Pathways to Higher Education Project

Center for Advancement of Postgraduate
Studies and Research in Engineering Sciences,
Faculty of Engineering - Cairo University
(CAPSCU)

Cairo University

Statistical Data Analysis

Prof. Dr. Amany Mousa

Statistical Data Analysis

by

Prof. Dr. Amany Mousa
Professor of Statistics
Institute of Statistical Studies and Research,
Cairo University

Cairo
2005

Statistical Data Analysis

First Published 2005

Published by Center for Advancement of Postgraduate Studies and Research in Engineering Sciences, Faculty of Engineering - Cairo University (CAPSCU)
Tel: (+202) 5716620, (+202) 5678216
Fax: (+202) 5703620
Web-site: www.capscu.com
E-mail: capscu@tedata.net.eg

Deposit No. 2372/2005

ISBN 977-223-955-8

Acknowledgment

On behalf of Pathways to Higher Education Management Team in Egypt, the Project Coordinator wishes to extend his thanks and appreciation to the Ford Foundation (FF) for its full support to reform higher education, postgraduate studies and research activities in Egypt. The Management Team extend their special thanks and appreciation to Dr. Bassma Kodmani, Senior Project Officer at the Ford Foundation office in Cairo, who helped initiate this endeavor, and who spared no effort to support the Egyptian overall reform activities, particularly research and quality assurance of the higher education system. Her efforts were culminated by the endorsement to fund our proposal to establish the Egyptian Pathways to Higher Education project by the Ford Foundation Headquarters in New York.

The role of our main partner, the Future Generation Foundation (FGF), during the initial phase of implementation of the Pathways to Higher Education Project is also acknowledged. The elaborate system of training they used in offering their Basic Business Skills Acquisition (BBSA) program was inspiring in developing the advanced training program under Pathways umbrella. This partnership with an NGO reflected a truly successful model of coordination between CAPSCU and FGF, and its continuity is mandatory in support of our young graduates interested in pursuing research activities and/or finding better job opportunities.

The contribution of our partner, The National Council for Women (NCW), is appreciated. It is worth mentioning that the percentage of females graduated from Pathways programs has exceeded 50%, which is in line with FF and NCW general objectives. The second phase of the project will witness a much more forceful contribution from the NCW, particularly when implementing the program on the governorates level as proposed by CAPSCU in a second phase of the program.

We also appreciate the efforts and collaborative attitude of all colleagues from Cairo University, particularly the Faculties of Commerce, Art, Mass Communication, Law, Economics and Political Sciences, and Engineering who contributed to the success of this project.

Finally, thanks and appreciation are also extended to every member of the Center for Advancement of Postgraduate Studies and Research in Engineering Sciences (CAPSCU), Steering Committee members, trainers, supervisors and lecturers who were carefully selected to oversee the successful implementation of this project, as well as to all those who are contributing towards the accomplishment of the project objectives.

Pathways Steering Committee Members

SN	Member Name	Title	Institution
1	Dr. Ahmed Aboulwafa Mohamed	Professor and Chief of the Department of Public International Law, Faculty of Law and Ex-Vice Dean for Postgraduate Studies, Faculty of Law	CU
2	Dr. Ahmed Farghally	Professor of Accounting and Dean of the Faculty of Commerce	CU
3	Dr. Ali Abdel Rahman	President of Cairo University	CU
4	Dr. Bassma Kodmani	Senior Program Officer, Governance and International Cooperation, Ford Foundation, Cairo Office	FF
5	Dr. Fouad Khalaf	Ex-Project Manager, Project Consultant and Local Coordinator of TEMPUS Risk Project	CU
6	Dr. Hoda Rashad	Professor and Director of Social Research Center, American University in Cairo (AUC)	NCW
7	Dr. Kamel Ali Omran	Professor of Human Resources and Organizational Behavior, Business Administration and Ex-Vice Dean for Postgraduate Studies, Faculty of Commerce	CU
8	Dr. Mahmoud Fahmy El Kourdy	Professor of Social Science and Ex-Vice Dean for Students Affairs, Faculty of Arts	CU
9	Mr. Moataz El-Alfy	Vice Chairman of Future Generation Foundation	FGF
10	Mr. Mohamed Farouk Hafeez	Secretary General and Board Member, Future Generation Foundation	FGF
11	Dr. Mohamed K. Bedewy	Dean of the Faculty of Engineering and Chairman of CAPSCU Board	CAPSCU
12	Dr. Mohamed M. Megahed	Director of CAPSCU	CAPSCU
13	Dr. Mohsen Elmahdy Said	Project Coordinator	CU
14	Dr. Salwa Shaarawy Gomaa	Professor of Public Policy and Ex-Director of Public Administration Research & Consultation Center (PARC), Faculty of Economics Political Sciences	NCW & CU
15	Dr. Sami El Sherif	Vice Dean for Students Affairs, Faculty of Mass Communication	CU
16	Dr. Sayed Kaseb	Project Manager	CU
17	Dr. Zeinab Mahmoud Selim	Professor of Statistics and Ex-Vice Dean for Students Affairs, Faculty of Economics and Political Sciences	CU

CU	**Cairo University**	**NCW**	**National Council for Women**
FF	**Ford Foundation**	**FGF**	**Future Generation Foundation**

CAPSCU Center for Advancement of Postgraduate Studies and Research in Engineering Sciences, Faculty of Engineering - Cairo University

Publisher Introduction

The Faculty of Engineering, Cairo University is a pioneer in the field of learning and continual education and training. The Center for Advancement of Postgraduate Studies and Research in Engineering Sciences, Faculty of Engineering - Cairo University (CAPSCU) is one of the pillars of the scientific research centers in the Faculty of Engineering. CAPSCU was established in 1974 in cooperation with UNIDO and UNESCO organizations of the United Nations. Since 1984, CAPSCU has been operating as a self-financed independent business unit within the overall goals of Cairo University strategy to render its services toward development of society and environment.

CAPSCU provides consultation services for public and private sectors and governmental organizations. The center offers consultation on contractual basis in all engineering disciplines. The expertise of the Faculty professors who represent the pool of consultants to CAPSCU, is supported by the laboratories, computational facilities, library and internet services to assist in conducting technical studies, research and development work, industrial research, continuous education, on-the-job training, feasibility studies, assessment of technical and financial projects, etc.

Pathways to Higher Education (PHE) Project is an international grant that was contracted between Cairo University and Ford Foundation (FF). During ten years, FF plans to invest 280 million dollars to develop human resources in a number of developing countries across the world. In Egypt, the project aims at enhancing university graduates' skills. PHE project is managed by CAPSCU according to the agreement signed in September 22nd, 2002 between Cairo University and Ford Foundation, grant No. 1020 - 1920.

The partners of the project are Future Generation Foundation (FGF), National Council for Women (NCW) and Faculties of Humanities and Social Sciences at Cairo University. A steering committee that includes representatives of these organizations has been formed. Its main tasks are to steer the project, develop project policies and supervise the implementation process.

Following the steps of CAPSCU to spread science and knowledge in order to participate in society development, this training material is published to enrich the Egyptian libraries. The material composes of 20 subjects especially prepared and developed for PHE programs.

Dr. Mohammad M. Megahed
CAPSCU Director
April 2005

Foreword by the Project Management

Pathways to Higher Education, Egypt (PHE) aims at training fresh university graduates in order to enhance their research skills to upgrade their chances in winning national and international postgraduate scholarships as well as obtaining better job.

Pathways steering committee defined the basic skills needed to bridge the gap between capabilities of fresh university graduates and requirements of society and scientific research. These skills are: mental, communication, personal and social, and managerial and team work, in addition to complementary knowledge. Consequently, specialized professors were assigned to prepare and deliver training material aiming at developing the previous skills through three main training programs:

1. Enhancement of Research Skills
2. Training of Trainers
3. Development of Leadership Skills

The activities and training programs offered by the project are numerous. These activities include:

1. Developing training courses to improve graduates' skills
2. Holding general lectures for PHE trainees and the stakeholders
3. Conducting graduation projects towards the training programs

Believing in the importance of spreading science and knowledge, Pathways management team would like to introduce this edition of the training material. The material is thoroughly developed to meet the needs of trainees. There have been previous versions for these course materials; each version was evaluated by trainees, trainers and Project team. The development process of both style and content of the material is continuing while more courses are being prepared.

To further enhance the achievement of the project goals, it is planned to dedicate complete copies of PHE scientific publications to all the libraries of the Egyptian universities and project partners in order to participate in institutional capacity building. Moreover, the training materials will be available online on the PHE website, www.Pathways-Egypt.com.

In the coming phases, the partners and project management team plan to widen project scope to cover graduates of all Egyptian universities. It is also planned that underprivileged distinguished senior undergraduates will be included in the targeted trainees in order to enable their speedy participation in development of society.

Finally, we would like to thank the authors and colleagues who exerted enormous efforts and continuous work to publish this book. Special credit goes to Prof. Fouad Khalaf for playing a major role in the development phases and initiation of this project. We greatly appreciate the efforts of all members of the steering committee of the project.

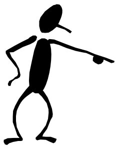
Dr. Sayed Kaseb — **Dr. Mohsen Elmahdy Said**

Project Manager — **Project Coordinator**

Table of Contents

Chapter 1: Introduction

Introduction



The collection processing, interpretation and presentation of numerical data all belong to the domain of statistics. These tasks include the calculation of football goals averages, collecting data on births and deaths, evaluating the effectiveness of commercial products, and forecasting the weather. Statistical information is presented to us constantly on radio and television. Our enthusiasm for statistical facts is encouraged by national newspapers such as the daily journals and magazines.

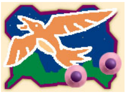
The word "statistics" is used in several ways. It can refer not only to the mere tabulation of numeric information, as in reports of stock market transactions, but also to the body of techniques used in processing or analyzing data.

The word "statistician" is also used in several ways. The term can be applied to those who simply collect information, as well as to those who prepare analyses or interpretations, and it is also applied to scholars who develop the mathematical theory on which statistics is based.

In Sections 1.1 and 1.2, we discuss the recent growth of statistics and its ever widening range of applications. In Section 1.3 we explain the distinction between the two major branches of statistics, descriptive statistical inference, and in the optional Section 1.4 we discuss the nature of various kinds of data and in connection with this warn the reader against the indiscriminate mathematical treatment of statistical data.

1.1 The Growth of Modern Statistics

The Growth of Modern Statistics



There are several reasons why the scope of statistics and the need to study statistics have grown enormously in the last fifty or so years. One reason is the increasingly quantitative approach employed in all the sciences as well as in business and many other activities which directly affect our lives. This includes the use of mathematical techniques in the evaluation of anti-pollution control, in inventory planning, in the analysis of traffic patterns, in the study of the effects of various kinds of medications, in the evaluation of teaching techniques, in the analysis of competitive behavior of businessmen and governments, in the study of diet and longevity, and so forth. The availability of powerful computers has greatly increased our ability to deal with numerical information. Many types of computers are also inexpensive, so that sophisticated statistical work can be done by small businessmen, college students, and even high-school students.

The other reason is that the amount of data that is collected, processed, and disseminated to the public for one reason or another has increased almost beyond comprehension, and what part is "good" statistics and what part is "bad" statistics is anybody's guess. To act as watchdogs, more persons with some knowledge of statistics are needed to take an active part in the collection of the data, in the analysis of the data, and what is equally important, in all of the preliminary planning. Without the latter, it is frightening to think of all the things that can go wrong in the compilation of statistical data. The results of costly surveys can be useless if questions are ambiguous or asked in the wrong way, if they are asked of the wrong persons, in the wrong place, or at the wrong time. Much of this is just common sense, as is illustrated by the following examples:

Example 1

Example
To determine public sentiment about the continuation of a certain government program an interviewer asks: "Do you feel that this wasteful program should be stopped?" Explain why this will probably not yield the desired information.

Solution 1

Solution:
The interviewer is "begging the question" by suggesting, in fact, that the program is wasteful.

Example 2

Example
To study consumer reaction to a new convenience food, a house to house survey is conducted during week day mornings, with no provisions for return visits in case no one is home. Explain why this may well yield misleading information.

Solution 2

Solution:
This survey will fail to reach those who are most likely to use the product: single persons and married couples with both spouses employed.

Although much of the above- mentioned growth of statistics began prior to the "computer revolution," the widespread availability and use of computers have greatly accelerated the process. In particular, computers enable us to handle, analyze and dissect large masses of data, and they enable us to perform calculations which previously had been too cumbersome even to contemplate. Our objective in these notes will be your gaining an understanding of the ideas of statistics. Access to a computer is not critical for this objective. Computer uses are occasionally illustrated in these notes, but nearly all the exercises can be done with nothing more than a four – function calculator.

The Study of Statistics

1.2 The Study of Statistics

The subject of statistics can be presented at various levels of mathematical difficulty, and it may be directed toward applications in various fields of inquiry. Accordingly, many textbook have been written on business statistics, educational statistics, medical statistics, psychological statistics … and even on statistics for historians. Although problems arising in these various disciplines will sometimes require special statistical techniques, none of the basic methods discussed in this text is restricted to any particular field of application. In the same way in which 2 + 2 = 4 regardless of whether we are adding dollar amounts, horses, or trees, the methods we shall present provide statistical models which apply regardless of whether the data are IQ's, tax payments, reaction times, humidity readings, test scores, and so on. To illustrate this further, consider the following situations.

1 : In a random sample of 200 retired persons, 137 stated that they prefer living in an apartment to living in a one - family home. At the 0.05 level of significance does this refute the claim that 60 percent of all retired persons prefer living in an apartment to living in a one – family home?

The question asked here should be clear, and it should also be clear that the answer would be of interest mainly to social scientists or to persons in the construction industry. However, if we wanted to cater to the special interests of students of biology, engineering education or ecology, we might rephrase the situation as follows:

2 : In a random sample of 200 citrus trees exposed to a 20^o frost, 137 showed some damage to their fruit. At the 0.05 level of significance does this refute the claim that 60 percent of citrus trees exposed to a 20^o frost will show some damage to their fruit?

3 : In a random sample of 200 transistors made by a given manufacturer, 137 passed an accelerate performance test. At the 0.05 level of significance does this refute the claim that 60 percent of all transistors made by a given manufacturer will pass the test?

4 : In a random sample of 200 high school seniors in a large city, 137 said that they will go on to college. At the 0.05 level of significance does this refute the claim that 60 percent of all high school seniors in a large city will go to college?

5 : In a random sample of 200 cars tested for the emission of pollutants, 137 failed to meet a state's legal standards. At the 0.05 level of significance does this refute the claim that 60 percent of all cars tested in this state will fail to meet legal emission standards?

So far as the work in these notes is concerned, the statistical treatment of all these versions is the same, and with some imagination the reader should be able to rephrase it for virtually any field of specialization. As some authors do, we could present, and so designate, special problems for readers with special interests, but this would defeat our goal of impressing upon the reader the importance of statistics in all of science, business, and everyday life. To attain this goal we have included in this text exercises covering a wide spectrum of interests.

To avoid the possibility of misleading anyone with our various versions, let us make it clear that we cannot squeeze all statistical problems into the same mold. Although the methods we shall study in these notes are all widely applicable, it is always important to make sure that the statistical model we are using is the right one.

Descriptive Statistics and Statistical Inference

1.3 Descriptive Statistics and Statistical Inference

The origin of modern statistics can be traced to two areas of interest which, on the surface have very little in common: government (political science), and games of chance.

Governments have long used census data to count persons and property, and the problem of describing, summarizing and analyzing census data has led to the development of methods which, until recently, constituted about all there was to the subject of statistics. **These methods, which at first consisted primarily of presenting data in the form of tables and charts, make up what we now call descriptive statistics.** This includes anything done to data which is designed to summarize or describe, without going any further; that is, without attempting to infer anything that goes beyond the data, themselves. For instance, if tests performed on six small cars imported in 1986 showed that they were able to accelerate from 0 to 60 miles per hour in 18.7, 19.2, 16.2, 12.3, 17.5, and 13.9 seconds, and we report that half of them accelerated from 0 to 60 mph in less than 17.0 seconds, our work belongs to the domain of descriptive statistics. This would also be the case if we claim that these six cars averaged

$$\frac{18.7+19.2+16.2+12.3+17.5+13.9}{6}=16.3\,\text{seconds},$$

But is not if we conclude that half of all cars imported that year could accelerate from 0 to 60 mph in less than 17.0 seconds.

Although descriptive statistics is an important branch of statistics and it continues to be widely used, statistics information usually arises from samples (from observations made on only part of a large set of items), and this means that its analysis requires generalizations which go beyond the data. **As a result, the most important feature of the recent growth of statistics has been a shift in emphasis from methods which merely describe to methods which serve to make generalizations; that is, a shift in emphasis from descriptive statistics to the methods of statistical inference.**

Such methods are required, for instance, to predict the operating life span of a hand – held calculator (on the basis of the performance of several such calculators);to estimate the 1995 assessed value of all privately owned property in Cairo (on the basis of business trends, population projections, and so for the); to compare the effectiveness of two reducing diets (on the basis of the weight losses of persons who have been on the diets); to determine the most effective dose of a new medication (on the basis of tests performed with volunteer patients from selected hospitals); or to predict the flow of traffic on a freeway which has not yet been built (on the basis of past traffic counts on alternative routes).

In each of the situations described in the preceding paragraph, there are uncertainties because there is only partial, incomplete, or indirect information; therefore, the methods of statistical inference are needed to judge the merits of our results, to choose a "most promising" prediction, to select a "most reasonable" (perhaps, a "potentially most profitable") course of action.

In view of the uncertainties, we handle problems like these statistical methods which find their origin in games of chance. Although the mathematical study of games of chance dates to the seventeenth century, it was not until the early part of the nineteenth century that the theory developed for "heads or tails," for example, or "red or black" or even or odd," was applied also to real-life situations where the outcomes were "boy or girl," "life or death," "pass or fail," and so forth. Thus, *probability theory was applied to many problems in the behavioral, natural, and social sciences, and nowadays it provides an important tool for the analysis of any situation (in science, in business, or in everyday life) which in some way involves an element of uncertainty of chance. In particular, it provides the basis for the methods which we use when we generalize from observed data, namely, when we use the methods of statistical inference.*

In recent years, it has been suggested that the emphasis has swung too far from descriptive statistics to statistical inference, and that more attention should be paid to the treatment of problems requiring *only descriptive techniques.* To accommodate these needs, some new descriptive methods have recently been developed under the general heading of exploratory data analysis. Two of these will be presented.

The Nature of Statistical Data

1.4 The Nature of Statistical Data

Statistical data are the raw material of statistical investigations – they arise whenever measurements are made or observations are recorded. *They may be weights of animals, measurements of personality traits, or earthquake intensities, and they may be simple "yes or no" answer or descriptions of persons' marital status as single, married, widowed, or divorced.* Since we said that statistics deals with numerical data, this requires some explanation, because "yes or no" answers and descriptions of marital status would hardly seem to qualify as being numerical. Observe, however, that we can record "yes or no" answers to a question as 0 (or as 1 and 2, or perhaps as 29 and 30 if we are referring to the 15th "yes or no" question of a test), and that we can record a person's marital status as 1, 2, 3, or 4, depending on whether the person is single, married, widowed, or divorced. In this artificial or nominal way, categorical (qualitative or descriptive) data can be made into numerical data, and if we thus code the various categories, we refer to the numbers we record as nominal data.

Nominal data are numerical in name only, because they do not share any of the properties of the numbers we deal with in ordinary arithmetic. For instance, if we record marital status as 1, 2, 3, or 4, as suggested above, we cannot write $3 > 1$ or 2,4, and we cannot write $2 - 1 = 4 - 3$, $1 + 3 = 4$, or $4 \div 2 = 2$. It is important, therefore, always to check whether mathematical calculations performed in a statistical analysis are really legitimate.

Let us now consider some examples where data share some, but not necessarily all, of the properties of the numbers we deal with in ordinary arithmetic. For instance, in mineralogy the hardness of solids is sometimes determined by observing "what scratches what." If one mineral can scratch another it receives a higher hardness number, and on Mohs' scale the numbers form 1 to 10 are assigned, respectively, to talc, gypsum, calcite, fluorite apatite, feldspar, quartz, topaz, sapphire, and diamond. With these numbers, we can write $6 > 3$, for Example or $7 < 9$, since feldspar is harder than calcite and quartz is softer than sapphire. On the other hand, we cannot write $10 - 9 = 2 -1$, for Example because the difference in hardness between diamond and sapphire is actually much greater than that between gypsum and talc. Also, it would be meaningless to say that topaz is twice as hard as fluorite simply because their respective hardness numbers on Mohs' scale are 8 and 4.

If we cannot except set up inequalities as was the case in the proceeding example, **we refer to the data as ordinal data.** In connection with ordinal data > does not necessarily mean "greater than" it may be used to denote "happier than," "preferred to," "more difficult than," "tastier than," and so forth.

If we can also form differences, but not multiply or divide, we refer to the data as interval data. To give an example, suppose we are given the following temperature readings Fahrenheit "63^{o}, 67^{o}, 91^{o}, 107^{o}, 126^{o}, and 131^{o}." Here, we can write $107\ ^{o}$ is warmer than $68\ ^{o}$ and that 91^{o} is colder than 131^{o}. Also, we can write $68\ ^{o}$-63^{o} = 131^{o} - 126^{o}, since equal temperature differences are equal in the sense that the same amount of heat is required to raise the temperature of an object from 63^{o} to $68\ ^{o}$ as from 126^{o} to 131^{o}. On the other hand, it would not mean much if we say that $126\ ^{o}$ is twice as hot as 63^{o}, even though 126 ÷ 63 = 2. To show why, we have only to change to the Celsius scale, where the first temperature becomes $\frac{5}{9}(126-32)=52.2^{0}$, the second temperature becomes $\frac{5}{9}(63-32)=17.2^{0}$, and the first figure is now more than three times the second. This difficulty arises because the Fahrenheit and Celsius scales both have artificial origins (zeros); in other words, the number 0 of neither scale is indicative of the absence of whatever quantity we are typing to measure.

If we can also form quotients, we refer to the data as ratio data, and such data are not difficult to find. They include all the usual measurements (or determinations) of length, height, money amounts, weight, volume, area, pressure, elapsed time (though not calendar time), sound intensity, density, brightness, velocity, and so on.

The distinction we have made here between nominal, ordinal, interval, and ratio data is important, for as we shall see, the nature of a set of data may suggest the use of particular statistical techniques. To emphasize the point that what we can and cannot do arithmetically with a given set of data depends on the nature of the data, consider the following scores which four students obtained in the three parts of a comprehensive history test.

Students	Tests			
	A	B	C	Total
L	89	51	40	180
T	61	56	54	171
H	40	70	55	183
R	13	77	72	162

The totals for the four students are 180, 171, 165, and 162, so that L scored highest, followed by T, H, and R.

Suppose now that somebody proposes that we compare the overall performance of the four students by ranking their scores from high to low for each part of the test, and then average their ranks. What we get is shown in the following table:

	A	B	C	Average Rank
L	1	4	4	3
T	2	3	3	$2\frac{2}{3}$
H	3	2	2	$2\frac{1}{3}$
R	4	1	1	2

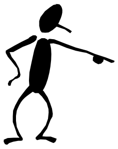
Here L's average rank was calculated as $\frac{1+4+4}{3}=\frac{9}{3}=3$, T's as $\frac{2+3+3}{3}=\frac{8}{3}=2\frac{2}{3}$, and so forth.

Now, if we look at the average ranks, we find that R came out best, followed by H, T, and L, so that the order has been reversed from what it was before. How can this be? Well, strange things can happen when we average ranks. For instance, when it comes to their ranks, L's outscoring T by 28 points in counts just as much as T's outscoring her by 5 points in B, and T's outscoring H by 21 points in A counts just as much as H's outscoring him by a single point in C. We conclude that, perhaps, we should not have averaged the ranks, but it might also be pointed out that, perhaps, we should not even have totaled the original scores. The variation of A scores, which go from 13 to 89, is much greater than that of other two kinds of scores, and this strongly affects the total scores and suggests a possible shortcoming of the procedure. We shall not go into this here, as it has been our *goal merely to alert the reader against the indiscriminate use of statistical techniques.*

Chapter 2: Data Collection and Sampling

Introduction

2.1 Introduction



In Chapter 1, we briefly introduced the concept of statistical inference-the process of inferring information about a population from a sample. Because information about populations can usually be described by parameters, the statistical technique used generally deals with drawing inferences about population parameters from sample statistics. (Recall that a parameter is a measurement about a population, and a statistic is a measurement about a sample.)

Working within the covers of a statistics textbook, we can assume that population parameters are known. In real life, however, calculating parameters become prohibitive because populations tend to be quite large. As a result, most population parameters are unknown. For example, in order to determine the mean annual income of blue-collar workers, we would have to ask each blue-collar worker what his or her income is and then calculate the mean of all the responses. Because this population consists of several million people, the task is both expensive and impractical. If we are willing to accept less than 100% accuracy, we can use statistical inference to obtain an estimate.

Rather than investigating the entire population, we select a sample of workers, determine the annual income of the workers in this group, and calculate the sample mean. While there is very little chance that the sample mean and the population mean are identical, we would expect them to be quite close. However, for the purposes of statistical inference, we need to be able to measure how close the sample mean is likely to be to the population mean. In this chapter, however, we will discuss the basic concepts and techniques of sampling itself. But first we will take a look at various sources for collecting data.

Sources of Data

2.2 Sources of Data

The validity of the results of a statistical analysis clearly depends on the reliability and accuracy of the data used. Whether you are actually involved in collecting the data; performing a statistical analysis on the data, or simply reviewing the results of such an analysis, it is important to realize that the reliability and accuracy of the data depend on the method of collection. **Three of the most popular sources of statistical data are published data, data**

collected from observational studies, and data collected from experimental studies.

Published Data

2.2.1 Published Data

The use of published data is often preferred due to its convenience, relatively low cost, and reliability (assuming that it has been collected by a reputable organization). There is an enormous amount of published data produced by government agencies and private organizations, available in printed form, on data tapes and disks, and increasingly on the Internet. **Data published by the same organization that collected them are called primary data.** *An Example of primary data would be the data published by Egypt Bureau of the Census, which collects data on numerous industries as well as conducts the census of the population every ten years.* Statistics agency is the central statistical agency, collecting data on almost every aspect of social and economic life in the country. These primary sources of information are invaluable to decision makers in both the government and private sectors.

Secondary data refers to data that are published by an organization different from the one that originally collected and published the data. *A popular source of secondary data is The Statistical Book of Egypt, which compiles data from several primary government sources and is updated annually. Another Example of a secondary data source is Central Bank which has a variety of financial data tapes that contain data compiled from such primary sources as the Stock Exchange.* Care should be taken when using secondary data, as errors may have been introduced as a result of the transcription or due to misinterpretation of the original terminology and definitions employed.

An interesting Example of the importance of knowing how data collection agencies define their terms appeared in an article in *The Globe and Mail* (February 12, 1996). The United States and Canada had similar unemployment rates up until the 1980s, at which time Canada's rate started to edge higher than the U.S. rate. By February 1996, the gap had grown to almost four percentage points (9.6% in Canada compared with 5.8% in the United States). Economists from the United States and Canada met for two days to compare research results and discuss possible reasons for this puzzling gap in jobless rates. The conference organizer explained that solving this mystery matters because "we have to understand the nature of unemployment to design policies to combat it." An Ohio State University economist was the first to notice a difference in how officials from the two countries define unemployment. "If jobless people say they are searching for work, but do nothing more than read job advertisements in the newspaper, Canada counts them as unemployed. U.S. officials dismiss such 'passive' job hunters and count them as being out of the labor force altogether, so they are not counted among the jobless." Statistics Canada reported that this difference in definitions accounted for almost one-fifth of the difference between the Canadian and U.S. unemployment rates.

Observational and Experimental Studies

2.2.2 Observational and Experimental Studies

If relevant data are not available from published sources, it may be necessary to generate the data by conducting a study. This will especially be the case when data are needed concerning a specific company or situation. The difference between two important types of studies – observational and experimental – is best illustrated by means of an example.

Example

Example (1)

Six months ago, the director of human resources for a large mutual fund company announced that the company had arranged for its salespeople to use a nearby fitness center free of charge. The director believes that fitter salespeople have more energy and an improved appearance, resulting in higher productivity. Interest and participation in the fitness initiative were high initially, but after a few months had passed, several employees stopped participating. Those who continued to exercise were committed to maintaining a good level of fitness, using the fitness center about three times per week on average.

The director recently conducted an observational study and determined that the average sales level achieved by those who regularly used the fitness center exceeded that of those who did not use the center. The director was tempted to use the difference in productivity levels to justify the cost to the company of making the fitness center available to employees. But the vice-president of finance pointed out that the fitness initiative was not necessarily the *cause* of the difference in productivity levels. Because the salespeople who exercised were self-selected - they determined themselves whether or not to make use of the fitness center-it is quite likely that the salespeople who used the center were those who were more ambitious and disciplined. These people would probably have had higher levels of fitness and productivity even without the fitness initiative. We therefore cannot necessarily conclude that fitness center usage led to higher productivity. It may be that other factors, such as ambition and discipline, were responsible both for higher fitness center usage and higher productivity.

The director and vice-president then discussed the possibility of conducting **an experimental study**, designed to control which salespeople made regular use of the fitness center. The director would randomly select 60 salespeople to participate in the study. Thirty of these would be randomly selected and persuaded to use the fitness center on a regular basis for six months. The other 30 salespeople selected would not be approached, but simply would have their sales performances monitored along with those using the fitness center regularly. Because these two groups were selected at random, we would expect them to be fairly similar in terms of original average fitness level, ambition, discipline, age, and other factors that might affect performance. From this experimental study, we would be more confident that any significantly higher level of productivity by the group using the fitness center regularly would be due to the fitness initiative rather than other factors.

The point of the preceding Example is to illustrate the difference between observational study and an experimental (or controlled) study. In the observational study, a survey simply was conducted to observe and record the average level for each group, without attempting to control any of the factors that might influence the sales levels. In the experimental study, the director controlled one factor (regular use of the fitness center) by *randomly* selecting who would be persuaded to use the center regularly, thereby reducing the influence of other factors on the difference between the sales levels of the two groups.

Although experimental studies make it easier to establish a cause–and–effect relationship between two variables, observational studies are used predominately in business and economics. *More often than not, surveys are conducted to collect business and economic data (such as consumer preferences or unemployment statistics), with no attempt to control any factors that might affect the variable of interest.*

Surveys

1.Public Surveys
2. Private Surveys

Surveys

One of the most familiar methods of collecting primary data is the survey, which solicits information from people concerning such things as their income, family size, and opinions on various issues. We're all familiar, for example, with opinion that accompany each political election. The Gallup poll and the Harris survey *are* two well-known **surveys of public opinion** whose results are often reported in the media. **But the majority of surveys are conducted for private use. Private surveys are used extensively by market researchers to determine the preferences and attitudes of consumers and voters.** The results can be used for a variety of purposes, from helping to determine the target market for an advertising campaign to modifying a candidate's platform in an election campaign. As an illustration, consider a television network that has hired a market research firm to provide the network with a profile of owners of luxury automobiles, including what they watch on television and at what times. The network could then use this information to develop a package of recommended time slots for Cadillac commercials including costs; that it would present to General Motors. It is quite likely that many, students reading this notes will one day be marketing executives who will "live and die" by such market research data.

Many researchers feel that the best way to survey people is by means of a personal interview, which involves an interviewer soliciting information from respondent by asking prepared questions. ***A personal interview*** *has the advantage of having a higher expected response rate than other methods of data collection. In addition, there will probably be fewer incorrect responses resulting from respondents misunderstanding some questions, because the interviewer can clarify misunderstandings when asked to. But the interviewer must also be careful not to say too much, for fear of biasing the response. To avoid introducing such biases, as well as to reap the potential benefits of a personal interview, interviewer must be well trained in proper interviewing techniques and well informed on the purpose of the study. The main disadvantage of personal interviews is that they are expensive, especially when travel is involved.* ***A telephone interview*** *is usually less expensive, but it is also less personal and has a lower expected response rate.*

A third popular method of data collection is the ***self - administered questionnaire****, which is usually mailed to a sample of people selected to be surveyed. This is a relatively inexpensive method of conducting a survey and is therefore attractive when the number of people to be surveyed is large. But self-administered questionnaires usually have a low response and may have a relatively high number of incorrect responses due to respondents misunderstanding some questions.*

Whether a questionnaire is self – administered or completed by an interviewer, it must be well designed. Proper questionnaire design takes knowledge, experience, time, and money. **Some basic points to consider regarding questionnaire design follow.**

1- First and foremost, **the questionnaire should be kept as short as possible** to encourage respondents to complete it. Most people are unwilling to spend much time filling out a questionnaire.

2- **The questions themselves should also be short, as well as simply and clearly worded,** to enable respondents to answer quickly, correctly, and without ambiguity. Even familiar terms, such as "unemployed" and "family," must be defined carefully because several interpretations are possible.

3- **Questionnaires often begin with simple demographic questions** to help respondents get started and become comfortable quickly.

4- **Dichotomous questions** (questions with only two possible responses, such as "yes" and "no') **and multiple – choice questions** are useful and popular because of their simplicity, but they, too, have possible shortcomings. For example, a respondent's choice of yes or no to a question may depend on certain assumptions not stated in the question. In the case of a multiple- choice question, a respondent may feel that none of the choices offered is suitable.

5- **Open-ended questions** provide an opportunity for respondents to express opinions more fully, but they are time-consuming and more difficult to tabulate and analyze.

6- **Avoid using leading questions,** such as "Wouldn't you agree that the statistics exam was too difficult?" These types of questions tend to lead the respondent to a particular answer.

7- **Time permitting,** it is useful to pretest a questionnaire on a small number of people in order to uncover potential problems, such as ambiguous wording.

8- Finally, **when preparing the questions, think about how you intend to tabulate and analyze the responses.** *First determine whether you are soliciting values (i.e., responses) for a qualitative variable or a quantitative variable. Then consider which type of statistical techniques – descriptive or inferential – you intend to apply to the data to be collected, and note the requirements of the specific techniques to be used.* Thinking about these questions will help to assure that the questionnaire is designed to collect the data you need.

Whatever method is used to collect primary data, we need to know something about sampling, the subject of the next section.

Sampling

2.3 Sampling

The chief motive for examining a sample rather than a population is cost. *Statistical inference permits us to draw conclusions about a population parameter based on a sample that is quite small in comparison to the size of the population.* For example, television executives want to know the proportion of television viewers who watch a network's programs. Because 100 million people may be watching television in the world on a given evening, determining the actual proportion of the population that is watching certain programs is impractical and prohibitively expensive. The ratings provide approximations of the desired information by observing what is watched by a sample of 1,000 television viewers. The proportion of households watching a particular program can be calculated for the households in the sample. This sample proportion is then used as an estimate of the proportion of all households (the population proportion) that watched the program.

Another illustration of sampling can be taken from the field of quality control. In order to ensure that a production process is operating properly, the operations manager needs to know what proportion of items being produced is defective. If the quality - control technician must destroy the item in order to determine whether it is defective, then there is no alternative to sampling: a complete inspection of the product population would destroy the entire output of the production process.

We know that the sample proportion of television viewers or of defective items is probably not exactly equal to the population proportion we want to estimate. Nonetheless, the sample statistic can come quite close to the parameter it is designed to estimate if the target population (the population about which we want to draw inferences) and the sampled population (the actual population from which the sample has been taken) are the same. In practice, these may not be the same.

Sampling Plans

2.4 Sampling Plans

Our objective in this section is to introduce three different sampling plans: simple random sampling, stratified random sampling, and cluster sampling. We begin our presentation with the most basic design.

Simple Random Sampling

2.4.1 Simple Random Sampling

Simple Random Sample

A simple random sample is a sample selected in such a way that every possible sample with the same number of observations is equally likely to be chosen.

One way to conduct a simple random sample is to assign a number to each element in the population, write these numbers on individual slips of paper, toss them into a hat, and draw the required number of slips (the sample size, *n)* from the hat. This is the kind of procedure that occurs in raffles, when all the ticket stubs go into a large, rotating drum from which the winners are selected.

Sometimes the elements of the population are already numbered. For example, virtually all adults have Social Security numbers or Social Insurance numbers; all employees of large corporations have employee numbers; many people have driver's license numbers, medical plan numbers, student numbers, and so on. In such cases, choosing which sampling procedure to use is simply a matter of deciding how to select from among these numbers.

In other cases, *the existing form of numbering has built-in flaws that make it inappropriate as a source of samples.* Not everyone has a phone number, for example, so the telephone book does not list all the people in a given area. Many households have two (or more) adults, but only one phone listing. Couples often list the phone number under the man's name, so telephone listings are likely to be disproportionately male. Some people do not have phones, some have unlisted phone numbers, and some have more than one phone; these differences mean that each element of the population does not have an equal probability of being selected.

After each element of the chosen population has been assigned a unique number, sample numbers can be selected at random. A random - number table can be used to select these sample numbers. Alternatively, we can employ a software package to generate random numbers. Both Minitab and Excel have this capability.

Example

Example (2)

A government income-tax auditor has been given responsibility for 1,000 returns. A computer is used to check the arithmetic of each return. However, to determine if the returns have been completed honestly, the auditor must check each entry and confirm its veracity. Because it takes, on average, one hour to completely audit a return and she has only one week to complete the task, the auditor has decided to randomly select 40 returns. The returns are numbered from 1 to 1,000. Use a computer random -number generator to select the sample for the auditor.

Solution

Solution:

There are several software packages that can produce the random numbers we need. Minitab and Excel are two of these.

Minitab Output for Example (2)

173	184	953	896	82	388	232	962	391	95
259	544	588	754	870	700	893	690	320	28
312	183	271	587	922	759	929	526	112	43
811	480	984	991	100	367	655	877	59	642
654	859	478	633	157	470	615	32	258	887

We generated 50 numbers between 1 and 1,000 and stored them in column 1. Although we needed only 40 random numbers, we generated 50 numbers because it is likely that some of them will be duplicated. We will use the first 40 unique random number to select our sample.

Excel Output for Example (2)

165	78	120	987	705	827	725	466	759	361
504	545	578	820	147	276	237	764	85	528
160	357	44	971	269	517	711	721	192	926
832	661	426	173	909	973	856	813	152	915
544	622	830	382	198	830	700	256	210	621

We generated 50 numbers between 1 and 1,000 and stored them in column 1. Although we needed only 40 random numbers, we generated 50 numbers because it is likely that some of them will be duplicated. We will use the first 40 unique random number to select our sample.

Stratified Random Sampling

Stratified Random Sampling

In making inferences about a population, we attempt to extract as much information as possible from a sample. The basic sampling plan, simple random sampling, often accomplishes this goal at low cost. Other methods, however, can be used to increase the amount of information about the population. One such procedure is stratified random sampling.

Stratified Random Sample

Stratified Random Sample

A stratified random sample is obtained by separating the population into mutually exclusive sets, or strata, and then drawing simple random sample from each stratum.

Examples of criteria for separating a population into strata (and of the strata themselves) follow.

1 - Sex
  Male
  Female

2 - Age
  Under 20
  20-30
  31-40
  41-50
  51-60
  Over 60

3 - Occupation
  Professional
  Clerical
  Blue-collar other

4 - Household income
  Under $15,000 $15,000-$29,999
  $30,000-$50,000
  Over $50,000

To illustrate, suppose a public opinion survey is to be conducted in order to determine how many people favor a tax increase. A stratified random sample could be obtained by selecting a random sample of people from each of the four income groups described above. We usually stratify in a way that enables us to obtain particular kinds of information. In this example, we would like to know if people in the different income categories differ in their opinions about the proposed tax increase, since the tax increase will affect the strata differently. We avoid stratifying when there is no connection between the survey and the strata. For example, little purpose is served in trying to determine if people within religious strata have divergent opinions about the tax increase.

One advantage of stratification is that, besides acquiring information about the entire population, we can also make inferences within each stratum or compare strata. For instance, we can estimate what proportion of the lowest income group favors the tax increase, or we can compare the highest and lowest income groups to determine if they differ in their support of the tax increase.

Any stratification must be done in such a way that the strata are mutually exclusive: each member of the population must be assigned to exactly one stratum.

After the population has been stratified in this way, we can employ simple random sampling to generate the complete sample. There are several ways to do this. For example, we can draw random samples from each of the four income groups according to their proportions in the population. Thus, if in the population the relative frequencies of the four groups are as listed below, our sample will be stratified in the same proportions. If a total sample of 1,000 is to be drawn, we will randomly select 250 from stratum 1,400 from stratum 2,300 from stratum 3, and 50 from stratum 4.

Stratum	Income Categories	Population Proportions
1	under $15,000	25%
2	15,000-29,999	40
3	30,000-50,000	30
4	over 50,000	5

The problem with this approach, however, is that if we want to make inferences about the last stratum, a sample of 50 may be too small to produce useful information. In such cases, we usually increase the sample size of the smallest stratum (or strata) to ensure that the sample data provide enough information for our purposes. An adjustment must then be made before we attempt to draw inferences about the entire population. This procedure is beyond the level of these notes. We recommend that anyone planning such a survey consult an expert statistician or a reference book on the subject. Better still, become an expert statistician yourself by taking additional statistics courses.

Cluster Sampling

2.4.2 Cluster Sampling

Cluster Sample

Cluster Sample

A cluster sample is a simple random sample of groups or clusters of elements. *Cluster sampling is particularly useful when it is difficult or costly to develop a complete list of the population members (making it difficult and costly to generate a simple random sample). It is also useful whenever the population elements are widely dispersed geographically.* For example, suppose we wanted to estimate the average annual household income in a large city. To use simple random sampling, we would need a complete list of households in the city from which to sample. To use stratified random sampling, we would need the list of households, and we would also need to have each household categorized by some other variable (such as age of household head) in order to develop the strata. A less expensive alternative would be to let each block within the city represent a cluster. A sample of clusters could then be randomly selected, and every household within these clusters could be questioned to determine income. By reducing the distances the surveyor must cover to gather data, cluster sampling reduces the cost.

But cluster sampling also increases sampling error, because households belonging to the same cluster are likely to be similar in many respects, including household income. This can be partially offset by using some of the cost savings to choose a larger sample than would be used for a simple random sample.

Sample Size

Sample Size

Whichever type of sampling plan you select, you still have to decide what size of sample to use. In determining the appropriate sample size, we can rely on our intuition, which tells us that the larger the sample size is, the more accurate we can expect the sample estimates to be.

Errors Involved in Sampling

2.5 Errors Involved in Sampling

Two major types of errors can arise when a sample of observations is taken from a population: **sampling error** and **nonsampling error**. Managers reviewing the results of sample surveys and studies, as well as researchers, who conduct the surveys and studies, should understand the sources of these errors.

Sampling Error

Sampling Error

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample. **Sampling error is an error that we expect to occur when we make a statement about a population that is based only on the observations contained in a sample**

taken from the population. To illustrate, consider again the Example described in which we wish to determine the mean annual income of blue-collar workers. As was stated there, we can use statistical inference to estimate the mean income (μ) of the population if we are willing to accept less than 100% accuracy. If we record the incomes of a sample of the workers and find the mean $(\overline{X})$ of this sample of incomes, this sample mean is an estimate of the desired population mean. But the value of $\overline{X}$ will deviate from the population mean (μ) simply by chance, because the value of the sample mean depends on which incomes just happened to be selected for the sample. The difference between the true (unknown) value of the population mean (μ) and its sample estimate $\overline{X}$ is the sampling error. The size of this deviation may be large simply due to bad luck that a particularly unrepresentative sample happened to be selected. The only way we can reduce the expected size of this error is to take a larger sample.

Given a fixed sample size, the best we can do is to state the probability that the sampling error is less than a certain amount. It is common today for such a statement to accompany the results of an opinion poll. If an opinion poll states that, based on sample results, Candidate Kreem has the support of 54% of eligible voters in an upcoming election, that statement may be accompanied by the following explanatory note: This percentage is correct to within percentage points, 29 times out of 30. This statement means that we have a certain level of confidence (95%) that the actual level of support for Candidate Kreem is between 51 % and 57%.

Nonsampling Error

Nonsampling Error

Nonsampling error is more serious than sampling error, because taking a larger sample won't diminish the size, or the possibility of occurrence, of this error. Even a census can (and probably will) contain nonsampling errors. Nonsampling errors are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.

Three Types of Nonsampling Errors

Three Types of Nonsampling Errors

1- Errors in Data Acquisition

1- Errors in Data Acquisition. *These types of errors arise from the recording of incorrect responses.* This may be the result of incorrect measurements being taken because of faulty equipment, mistakes made during transcription from primary sources, inaccurate recording of data due to misinterpretation of terms, or inaccurate responses to questions concerning sensitive issues such as sexual activity or possible tax evasion.

2- Nonresponse Error

2- Nonresponse Error*. Nonresponse error refers to error (or bias) introduced when responses are not obtained from some members of the sample.* When this happens, the sample observations that are collected may not be representative of the target population, resulting in biased results. Nonresponse can occur for a number of reasons. An interviewer may be unable to contact a person listed in the sample, or the sampled person may refuse to respond for some reason. In either case, responses are not obtained from a sampled person, and bias is introduced. The problem of nonresponse is even greater when self-administered questionnaires are used rather than an interviewer, who can attempt to reduce the nonresponse rate by means of callbacks. As noted earlier, a high nonresponse rate, resulting in a biased, self-selected sample.

3- Selection Bias

***3- Selection Bias**. Selection bias occurs when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample.* Together with nonreponse error, selection biases played a role in pool being so wrong, as voters without telephones or without a subscription were excluded from possible inclusion in the sample taken.

Summary

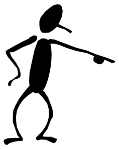
2.6 Summary

Because most populations are very large, it is extremely costly and impractical to investigate each member of the population to determine the values of the parameters. As a practical alternative, we take a sample from the population and use the sample statistics to draw inferences about the parameters. Care must be taken to ensure that the sampled population is the same as the target population.

We can choose from among several different sampling plans, including simple random sampling, stratified random sampling, and cluster sampling. Whatever sampling plans used, it is important to realize that both sampling error and nonsampling error will occur, and to understand what the sources of these errors are.

Chapter 3: Summarizing Data Listing and Grouping

Introduction



In recent years the collection of statistical data has grown at such a rate that it would be impossible to keep up with even a small part of the things that directly affect our lives unless this information is disseminated in "predigested" or summarized form. The whole matter of putting large masses of data into a usable form has always been important, but it has multiplied greatly in the last few decades. This has been due partly to the development of computers, which was previously left undone because it would have taken months or years, and partly to the deluge of data generated by the increasingly quantitative approach of the sciences, especially the behavioral and social sciences, where nearly every aspect of human life is nowadays measured in one way or another.

The most common method of summarizing data is to present them in condensed form in tables or charts, and at one time this took up the better part of an elementary course in statistics. Nowadays, there is so much else to learn in statistics that very little time is devoted to this kind of work. In a way this is unfortunate, because one does not have to look far in newspapers, magazines, an even professional journal to find unintentionally or intentionally misleading statistical charts.

In Sections 3.1 and 3.2 we shall present ways of listing data so that they present a good overall picture and, hence, are easy to use. By listing we are referring to any kind of treatment that preserves the identity of each value (or item). In other words, we rearrange but do not change. A speed of 63 mph remains a speed of 63 mph, a salary of $75,00 and when sampling public opinion, a National Party remains a National and a Wafdy remains a Wafdy. In Sections 3.3 and 3.4, we shall discuss ways of grouping data into a number of classes, intervals, or categories and presenting the result in the form of a table or a chart. This will leave us with data in a relatively compact and easy-to-use form, but it does entail a substantial loss of information. Instead of a person's weight, we may know only that he or she weights anywhere from 160 to 169 pounds, and instead of an actual pollen count we may know only that it is medium (11-25 parts per cubic meter).

Listing Numerical Data

3.1 Listing Numerical Data

Listing and thus, organizing the data is usually the first task in any kind of statistical analysis. As a typical situation, consider the following data, representing the lengths (in centimeters) of 60 sea trout caught by a commercial trawler in Bay Area :

19.2	19.6	17.3	19.3	19.5	20.4	23.5	19.0	19.4	18.4
19.4	21.8	20.4	21.0	21.4	19.8	19.6	21.5	20.2	20.1
20.3	19.7	19.5	22.9	20.7	20.3	20.8	19.8	19.4	19.3
19.5	19.8	18.9	20.4	20.2	21.5	19.9	21.7	19.5	20.9
18.1	20.5	18.3	19.5	18.3	19.0	18.2	21.9	17.0	19.7
20.7	21.1	20.6	16.6	19.4	18.6	22.7	18.5	20.1	18.6

The mere gathering of this information is so small task, but it should be clear that more must be done to make the numbers comprehensible.

What can be done to make this mass of information more usable? Some persons find it interesting to locate the extreme values, which are 16.6 and 23.5 for this list. Occasionally, it is useful to sort the data in an ascending or descending order. The following list gives the lengths of the trout arranged in an ascending order.

16.6	17.0	17.3	18.1	18.2	18.3	18.3	18.4	18.5	18.6
18.6	18.9	19.0	19.0	19.2	19.3	19.3	19.4	19.4	19.4
19.4	19.5	19.5	19.5	19.5	19.5	19.6	19.6	19.7	19.7
19.8	19.8	19.8	19.9	20.1	20.1	20.2	20.2	20.3	20.3
20.4	20.4	20.4	20.5	20.6	20.7	20.7	20.8	20.9	21.0
21.1	21.4	21.5	21.5	21.7	21.8	21.9	22.7	22.9	23.5

Sorting a large set of numbers in an ascending or descending order can be a surprisingly difficult task. It is simple, though, if we can use a computer or a graphing calculator. In that case, entering the data is the most tedious part. Then, with a graphing calculator we press STAT and 2, fill in the list where we put the data, press ENTER, and the display screen spells out DONE.

If a set of data consists of relatively few values, many of which are repeated, we simply count how many times each value occurs and then present the results in the form of a Table or a dot diagram. In such a diagram we indicate by means of dots how many times each value occurs.

Example

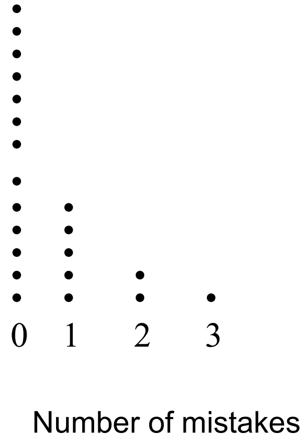
Example (1)
An audit of twenty tax returns revealed 0, 2, 0, 0, 1, 3, 0, 0, 0, 1, 0, 1, 0, 0, 2, 1, 0, 0, 1, and 0 mistakes in arithmetic.

(a) Construct a table showing the number of tax returns with 0, 1, 2, and 3, mistakes in arithmetic.
(b) Draw a dot diagram displaying the same information

Solution

Solution:
Counting the number of 0's, 1's, 2's and 3's we find that they are, respectively, 12, 5, 2, and 1. This information is displayed as follows, in tabular form on the left and n graphical form on the right.

Number of mistakes	Number of the returns
0	12
1	5
2	2
3	1



Number of Mistakes

There are various ways in which dot diagram can be modified, for instance, instead of dots we can use other symbols such as x's, ★'s, or✧'s. Also, we could align the dots horizontally rather than vertically.

The methods we used to display relatively few numerical values, many of which are repeated, can also be used to display categorical data.

Example

Example (2)
The faculty of a university's mathematics department consists of four professors, six associate professors, eleven assistant professors, and nine instructors. Display this information in the form of a horizontally aligned dot diagram.

Solution

Solution:

Faculty Rank											
Professor	✧	✧	✧	✧							
Associate professor	✧	✧	✧	✧	✧	✧					
Assistant professor	✧	✧	✧	✧	✧	✧	✧	✧	✧	✧	✧
Instructor	✧	✧	✧	✧	✧	✧	✧	✧	✧		

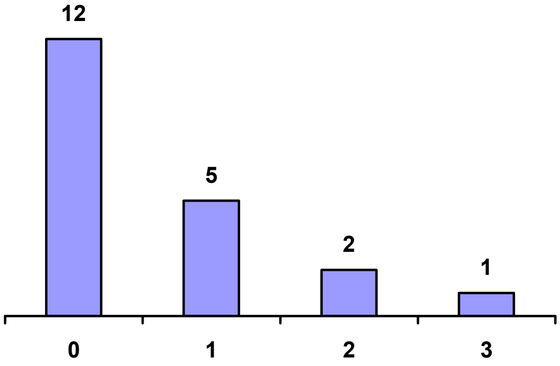
Another way of modifying dot diagram is to replace the numbers of dots with rectangles lengths are proportional to the respective numbers of dots. Such diagrams are referred to as bar charts, and the rectangles are often supplemented with the corresponding frequencies (number of symbols) as shown in the next Figure of Example 3.

Example

Example (3)
Draw a bar chart for the data of Example 3.1; that is, for the numbers of mistakes in arithmetic in the twenty tax returns.

Solution 3

Solution:



Bar Chart of Mistakes in Arithmetic in Tax Returns

3.2 Stem-And-Leaf-Display

Stem-And-Leaf-Display

Dot diagrams are impractical and ineffective when a set of data contains many different values or categories, or when some of the values or categories require too many dots to yield a coherent picture. To give an example, consider the first –round scores in PGA tournament, where the lowest score was a 62, the highest score was an 88, and 27 of the 126 golfers shot a par 72. This illustrates both of the reasons cited previously for not using dot diagrams. There are too may different values from 62 to88, and at least one of them, 72 requires too many dots.

In recent years, **an alternative method of listing data has been proposed for the exploration of relatively small sets of numerical data. It is called a stem-and leaf display and it also yields a good overall picture of the data without any appreciable loss of information.** Again, each value retains its identify, and the only information we lose is the order in which the data were obtained.

To illustrate this technique consider the following data on the number of rooms occupied each day in a resort hotel during a recent month of June:

55	49	37	57	46	40	64	35	73	62
61	43	72	48	54	69	45	78	46	59
40	58	56	52	49	42	62	53	46	81

The smallest and largest values are 35 and 81, so that a dot diagram would require that we allow for 47 possible values. Actually, only 25 of the values occur, but in order to avoid having to allow for that many possibilities, let us combine all the values beginning with a 3, all those beginning with a 4, all those beginning with a 5 and so on. This would yield

37	35									
49	46	40	43	48	45	46	40	49	42	46
55	57	54	59	58	56	52	53			
64	62	61	69	62						
73	72	78								
81										

This arrangement is quite informative, but it is not the kind of diagram we use in actual practice. To simplify it further, **we show the first digit only once for each row, on the left and separated from the other digits by means of a vertical line.** *This leaves us with*

3	7	5									
4	9	6	0	3	8	5	6	0	9	2	6
5	5	7	4	9	8	6	2	3			
6	4	2	1	9	2						
7	3	2	8								
8	1										

And this is what we refer to as a stem-and leaf display. *In this arrangement, each row is called a stem, each number on a stem to the left of the vertical line is called a stem label, and each number on a stem to the right of the vertical line is called a leaf.* As we shall see later, there is a certain advantage to arranging the leaves on each stem according to size, and for our data this would yield

3	5	7									
4	0	0	2	3	5	6	6	6	8	9	9
5	2	3	4	5	6	7	8	9			
6	1	2	2	4	9						
7	2	3	8								
8	1										

A stem-and-leaf display is actually a hybrid kind of arrangement obtained in part by grouping and in part by listing. The values are grouped into the six stems, and yet each value retains its identity Thus, from the preceding stem-and-leaf display, we can reconstruct the original data as 35, 37, 40, 40, 42, 43, 45, 46, 46, 46, 48, 49, 49, 52, 53, …, and 81, though not in their original order.

There are various ways in which stem-and-leaf displays can be modified For instance, the stem labels or the leaves could be two-digit numbers, so that

24 | 0 2 5 8 9

would represent the numbers 240, 242, 245, 248, and 249, and

2 | 31 45 70 88

Would represent the numbers 231, 245, 270, and 288.

Now suppose that in the room occupancy Example we had wanted to use more than six stems. Using each stem label twice, if necessary, once to hold the leaves from 0 to 4 and once to hold the leaves from 5 to 9, we would get

3	5	7					
4	0	0	2	3			
4	5	6	6	6	8	9	9
5	2	3	4				
5	5	6	7	8	9		
6	1	2	2	4			
6	9						
7	2						
7	8						
8	1						

Frequency Distributions

3.3 Frequency Distributions

When we deal with large sets of data, and sometimes even when we deal with not so large sets of data, it can be quite a problem to get a clear picture of the information that they convey. As we saw in Sections 3.1 and 3.2, *this usually requires that we rearrange and/or display the raw (untreated) data in some special form. Traditionally, this involves a frequency distribution or one of its graphical presentations, where we group or classify the data into a number of categories or classes.*

Following are two examples. A recent study of their total billings (rounded to the nearest dollar) yielded data for a sample of 4,757 law firms. Rather than providing printouts of the 4,757 values, the information is disseminated by means of the following table:

Total billings	Number of law firms
Less than $300,000	2,405
$300,000 to $499,999	1,088
$500,000 to $749,999	271
$750,000 to $999,999	315
$1,000,000 or more	678
Total	4,757

This distribution does not show much detail, but it may well be adequate for most practical purposes. This should also be the case in connection with the following table, which summarizes the 2,439 complaints received by an airline about comfort-related characteristics of its airplanes:

Nature of complaint	Number of complaints
Inadequate leg room	719
Uncomfortable seats	914
Narrow aisles	146
Insufficient carry-on facilities	218
Insufficient restrooms	58
Miscellaneous other complaints	384
Total	2,439

When data are grouped according to numerical size, as in the first example, **the resulting table is called a numerical or quantitative distribution. When they are grouped into nonnumerical categories,** as in the second example, **the resulting table is called a categorical or qualitative distribution.**

Frequency distributions present data in a relatively compact form, give a good overall picture, and contain information that is adequate for many purposes, but, as we said previously, there is some loss of information. Some things that can be determined from the original data cannot be determined from a distribution. For instance, in the first Example the distribution does not tell us the exact size of the lowest and the highest billings, nor does it provide the total of the billings of the 4,757 law firms. Similarly, in the second Example we cannot tell how many of the complaints about uncomfortable seats pertained to their width or how many complains about insufficient carry-on facilities applied to particular size luggage. Nevertheless, frequency distributions present information in a generally more usable form, and the price we pay for this-the loss of certain information-is usually a fair exchange.

The construction of a frequency distribution consists essentially of three steps:

1- Choosing the classes (intervals or categories)
2- Sorting or tallying the data into these classes
3- Counting the number of items in each class

Since the second and third steps are purely mechanical, we concentrate here on the first, namely, that of choosing a suitable classification.

For numerical distributions, this consists of deciding how many classes we are going to use and from where to where each classes should go, both of these choices are essentially arbitrary, **but the following rules are usually observed:**

We seldom use fewer than 5 or more than 15 classes; the exact number we use in a given situation depends largely on how many measurements or observations there are.

Clearly, *we would lose more than we gain if we group five observations into 12 classes with most of them empty, and we would probably discard too much information if we group a thousand measurements into three classes.*

We *always make sure that each item (measurement or observation) goes into one and only one class.*

To this end, *we must make sure that the smallest and largest values fall within the classification, that none of the values can fall into a gap between successive classes, and that the classes do not overlap, namely, that successive classes have no values in common.*

Whenever possible, *we make the classes cover equal ranges of values.*

Also, if we can, *we make these ranges multiples of numbers that are easy to work with, such as 5, 10, or 100, since this will tend to facilitate the construction and the use of a distribution.*

If we assume that the law firm billings were all rounded to the nearest dollar**,** only the third of these rules was violated in the construction of the distribution on page 21. However, had the billings been given to the nearest cent, then a billing of, say, $499,999.54 would have fallen between the second class and the third class, and we would also have violated the second rule. The third rule was violated because the classes do not all cover equal ranges of values; in fact, the first class and the last class have, respectively, no specified lower and upper limits.

Classes of the "less than," "or less," "more than," or "or more" variety are referred to as open classes, and they are used to reduce the number of classes that are needed when some of the values are much smaller than or much greater than the rest. Generally, open classes should be avoided, however, because they make it impossible to calculate certain values of interest, such as averages or totals.

Insofar as the second rule is concerned, we have to watch whether the data are given to the nearest dollar or to the nearest cent, whether they are given to the nearest inch or 10 the nearest tenth of an inch, whether they are given to the nearest ounce or to the nearest hundredth of an ounce, and so on. For instance, if we want to group the weights of certain animals, we might use the first of the following classifications when the weights are given to the nearest kilogram, the second when the weights are given to the nearest tenth of a kilogram, and the third when the weights are given to the nearest hundredth of a kilogram:

Weight (Kilograms)	Weight (Kilograms)	Weight (Kilograms)
10-14	10.0-14.9	10.0-14.9
15-19	15.0-1909	15.0-1909
20-24	20.0-24.9	20.0-24.9
25-29	25.0-29.9	25.0-29.9
30-34	30.0-34.9	30.0-34.9
etc.	etc.	etc.

To illustrate what we have been discussing in this section, let us now go through the actual steps of grouping a set of data into a frequency distribution.

Example

Example (4)
Based on 1997 figures, the following are 11.0 "waiting times" (in minutes) between eruptions of the Old Faithful Geyser m Yellowstone National Park:

81	83	94	73	78	94	73	89	112	80
94	89	35	80	74	91	89	83	80	82
91	80	83	91	89	82	118	105	64	56
76	69	78	42	76	82	82	60	73	69
91	83	67	85	60	65	69	85	65	82
53	83	62	107	60	85	69	92	40	71
82	89	76	55	98	74	89	98	69	87
74	98	94	82	82	80	71	73	74	80
60	69	78	74	64	80	83	82	65	67
94	73	33	87	73	85	78	73	74	83
83	51	67	73	87	85	98	91	73	108

Construct a frequency distribution.

Solution

Solution:
Since the smallest value is 33 and the largest value is 118, we have to cover an interval of 86 values and a convenient choice would be to use the nine classes 30 -39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, 80 - 89, 90 - 99, 100 - 109, and 110-119. These classes will accommodate all of the data, they do not overlap, and they are all of the same size. There are other possibilities (for instance, 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65 - 74, 75 - 84, 85 - 94, 95 - 104, 105 - 114, and 115 - 124), but it should be apparent that our first choice will facilitate the tally.

We now tally the 110 values and get the result shown in the following table:

<table>
<tr><th colspan="10">Waiting between
eruption (minutes)
Frequency</th><th>Tally</th></tr>
<tr><td>30-39</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2</td></tr>
<tr><td>40-49</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2</td></tr>
<tr><td>50-59</td><td>||||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>4</td></tr>
<tr><td>60-69</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td></td><td></td><td></td><td></td><td></td><td>19</td></tr>
<tr><td>70-79</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td></td><td></td><td></td><td></td><td>24</td></tr>
<tr><td>80-89</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td></td><td>39</td></tr>
<tr><td>90-99</td><td>||||</td><td>||||</td><td>||||</td><td></td><td></td><td></td><td></td><td></td><td></td><td>15</td></tr>
<tr><td>100-109</td><td>|||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>3</td></tr>
<tr><td>110-119</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Total</td><td>110</td></tr>
</table>

The numbers given in the right-hand column of this table, which show how many values fall into each class, are called the class frequencies. The smallest and largest values that can go into any given class are called its class limits, and for the distribution of the waiting times between eruptions they are 30 and 39, 40 and 49,50 and 59,. .., and 110 and 119. More specifically, **30, 40, 50, ..., and 110 are called the lower class limits, and 39,49,59, ..., and 119 are called the upper class limits.**

The amounts of time that we grouped in our Example were all given to the nearest minute, so that 30 actually includes everything from 29.5 to 30.5,39 includes everything from 38.5 to 39.5, and the class 30-39 includes everything from 29.5 to 39.5. Similarly, the second class includes everything from 39.5 to 49.5... and the class at the bottom of the distribution includes everything from 109.5 to 119.5. It is customary to refer to 29.5, 39.5, 49.5... and 119.5 as the class boundaries or the real class limits of the distribution. Although 39.5 is the upper boundary of the first class and also the lower boundary of the second class, 49.5 is the upper boundary of the second class and also the lower boundary of the third class, and so forth, there is no cause for alarm. The class boundaries are by choice impossible values that cannot occur among the data being grouped. If we assume again that the law firm billings grouped in the distribution on page 21 were all rounded to the nearest dollar, the class boundaries $299,999.50, $499,999.50, $749,999.50, and $999,999.50 are also impossible values.

We emphasize this point because, to avoid gaps in the continuous number scale, some statistics texts, some widely used computer programs, and some graphing calculators (MINITAB, for example, and the TI-83) include in each class its lower boundary, and the highest class also includes its upper boundary. They would include 29.5 but not 39.5 in the first class *of* the preceding distribution of waiting times between eruptions of Old Faithful. Similarly, they would include 39.5 but not 49.5 in the second class,. .., but 109.5 as well as 119.5 in the high boundaries are impossible values that cannot occur among the data being grouped. Especially for this reason, the use of impossible class boundaries can- not be.

Numerical distributions also have what we call class marks and classes intervals. Class marks are simply the midpoints of the classes, and they are found by adding the lower and upper limits of a class (or its lower and upper boundaries) and dividing by 2. A class interval is merely the length of a class, or the range of values it can contain, and it is given by the difference between its boundaries. If the classes of a distribution are all equal in length, their common class interval, which we call the class interval or the distribution, is also given by the difference between any two successive class marks. Thus, the class marks of the waiting-time distribution are 34.5, 44.5, 54.5, ..., and 114.5, and the class intervals and the class interval of the distribution are all equal to 10.

There are essentially two ways in which frequency distributions can be modified to suit particular needs. One way is to convert a distribution into a percentage distribution by dividing each class frequency by the total number of items grouped, and then multiplying by 100.

Example

5

Example (5)

Convert the waiting-time distribution of Example 2.4 into a percentage distribution.

Solution:

Solution

5

The first class contains $\frac{2}{110}.100 = 1.82\%$ of the data (rounded to two decimals), and so does the second class. The third class contains $\frac{4}{110}.100 = 3.64\%$ of the data, the fourth class contains $\frac{19}{110}.100 = 17.27\%$ of the data,..., and the bottom class again contains 1.82% of the data. These results are shown in the following table:

Waiting times between eruptions (minutes)	Percentage
30-39	1.82
40-49	1.82
50-59	3.64
60-69	17.27
70-79	21.82
80-89	35.45
90-99	13.64
110-109	2.73
110-119	1.82

The percentages total 100.01, with the difference, of course, due to rounding.

The other way of modifying a frequency distribution is to convert it into a "less than," "or less," "more than," or "or more" cumulative distribution. To construct a cumulative distribution, we simply add the class frequencies, starting either at the top or at the bottom of the distribution.

Example 6

Example (6)
Convert the waiting-time distribution of Example 6 into a cumulative "less than" distribution.

Solution 6

Solution:
Since none of the values is less than 30, 2 of the values are less than 40, 2 + 2 = 4 of the values are less than 50, 2 + 2 + 4 = 8 of the values are less than 60, ..., and all 110 of the values are less than 120, we get

Waiting times between eruptions (minutes)	Cumulative Frequency
Less than 30	0
Less than 40	2
Less than 50	4
Less than 60	8
Less than 70	27
Less than 80	51
Less than 90	90
Less than 100	105
Less than 110	108
Less than 120	110

Note that instead of "less than 30" we could have written "29 or less," instead of "less than 40" we could have written "39 or less," instead of "less than 50" we could have written "49 or less," and so forth.

In the same way we can also convert a percentage distribution into a cumulative percentage distribution. We simply add the percentages instead of the frequencies, starting either at the top or at the bottom of the distribution.

So far we have discussed only the construction of numerical distributions, but the general problem of constructing categorical (or qualitative) distributions is about the same. Here again we must decide how many categories (classes) to use and what kind of items each category is to contain, making sure that all the items are accommodated and that there are no ambiguities. Since the categories must often be chosen before any data are actually collected, it is usually prudent to include a category labeled "others" or "miscellaneous."

For categorical distributions, we do not have to worry about such mathematical details as class limits, class boundaries, and class marks. On the other hand, there is often a serious problem with ambiguities and we must be very careful and explicit in defining what each category is to contain. For instance, if we had to classify items sold at a supermarket into "meats," "frozen foods," "baked goods," and so forth, it would be difficult to decide, for example, where to put frozen beef pies. Similarly, if we had to classify occupations, it would be difficult to decide where to put a farm manager, if our table contained (without qualification) the two categories "farmers" and "managers." For this reason, it is advisable, where possible, to use standard categories developed by the Bureau of the Census and other government agencies.

Graphical Presentation

3.4 Graphical Presentation

When frequency distributions are constructed mainly to condense large sets of data and present them in an "easy to digest" form, it is usually most effective to display them graphically. As the saying goes, a picture speaks louder than thousand words, and this was true even before the current proliferation of computer graphics. Nowadays, each statistical software package strives to outdo is competitors by means of more and more elaborate pictorial presentations of statistical data.

For frequency distributions, the most common form of graphical presentation is the histogram, like the one shown in Figures 3.1 and 3.2. *Histograms are constructed by representing the measurements or observations that are grouped* (in Figures 3.1-3.2 the waiting times between eruptions of old Faithful) *on a horizontal scale, the class frequencies on a vertical scale, and drawing rectangles whose bases equal the class intervals and whose heights are the corresponding class frequencies.*

The marketing on the horizontal scale of histogram can be the class limits as in Figures 3.1-3.2 *the class marks, the class boundaries, or arbitrary key values. For practical reasons, it is usually preferable to show the class limits, even though the rectangles actually go from one class boundary to the next.* After all, they tell us what values go into each class. *Note that histograms cannot be drawn for distributions with open classes and that they require special care when the class intervals are not all equal.*

The data that led to Figure 3.1 were easy to group because there were only 110 values in the sample. For really large sets of data, it may be convenient to construct histograms directly from raw data by using a suitable computer package or graphing calculator. We said that it may be convenient to use a computer package or a graphing calculator – in actual practice, just entering the data in a computer or a calculator can be more work than tallying the data manually and drawing the rectangles.

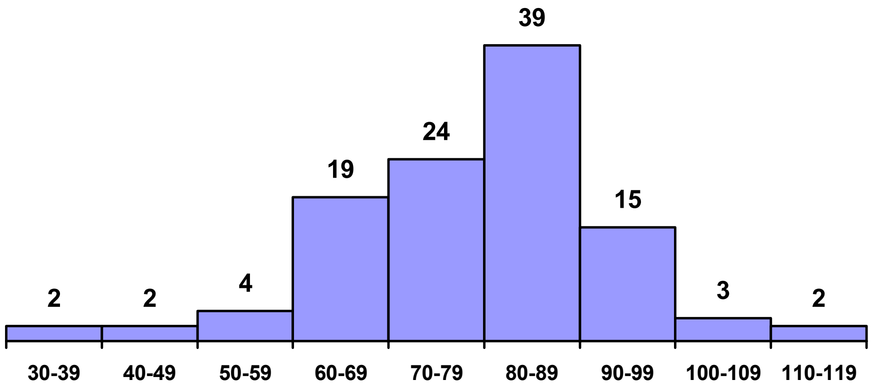


Figure 3.1: Histogram of waiting times between eruptions of old faithful geyser

Also referred to at times as histograms are bar charts (see Section 2.1), *such as the one shown in Figure 3.2. The heights of the rectangles, or bars again represent the class frequency but there is no pretense of having a continuous horizontal scale.*

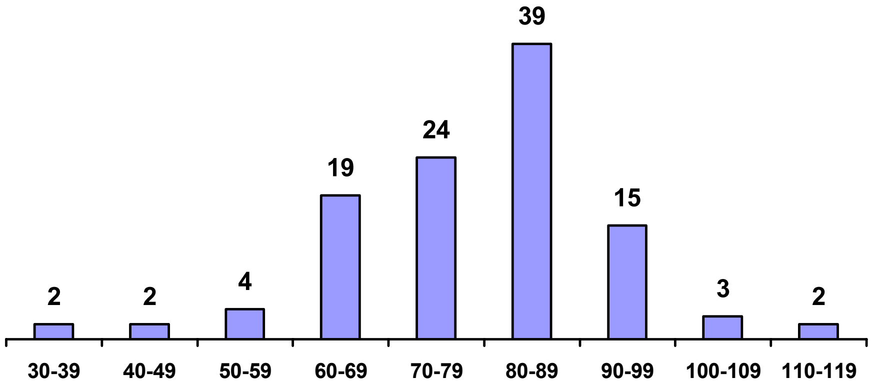


Figure 3.2: Bar Chart of distribution of waiting times between eruptions of old faithful geyser

Measures of Association

3.5 Measures of Association

In Chapter 2, we presented scatter diagrams, which graphically depict variables that are related. In this section, **we present two numerical measure linear relationships depicted in a scatter diagram. The two measures are covariance and the coefficient of correlation.**

Covariance

Covariance

If we have all the observations that constitute a population, we can compute population covariance. It is defined as follows.

Population covariance = $COV(X,Y)=\dfrac{\sum(x_i-\mu_x)(y_i-\mu_x)}{N}$

Where μ_x *is the population mean of the first variable,* $X;\mu_y$ *is the population mean of the second variable, Y*; and *N is the size of the population.* The sample covariance is defined similarly, *where n is the number of pairs of observation sample.*

Sample covariance = $\mathrm{cov}(X,Y)=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

For convenience, we label the population covariance COV(X,Y) and the sample covariance COV(X,Y). To illustrate how covariance measures association, the following three sets of sample data are given.

	x	y	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$
Set 1	2 6 7	13 20 27	-3 1 2	-7 0 7	21 0 14
	$\bar{x}=5$	$\bar{y}=20$			17.5 = cov(X,Y)

	x	y	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$
Set 2	2 6 7	27 20 13	-3 1 2	7 0 -7	-21 0 -14
	$\bar{x}=5$	$\bar{y}=20$			-17.5 = cov(X,Y)

	x	Y	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$
Set 3	2 6 7	20 27 13	-3 1 2	0 7 -7	0 7 -14
	$\bar{x}=5$	$\bar{y}=20$			-3.5 = cov(X,Y)

In set 1, as *x* increases, so does *y.* In this case, when x is larger than its mean, and y is at least as large as its mean, thus $(x_i-\bar{x})$ and $(y_i-\bar{y})$ have the same sign or zero, which means that the product is either positive or zero. Consequently, the covariance is a positive number. In general, if two variables move in the same direction (both increase or both decrease), the covariance will be a large positive number. Figure 3.3 depicts a scatter diagram of one such case.

Next, consider set 2. As x increases, y decreases. Thus, the signs of $(x_i-\bar{x})$and $(y_i-\bar{y})$are opposite. As a result, the covariance is a negative number. If, as one variable increases, the other generally decreases, the covariance will be a large negative number. See Figure 3.4 for an illustrative scatter diagram.

Now consider set 3. As x increases, y exhibits no particular pattern. One product is positive, one is negative, and the third is zero.

Consequently, the covariance is a small number. Generally speaking, if the two variables are unrelated (as one increases, the other shows no pattern), the covariance will be close to zero (either positive or negative). Figures 3.5, 3.6, 3.7, 3.8 describe the movement of two unrelated variables.

As a measure of association, covariance suffers from a major drawback. It is usually difficult to judge the strength of the relationship from the covariance. For example, suppose that you have been told that the covariance of two variables is 250. What does this tell you about the relationship between the two variables? The sign, which is positive, tells you that as one increases, the other also generally increases. However, *the degree to which the two variables move together is difficult to ascertain because we don't know whether 250 is a large number. To over-come this shortcoming, statisticians have produced another measure of association, which is based on the covariance.* **It is called the coefficient of correlation.**

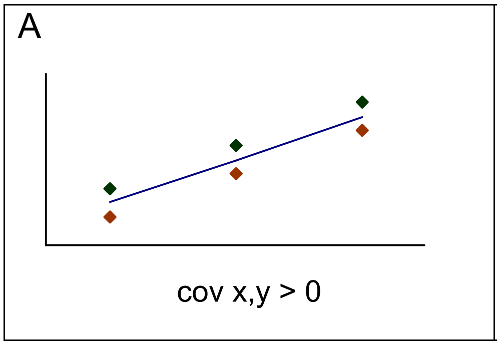


Figure 3.3

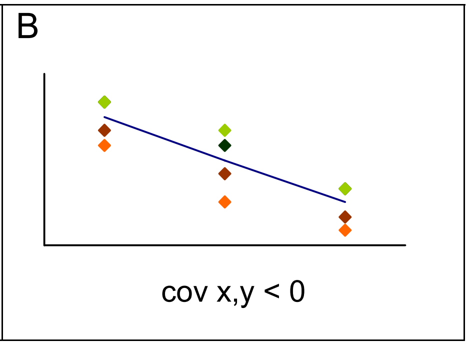


Figure 3.4

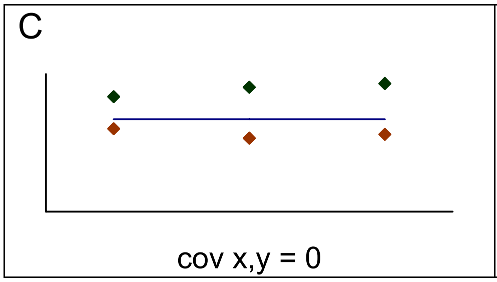


Figure 3.5:

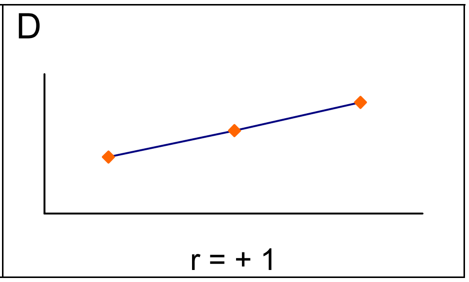


Figure 3.6

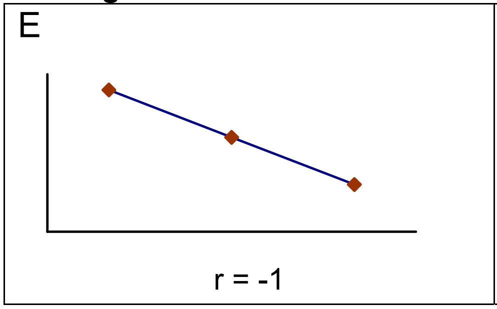


Figure 3.7

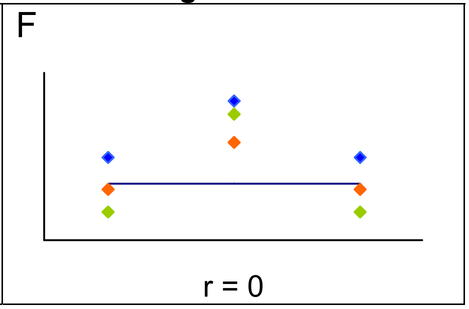


Figure 3.8

Coefficient of Correlation

Coefficient of Correlation

The coefficient of correlation is the covariance divided by the standard deviation of X and Y. The population coefficient of correlation is labeled ρ Greek and **is defined as**

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y}$$

Where σ_x *and* σ_y *are the standard deviations of X and Y, respectively.*

We label **the sample coefficient of correlation r**, which **we define as**

$$r = \frac{\operatorname{cov}(X,Y)}{S_x S_y}$$

Where S_x and S_y are the sample standard deviations of X and Y, respectively.

Solution

Solution:

We begin by calculating the sample means and standard deviations.

$$\bar{x} = 18.0$$

$$S_x = 4.02$$

$$\bar{y} = 217.0$$

$$S_y = 63.9$$

We then compute the deviations from the mean for both x and y, ant, their products. The following Table describes these calculations.

X	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
20.0	219	2.0	2.0	4.0
14.8	190	-3.2	-27.0	86.4
20.5	199	2.5	-18.0	-45.0
12.5	121	-5.5	-96.0	528.0
18.0	150	0.0	-67.0	0.0
14.3	198	-3.7	-19.0	70.3
24.9	334	6.9	117.0	807.3
16.5	188	-1.5	-29.0	43.5
24.3	310	6.3	93.0	585.9
20.2	213	2.2	-4.0	-8.8
22.0	288	4.0	71.0	284.0
19.0	312	1.0	95.0	95.0
12.3	186	-5.7	-31.0	176.7
14.0	173	-4.0	-44.0	176.0
16.7	174	-1.3	-43.0	55.9

Total = 2,859.2

Thus,

$$\operatorname{cov}(X,Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{2{,}859.2}{14} = 204.2$$

The coefficient of correlation is

$$r = \frac{\mathrm{cov}(X,Y)}{S_x Y_x} = \frac{204.2}{4.02 \times 63.9} = .796$$

USING THE COMPUTER

Excel Output for the Example

Excel Output for the Example

1		Size	Price
2	Size	15.0667	
3	Price	190.6133	3808.667

Excel prints the population covariance and variances. Thus, cov(X,Y) = 109.6133, σ^2=15.06667, and σ_y^2= 3,808.667. To compute the corresponding sample statistics, multiply each by n/(n-1). Therefore, the sample covariance is cov (X, Y) = 190.6133 ×(15/14) = 204.2286.

COMMANDS	COMMANDS FOR EXAMPLE
1 type or import the data into two columns 2 click Tools, Data Analysis …, and Covariance 3 Specify the coordinates of the data	Open file A1:B16

From the output, we observe that r = .795716

Commands

Commands

Repeat the steps above, except click Correlation instead of Covariance.

	A	B	C
1		Size	Price
2	Size	1	
3	Price	0.795716	1

The covariance provides very little useful information other than telling us that the two variables are positively related. The coefficient of correlation informs us that there is a strong positive relationship. This information can be extremely useful to real estate agents, insurance brokers, and all potential home purchasers.

Excel Output for the Example

Excel Output for the Example

	A	B	C
1		Odometer	Price
2	Odometer	1	
3	Price	-0.806307604	1

Excel prints the coefficient of correlation. The test can manually

COMMANDS	COMMANDS FOR EXAMPLE
1 type or import the data into adjacent columns 2 click Tools, Data Analysis …, and Correlation 3 Specify the input range. Click Labels in First row (if necessary). Click OK	Open file A1:B101

Interpreting the results

There is overwhelming evidence to infer that the two variables are correlated.

Spearman Rank Correlation Coefficient

Spearman Rank Correlation Coefficient

In the previous sections of this chapter, we have dealt only with quantitative variables and have assumed that all of the conditions for the validity of the hypothesis tests and confidence interval estimates have been met. In many situations, however, one or both variables may be ranked, or if both variables are quantitative, the normality requirement may not be satisfied. In such cases, we measure and test to determine if a relationship exists by employing a nonparametric technique, the Spearman rank correlation coefficient.

The Spearman rank correlation coefficient is calculated like all of the previously introduced nonparametric methods by first ranking the data. We then calculate the Pearson correlation coefficient of the ranks.

The population Spearman correlation coefficient is labeled ρ_s *, and the sample statistics used to estimate its value is labeled* r_s .

Sample Spearman Rank Correlation Coefficient

Sample Spearman Rank Correlation Coefficient

$$r_s = \frac{SS_{ab}}{\sqrt{SS_a . SS_b}}$$

Where *a and b are the ranks of the data.*

Summarizing Two-Variable Data

3.6 Summarizing Two-Variable Data

So far we have dealt only with situations involving one variable- the room occupancies in Section 2.2, the waiting times between eruptions of Old Faithful in Example 2.4, and so on. In actual practice, many statistical methods apply to situations involving two variables, and some of them apply even when the number of variables cannot be counted on one's fingers and toes not quite so extreme would be a problem in which we want to study the values of one-family homes, taking into consideration their age, their location, the number of bedrooms, the number of baths, the size of the garage, the type of roof, the number of fireplaces, the lot size, the value of nearby properties, and the accessibility of schools.

Leaving some of this work to later work and, in fact, most of it to advanced courses in statistics, we shall treat here only the display, listing, and grouping of data involving two variables; that is, problem dealing with the display of paired data. **In most of these problems, the main objective is to see whether there is a relationship, and if so what kind of relationship, so that we can predict one variable, denoted by the letter y, in terms of other variable denoted by the letter x .**For instance, the x's might be family incomes and the y's might be family expenditures on medical care, they might be annealing temperatures and the hardness of steel, or they might be the time that has elapsed since the chemical treatment of a swimming pool and the remaining on concentration of chlorine.

Pairs (x, y), in the same way which we denote points in the plane, with x, and y being their x- and y-coordinates. When we actually plot the points corresponding to paired values of x and y, we refer to the resulting graph as a scatter diagram, ***a scatter plot, or a scatter gram.*** *As their name implies,* **such graphs are useful tools in the analysis of whatever relationship there may exist between the x's and the y's namely, judging whether there are any discernible patterns.**

Example

7

Example (7)

Raw materials used in the production of synthetic fiber are stored in a place that has no humidity control. Following are measurement of the relative humidity in the storage place, x, and the moisture content of a sample of the raw material, y, on 15 days

X (Percent)	Y (Percent)		X (Percent)	Y (Percent
36	12		3	14
27	11		32	13
24	10		19	11
50	17		34	12
1	10		38	17
23	12		21	8
45	18		16	7
44	16			

Construct a scatter gram.

Solution

7

Solution

Scatter grams are easy enough to draw, yet the work can be simplified by using appropriate computer software or a graphing calculator. The one shown in Figure 3.9 was reproduced from the display screen of a TI-83 graphing calculator.

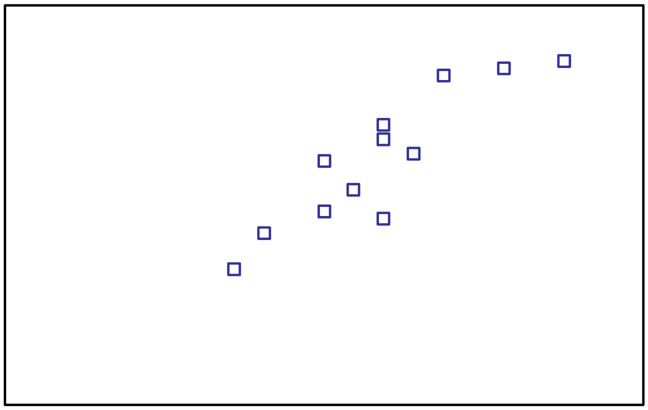


Figure 3.9: Scatter gram of humidity and water content data

As can be seen from the diagram the points are fairly widely scattered, yet there is evidence of an upward trend that is, increase in the water content of the raw material seem to go with increase in humidity. In Figure 3.9 the dots are squares with their centers removed, but they can also be circles, x's dots, or other kinds of symbols (The units are not marked to either scale, but on the horizontal axis the tick marks are at 10, 20, 30, 40, and 50, and on the vertical axis they are at 5, 10, 15, and 20).

Some difficulties arise when two or more of the data points are identical. In that case, the TI-83 graphing calculator shows only one point and so do some of the printouts obtained with statistical software. However, **MINITAB has a special scatter gram to take care of situations like this. Its so called character plot prints the number 2 instead of the symbol x or ★ to indicate that there are two identical data points, and it would print a 3 if there were three.** *This is illustrated by the following example.*

Example

Example (8)

Following are the scores which 40 students obtained on both parts of the test, with the scores on the even-numbered problems denoted by x and the scores on the odd-numbered problems denoted by y.

x	y	x	y	x	y	x	y
40	39	32	23	37	34	32	28
45	45	45	35	41	38	40	34
27	24	42	36	35	33	37	37
42	39	44	42	34	30	47	45
42	9	41	35	38	40	44	40
49	40	48	45	42	34	35	35
36	28	44	39	32	35	44	35
39	39	40	28	38	27	43	38
43	38	50	48	36	37	37	35
39	34	37	39	43	42	43	33

Choosing the five classes 26-30, 31-35, 36-40, 41-45, and 46-50 for x and the six classes 21-25, 26-30, 31-35, 36-4041-45, and 46-50 for y, group these data into a two-way frequency distribution.

Solution

Solution

Performing the tally, we find that the first of values, 40 and 39, goes into the cell belonging to the third column and the fourth row, the second pair of values, 45 and 45, goes into the cell belonging to the fourth column and the fifth row, and so on. We thus get.

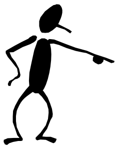
		x				
		26-30	31-35	36-40	41-45	46-50
	21-25	|	|			
	26-30		||	|||	|	
	31-35		|||	||||	||||	
y	36-40			|||| |	|||| ||	|
	41-45				|||	||
	46-50					|

and, hence, the following two-way frequency distribution :

		x				
		26-30	31-35	36-40	41-45	46-50
	21-25	1	1			
	26-30		2	3	1	
	31-35		3	4	5	
y	36-40			6	7	1
	41-45				3	2
	46-50					1

Chapter 4: Summarizing Data: Measures of Location

Introduction



When we are about to describe a set of data, it is a sound advice to say neither too little nor too much. Thus, depending on the nature of the data and the purpose we have in mind, statistical descriptions can be very brief or very elaborate. Sometimes we present data just as they are and let them speak for themselves; on other occasions we may just group the data and present their distribution in tabular or graphical form. Most of the time, though, we have to describe data in various other ways.

It is often appropriate to summarize data by means of a few well-chosen numbers that, in their way, are descriptive of the entire set. Exactly what sort of numbers we choose depends on the particular characteristics we want to describe. In one study we may be interested in a value that somehow describes the middle or the most typical of a set of data; in another we may be interested in the value that is exceeded only by 25% of the data; and in still another we may be interested in the length of the interval between the smallest and the largest values among the data. *The statistical measures cited in the first two situations come under the heading of measures of location and the one cited in the third situation fits the definition of a measure of variation.*

In this chapter, we shall concentrate on measures of location, and in particular on measures of central location, which in some way describe the center or the middle of a set of data. Measures of variation and some other kinds of statistical descriptions will be discussed in next Chapter.

Populations And Samples

4.1 Populations and Samples

When we stated that the choice of a statistical description may depend on the nature of the data, we were referring among other things to the following distinction:

If a set of data consists of all conceivably possible (or hypothetically possible) observations of a given phenomenon, we call it a population; if a set of data consists of only a part of these observations, we call it a sample.

Here, we added the phrase "hypothetically possible" to take care of such clearly hypothetical situations as where we look at the outcomes (heads or tails) of 12 flips of a coin as a sample from the potentially unlimited number of flips of the coin, where we look at the weights of ten 30-day-old lambs as a sample of the weights of all (past, present, and future) 30-day-old lambs raised at a certain farm, or where we look at four determination of the uranium content of an ore as a sample of the many determinations that could conceivably be made. In fact, we often look at the results of an experiment as a sample of what we might get if, the experiment were repeated over and over again.

Originally, statistics dealt with the description of human populations, census, counts and the like, but as it grew in scope, the term "population" took on the much wider connotation given to it in the preceding distinction between populations and samples. Whether or not it sounds strange to refer to the heights of all the trees in a forest or the speeds of all the cars passing a checkpoint as populations is beside the point-in statistics, "population" is a technical term with a meaning of its own.

Although we are free to call any group of items a population, what we do in practice depends on the context in which the items are to be viewed. Suppose, for instance, that we are offered a lot of 400 ceramic tiles, which we may or may not buy depending on their strength. If we measure the breaking strength of 20 of these tiles in order to estimate the average breaking strength of all the tiles, these 20 measurements are a sample from the population that consists of the breaking strengths of the 400 tiles. In another context, however, if we consider entering into a long-term contract calling for the delivery of tens of thousands of such tiles, we would look upon the breaking strengths of the original 400 tiles only as a sample. Similarly, the complete figures for a recent year, giving the elapsed times between the filing and disposition of divorce suits in a County, can be looked upon as either a population or a sample. If we are interested only in a County and that particular year, we would look upon the data as a population; on the other hand, if we want to generalize about the time that is required for the disposition of divorce suits in: the entire Country, in some other County, or in some other year, we would look upon the data as a sample.

As we have used it here, the word "sample" has very much the same meaning as it has in everyday language. A newspaper considers the attitudes of 150 readers toward a proposed school bond to be a sample of the attitudes of all its readers toward the bond; and a consumer considers a box of Mrs. See's candy a sample of the firm's product. Later, *we shall use the word "sample" only when referring to data that can reasonably serve as the basis for valid generalizations about the populations from which they came; in this more technical sense, many sets of data that are popularly called samples are not samples at all.*

In this chapter and in the next one we shall describe things statistically without making any generalizations. For future reference, though, it is important to distinguish even here between populations and samples. Thus, we shall use different symbols depending on whether we are describing populations or samples.

The Mean

4.2 The Mean

The most popular measure of central location is what the lay person calls an "average" and what the statistician calls an arithmetic mean, or simply a mean. **It is defined as follows:**
The mean of n numbers is their sum divided by n

It is all right to use the word "average," and on occasion we shall use it ourselves, but there are other kinds of averages in statistics and we cannot afford to speak loosely when there is any risk of ambiguity.

Example **1**

Example (1)
From 1990 through 1994, the combined seizure of drugs the Drug Enforcement Administration, Custom's Service added up to 1,794, 3,030, 2,551, 3,514, and 2,824 pounds. Find the mean seizure of drugs for the given five-year period.

Solution **1**

Solution:
The total for the five years is:

$$1{,}794 + 3{,}030 + 2{,}551 + 3{,}514 + 2{,}824 = 13{,}713$$

Pounds, so that the mean is $\frac{13{,}713}{5} = 2{,}742.6$ pounds.

Example **2**

Example (2):
In the 9th through 97th Congress of Egypt, there were, respectively, 67, 71, 78, 82, 96, 110, 104, and 92 Representatives at least 60 years old at the beginning of the first session. Find the mean.

Solution **2**

Solution:
The total of these figures is 67+ 71 + 78 + 82 + 96 + 110 + 104 + 92 = 700. Hence, the mean is $\frac{700}{8} = 87.5$.

Since we shall have occasion to calculate the means of many different sets of sample data, it will be convenient to have a simple formula that is always applicable. This requires that *we represent the figures to be averaged by some general symbol such as x, y, or z; the number of values in a sample, the sample size, is usually denoted by the letter n. Choosing the letter x, we can refer to the n values in a sample as x_1, x_2, …, and x_n (which read "x sub-one," "x sub-two," ..., and "x sub-n"), and write*

Sample mean = $\frac{x_1 + x_2 + x_3 + ... + x_n}{n}$

This formula will take care of any set of sample data, but it can be made more compact by assigning the sample mean the symbol $\bar{x}$ (which reads “x bar") and using the $\sum$ notation. The symbol $\sum$ is capital sigma*,* the Greek letter for S. In this notation we let $\sum x$; stand for "the sum of the x's" (that is, $\sum x = x_1 + x_2 + ... + x_n$), and we can write

$$\boxed{\bar{x} = \frac{\sum x}{n}}$$

If we refer to the measurements as y's or z's, we write their mean as $\bar{y}$ or $\bar{z}$. In the formula for $\bar{x}$ the term $\sum x$ does not state explicitly which values of x are added; let it be understood, however, that $\sum x$ always refers to the sum of all the x's under consideration in a given situation.

The number of values in a population, the population size, is usually denoted by N. **The mean of a population of N items** *is defined in the same way as the mean of a sample.* **It is the sum of the N items,** $x_1 + x_2 + x_3 + ... + x_N$ **or** $\sum x$ **divided by N.**

Assigning the population mean the symbol μ (mu, the Greek letter for lowercase m) we write

$$\boxed{\mu = \frac{\sum x}{N}}$$

With the reminder that $\sum x$ is now the sum of all N values of x that constitute the population.

Also, *to distinguish between descriptions of populations and descriptions of samples, we not only use different symbols such as* μ *and* $\bar{x}$*, but we refer to a description of a population as a parameter and a description of a sample as a statistic. Parameters are usually denoted by Greek letters.*

To illustrate the terminology and notation just introduced, suppose that we are interested in the mean lifetime of a production lot of N = 40,000 light bulbs. Obviously, we cannot test all of the light bulbs for there would be none left to use or sell, so we take a sample, calculate *X,* and use this quantity as an estimate of μ.

Example **3**

Example (3)

If n = 5 and the light bulbs in the sample last 967, 949, 952, 940, and 922 hours, what can we conclude about the mean lifetime of the 40,000 light bulbs in the production lot?

Solution **3**

Solution:

The mean of this sample is

$$\overline{x} = \frac{967 + 949 + 952 + 940 + 922}{5} = 946 \text{ hours}$$

If we can assume that the data constitute a sample in the technical sense (namely, a set of data from which valid generalizations can be made), we estimate the mean of all 40,00 light bulbs as μ = 946 hours.

For nonnegative data, the mean not only describes their middle, but it also puts some limitation on their size. If we multiply by n on both sides of the equation $\overline{x} = \frac{\sum x}{n}$, we find that $\sum x = n \cdot \overline{x}$ and, hence, that no part, or subset of the data can exceed $n \cdot \overline{x}$.

Example ***4***

Example (4)

If the mean salary paid to three NBA players for the 1998-1999 season is $2,450,000, can:

a. Anyone of them receive an annual salary of $4,000,000;
b. Any two of them receive an annual salary of $4,000,000?

Solution ***4***

Solution:

The combined salaries of the three players total 3(2,450,000) = $7,350,000.

a. If one of them receives an annual salary of $4,000,000, this would leave 7,350,000 - 4,000,000 = $3,350,000 for the other two players, so this could be the case.
b. For two of them to receive an annual salary of $4,000,000 would require 2(4,000,000 = $8,000,000, which exceeds the total paid to the three players. Hence, this cannot be the case.

Example ***5***

Example (5)

If six high school juniors averaged 57 on the verbal part of the PSAT/MSQT test, at most how many of them could have scored 72 or better on the test?

Solution **5**

Solution:
Since n = 6 and x = 57, it follows that their combined scores total 6(57) = 342. Since 342 = 4 x 72 + 54, we find that at most four of the six students could have scored 72 or more.

The popularity of the mean as a measure of the "middle" or "center" of a set of data is not accidental. Anytime we use a single number to describe some aspect of a set of data, there are certain requirements, or desirable features, that should be kept in mind. Aside from the fact that the mean is a simple and familiar measure, **the following are some of it- noteworthy properties:**

1- The mean can be calculated for any set of numerical data, so it always exists.
2- Any set of numerical data has one and only one mean, so it is always unique.
3- The mean lends itself to further statistical treatment; for instance, as we shall see, the means of several sets of data can always be combined into the overall mean of all the data.
4- The mean is relatively reliable in the sense that means of repeated samples drawn from the same population usually do not fluctuate, or vary, as widely as other statistical measures used to estimate the mean of a population.

Finally, let us consider another property of the mean that, on the surface, seems desirable.

5- *The mean takes into account each item in a set of data.*

Note, however, that samples may contain very small or very large values that are so far removed from the main body of the data that the appropriateness of including them in the sample is questionable. Such values may be due to chance, they may be due to gross errors in recording the data, gross errors in calculations, malfunctioning of equipment, or other identifiable sources of contamination. In any case, when such values are averaged in with the other values, they can affect the mean to such an extent that it is debatable whether it really provides a useful, or meaningful, description of the "middle" of the data.

Example **6**

Example (6)
The editor of a book on nutritional values needs a figure for the calorie count of a slice of a l2-inch pepperoni pizza. Letting a laboratory with a calorimeter do the job, she gets the following figures for the pizza from six different fast- food chains: 265, 332, 340, 225, 238, and 346.

a) Calculate the mean, which the editor will report in her book.

b) Suppose that when calculating the mean, the editor makes the mistake of entering 832 instead of 238 in her calculator. How much of an error would this make in the

c) Figure that she reports in her book?

Solution
6

Solution:

a) The correct mean is:

$$\bar{x} = \frac{265+332+340+225+238+346}{6} = 291$$

(b) The correct mean is:

$$\bar{x} = \frac{265+332+340+225+238+346}{6} = 390$$

So that her error would be a disastrous 390-291 = 99.

Example
7

Example (7)

The ages of six students who went on a geology field trip are 16, 17, 15, 19, 16, and 17, and the age of the instructor who went with them is 54. Find the mean age of these seven persons.

Solution
7

Solution:

The mean is:

$$\bar{x} = \frac{16+17+15+19+16+17+54}{7} = 22$$

But any statement to the effect that the average age of the group is 22 could easily be misinterpreted. We might well infer incorrectly that most of the persons who went on the field trip are in their low twenties.

To avoid the possibility of being misled by a mean affected by a very small value or a very large value, we sometimes find it preferable to describe the middle or center of a set of data with a statistical measure other than the mean; perhaps, with the median, which we shall discuss.

The Weighted Mean

4.3 The Weighted Mean

When we calculate a mean, we may be making a serious mistake if we overlook the fact that the quantities we are averaging are not all of equal importance with reference to the situation being described. Consider, for example, a cruise line that advertises the following fares for single-occupancy cabins on an 11-day cruise:

Cabin category	Fare
Ultra deluxe(outside)	$7,870
Deluxe (outside)	$7,080
Outside	$5.470
Outside (shower only)	$4,250
Inside (shower only)	$3.46

The mean of these five fares is

$$\bar{x} = \frac{7.870 + 7{,}080 + 5{,}470 + 4{,}250 + 3{,}460}{5} = \$5{,}626$$

But we cannot very well say that the average fare for one of these single occupancy cabins is $5,626. To get that figure, we would also have to know how many cabins there are in each of the categories. Referring to the ship's deck plan, where the cabins are color-coded by category, we find that there are, respectively, 6, 4, 8, 13, and 22 cabins available in these five categories. If it can be assumed that these 53 cabins will all be occupied, the cruise line can expect to receive a total of:

$$6(7{,}870)+4(7{,}080)+8(5{,}470)+13(4{,}250)+22(3{,}460)=250{,}670$$

for the 53 cabins and, hence, on the average $\frac{250{,}670}{53} \approx \$4{,}729.62$ per cabin.

To give quantities being averaged their proper degree of importance, it is necessary to assign them (relative importance) weights and then calculate a weighted mean. In general, the weighted mean $\bar{x}_w$ of a set of numbers x_1, x_2, x_3, … and x_n, whose relative importance is expressed numerically by a corresponding set of numbers w_1, w_2, w_3, … and w_n is given by:

Weighted mean

$$\boxed{\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum w \cdot x}{\sum w}}$$

Here $\sum w \cdot x$ **is the sum of the products obtained by multiplying each x by the corresponding weight, and** $\sum w$ **is simply the sum of the weights.** *Note that when the weights are all equal, the formula for the weighted mean reduces to that for the ordinary (arithmetic) mean. ..*

Example
8

Example (8)
The following Table shows the number of households in the five Pacific states in 1990, and the corresponding percentage changes in the number of households 1990-1994:

	Number of households (1,000)	Percentage change
Washington	1,872	9.1
Oregon	1,103	8.3
California	10,381	4.5
Alaska	189	10.3
Hawaii	356	7.1

Calculate the weighted mean of the percentage changes using the 1990 numbers of households as weights.

Solution
8

Solution:
Substituting x_1 = 9.1, x_2 = 8.3, X_3 = 4.5, x_4 = 10.3, X_s = 7.1, W_l = 1,872, W_2 = 1,103, w_3 = 10,381, W_4 = 189, and W_s = 356 into the formula for the weighted mean, we get

$$\frac{9.1(1{,}872)+8.3(1{,}103)+4.5(10.381)+10.3(189)+7.1(356)}{1{,}872+1{,}103+10{,}381+189+356}$$

$$=\frac{77{,}378.9}{13.901}\approx 5.6\%$$

Note that we used the symbol ≈ to mean "approximately equal to." We use this symbol only for steps where numerical rounding occurs.

*A special application of the formula for the weighted mean arises when we must find **the overall mean,** or **grand mean**, of k sets of data having the means $\bar{x}_1, \bar{x}_2, \bar{x}_3, ...$ and $\bar{x}_k$ and consisting of $n_1, n_2, ..., n_3$, and n_k measurements or observations. The result is given by:*

Grand mean of combined data

$$\boxed{\bar{\bar{x}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + ... + n_k\bar{x}_k}{n_1 + n_2 + ... + n_k} = \frac{\sum n \cdot \bar{x}}{\sum n}}$$

Where the weights are the sizes of the samples, the numerator is the total of all the measurements or observations, and the denominator is the number of items ~ in the combined samples.

Example
9

Example (9)
There are three sections of a course in European history, with 19 students in the 1st section meeting MWF at 9 A.M., 27 in the 2nd section meeting MWF at 11 A.M., and 24 in the 3rd section meeting MWF at 1 P.M. If the students in the 9 A.M. section averaged 66 in the midterm examination, those in the 11 A.M. section averaged 71, and those in the 1 P.M. section averaged 63, what is the mean score for all three sections combined?

Solution **9**

Solution:

Substituting $n_1 = 19, n_2 = 27, n_3 = 24, \bar{x}_1 = 66, \bar{x}_2 = 71$ and $\bar{x}_3 = 63$ into the formula for the grand mean of combined data, we get

$$\bar{\bar{x}} = \frac{19*66+27*71+24*63}{19+27+24} = \frac{4,683}{70} = 66.9$$

Or 67 rounded to the nearest integer.

The Median

4.4 The Median

To avoid the possibility of being misled by one or a few very small or very large values, we sometimes describe the "middle" or "center" of a set of data with statistical measures other than the mean. One of these, the median of n values requires that we first arrange the data according to size. Then **it is defined as follows:**

The median is the value of the middle item when n is odd, and the mean of the two middle items when n is even.

In either case, when no two values are alike, the median is exceeded by as many values as it exceeds. When some of the values are alike, this may not be the case.

Example **10**

Example (10)
In five recent weeks, a town reported 36, 29, 42, 25, and 29 burglaries. Find the median number of burglaries for these weeks.

Solution **10**

Solution:
The median is not 42, the third (or middle) item, because the data must first be arranged according to size. Thus, we get:

25 29 29 36 42

and it can be seen that the middle one, the median, is 29.

Note that in this Example there are two 29's among the data and that we did not refer to either of them as the median - *the median is a number and not necessarily a particular measurement or observation.*

Example **11**

Example (11)
In some cities, persons cited for minor traffic violations can attend a class in defensive driving in lieu of paying a fine. Given that 12 such classes in Phoenix, Arizona, were attended by 37, 32, 28, 40, 35, 38, 40, 24, 30, 37, 32, and 40 persons, find the median of these data.

Solution ***11***

Solution:
Ranking these attendance figures according to size, from low to high, we get

	24	28	30	32	32	35	37	37	38	40	40	40

and we find that the median is the mean of the two values nearest the middle namely, $\frac{35+37}{2}=36$

Some of the values were alike in this example, but not affect the median, which exceeds six of the values and is exceeded by equally many. The situation is quite different, however, in the Example that follows.

Example ***12***

Example (12)
On the seventh hole of a certain golf course, a par four, nine golfers scored par, birdie (one below par), par, par, bogey (one above par), eagle (two below par), par, birdie, birdie. Find the median.

Solution ***12***

Solution:
Ranking these figures according to size, from low to high, we get

2 3 3 3 4 4 4 4 5

and it can be seen that the fifth value, the median, is equal to par 4.

This time the median exceeds four of the values but is exceeded by only one, and it may well be misleading to think of the median, 4, as the middle of the nine scores. It is not exceeded by as many values as it exceeds, but by definition the median is 4.

The symbol that we use for the median of n sample values x_1, x_2, x_3, …, and x_n (and, hence, $\tilde{y}$ or $\tilde{z}$ if we refer to the values of y's or z's) is μ. If a set of data constitutes a population, we denote its median by $\tilde{\mu}$.

Thus, we have a symbol for the median, but no formula; there is only a formula for the median position. Referring again to data arranged according to size, usually ranked from low to high, we can write

Median position	**The median is the value of the** $\frac{n+1}{2}$ **th item**

***Example* 13**

Example (13)
Find the median position for

(a) n = 17; (b) n = 41.

***Solution* 13**

Solution:
With the data arranged according to size (and counting from either end)

(a) $\frac{n+1}{2} = \frac{17+1}{2} = 9$ and the median is the value of the 9th item;

(b) $\frac{n+1}{2} = \frac{41+1}{2} = 21$ and the median is the value of the 21st item.

***Example* 14**

Example (14)
Find the median position for

(a) n = 16; (b) n = 50.

***Solution* 14**

Solution:
With the data arranged according to size (and counting from either end)

(a) $\frac{n+1}{2} = \frac{16+1}{2} = 8.5$ and the median is the mean of the values of the 8th and 9th items;

(b) $\frac{n+1}{2} = \frac{50+1}{2} = 25.5$ and the median is the mean of the values of the 25th and 26th items.

It is important to remember that $\frac{n+1}{2}$ *is the formula for the median position and not a formula for the median, itself. It is also worth mentioning that determining the median can usually be simplified, especially for large sets of data, by first presenting the data in the form of a stem-and-leaf display.*

***Example* 15**

Example (15)
We gave data on the number of rooms occupied each day in a resort hotel during the month of June, and we displayed these data as follows:

2	3	57
6	4	0023
13	4	5666899
(3)	5	234
14	5	56789

9	6	1224
5	6	9
4	7	23
2	7	8
1	8	1

Use this double-stem display to find the median of these room-occupancy data.

Solution **15**

Solution:

When we gave this display in earlier, we did not explain *the significance of the figures in the column to the left of the stem labels.* As can easily be verified, *they are simply the accumulated numbers of leaves counted from either end.* Furthermore, *the parentheses around the 3 are meant to tell us that the median of the data are on that stem (or else are the mean of two values on that stem).*

Since n = 30 for the given table, the median position is $\frac{30+1}{2} = 15.5$, so that the median is the mean of the fifteenth and sixteenth largest values among the data. Since 2 + 4 + 7 = 13 of the values are represented by leaves on the first three stems, the median is the mean of the values represented by the second and third leaves on the fourth stem. These are 53 and 54, and hence the median of the room-occupancy data is $\frac{53+54}{2} = 53.5$. Note that this illustrates why we said that it is generally advisable to arrange the leaves on each stem, so that they are ranked from low to high.

As a matter of interest, let us also mention that the mean of the room-occupancy data is 55.7. It really should not come as a surprise that the median does not equal the mean-it defines the middle of a set of data in a different way. *The median is average in the sense that it splits the data into two parts so that, unless there are duplicates, there are equally many values above and below the median. The mean, on the other hand, is average in the sense that if each value is replaced by some constant k while the total remains unchanged, this number k will have to be the mean. (This follows directly from the relationship,* $n \cdot \bar{x} = \sum x \cdot$*) In this sense, the mean has also been likened to a center of gravity.*

The median shares some, but not all, of the properties of the mean. ***Like the mean,*** *the median always exists and it is unique for any set of data. Also like the mean, the median is simple enough to find once the data have been arranged according to size, but as we indicated earlier, sorting a set of data manually can be a surprisingly difficult task.*

Unlike the mean, *the medians of several sets of data cannot generally be combined into an overall median of all the data, and in problems of statistical inference the median is usually less reliable than the mean. This is meant to say that the medians of repeated samples from the same population will usually vary more widely than the corresponding means. On the other hand, sometimes the median may be preferable to the mean because it is not so easily, or not at all, affected by extreme (very small or very large) values.* For instance, we showed that incorrectly entering 832 instead of 238 into a calculator caused an error of 99 in the mean. As the reader will be asked to verify, the corresponding error in the median would have been only 37.5.

Finally, *also* ***unlike the mean,*** *the median can be used to define the middle of a number of objects, properties, or qualities that can be ranked, namely, when we deal with ordinal data.* For instance, we might rank a number of tasks according to their difficulty and then describe the middle (or median) one as being of "average difficulty." Also, we might rank samples of chocolate fudge according to their consistency and then describe the middle (or median) one as having "average consistency."

Besides the median and the mean there are several other measures of central location; for example, ***the midrange described*** *and* ***the mid quartile****. Each describes the "middle" or "center" of a set of data in its own way, and it should not come as a surprise that their values may well all be different. Then there is also* ***the mode.***

Other Fractiles

4.5 Other Fractiles

The *median is but one of many fractiles that divide data into two or more parts, as nearly equal as they can be made. Among them we also find quartiles, deciles, and percentiles, which are intended to divide data into four, ten, and a hundred parts. Until recently, fractiles were determined mainly for distributions of large sets of data.*

In this section, we shall concern ourselves mainly with a problem that has arisen in exploratory data analysis - in the preliminary analysis of relatively small sets of data. It is the problem of dividing such data into four nearly equal parts, where we say "nearly equal" because there is no way in which we can divide a set of data into four equal parts for, say, n = 27 or n = 33. Statistical measures designed for this purpose have traditionally been referred to as the three quartiles, Q_1, Q_2, and Q_3, and there is no argument about Q_2, which is simply the median. On the other hand, there is some disagreement about the definition of Q_1, and Q_3.

As we shall define them, *the quartiles divide a set of data into four parts such that there are as many values less than Q_1 as there are* *between Q_1 and Q_2 between Q_2 and Q_3, and greater than Q_3. Assuming that no two values are alike, this is accomplished by letting Q_1 be the median of all the values less than the median of the whole set of data, and Q_3 be the median of all the values greater than the median of the whole set of data.*

Example **16**

Example (16)
Following are the high-temperature readings in twelve European capitals on a recent day in the month of June: 90, 75, 86, 77, 85,72,78,79,94,82,74, and 93. Find Q_1, Q_2 (the median), and Q_3.

Solution **16**

Solution:
For n = 12 the median position is $\frac{12+1}{2}=6.5$ and, after arranging the data according to size, we find that the sixth and seventh values among

	72	74	75	77	78	79	82	85	86	90	93	94

are 79 and 82. Hence the median is $\frac{79+82}{2}=80.5$. For the six values below 80.5 the median position is $\frac{6+1}{2}=3.5$, and since the third and fourth values are 75 and 77, $Q_1=\frac{75+77}{2}=76$ Counting from the other end, the third and fourth values are 90 and 86, and $Q_3=\frac{90+86}{2}=88$. As can be seen from the data and also from figure 4.1, there are three values below 76, three values between 76 and 80.5, three values between 80.5 and 88, and three values above 88.

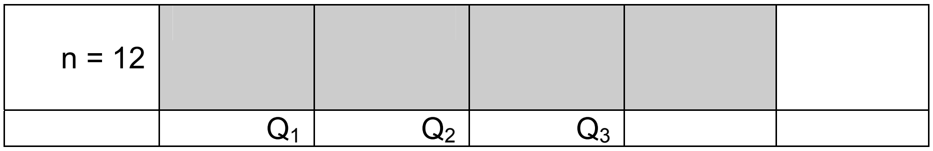


Figure 4.1: Three quartiles of Example 3.16

Everything worked nicely in this example, but n = 12 happened to be a multiple of 4, which raises the question whether our definition of Q_1 and Q_3 will work also when this is not the case.

Example **17**

Example (17)
Suppose that the city where the high temperature was 77 failed to report, so that we are left with the following 11 numbers arranged according to size:

	72	74	75	78	79	82	85	86	90	93	94

Solution 17

Solution:

For n = 11 the median position is $\frac{11+1}{2}=6$ and, referring to the preceding data, which are already arranged according to size, we find that the median is 82. For the five values below 82 the median position is $\frac{5+1}{2}=3$, and Q_1, the third value, equals 75. Counting from the other end, Q_3, the third value, equals 90. As can be seen from the data and also from figure 4.2, there are two values below 75, two values between 75 and 82, two values between 82 and 90, and two values above 90. Again, this satisfies the requirement for the three quartiles, Q_1, Q_2, and Q_3.

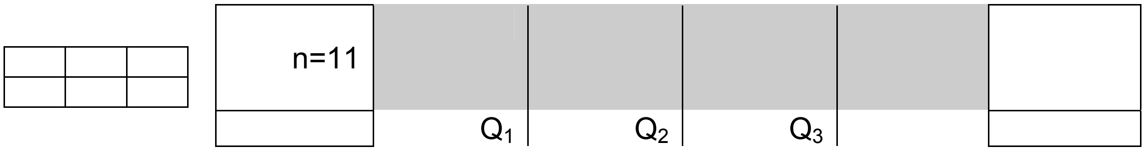


Figure 4.2: Three quartiles of Example 3.17

If some of the values are alike, we modify the definitions of Q_1 and Q_3 by replacing "less than the median" by "to the left of the median position" and "greater than the median" by "to the right of the median position". For instance, for Example (12), we already showed that the median, the fifth value, equals 4. Now, the median of the four values to the left of the median position, Q_1, equals 3, and the median of the four values to the right of the median position, Q_3, equals 4.

Quartiles are not meant to be descriptive of the "middle" or "center" of a set of data, and we have given them here mainly because, like the median, they are fractiles and they are determined in more or less the same way. The midquartile $\frac{Q_1+Q_3}{2}$ *has been used on occasion as another measure of central location.*

The information provided by the median, the quartiles Q_1 and Q_3, and the smallest and largest values is sometimes presented in the form of a box plot. Originally referred to somewhat whimsically as ***a box-and-whisker plot,*** such a display consists of a rectangle that extends from Q_1 to Q_3, lines drawn from the smallest value to Q_1 and from Q_3 to the largest value, and a line at the median that divides the rectangle into two parts. In practice, box plots are sometimes embellished with other features, but the simple form shown here is adequate for most purposes.

Example 18

Example (18)

In Example 15 we used the following double-stem display to show that the median of the room occupancy data, originally given before is 53.5:

2	3	57
6	4	0023
13	4	5666899
(3)	5	234
14	5	56789
9	6	1224
5	6	9
4	7	23
2	7	8
1	8	1

(a) Find the smallest and largest values.
(b) Find Q_1 and Q_3.
(c) Draw a box plot.

Solution
18

Solution:

a) As can be seen by inspection the smallest value is 35 and the largest value is 81.

b) For n = 30 the median position is $\frac{30+1}{2} = 15.5$ and, hence, for the 15 values below 53.5 the median position is $\frac{15+1}{2} = 8$. It follows that Q_1 the eighth value, is 46. Similarly, Q_3, the eighth value from the other end, is 62.

c) Combining all this information, we obtain the box plot shown in Figure 4.3.

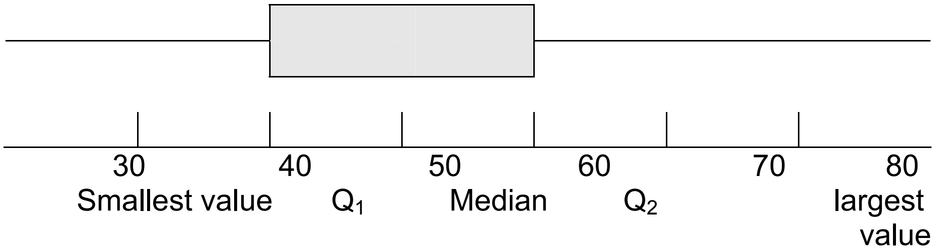


Figure 4.3: Box plot of room occupancy data

Box plots can also be constructed with appropriate computer software or a graphing calculator. Using same data as in Example 18, we reproduced the one shown in Figure 4.4 from the display screen of a TI-83 graphing calculator.

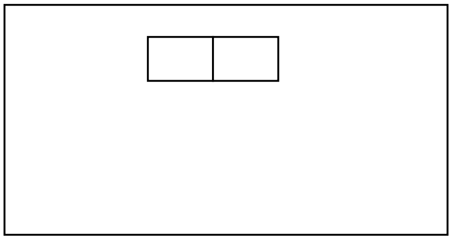


Figure 4.4: Box plot of room occupancy data (TI-83 graphing)

The Mode

4.6 The Mode

Another measure that is sometimes used to describe the middle or center of a set of data is the mode, which is defined simply as the value that occurs with the highest frequency and more than once. **Its two main advantages are that it requires no calculations, only counting, and it can be determined for qualitative, or nominal, data.**

***Example* 19**

Example (19)

The 20 meetings of a square dance club were attended by 22, 24, 23, 24, 27, 25, 20, 24, 26, 28, 26, 23, 21, 24, 24, 25, 23, 28, 26, and 25 of its members. Find the mode.

***Solution* 19**

Solution:
Among these numbers, 20, 21, 22, and 27 each occurs once, 28 occurs twice, 23, -25, and 26 each occurs three times; and 24 occurs 5 times. Thus, the modal attendance is 24.

***Example* 20**

Example (20)
In Example 12, we gave the scores of nine golfers on a par-four hole as 2, 3, 3, 3, 4, 4, 4, 4, and 5. Find the mode.

***Solution* 20**

Solution:
Since these data are already arranged according to size, it can easily be seen that 4, which occurs four times, is the modal score.

As we have seen in this chapter, there are various measures of central location that describe the middle of a set of data. What particular "average" should be used in any given situation can depend on many different things and the choice may be difficult to make. Since the selection of statistical descriptions often contains an element of arbitrariness, some persons believe that the magic of statistics can be used to prove nearly anything. Indeed, a famous nineteenth-century British statesman is often quoted as saying that there are three kinds of lies: lies, damned lies, and statistics.

The Description of Grouped Data

4.7 The Description of Grouped Data

In the past, considerable attention was paid to the description of grouped data, because it usually simplified matters to group large sets of data before calculating various statistical measures. This is no longer the case, since the necessary calculations can now be made in a matter of seconds with the use of computers or even hand-held calculators. Nevertheless, we shall devote this section to the description of grouped data, since many kinds of data (for example, those reported in government publications) are available only in the form frequency distributions.

As we have already seen, the grouping of data entails some loss of information. Each item loses its identity, so to speak; we know only how many values there are in each class or in each category. This means that we shall have to be satisfied with approximations. *Sometimes we treat our data as if all the values falling into a class were equal to the corresponding class mark, and we shall do so to define the mean of a frequency distribution. Sometimes we treat our data as if all the values falling into a class are spread evenly throughout the corresponding class interval, and we shall do so to define the median of a frequency distribution. In either case, we get good approximations since the resulting errors will tend to average out.*

To give a general formula for the mean of a distribution with k classes, let us denote the successive class marks by x_1, x_2 ..., and x_k, and the corresponding class frequencies by f_1, f_2, …, and f_k. Then, **the sum of all the measurements is approximated by:**

$x_1 \cdot f_1 + x_2 \cdot f_2 + ... x_k \cdot f_k + = \sum x \cdot f$ and **the mean of the distribution is given by**

Mean of grouped data	$\bar{x} = \dfrac{\sum x \cdot f}{n}$

Here n is *the size of the sample, f_1 + f_2 + ...+ f_3, and to write a corresponding formula for the mean of a population we substitute μ for $\bar{x}$ and N for n.*

Example **21**

Example (21)
Find the mean for the distribution of the waiting times between eruptions of Old Faithful Geyser that was obtained in Example before.

Solution **21**

Solution:
To get $\sum x \cdot f$, we perform the calculations shown in the following table, where the first column contains the class marks, the second column consists of the class frequencies shown on page 24, and the third column contains the products x. f:

Class Mark x	Frequency f	$x \cdot f$
34.5	2	69.0
44.5	2	89.0
54.5	4	218.0
64.5	19	1,225.5
74.5	24	1,788.0
84.5	39	3,295.5
94.5	15	1,417.5
104.5	3	313.5
114.5	2	229.0
	110	8,645.0

Then, substitution into the formula yields $\overline{x} = \frac{8{,}645.0}{110} = 78.59$ rounded to two decimals.

To check on the grouping error, namely, the error introduced by replacing each value within a class by the corresponding class mark, we can calculate $\overline{x}$ for the original data, or use the same computer software. Having already entered the data, we simply change the command to MEAN C1 and we get 78.273, or 78.27 rounded to two decimals. Thus, the grouping error is only 78.59 -78.27 = 0.32, which is fairly small.

When dealing with grouped data, we can determine most other statistical measures besides the mean, but we may have to make different assumptions and / or modify the definitions. For instance, for the median of a distribution we use the assumption (namely, the assumption that the values within a class are spread evenly throughout the corresponding class interval). Thus, with reference to a histogram

The median of a distribution is such that the total area of the rectangles to its left equals the total area of the rectangles to its right.

To find the dividing line between the two halves of a histogram (each of which represents $\frac{n}{2}$ of the items grouped), we must count $\frac{n}{2}$ of the items starting at either end of the distribution. How this is done is illustrated by the following Example and Figure 4.5.

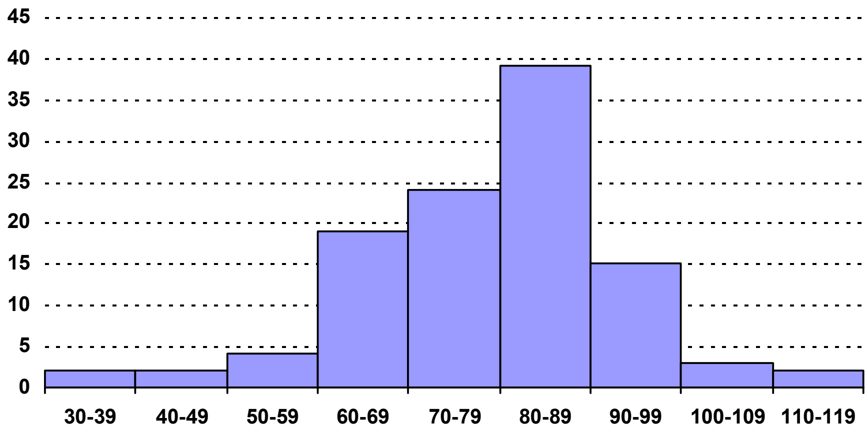


Figure 4.5: Median of distribution of eruptions of old faithful

Example 22

Example (22)

Find the median of the distribution of the waiting times between eruptions of Old Faithful.

Solution 22

Solution:

Since $\frac{n}{2}=\frac{110}{2}=55$, we must count 55 of the items starting at either end. Starting at the bottom of the distribution (that is, beginning with the smallest values), we find that 2 + 2 + 4 + 19 + 24 = 51 of the values fall into the first five classes. Therefore, we must count 55 -51 = 4 more values from among the values in the sixth class. Based on the assumption that the 39 values in the sixth class are spread evenly throughout that class, we accomplish this by adding $\frac{4}{39}$ of the class interval of 10 to 79.5, which is its lower class boundary. This yields:

$$\tilde{x} = 79.5 + \frac{4}{39}\cdot 10 = 80.53$$

Rounded to two decimals.

In general, *if L is the lower boundary of the class into which the median must fall, f is its frequency, c is its class interval, and j is the number of items we still lack when we reach L, then the median of the distribution is given by*

Median of grouped data	$\hat{x} = L + \frac{j}{f}\cdot c$

If we prefer, we can find the median of a distribution by starting to count at the other end (beginning with the largest values) and subtracting an appropriate fraction of the class interval from the upper boundary of the class into which the median must fall.

Example (23)

Example **23**

Use this alternative approach to find the median of the waiting times between eruptions of Old Faithful.

Solution:

Solution **23**

Since 2 + 3 + 15 = 20 of the values fall above 89.5, we need 50 – 20 = 35 of the 39 values in the next class to reach the median. Thus, we write

$\tilde{x} = 89.5 - \frac{35}{39} \cdot 10 = 80.35$ and the result is, of course, the same.

Note that the median of a distribution can be found regardless of whether the class intervals are all equal. In fact, it can be found even when either or both classes at the top and at the bottom of a distribution are open, so long as the median does not belong to either class.

The method by which we found the median of a distribution can be also used to determine other fractiles. For instance Q_1 and Q_3 are defined for grouped data so that 25% of the total area of the rectangles of the histogram lies to the left of Q_1 and 25% lies to the right of Q_3. Similarly, the nine deciles (which are intended to divide a set of data into ten equal parts) are defined for grouped data so that 10 percent of the total area of the rectangles of the histogram lies to the left of D_1, 10 percent lies between D_1 and D_2, …, and 10 percent lies to the right of D_9. And finally, the ninety-nine percentiles (which are intended to divide a set of data into a hundred equal parts) are defined for grouped data so that 1 percent of the total area of the rectangles of the histogram lies to the left of P_1, 1 percent lies between P_1 and P_2, … and 1 percent lies to the right of P_{99}. Note that D_s and P_{50} are equal to the median and that P_{25} equals Q_1 and P_{75} equals Q3.

Example (24)

Example **24**

Find Q_1 and Q_3 for the distribution of the waiting times between eruptions of Old Faithful.

Solution:

Solution **24**

To find Q_1 we must count $\frac{110}{4} = 27.5$ of the items starting at the bottom of the distribution. Since there are 2+2+4+19 = 27 values in the first four classes, we must count 27.5 – 27 = 0.5 of the 24 values in the fifth class to reach Q_1. This yields:

$$Q_1 = 69.5 + \frac{0.5}{24}.10 \approx 69.71$$

Since 2+3+15=20 of the values fall into the last three classes, we must count 27.5 - 20 = 7.5 of the 39 values in the next class to reach Q_3. Thus, we write

$$Q_3 = 89.5 + \frac{7.5}{39}.10 \approx 87.58$$

Example **25**

Example (25)
Find D_2 and P_8 for the distribution of the waiting times between eruptions of Old Faithful.

Solution **25**

Solution:

To find D_2 we must count $110 \cdot \frac{2}{10} = 22$ of the items starting at the bottom of the distribution. Since there are 2+2+4=8 values in the first three classes, we must count 22-8= 14 of the 19 values of the fourth class to reach D_2. This yields

$$D_2 = 59.9 + \frac{14}{19} \cdot 10 \approx 66.87$$

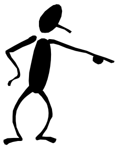
Since 2+3+15=20 of the values fall into the last three classes, we must count 22-20 = 2 of the 39 values in the next class to reach P_8. Thus, we write

$$P_8 = 89.5 + \frac{2}{39} \cdot 10 \approx 88.99$$

Note that when we determine a fractile of a distribution, the number of items we have to count and the quantity j in the formula on page 73 need not be a whole number.

Chapter 5: Summarizing Data: Measures of Variation

Introduction



One aspect of most sets of data is that the values are not all alike; indeed, the extent to which they are unalike, or vary among themselves, is of basic importance in statistics. Consider the following examples:

In a hospital where each patient's pulse rate is taken three times a day, that of patient *A* is 72, 76, and 74, while that of patient B is 72, 91, and 59. The mean pulse rate of the two patients is the same, 74, but observe the difference in variability. Whereas patient A's pulse rate is stable, that of patient B fluctuates widely.

A supermarket stocks certain 1-pound bags of mixed nuts, which on the average contain 12 almonds per bag. If all the bags contain anywhere from 10 to 14 almonds, the product is consistent and satisfactory, but the situation is quite different if some of the bags have no almonds while others have 20 or more.

Measuring variability is of special importance in statistical inference. Suppose, for instance, that we have a coin that is slightly bent and we wonder whether there is still a fifty-fifty chance for heads. What If we toss the con 100 times and get 28 heads and 72 tails? Does the shortage of heads-only 28 where we might have expected 50-imply that the count is not "fair?" To answer such questions we must have some idea about the magnitude of the fluctuations, or variations, that are brought about by chance when coins are tossed 100 times.

We have given these three examples to show the need for measuring the extent to which data are dispersed, or spread out; the corresponding measures that provide this information are called measures of variation. In Sections 1 through 3 we present the most widely used measures of variation and some of their special applications. Some statistical descriptions other than measures of location and measures of variation are discussed in Section 4.5.

The Range

5.1 The Range

To introduce a simple way of measuring variability, let us refer to the first of the three examples cited previously, where the pulse rate of patient A varied from 72 to 76 while that of patient B varied from 59 to 91. These extreme (smallest and largest) values are indicative of the variability of the two sets of data, and just about the same information is conveyed if we take the differences between the respective

extremes. So, let us make the following definition:
The range of 8 set of data is the difference between the largest value and the smallest.

For patient A the pulse rates had a range of 76 -72 = 4 and for patient B they had a range of 91 -59 = 32, and for the waiting times between eruptions of Old Faithful in Example 2.4, the range was 118 -33 = 85 minutes.

Conceptually, the range is easy to understood, its calculation is very easy, and there is a natural curiosity about the smallest and largest values. Nevertheless, it is not a very useful measure of variation - its main shortcoming being that it does not tell us anything about the dispersion of the values that fall between the two extremes. For example, each of the following three sets of data

Set A: 5 18 18 18 18 18 18 18 18 18

Set B: 5 5 5 5 5 18 18 18 18 18

Set C: 5 6 8 9 10 12 14 15 17 18

has a range of 18 -5 = 13, but their dispersions between the first and last values are totally different

In actual practice, the range is used mainly as a "quick and easy" measure of variability; for instance, in industrial quality control it is used to keep a close check on raw materials and products on the basis of small samples taken at regular intervals of time.

Whereas, the range covers all the values in a sample, a similar measure of variation covers (more or less) the middle 50 percent. It is the inter quartile range: Q_3 –Q_1, where Q_1 and Q_3 may be defined as before. For instance, for the twelve temperature readings in Example 3.16 we might use 88 -76 = 12 and for the grouped data in Example 3.24 we might use 87.58 -69.71 = 17.87. Some statisticians also use the semi-inter quartile range $\frac{1}{2}(Q_3 - Q_1)$, which is sometimes referred to as the quartile deviation.

The Variance and The Standard Deviation

5.2 The Variance and the Standard Deviation

To define the standard deviation, by far the most generally useful measure of variation. Let us observe that the dispersion of a set of data is small if the values are closely bunched about their mean, and that it is large if the values are scattered widely about their mean. Therefore, it would seem reasonable to measure the variation of a set of data in terms of the amounts by which the values deviate from their mean. If a set of numbers

x_1, x_2, x_3, … and x_n

constitutes a sample with the mean $\overline{x}$, *then the differences*

$$x_1 - \overline{x},\ x_2 - \overline{x},\ x_3 - \overline{x},\ \ldots,\ \text{and}\ x_n - \overline{x}$$

are called the deviation from the mean, and we might use *their average (that is, their mean) as a measure of the variability of the sample.* Unfortunately, this will not do. *Unless the x's are all equal, some of the deviations from the mean will be positive, some will be negative, the sum of deviations from the mean,* $\sum(x - \overline{x})$*, and hence also their mean, is always equal to zero.*

Since we are really interested in the magnitude of the deviations, and not in whether they are positive or negative, we might simply ignore the signs and define a measure of variation in terms of the absolute values of the deviations from the mean. Indeed, if we add the deviations from the mean as if they were all positive or zero and divide by n, we obtain the statistical measure that is called the mean deviation. This measure has intuitive appeal, but because the absolute values if leads to serious theoretical difficulties in problems of inference, and it is rarely used.

An alternative approach is to work with the squares of the deviations from the mean, as this will also eliminate the effect of signs. Squares of real numbers cannot be negative; in fact, squares of the deviations from a mean are all positive unless a value happens to coincide with the mean. Then, if we average the squared deviation from the mean and take the square root of the result (to compensate for the fact that the deviations were squared), we get

$$\sqrt{\frac{\sum(x - \overline{x})^2}{n}}$$

and this is how, traditionally, the standard deviation used to be defined. **Expressing literally what we have done here mathematically, it is also called the root-mean-square deviation.**

Nowadays*, it is customary to modify this formula by dividing the sum of the squared deviations from the mean by n-1 instead of n.* Following this practice, which will be explained later, let us define the sample standard deviation, denoted by s, as

Sample Standard Deviation

$$\boxed{s = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}}$$

And its square, the sample variance, as

Sample Variance

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

These formulas for the standard deviation and the variance apply to samples, but if we substitute μ for $\overline{x}$ and N for n, we obtain analogous formulas for the standard deviation and the variance of a population. It is customary to denote the population standard deviation by σ (sigma, the Greek letter for lower case) when dividing by N, and by S when dividing by N-1. Thus, for σ we write

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

and the population variance is σ^2.

Ordinarily, **the purpose of calculating a sample statistics (such as the mean, the standard deviation, or the variance) is to estimate the corresponding population parameter.** If we actually took many samples from a population that has the mean μ, calculated the sample means $\overline{x}$, and then averaged all these estimated of μ, we should find that their average is very close to μ. However, if we calculated the variance of each sample by means of the formula $\frac{\sum(x-\overline{x})^2}{n}$ and then averaged all these supposed estimates of σ^2. Theoretically, it can be shown that we can compensate for this by dividing by n-1 instead of n in the formula for s^2. Estimators, having the desirable property that their values will, on the average, equal the quantity they are supposed to estimate are said to be unbiased; otherwise, they are said to be biased. So, we say that $\overline{x}$ is an unbiased estimator of the population mean μ and that s^2 is an unbiased estimator of the population variance σ^2. It does not follow from this that s is also an unbiased estimator of σ , but when n is large the bias is small and can usually be ignored.

In calculating the sample standard deviation using the formula by which it is defined, we must (1) find $\overline{x}$, (2) determine the n deviations from the mean $x-\overline{x}$, (3) square these deviations, (4) add all the squared deviations, (5) divide by n-1, and (6) take the square root of the result arrived at in step 5. In actual practice, this formula is rarely used – there are various shortcuts – but we shall illustrate it here to emphasize what is really measured by a standard deviation.

Example 1

Example (1)

A bacteriologist found 8, 11, 7, 13, 10, 11, 7, and 9 microorganism of a certain kind in eight cultures. Calculate s.

Solution 1

Solution:
First calculating the mean, we get

$$\overline{x} = \frac{8+11+7+13+10+11+7+9}{8} = 9.5$$

and then the work required to find $\sum(x-\overline{x})^2$ may be arranged as in the following table:

x	$x-\overline{x}$	$(x-\overline{x})^2$
8	-1.5	2.25
11	1.5	2.25
7	-2.5	6.25
13	3.5	12.25
10	0.5	0.25
11	1.5	2.25
7	-2.5	6.25
9	-0.5	0.25
	0.0	32.00

Finally, dividing 32.00 by 8 -1 = 7 and taking the square root (using a simple handheld calculator), we get

$$s = \sqrt{\frac{32.00}{7}} \approx \sqrt{4.57} = 2.14$$

rounded to two decimals

Note in the preceding Table that the total for the middle column is zero; since this must always be the case; it provides a convenient check on the calculations.

It was easy to calculate s in this Example because the data were whole numbers and the mean was exact to one decimal. Otherwise, the calculations required by the formula defining s can be quite tedious, and, unless we can get *s* directly with a statistical calculator or a computer, it helps to use the formula

Computing formula for the sample standard deviation

$$\boxed{s = \sqrt{\frac{S_{xx}}{n-1}} \text{ where } S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}}$$

Example 2

Example (2)
Use this computing formula to rework Example (1).

Solution 2

Solution:
First we calculate $\sum x$ and $\sum x^2$, getting

$$\sum x = 8+11+7+13+10+11+7+9 = 76$$

and

$$\sum x^2 = 64 + 121 + 49 + 169 + 100 + 121 + 49 + 81 = 754$$

Then, substituting these totals and n = 8 into the formula for S_{xx}, and n-1 = 7 and the value obtained for S_{xx} into the formula for s, we get

$$S_{xx} = 754 - \frac{(76)^2}{8} = 32$$

and, hence, $s = \sqrt{\frac{32}{7}} = 2.14$ rounded to two decimals. This agrees, as it should, with the result obtained before.

As should have been apparent from these two examples, **the advantage of the computing formula is that we got the result without having to determine $\bar{x}$ and work with the deviations from the mean.** Incidentally, the computing formula can also be used to find σ with the n in the formula for S_{xx} and the n -1 in the formula for s replaced by N.

In the introduction to this chapter we gave three examples in which knowledge about the variability of the data was of special importance. This is also the case when we want to compare numbers belonging to different sets of data. To illustrate, suppose that the final examination in a French course consists of two parts, vocabulary and grammar, and that a certain student scored 66 points in the vocabulary part and 80 points in the grammar part. At first glance it would seem that the student did much better in grammar than in vocabulary, but suppose that all the students in the class averaged 51 points in the vocabulary part with a standard deviation of 12, and 72 points in the grammar part with a standard deviation of 16. Thus, we can argue that the student's score in the vocabulary part is $\frac{66-51}{12} = 1.25$ standard deviations above the average for the class, while her score in the grammar part is only $\frac{80-72}{16} = 0.50$ standard deviation above the average for the class. Whereas the original scores cannot be meaningfully compared, these new scores, expressed in terms of standard deviations, can. Clearly, the given student rates much higher on her command of French vocabulary than on her knowledge of French grammar, compared to the rest of the class.

What we have done here consists of converting the grades into standard units or z-scores. It general, if x is a measurement belonging to a set of data having the mean $\bar{x}$ (or μ) and the standard deviation s (or σ), then its value in standard units, denoted by z, is

Formula for Converting to Standard Units	$z = \frac{x - \bar{x}}{s}$ or $z = \frac{x - \mu}{\sigma}$

Depending on whether the data constitute a sample or a population. In these units, z tells us how many standard deviations a value lies above or below the mean of the set of data to which it belongs. Standard units will be used frequently in application.

Example 3

Example (3)

Mrs. Clark belongs to an age group for which the mean weight is 112 pounds with a standard deviation of 11 pounds, and Mr. Clark, her husband, belongs to an age group for which the mean weight is 163 pounds with a standard deviation of 18 pounds. If Mrs. Clark weighs 132 pounds and Mr. Clark weighs 193 pounds, which of the two is relatively more overweight compared to his / her age group?

Solution 3

Solution:
Mr. Clark's weight is 193 -163 = 30 pounds above average while Mrs. Clark's weight is "only" 132 -112 = 20 pounds above average, yet in standard units we get $\frac{193-163}{18} \approx 1.67$ for Mr. Clark and $\frac{132-112}{11} \approx 1.82$ for Mrs. Clark.

Thus, relative to them age groups Mrs. Clark is somewhat more overweight than Mr. Clark.

A serious disadvantage of the standard deviation as a measure of variation is that it depends on the units of measurement. For instance, the weights of certain objects may have a standard deviation of 0.10 ounce, but this really does not tell us whether it reflects a great deal of variation or very little variation. If we are weighing the eggs of quails, a standard deviation of 0.10 ounce would reflect a considerable amount of variation, but this would not be the case if we are weighing, say, 100-pound bags of potatoes. What we need in a situation like this is a measure of relative variation such as the coefficient of variation, defined by the following formula:

Coefficient of variation	$V = \frac{s}{\bar{x}} \cdot 100\%$ or $V = \frac{\sigma}{\mu} \cdot 100\%$

The coefficient of variation expresses the standard deviation as a percentage of what is being measured, at least on the average.

Example 4

Example (4)
Several measurements of the diameter of a ball bearing made with one micrometer had a mean of 2.49mm and a standard deviation of 0.012mm, and several measurements of the unstretched length of a spring made with another micrometer had a mean of 0.75 in. with a standard deviation of 0.002 in. Which of the two micrometers is relatively more precise?

Solution

4

Solution:
Calculating the two coefficients of variation, we get

$$\frac{0.012}{2.49}\cdot 100\% \approx 0.48\% \qquad \text{and} \qquad \frac{0.002}{0.75}\cdot 100\% \approx 0.27\%$$

Thus, the measurements of the length of the spring are relatively less variable, which means that the second micrometer is more precise.

The Description of Grouped Data

5.3 The Description of Grouped Data

As we saw in before, the grouping of data entails some loss of information. Each item has lost its identity and we know only how many values there are in each class or in each category. *To define the standard deviation of a distribution we shall have to be satisfied with an approximation and, as we did in connection with the mean, we shall treat our data as if all the values falling into a class were equal to the corresponding class mark. Thus, letting x_1, $x_,$ ..., and x_k denote the class marks, and f_1, f_2, ..., and f_k the corresponding class frequencies, we approximate the actual* ***sum of all the measurements or observations*** *with*

$\Sigma x.f = x_1f_1 + x_2f_2 + \ldots.. x_kf_k$ and ***the sum of their squares*** *with*

$$\sum x^2 \cdot f = x^2 1f_1 + x_2^2 f_2 + \ldots x_k^2 f_k$$

Then, *we write the computing formula for the* ***standard deviation of grouped sample data*** *as*

$$S = \sqrt{\frac{S_{xx}}{n-1}} \quad \text{where} \quad S_{xx} = \sum x^2 \cdot f - \frac{\left(\sum x \cdot f\right)^2}{n}$$

Which is very similar to the corresponding computing formula for s for ungrouped data. To obtain a corresponding computing formula for σ, we replace n by N in the formula for S_{xx} and n -1 by N in the formula for s.

When the class marks are large numbers or given to several decimals, we can simplify things further by using the coding suggested below. **When the class intervals are all equal,** and only then, **we replace the class marks with consecutive integers, preferably with 0 at or near the middle of the distribution.** Denoting the coded class marks by the letter u, we then calculate S_{xx} and substitute into the formula

$$S_u = \sqrt{\frac{S_{uu}}{n-1}}$$

This kind of coding is illustrated by Figure 5.1, where we find that if u varies (is increased or decreased) by 1, the corresponding value of x varies (is increased or decreased) by the class interval c. Thus, to change s_u from the u-scale to the original scale of measurement, the x-scale, we multiply it by c.

x-2c	x-c	x	x+c	x+2c	x-scale
-2	-1	0	1	2	u-scale

Figure 5.1: Coding the class marks of a distribution

Example (5)

With reference to the distribution of the waiting times between eruptions of Old Faithful shown in before, calculate its standard deviation

(a) Without coding;

(b) With coding.

Solution:

(a)	x	F	x.f	x^2.f
	34.5	2	69	2,380.5
	44.5	2	89	3,960.5
	54.5	4	218	11,881
	64.5	19	1,22.5	79,044.75
	74.5	24	1,788	133,206
	84.5	39	3,295.5	278,469.75
	94.5	15	1,417.5	133,760.75
	104.5	3	313.5	32,760.75
	114.5	2	229	26.220.5
		110	8,645	701,877.5

so that

$$S_{xx} = 701{,}877.5 - \frac{(8.645)^2}{110} = 22{,}4593.1$$

and

$$s = \sqrt{\frac{22{,}459.1}{109}} \approx 14.35$$

(b)	u	F	u.f	U^2.f
	-4	2	-8	32
	-3	2	-6	18
	-2	4	-8	16
	-1	19	-19	19
	0	24	0	0
	1	39	39	39
	2	15	30	60
	3	3	9	27
	4	2	8	32
		110	45	243

so that

$$S_{uu} = 243 - \frac{(45)^2}{110} = 224.59$$

and

$$s_u = \sqrt{\frac{224.59}{109}} \approx 1.435$$

Finally, s = 10(1.435) = 1435, which agrees, as it should, with the result obtained in part (a). This clearly demonstrates how the coding simplified the calculations.

Some Further Descriptions

5.4 Some Further Descriptions

So far we have discussed only statistical descriptions that come under the general heading of measures of location or measures of variation. Actually, there is no limit to the number of ways in which statistical data can be described, and statisticians continually develop new methods of describing characteristics of numerical data that are of interest in particular problems. In this section we shall consider briefly the problem of describing the overall shape of a distribution.

*Although frequency distributions can take on almost any shape or form, most of the distributions we meet in practice can be described fairly well by one or another of few standard types. Among these, foremost in importance is **the aptly described symmetrical bell-shaped distribution**.* The two distributions shown in Figure 5.2 can, by a stretch of the imagination, be described as bell shaped, but they are not symmetrical. Distributions like these, having a "tail" on one side or the other, are said to be skewed; if the tail is on the left we say that they are negatively skewed and if the tail is on the right we say that they are positively skewed. Distributions of incomes or wages are often positively skewed because of the presence of some relatively high values that are not offset by correspondingly low values.

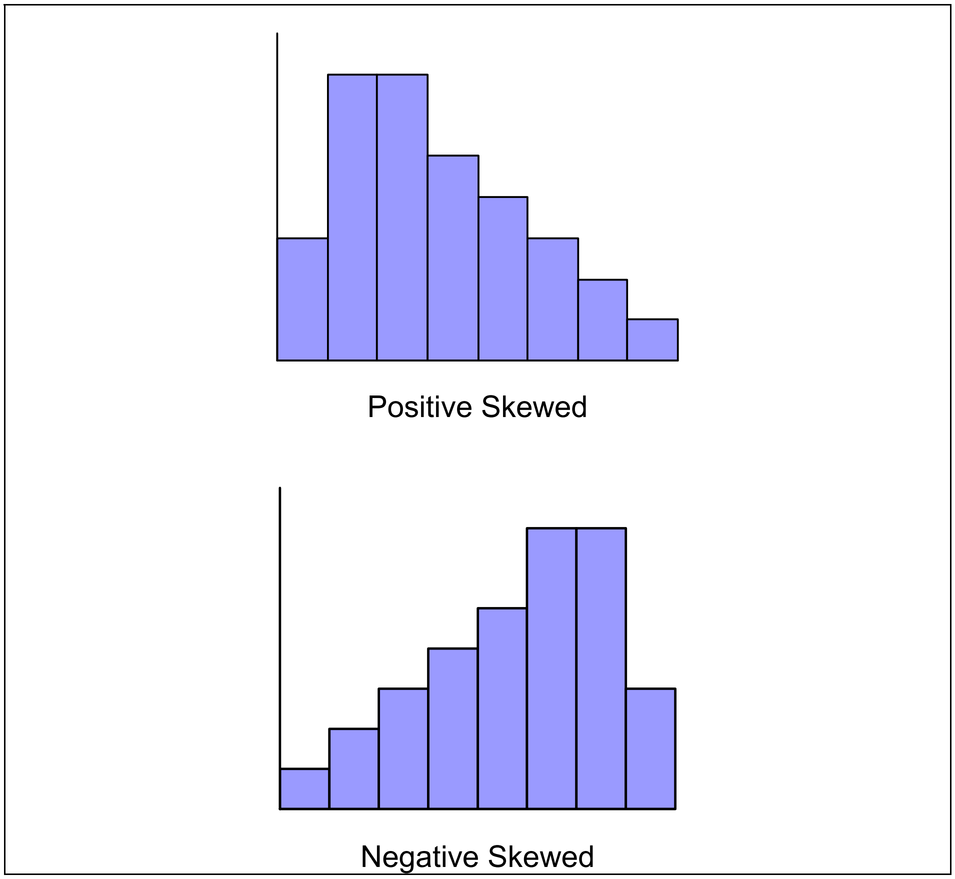


Figure 5.2: Skewed distributions.

The concepts of symmetry and skewness apply to any kind of data, not only distributions. Of course, for a large set of data we may just group the data and draw and study a histogram, but if that is not enough, we can use anyone of several statistical measures of skewness. A relatively easy one is based on the fact that when there is perfect symmetry, the mean and the median will coincide. When there is positive skewness and some of the high values are not offset by correspondingly low values, as shown in Figure 5.3, the mean will be greater than the median; when there is a negative skewness and some of the low values are not offset by correspondingly high values, the mean will be smaller than the median.

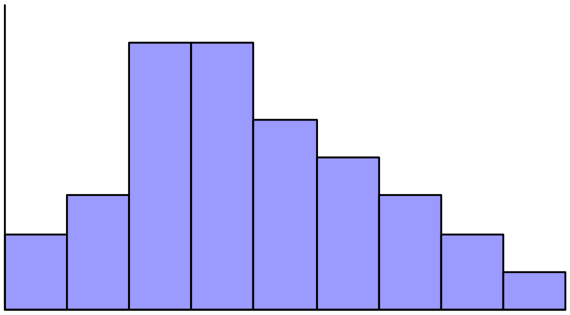


Figure 5.3: Mean and median of positively skewed distribution

This relationship between the median and the mean can be used to define a relatively simple measure of skewness, called ***the Pearsonian coefficient of skewness****. It is given by*

Pearsonian coefficient of skewness

$$\boxed{\mathrm{SK} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}}$$

For a perfectly symmetrical distribution, such the mean and the median coincide and SK = 0. In general, values of the Pearsonian coefficient of skewness must fall between -3 and 3, and it should be noted that division by the standard deviation makes SK independent of the scale of measurement.

Example

Example (6)

Calculate *SK* for the distribution of the waiting times between eruptions of Old Faithful, using the results of Examples 3.21, 3.22, and 4.7, where we showed $\bar{x} = 78.59, \tilde{x} = 80.53$, and s = 14.35.

Solution

Solution:

Substituting these values into the formula for SK, we get

$$\mathrm{SK} = \frac{3(78.59 - 80.53)}{14.35} \approx -0.41$$

Which shows that there is a definite, though modest, negative skewness. This is also apparent from the histogram of the distribution, shown originally and here again in Figure 5.4, reproduced from the display screen of a TI-83 graphing calculator.

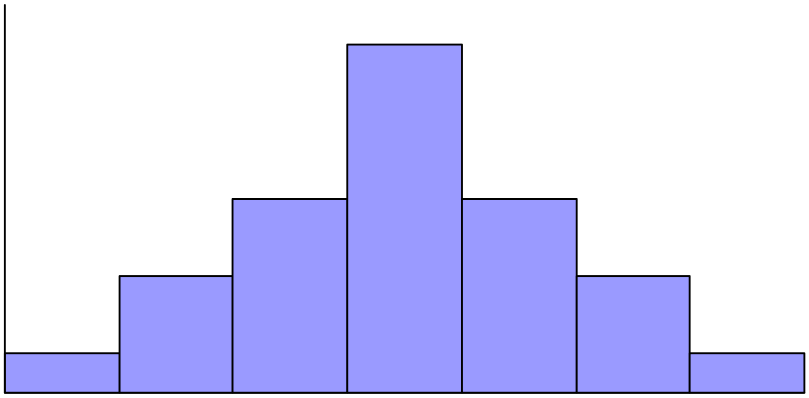


Figure 5.4: Histogram of distribution of waiting times between eruptions of old faithful

When a set of data is so small that we cannot meaningfully construct a histogram, a good deal about its shape can be learned from a box plot (defined originally). *Whereas the Pearsonian coefficient is based on the difference between the mean and the median, with a box plot we* *judge the symmetry or skewness of a set of data on the basis of the position of the median relative to the two quartiles, Q_1 and Q_3. In particular, if the line at the median is at or near the center of the box, this is an indication of the symmetry of the data; if it is appreciably to the left of center, this is an indication that the data are positively skewed; and if it is appreciably to the right of center, this is an indication that the data are negatively skewed. The relative length of the two "whiskers," extending from the smallest value to Q_l and from Q_3 to the largest value, can also be used as an indication of symmetry or skewness.*

Example 7

Example (7)
Following are the annual incomes of fifteen CPAs in thousands of dollars: 88, 77, 70, SO, 74, 82, 85, 96, 76, 67, 80, 75, 73, 93, and 72. Draw a box plot and use it to judge the symmetry or skewness of the data.

Solution 7

Solution:
Arranging the data according to size, we get

67	70	72	73	74	75	76	77
80	80	82	85	88	93	96	

It can be seen that the smallest value is 67; the largest value is 96; the median is the eighth value from either side, which is 77; Q_1 is the fourth value from the left, which is 73; and Q_3 is the fourth value from the right, which is 85. All this information is summarized by the MINITAB printout of the box plot shown in Figure 5.5. As can be seen, *there is a strong indication that the data are positively skewed. The line at the median is well to the left of the center of the box and the "wisker" on the right is quite a bit longer than the one on the left.*

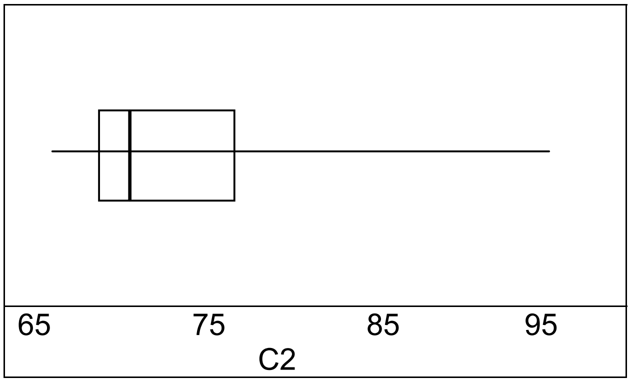


Figure 5.5: Box plot of incomes of the CPAs.

Besides the distributions we have discussed in this section, two others sometimes met in practice are the reverse **J-shaped and U-shaped distributions shown in Figure 5.6.** As can be seen from this figure, *the names of these distributions literally describe their shapes.* Examples of such distribution may be found in real life situations.

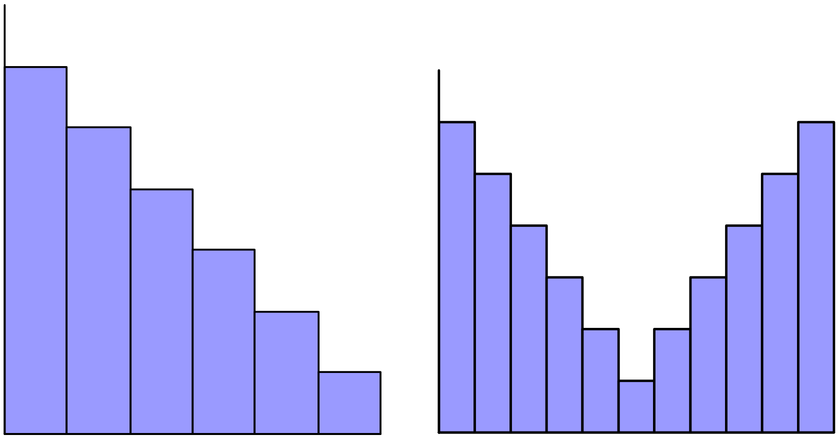


Figure 5.6: Reverse J-shaped and U-shaped distributions

Chapter 6: Simple Linear Regression and Correlation

Introduction

6.1 Introduction

This objective of this chapter is to analyze the relationship among quantitative variables. Regression analysis is used to predict the value of one variable on the basis of other variables. This technique may be the most commonly used statistical procedure because, as you can easily appreciate, almost all companies and government institutions forecast variables such as product demand, interest rates, inflation rates, prices of raw materials, and labor costs by using it.

The technique involves developing a mathematical equation that describes the relationship between the variable to be forecast, which is called the dependent variable, and variables that the statistician believes are related to the dependent variable. *The dependent variable is denoted y, while the related variables are called independent variables and are denoted x_1, x_2, …x_k (where k is the number of independent variables).*

If we are interested only in determining whether a relationship exists, we employ correlation analysis. We have already introduced this technique. We presented the graphical method to describe the association between two quantitative variables - the scatter diagram. We introduced the coefficient of correlation and covariance.

Because regression analysis involves a number of new techniques and concepts. In this chapter, we present techniques that allow us to determine the relationship between only two variables.

Here are three examples of regression analysis.

Example

Example (1)
The product manager in charge of a particular brand of children's breakfast cereal would like to predict the demand for the cereal during the next year. In order to use regression analysis, she and her staff list the following variables as likely to affect sales.

Price of the product
Number of children 5 to 12 years of age (the target market)
Price of competitor's products
Effectiveness of advertising (as measured by advertising exposure)
Annual sales this year
Annual sales in previous years

Example

Example (2)
A gold speculator is considering a major purchase of gold bullion. He would like to forecast the price of gold two years from now (his planning horizon) using regression analysis. In preparation, he produces the following list of independent variables.

Interest rates
Inflation rate
Price of oil
Demand for gold jewelry
Demand for industrial and commercial gold
Dow Jones Industrial Average

Example

Example (3)
A real estate agent wants to more accurately predict the selling price of houses. She believes that the following variables affect the price of a house.

Size of the house (number of square feet)
Number of bedrooms
Frontage of the lot
Condition
Location

In each of these examples, the primary motive for using regression analysis is forecasting. Nonetheless, analyzing the relationship among variables can also be quite useful in managerial decision making. For instance, in the first application, the product manager may want to know how price is related to product demand so that a decision about a prospective change in pricing can be made.

Another application comes from the field of finance. *The capital asset pricing model analyzes the relationship between the returns of a particular stock and the behavior of a stock index. Its function is not to predict the stock's price but to assess the risk of the stock versus the risk of the stock market in general.*

Regardless of why regression analysis is performed, **the next step in the technique is to develop a mathematical equation or model that accurately describes the nature of the relationship that exists between the dependent variable and the independent variables.** *This stage – which is only a small part of the total process – is described in the next section. Only when we're satisfied with the model do we use it to estimate and forecast.*

Model

6.2 Model

The job of developing a mathematical equation can be quite complex, because we need to have some idea about the nature of the relationship between each of the independent variables and the dependent variable. For example, the gold speculator mentioned in

Example 2 needs to know how interest rates affect the price of gold. If he proposes a linear relationship, that may imply that as interest rates rise (or fall), the price of gold will rise or fall. A quadratic relationship may suggest that the price of gold will increase over a certain range of interest rates but will decrease over a different range. Perhaps certain combinations of values of interest rates and other independent variables influence the price in one way, while other combinations change it in other ways. The number of different mathematical models that could be proposed is virtually infinite.

You might have encountered various models in previous courses. For instance, the following represent relationships in the natural sciences.

$E = mc^2$, where E = Energy, m = Mass, and c = Speed of light
$F = ma$, where F = Force, m = Mass, and a = Acceleration
$S = at^2/2$, where S = Distance, t = Time, and a = Gravitational acceleration

In other business courses, you might have seen the following equations.

Profit = Revenue - Costs
Total cost = Fixed cost + (Variable cost X Number of units produced)

The above are all examples of deterministic models, so named because - except for small measurement errors - *such equations allow us to determine the value of the dependent variable (on the left side of the equation) from the value of the independent variables.* In many practical applications of interest to us, deterministic models are unrealistic. For example, is it reasonable to believe that we can determine the selling price of a house solely on the basis of its size? Unquestionably, the size of a house affects its price, but many other variables (some of which may not be measurable) also influence price. What must be included in most practical models is a method that represents the randomness that is part of a real-life process. Such a model is called probabilistic.

To create a probabilistic model, we start with a deterministic model that approximates the relationship we want to model. We then add a random term that measures the error of the deterministic component. Suppose that in Example 3 described above, the real estate agent knows that the cost of building a new house is about $75 per square foot and that most lots sell for, about $25,000. The approximate selling price would be

$$Y=25,000+75x$$

Where y = Selling price and *x* = Size of the house in Square feet. A house of 2,000 square feet would therefore be estimated to sell for y =25,000 + 75(2,000) = 175,000)

We know, however, that the selling price is not likely to be exactly \$175,000. Prices may actually range from \$100,000 to \$250,000. In other words, the deterministic model is not really suitable. To represent this situation properly, we should use the probabilistic model.

$Y = 25{,}000+75x + \in$

Where $\in$ **(the Greek letter epsilon) represents the random term (also called the error variable) – the difference between the actual selling price and the estimated price based on the size of the house.** The random term thus accounts for all the variables, measurable and immeasurable, that are not part of the model. The value of $\in$ will vary from one scale to the next, even if x remains constant. That is, houses of exactly the same size will sell for different prices because of differences in location, selling season, decorations, and other variables.

In the regression analysis, we will present only ***probabilistic models****.* Additionally, to simplify the presentation, *all models will be* ***linear***. In this chapter, we restrict *the number of independent variables to one*. *The model to be used in this chapter is called* ***the first-order linear model*** *– sometimes called the* ***simple linear regression model****.*

First-Order Linear Model

$$y = \beta_o + \beta_1 x + \in$$

where

y= dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line *(defined as the ratio rise/run or change in y/change in x)*
$\in$ = error variable

Figure 6.1 depicts the deterministic component of the model.

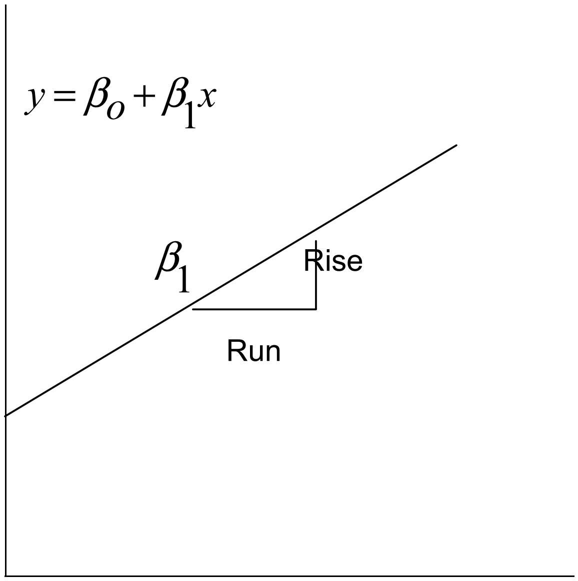


Figure 6.1: First-order linear model, deterministic component

The problem objective addressed by the model is to analyze the relationship between two variables, x and y, both of which must be quantitative. To define the relationship between x and y, we need to know the value of the coefficients of the linear model β_0 *and* β_1*. However, these coefficients are population parameters, which are almost unknown*. In the next section, we discuss how these parameters are estimated.

Least Squares Method

6.3 Least Squares Method

We estimate the parameters β_0 and β_1 in a way similar to the methods used to estimate all the other parameters discussed in this notes. We draw a random sample from the populations of interest and calculate the sample statistics we need. *Because* β_0 *and* β_1 *represent the coefficients of a straight line, their estimators are based on drawing a straight line through the sample data.* To see how this is done, consider the following simple example.

Example

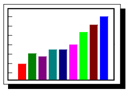
Example (4)
Given the following six observations of variables *x* and y,-determine the straight line that fits these data.

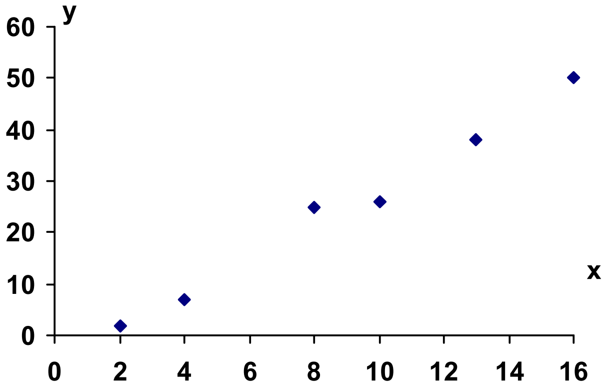
x	2	4	8	10	13	16
y	2	7	25	26	38	50

Solution

Solution:
As a first step we graph the data, as shown in the figure. Recall that this graph is called a scatter diagram. The scatter diagram usually

reveals whether or not a straight line model fits the data reasonably well. Evidently, in this case a linear model is justified. Our task is to draw the straight line that provides the best possible fit.





Scatter Diagram for Example

We can define what we mean by best in various ways. For example, we can draw the line that minimizes the sum of the differences between the line and the points. Because some of the differences will be positive (points above the line), and others will be negative (points below the line), a canceling effect might produce a straight line that does not fit the data at all. To eliminate the positive and negative differences, we will draw the line that minimizes the sum of squared differences. That is, we want to determine the line that minimizes.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where y_i represents the observed value of y and $\hat{y}_i$ represents the value of y calculated from the equation of the line. That is,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The technique that produces this line is called ***the least squares method.*** *The line itself is called* ***the least squares line****, or the* ***regression line****. The "hats" on the coefficients remind us that they are estimators of the parameters* β_0 *and* β_1 *.*

By using calculus, we can produce formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ .Although we are sure that you are keenly interested in the calculus derivation of the formulas, we will not provide that, because we promised to keep the mathematics to a minimum. Instead, we offer the following, which were derived by calculus.

Calculation of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

where

$$SS_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$$

$$SS_x = \sum (x_i - \overline{x})^2$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

and where

$$\hat{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \overline{x} = \frac{\sum x_i}{n}$$

The formula for SS_x should look familiar; it is the numerator in the calculation of sample variance s^2. *We introduced the SS notation; it stands for sum of squares. The statistic SS_x is the sum of squared differences between the observations of x and their mean. Strictly speaking, SS_{xy} is not a sum of squares.*

The formula for SS_{xy} may be familiar; it is the numerator in the calculation for covariance and the coefficient of correlation.

Calculating the statistics manually in any realistic Example is extremely time consuming. Naturally, we recommend the use of statistical software to produce the statistics we need. However, it may be worthwhile to manually perform the calculations for several small-sample problems. Such efforts may provide you with insights into the working of regression analysis. To that end we provide shortcut formulas for the various statistics that are computed in this chapter.

Shortcut Formulas for SS_x and SS_{xy}

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

As you can see, to estimate the regression coefficients by hand, **we need to determine the following summations.**

Sum of x: $\sum x_i$
Sum of y: $\sum y_i$
Sum of x-squared: $\sum x_i^2$
Sum of x times y: $\sum x_i y_i$

Returning to our Example we find

$$\sum x_i = 53$$
$$\sum y_i = 148$$
$$\sum x_i^2 = 609$$
$$\sum x_i y_i = 1{,}786$$

Using these summations in our shortcut formulas, we find

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 609 - \frac{(53)^2}{6} = 140.833$$

and

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 1{,}786 - \frac{53 \times 148}{6} = 478.667$$

Finally, we calculate

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{478.667}{140.833} = 3.399$$

and

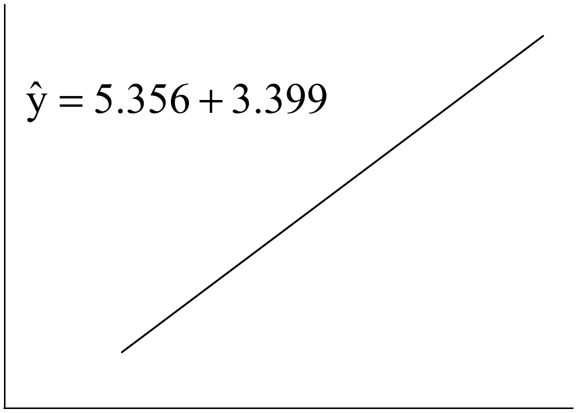
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{148}{6} = 3.399 \times \left(\frac{53}{9} = -5.356\right)$$

Thus, the least squares line is

$$\hat{y} = -5.356 + 3.399x$$

The next figure describes the regression line. As you can see, the line fits the data quite well. *We can measure how well by computing the value of the minimized sum of squared differences. The differences between the points and the line are called residuals, denoted r_i . That is,*

$$r_i = y_i - \hat{y}_i$$



Scatter Diagram with Regression Line Example

The residuals are the observed values of the error variable. *Consequently, the minimized sum of squared differences is called* ***the sum of squares for error denoted SSE.***

Sum of Squares for Error

$$SSe = \sum (y_i - \hat{y}_i)^2$$

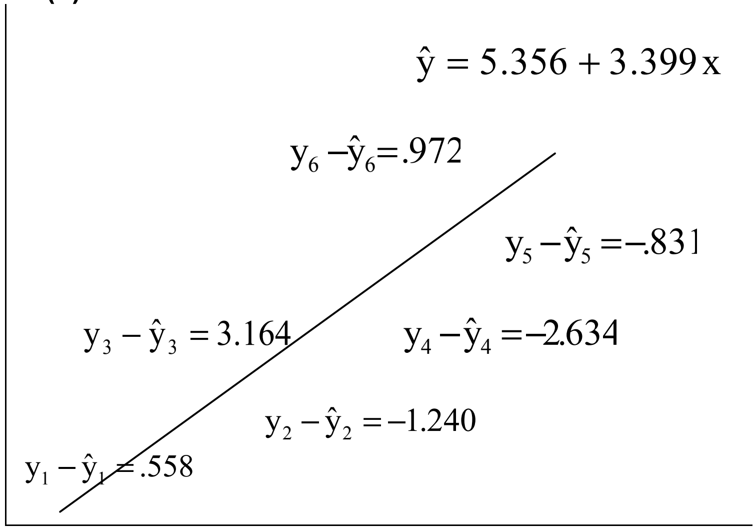
The calculation of SSE Y_i' this Example is shown in the next figure. *Notice that we compute Yi by substituting Xi into the formula for the regression line. The residuals are the differences between the observed values* y_i *and the computed values* $\hat{y}_i$*. The following Table describes the calculation of SSE.*

				RESIDUAL	RESIDUAL SQUARED
i	x_i	y_i	$\hat{y}_i = -5.356 + 3.399x_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	2	2	1.442	0.558	0.3114
2	4	7	8.240	-1.240	1.5376
3	8	25	21.836	3.164	10.0109
4	10	26	28.634	-2.634	6.9380
5	13	38	38.831	-0.831	0.6906
6	16	50	49.028	0.972	0.9448
				$\sum (y_i - \hat{y}_i)^2 = 20.4332$	

Thus, SSE = 20.4332. No other straight line will produce a sum of squared errors as small as 20.4332. In that sense, the regression line fits the data best. The sum of squares for error is an important statistic because it is the basis for other statistics that assess how well the linear model fits the data.

Example

Example (5)


$\hat{y} = 5.356 + 3.399x$
$y_6 - \hat{y}_6 = .972$
$y_5 - \hat{y}_5 = -.831$
$y_3 - \hat{y}_3 = 3.164$
$y_4 - \hat{y}_4 = -2.634$
$y_2 - \hat{y}_2 = -1.240$
$y_1 - \hat{y}_1 = .558$

We now apply the technique to a more practical problem.

Example

Example (6)

Car dealers across North America use the "Red Book" to help them determine the value of used cars that their customers trade in when purchasing new cars. The book, which is published monthly, lists the trade-in values for all basic models of cars. It provides alternative values of each car model according to its condition and optional features. The values are determined on the basis of the average paid at recent used-car auctions. (These auctions are the source of supply for many used-car dealers.) However, the Red Book does not indicate the value determined by the odometer reading, despite the fact that a critical factor for used – car buyers is how far the car has been driven. To examine this issue, a used-car dealer randomly selected 100 three year of Ford Taurusses that were sold at auction during the past month. Each car was in top condition and equipped with automatic transmission, AM/FM cassette tape player, and air conditioning. The dealer recorded the price and the number of miles on the odometer. These data are summarized below. The dealer wants to find the regression line.

Solution

Solution:

Notice that the problem objective is to analyze the relationship between two quantitative variables. Because we want to know how the odometer reading affects selling price, we identify the former as the independent variable, which we used, and the latter as the dependent variable, which we label y.

Solution by Hand

Solution by Hand:

To determine the coefficient estimates, we must compute SS_x and SS_{xy}. They are

$$SS_x = \sum (x_i - \bar{x})^2 = 4,309,340,160$$

and

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 134{,}269.2960$$

Using the sums of squares, we find the slope coefficient.

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{-134{,}269{,}296}{4{,}309{,}340{,}160} = -.0311577$$

To determine the intercept, we need to find *x* and *y.* They are

$$\bar{y} = \frac{\sum y_i}{n} = \frac{541{,}141}{100} = 5{,}411.41$$

and

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3{,}600{,}945}{100} = 36{,}009.45$$

Thus,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5{,}411.41 - (-.0311577)(36{,}009.45) = 6{,}533.38$$

The sample regression line is

$$\hat{y} = 6{,}533 - 0.0312x$$

Interpreting The Coefficient

Interpreting The Coefficient

The coefficient $\tilde{\beta}_1$ is - 0.0312 which means that for each additional mile on the odometer, the price decreases by an average of $0.0312 (3.12 cents).

The intercept is $\tilde{\beta}_0 = 6{,}533$. Technically, **the intercept is the point at which the regression line and the y-axis intersect.** This means that when x = 0 (i.e., the car was not driven at all) the selling price is $6,533. We might be tempted to interpret this number as the price of cars that have not been driven. However, in this case, the intercept is probably meaningless. Because our sample did not include any cars with zero miles on the odometer we have no basis for interpreting $\tilde{\beta}_0$. As a general rule, we cannot determine the value of y for a value of x that is far outside the range of the sample values of x. In this example, the smallest and largest values of x are 19,057 and 49,223, respectively. Because x = 0 is not in this interval we cannot safely interpret the value of *11* when x = o.

Assessing the Model

6.4 Assessing the Model

The least squares method produces the best straight line. However, *there may in fact be no relationship or perhaps a nonlinear (e.g., quadratic) relationship between the two variables.* If so, the use of a linear model is pointless. Consequently, it is important for us to assess how well the linear model fits the data. If the fit is poor, we should discard the linear model and seek another one.

Using the Regression Equation

6.5 Using the Regression Equation

Using the techniques in Section 5, we can assess how well the linear model fits the data. *If the model fits satisfactorily, we can use it to forecast and estimate values of the dependent variable.* To illustrate, suppose that in Example 2, the used car dealer wanted to predict the selling price of a three-year-old Ford Taurus with 40,000 miles on the odometer. Using the regression equation, with *x* = 40,000, we get

$$\hat{y} = 6{,}533 - 0.0312x = 6{,}533 - 0.0312(40{,}000) = 5{,}285$$

Thus, the dealer would predict that the car would sell for $5,285.

Coefficients of Correlation

6.6 Coefficients of Correlation

*When we introduced the coefficient of correlation (also called **the Pearson coefficient of correlation**), we pointed out that it is used to measure the strength of association between two variables.*

Chapter 7: Cross Table Analysis

Chi-Squared Test of a Contingency Table

7.1 Chi-Squared Test of a Contingency Table

The chi-squared test is used to determine if there is enough evidence to infer that two are related and to infer that differences exist among two qualitative variables. Completing both objectives entails to two different criteria. The following is an Example to see how this is done.

Example

Example (1)
One of the issues that came up in a recent national election in many future elections) is how to deal with a sluggish should governments cut spending, raise taxes, inflate the more money), or do none of the above and let the deficit rise politicians need to know which parts of the electorate suppose that a random sample of 1,000 people was asked which and their political affiliations. The possible responses to the affiliation were Democrat, Republican, and Independent. The responses were summarized in cross-classification table, shown below. Do this conclude that political affiliation affects support for the ecology.

	Political Aff.	
Economic Opinions	Democrat	Republican
Cut spending	101	282
Raise taxes	38	67
Inflate the economy	131	88
Let deficit increase	61	90

Solution

Solution:
One way to solve the problem is to consider the contingency table. The variables are economic affiliation. Both are qualitative. The values of economic "raise taxes," "inflate the economy," and "let deficit increase political affiliation are "Democrat," "Republican" and "Independent" objective is to analyze the relationship between the two variables. Specifically, we want to know whether one variable affects the other.

Another way of addressing the problem is to determine whether differences exist among Democrats, Republicans, and Independents. In other words, we treat each political group as a separate population. Each population has four possible values, represented by the four economic options. (We can also answer the question by treating the economic options as populations and the political affiliations as the values of the random variable.) Here the problem objective is to compare three populations.

As you will shortly discover, *both objectives lead to the same test. Consequently, we can address both objectives at the same time.*

The null hypothesis will specify that there is no relationship between the two variables. We state this in the following way.

H_o: The two variables are independent.

The alternative hypothesis specifies that one variable affects the other, which is expressed as

H_A: The two variables are dependent.

If the null hypothesis is true, political affiliation and economic option are independent of one another. *This means that whether someone is a Democrat, Republican, or Independent does not affect his economic choice. Consequently, there is no difference among Democrats, Republicans, and Independents in their support for the four economic options. If the alternative hypothesis is true, political affiliation does affect which economic option is preferred. Thus, there are differences d is likely to among the three political groups.*

The test statistic is

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Where k is the number of cells in the contingency table. *The null hypothesis for the chi-squared test of a contingency table only states that the two variables are independent. However, we need the probabilities in order to compute the expected values (e_j), which in turn permits us to calculate the value of the test statistic. (The entries in the contingency table are the observed values, o_i. The question immediately arises: from where do we get the probabilities? The answer is that they will come from the data after we assume that the null hypothesis is true.*

If we consider each political affiliation to be a separate population, each column of the contingency table represents an experiment with four cells. If the null hypothesis is true, the three experiments should produce similar proportions in each cell. We can estimate the cell probabilities by calculating the total in each row and dividing by the sample size. Thus,

P(cut spending) $\approx \frac{444}{1{,}000}$

P(raise taxes) $\approx \frac{250}{1{,}000}$

P(let deficit increase) $\approx \frac{176}{1{,}000}$

We can calculate the expected values for each cell in the three by multiplying these probabilities by the total number of political group. By adding down each column, we find that there are residents who identified themselves as Democrats (331), 527 as Republicans and 142 as independents.

Expected Values of the Economic Options of Democrats

EONOMIC OPTION	EXPECTED VALUE
Cut spending	$331\times\frac{444}{1,000}=146.96$
Raise Taxes	$331\times\frac{130}{1,000}=143.03$
Inflate economy	$331\times\frac{250}{1,000}=82.75$
Let deficit increase	$331\times\frac{176}{1,000}=58.26$

Expected Values of the Economic Options of Republicans

EONOMIC OPTION	EXPECTED VALUE
Cut spending	$527\times\frac{444}{1,000}=233.99$
Raise Taxes	$527\times\frac{130}{1,000}=68.51$
Inflate economy	$527\times\frac{250}{1,000}=131.75$
Let deficit increase	$527\times\frac{176}{1,000}=92.75$

Expected Values of the Economic Options of Independents

EONOMIC OPTION	EXPECTED VALUE
Cut spending	$142\times\frac{444}{1,000}=63.05$
Raise taxes	$142\times\frac{130}{1,000}=18.46$
Inflate economy	$142\times\frac{250}{1,000}=35.50$
Let deficit increase	$142\times\frac{176}{1,000}=24.99$

Notice that the expected values are computed by multiplying the column total by the row total and dividing by the sample size.

Expected Frequencies for a Contingency Table

7.2 Expected Frequencies for a Contingency Table

The expected frequency of the cell in column j and row i is

$$e_{ij} = \frac{(\text{Columnj total})(\text{Rowi total})}{\text{Samplesize}}$$

The expected cell frequencies are shown in parentheses in the Table below, the expected cell frequencies should satisfy the rule of five.

Contingency Table of Example 3

ECONOMIC	POLITICAL AFFILIATION		
OPTIONS	DEMOCRATE	REPUBLIC	INDEPENDENT
Cut spending	101 (146.96)	282 (233.99)	61 (63.05)
Raise Taxes	38 (43.03)	67 (68.51)	25 (18.46)
Inflate economy	131 (82.75)	88 (131.75)	31 (35.50)
Let deficit increase	61 (58.26)	90 (92.75)	25 (24.99)

We can now calculate the value of the test statistic. It is

$$x^2 = \sum_{i=1}^{12} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(101 - 146.96)^2}{146.96} + \frac{(38 - 43.03)^2}{43.03} + \frac{(131 - 82.75)^2}{82.75} + \frac{(61 - 58.26)^2}{58.26}$$

$$+ \frac{(282 - 233.99)^2}{233.99} + \frac{(67 - 68.51)^2}{68.51} + \frac{(88 - 131.75)^2}{131.75} + \frac{(90 - 92.75)^2}{90.75}$$

$$+ \frac{(61 - 63.05)^2}{63.05} + \frac{(25 - 18.46)^2}{18.46} + \frac{(31 - 35.50)^2}{35.50} + \frac{(25 - 24.99)^2}{24.99}$$

$$= 70.675$$

Notice that we continue to use a single subscript in the formula of the test statistic when we should use two subscripts, one for the rows and one for the columns. *We feel that it is clear that for each cell, we need to calculate the squared difference between the observed and expected frequencies divided by the expected frequency*. We don't believe that the satisfaction of using the mathematically correct notation would overcome the unnecessary complication.

Rejection Region

Rejection Region

To determine the rejection region, we need to know the number of degrees of freedom associated with this x2 - statistic. The number of degrees of freedom for a contingency Table with r rows and c columns is

For Example 3, the number of degrees of freedom is

d.f. = (r -l)(c -1) = (4 -1)(3 -1) = 6

If we use a 5% significance level, the rejection region is

$$x^2 > x^2_\alpha, x^2.05,6 = 12.5916$$

Because x^2 = 70.675, we reject the null hypothesis and conclude that evidence of a relationship between political affiliation and support for nomic options. It follows that the three political affiliations differ in their for the four economic options. We can see from the data that Republicans favor cutting spending, whereas Democrats prefer to inflate the economy.

Example

2

Example (2)

The operations manager of a company that manufactures shirts whether there are differences in the quality of workmanship am shifts. She randomly selects 600 recently made shirts and scarf. Each shirt is classified as either perfect or flawed, and the shift also recorded. The accompanying Table summarizes the number into each cell. Do these data provide sufficient evidence at the 5 to infer that there are differences in quality among the three?

Contingency Table Classifying Shirts

		SHIFT
SHIFT CONDITION	1	2
Perfect	240	191
Flawed	10	9

Solution

2

Solution:

The problem objective is to compare three populations (the shirt three shifts). The data are qualitative because each shirt will be perfect or flawed. This problem - objective / data - type combination statistical procedure to be employed is the chi-squared test of a. The null and alternative hypotheses are as follows.

H_o: The two variables are independent.

H_A: The two variables are dependent.

Test statistics:

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad \text{d.f.} = (r - 1)(c - 1)$$

We calculated the row and column totals and used them to determine the expected values. For example, the expected number of perfect shirts produced in shift 1 is

$$e_1 = \frac{250 \times 570}{600} = 237.5$$

The remaining expected values are computed in a like manner. The original Table and expected values are shown in the Table below.

SHIRT	SHIFT			
CONDITION	1	2	3	TOTAL
Perfect	240 (237.5)	191 (190.0)	139 (142.5)	570
Flawed	10 (12.5)	9 (10.0)	11 (7.5)	30
TOTAL	250	200	150	600

The value of the test statistic is

$$x^2 = \sum_{i=1}^{6} = \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(240 - 237.5)^2}{237.5} + \frac{(10 - 12.5)^2}{12.5} + \frac{(191 - 190.0)^2}{190.0} + \frac{(9 - 10.0)^2}{10.0}$$

$$+ \frac{(139 - 142.5)^2}{142.5} + \frac{(11 - 7.5)^2}{7.5}$$

$$= 2.36$$

Conclusion: Do not reject the null hypothesis

We can measure how strong is the relationship between the two variables using (sort of a correlation coefficient called contingency coefficient CCC)

$$CC = \sqrt{\frac{x^2}{x^2 + n}}$$

Chapter 8: Price and Quantity Indexes

Introduction

Price and quality indexes are summary measures of relative price and quality changes over time in a set of items. For example, the consumer price index summarizes relative price changes in goods and services purchased by urban households; the producer price indexes measure changes in prices received in primary markets by producers of commodities in all stages of processing; and the index of industrial production measures relative changes in output in manufacturing, mining and utilities.

Price and quality indexes not only serve as summary measures of price and quality changes but also are employed in many analyses. *Regression models, for instance, frequently contain price or quality indexes as independent variables. Another use of price indexes is to adjust time series expressed in monetary units, such as annual sales revenues during the past 20 years, so that changes in the series other than price changes can be studied.*

In this chapter, we discuss price and quality indexes in their several roles. First we take up the construction of price indexes various uses of them. Then we discuss quantity indexes.

Percent Relatives and Link Relatives

Percent Relatives and Link Relatives

We begin with the measurement of price changes for a single item, such as a gallon of 89 octane unleaded gasoline or a $3\frac{1}{2}$ inch high-density computer disk.

Present Relatives

Present Relatives

Percent relatives are useful for studying the pattern of relatives changes in the price of an item over time.

Example: Chemical Compound

Example: Chemical Compound

The average price of a chemical compound for each year in the period 1988 – 1991 is shown in the Table below. The corresponding percent relatives, calculated on the base period 1988 are shown in column 2. For instance, the percent relative for 1989 is 100 (22/20) = 110. This relative tells us that the price of the compound in 1989 was 110 percent as great as in 1988.

Calculation of percent relative and link relatives series – chemical compound example

Year	(1)	(2)	(3)	(4)
	Average Price (dollars)	Percent Relatives		Link Relative
		1988=100	1991=100	
1988	20	100	40	-
1989	22	110	44	110
1990	36	180	72	164
1991	50	250	100	139

The base period for a percent relatives series is usually identified for reporting purposes in the manner shown in the above Table; for example, 1988 = 100. The base period may be any period that facilitates the comparisons of interest. For instance, the choice of 1988 in the chemical compound Example facilitates the study of the price increases during a period when demand was increasing rapidly because of a new industrial use for the compound.

Interpretation of Percent Relatives

Interpretation of Percent Relatives

Percent relatives must be interpreted with care. We discuss briefly three important considerations.

1. *The absolute magnitudes of percent relatives are affected by the choice of the base period, but their proportional magnitudes are not affected by this choice.* We illustrate this for the chemical compound Example by presenting in column 3 of the above Table the percent relatives calculated on the base period 1991. The percent relative for 1988 now is 100(20/50 = 40, not 100 as for the 1988 base period series. However, the proportional magnitudes of the percent relatives in columns 2 and 3 are the same. For instance, the percent relatives for 1989 and 1988, with base period 1988, have the ratio 110/100 = 1.1. For the percent relatives with the 1991 base period, this ratio is the same; that is, 44/40 = 1.1. Similarly for both series of percent relatives, the ratio of the 1990 price relative to the 1989 price relative is 1.64.

2. *A comparison of percent relatives for two different price series indicates nothing about the actual prices unless we know the actual magnitudes in the base period.* For example, information that the percent relatives for large "grade-A" eggs last month were 120 for Montreal and 123 for Toronto tells us nothing about how the egg prices in the two cities compared last month unless we know the prices in the base period.

3. *In analyzing percent relatives, we distinguish between percent points of change and percent change.* To illustrate the difference, refer to the percent relatives with base period 1988 in column 2 of the Table for the chemical compound example. Between 1990 and 1991, the

percent points of change were 250 - 180 = 70. On the other hand, the percent change was 100[(250-180)/180] = 100(70/180) = 38.9. Thus, a percent point of change refers to the absolute change in the percent relatives and is dependent on the choice of the base period. Percent change, in contrast, refers to the relative change in percent relative and, as we noted previously, is not affected by the choice of the base period.

Link Relatives

Link Relatives

Link relatives are useful in studying relative period-to-period changes instead of change from a fixed base period. For instance, link relatives enable us to study whether or not the price of an item is increasing at a constant rate, an increasing rate, or a decreasing rate.

A link relative series express the value of the series in each period as percent of the value in the immediately preceding period.

Example

1

Example

The link relatives for the chemical compound Example are given in column 4 in the same Table of the above Example. For instance, the link relative for 1989 is 100(22/20) = 110. This tells us that the price in 1989 was 110 percent as great as in 1988. The link relative for 1990 is 100(36/22) = 164, indicating that the price in 1990 was 164 percent as great as in 1989. Note that the percent increase from 1990 to 1991 was smaller than that from 1989 to 1990 (column 4), even though the amount of price increase ($14) did not decline (column 1).

Comment

Comment

We have illustrated percent relatives and link relatives for price series, but these relatives can be utilized for any time series (for example, annual sales or monthly production).

Price Index by Method of Weighted Aggregates

8.1 Price Index by Method of Weighted Aggregates

Price indexes are summary measures that combine the price changes for a group of items, using weights to give each item its appropriate importance. *The consumer price index is such an index measuring the combined effect of price changes in many goods and services purchased by urban households.* ***Two basic methods*** *are widely used for calculating price indexes:* ***the method of weighted aggregates and the method of weighted average of relatives.*** We shall explain each in turn, using the following illustration.

Example: School Maintenance Supplies

Example: School Maintenance Supplies

Officials of a large school district needed to develop a price index for maintenance supplies for the school buildings in the district. In compiling this index, the officials could not include all supply items in the index because the number of supply items used is very large. Instead, a sample of supply items was selected to represent all items used. The selected items are shown in Tabulated below, column 1. Also shown in this table, in columns 3, 4, and 5, are the unit prices of these items for 1988, 1989 and 1990, respectively. Finally, column 2 of the Table presents typical quantities consumed annually for each of the selected supply items.

Method of Weighted Aggregates

Method of Weighted Aggregates

The method of weighted aggregates for compiling a price index simply compares the cost of the typical quantities consumed at the prices of a given period with the corresponding cost in the base period. We shall first explain the notation and terminology used we illustrate the method of calculation by an example.

Calculation of price index series by method of weighted aggregates-school maintenances supplies example

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Schedule of Items	**Quantity** Q_{ia}	**Unit Price**			**$P_{10}Q_{12}$** **(3) x (2)**	**$P_{11}Q_{12}$** **(4)x (2)**	**$P_{12}Q_{1a}$** **5x2**
		1998 **P_{10}**	**1989** **P_{11}**	**1990** **P_{12}**			
1 Window glass	15 sheets	15.55	15.86	16.21	233.25	237.90	243.15
2 Fluorescent tube	130 boxes	39.17	39.81	40.55	5.092.1	5.175.3	5.271.5
3 Floor detergent	290 cans	21.95	23.55	24.90	6.365.5	6.829.5	7.221.0
4 Floor finish	100 cans	49.39	52.27	25.53	4.939.0	5.227.0	5.253.0
5 Latex interior paint	175 cans	11.13	11.50	11.56	1.94.75	2.012.5	2.023.0
6 Mop head	200 pieces	2.86	2.93	2.99	572.00	586.00	598.00
			Total		19.149.6	20.068.2	20.609.65

$$I_{88} = \left(\frac{19.149.60}{19.149.60}\right) = 100.0$$

$$I_{89} = \left(\frac{20.068.20}{19.149.60}\right) = 100.0$$

$$I_{90} = \left(\frac{20.609.65}{19.149.60}\right) = 100.0$$

The schedule of items is the list of items included in the price index. Often, the schedule of items consists of a sample of all items of interest, as in the school maintenance supplies example. Sometimes, it is

The price of the I^{th} item in the schedule in any given period t is denoted by P_{it} and the price of this item in the base period is denoted by P_{io}. The typical quantity consume the ith item is denoted by Q_{ia} where the subscript a stands for an average or typical period. These quantities are used as weights to reflect the importance of each item.

The method of weighted aggregates compares the cost of the typical quantities period t prices with the cost of the same quantities at base period prices. **The cost at period *t* prices is:**

$$\textbf{Cost at period } t \textbf{ prices} = \sum_{i} P_{it} A_{ia}$$

Where the summation is over all the items in the schedule. **The cost at base period price is :**

$$\textbf{Cost at period 0 prices} = \sum_{i} P_{io} A_{ia}$$

The price index for period *t* by the method of weighted aggregates is the ratio of these two costs, expressed as a percent. The price index for period *t* will be denoted by 1_{it}.

$$I_{it} = 100 \frac{\sum P_{it} Q_{ia}}{\sum P_{ia} Q_{ia}}$$

Where:

P_{io} *is the unit price of the ith item in period o (based period).*

P_{it} *is the unit price of the ith item in period t (given period).*

Q_{ia} *is the quantity weight assigned to the ith item*

Example

2

Example

For the schools maintenance supplies example, the price index series for 1988, 1989, and 1990 is calculated by the method of weighted aggregates. Columns 6. 7 and 8 contain the costs of the typical quantities at the prices of each year. For example, the cost of 15 sheets of window glass at 1988 prices is 15(15.55) = 233.25 while at 1989 prices the cost is 15(15.86) = 237.90. The aggregate costs of the schedule of items at the prices of the three periods are shown at the bottoms of columns 6, 7, and 8, respectively. The index numbers are calculated at the bottom of the table.

Since the same items and quantities are priced in each period, the differences in the index are attributable wholly to changes in the prices of the items. Thus, the price index I_{89} = 104.8 indicates that the prices of school maintenance supplies increased, in terms of their aggregate effect, by 4.8 percent between 1988 and 1989. Since the index for 1990 (I_{90} = 107.6) is above the index for 1989 (I_{89} = 104.8), the prices, in their aggregate effect, increased still more between 1989 and 1990. We calculate the percent price increase between 1989 and 1990 by expressing the percent point change, 107.6 - 104.8 = 2.8, as a percent of the 1989 index and obtain 100 (2.8/104.8) = 2.7 percent as the relative price increase between 1989 and 1990.

Comments

Comments

1. *Just as with percent relatives, the proportional magnitudes of the index numbers in the Table are not affected by which period is chosen as the base period.*
2. *The period from which the quantity weights are derived is called the weight period. The base period for the price index and the weight period need not coincide.*
3. *When the typical quantity weights Q_{it} are the quantities consumed in the base period - that is, Q_{ot} the weighted aggregates price index is called a Laspeyres price index.*

Price Index by Method of Weighted Average of Relatives

8.2 Price Index by Method of Weighted Average of Relatives

A second method of compiling a price index series is the method of weighted average of relatives. *This method utilizes the percent relatives of the prices for the item in the schedule. The index for period t is simply a weighted average of the percent relatives for all schedule items. The weights required by this method are value weights rather than quantity weights. We shall denote the value weight for the I^{th} item by V_{ia}.*

The price index it for period t calculated by the method of weighted average of relatives is

$$IT = \sum_{i} \frac{\left(100 \frac{P_{it}}{P_{ia}}\right) V_{ia}}{\sum V_{ia}}$$

Where:

P_{io} *is the unit price of the i_{th} item in period o (based period).*

P_{it} *is the unit price of the i_{th} item in period t (given period).*

V_{ia} *is the value weight assigned to i_{th} item.*

The value weights V_{ia} are dollar values that reflect the importance of each item in the schedule. For example, the value weights may be obtained from the typical quantities consumed, Q_{ia} by multiplying these quantities by typical prices.

Example

3

Example

For the school maintenance supplies example, we shall calculate a price index series by the method of weighted average of relatives with base period 1988. The value weights will be based on the typical quantities and the unit prices for the base period 1988 in the Table that is: $V_{ia} = P_{io} Q_{ia}$

The necessary calculations are tabulated below. Column 1 in the Table repeats the schedule of Items, and column 2 contains the value weights as obtained from the Table. For example, the value weight for window glass is $V_{ia} = P_{io}\, Q_{ia}$ = 15.55(15) = 233.25. The sum of the value weight is $\sum V_{ia} = 19.149.60$

Columns 3, 4, and 5 contain the percent relatives for the prices of each schedule Item in the three years based on the data in the Table. Note that all percent relatives are expressed on a 1988 base because the price index is to have 1988 as the base period. For example, the 1989 percent relative for window glass is 100(15.86/15.55) = 101.99.

We now take a weighed average of the percent relatives for each year. The weighting is done in columns 6, 7, and 8, and the indexes are obtained at the bottom of the table. The index I_{89} = 104.8 for 1989 indicates that prices of school maintenance supplies increased, in terms of their aggregate effect, by 4.8 percent between 1988 and 1989.

Comment

Comment

The price indexes in Tables 26.2 and 26.3 are identical. This happened because we used base year prices P_{10}, in obtaining the value weights. In that case, the method of weighted average of relatives index (26.4) reduces to the method of weighted aggregates index (26.3) as follows:

$$\frac{\sum_i \left(100\,\frac{P_{it}}{P_{io}}\right) V_{ia}}{\sum_i V_{ia}} = \frac{\sum_i \left(100\,\frac{P_{it}}{P_{io}}\right) P_{io}\, V_{ia}}{\sum_i P_{io}\, Q_{ia}} = 100\,\frac{\sum_i P_{it}\, V_{ia}}{\sum_i P_{io}\, Q_{ia}}$$

When prices other than base period prices are used for obtaining the value weights V_{ia}, the two methods will not lead to identical results. Generally, though, the differences in the indexes calculated by the two methods will not be great.

We now take up several considerations that arise in compiling index numbers series.

Considerations in Compiling Index Series

8.3 Considerations in Compiling Index Series

Choice of Base Period: We now take up several considerations that arise in compiling index numbers series. *Any period within the coverage of the index series can be used as the base period. Of course, when a special-purpose index is designed to measure price changes occurring since particular period, such as since the lifting of price controls, that period would be taken as the base period. Whenever possible, the base period should involve relatively normally standard conditions, because many index users assume that the base period represent, such conditions. Sometimes an interval of two or more years is selected as the base period. In this case, the index numbers are calculated so that* *the indexes for the base period year average to 100. The base period is then reported in a form such as 1990-1991 = 100.*

Calculation of price index series by method of weighted average of relatives-school maintenance supplies example

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Schedule of Items	Value Weight	Unit Price			Percent Relative x Value Weight		
	$V_{ia} = P_{io} Q_{ia}$	1998 $100\left(\frac{P_{it}}{P_{io}}\right)$	1989 $100\left(\frac{P_{ii1}}{P_{io}}\right)$	1990 $100\left(\frac{P_{it}}{P_{io}}\right)$	1998 $100\left(\frac{P_{it}}{P_{io}}\right)V_{ia}$ (3)x(2)	1989 $100\left(\frac{P_{it}}{P_{io}}\right)V_{ia}$ (4)x(2)	1990 $100\left(\frac{P_{it}}{P_{io}}\right)V_{ia}$ (5)x(2)
Window glass	233.25	100.0	101.99	104.24	23.325	23.789	24.314
Fluorescent tube	5.092.10	100.0	101.63	103.52	509.210	517.510	527.134
Floor detergent	60365.5	100.0	107.29	113.44	636.550	682.954	722.102
Floor finish	4.939.00	100.0	105.83	106.36	493.900	522.694	525.312
Latex interior paint	1.947.75	100.0	103.32	103.86	194.775	201.242	202.293
Mop head	572.00	100.0	102.45	104.55	57.200	58.601	59.803
	19.149.6			Total	1914.96	2.006.79	2.060.958

$$I_{88} = \left(\frac{1.914\ .960}{19\ .149\ .60}\right) = 100\ .0$$

$$I_{89} = \left(\frac{2.006\ .790}{19\ .149\ .60}\right) = 104\ .8$$

$$I_{90} = \left(\frac{2.060\ .9585}{19\ .149\ .60}\right) = 107\ .6$$

Shifting of Base Period

Shifting of Base Period

Sometimes it is necessary to shift the base period of an index series that is already compiled, as when two index series on different base periods need to be placed on a common base period to facilitate comparison. Usually it is not possible to recalculate a published index on the new base period because the needed data are not available. A shortcut method can be employed; however, that does not entail recalculation of the index.

To illustrate the shortcut procedure, let us consider the following price index series with 1989 as the base period:

Year	1986	1987	1988	1989	1990	1991
Price Index (1989 = 100):	75	88	92	100	110	122

Suppose the base period is to be shifted to 1986. We simply divide each index number by *0.75,* the index number for the new base period in decimal form. The index number for 1986 becomes *75/0.75* = 100.0, that for 1987 becomes 88/0.75 = 117.3, and so on. The new price index series with base period 1986 is:

Year	1986	1987	1988	1989	1990	1991
Price Index (1986 = 100):	100.0	117.3	122.7	133.3	146.7	162.7

Because the proportional magnitudes of the index numbers are not affected by this shifting, the new series conveys the same information about year-to-year relative price changes as the original series.

The shortcut procedure yields results identical to those obtained by recalculating the index series on the new base period when the index series is calculated by the method of weighted aggregates with fixed quantity weights. With most other formulas, however, the shortcut procedure only provides approximate results.

Splicing

Splicing

When an index series is revised, the new series can be joined or spliced to the older series to yield a single continuous series by a procedure similar to that just described for shifting the base period of an index series. Suppose the following two price index series are to be joined:

Year	1983	1984	1985	1986	1987	1988	1989	1990	1991
Old Series (1983 = 100)	100	103	106	113	130				
Revised Series (1987 = 100)					100	120	126	131	134

To splice the two series together, we simply shift the level of one of the series so that both have a common value for 1987. For instance, to obtain a combined series with 1983 as the base period, we multiply every index in the revised series by 1.30. The spliced series is:

Year	1983	1984	1985	1986	1987	1988	1989	1990	1991
Spliced Series (1983 = 100):	100	103	106	113	130	156	164	170	174

Collection of Price and Quantity Data

Collection of Price and Quantity Data

All of the data-collection methods discussed is used in collecting data for price indexes. *Price and quantity data are obtained by* *observation, interview, and self- enumeration. Often sampling is used, as in the consumer price index where probability samples of retail outlets and of items carried by the selected outlets are employed to obtain price data. Sampling is also used for the consumer price index in periodic consumer expenditures surveys to obtain quantity data for weights.*

In collecting price data for an item over a period of years, it is important that the quality and other price-determining characteristics be held as constant as possible. This typically requires that detailed specifications be developed and adhered to in pricing the item. Here if is an Example of such a specification:

Interior latex paint. Professional - or commercial - grade latex interior house paint, matte or flat finish off white, first - line or quality.

Note that color, finish, and quality are specified because each of these characteristics can affect the price.

A different problem in collecting price data in ongoing index series is the adjustment for quality changes in the items covered by the index. In principle, a price change that is caused by a quality change should not be reflected in the price index. In practice, minor quality changes are ignored. On the other hand, attempts are made in better price indexes to adjust for important quality changes. However, the conceptual and measurement problems are difficult. For instance, if an oven door is modified by the manufacturer to exclude the glass window while the list price remains unchanged, should this be treated as a price increase and reflected in the index? If so, what should be the magnitude of the imputed increase? Extensive research has been undertaken on how problems connected with quality changes should be handled.

Another issue in compiling an index series concerns weights. In some indexes, the schedule includes all or almost all of the items used in the activity. For instance, in a price index for highway construction, where a relatively small number of major items account for most of the total cost, it suffices for practical purposes to limit the schedule to these major items. Here, each weight reflects the importance of the particular item.

Often, however, the schedule must be limited to a sample of items, since thousands of items may be involved in the activity and it is neither feasible nor necessary to include them all. Each item is then selected for the schedule to represent a class of related items. In such cases, the weight for an item normally reflects the importance of the entire class represented by the item. Consider, for instance, a particular type of insulation that was selected for a price index for residential building construction to represent all types of insulation and vapor barriers. The weight for the insulation the schedule here needs to represent the importance of the entire class of insulation and vapor barriers.

Maintenance of Price Index Series

Maintenance of Price Index Series

A key issue in the maintenance of price index series involves changes in the items included in the index and in the weights assigned to the items to keep them up to date. Should the schedule of items and the weights be changed frequently so that they are always up to date or should they be held fixed over a relatively long sequence of years? *If frequent changes are made, changes in the price index series over time will reflect both changes in price, and changes in the composition of the index. Another disadvantage of frequent revision is that it may be very expensive. On the other hand an index can become badly outdated if the schedule of items and the weights are held fixed over too long a period.*

The procedure generally followed in practice is to hold the schedule of items and the weights essentially fixed for 5 to 10 years and then to revise them. The schedule and weights need not be held exactly fixed, since procedures are available for making inter adjustments without disturbing the index. For instance, a new item can be substituted for an item that has been taken off the market. And seasonal items, such as fresh produce, can be included in the schedule in season.

In addition to periodic updating of the schedule of items and the weights, revisions often also include upgrading of data collection and data handling procedures and modernizing of definitions. A new series of index numbers is then initiated with each revision. The new series can be joined to the preceding series, if desired, by the splicing procedure already explained.

Uses of Price Indexes

8.4 Uses of Price Indexes

Price indexes play important roles in a variety of applications. We now consider two important uses of price indexes.

Measuring Real Earnings

Measuring Real Earnings

As prices change, so does the quantity of goods and services that can be purchased by a fixed sum of money. In economic analysis, it is frequently important to measure change in real earnings (that is, in the quantity of goods and services that can be purchased). ***Price indexes are a basic tool in making this measurement.***

Example: Weekly Earnings

Example: Weekly Earnings

Average weekly dollar earnings of production or non-supervisory workers on non-agricultural payrolls in the United States during the period 1985-1989 are shown in the Table below. These earnings data are expressed in *current dollars.* For example, 1985 earnings are expressed in terms of the buying power of workers' dollars in 1985, and earnings in 1986 are expressed in terms of the buying power of workers' dollars in 1986. We see that earnings in current dollars increased steadily during the period.

The extent to which the prices of goods and services purchased for daily living by the workers and their dependents changed during the period is shown in column 2 by the consumer price index for urban wage earners and clerical workers. This index is expressed on a 1985 base period in the table. Note that prices in 1986 stood at 101.6 percent of their 1985 level. Thus, on the average, families had to spend $1.016 in 1986 for every $1 spent in 1985 in purchasing goods and services for daily living. Consequently, the average weekly earnings of $304.85 in 1986 were equivalent in purchasing power to 304.85/1.016 = $300.05 at 1985 prices. Since 1985 average weekly earnings were $299.09, we see that most of the increase in average weekly earnings between 1985 and 1986 (from $299.09 to $304.85) was offset by price increases and that real earnings increased only slightly.

Earnings in current and 1985 dollars - weekly earnings Example

Year	(1)	(2)	(3)	(4)	(5)
	Weekly Earnings	Consumer Price Index (1985=100)	Weekly Earnings in 1985 Dollars $(1)\div\frac{(2)}{100}$	Relatives for Column 3	
				Link Relative	Percent Relative (1985=100)
1985	299.09	100.0	299.09	-	100.0
1986	304.85	101.6	300.05	100.3	100.3
1987	312.50	105.2	297.05	99.0	99.3
1988	322.36	109.4	394.66	99.2	98.5
1989	335.20	114.7	292.24	99.2	97.7

Sourrc: Basic data from Monthly Labor Review

The average weekly earnings in the other years at 1985 prices are obtained by the same procedure: Each current earnings figure is divided by the decimal value of the price index for that year. Column 3 of the Table contains the average weekly earnings data expressed at 1985 prices. These data are said to be in constant dollars or 1985 dollars since they are expressed in terms of the purchasing power of the dollar in 1985. Relative changes in constant-dollars earnings indicate the relative changes in purchasing power associated with the money earnings - that is, the relative changes in real earnings. To show these relative changes explicitly, we have expressed the constant - dollars data as link relatives in column 4 in the Table and as percent relatives in column 5 with 1985 = 100. The link relatives show that real earnings increased slightly between 1985 and 1986 and then declined a little each year between 1986 and 1989. The percent relatives show that real earnings in 1989 were only at 97.7 percent of their 1985 level.

Comment

Comment

We noted earlier that proportional magnitudes in an index series are not affected by the choice of the base period but that the absolute magnitudes are affected. The same holds for constant – dollars series. Thus, in the weekly earnings example, if the consumer price index had been expressed on a 1989 base, the constant-dollars earnings would have differed from column 3 of its Table, but the percent relatives and the link relatives based on these 1989 constant-dollars data would have remained exactly the same.

Measuring Quantity Changes

Measuring Quantity Changes

Many important business and economic series on volume of activity are expressed in current dollars. Changes in the dollar volume reflect quantity changes, or price changes; or both. Often, there is interest in the quantity changes alone. For example, if retail sales this year are 5 percent higher than last year's sales, is this due to price increases only or did the physical volume of goods sold increase? *We can measure relative changes in quantities by the use of price indexes. The procedure is similar to that for expressing current-dollars earnings as constant-dollars earnings.*

Appliance Sales Example

Annual sales (in $ thousands) by a manufacturer of household appliances during 1989-1992 are shown in the Table below. A relevant price index for the products of the manufacturer is shown in column 2. The constant-dollars sales in 1989 dollars are shown in column 3. The calculations parallel those for obtaining constant dollars earnings in the previous Table. The link relatives and the percent relatives (1989 = 100) of the constant-dollars sales are shown in columns 4 and 5, respectively. We see that the quantity of household appliances sold increased by 33 percent between 1989 and 199 and that the annual rate of increase was about 10 percent per year.

Manufacturer's sales in current and 1989 dollars appliance sales Example

Appliance Sales Example

Year	(1)	(2)	(3)	(4)	(5)
	Annual Sales ($ thousand)	Product Price Index (1989=100)	Sales in 1989 Dollars($ thousand) $(1) \div \frac{(2)}{100}$	Relatives for Column 3	
				Link Relative	Percent Relative (1989=100)
1989	38.500	100	38.500	-	100.0
1990	43.538	103	42.270	109.8	109.8
1991	49.050	105	46.714	110.5	121.3
1992	54.950	107	51.355	109.9	133.4

Comments

Comments

1. *When current-dollars series are converted into constant dollars, the price index must be relevant to the series. For instance, we would not use a food price index to adjust a series appliances*

sale. or the consumer price index to adjust the sales of a steel manufacturer

2. *When a current-dollars series is converted into a constant -dollars series, the latter is often refined to as a deflaled series.*

8.5 Quantity Indexes

Quantity Indexes

A quantity index is a summary measure of relative changes over time in the quantities set of items - for instance, in the quantities of automobiles, trucks, and other imported annually by the United States. *In such index series, the quantities can from one period to another but the prices or other weights remain fixed. The index mulas are analogous to those for price indexes.*

The quantity index I_t, for period t calculated by the method of weighted aggregates is

$$I_t = 100 \frac{\sum Q_{it} P_{ia}}{\sum Q_{io} P_{ia}}$$

The quantity index for period t calculated by method of weighted average of relatives is

$$I_t = \frac{\sum \left(100 \frac{Q_{it}}{Q_{io}} \right) V_{ia}}{\sum V_{ia}}$$

A problem in constructing quantity indexes arises in the choice of weights.

The usual price or value weights may not be appropriate here. Consider a company that wishes to measure quantity changes in it outputs of different electronic circuits assembled purchased components. Here, price weights or value weights for the different circuits would reflect mainly the prices or values of the purchased components and not the quantities of assembly work performed by the company per se. Weights derived from the typical number of hours expended in assembling each circuit or form the value added to each circuit in the assembly would be more appropriate. Value added here would be the value of the circuit when shipped from the plant less than the cost of the purchased components and other purchased goods and services (for example, fuel, electricity, containers) used in assembling the circuit and packing it for shipment. The index of industrial production utilizes value-added weights in many of its segments.

Glossary

Glossary

Census: A survey that includes all members of the population.

Continuous variable: A (quantitative) variable that can assume any numerical value over a certain interval or intervals.

Data or data set: Collection of observations or measurements on a variable.

Descriptive statistics: Collection of methods that are used for organizing, displaying, and describing data using tables, graphs, and summary measures.

Discrete variable: A (quantitative) variable whose values are countable.

Element or member: A specific subject or object included in a sample or population.

Inferential statistics: Collection of methods that help make decisions about a population based on sample results.

Interval scale: Data that can be ranked and for which we can find the difference between two values are said to have an interval scale.

Nominal scale: Data that are divided into different categories that are used for identification purposes only are said to have a nominal scale.

Measures of dispersion: Measures that give the spread of a distribution. The range, variance, standard deviation, and coefficient of variation are four such measures.

Measures of position: Measures that determine the position of a single value in relation to other values in a data set. Quartiles, percentiles, and percentile rank are examples of measures of position.

Median: The value of the middle term in a ranked data set. The median divides a ranked data set into two equal parts.

Mode: A value (or values) that occurs with highest frequency in a data set.

Multimodal distribution: A distribution that has more than two modes. Bimodal is a special case of a multimodal distribution with two modes.

Outliers or extreme values: Values those are very small or very large relative to the majority of the values in a data set.

Parameter: A summary measure calculated for population data.

Percentile rank: The percentile rank of a value gives the percentage of values in the data set that are smaller than this value.

Percentiles: Ninety-nine values that divide a ranked data set into 100 equal parts.

Quartiles: Three summary measures that divide a ranked data set into four equal parts.

Range: A measure of spread obtained by taking the difference between the largest and the smallest values in a data set.

Second quartile: Middle or second of the three quartiles that divide a ranked data set into four equal parts. About 50% of the values in the data set are smaller and about 50% are larger than the second quartile. The second quartile is the same as the median.

Observation or measurement: The value of a variable for an element.

Ordinal scale: Data that can be divided into different categories that can be ranked are said to have an ordinal scale.

Population or target population: The collection of all elements whose characteristics are being studied.

Qualitative or categorical data: Data generated by a qualitative variable.

Qualitative or categorical variable: A variable that cannot assume numerical values but is classified into two or more categories.

Quantitative data: Data generated by a quantitative variable.

Quantitative variable: A variable that can be measured numerically.

Random sample: A sample drawn in such a way that each element of the population has some chance of being included in the sample.

Ratio scale: Data that can be ranked and for which all arithmetic operations can be performed are said to have a ratio scale.

Representative sample: A sample that contains the characteristics of the corresponding population.

Sample: A portion of the population of interest.

Sample survey: A survey that includes elements of a sample.

Statistics: Collection of methods that are used to collect, analyze, present, and interpret data and to make decisions.

Survey: Collecting data on the elements *of* a population or sample.

Variable: A characteristic under study or investigation that assumes different values for different elements.

Bimodal distribution: A distribution that has two modes.

Box-and-whisker plot: A plot that shows the center, spread, and skewness of a data set by drawing a box and two whiskers using the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

Coefficient of variation: A measure of relative variability that expresses standard deviation as a percentage of the mean.

Empirical rule: For a specific bell-shaped distribution, about 68% of the observations fall in the interval $(\mu-\sigma)$ to $(\mu+\sigma)$, about 95% fall in the interval $(\mu-2\sigma)$ to $(\mu+2\sigma)$, and about 99.7% fall in the interval $(\mu-3\sigma)$ to $(\mu+3\sigma)$.

First quartile: The value in a ranked data set such that about 25% of the measurements are smaller than this value and about 75% are larger. It is the median of the values that are smaller than the median of the whole data set.

Inter quartile range: The difference between the third and the first quartiles.

Mean A measure of central tendency: calculated by dividing the sum of all values by the number of values in the data set.

Measures of central tendency: Measures that describe the center of a distribution. The mean, median, and mode are three of the measures of central tendency.

Standard deviation: A measure of spread that is given by the positive square root of the variance.

Statistic: A summary measure calculated for sample data.

Third quartile: Third of the three quartiles that divide a ranked data set into four equal parts. About 75% of the values in a data set are smaller than the value of the third quartile and about 25% are larger. It is the median of the values that are greater than the median of the whole data set.

Unimodal distribution: A distribution that has only one mode.

Variance: A measure of spread.

Coefficient of determination: A measure that gives the proportion (or percentage) of the total variation in a dependent variable that is explained by a given independent variable.

Degrees of freedom for a simple linear regression model: Sample size minus 2, that is, *n* -2.

Dependent variable: The variable to be predicted or explained.

Deterministic model: A model in which the independent variable determines the dependent variable exactly. Such a model gives an exact relationship between two variables.

Estimated or predicted value of y: The value of the dependent variable, denoted by $\hat{y}$, that is calculated for a given value of *x* using the estimated regression model.

Independent or explanatory variable: The variable included in a model to explain the variation in the dependent variable.

Least squares estimates of *A* and *B*: The values of *a* and *b* that are calculated by using the sample data.

Least squares method: The method used to fit a regression line through a scatter diagram such that the error sum of squares is minimum.

Least squares regression line: A regression line obtained by using the least squares method.

Linear correlation coefficient: A measure of the strength of the linear relationship between two variables.

Linear regression model: A regression model that gives a straight line relationship between two variables.

Multiple regression model: A regression model that contains two or more independent variables.

Negative relationship between two variables: The value of the slope in the regression line and the correlation coefficient between two variables are both negative.

Nonlinear (simple) regression model: A regression model that does not give a straight line relationship between two variables.

Population parameters for a simple regression model: The values of *A* and *B* for the regression model y = A + bx + $\in$ that are obtained by using population data.

Positive relationship between two variables: The value of the slope in the regression line and the correlation coefficient between two variables are both positive.

Prediction interval: The confidence interval for a particular value of *y* for a given value of *x.* Probabilistic or statistical model A model in which the independent variable does not determine the dependent variable exactly.

Random error term ($\in$**)**: The difference between the actual and predicted values of y.

Scatter diagram or scatter gram: A plot of the paired observations of x and y.

Simple linear regression: A regression model with one dependent and one independent variable that assumes a straight line relationship.

Slope: The coefficient of *x* in a regression model that gives the change in y for a change of one unit in x.

SSE: (error sum of squares) The sum of the squared differences between the actual and predicted values of *y.* It is that portion of the SST that is not explained by the regression model.

SSR (regression sum of squares): That portion of the SST that is explained by the regression model.

SST (total sum of squares): The sum of the squared differences between actual y values and y.

Standard deviation of errors: A measure of spread for the random errors.

Y-Intercept: The point at which the regression line intersects the vertical axis on which the dependent variable is marked. It is the value of *y* when *x* is zero.

References

References

1. Neeter, J, Waserman, Whitmare, (1993): Applied Statistics. 4th Edition, Louise Richardson.

2. Keller, G and Waracck, B (2001): Statistics for Management and Economics 6th Edition Duxbury.

3. Freund, J (2001) Modern Elementary Statistics 10th Edition, Printice Hall.

Pathways to Higher Education Project

Pathways Mission

Training fresh university graduates in order to enhance their research skills to upgrade their chances in winning national and international postgraduate scholarships as well as obtaining better job.

Partners

- CAPSCU, Cairo University
- Ford Foundation, USA
- Future Generation Foundation, FGF
- National Council for Women, NCW
- Cairo University: Faculties of Commerce, Arts, Mass Communication, Law, Economics & Political Science, and Engineering

Training Programs

- Enhancement of Research Skills
- Training of Trainers
- Development of Leadership Skills

Publications of Training Programs

1- Planning and Controlling
2- Systems and Creative Thinking
3- Research Methods and Writing Research Proposals
4- Statistical Data Analysis
5- Teams and Work Groups
6- Risk Assessment and Risk Management
7- Argumentation: Techniques of Measurement and Development
8- Communication Skills
9- Negotiation Skills
10- Analytical Thinking
11- Problem Solving and Decision Making
12- Stress Management
13- Accounting for Management and Decision Making
14- Basics of Managerial Economics
15- Economic Feasibility Studies
16- Health, Safety and Environment
17- Wellness Guidelines: Healthful Life
18- Basic Arabic Language Skills for Scientific Writing
19- General Lectures Directory
20- Enhancement of Research Skills Graduation Projects Directory

Project Web-site

www.Pathways-Egypt.com

Published by: CAPSCU – Center for Advancement of Postgraduate Studies and Research in Engineering Sciences, Faculty of Engineering - Cairo University
Tel: (+202) 5716620, (+202) 5678216 Fax: (+202) 5703620
Web-site: www.capscu.com E-mail: capscu@tedata.net.eg