STATISTICAL METHODS IN CIVIL ENGINEERING

LECTURE NOTES

PROF. DR. ÖMER YÜKSEK

KTÜ CIVIL ENG. DEPT.

TRABZON

CHAPTER 1

INTRODUCTION

1.1. STATISTICS IN ENGINEERING

1.1.1. General Introduction

Probabilistic methods are with increasing frequency used in the design of civil structures such as channels, storm surge barriers, bridges, buildings, etc. The methods are being applied directly, or are translated to relatively simple design rules with safety coefficients. In both cases the foundation of the calculations is given by the statistical distribution functions of the strength and load variables.

In the application of probabilistic methods, the availability of useful calculation models and adequate statistical distributions are required. The development of calculation models has had a lot of attention during the last years. Methods for finding the best estimate, as well as a quantification of the uncertainty in the estimate, should be described. How to deal with uncertainties is an essential part of statistics science.

Probability and statistics are concerned with events which occur *by chance*. Examples include occurrence of accidents, errors of measurements, production of defective items from a production line. In each case one may have some knowledge of the likelihood of various possible results, but she/he cannot predict with any certainty the outcome of any particular trial. Probability and statistics are used throughout engineering. Civil engineers use statistics and probability to test and account for variations in materials and goods.

1.1.2. Some Important Terms

a. *Probability* is an area of study which involves predicting the relative likelihood of various outcomes. It is a mathematical area which has developed over the past three or four centuries. Its usefulness for describing errors of scientific and engineering measurements was soon realized. Engineers study probability for its many practical uses, ranging from quality control and quality assurance to communication theory in electrical engineering. Engineering measurements are often analyzed using statistics and a good knowledge of probability is needed in order to understand statistics.

b. *Statistics* is a word with a variety of meanings. To the man in the street it most often means simply a collection of numbers, such as the number of people living in a country or city, a stock exchange index, or the rate of inflation. These all come under the heading of *descriptive statistics*. Another type of statistics will engage people attention to a much greater extent. That is *inferential statistics* or statistical inference.

c. *Chance* is a necessary part of any process to be described by probability or statistics. Sometimes that element of chance is due partly or even perhaps entirely to lack of knowledge of the details of the process. For example, if one had complete knowledge of the composition of every part of the raw materials used to make bolts, and of the physical processes and conditions in their manufacture, in principle she/he could predict the diameter of each bolt. But in practice one generally lack that complete knowledge, so the diameter of the next bolt to be produced is an unknown quantity described by a random variation. Under these conditions the distribution of diameters can be described by probability and statistics. If one wants to

improve the quality of those bolts and to make them more uniform, one will have to look into the causes of the variation and make changes in the raw materials or the production process. But even after that, there will very likely be a random variation in diameter that can be described statistically. Relations which involve chance are called *probabilistic* or *stochastic* relations. These are contrasted with deterministic relations, in which there is no element of chance. For example, Bernoulli's Law and Newton's Second Law involve no element of chance, so they are deterministic.

d. Another term which requires some discussion is *randomness*. A *random* action cannot be predicted and so is due to chance. A *random sample* is one in which every member of the population has an equal likelihood of appearing. Just which items appear in the sample is determined completely by chance. If some items are more likely to appear in the sample than others, then the sample is not random.

1.1.3. The Engineering Method and Statistical Thinking

An engineer is someone who solves problems of interest to society by the efficient application of scientific principles. Engineers accomplish this by either refining an existing product or process or by designing a new product or process that meets customers' needs. The engineering, or scientific, method is the approach to formulating and solving these problems. The steps in the engineering method are as follows:

1. Develop a clear and concise description of the problem.
2. Identify the important factors that affect this problem or that may play a role in its solution.
3. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied. State any limitations or assumptions of the model.
4. Conduct appropriate experiments and collect data to test or validate the tentative model or conclusions made in steps 2 and 3.
5. Refine the model on the basis of the observed data.
6. Manipulate the model to assist in developing a solution to the problem.
7. Conduct an appropriate experiment to confirm that the proposed solution to the problem is both effective and efficient.
8. Draw conclusions or make recommendations based on the problem solution.

The engineering method steps are shown in Fig. 1.1. Steps 2–4 in Fig. 1.1 are enclosed in a box, indicating that several cycles or iterations of these steps may be required to obtain the final solution. Consequently, engineers must know how to efficiently plan experiments, collect data, analyze and interpret the data, and understand how the observed data are related to the model they have proposed for the problem under study.

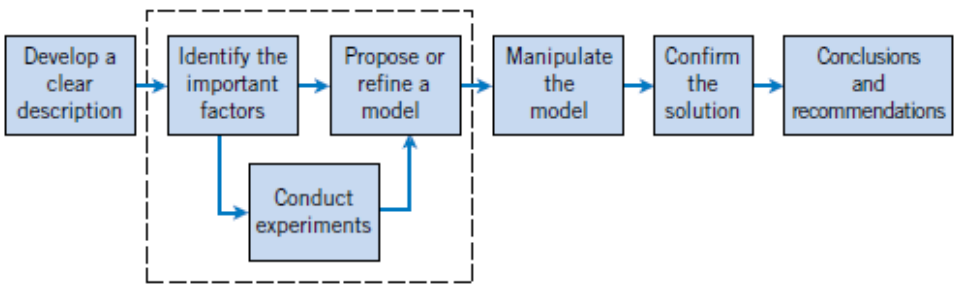


Figure 1.1. The Engineering Method Steps

The field of statistics deals with the collection, presentation, analysis, and use of data to make decisions, solve problems, and design products and processes. Because many aspects of engineering practice involve working with data, obviously some knowledge of statistics is important to any engineer. Specifically, statistical techniques can be a powerful aid in designing new products and systems, improving existing designs; and designing, developing, and improving production processes.

Statistical methods are used to help people describe and understand variability. By variability, one mean that successive observations of a system or phenomenon do not produce exactly the same result. Everybody encounters variability in her/his everyday lives, and statistical thinking can give him/her a useful way to incorporate this variability into decision-making processes. For example, consider the gasoline mileage performance of a car. Do one always gets exactly the same mileage performance on every tank of fuel? Of course not; in fact, sometimes the mileage performance varies considerably. This observed variability in gasoline mileage depends on many factors, such as the type of driving that has occurred most recently (city versus highway), the changes in condition of the vehicle over time (which could include factors such as tire inflation, engine compression, or valve wear), the brand and/or octane number of the gasoline used, or possibly even the weather conditions that have been recently experienced. These factors represent potential sources of variability in the system. Statistics gives people a framework for describing this variability and for learning about which potential sources of variability are the most important or which have the greatest impact on the gasoline mileage performance.

A convenient way to think of a random variable, say *X*, that represents a measurement, is by using the model

$$X = \mu + \varepsilon \tag{1.1}$$

where μ is a constant and ε is a random disturbance. The constant remains the same with every measurement, but small changes in the environment, test equipment, differences in the individual parts themselves, and so forth change the value of ε. If there were no disturbances, ε would always equal zero and *X* would always be equal to the constant μ. However, this never happens in the real world, so the actual measurements *X* exhibit variability. One often needs to describe, quantify and ultimately reduce variability.

1.2. RELIABILITY ENGINEERING

Failures of major engineering systems always raise public concern on the safety and reliability of engineering infrastructure. Decades ago quantitative evaluations of the reliability of complex infrastructure systems were not practical, if not impossible. Engineers had to resort to the use of a safety factor mainly determined through experience and judgment. Without exception, failures of hydrosystem infrastructure (e.g., dams, levees, and storm sewers) could potentially pose significant threats to public safety and inflict enormous damage on properties and the environment. The traditional approach of considering occurrence frequency of heavy rainfalls or floods, along with an arbitrarily chosen safety factor, has been found inadequate for assessing the reliability of hydrosystem infrastructure and for risk-based cost analysis and decision making. In the past two decades or so, there has been a steady growth in the development and application of reliability analysis in hydrosystems engineering and other disciplines.

Hydrosystems is the term used to describe collectively the technical areas of hydrology, hydraulics and water resources. The term has now been widely used to encompass various water resource systems including surface water storage, groundwater, water distribution, flood control, drainage, and others. In many hydrosystem infrastructural engineering and management problems, both quantity and quality aspects of water and other environmental issues have to be addressed simultaneously. Due to the presence of numerous uncertainties, the ability of the system to achieve the goals of design and management decisions cannot be assessed definitely. It is almost mandatory for an engineer involved in major hydrosystem infrastructural design or hazardous waste management to quantify the potential risk of failure and the associated consequences.

Occasionally, failures of engineering systems catch public attention and raise concern over the safety and performance of the systems. The cause of the malfunction and failure could be natural phenomena, human error, or deficiency in design and manufacture. Reliability engineering is a field developed in recent decades to deal with such safety and performance issues.

Based on their setup, engineering systems can be classified loosely into two types, namely, manufactured systems and infrastructural systems. Manufactured systems are those equipment and assemblies, such as pumping stations, cars, computers, airplanes, bulldozers, and tractors, that are designed, fabricated, operated, and moved around totally by humans. Infrastructural systems are the structures or facilities, such as bridges, buildings, dams, roads, levees, sewers, pipelines, power plants, and coastal and offshore structures, that are built on, attached to, or associated with the ground or earth. Most civil, environmental, and agricultural engineering systems belong to infrastructural systems, whereas the great majority of electronic, mechanical, industrial, and aeronautical/aerospace engineering systems are manufactured systems.

The major causes of failure for these two types of systems are different. Failure of infrastructures usually is caused by natural processes, such as geophysical extremes of earthquakes, tornadoes, hurricanes or typhoons, heavy rain or snow, and floods, that are beyond human control. Failure of such infrastructural systems seldom happens, but if a failure occurs, the consequences often are disastrous. Replacement after failure, if feasible, usually involves so many changes and improvements that it is essentially a different, new system. On the other hand, the major causes of failure for manufactured systems are wear and tear, deterioration, and improper operation, which could be dealt with by human abilities but may not be economically desirable. Their failures usually do not result in extended major calamity. If failed, they can be repaired or replaced without affecting their service environment.

The performance of a hydrosystem engineering infrastructure, function of an engineering project, or completion of an operation all involve a number of contributing components, and most of them, if not all, are subject to various types of uncertainty (Fig. 1.2). Reliability and risk, on the other hand, generally are associated with the system as a whole. Thus methods to account for the component uncertainties and to combine them are required to yield the system reliability. Such methods usually involve the use of a logic tree. The reliability of an engineering system may be considered casually, such as through the use of a subjectively decided factor of safety. Today, reliability also may be handled in a more comprehensive and systematic manner through the aid of probability theory.

The basic idea of reliability engineering is to determine the failure probability of an engineering system, from which the safety of the system can be assessed or a rational decision can be made on the design, operation, or forecasting of the system, as depicted in Fig. 1.3.

An infrastructure is a functioning system formed from a combination of a number of components. From the perspective of reliability analysis, infrastructure systems can be classified in several ways. Infrastructures may follow different paths to failure. The ideal and simplest type is the case that the resistance and loading of the system are statistically independent of time, or a stationary system. Most of the existing reliability analysis methods have been developed for such a case. A more complicated but realistic case is that for which the statistical characteristics of the loading or resistance or both are changing with time, e.g., floods from a watershed under urbanization, rainfall under the effect of global warming, sewer or water supply pipes with deposition, and fatigue or elastic behavior of steel structure members. For some infrastructures, the statistical characteristics of the system change with space or in time (or both), e.g., a reach of highway or levee along different terrains.

1.3. DEFINITIONS OF RELIABILITY AND RISK

In view of the lack of generally accepted rigorous definitions for risk and reliability, it will be helpful to define these two terms in a manner amenable to mathematical formulation for their quantitative evaluation for engineering systems. Risk is defined as the probability of failure to achieve the intended goal. Reliability is defined mathematically as the complement of the risk. In some disciplines, often the non-engineering ones, the word risk refers not just to the probability of failure but also to the consequence of that failure, such as the cost associated with the failure. Nevertheless, to avoid possible confusion, the mathematical analysis of risk and reliability is termed herein reliability analysis.

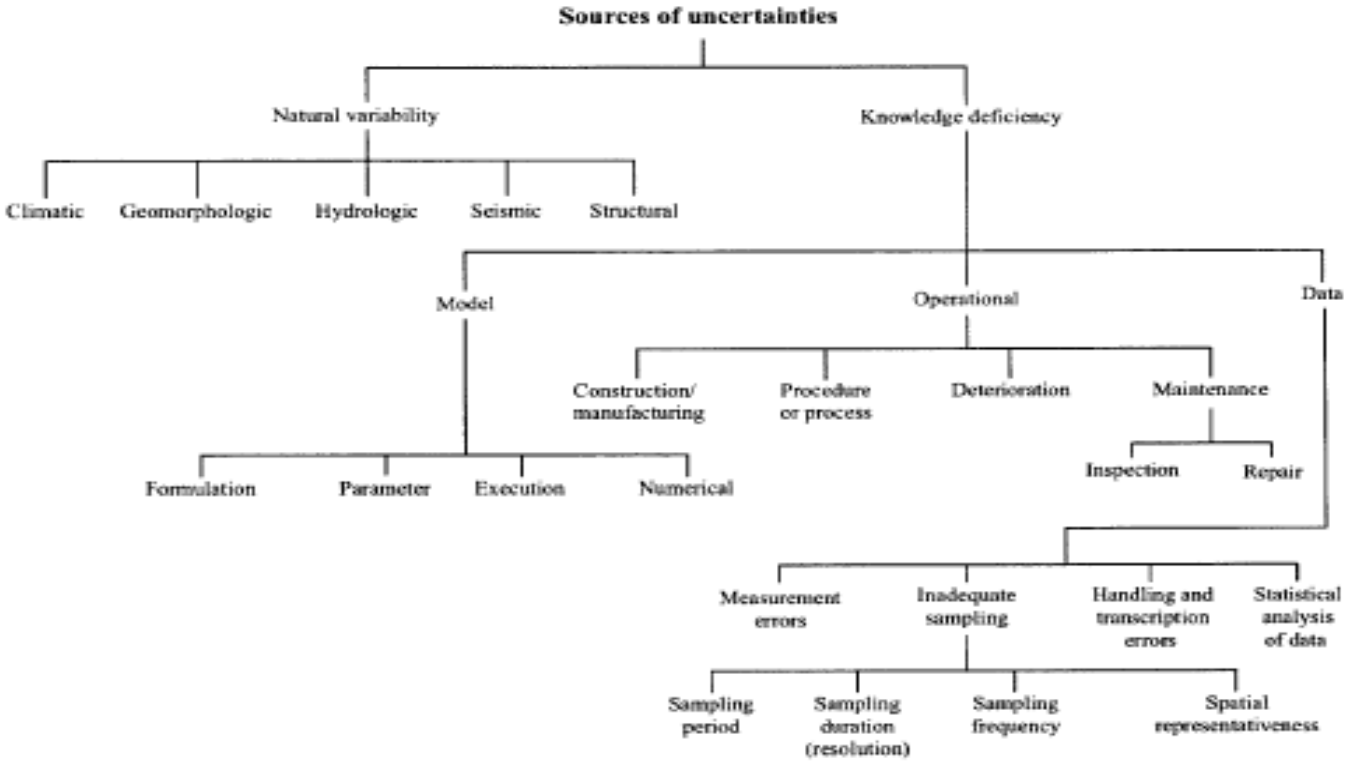


Figure 1.2. Sources of Uncertainties

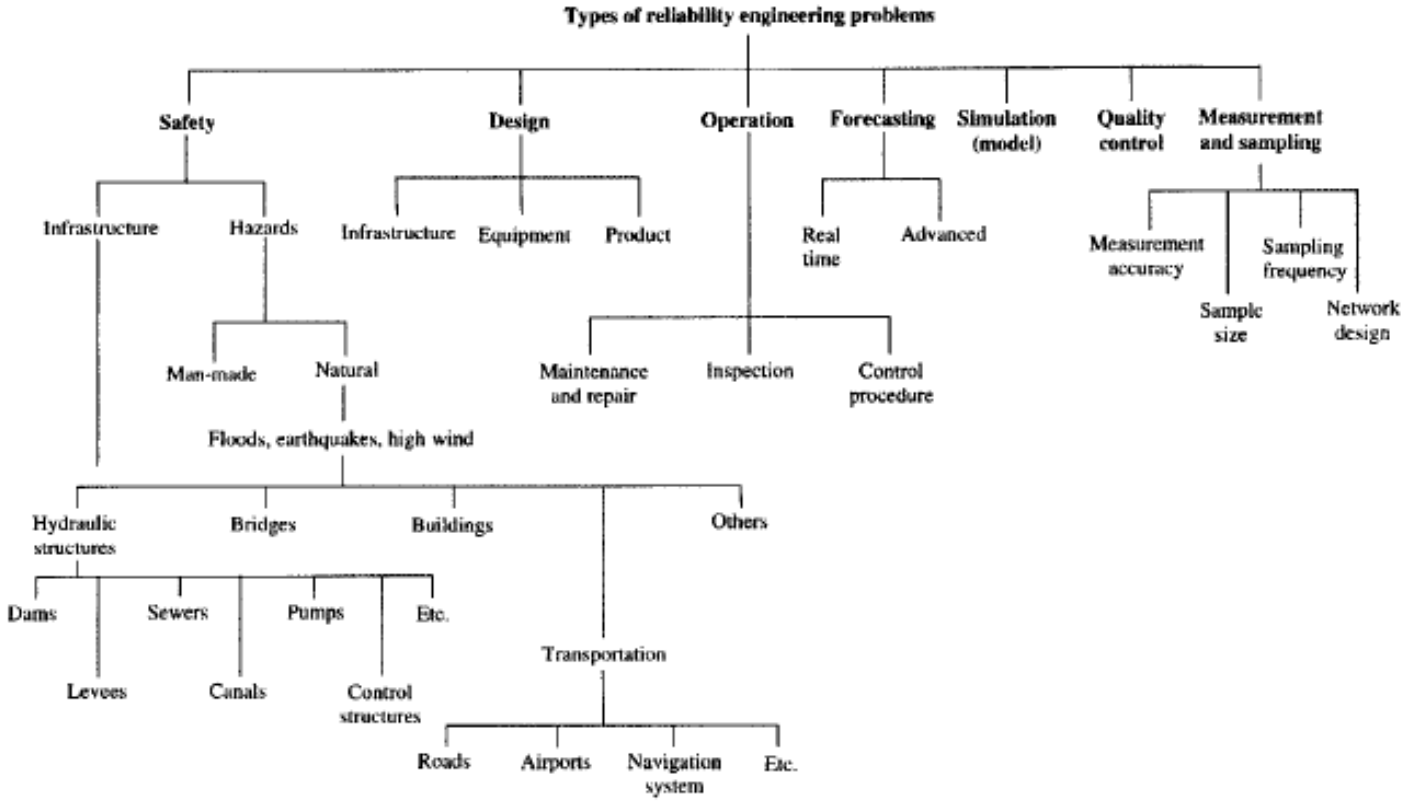


Figure 1.3. Types of Reliability Engineering Problems

Failure of an engineering system can be defined as a situation in which the load L (external forces or demands) on the system exceeds the resistance R (strength, capacity, or supply) of the system. The reliability p_s of an engineering system is defined as the probability of non-failure in which the resistance of the system exceeds the load; that is,

$$p_s = P(L \le R) \tag{1.2a}$$

in which P (·) denotes probability. Conversely, the risk is the probability of failure when the load exceeds the resistance. Thus the failure probability (risk) p_f can be expressed mathematically as

$$p_f = P(L > R) = 1 - p_s \tag{1.2b}$$

1.4. MEASURES OF RELIABILITY

In engineering design and analysis, loads usually arise from natural events, such as floods, storms, or earthquakes, that occur randomly in time and in space. The conventional practice for measuring the reliability of a hydrosystems engineering infrastructure is the return period or recurrence interval (T). The return period is defined as the long-term average (or expected) time between two successive failure-causing events. Simplistically, the return period is equal to the reciprocal of the probability of the occurrence of the event in any one-time interval (T=1/p).

In fact, the conventional interpretation of return period can be generalized as the average time period or mean time of the system failure when all uncertainties affecting load and resistance are considered. In other words, the return period can be calculated as the reciprocal of the failure probability computed by Eq. (1.1b). Two other types of reliability measures that consider the relative magnitudes of resistance and anticipated load (called design load) are used frequently in engineering practice. One is the safety margin (SM), defined as the difference between the resistance and the anticipated load, that is,

$$SM = R - L \tag{1.3a}$$

The other is called the safety factor (SF), a ratio of resistance to load defined as

$$SF = R / L \tag{1.3b}$$

CHAPTER 2

BASIC CONCEPTS OF STATISTICS

2.1. INTRODUCTION

2.1.1. Definition of Statistics

Statistics can be defined in two senses: Plural (as Statistical Data) and singular (as Statistical Methods).

Plural Sense: Statistics are *collection of facts (figures).* This meaning of the word is widely used when reference is made to facts and figures on sales, employment or unemployment, accident, weather, death, education, etc. E.g.: Sales Statistics, Labor Statistics, Employment Statistics, etc. In this sense the word Statistics serves simply as *data*. But not all numerical data are statistics.

Singular sense: Statistics is the science that deals with the methods of data collection, organization, presentation, analysis and interpretation of data. It refers the subject area that is concerned with *extracting relevant information from available data* with the aim to make sound decisions. According to this meaning, statistics is concerned with the development and application of methods and techniques for collecting, organizing, presenting, analyzing and interpreting statistical data.

2.1.2. Science of Statistics

Statistics is a branch of applied mathematics, which tries to solve the problems about random variables. The goal of the statistics is to provide right (true) conclusions by using insufficient (deficient) data and information. The events may be classified as *deterministic* (certain) and *probabilistic* (random) events. The scope of the statistics science is only probabilistic events. A random or probabilistic event is described as, an event which can not be certainly determined whether or not to occur, or if occurs, which value it will take. Engineering problems have generally random characteristics. For example, it is impossible to certainly predict that how much precipitation will take place next year in a location, this value may be predicted by some probabilities (80, 90 or 99 percent).

The purpose of statistics is to develop and apply methodology for extracting useful knowledge from experiments, measurements and data. Statisticians provide crucial guidance in determining what information is reliable and which predictions can be trusted. There is a general perception that statistical knowledge is frequently intentionally misused, by finding ways to interpret the data that are favorable to the presenter. A famous quote is "There are three types of lies; lies, damn lies, and statistics."

2.1.3. Classification of Statistics

Based on the scope of the decision, statistics can be classified into two; Descriptive and Inferential Statistics.

Descriptive Statistics refers to the procedures used to *organize and summarize masses of data*. It is concerned with describing or summarizing the most important features of the data. It deals only the characteristics of the collected data without going beyond it. That is, this part deals with only describing the data collected without going any further: that is without attempting to infer(conclude) anything that goes beyond the data themselves. The methodology of descriptive statistics includes the methods of organizing (classification, tabulation, Frequency Distributions) and presenting (Graphical and Diagrammatic

Presentation) data and calculations of certain indicators of data like Measures of Central Tendency and Measures of Dispersion (Variation) which summarize some important features of the data.

Inferential (Inductive) Statistics includes the methods used to find out something about a population, based on the sample. It is concerned with *drawing statistically valid conclusions about the characteristics of the population based on information obtained from sample*. In this form of statistical analysis, *inferential statistics is linked with probability theory* in order to generalize the results of the sample to the population. Performing *hypothesis testing*, determining *relationships between variables* and making *predictions* are also inferential statistics.

2.1.4. Stages in Statistical Investigation

According to the singular sense definition of statistics, a statistical study (statistical investigation) involves five stages: Collection, Organization, Presentation, Analysis and Interpretation of data.

a. Collection of Data: This is the first stage in any statistical investigation and involves the process of obtaining (gathering) a set of related measurements or counts to meet predetermined objectives. The data collected may be primary data (data collected directly by the investigator) or it may be secondary data (data obtained from intermediate sources such as newspaper s, journals, official records, etc).

b. Organization of Data: It is usually not possible to derive any conclusion about the main features of the data from direct inspection of the observations. The second purpose of statistics is *describing the properties of the data in a summary form*. This stage of statistical investigation helps to have a clear understanding of the information gathered and includes editing (correcting), classifying and tabulating the collected data in a systematic manner. Thus the first step in the organization of data is *editing*. It means correcting (adjusting) omissions, inconsistencies, irrelevant answers and wrong computations in the collected data. The second step of the organization of data is *classification* that is arranging the collected data according to some common characteristics. The last step of the organization of data is presenting the classified data in tabular form, using rows and columns (*tabulation*).

c. Presentation of Data: The purpose of data presentation is to have an overview of what the data actually looks like, and to facilitate statistical analysis. Data presentation can be done using Graphs and Diagrams which have great memorizing effect and facilitates comparison.

d. Analysis of Data: The analysis of data is the *extraction of summarized and comprehensive numerical description* in order to reach conclusions or provide answers to a problem. The problem may require simple or sophisticated mathematical expressions.

e. Interpretation of Data: This is the last stage of statistical investigation. Interpretation involves drawing conclusions from the data collected and analyzed in order to make decision.

2.1.5. Definition of Some Statistical Terms

A list of terms to be used in statistics are briefly defined as follows:

Population is a community, which is made of all of the components with a particular character. A population is a totality of things, objects, peoples, etc about which information is being

Sample is a group of components which are supposed to represent the population. A sample is a subset or part of a population selected to draw conclusions about the population.

Sampling is the process of selecting a sample from the population.

Parameter is numerical values of population. Parameter is a descriptive measure (value) computed from the population. It is the population measurement used to describe the population.

Statistics is numerical values of sample. It is a measure used to describe the sample; it is a value computed from the sample. The parameter and statistic values are generally different. One of the main goals of statistics science is to predict the statistic values as close as possible to parameter value.

Census Survey is the process of examining the entire population. It is the total count of the population.

Sampling Frame is a list of people, items or units from which the sample is taken.

Data are a collection of related facts and figures from which conclusions may be drawn.

Variable is a certain characteristic which changes from object to object and time to time.

Sample Size is the number of elements or observation to be included in the sample.

Collected is the totality of observations with which the researcher is concerned.

2.1.6. Applications, Uses and Limitations of Statistics

a. Applications of Statistics in Engineering:

In this modern time, statistical information plays a very important role in a wide range of fields. Today, statistics is applied in almost all fields of human endeavor:

- In Scientific Research: Statistics is used as a tool in a scientific research. Statistical formulas and concepts are applied on a data which are results of an experiment.
- In Quality Control: Statistical methods help to check whether a product satisfies a given standard.
- For Decision Making: Statistics helps to enhance the power of decision making in the face of uncertainty by providing sufficient information.
- Reliability Engineering is the study of the ability of a system or component to perform its required functions under stated conditions for a specified period of time.
- The application of probability theory, which includes mathematical tools for dealing with large populations, to the field of mechanics, which is concerned with the motion of particles or objects when subjected to a force.
- The field of statistics deals with the collection, presentation, analysis, and use of data to: Such as make decisions, solve problems and design products and processes. It is the science of learning information from data.

b. Uses of Statistics in Engineering:

- Design of Experiments (DOE) uses statistical techniques to test and construct models of engineering components and systems.
- Quality control and process control use statistics as a tool to manage conformance to specifications of manufacturing processes and their products.
- Time and methods engineering uses statistics to study repetitive operations in manufacturing in order to set standards and find optimum (in some sense) manufacturing procedures.
- Reliability engineering uses statistics to measures the ability of a system to perform for its intended function (and time) and has tools for improving performance.
- Probabilistic design uses statistics in the use of probability in product and system design.
- Every structural design, every safety factor, every hydrological analysis, every mechanical analysis, everything, even the materials used are based on statistics. The results gotten from the analysis are projected to other conditions, and the probability of them to interact together (for example, earthquake, wind and max load or having the highest flow and rain).
- Condenses and summarizes masses of data and presents facts in numerical and definite form.

- Facilitates comparison: Statistical devises such as averages, percentages, ratios, etc. are used for this purpose.
- Formulating and testing hypothesis.
- Forecasting: Statistical methods help in studying past data and predicting future trends.

c. Limitations of Statistics:

- It cannot deal with a single observation; rather it deals aggregate of facts.
- Statistical methods are not applicable to qualitative character; it deals only with quantitative characteristics.
- Statistical results are true on average; i.e. for the majority of case. Laws of statistics are not universally true like the laws of physics, chemistry and mathematics.
- Statistics are liable to be misused or misinterpreted. This may be due to incomplete information, inadequate and faulty procedures during data collection and sample selection and mainly due to ignorance (lack of knowledge).

2.2. FREQUENCY ANALYSIS

2.2.1. Statistical Sample

The population, consisting of all the observations belonging to a random variable should be observed in order to determine exactly the probability distribution of the random variables. However, in practise only a statistical sample, which has finite number of elements, can be obtained from the population. A sample is a set of observations or measurements collected to determine or estimate the statistical properties of a random variable. Each element in the sample is an event, belonging to the random variable or is a value the random variable has taken. Since the probability distribution function of the random variable and the parameters of this distribution can only be estimated depending on the limited sample in hand, a sample must be analyzed in an optimum way. Statistics is the science which obtains all possible information from samples and arrives at conclusions about the statistical properties of the population, by using the obtained information. The samples to be used in statistical studies should be adequate both *qualitatively* and *quantitatively*.

For qualitatively adequacy, a sample should realize the following conditions:

The data in the sample should be *homogenous*; in other words, all data should indeed be elements of the population of the same random variable. Otherwise, the statistical calculations will have no significance. For example, in the case the flow of a river is controlled by a dam; it would not be correct to evaluate the flows downstream of the dam, measured before and after the dam construction, because these flows would not be homogenous.

- There should be no *systematic error* in the measurement of the elements of the sample. In order to meet this condition, the people responsible for sampling should be aware of both subject and material; also, the measuring techniques to obtain the data should be adequate.
- *Random errors* should be minimized. This necessity can partly be realized by random (neutral) sampling within the population.

The sample being quantitatively adequate means that, the number of elements in the sample is sufficiently large. Although an exact limit can not be given for the *sufficient number*, it can easily be said that, as the number of elements in the sample increase, more reliable results about the properties of the population can be obtained. In statistics, samples having less than 25 to 30 elements are called *small samples* and in some cases it is not suitable to use expressions that are valid for large samples in the analysis of small samples.

2.2.2. Types of Variables

Variables can be classified as *quantitative* and *qualitative* variables. If a variable can be stated by a number, it is a quantitative (numerical) variable. The other variables, which have no numerical character, are qualitative variables. Since in statistics only quantitative variables are used, different classification as *discrete* and *continuous* variables is made.

Discrete Variable: The variables of which number of components is limited (with small sample space) are called as discrete variables. All of the qualitative variables are discrete. For example, the names of people, colours of cars are discrete variables. The quantitative variables, which can be stated only by integer numbers, in other words, the variables that are obtained by enumerating (1, 2, 3, ..) are also discrete variables. For example, the numbers of students in a class or rainy days in a year are discrete variables.

Continuous Variable: If the number of components of a variable is unlimited or infinitive (with large sample space), this variable is named as continuous. The quantitative variables that are obtained by measuring are continuous and may have any fractional number (2.34, 145.036, …). Discharges of a stream, weights and lengths of people are examples of continuous variables.

2.2.3. Frequency Analysis of Discrete Variables

In a sample with N component, if an $X = x_i$ event is occurred N_i times, its frequency is

$$f(x_i) = f_i = N_i / N \tag{2.1}$$

If these $f(x_i)$ frequency values are plotted in ordinate and x_i values are shown in abscissa, then *"frequency graph or histogram"* of the event is obtained. By adding or cumulating the frequency values and plotting them versus x_i values, then *"cumulative frequency graph"* is obtained. Cumulative frequencies are calculated as follows:

$$F(x_i) = \sum_{j=1}^{i} N_j / N = \sum_{j=1}^{i} f(x_j) \tag{2.2}$$

2.2.4. Frequency Analysis of Continuous Variables

Since total number of a continuous random variable is theoretically infinitive (in other words, sample space is very large), and total probability is equal to unity, the probability of a simple event is approximated to zero ($1/\infty \cong 0$). So, the probability of combined events between (x) and $(x+dx)$ interval is calculated, instead of probability of a simple event, and a *probability density function (p.d.f)* is defined as (Figure 2.1a):

$$f(x)dx = P\left(x \le X \le x + dx\right) \tag{2.3}$$

The probability of X being between $(x_1 - x_2)$ interval is calculated as

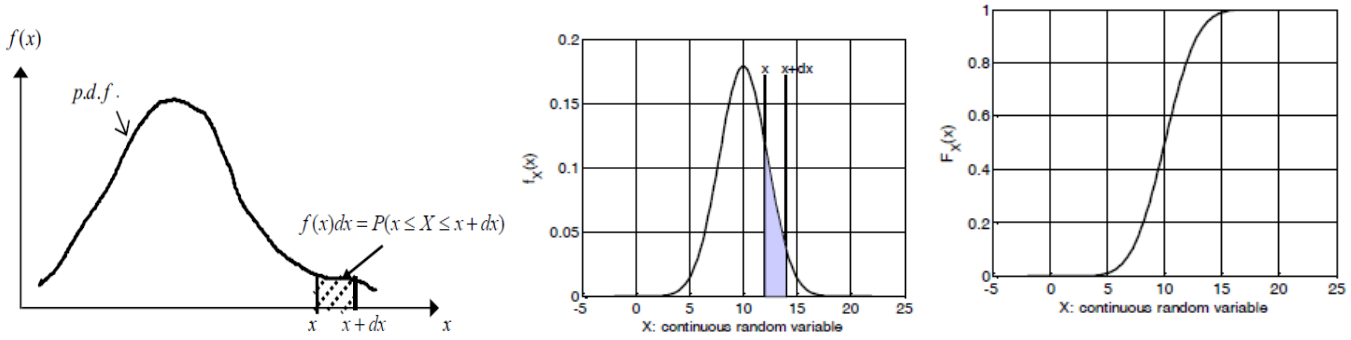
$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)dx \tag{2.4}$$

The cumulative distribution function of a real-valued random variable is the function given by (Figure 2.1b):

$$F_x(x) = P(X \le x) \tag{2.5}$$

By differentiating this function, the probability density function (pdf) is obtained.

$$f_x(x) = \frac{dF_x(x)}{dx} \tag{2.6}$$



a. Probability Density Functions b. Cumulative Frequency Distribution

Figure 2.1. Probability Density and Cumulative Frequency Distributions

Frequency analysis of a continuous random variable can be performed in two ways depending on the size of the sample; whether it is a small or a large sample.

2.2.4.1. Frequency Analysis of Large Samples:

If the sample is large (at least have 25 to 30 samples), because of some difficulties in processing all of the data, the range of the random variable is divided into appropriate number of *class intervals*. An important point is to choose the number o class intervals (m). This number should be increased as the number of elements in the sample increases. Generally, the number of class intervals is kept between 10 and 20. If too few intervals are used in the analysis, a large amount of the information in the samples may be lost. On the other hand, if too many intervals are used then both more effort than required will be needed and the histogram will have a quite irregular shape, because very few or no observations will fall into some class intervals. The following empirical formula can be used to determine the number of class intervals:

$$m \cong 1 + 3.3 \log_{10} N \tag{2.7}$$

2.2.4.2. Frequency Analysis of Small Samples:

If the number of elements in the sample is small, it is not appropriate to classify the data. In this case, the objective is to determine the cumulative frequency distribution only. The *ordered sample* is obtained by listing the elements of the sample from smaller values to large values as:

$$x_1 \le x_2 \le \ldots \le x_m \le \ldots \le x_N \tag{2.8}$$

One can calculate the frequency of the random variable remaining equal to or smaller than x_m as:

$$P(X \le x_m = F(x_m) = m/N \tag{2.9a}$$

However, if this equation is applied, the frequency remaining equal to or smaller than the x_N, the frequency value is equal to 1. Since elements greater than the x_N value may exist, it is not correct to use this expression, which implies that the random variable (X) would never exceed the x_N . Various formulas called *plotting position formulas* have been proposed to eliminate this inconvenient aspect. The most popular among them is as follows:

$$F(x_m) = m/(N+1) \tag{2.9b}$$

2.2.5. Parameters of Random Variables

The numbers, which express certain properties of the random variable's distribution function, are called the *parameters of the distribution*. Their estimation from the data is much easier compared to the estimation and use of the distribution function.

2.2.5.1. Central Parameters

a. The Mean: The most commonly used central parameter is called *mean (expected value, arithmetic mean).* The mean in a population is calculated as:

$$\mu = \sum x / \sum N \tag{2.10a}$$

In a sample, the mean is calculated as:

$$\bar{x} = \sum x / N \tag{2.10b}$$

If the data are classified, then *weighted mean* is employed:

$$\bar{x} = \sum (x_i N_i) / N \tag{2.10c}$$

b. The Mode: The mode is the most frequently occurring score value.

c. The Median: The median is the middle of a distribution: half the scores are above the median and half are below the median.

d. The Geometric Mean: The geometric mean is the N th root of the product of the scores and is given as follows:

$$G.M = \sqrt[N]{x_1 * x_2 * ..... * x_n} \tag{2.11}$$

e. The Harmonic Mean: Harmonic mean is used in calculating the mean slope of a highway, stream etc:

$$H.M. = \frac{N}{1/x_1 + 1/x_2 + ... + 1/x_N} \tag{2.12}$$

2.2.5.2. Variation Parameters

A deviation score is a measure of by how much each point in a frequency distribution lies above or below the mean for the entire dataset as $x - \bar{x}$, where x is raw score and $\bar{x}$ is the mean. A list of v*ariation parameters*, which are used determine variations of all of the data around the centre of distribution, are presented in the following:

a. The Range: The range is the difference between the maximum and minimum values of a series. Example: The range of 35, 12, 34, 76, 87, 39, 48 is 87-12=75.

b. The Variance and Standard Deviation: *The variance* and the closely-related *standard deviation* are measures of how spread out a distribution is. In other words, they are measures of variability. In order to define the amount of deviation of a dataset from the mean, calculate the mean of all the deviation scores, i.e. the variance. The variance is computed as the average squared deviation of each number from its mean. For a continuous variance is defined as:

$$Var_x = \int (x - \bar{x})^2 * f(x) dx \tag{2.13a}$$

and is calculated by a sample as:

$$Var_x = \sum (x_i - \bar{x})^2 / N = \sum x_i^{\,2} / N - \bar{x}^2 \tag{2.13b}$$

If the number of data is less than 30 $(N < 30)$, in this equation $(N - 1)$ is used instead of (N).If the data is classified, the variance is calculated as:

$$Var_x = \sum\left[(x_i - \bar{x})^2 * N_i\right] / N \qquad (2.13c)$$

Since the variance has the square dimension of the variation, in order to make dimensional homogeneity, the square root of variance, the standard deviation, is commonly used.

$$S_x = \sqrt{Var_x} \qquad (2.13d)$$

c. Coefficient of Variance: It would not be proper to compare the standard deviations of two variables with different averages in order to understand which variable has more variation; therefore, a dimensionless coefficient, *coefficient of variance* is calculated as follows:

$$V.C. = S_x / \bar{x} \qquad (2.14)$$

2.2.5.3. Skewness Parameter

There are many different-shaped frequency distributions: J-shaped, Normal (Symmetrical), Rectangular, Bimodal, Positive (Right) and Negative (Left) Skew (Figure 2.2).

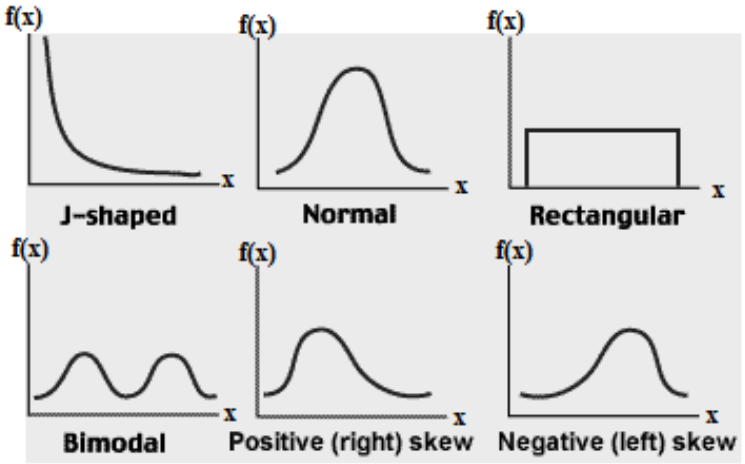


Figure 2.2 Various Shaped Frequency Distributions

Skewness coefficient is employed to determine the shape of whether the distribution is symmetrical, right or left skew. Skewness coefficient is positive, negative and zero for right and left skewed and symmetrical distributions, respectively. Skewness coefficient is calculated as follows:

$$C_s = \frac{N^2}{(N-1)(N-2)} * \frac{m_3}{m_2^{1.5}} \qquad (2.15a)$$

Where, m_2 is the second central moment (variance), m_3 (the third central moment) is calculated:

$$m_3 = \sum (x_i - \bar{x})^3 / N \qquad (2.15b)$$

If the data are classified, m_3 is found as:

$$m_3 = \left\{\sum\left[(x_i - \bar{x})^3 N_i\right]\right\} / N \qquad (2.15c)$$

2.3. EXAMPLES

Example 2.1: Calculate the harmonic mean, standard deviation and skewness coefficient for the following data. 25, 34, 43, 19, 36, 26, 38, 17, 25

Solution:

										$\sum$
x										
$x-\bar{x}$										
$(x-\bar{x})^2$										
$(x-\bar{x})^3$										

Example 2.2: 32 yearly mean discharge values of a stream (m^3/s) are given as follows. By taking class interval as 3 m^3/s, classify the data, obtain the frequency values and considering mid values of all classes, calculate the mean and standard deviation:

28 19 16 11 19 20 17 15 13 16 24 13 18 20 23 20

15 13 10 17 21 19 18 24 12 21 25 26 13 18 27 14

Solution:

GROUP								$\sum$
N_i								
f_i								
x_i								
$x_i N_i$								
$(x_i-\bar{x})^2 * N_i$								

Example 2.3: The following data are the mean annual precipitation heights (cm) of a city. Calculate the mean, standard deviation and skewness coefficient by

a. Not classifying the data.

68 45 78 74 54 98 87 74 90 75 79 67 80 83 85 92 82 73 87 82 73 90 85 77 73 78 85 95 74 79 88 87

Solution:

N_i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$		N_i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$
1	68					17	82			
2	45					18	73			
3	78					19	87			
4	74					20	82			
5	54					21	73			
6	98					22	90			
7	87					23	85			
8	74					24	77			
9	90					25	73			
10	75					26	78			
11	79					27	85			
12	67					28	95			
13	80					29	74			
14	83					30	79			
15	85					31	88			
16	92					32	87			
TOTAL										

b. Classifying the data into 10 groups as (42-47), (48-53) ….. (98-103) and taking into consideration of mid values of each group.

Solution:

Group	N_i	Mid. Val. = x_i	x_iN_i	$(x_i - x)^2 * N_i$	$(x_i - x)^3 * N_i$

c. Calculate and draw the frequency histogram for the classified data.

Solution:

Group										
N_i										
$f_i = N_i/N)*100$										
$\sum$										

CHAPTER 3

PROBABILITY

3.1. CONCEPT OF PROBABILITY

3.1.1. Definitions

Probability, as a specific term, is a measure of the likelihood that a particular event will occur. If one is certain that an event will occur, its probability is 1 or 100%. If it certainly will not occur, its probability is zero. The first situation corresponds to an event which occurs in every trial, whereas the second corresponds to an event which never occurs. At this point one might be tempted to say that probability is given by relative frequency, the fraction of all the trials in a particular experiment that give an outcome meeting the stated requirements. But in general that would not be right. Why? Because the outcome of each trial is determined by chance. Say one toss a fair coin, one which is just as likely to give heads as tails. It is entirely possible that six tosses of the coin would give six heads or six tails, or anything in between, so the relative frequency of heads would vary from zero to one. If it is just as likely that an event will occur as that it will not occur, its true probability is 0.5 or 50%. But the experiment might well result in relative frequencies all the way from zero to one. Then the relative frequency from a small number of trials gives a very unreliable indication of probability. As an illustration, suppose the weather man on TV says that for a particular region the probability of precipitation tomorrow is 40%. Let us consider 100 days which have the same set of relevant conditions as prevailed at the time of the forecast. According to the prediction, precipitation the next day would occur at any point in the region in about 40 of the 100 trials. (This is what the weather man predicts, but we all know that the weather man is not always right!)

Although one cannot make an infinite number of trials, in practice she/he can make a moderate number of trials, and that will give some useful information. The *relative frequency* of a particular event, or the proportion of trials giving outcomes which meet certain requirements, will give an *estimate* of the probability of that event. The larger the number of trials, the more reliable that estimate will be. Another type of probability is the subjective estimate, based on a person's experience. To illustrate this, say a geological engineer examines extensive geological information on a particular property. He chooses the best site to drill an oil well, and he states that on the basis of his previous experience he estimates that the probability the well will be successful is 30%. (Another experienced geological engineer using the same information might well come to a different estimate.) This, then, is a subjective estimate of probability. The executives of the company can use this estimate to decide whether to drill the well.

Another approach is possible in certain cases. This includes various gambling games, such as tossing an unbiased coin; drawing a colored ball from a number of balls, identical except for color, which are put into a bag and thoroughly mixed; throwing an unbiased die; In each of these cases we can say before the trial that a number of possible results are *equally likely*. This is the *classical* or *a priori* approach. The phrase "a priori" comes from Latin words meaning coming from what was known before. This approach is often simple to visualize, so giving a better understanding of probability.

3.1.2. Basic Principles

Random variables cannot be studied by a deterministic approach; instead, a probabilistic approach is required in their analysis. If a random variable is studied, the outcome of a future event can never be determined *with certainty*. It is possible only to estimate the *chance* of the variable to assume a certain value. The chance of the occurrence of a random variable is defined its *probability*. Total number of all of the observations in an event or experiment is called *total frequency*.

The expected (anticipated) value of an event under normal conditions is named as *expected frequency*. The probability of the event is determined as (expected frequency)/(total frequency). For example, when a coin is tossed (thrown), under normal conditions the probabilities of head and tail are equal. If a coin is tossed 10 times, it is expected that 5 times head and 5 times tail to occur; the probability of either head or tail is $5/10 = 0.5$. Denoting the random variable by a capital letter and its value in an observation by the corresponding small letter, it can be written:

$$P(X = x_i) = p_i \tag{3.1}$$

As it was previously explained, the frequency of an event is $f(x_i) = f_i = N_i / N$. The probability of this event is defined as limit of its frequency as the number of observations approaches infinity:

$$p_i = \underset{N \to \infty}{Lim}\left(N_i / N\right) \tag{3.2}$$

For example, the probability of head (or tail) is 0.5, which does not mean that in 10 tossing; certainly 5 heads and 5 tails will be observed. However, if the tossing is repeated more and more times, it is expected that the numbers of heads and tails will approach to each other. In 1 000 000 tossing for example, it is highly probable that nearly 500 000 heads and 500 000 tails will occur.

The basic axiom of the probability theory states that each random event has a certain probability that varies in the range of 0 to 1. $p_i = 0$ implies that the event $X = x_i$ will never occur (impossible), $p_i = 1$ means that the event will occur certainly (in all observations). For example, in a dice throwing, the probability of a number between 1 and 6 is 1, and the probability of 0 or 7 is 0. The occurrence of an event is called as *success* and non-occurrence is called as *failure*. If the probabilities of success and failure are displayed as p and q, the equations can be written as:

$$p + q = 1, \quad p = 1 - q, \quad q = 1 - p \tag{3.3}$$

For example, the success and failure probabilities 2 in a dice are 1/6 and 1-1/6=5/6.

3.2. BASIC RULES OF COMBINING PROBABILITIES

3.2.1. Joint, Disjoint, Independent and Dependent Events

In the case of compound (combined) events, which have two or more simple events, if occurrence of one of these events does not prevent the other(s) from occurring, these events are called as *joint events*. In other words, if these events can occur simultaneously, they are joint events. Joint events have one or more than one *joint* components. For example, in a dice tossing event, the events of (an odd number) and (a number less than 4) are joint events; because they have joint (partner) components; as is seen, the numbers of 1 and 3 are joint components. If occurrence of one of these events prevents the other(s) from occurring, these events are called as dis*joint (mutually exclusive) events*. In other words, if these events can not occur simultaneously, they are disjoint events. Disjoint events have no *joint* component. For example, in a dice tossing event, the events of (an odd number) and (an even number) are disjoint events; because they have not joint (partner) components.

In a compound event, if occurrence of an event does not affect the occurrence of the other events, these events are called as *independent events*. For example, in the event of tossing two coins, the occurrence of tail in each tossing does not affect the other and therefore these events are independent. If the occurrence of the events affects each other, the events are *dependent*.

3.2.2. Addition Rule ("or")

a. Disjoint (Mutually Exclusive) Events: In the case of two disjoint events, the occurrence probability of *(one or another)*, is the sum of their individually occurrence probabilities. The equation is:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) \tag{3.4a}$$

For example, in a dice tossing event, the probability of (3 or 4) is

$P(3 \cup 4) = P(3) + P(4) = 1/6 + 1/6 = 1/3$

In case of three or more events, the same rule is valid. In three disjoint events the rule is:

$$P(A \text{ or } B \text{ or } C) = P(A \cup B \cup C) = P(A) + P(B) + P(C) \tag{3.4b}$$

For example, the probability of even number in a dice is

$P(even) = P(2) + P(4) + P(6) = 1/6 + 1/6 + 1/6 = 1/2$

b. Joint Events: In the case of two joint events, the occurrence probability of *(one or another)*, is obtained by subtracting the probability of joint event from the sum of their individually occurrence probabilities. In Figure 3.2b, the following equation may be obtained:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.5a}$$

For example: In a dice, calculate the probability of (an odd number) or (a number less than 3)

Solution: $P(\text{ odd number}) = P(A) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2,$

$(P \text{ less than } 3) = P(B) = P(1) + P(2) = 1/6 + 1/6 = 1/3,\ P(A \cap B) = P(1) = 1/6$

From Equation (2.18a) $P(A \text{ or } B) = 1/2 + 1/3 - 1/6 = 2/3$

In the case of three joint events, the probability is calculated as follows:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \tag{3.5b}$$

3.2.3. Multiplication Rule ("and")

a. Independent Events: The probability of independent events is equal to multiplication of their individual probabilities. In other words; the probability of occurrence of more than one event together is the *product* of the probabilities of the separate events.

$$P(A \text{ and } B) = P(A \cap B) = P(A) * P(B) \tag{3.6a}$$

$$P(A \text{ and } B \text{ and } C) = P(A \cap B \cap C) = P(A) * P(B) * P(C) \tag{3.6b}$$

For example, if a dice is thrown twice, the probability of 6 in both event is $P(6,6) = 1/6 * 1/6 = 1/36$.

In the case of independent events more than 3, the same rule is valid.

b. Dependent Events: If the events are *not independent*, one event affects the probability for the other event. In this case *conditional probability* must be used. The conditional probability of B given that A occurs, or *on condition* that A occurs, is written P(B/A). This is read as the probability of B given A, or the probability of B on condition that A occurs. Conditional probability can be found by considering only those events which meet the condition, which in this case is that A occurs. Among these events, the probability that B occurs is given by the conditional probability, P(B/A).

The multiplication rule for the occurrence of both A *and* B together when they are not independent is the product of the probability of one event and the conditional probability of the other:

$$P(A \cap B) = P(A) * P(B/A) = P(B) * P(A/B) \tag{3.7a}$$

This implies that conditional probability can be obtained by:

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \; P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{3.7b}$$

Example: Knowing that when a dice is tossed, the number on top is greater than 2, what is the probability that the number on the top of the dice is an even number?

Solution: The event that number is greater than 2 is A and the event of even number is B, then

A is (3, 4, 5, 6) number of elements of A is 4 and P(A) = 4/6,

B is (2, 4, 6), number of elements of B is 2 and P(B) = 2/6,

Event of $(A \cap B)$ = (4, 6), number of elements is 2 and $P(A \cap B) = 2/6$

Then, the probability is

$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{2/6}{4/6} = \frac{1}{2}$ is found.

Addition and multiplication rules are summarized in Table 3.1.

Table 3.1. Addition and Multiplication Rules

Rule	Type of Events	
Addition ("*or*")	Disjoint Events $P(A \cup B) = P(A) + P(B)$	Joint Events $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Multiplication ("*and*")	Independent Events $P(A \cap B) = P(A) * P(B)$	Dependent Events $P(B/A) = \frac{P(A \cap B)}{P(A)}, \; P(A/B) = \frac{P(A \cap B)}{P(B)}$

3.3. EXAMPLES

Example 3.1: The probabilities of earthquake, material failure and foundation failure of a building are 3, 6, and 8 percent, respectively. Assuming that these three events are independent:

a. Calculate the probability of (earthquake or foundation failure)

b. The building will be destroyed (demolished) even if one of these three events is occurred; calculate the probability of the building to be destroyed.

Solution:

Example 3.2: The probabilities of flood (F), earthquake (E), and hurricane (H) in a region are 10%, 5%, and 8%, respectively. Calculate the probabilities of:

a. No event occurrence, **b.** (F or E) and (F or E or H).

Solution:

Example 3.3. The water of a city is transmitted by A, B and C pipes. The discharges of the pipes are Q_A = 10 l/s, Q_B = 15 l/s and Q_C = 25 l/s. The failure probabilities of these pipes are P_A=0.02, P_B=0.04 and P_C=0.06 and their failures are independent. Calculate the probabilities for the city:

a. Transmission of all of the water, **b.** Being the transmitted discharge at least 40 l/s, **c.** Being without water.

Solution:

CHAPTER 4

IMPORTANT PROBABILITY FUNCTIONS

4.1. PROBABILITY DISTRIBUTIONS

The behavior of a random variable is characterized by its probability distribution, that is, by the way probabilities are distributed over the values it assumes. A probability distribution function and a probability mass function are two ways to characterize this distribution for a discrete random variable. They are equivalent in the sense that the knowledge of either one completely specifies the random variable. The corresponding functions for a continuous random variable are the probability distribution function, defined in the same way as in the case of a discrete random variable, and the probability density function.

Given a random experiment with its associated random variable and given a real number, x let us consider the probability of the event $P(X \le x)$. This probability is clearly dependent on the assigned value x. The following function is defined as the Probability Distribution Function (PDF), or simply the *distribution function* of X.

$$F_X(x) = P(X \le x) \tag{4.1}$$

In Equation (4.1), subscript X identifies the random variable. This subscript is sometimes omitted when there is no risk of confusion. Let us repeat that $F_X(x)$ is simply P(A), the probability of an event A occurring, the event being $P(X \le x)$.

The PDF is thus the probability that X will assume a value lying in a subset of S, the subset being point x and all points lying to the 'left' of x. As x increases, the subset covers more of the real line, and the value of PDF increases until it reaches 1. The PDF of a random variable thus accumulates probability as x increases, and the name *Cumulative Distribution Function* (CDF) is also used for this function.

In view of the definition and the discussion above, one gives below some of the important properties possessed by a PDF.

- It exists for discrete and continuous random variables and has values between .0 and 1.
- It is a nonnegative, continuous-to-the-left, and nondecreasing function of the real variable Moreover, one has

$$F_X(-\infty) = 0 \text{ and } F_X(+\infty) = 1 \tag{4.2}$$

- If and are two real numbers such that a<b, then

$$P(a < X \le b) = F_X(b) - F_X(a) \tag{4.3a}$$

This relation is a direct result of the identity

$$P(X \le b) = P(X \le a) + P(a < X \le b) \tag{4.3b}$$

One can see from Eq. (4.3a) that the probability of X having a value in an arbitrary interval can be represented by the difference between two values of the PDF. Generalizing, probabilities associated with any sets of intervals are derivable from the PDF.

Example: Let a discrete random variable X assume -1, 1, 2, and 3 with probabilities 1/4, 1/8, 1/8 and 1/2, respectively. One then has

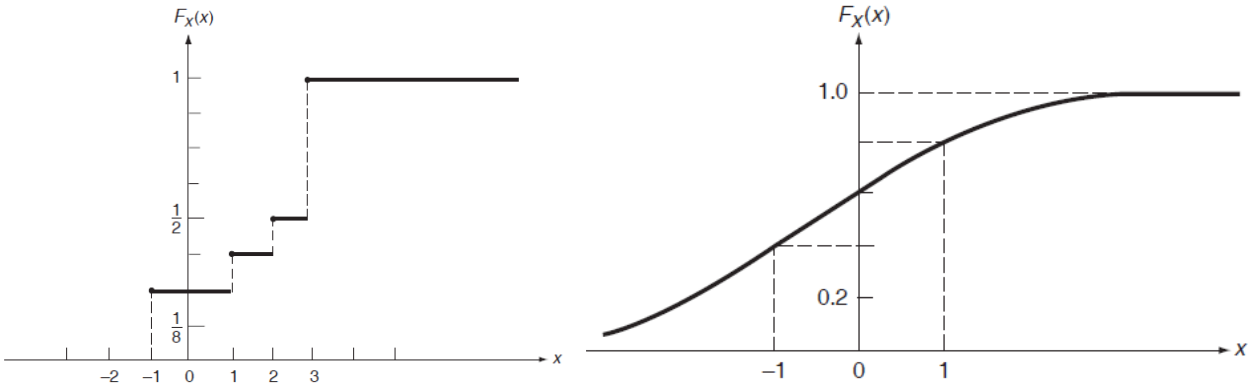
$$F_X(x) = \begin{cases} 0, & \text{for } x < -1; \\ \frac{1}{4}, & \text{for } -1 \le x < 1; \\ \frac{3}{8}, & \text{for } 1 \le x < 2; \\ \frac{1}{2}, & \text{for } 2 \le x < 3; \\ 1, & \text{for } x \ge 3. \end{cases}$$

This function is plotted in Fig. (4.1a). It is typical of PDFs associated with discrete random variables, increasing from 0 to 1 in a 'staircase' fashion.

A continuous random variable assumes a none numerable number of values over the real line. Hence, the probability of a continuous random variable assuming any particular value is zero and therefore no discrete jumps are possible for its PDF. A typical PDF for continuous random variables is shown in Fig. (4.1b). It has no jumps or discontinuities as in the case of the discrete random variable. The probability of X having a value in a given interval is found by using Eq. (4.3a), and it makes sense to speak only of this kind of probability for continuous random variables. For example, in Fig. (4.2b)

$$P(-1 < X \le 1) = F_X(1) - F_X(-1) = 0.8 - 0.4 = 0.4$$

Clearly, P(x=a) = 0, for any a.



a. Random Variable b. Continuous Variable

Figure 4.1. Probability Distribution Functions of X, $F_X(x)$ for Discrete and Random Variables

For a continuous random variable X, its PDF, $F_X(x)$ is a continuous function of x and the derivative

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{4.4}$$

exists for all x. The function is called the *probability density function* (pdf), or simply the *density function* of X. (Note the use of upper-case and lower-case letters, PDF and pdf, to represent the distribution and density functions, respectively).

Since is monotone nondecreasing, one clearly has for all x:

$$f_X(x) \ge 0 \tag{4.5}$$

Additional properties of can be derived easily from Eq. (4.4); these include

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1 \tag{4.6a}$$

and

$$P(a < X \le b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx \tag{4.6b}$$

An example of pdfs has the shape shown in Fig. (4.2). As indicated by Eq. (4.6a and 4.6b), the total area under the curve is unity and the shaded area from to gives the probability $P(a < X \le b)$. One again observes that the knowledge of either pdf or PDF completely characterizes a continuous random variable. The pdf does not exist for a discrete random variable since its associated PDF has discrete jumps and is not differentiable at these points of discontinuity.

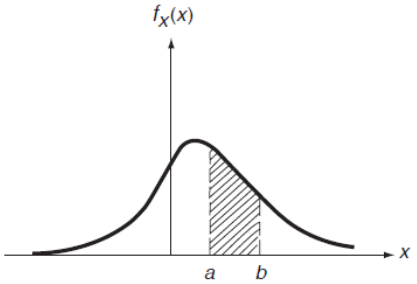


Figure 4.2. A Probability Density Function, $f_X(x)$

Using the mass distribution analogy, the pdf of a continuous random variable plays exactly the same role as the *probability mass function* (pmf) of a discrete random variable. The function $f_X(x)$can be interpreted as the mass density (mass per unit length). There are no masses attached to discrete points as in the discrete random variable case. The use of the term is *density function* therefore appropriate here for $f_X(x)$.

Many of the probability problems encountered in application fit some known probability functions. Since the features of these functions are known, various problems can be easily solved by using them. Probability functions in engineering practice can be classified as related to discrete variables and continuous variables.

4.2. FUNCTIONS OF DISCRETE VARIABLES

4.2.1. Binom Distribution

The events that have two alternatives are commonly encountered in practice. One of these alternatives is success and the other is fail, of which probabilities are p and q, respectively (p+q=1). For example, in a coin tossing event, the probabilities of tail and head are equal to 0.5. Whether or not a flood will be encountered in a year is another example. Let a sample with N component is taken from these variables; in other words, let N independent trials are performed. These are called *Bernoulli Trials*. The probability of x successes (x an integer between 0 and N) of an event, of which probability is p, in N trials fits Binom Distribution and can be computed as:

$$P(x) = \frac{N!}{x!(N-x)!} p^x q^{N-x} \tag{4.7a}$$

Binom Distribution has the following parameters:

$$\text{Expected value = Mean : } E_x = \bar{x} = Np \text{ , Variance: } V_x = Npq \tag{4.7b}$$

4.2.2. Poisson Distribution

Poisson distribution is a limit case of Binom and is successfully applied for calculating the probabilities of seldom (with have very small probability) simple events (heavy rain, flood, hurricane etc). The probability of x successes in N trials is calculated by:

$$P(X = x) = \frac{\lambda^x}{e^{\lambda} x!} \tag{4.8a}$$

Poisson distribution has one parameter (λ)

$$E_x = \bar{x} = V_x = \lambda = Np \tag{4.8b}$$

4.3. FUNCTIONS OF CONTINUOUS VARIABLES

4.3.1. Normal Distribution

A number of random variables encountered in practical applications fit to the *normal (Gaussian) distribution* with the following probability density function (pdf):

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{4.9}$$

It has two parameters; μ, mean and σ, standard deviation; it is symmetrical (C_s=0).Total area covered by normal curve is equal to 1. The probability of x between (x_1, x_2) interval is equal to the area between these values. The probabilities of the normal variable to remain in the intervals around the mean one, two and three standard deviations are equal to 0.6826, 0.9544 and 0.9974 (≅ 1), respectively (Figure 4.3).

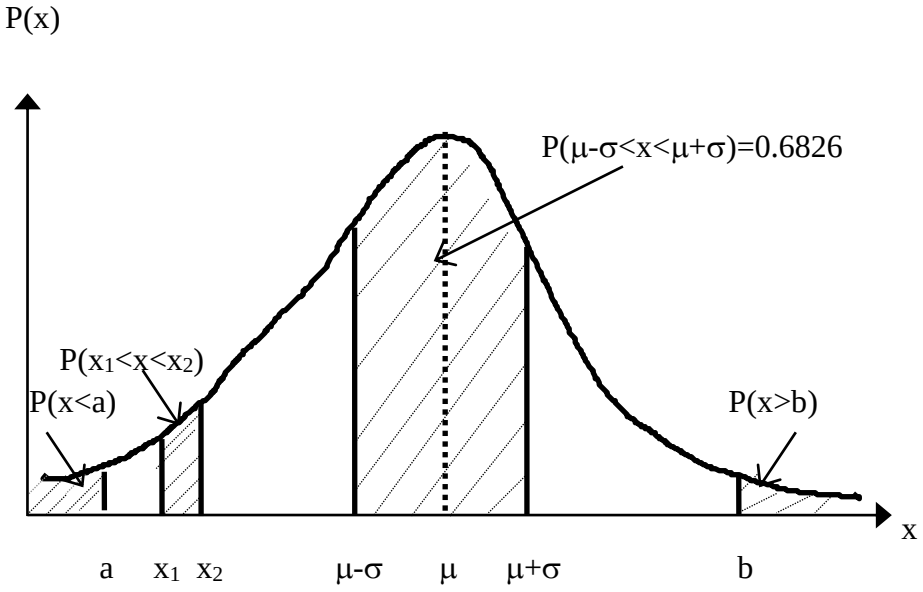


Figure 4.3. Properties of Normal Distribution

$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} P(x)dx = 1$$

$$P(x < \mu) = P(x > \mu) = \int_{-\infty}^{\mu} P(x)dx = \int_{\mu}^{\infty} P(x)dx = 0.5$$

$$P(x < a) = \int_{-\infty}^{a} P(x)dx, P(x > b) = \int_{b}^{\infty} P(x)dx, P(x_1 < x < x_2) = \int_{x_1}^{x_2} P(x)dx \quad (4.10)$$

$$P(\mu - 3\sigma < x < \mu + 3\sigma) = \int_{\mu-3\sigma}^{\mu+3\sigma} P(x)dx = 0.9974 \cong 1$$

The probability distribution function $F(x)$ of the normal distribution is tabulated numerically, by standardizing the random variable as follows:

$$z = \frac{x - \mu}{\sigma} \cong \frac{x - \bar{x}}{S_x} \quad (4.11)$$

The distribution of z is called as *standard normal distribution*. All of the rules of normal distribution are also valid for standard normal distribution. The probability density function of standard normal distribution is:

$$P(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (4.12)$$

The tabulated values of normal distribution can be used for calculating the probability of z is between zero and a positive z_1 (Table 4.1). Since the distribution is symmetrical, the same probabilities are valid for negative z values. The probability can be calculated by the following equation:

$$p = 0.5 - 0.5 * \exp\left(\frac{-[(83z + 351)z + 562]}{\frac{703}{z} + 165}\right) \quad (4.13a)$$

Similarly, the related z value for p probability is calculated as:

$$z = \frac{(0.5 + p)^{0.135} - (0.5 - p)^{0.135}}{0.1975} \quad (4.13b)$$

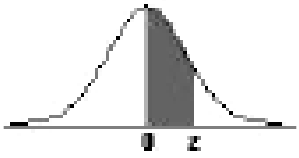
4.3.2. Lognormal Distribution

All of the variables do not fit to normal distribution; however, it is often attempted to transform a non-normal random variable because the normal distribution has well known properties and is easy to use. The most commonly used transformation is the logarithmic transformation. If the transformed variable

$$Y = \ln X \quad (4.14)$$

fits to the normal distribution, then the distribution of the original variable X is called *lognormal*. The distribution is defined only for positive values.

Table 4.1. Standard Normal Distribution



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

4.3.3. Gumbel Distribution

Gumbel (extreme value type I) distribution has two forms. One is based on the smallest extreme and the other is based on the largest extreme. The general formula for the probability density function of the maximum Gumbel distribution is

$$f(x_i) = \frac{1}{a} e^{-(x-x_0)} e^{-e^{-(x-x_0)}} \tag{4.15}$$

where x_0 is the location parameter and a is the scale parameter. The formula for the cumulative distribution function of the Gumbel distribution is

$$F(x_i) = P(x \geq x_i) = 1 - e^{-e^{-a(x-x_0)}} \tag{4.16}$$

$$a = \frac{S_N}{S_x}, \quad x_0 = \bar{x} - y_N \frac{S_x}{S_N} \tag{4.17}$$

Let $y = a(x - x_0)$, then one can obtain,

$$F(x_i) = P(x \geq x_i) = 1 - e^{-e^{-y}} \quad \Rightarrow 1 - F(x_i) = P(x < x_i) = e^{-e^{-y}} \tag{4.18a}$$

From this equation, one can obtain

$$y = -\ln\{-\ln[1 - F(x_i)]\} \tag{4.18b}$$

For example, $F(x_i) = 0.01 \Rightarrow y = -\ln\{-\ln[1-0.01]\} = 4.600$, $F(x) = 0.1 \Rightarrow y = -\ln\{-\ln[1-0.1]\} = 2.250$

y value is calculated as $y = a(x - x_0) = \frac{S_N}{S_x}\left[x - (\bar{x} - y_N \frac{S_x}{S_N}\right] \Rightarrow$

$$y_i = \frac{S_N}{S_x}\left[x_i - \left(\bar{x} - y_N \frac{S_x}{S_N}\right)\right] = \frac{S_N}{S_x}\left(x_i - \bar{x} + S_x \frac{y_N}{S_N}\right) \Rightarrow x_i = y_i \frac{S_x}{S_N} + \bar{x} - S_x \frac{y_N}{S_N} \tag{4.19}$$

Where, $\bar{x}$ and S_x are the mean and standard deviation of the data, respectively. The values of S_N and y_N are tabulated as follows according to the number of sample (N) and are given in Table 4.2 .

Table 4.2. y_N and S_N Values for Gumbel Distribution

N	10	15	20	25	30	35	40	50	75	100	200
y_N	0.495	0.513	0.524	0.531	0.536	0.540	0.544	0.549	0.556	0.560	0.567
S_N	0.950	1.021	1.063	1.092	1.112	1.129	1.141	1.161	1.190	1.207	1.236

4.3.4. Pearson Type III Distribution

The x value, of which exceedance probability is p is calculated as

$$x = \bar{x} + KS_x, \quad K = \frac{2}{C_s}\left(1 + \frac{zC_s}{6} - \frac{C_s^2}{36}\right)^3 - \frac{2}{C_s} \tag{4.20}$$

Where, C_s is the skewness coefficient and z is the standard normal value.

If $y = \ln x$ values fit Pearson Type III Distribution, then the distribution of the original variable X is called *Log Pearson Type III Distribution.*

Log Normal (LN,) Gumbel (G), Pearson (P III) and Log-Pearson (LP III) Type III Distributions are often used in extreme values (floods, droughts, heavy rains etc.).

4.4. EXAMPLES

Example 4.1: The spillway of a dam is designed according to a flood with annual probability is 0.33 percent. Calculate the non-occurring probability and 1 times and 2 times occurring probabilities of this flood in 45 years according to both Binom and Poisson Distributions.

Solution:

Binom Distribution:

p = 0.0033, q = 1 – 0.0033 = 0.9967, N=45

Eq. (4.7a): $P(x) = \frac{N!}{x!(N-x)!} p^x q^{N-x} = \frac{45!}{x!(45-x)!} * 0.0033^x * 0.9967^{45-x}$

Non-occurring = 0 times occurring x = 0 $\Rightarrow P(0) = \frac{45!}{0!(45-0)!} * 0.0033^0 * 0.9967^{45-0} = 0.8618$

x = 1 times occurring $\Rightarrow P(1) = \frac{45!}{1!(45-1)!} * 0.0033^1 * 0.9967^{45-1} = 0.1283$

x = 2 times occurring $\Rightarrow P(2) = \frac{45!}{2!(45-2)!} * 0.0033^2 * 0.9967^{45-2} = 0.00935$

Poisson Distribution:

Eq. (4.8a): $P(x) = \frac{\lambda^x}{e^\lambda x!}$, Eq. (4.8b): $\lambda = Np = 45 * 0.0033 = 0.1485,\ P(x) = \frac{\lambda^x}{e^\lambda x!} = \frac{0.1485^x}{e^{0.1485} * x!}$

Non-occurring = 0 times occurring x = 0 $\Rightarrow P(0) = \frac{0.1485^0}{e^{0.1485} * 0!} = 0.8620$

x = 1 times occurring $\Rightarrow P(1) = \frac{0.1485^1}{e^{0.1485} * 1!} = 0.1280$

x = 2 times occurring $\Rightarrow P(2) = \frac{0.1485^2}{e^{0.1485} * 2!} = 0.0095$

Example 4.2: Calculate the probability of a traffic accident occurring 0 times and 5 times in 50 years, according to Binomial and Poisson distributions, with a 20% probability to occur in any year.

Solution:

Binom Distribution:

$P(0) = \frac{50!}{0!(50-0)!} 0.2^0 * 0.8^{50} = 1.427 * 10^{-5}$, $P(5) = \frac{50!}{5!(50-5)!} 0.2^5 * 0.8^{45} = 0.0295$

Poisson Distribution:

$\lambda = 50 * 0.20 = 10, \Rightarrow P(0) = \frac{0.10^1}{e^{10} * 1!} = 4.54 * 10^{-5}, ,\ P(5) = \frac{10^5}{e^{10} 5!} = 0.0378$

Example 4.3: Calculate the probability of z variable is between -2 and 2.

Solution:

Example 4.4: Calculate of area of between x = 4 and x = 9 in a Normal Distribution with a mean of 5 and standard deviation of 2. In other worlds, calculate the probability of the value is between 4 and 9.

Solution:

Example 4.5: Total annual precipitation height of a gauge station fit Normal Distribution with a mean of 90 cm and standard deviation of 25 cm. Calculate the probabilities of total annual precipitation in any year are: **a.** Less than 70 cm, **b.** Between 80 and 105 cm, **c.** Between 60 and 130 cm, **d.** More than 140 cm

Solution:

Example 4.6: Mean and standard deviation values of total annual precipitation heights of a city are 60 cm and 15 cm, respectively. Assuming that the data fit Normal Distribution:

a. Determine the interval which contains 95 percent of the data (95% confidence interval),

b. A year with less than 45 cm precipitation height is called as "drought year" and with greater than 90 cm precipitation height is called as "flood year". Calculate the expected numbers of both drought and flood years in 50 years.

c. Estimate the maximum and minimum annual precipitation heights in 50 years.

Solution:

b.

c.

Example 4.7: The mean annual discharge values of a stream fit Normal Distribution with a mean of 70 m^3/s and a standard deviation of 10 m^3/s. With the 80 yearly duration:

a. Estimate the number of years with less than 78 m^3/s mean discharge,
b. Calculate the 92 percent confidence interval of mean of the population.

Solution:

Example 4.8: Yearly maximum discharges of a stream (m^3/s) are given as follows. Calculate the probability that discharges are between (55 and 70) m^3/s, greater than 75 m^3/s; the discharge values of which exceedance probabilities are 1 and 0.1 percent; by using the distributions of: **a.** Gumbel (G), **b.** Log Normal (LN), **c.** Log Pearson Type III (LPT)

68 76 39 48 57 71 63 56 37 54 59 70 54 62 68

Solution:

CHAPTER 5

SAMPLING DISTRIBUTIONS

5.1. THE CONCEPT OF SAMPLING DISTRIBUTION

In chapter 2.2.5, several summary statistics were presented which described key attributes of a dataset. They were sample estimates (such as $\bar{x}$ and S_x) of true and unknown population parameters (such as μ and σ). In this chapter, descriptions of the uncertainty or reliability of sample estimates are presented. As an alternative to reporting a single estimate, the utility of reporting a range of values called an *interval estimate* is demonstrated. The sample mean and standard deviation estimate the corresponding points of a population. Such estimates are called *point estimates*. By themselves, point estimates do not portray the reliability, or lack of reliability (variability), of these estimates. For example, suppose that two data sets X and Y exist, both with a sample mean of 5 and containing the same number of data. The Y data all cluster tightly around 5, while the X data are much more variable. The point estimate of 5 for X is much less reliable than that for Y because of the greater variability in the X data. In other words, more caution is needed when stating that 5 estimates the true population mean of X than when stating this for Y. Reporting only the point estimate of 5 fails to give any hint of this difference.

As an alternative to point estimates, *interval estimates* are intervals which have a stated probability of containing the true population value. The intervals are wider for data sets having greater variability. Thus in the above example an interval between 4.7 and 5.3 may have a 95% probability of containing the (unknown) true population mean of Y. It would take a much wider interval, say between 2.0 and 8.0, to have the same probability of containing the true mean of X. The difference in the reliability of the two estimates is therefore clearly stated using interval estimates. Interval estimates can provide a piece of information which point estimates cannot: A statement of the probability that the interval contains the true population value (its reliability).

The real value of any β parameter of a random variable is never certainly determined, because it is impossible to observe the whole population of this variable. However, it is possible to calculate a value of b statistics which is an estimate of this parameter from a sample, to be supposed to represent the population. Statistics b is highly probably not equal the parameter β, it is the best estimate of β which can be obtained from the sample at hand. The calculated b statistics (for example the mean values, $\bar{x}_i$) from various samples drawn from the same population are only estimates of the corresponding parameter (for example mean value of the population, μ). The values of a statistics calculated from different samples have a distribution, because any statistic value can be treated as a random variable. The probability distribution of the values of any statistic to be calculated of same population with same size is called the *sampling distribution* of this statistics.

To know the sampling distribution of a statistic is important for the following reason: As mentioned above, the statistics determined from the sample at hand is not equal to the real value of the population parameter. Without observing the whole population, it is impossible to determine the value of parameter with an absolute correctness. However, it can be determined the interval in which the unknown parameter will remain around the calculated statistic with a given probability. For this, the sampling distribution of that statistics must be known.

The sampling distribution $f(b)$ of the b statistics of the β parameter calculated from samples of magnitude N is presented in Figure 5.1. The expected value of this distribution $E(b)$ will be the b_0 value calculated from the sample at hand. For determining (b_1, b_2) interval in which the unknown β parameter will remain with a given P_c probability, such a symmetrical interval (b_1, b_2) is chosen around b_0 that, the percentage of the sampling distribution within this interval is P_c. Here, P_c is the probability that the interval includes the true value and is called the *confidence level* and the interval (b_1, b_2) is named the *confidence interval* at this confidence level and $\alpha = 1 - P_c$ is the probability that this interval will not cover the true value and is called the *significance level*. Values as 0.90, 0.95 and 0.99 are used for P_c in practice.

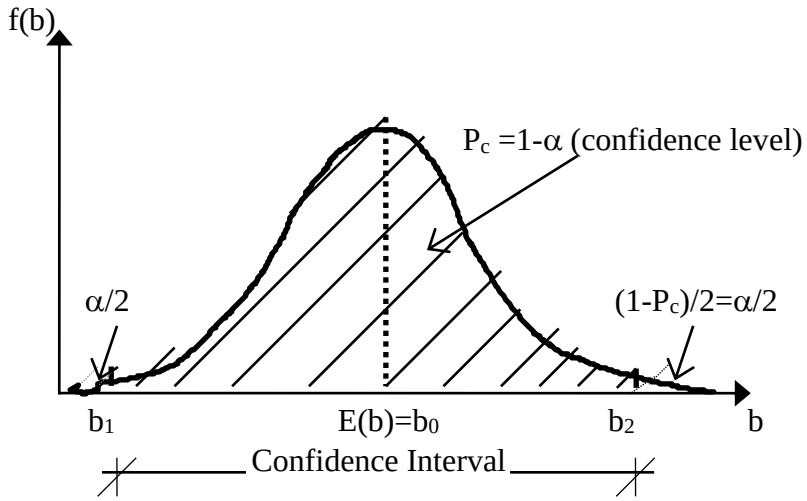


Figure 5.1. Confidence Level and Confidence Interval

As the number of the elements of the sample (N) increases, the confidence interval corresponding to a certain confidence level gets narrower. In other words, the confidence interval within which the parameter will remain with a certain probability is smaller for large samples, expressing that the error in parameter estimation is reduced. In order to estimate the parameter of the population, the sampling distribution related to the parameter is used. Sampling distributions are determined theoretically for some statistics.

5.2. ESTIMATION OF CONFIDENCE INTERVALS OF POPULATION PARAMETERS

In the estimation of population parameters with a confidence level, two different methods are used according to asymptotic and exact distributions. In general, it is adequate to estimate the confidence intervals of only two important parameters, the mean and the variance (or the standard deviation).

5.2.1. Estimation of Confidence Interval of Mean

a. In the Event of Asymptotic Distribution:

If the number of elements is large $(N \geq 30)$, sampling distribution of means that are calculated from different samples fit to Normal Distribution. Accordingly, this sampling distribution has a mean, a standard deviation; moreover, its values within a confidence interval can be estimated by Normal Distribution. The confidence interval of the mean of sampling distribution's mean is calculated as follows:

$$P_c = P(b_1 < \mu < b_2) = P\left[\left(\bar{x} - zS_{\bar{x}}\right) < \mu < \left(\bar{x} + zS_{\bar{x}}\right)\right] \tag{5.1}$$

This equation implies that, the mean of population (μ) will be within the (b_1, b_2) interval with P_c probability; in other words, the confidence interval with P_c probability of μ is (b_1, b_2). In Equation 5.1, the z value is obtained from Normal Distribution Table corresponding the value of $\left(0.5 - \alpha/2\right) = \left[0.5 - \left(1 - P_c\right)/2\right] = \left(P_c/2\right)$. For example, the z value for 0.95 confidence level is obtained as to correspond the value of $\left(0.95/2\right) = 0.475 \Rightarrow z = 1.96$. The $S_{\bar{x}}$ value in (5.1) the standard deviation of sampling distribution of x variable and is calculated as follows:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{N}} \tag{5.2a}$$

If this value is inserted in (5.1), the following equation is obtained:

$$P_c = P(b_1 < \mu < b_2) = P\left[\left(\bar{x} - zS_x/\sqrt{N}\right) < \mu < \left(\bar{x} + zS_x/\sqrt{N}\right)\right] \tag{5.2b}$$

In other words, the lower and upper limits of confidence interval of the mean the population (μ) are:

Lower limit: $b_1 = \bar{x} - z\frac{S_x}{\sqrt{N}}$, upper limit: $b_2 = \bar{x} + z\frac{S_x}{\sqrt{N}}$ (5.2c)

b. In the Event of Exact Distribution:

If the amplitude of the data is small $(N < 30)$, the confidence interval of the mean is calculated by t (student) distribution as follows:

$$t = \frac{x - \bar{x}}{S_x/\left(\sqrt{N-1}\right)} \tag{5.3}$$

The t distribution is symmetric and the probability of t is greater than a given value of t_0 is tabulated according to $(N-1)$ degree of freedom for various confidence levels (Table 5.1). For large N values, t distribution approaches Normal Distribution. The confidence interval of parameter μ is given as:

$$P_c = P(b_1 < \mu < b_2) = P\left[\left(\bar{x} - tS_x/\sqrt{N-1}\right) < \mu < \left(\bar{x} + tS_x/\sqrt{N-1}\right)\right] \tag{5.4a}$$

In other words, the lower and upper limits of confidence interval of the mean of the population (μ) are:

Lower limit: $b_1 = \bar{x} - t\frac{S_x}{\sqrt{N-1}}$, upper limit: $b_2 = \bar{x} + t\frac{S_x}{\sqrt{N-1}}$ (5.4b)

As is seen, when this equation is compared to (5.2c), t and $N-1$ values have come instead of z and N, respectively.

TABLE 5.1. t (STUDENT) DISTRIBUTION (d_f: DEGREE OF FREEDOM)

	P_c					P_c			
d_f	0.90	0.95	0.98	0.99	d_f	0.90	0.95	0.98	0.99
1	6.314	12.706	31.821	63.657	2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841	4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032	6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499	8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250	10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106	12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012	14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947	16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898	18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861	20	1.725	2.086	2.528	2.845
22	1.717	2.074	2.508	2.819	24	1.711	2.064	2.492	2.797
26	1.706	2.056	2.479	2.779	28	1.701	2.048	2.467	2.763
30	1.697	2.042	2.457	2.750	120	1.658	1.980	2.358	2.617

5.2.2. Estimation of Confidence Interval of Variance

Chi-Square (χ^2) distribution is used in the estimation of confidence interval of the variance. The probability of χ^2 is greater than a given value of $\chi_0^{\,2}$ is tabulated according to $(N-1)$ degree of freedom for various confidence levels (Table 5.2). Confidence interval of the variance of the population (σ^2) is found as:

$$P_c = P\left(b_1 < \sigma^2 < b_2\right) = P\left(\frac{NS^2{}_x}{\chi^2{}_{\alpha/2}} < \sigma^2 < \frac{NS^2{}_x}{\chi^2{}_{1-\alpha/2}}\right) \tag{5.5a}$$

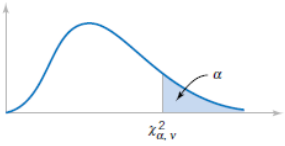
Confidence interval of the standard deviation of the population (σ) is found as follows:

$$P_c = P\left(b_1 < \sigma < b_2\right) = P\left[\left(\sqrt{\frac{NS^2{}_x}{\chi^2{}_{\alpha/2}}}\right) < \sigma < \left(\sqrt{\frac{NS^2{}_x}{\chi^2{}_{1-\alpha/2}}}\right)\right] \tag{5.5b}$$

In other words, the lower and upper limits of confidence interval of the standard deviation of the population (σ) are:

Lower limit: $b_1 = \sqrt{\frac{NS^2{}_x}{\chi^2{}_{\alpha/2}}}$, upper limit: $b_2 = \sqrt{\frac{NS^2{}_x}{\chi^2{}_{1-\alpha/2}}}$ (5.5c)

TABLE 5.2. χ^2 DISTRIBUTION ($d_f = \upsilon$:DEGREE OF FREEDOM)



d_f = υ	α = 0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
	P_C = 0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
1	-	-	0.004	0.0158	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.490	4.865	15.897	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	22.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.075	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

5.3. EXAMPLES

Example 5.1: The mean and standard deviation values of compressive strength of 50 samples obtained from a concrete mass are 240 kg/cm^2 and 65 kg/cm^2, respectively.

a. Determine the interval which contains 95 percent of mean compressive strength of concrete,

b. Calculate the probability that the mean compressive strength is greater than 260 kg/cm^2.

Solution:

Example 5.2: Experiments were performed for 21 steel samples. It is found that mean and standard deviation of yielding values in the experiments are 8490 kg and 100 kg, respectively. Calculate the confidence intervals of population mean for **a.** 90 % and **b.** 98 % confidence levels.

Solution:

Example 5.3: a. 20 and **b.** 10 samples are taken from concrete, produced a batch plant. Both groups have a 100 kg/cm^2 standard deviation. Calculate the 95% confidence interval of population standard deviation for both groups.

Solution:

Example 5.4. The mean and standard deviation values of 20 samples taken from a concrete mass are 280 kg/cm^2 and 15 kg/cm^2, respectively. Calculate the confidence interval of population of **a.** mean and **b.** standard deviation for 5 % significance level.

Solution:

CHAPTER 6

STATISTICAL HYPOTHESIS

6.1. INTRODUCTION

In the previous chapter we illustrated how to construct a confidence interval estimate of a parameter from sample data. However, many problems in engineering require that we decide whether to accept or reject a statement about some parameter. The statement is called a *hypothesis*, and the decision-making procedure about the hypothesis is called *hypothesis testing*. This is one of the most useful aspects of statistical inference, since many types of decision-making problems, tests, or experiments in the engineering world can be formulated as hypothesis-testing problems. Statistical hypothesis testing and confidence interval estimation of parameters are the fundamental methods used at the data analysis stage of a comparative experiment, in which the engineer is interested, for example, in comparing the mean of a population to a specified value. These simple comparative experiments are frequently encountered in practice and provide a good foundation for the more complex experimental design problems. Since we use probability distributions to represent populations, a statistical hypothesis may also be thought of as a statement about the probability distribution of a random variable. The hypothesis will usually involve one or more parameters of this distribution. Hypotheses are always statements about the population or distribution under study, not statements about the sample.

A procedure leading to a decision about a particular hypothesis is called a *test of a hypothesis*. Hypothesis testing procedures rely on using the information in a random sample from the population of interest. If this information is consistent with the hypothesis, we will conclude that the hypothesis is true; however, if this information is inconsistent with the hypothesis, we will conclude that the hypothesis is false. We emphasize that the truth or falsity of a particular hypothesis can never be known with certainty, unless we can examine the entire population. This is usually impossible in most practical situations. Therefore, a hypothesis-testing procedure should be developed with the probability of reaching a wrong conclusion in mind.

One of the main goals of the statistic is to perform correct estimations about population by using inadequate data obtained from small samples. During this process, some assumes, called *statistical hypothesis* are done and after some tests, their validity is accepted or rejected. Scientists collect data in order to learn about the processes and systems those data represent. Often they have prior ideas, called hypotheses, of how the systems behave. One of the primary purposes of collecting data is to test whether those hypotheses can be substantiated, with evidence provided by the data.

Statistical tests are the most quantitative ways to determine whether hypotheses can be substantiated, or whether they must be modified or rejected outright. The acceptance of a hypothesis does not imply that it is certainly true and its rejection does not mean that it is certainly false. The determination of a hypothesis and evaluating it as true or false can be expressed by some probabilities (90, 93, 99 percent, etc).

One important use of hypothesis tests is to evaluate and compare groups of data. For example, water quality has been compared between two or more aquifers, and some statements made as to which are different. Rather than using hypothesis tests, the results are sometimes expressed as the author's educated opinions "it is clear that development has increased well yield." Hypothesis tests have at least two advantages over educated opinion:

i. They insure that every analyst of a data set using the same methods will arrive at the same result. Computations can be checked on and agreed to by others,

ii. They present a measure of the strength of the evidence (the p-value).

6.2. STRUCTURE OF HYPOTHESIS TESTS

6.2.1. Choose an Appropriate Test

Test procedures are selected based on the data characteristics and study objectives. The second criterion is the objective of the test. Hypothesis tests are available to detect differences between central values of two groups, three or more groups, between spreads of data groups, and for covariance between two or more variables, among others. The third selection criterion is the choice between parametric or nonparametric tests. This should be based on the expected distribution of the data involved. If similar data in the past were normally distributed, a parametric procedure would usually be selected. If data were expected to be non-normal, or not enough is known to assume any specific distribution, nonparametric tests would be preferred.

6.2.2. Establish the Null and Alternative Hypotheses

The null and alternate hypotheses should be established prior to collecting data. These hypotheses are a concise summary of the study objectives, and will keep those objectives in focus during data collection.

The null hypothesis (H_0) is what is assumed to be true about the system under study prior to data collection, until indicated otherwise. It usually states the null situation – no difference between groups, no relation between variables. One may suspect, hope, or root for either the null or alternate hypothesis, depending on one's vantage point. But the null hypothesis is what is assumed true until the data indicate that it is likely to be false. For example, an engineer may test the hypothesis that wells upgradient and downgradient of a hazardous waste site have the same concentrations of some contaminant. They may hope that downgradient concentrations are higher (the company gets a new remediation project), or that they are the same (the company did the original site design). In either case, the null hypothesis assumed to be true is the same: concentrations are similar in both groups of wells.

The alternate hypothesis (H_1) is the situation anticipated to be true if the evidence (the data) show that the null hypothesis is unlikely. It is in some cases just the negation of H_0, such as "the 100-year flood is not equal to the design value." H_1 may also be more specific than just the negation of H_0 "the 100-year flood is greater than the design value".

6.2.3. Decide on an Acceptable Error Rate α

The α-value, or significance level, is the probability of incorrectly rejecting the null hypothesis H_0 when it is in fact true, called a "Type I error"). Table 6.1. shows that this is one of four possible outcomes of a hypothesis test. The significance level is the risk of a Type I error deemed acceptable by the decision maker. It is a " management tool" dependent not on the data, but on the objectives of the study. Statistical tradition uses a default of 5% (0.05) for α, but there is no reason why other values should not be used. Before testing the hypothesis, a significance level (α) is chosen. This level represents the probability of type I error. If the probability of type I error decreases, the probability of type II error increases, Therefore, α value can not be chosen as small as desired.

Table 6.1. Four Possible Results of Hypothesis Testing.

Decision	Unknown True Situation	
	H_0 is True	H_0 is False
Accept H_0 (Reject H_1)	Correct Decision Probability = 1-α	Type II Error Probability = β
Reject H_0 (Accept H_1)	Type I Error Probability = α	Correct Decision Probability = 1-β

6.2.4. Make the Decision to Reject H_0 or Not

When the p-value is less than the decision criteria (the α level), H_0 is rejected. When the p value is greater than α, H_0 is not rejected. The null hypothesis is never accepted, or proven to be true. It is assumed to be true until proven otherwise, and is not rejected when there is insufficient evidence to do so.

6.3. HYPOTHESIS TESTS FOR PARAMETERS

Hypothesis tests may be performed for values of parameters of population. The tests for the mean and the standard deviation, the most widely used parameters in practice, are as follows:

6.3.1. Test of Mean

In order to test whether the means of two different populations are the same or not (H_0: $\mu_1 = \mu_2$), samples are taken from the populations with N_1 and N_2 quantities, their means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations S_1 and S_2 are calculated. The standard normal value (z) is found from normal distribution corresponding to $(0.5-\alpha/2)$; the following values are calculated:

$$S_{\bar{x}} = \sqrt{\frac{(N_1 S_2{}^2 + N_2 S_1{}^2)}{N_1 N_2}}, \qquad \Delta\bar{x} = \bar{x}_1 - \bar{x}_2 \tag{6.1}$$

If $|\Delta\bar{x}| \le zS_{\bar{x}}$, then the H_0 is accepted, if $|\Delta\bar{x}| > zS_{\bar{x}}$ then it is rejected.

6.3.2. Test of Standard Deviation

In order to test whether the standard deviations of two different populations are the same or not, (H_0: $\sigma_1 = \sigma_2$) F (Fisher) test is used. The computed F is calculated as follows:

$$\text{If } S_1 > S_2 \Rightarrow F_c = \frac{S_1^{\,2}}{S_2^{\,2}}, \quad \text{if } S_2 > S_1 \Rightarrow F_c = \frac{S_2^{\,2}}{S_1^{\,2}} \tag{6.2}$$

If the computed F_c is less than its tabulated value (F_t) then H_0 hypothesis is accepted, and vice versa. The tabulated F is found from F Table (Table 6.1), with degrees of freedom (N_1-1) and (N_2-1) on numerator and denominator, respectively. Here, N_1 and N_2 are the numbers of data of sample which have greater and smaller variance, respectively.

TABLE 6.1. F DISTRIBUTION ($\alpha = 0.01$)

$n \downarrow, m \rightarrow$

	5	6	7	8	9	10	15	20	30	40	50	100
5	11.0	10.7	10.5	10.3	10.2	10.1	9.72	9.55	9.38	9.29	9.24	9.13
6	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.14	7.09	6.99
7	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.91	5.86	5.75
8	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.12	5.07	4.96
9	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.57	4.52	4.42
10	5.64	5.39	5.20	5.05	4.94	4.85	4.56	4.41	4.25	4.17	4.12	4.01
12	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.62	3.57	3.47
14	4.70	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.27	3.22	3.11
16	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10	3.02	2.97	2.86
18	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92	2.84	2.78	2.68
20	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.69	2.64	2.54
25	3.86	3.63	3.46	3.33	3.22	3.13	2.86	2.70	2.54	2.46	2.40	2.29
30	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.30	2.25	2.13
40	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.11	2.06	1.94
60	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.94	1.88	1.75
80	3.26	3.04	2.87	2.74	2.64	2.55	2.27	2.12	1.94	1.85	1.79	1.66
100	3.21	2.99	2.82	2.69	2.59	2.50	2.20	2.07	1.89	1.80	1.73	1.60

6.4. HYPOTHESIS TESTS FOR PROBABILITY DISTRIBUTIONS

In order to test the frequency distribution of a sample whether or not fit a theoretical distribution (for example normal distribution), test of statistical hypothesis is the most reliable method.

6.4.1. Chi Square Test

If, the observed sample with N components is classified into m classes; each of the class has O_i components and the corresponding theoretical distribution has e_i component; then χ^2 is calculated as follows:

$$\chi^2_c = \sum_{i=1}^{m}\left[\frac{(O_i - e_i)^2}{e_i}\right] \tag{6.3}$$

If $\chi^2_c \le \chi^2_t$ then the sample fits the theoretical distribution, and vice versa. If $\chi^2_t = 0$, exact fitting exist between the sample and the related distribution. χ^2_t is read for α. In reading the tabulated χ^2 values, the degree of freedom is m-2 for Poisson and m-3 for other distributions.

6.4.2. Probability Plot Correlation Coefficient Test

In this test, the correlation coefficient (r) between the theoretical values of a distribution (x_t) and observed values (x_o) is calculated. If the calculated value is greater than or equal to the critical value given in Table 6.2, then it is assumed that the observed data fit the related distribution. Correlation coefficient is found as:

$$r = \frac{N\sum xy - \sum x \sum y}{\left\{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]\right\}^{0.5}} = \frac{\sum\left[(x-\bar{x})(y-\bar{y})\right]}{NS_xS_y} = \frac{\sum xy - N\bar{x}\bar{y}}{NS_xS_y} \tag{6.4}$$

The values are ranked from smaller to greater, the probability of non exceedance of each data is calculated as following, where, i is the rank number.

$$F(x < x_i) = 1 - p = \frac{i - 0.40}{N + 0.20} \tag{6.5}$$

Table 6.2. Critical r Values

N	G, P AND LP III		N, LN	
	α=0.01	α=0.05	α=0.01	α=0.05
10	0.8630	0.9084	0.879	0.918
20	0.9060	0.9390	0.926	0.951
30	0.9191	0.9526	0.947	0.964
40	0.9286	0.9594	0.959	0.972
50	0.9389	0.9646	0.966	0.977
60	0.9467	0.9685	0.971	0.980
70	0.9506	0.9720	0.975	0.983
80	0.9525	0.9747	0.978	0.985
90	0.9554	0.9764	0.980	0.986
100	0.9596	0.9779	0.982	0.987

6.5. EXAMPLES

Example 6.1. Lighting is applied to decrease traffic accidents in a junction. Accident numbers are given below before lighting for 12 months (x_1) and after lighting for 8 months (x_2). Decide whether or not the lighting has affected **a.** the mean and **b.** standard deviation of traffic accident for $\alpha = 0.01$.

x_1	5	8	11	7	6	9	6	8	6	7	7	10
x_2	3	6	7	5	11	8	7	2	-	-	-	-

Solution:

Example 6.2: The number (N), mean ($\bar{x}$) and standard deviations (S_x) of precipitation height values before and after dam construction are given as follows. Determine whether or not dam construction has changed **a.** the mean and **b.** the standard deviation of precipitation heights for α=0.01.

	N	$\bar{x}$ (cm)	S_x (cm)
Before the dam	20	100	23
After the dam	25	116	10

Solution:

a.

b.

Example 6.3: Numbers of occurring (x) and observation (O) of a flood, of which probability is 5 percent, are given as follows, for a 40 yearly observation period. Determine whether the data fit Binom and Poisson Distributions for 5% significance level.

Solution:

Example 6.4: The 80 yearly annual total precipitation height values of a region fit Normal Distribution with a mean of 70 cm and a standard deviation of 10 cm. It is aimed to test whether or not the data fit Normal Distribution. There are 50 data between (60-75) cm, calculate the chi-square (χ^2) value.

Solution:

Example 6.5: The distribution of marks taken by 80 students in an exam is given as follows. The mean is 50 and the standard deviation is 16. Determine whether or not the marks fit normal distribution for both 1% and 5% significance levels.

GROUP	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
NO OF STUD. (O_i)	1	5	8	10	14	15	11	8	6	2

Solution:

Number of data N = 80, number of groups m = 10, $\bar{x} = 50, S_x = 16$, N = 80, $z = \dfrac{x - \bar{x}}{S_x} = \dfrac{x - 50}{16}$,

For example, for the first group (0-10), the values are 0 and 10 and the corresponding z values are calculated as follows:

If the data fits Normal Distribution, then, the probability of the data is between (0 - 10) is calculated as was given in Normal Distribution:

$$x_1 = 0 \Rightarrow z_1 = \frac{0 - 50}{16} = -3.13,\ x_2 = 10 \Rightarrow z_2 = \frac{10 - 50}{16} = -2.50.$$

$$P(0 < x < 10) = P(-3.13 < z < -2.50) = 0.4991 - 0.4938 = 0.0053$$

Expected value $e_i = N*p_i = 80*0.0053 = 0.424$, observed value (no of student) $O_i = 1$

$$\chi^2{}_c = \sum_{i=1}^{m}\left[\frac{(O_i - e_i)^2}{e_i}\right] = \frac{\left(1 - 0.424\right)^2}{0.424} = 0.782 \text{ is obtained.}$$

For the second group (10-20), values are 10 and 20 and the corresponding z values are calculated as follows:

If the data fits Normal Distribution, then, the probability of the data is between (10 - 20) is calculated as was given in Normal Distribution:

$$x_1 = 10 \Rightarrow z_1 = \frac{10 - 50}{16} = -2.50,\ x_2 = 20 \Rightarrow z_2 = \frac{20 - 50}{16} = -1.88$$

$$P(10 < x < 20) = P(-2.50 < z < -1.88) = 0.4938 - 0.4699 = 0.0239$$

Expected value $e_i = N*p_i = 80*0.0239 = 1.912$, observed value (no of student) $O_i = 5$

$$\chi^2{}_c = \sum_{i=1}^{m}\left[\frac{(O_i - e_i)^2}{e_i}\right] = \frac{\left(5 - 1.912\right)^2}{1.912} = 4.987 \text{ is obtained.}$$

Similar calculations are made for all of the groups and the results are presented in the table:

GROUP	0-10	10-20	20-30	30-40	40-50
NO OF ST. (O_i)	1	5	8	10	14
x_1	0	10	20	30	40
x_2	10	20	30	40	50
z_1	-3.13	-2.50	-1.88	-1.25	-0.63
z_2	-2.50	-1.88	-1.25	-0.63	0.00
P_i	0.0053	0.0239	0.0755	0.1587	0.2357
$e_i=80*P_i$	0.424	1.912	6.040	12.696	18.856
$\chi^2_c = \sum_{i=1}^{m}\left[\frac{(O_i - e_i)^2}{e_i}\right]$	0.782	4.987	0.636	0.572	1.251

GROUP	50-60	60-70	70-80	80-90	90-100
NO OF ST. (O_i)	15	11	8	6	2
x_1	50	60	70	80	90
x_2	60	70	80	90	100
z_1	0.00	0.63	1.25	1.88	2.50
z_2	0.63	1.25	1.88	2.50	3.13
P_i	0.2357	0.1587	0.0755	0.0239	0.0053
$e_i = 80*P_i$	18.856	12.696	6.040	1.912	0.424
$\chi^2_c = \sum_{i=1}^{m}\left[\frac{(O_i - e_i)^2}{e_i}\right]$	0.789	0.227	0.636	8.740	5.858

$\chi^2_c = \sum_{i=1}^{m}\left[\frac{(O_i - e_i)^2}{e_i}\right] = 0.782 + 4.987 + ... + 8.740 + 5.858 = 24.478$ is calculated.

If $\chi^2_c \le \chi^2_t$ then the sample fits the theoretical distribution, and vice versa. χ^2_t is read for 1-α . In reading the tabulated χ^2 values, the degree of freedom is d_f = m-3 for Normal distribution. d_f = m – 3 = 10 – 3 =7 for 1% significance level (α = 0.01), 1 - α = 0.99, From Table 5.2 $\Rightarrow \chi^2_t = 18.475$ is read.

Conclusion: Since $\chi^2_c = 24.878 > \chi^2_t = 18.475 \Rightarrow$ The data do not fit Normal Distribution for 5% significance level.

for 5% significance level (α = 0.05), 1 - α = 0.95, From Table 5.2 $\Rightarrow \chi^2_t = 14.067$ is read.

Conclusion: Since $\chi^2_c = 24.878 > \chi^2_t = 14.0675 \Rightarrow$ The data do not fit Normal Distribution for 1% significance level.

Example 6.6: Classified values of 50 yearly mean annual discharges of a stream (m^3/s) are given as follows. The mean is 24.0 m^3/s and the standard deviation is 3.80 m^3/s.

a. Determine the 90 % confidence interval of mean of discharges for the stream population.

b. Suppose (assume) that the data fit normal distribution. If 100 yearly observations were made, estimate the number of years which have greater than 20 m^3/s discharge,

Solution:

c. Determine whether or not the data fit normal distribution for 1 % significance level.

Solution:

Example 6.7: The annual maximum discharge values of a gauging station are given below. Test the data whether or not fit Gumbel and Log Normal distributions by chi square (classifying into 10 groups) and probability plot correlation coefficient tests for α =0.01 and α =0.05.

22.4 25.3 26.5 17.3 24.0 48.4 27.7 33.5 41.3 24.5 38.4 28.4 13.5 16.4 22.2 23.5 22.8 21.4 29.0 34.4

21.5 42.4 47.0 47.3 29.3 44.6 38.0 33.3 53.1 27.0 23.6 34.4 29.0 42.3 51.5 36.4 24.5 47.0 34.5 27.6

Solution:

a. BASIC CALCULATIONS:

The ranked values of x from smaller to greater for the original values (for Gumbel Distribution):

13.5 16.4 17.3 21.4 21.5 22.2 22.4 22.8 23.5 23.6 24.0 24.5 24.5 25.3 26.5 27.0 27.6 27.7 28.4 29.0

29.0 29.3 33.3 33.5 34.4 34.4 34.5 36.4 38.0 38.4 41.3 42.3 42.4 44.6 47.0 47.0 47.3 48.4 51.5 53.1

Number of data: N = 40, the mean $\bar{x} = 31.88$, standard deviation $S_x = 10.313$are calculated.

The ranked values of x from smaller to greater for the logarithmic values (y = lnx) (for Normal Distribution):

2.60 2.80 2.85 3.06 3.07 3.10 3.11 3.13 3.16 3.16 3.18 3.20 3.20 3.23 3.28 3.30 3.32 3.32 3.35 3.37

3.37 3.38 3.51 3.51 3.54 3.54 3.54 3.59 3.64 3.65 3.72 3.74 3.75 3.80 3.85 3.85 3.86 3.88 3.94 3.97

Number of data: N = 40, the mean $\bar{x} = 3.4105$, standard deviation $S_x = 0.3316$are calculated.

b. CHI SQUARE TESTS:

b1: GUMBEL DISTRIBUTION:

$\bar{x} = 31.88, S_x = 10.313$, N = 40, From Gumbel Distribution Table 4.2 $\Rightarrow y_N = 0.544, S_N 1.141$,

Eq. (4.19) $\Rightarrow y_i = \frac{S_N}{S_x}\left(x_i - \bar{x} + S_x \frac{y_N}{S_N}\right) = \frac{1.141}{10.313}\left(x_i - 31.88 + 10.313\frac{0.544}{1.141}\right) = 0.1106(x_i - 26.963)$

Eq. (4.18a) $\Rightarrow 1 - F(x_i) = P(x < x_i) = e^{-e^{-y}} = e^{-e^{-0.1106(x_i - 26.963)}}$

Number of groups m = 10, the range is $x_{max} - x_{min}$ = 53.1 – 13.5 = 39.6,

The group interval = The range/number of groups = 39.6/10 = 3.96,

The first group: (13.5) and (13.5 + 3.96 = 17.46) $\Rightarrow$ (13.5 - 17.46),

The observed data between (13.5 and 17.46) are (13.5, 16.4 and 17.3) and number of data is O_i = 3

$P(x < 13.5) = e^{-e^{-0.1106(13.5 - 26.963)}} = 0.0119$, $P(x < 17.46) = e^{-e^{-0.1106(17.46 - 26.963)}} = 0.0572$

P(13.5<x<17.46) = 0.0572 – 0.0119 = 0.0453, expected value $e_i = NP_i = 40*0.0453 = 1.812$

$\chi^2_c = \frac{(O_i - e_i)^2}{e_i} = \frac{(3 - 1.812)^2}{1.812} = 0.779$

The second group (17.46) and (17.46 + 3.96 = 21.42) $\Rightarrow$(17.46 – 21.42),

The observed data between (17.46 and 21.42) is (21.4) and number of data is $O_i = 1$,

$P(x<17.46) = e^{-e^{-0.1106(17.46-26.963)}} = 0.0572$, $P(x<21.42) = e^{-e^{-0.1106(21.42-26.963)}} = 0.1579$,

P(17.46<x<21.42) = 0.1579 – 0.0572 = 0.1007, expected value $e_i = 40*0.1107 = 4.028$

$$\chi^2_c = \frac{(O_i - e_i)^2}{e_i} = \frac{(1-4.028)^2}{4.028} = 2.276$$

Similar calculations are made for all of the groups and presented in the following table:

GROUP										
O_i										
x_1										
x_2										
P_2										
P_1										
$\Delta P = P_2 - P_1$										
$e_i = \Delta P*40$										
$\chi^2_c = \frac{(O_i - e_i)^2}{e_i}$										

b_2: LOG NORMAL DISTRIBUTION:

Number of data: N = 40, the mean $\bar{x} = 3.4105$, standard deviation $S_x = 0.3316$

Eq. (4.11) $\Rightarrow z = \dfrac{x - \bar{x}}{S_x} = \dfrac{x - 3.4105}{0.3316}$,

Number of groups m = 10, range is $x_{max} - x_{min} = 3.97 - 2.60 = 1.37$, the group interval = 1.37/10 = 0.137,

The first group: (2.60) and (2.60 + 0.137 = 2.737) $\Rightarrow$ (2.60 – 2.737),

The observed data between (2.60 and 2.737) is (2.60) and number of data is $O_i = 1$

For the first group: $\Rightarrow z_1 = \dfrac{2.60 - 3.4105}{0.3316} = -2.44$, $z_2 = \dfrac{2.737 - 3.4105}{0.3316} = -2.03$

$P(2.60 < x < 2.737) = P(-2.44 < z < -2.03) = 0.4927 - 0.4788 = 0.0139$,

The expected value $e_i = 40*0.0139 = 0.556$, $\chi^2_c = \dfrac{(O_i - e_i)^2}{e_i} = \dfrac{(1 - 0.556)^2}{0.556} = 0.355$

The second group: (2.737) and (2.737 + 0.137 = 2.874) $\Rightarrow$ (2.737 – 2.874),

The observed data between (2.737 and 2.874) are (2.80 and 2.85) and number of data is $O_i = 2$

For the second group: $z_1 = \dfrac{2.737 - 3.4105}{0.3316} = -2.03$ $z_2 = \dfrac{2.874 - 3.4105}{0.3316} = -1.62$

$P(2.737 < x < 2.874) = P(-2.03 < z < -1.62) = 0.4788 - 0.4474 = 0.0314$,

The expected value $e_i = 40*0.0314 = 1.256$, $\chi^2_c = \dfrac{(O_i - e_i)^2}{e_i} = \dfrac{(2 - 1.256)^2}{1.256} = 0.441$.

Similar calculations are made for all of the groups and presented in the following table:

GROUP										
O_i										
z_1										
z_2										
P_1										
P_2										
$\Delta P = P_2 - P_1$										
$e_i = \Delta P*40$										
$\chi^2_c = \dfrac{(O_i - e_i)^2}{e_i}$										

c. PROBABILITY PLOT CORRELATION COEFFICIENT TESTS:

The correlation coefficient (r) between the theoretical values of a distribution (x_t) and observed values (x_o) is calculated. If the calculated value is greater than or equal to the critical value given in Table 5.2, then it is assumed that the observed data fit the related distribution. The correlation coefficient between x and y is calculated by Eq. (5.4):

$$r = \frac{N\sum xy - \sum x \sum y}{\left\{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]\right\}^{0.5}} = \frac{\sum\left[(x-\bar{x})(y-\bar{y})\right]}{NS_xS_y} = \frac{\sum xy - N\bar{x}\bar{y}}{NS_xS_y}$$

The values are ranked from smaller to greater, the probability of non exceedance of each data is calculated by Eq. (5.5) where, i is the rank number.

$$F(x < x_i) = 1 - p = \frac{i - 0.40}{N + 0.20}$$

When the data are ranked from minimum value to maximum value, N = 40, for the minimum value i = 1, for the second value i = 2, ... for the maximum value i = 40.

For the minimum value the probability of value less than minimum value is $F(x < x_i) = \frac{1-0.40}{40+0.20} = 0.0149$,

For the second value the probability of value less than second value is $F(x < x_i) = \frac{2-0.40}{40+0.20} = 0.0398$,

For the maximum value the probability of value less than maximum value is

$$F(x < x_i) = \frac{40-0.40}{40+0.20} = 0.9851.$$

c_1: GUMBEL DISTRIBUTION:

N = 40 $\Rightarrow y_N = 0.544, S_N 1.141,\ \bar{x} = 31.88,\ S_x = 10.313$

Eq. (3.14a) $\Rightarrow P(x < x_i) = e^{-e^{-y}} \Rightarrow y = -\ln\left(-\ln(-(P(x < x_i)))\right)$,

For example for the minimum value $y = -\ln\left(-\ln(-(0.0149))\right) = -1.437$,

For the second value $y = -\ln\left(-\ln(-(0.0398))\right) = -1.171$,

For the maximum value $y = -\ln\left(-\ln(-(0.9851))\right) = 4.199$.

Eq. (3.15a) $x_i = y_i \frac{S_x}{S_N} + \bar{x} - S_x \frac{y_N}{S_N} = \frac{10.313}{1.141} y_i + 31.88 - \frac{0.544}{1.141} 10.313 \Rightarrow x_i = 9.039 y_i + 26.963$

For the minimum value, $y_1 = -1.437 \Rightarrow x_1 = 9.039 * (-1.437) + 26.963 = 13.99$

This is the expected value for the minimum discharge, e_1 = 13.99, the observed minimum discharge is O_1 =13.5

For the second value $y_2 = -1.171 \Rightarrow x_2 = 9.039 * (-1.171) + 26.963 = 16.39$

This is the expected value for the second discharge, e_2 = 16.39, the observed minimum discharge is O_2 = 16.4

For the maximum value, $y_{40} = 4.199 \Rightarrow x_1 = 9.039*(4.199) + 26.963 = 64.9$

This is the expected value for the maximum discharge, e_{40}= 64.9, the observed minimum discharge is O_{40} =53.1

Similar calculations are made for all of the groups and presented in the following table:

i	1	2	3	4	5	6	7	8	9	10	11
P_i											
O_i											
y_i											
e_i											

i	12	13	14	15	16	17	18	19	20	21	22
P_i	0.2886	0.3134	0.3383	0.3632	0.3881	0.4129	0.4378	0.4627	0.4876	0.5124	0.5373
O_i	24.5	24.5	25.3	26.5	27.0	27.6	27.7	28.4	29.0	29.0	29.3
y_i	-0.217	-0.149	-0.080	-0.013	0.055	0.123	0.191	0.260	0.331	0.402	0.476
e_i	25.00	25.62	26.24	26.85	27.46	28.07	28.69	29.32	29.96	30.60	31.27

i	23	24	25	26	27	28	29	30	31	32	33
P_i	0.5622	0.5871	0.6119	0.6368	0.6617	0.6866	0.7114	0.7363	0.7612	0.7861	0.8109
O_i	33.3	33.5	34.4	34.4	34.5	36.4	38.0	38.4	41.3	42.3	42.4
y_i	0.552	0.631	0.711	0.796	0.884	0.978	1.077	1.184	1.299	1.424	1.563
e_i	31.95	32.66	33.39	34.15	34.96	35.80	36.70	37.66	38.70	39.84	41.09

i	34	35	36	37	38	39	40
P_i	0.8358	0.8607	0.8856	0.9104	0.9353	0.9602	0.9851
O_i	44.6	47.0	47.0	47.3	48.4	51.5	53.1
y_i	1.718	1.897	2.108	2.366	2.705	3.204	4.199
e_i	42.49	44.11	46.02	48.35	51.41	55.92	64.92

c2: LOG NORMAL DISTRIBUTION:

Number of data: N = 40, the mean $\bar{x} = 3.4105$, standard deviation $S_x = 0.3316$

Eq. (3.7) $\Rightarrow z = \dfrac{x - \bar{x}}{S_x} = \dfrac{x - 3.4105}{0.3316} \Rightarrow x = 0.3316z + 3.4105,$

For the minimum value, P = 0.0149 < 0.5 $\Rightarrow z < 0$, from table 1, P = 0.5 – 0.0149 = 0.4851 $\Rightarrow z = -2.17$

$\Rightarrow x = 0.3316(-2.17) + 3.4105 = 2.69$

This is the expected value for the minimum discharge, e_1 = 2.69, the observed minimum value is O_1 =2.60,

For the second value, P = 0.0398 < 0.5 $\Rightarrow z < 0$, from table 1, P = 0.5 – 0.0398 = 0.4602 $\Rightarrow z = -1.75$

$\Rightarrow x = 0.3316(-1.75) + 3.4105 = 2.83$

This is the expected value for the second discharge, e_2 = 2.69, the observed second value is O_2 =2.80,

For the maximum value, P = 0.9851 > 0.5 $\Rightarrow z > 0$, from table 1, P = 0.9851 – 0.5 = 0.4851 $\Rightarrow z = 2.17$

$\Rightarrow x = 0.3316(2.17) + 3.4105 = 4.13$

This is the expected value for the maximum discharge, e_{40} = 4.13, the observed maximum value is O_{40} = 3.97,

Similar calculations are made for all of the groups and presented in the following table:

i	1	2	3	4	5	6	7	8	9	10	11
P_i											
O_i											
z_i											
e_i											

i	12	13	14	15	16	17	18	19	20	21	22
P_i	0.2886	0.3134	0.3383	0.3632	0.3881	0.4129	0.4378	0.4627	0.4876	0.5124	0.5373
O_i	3.20	3.20	3.23	3.28	3.30	3.32	3.32	3.35	3.37	3.37	3.38
z_i	-0.56	-0.49	-0.42	-0.35	-0.30	-0.22	-0.16	-0.09	-0.03	0.03	0.09
e_i	3.22	3.25	3.27	3.29	3.31	3.34	3.36	3.38	3.40	3.42	3.44

i	23	24	25	26	27	28	29	30	31	32	33
P_i	0.5622	0.5871	0.6119	0.6368	0.6617	0.6866	0.7114	0.7363	0.7612	0.7861	0.8109
O_i	3.51	3.51	3.54	3.54	3.54	3.59	3.64	3.65	3.72	3.74	3.75
z_i	0.16	0.22	0.30	0.35	0.42	0.49	0.56	0.63	0.71	0.79	0.88
e_i	3.46	3.48	3.50	3.52	3.54	3.56	3.58	3.61	3.63	3.66	3.68

i	34	35	36	37	38	39	40
P_i	0.8358	0.8607	0.8856	0.9104	0.9353	0.9602	0.9851
O_i	3.80	3.85	3.85	3.86	3.88	3.94	3.97
z_i	0.98	1.09	1.20	1.34	1.52	1.75	2.17
e_i	3.72	3.75	3.81	3.85	3.92	3.99	4.13

CHAPTER 7

REGRESSION ANALYSIS

7.1. GENERAL INFORMATION

7.1.1. Introduction

Most of the variables in engineering are statistically dependent of each other, thus there is a relation between these variables. For example, there is relation between strength of concrete and its cement ratio or between flow and precipitation of a basin. One of these variables (for example precipitation) is defined as *independent variable* and the other variable (flow) is called *dependent variable* of the independent variable. The reason for such a relation is either one variable is affected by the other or both variables are affected by other variable(s). As an example, the relation between precipitation and flow in a basin originates because flow takes place due to effect (as a consequence) of precipitation. The relation between flows in neighboring basins arises due to fact that the flows are affected by the precipitation of that region.

The relations have not a deterministic (functional) character; in other words, when one of the variables takes a certain value, the other will not always take the same value. This value will change more or less in various observations with the effect of other variables which have not been considered in the relation. Moreover, the estimations depend on the very limited data of the sample, of which degree of representation to the population is questionable. For example, when flow of one of the neighboring basins takes a certain value, the flow of the other basin does not always take the same value. Nevertheless, the determination of the *existence* and the *form* of a nonfunctional relationship between the variables has a great importance in practice. Because, by using this relationship, it is possible to estimate a future value of a variable depending on known value(s) of another (or more than one) variable(s). While this estimate will not be the *exact* future value of the variable under consideration, it will be the *best* estimate *closest* to this value. The mathematical expression showing a relation of the above mentioned type is called the *regression equation*. The aims of the regression analysis are:

- To check whether there is a significant relation between the variables under consideration, and if there is one, to obtain regression equation expressing this relation,
- To study and evaluate the reliability of the estimates to be made by using this equation.

The relation between flows of two neighboring basins recorded for same years is an example to the usage of regression analysis in Civil Engineering. If the regression equation can be obtained, the missing data unrecorded in the past in one of the basins can be *estimated* by using this equation. The estimated values are the best ones, though they are not equal to the real values which would have been observed.

7.1.2. Types of Regression

Depending on the number of independent variable(s) and the type of relation, there are three types of regression equations:

a. Simple Linear Regression: There is one dependent and one independent variables and the relation between these variables is linear. This is the simplest type of regression equations.

b. Multivariate Linear Regression: The number of independent variables is greater than one and the relation is linear.

c. Nonlinear Regression: The relation between dependent variable and independent variable(s) is nonlinear (for example parabolic, exponential etc). The regression is transformed into linear form by using appropriate transformation techniques.

7.2. SIMPLE LINEAR REGRESSION

Simple linear regression equation is as follows:

$$y = a + bx \tag{7.1}$$

Where, y is dependent and x is independent variables, a and b are regression coefficients. When x and y values are plotted on a coordinate, if the data are scattered around a line, it is supposed that there is a linear regression between x and y (Figure 7.1).

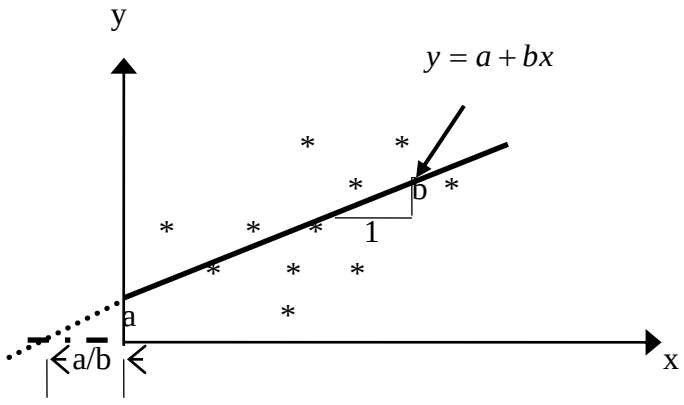


Figure 7.1 Regression Line

7.2.1. Correlation Coefficient

Correlation coefficient (r) shows the degree of the relation and the reliability of the regression equation between x and y values. r values change between the interval of -1 and +1. As the reliability of the regression equation increases, the absolute value of r approaches 1, which also explains that the relation and dependence between x and y are strong. If r value is positive (negative) a positive (negative) relation there exist between x and y; if it is equal to zero $(r = 0)$, it is obvious that there is no relation. $r = +1$ and $r = -1$ imply that, there is exact positive and negative (the values of a variable increase while values of other variable decrease) relations. When the observed x and y values are plotted, except the situation $|r| = 1$, a scatter is observed. As the scattering increases, $|r|$ approaches zero; in the case of haphazard scattering correlation coefficient is nearly equal to zero (Figure 7.2).

An estimation value for correlation coefficient of the population (ρ) is calculated as follows:

$$r = \frac{N\sum xy - \sum x \sum y}{\left\{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]\right\}^{0.5}} = \frac{\sum\left[(x-\bar{x})(y-\bar{y})\right]}{NS_xS_y} = \frac{\sum xy - N\bar{x}\bar{y}}{NS_xS_y} \qquad (7.2)$$

Note that, in this equation, during the calculation of S_x and S_y values, even if the number of data is less than 30, $S_x{}^2 = \sum(x-\bar{x})^2 / N$ and $S_y{}^2 = \sum(y-\bar{y})^2 / N$ equations should be used and the value of (N-1) should not be used instead of N.

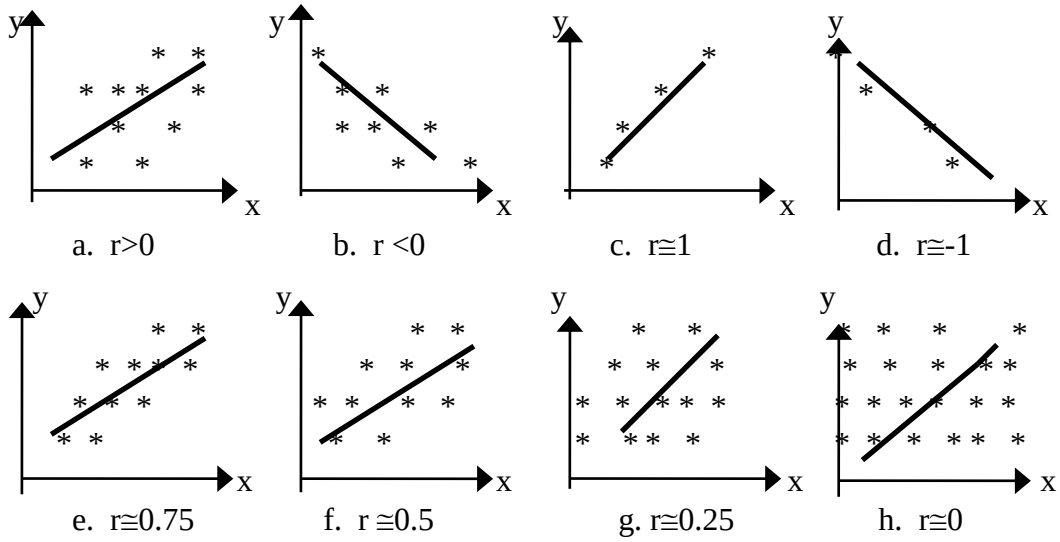


Figure 7.2. Correlation Coefficient Values for Various Scattering Scenarios

Even if the correlation coefficient calculated from the sample is different from zero (r≠0), as a result of sampling errors, it is possible that the correlation coefficient of the population is equal to zero (ρ=0). In order to check this, the calculated r values are compared to the critical values (r_{cr}), which are tabulated according to significance level (α) and degree of freedom (d_f = N-2) (Table 7.1). If $|r| \geq r_{cr}$, then it is concluded that the correlation coefficient is not equal to zero and that the regression equation is reliable.

Table 7.1. Critical Values of Correlatıon Coefficıent (d_f=N-2)

	α				α	
d_f	0.05	0.01		d_f	0.05	0.01
1	.999	.999		2	.950	.999
3	.878	.959		4	.811	.917
5	.754	.875		6	.707	.834
7	.666	.798		8	.632	.765
9	.602	.735		10	.576	.708
11	.553	.684		12	.532	.661
13	.514	.641		14	.497	.623
15	.482	.606		16	.468	.590
17	.456	.575		18	.444	.561
19	.433	.549		20	.423	.537
22	.404	.515		24	.388	.496
26	.374	.479		28	.361	.463
30	.349	.304		40	.304	.393
50	.273	.354		100	.195	.254

7.2.2. Regression Equation

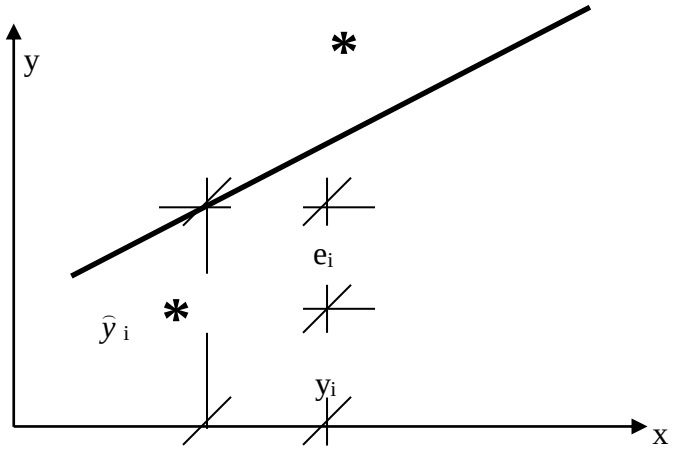
The regression coefficients of simple linear regression equation (a and b) are calculated as follows: The sum of the squares of the vertical distance values (e) between the data and regression line should be minimum (Figure 7.3). Therefore, the following function should be minimal:

$$\sum e^2 = (y - \hat{y})^2 = \sum [y - (a + bx)]^2 = \sum (y - a - bx)^2 \tag{7.3}$$

The partial derivatives of a and b should be equal to zero, separately:

$$\frac{\partial e}{\partial a} = 2\sum (y - a - bx)(-1) = 2\sum (-y + a + bx) = 0 \tag{7.4a}$$

$$\frac{\partial e}{\partial b} = 2\sum (y - a - bx)(-x) = 2\sum (-xy + ax + bx^2) = 0 \tag{7.4b}$$



$\hat{y}_i = a + bx_i$,

$\hat{y}_i \pm e_i = y_i \Rightarrow y_i \pm e_i = {}_i$

$e_i = y_i \mp \hat{y}_i$

Figure 7.3 Vertical Distances from Regression Equation

From these equations one can obtain:

$$\sum (-y + a + bx) = 0 \quad \Rightarrow \quad -\sum y + Na + b\sum x = 0 \tag{7.5a}$$

$$\sum (-xy + ax + bx^2) = 0 \quad \Rightarrow \quad -\sum xy + a\sum x + b\sum x^2 = 0 \tag{7.5b}$$

The following equations are obtained from these equations:

$$\sum y = Na + b\sum x \tag{7.6a}$$

$$\sum xy = a\sum x + b\sum x^2 \tag{7.6b}$$

These equations are called *normal equations*. By solving them, a and b are calculated as:

$$b = \frac{\sum [(x - \bar{x})(y - \bar{y})]}{\sum (x - \bar{x})^2} = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2} = r\frac{S_y}{S_x}, \tag{7.7a}$$

$$a = \bar{y} - b\bar{x} \tag{7.7b}$$

Note that; in the last expression of (7.7a), since the standard deviation values are always positive, b values may be positive (negative) if r values are positive (negative).

NOTE: If (7.1) ($y = a + bx$) is summed from 1 to N, the following equation is obtained;

$$\sum y = \sum a + b\sum x = Na + b\sum x \tag{7.6a}$$

and if it is multiplied by x and ($xy = ax + bx^2$) and is summed the following equation is obtained:

$$\sum xy = \sum ax \sum bx^2 = a\sum x + b\sum x^2 \tag{7.6b}$$

7.3. MULTIVARIATE LINEAR REGRESSION

The multivariate regression equation is as follows:

$$y = a_0 + a_1x_1 + a_2x_2 + ... + a_kx_k = a_0 + \sum_{i=1}^{K} a_ix_i \tag{7.8}$$

If the number of independent variable is K=2, then the equation is as follows:

$$y = a_0 + a_1x_1 + a_2x_2 \tag{7.9}$$

If this equation is summed from 1 to N, then is multiplied by x_1 and x_2 and summed from 1 to N, the following normal equations are obtained:

$$y = a_0 + a_1x_1 + a_2x_2 \quad \Rightarrow \quad \sum y = Na_0 + a_1\sum x_1 + a_2\sum x_2 \tag{7.10a}$$

$$yx_1 = a_0x_1 + a_1x_1^2 + a_2x_1x_2 \quad \Rightarrow \quad \sum yx_1 = a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_1x_2 \tag{7.10b}$$

$$yx_2 = a_0x_2 + a_1x_1x_2 + a_2x_2^2 \quad \Rightarrow \quad \sum yx_2 = a_0\sum x_2 + a_1\sum x_1x_2 + a_2\sum x_2^2 \tag{7.10c}$$

By solving these equations, the regression coefficients a_0, a_1, a_2 are calculated and the regression equation is obtained.

Multivariate correlation coefficient is calculated as:

$$R = \left(1 - \frac{S_e^2}{S_y^2}\right)^{0.5} \tag{7.11}$$

S_e^2 is the variance of the vertical distances from the regression equation (e) and is calculated as:

$$S_e^2 = \frac{\sum e^2}{N - K - 1} = \frac{\sum (y_r - y_c)^2}{N - K - 1} \tag{7.12a}$$

Since the number of independent variable K=2, the equation is obtained as:

$$S_e^2 = \frac{\sum e^2}{N - 3} = \frac{\sum (y_r - y_c)^2}{N - 3} \tag{7.12b}$$

Here, e values are the differences between real and calculated y values (y_r and y_c); S_y^2 is the variance of real y values.

NOTE: If the number of independent variables is greater than 2, similar calculations are made. For example: In the case of 3 independent variables (K=3), the regression equation is as follows:

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 \tag{7.13}$$

Sum the equation from 1to N: $\sum y = Na_0 + a_1 \sum x_1 + a_2 \sum x_2 + a_3 \sum x_3$

Multiply by x_1 and sum $\sum x_1 y = a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_1 x_2 + a_3 \sum x_1 x_3$

Multiply by x_2 and sum $\sum x_2 y = a_0 \sum x_2 + a_1 \sum x_1 x_2 + a_2 \sum x_2^2 + a_3 \sum x_2 x_3$

Multiply by x_3 and sum $\sum x_3 y = a_0 \sum x_3 + a_1 \sum x_1 x_3 + a_2 \sum x_2 x_3 + a_3 \sum x_3^2$ (7.14)

By solving these equations, a_0, a_1, a_2, a_3 are found.

7.4. NONLINEAR REGRESSION

The linear regression model is the model most frequently used because of its simplicity; however, its usage may be erroneous in the case of very small correlation coefficients and then nonlinear regression analysis should be employed. There are numerous types of nonlinear regression. In the analysis of all of them, the regression is transformed into linear form by using appropriate transformation techniques. The normal equations are established and both regression coefficients $(a_0, a_1, ..a_K)$ and correlation coefficient (r) are calculated. Some of the nonlinear regression types that are frequently encountered and their normal equations are given in the following. In the analysis of other nonlinear regressions, similar principles are employed:

7.4.1. Polynomial Function

The most common form of a polynomial is as follows:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_K x^K \tag{7.15a}$$

If K=2, the polynomial is called *parabola*:

$$y = a_0 + a_1 x + a_2 x^2 \tag{7.15b}$$

If the transformations of $x_1 = x$ and $x_2 = x^2$ are made, the equation is obtained $y = a_0 + a_1 x_1 + a_2 x_2$, as a multivariate linear regression (Eq. 7.9). Normal equations are found by Eq. (7.10 a, b c) and regression coefficients (a_0, a_1, a_2) are calculated. The multivariate correlation coefficient is calculated by (7.11). In the cases of K is greater than 2, similar procedure is employed.

7.4.2. Exponential Function

Exponential function is as follows:

$$y = ab^x \tag{7.16a}$$

This equation is made linear by using the following transformation:

$$\log y = \log a + x \log b \tag{7.16b}$$

Let $Y = \log y$, $A = \log a$ and $B = \log b$, the equation is transformed to a linear one:

$$Y = A + Bx \tag{7.16c}$$

Correlation coefficient is calculated by Eq. (6.2). If the Eq. (6.16c) is summed and then multiplied by x and summed, its normal equations are obtained as:

$$\sum Y = NA + B\sum x, \quad \sum xY = A\sum x + B\sum x^2 \tag{7.16d}$$

By using these equations, firstly A and B and then $a = 10^A$ and $b = 10^B$ regression coefficients are found and inserted in Eq. (7.16a), thus, the regression equation is obtained.

7.4.3. Hyperbolic Function

Hyperbolic function is as follows:

$$y = ax^b \Rightarrow \quad \log y = \log a + b \log x \tag{7.17a}$$

If $Y = \log y$, $A = \log a$ and $X = \log x$ are written, then the equation is linearized as:

$$Y = A + bX \tag{7.17b}$$

Correlation coefficient is calculated by Eq. (7.2). If Eq. (7.17b) is summed and then multiplied by X and summed, normal equations are obtained as follows:

$$\sum Y = NA + b\sum X, \quad \sum XY = A\sum X + b\sum X^2 \tag{7.17c}$$

By using these equations, A (and $a = 10^A$) and b are found, by inserting these values in Eq. (7.17a), the regression equation is obtained.

7.4.4. Geometric Function

Geometric function is as follows:

$$y = \frac{1}{a + bx} \quad \Rightarrow \quad \frac{1}{y} = a + bx \tag{7.18a}$$

If $Y = 1/y$ is transformation is made, Eq. (7.18b) is obtained as a linear one:

$$Y = a + bx \tag{7.18b}$$

Correlation coefficient is calculated by Eq. (7.2). If Eq. (7.18b) is summed and then multiplied by x and summed, normal equations are obtained as follows:

$$\sum Y = Na + b\sum x, \quad \sum xY = a\sum x + b\sum x^2 \tag{7.18c}$$

By using these equations, a and b are found, by inserting these values in Eq. (7.18a), the regression equation is obtained.

7.5. EXAMPLES

Example 7.1: There is simple linear regression between the following x and y values. Calculate the correlation coefficient, test its reliability (whether correlation coefficient of the population is equal to zero) for 0.01 and 0.05 significance levels, and if it is suitable, calculate the regression equation; by using this equation estimate the y value for x = 5, and x value for y =14.

									Solution: $\sum$
x	22	19	16	13	9	7	4		90
y	8	11	15	17	15	16	20		102

Eq. (7.2) $\Rightarrow r = \dfrac{N\sum xy - \sum x \sum y}{\left\{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]\right\}^{0.5}} = \dfrac{\sum\left[(x-\bar{x})(y-\bar{y})\right]}{NS_xS_y} = \dfrac{\sum xy - N\bar{x}\bar{y}}{NS_xS_y}$

Calculation of correlation coefficient: Only the last equation ($r = \dfrac{\sum xy - N\bar{x}\bar{y}}{NS_xS_y}$) will be used.

$N = 7, \bar{x} = 90/7 = 12.857, S_x = 6.081, \bar{y} = 102/7 = 14.571, S_y = 3.659$

Example 7.2: Linear regression is supposed between the 7 daily and 28 daily strength values (kg/cm^2) of a concrete population. The measured 7 daily (x) and 28 daily (y) strength values of 10 samples are given below.

a. Calculate correlation coefficient and test its reliability for α=0.01 and α=0.05 significance levels,

b. Obtain the regression equation and estimate 28 daily (y) strength values for x=280 kg/cm^2 and x= 300 kg/cm^2 and also estimate x values for y= 210 kg/cm^2 and 250 kg/cm^2.

x	230	245	225	256	238	219	227	243	234	228
y	290	305	280	315	286	280	284	302	289	275

Data		**Solution: a.**							
x	y	xy	x^2	y^2	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x}).(y-\bar{y})$
230	290	66 700	52 900	84 100	-4.5	-0.6	20.25	0.36	2.7
245	305	74 425	60 025	93 025	10.5	14.4	110.25	207.36	151.2
225	280	63 000	50 625	78 400	-9.5	-10.6	90.25	112.36	100.7
256	315	80 640	65 536	99 225	21.5	24.4	462.25	595.36	524.6
238	286	68 068	56 644	81 796	3.5	-4.6	12.25	21.16	-16.1
219	280	61 320	47 961	78 400	-15.5	-10.6	240.25	112.36	164.3
227	284	64 468	51 529	80 656	-7.5	-6.6	56.25	43.56	49.5
243	302	73 386	59 049	91 204	8.5	11.4	72.25	129.96	96.9
234	289	67 626	54 756	83 521	-0.5	-1.6	0.25	2.56	0.8
228	275	62 700	51 984	75 625	-6.5	-15.6	42.25	243.36	101.4
				$\sum$					
2345	2906	682633	551009	845952	0	0	1106.5	1468.4	1176.0

N = 10, $\bar{x} = 234.5$, $\bar{y} = 290.6$, S_x = 10.52, S_y = 12.12

Eq. (7.2)

$$\Rightarrow r = \frac{N\sum xy - \sum x \sum y}{\left\{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]\right\}^{0.5}} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{NS_xS_y} = \frac{\sum xy - N\bar{x}\bar{y}}{NS_xS_y} =$$

$$= \frac{10*682633 - 2345*2906}{\left\{\left[10*55109 - 2345^2\right]\left[10*845952 - 2906^2\right]\right\}^{0.5}} = \frac{1176}{10*10.52*12.52} = \frac{682633 - 10*234.5*290.6}{10*10.52*12.52}$$

$\Rightarrow r = 0.923,$ Absolute value $|r| = 0.923$, from Table 7.1, $d_f = N - 2 = 10 - 2 = 8$,

For 0.01 significance level (α =0.01) $\Rightarrow r_{cr} = 0.765$, $|r| = 0.923 > r_{cr}$, therefore, the correlation coefficient of the population is not equal to zero, the this value is reliable,

For 0.05 significance level (α =0.05) $\Rightarrow r_{cr} = 0.632$, $|r| = 0.923 > r_{cr}$, therefore, the correlation coefficient of the population is not equal to zero, the this value is reliable.

Calculation of regression coefficient and obtaining of regression equation:

Eq. (7.1): $y = a + bx$

Normal equations (6.6a and 6.6b): $\sum y = Na + b\sum x \Rightarrow 2906 = 10a + 2345b$,

$$\sum xy = a\sum x + b\sum x^2 \Rightarrow 682633 = 2345a + 551009b$$

By solving these equations one may obtain a = 41.327 and b = 1.063 are obtained.

Or by using Eq. (7.7a)

$$b = r\frac{S_y}{S_x} = 0.923\frac{12.12}{10.52} = 1.063,\ a = \bar{y} - b\bar{x} = 290.6 - 1.063*234.5 = 41.327$$

Regression Equation: $y = a + bx = 41.327 + 1.063x$,

For x = 210 $\Rightarrow y = 41.327 + 1.063*210 = 264.6 kg/cm^2$,

For x = 250 $\Rightarrow y = 41.327 + 1.063*250 = 307.1 kg/cm^2$

For y = 280 $\Rightarrow y = 280 = 41.237 + 1.063x \Rightarrow x = 224.5 kg/cm^2$ and

For y = 300 $\Rightarrow y = 300 = 41.237 + 1.063x \Rightarrow x = 243.3 kg/cm^2$

values are obtained.

Example 7.3: The equation among the following data is $y = x_1 + 2x_2 + 3x_3 - 4x_4$. Calculate the correlation coefficient and test its reliability for α=0.05 and α=0.01 significance levels.

GIVEN DATA	x_1	1	2	2	3	3	3	3
	x_2	2	2	3	3	3	4	4
	x_3	3	4	3	4	5	4	5
	x_4	4	4	5	5	4	3	6
	$y = y_r$	-3	3	-3	0	6	10	2
	Solution:							
CALCU-LATIONS								

Example 7.4: Multiple linear regression is supposed between the mean annual precipitation height values of A gauge station (y) and B (x_1) and C (x_2) gauge stations ($y = a_0 + a_1x_1 + a_2x_2$). Observed values for 6 years (cm) are given below.

a. Obtain the regression equation, estimate the mean annual precipitation value (y) in A gauge station for x_1=150 cm and x_2= 80 cm.

b. Calculate the multiple correlation coefficient.

y	120.3	140.5	135.2	105.4	138.3	146.5
x_1	183.5	210.4	201.3	163.2	214.0	260.3
x_2	96.3	101.5	98.0	88.5	103.1	118.2

Solution: a.

	x_1	x_2	y	x_1* x_2	x_1*y	x_2*y	x_1^2	x_2^2
	183.5	96.3	120.3	17671	22075	11585	33672	9274
	210.4	101.5	140.5					
	201.3	98.0	135.2					
	260.3	118.2	146.5					
$\sum$	1233	605.6	786.2	126028	163783	79998	258648	61613

$\sum y = Na_0 + a_1\sum x_1 + a_2\sum x_2$ $\Rightarrow$ 786.2 = 6a_0 + 1233a_1 + 605.6a_2

$\sum yx_1 = a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_1x_2$ $\Rightarrow$ 163783 = 1233a_0 + 258648a_1 + 126028a_2

$\sum yx_2 = a_0\sum x_2 + a_1\sum x_1x_2 + a_2\sum x_2^2$ $\Rightarrow$ 79998 = 606a_0 + 126028a_1 + 61613a_2

By solving these equations, one can obtain a_0 = 183.984 a_1 = 1.465 a_2 = -3.506
and regression equation is $y = 183.984 + 1.465x_1 - 3.506x_2$

$y = 183.984 + 1.465*150 - 3.506*80 = 123.254cm$

b.

y_r	x_1	x_2	y_c	e	e^2
120.3	183.5	96.3	115.18	5.12	26.21
140.5	210.4	101.5			
135.2	201.3	98.0			
105.4					
138.3					
146.5					
146.5	260.3	118.2	150.91	-4.41	19.45
			$\sum$	0.0	122.57

S_y^2 = 195.2 , $S_e^2 = \frac{\sum e^2}{N-K-1} = \frac{\sum (y_r - y_c)^2}{N-K-1} = \frac{122.57}{6-2-1} = 40.86,$

$R = \left(1 - \frac{S_e^2}{S_y^2}\right)^{0.5} = \left(1 - \frac{40.86}{195.2}\right)^{0.5} = 0.889$

Example 7.5: There is an equation of $y = 2 - 3x_1 + 2x_2 + x_3$ among the following data. Obtain the correlation coefficient.

x_1	1	2	3	4	5	6
x_2	2	4	2	2	10	10
x_3	2	3	1	2	2	1
y	6	5	-3	-7	10	9
Solution:						
y_c						
e						
e^2						

$\sum e^2 = 32$, $S_y{}^2 = 38.889$, $\sum e^2 = 32$, $S_e{}^2 = \frac{32}{6-3-1} = 16$, $R = \left(1 - \frac{16}{38.889}\right)^{0.5} = 0.767$

Example 7.6. There is a regression equation of $y = e^{a+bx^3}$ among the following data. Calculate the correlation coefficient; test its reliability for 0.01 and 0.05 significance levels. Obtain the regression equation, calculate y value for x = 0.7 and x value for y = 1.7

x	-0.5	0.0	0.5	1.0	1.2	1.3
y	2.1	2.0	1.9	1.5	1.3	1.2
Solution: $y = e^{a+bx^3} \Rightarrow \ln y = a + bx^3, Y = \ln y, X = x^3 \Rightarrow Y = a + bX$						
$X = x^3$	-0.125	0.000	0.125	1.000	1.728	2.197
Y = lny	0.742	0.693	0.642	0.405	0.262	0.182
X*Y	-.0927	.000	.0802	.4055	.4534	.4006

$\sum X = 4.925, \sum Y = 2.927, \sum XY = 1.2469$, $\bar{X} = 0.8208, S_X = 0.8946, \bar{Y} = 0.4878, S_Y = 0.2166$

$r = \frac{1.2469 - 6 * 0.8208 * 0.4878}{6 * 0.8946 * 0.2166} = -0.994, |r| = 0.994$, N = 6, d_f = N – 2 = 6 – 2 =4, from Table 7.1 $\Rightarrow$ for

α =0.05 r_c = 0.811 < $|r| = 0.994 \Rightarrow$ It is reliable,

$\Rightarrow$ for α =0.01 r_c = 0.917 < $|r| = 0.994 \Rightarrow$ It is reliable

$b = r\frac{S_Y}{S_X} = -0.994\frac{0.2166}{0.8946} = -0.2407$, $a = \bar{Y} - b\bar{X} = 0.4878 - (-0.2407) * 0.8208 = 0.6853$

The regression equation is: $y = e^{a+bx^3} = e^{0.6853 - 0.2407x^3}$,

For $x = 0.7 \Rightarrow y = e^{0.6853 - 0.2407 * 0.7^3} = 1.827$,

for $y = 1.7 = e^{0.6853 - 0.2407 * 0.7^3} \Rightarrow \ln 1.7 = 0.5306 = 0.6853 - 0.2407 * x^3 \Rightarrow x = 0.863$

Example 7.7: There is a regression equation of $y = \ln(a + bx^2)$ between the following x and y values. Determine the regression equation and estimate y value for x=0.5. Calculate the correlation coefficient and test its reliability for α=0.05 and α=0.01.

x	1.0	1.5	2.0	2.5	3.0
y	0.7	0.5	0.25	0.0	-0.3

Solution:

$y = \ln(a + bx^2) \Rightarrow e^y = a + bx^2$, if the transformations of $Y = e^y$ and $X = x^2$ are made, the following equation is obtained: $e^y = Y = a + bx^2 + a + bX$, Then, a simple linear equation between $X = x^2$ and $Y = e^y$ is carried out.

Example 7.8: There is

a. Linear, **b.** Equation of $y = \log(a + bx^{0.5})$

between the following variables. For both of the equations, calculate the correlation coefficients; obtain the regression equations and estimate y values for x=2.0 and x values for y=0.5.

x	1.3	1.5	1.8	2.4	2.7
y	0.8	0.63	0.55	0.43	0.35

Solution:

Example 7.9: Obtain the **a.** Exponential, **b.** Hyperbolic functions for the following x and y values. Calculate correlation coefficients and determine which one more reliable is.

Solution:

a. Exponential function: Eq. (7.16a) $\Rightarrow y = ab^x$

This equation is made linear by using the following transformation: $\log y = \log a + x \log b$

Let $Y = \log y$, $A = \log a$ and $B = \log b$, the equation is transformed to linear: $Y = A + Bx$

								$\sum$
GIVEN DATA	x	0.5	1.0	1.5	2.0	2.5	3.0	10.5
	y	2	8	18	30	70	100	228
SOLU-TION:	Y = log(y)	0.301	0.903	1.255	1.477	1.845	2.000	7.782
	x*Y	0.151	0.903	1.883	2.954	4.613	6.000	16.503

Number of data N = 6, basic calculations on x and Y yield:

$\bar{x} = 10.5/6 = 1.75$, $S_x = 0.854$, $\bar{Y} = 7.782/6 = 1.297$, $S_Y = 0.574$, $\sum x*Y = 16.503$

$$r = \frac{\sum xY - N\bar{x}\bar{Y}}{NS_xS_Y} = \frac{16.503 - 6*1.75*1.297}{6*0.854*0.574} = 0.981$$

$$B = \log b = r\frac{S_Y}{S_x} = 0.981\frac{0.574}{0.854} = 0.6592 \Rightarrow b = 10^B = 10^{0.6592} = 4.562$$

$$A = \log a = \bar{Y} - B\bar{x} = 1.297 - 0.6592*1.75 = 0.1434 \Rightarrow a = 10^A = 10^{0.1434} = 1.391$$

Regression equation: $y = ab^x = 1.391*4.562^x$

b. Hyperbolic Function: Eq. (7.17a) $\Rightarrow y = ax^b$

This equation is made linear by using the following transformation: $\log y = \log a + b \log x$

Let $Y = \log y$, $A = \log a$ and $X = \log x$, the equation is transformed to linear: $Y = A + bX$

								$\sum$
GIVEN DATA	x	0.5	1.0	1.5	2.0	2.5	3.0	-
	y	2	8	18	30	70	100	-
SOLU-TION:	Y = log(y)	0.301	0.903	1.255	1.477	1.845	2.000	7.782
	X = log(x)	-0.301	0.000	0.176	0.301	0.398	0.477	1.051
	X*Y	-0.0906	0.0000	0.2209	0.4446	0.7343	0.9540	2.263

Number of data N = 6, basic calculations on X and Y yield:

$\bar{X} = 1.051/6 = 0.175$, $S_X = 0.263$, $\bar{Y} = 7.782/6 = 1.297$, $S_Y = 0.574$, $\sum X*Y = 2.263$

$$r = \frac{\sum XY - N\bar{X}\bar{Y}}{NS_XS_Y} = \frac{2.263 - 6*0.175*1.297}{6*0.263*0.574} = 0.995$$

$$b = r\frac{S_Y}{S_X} = 0.995\frac{0.574}{0.263} = 2.171$$

$$A = \log a = \bar{Y} - b\bar{X} = 1.297 - 2.171*0.175 = 0.9173 \Rightarrow a = 10^A = 10^{0.9173} = 8.265$$

Regression equation: $y = ax^b = 8.265*x^{2.171}$

Conclusion: Since the r value is greater, hyperbolic function is more reliable.

Example 7.10: The relation between discharge (Q, m^3/s) and water depth (h, m) (rating curve) in a stream is $Q = ah^b$. Obtain the regression equation considering the following data. Calculate Q value for h = 5 m and h value for Q = 100 m^3/s.

Solution: The type of the regression equation $Q = ah^b$ is hyperbolic function: $\Rightarrow y = ax^b$, so the equation is made linear by using the following transformation: $\log Q = \log a + b \log h$

Let $Y = \log Q$, $A = \log a$ and $X = \log h$, the equation is transformed to linear: $Y = A + bX$

								$\sum$
GIVEN DATA	h (m)	1.25	1.50	2.00	2.50	3.00	3.50	-
	Q (m^3/s)	4	7	13	25	44	70	-
SOLU-TION:								