INTRODUCTION TO STATISTICS AND PROBABILITY

Nafy Aidara
Foreword

The African Virtual University (AVU) is proud to participate in increasing access to education in African countries through the production of quality learning materials. We are also proud to contribute to global knowledge as our Open Educational Resources are mostly accessed from outside the African continent.

This module was developed as part of a diploma and degree program in Applied Computer Science, in collaboration with 18 African partner institutions from 16 countries. A total of 156 modules were developed or translated to ensure availability in English, French and Portuguese. These modules have also been made available as open education resources (OER) on oer.avu.org.

On behalf of the African Virtual University and our patron, our partner institutions, the African Development Bank, I invite you to use this module in your institution, for your own education, to share it as widely as possible and to participate actively in the AVU communities of practice of your interest. We are committed to be on the frontline of developing and sharing Open Educational Resources.

The African Virtual University (AVU) is a Pan African Intergovernmental Organization established by charter with the mandate of significantly increasing access to quality higher education and training through the innovative use of information communication technologies. A Charter, establishing the AVU as an Intergovernmental Organization, has been signed so far by nineteen (19) African Governments - Kenya, Senegal, Mauritania, Mali, Cote d’Ivoire, Tanzania, Mozambique, Democratic Republic of Congo, Benin, Ghana, Republic of Guinea, Burkina Faso, Niger, South Sudan, Sudan, The Gambia, Guinea-Bissau, Ethiopia and Cape Verde.

The following institutions participated in the Applied Computer Science Program: (1) Université d’Abomey Calavi in Benin; (2) Université de Ouagadougou in Burkina Faso; (3) Université Lumière de Bujumbura in Burundi; (4) Université de Douala in Cameroon; (5) Université de Nouakchott in Mauritania; (6) Université Gaston Berger in Senegal; (7) Université des Sciences, des Techniques et Technologies de Bamako in Mali (8) Ghana Institute of Management and Public Administration; (9) Kwame Nkrumah University of Science and Technology in Ghana; (10) Kenyatta University in Kenya; (11) Egerton University in Kenya; (12) Addis Ababa University in Ethiopia (13) University of Rwanda; (14) University of Dar es Salaam in Tanzania; (15) Université Abdou Moumouni de Niamey in Niger; (16) Université Cheikh Anta Diop in Senegal; (17) Universidade Pedagógica in Mozambique; and (18) The University of the Gambia in The Gambia.

Bakary Diallo

The Rector

African Virtual University
Production Credits

Author
Nafy Aidara

Peer Reviewer
Robert Oboko

AVU - Academic Coordination
Dr. Marilena Cabral

Overall Coordinator Applied Computer Science Program
Prof Tim Mwololo Waema

Module Coordinator
Florence Tushabe

Instructional Designers
Elizabeth Mbasu
Benta Ochola
Diana Tuel

Media Team
Sidney McGregor    Michal Abigael Koyier
Barry Savala       Mercy Tabi Ojwang
Edwin Kiprono      Josiah Mutsogu
Kelvin Muriithi    Kefa Murimi
Victor Oluoch Otieno Gerisson Mulongo
Introduction to Statistics and Probability

Copyright Notice

This document is published under the conditions of the Creative Commons Attribution http://creativecommons.org/licenses/by/2.5/

Module Template is copyright African Virtual University licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. CC-BY, SA

Supported By

AVU Multinational Project II funded by the African Development Bank.
# Table of Contents

**Foreword** 2

**Production Credits** 3

**Copyright Notice** 4

**Supported By** 4

**Course Overview** 8

- **Prerequisites** ............................................... 8
- **Materials** ..................................................... 8
- **Course Goals** .................................................. 8
- **Units** .......................................................... 9
- **Assessment** .................................................... 9
- **Schedule** ....................................................... 10
- **Readings and Other Resources** ............................. 10

**Unit 0: Pre-Assessment** 12

- **Unit Introduction** ............................................ 12
- **Unit Objectives** .............................................. 12
- **Key Terms** .................................................. 13
- **Unit Assessment** ............................................ 15
- **Grading Scheme** ............................................. 22
- **Unit Readings and Other Resources** ..................... 22

**Unit 1: Basic Statistics and its Application in ACS** 23

- **Unit Introduction** ............................................ 23
- **Unit Objectives** .............................................. 23
- **Key Terms** .................................................. 23
- **Learning Activities** ....................................... 25
  - Activity 1 statistics ........................................ 25
  - Activity Details ............................................. 25
  - Activity 2 Types of statistical data ...................... 31
  - Activity Details ............................................. 31
Course Overview

Welcome to Introduction to Probability and Statistics

The aim of the course is to equip students with basic knowledge in probability and statistics needed for their studies in ACS. In modern computer science, software engineering, and other fields, the need arises to make decisions under uncertainty. Probability and Statistics helps computer science students solve problems and make decisions in uncertain conditions, compute probabilities and forecasts, and evaluate performance of computer systems and networks. At the end of the course, students should be able to apply Probability & Statistics in the context of ACS. Indeed, they will be able to use statistical concepts, probabilistic calculations, methods of observations, sampling techniques, analysis and classification of variables in interpreting data, and inferring design variables results.

Prerequisites

Basic Mathematics and Calculus, Basic IT skills (spreadsheet)

Materials

The materials required to complete this course are:

- Lecture notes, calculators, computers and Internet connectivity

Course Goals

Upon completion of this course the learner should be able to:

- Analyze counting problems in computer science
- Apply basic concepts of probability;
- Apply random situations involving the concept of chance;
- Design studies using descriptive statistical methods to interpret and analyze the results from these formulating conjectures;
- Select appropriate statistical data analysis methods;
- Interpret information of a statistical nature;
Units

Unit 0: Pre-Assessment
This unit will help students assess their level of knowledge in probability and statistics and also evaluate their competencies in basic mathematics. It is not compulsory but could serve as a guide for both the teachers and the students to identify the knowledge gaps.

Unit 1: Basic Statistics and introduction to SPSS
This unit will introduce the student to the basic concepts of statistics and the use of statistical software in computer science problems solving.

Unit 2: Basic Probability and its applications in ACS
This unit will introduce students to the basic concepts in probability.

Unit 3: Linear Regression and Its Applications in ACS

Unit 4: Practical Applications of probability and statistics in ACS
Specific examples in applied computer science will be provided in this unit.

Assessment
Formative assessments, used to check learner progress, are included in each unit.
Summative assessments, such as final tests and assignments, are provided at the end of each module and cover knowledge and skills from the entire module.
Summative assessments are administered at the discretion of the institution offering the course.
The suggested assessment plan is as follows:

<table>
<thead>
<tr>
<th></th>
<th>Assignments</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Quizzes</td>
<td>10%</td>
</tr>
<tr>
<td>3</td>
<td>Tests</td>
<td>30%</td>
</tr>
<tr>
<td>4</td>
<td>Final Examination</td>
<td>50%</td>
</tr>
</tbody>
</table>
# Schedule

<table>
<thead>
<tr>
<th>Unit</th>
<th>Activities</th>
<th>Estimated time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1</td>
<td>Learning Activity 1</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 2</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 3</td>
<td>10 Hours</td>
</tr>
<tr>
<td>Unit 2</td>
<td>Learning Activity 1</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 2</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 3</td>
<td>10 Hours</td>
</tr>
<tr>
<td>Unit 3</td>
<td>Learning Activity 1</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 2</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 3</td>
<td>10 Hours</td>
</tr>
<tr>
<td>Unit 4</td>
<td>Learning Activity 1</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 2</td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>Learning Activity 3</td>
<td>10 Hours</td>
</tr>
</tbody>
</table>

# Readings and Other Resources

The readings and other resources in this course are:

**Unit 0**

Required readings and other resources:


Optional readings and other resources:

- [http://mathworld.wolfram.com/Probability](http://mathworld.wolfram.com/Probability): Wolfram is a useful site that provides insights in number theory while providing new challenges and methodology in number theory.

- [http://en.wikipedia.org/wiki/Probability](http://en.wikipedia.org/wiki/Probability): Mathsguru is a website that helps learners to understand various branches of number theory module. It is easy to access through Google search and provides very detailed information on various probability questions.

Unit 1

Required readings and other resources:


Optional readings and other resources:

- [http://mathworld.wolfram.com/Probability](http://mathworld.wolfram.com/Probability): Wolfram is a useful site that provides insights in number theory while providing new challenges and methodology in number theory.

- [http://en.wikipedia.org/wiki/Probability](http://en.wikipedia.org/wiki/Probability): Mathsguru is a website that helps learners to understand various branches of number theory module. It is easy to access through Google search and provides very detailed information on various probability questions.

Unit 2

Optional readings and other resources:

- [http://mathworld.wolfram.com/Probability](http://mathworld.wolfram.com/Probability): Wolfram is a useful site that provides insights in number theory while providing new challenges and methodology in number theory.

- [http://en.wikipedia.org/wiki/Probability](http://en.wikipedia.org/wiki/Probability): Mathsguru is a website that helps learners to understand various branches of number theory module. It is easy to access through Google search and provides very detailed information on various probability questions.
Unit 0: Pre-Assessment

Unit Introduction

The purpose of this unit is to determine your grasp of knowledge related to this course. Students will be able to assess their understanding in basic algebra and arithmetic prior to the beginning of the course. Students evaluate their basic skills in calculus and IT.

Unit Objectives

Upon completion of this unit you should be able to:

- Define the basic concepts in arithmetic
- Draw graph from a specific data set
- Analyse and Interpret data.
- Draw conclusions from the interpretations.
## Key Terms

**Probability**: provides mathematical models for random phenomena and experiments, such as: gambling, stock market, packet transmission in networks, electron emission, noise, statistical mechanics, etc.

**Statistics**: Statistics is a very broad subject, with applications in a vast number of different fields. In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information.

**Population**: A (statistical) population is the set of measurements (or record of some qualitative trait) corresponding to the entire collection of units for which inferences are to be made.

**Sample**: A sample from statistical population is the set of measurements that are actually collected in the course of an investigation.

**Parameter**: A parameter is an unknown numerical summary of the population. A statistic is a known numerical summary of the sample which can be used to make inference about parameters.

**Probability**: provides mathematical models for random phenomena and experiments, such as: gambling, stock market, packet transmission in networks, electron emission, noise, statistical mechanics, etc.

Descriptive Statistics: deals with procedures used to summarize the information contained in a set of measurements.

---

Measures of central tendency: There are different ways to measure the central tendency of a data set: Mean, Mode and Median.

**Mean**: It represents the average measure of the data set.

**Median**: It represents the middle value of the data set

**Mode**: This is the value that occurs the most frequently.

**Variance**: It is the nonnegative integer that measure the amount of variation within the data set.
**Standard Deviation**: The square root of the variance.

**Percentiles**: dividing the ordered data set by the use of ordinates so that the amount of data points less than the ordinate is some percentage of the total amount of observations. The values corresponding to such areas are called the percentiles.

**Interquartile ranges**: This is the difference between the first quartile and the third quartile.

**Skewness**: If the distribution of the data set ical about any value, it is said to be skewed. If the data set has a few more lower values, it is said to be skewed to the left. If it ha more higher values, it is said to be skewed to the right.

**Random variable**: Suppose that to each sample point we assign a number. We then have a function defined on the sample space. This function is called a random variable. It is usually denoted by X or Y.

**Discrete Random Variable**: A random variable that takes on a finite or countably infinite number of values is called a discrete random variable.

**Continuous Random Variable**: A random variable that takes on an uncountably infinite number of values is called a non discrete/continuous random variable.
Unit Assessment

Check your understanding!

Diagnostic Test

Instructions

The following tests are intended to diagnose weaknesses that you might have in the areas under the four units. This is a self-administered test. As you begin your studies on statistics and probability, take this course pre-assessment to help you understand better what you may already know about the subject and how best to study the course material. If you find you do well on parts of this self-test, you can move more quickly over that subject in the study guide and leave yourself more time for other sections. This approach will help you determine which areas to review in order to be successful in this course. If you don’t do well at first, do not worry. After taking the test you can check your answers against the given answers and, if necessary, refresh your skills. You are not required to hand in your work.

For multiple choice and true/false questions, circle the best answer.

1. _________ is the science that involves collection, analysis, and interpretation of data in order to make inferences about populations.

   A. Probability
   B. Statistics
   C. Management
   D. None of the above

2. Out of 100 numbers, 20 were 4s, 40 were 5s, 30 were 6s and the remainder were 7s. Find the arithmetic mean of the numbers.

   A. 0.22
   B. 0.53
   C. 2.20
   D. 5.30

3. The following set of raw data shows the length, in millimeters, measured to the nearest mm, of each of 40 leaves taken from plants of a certain species. This is the table of frequency distribution.
<table>
<thead>
<tr>
<th>Length (mm)</th>
<th>Frequency (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 29</td>
<td>2</td>
</tr>
<tr>
<td>30 – 34</td>
<td>4</td>
</tr>
<tr>
<td>35 – 39</td>
<td>7</td>
</tr>
<tr>
<td>40 – 44</td>
<td>10</td>
</tr>
<tr>
<td>45 – 49</td>
<td>8</td>
</tr>
<tr>
<td>50 – 54</td>
<td>6</td>
</tr>
<tr>
<td>55 – 59</td>
<td>3</td>
</tr>
</tbody>
</table>

4. Find the mean of the distribution.
   A. 2
   B. 3
   C. 4
   D. 5

5. Find the mode of the following data: 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, and 4.
   A. 4
   B. 5
   C. 6
   D. 8

6. The range of the values a probability can assume is
   A. From 0 to 1
   B. From -1 to +1
   C. From 1 to 100
   D. From 0 to 5
7. The grades (on a scale of 100) of 69 students are depicted on a stem and leaf diagram. Determine the median.

5 | 778
6 | 1122334
6 | 55566677799999
7 | 00112333344
7 | 55566677779999
8 | 00222244
8 | 5567788999
9 | 1
A. 5
B. 7
C. 9
D. 2

8. If two children are picked from a group of ten. Determine the probability of picking two children
A. 0.2
B. 0.6
C. 0.4
D. 0.3

9. A box contains three coins with a head on both sides, four coins with a tail on both sides, and two fair coins. If one of these nine coins is selected at random and tossed once, what is the probability that a head will be obtained?
A. 3/9
B. 4/9
C. ½
D. 2/9
10. A nation-wide professional qualifying exam has a mean score of 600 with a standard deviation of 50. A random sample of 100 examinees was selected. The sample mean was 630. Determine the standardized value or z-score of the sample mean in this instance.

A. 0.6  
B. 0.2  
C. 0.45  
D. 0.5

11. What type of graphical representation would be best to display the following data? Leslie sold 47 hot dogs. Kelly sold 32 hot dogs. Jessie sold 30 hot dogs. Carlos sold 4 hot dogs. You want to show a comparison among these people to determine who sold the most hot dogs and who sold the least hot dogs.

A. Bar graph  
B. Line graph  
C. Stem and leaf plot  
D. Histogram

12. Using the data below, approximately what percentage of students buys either hot dogs or hamburgers from the cafeteria at lunch?

<table>
<thead>
<tr>
<th>Food Bought</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamburgers</td>
<td>241</td>
</tr>
<tr>
<td>Hotdogs</td>
<td>361</td>
</tr>
<tr>
<td>Pizzas</td>
<td>129</td>
</tr>
<tr>
<td>Salad</td>
<td>45</td>
</tr>
<tr>
<td>Sandwich</td>
<td>63</td>
</tr>
<tr>
<td>Nothing</td>
<td>84</td>
</tr>
<tr>
<td>Total</td>
<td>923</td>
</tr>
</tbody>
</table>
13. What type of graphical representation would be best to display the following data?
{3,5,8,5,2,34,8,9,16,21} You want to show which numbers in the following set are outliers and you want to show the mean of the following set.
A. Line graph
B. Bar Graph
C. Stem and leaf plot
D. Line plot

14. If \( P(A) = \frac{1}{2} \) and \( P(B) = \frac{1}{3} \). Find \( P(A \cup B) \).
A. \( \frac{1}{6} \)
B. \( \frac{2}{3} \)
C. \( \frac{5}{6} \)
D. \( \frac{1}{3} \)

15. The mean age of 5 persons in a room is 30 years. A 36-year-old person walks in. What is the mean age of the persons in the room now?
A. 35
B. 34
C. 31
D. 30
16. 99.7% of the data falls within __________ standard deviations of the mean on the normal distribution.
A. 1  
B. 2  
C. 3  
D. 4

17. A z-value of __________ is used for a 90% confidence interval.
A. 1  
B. 2  
C. 1.5  
D. 3

18. The normal distribution is a __________-shaped distribution.
A. Oval  
B. Bell  
C. Circ  
D. None of the above

19. The table below gives the probability density function of a random variable X. Find the expected value of X.

<table>
<thead>
<tr>
<th>X</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>0.1</td>
</tr>
<tr>
<td>-2</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Unit 0: Pre-Assessment

A. 0.3
B. 1.2
C. 2.5
D. None of the above

20. The loss due to a fire in a commercial building is modeled by a random variable X with density function \( f(x) \) if and 0 elsewhere. Given that a fire loss exceeds 8, what is the probability that it exceeds 16?

A. 1/25
B. \( \frac{1}{8} \)
C. 1/9
D. \( \frac{1}{3} \)

Answers:
1. B
2. C
3. A
4. B
5. D
6. A
7. B
8. B
9. C
10. B
11. A
12. D
13. D
14. C
15. A
Grading Scheme

Each correct answer will carry one (1) mark.

Unit Readings and Other Resources

The readings in this unit are to be found at the course-level section “Readings and Other Resources”.
Unit Introduction

This unit provides a set of concepts and methods that are designed to enable students to be able, through correct application, to interpret and analyze sample data, apply the appropriate techniques for data analysis and critically interpret the results.

Unit Objectives

Upon completion of this unit you should be able to:

1. The concept of Statistics and its divisions;
2. Identify the concept of population, sample and random experiment;
3. Describe the organization and type of statistical data
4. Identify the different types of sampling
5. Compute the mean, mode and the measures of variation

Key Terms

Statistics: The branch of mathematics that investigates the processes of obtaining, organizing and analyzing data on a population or on a collection of all beings, and the methods to draw conclusions and make inferences or predictions based on these data.

Descriptive Statistics: Responsible for the organization and description of the information

Inductive and Inferential Statistics: Understands the generalization process, from the analysis and interpretation of sample data.

Population or Universe: fundamental set of all elements with at least one common feature.

Sample or Event: Subset of a universe.
**Experiment Random or not deterministic:** Statistical Observation of any phenomenon such as the toss of a coin (not addicted) to observe the end result, heads or tails.

**Variables:** It is a symbol representing certain characteristic of a population or sample. In other words, a variable is a characteristic of the population that can be measured according to some scale.

**Quantitative Variables:** are the characteristics that can be measured in a quantitative scale that is, having numerical values that make sense. It can be continuous or discrete.

**Discrete Variables:** measurable characteristics that can assume only a finite or infinite number of values countable and thus make sense only integer values. They are usually the result from counting. Examples: number of children, number of bacteria per liter of milk, number of cigarettes smoked per day.

**Continuous variables:** measurable characteristics that take values on a continuous scale (on the real line), for which fractional values make sense. Usually should be measured using an instrument. Examples: weight (balance), height (ruler), time (clock), blood pressure, age.

**Qualitative Variables (or categorical):** features do not have quantitative values, but rather, are defined by several categories, i.e., represent a classification of individuals. It can be nominal or ordinal.

**Nominal variables:** there is no ordering among the categories. Examples: sex, eye color, smoker / non-smoker, patient / healthy.

**Ordinal variables:** there is an ordering between the categories. Examples: education (1st, 2nd, 3rd degree), disease stage (initial, intermediate, terminal), month of observation (January, February, ..., December).

**Sampling:** set of procedures by which you select a sample of a population.
**Sampling probabilistic** - Procedure in which all elements of the population have a known and a probability of zero to integrate sample.

Not probabilistic sampling **Intentional Sampling** - Sampling Probabilistic not subject to specific objectives of the investigator.

**Sampling unintentional** - Sampling Probabilistic not governed by criteria of convenience and / or availability of respondents.

**Raw data**: set of data that have not been organized numerically obtained after the critical values.

**Role**: It is an arrangement of raw data in ascending order.

**Total Width (AT)**: the difference between the highest and the lowest value observed.

---

**Learning Activities**

**Activity 1 statistics**

**Introduction**

Every day we are exposed to a large amount of numerical information. Depending on the situation, why are consumers of numerical information, now we need to produce them. Thus, we need knowledge and training to understand the construct them. Procedures, techniques and statistical methods are fundamental to aid the implementation of these tasks. In summary, the statistic is SCIENCE The data, a science for the producer and the consumer of numerical information. It involves collection, classification, summarization, organization, analysis and interpretation of data.

**Activity Details**

Statistics is a part of applied mathematics that provides methods for collection, organization, presentation, analysis and interpretation of data.

Statistics is divided into two areas:

- Descriptive Statistics is part of the statistic that takes care of the collection, organization and description of the observed data;
- Inferential Statistics (Inductive Statistics) is the part of the statistic that tries to generalize findings to a population from the analysis and interpretation of data from a sample.
The statistic is present in many activities that directly affect our lives, for example:

- The analysis of traffic problems;
- The study of the effects of various drugs;
- Quality control of products;
- The evaluation of teaching techniques;
- The behavior;
- The study

The statistical scope has expanded considerably. One reason is the increasing use of statistics in various sectors such as agriculture, education, politics, ENGINEERING, psychology, economics and administration. Another reason why the development of statistics in recent years is the technological developments that facilitated our ability to handle information.

The analysis of a statistical problem is done over several phases:
i. Problem Definition: Know exactly what you want to search; the purpose of establishing the analysis and definition of the population.

ii. Sampling and Data Collection: Operational phase. The process of selection and systematic recording of data, with a particular purpose. Data can be primary (published by the person or organization) or secondary (when they are published by another organization).

iii. Treatment and Presentation of Data: Summary of data through its counting and grouping. It is the data classification, using tables or graphics.

iv. Analysis and Data Interpretation: The last phase of statistical work is the most important and delicate. It is essentially linked to the calculation of measures and coefficients, whose main purpose is to describe the phenomenon of behavior under study (descriptive statistics). In inductive statistical data interpretation are based on probability theory.

**Conclusion**

This is an introductory activity based on the basic concepts of statistics, for the same TRANSMISSION trainees must carry out various tasks such as:

1st Activity: Reading the references mandatory;

2nd Activity: Investigation of different concepts on the Internet;

3rd Activity: Group work - Prepare a summary of 50 words, explaining the circumstances in which sampling is preferable to a census (all members of the public)? Exemplifies.

Activity 4: application exercises Resolution.

**Practical Exercise**

1. Choose the correct alternative:

   a. Population or universe is:
      
      i. Set people.
      
      ii. Individuals presenting a special feature.
      
      iii. Set all individuals with a common characteristic object of study.
      
   b. The variable is discrete when:
      
      i. Given two real values, we can find at least one value among them.
      
      ii. Given two real values, we cannot find values between them.
      
      iii. Given two real values, the difference between them is zero.
c. The main stages of the statistical method are:
   i. Data collection, sampling, tabular presentation and layout and definition of problems.
   ii. Sampling, tabular presentation, verification of data, interpretation of data and planning.
   iii. Problem definition, planning, data collection, calculation, data presentation, analysis and interpretation of data.

d. Part of the population removed to analyze it is called:
   i. Universe;
   ii. Party;
   iii. Piece;
   iv. Raw Data;
   v. Sample.

2. The intention was to make a study of the number of siblings of students in the 10th grade of a secondary school.

For this, a survey was carried out to which 60 students answered. Indicate:
   a) The study population
   b) The chosen sample;
   c) The study variable and rate it

3. In a survey about the time (in hours) that Guineans are connected to the internet, were interviewed 2500 people.

Assuming, in Bissau city there are about 400 000 people, identifying the population and the sample in this situation.

4. The director of a college, in which they are enrolled 280 boys and 320 girls, wanting to know the conditions of extra-school life of their students and not having time to interview all families, decided to do a survey by sampling 10% of these clients. Get, this director, the elements of the sample.

5. A city X has the following table of his faculties:
<table>
<thead>
<tr>
<th>SCHOOL</th>
<th>NUMBER OF STUDENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MALE</td>
</tr>
<tr>
<td>A</td>
<td>80</td>
</tr>
<tr>
<td>B</td>
<td>102</td>
</tr>
<tr>
<td>C</td>
<td>110</td>
</tr>
<tr>
<td>D</td>
<td>134</td>
</tr>
<tr>
<td>E</td>
<td>150</td>
</tr>
<tr>
<td>F</td>
<td>300</td>
</tr>
<tr>
<td>Total</td>
<td>876</td>
</tr>
</tbody>
</table>

Get a stratified proportional sample of 120 students.

6. A population is divided into three levels, with sizes respectively, \( n_1 = 40 \), \( n_2 = n_3 = 100 \) and 60. Knowing that, when held a proportionate stratified sampling, nine elements of the sample were taken from the 3rd layer, determine the total number of sample elements.

7. Show how could a sample of 32 elements of a population consisting of 2,432 ordered elements.

In the general ranking, which of the following elements would be chosen to belong to the sample, given that the element of order 1420 belongs to it?

1.648°, 290°, 725°, 2.025°, 1.120°.

8. Identify which of these sample types are used: random, systematic, convenience, stratified or cluster

a. News on TV - A news reporter of the Globe network analyzes the reaction to an impressive history interviewing people passing in front of his studio.

b. Telephone surveys - In a survey on the operation of the 1059 MTN people, the subjects of the interview were selected using a computer to randomly generate telephone numbers, which were then dialed.

c. Car ownership - A researcher at General Motors divided all cars registered in categories of subcompact, compact, medium, intermediate and large. He’s searching 200 car owners in each category.

d. Drink between Students - Motivated by the fact that a student has died from excessive drinking, college did a study of student drinking habit, randomly selecting 10 different classes and interviewing all the students in each class.
e. Of Sobriety Checkpoint - The author was an observer from one point police sobriety checks, in which every fifth driver was stopped and interviewed. (He witnessed the arrest of a former student).

f. Urn of Boca Research - A news network is planning a survey in which 100 polling stations will be selected at random and all voters will be interviewed on leaving the site.

Anthropometry

g. A student obtains statistical data on height / weight interviewing family members.

h. Medical Research - A researcher at ENA examines all heart patients from each of the 30 hospitals randomly selected.

9. In a binder with 500 numbered chips and ordered from 1 to 500, select 10 records for a search.

Answers

1a) iii  b) ii  c) iii  d) v

<table>
<thead>
<tr>
<th>Sex</th>
<th>Number of students</th>
<th>Rate of sampling fraction</th>
<th>Number of selected students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>876</td>
<td>0.07</td>
<td>57.41</td>
</tr>
<tr>
<td>Female</td>
<td>955</td>
<td>0.07</td>
<td>62.59</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>120</td>
</tr>
</tbody>
</table>

6 – 30

7 - 1.648º

8. Note that $k = \frac{500}{10} = 50$. A record draws between 1 and 50, for example, the plug 17 will be selected the next number $17 + 50 = 67$, and so forth, and therefore, the sample is composed of records:

17, 67, 117, 167, 217, 267, 317, 367, 417, 467
Activity 2 Types of statistical data

Introduction

Statistical data are the submission of observations of a particular variable, either qualitative or quantitative nature, in order to describe the entire set of units observed in summary or summary form. Statistical data form the basis of all study and statistical analysis of the discussed context, give that the statistical data are considered as the primary ingredient to any investigation. Statistical data can be collected from existing sources or collected through survey and experimental studies; they can be of different types and therefore need to be treated with different statistical methods.

Activity Details

The description and interpretation of data is an essential part of statistics. The quality of the solution of the statistical problem is directly related to the quality of the data obtained. Therefore, in terms of utilization, appropriate methods for data collection depend to the problem to be studied.

Figure representing the types of statistical data.

Conclusion

Exercises: Application and concentration

1. In a study in a school, data was collected for the following variables:

   (a) age (E) time spent daily in study
   (b) grade (F) distance from home to school
   (c) sex (G) study site
(d) note in the discipline of Mathematics (H) number of siblings

a) The indicated variables, which are quantitative and which are qualitative?

b) Of quantitative variables, which are continuous say.

2. Rate the variable as qualitative (nominal or ordinal) or quantitative (discrete or continuous):

   d. Variable: number of ports.
   e. Population: animals in a zoo.
   f. Variable: predominant color.
   g. Population: candidates for a job opening.
   h. Variable: Education.
   i. Population: soccer players in a club.
   j. Variable: position in which they operate.
   l. Variable: preparation time.

3. A network of shops are doing a survey on customer satisfaction. One of the questions that the client must answer is: “Are you satisfied with our service.” The categories of responses were.

<table>
<thead>
<tr>
<th>Satisfied</th>
<th>Dissatisfied</th>
<th>undecided</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 1 1</td>
<td>1 0 2 2 2 2 0</td>
<td></td>
</tr>
<tr>
<td>0 1 0</td>
<td>1 2 0 2 2 2 2</td>
<td></td>
</tr>
<tr>
<td>1 2 2</td>
<td>1 2 2 0 0 2 2</td>
<td></td>
</tr>
<tr>
<td>2 1 1</td>
<td>1 1 2 0 2 1 1</td>
<td></td>
</tr>
<tr>
<td>1 1 1</td>
<td>0 0 2 1 2 0 1</td>
<td></td>
</tr>
</tbody>
</table>
A sample of 50 people who answered the questionnaire, provided the following answers to specific question (to help the procedures of the results via computer a numerical scale was used where: 2 = satisfied, 1 = dissatisfied, 0 = undecided)

These data are quantitative or qualitative. Review your answer.

4. A survey of a newspaper surveyed 2500 adults, “Are you satisfied with the economic situation in the country today.” The response categories were dissatisfied, satisfied and indecisive.

   a. What is the size of the survey sample?
   b. The data collected were qualitative or quantitative?
   c. Of those who responded, 28% said they were dissatisfied with the economic situation. How many individuals have provided this answer?

5. Classify the following variables as qualitative and quantitative and the latter is discrete or continuous:

   a) No. passengers on the bus from Bafatá line;
   b) Education of a group of people;
   c) The average weight of newborns from maternity;
   d) Altitude above sea level;
   e) A survey conducted with 1,015 people indicates that 40 of them are subscribers to a broadband Internet service;
   f) The electronic indicates that the player radar last snapped ball 82,3mi / h;
   g) The time spent for a person to make a Bissau drive to Gabu is approximately 2: 40h at an average speed of 100km / h.
   h) The students eye color;
   i) Production of cashew nuts in Guinea-Bissau;
   j) Number of defects on TV equipment;
   k) The point obtained in each play of a given
Answers

1.a. Quantitative (A), (D), (E), (F), (H)

b. Qualitative: (B) (C) (G)

c. Are continuous quantitative variables (E), (F) and optionally (A); the variable Age is also continuous, it can take any value in a range, although it is usually treated as discrete)

2. a) continuous Quantitative;

b) Quantitative discreet;

c) Qualitative ordinal;

d) Qualitative

ordinal;

e) nominal Quantitative;

f) Continuous Quantitative

3. Are quantitative were converted into numerical scale according to a number

4.a. 2 500 adults

b. qualitative

c. 700 adults

5. a. Quantitative discrete

b. qualitative

c. quantitative continuous

d. quantitative continuous

e. quantitative discrete

f. qualitative

g. quantitative continuous

h. qualitative

i. quantitative continuous

j. quantitative continuous

k. quantitative discrete
Activity 3 Tabular and graphic representation

Introduction

Descriptive statistics, whose basic objective is to synthesize a series of related values, thereby enabling to have a global view of the variation of these values, organizes and describes the data in three ways: through tables, graphs and descriptive measures.

The table is a representation summarizing a set of observations, while the graphics are presentations of data, whose goal is to produce a faster and more vivid impression of the phenomenon under study.

Activity Details

Table

It is very common these days due to the use of computers, conduct research where data collection results in large collections (quantity) of data for analysis and it becomes almost impossible to understand them, as to (s) particular (s) objective (s) of study if these data are not summarized. In other words, the manner in which data were collected do not allow an easy and quick way, which extract information.

Table of Elements

Every table should be simple, clear, objective and self-explanatory

The provision of a table can be generalized as shown in the Figure below.

Table X Title answering the questions: what, where and when?

<table>
<thead>
<tr>
<th>indicator column</th>
<th>Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online content</td>
<td>Cell</td>
</tr>
</tbody>
</table>

Source: Data source

Note: Insightful information table Body

It is noteworthy that the tables must be numbered in ascending order or in which they appear in the text, as in the case of scientific papers; the top and bottom edges must be closed with horizontal lines while the left and right not, or cannot be closed by vertical lines separating the columns in the body of the table. It is also necessary that the number of decimal places to be standardized.
**Frequency Distribution**

When studying a mass of data is frequent interest summarize the variable information.

Steps for building a frequency distribution:

1. Find the values that can be assumed by the variable;
2. Arrange the values in ascending order, in the left column of your table;
3. Make a number of times each value appears consolidation;
4. Enter the numbers found in step 3 in the column next to “values” in the column named “Frequency”;

Relative frequency or percentage
Cumulative absolute frequency, denoted by $F_{ai}$. These frequencies are obtained by adding the absolute frequency of the value considered, the previous absolute frequencies to the same value.

Cumulative relative frequency

\[
F_{ri} = \frac{F_i}{n}
\]

Distribution in classes or interval

"The distribution of frequencies of classes is suitable for quantitative provide continuous or discrete data with a large number of possible values"

It is necessary to divide the data into intervals or ranges of values that are called classes. A class is a line of frequency distribution. The lowest class is called lower bound ($l_\text{l}$) and the highest value of the class is called upper limit ($l_\text{u}$). The range or class can be represented the following ways:

a. $l_\text{l} | ____ l_\text{u}$, where the lower limit of the class is included in the absolute frequency count but not higher;

b. read $____ | l_\text{u}$, where the upper limit of the class is included in the count but not less;

c. $l_\text{l} | ____ | l_\text{u}$, where both the lower limit and the upper are included in the count;

d. $____ | l_\text{u} l_\text{l}$, where the limits are not part of the count.

Methods for determining the number of classes

**Rule Sturge**

\[
K = 1 + 3.33 \log_{10} n
\]

**Square Rule or square root**

where:

\[
K = \sqrt{N} \quad \text{number of classes}
\]

\[
N = \sqrt{N} \quad \text{total number of observations}
\]

**Example**

Age of students attending the course Statistical Inference, a course in Statistics at the University of Amilcar Cabral Bissau, 03.21.2014.
Range = 37 - 18 = 19 years

<table>
<thead>
<tr>
<th>AGE</th>
<th>Fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
</tr>
</tbody>
</table>

Range = 37 - 18 = 19 years

\[ k = \sqrt{\frac{22}{5}} = 4.69 \approx \text{class } 5 \]

\[ i = \frac{19}{5} = 3.8 \approx 4 \text{ years} \]

<table>
<thead>
<tr>
<th>Age</th>
<th>Xi</th>
<th>fi</th>
<th>fi %</th>
<th>Fan</th>
<th>fi%</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 l---22</td>
<td>20</td>
<td>11</td>
<td>0,5</td>
<td>11</td>
<td>0,5</td>
</tr>
<tr>
<td>22 l---26</td>
<td>24</td>
<td>6</td>
<td>0,2727</td>
<td>17</td>
<td>0,7727</td>
</tr>
<tr>
<td>26 l---30</td>
<td>28</td>
<td>2</td>
<td>0,0909</td>
<td>19</td>
<td>0,8636</td>
</tr>
<tr>
<td>30 l---34</td>
<td>32</td>
<td>1</td>
<td>0,0455</td>
<td>20</td>
<td>0,9091</td>
</tr>
<tr>
<td>34 l---38</td>
<td>36</td>
<td>2</td>
<td>0,0909</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1.4 - Statistical Measures

Another way to summarize the data in a quantitative variable, plus tables and graphs, is present them in the form of numeric values, called descriptive measures. These measures are calculated from population data, are called parameters and calculated from sample data are called estimators or statistics.

The descriptive measures are: measures position (measure of central tendency), measures of dispersion, skewness and kurtosis measures.

1. Arithmetic mean

- Mostly used measure of central tendency;
- Is defined as the sum of the expected values of all observations (observation is an element of a sample) divided by the number of observations;
- The symbol $\mu$ (mi) will be used to denote mean of a population;
- The symbol is used to denote the mean of a sample;

$$
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
$$
Example: A sample of 3 newborns weight: 2.75kg, 3.25kg and 3.80 kg.

Here, $n$, the size of the sample is equal to 3. $x_1$, first observation is 2.75kg; $x_2$, second observation is 3.25kg; $x_3$ is 3.80kg.

$$= (2.75 + 3.25 + 3.80) = 9.80 / 3 = 3.27,$$ that is, average weight is 3.27 kg.

Weighted arithmetic mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

<table>
<thead>
<tr>
<th>Number of children ($x_i$)</th>
<th>Simple Absolute frequency ($f_i$)</th>
<th>$x_i f_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>$\Sigma=78$</td>
</tr>
</tbody>
</table>

$$\bar{X} = \frac{78}{34} = 2.29 \approx 2$$
2. Median Arithmetic. Grouped Data

Calculate the average height of infants according to the table below.

Where:

\( Pm: \) Midpoint of classes

\[ Pm = \frac{L_i + L_s}{2} \]

\( f_i: \) Simple Absolute frequency

Example: Calculate the average height of infants according to the table below.

<table>
<thead>
<tr>
<th>Estaturas (cm)</th>
<th>fi</th>
<th>PM</th>
<th>PM*fi</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>---</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>54</td>
<td>---</td>
<td>58</td>
<td>9</td>
</tr>
<tr>
<td>58</td>
<td>---</td>
<td>62</td>
<td>11</td>
</tr>
<tr>
<td>62</td>
<td>---</td>
<td>66</td>
<td>8</td>
</tr>
<tr>
<td>66</td>
<td>---</td>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>70</td>
<td>---</td>
<td>74</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>( \Sigma = 40 )</td>
<td>( \Sigma = 2440 )</td>
<td></td>
</tr>
</tbody>
</table>

Median (md)

Other measure used to indicate the center of a distribution;

Ordered elements in the sample, the median is the value (or not the sample belongs) which divides in half, i.e. 50% of the sample elements are less than or equal to the median and 50% are greater than or equal to the median

Example: Given the variable \( x = \{1, 3, 0, 2.4\} \), the mean is 2.

To calculate the average median of a data set is due to:
1) order the set; in the above example: \( x = \{0,1,2,3,4\} \)

2) Verify that there is an odd or even number of values in the set; in the above example: 5

Observations - odd

3) If it is odd the median is the value that occupies the central position and if even will be the average of the two central positions.

\[
Md = Li + \left( \frac{N - \text{Fant}}{2} \right) * i
\]

\( Li \) class = lower limit

\( N \) = total number of elements

\( \text{Fant} \) = cumulative frequency

\( f_i \) = frequency of previous class

\( i \) = amplitude of the class

Example:

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>140-144</td>
<td>3</td>
</tr>
<tr>
<td>145-150</td>
<td>5</td>
</tr>
<tr>
<td>151-155</td>
<td>7</td>
</tr>
<tr>
<td>156-160</td>
<td>2</td>
</tr>
<tr>
<td>161-165</td>
<td>1</td>
</tr>
</tbody>
</table>

- Create class
- Frequency Table (accumulated / Relative)
- Media, Mode, median
Frequency table

<table>
<thead>
<tr>
<th>Classes</th>
<th>f</th>
<th>Facum</th>
<th>X</th>
<th>%</th>
<th>% rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>140 - 144</td>
<td>5</td>
<td>5</td>
<td>710</td>
<td>0,10</td>
<td>10 %</td>
</tr>
<tr>
<td>144 - 148</td>
<td>4</td>
<td>9</td>
<td>584</td>
<td>0,08</td>
<td>8 %</td>
</tr>
<tr>
<td>148 - 152</td>
<td>4</td>
<td>13</td>
<td>600</td>
<td>0,08</td>
<td>8 %</td>
</tr>
<tr>
<td>152 - 156</td>
<td>2</td>
<td>15</td>
<td>308</td>
<td>0,04</td>
<td>4 %</td>
</tr>
<tr>
<td>159 - 160</td>
<td>5</td>
<td>20</td>
<td>790</td>
<td>0,10</td>
<td>10 %</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td>2992</td>
<td>0,40</td>
<td>40 %</td>
</tr>
</tbody>
</table>

Median: to get the result the average is necessary to add part-mean (x), and then divided by the sum frequency (f)

\[ M = \frac{3016}{14} = 149.6 \]

Median: to find the result you must use the formula of the precise calculation of the median.

\[
Md = Li + \left( \frac{N - \text{Facum}}{2} \right) \times i
\]

\[ Md = 149 \]

Mode

Mode is the value that appears most frequently in a distribution.

Examples:

Let \( x = \{0 \ 1 \ 0 \ 2 \ 3 \ 4 \ 0 \ 3 \ 2 \ 5 \ 6\} \), Mode is 0.

Let \( x = \{3 \ 1 \ 2 \ 3 \ 3 \ 4 \ 5 \ 1.5 \ 2 \ 1.5 \ 0 \ 4 \ 1.5 \ 1 \ 5 \ 6\} \) mode is 1.5. The series is unimodal.

Let \( X \{2, 3, 4, 4, 5, 6, 7, 7, 8, 9\} \) has two modes: 4 and 7. The series is bimodal.

Mode for grouped data in class

In order to calculate the mode of grouped data, you need to:

- Find the modal class. The modal class is the class interval that has the largest frequency.
- Find the lower class boundary of the modal class (\( = L_b \))
- Find the difference of frequency between the modal class to its upper class (\( = a \)).
• Find the difference of frequency between the modal class to its lower class \((= b)\).

\[
\text{Add the } Lb \text{ to products } \frac{a}{a + b} \text{ by } C \text{, then add it to .}
\]

**Formula Method**

\[
Mo = Lb_{mo} + \frac{a}{a + b} \times C
\]

- \(Lb\) = lower limit of the modal class
- \(b\) = frequency of the modal class - frequency of previous class to the modal class
- \(a\) = frequency of the modal class - frequency of posterior class to the modal class
- \(C\) = amplitude of the modal class

**Separatrices**

The series are equally separated. These measures are - the quartiles, deciles and percentiles.

**Quartiles**

We call quartiles the values of a series that fall into four (4) equal parts.

Three quartiles are therefore identified (Q1, Q2 and Q3) to divide the series into four equal parts.

**Note:** The quartile 2 (Q2) will always be equal to the median of the series.

Example: Calculate the quartiles of the series: \(\{5, 2, 6, 9, 10, 13, 15\}\)

1. The first step to be taken is the sort (ascending or descending) of the values:

   \(\{2, 5, 6, 9, 10, 13, 15\}\)

2. The value that divides the above series into two equal parts is greater than 9, then the Median = Q2 = 9 that will be.
4. The steps for determining the Q1 of a set are as follows:

- Determine on the (adding column fi);
- Calculate the value of \((n / 4)\) (whether \(n\) is even or odd!);
- To build the college column;
- To compare the value of \((n / 4)\) with the college’s values, starting from the first college class (the top!) And asking the question: “This college is greater than or equal to \((n / 4)\) ?” If the answer is NO, the college spent the next class. When the answer is YES, we will stop and try the corresponding class! This will be our Class of First Quartile.
- Finally, we will apply the formula for Q1, extracting the data Q1 of this class, we just found! Again the formula:

\[
Q1 = \text{inf} + \left[ \frac{(n / 4)}{fi} \right] \cdot h
\]

Example: For the set below, determine the value of the third quartile!

<table>
<thead>
<tr>
<th>(X_i)</th>
<th>(f_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 10</td>
<td>2</td>
</tr>
<tr>
<td>20-Oct</td>
<td>5</td>
</tr>
<tr>
<td>20 - 30</td>
<td>8</td>
</tr>
<tr>
<td>30 - 40</td>
<td>6</td>
</tr>
<tr>
<td>40 - 50</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(n = 24)</td>
</tr>
</tbody>
</table>

Step 1) We will find and calculate \(n \cdot (3n / 4)\):

Hence, we find that \(n = 24\) and therefore \((3n / 4) = 18\)

Step 2) builds the college:
Step 3) we compared the college values with the value of $(3n / 4)$, asking the question of practice, adapted to the third quartile:

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$f_i$</th>
<th>$f_{ac}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10 – 20</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>20 – 30</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>30 – 40</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>40 – 50</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>$n = 24$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As the YES answer came in the fourth college class (30! --- 40), we will say that this will be our third class Quartile!

Step 4) we will apply the formula of $Q_3$, using data from the Class of $Q_3$, just identified!

$$Q_3 = l_{inf} + \left[ \left( \frac{3n}{4} \right) - f_{ac,ant} \right] \cdot h$$
Decision

These are values that divide the ordered data set (list) within ten (10) equal parts.

- First Decile (D1) - set value so the data series that 10% of the observations are smaller than him and 90% are greater.
- According Decile (D2) - set value so the data series that 20% of the observations are smaller than him and 80% are greater.
- Ninth Decile (D9) - set value so the data series that 90% of the observations are smaller than him and 10% are greater.

\[ Q3 = 30 + \left[ \frac{18 - 15}{6} \right] \cdot 10 \]

\[ Q3 = 35 \]

Steps taken to calculate the First Decile:

Determined the n (adding column \( f_i \));

Calculate the value of \( \frac{n}{10} \) (whether \( n \) is even or odd!);

Build the college column;

compare the value of \( \frac{n}{10} \) with the college's values, starting from the first college class (the top!) and asking the question: “This college is greater than or equal to \( \frac{n}{10} \)?” If the answer is NO, the college spent the next class. When the answer is YES, we will stop and try the corresponding class! This will be our third class quartile.

Finally, we will apply the formula of Q3, extracting the data Q1 of this class, we just found!

Example: For the set below, determine the value of the first decile!
Step 1) we will find and calculate \( n \) (\( n / 10 \)):

<table>
<thead>
<tr>
<th>( Xi )</th>
<th>( fi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 !--- 10</td>
<td>2</td>
</tr>
<tr>
<td>10 !--- 20</td>
<td>5</td>
</tr>
<tr>
<td>20 !--- 30</td>
<td>8</td>
</tr>
<tr>
<td>30 !--- 40</td>
<td>6</td>
</tr>
<tr>
<td>40 !--- 50</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ n = 24 \]

Hence, we find that \( n = 24 \) and therefore \( (n / 10) = 2.4 \)

Steps 2) build the college:

<table>
<thead>
<tr>
<th>( Xi )</th>
<th>( fi )</th>
<th>( fac )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 !--- 10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10 !--- 20</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>20 !--- 30</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>30 !--- 40</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>40 !--- 50</td>
<td>3</td>
<td>24</td>
</tr>
</tbody>
</table>

\[ n = 24 \]

Step 3) we compared the college values with the value of \( (n / 10) \), asking the question of practice, adapted to the first decile:
We think, therefore, that the corresponding class (10! --- 20) will be our Class of First Decile!

Step 4) we apply the formula of the First Decile:

To the E: D1 = 10.8

**Percentile or centile**

Percentiles call or centiles as the ninety-nine values that separate a series of 100 equal parts. Indicated: P1, P2, ..., P99.

<table>
<thead>
<tr>
<th>Xi</th>
<th>fi</th>
<th>fac</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 !--- 10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10 !--- 20</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>20 !--- 30</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>30 !--- 40</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>40 !--- 50</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>n=24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Finally, we will apply the formula of PX, PX extracting the data from this class, we just found!

Here is the formula:

**MEASURES dispersion or variability**
Measures of dispersion absolute

a) Total range: It is the only measure of dispersion that has on average the reference point.

When the data are not grouped the total amplitude is the difference between the largest and the smallest observed value: \( AT = \text{Maximum } X - \text{X min.} \)

Example: For the values 40, 45, 48, 62 and 70 will be the total amplitude \( AT = 70 - 40 = 30 \)

When data are grouped without class intervals still have \( AT = \text{maximum } X - \text{X min.} \)

Example:
\( AT = 4 - 0 = 4 \)

\[
\begin{array}{|c|c|}
\hline
X_i & f_i \\
\hline
0  & 2  \\
1  & 6  \\
2  & 5  \\
4  & 3  \\
\hline
\end{array}
\]

With class intervals the total amplitude is the difference between the upper limit of the last class and the lower limit of the first class. Then \( \text{Range} = \text{Lmaximum - L min} \)

Example:

\[
\begin{array}{|c|c|}
\hline
\text{Classes} & f_i \\
\hline
4 |---------- 6 & 6 \\
6 |---------- 8 & 2 \\
8 |---------- 10 & 3 \\
\hline
\end{array}
\]

\( \text{Range} = 10 - 4 = 6 \)

The full range is inconvenient and only consider the two extreme values of the series, neglecting the set of intermediate values. Makes use of the full range when you want to determine the amplitude of the temperature in a day, quality control or as a quick calculation measure without much accuracy.

b) Deviation quartile

Also called semi-interquartilica range and is based on quartiles.

Symbol: \( Dq \) and Formula: \( Dq = (Q3 - Q1) / 2 \)
Remarks:

1 - The quartile deviation has the advantage the fact that it is an easy measure to calculate and interpret. Besides, is not affected by extreme, large or small values and is recommended, therefore, when between the data contained extreme values are not considered representative.

2- The quartile deviation should be used preferably when the measure of central tendency is the median.

3- It is a measure insensitive to the distribution of smaller items that Q1, between Q1 and Q3 and higher than Q3.

Example: For the values 40, 45, 48, 62 and 70 the quartile deviation is:

\[ Q_1 = \frac{(45 + 40)}{2} = 42.5 \quad Q_3 = \frac{(70 + 62)}{2} = 66 \quad D_q = \frac{(66 - 42.5)}{2} = 11.75 \]

**Standard Deviation**

It is the measure most commonly used dispersion because it takes into consideration all the values of the variable under study. It is an indicator of variability quite stable. The standard deviation is based on deviations around the mean and its basic formula can be translated as: the square root of the arithmetic mean of the squares of deviations and is represented by \( S \).

<table>
<thead>
<tr>
<th>Classes</th>
<th>( f_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

The above formula is used when dealing with a population of non-grouped data.

Example: Calculate the standard deviation of the population represented by - 4, -3, -2, 3, 5
We know that $n = 5$ and $62.8 / 5 = 12.56$.

The square root of 12.56 is the standard deviation $= 3.54$

FMC: Pearson’s Coefficient of Variation

$$CV = \frac{S}{\bar{X}} \times 100$$

If: $CV < 15\%$ for low dispersion

If $15\% \leq C.V. < 30\%$ for average dispersion

Where: $RC \geq 30\%$ for high dispersion

**Measures asymmetry:**

It is an indicator of the shape of the data distribution

Pearson’s coefficient

$$AS = \frac{3\left(\bar{X} - M_d\right)}{S}$$
AS = 0 → distribution is symmetrical;

![Symmetrical Distribution](image1)

AS > 0 → positive distribution is asymmetric;

![Asymmetric Distribution](image2)

AS < 0 → distribution is asymmetrical negative.

![Asymmetrical Negative Distribution](image3)
Measures kurtosis:

The degree of flattening of the distribution, is an indicator of the shape of this distribution.

Coefficient of kurtosis

- **leptokurtic**: when the distribution has a frequency curve rather closed, with the data strongly concentrated around its center, $C < 0.263$.
- **mesokurtic**: when data is fairly concentrated around its center, $C = 0.263$.
- **platykurtic**: when the distribution has a frequency curve more open with data weakly concentrated around its center, $C > 0.263$.

Exercise Application and concentration

1. Consider a sample comprised of discrete data:

   9, 8, 5, 4, 5, 6, 2, 2, 4, 3, 4, 7, 9, 5, 6, 7, 1, 4, 7, 2, 4, 6, 3, 5, 7, 9, 5, 1, 4, 8, 2, 9

2. Consider a set of values of measured results. It could be, for example, the age of the students in the class U of Statistics discipline.
Age (in months) of the students in the class U - Discipline Statistics

230 234 276 245 345 240 270 310 368 369
334 268 288 336 299 236 239 355 330 247
287 344 300 244 303 248 251 265 246 266
240 320 308 299 312 324 289 320 264 275
252 298 315 255 274 264 263 230 303 281

**Answers**

1.
Conclusion

**Practical Exercise:**

Group work:

Formative Evaluation

1. The data represent 60 family incomes W Subdivision (data at $1,000)

<table>
<thead>
<tr>
<th>Idades</th>
<th>Número de alumnos</th>
<th>Fi</th>
<th>fri</th>
<th>%</th>
<th>Fri</th>
<th>%</th>
<th>i</th>
<th>Iri</th>
<th>Irs</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td>----</td>
<td>250</td>
<td>12</td>
<td>12</td>
<td>0.24</td>
<td>24%</td>
<td>0.24</td>
<td>24%</td>
<td>240</td>
</tr>
<tr>
<td>250</td>
<td>----</td>
<td>270</td>
<td>9</td>
<td>21</td>
<td>0.18</td>
<td>18%</td>
<td>0.42</td>
<td>42%</td>
<td>260</td>
</tr>
<tr>
<td>270</td>
<td>----</td>
<td>290</td>
<td>8</td>
<td>29</td>
<td>0.16</td>
<td>16%</td>
<td>0.58</td>
<td>58%</td>
<td>280</td>
</tr>
<tr>
<td>290</td>
<td>----</td>
<td>310</td>
<td>7</td>
<td>36</td>
<td>0.14</td>
<td>14%</td>
<td>0.72</td>
<td>72%</td>
<td>300</td>
</tr>
<tr>
<td>310</td>
<td>----</td>
<td>330</td>
<td>6</td>
<td>42</td>
<td>0.12</td>
<td>12%</td>
<td>0.84</td>
<td>84%</td>
<td>320</td>
</tr>
<tr>
<td>330</td>
<td>----</td>
<td>350</td>
<td>5</td>
<td>47</td>
<td>0.1</td>
<td>10%</td>
<td>0.94</td>
<td>94%</td>
<td>340</td>
</tr>
<tr>
<td>350</td>
<td>----</td>
<td>370</td>
<td>3</td>
<td>50</td>
<td>0.06</td>
<td>6%</td>
<td>1</td>
<td>100%</td>
<td>360</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>1</td>
<td>100%</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>3</td>
<td>13</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>7</td>
<td>12</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>11</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>16</td>
<td>13</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is asked:

a) Obtaining descriptive statistics for Microsoft Excel use.

b) Interpret the results

Is asked:

a) Obtaining descriptive statistics for Microsoft Excel use.

b) Interpret the results
b) The average income of all the 60 households is $7,677.97

The standard error = devio ratio between the standard and the square root of n, where n = 60, is $590.

50% of households have incomes of less than $8, and the remaining 50%, above this amount (Note that the median is 8,000)

The fashion equal to $3,000, means that the most common income of 60 families group is $3,000

The dispersion around the mean, as measured by deviation, is $4,533.1.

The measurement urtose evaluates the degree of flattening of the distribution, and indicates that the distribution is leptokurtic because the coefficient is negative

It is mildly asymmetric to direct the coefficient is 0.28648

The total amplitude is equal to $15,000 (16000-1000)

The lower income is $1,000, while the highest is $16,000.

The sum of all income reaches 453,000.
Unit Summary

The teacher will indicate a practical group work field

Unit Assessment

1. Establish which of the following data are discrete and which are continuous:
   a) Number of shares sold on the stock exchange
   b) Temperatures recorded every half hour in a weather station
   c) Length of parts produced by certain machine
   d) Diameters 1000 fasteners produced by a factory
   e) Number of people in the carnival of Brazil

2. What is a random trial and in what circumstances should be used?

3. What is probability sampling and when it should be used?

4. A study should be done to determine the annual use in schools. For this, the Ministry of National Education, has a population of 7000 students spread over 4 levels of education: 3000 are of primary education, secondary education 2000, 1500 high school and 500 are higher education. The direction of statistical studies of the ministry estimated that the sample must be at least 700 students to be considered representative.

   a) Determine the rate or sampling fraction;
   b) Determine, using proportional stratified sampling, the number of students to be extracted in each stratum of education.

5. To conduct a study on the time spent, in minutes, for 60 elements of a karting club in a 20-lap circuit, there was the time spent by 16 of these elements. The results were as follows:

   14.1 13.5 15.0 16.2 17.6 18.7 13.1 15.4
   16.6 17.2 14.8 15.9 18.0 16.3 14.9 14.3

   a) State:
   b) the population;
   c) sample.
   d) Indicate the study variety and rate it.
   e) Indicate four values that the statistical variable can take
Answers

1. Discrete, Continuous, Continuous, Continuous, Discrete.

2. Each member of the population possesses a certain probability of being selected. Usually have the same probability. Thus, if N is the population size, the probability of each element is selected is $1 / N$.

3. a) $f = n / N = 700/7000 = 0.10$ or 10%

   b) Primary Education $N1 = 3000$ students

      - A level $N2 = 2000$ students
      - High School $N3 = 1500$ students
      - Higher Education $N4 = 500$ students

4. Calculating the number of elements to be selected in each stratum of education $n_i = f * N$

   - Elementary School $n1 = 0.10 * 3000 = 300$
   - Secondary education $n2 = 0.10 * 2000 = 200$
   - High school $n3 = 0.10 * 1500 = 150$
   - Higher education $n4 = 0.10 * 500 = 50$

   The sum of $300 + 200 + 150 + 50 = 700$ units in the sample

5. a) 60 elements,

   b) 16 elements,

   c) time spent in minutes

   d) Continuous Quantitative

   e) 16.6, 17.2, 14.8 and 15.9

---

**Número e tipo de avaliação**

<table>
<thead>
<tr>
<th>Número e tipo de avaliação</th>
<th>Coeficiente</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Trabalhos em grupo</td>
<td>10%</td>
</tr>
<tr>
<td>2 Trabalhos individuais</td>
<td>15%</td>
</tr>
<tr>
<td>4 Solução de exercícios</td>
<td>20%</td>
</tr>
<tr>
<td>3 Trabalho de pesquisa</td>
<td>30%</td>
</tr>
<tr>
<td>1 Exame final</td>
<td>25%</td>
</tr>
</tbody>
</table>
Readings and other required resources:

Lecture 1


Lecture 2


Readings and other optional features:


Internet Resources

www.youtube.com/watch?v=B4L3G30XB7I. 7min Statistics - Basics - Video Lesson

www.portalaction.com.br/content/estatistica-básica

It is a statistical software developed for students with easy to use, comprehensive and reliable. The system was developed under Action R platform, one of the most widely used statistical systems.

The Action system is a great improvement compared to the statistical software:

2. Allows you to work with Excel in an integrated manner;

3. It is easy to install, creative and covers the main needs of the statistical user;

4. It is becoming more intuitive, easier to use than ever, with a lot of features Action system is an open and democratic system for the use of statistics:

- This program is free software; you can use it under the terms of the GNU General Public License;
- No language barrier - is available in Portuguese and English;
- First statistical system that uses the R and Excel platform in an integrated manner, all to facilitate and expedite their statistical analyzes;
ALEA - Local Action of Applied Statistics - is within the scope of Education, Information Society, the Statistical Information, Training for Citizenship and Literacy Statistics as a contribution to the development and availability of support instruments to the teaching of Statistics for students and teachers of Basic and Secondary Education, the main support a web site.

Improving statistical literacy is thus an important condition for, on the one hand, ensure better provision of a public utility and, on the other hand, foster environments and diverse learning experiences using new information technologies.

http://www.infoescola.com/estatistica/distribuicao-de-frequencias/

**Internet Resources**


It is a democratic feature that is frequently updated, easily accessible through Google, contains book Probability and Statistics, graduates can find on this site many problems of probability and statistics content.


Site consists of various items of Probability and Statistics, is very practical and easy to access. This site specifically investigate the theme “coleita methods of data”

[http://www.alea.pt](http://www.alea.pt)

This site presents different statistical concepts with their examples, exercises and intelligent didactical games.

[http://www.pordata.pt/Portugal](http://www.pordata.pt/Portugal)

[www.youtube.com/watch?v=UzBpykJhpyw](www.youtube.com/watch?v=UzBpykJhpyw)

Lesson on calculating the mean, median and mode for frequency distribution table for grouped data

**Media:** to get the result the average is necessary to add part-mean \(x\), and then divided by the sum frequency \(f\)

\[
M = \frac{3016}{149.6}
\]

Median: to find the result you must use the formula of the precise calculation of the median.
Unit Summary

We discussed basic statistics and introduction to SPSS. We provided the meaning of statistics, types of statistics, measures of central tendencies and presentation of data.

Unit Assessment

Check your understanding!

Unit Readings and Other Resources

The readings in this unit are to be found at course level readings and other resources.
Unit 3: Linear Regression

Total hours: 40 hours

Introduction to Unit

In prior units, description and the statistical inferences were treated only in terms of a variable. Thus, when the sample had a business, we considered a variable at a time, for example, the billing. However when we get a sample of companies, there are several variables that can be observed in each sampled unit: number of employees, wages, etc. In the first case, each observation unit is associated with the measurement of a variable X; then, each unit is associated with the measurements of several variables X, Y, Z, etc.

Unit Goals

Students will be able to:

1. Identify values of a dependent variable (Y) as a function of the independent variable (X).
2. Describe how changes in X can affect Y.
3. Analyze the simple linear regression model.

Key Terms

- **Regression analysis**: Allow describe using a mathematical model, relationship between two variables, starting from n observations of the same.
- **Least squares method**: Method that allows the adjustment of a straight line to the observed data.
- **Dependent variable (RV)**: Measures the phenomenon that is studied and we want to explain. Are those whose effects are expected according to the causes. They are located usually at the end of causal process and are always set in the case or towards Statistics.
- **Independent variable (VI)**: Those candidates' variables to explain the (s) variable (s) dependent (s), the effects of which we want to measure. Here we must be careful because even finding relationship between variables that do not necessarily mean causation.
Learning activities

Activity 1 The simple linear regression model

Total hours: 40 hours

Introduction

Whenever we wish to study particular variable in another function always do a regression analysis.

We can say that the regression analysis aims to describe, through a mathematical model, the relationship between two variables, starting from n observations of the same.

Details of the activity

1st Activity: Reading the obligatory references; in order to explain the linear regression model

2nd Activity: Build the scatter diagram by way of example, with the use of Microsoft Excel and SPSS

3rd Activity: Interpret the scatter plot

The simple linear model regression

In the decision-making process is often necessary to make predictions. At the same time, it is much easier to make decisions on certain variable when it is possible to establish a link between this and another variable whose behavior is known.

In order to make predictions about a variable from another variable there must be between the two a cause and effect relationship, i.e. the variation of a variation can be attributed to another. The first step in the regression study consists precisely in establishing whether the relationship variable is not merely accidental.

Once you have established the existence of a possible causal relationship between the variables, the next step is to study the type of relationship. To do this, should first be making a scatter plot of the observed data.

Chart Wasting

Graph where each point represents one to observed values (Xi, Yi) corresponding respectively to the values of the independent and dependent variables. The scatter diagram has a dual function:

- Helps determine whether there is any relationship between the variables and
- Allows you to identify what is the most appropriate equation to describe this relationship.
The relationships between variables can be of various types: linear, exponential, logarithmic, power, logistics, etc.

Graphic: different relationships between variables

A: Negative linear relationship
B: Positive linear relationship
C: Ratio of Absence
D: Nonlinear Relationship

The simplest relation is linear type, and you can make many of the linear non-linear relationships previously identified. The type of linear relationship between two variables can be described mathematically by the following equation:

\[ Y = a + bX + e \]

Simple linear regression model

Where:

- \( Y \) is the explained or dependent variable;
- \( X \) is the explanatory variable or independent;
- A residual variable type that includes other \( Y \) explanatory factors not included in \( X \) and even measurement errors;
- \( a \) and \( b \) are constants: is the intercept of the straight line with the vertical axis and \( b \) is the slope of the line.
Example:

Suppose, for example, that an economist studies the relationship between unit labor cost and the price index of producer in order to make predictions about the last variable from known values of the former. For such data is available from 1984 to 1990:

Growth of unit labor costs and producer prices

<table>
<thead>
<tr>
<th>year</th>
<th>Growth of unit cost</th>
<th>Price growth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>labor (%)</td>
<td>the producer (%)</td>
</tr>
<tr>
<td>1984</td>
<td>7,8</td>
<td>10,8</td>
</tr>
<tr>
<td>1985</td>
<td>5,7</td>
<td>4,4</td>
</tr>
<tr>
<td>1986</td>
<td>6,1</td>
<td>6,5</td>
</tr>
<tr>
<td>1987</td>
<td>7,7</td>
<td>7,8</td>
</tr>
<tr>
<td>1988</td>
<td>11,2</td>
<td>11,1</td>
</tr>
<tr>
<td>1989</td>
<td>11,2</td>
<td>13,5</td>
</tr>
<tr>
<td>1990</td>
<td>8,3</td>
<td>9,2</td>
</tr>
</tbody>
</table>

Once the economist is interested in predicting changes in the production price, this is defined as the dependent variable and will be called Y. prediction will be based on the independent variable called X and that in this concrete example is the growth of unit labor cost. In SPSS.

1 From the menu bar choose:

Scatter Graphs →....

Select simple scatter
• Select the dependent variable for the Y axis.
• Select the independent variable to the axis X.
To identify the points, use the Data mode ID option in the Chart Editor window.
The relationship between the two variables is linear, positive type, because the higher the growth of the labor cost, the greater the increase in the price of the manufacturer, i.e., where the two variables vary in the same direction.
Conclusion

Linear regression permit find the line that best represents the relationship between two variables.

Example: Simple Linear Regression - Excel software Use

We have a set of 5,000 observations of the variables X and Y has been registered and the equation of the straight line was obtained.

See the column “coefficients” to make sure the values of the parameters a and b of the line.

Formative Evaluation

Preparation of research papers, individual, based on the regression analysis applied to different areas or professions (medicine, agriculture, economics, etc...), so that the students feel encouraged to use the resources available on the Internet. All materials produced in this area will be an integral part of the individual electronic portfolio forming.

Exercise Application and concentration

1. An interviewer’s administrator wants to develop a model to predict the number of interviews in a given day. He believes the interviewer’s experience (measured in weeks worked) is determining the number of interviews. A sample of 10 interviewers gave the following data:

<table>
<thead>
<tr>
<th>Week experience</th>
<th>15</th>
<th>41</th>
<th>58</th>
<th>18</th>
<th>37</th>
<th>52</th>
<th>28</th>
<th>24</th>
<th>45</th>
<th>33</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of interviews</td>
<td>4</td>
<td>9</td>
<td>12</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>
Naming \( Y \) = number of interviews and \( X \) = Weeks experience, we can build the scatter diagram in SPSS.

2. A study was conducted in order to investigate whether there is any relationship between agricultural production and energy consumption and ultimately, it is possible to predict agricultural production from energy consumption.

Considering the data for 9 years.

<table>
<thead>
<tr>
<th>Year</th>
<th>Agricultural production index ((1997 = 100))</th>
<th>Contents of energy consumption ((1997 = 100))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>1999</td>
<td>104</td>
<td>112</td>
</tr>
<tr>
<td>2002</td>
<td>111</td>
<td>121</td>
</tr>
<tr>
<td>2004</td>
<td>127</td>
<td>131</td>
</tr>
<tr>
<td>2006</td>
<td>133</td>
<td>1374</td>
</tr>
<tr>
<td>2008</td>
<td>139</td>
<td>162</td>
</tr>
<tr>
<td>2010</td>
<td>144</td>
<td>185</td>
</tr>
<tr>
<td>2012</td>
<td>144</td>
<td>193</td>
</tr>
<tr>
<td>2013</td>
<td>173</td>
<td>219</td>
</tr>
</tbody>
</table>

a) Which of the variables should be considered explanatory?

b) Draw a scatter plot and describe the type of relationship between the variables

2. The table below shows the frequency of the average pulse in different age periods:
a) Find the linear regression equation.

4. A sample of plants has led to an industry:

<table>
<thead>
<tr>
<th>Total Cost Y</th>
<th>Production X</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>51</td>
<td>6</td>
</tr>
<tr>
<td>70</td>
<td>11</td>
</tr>
<tr>
<td>61</td>
<td>8</td>
</tr>
</tbody>
</table>

a) Find the linear regression equation.

b) What are the economic meanings of “a” and “b”?  

c) Find the coefficient of determination (or explanation). 

d) Test the existence of regression to a 5% significance level. 

e) Determine a prediction interval (90%) for a given X Y mean = 10
Answers

1.

3. a) Index of consumption of energy

b) 

<table>
<thead>
<tr>
<th>g1</th>
<th>SO</th>
<th>MQ</th>
<th>F</th>
<th>&gt; significância</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>814,881</td>
<td>814,881</td>
<td>120,3693</td>
<td>3,41E-05</td>
</tr>
<tr>
<td>6</td>
<td>40,61905</td>
<td>6,769841</td>
<td>855,5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ciencias</th>
<th>mpp</th>
<th>stdart</th>
<th>valor P</th>
<th>5% inferior</th>
<th>5% superior</th>
<th>95,0 inferior</th>
<th>95,0 superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,2962778</td>
<td>2,027378</td>
<td>55,77225</td>
<td>2,23E-09</td>
<td>108,1106</td>
<td>118,0322</td>
<td>108,1106</td>
<td>118,0322</td>
</tr>
<tr>
<td>2,380952</td>
<td>0,20074</td>
<td>-10,9713</td>
<td>3,41E-05</td>
<td>-2,69357</td>
<td>-1,71119</td>
<td>-2,69357</td>
<td>-1,71119</td>
</tr>
</tbody>
</table>
### Estatística de regressão

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>múltiplo</td>
<td>0,97597</td>
<td>1366</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adrado de</td>
<td>0,95252</td>
<td>0108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adrado de</td>
<td>0,944606</td>
<td>6793</td>
<td></td>
<td></td>
</tr>
<tr>
<td>o-padrão</td>
<td>2,601891</td>
<td>8184</td>
<td></td>
<td></td>
</tr>
<tr>
<td>servações</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### OVA

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$g^2$</td>
<td>$S^2$</td>
<td>$M^2$</td>
<td></td>
</tr>
<tr>
<td>regressão</td>
<td>1</td>
<td>814,881</td>
<td>814,881</td>
<td>$120,3693$</td>
</tr>
<tr>
<td>residual</td>
<td>6</td>
<td>40,61905</td>
<td>6,769841</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>7</td>
<td>855,5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeficientes</td>
<td>erro-padrão</td>
<td>$Stat$</td>
<td>$valor P$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>intercepto</td>
<td>113,0714286</td>
<td>2,027378</td>
<td>55,77225</td>
<td>$2,23E-09$</td>
<td>$108,1106$</td>
</tr>
<tr>
<td>del X 1</td>
<td>-2,202380952</td>
<td>0,20074</td>
<td>-10,9713</td>
<td>$3,41E-05$</td>
<td>$-2,69357$</td>
</tr>
</tbody>
</table>
### Estatística de regressão

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R múltiplo</td>
<td>0,968988</td>
</tr>
<tr>
<td>R² Quadrado</td>
<td>0,978098</td>
</tr>
<tr>
<td>R² quadrado ajustado</td>
<td>0,970797</td>
</tr>
<tr>
<td>Erro padrão</td>
<td>2,462619</td>
</tr>
<tr>
<td>Observações</td>
<td>5</td>
</tr>
</tbody>
</table>

### ANOVA

<table>
<thead>
<tr>
<th></th>
<th>gl</th>
<th>SQ</th>
<th>MQ</th>
<th>F</th>
<th>F de significação</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressão</td>
<td>1</td>
<td>812,6036</td>
<td>812,6036</td>
<td>133,9719</td>
<td>0,001385</td>
</tr>
<tr>
<td>Resíduo</td>
<td>3</td>
<td>18,19543</td>
<td>6,065476</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>830,8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Coeficientes

<table>
<thead>
<tr>
<th></th>
<th>Coeficiente</th>
<th>Erro padrão</th>
<th>Stat t</th>
<th>valor-P</th>
<th>95% inferiores</th>
<th>95% superiores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interseção</td>
<td>26,27679</td>
<td>3,211966</td>
<td>8,180904</td>
<td>0,003621</td>
<td>16,05487</td>
<td>36,4987</td>
</tr>
<tr>
<td>Variável X</td>
<td>4,258929</td>
<td>0,357954</td>
<td>11,57462</td>
<td>0,001385</td>
<td>3,087934</td>
<td>5,429923</td>
</tr>
</tbody>
</table>

---

76
Activity 2 Least squares

Introduction

The least squares method enables the adjustment of a straight line to observed data so that minimizes the sum of squared distances between the observed values and the fitted line, these measures the vertical distance and exactly corresponding to the differences between the observed values Y and the adjusted values of

1st Activity: Construct scatter diagram with an example

When adjusting one regression line to the observed data, the linear relationship between the two variables becomes great because they cancel all purposes residual variable. Adjusted straight, will then have the mathematical form:

Regression line

Deviation between the observed values and the adjusted values

More specifically, for a given value of the independent variable $X_i$ will have two values for $Y$: an observed value $Y_i$ and another which is given by the set straight $Y_i$. The difference between the two, how easily it shows exactly matches the random effect will residue:

Being $Y_i = a + bX_i + e_i$ and $Y_i = a + bX_i$,

Then $Y_i - Y_i = (a + bX_i + e_i) - (a + bX_i) = e_i$. 
By applying the method of least squares is intended to fit a straight line which minimizes the square sum of the waste, i.e., find values for the constants $a$ and $b$ that make this minimum sum:

The least squares method allows us to find a regression line, whose coefficients are given by:

- Intercept of the regression line
- Slope of the regression line

And that, for any line, has the advantage of being an optimal solution in that makes minimum distances between observed values for $Y$ and the straight set.

**Example**

For example presented in the previous activity, we intend to fit a least squares straight to the observed data in order to predict changes in producer prices. First you need to calculate $b$ for it must be known the following sums: and. Once be calculated by knowing the average of $X$ and $Y$, calculating the immediate.

<table>
<thead>
<tr>
<th>Year</th>
<th>$Y_i$</th>
<th>$X_i$</th>
<th>$X_iY_i$</th>
<th>$X_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>10.8</td>
<td>7.8</td>
<td>84.24</td>
<td>60.84</td>
</tr>
<tr>
<td>1985</td>
<td>4.4</td>
<td>5.7</td>
<td>25.08</td>
<td>32.49</td>
</tr>
<tr>
<td>1986</td>
<td>6.5</td>
<td>6.1</td>
<td>39.65</td>
<td>37.21</td>
</tr>
<tr>
<td>1987</td>
<td>7.8</td>
<td>7.7</td>
<td>60.06</td>
<td>59.29</td>
</tr>
<tr>
<td>1988</td>
<td>11.1</td>
<td>11.2</td>
<td>124.32</td>
<td>125.44</td>
</tr>
<tr>
<td>1989</td>
<td>13.5</td>
<td>11</td>
<td>148.5</td>
<td>121</td>
</tr>
<tr>
<td>1990</td>
<td>9.2</td>
<td>8.3</td>
<td>76.36</td>
<td>68.89</td>
</tr>
</tbody>
</table>

$\sum Y_i = 63.3$ $\sum X_i = 57.8$ $\sum X_i^2 = 505.16$

Number of observations $n = 7$

$$b = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{7 \times 558.21 - (63.3 \times 57.8)}{7 \times 505.16 - (57.8^2)}$$

$$= \frac{248.73}{195.28} = 1.274$$

$$a = \frac{\sum Y_i}{n} - b \frac{\sum X_i}{n} = \frac{63.3}{7} - 1.274 \times \frac{57.8}{7} = -1.47$$
Graphically, this regression line has the following configuration

![Graph showing regression line](image)

**Conclusion**

How to Interpret the coefficients a and b?

In a regression line, the calculated value for the coefficient gives us exactly the point of intersection of the line with the axis of the dependent variable. This value can be positive or negative and indicates the very general position of the regression line. Its simplest interpretation is as follows:

It corresponds to the value of the dependent variable Y when the effect of the independent variable X is null. For this reason the coefficient is referred to as constant. The b coefficient of the regression line exactly matches the slope of this line

\[ Y = a + bX \]

and therefore gives us know setting this straight: if it is positive the line will have a positive slope and vice versa: the greater its value, the stronger this slope. The coefficient b represents the expected variation of the dependent variable Y for each unit of the independent variable X. For example, if Y represents sales of a product X, and the advertising expenditure for the same product, if b = 8.0, this means that by $1,000 of increased spending on advertising, sales will increase $8,000.

**Exercise**

1. Consider the data:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>


a) Calculate $x_iy_i$

b) Calculate $X_iX_i$

c) Calculate $X$

d) Find $b$

e) Find $a$

2. The data below refer to the volume of rainfall (mm) and the volume of milk production type C (million liters) in a certain region of the country.

Set the data using a linear model

Granted, in 1980, a 24 mm rainfall, which should be expected volume of the milk type C?

<table>
<thead>
<tr>
<th>Year</th>
<th>Milk production (1,000,000 l)</th>
<th>Rainfall Index (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td>1971</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>1972</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>1973</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>1974</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>1975</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td>1976</td>
<td>32</td>
<td>27</td>
</tr>
<tr>
<td>1977</td>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>1978</td>
<td>30</td>
<td>26</td>
</tr>
<tr>
<td>1979</td>
<td>30</td>
<td>25</td>
</tr>
</tbody>
</table>

3. Consider the data

<table>
<thead>
<tr>
<th>X</th>
<th>-5</th>
<th>-3</th>
<th>0</th>
<th>1</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0,8</td>
<td>1,1</td>
<td>2,5</td>
<td>3,1</td>
<td>5,0</td>
<td>4,7</td>
</tr>
</tbody>
</table>

a) Build the scatter diagram.

b) Find the line of least squares

4. For a company to remain competitive, spending on research and development (R & D) are essential. To determine the optimal level of spending on R & D and its effect on the value of the company, was applied simple linear regression analysis, where:
Y = ratio of prices and earnings

X = ratio of expenditure on R & D and sales

The following data of 20 companies used in the study are:

<table>
<thead>
<tr>
<th>Empresas</th>
<th>y</th>
<th>x</th>
<th>Empresas</th>
<th>y</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5,6</td>
<td>0,003</td>
<td>11</td>
<td>8,4</td>
<td>0,058</td>
</tr>
<tr>
<td>2</td>
<td>7,2</td>
<td>0,004</td>
<td>12</td>
<td>11,1</td>
<td>0,058</td>
</tr>
<tr>
<td>3</td>
<td>8,1</td>
<td>0,009</td>
<td>13</td>
<td>11,1</td>
<td>0,067</td>
</tr>
<tr>
<td>4</td>
<td>9,9</td>
<td>0,021</td>
<td>14</td>
<td>13,2</td>
<td>0,080</td>
</tr>
<tr>
<td>5</td>
<td>6,0</td>
<td>0,023</td>
<td>15</td>
<td>13,4</td>
<td>0,080</td>
</tr>
<tr>
<td>6</td>
<td>8,2</td>
<td>0,030</td>
<td>16</td>
<td>11,5</td>
<td>0,083</td>
</tr>
<tr>
<td>7</td>
<td>6,3</td>
<td>0,035</td>
<td>17</td>
<td>9,8</td>
<td>0,091</td>
</tr>
<tr>
<td>8</td>
<td>10,0</td>
<td>0,037</td>
<td>18</td>
<td>16,1</td>
<td>0,092</td>
</tr>
<tr>
<td>9</td>
<td>8,5</td>
<td>0,044</td>
<td>19</td>
<td>7,0</td>
<td>0,064</td>
</tr>
<tr>
<td>10</td>
<td>13,2</td>
<td>0,051</td>
<td>20</td>
<td>5,9</td>
<td>0,028</td>
</tr>
</tbody>
</table>

a) Build the scatter plot

b) Adjust the line of least squares

c) Using the equation obtained to predict the value of y when x = $0.070

solution

1.a) 15
b) 10
c) 1
d) 1.50
e) 0.70
2.

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X</th>
<th>X^2</th>
<th>XY</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>23</td>
<td>529</td>
<td>598</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>21</td>
<td>441</td>
<td>525</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>28</td>
<td>784</td>
<td>868</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>27</td>
<td>729</td>
<td>783</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>23</td>
<td>529</td>
<td>621</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>28</td>
<td>784</td>
<td>868</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>27</td>
<td>729</td>
<td>864</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>22</td>
<td>484</td>
<td>616</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>26</td>
<td>676</td>
<td>780</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>625</td>
<td>750</td>
<td></td>
</tr>
</tbody>
</table>

\[ y = 289 \quad x = 250 \quad x^2 = 6310 \quad xy = 7273 \]

Determine the value of the parameter \( b \)

\[
b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}
\]

\[
b = \frac{(10 \times 7273) - (250 \times 289)}{(10 \times 6310) - 2502} = 0.8
\]

II - Determining the value of the parameter

\[
a = \frac{\sum y}{n} - b \times \frac{\sum x}{n}
\]
$a = 289 - 0.8 \cdot 250 = 8.9$

III - Equation of Straight Adjusted

$y = a + bx$

$y = 8.9 + 0.8x$

b) setting $x = 24$ mm have $y = 8.9 + 28.1 = 0.8x24$.

According to the model, we can expect 28.1 million liters produced for a 24 mm rainfall.

3. a)

\[ Y = 3.34 + 0.58x. \]
4.

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.726\textsuperscript{a}</td>
<td>.527</td>
<td>.501</td>
<td>2.0738</td>
</tr>
</tbody>
</table>

\textit{a. Predictors: (Constant), X}

**ANOVA\textsuperscript{b}**

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>86,404</td>
<td>1</td>
<td>86,404</td>
<td>20,090</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>77,414</td>
<td>18</td>
<td>4,301</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>163,818</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textit{a. Predictors: (Constant), X}

\textit{b. Dependent Variable: Y}
**Unit Summary**

Linear regression is a statistical tool used to predict future values from past values. The linear regression trend line uses squares method to draw a straight line through prices in order to minimize the distance between them and the resulting trend line.

Linear Regression is a statistical issue of great importance and applicability, not only the disciplines and related professions such as mathematics, engineering, statistics among others, but also realize its application in various areas such as medicine, pharmacology and even in music. Studying this topic will help the individual to improve their statistical perception providing you a complete logical reasoning.

**Unit Assessment**

- Construction and interpretation of the scatter gram, the use of Excel and SPSS
- Exercises on the method of least squares
- Graphical straight adjustment Representation
- Application of mathematical model

Assessment tools

- Construction of PORTFOLIOS
- Resolution of exercises
- Reporting

**Evaluation criteria**

<table>
<thead>
<tr>
<th>Number and type of evaluation</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Group work</td>
<td>10%</td>
</tr>
<tr>
<td>2 individual work</td>
<td>15%</td>
</tr>
<tr>
<td>4 Solving exercises</td>
<td>20%</td>
</tr>
<tr>
<td>3 Research work</td>
<td>15%</td>
</tr>
<tr>
<td>1 Case Study</td>
<td>15%</td>
</tr>
<tr>
<td>1 Final exam</td>
<td>25%</td>
</tr>
</tbody>
</table>
Evaluation exercise

Considering the data in the following table:

<table>
<thead>
<tr>
<th>Provided units</th>
<th>heat</th>
</tr>
</thead>
<tbody>
<tr>
<td>11,878</td>
<td>179476</td>
</tr>
<tr>
<td>13,087</td>
<td>190724</td>
</tr>
<tr>
<td>11,623</td>
<td>173965</td>
</tr>
<tr>
<td>13,474</td>
<td>196530</td>
</tr>
<tr>
<td>11,584</td>
<td>172064</td>
</tr>
<tr>
<td>10,949</td>
<td>162246</td>
</tr>
<tr>
<td>14,52</td>
<td>212716</td>
</tr>
<tr>
<td>4,056</td>
<td>59639</td>
</tr>
<tr>
<td>14,344</td>
<td>211407</td>
</tr>
<tr>
<td>13,316</td>
<td>194961</td>
</tr>
<tr>
<td>15,852</td>
<td>233603</td>
</tr>
<tr>
<td>13,26</td>
<td>194932</td>
</tr>
<tr>
<td>14,69</td>
<td>213024</td>
</tr>
</tbody>
</table>

By using the SPSS

a) Determine the equation of the regression line

b) Build the scatter diagram.
Solution

a) \( Y = -0.869 + 7.20 \times 10^{-5} x \)

Lecture and other resources

- [www.youtube.com/watch?v=x42skwrbiek](https://www.youtube.com/watch?v=x42skwrbiek)
- [https://www.google.pt/#q=Regres%C3%A3o+Linear+simples](https://www.google.pt/#q=Regres%C3%A3o+Linear+simples)
- [slideshare.net/monica_lima/regresso-linear-simples](http://www.youtube.com/watch?v=L_grlTzMd7c)
Unit 4: Applications of Probability and Statistics in ACS

Unit Introduction
Statistics may have little to offer the search architectures in a data mining search, but a great deal to offer in evaluating hypotheses in the search, in evaluating the results of the search, and in applying the results in computing. Probability and statistics have a widespread applications in the area of applied science e.g artificial intelligence, Virtual reality.

Unit Objectives
Upon completion of this unit you should be able to:

1. Apply probability and statistics in machine learning.
2. Solve problems in ACS using Bayesian methods.
3. Define Supervised and unsupervised learning.
4. Identify the different types of classification and give examples.
5. Define clustering and gives examples.

Key Terms

**Machine Learning**: It builds statistical models of data in order to recognize complex patterns and to make decisions based on observed data. Examples include: classification (recognition of faces or handwritings, predictions (stock market, election), data mining, etc.

**Labeled Data**: there is a specially designated attribute and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen. Data of this kind is called labeled.

**Unlabeled Data**: Data that does not have any specially designated attribute is called unlabelled.

**Supervised learning**: Data mining using labeled data is known as supervised learning.

**Unsupervised learning**: Data mining of unlabelled data is known as unsupervised learning.
Classification: is a task that occurs very frequently in everyday life. Essentially it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes.

Nearest Neighbor Matching: This method relies on identifying (say) the five examples that are ‘closest’ in some sense to an unclassified one.

Training Data Set: The training set constitutes the results of a sample of trials that we can use to predict the classification of other (unclassified) instances.

Instance: An instance comprises the values of a number of attributes and the corresponding classification.

Classification Tree: One way of generating classification rules is via an intermediate tree-like structure called a classification tree or a decision tree.

Neural Network: This is a complex modeling technique based on a model of a human neuron.

Association Rules: A training set is use to find any relationship that exists amongst the values of variables, generally in the form of rules known as association rules.

Clustering: Clustering algorithms examine data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased or prior claims experience.

Learning Activities

Activity 1 Naive Bayesian and k-means neighbor

Introduction

Classification is a task that occurs very frequently in everyday life. Essentially it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes. The term ‘mutually exhaustive and exclusive’ simply means that each object must be assigned to precisely one class, i.e. never to more than one and never to no class at all. For example, a hospital may want to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness, an opinion polling company may wish to classify people interviewed into those who are likely to vote for each of a number of political parties or are undecided, or we may wish to classify a student project as distinction, merit, pass or fail.
Classification for supervised learning can be done in two ways: Naive Baye's and K-means neighbor

**Activity Details**

**Lesson 1: Naïve Baye's classifier**

The Naive Bayes algorithm gives us a way of combining the prior probability and conditional probabilities in a single formula, which we can use to calculate the probability of each of the possible classifications in turn.

For example, a fruit may be considered to be an apple if it is red, round, and about 3” in diameter. A naive Baye's classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness and diameter features.

**Example 1:**

Given a set of k mutually exclusive and exhaustive classifications $C_1, C_2, \ldots, C_k$ which have prior probabilities $P(C_1), P(C_2), \ldots, P(C_k)$, respectively, and n attributes for a given instance have values $V_1, V_2, \ldots, V_n$ respectively, the posterior probability of class $i$ occurring for the specified instance can be shown to be proportional to $P(C_i) \times P(A_1 = v_1 \text{ and } A_2 = v_2 \ldots \text{ and } A_n = v_n | C_i)$. Making the assumption that the attributes are independent, the value of this expression can be calculated using the product $P(C_i) \times P(A_1 = v_1 | C_i) \times P(A_2 = v_2 | C_i) \times \ldots \times P(A_n = v_n | C_i)$. We calculate this product for each value of i from 1 to k and choose the classification that has the largest value.

**Example 2:**

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contain a continuous attribute $X$. We first segment the data by the class, and then compute the mean and variance of $X$ in each class. Let $\mu_c$ be the mean of the value $X$ associated with class $c$, and let $\sigma_c^2$ be the variance of the values in $X$ associated with class $c$. Then, the probability distribution of some value given a class, $P(X = v | c)$, can be computed by into the equation for a Normal Distribution parameterized by $\mu_c$ and $\sigma_c^2$. That is,

$$P(X = v | c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$
For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting.

Exercise: Classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size.

Solution:

Using the training set below

<table>
<thead>
<tr>
<th>sex</th>
<th>height (feet)</th>
<th>weight (lbs)</th>
<th>foot size (inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>6</td>
<td>180</td>
<td>12</td>
</tr>
<tr>
<td>male</td>
<td>5.92 (5’11”)</td>
<td>190</td>
<td>11</td>
</tr>
<tr>
<td>male</td>
<td>5.58 (5’7”)</td>
<td>170</td>
<td>12</td>
</tr>
<tr>
<td>male</td>
<td>5.92 (5’11”)</td>
<td>165</td>
<td>10</td>
</tr>
<tr>
<td>female</td>
<td>5</td>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>female</td>
<td>5.5 (5’6”)</td>
<td>150</td>
<td>8</td>
</tr>
<tr>
<td>female</td>
<td>5.42 (5’5”)</td>
<td>130</td>
<td>7</td>
</tr>
</tbody>
</table>

Assuming that we have equi-probable classes so $P(\text{male}) = P(\text{female}) = 0.5$. This prior probability distribution might be based on our knowledge on frequency in the training set.

**Testing**

Below is a sample to be classified as a male or female.

<table>
<thead>
<tr>
<th>sex</th>
<th>height (feet)</th>
<th>weight (lbs)</th>
<th>foot size (inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>6</td>
<td>130</td>
<td>8</td>
</tr>
</tbody>
</table>

We wish to determine which posterior is greater, male or female. For the classification as male the posterior probability is given by

For the classification as female the posterior probability is given by

The evidence (also termed normalizing constant) may be calculated:

$$
evidence = P(\text{male}) p(\text{height} | \text{male}) p(\text{weight} | \text{male}) p(\text{footsize} | \text{male}) + P(\text{female}) p(\text{height} | \text{female}) p(\text{weight} | \text{female}) p(\text{footsize} | \text{female})$$
We now determine the probability distribution for the sex of the sample.

\[
P(\text{male}) = 0.5
\]

\[
P(\text{height} | \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(6 - \mu)^2}{2\sigma^2}\right) \approx 1.5789
\]

where \(3.5033 \times 10^{-2}\) and \(\sigma^2 = 3.5033 \times 10^{-2}\) are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 is OK here – it is a probability density rather than a probability, because height is a continuous variable.

\[
P(\text{weight} | \text{male}) = 5.9881 \times 10^{-6}
\]

\[
P(\text{footsize} | \text{male}) = 1.3112 \times 10^{-3}
\]

Posterior numerator (male) = their product = \(6.1984 \times 10^{-9}\)

\[
P(\text{female}) = 0.5
\]

\[
P(\text{height} | \text{female}) = 2.2346 \times 10^{-1}
\]

\[
P(\text{weight} | \text{female}) = 1.6789 \times 10^{-2}
\]

\[
P(\text{footsize} | \text{female}) = 2.8669 \times 10^{-1}
\]

Posterior numerator (female) = their product = \(5.3778 \times 10^{-4}\)

Since posterior numerator is greater in the female case, we predict the sample is female.

Lesson 2: k-nearest neighbor (k-NN)

In practice there are likely to be many more instances in the training set but the same principle applies. It is usual to base the classification on those of the k nearest neighbors (where k is a small integer such as 3 or 5), not just the nearest one. The method is then known as k-Nearest Neighbor or just k-NN classification.

Basic k-Nearest Neighbor Classification Algorithm

- Find the k training instances that are closest to the unseen instance.
- Take the most commonly occurring classification for these k instances.

Supposing we have a training set with just two instances such as the following:
There are six attribute values, followed by a classification (positive or negative). We are then given a third instance

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>6.4</td>
<td>8.3</td>
<td>low</td>
<td>negative</td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>18.2</td>
<td>4.7</td>
<td>high</td>
<td>positive</td>
<td></td>
</tr>
</tbody>
</table>

What should its classification be?

Even without knowing what the six attributes represent, it seems intuitively obvious that the unseen instance is nearer to the first instance than to the second. In the absence of any other information, we could reasonably predict its classification using that of the first instance, i.e. as ‘negative’.

We can illustrate k-NN classification diagrammatically when the dimension (i.e. the number of attributes) is small. The following example illustrates the case where the dimension is just 2.
Table 1 show a training set with 20 instances, each giving the values of two attributes and an associated classification. How can we estimate the classification for an ‘unseen’ instance where the first and second attributes are 9.1 and 11.0, respectively? For this small number of attributes we can represent the training set as 20 points on a two-dimensional graph with values of the first and second attributes measured along the horizontal and vertical axes, respectively. Each point is labeled with a + or − symbol to indicate that the classification is positive or negative, respectively. The result is shown in Figure 2.

<table>
<thead>
<tr>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>6.3</td>
<td>−</td>
</tr>
<tr>
<td>1.4</td>
<td>8.1</td>
<td>−</td>
</tr>
<tr>
<td>2.1</td>
<td>7.4</td>
<td>−</td>
</tr>
<tr>
<td>2.6</td>
<td>14.3</td>
<td>+</td>
</tr>
<tr>
<td>6.8</td>
<td>12.6</td>
<td>−</td>
</tr>
<tr>
<td>8.8</td>
<td>9.8</td>
<td>+</td>
</tr>
<tr>
<td>9.2</td>
<td>11.6</td>
<td>−</td>
</tr>
<tr>
<td>10.8</td>
<td>9.6</td>
<td>+</td>
</tr>
<tr>
<td>11.8</td>
<td>9.9</td>
<td>+</td>
</tr>
<tr>
<td>12.4</td>
<td>6.5</td>
<td>+</td>
</tr>
<tr>
<td>12.8</td>
<td>1.1</td>
<td>−</td>
</tr>
<tr>
<td>14</td>
<td>19.9</td>
<td>−</td>
</tr>
<tr>
<td>14.2</td>
<td>18.5</td>
<td>−</td>
</tr>
<tr>
<td>15.6</td>
<td>17.4</td>
<td>−</td>
</tr>
<tr>
<td>15.8</td>
<td>12.2</td>
<td>−</td>
</tr>
<tr>
<td>16.6</td>
<td>6.7</td>
<td>+</td>
</tr>
<tr>
<td>17.4</td>
<td>4.5</td>
<td>+</td>
</tr>
<tr>
<td>18.2</td>
<td>6.9</td>
<td>+</td>
</tr>
<tr>
<td>19</td>
<td>3.4</td>
<td>−</td>
</tr>
<tr>
<td>19.6</td>
<td>11.1</td>
<td>+</td>
</tr>
</tbody>
</table>
Using the training dataset in the table above, the figure below is obtained.

A circle has been added to enclose the five nearest neighbors of the unseen instance, which is shown as a small circle close to the centre of the larger one.

The five nearest neighbors are labeled with three + signs and two − signs, so a basic 5-NN classifier would classify the unseen instance as ‘positive’ by a form of majority voting. There are other possibilities, for example the ‘votes of each of the k nearest neighbors can be weighted, so that the classifications of closer neighbors are given greater weight than the classifications of more distant ones.

We can represent two points in two dimensions (‘in two-dimensional space’ is the usual term) as

\((a_1, a_2)\) and \((b_1, b_2)\) and visualize them as points in a plane.

When there are three attributes we can represent the points by \((a_1, a_2, \ldots, a_n)\) and \((b_1, b_2, \ldots, b_n)\) and think of them as points in a room with three axes at right angles. As the number of dimensions (attributes) increases it rapidly becomes impossible to visualize them, at least for anyone who is not a physicist (and most of those who are).

When there are \(n\) attributes, we can represent the instances by the points \((a_1, a_2, \ldots, a_n)\) and \((b_1, b_2, \ldots, b_n)\) in ‘\(n\)-dimensional space’.
Conclusion

The Naive Bayes approach is a very popular one, which often works well. However it has a number of potential problems, the most obvious one being that it relies on all attributes being categorical. In practice, many datasets have a combination of categorical and continuous attributes, or even only continuous attributes. This problem can be overcome by converting the continuous attributes to categorical ones, using a different method.

Group Activity

In a group of five students, answer the following questions. It should be submitted after one week. The total mark is out of 50.

1. Using the Naive Bayes classification algorithm with the train dataset in the table below, calculate the most likely classification for the following unseen instances.

Testing Set:

| weekday | summer | high  | heavy | ????
|---------|--------|-------|-------|------
| sunday  | summer | normal| slight| ????


Training Data set:

<table>
<thead>
<tr>
<th>Day</th>
<th>Season</th>
<th>Wind</th>
<th>Rain</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>weekday</td>
<td>Spring</td>
<td>none</td>
<td>none</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Winter</td>
<td>none</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Winter</td>
<td>none</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Winter</td>
<td>high</td>
<td>heavy</td>
<td>late</td>
</tr>
<tr>
<td>Saturday</td>
<td>Summer</td>
<td>normal</td>
<td>none</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Autumn</td>
<td>normal</td>
<td>none</td>
<td>Very late</td>
</tr>
<tr>
<td>holiday</td>
<td>Summer</td>
<td>high</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>Sunday</td>
<td>Summer</td>
<td>normal</td>
<td>none</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Winter</td>
<td>high</td>
<td>heavy</td>
<td>Very late</td>
</tr>
<tr>
<td>weekday</td>
<td>Summer</td>
<td>none</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>Saturday</td>
<td>Spring</td>
<td>high</td>
<td>heavy</td>
<td>cancelled</td>
</tr>
<tr>
<td>weekday</td>
<td>Summer</td>
<td>high</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>Saturday</td>
<td>Winter</td>
<td>normal</td>
<td>none</td>
<td>late</td>
</tr>
<tr>
<td>weekday</td>
<td>Summer</td>
<td>high</td>
<td>none</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Winter</td>
<td>normal</td>
<td>heavy</td>
<td>Very late</td>
</tr>
<tr>
<td>Saturday</td>
<td>Autumn</td>
<td>high</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Autumn</td>
<td>none</td>
<td>heavy</td>
<td>On time</td>
</tr>
<tr>
<td>holiday</td>
<td>Spring</td>
<td>normal</td>
<td>slight</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Spring</td>
<td>normal</td>
<td>none</td>
<td>On time</td>
</tr>
<tr>
<td>weekday</td>
<td>Spring</td>
<td>normal</td>
<td>slight</td>
<td>On time</td>
</tr>
</tbody>
</table>

Using the training set shown in Table 1 and the Euclidean distance measure, calculate the 5-nearest neighbors of the instance with first and second attributes 9.1 and 11.0, respectively.

Activity 2 Decision Tree for classification

Introduction

In this activity students will learn how to represent data graphically and derive information from it.
**Activity Details**

Decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

**Example**

We have a dataset in the form of a table containing students’ grades on five subjects (the values of attributes SoftEng, ARIN, HCI, CSA and Project) and their overall degree classifications. The row of dots indicates that a number of rows have been omitted in the interests of simplicity. We want to find some way of predicting the classification for other students given only their grade ‘profiles’.

<table>
<thead>
<tr>
<th>Soft Eng</th>
<th>ARIN</th>
<th>HCI</th>
<th>CSA</th>
<th>Project</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>Second</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>Second</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>Second</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>First</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>First</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>Second</td>
</tr>
<tr>
<td>..........</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>..........</td>
<td>..........</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>First</td>
</tr>
</tbody>
</table>

One way of generating classification rules is via an intermediate tree-like structure called a classification tree or a decision tree.

The figure below shows a possible decision tree corresponding to the degree classification data.
Figure 3: Decision Tree for Degree Classification Data

The decision tree can be used for classification purpose. Refer to Principles of Data Mining – Max Bramer (Page 42 – 47).

**For reading purpose, refer to**


**Conclusion**

In this activity, the use of tree diagram in computer science is illustrated. It is a very useful way in organizing the data set presented to them.

**Group Activity**

This is an activity to be done by a group of five (5). Its purpose is to help the students better understand and apply the concepts learned previously.

A golfer decides whether or not to play each day on the basis of the weather. The table below shows the results of two weeks (14 days) of observations of weather conditions and the decision on whether or not to play.
<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temp (°F)</th>
<th>Humidity</th>
<th>Windy</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>75</td>
<td>70</td>
<td>1</td>
<td>play</td>
</tr>
<tr>
<td>Sunny</td>
<td>80</td>
<td>90</td>
<td>1</td>
<td>Don’t play</td>
</tr>
<tr>
<td>Sunny</td>
<td>85</td>
<td>85</td>
<td>0</td>
<td>Don’t play</td>
</tr>
<tr>
<td>Sunny</td>
<td>72</td>
<td>95</td>
<td>0</td>
<td>Don’t play</td>
</tr>
<tr>
<td>Sunny</td>
<td>69</td>
<td>70</td>
<td>0</td>
<td>play</td>
</tr>
<tr>
<td>Overcast</td>
<td>72</td>
<td>90</td>
<td>1</td>
<td>play</td>
</tr>
<tr>
<td>Overcast</td>
<td>83</td>
<td>78</td>
<td>0</td>
<td>play</td>
</tr>
<tr>
<td>Overcast</td>
<td>64</td>
<td>65</td>
<td>1</td>
<td>play</td>
</tr>
<tr>
<td>Overcast</td>
<td>81</td>
<td>75</td>
<td>0</td>
<td>play</td>
</tr>
<tr>
<td>Rain</td>
<td>71</td>
<td>80</td>
<td>1</td>
<td>Don’t play</td>
</tr>
<tr>
<td>Rain</td>
<td>65</td>
<td>70</td>
<td>1</td>
<td>Don’t play</td>
</tr>
<tr>
<td>Rain</td>
<td>75</td>
<td>80</td>
<td>0</td>
<td>play</td>
</tr>
<tr>
<td>Rain</td>
<td>68</td>
<td>80</td>
<td>0</td>
<td>play</td>
</tr>
<tr>
<td>Rain</td>
<td>70</td>
<td>96</td>
<td>0</td>
<td>play</td>
</tr>
</tbody>
</table>

With the help of a decision tree, assuming the golfer is acting consistently, what are the rules that determine the decision whether or not to play each day? If tomorrow the values of Outlook, Temperature, Humidity and Windy were sunny, 74°F, 77% and false respectively, what would the decision be?
Activity 3 Clustering

Introduction

Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters.

In many fields there are obvious benefits to be had from grouping together similar objects. There are various ways of clustering:

1. Euclidean distance
2. k-means clustering

Activity Details

Lesson 1: Euclidean Distance

Typically, the basic data used to form clusters is a table of measurements on several variables where each column represents a variable and a row represents an object often referred to in statistics as a case. Thus the set of rows are to be grouped so that similar cases are in the same group. The number of groups may be specified or has to be determined from the data.

A popular distance measure based on variables that take on continuous values is to standardize the values by dividing by the standard deviation (sometimes other measures such as range are used) and then to compute the distance between objects using the Euclidean metric. The Euclidean distance $d_{ij}$ between two cases, i and j with variable values $(x_{i1}, x_{i2}, \ldots, x_{ip})$ and $(x_{j1}, x_{j2}, \ldots, x_{jp})$ is defined by:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Example: Public Utilities Data (corporate data on 22 US public utilities)
<table>
<thead>
<tr>
<th>No.</th>
<th>Company</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arizona Public Service</td>
<td>1.06</td>
<td>9.2</td>
<td>151</td>
<td>54.4</td>
<td>1.6</td>
<td>9077</td>
<td>0</td>
<td>0.628</td>
</tr>
<tr>
<td>2</td>
<td>Boston Edison Company</td>
<td>0.89</td>
<td>10.3</td>
<td>202</td>
<td>57.9</td>
<td>2.2</td>
<td>5088</td>
<td>25.3</td>
<td>1.555</td>
</tr>
<tr>
<td>3</td>
<td>Central Louisiana Electric Co.</td>
<td>1.43</td>
<td>15.4</td>
<td>113</td>
<td>53</td>
<td>3.4</td>
<td>9212</td>
<td>0</td>
<td>1.058</td>
</tr>
<tr>
<td>4</td>
<td>Commonwealth Edison Co.</td>
<td>1.02</td>
<td>11.2</td>
<td>168</td>
<td>56</td>
<td>0.3</td>
<td>6423</td>
<td>34.3</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>Consolidated Edison Co. (NY)</td>
<td>1.49</td>
<td>8.8</td>
<td>1.92</td>
<td>51.2</td>
<td>1</td>
<td>3300</td>
<td>15.6</td>
<td>2.044</td>
</tr>
<tr>
<td>6</td>
<td>Florida Power and Light</td>
<td>1.32</td>
<td>13.5</td>
<td>111</td>
<td>60</td>
<td>-2.2</td>
<td>11127</td>
<td>22.5</td>
<td>1.241</td>
</tr>
<tr>
<td>7</td>
<td>Hawaiian Electric Co.</td>
<td>1.22</td>
<td>12.2</td>
<td>175</td>
<td>67.6</td>
<td>2.2</td>
<td>7642</td>
<td>0</td>
<td>1.652</td>
</tr>
<tr>
<td>8</td>
<td>Idaho Power Co.</td>
<td>1.1</td>
<td>9.2</td>
<td>245</td>
<td>57</td>
<td>3.3</td>
<td>13082</td>
<td>0</td>
<td>0.309</td>
</tr>
<tr>
<td>9</td>
<td>Kentucky Utilities Co.</td>
<td>1.34</td>
<td>13</td>
<td>168</td>
<td>60.4</td>
<td>7.2</td>
<td>8406</td>
<td>0</td>
<td>0.862</td>
</tr>
<tr>
<td>10</td>
<td>Madison Gas &amp; Electric Co.</td>
<td>1.12</td>
<td>12.4</td>
<td>197</td>
<td>53</td>
<td>2.7</td>
<td>6455</td>
<td>39.2</td>
<td>0.623</td>
</tr>
<tr>
<td>11</td>
<td>Nevada Power Co.</td>
<td>0.75</td>
<td>7.5</td>
<td>173</td>
<td>51.5</td>
<td>6.5</td>
<td>17441</td>
<td>0</td>
<td>0.768</td>
</tr>
<tr>
<td>12</td>
<td>New England Electric Co.</td>
<td>1.13</td>
<td>10.9</td>
<td>178</td>
<td>62</td>
<td>3.7</td>
<td>6154</td>
<td>0</td>
<td>1.897</td>
</tr>
<tr>
<td>13</td>
<td>Northern States Power Co.</td>
<td>1.15</td>
<td>12.7</td>
<td>199</td>
<td>53.7</td>
<td>6.4</td>
<td>7179</td>
<td>50.2</td>
<td>0.527</td>
</tr>
<tr>
<td>14</td>
<td>Oklahoma Gas and Electric Co.</td>
<td>1.09</td>
<td>12</td>
<td>96</td>
<td>49.8</td>
<td>1.4</td>
<td>9673</td>
<td>0</td>
<td>0.588</td>
</tr>
<tr>
<td>15</td>
<td>Pacific Gas &amp; Electric Co.</td>
<td>0.96</td>
<td>7.6</td>
<td>164</td>
<td>62.2</td>
<td>-0.1</td>
<td>6468</td>
<td>0.9</td>
<td>1.4</td>
</tr>
<tr>
<td>16</td>
<td>Puget Sound Power &amp; Light Co.</td>
<td>1.16</td>
<td>9.9</td>
<td>252</td>
<td>56</td>
<td>9.2</td>
<td>15991</td>
<td>0</td>
<td>0.62</td>
</tr>
<tr>
<td>17</td>
<td>San Diego Gas &amp; Electric Co.</td>
<td>0.76</td>
<td>6.4</td>
<td>136</td>
<td>61.9</td>
<td>9</td>
<td>5714</td>
<td>8.3</td>
<td>1.92</td>
</tr>
<tr>
<td>18</td>
<td>The Southern Co.</td>
<td>1.05</td>
<td>12.6</td>
<td>150</td>
<td>56.7</td>
<td>2.7</td>
<td>10140</td>
<td>0</td>
<td>1.108</td>
</tr>
<tr>
<td>19</td>
<td>Texas Utilities Co.</td>
<td>1.16</td>
<td>11.7</td>
<td>104</td>
<td>54</td>
<td>-2.1</td>
<td>13507</td>
<td>0</td>
<td>0.636</td>
</tr>
<tr>
<td>20</td>
<td>Wisconsin Electric Power Co.</td>
<td>1.2</td>
<td>11.8</td>
<td>148</td>
<td>59.9</td>
<td>3.5</td>
<td>7297</td>
<td>41.1</td>
<td>0.702</td>
</tr>
<tr>
<td>21</td>
<td>United Illuminating Co.</td>
<td>1.04</td>
<td>8.6</td>
<td>204</td>
<td>61</td>
<td>3.5</td>
<td>6650</td>
<td>0</td>
<td>2.116</td>
</tr>
<tr>
<td>22</td>
<td>Virginia Electric &amp; Power Co.</td>
<td>1.07</td>
<td>9.3</td>
<td>1784</td>
<td>54.3</td>
<td>5.9</td>
<td>10093</td>
<td>26.6</td>
<td>1.306</td>
</tr>
</tbody>
</table>
We are interested in forming groups of similar utilities. The objects to be clustered are the utilities. There are 8 measurements on each utility described in the table below. An example where clustering would be useful is a study to predict the cost impact of deregulation. To do the requisite analysis economists would need to build a detailed cost model of the various utilities. The objects to be clustered are the utilities and there are 8 measurements on each utility.

The idea behind this set of techniques is to start with each cluster comprising of exactly one object and then progressively agglomerating (combining) the two nearest clusters until there is just one cluster left consisting of all the objects.

Nearness of clusters is based on a measure of distance between clusters. All agglomerative methods require as input a distance measure between all the objects that are to be clustered. This measure of distance between objects is mapped into a metric for the distance between clusters (sets of objects) metrics for the distance between two clusters.

The results of the distance are calculated using SPSS and are shown in the table 1 below.

### 2. Nearest neighbor clustering

Here the distance between two clusters is defined as the distance between the nearest pair of objects with one object in the pair belonging to a distinct cluster. If cluster A is the set of objects \( A_1, A_2, \ldots, A_m \) and cluster B is \( B_1, B_2, \ldots, B_n \), the single linkage distance between A and B is \( \sqrt{\frac{m+n}{mn}} \sum_{i=1}^{m+n} \text{distance}(A_i, B_j) \). This method has a tendency to cluster together at an early stage objects that are distant from each other in the same cluster because of a chain of intermediate objects in the same cluster. Such clusters have elongated sausage-like shapes when visualized as objects in space.

| X1: Fixed-charge covering ratio (income/debt) |
| X2: Rate of return on capital |
| X3: Cost per KW capacity in place |
| X4: Annual Load Factor |
| X5: Peak KWH demand growth from 1974 to 1975 |
| X6: Sales (KWH use per year) |
| X7: Percent Nuclear |
| X8: Total fuel costs (cents per KWH) |
3. Group Average (also called average linkage).

Here the distance between two clusters is defined as the average distance between all possible pairs of objects with one object in each pair belonging to a distinct cluster. If cluster $A = A_1, A_2, \ldots, A_m$ and cluster $B = B_1, B_2, \ldots, B_n$, the single linkage distance between $A$ and $B$ is 
$$d(A, B) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d(A_i, B_j)$$

The nearest neighbor clusters for the utilities are displayed in Figure 1 below in a useful graphic format called a Dendogram. For any given number of clusters we can determine the cases in the clusters by sliding a vertical line from left to right until the number of horizontal intersections of the vertical line equals the desired number of clusters. For example, if we wanted to form 6 clusters we would find that the clusters are: {1, 18, 14, 19, 9, 10, 13, 4, 20, 2, 12, 21, 7, 15, 22, 6}; {3}; {8, 16}; {17}; {11}; and {5}. Notice that if we wanted 5 clusters they would be the same as for six with the exception that the first two clusters above would be merged into one cluster. In general all hierarchical methods have clusters that are nested within each other as we decrease the number of clusters we desire.
Table 1: Distances based on standardized variable values.

For the average linkage, SPSS is used to construct a hierarchical cluster. It is illustrated in Figure 2.

Figure 1: Dendogram - Single Linkage

Figure 2: Average Linkage

Average Linkage (Between Groups)

Agglomeration Schedule
<table>
<thead>
<tr>
<th>Stage</th>
<th>Cluster Combined</th>
<th>Coefficients</th>
<th>Stage Cluster First Appears</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster 1</td>
<td>Cluster 2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>1905.226</td>
</tr>
</tbody>
</table>
Conclusion

This unit illustrates how important probability and statistics is important to computer sciences. There are various applications statistics into it specially in analyzing the results obtained in computing. Few examples have been used and the results are presented using SPSS.

Unit Summary

In this unit, students were presented with some basic applications of probability and statistics in applied computer science. It illustrates also the idea that SPSS could be used to solve most of the statistical problems in computer science. More applications of could be find in the reference text. Throughout the unit, students did learn about classification, nearest neighbor, Naïve Baye’s, etc.
Assignment

Instructions

After reading the case use the data provided to answer the questions that follow

Case: German Credit The German Credit data set (available at ftp.ics.uci.edu/pub/machine-learning-databases/statlog/) contains observations on 30 variables for 1000 past applicants for credit. Each applicant was rated as “good credit” (700 cases) or “bad credit” (300 cases). New applicants for credit can also be evaluated on these 30 “predictor” variables. We want to develop a credit scoring rule that can be used to determine if a new applicant is a good credit risk or a bad credit risk, based on values for one or more of the predictor variables. All the variables are explained in Table 1.1. (Note: The original data set had a number of categorical variables, some of which have been transformed into a series of binary variables so that they can be appropriately handled by SPSS. Several ordered categorical variables have been left as is; to be treated by SPSS as numerical. The data has been organized in the spreadsheet German CreditI.xls)

The data set and the problem (2) can be obtained from this link: http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/assignments/

Table 1.1 Variables for the German Credit data.

<table>
<thead>
<tr>
<th>Codelist</th>
<th>Var. #</th>
<th>Variable Name</th>
<th>Description</th>
<th>Variable Type</th>
<th>Code Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>OBS#</td>
<td>Observation No.</td>
<td>Categorical</td>
<td>0 : &lt; 0 DM</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>CHK_ACCT</td>
<td>Checking account status</td>
<td>Categorical</td>
<td>1: 0 &lt; ...&lt; 200 DM</td>
<td>1: 0 &lt; ...&lt; 200 DM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 : =&gt; 200 DM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3: no checking account</td>
</tr>
<tr>
<td>3.</td>
<td>DURATION</td>
<td>Duration of credit in months</td>
<td>Numerical</td>
<td>0 : &lt; 0 DM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Credit history</td>
<td>Categorical</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>--------</td>
<td>----------------</td>
<td>-------------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>HISTORY</td>
<td>Credit history</td>
<td>Categorical</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0: no credits taken</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1: all credits at this bank paid back duly</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2: existing credits paid back duly till now</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3: delay in paying off in the past</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4: critical account</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>NEW.Car</td>
<td>Purpose of credit</td>
<td>Binary</td>
<td>car (new) 0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>USED.Car</td>
<td>Purpose of credit</td>
<td>Binary</td>
<td>car (used) 0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>FURNITURE</td>
<td>Purpose of credit</td>
<td>Binary</td>
<td>furniture/equipment 0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>RADIO/TV</td>
<td>Purpose of credit</td>
<td>Binary</td>
<td>radio/television 0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>EDUCATION</td>
<td>Purpose of credit</td>
<td>Binary</td>
<td>education 0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>RETRAINING</td>
<td>Purpose of credit</td>
<td>Binary</td>
<td>retraining 0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>AMOUNT</td>
<td>Credit amount</td>
<td>Numerical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>SAV_ACCT</td>
<td>Average balance in savings account</td>
<td>Categorical</td>
<td>0 : &lt; 100 DM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 : 100&lt;= ... &lt; 500 DM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 : 500&lt;= ... &lt; 1000 DM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 : =&gt;1000 DM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 : unknown/ no savings account</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EMPLOYMENT</td>
<td>Present employment since</td>
<td>Categorical</td>
<td>0 : unemployed</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------</td>
<td>---------------------------</td>
<td>-------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>1: &lt; 1 year</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 : 1 &lt;= ... &lt; 4 years</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 : 4 &lt;=... &lt; 7 years</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 : &gt;= 7 years</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>INSTALL_RATE</td>
<td>Installment rate as % of disposable income</td>
<td>Numerical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>MALE_DIV</td>
<td>Applicant is male and divorced</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>MALE_SINGLE</td>
<td>Applicant is male and single</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>MALE_MAR_WID</td>
<td>Applicant is male and married or a widower</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>CO-APPLICANT</td>
<td>Application has a co-applicant</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>GUARANTOR</td>
<td>Applicant has a guarantor</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>PRESENT_RESIDENT</td>
<td>Present resident since - years</td>
<td>Categorical</td>
<td>0: &lt;= 1 year</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1:&lt;...&lt;=2 years</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2:&lt;...&lt;=3 years</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3:&gt;=4years</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>REAL_ESTATE</td>
<td>Applicant owns real estate</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>PROP_UNKN_NONE</td>
<td>Applicant owns no property (or unknown)</td>
<td>Binary</td>
<td>0: No, 1: Yes</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>AGE</td>
<td>Age in years</td>
<td>Numerical</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Introduction to Statistics and Probability**

110
The consequences of misclassification have been assessed as follows: the costs of a false positive (incorrectly saying an applicant is a good credit risk) outweigh the cost of a false negative (incorrectly saying an applicant is a bad credit risk) by a factor of five. This can be summarized in the following table.
The opportunity cost table was derived from the average net profit per loan as shown below:

![Table 1.3 Opportunity Cost Table (in deutch Marks)](image)

Table 1.4 Average Net Profit

Let us use this table in assessing the performance of the various models because it is simpler to explain to decision-makers who are used to thinking of their decision in terms of net profits.

1. Review the predictor variables and guess from their definition at what their role might be in a credit decision. Are there any surprises in the data?

2. Divide the data randomly into training (60%) and validation (40%) partitions, and develop classification models using the following data mining techniques in SPSS:

   - Classification trees
   - Neural networks
   - Discriminant Analysis.

**Grading Scheme**

The assignment is marked out of 50.

**Answer**

Solutions and answers will be provided by the instructor.
Unit Readings and Other Resources

The readings in this unit are to be found at course level readings and other resources.

Module Summary

In modern computer science, software engineering, and other fields, the need arises to make decisions under uncertainty. Presenting probability and statistical methods, simulation techniques, and modeling tools, Probability and Statistics for applied Computer Science helps students solve problems and make optimal decisions in uncertain conditions, select stochastic models, compute probabilities and forecasts, and evaluate performance of computer systems and networks.

After introducing probability and distributions, this easy-to-follow module provides two course options. The first approach is a probability-oriented course that begins with stochastic processes, Markov chains, and queuing theory, followed by computer simulations and Monte Carlo methods. The second approach is a more standard, statistics-emphasized course that focuses on statistical inference, estimation, hypothesis testing, and regression. The Module is illustrated throughout with numerous examples, exercises, figures, and tables that stress direct applications in computer science and software engineering.

By the end of this course, advanced undergraduate and beginning graduate students should be able to read a word problem or a corporate report, realize the uncertainty involved in the described situation, select a suitable probability model, estimate and test its parameters based on real data, compute probabilities of interesting events and other vital characteristics, and make appropriate conclusions and forecasts.
Module Course Assessment

Identify the choice that best completes the statement or answers the question.

1. A random sample of 1000 people was taken. Four hundred fifty of the people in the sample favored Candidate A. The 95% confidence interval for the true proportion of people who favors Candidate A is
   a. 0.419 to 0.481
   b. 0.40 to 0.50
   c. 0.45 to 0.55
   d. 1.645 to 1.96

   In order to estimate the average time spent on the computer terminals per student at a local university, data were collected for a sample of 81 business students over a one-week period. Assume the population standard deviation is 1.8 hours.

   2. Refer to Exhibit 8-1. With a 0.95 probability, the margin of error is approximately
      a. 0.39
      b. 1.96
      c. 0.20
      d. 1.64

   6. The probability of committing a Type I error when the null hypothesis is true is
      a. the confidence level
      b. b
      c. greater than 1
      d. the Level of Significance

   7. The probability of making a Type I error is denoted by
      a. a
      b. b
      c. 1 - a
      d. 1 - b
8. For a one-tailed test (upper tail), a sample size of 26 at 90% confidence, \( t = \)
   a. 1.316  
   b. -1.316  
   c. -1.740  
   d. 1.740

9. For a one-tailed test (lower tail) with 22 degrees of freedom at 95% confidence, the value of \( t = \)
   a. -1.383  
   b. 1.383  
   c. -1.717  
   d. -1.721

Problem
A random sample of 49 lunch customers was taken at a restaurant. The average amount of time the customers in the sample stayed in the restaurant was 45 minutes with a standard deviation of 14 minutes.

a. Compute the standard error of the mean.

b. With a .99 probability, what statement can be made about the size of the margin of error?

c. Construct a 99% confidence interval for the true average amount of time customers spent in

The monthly incomes from a random sample of workers in a factory are shown below.

115
### Monthly Income (In $1,000)

<table>
<thead>
<tr>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
</tr>
<tr>
<td>7.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>6.0</td>
</tr>
<tr>
<td>6.0</td>
</tr>
<tr>
<td>7.0</td>
</tr>
<tr>
<td>9.0</td>
</tr>
</tbody>
</table>

a. Compute the standard error of the mean (in dollars).

b. Compute the margin of error (in dollars) at 95% confidence.

c. Compute a 95% confidence interval for the mean of the population. Assume the population has a normal distribution. Give your answer in dollars.

The proprietor of a boutique in New York wanted to determine the average age of his customers. A random sample of 53 customers revealed an average age of 28 years with a standard deviation of 4 years. Determine a 98% confidence interval estimate for the average age of all his customers.

A coal company wants to determine a 95% confidence interval estimate for the average daily tonnage of coal that they mine. Assuming that the company reports that the standard deviation of daily output is 200 tons, how many days should they sample so that the margin of error will be 39.2 tons or less?

The average score of a sample of 87 senior business majors at UTC who took the Graduate Management Admission Test was 510 with a standard deviation of 36. Provide a 98% confidence interval for the mean of the population.
In order to determine the average weight of carry-on luggage by passengers in airplanes, a sample of 25 pieces of carry-on luggage was collected and weighed. The average weight was 18 pounds. Assume that we know the standard deviation of the population to be 7.5 pounds.

a. Determine a 97% confidence interval estimate for the mean weight of the carry-on luggage.

b. Determine a 95% confidence interval estimate for the mean weight of the carry-on luggage.

In order to determine the average price of hotel rooms in Atlanta, a sample of 64 hotels was selected. It was determined that the average price of the rooms in the sample was $108.50 with a standard deviation of $16.

a. Formulate the hypotheses to determine whether or not the average room price is significantly different from $112.

b. Compute the test statistic.

c. At 95% confidence using the p-value approach, test the hypotheses. Let $a = 0.1$.

Course References

Optional readings and other resources:

- [http://mathworld.wolfram.com/Probability](http://mathworld.wolfram.com/Probability): Wolfram is a useful site that provides insights in number theory while providing new challenges and methodology in number theory.
- [http://en.wikipedia.org/wiki/Probability](http://en.wikipedia.org/wiki/Probability): Mathsguru is a website that helps learners to understand various branches of number theory module.

Required readings and other resources:

- [http://en.wikipedia.org/wiki/Probability](http://en.wikipedia.org/wiki/Probability): Mathsguru is a website that helps learners to understand various branches of number theory module. It is easy to access through Google search and provides very detailed information on various probability questions.
• http://en.wikipedia.org/wiki/Naive_Bayes_classifier
• http://archive.ics.uci.edu/ml/
The African Virtual University
Headquarters
Cape Office Park
Ring Road Kilimani
PO Box 25405-00603
Nairobi, Kenya
Tel: +254 20 2528333
contact@avu.org
oer@avu.org

The African Virtual University Regional Office in Dakar
Université Virtuelle Africaine
Bureau Régional de l’Afrique de l’Ouest
Sicap Liberté VI Extension
Villa No.8 VDN
B.P. 50609 Dakar, Sénégal
Tel: +221 338670324
bureauregional@avu.org