. The results are as indicated in the table. For these data:

Test No.	*Pile Capacity (ton)*
1	82
2	75
3	95
4	90
5	88
6	92
7	78
8	85
9	80

(a) Estimate the mean and standard deviation of the individual pile capacity to be used at the site.

(b) At the 5% significance level, should the piles be accepted based on the results of the nine test piles? That is, perform hypothesis test, with the null hypothesis that the mean pile capacity is 80 tons.

Ans. $\bar{x} = \frac{\Sigma x_i}{n} = 85\ ton, s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}} = 6.76\ ton$

$t = 2.22 < t_{\frac{\alpha}{2}} = 2.306$

HOME WORK 8.3-2

Concrete placed on a structure was subsequently cored after 28 days, and the following results were obtained of the compressive strengths from five test specimens.

TEST NO.	COMPRESSIVE STRENGTH PSI
1	4142
2	3405
3	3402
4	4039
5	3372

(a) Estimate the mean and standard deviation of the concrete compressive strength.

(b) If the required minimum compressive strength is 3500 psi, perform a one-sided hypothesis test at the 2% significance level.

Ans. $\bar{x} = \frac{\Sigma x_i}{n} = 3672 \text{ psi}$, $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = 384\, psi$

$t = 1.00 < t_{\frac{\alpha}{2}} = 1.533$

HOME WORK 8.3-3

The ***daily dissolved oxygen concentration*** (***DO***) for a location A downstream from an industrial plant has been recorded for the past 10 consecutive days as tabulated.

DAY	DO (MG/L)
1	1.8
2	2
3	2.1
4	1.7
5	1.2
6	2.3
7	2.5
8	2.9
9	1.9
10	2.2

(a) Estimate the mean and standard deviation of the daily dissolved oxygen concentration (DO).

(b) Suppose the minimum concentration of DO required by the ***Environmental Protection Agency*** is 2.2 mg/l. Perform a hypothesis test to determine whether the stream quality satisfies the EPA standard at the significance level of 5%.

Ans. $\bar{x} = \frac{\Sigma x_i}{n} = 2.06\,\frac{\text{mg}}{\text{l}}$, $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = 0.46\,\frac{\text{mg}}{\text{l}}$

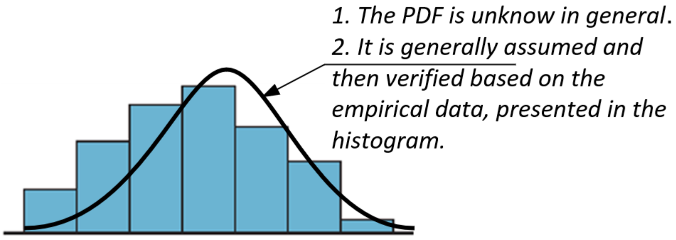
$|t| = 0.962 < t_{\frac{\alpha}{2}} = 2.262$

Contents

CHAPTER 9

DETERMINATION OF PROBABILITY DISTRIBUTION MODELS

9.1 INTRODUCTION



1. The PDF is unknow in general.
2. It is generally assumed and then verified based on the empirical data, presented in the histogram.

Figure 9.1-1: The main concern of Chapter 9: how to assume and verify the probability density function, PDF.

- Probability distribution model is unknown in general:
  - The probability distribution function, PDF, model appropriate to describe a random phenomenon is ***generally not known***; i.e., the ***functional form of the probability distribution is not defined***.
  - The final decision about selecting of PDF consists of the following two steps in general:
    - Select a trial PDF,
    - Verify of the selected PDF.
  - These steps have been briefly discussed in below to give the big picture of the chapter.

9.1.1 SELECT A TRIAL PDF

- Physical estimation of the PDF:
  Under certain circumstances, the basis or properties of the ***physical process underlying the random phenomenon may suggest the form of the required distribution***. For example,
  - If a process is composed of ***the sum of many individual effects, the Gaussian distribution*** may be appropriate on the basis of the ***central limit theorem***.
  - If the ***extremal conditions of a physical process are of interest***, one of the ***asymptotic extreme-value distributions*** may be a suitable model.
- Empirical estimation of the PDF:
  In many cases, the required probability distribution may need to be ***determined empirically*** based on the available observational data. For example,
  - If the frequency diagram, histogram, for a set of data can be constructed the required distribution model may be ***inferred by visually comparing a particular PDF with the corresponding frequency diagram***, see ***Figure 9.1-2***.
  - Alternatively, the available data may be plotted on probability papers prepared for specific distributions. If the data points plot approximately with a linear trend on one of these papers, the distribution associated with this paper may be an appropriate distribution model, see ***Figure 9.1-3***.

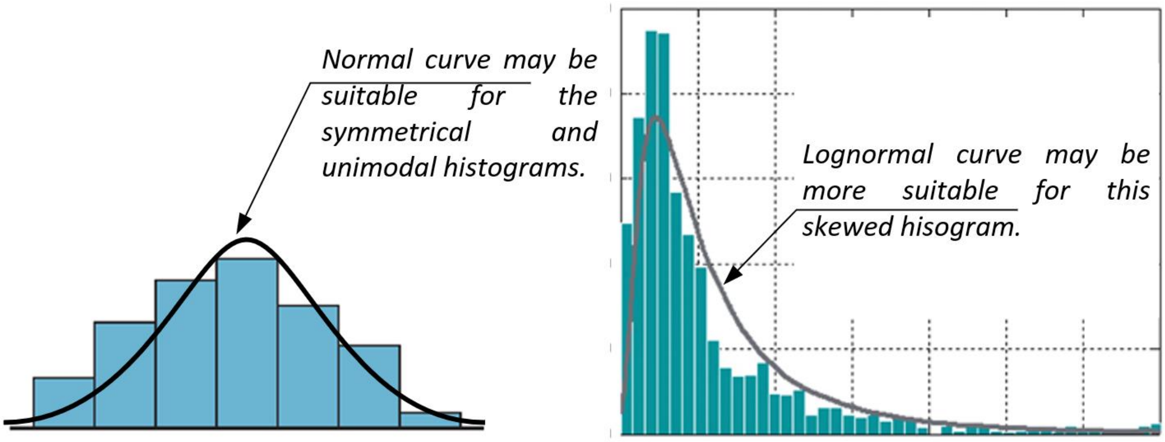


Figure 9.1-2: PDF added to the histograms for visual selection of a trial model.

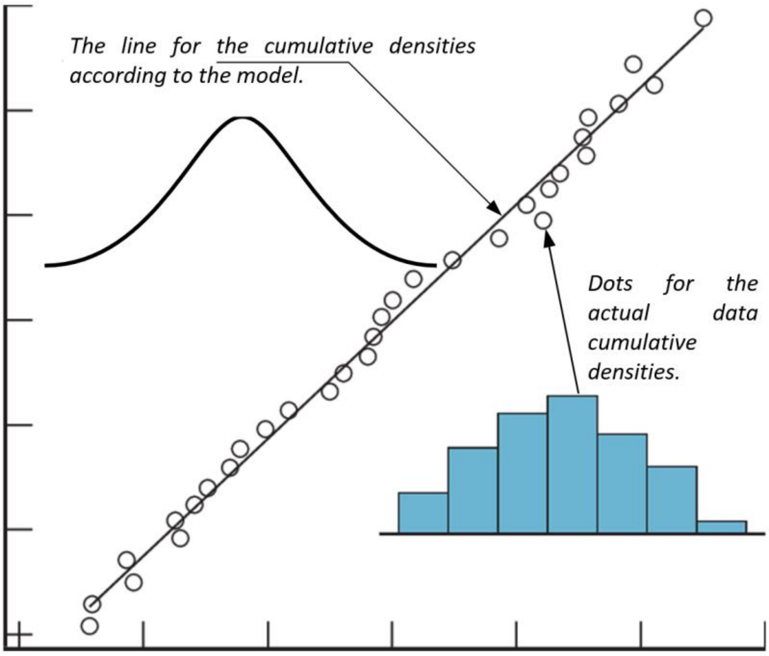


Figure 9.1-3: Using the probability paper to select a trial PDF.

- Using approximate model due to its mathematical convenience:
  - In practice, the choice of ***the appropriate distribution model may be dictated by mathematical convenience***. For example, because of:
    - The ***mathematical simplifications possible with the normal distribution***, and
    - The ***wide availability of probability information*** (such as probability tables) for this distribution,

    the normal or lognormal distribution is frequently used to model nondeterministic problems.
  - When the mathematically convenient models can be used?
    - These approximated models can be used at times even when there is ***no clear-cut basis for such a model*** and ***when the information is needed only for relative purposes***.
    - However, when the form of a distribution is important, particularly when ample data are available, the methods described in this chapter should provide the tools needed for its determination.

9.1.2 Verify of the selected PDF

- The assumed or trail PDF ***can be verified***, or ***disproved***, based on certain statistical tests, known as ***goodness-of-fit tests*** for distribution.

- Such tests can be used to discriminate ***the relative validity of the different distributions***.
- Two such goodness-of-fit tests are commonly used for these purposes:
  - The Chi-square, χ^2, test,
  - The Kolmogorov-Smirnov (K-S) test;
- A third test known as the ***Anderson-Darling test*** is particularly useful when the tails of a distribution are important.

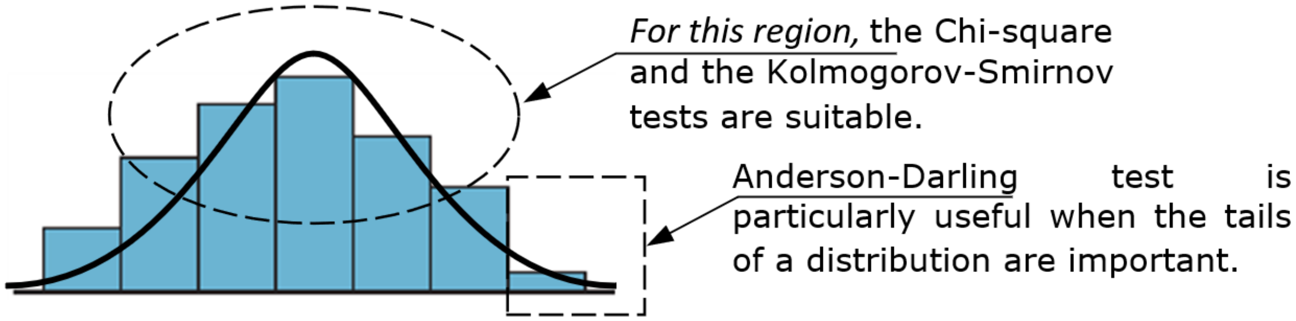

For this region, the Chi-square and the Kolmogorov-Smirnov tests are suitable.
Anderson-Darling test is particularly useful when the tails of a distribution are important.

Figure 9.1-4: Three different goodness of fit tests.

9.1.3 Chapter Layout

- Chapter layout is presented in ***Figure 9.1-5***.
- Logically a section for selection of PDF have to presented firstly while a testing section to be presented secondly.
- But to be compatible with the undergraduate curriculum, this chapter starts with the testing aspect as in ***Section 9.2*** and then presents the selection aspects as in ***Section 9.3***.

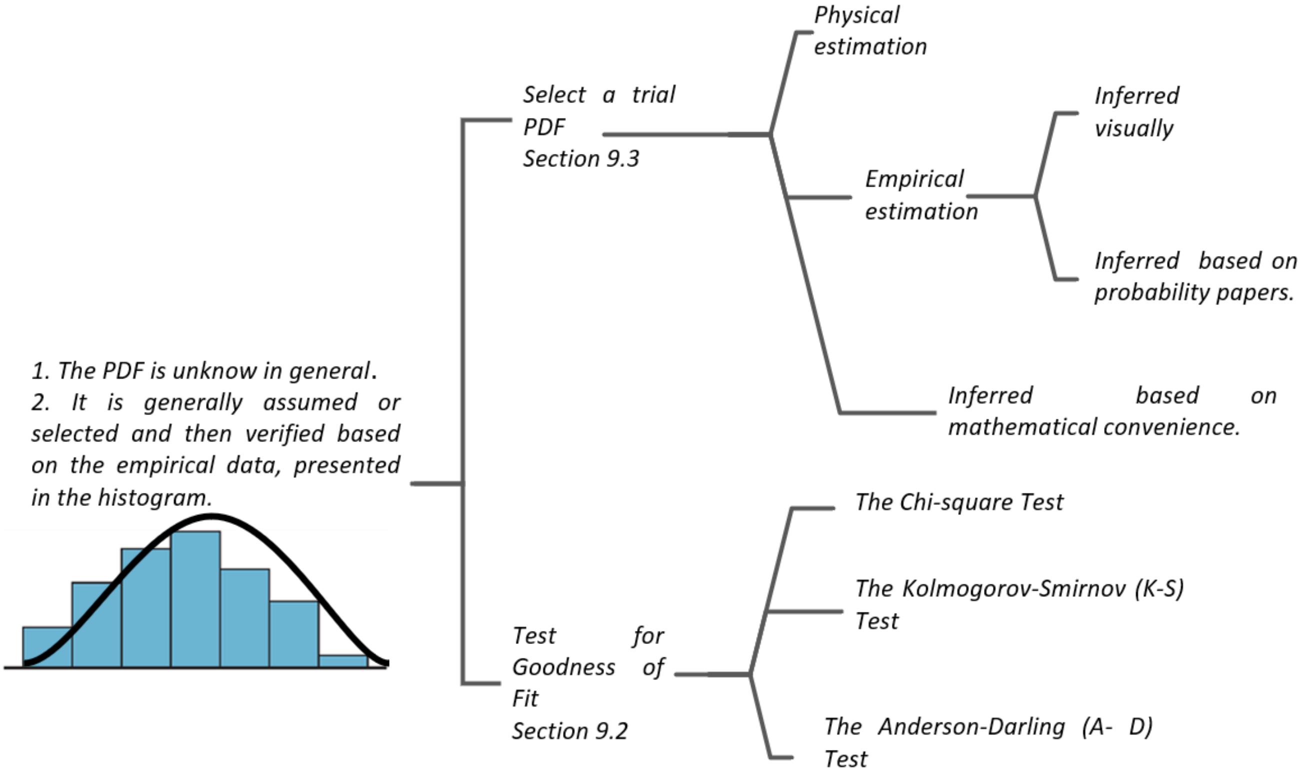

1. The PDF is unknow in general.
2. It is generally assumed or selected and then verified based on the empirical data, presented in the histogram.
Select a trial PDF Section 9.3
Physical estimation
Empirical estimation
Inferred visually
Inferred based on probability papers.
Inferred based on mathematical convenience.
Test for Goodness of Fit Section 9.2
The Chi-square Test
The Kolmogorov-Smirnov (K-S) Test
The Anderson-Darling (A- D) Test

Figure 9.1-5: Layout for Chapter 9.

9.2 Test for Goodness of Fit

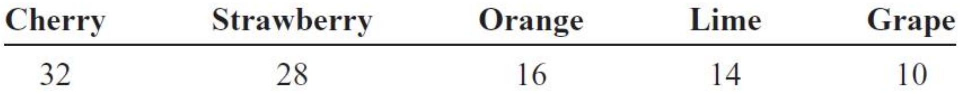
9.2.1 The Chi-square Test

- ***In addition to being used to test a single variance*** as discussed in Chapter 7, the ***chi-square***, χ^2, statistic can be used to ***see whether a frequency distribution fits a specific pattern***.
- For example,
  - To meet customer demands, a manufacturer of running shoes may wish to see whether buyers show a preference for a specific style.
  - A traffic engineer may wish to see whether accidents occur more often on some days than on others, so that she can increase police patrols accordingly.
  - An emergency service may want to see whether it receives more calls at certain times of the day than at others, so that it can provide adequate staffing.
- When ***you are testing to see whether a frequency distribution fits a specific pattern***, you can ***use the chi-square <u>goodness-of-fit test</u>***.

9.2.1.1 Basic Example

- For example, ***suppose as a market analyst you wished to see whether consumers have any preference among five flavors of a new fruit soda***. A sample of 100 people provided these data:

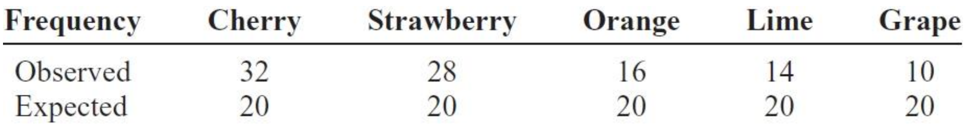
Table 9.2-1: Preference among five flavors of a new fruit soda.



Cherry	Strawberry	Orange	Lime	Grape
32	28	16	14	10

- Since the frequencies for each flavor were obtained from a sample, these actual frequencies are called the ***observed frequencies***. The frequencies obtained by calculation (***as if there were no preference***) are called the ***expected frequencies***. A completed table for the test is shown.

Table 9.2-2: Observed versus expected frequencies a new fruit soda.



Frequency	Cherry	Strawberry	Orange	Lime	Grape
Observed	32	28	16	14	10
Expected	20	20	20	20	20

- The observed frequencies will ***usually differ from the expected frequencies due to sampling error***; that is, the values differ from sample to sample. However, the question is: ***Are these differences significant (a preference exists), or are they due to chance***? The ***chi-square goodness-of-fit test will enable the researcher to determine the answer***.
- The Degrees of Freedom:
  - In the ***goodness-of-fit test***, ***the degrees of freedom are equal to the number of categories minus 1***.
  - For this example, there are five categories (cherry, strawberry, orange, lime, and grape); hence, the degrees of freedom are $5 - 1 = 4$.
  - This is so because the number of subjects in each of the first four categories is free to vary. However, in order for the sum to be 100—the total number of subjects—the number of subjects in the last category is fixed.

- Formula for the Chi-Square Goodness-of-Fit Test
  Based on two following assumptions.
  - The data are obtained from a random sample.
  - The expected frequency for each category must be 5 or more.

  The χ^2 formula for goodness-of-fit test would be:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad \blacksquare \qquad \textbf{Eq. 9.2-1}$$

  with degrees of freedom equal to the ***number of categories minus 1***, and where
  O is observed frequency,
  E is expected frequency.

- Difference between $(O-E)^2$ have been determined in terms of table below:

O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
32	20	12	144	7.2
28	20	8	64	3.2
16	20	-4	16	0.8
14	20	-6	36	1.8
10	20	-10	100	5
			Σ	**18**

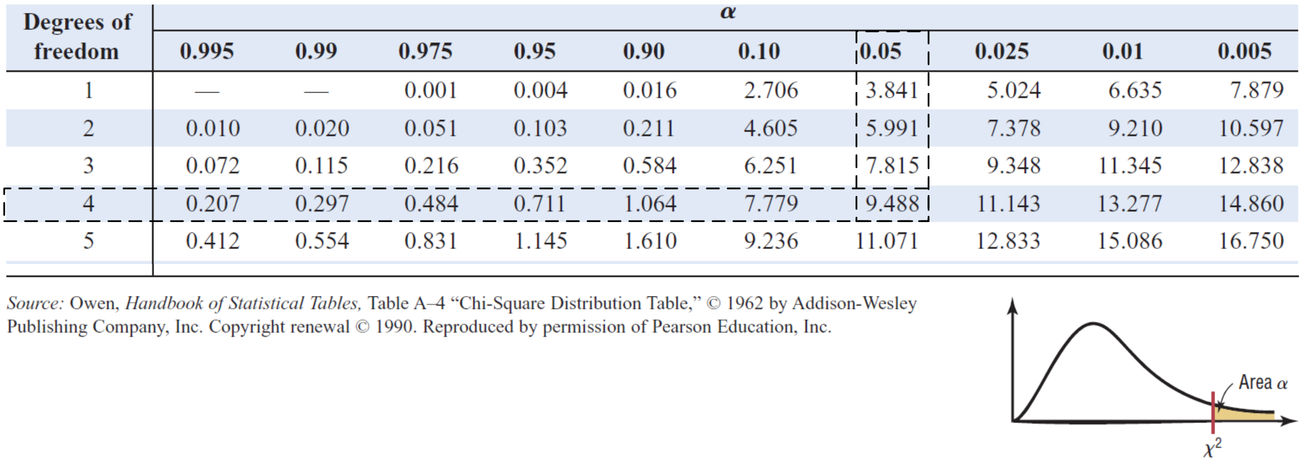
The χ^2 would be:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 18.0$$

With $\alpha = 0.05$ and DOF of 5-1=4 and from ***Table 9.2-3***, for convenience this table has been repriced from ***Chapter 7***,see ***Figure 9.2-1*** below, the critical value would be 9.488.
As,

$$\chi^2 = 18.0 > 9.488$$

therefore, the observed frequencies significantly differ from the expected ones and ***there is a preference among five flavors of a new fruit soda***.



Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750

Source: Owen, *Handbook of Statistical Tables,* Table A–4 “Chi-Square Distribution Table,” © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.

Figure 9.2-1: χ^2 Table for the basic example.

9.2.1.2 GRAPHICAL INTERPRETATIONS FOR GOODNESS-OF-FIT TEST

- To get some idea of why this test is called the goodness-of-fit test, examine graphs of the observed values and expected values. See ***Figure 9.2-2***. From the graphs, you can see whether the observed values and expected values are close together or far apart.

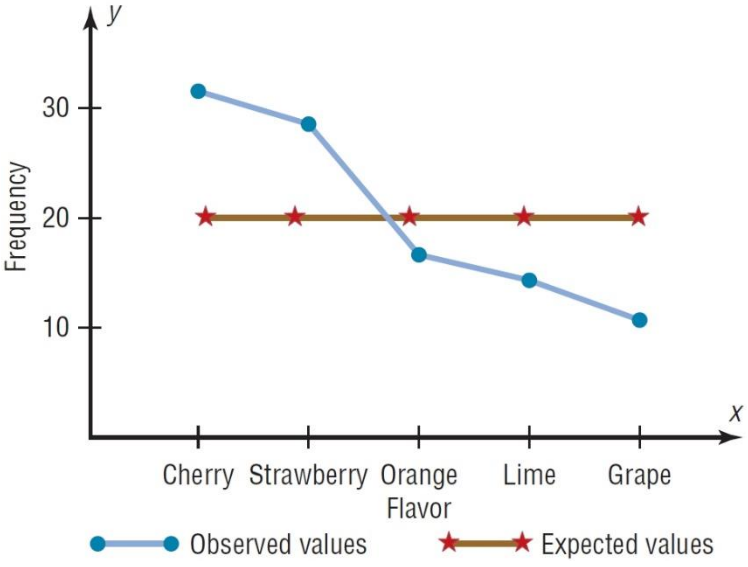


Figure 9.2-2: Graphs of the Observed and Expected Values for Soda Flavors.

- When the observed values and expected values are ***close together***, the ***chi-square test value will be small***. Then ***the decision will be there is "a good fit."*** See ***Figure 9.2-3***(a).
- When the observed values and the expected values are ***far apart***, the ***chi-square test value will be large***. Then ***there is "not a good fit."*** See ***Figure 9.2-3***(b).

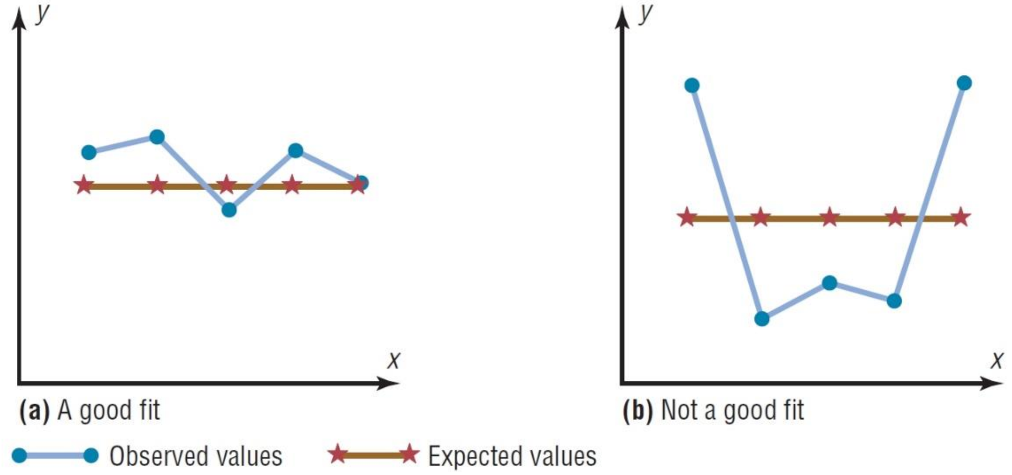


Figure 9.2-3: Results of the Goodness-of-Fit Test.

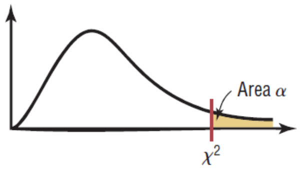
9.2.1.3 TABLE FOR THE CHI-SQUARE DISTRIBUTION

- For convenient and ready reference, the χ^2 table of ***Chapter 7*** has been reproduced in ***Table 9.2-3*** below.
- It is useful to recall that two different values are used in the formula because the distribution is not symmetric. One value is found on the left side of the table, and the other is on the right, see ***Figure 9.2-4*** below.

Table 9.2-3: The Chi-Square Distribution

Degrees of freedom	α									
	0.995	**0.99**	**0.975**	**0.95**	**0.90**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Source: Owen, *Handbook of Statistical Tables,* Table A–4 "Chi-Square Distribution Table," © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.



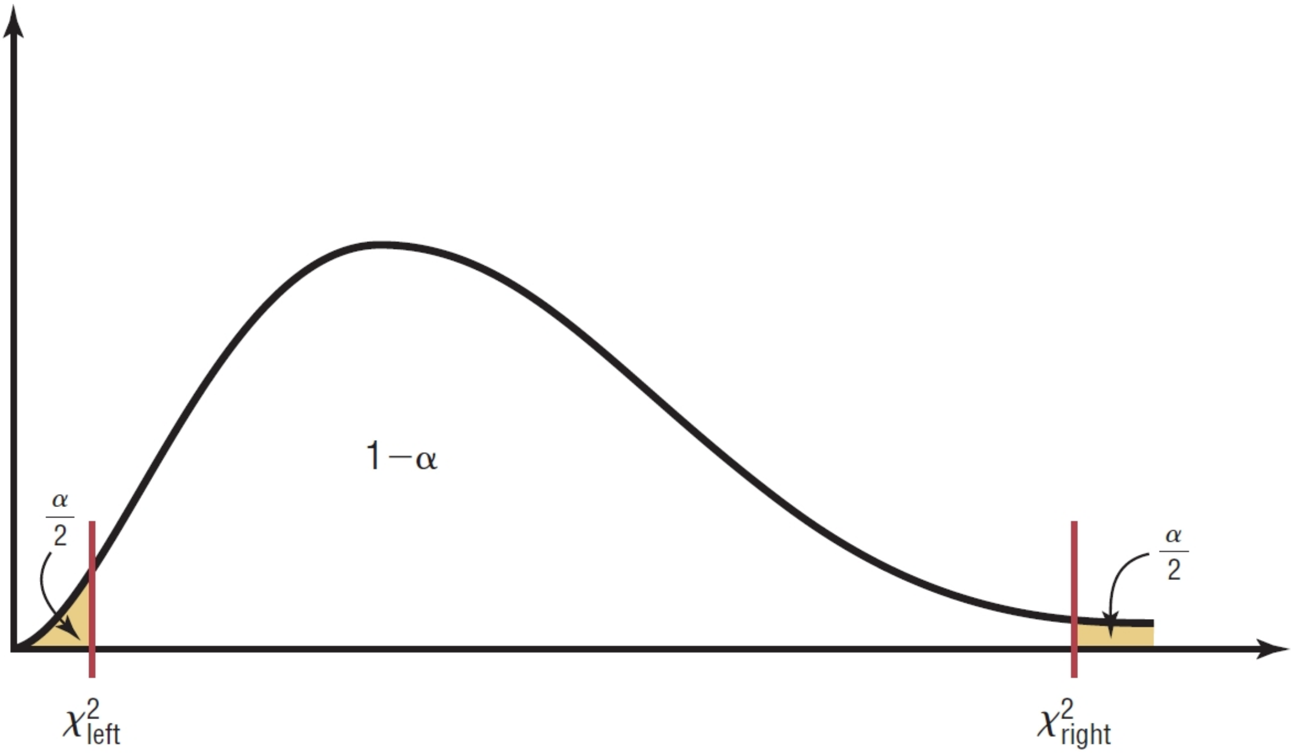


Figure 9.2-4: Chi-Square Distribution for d.f. = n − 1.

9.2.1.4 Examples

Example 9.2-1

A period of consecutive rainy days is called a ***wet run*** if ***the day immediately before the period and the day immediately after the period are dry***. Similarly, a period of days on which no rainfall is experienced is called a ***dry run if wet days precede and succeed it***.

The distribution of wet runs observed from January 1958 to May 1965 at Kew in London, England, is shown in ***Table 9.2-4***. Use χ^2, with α of 0.05, for a formal assessment of goodness-of-fit of observed frequencies to those of geometric distribution model.

Table 9.2-4: Observed and Expected Frequency distribution for wet runs.

Length of wet run in days	Observed Frequencies	Expected Frequencies According to Geometric Distribution
1	194	179.6
2	101	109.2
3	66	66.4
4	30	40.3
5	26	24.5
6	11	14.9
7	13	9.1
8	7	5.5
9	5	3.3
10	2	2.2

Solution

For graphical assessment for goodness of fit, observed and expected frequencies have been presented in ***Figure 9.2-5*** below which ***informally*** indicated that observed and expected frequencies have good fitness.

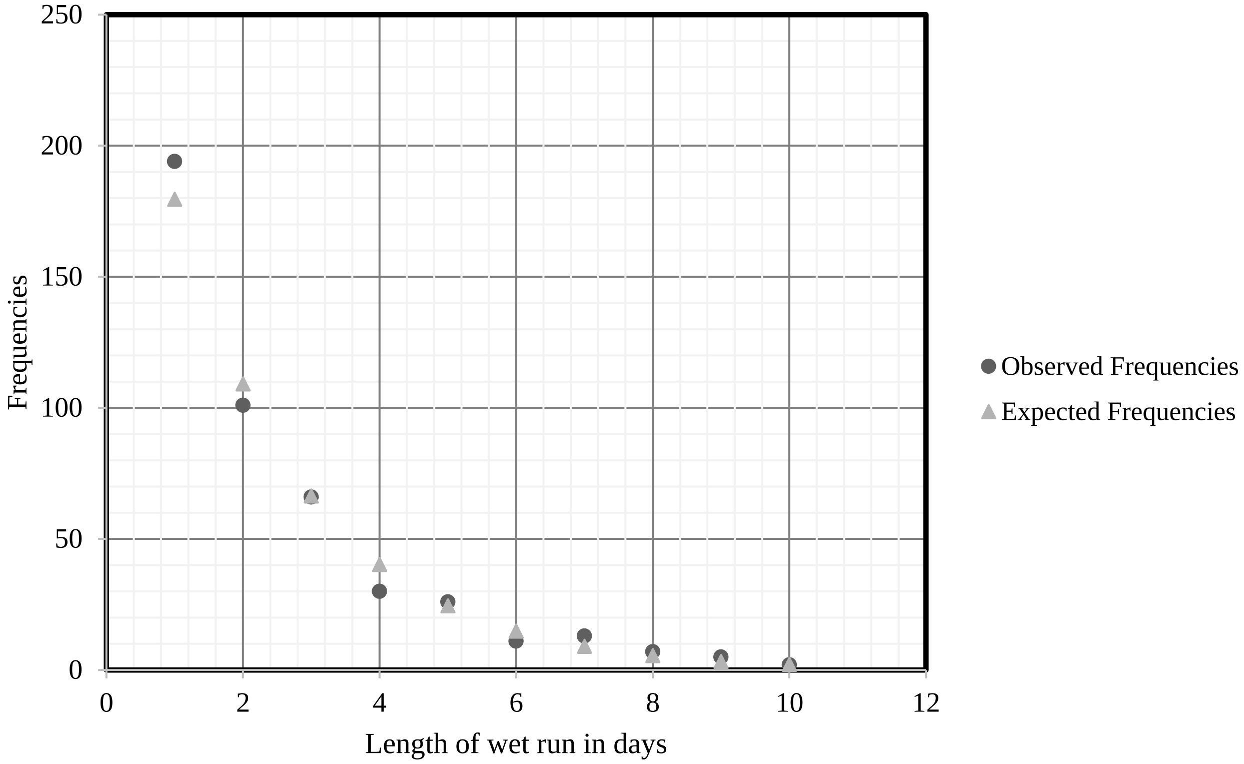


Figure 9.2-5: Observed and expected frequencies for Example 9.2-1.

For formal assessment of goodness of fit, determine χ^2 based on following relation and table below:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Length of wet run in days	Observed Frequency	Expected Frequencies According to Geometric Distribution	$\boldsymbol{O-E}$	$\boldsymbol{(O-E)^2}$	$\boldsymbol{\frac{(O-E)^2}{E}}$
1	194	179.6	14.4	207.36	1.15
2	101	109.2	-8.2	67.24	0.62
3	66	66.4	-0.4	0.16	0.00
4	30	40.3	-10.3	106.09	2.63
5	26	24.5	1.5	2.25	0.09
6	11	14.9	-3.9	15.21	1.02
7	13	9.1	3.9	15.21	1.67
8	7	5.5	1.5	2.25	0.41
9	5	3.3	1.7	2.89	0.88
10	2	2.2	-0.2	0.04	0.02
				Σ	**8.49**

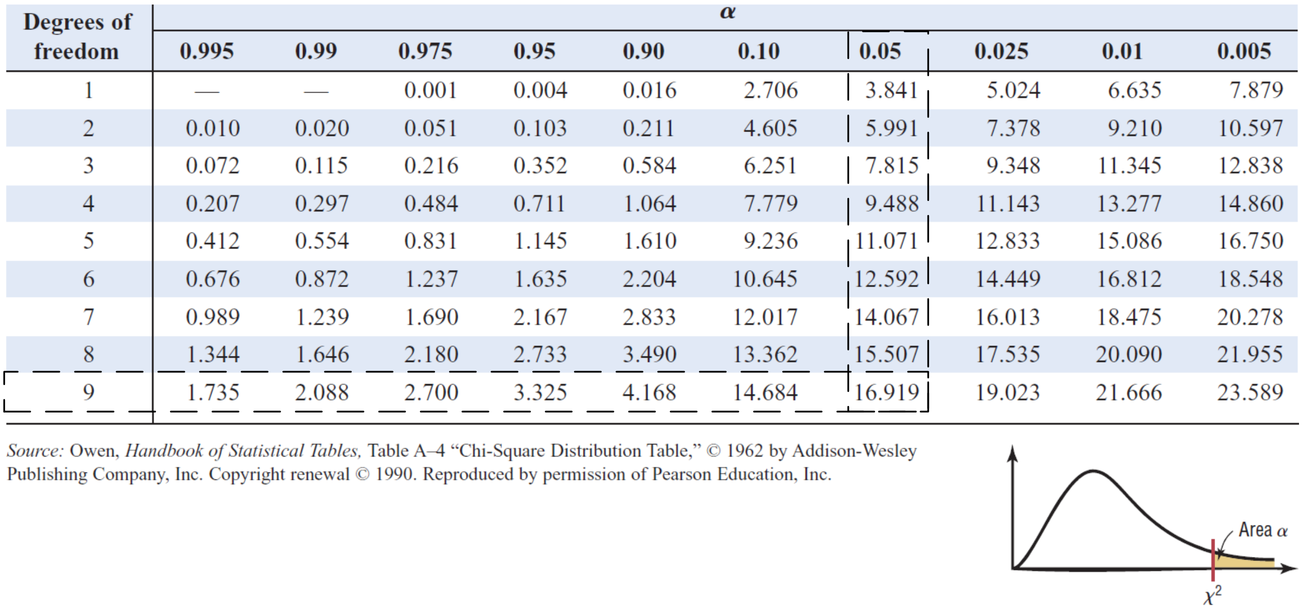
Based on ***Table 9.2-3*** and with α of 0.05 and DOF of 10-9=8, the critical value would be, see ***Figure 9.2-6*** below.

$\chi^2_{9,0.05} = 16.919$

As

$\chi^2 = 8.49 < \chi^2_{9,0.05}$

therefore, ***the difference between observed and expected frequencies is insignificant and they have good fit***.



Degrees of freedom	α 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589

Source: Owen, *Handbook of Statistical Tables,* Table A–4 "Chi-Square Distribution Table," © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.

Figure 9.2-6: χ^2 Table for Example 9.2-1.

Example 9.2-2

The annual number of rainstorms recorded at a number of stations have been observed. Poisson distribution with $\lambda = 1$ has been adopted to simulate the number of rainstorms. Observed and expected frequencies have been presented in ***Table 9.2-5***. Use graphical interpretations and χ^2 test, with $\alpha = 0.05$, for informal and formal test for goodness of fit.

Table 9.2-5: Observed and expected frequencies for Example 9.2-2.

Number of rainstorms per station per year	Observed Frequency	Expected Frequencies Based on Poisson Distribution with λ = 1
0	102	132.48
1	144	132.48
2	74	66.24
3	28	21.96
4	10	5.40
5	2	1.08

Solution

For graphical assessment for goodness of fit, observed and expected frequencies have been presented in ***Figure 9.2-7***. The figure ***informally*** indicates that there is a significant difference between the observed and expected frequencies.

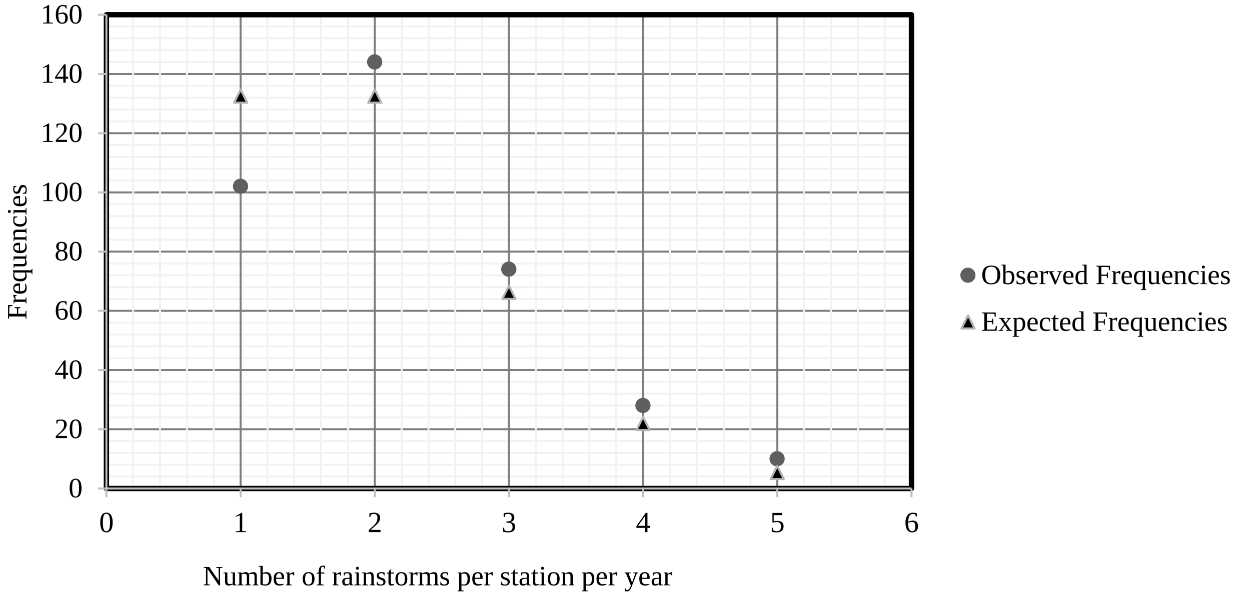


Figure 9.2-7: Observed and expected frequencies for Example 9.2-2.

For formal assessment of goodness of fit, determine χ^2 based on following relation and table below:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Number of rainstorms per station per year	Observed Frequency	Expected Frequencies Based on Poisson Distribution with λ = 1	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	102	132.48	-30.48	929.03	7.01
1	144	132.48	11.52	132.71	1.00
2	74	66.24	7.76	60.22	0.91
3	28	21.96	6.04	36.48	1.66
4	10	5.4	4.6	21.16	3.92
5	2	1.08	0.92	0.85	0.78
				Σ	**15.29**

The critical value for $\chi^2_{4,0.05}$ can be determined from ***Table 9.2-3***, see ***Figure 9.2-8*** below.

$\chi^2_{4,0.05} = 9.488$

As $\chi^2 = 15.29 > \chi^2_{4,0.05}$ therefore, ***the difference between observed and expected frequencies is significant and they have no good fit***.

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750

Source: Owen, *Handbook of Statistical Tables,* Table A–4 "Chi-Square Distribution Table," © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reproduced by permission of Pearson Education, Inc.

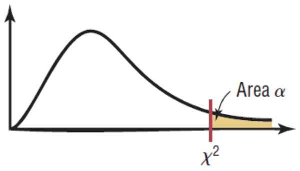


Figure 9.2-8: χ^2 Table for Example 9.2-2.

9.2.1.5 FURTHER NOTES ON THE DEGREE OF FREEDOM

- Most of the proposed models have parameters. For example, the Normal model has two parameters of μ and σ.
- When these parameters are assumed based on previous experience and do not determined from the sample data, the degree of freedom would be $Number\ of\ Classes - 1$. When these parameters are determined from the sample itself, a smaller degree of freedom should be adopted.
- For example, assume that a Normal model is adopted and its mean has been determined based on sample while its standard deviation is assumed based on a previous experience, the degree of freedom would be $Number\ of\ Classes - 2$.
  On the other hand, if μ and σ have been estimated based on the sample data, the degree of freedom would be $Number\ of\ Classes - 3$.
- In this chapter, if no otherwise is mentioned, model parameters have been assumed to be determined based on a previous experience.

9.2.1.6 HOMEWORK PROBLEMS

Problem 9.2-1

To test the ***honesty*** or ***fairness*** of a die, it has been rolled for 60 times. The observed frequency and the theorical uniform frequency have been presented in the indicated table. Use the χ^2 test with $\alpha = 0.05$ to check the honesty of this die.

	Observed Freq.	*Expected Freq.*
Face	O	E
1	6	10
2	16	10
3	7	10
4	4	10
5	17	10
6	11	10

Ans.

$\because \chi^2 = 14.7 > \chi^2_{5,0.05} = 11.07$

Therefore, the die is dishonest or unfair.

Problem 9.2-2

Severe thunderstorms have been recorded at a given station over a period of 66 years. During the period, the frequencies of severe thunderstorms observed are as indicated in the table.

Thunderstorms Number	*Observed Freq.*	*Expected Freq.*	
	O	E	(O-E)
0	20	19.94	0.06
1	23	23.87	-0.87
2	15	14.29	0.71
3	8	7.9	-1.9

A Poisson distribution with mean annual occurrence of $\mu = 1.197$ with the indicated expected frequencies has been used to model the phenomenon. Use the χ^2 test with $\alpha = 0.05$ to test the goodness of the proposed model.

Ans.

$\because \chi^2 = 0.0684 < \chi^2_{3,0.05} = 7.81$

Therefore, the proposed Poisson model has a good fit.

Problem 9.2-3

The histogram for concrete crushing strength with two proposed models of Normal and Lognormal have been presented in ***Figure 9.2-9***.

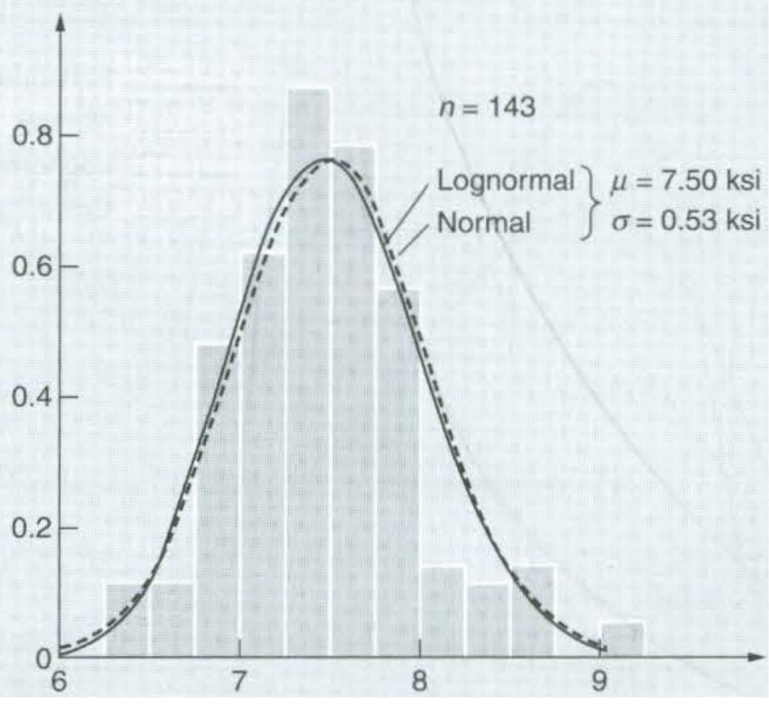


Figure 9.2-9: Histogram of crushing strength of concrete cubes with two proposed models for Problem 9.2-3.

Observed and theoretical frequencies for the Normal model have been presented in the table, use the χ^2 test with $\alpha = 0.05$ to check the goodness of the Normal model.

Ans.

$\because \chi^2 = 10.71 < \chi^2_{6,0.05} = 12.59$

Therefore, the proposed Normal model has a good fit.

Compressive Strength in ksi		**Observed Frequency**	**Expected Frequency**
Lower Limit	Upper Limit	O	E
<6.75		9	11.1
6.75	7	17	13.2
7	7.25	22	21.1
7.25	7.5	31	26.1
7.5	7.75	28	26.1
7.75	8	20	21
8	8.25	9	20.2
	>8.50	7	4.2

Problem 9.2-4

Resolve the ***Problem 9.2-3*** above with using the theoretical frequencies from the Lognormal model as indicated in the table. Which model seems more match and representative.

Ans.

$\because \chi^2 = 7.70 < \chi^2_{6,0.05} = 12.59$

Therefore, the proposed Lognormal model has a good fit.

Compressive Strength in ksi		**Observed Frequency**	**Expected Frequency**
Lower Limit	Upper Limit	O	E
<6.75		9	9.9
6.75	7	17	14
7	7.25	22	22.1
7.25	7.5	31	26.9
7.5	7.75	28	25.6
7.75	8	20	19.8
8	8.25	9	19.4
	>8.50	7	5.3

As $\chi^2_{For\ Lognormal} = 7.70 < \chi^2_{Normal} = 10.71$, therefore the lognormal model is superior to the normal model.

9.5 CONTENTS

CHAPTER 10

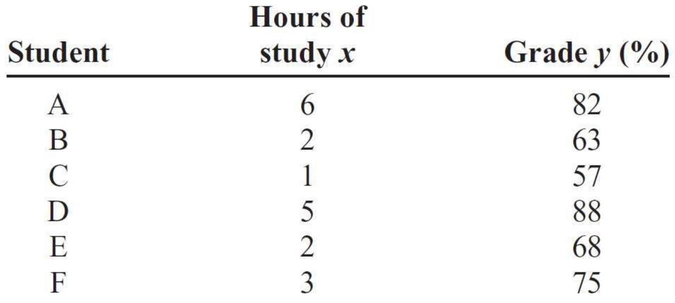
CORRELATION AND REGRESSION

10.1 INTRODUCTION

- Statistics involves determining whether a relationship exists between two or more numerical or quantitative variables.
- For example,
  - A businessperson may want to know whether the volume of sales for a given month is related to the amount of advertising the firm does that month.
  - Educators are interested in determining whether the number of hours a student studies is related to the student's score on a particular exam.
  - Medical researchers are interested in questions such as, is caffeine related to heart damage? or Is there a relationship between a person's age and his or her blood pressure?
  - A zoologist may want to know whether the birth weight of a certain animal is related to its life span.
  - A civil engineer interests whether the first-crack load and the failure load of a beam are related.
- These are only a few of the many questions that can be answered by using the techniques of ***correlation*** and ***regression*** analysis.
- Correlation ***is a statistical method used to determine whether a linear relationship between variables exists***.
- Regression ***is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear***.
- The purpose of this Chapter is to present ***meaning***, ***calculations***, and ***applications*** of ***Correlation*** and ***Regression***.

10.2 SCATTER PLOT

- In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables.
- For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of students, determine the hours each studied, and obtain their grades on the exam. A table can be made for the data, as shown here.



Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

10.2.1 INDEPENDENT AND DEPENDENT VARIABLES

- In above examples:
  - The number of hours of study is the ***independent variable*** and is designated as the ***x variable***.
  - The ***dependent variable*** is the ***variable that cannot be controlled or manipulated***.
  - The grade the student received on the exam is the dependent variable, designated as the ***y variable***.
  - Implicit assumption regarding to independent and dependent variables:

    The reason for this distinction between the variables is that:
    - You assume that the grade the student earns depends on the number of hours the student studied.
    - Also, you assume that, to some extent, the student can regulate or control the number of hours he or she studies for the exam.
  - Implicit assumption is not always clear:

    The determination of the x and y variables is not always clear-cut and is sometimes an arbitrary decision.
  - An example with a clear implicit assumption:

    For example, if a researcher studies the effects of age on a person’s blood pressure, the researcher can generally assume that age affects blood pressure. Hence, the variable age can be called the independent variable, and the variable blood pressure can be called the dependent variable.
  - An example with a not clear implicit assumption:

    On the other hand, if a researcher is studying the attitudes of husbands on a certain issue and the attitudes of their wives on the same issue, it is difficult to say which variable is the independent variable and which is the dependent variable.

    In this study, the researcher can arbitrarily designate the variables as independent and dependent.
- Axes for Independent and Dependent Variables:
  - The independent and dependent variables can be plotted on a graph called a ***scatter plot***.

- The ***independent variable x is plotted on the horizontal axis***, and ***the dependent variable y is plotted on the vertical axis***.

10.2.2 Formal Definition of Scatter Diagram

A ***scatter plot*** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y.

10.2.3 Examples

Example 10.2-1

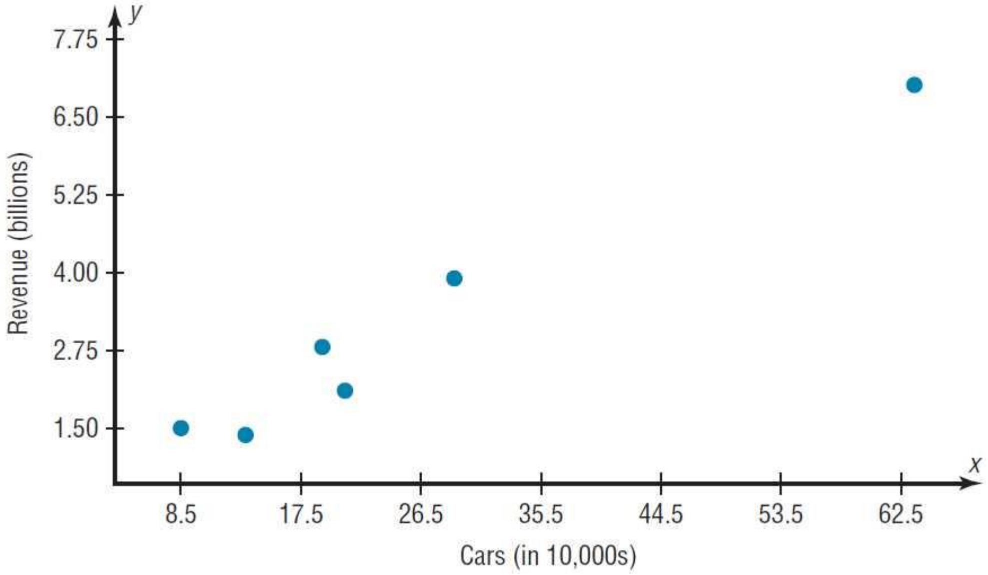
Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure below.



The plot shown in figure above suggests a positive relationship, since as the number of cars rented increases, revenue tends to increase also.

Example 10.2-2

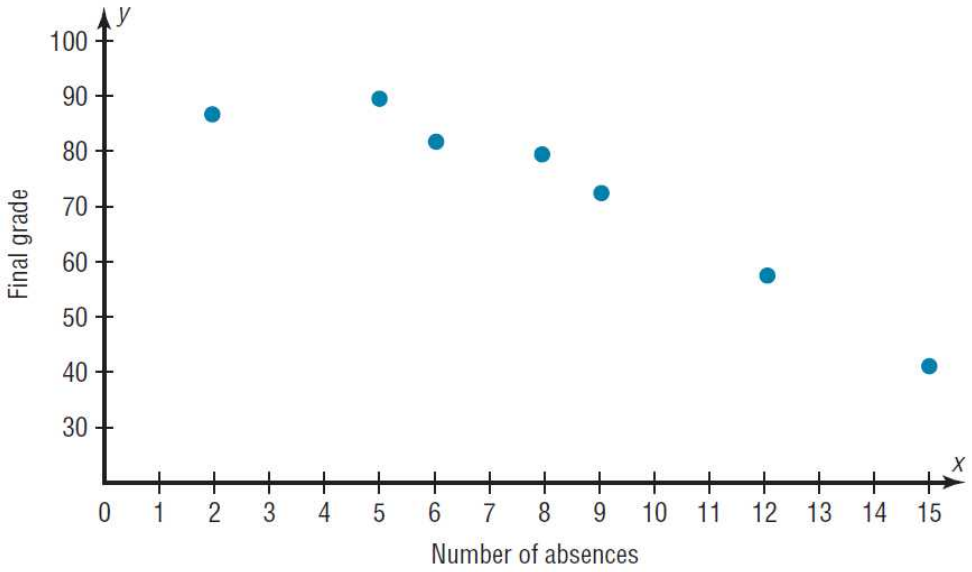
Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure below.



The plot of the data shown in Figure above suggests a negative relationship, since as the number of absences increases, the final grade decreases.

Example 10.2-3

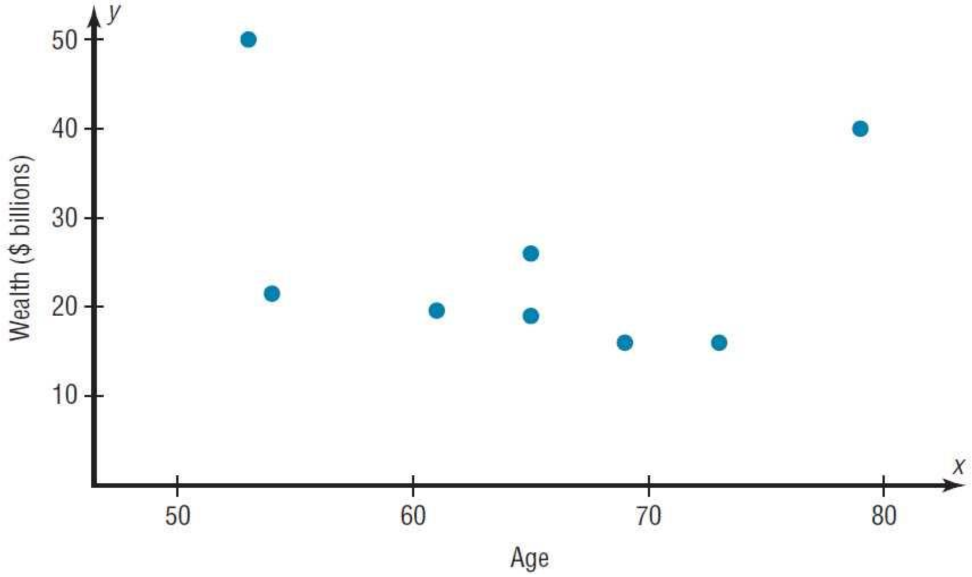
A researcher wishes to see if there is a relationship between the ages and net worth of the wealthiest people in America. The data for a specific year are shown.

Person	Age x	Net wealth y (\$ billions)
A	73	16
B	65	26
C	53	50
D	54	21.5
E	79	40
F	69	16
G	61	19.6
H	65	19

Solution

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure below.



The plot of the data shown in Figure above shows no specific type of relationship, since no pattern is visible.

10.3 Correlation Coefficient, Descriptive Aspects

10.3.1 Definition

- Correlation coefficient is a measure used to determine the strength of the ***linear relationship between two variables***.
- Symbols
  - The sample correlation coefficient is r.
  - The population correlation coefficient is ρ (Greek letter rho).
- There are several types of correlation coefficients. The one explained in this section is called the ***Pearson product moment correlation coefficient (PPMC)***, named after statistician Karl Pearson, who pioneered the research in this area.

Karl Pearson (27 March 1857 – 27 April 1936) was an English ***mathematician*** and ***biostatistician***. He has been credited with establishing the discipline of ***mathematical statistics***. He founded the world's ***first university statistics*** department at ***University College London*** in 1911 and contributed significantly to the field of ***biometrics*** and ***meteorology***.

10.3.2 Range of the Correlation Coefficient

- The range of the correlation coefficient is from -1 to +1.
- If there is a strong positive linear relationship between the variables, the value of r will be close to +1.
- If there is a strong negative linear relationship between the variables, the value of r will be close to -1.
- When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.

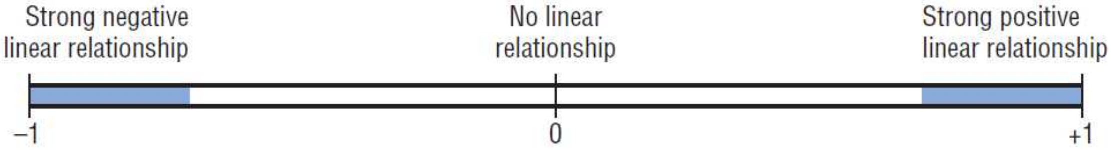

Strong negative linear relationship
No linear relationship
Strong positive linear relationship
−1
0
+1

Figure 10.3-1: Range for correlation coefficient, r.

10.3.3 Relationship between the Correlation Coefficients and their Corresponding Scatter Plots

The graphs in ***Figure 10.3-2*** below shows the relationship between the correlation coefficients and their corresponding scatter plots. Notice that:

- As the value of the correlation coefficient increases from 0 to +1 (parts a, b, and c), data values become closer to an increasingly strong relationship. ***It is useful to note that*** $\mathrm{r}=1$ ***means that all points are located exactly on a line with positive slope, but it says nothing more about specific value of the positive slope***.
- As the value of the correlation coefficient decreases from 0 to -1 (parts d, e, and f), the data values also become closer to a straight line. ***The*** $\mathrm{r}=-1$ ***means that all points are located exactly on a line with negative slope, but it says nothing more about specific value of the negative slope***.
- In general, data with $r=+1$ or $r=-1$ are ***perfectly correlated data that are rare or even impossible to occur in the physical world***.

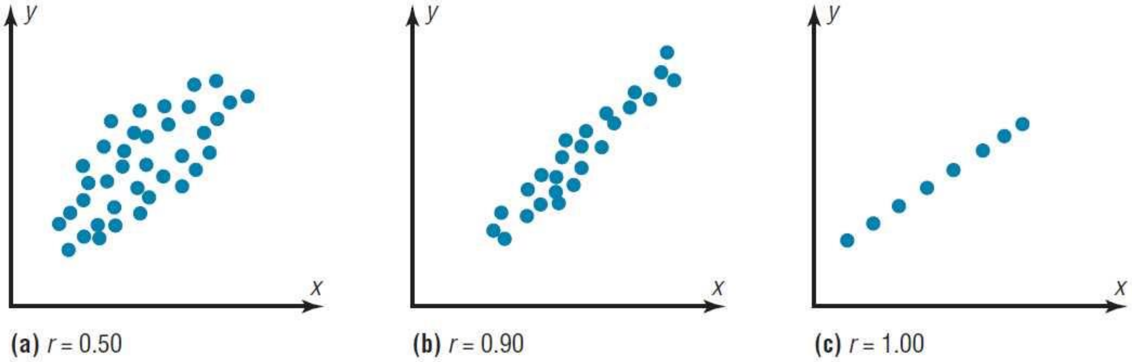


Figure 10.3-2: Different scatter plates and corresponding correlation coefficients, r.

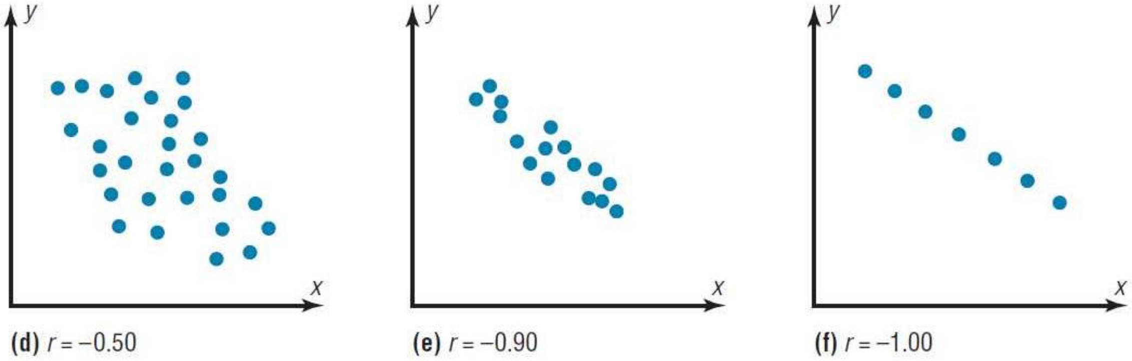


Figure 10.3-2: Different scatter plates and corresponding correlation coefficients, r. Continued.

10.3.4 Formula for the Correlation Coefficient

- There are several ways to compute the value of the correlation coefficient. One method is to use ***Pearson product moment correlation coefficient (PPMC)*** formula shown here.

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) \qquad \textbf{Eq. 10.3-1}$$

- Terms $\left(\frac{x-\bar{x}}{s_x}\right)$ and $\left(\frac{y-\bar{y}}{s_y}\right)$ represents standard scores for x and y respectively.
- Notes on the Formula
  - The best way to discover the direct or indirect linear relation between two variables is to multiply their corresponding values, hence Pearson
    - used product and no other algebraic operation in the relation of ***Eq. 10.3-1*** above.
    - used the standard scores in his formula to avoid possible misleading due to different scales and different order of the correlated variables,.
    - normalized his formula in such a way that it would be bounded between $+1$ and -1.
  - It measures the strength of ***linear*** relations. To have a linear relation, a scatter diagram shall have a ***football shape*** as indicated in ***Figure 10.3-2***.
  - Linear relation may be significantly affected by ***outlier*** if any.
  - The correlation coefficient is ***nondimensional value*** as it has been written in term of standard scores that are nondimensional in nature.
  - To test that the maximum value of r is $+1$, try to substitute same variable for x, and y. You would note:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{x-\bar{x}}{s_x}\right)\right) = \frac{1}{s_x^2} \times \frac{\Sigma(x-\bar{x})^2}{n-1}$$

But, as discussed in ***Chapter 3***,

$$s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

Therefore, the correlation coefficient, r, would be:

$$r = \frac{1}{s_x^2} \times s_x^2 = 1$$

As there is no relation stronger than the relation between a variable and itself, hence there is no r value greater than $+1$. In same reasoning, one can show that the minimum r value is -1.

- As the relation has been written in terms of standard scores, therefore it is ***insensitive for changing in scaling***, and units changing, see ***Example 10.3-7*** below.
- It is ***insensitive to the order of variables***, i.e. any one of two variables can be selected as x variable or as y variable, see ***Example 10.3-9*** below.

- The formula looks somewhat complicated, but using a table to compute the values, as shown in examples, makes it somewhat easier to determine the value of r.

10.3.5 Second Form of Correlation Formula

- With an algebraic manipulation, correlation formula ***Eq. 10.3-1*** above can be rewritten as indicated in below:

$$r = \frac{\left(\frac{1}{n-1}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}}{s_x s_y} \qquad \textbf{Eq. 10.3-2}$$

- The formula ***Eq. 10.3-2*** is totally equivalent to ***Eq. 10.3-1*** and it may be simpler from computational point of view.

10.3.6 Rounding Rule for the Correlation Coefficient Round

The value of r to three decimal places.

10.3.7 Examples

Example 10.3-1

Show that when data is exactly located on a line, the correlation coefficient, r, of ***Eq. 10.3-1*** would be ± 1.

Solution

To be located on a line, data should fallow:

$y_i = k_1 + k_2 x_i$ **(a)**

From the ***linear transformation*** of ***Chapter 3*** it is known:

$\bar{y} = k_1 + k_2 \bar{x}$ **(b)**

$s_y = |k_2| s_x$ **(c)**

Substitute ***Eq. a***, ***Eq. b***, and ***Eq. c*** into ***Eq. 10.3-1*** to have:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{\cancel{k_1}+k_2x-(\cancel{k_1}+k_2\bar{x})}{|k_2|s_x}\right)\right)$$

$$= \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{k_2x-k_2\bar{x}}{|k_2|s_x}\right)\right) = \frac{\mathrm{k}_2}{|k_2|}\left(\frac{1}{(n-1)}\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{x-\bar{x}}{s_x}\right)\right)$$

$$= \frac{\mathrm{k}_2}{|k_2|}\left(\frac{1}{s_x^2}\times\frac{\sum(x-\bar{x})^2}{(n-1)}\right) = \frac{\mathrm{k}_2}{|k_2|}\left(\frac{s_x^2}{s_x^2}\right) = \frac{\mathrm{k}_2}{|k_2|}$$

Therefore, when data located on a line with slope of k_2, the correlation coefficient would be:

$$r = \frac{\mathrm{k}_2}{|k_2|}$$

When the slope is positive, $|k_2| = k_2$ and the correlation coefficient would be $r = +1$. While when the slope is negative, $|k_2| = -k_2$ and the correlation coefficient would be $r = -1$. The poof shows the fact implicitly indicated in ***Eq. 10.3-1***, i.e. the line intercept, k_1, has no effect on the correlation coefficient, r.

Example 10.3-2

Plot the scatter diagram and compute the correlation coefficient for the data shown for car rental companies in the United States for a recent year.

Table 10.3-1: Data for Example 10.3-2.

Company	Cars (in ten thousand), x	Revenue (in billion), y
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Solution

The scatter plot for these data is shown in ***Figure 10.3-3*** below. The plot indicates that r would be positive and relatively close to 1. This indication is useful for a cross check of the calculations of r.

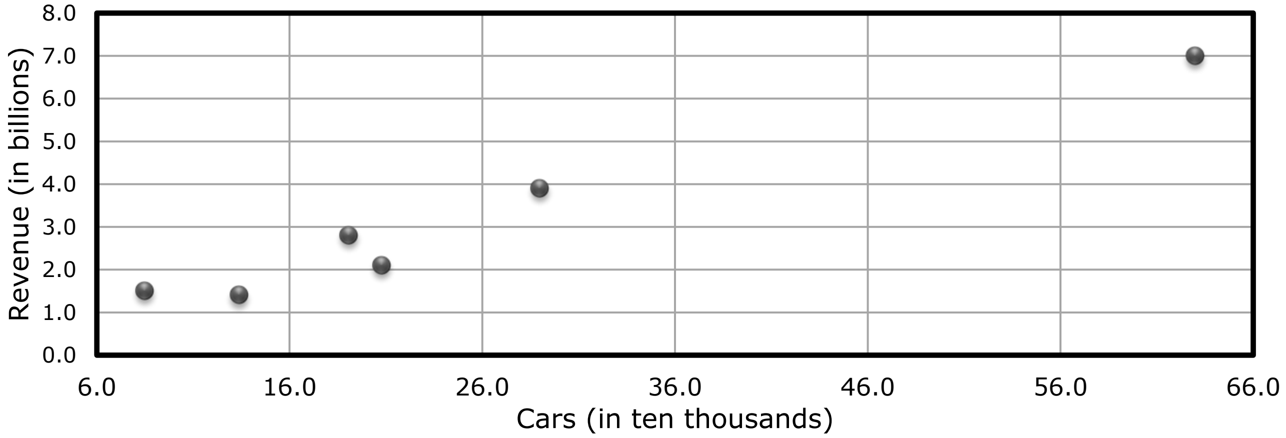


Figure 10.3-3: Scatter plot for Example 10.3-2.

Compute,

$$\bar{x}=\frac{\Sigma x}{n}=25.6, \bar{y}=\frac{\Sigma y}{n}=3.1$$

$$s_x=\sqrt{\frac{(x-\bar{x})^2}{n-1}}=19.6, s_y=\sqrt{\frac{(y-\bar{y})^2}{n-1}}=2.12$$

Standard scores for x and y and summation of their multiplication are shown in ***Table 10.3-2*** below. The correlation coefficient, r, is:

$$r=\frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right)=\frac{4.910}{6-1}=0.982$$

The value of r confirms positive relation noticed in the scatter plot of data.

Table 10.3-2: Calculation steps of r of Example 10.3-2.

Company	Cars (in ten thousand), x	Revenue (in billions), y	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
A	63.0	7.0	1.91	1.8	3.5
B	29.0	3.9	0.17	0.4	0.1
C	20.8	2.1	-0.25	-0.5	0.1
D	19.1	2.8	-0.33	-0.1	0.0
E	13.4	1.4	-0.62	-0.8	0.5
F	8.5	1.5	-0.88	-0.8	0.7
				Σ	4.9

Example 10.3-3

Plot the scatter diagram and compute the correlation coefficient for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

Table 10.3-3: Data for Example 10.3-3.

Student	Number of Absences, x	Final Grades, y
A	6.0	82.0
B	2.0	86.0
C	15.0	43.0
D	9.0	74.0
E	12.0	58.0
F	5.0	90.0
G	8.0	78.0

Solution

Scatter plot for data is presented in ***Figure 10.3-4***. The figure indicates that r would be negative and close to -1.

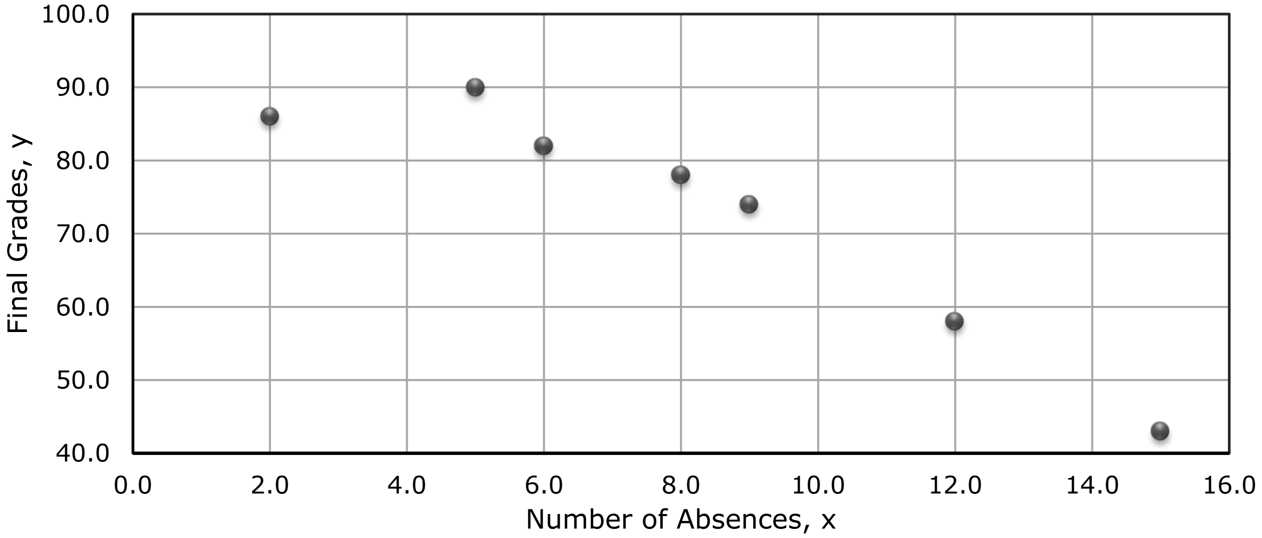


Figure 10.3-4: Scatter plot for Example 10.3-3.

Compute,

$$\bar{x} = \frac{\Sigma x}{n} = 8.1, \bar{y} = \frac{\Sigma y}{n} = 73.0$$

$$s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = 4.37, s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = 16.78$$

Standard scores for x and y and summation of their multiplication are shown in Table below. The correlation coefficient, r, is:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{-5.67}{7-1} = -0.945$$

The value of r confirms negative relation noticed in the scatter plot of data.

Table 10.3-4: Calculation steps of r of *Example 10.3-3*.

STUDENT	NUMBER OF ABSENCES, X	FINAL GRADES, Y	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
A	6.0	82.0	-0.49	0.5	-0.3
B	2.0	86.0	-1.40	0.8	-1.1
C	15.0	43.0	1.57	-1.8	-2.8
D	9.0	74.0	0.20	0.1	0.0
E	12.0	58.0	0.88	-0.9	-0.8
F	5.0	90.0	-0.72	1.0	-0.7
G	8.0	78.0	-0.03	0.3	0.0
				Σ	-5.67

Example 10.3-4

Plot the scatter diagram and compute the correlation coefficient for the data obtained in a study to see if there is a relationship between the ages and net worth of the wealthiest people in America. The data for a specific year are shown.

Table 10.3-5: Data for Example 10.3-4.

Person	*Age, x*	*Net wealth, y, ($ Billions)*
A	73	16.0
B	65	26.0
C	53	50.0
D	54	21.5
E	79	40.0
F	69	16.0
G	61	19.6
H	65	19.0

Solution

Scatter plot for data is presented in ***Figure 10.3-5*** below. The figure indicates no clear linear relation.

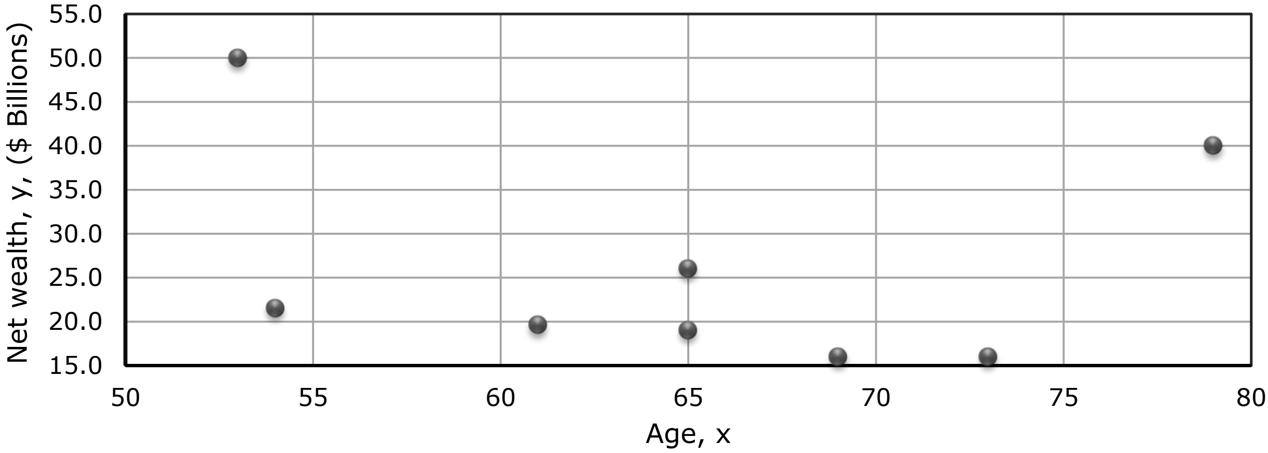


Figure 10.3-5: Scatter plot for Example 10.3-4.

Compute,

$$\bar{x} = \frac{\Sigma x}{n} = 64.9, \bar{y} = \frac{\Sigma y}{n} = 26.0$$

$$s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = 8.92, s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = 12.43$$

Standard scores for x and y and summation of their multiplication are shown in Table below. The correlation coefficient, r, is:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{-1.23}{8-1} = -0.176$$

As concluded from scatter plot, correlation coefficient, r, indicates that there is no clear relation between two variables.

Table 10.3-6: Calculation steps of r of Example 10.3-4.

Person	Age, x	Net wealth, y, ($ Billions)	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
A	73	16.0	0.91	-0.8	-0.7
B	65	26.0	0.01	0.0	0.0
C	53	50.0	-1.33	1.9	-2.6
D	54	21.5	-1.22	-0.4	0.4
E	79	40.0	1.58	1.1	1.8
F	69	16.0	0.46	-0.8	-0.4
G	61	19.6	-0.43	-0.5	0.2
H	65	19.0	0.01	-0.6	0.0
				Σ	-1.23

10.3.8 Civil Engineering Examples

Example 10.3-5

Table 10.3-7: Concrete compressive concrete, cubical and corresponding cylindrical specimens, for Example 10.3-5.

Cube compressive strength, $\boldsymbol{f_{cu}}$, MPa	Cylinder compressive strength, $\boldsymbol{f_c}'$, MPa
9.00	6.90
15.20	11.70
20.00	15.20
24.80	20.00
27.60	24.10
29.00	26.20
29.60	26.90
35.80	31.70
36.50	34.50
42.10	36.50
44.10	40.70
48.30	44.10
52.40	50.30

Concrete compressive strength is either estimated based on a cubical specimen with side of 150mm or based on cylindrical specimen with dimeter of 150mm and height of 300mm. British, Euro, Indian, and many other specifications adopt the cubical specimen to estimate concrete compressive strength, while American standards adopt the cylindrical specimen. ***Table 10.3-7*** above shows cubical and corresponding cylindrical compressive strength for 13 specimens. Draw scatter diagram and determine the correlation coefficient, r, to show if there is a linear relationship between concrete compressive strength based on cubical and cylindrical specimen.

Solution

The scatter diagram in ***Figure 10.3-6*** below shows that there is a strong positive relationship between cubical and corresponding cylindrical compressive strengths of concrete. In Iraq, the indicated linear relation is useful to transform the cubical strength, f_{cu}, that is obtained from testing to the cylindrical strength, f_c', adopted in design. This may interpret adopting cubical strength as x value.

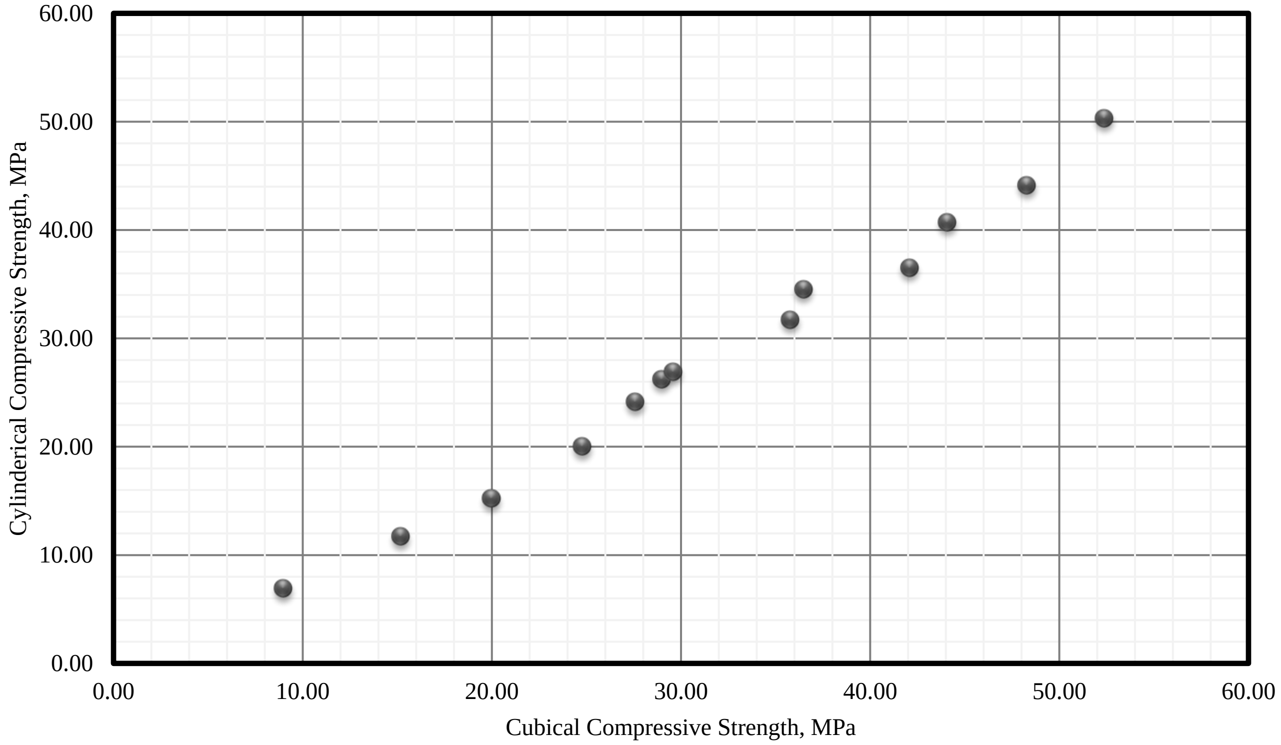


Figure 10.3-6: Scatter diagram for cylindrical versus cubical compressive strength of the concrete of Example 10.3-5.

Standard scores for x and y and summation of their multiplication are shown in ***Table 10.3-8*** below. The correlation coefficient, r, is:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{11.95}{13-1} = 0.99$$

The value of r confirms positive relation noticed in the scatter plot of data.

Table 10.3-8: Calculation steps of r of Example 10.3-5.

	$x = f_{cu}\,MPa$	$y = f'_c, MPa$	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
	9.00	6.90	-1.77	-1.66	2.94
	15.20	11.70	-1.29	-1.29	1.66
	20.00	15.20	-0.92	-1.02	0.94
	24.80	20.00	-0.55	-0.65	0.35
	27.60	24.10	-0.33	-0.33	0.11
	29.00	26.20	-0.22	-0.17	0.04
	29.60	26.90	-0.18	-0.11	0.02
	35.80	31.70	0.30	0.26	0.08
	36.50	34.50	0.36	0.47	0.17
	42.10	36.50	0.79	0.63	0.50
	44.10	40.70	0.95	0.95	0.90
	48.30	44.10	1.27	1.22	1.55
	52.40	50.30	1.59	1.70	2.69
Mean values	31.9	28.4		∑	11.95
Standard deviation	12.9	12.9			

Example 10.3-6

Ten timber beams were tested on a span of 4 ft by a single concentrated load at mid-span to determine if a relationship exists between deflection at working load and ultimate load. Draw scatter diagram, compute correlation coefficient, r, and comment on possible relation.

Table 10.3-9: Data for Example 10.3-6.

Specimen	Deflection at Working load, in.	Ultimate Load, lb
1	0.160	1750
2	0.130	2350
3	0.155	2050
4	0.134	2100
5	0.135	1525
6	0.123	2000
7	0.168	1450
8	0.130	2100
9	0.150	1475
10	0.132	1675

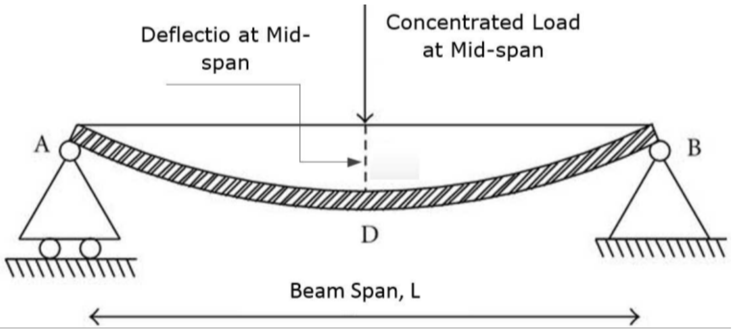


Figure 10.3-7: Beam for Example 10.3-6.

Solution

Scatter diagram for ultimate load and deflection is presented below.

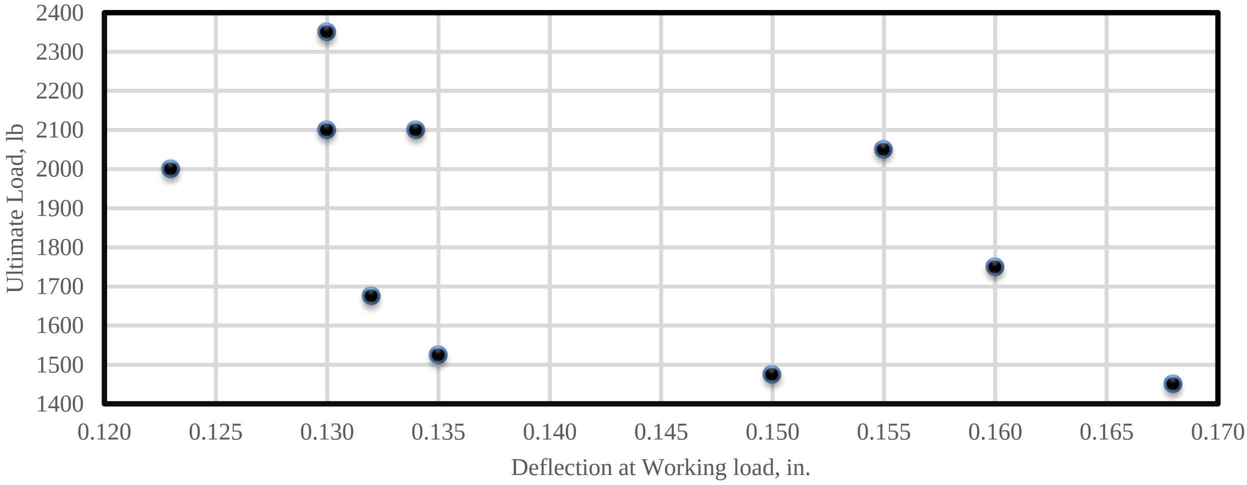


Figure 10.3-8: Scatter plot for Example 10.3-6.

$$\bar{x} = \frac{\Sigma x}{n} = 0.1417, \qquad \bar{y} = \frac{\Sigma y}{n} = 1847.5$$

$$s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = 0.015, \qquad s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = 313.5$$

Standard scores for x and y and summation of their multiplication are shown in Table below. The correlation coefficient, r, is:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{-4.69}{10-1} = -0.521$$

Unfortunately, there is no clear relation between deflection at service load and ultimate load. If this relation exists, nondestructive deflection test can be used to estimate beam ultimate load.

Table 10.3-10: Calculation steps of r of Example 10.3-6.

Deflection at Working load, in.	Ultimate Load, lb	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
0.160	1750	1.200	-0.311045	-0.37316
0.130	2350	-0.767	1.6030772	-1.2296
0.155	2050	0.872	0.6460162	0.563271
0.134	2100	-0.505	0.8055264	-0.40662
0.135	1525	-0.439	-1.028841	0.451903
0.123	2000	-1.226	0.486506	-0.59642
0.168	1450	1.724	-1.268106	-2.18642
0.130	2100	-0.767	0.8055264	-0.61786
0.150	1475	0.544	-1.188351	-0.64661
0.132	1675	-0.636	-0.55031	0.349947
			Summation	-4.69157

Example 10.3-7

To show effect of scaling, changing units, resolve previous example after changing units for ultimate load and deflection into kN and mm respectively.

Solution

It can be shown that:

$1\ inch = 25.4\ mm$

$1\ lb = 4.45 \times 10^{-3}\ N$

With these units, data would be and indicated in ***Table 10.3-11***.

$$\bar{x} = \frac{\Sigma x}{n} = 3.599\ mm, \qquad \bar{y} = \frac{\Sigma y}{n} = 8.238\ kN$$

$$s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = 0.381\ mm, \qquad s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = 1.398\ kN$$

Standard scores for x and y and summation of their multiplication are shown in ***Table 10.3-12***. It useful to note that standard scores for x, $\left(\frac{x-\bar{x}}{s_x}\right)$, and for y, $\left(\frac{y-\bar{y}}{s_y}\right)$, are both dimensionless quantity and have same values irrespective of the unit system. Therefore, their summation, $\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$, and in turn r value still the same irrespective of adopted unit system,

Table 10.3-11: Metric data for Example 10.3-7.

Deflection at Working load, mm	Ultimate Load, kN
4.06	7.80
3.30	10.48
3.94	9.14
3.40	9.36
3.43	6.80
3.12	8.92
4.27	6.47
3.30	9.36
3.81	6.58
3.35	7.47

Table 10.3-12: Calculation steps of r of Example 10.3-7.

Deflection at Working load, mm	Ultimate Load, kN	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
4.06	7.80	1.200	-0.311045	-0.37316
3.30	10.48	-0.767	1.6030772	-1.2296
3.94	9.14	0.872	0.6460162	0.563271
3.40	9.36	-0.505	0.8055264	-0.40662

Deflection at Working load, mm	Ultimate Load, kN	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
3.43	6.80	-0.439	-1.028841	0.451903
3.12	8.92	-1.226	0.486506	-0.59642
4.27	6.47	1.724	-1.268106	-2.18642
3.30	9.36	-0.767	0.8055264	-0.61786
3.81	6.58	0.544	-1.188351	-0.64661
3.35	7.47	-0.636	-0.55031	0.349947
			Summation	-4.69157

The correlation coefficient, r, is:

$$r=\frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right)=\frac{-4.69}{10-1}=-0.521$$

Example 10.3-8

Measurements indicated in ***Table 10.3-13*** taken on farmlands of the amounts of soil washed away by erosion suggest a relationship with flow rates. Draw a scatter plot, and compute correlation coefficient for the data. Comment on the results.

Table 10.3-13: Data for Example 10.3-8.

Flow (L/s)	Soil eroded (kg)
0.31	0.82
0.85	1.95
1.26	2.18
2.47	3.01
3.75	6.07

Solution

Scatter diagram for ultimate load and deflection is presented below.

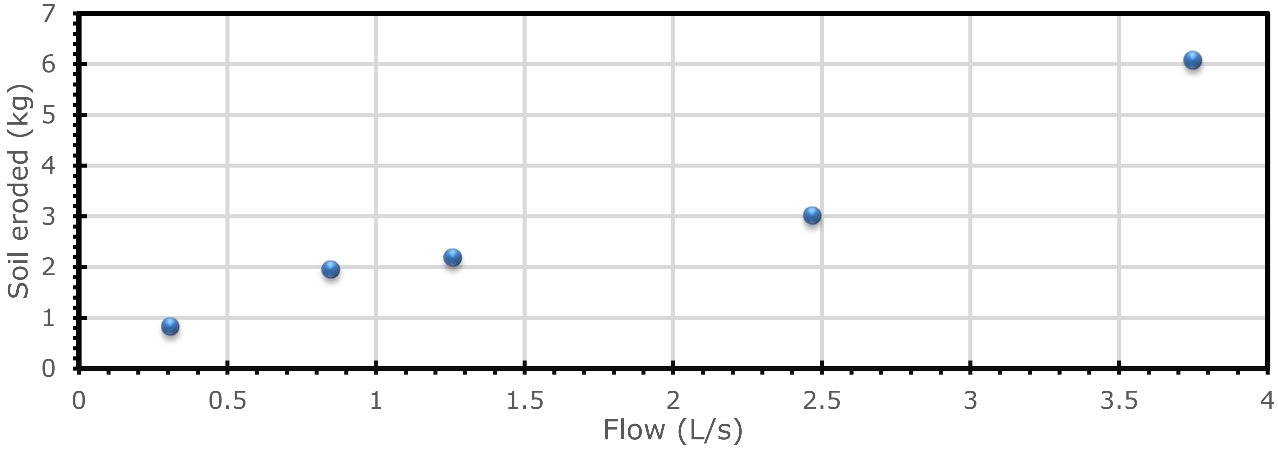


Figure 10.3-9: Scatter plot for Example 10.3-8.

$$\bar{x}=\frac{\Sigma x}{n}=1.728,\qquad \bar{y}=\frac{\Sigma y}{n}=2.806$$

$$s_x=\sqrt{\frac{(x-\bar{x})^2}{n-1}}=1.382,\qquad s_y=\sqrt{\frac{(y-\bar{y})^2}{n-1}}=1.985$$

Standard scores for x and y and summation of their multiplication are shown in Table below. The correlation coefficient, r, is:

$$r=\frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right)=\frac{3.87}{5-1}=0.967$$

Table 10.3-14: Calculation steps of r of Example 10.3-8.

Flow (L/s)	Soil eroded (kg)	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
0.31	0.82	-1.026050278	-1.00040159	1.026462331
0.85	1.95	-0.635311809	-0.43119021	0.273940234
1.26	2.18	-0.338640007	-0.31533303	0.106784379
2.47	3.01	0.536903601	0.102760284	0.055172367
3.75	6.07	1.463098493	1.644164548	2.405574673
			Σ	3.867933984

From the diagram and correlation coefficient, one concludes that there is a clear relation, which can be used in future to estimate amount of soil eroded for a specific water flow.

Example 10.3-9

To show effects of selection independent variable, x, and dependent variable, y, on strength of linear relationship between two variables expressed in terms of correlation coefficient, r, resolve previous example with adopting soil eroded as independent variable, x, and water flow as dependent variable, y.

Solution

As switching between independent variable and dependent variable, only change the order of standard scores in the relation, then it has no effect on strength of linear relation between two variables, i.e. the value of r.

Then for purpose of computing correlation coefficient, r, it does not a matter which variable is selected as independent and which one is dependent.

10.3.9 Software Oriented Examples

- When data size is relatively large, the correlation coefficient can be determined with a software aid.
- Excel software can be used to determine the correlation coefficient in one of the following two approaches:
  - Its spreadsheets can be used to prepared the necessary table to compute the correlation coefficient, r, that can subsequently be determined based on ***Eq. 10.3-1*** or ***Eq. 10.3-2***.
  - It can be used directly to determine the correlation coefficient based on adding a ***trend*** to a plotted scatter diagram.
- The following examples show the two aforementioned approach.

Example 10.3-10

For data in ***Table 10.3-15***, which represents the density and compressive strength at 28 days draw a scatter diagram; compute correlation coefficient, r, and comments on possible relation between density and compressive strength.

Table 10.3-15: Data for Example 10.3-10.

Date	Density (kg/m^3)	Compressive strength (MPa)
21 September 1992	2437	60.5
29 June 1992	2437	60.9
26 June 1992	2425	59.8
14 April 1992	2427	53.4
31 March 1992	2428	56.9
19 March 1992	2448	67.3
9 March 1992	2456	68.9
7 February 1992	2436	49.9
28 January 1992	2435	57.8
18 December 1991	2446	60.9
6 December 1991	2441	61.9
6 December 1991	2456	67.2
6 December 1991	2444	64.9
5 December 1991	2447	63.4
4 December 1991	2433	60.5
3 December 1991	2429	68.1
2 December 1991	2435	68.3
22 October 1991	2471	65.7
18 October 1991	2472	61.5
14 October 1991	2445	60.0
9 October 1991	2436	59.6
7 October 1991	2450	60.5
3 October 1991	2454	59.8
2 October 1991	2449	56.7
30 September 1991	2441	57.9
27 September 1991	2457	60.2
23 September 1991	2447	55.8
20 September 1991	2436	53.2
17 September 1991	2458	61.1
13 September 1991	2415	50.7
10 September 1991	2448	59.0
9 September 1991	2445	63.3
6 September 1991	2436	52.5
3 September 1991	2469	54.6
2 September 1991	2455	56.3
29 August 1991	2473	64.9
23 August 1991	2488	69.5
12 July 1991	2454	58.9
9 July 1991	2427	54.4
8 July 1991	2411	58.8

Solution

Scatter plot for data presented in ***Figure 10.3-10*** can be drawn with Excel and trend can be added to automatically determined the correlation coefficient, r. As indicated in Figure 10.3-10, Excel gives the square value of the correlation coefficient, indicated with R^2. Therefore, the r would be:

$$r = \sqrt{R^2_{From\ Excel\ Scatter\ Plot}} = \sqrt{0.1905} = 0.436$$

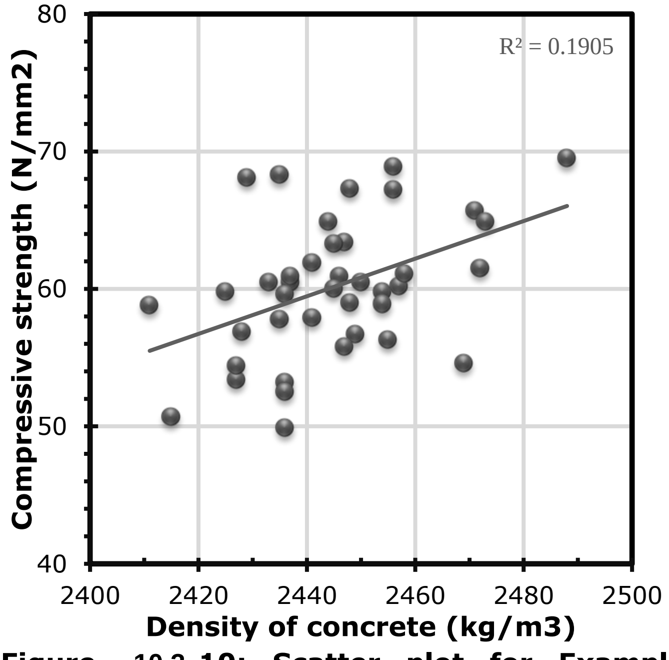


Figure 10.3-10: Scatter plot for Example 10.3-10.

On the other hand, the Excel spreadsheet can be used to prepare and execute the calculations presented in ***Table 10.3-16***.

$$\bar{x} = \frac{\Sigma x}{n} = 2444.9, \bar{y} = \frac{\Sigma y}{n} = 60.14, s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = 15.99, s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = 5.01$$

Standard scores for x and y and summation of their multiplication are shown in ***Table 10.3-16***. The correlation coefficient, r, is:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{17.020}{40-1} = 0.436$$

Table 10.3-16: Calculation steps of r of Example 10.3-10.

Density of Concrete (kg/m3), x	Compressive Strength (N/mm2), y	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
2437	60.5	-0.495491816	0.07228481	-0.035816532
2437	60.9	-0.495491816	0.152047358	-0.075338222
2425	59.8	-1.245763335	-0.06729965	0.083839437
2427	53.4	-1.120718082	-1.343500427	1.505685222
2428	56.9	-1.058195455	-0.645578127	0.68314784
2448	67.3	0.192257077	1.428248135	0.274590811
2456	68.9	0.692438089	1.747298329	1.209895917
2436	49.9	-0.558014442	-2.041422727	1.139143365
2435	57.8	-0.620537069	-0.466112393	0.289240018
2446	60.9	0.067211824	0.152047358	0.01021938
2441	61.9	-0.245401309	0.35145373	-0.086247205
2456	67.2	0.692438089	1.408307498	0.975165753
2444	64.9	-0.05783343	0.949672844	-0.054922838
2447	63.4	0.12973445	0.650563287	0.08440047
2433	60.5	-0.745582322	0.07228481	-0.053894276
2429	68.1	-0.995672828	1.587773232	-1.580902665
2435	68.3	-0.620537069	1.627654507	-1.010019957
2471	65.7	1.630277488	1.109197941	1.808300433
2472	61.5	1.692800115	0.271691181	0.459918862
2445	60	0.004689197	-0.027418376	-0.00012857
2436	59.6	-0.558014442	-0.107180925	0.059808504
2450	60.5	0.31730233	0.07228481	0.022936139
2454	59.8	0.567392836	-0.06729965	-0.038185339
2449	56.7	0.254779703	-0.685459402	-0.174641143
2441	57.9	-0.245401309	-0.446171756	0.109491133
2457	60.2	0.754960716	0.012462898	0.009408999
2447	55.8	0.12973445	-0.864925136	-0.112210587
2436	53.2	-0.558014442	-1.383381702	0.771946969
2458	61.1	0.817483343	0.191928632	0.15689846
2415	50.7	-1.870989601	-1.88189763	3.521010896
2448	59	0.192257077	-0.226824747	-0.043608663
2445	63.3	0.004689197	0.63062265	0.002957114
2436	52.5	-0.558014442	-1.522966162	0.849837113
2469	54.6	1.505232235	-1.104212782	-1.662096673
2455	56.3	0.629915463	-0.76522195	-0.482025139
2473	64.9	1.755322741	0.949672844	1.66698234
2488	69.5	2.69316214	1.866942152	5.027977923
2454	58.9	0.567392836	-0.246765385	-0.140012911
2427	54.4	-1.120718082	-1.144094056	1.282206896
2411	58.8	-2.121080107	-0.266706022	0.565704837
			Σ	17.020

At first sight, there is no well-defined relationship between the two sets of observations although one would expect a density that is higher or lower than average to be associated with a compressive strength of concrete that is correspondingly higher or lower than its average. If a relation exists between concrete density and strength, it would be very useful as it is simpler to measure density than measure the strength especially for existing structures.

Example 10.3-11

Measurements of engineering interest have been recorded during earthquakes in Japan and in other parts of the world since 1800. One of the critical recordings is of apparent relative density, RDEN. After the commencement of a strong earthquake, a saturated fine, loose sand undergoes vibratory motion and consequently the sand may liquefy without retaining any shear strength, thus behaving like a dense liquid. This will lead to failures in structures supported by the liquefied sand. These are often catastrophic. The standard penetration test is used to measure RDEN. Another measurement taken to estimate the prospect of liquefaction is that of the intensity at which the ground shakes. This is the peak surface acceleration of the soil during the earthquake, ACCEL.

Table 10.3-17: Data for Example 10.3-11.

RDEN	ACCEL	RDEN	ACCEL	RDEN	ACCEL
(%)	(units of g)	(%)	(units of g)	(%)	(units of g)
53	0.219	30	0.138	50	0.313
64	0.219	72	0.422	44	0.224
53	0.146	90	0.556	100	0.231
64	0.146	40	0.447	65	0.334
65	0.684	50	0.547	68	0.419
55	0.611	55	0.204	78	0.352
75	0.591	50	0.170	58	0.363
72	0.522	55	0.170	80	0.291
40	0.258	75	0.192	55	0.314
58	0.250	53	0.292	100	0.377
43	0.283	70	0.299	100	0.434
32	0.419	64	0.292	52	0.350
40	0.123	53	0.225	58	0.334

Compute the sample mean $\bar{x}$, and standard deviation s, for RDEN and ACCEL. Plot the scatter diagram and calculate the correlation coefficient r. What conclusions can be reached?

Solution

Excel can be used to draw the scatter diagram shown in Figure 10.3-11 and to add a trend line to indicate the R^2 value. Then the correlation coefficient, r, would be:

$$r = \sqrt{R^2_{From\ Excel\ Scatter\ Plot}} = \sqrt{0.0788} = 0.2807$$

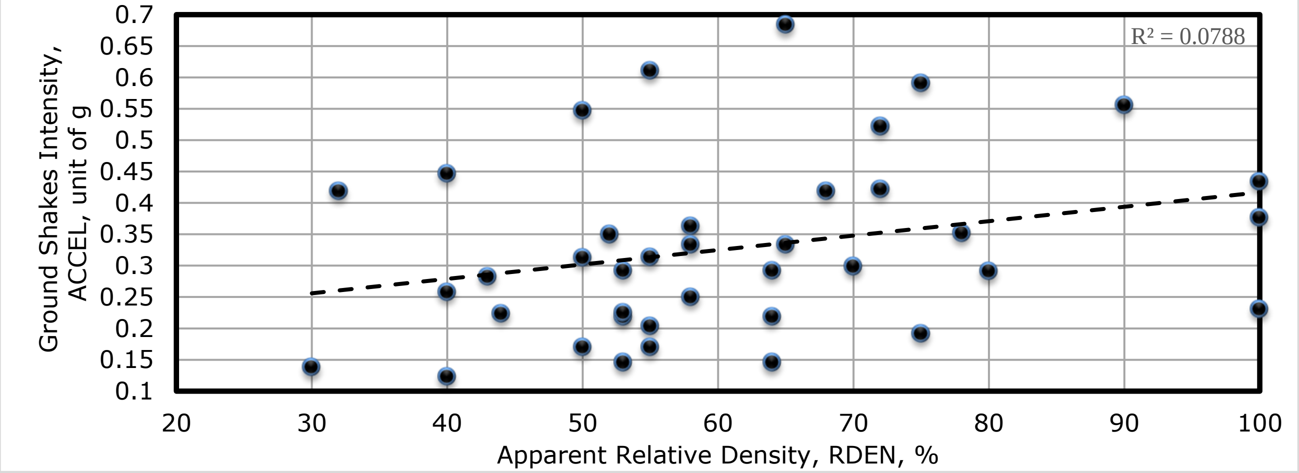


Figure 10.3-11: Scatter plot for Example 10.3-11.

Excel spreadsheet can be used to prepare the calculations indicated in Table 10.3-18 that can be subsequently adopted in computing of r based on ***Eq. 10.3-1***.

$$\bar{x} = \frac{\Sigma x}{n} = 61.0, \bar{y} = \frac{\Sigma y}{n} = 0.3272, s_x = \sqrt{\frac{(x-\bar{x})^2}{n-1}} = 17.39, s_y = \sqrt{\frac{(y-\bar{y})^2}{n-1}} = 0.1423$$

Standard scores for x and y and summation of their multiplication are shown in Table below. The correlation coefficient, r, is:

$$r = \frac{1}{n-1}\left(\Sigma\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right) = \frac{10.7}{39-1} = 0.281$$

Table 10.3-18: Calculation steps of r of Example 10.3-11.

RDEN, %	ACCEL, (unit of g)	$\left(\frac{x-\bar{x}}{s_x}\right)$	$\left(\frac{y-\bar{y}}{s_y}\right)$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
53	0.219	-0.460	-0.760	0.3
64	0.219	0.173	-0.760	-0.1
53	0.146	-0.460	-1.273	0.6
64	0.146	0.173	-1.273	-0.2
65	0.684	0.230	2.506	0.6
55	0.611	-0.345	1.994	-0.7
75	0.591	0.805	1.853	1.5
72	0.522	0.633	1.368	0.9
40	0.258	-1.208	-0.486	0.6
58	0.25	-0.173	-0.542	0.1
43	0.283	-1.035	-0.311	0.3
32	0.419	-1.668	0.645	-1.1
40	0.123	-1.208	-1.434	1.7
30	0.138	-1.783	-1.329	2.4
72	0.422	0.633	0.666	0.4
90	0.556	1.668	1.607	2.7
40	0.447	-1.208	0.842	-1.0
50	0.547	-0.633	1.544	-1.0
55	0.204	-0.345	-0.865	0.3
50	0.17	-0.633	-1.104	0.7
55	0.17	-0.345	-1.104	0.4
75	0.192	0.805	-0.950	-0.8
53	0.292	-0.460	-0.247	0.1
70	0.299	0.518	-0.198	-0.1
64	0.292	0.173	-0.247	0.0
53	0.225	-0.460	-0.718	0.3
50	0.313	-0.633	-0.100	0.1
44	0.224	-0.978	-0.725	0.7
100	0.231	2.243	-0.676	-1.5
65	0.334	0.230	0.048	0.0
68	0.419	0.403	0.645	0.3
78	0.352	0.978	0.174	0.2
58	0.363	-0.173	0.251	0.0
80	0.291	1.093	-0.254	-0.3
55	0.314	-0.345	-0.093	0.0
100	0.377	2.243	0.350	0.8
100	0.434	2.243	0.750	1.7
52	0.35	-0.518	0.160	-0.1
58	0.334	-0.173	0.048	0.0
			Σ	10.7

From scatter plot and r value, one concludes that there is no clear linear relation between apparent relative density of soil and ground shanking intensity.

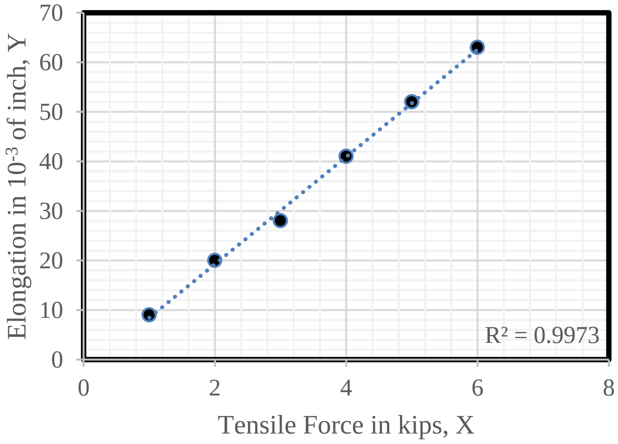
10.3.10 Homework Problems

Home Work 10.3-1

A tensile load test was performed on an aluminum specimen. The applied tensile force and the corresponding elongation of the specimen at various stages of the test are recorded as shown in the table. We may assume that the force- elongation relation over the range of the applied loads is linear. On this basis, determine the correlation coefficient, r, to show if the data satisfies this assumption.

Ans. $r = 0.9986$

Tensile Force in kips, X	Elongation in 10^{-3} of inch, Y
1	9
2	20
3	28
4	41
5	52
6	63



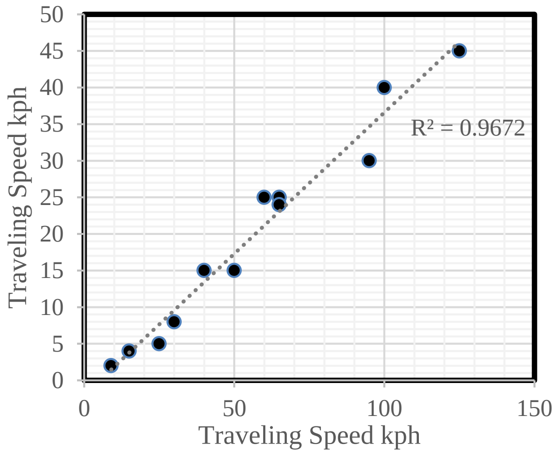
Home Work 10.3-2

The distance Y necessary for stopping a vehicle is a function of the speed of travel of the vehicle X. Suppose the following set of data were observed for 12 vehicles traveling at different speed as shown in the table.

- Plot the stopping distance versus the speed of travel.
- Determine the correlation coefficient factor, r, to show if there is a strong relation or not.

Ans. $r = 0.983$

Vehicle No.	Traveling Speed kph	Stopping Distance, m
1	40	15
2	9	2
3	100	40
4	50	15
5	15	4
6	65	25
7	25	5
8	60	25
9	95	30
10	65	24
11	30	8
12	125	45



10.4 REGRESSION, DESCRIPTIVE ASPECTS

10.4.1 INTRODUCTION

In summary, in studying relationships between two variables following steps are followed:

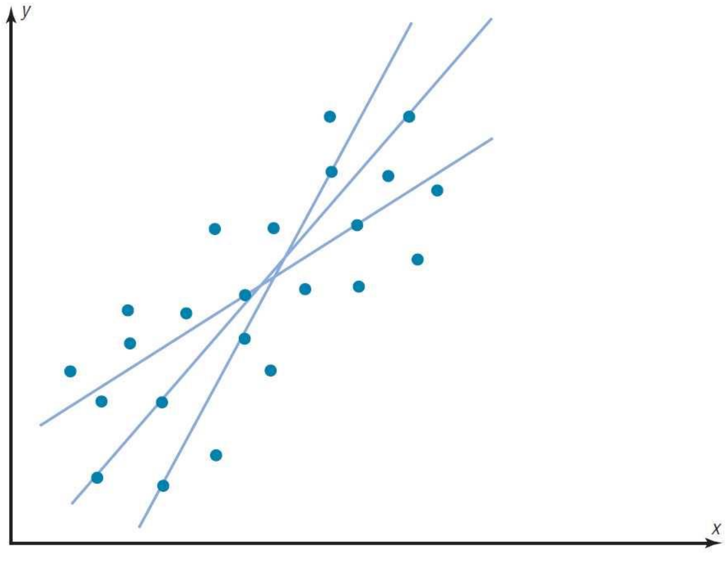
- Collect the data and then construct a scatter plot as presented in ***Article 10.3*** above. The purpose of the scatter plot is to determine the nature of the relationship. The possibilities include:
    - a positive linear relationship,
    - a negative linear relationship,
    - a curvilinear relationship,
    - no discernible relationship.
- After the scatter plot is drawn, the next steps are to compute the value of the correlation coefficient as discussed in ***Article 10.3***.
    - If the value of the correlation coefficient is significant, the next step is to determine the equation of the, best line, usually called ***regression line***, which is the data's line of best fit.
    - Determining the regression line when r ***is not significant and then making predictions using the regression line are meaningless***.

10.4.2 PURPOSE OF REGRESSION LINE

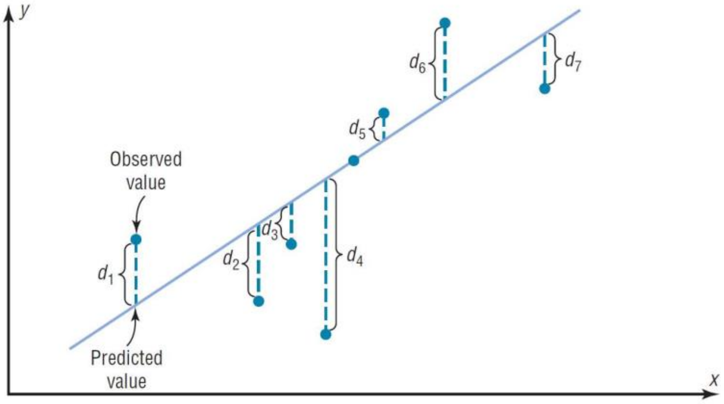
The purpose of the regression line is to enable the researcher to see the trend and make predictions based on the data.

10.4.3 LINE OF BEST FIT

- ***Figure 10.4-1*** shows a scatter plot for the data of two variables. It shows that several lines can be drawn on the graph near the points.
- Given a scatter plot, you must be able to draw the line of best fit.
- ***Best-fit*** means that ***the sum of the squares of the vertical distances from each point to the line is at a minimum*** see ***Figure 10.4-2***.
- Therefore, best-fit line usually computed based on a numerical technique called as ***least squares method***. This method is out of our scope for this freshman course and will be studies in junior course on numerical analysis.
- For time being, it is enough to know that ***regression line passing through*** $\bar{y}$ ***value for each*** x ***value***.



igure 10.4-1: Scatter plot with three ines fit to the data.



igure 10.4-2: Line of best fit for a set f data points.

10.4.4 Determination of the Regression Line Equation

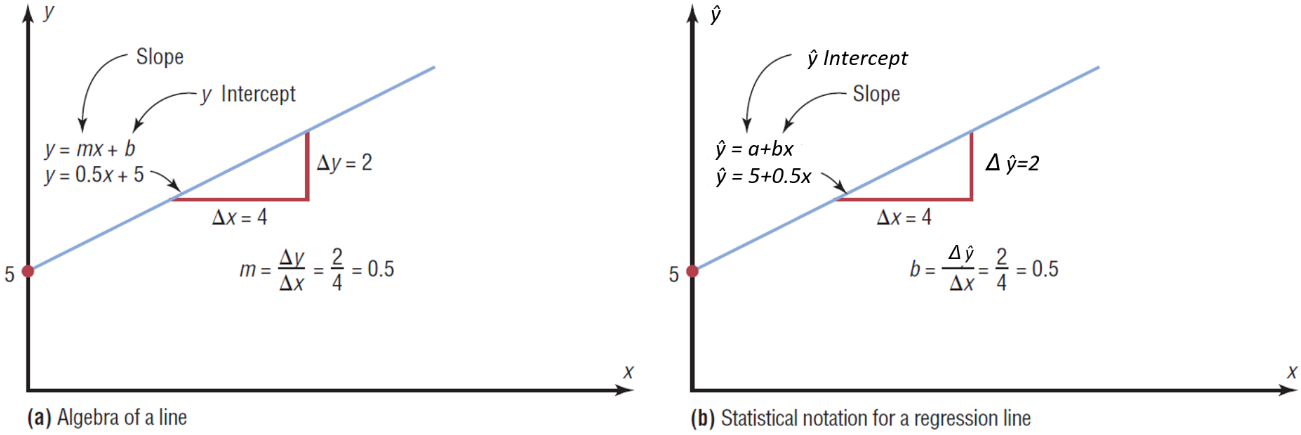
- In algebra, the equation of a line is usually given as:

  $y = mx + b$ **Eq. 10.4-1**

  where m is the ***slope of the line*** and b is the y ***intercept***.
- In statistics, the equation of the regression line is written as:

  $\hat{y} = a + bx$ **Eq. 10.4-2**

  where a is the $\hat{y}$ intercept and b is the slope of the line. See ***Figure 10.4-3***.


y
Slope
y Intercept
y = mx + b
y = 0.5x + 5
Δy = 2
Δx = 4
$m = \frac{\Delta y}{\Delta x} = \frac{2}{4} = 0.5$
5
x
ŷ
ŷ Intercept
Slope
ŷ = a+bx
ŷ = 5+0.5x
Δ ŷ=2
Δx = 4
$b = \frac{\Delta \hat{y}}{\Delta x} = \frac{2}{4} = 0.5$
5
x

(a) Algebra of a line

(b) Statistical notation for a regression line

Figure 10.4-3: A line as represented in algebra and in statistics.

- There are several methods for finding the equation of the regression line. Two formulas are given here.
- The ***mathematical development of these formulas is beyond the scope of this course***.

  $$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$ **Eq. 10.4-3**

  $$b = \frac{n(\Sigma \mathrm{xy}) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$ **Eq. 10.4-4**

  where a is the $\hat{y}$ intercept and b is the slope of the line.

10.4.5 EXAMPLES

Example 10.4-1

Find the equation of the regression line for the data of ***Example 10.3-2***, data for car rental companies. The data have been represented in below for convenience.

Company	Cars (in ten thousand), x	Revenue (in billions), y
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Solution

Parameters for ***Eq. 10.4-3*** and ***Eq. 10.4-4*** have been determined with referring to table below:

Company	Cars (in ten thousands), x	Revenue (in billions), y	x^2	xy
A	63.0	7.0	3969.00	441.0
B	29.0	3.9	841.00	113.1
C	20.8	2.1	432.64	43.7
D	19.1	2.8	364.81	53.5
E	13.4	1.4	179.56	18.8
F	8.5	1.5	72.25	12.8
Summation	**153.8**	**18.7**	**5859.3**	**682.8**

Therefore, a and b would be:

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{18.7 \times 5859.3 - 153.8 \times 682.8}{6 \times 5859.3 - 153.8^2} = 0.396$$

$$b = \frac{n(\Sigma \mathrm{xy}) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.8) - (153.8)(18.7)}{6(5859.3) - (153.8)^2} = 0.106$$

Hence, the equation of the regression line of ***Eq. 10.4-2*** would be:

$\hat{y} = a + bx = 0.396 + 0.106x$

Graphically on the scatter diagram, the regression line would be as indicated in ***Figure 10.4-4***.

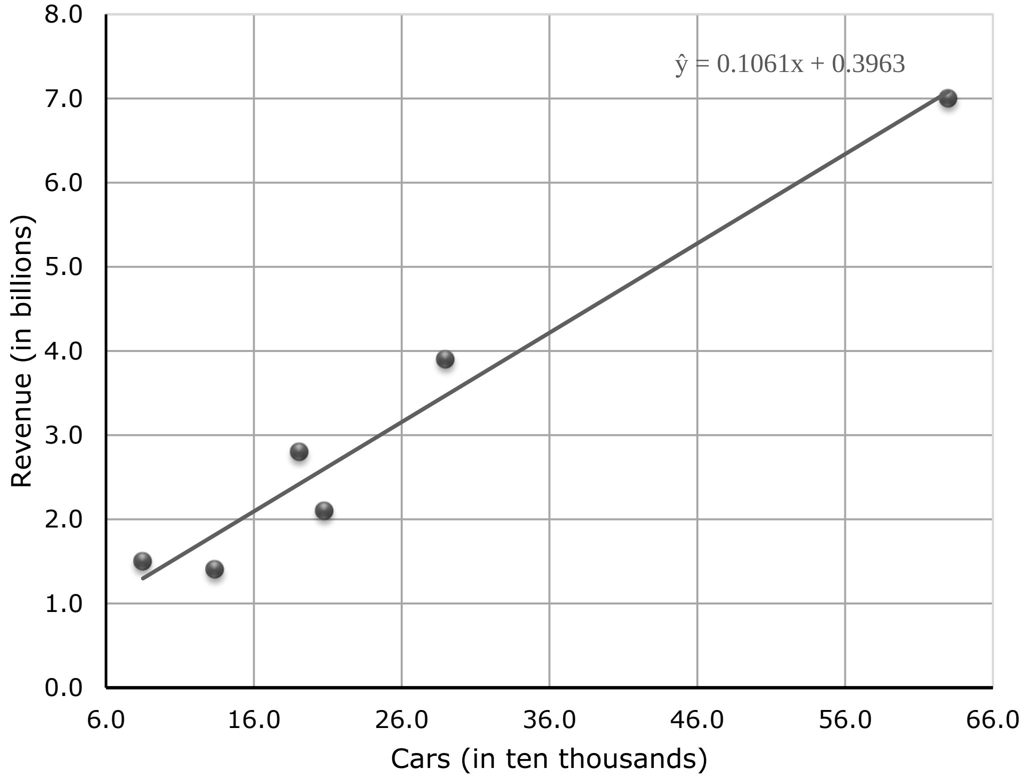


Figure 10.4-4: Scatter plot for Example 10.4-1 with best lint fit.

Example 10.4-2

Data cube, f_{cu} , and cylinder, $f_c{}'$, compressive strength of ***Example 10.3-5***, has been represented in **Table 10.4-1**. Use the data to determine the equation of the regression line. What would be the cylinder strength if the concrete has a cube strength of $28\,MPa$.

Table 10.4-1: Concrete compressive concrete, cubical and corresponding cylindrical specimens, for Example 10.4-2.

Cube compressive strength, f_{cu}, MPa x	Cylinder compressive strength, $f_c{}'$, MPa y
9.00	6.90
15.20	11.70
20.00	15.20
24.80	20.00
27.60	24.10
29.00	26.20
29.60	26.90
35.80	31.70
36.50	34.50
42.10	36.50
44.10	40.70
48.30	44.10
52.40	50.30

Solution

Parameters for ***Eq. 10.4-3*** and ***Eq. 10.4-4*** have been determined with referring to table below.

	Cube, MPa, x	Cylinder, MPa, y	x^2	xy
	9.00	6.90	81.00	62.10
	15.20	11.70	231.04	177.84
	20.00	15.20	400.00	304.00
	24.80	20.00	615.04	496.00
	27.60	24.10	761.76	665.16
	29.00	26.20	841.00	759.80
	29.60	26.90	876.16	796.24
	35.80	31.70	1281.64	1134.86
	36.50	34.50	1332.25	1259.25
	42.10	36.50	1772.41	1536.65
	44.10	40.70	1944.81	1794.87
	48.30	44.10	2332.89	2130.03
	52.40	50.30	2745.76	2635.72
∑	**414.4**	**368.8**	**15215.8**	**13752.5**

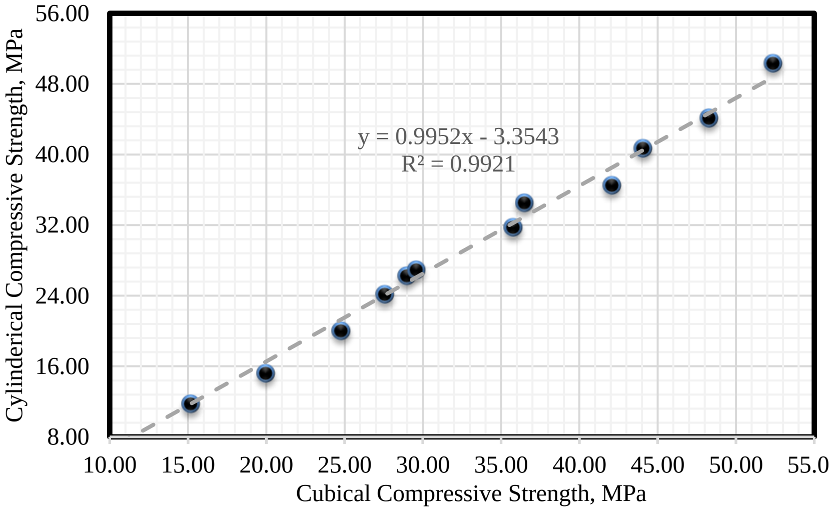
$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{368.8 \times 15215.8 - 414.4 \times 13752.5}{13 \times 15215.8 - 414.4^2} = -3.35$$

$$b = \frac{n(\Sigma \mathrm{xy}) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{13 \times (13752.5) - (414.4)(368.8)}{13 \times (15215.8) - (414.4)^2} = \; 0.995$$

Hence, the equation of the regression line of ***Eq. 10.4-2*** would be:

$$\hat{y} = a + bx = -3.35 + 0.995x$$

Graphically on the scatter diagram, the regression line would be as indicated.



For a cube strength of $f_{cu} = 28\,MPa$, the cylinder strength would be:

$$f_c' = -3.35 + 0.995 \times 28 = 24.5\,MPa$$

Example 10.4-3

Is it meaningful to determine the regression line for the deflection-ultimate load data of ***Example 10.3-6***? From correlation analysis of this example, the data has a correlation coefficient, r, of -0.521.

Solution

As there is no clear relation between deflection at service load and ultimate load, therefore it is meaningless to determine the regression line.

10.4.6 Homework Problems

Home Work 10.4-1

Elongation-tensile force data for the aluminum specimens of ***Home Work 10.3-1*** has been represented as indicated. Determine the equation of regression line.

Tensile Force in kips, X	**Elongation in 10^{-3} of inch, Y**
1	9
2	20
3	28
4	41
5	52
6	63

Ans.

$\Delta = -2.4 + 10.83F$

where

Δ is the elongation in inch, and F is the force in kips.

Home Work 10.4-2

Date for traveling speed versus stopping distance of ***Home Work 10.3-2*** has been represented as indicated. Determine the regression line.

Vehicle No.	Traveling Speed kph	Stopping Distance, m
1	40	15
2	9	2
3	100	40
4	50	15
5	15	4
6	65	25
7	25	5
8	60	25
9	95	30
10	65	24
11	30	8
12	125	45

Ans.

$\Delta = 0.3861V - 2.0108$

where Δ is the stopping distance in m and V is the speed in kph.

10.12 CONTENTS