Statistics 401: An Introduction to Statistics for Engineers and Scientists

Michael G. Akritas

Penn State University

Spring 2008

Library of Congress Cataloging

An Introduction to Statistics for Engineers and Scientists/Michael G. Akritas

Contents

Chapter 1

Basic Statistical Concepts

1.1 Why Statistics?

Statistics deals with collecting, processing, summarizing, analyzing and interpreting data. On the other hand, engineering and industrial management deal with such diverse issues as solving production problems, effective use of materials and labor, development of new products, quality improvement and reliability and, of course, basic research. The usefulness of statistics as a tool for dealing with the above problems is best seen by considering some examples of specific engineering tasks or studies, which we now mention.

Example 1.1.1. Examples of specific engineering studies include:

1. estimating the coefficient of thermal expansion of a metal,

2. comparing the hardness of two or more cement mixtures,

3. comparing the effectiveness of three cleaning products in removing four different types of stains,

4. predicting failure time on the basis of stress applied,

5. studying the relation between vibrations of high-speed train tracks and speed of the train,

6. assessing the effectiveness of a new traffic regulatory measure in reducing the weekly rate of accidents,

7. testing a manufacturer's claim regarding the quality of his/her product,

8. studying the relation between salary increases and employee productivity in a large corporation,

9. estimating the proportion of the US citizens of age 18 and over who are in favor of expanding solar energy sources, and

10. determining whether the content of lead in the water of a certain lake is within the safety limit.

The reason why tasks like the above require statistics is **variability**. Thus, if the hardness of all preparations of the same cement mixture were the same, the task of comparing the hardness of two cement mixtures of study example 2 would not require statistics; it would suffice to compare the hardness of one preparation from each of the two cement mixtures. However, the hardness of different preparations of the same cement mixture (even when done by the same preparer) will differ, complicating thus the comparison problem. An appreciation of the complications caused by variability begins by realizing that the problem of study example 2, as stated, is ambiguous. Indeed, if the hardness of preparations of the same cement mixture differ, then what does it mean to compare the hardness of different cement mixtures? A more precise statement of the problem would be to compare the *average* or *mean* hardness of the different cement mixtures. (A similar comment applies to the other aforementioned study examples (with the exception of study example 4 which deals with prediction). For example, the estimation problem in study example 1 is stated more precisely by referring to the average thermal expansion.) Moreover, the familiar words of average and mean have a technical meaning in statistics, and full understanding of it requires a clear distinction between the concepts of population and sample. These concepts are discussed in the next section.

In summary, statistics is an indispensable tool for dealing with the variability which is inherent in most engineering and scientific studies.

1.2 Populations and Samples

As the examples of engineering studies mentioned in Example 1.1.1 indicate, the use of statistics is required whenever the study involves the investigation of a certain **characteristic** or characteristics of members (objects or subjects) in a certain **population** or populations. The members of a population will also be called **population units**.

Example 1.2.1. a) In study example 1 the characteristic under investigation is the average thermal expansion of a metal in the population of all specimens of the particular metal.
b) In the study example 2 we have two or more populations, one for each type of cement mixture, and the characteristic under investigation is hardness. Population units are the cement preparations.
c) In study example 8 we have two characteristics, salary increase and productivity, for each subject in the population of employees of a large corporation.

As is clear from these examples, the characteristic of interest will differ among members of the same population. This is called the **inherent** or **intrinsic** variability of a population. Because of the intrinsic variability, full understanding of a particular characteristic in a population requires examination of all members of the population (**census**). For example, full understanding the relation between salary and productivity, as it applies to the population of employees of a large corporation (study example 8), requires obtaining information on these two characteristics for all employees of the particular large corporation. However, census is not conducted typically due to cost and time considerations.

Example 1.2.2. a) Cost and time considerations make it impractical to conduct a census of all US citizens of age 18 and over, in order to determine the proportion of these citizens who are in favor of expanding solar energy sources.
b) Cost and time considerations make it impractical to analyze all the water in a lake in order to determine the lake's content of lead.

Moreover, census is often not feasible because the population is *hypothetical* or *conceptual*, in the sense that not all members of the population are available for examination.

Example 1.2.3. a) If the objective is to study the quality of a product (as is the case in study example 3), the relevant population does not consist only of the available supply of this product but, also that which will be produced in the future. Thus, the relevant population is hypothetical.
b) In a study aimed at reducing the weekly rate of accidents (example 6) the relevant population does not consist only of the one-week time periods on which records have been kept, but also of future one-week periods. Thus, the relevant population is hypothetical.

In studies where it is either impractical of infeasible to conduct census (which is the vast majority of cases), answers to questions regarding a characteristic under investigation are obtained by **sampling** the population. Sampling refers to the process of selecting a number of population units and recording their characteristic(s). For example, determination of the proportion of US citizens of age 18 and over who are in favor of expanding solar energy sources, is based on a sample of such citizens. Similarly, the determination of whether or not the content of lead in the water of a certain lake is within the safety limit must be based on water samples. The good news is that, if the sample is suitably drawn from the population, then the **sample properties** of the characteristic of interest resemble (though they are not identical to) the **population properties**.

Example 1.2.4. a) A **sample proportion** (i.e., the proportion in a chosen sample) of US citizens who favor an expansion of solar energy, approximates (though, in general, will differ from) the **population proportion**.
b) The average concentration of lead in water samples (**sample average**), approximates (but is, in general, different from) the average concentration in the entire lake (**population average**).
c) The relation between salary and productivity that a sample of employees suggests, approximates (but is, in general, different from) the relation in the entire population of employees of a large corporation.

Samples properties of the characteristic of interest also differ from sample to sample. This is a consequence of the intrinsic variability of the population from which they are drawn. For example, the number of US citizens, in a sample of size 20, who favor expanding solar energy, will (most likely) be different from the corresponding number in a different sample of 20 US citizens. (See also the examples in Section 1.6.2.) The term **sampling variability** is used to describe such differences in the characteristic of interest from sample to sample.

It is the role of statistics to account for this sampling variability in assessing the accuracy with which sample properties of the characteristic of interest approximate corresponding population properties.

1.2.1 Exercises

1. A car manufacturer wants assess customer satisfaction for cars sold during the previous year.

   (a) Describe the population involved.

(b) Is the population involved hypothetical or not?

2. An automobile assembly line is manned by two shifts a day. The first shift amounts for two thirds of the overall production. Quality control engineers want to compare the average number of non-conformances per car in each of the two shifts.

   (a) Describe the population(s) involved.

   (b) Is the population(s) involved hypothetical or not?

3. A consumer magazine article titled ”How Safe is the Air in Airplanes” reports that the air quality, as quantified by the degree of staleness, was measured on 175 domestic flights.

   (a) Identify the population of interest, and the population units.

   (b) Identify the sample.

4. In an effort to determine the didactic benefits of computer activities, when used as an integral part of teaching Stat 401, one section is taught using the traditional method, while another is taught with computer activities. In the end of the semester, each student's score on the same test is recorded. To eliminate unnecessary variability, both sections were taught by the same professor.

   (a) Is there one or two populations involved in the above study?

   (b) Describe the population(s) involved.

   (c) Is (are) the population(s) involved hypothetical or not?

   (d) What is (are) the sample(s) in this study?

1.3 Statistical Inference

A sample can be thought of as a window which provides a glimpse into the population. However, as the previous section emphasized, due to the ubiquitous intrinsic variability, a sample cannot yield precise information regarding the population of interest. Statistical inference is the methodology used for dealing with the uncertainty issues, which arise in the process of extrapolating to the population the information contained in the sample. It provides a set of tools which help decision makers choose actions in the absence of precise knowledge about the population. For example, city officials might want to know whether a new industrial plant is pushing the average air pollution beyond the acceptable limits. Air samples are taken and the air pollution is measured in each sample. The sample average, or sample mean, of the air pollution measurements must then be used to decide if the overall, i.e. the population, average air pollution is above a critical limit that would justify taking corrective action.

As we will see in later chapters, statistical inference, mainly takes the form of **estimation** (both **point** and, the more useful, **interval** estimation) of certain **population parameters**, and of **testing** various **hypotheses** regarding the value of certain population parameters. For example, estimation would be used in the task of estimating the coefficient of thermal expansion of a metal (Example 1.1.1, part 1), while the task of testing a manufacturer's claim regarding the quality of his/her product (Example 1.1.1, part 7) involves hypothesis testing. Finally, the principles of statistical inference are also used in the problem of prediction, which arises, for example, in situations where we would like to predict the failure time on the basis of the stress applied (Example 1.1.1, part 4).

Remark 1.3.1. The term "population parameter" refers to a population property which is quantified, such as the population mean and population proportion that were referred to in Example 1.2.4 (see also Section 1.6). As another example, the relation between employee's productivity and salary increase, is a population property that we have not quantified yet, though, in later chapters, parameters that offer different quantifications of a relation between two variables will be introduced.

As mentioned in the end of Section 1.2, it is the role of statistics to assess the accuracy with which sample properties of the characteristic of interest approximate corresponding population properties. Even with such assessment, there is the risk of making a wrong decision. For example, there is a risk that city officials might decide that the average air pollution exceeds the acceptable limit, when in fact it does not, or, conversely, that the average air pollution does not exceed the acceptable limit, when in fact it does. Thus, statistical inference is also concerned with the appraisal of the risks, and consequences, to which one might be exposed by making generalizations beyond the sample, especially when such generalizations lead to decisions and actions. This includes appraisal of a) the probability of making wrong decision, b) the possibility of obtaining estimates that do not lie within permissible limits, and c) the chances of making incorrect predictions.

1.4 Some Sampling Concepts

1.4.1 Representative Samples

It should be intuitively clear that statistical inference (i.e. the extrapolation of sample information to the population) can only be valid if the sample is **representative** of the population. For example, extrapolation to the population of US citizens, of voting age, the information of a sample which consists of those who work in the oil industry, will unavoidably lead to wrong conclusions about the prevailing public opinion regarding the use of solar energy.

A famous (or infamous) example, that demonstrates what can go wrong when a non-representative sample is used, is the Literary Digest poll of 1936. The magazine Literary Digest had been extremely successful in predicting the results in US presidential elections, but in 1936 it predicted a 3-to-2 victory for Republican Alf Landon over the Democratic incumbent Franklin Delano Roosevelt. The blunder was due to the use of a non-representative sample. See subsection 1.4.4 for further discussion. It is worth mentioning that the prediction of the Literary Digest magazine was wrong even though it was based on 2.3 million responses (out of 10 million questionnaires sent). On the other hand Gallup correctly predicted the outcome of that election by surveying only 50,000 people.

The notion of representativeness of a sample, though intuitive, is hard to pin down. This is because there is no way to tell just by looking at a sample whether or not it is representative. Thus we adopt an indirect definition and say that a sample is representative if it allows for valid statistical inference.

1.4.2 Simple Random Sampling, and Stratified Sampling

The only assurance that the sample will be representative comes from the method used to select the sample. The most straightforward method for obtaining representative samples is called **simple random sampling**. A sample of size n, selected from some population, is a simple random sample if the selection process ensures that every sample of size n has an equal chance of being selected. In particular, every member of the population has the same chance of being included in the sample.

A common way to select a simple random sample, of size n, from a finite population consisting of N units, is to number the population units from $1, \ldots, N$, use a **random**

number generator to randomly select n of these numbers, and form the sample from the units that correspond to the n selected numbers. A random number generator for selecting a simple random sample, simulates the process of writing each number from $1, \ldots, N$ on slips of paper, putting the slips in a box, mixing them thoroughly, selecting one slip at random and recording the number on the slip. The process is repeated (without replacing the selected slips in the box) until n distinct numbers from $1, \ldots, N$ have been selected.

Example 1.4.1. Sixty KitchenAid professional grade mixers are manufactured per day. Prior to shipping, a simple random sample of 12 must be selected from each day's production, and carefully rechecked for possible defects. Describe a software-based procedure for obtaining a simple random sample of 12 mixers from a day's production of 60 mixers.

Solution. As a first step to selecting a random sample of 12, we number the mixers from 1 to 60. We will use the random number generator of the statistical software Minitab to draw a simple random sample of 12 from the numbers 1-60. Click and enter, in succession,

> **Calc > Make Patterned Data > Simple Set of Numbers**
>
> **Enter C1 in "Store patterned data in:", Enter 1 in "From first value:",**
>
> **Enter 60 in "To last value:", Enter 1 in "In steps of:",**
>
> **Enter 1 in "List each value BLANK times";, Enter 1**
>
> **in "List the whole sequence BLANK times"; OK.**

This will generate the numbers 1-60 in column C1 of the Minitab *Worksheet*. Then click and enter, in succession,

> **Calc > Random Data > Sample from Columns > Enter 12 in "Sample**
>
> **BLANK from column(s):" Enter C1, Enter C2 in "Store samples in:",**
>
> **DO NOT select the option "Sample with replacement", OK**

These steps will generate 12 numbers, which can be thought of as having being selected by random drawings of a slip from a box containing 60 slips numbered from 1 to 60, as described in the paragraph preceding this example. A set of 12 numbers thus obtained is

```
6 8 57 53 31 35 2 4 16 7 49 41
```

This set of random numbers specifies the desired sample of size $n = 12$ mixers from a day's production of 60.

Clearly, the above technique cannot not be used with hypothetical/infinite populations. However, measurements taken according to a set of well defined instructions can assure that the essential properties of simple random sampling hold. For example, in comparing the hardness of cement mixtures, guidelines can be established for the mixture preparations and the measurement process to assure that the sample of measurements taken is representative.

As already mentioned, simple random sampling guarantees that every population unit has the same chance of being included in the sample. However, the mere fact that every population unit has the same chance of being included in the sample, does not guarantee that the sampling process is simple random.

Example 1.4.2. Suppose we want to select a representative sample of 10 from a group of 100 undergraduate students consisting of 50 male and 50 female students. We do that by assigning numbers 1-50 to the male students and use a random number generator to select 5 of them, and then repeat the same for the female students. This way, every student has the same chance (1 out of 10) of being selected. However, this sampling is not simple random because it excludes, for example, samples consisting of 4 male and 6 female students. Thus, the condition for simple random sampling, namely that each sample of size 10 have equal chance of being selected, is violated.

This sample selection method of Example 1.4.2 is an example of what is called *stratified* sampling. Stratified sampling can be used whenever the population of interest consists of well defined subgroups, or sub-populations, which are called **strata**. Essentially, a stratified sample consists of simple random samples from each of the strata. Depending on the variable of interest, examples of strata are the sub-populations defined by ethnic groups, types of cars, age of equipment, different labs where water samples are sent for analysis, and so forth. Stratified samples are also representative, i.e., they allow for valid statistical inference. In fact, if the members of each stratum tend to be more homogenous (i.e. similar) with respect to the variable of interest, than members belonging in different strata, then stratified sampling is preferable to simple random sampling, as it can provide more accurate information regarding the entire population.

1.4.3 Sampling With and Without Replacement

In sampling from a finite population, one can choose to do the sampling *with replacement* or *without replacement.* Sampling with replacement means that after a unit is selected and its characteristic is measured and included in the sample, it is replaced back into the population and may therefore be selected again. Tossing a fair coin can be thought of as sampling with replacement from the population {Heads, Tails}. In sampling without replacement, each unit can be included only once in the sample. Thus, simple random sampling is sampling without replacement.

Sampling with replacement is somewhat simpler conceptually, because each selected unit is drawn from the same (the original) population of N units. On the other hand, it should be intuitively clear that allowing the possibility of including a population unit more than once in the sample will not enhance the representativeness of the sample. However, when the population size is very large, sampling with and without replacement are essentially equivalent, and, in such cases, we can pretend that a simple random sample is obtained though sampling with replacement (the likelihood of a unit being included twice is negligible). Finally, sampling with replacement is used in the context of the **bootstrap**, a very common and useful approach to statistical inference.

1.4.4 Non-representative Sampling

Non-representative samples arise whenever the sampling plan is such that a part, or parts, of the population of interest are either excluded from, or systematically under-represented in, the sample.

Typical non-representative samples are the so-called *self-selected* and *convenience* samples. As an example of a self-selected sample, consider a magazine which surveys its readers by using information from cards that were returned to make statements like '40% of readers have purchased cell phones with digital camera capabilities'. In this case, readers who like to update and try new technology are more likely to respond indicating their purchases. Thus, the proportion of purchasers of cell phones with digital camera capabilities in the sample of returned cards will likely be much higher than it is amongst all readers. As an example of a convenience sample, consider using the students in your statistics class as a sample of students at your university. However, not all majors require a statistics course and most students take statistics in their sophomore or junior year.

The **selection bias**, i.e. the systematic exclusion of some part of the population of interest, is inherent in self-selected and convenience samples. Selection bias, which the method of simple random sampling avoids, is the typical cause of non-representative samples.

For its 1936 poll, the Literary Digest used a sample of 10 million to people selected mainly from magazine subscribers, car owners, and telephone directories. In 1936, those who owned telephones or cars, or subscribed to magazines, were more likely to be wealthy individuals who were not happy with the Democratic incumbent. Thus, it was convenience sampling that excluded part of the population. Moveover, only 2.3 million responses were returned from the 10 million questionnaires that were sent. Obviously, those who felt strongly about the election, and that included a majority of those who wanted change, were more likely to respond. Thus, the Literary Digest sample was self-selected, in addition to being a sample of convenience. [The Literary Digest went bankrupt, while Gallup survived to make another blunder another day (in the 1948 Dewey-Truman contest).]

Other sampling methods which, in some situations, can be less costly or easier to implement or more informative than simple random sampling, do exist. Stratified sampling, an example of which we saw above, is one of them. But in this book we will mainly assume that the samples are simple random samples, with occasional passing reference to stratified sampling.

1.4.5 Exercises

1. The person designing the study of Exercise 1.2.1-4, aimed at determining the didactic benefits of computer activities, can make one of the two choices: (i) make sure that the students know which of the two sections will be taught with computer activities, so they can make an informed choice, or (ii) not make available any information regarding the teaching method of the two sections. Which of these two choices provide a closer approximation to simple random sampling?

2. A type of universal remote for home theater systems is manufactured in three distinct locations. 20% of the remotes are manufactured in location A, 50% in location B, and 30% in location C. The quality control team (QCT) wants to inspect a simple random sample (srs) of 100 remotes to see if a recently reported problem with the menu feature has been corrected. The QCT requests that each location send to the QC Inspection Facility a srs of remotes from their recent production as follows: 20 from location A, 50 from B and 30 from C.

   (a) Does the sampling scheme described above produce a simple random sample of size 100 from the recent production of remotes?

(b) Justify your answer in part a). If your answer is no, then what kind of sampling is it?

3. A civil engineering student, working on his thesis, plans a survey to determine the proportion of all current drivers that regularly wear seat belt. He decides to interview a his classmates in the three classes he is currently enrolled.

   (a) What is the population of interest?

   (b) Do the student's classmates constitute a simple random sample from the population of interest?

   (c) What name have we given to the sample that the student collected?

   (d) Do you think that this sample proportion is likely to overestimate, or underestimate the true proportion of all drivers who regularly wear seatbelt?

4. A car manufacturer wants information about customer satisfaction for cars sold during the previous year. The particular manufacturer makes three different types of cars. Describe and discuss two different random sampling methods that might be employed.

5. A particular product is manufactured in two facilities, A and B. Facility B is more modern and accounts for 70% of the total production. A quality control engineer wishes to obtain a simple random sample of 50 from the entire production during the past hour. A coin is flipped and each time the flip results in heads, he/she selects an item at random from those produced in facility A, and each time the flip results in tails, he/she selects an item at random from those produced in facility B. Does this sampling scheme result in simple random sampling? Explain your answer.

6. An automobile assembly line is manned by two shifts a day. The first shift amounts for two thirds of the overall production. The task of quality control engineers is to monitor the number of non-conformances per car. Each day a simple random sample of six cars from the first shift, and a simple random sample of three cars from the second shift is taken, and the number of non-conformances per car in recorded. Does the sampling scheme described above produce a simple random sample of size nine from the day's production? Justify your answer.

1.5 Random Variables and Statistical Populations

The characteristics of interest in all study examples given in Section 1.1 are **quantitative** in the sense that they can be measured and thus can be expressed as numbers. Though quantitative characteristics are more common, **categorical** or **qualitative** characteristics also arise. Two examples of qualitative characteristics are gender, and type of car. Since statistical procedures are applied on numerical data sets, numbers are assigned for expressing qualitative characteristics. For example, -1 can be used to denote that a subject is male, and $+1$ to denote that it is female.

A quantitative or qualitative characteristic expressed as a number is called a **variable**. Variables can be **univariate**, **bivariate** or **multivariate** depending on whether one or two or more characteristics are measured, or recorded, on each population unit.

Example 1.5.1. a) In a study aimed at determining the relation between productivity and salary increase, two characteristics are recorded on each population unit (productivity and salary increase), resulting in a bivariate variable.
b) Consider the study which surveys US citizens aged 18 and over regarding their opinion on solar energy. If an additional objective of the study is to determine how this opinion varies among different age groups, then the age of each individual in the sample is also recorded, resulting in a bivariate variable. If, in addition, the study aims to investigate how this opinion varies between genders, then the gender of each individual in the sample is also recorded, resulting in a multivariate variable.
c) Consider the environmental study which measures the content of lead in water samples from a lake, in order to determine if the concentration of lead exceeds the safe limits. If other contaminants are also of concern, then the content of these other contaminants is also measured in each water sample, resulting in a multivariate variable.

Due to the intrinsic variability, the value of the (possibly multivariate) variable varies among population units. It follows that when a population unit is randomly sampled from a population, its value is not known a-priori. The value of the variable of a population unit that will be randomly sampled will be denoted by a capital letter, such as X. The fact that X is not known a-priori, justifies the terminology **random variable** for X.

A random variable, X, denotes the value of the variable of a population unit that will be sampled.

The population from which a random variable was drawn will be called the **underlying population** of the random variable. Such a terminology is particularly helpful in studies involving several populations, as are all studies that compare the performance of two or more methods or products; see, for example, the study example 2 of Example 1.1.1.

Finally, we need a terminology for the entire collection of values that the variable under investigation takes among the units in the population. Stated in different terms, suppose that each unit in the population is labeled by the value of the variable under investigation, and the values in all labels are collected. This collection of values is called the **statistical population**. Note that, if two (or more) population units have the same value of the variable, then this value appears two (or more) times in the statistical population.

Example 1.5.2. Consider the study that surveys US citizens, of age 18 and over, regarding their opinion on solar energy. Suppose that the opinion is rated on a scale from $0, 1, \ldots, 10$, and imagine each member of the population labeled by the value of their opinion. The statistical population contains as many 0's as there are people with opinion rated 0, as many 1's as there are people whose opinion is rated 1, and so forth.

The word 'population' will be used to refer either to the population of units or to the statistical population. The context, or an explanation, will make clear which is the case.

1.5.1 Exercises

1. Consider the following examples of populations, together with the variable/characteristic measured on each population unit.

   (a) All undergraduate students currently enrolled at Penn State. Variable: major type.

   (b) All campus restaurants. Variable: seating capacity.

   (c) All books in Penn State libraries. Variable: frequency of check-out.

   (d) All steel cylinders manufactured in a given month. Variable: diameter of cylinder.

   For each of the above examples:

   (1) Describe the statistical population.

   (2) Is the variable of interest is quantitative or qualitative.

(3) Specify another variable that could be measured on the population units.

2. In a population of 500 tin plates, the number of plates with 0, 1 and 2 scratches is $N_0 = 190$, $N_1 = 160$ and $N_2 = 150$.

   (a) Identify the variable of interest and the statistical population.

   (b) Is the variable of interest quantitative or qualitative?

   (c) Is the variable of interest univariate or multivariate?

3. At the final assembly point of BMW cars in Graz, Austria, quality control inspectors record the number of non-conformances in the engine and transmission that arrive from Germany and France, respectively.

   (a) Is the variable of interest quantitative or qualitative?

   (b) Describe the statistical population.

4. In Exercise 1.2.1-3 a consumer magazine article reports that the air quality, as quantified by the degree of staleness, was measured on 175 domestic flights.

   (a) Identify the variable of interest and the statistical population.

   (b) Is the variable of interest quantitative or qualitative?

   (c) Is the variable of interest univariate or multivariate?

5. A car manufacturing company, which makes three different types of cars, wants information about customer satisfaction for cars sold during the previous year. Each customer is asked for the type of car he/she bought last year, and to rate his/her level of satisfaction on a scale from 1-6.

   (a) Identify the variable recorded and the statistical population.

   (b) Is the variable of interest univariate?

   (c) Is the variable of interest quantitative or qualitative?

1.6 Proportions, Averages and Variances

Having labeled each member of the population by that member's value of the variable of interest, scientists are typically interested in learning about certain quantifiable aspects, or parameters, of the resulting (statistical) population. The most common parameters are the proportion, the average and the variance. In this section we discuss the population version of these parameters for finite populations. Sample versions of these parameters will also be discussed.

1.6.1 Population Proportion and Sample Proportion

When the variable of interest is categorical, such as *Male* or *Female*, *Defective* or *Non Defective*, *Strength of opinion*, *Type of Car* etc, then interest lies in the proportion of population units in each of the categories. If the population has N units, and N_i units are in category i, then the **population proportion** of category i, is

$$p_i = \frac{\#\{\text{population units of category } i\}}{\#\{\text{population units}\}} = \frac{N_i}{N}.$$

If a sample of size n is taken from this population, and n_i sample units are in category i, then the **sample proportion** of category i, is

$$\widehat{p}_i = \frac{\#\{\text{sample units of category } i\}}{\#\{\text{sample units}\}} = \frac{n_i}{n}.$$

Example 1.6.1. A car manufacturer receives a shipment of $N = 10,000$ navigation systems that are to be installed as a standard features in the next line of luxury cars. Of concern is a type of satellite reception malfunction. If $N_1 = 100$ systems have this malfunction (category 1) and $N_2 = 9,900$ do not (category 2), then the population proportions for the two categories are

$$p_1 = \frac{100}{10,000} = 0.01, \quad p_2 = \frac{9,900}{10,000} = 0.99.$$

If a sample of $n = 1,000$ is found to have $n_1 = 8$ systems with the malfunction and $n_2 = 992$ without the malfunction, then the sample proportions for the two categories are

$$\widehat{p}_1 = \frac{8}{1,000} = 0.008, \quad \widehat{p}_2 = \frac{992}{1,000} = 0.992.$$

As already suggested in Example 1.2.4, sample properties of the variable of interest, approximate (though, in general, will not be identical to) corresponding population properties. In particular,

The sample proportion $\widehat{\mathrm{p}}$ approximates, but is, in general, different from the population proportion p.

The following example, will use Minitab to obtain the sample proportions of five samples of size 1,000 from the population of 10,000 navigation systems of Example 1.6.3 as a further demonstration of the quality of the approximation of p by $\widehat{p}$, and as a demonstration of the sampling variability of $\widehat{p}$.

Example 1.6.2. Consider the population of 10,000 navigation systems of Example 1.6.1. Assign the value zero to each of the 9,900 systems with no reception malfunction and the number 1 to each of the 100 systems with reception malfunction. Thus the corresponding statistical population consists of 9,900 0's and 100 1's. To use Minitab to obtain a random sample of 1,000 from the (statistical) population, we first enter the 9,900 0's and the 100 1's in column C1, using the commands:

Calc > Make Patterned Data > Simple Set of Numbers >

Enter C1 in ”Store patterned data in:”, Enter 0 in ”From first value:”,

Enter 0 in ”To last value:”, Enter 1 in ”In steps of:”,

Enter 9900 in ”List each value BLANK times”, OK

With a similar set of commands, enter 1 in the first 100 rows of column C2. Then join the columns C1 and C2 to represent the statistical population, using the commands:

Data > Stack > Columns > Enter C1 C2 in ”Stack the following columns:”, Check ”Column of current worksheet” and enter C3, OK

Having the statistical population we can draw a random sample of size 1000 using the commands:

Calc > Random Data > Sample from Columns > Enter 1000 in ”Sample BLANK from column(s):” Enter C3, Enter C4 in ”Store samples in:”, OK

To calculate the proportion of 1's (systems with reception malfunction) in the sample, use the commands:

Calc > Column Statistics > Select ”Mean”, Enter C4 in ”Input Variable”, OK

Repeating the last two sequences of commands five times gives the following sample proportions:

```
0.013, 0.012, 0.008, 0.0014, 0.01.
```

1.6.2 Population Average and Sample Average

Consider a population consisting of N units, and let $v_1, v_2, \ldots, v_N$ denote the values in the statistical population corresponding to some variable. Then the **population average** or **population mean**, denoted by μ, is simply the arithmetic average of all numerical values in the statistical population. That is,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} v_i.$$

If random variable X denotes the value of the variable of a randomly selected population unit, then a synonymous terminology for the population mean is **expected value** of X, or **mean value** of X, and is denoted by μ_X or $E(X)$.

Example 1.6.3. Suppose that a company has $N = 10,000$ employees, and a study on employee productivity is initiated. Suppose further that the productivity of each employee is rated on a scale from 1 - 5. Thus, the statistical population consists of the productivity ratings of the 10,000 employees, which we denote by $v_1, v_2, \ldots, v_{10{,}000}$. Finally, suppose that

$$\begin{aligned}
v_i &= 1, \quad i = 1, \ldots, 300, \\
v_i &= 2, \quad i = 301, \ldots, 1,000, \\
v_i &= 3, \quad i = 1,001, \ldots, 5,000, \\
v_i &= 4, \quad i = 5,001, \ldots, 9,000, \\
v_i &= 5, \quad i = 9,001, \ldots, 10,000.
\end{aligned}$$

Thus, 300 of the employees are rated 1, 700 are rated 2, 4000 are rated 3, 4000 are rated 4 and 1000 are rated 5, and the population values have been listed in increasing (non-decreasing) order. Then, the population average rating is

$$\begin{aligned}
\mu &= \frac{1}{10,000}\sum_{i=1}^{10{,}000} v_i = \frac{1}{10,000}(1 \times 300 + 2 \times 700 \\
&\quad +3 \times 4,000 + 4 \times 4,000 + 5 \times 1,000) = 3.47.
\end{aligned}$$

The following Example 1.6.4 demonstrates the simple, but very important fact, that *proportions can be expressed as means*, or, in other words

$$\boxed{\textbf{a proportion is a special case of mean.}}$$

Example 1.6.4. Consider a study aimed at determining the proportion of US citizens of voting age in a certain county, who are in favor of expanding solar energy. The characteristic of interest here is qualitative (yes or no), but we can convert it to a variable by setting 0 for ”no” and 1 for ”yes”. Suppose that the particular county has $60,000$ US citizens of voting age, and that $36,000$ of them are in favor of expanding solar energy. The statistical population here consists of $24,000$ 0’s and $36,000$ 1’s. Thus, denoting the values of the statistical population by $v_1, v_2, \ldots, v_{60,000}$, and after possible re-labeling so the values are in non-decreasing order, the values v_i are

$$\begin{aligned} v_i &= 0, \quad i = 1, \ldots, 24,000, \\ v_i &= 1, \quad i = 24,001, \ldots, 60,000. \end{aligned}$$

Thus, the population mean value, which is the same as the population proportion of those in favor of expanding solar energy, is

$$\mu = p = \frac{1}{60,000} \sum_{i=1}^{60,000} v_i = \frac{36,000}{60,000} = 0.6.$$

As described in Section 1.2, samples offer a useful way of surveying a population whenever census is too costly and/or time consuming. Since this is the case in the vast majority of real life applications, we basically rely on samples in order to determine the population properties of the characteristic (variable) of interest.

If a sample of size n is randomly selected from the population, and if $x_1, x_2, \ldots, x_n$ denote the variable values corresponding to the sample units (note that a different symbol is used to denote the sample values), then the **sample average** or **sample mean** is simply

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Example 1.2.4 already highlights the fact that sample properties of the variable of interest, approximate (though, in general, will not be identical to) corresponding population properties. In particular,

The sample mean $\overline{\mathrm{x}}$ approximates, but is,
in general, different from the population mean $\boldsymbol{\mu}$.

Moreover, due to sampling variability, the value of the sample mean differs from sample to sample, where as there is only one population mean. The distinction between a sample mean and the population mean is important as it is simple, but often a cause of confusion to students studying statistical inference.

Example 1.6.5. Consider the 10,000 employees of Example 1.6.3, suppose that a sample of 10 employees is randomly selected, and that their productivity ratings are recorded. Suppose also that, after possible re-labeling so the values are in non-decreasing order, the sample values are

$$x_1 = 2, \quad x_2 = x_3 = x_4 = 3,$$
$$x_5 = x_6 = x_7 = x_8 = 4, \quad x_9 = x_{10} = 5.$$

Then the sample average rating is

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 3.7,$$

which is different from the value of the population mean, 3.47, that we saw in Example 1.6.3. Moreover, it should be clear that, if another sample of 10 employees were to be drawn at random (either by the same or a different investigator) then, more likely than not, the resulting sample mean would be different from 3.7.

Example 1.6.6. Consider the $60,000$ US citizens of voting age of Example 1.6.4, suppose that a sample of 50 such citizens is randomly selected, and that their opinion is recorded. Suppose also that, after possible re-labeling the values in increasing order, the sample values are

$$\begin{aligned} x_i &= 0, \quad i = 1, \ldots, 22, \\ x_i &= 1, \quad i = 23, \ldots, 50. \end{aligned}$$

Then, the sample average, which, in this case, is the proportion of those in the sample in favor of solar energy, is

$$\overline{x} = \widehat{p} = \frac{1}{50}\sum_{i=1}^{50} x_i = \frac{28}{50} = 0.56.$$

which is different from the population proportion of 0.6. Moreover, it should be clear that, if another sample of 50 citizens were to be drawn at random then, more likely than not, the resulting sample proportion would be different from 0.56.

The above exposition assumes a finite population. The definition of population mean or average for an infinite or conceptual population, such as that of the cement mixtures

Example 1.1.1, part 2, will be given in a later section. The definition of sample mean remains the same regardless of whether or not the sample has been drawn from a finite or an infinite population.

The above exposition also assumes univariate measurement on each population unit. The population mean and the sample mean for bivariate or multivariate variables is given by averaging each coordinate separately.

1.6.3 Population Variance and Sample Variance

The population variance, and *standard deviation*, offer a quantification of the intrinsic variability of the population. Quantifications of the intrinsic variability are of interest, as a quality measure in manufacture. Indeed, if the product characteristics of should vary as little as possible from one product to the other. For example, while high average gas mileage is a desirable characteristic of a car, it is also desirable the achieved gas mileage of different cars of the same type is approximately the same.

Consider a population consisting of N units, and let $v_1, v_2, \ldots, v_N$ denote the values in the statistical population corresponding to some variable. Then the **population variance**, denoted by σ^2, is defined as

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \mu)^2, \tag{1.6.1}$$

where μ is the population mean. If the random variable X denotes the value of the variable of a randomly selected population unit, then the population variance is also called the variance of the random variable X, and we write σ_X^2, or $\mathrm{Var}(X)$.

The variance of a random variable X, or of its underlying population, indicates/quantifies extent to which the values in the statistical population differ from the population mean. As its expression in (1.6.1) indicates, the population variance is the average squared distance of members of the (statistical) population from the population mean. As it is an average square distance, it goes without saying that, the variance of a random variable can never be negative.

Some simple algebra reveals the following alternative expression for the population variance.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} v_i^2 - \mu^2. \tag{1.6.2}$$

Expression (1.6.2) is more convenient for calculating the variance.

The positive square root of the population variance is called the population **standard deviation**:

$$\boxed{\textbf{Standard deviation is the positive square root of the variance: } \boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}^2}.} \tag{1.6.3}$$

If a sample of size n is randomly selected from the population, and if $x_1, x_2, \ldots, x_n$ denote the variable values corresponding to the sample units, then the **sample variance** is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{1.6.4}$$

The positive square root of the sample variance is called the **sample standard deviation**:

$$S = \sqrt{S^2}. \tag{1.6.5}$$

Remark 1.6.1. The definition of sample variance involves the **deviations** of each observation from the sample mean:

$$x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x}.$$

Because these deviations sum to zero, i.e.

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0,$$

there are $n-1$ independent quantities (deviations) that determine S^2. This is also expressed by saying that there are $n-1$ **degrees of freedom**. As it turns out, dividing by $n-1$ results in S^2 having a desirable property as an estimator of the population variance σ^2. This is discussed at a later chapter (see Proposition **??**).

A computational formula for S^2 is

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]. \tag{1.6.6}$$

Example 1.6.7. Consider a study, mentioned in Example 1.6.4, aimed at determining the proportion of US citizens of voting age in a certain county, who are in favor of expanding solar energy. Recall that the qualitative characteristic of interest (yes or no), was converted to a variable by setting 0 for "no" and 1 for "yes", so that the statistical population consisted of 24, 000 0's and 36, 000 1's. Using the same labeling of the values as before, which is

$$\begin{aligned} v_i &= 0, \quad i = 1, \ldots, 24,000, \\ v_i &= 1, \quad i = 24,001, \ldots, 60,000, \end{aligned}$$

and using the computational formula (1.6.2), the population variance is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} v_i^2 - \mu^2 = \frac{36,000}{60,000} - (0.6)^2 = 0.6(1-0.6) = 0.24.$$

Example 1.6.8. Consider the sample of size 50 from the the statistical population consisting of 24, 000 0's and 36, 000 1's, given in Example 1.6.8, which is

$$\begin{aligned} x_i &= 0, \quad i = 1, \ldots, 22, \\ x_i &= 1, \quad i = 23, \ldots, 50. \end{aligned}$$

Using the computational formula (1.6.6) for the sample variance, we have

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] = \frac{1}{49}\left[28 - \frac{1}{50}28^2\right] = 0.25.$$

The definition of the population variance and standard deviation given in this section assume a finite population. The more general definition, applicable to any population, will be given in a later chapter. The definition of the sample variance and standard deviation remains the same regardless of whether or not the sample has been drawn from a finite or an infinite population.

1.6.4 Exercises

1. **(a)** Use the information given in Example 1.6.3 to calculate the population variance and standard deviation.

   (b) Use the data given in Example 1.6.5 to calculate the sample variance and standard deviation.

2. Consider the population of 500 tin plates described in Exercise 1.5.1-2.

   (a) Find the population mean (or the expected value of the variable of interest.)

(b) Find the population variance (or the variance of the variable of interest).

3. Consider a s.r. sample of $n = 100$ from the population of 500 tin plates described in Exercise 1.5.1-2. Suppose that the number of plates with 0, 1 and 2 scratches in the sample is $n_0 = 38$, $n_1 = 33$ and $n_2 = 29$.

   (a) Find the sample mean of the variable of interest.

   (b) Find the sample variance of the variable of interest.

4. Consider a statistical population consisting of the N values $v_1, \ldots, v_N$, and let μ_V, σ_V^2, σ_V denote the population mean value, variance and standard deviation.

   (a) Suppose that the v_i values are coded to $w_1, \ldots, w_N$, where $w_i = c_1 + v_i$, where c_1 is a known constant. Show that the mean value, variance and standard deviation of the statistical population $w_1, \ldots, w_N$ are

   $$\mu_W = c_1 + \mu_V, \ \sigma_W^2 = \sigma_V^2, \ \sigma_W = \sigma_V.$$

   (b) Suppose that the v_i values are coded to $w_1, \ldots, w_N$, where $w_i = c_2 v_i$, where c_2 is a known constant. Show that the mean value, variance and standard deviation of the statistical population $w_1, \ldots, w_N$ are

   $$\mu_W = c_2 \mu_V, \ \sigma_W^2 = c_2^2 \sigma_V^2, \ \sigma_W = |c_2| \sigma_V.$$

   (c) Suppose that the v_i values are coded to $w_1, \ldots, w_N$, where $w_i = c_1 + c_2 v_i$, where c_1, c_2 are known constants. Show that the mean value, variance and standard deviation of the statistical population $w_1, \ldots, w_N$ are

   $$\mu_W = c_1 + c_2 \mu_V, \ \sigma_W^2 = c_2^2 \sigma_V^2, \ \sigma_W = |c_2| \sigma_V.$$

5. Consider a sample $x_1, \ldots, x_n$ from some statistical population, and let $\overline{x}$, s_X^2, and s_X denote the sample mean, sample variance and and sample standard deviation.

   (a) Suppose that the x_i values are coded to $y_1, \ldots, y_n$, where $y_i = c_1 + x_i$, where c_1 is a known constant. Show that the sample mean, sample variance and sample standard deviation of $y_1, \ldots, y_n$ are

   $$\overline{y} = c_1 + \overline{x}, \ s_Y^2 = s_X^2, \ s_Y = s_X.$$

   (b) Suppose that the x_i values are coded to $y_1, \ldots, y_n$, where $y_i = c_2 x_i$, where c_2 is a known constant. Show that the sample mean, sample variance and sample standard deviation of $y_1, \ldots, y_n$ are

   $$\overline{y} = c_2 \overline{x}, \ s_Y^2 = c_2^2 s_X^2, \ s_Y = |c_2| s_X.$$

(c) Suppose that the x_i values are coded to $y_1, \ldots, y_n$, where $y_i = c_1 + c_2 x_i$, where c_1, c_2 are known constants. Show that the sample mean, sample variance and sample standard deviation of the statistical population $y_1, \ldots, y_n$ are

$$\overline{y} = c_1 + c_2\overline{x},\ s_Y^2 = c_2^2 s_X^2,\ s_Y = |c_2| s_X.$$

6. Consider the sample $X_1 = 81.30031$, $X_2 = 81.3015$, $X_3 = 81.3006$, $X_4 = 81.3011$, $X_5 = 81.2997$, $X_6 = 81.3005$, $X_7 = 81.3021$. Code the data by subtracting 81.2997 and multiplying by 10,000. Thus the coded data are 4, 18, 9, 14, 0, 8, 24. It is given that the sample variance of the coded data is $S_Y^2 = 68.33$. Find the sample variance of the original data. [**Hint.** Reverse the code to express the x-values in terms of the y-values, and then apply the formula in the above Exercise 5.]

1.7 Statistical Experiments and Observational Studies

When the objective of the investigation is to learn about a particular population of interest, then getting a representative sample suffices, as a first step, for a valid statistical analysis. Often, however, the study is *comparative* in nature, and thus it involves more than one target population.

Example 1.7.1. a) A study whose objective is to compare the strength of cement from three different cement mixtures is a comparative study. Here each cement mixture corresponds to a different population (of strength measurements).
b) A study whose objective is to compare the cleaning effectiveness of two cleaning products on four different types of stains is a comparative study. Here each combination of type of cleaning product and type of stain corresponds to a different population (of cleaning effectiveness measurements).
c) A study whose objective is to compare how five different temperature levels and five different humidity levels affect the yield of a chemical reaction is a comparative study. Here each temperature-humidity combination corresponds to a different population (of yields).

Comparative studies have their own jargon. Thus, the study in Example 1.7.1 a) involves one **factor**, which is *cement mixture*, and this factor enters the study at three **levels**; the **response variable** in this study is *cement hardness*. The study in Example 1.7.1

b) involves two factors, *cleaning product*, and *stain*; the factor cleaning product enters the study at two levels, and the factor stain enters the study at four levels; the response variable is the degree of stain removal. The study in Example 1.7.1 c) involves two factors, *temperature* and *humidity*; both factors enter the study at five levels; the response variable is the yield. Moreover, a factor-level combination (e.g. a combination of a temperature level and and humidity level) is also called a **treatment**; thus different treatments correspond to different populations. In comparative studies with more than one factor, it is of interest to also examine how the different factors **interact** in influencing the response, in addition to comparing the populations.

The definition of the population that corresponds to each treatment in a comparative study, must include the **experimental unit**, i.e. the unit on which measurements are made.

Example 1.7.2. a) In the study that compares the strength of different cement mixtures, experimental units are the cement preparations, which are made under precise guidelines.
b)In the study that compares the cleaning effectiveness of cleaning products on different types of stains, the experimental units are pieces of fabric.
c) In the study that compares the effect of temperature and humidity on the yield of a chemical reaction, experimental units are materials used in the reaction.

In order to avoid comparing apples with oranges, the experimental units assigned to different treatments must be as homogenous as possible. For example, if the age of fabric is a factor which affects the response in Example 1.7.1 b) above, then, unless the ages of the fabrics that are assigned to different treatments are homogenous, the comparison of treatments will be distorted. To guard against such distorting effects of other possible factors or, in technical parlance, to avoid **confounding** with other factors, it is recommended that the allocation of units to treatments be **randomized**.

Example 1.7.3. a) Randomizing the allocation of fabric pieces to the different treatments (i.e. combinations of cleaning product and type of stain) avoids confounding the factors of interest (*cleaning product* and *stain*) with the potentially influential factor *age*.
b) In the study of Example 1.7.1 c), the acidity of the materials used in the reaction might be another factor affecting the yield. Randomizing the allocation of the experimental units, i.e. the materials, to the different treatments, avoids confounding the factors of interest (*temperature* and *humidity*) with the potentially influential factor *acidity*.

A study is called a **statistical experiment** if the investigator controls the allocation of units to treatments or factor-level combination, and this allocation is done in a randomized fashion. The studies mentioned in Example 1.7.1 are statistical experiments if the allocation of units (materials for the cement mixture, fabric pieces, and materials for the chemical reaction) to treatments is randomized.

When the allocation of units to treatments is not controlled, the study is called **observational**. Example 8 of Section 1.1 is an example of an observational study, since salary increases are not assigned in a randomized fashion. Observational studies cannot be used for establishing **causation**. For example, even a strong relation between salary increases and employee productivity does not imply that salary increases cause increased productivity (it could be vice-versa). Similarly a strong relation between spanking and anti-social behavior in children does not imply that spanking causes anti-social behavior (it could be vice-versa). Causality can be established only through experimentation. This is why experimentation plays a key role in industrial production and especially in the area of quality improvement of products, as was vigorously advocated by W. Edwards Deming (1900-1993). Having said that, one should keep in mind that observational studies have yielded several important insights and facts. For example, studies involving the effect of smoking on health are observational, but the link they have established between the two is one of the most important issues of public health.

1.7.1 Exercises

1. An experiment to assess the effect of watering on the life time of a certain type of root system incorporates three watering regimens and will be carried out in three different locations. The researcher proceeds to assign a different watering regimen to each of the three locations. It is known that factors such as temperature and soil conditions, which may differ among the three locations, also affect the life time of the root systems. Comment on whether or not the above allocation of treatments to units (root systems) will avoid confounding with the location factor. Explain your answer and describe a possibly better allocation of treatments to units.

2. In the context of the above Exercise 1.7.1-1 suppose that it is known that a factor influencing the survival of the root systems is whether or not the depth where the root systems grew is less than 4cm. Suppose also that the experimenter does not control the depth.

   (a) Describe a design that would be appropriate for analyzing the effect of the tree watering regimens and possible interactions with the depth factor.

   (b) What is considered a "treatment" in this design?

(c) Comment on an allocation of treatments to units that will avoid confounding with the location factor.

3. Rural roads with little or no evening lighting use reflective paint to mark the lanes on highways. It is suspected that the currently used paint does not maintain its reflectivity over long periods of time. Three new types of reflective paint are now available and a study is initiated to compare all four types of paint. The researcher in charge of the study identifies four locations of the highway and for each location he/she designates four sections of length six feet to serve as the experimental units on which treatments (types of paint) will be applied. The researcher proceeds to randomly assign a type of paint to each of the four locations, and uses the assigned paint type to each of the four segments of that location. It is known that factors such as traffic volumes and road conditions, which differ among the four locations, affect the duration of the reflectivity of the paints. Comment on whether or not the above allocation of treatments to units (road segments) will avoid confounding with the location factor. Explain your answer and describe a possibly better allocation of treatments to units.

1.8 The Role of Probability

Probability is an indispensable tool for statistical inference. The discussion at the end of Section 1.3 already hints at this close connection. However, the central role of probability in statistics is not immediately obvious. This is because, in a probability problem, the properties of the population of interest are assumed known, whereas statistics is concerned with learning the properties of the population from that of the sample.

Example 1.8.1. a) Suppose that we know that 75% of all batteries last more than 1500 hours of operation and want to know what are the chances that in a sample of 50 batteries at least 30 will last more than 1500 hours. Then this is a probability problem.
b) Suppose that out of a sample of 50 batteries only 30 are found to last more than 1500 hours and want to know if this provides enough evidence to conclude that the proportion of all batteries that last more than 1500 hours is less than 75%. This is an example of a statistical inference question.

Thus probability uses properties of the population to infer those of the sample, while statistical inference does the opposite. In spite of this difference, statistical inference itself would not be possible without probability.

It should be emphasized that in answering probability or statistical inference questions, the sample must be representative of the population. In other words, the sample should be simple random, or stratified, or some other well defined sampling scheme. Taking into consideration the particular type of sampling scheme is critical for obtaining correct answers to probability or statistical inference questions. Thus, in Example 1.8.1 b), the answers obtained would depend on whether the sample of batteries was simple random, or stratified according to battery type (Duracell, Energizer, etc). Answers obtained using the wrong assumption about the sampling process can be misleading.

1.9 Approaches to Statistical Inference

The main approaches to statistical inference can be classified into *parametric*, *robust*, *nonparametric* and *Bayesian.*

The parametric approach relies on modeling aspects of the mechanism underlying the generation of the data.

Example 1.9.1. In a study aimed at predicting failure time on the basis of stress applied the mechanism underlying the generation of the data that will be collected includes aspects such as the relation between the average failure time and stress applied, and the *distribution* of the *intrinsic error* of the data (these are a technical term to be clarified in a later chapter). A parametric approach might model the relation between the average failure time and stress with a linear function, and might specify a particular type of distribution for the intrinsic error.

In the parametric approach, models are described in terms of unknown *model parameters.* Hence the name *parametric.* In the above example, the slope and intercept of the linear function which models the relation between failure time and stress are model parameters; specification of the (intrinsic) error distribution typically introduces further model parameters. In the parametric approach, model parameters are assumed to coincide with population parameters, and thus they become to focus of the statistical inference. If the assumed parametric model is a good approximation to the data-generation mechanism, then the parametric inference is not only valid but can be highly efficient. However, if the approximation is not good, the results can be distorted. It has been shown that even small deviations of the data-generation mechanism from the specified model can lead to large biases.

The robust approach is still parametric in flavor, but its main concern is with procedures that guard against aberrant observations such as outliers.

The nonparametric approach is concerned with procedures that are valid under minimal modeling assumptions. Some procedures are both nonparametric and robust, so there is overlap between these two approaches. In spite of their generality, the efficiency of nonparametric procedures are typically very competitive compared to parametric ones that employ correct model assumptions.

The Bayesian approach is quite different from the first three as it relies on modeling prior beliefs/information about aspects of the population. The increase of computational power and efficiency of algorithms have made this approach attractive for dealing with some complex problems in different areas of application.

In this book we will develop, in a systematic way, parametric and nonparametric procedures, with passing reference to robustness issues, for the most common ('bread-and-butter') applications of statistics in the sciences and engineering.

1.10 Computer Activities

1. A distributor has just received a shipment of ninety draining pipes from a major manufacturer of such pipes. The distributor wishes to select a sample of size five to carefully inspect for defects. Describe a method for obtaining a simple random sample of five pipes from the shipment of ninety pipes. Use Minitab to implement the method (see Example 1.4.1).

2. Consider the information on the statistical population of productivity ratings given in Example 1.6.3, and use Minitab commands similar to those of Example 1.6.2 to obtain a simple random sample of size 50. Repeat this for a total of five times, computing for each sample the sample mean and sample variance. Compare the sample means and sample variances of the five samples with the population mean and the population variance obtained in the above Exercise 1.6.4-1.

3. (**Simulating rolls of a die. Checking the estimation error $|\overline{x} - \mu|$.**) In this activity we will select a sample with replacement from the finite population $1, \ldots, 6$, compute the sample mean, and check how well it approximates the true population mean of 3.5. [The above sampling with replacement, can also be thought of as simple random sampling from the hypothetical population of all throws of a die; as we will see later, 3.5 is also the mean of this hypothetical population.]

(a) Enter the numbers $1, \ldots, 6$ in column C1 and sample with replacement 100 times, storing the sampled numbers in C2 (see Minitab commands in Example 1.4.1, but make sure the option "Sample with replacement" is selected).

(b) Use the commands:

Stat > Basic Statistics > Display Descriptive Statistics

Enter C2 in "Variables", Click on Statistics and select "mean". OK, OK.

4. (**Simulating rolls of a die. Checking the estimation error $|\widehat{p} - p|$.**) In this activity we will select a sample with replacement from the finite population $1, \ldots, 6$, and check how well the proportion that each number appears approximates $1/6$.

   (a) Enter the numbers $1, \ldots, 6$ in column C1 and sample with replacement 100 times, storing the sampled numbers in C2 (see Minitab commands in Example 1.4.1, but make sure the option "Sample with replacement" is selected).

   (b) Use the commands:

   Stat > Basic Statistics > Display Descriptive Statistics

   Enter C2 in "Variables" and in "By variables (optional)", Click on Statistics and select "N total" "Percent", "Cumulative percent". OK, OK.

Chapter 2

Introduction to Probability

2.1 Overview

Probability theory arose from the need to quantify the likelihood of occurrence of certain events associated with games of chance. Today, probability theory is applied much more widely as it models a wide variety of **chance phenomena**. By chance phenomena we mean any actions, processes or situations the outcomes of which are random. Thus, kicking a soccer ball, and stock market fluctuations, are both chance phenomena. An expression synonymous to chance phenomena is **probabilistic** or **random experiments**.

The random experiments we will deal with are observational studies and statistical experiments, including the random sampling methods discussed in Chapter 1. To introduce probability concepts, and to demonstrate probability calculations, we will also talk about such probabilistic experiments as picking a card from a deck, or rolling a die.

In this chapter we introduce, at an elementary level, the basic concepts of probability theory, including conditional probability and the notion of independence, and describe common techniques for calculating probabilities. This, and the specific probability models which will be discussed in Chapters 3, and 4 will provide the needed probabilistic background for discussing statistical inference. In this chapter the term 'experiment' will be used in its wider sense, i.e. to indicate a probability experiment.

2.2 Sample Spaces

The purpose of this section is to point out that the random experiments we will consider can be thought of as sampling experiment from some population. In fact, we will see that each random experiment can be conceptualized as a sampling experiment from several populations. A useful way of conceptualizing a random experiment as a sampling experiment involves the concept of *sample space*, and random sampling (but not necessarily simple random sampling) from the sample space. The following simple examples illustrating these points.

Example 2.2.1. a) **(Sampling from a finite or infinite population?)** Consider the probabilistic experiment of selecting at random a card from a deck of cards, so each card has the same chance of being selected. Thus, this is as a simple random sample of size 1 from the finite population of 52 cards. Alternatively, however, this experiment can be thought of as obtaining a simple random sample of size one from the infinite hypothetical population of all selections of one card from a deck of cards. Thus, either a finite or an infinite population can be thought of as the population corresponding to this probabilistic experiment.
b) **(Random but not simple random sampling!)** Consider the probabilistic (also statistical sampling) experiment of selecting, through simple random sampling, one US citizen aged 18 and over and recording his/her opinion regarding solar energy on the scale $0, 1, \ldots, 10$. Here we can think of the population corresponding to this experiment as the population of units, i.e. the population of US citizens aged 18 and over. Alternatively, we can think of this experiment as obtaining a simple random sample of size one from the corresponding *statistical* population which, recall discussion in Section 1.2, consists of as many 0's as there are people with opinion rated 0, as many 1's as there are people with opinion rated 1, and so forth. Finally, we can think of this experiment as obtaining a random (but not simple random!) sample from the population of numbers $\{0, 1, \ldots, 10\}$.
c) **(Simple random sample of size n, or of size 1?)** A distributor receives a new shipment of 20 ipods. He wants to draw a simple random sample of five ipods and thoroughly inspect the click wheel of each of them. This experiment, which involves sampling without replacement five times from the 20 ipods, can also be thought as a simple random sample of size 1 from the population of all groups of five ipods that can be formed from the 20 ipods.

In all of the above examples, one of the populations we considered is the population of all possible outcomes of the experiment. Thus, in Example 2.2.1a) the population is the collection of 52 cards, which is the collection of all possible outcomes of the experiment; in Example 2.2.1b) one of the populations we considered is the set of numbers $\{0, 1, \ldots, 10\}$ which again is the collection of all possible outcomes of the experiment; in Example 2.2.1c) the collection of all possible outcomes of the experiment is all groups of 5 ipods.

Definition 2.2.1. *The set of all possible outcomes of an experiment is called the* **sample space** *of the experiment, and will be denoted by $\mathcal{S}$. When the sample space is thought of as the population which is sampled it is also called the* **sample space population***.*

Thus, in Example 2.2.1b), the sample space is

$$\mathcal{S} = \{0, 1, \ldots, 10\}.$$

When we think of the experiment as a random (but not simple random) sampling from the sample space, $\mathcal{S}$ is the sample space population.

We conclude this section with some more examples of random experiments and sample spaces, all of relevance to engineering.

Example 2.2.2. a) Consider a production line producing a certain type of fuse. Select two fuses for close examination and classify each as non-defective or defective. The sample space of this experiment can be

$$\mathcal{S}_1 = \{NN, ND, DN, DD\}$$

b) Consider the same experiment as in part a), except that now the experimenter records the number of defective fuses as outcome of the experiment. In this case, the sample space is

$$\mathcal{S}_2 = \{0, 1, 2\}$$

Thus, if the outcome is 0, then none of the two examined fuses are defective, if the outcome is 1 then either the first or the second or the second is defective (but not both), and if the outcome is 2 then both fuses are defective.
c) Consider again the production line of part a), but now each fuse that comes off the production line is examined until the second defective fuse is found. If the investigator wishes to record the number of fuses examined, then the sample space is

$$\mathcal{S}_1 = \{2, 3, \ldots\}$$

Alternatively, the investigator may choose to record the number of non-defective fuses he/she will examine prior to finding the second defective fuse. In this case, the sample space is

$$\mathcal{S}_2 = \{0, 1, 2, 3, \ldots\}$$

Note that there is complete (i.e. one-to-one) correspondence between the members of the two sample spaces. For example, outcome 2 of sample space $\mathcal{S}_1$ means that the first two fuses were defective, which is also what outcome 0 of sample $\mathcal{S}_2$ means.

Remark 2.2.1. *The above example demonstrates that different variables (and thus different sample spaces) can be associated with the same experiment. Thus, parts a) and b) of Example 2.2.2 both deal with exactly the same experiment, but the variable in part b) is coarser than the variable in part a).*

2.2.1 Exercises

1. Consider the probabilistic experiment of rolling two dice.

   (a) Write down the sample space of this experiment.

   (b) Conceptualize this experiment as taking a s.r. sample of size one from a finite population. What population is this?

   (c) Conceptualize this experiment as taking a s.r. sample of size two from an infinite population. What population is this?

   (d) Conceptualize this experiment as taking a s.r. sample of size one from an infinite population. What population is this?

2.3 Events and Set Operations

In experiments with many possible outcomes, investigators often classify individual outcomes into distinct categories. This is done for convenience in summarizing and interpreting the results. For example, in the context of the experiment of Example 2.2.1b) of the previous section, the investigator may wish to classify the opinion ratings into low ($L = \{0, 1, 2, 3\}$), medium ($M = \{4, 5, 6\}$) and high ($H = \{7, 8, 9, 10\}$). Such classification can be very convenient in reporting the results especially if the experiment involves a large sample size.

Such collections of individual outcomes, or subsets of the sample space, are called **events**. An event consisting of only one outcome is called a **simple event**. Events can be described either by listing the individual outcomes comprising them, as done in the previous paragraph, or in a descriptive manner. Thus, in the context of the experiment of Example 2.2.2c), an event of interest might be described by the statement 'two of the first four fuses to be examined is defective', which is the event $A = \{2, 3, 4\}$, when the sample space is $\mathcal{S}_1$, i.e. the event that either 2 or 3 or 4 fuses are examined before the second defective fuse is found.

We say that a particular event A has **occurred** if the outcome of the experiment is a member of (i.e. contained in) A. In this parlance, the sample space of an experiment is an event which always occurs when the experiment is performed.

Because events are sets, the usual set operations are relevant for probability theory. A quick review, with corresponding illustrations known as *Venn diagrams*, is given below. Note that the complement of an event A, which is denoted here by A', is also commonly denoted by A^c. Verbally, the union $A_1 \cup \cdots \cup A_k$ is also referred to either as the event where A_1 or A_2 or … or A_k happens (where 'or' is used in its non-exclusive sense), or as the event where *at least one* of $A_1, \ldots, A_k$ happens. The intersection $A_1 \cap \cdots \cap A_k$ is also referred to either as the event where A_1 and A_2 and … and A_k happens, or as the event where *all* of $A_1, \ldots, A_k$ happen. Finally, two events, A, B, are called **disjoint** if they have no outcomes in common, or, if they cannot occur together; in mathematical notation, A, B are disjoint if $A \cap B = \emptyset$, where $\emptyset$ denotes the empty set. The empty event can be

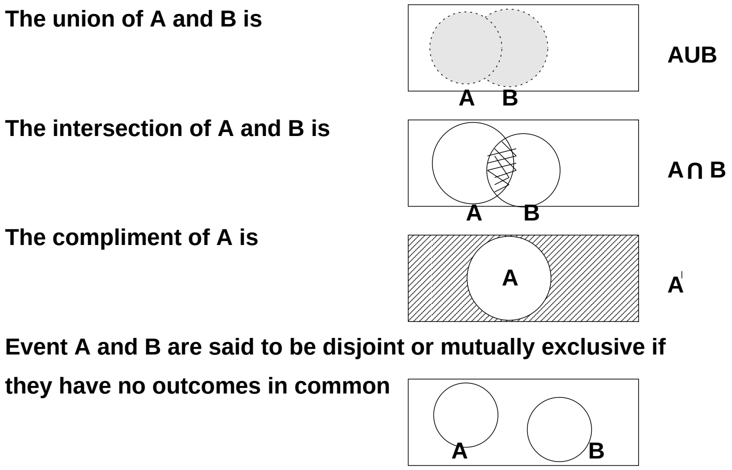

The union of A and B is
A B
AUB
The intersection of A and B is
A B
A∩ B
The compliment of A is
A
A'
Event A and B are said to be disjoint or mutually exclusive if
they have no outcomes in common
A
B

Figure 2.1: Illustration of Basic Set Operations

thought of as the complement of the sample space, $\emptyset = \mathcal{S}'$.

Example 2.3.1. Consider a sampling inspection experiment where 20 items are randomly selected for inspection from a product shipment to determine whether they meet quality standards. Let A_1 denote the event in which no item fails inspection, and A_2 denote the event in which exactly one item fails inspection. Then the event $B_1 = A_1 \cup A_2$, which is also described as the event in which A_1 or A_2 happens, is the event where at most one item fails inspection. Moreover, the event $B_2 = A_1 \cap A_2$, which is also described as the event in which A_1 and A_2 happen, is the empty event, since A_1, A_2 are disjoint.

We finish this section with a brief mention of the basic laws of union and intersection.

Commutative Laws:

$$A \cup B = B \cup A, \; A \cap B = B \cap A.$$

Associative Laws:

$$(A \cup B) \cup C = A \cup (B \cup C), \; (A \cap B) \cap C = A \cap (B \cap C).$$

Distributive Laws:

$$\begin{aligned}(A \cup B) \cap C &= (A \cap C) \cup (B \cap C)\\ (A \cap B) \cup C &= (A \cup C) \cap (B \cup C)\end{aligned}$$

2.3.1 Exercises

1. Consider the experiment that measures the rise time of a reactor (in minutes and fractions of minutes). Let the sample space be the positive real numbers ($\mathcal{S} = \{x|x > 0\}$). Define the events A and B as follows,

   $$A = \{x|x < 72.5\}, \quad \text{and} \quad B = \{x|x > 52.5\}.$$

   Describe each of the events:

   (a) A', (b) $A \cap B$, and (c) $A \cup B$

   both in words and as subsets of $\mathcal{S}$.

2. An US-based engineering firm is considering the possibility of establishing a branch office in Toronto and one in Mexico City. Define events A and B as follows:

   A = {The engineering firm will establish a branch office in Toronto}
   B = {The engineering firm will establish a branch office in Mexico City}

   For each part below draw a Venn diagram depicting events A and B, and shade the event described.

(a) The firm establishes branch office in both cities.

(b) The firm establishes branch office in neither of the cities.

(c) The firm establishes branch office in exactly one of the cities.

3. Disks of polycarbonate plastic from a supplier are analyzed for scratch and shock resistance. The results from 100 disks are summarized below

		shock resistance	
		high	low
scratch	high	70	9
resistance	low	16	5

(a) List the sample space, $\mathcal{S}$, of this experiment.

(b) Give the subset of $\mathcal{S}$ that corresponds to the event that the selected disk has low shock resistance. How many disks have low shock resistance?

(c) Give the subset of $\mathcal{S}$ that corresponds to the event that the selected disk has low shock resistance or low scratch resistance. How many disks have low shock resistance or low shock resistance?

4. In a literature class, 30% of students speak French, 35% speak German and 50% of students speak French OR German. Let F denote the event that a randomly selected student speaks French and let G denote the corresponding event for German. What is the proportion of students who

(a) Make a Venn Diagram showing events F and G and shade the event that a randomly selected student speaks French but not German.
Bonus Question: What percent of students speak only French?

(b) Make a Venn Diagram showing events F and G and shade the event that a randomly selected student speaks French AND German?
Bonus Question: What percent of students speak French and German?

5. Suppose that in State College 40% of the households subscribe to newspaper A, 30% to newspaper B, and 60% of the households subscribe to at least one of the two newspapers. Let A denote the event that a randomly chosen household subscribe to newspaper A, and let B denote the corresponding event for newspaper B.

(a) Make a Venn Diagram showing events A and B and shade the region representing the 60% of the households which subscribe to at least one of the two newspapers.

(b) Make a Venn Diagram showing events A and B, shade the event that a randomly selected household subscribes only to newspaper A (that is, subscribes to A but not to B).

Bonus Question: What percent of households subscribe only to newspaper A?

(c) Make a Venn Diagram showing events A and B, shade the event that a randomly selected household subscribes to both newspapers.

Bonus Question: What percent of households subscribe to both newspapers?

2.4 Probability

2.4.1 Definition and Interpretation of Probability

In any given experiment we might be interested in assessing the 'likelihood' of occurrence of any one of the possible outcomes, or, more generally, the 'likelihood' of occurrence of some event A. This 'likelihood' of occurrence of an event A will be called the *probability* of A and will be denoted by $P(A)$.

As probability is almost a household word, most people have an intuitive understanding of its meaning. For example, when a coin is tossed most people would agree that the probability of heads (H) is 0.5 (which is what might also be referred to as the chances being 50-50, or as happening 50% of the time). But how is a probability of 0.5, or 50-50 chances, to be interpreted in terms of practical implications? For example, does it mean that in $n = 100$ tosses we are assured of 50 H? To this question most would answer no and they would be correct. However, here are some other related questions that might not be so easy to answer. If there are 45 H in 100 tosses can we say that $P(H)$ is not equal to 0.5? What about if there are 40? Or 30? Note that these type of questions fall in the realm of statistical inference, as was briefly introduced in Chapter 1, and thus they will be answered in a later chapter. [1]

Returning to the issue of interpreting probability, if $P(H) = 0.5$ does not guarantee 50 H in 100 tosses, what does $P(H) = 0.5$ really mean? We will adopt a **limiting relative frequency** interpretation of $P(H)$, according to which the *proportion* of H's in a long

[1] To see that these questions are indeed statistical inference questions, note that the experiment of tossing a coin 100 times corresponds to taking a simple random sample of size 100 from the (hypothetical) population of all coin tosses, and from the sample information we want to find out if the proportion of heads in the population of all coin-tosses is indeed 0.5. Thus we are seeking to infer a population characteristic from that of a sample, which is what statistical inference is about.

sequence of coin-tosses will be close to $P(H)$, and in an infinite sequence of coin-tosses the proportion of H's equals $P(H)$. In general, we have the following

Definition 2.4.1. *Let A be an event defined in connection to an experiment, and let $N_n(A)$ denote the number of occurrences of A in n repetitions of the experiment. Then the* **probability** *$P(A)$ of A is defined as*

$$\frac{N_n(A)}{n} \to P(H),$$

as $n \to \infty$. The ratio $N_n(A)/n$ is called the **relative frequency** *of occurrence of the event A; thus the probability of A is defined to be the limiting relative frequency of A.*

Remark 2.4.1. In the above definition the n repetitions of the experiment need to be *independent*, a concept that will be introduced in Section 2.5 below. A definition using terms we have introduced would refer to a simple random sample of size n from the conceptual population of all repetitions of the experiment, and would define $N_n(A)$ as the number of experiments in the sample where A occurs.

2.4.2 Assignment of Probabilities

So far we have discussed the interpretation of probability, but not how probabilities are assigned to events. This depends on the particular experiment. However, for experiments which have a finite number of equally likely outcomes, i.e. those that take the form of simple random sampling from a finite sample space, the assignment of probabilities to events is straightforward and intuitive: If we denote by N the finite number of outcomes of such an experiment, and denote by $N(A)$ the number of outcomes that constitute the event A, then the probability of A is

$$P(A) = \frac{N(A)}{N}. \boxed{\begin{array}{l}\text{Assignment of probabilities in the case of}\\ \text{finite many and equally likely outcomes}\end{array}} \tag{2.4.1}$$

Example 2.4.1. a) Roll a die once. Find the probability of the event that the outcome is either an even number or a 3.

b) Roll two dice separately (or one die twice). Find the probability of the event that the sum of the two sides is seven.

Solution. a) Let $A =$ the outcome is either an even number or a 3. Here $N = 6$, $N(A) = 4$ and thus $P(A) = 4/6$ or two thirds.

b) Let $A = \{\text{sum of two sides=7}\}$. Here $N = 36$ and A consists of (1,6), (2,5), (3,4), (4,3), (5,2) and (6,1). Thus $N(A) = 6$, and $P(A) = \frac{6}{36} = \frac{1}{6}$.

While the method of assigning probabilities to events of experiments with equally likely outcomes is straightforward, the implementation of this method, if N is large and/or the event A is complicated, is not straightforward. For example, to find the probability that five cards, randomly selected from a deck of 52 cards, will form a full house (three of a kind and two of a kind) we need to be able to determine how many 5-card hands are possible and how many of those constitute a full house. Such determination requires specialized counting techniques which will be presented in Section 2.6 below.

2.4.3 Axioms and Properties of Probability

We now outline the axioms which govern any assignment of probabilities to events of an experiment, and their consequences.

Axiom 1: $P(A) \geq 0$, for all events A

Axiom 2: $P(\mathcal{S}) = 1$

Axiom 3: (a) If $A_1, A_2, \ldots, A_n$ are disjoint

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

(b) If $A_1, A_2, \ldots$ are disjoint

$$P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

Example 2.4.2. In this example, we now demonstrate that the assignment of probabilities to events in the case of equally likely outcomes is consistent with the above axioms. First it is clear that the assignment is consistent with Axiom 1. To see there is consistency with Axiom 3, note that a) the probability of a simple event A_i (i.e. an event which consists of only one outcome, and thus $N(A_i) = 1$), is $P(A_i) = 1/N$, and b) if event A consists of n simple outcomes (thus $N(A) = n$) then, it can be thought of as the union of the disjoint events $A_1, \ldots, A_n$, where the A_i are simple events, each consisting of an outcome in the event A. Using these observations, if A is an event consisting of n simple outcomes, then, according to Axiom 3,

$$P(A) = P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) = \frac{n}{N}, \tag{2.4.2}$$

as it should be. As exercise, verify that Axiom 2 is also consistent with the assignment of probabilities in the case of equally likely outcomes, i.e. verify that this assignment implies $P(\mathcal{S}) = 1$.

Proposition 2.4.1. *The axioms imply the following properties of probability:*

1. $P(A) = 1 - P(A')$*, for any event* A*.*

2. If A *and* B *are disjoint,* $P(A \cap B) = 0$*.*

3. $P(A) = \sum_{\{all\ simple\ events\ E_i\ in\ A\}} P(E_i)$

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *holds for any two events* A, B*.*

5. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$*, holds for any three events* A, B, C*.*

Example 2.4.3. Consider the experiment where a Hershey Kiss is tossed and it is noted whether it lands on its base or on its side. Thus the sample space is $\mathcal{S} = \{B, S\}$. If we are given that the probability of the Kiss landing on its base is p, then using Axioms 2 and 3 we obtain

$$1 = P(\mathcal{S}) = P(B) + P(S),$$

and thus the probability that the Kiss will land on its side is $1 - p$. For example, if the probability of landing on its base is 0.6 then the probability of landing on its side is 0.4.

Example 2.4.4. Suppose that 60% of all households in a certain community subscribe to newspaper 1, 80% subscribe to newspaper 2 and 50% subscribe to both. If a household is selected at random find the probability it subscribes to: (a) at least one of the two newspapers, and (b) exactly one of the two.

Solution. Set $A = \{\text{subscribes to newspaper 1}\}$ and $B = \{\text{subscribes to newspaper 2}\}$. These events and resulting operations are illustrated in Figure 2.2. Then for (a) we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.8 - 0.5 = 0.9.$$

For (b) we have

$$P(A \cup B) - P(A \cap B) = P(A \cap B') + P(A' \cap B) = 0.1 + 0.3 = 0.4.$$

Additional examples using the axioms and properties of probabilities are given in the next section.

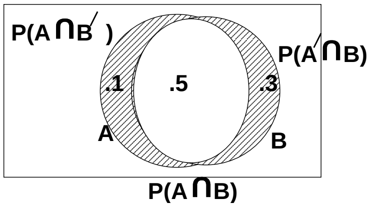


Figure 2.2: Illustration of Events for Example 2.4.4

2.4.4 Sample Space Populations and Probability Sampling

In Example 2.2.1,b) we saw that the simple random sampling of one US citizen aged 18 and over and recording his/her opinion regarding solar energy on the scale $0, 1, \ldots, 10$, can be thought of a random (but not simple random) sampling from the sample space population $\{0, 1, \ldots, 10\}$. The reason it is not simple random sampling is because each of the numbers in the sample space occur with probability equal to the proportion of the population having the corresponding opinion.

Definition 2.4.2. *The type of random sampling where different outcomes have different probability of occurring is referred to as* **probability sampling**.

Example 2.4.5. Consider rolling a die twice. The sample space consists of the pairs of numbers

$$\begin{aligned}&(1,1),(1,2),\ldots,(1,6)\\&(2,1),(2,2),\ldots,(2,6)\\&\vdots\\&(6,1),(6,2),\ldots,(6,6).\end{aligned}$$

Because each of these outcomes is equally likely, the rule of assigning probabilities gives that the probability of a 6 in neither roll is 25/36, the probability of rolling only one 6 is 10/36, and the probability of rolling two 6s is 1/36. Suppose now that we are only interested in recording the number of times a 6 occurs, so that the sample space is $\mathcal{S} = \{0, 1, 2\}$. This experiment can be thought of as a probability sampling experiment where selection from the sample space population $\{0, 1, 2\}$ is done at random according to the above probabilities.

2.4.5 Exercises

1. An electronics factory has machines A and B, each of which produces a single batch of 50 electrical components per hour. The probabilities that in any given hour machines A and B will produce a batch with no defectives are .95 and .92, respectively, and that they both produce a batch with no defectives is .88. Find the probabilities that in any given hour:

   (a) At least one machine produces a batch with no defective components.

   (b) Exactly one machine produces a batch with no defective components.

2. Suppose that in State College 40% of the households subscribe to newspaper A, 30% to newspaper B, and 60% of the households subscribe to at least one of the two newspapers. Let A denote the event that a randomly chosen household subscribe to newspaper A, and let B denote the corresponding event for newspaper B.

   (a) Find the probability that a randomly selected household subscribes to both newspapers.

   (b) Find the probability that a randomly selected household subscribes only to newspaper A (that is, subscribes to A but not to B).

3. In a literature class, 30% of students speak French, 35% speak German and 50% of students speak French OR German. What is the probability that a randomly selected student

   (a) speaks French AND German?

   (b) speaks French but not German?

4. An unfair die with six sides is rolled. Let X denote the outcome of the roll. The probability of outcome i is given by $P(X = i) = i/21$, $i = 1, 2, \ldots, 6$.

   (a) Show that Axiom 2 of probability is satisfied.

   (b) Find the probability that the outcome is an even number.

   (c) Use part 1 of Proposition 2.4.1 to find the probability that the outcome is strictly greater than one.

5. Consider the 100 disks of polycarbonate plastic, as described in Exercise 2.3.1,3. Thus, they are classified according to their scratch and shock resistance according to the table

		shock resistance	
		high	low
scratch	high	70	9
resistance	low	16	5

Suppose that one of the 100 disks is randomly selected.

(a) What is the probability that the selected disk has low shock resistance?

(b) What is the probability that the selected disk has low shock resistance or low scratch resistance?

6. Consider the game where two dice, die A and die B, are rolled. We say that die A wins, and write $A > B$, if the outcome of of rolling A is larger than that of rolling B. If both rolls result in the same number it is a tie.

   (a) Find the probability of a tie.

   (b) Find the probability that die A wins.

7. **Efron's Dice.** Using arbitrary numbers on the sides of dice can have surprising consequences. Efron's dice are sets of dice with with the property that for each dice there is one that dominates it (beats it with larger probability) when the game described in Exercise 6 is played. An example of Efron's dice is as follows:

   - Die A: four 4s and two 0s
   - Die B: six 3s
   - Die C: four 2s and two 6s
   - Die D: three 5's and three 1's

   (a) Specify the events $A > B$, $B > C$, $C > D$, $D > A$.

   (b) Find the probabilities that $A > B$, $B > C$, $C > D$, $D > A$.

8. **Let's Make a Deal.** In the game Let's Make a Deal, the host asks a participant to choose one of three doors. Behind one of the doors is a big prize (e.g. a car), while behind the other two doors are minor prizes (e.g. a blender). After the participant selects a door, the host opens one of the two doors and shows the participant a door that does not have the big prize. (The host does not show to the participant what is behind the door the participant chose.) The host asks the participant to either

   (a) stick with his/her original choice, or

   (b) select the other of the remaining two closed doors.

   Find the probability that the participant will win the big prize for each of the strategies a) and b).

2.5 Independent Events

Though the axioms and properties of probability do not contain explicit formulas for calculating the intersection of two events, part 3 of Proposition 2.4.1 contains implicitly the following formula

$$P(A \cap B) = P(A) + P(B) - P(A \cup B). \tag{2.5.1}$$

A different (and simpler) formula can be used when the events arise in connection with experiments which are performed *independently*, or with *independent* repetitions of the same experiment. By **independent experiments** or **independent repetitions** of the same experiment we mean that there is no mechanism through which the outcome of one experiment will influence the outcome of the other. In such a case, the probability of the intersection of two events is the product of each of their probabilities.

Example 2.5.1. Consider tossing a die twice. Let $A = \{\text{outcome of 1st toss is even}\}$ and $B = \{\text{outcome of 2nd toss is even}\}$. Then,

$$\begin{aligned} P(A \cap B) &= \frac{N(A \cap B)}{N} = \frac{9}{36} = \frac{1}{4} \\ &= \frac{1}{2}\frac{1}{2} \\ &= P(A)P(B) \end{aligned}$$

The probabilities in Example 2.5.1 are calculated under the assumption that all 36 outcomes of the experiment of tossing a die twice are equally likely. This assumption is equivalent to saying that the two tosses are performed independently. Note further that in Example 2.5.1, the event A is associated with the first toss while event B is associated with the second toss.

Having demonstrated that the probability of the intersection of events associated with independent experiments is the product of their probabilities, we now present the following definition.

Definition 2.5.1. *Events A and B are called* **independent** *if*

$$P(A \cap B) = P(A)P(B).$$

Example 2.5.2. Consider buying a washing machine and a dryer. Suppose that 30% of washing machines require warranty service, while 10% of the dryers do. What is the probability that both machines will require warranty service?

Solution. Assume that the two machines function independently. Thus operating the machines during their warranty period corresponds to performing two independent experiments. Let A denote the event that the washing machine requires warranty service, and B the event that the dryer requires warranty service. Since these events are associated with independent experiments they are independent. Therefore,

$$P(A \cap B) = P(A)P(B) = (0.3)(0.1) = 0.03$$

Note that with the information given in Example 2.5.2, it is not possible to evaluate the desired probability without the assumption of independence. For example, use of the formula (2.5.1) requires knowledge of the probability of the union of A and B; similarly use of the formulas in in Section 2.7 also requires additional knowledge. If the assumption of independence is incorrect, then so is the answer that was obtained in Example 2.5.2.

The concept of independence extends to more than two events:

Definition 2.5.2. *The events $A_1, \ldots, A_n$ are* **mutually independent** *if*

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k})$$

for any sub-collection $A_{i_1}, \ldots, A_{i_k}$ of k events chosen from $A_1, \ldots, A_n$.

Example 2.5.3. The system of components shown in Figure 2.3 below works as long as components 1 and 2 both work, or components 3 and 4 both work. Assume that the components work or fail independently, and each works with probability 0.9. Find the probability that the system works. *Solution.* Let $A_i,\ i = 1, 2, 3, 4$, denote the event that

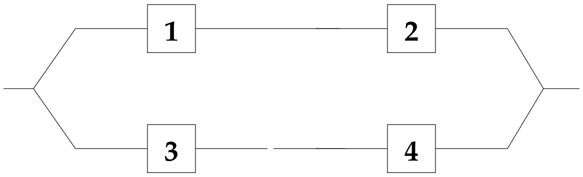


Figure 2.3: System of Four Components

the ith component works. Thus, $P(A_i) = 0.9,\ i = 1, 2, 3, 4$. We want the probability of $B = (A_1 \cap A_2) \cup (A_3 \cap A_4)$. Using part 4 of Proposition 2.4.1, and then the assumption of independence we have

$$P(B) = P(A_1 \cap A_2) + P(A_3 \cap A_4) - P(A_1 \cap A_2 \cap A_3 \cap A_4)$$
$$= (.9)(.9) + (.9)(.9) - (.9)(.9)(.9)(.9) = .9639$$

Example 2.5.4. Suppose that fuses are inspected as they come off a production line until the first defective fuse is found. If we record the number of fuses inspected, then the sample space of this experiment is $\mathcal{S} = \{1, 2, 3, \ldots\}$. Assume that each fuse which comes off the production line will be defective with a probability of 0.01, independently of other fuses. Find the probability of each outcome in the sample space and verify that these probabilities sum to 1, in agreement with Axiom 2, i.e. $P(\mathcal{S}) = 1$.

Solution. Let D denote the event that a selected item is defective and N be the event that a selected item is nondefective. Since outcome 1 occurs when the first selected item is defective, it follows that $P(1) = P(D) = 0.01$. Moreover, using the independence assumption it is easily seen that $P(2) = P(ND) = (.99)(.01)$, $P(3) = P(NND) = (.99)^2(.01)$, and so on.

Since $\mathcal{S} = \{1\} \cup \{2\} \cup \{3\} \cup \ldots$, Axiom 3(b) gives

$$P(\mathcal{S}) \;=\; P(1) + P(2) + P(3) + \ldots = 0.01[1 + 0.99 + (0.99)^2 + (0.99)^3 + \ldots] = 1,$$

in agreement with Axiom 2.

The last example of this section illustrates the type of situations where the calculation of the probability of an event is difficult whereas that of its complement is simple.

Example 2.5.5. Consider a system of 5 components connected in series

$$--\boxed{1}--\boxed{2}--\boxed{3}--\boxed{4}--\boxed{5}--$$

The system fails if one component fails. If components fail independently with a probability of 0.1, what is the probability of the event that the system fails?

Solution. Let A denote the event that the system fails. The probability of A is most easily computed using part 1 of Proposition 2.4.1. Indeed

$$P(A') = P \text{ (no component fails)} = (.9)^5 = .59$$

and thus $P(A) = 1 - P(A') = .41$. The difficulty of computing $P(A)$ directly is attributable to the fact that the event A is comprised of a multitude of sub-events such as the event that component 1 fails, or the event that components 1 and 2 fail together, etc.

2.5.1 Exercises

1. Suppose the events $A_1, A_2, A_3, A_4, A_5, A_6$, and A_7 partition the sample space of an experiment (i.e. they are mutually exclusive and exhaustive) and that they are equally likely. Calculate $P(A_1 \cap A_4)$ and $P(A_3 \cup A_5 \cup A_6)$.

2. Suppose $P(A) = 0.25, P(B) = 0.62, P(C) = 0.11$, and we are given that $P(A \cap B) = 0.17$, $P(A \cap C) = 0.02$, and $P(B \cup C) = 0.63$. Find a) $P(A \cup C)$ and b) $P(B \cap C)$.

3. Suppose 10% of all a certain type of software widgets have connectivity problems. Suppose that a simple random sample of 10 such widgets are installed. What is the probability that:

   (a) None of the 10 have connectivity problems?

   (b) The first widget installed has connectivity problems but the rest do not?

   (c) Exactly one of the 10 has connectivity problems?

4. In a fictitious recent poll of all U.S. citizens who had seen all 3 episodes of the Star Wars trilogy, 60% said their favorite episode is "Star Wars," 50% said they plan to see all 3 re-released episodes in the theater, and 25% said "Star Wars" is their favorite episode and they plan to see all 3 re-released episodes. Define the events
   $A = \{$respondent favors "Star Wars"$\}$ and
   $B = \{$respondent plans to see all 3 episodes$\}$.

   (a) Find the probability of $A \cup B$.

   (b) Find the probability of B given A.

   (c) Are A and B independent? Why or why not?

   (d) A total of 10 individuals who said they plan to see all three episodes have been selected as finalists for receiving prizes. The first 3 prizes are, respectively, \$10,000, \$5,000, and \$1,000. The remaining will get a Star Wars t-shirt. In how many ways can the first 3 winners be selected?

5. The probability that an American engineering firm will establish a branch office in Toronto is .7, the probability that it will establish a branch office in Mexico City is .4 and the probability it will establish a branch office in at least one of the cities is .8. Define events A and B as follows:
   A = {American engineering firm will establish a branch office in Toronto}
   B = {American engineering firm will establish a branch office in Mexico City}

It may be useful to draw Venn diagrams. For parts (a)-(c), what is the probability that a branch office will be established in

(a) in both cities?

(b) neither of the cities?

(c) exactly one of the cities?

(d) Are events A and B independent? Why or why not?

6. In the context of Exercise 2.3.1,3:

   (a) Let A be the event that the randomly selected disk has low shock resistance and B be the event that it has low scratch resistance. Are the events A and B independent? (Justify your answer.)

   (b) If two disks are selected at random from the 100 disks, find the probability that none of the two is of the low scratch resistance, low shock resistance type.

7. An automobile assembly line is manned by two shifts a day. The first shift amounts for two thirds of the overall production. The task of quality control engineers is to monitor the number of non-conformances per car. Each day a simple random sample of six cars from the first shift, and a simple random sample of three cars from the second shift is taken, and the number of non-conformances per car in recorded. Let A denote the event that all cars inspected from the first shift have zero non-conformances, and B denote the corresponding event for the second shift.

   (a) Suppose that the probability that an automobile produced in the first shift has zero non-conformances is 0.75, independently from other automobiles. The corresponding probability for the second shift is 0.9. Find $P(A)$ and $P(B)$.

   (b) With the information given in the previous question, find $P(A \cap B)$ and $P(A \cup B)$.

8. We select a student-athlete from a small private high school at random. Let A be the event {the student-athlete is male} and B be the event {the student-athlete prefers track}. We are told that the proportion of male-athletes who prefer track is the same as the proportion of student-athletes who prefer track.

   (a) This statement implies that (*choose one*):

A The events A and B are disjoint.

B The events A and B are independent.

C The events A and B are both disjoint and independent.

D The events A and B are neither disjoint nor independent.

(b) Does this statement imply that the proportion of female-athletes who prefer track is the same as the proportion of student-athletes who prefer track?

9. In the high school of Exercise 8, sixty-five percent of the student-athletes are male, and are allowed to participate in three varsity sports: football, basketball, and track. However, the female athletes are allowed to participate in only basketball and track. Thirty percent of the student-athletes say that football is their favorite sport in which to compete, while 50% prefer basketball and 20% prefer track.

   (a) Place this information in the table below, and complete the table under the assumption that the events A and B, defined in Exercise 8, are independent.

	Football	Basketball	Track	
Male				
Female				
				1

   (b) If a randomly selected student-athlete prefers basketball, what is the probability that the student-athlete is female?

2.6 Counting Techniques

In Section 2.4 we saw that in experiments with a finite number of equally likely outcomes the probability of an event is the number of outcomes that comprise the event divided by the total number of possible outcomes. Implementing this simple method, however, requires counting the number of outcomes in the sample space as well as the number of outcomes in the events of interest. In this section we present techniques that facilitate this counting.

The product rule: If a task can be completed in two stages, if stage 1 has n_1 outcomes, and if stage 2 has n_2 outcomes, regardless of the outcome in stage 1, then the task has $n_1 n_2$ outcomes.

Example 2.6.1. A president and a vice-president for student council will be selected from four finalists. How many outcomes are possible?

Solution. Let stage 1 be the selection of president, and stage 2 be the selection of vice-president. Then $n_1 = 4, n_2 = 3$. Thus there are $n_1 n_2 = 12$ possible outcomes.

Other examples of tasks where the product rule applies are:

Example 2.6.2. (a) Select a junior (2-year) college and then a state college from 5 junior and 3 state college.

(b) Select a plumber and an electrician, from those available in the yellow pages.

(c) Select a first and second winner in a certain competition, from a group of finalists.

Note that in (a) and (b) of Example 2.6.2, different groups of objects or subjects are involved in the two stages. For example, a junior college is selected in stage I of task (a) and a stage college is selected in stage II. On the other hand, in task (c), the selection in both stages is from the same group. In some cases where the selection is from the same group it may be that we do not care to distinguish between the two stages. For example, in the context of task (c), it may be that both winners take the same prize, so it does not make sense to talk of first and second winner. When the selection in all stages is from the same group of subjects or objects and we do distinguish the outcomes of the two stages, we talk about **ordered** outcomes. The product rule applies only for the enumeration of ordered outcomes, or when the selection in different stages is from different groups.

The general product rule: If a task can be completed in k stages and stage i has n_i outcomes, regardless of the outcomes the previous stages, then the task has $n_1 n_2 \dots n_k$ outcomes.

Examples of tasks that can be completed in more than two stages are given in Example 2.6.3 that follows. (Note that Examples 2.6.3(b),(c) generalize corresponding examples in Example 2.6.2.)

Example 2.6.3. (a) Select a president, a vice-president & treasurer.

(b) Select a plumber, an electrician, & a person for remodeling

(c) Select a 1st, 2nd & 3rd place winner.

When the stages involve selection from the same group of objects or subjects, as in (a) and (c) of Example 2.6.3, the number of outcomes at each stage is easily determined from the size of the group from which the selections are made. For example, if the 1st, 2nd & 3rd place winner are to be selected from a group of n finalists, then stages 1, 2 & 3 have

$$n_1 = n, \quad n_2 = n - 1, \quad , \text{and} \;\; n_3 = n - 2$$

outcomes, respectively. (As explained before, it should be kept in mind that the general rule applies only to when we care to distinguish the outcomes of the different stages.)

In general, if all selections come from the same group of objects or subjects and if the task consists of k stages, the number of outcomes at stages $1, 2, \ldots, k$ are

$$n_1 = n, n_2 = n - 1, \ldots, n_k = n - k + 1.$$

The resulting total number of outcomes of such a task, i.e.

$$n_1 n_2 \ldots n_k = n(n-1)\ldots(n-k+1)$$

is called the number of **permutations** of k objects selected from a group of n objects and is denoted by

$$P_{k,n} = n(n-1)\ldots(n-k+1). \tag{2.6.1}$$

Using the notation $m!$, which is read 'm factorial' and is defined as

$$m! = m(m-1)\ldots(2)(1), \tag{2.6.2}$$

we can write

$$P_{k,n} = \frac{n!}{(n-k)!} \tag{2.6.3}$$

Note: By convention we set $0! = 1$. This gives $P_{n,n} = n!$, as it should, because selecting n (ordered) objects out of n, is a task having n stages with $n_1 = n, n_2 = n-1, \ldots, n_n = 1$.

Example 2.6.4. An exam consists of four questions. The professor wishes to select a different TA for each question. There are 8 TAs. In how many ways can the TAs be chosen?

Solution. This can be solved in two ways:
(i) Apply the general product rule with stage 1 being the selection of the TA for question

1, stage 2 the selection of the TA for question 2, stage 3 the selection of the TA for question 3 and stage 4 the selection of the TA for question 4. Thus,

$$n_1 = 8, \; n_2 = 7, \; n_3 = 6, \; n_4 = 5,$$

and $n_1 n_2 n_3 n_4 = 1680$.

(ii) Recognize the answer as being the permutation of 4 subjects chosen out of 8 and write it directly

$$P_{4,8} = \frac{8!}{4!} = 8 \cdot 7 \cdot 6 \cdot 5 = 1680.$$

Now we turn attention to the case where we do not care to distinguish the selections at the different stages. Thus, if all winners receive the same prize we do not care to distinguish between 1st, 2nd & 3rd place winner; when the objective is to select three new members of a student council (instead of selecting a president, a vice-president & treasurer), we do not care to distinguish the outcomes of the three stages.

When we are interested in the group, or collection, of the subjects or objects without a predetermined assignment, or another type of ordering, we speak of **combinations**.

It should be clear that the number of possible combinations is smaller than the number of possible permutations. For example, the same group or combination of three winners, say Niki, George and Sophia, can result in $3! = 6$ permutations if the prizes are different and thus we distinguish between the 1st, 2nd & 3rd place winner. In general, from each combination of k objects or subjects there can result $k!$ permutations. Thus, the number of combinations of k objects or subjects selected from a group of n is

$$\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

The notation $C_{k,n}$ is also used in some books for the number of combinations, but it is less common than $\binom{n}{k}$.

Example 2.6.5. Consider the TA selection problem of Example D, except that now there are 10 questions to be graded and the professor does not care which TA grades which questions. In how many ways can 4 TAs be chosen?

Solution. Since there is no predetermined assignment for the TA chosen at each stage, the question pertains to the number of combinations of 4 chosen out of 8. This number is

$$\binom{8}{4} = \frac{P_{4,8}}{4!} = \frac{1680}{24} = 70.$$

If properly utilized, the product rules and the concepts of permutations and combinations can yield answers to complex problems as demonstrated by the following.

Example 2.6.6. Find the probability that a random hand of 5 cards, selected from a deck of 52 cards, is a full house.

Solution. First, the number of all 5-card hands is $\binom{52}{5} = \dfrac{52!}{5!47!} = 2,598,960$. To find the number of possible full houses that can be formed, think of the task of forming a full house as consisting of two stages. Stage 1 consists of choosing two cards of the same kind, and stage 2 consists of choosing three cards of the same kind. Since there are 13 kind of cards, stage 1 can be completed in $\binom{13}{1}\binom{4}{2} = (13)(6) = 78$ ways. (Think of the task of stage 1 as consisting of first selecting a kind, and then selecting two from the four cards of the selected kind.) For each outcome of stage 1, the task of stage 2 becomes that of selecting three of a kind from one of the remaining 12 kinds. This can be completed in $\binom{12}{1}\binom{4}{3} = 48$ ways. Thus there are $(78)(48) = 3,744$ possible full houses.

2.6.1 Exercises

1. From 5 biologists, 4 chemists, and 3 physicists,

   (a) how many committees of size 4 can be formed?

   (b) how many committees containing 2 biologists, 1 chemist, and 1 physicist can be formed?

2. Suppose that 8 of the 20 buses in a particular city have developed cracks on the underside of the main frame. Five buses are to be selected for thorough inspection.

   (a) How many possible selections are there?

   (b) How many selections contain exactly 4 buses with cracks?

   (c) If buses are chosen at random, what is the probability that exactly 4 of the 5 selected will contain cracks?

3. A drawer contains 6 black and 8 white shirts.

   (a) How many ways are there to select a sample of 5 shirts from the drawer?

   (b) If a sample of 5 shirts is randomly chosen, what is the probability that there are exactly 3 white shirts in the sample.

(c) If a sample of 5 shirts is randomly chosen, what is the probability that there are at least 3 white shirts in the sample.

4. In a shipment of 10 electronic components, 2 are defective. Suppose that 5 components are selected at random for inspection.

   (a) What is the probability that no defective components are included among the 5 selected?

   (b) What is the probability of exactly one defective component being included?

5. A manufacturer of shingles makes three different grades of asphalt shingles, high, medium, and low, each with a different warranty. Currently in the warehouse there are 10 palettes of high grade shingles, 25 palettes of medium grade shingles,and 30 palettes of low grade shingles. An order comes in for 5 palettes of low grade shingles. An inexperienced shipping clerk is unaware of the distinction in grades of asphalt shingles and he ships 5 randomly selected palettes.

   (a) How many different groups of 5 shingles are there? (Each shingle is identified from all others by its serial number.)

   (b) What is the probability that all of the shipped palettes are high grade?

   (c) What is the probability that all of the shipped palettes are of the same grade?

2.7 Conditional Probability

Many experiments consist of recording the value of more than one characteristic from each population unit, in other words, they have a multivariate outcome variable.

Example 2.7.1. a) A particular consumer product is being assembled on two production lines, which we label A and B. Consider now the experiment which consists of selecting, through simple random sampling, a product item, inspecting it for the presence of a particular defect and also recording whether it was assembled in line A or B.
b) Consider the experiment which consists of selecting, through simple random sampling, a person aged 18 or over, and recording the person's opinion on solar energy, as well as the person's age and gender.

In such cases, information about one characteristic might help improve our prediction about the other characteristic. The problem of prediction will be studied in more detail in later chapters. For now, (possible) improved prediction will mean (possible) change of the probability of an event having to do with the value of one of the characteristics, when some information on the value of the other characteristic of the unit is available. Thus, knowing which production line a particular product came from might lead to improved prediction about (i.e. update the probability of) the presence or absence of the defect; knowing a person's age and gender might improve our prediction about (i.e. update the probability of) that person feeling strongly about solar energy.

Knowing part of the outcome of the experiment effectively shrinks the underlying population and the corresponding sample space. It is possible, therefore, to consider that a simpler experiment has been performed instead. This is illustrated in the next example.

Example 2.7.2. a) Consider the experiment of Example 2.7.1a). If we know that a particular product item has come from production line A, then it is a representative of product items that are assembled in line A. As such the particular product item and can be regarded as having been selected in a different experiment. Namely, the experiment that selects, through simple random sampling, from the population of items that come from line A, and records only whether or not the defect is present (univariate outcome variable).

b) Consider the experiment of Example 2.7.1b). If we are told that the selected person is a 29 year old female, then that particular person can be regarded as having been obtained through simple random sampling from the population of 29 year old females, and recording only her opinion on solar energy (univariate response variable).

c) Consider the experiment of rolling a die and recording the outcome. Suppose we are given the information that the outcome is even. This information shrinks the sample space from $\{1, 2, 3, 4, 5, 6\}$ to $\{2, 4, 6\}$. [The experiment of rolling a die can be considered as performing simple random sampling from the sample space, in which case the population shrinks in the same way as the sample space; it might also be considered as performing simple random sampling from the conceptual population of all rolls of a die, in which case this population shrinks to those rolls of a die that result in an even number.] Regardless of how we conceptualize the population, the even outcome that has occurred can be regarded as having been obtained through simple random sampling from the reduced population.

We write $P(A|B)$ for the **conditional probability** of the event A given the information that event B has occurred.

The insight that a given piece of information shrinks the sample space, while the selection from the reduced population continues to be of the simple random sampling kind, is useful for understanding the calculation of conditional probabilities.

Example 2.7.3. A die is rolled and we are told that the outcome is an even number. What is the probability that the outcome is 2?

Solution. If A denotes the event that the outcome is 2, and B the event that the outcome is even, we are asked to find $P(A|B)$. Since the outcomes of the reduced sample space (see Example 2.7.2c)) are still equally likely, $P(A|B) = 1/3$.

Example 2.7.4. Consider Example 2.4.4 of Section 2.4. In this example, we recorded two characteristics from the randomly selected household: a) whether or not the household subscribes to newspaper 1, and b) whether or not the household subscribes to newspaper 2. Given that the selected household subscribes to newspaper 1, what is the probability it also subscribes to newspaper 2?

Solution. The information that the household subscribes to newspaper 1 shrinks the (finite and non-conceptual) population of all households in the community to those that subscribe to newspaper 1, and we can regard the selected household as a simple random sample of size one from this reduced population. It follows that the desired probability is the proportion of those that subscribe to both among those that subscribe to newspaper 1. We are given that, among all households, the proportion that subscribe to newspaper 1 is 60%, while those that subscribe to both is 50%. Thus the desired probability is $0.5/0.6 = 0.833$.

The solution to Example 2.7.4 suggests the following computational definition of conditional probability.

Definition 2.7.1. *For any two events A, B with $P(B) > 0$,*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

In the definition of conditional probability, the events A and B might themselves be unions or intersections of other events.

Example 2.7.5. From the population of cell phone users with a Verizon plan, a person is selected according to simple random sampling. Let A be the event that the chosen subscriber has friends or family added to his/her plan, B denote the event that the subscriber has unlimited text messaging, and C denote the event that the subscriber owns a cell phone with digital imaging capabilities. It is given that

$$P(A) = 0.37, \; P(B) = 0.23, \; P(C) = 0.14,$$

$$P(A\cap B)=0.13,\ P(A\cap C)=0.09,$$

$$P(B\cap C)=0.08,\ P(A\cap B\cap C)=0.05.$$

Find $P(C|A\cup B)$, and $P(B\cup C|A)$.

Solution. Applying the definition of conditional probability we have

$$P(C|A\cup B)=\frac{P(C\cap(A\cup B))}{P(A\cup B)}$$

Now, since $C\cap(A\cup B)=(C\cap A)\cup(C\cap B)$, the numerator above equals

$$\begin{aligned}&P(C\cap A)+P(C\cap B)-P((C\cap A)\cap(C\cap B))\\ =\;&0.09+0.08-0.05=0.12,\end{aligned}$$

since $(C\cap A)\cap(C\cap B)=A\cap B\cap C$. Applying the formula for the probability of the union of two events one more time we find $P(A\cup B)=0.47$. Thus,

$$P(C|A\cup B)=\frac{0.12}{0.47}=0.255.$$

Working similarly, we obtain

$$P(B\cup C|A)=\frac{P((B\cup C)\cap A)}{P(A)}=\frac{0.17}{0.37}=0.459.$$

The definition of $P(A|B)$ yields the following formula for calculating the probability of the intersection of two events.

MULTIPLICATION RULE:

$$P(A\cap B)=P(A|B)P(B),\quad\text{or}\quad P(A\cap B)=P(B|A)P(A)\tag{2.7.1}$$

The multiplication rule extends to more than two events. For example, the extension to three events is

$$\begin{aligned}P(A\cap B\cap C)&=P(A|B\cap C)P(B\cap C)&(2.7.2)\\&=P(A|B\cap C)P(B|C)P(C).&(2.7.3)\end{aligned}$$

Example 2.7.6. a) Suppose that 50% of TV sales are of brand 1, 30% are of brand 2 and 20% are of brand 3. It is known that 25% of brand 1 TVs require warranty repair work, as do 20% of brand 2 and 10% of brand 3. In this narrative, which of the percentages given correspond to conditional probabilities?

b) Find the probability that the next TV purchase is brand 1 TV which will need warranty repair work.

Solution. a) The probabilities in the first sentence are unconditional while those in the second sentence are conditional. To see this, note that the outcome here consists of recording two characteristics of the next purchaser: the brand of the purchased TV, and whether the purchased TV gets brought back for warranty repair work. It is seen that the second sentence makes probability statements regarding the second characteristic for each fixed value of the first characteristic. Thus they are conditional probabilities.

b) Let A denote the event where a brand 1 TV is bought, and B denote the event where the bought TV will need warranty repair work. We are being asked for the probability of the intersection of A and B. From the multiplication rule we obtain

$$P(A \cap B) = P(B|A)P(A) = (.25)(.5) = .125.$$

Example 2.7.7. Pick three cards from a deck. Find the probability that the 1st draw is an ace, the 2nd draw is a king and the third draw is a queen.

Solution. Let $A = \{$1st draw results in ace$\}$, $B = \{$2nd draw results in king$\}$, and $C = \{$3rd draw results in queen$\}$. Thus we want to calculate $P(A \cap B \cap C)$. From the multiplication rule for three events we have

$$\begin{aligned} P(A \cap B \cap C) &= P(A|B \cap C)P(B|C)P(C) \\ &= \frac{4}{50}\frac{4}{51}\frac{4}{52} = 0.000455. \end{aligned}$$

2.7.1 The Law of Total Probability

In this subsection we present a rule for calculating the probability of an event B which can arise in connection with some events, say $A_1, \ldots, A_k$, which are disjoint and make up the entire sample space, i.e. $A_i \cap A_j = \emptyset$, for all $i \neq j$, and $A_1 \cup A_2 \cup \ldots \cup A_k = \mathcal{S}$; see Figure 2.4 below. Thus, since the events $B \cap A_i,\ i = 1, \ldots, k$, are disjoint and they make up the event B, probability Axiom 3 implies that

$$P(B) = P(B \cap A_1) + \cdots + P(B \cap A_k).$$

Using the multiplication rule we get,

$$P(B) = P(A_1)P(B|A_1) + \cdots + P(A_k)P(B|A_k), \tag{2.7.4}$$

which is the **Law of Total Probability**.

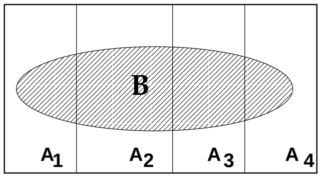


Figure 2.4: An Event B Arising in Connection with Events $A_1, \ldots, A_4$

Example 2.7.8. In the setting of Example 2.7.6, find the probability the next TV purchased which will need warranty repair work.

Solution. Define the events $B = \{\text{TV will need warranty repair}\}$, $A_1 = \{\text{TV is brand 1}\}$, $A_2 = \{\text{TV is brand 2}\}$, $A_3 = \{\text{TV is brand 3}\}$, and apply the formula for the Law of Total Probability. Thus

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) \\ &\quad + P(A_3)P(B|A_3) \\ &= (.5)(.25) + (.3)(.2) + (.2)(.1) = .205 \end{aligned}$$

Example 2.7.9. Draw two cards, sequentially, from a standard deck of 52 cards. What is the probability that the card in the second draw is an ace?

Solution. Let B denote the event that the second draw results in an ace, A_1 denote the event that the first draw results in an ace, and let A_2 be the complement of A_1. Then, according to the Law of Total Probability,

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) \\ &= \frac{4}{52}\frac{3}{51} + \frac{48}{52}\frac{4}{51} = \frac{4}{52}. \end{aligned}$$

In other words, Example 2.7.9 states that the probability of an ace in the second draw is the same as that of an ace in the first draw.

2.7.2 Bayes Theorem

Consider events B and $A_1, \ldots, A_k$ as in the previous subsection. Now, however, we ask a different question: Given that B has occurred, what is the probability that a particular

A_j has occurred? The answer is provided by the **Bayes theorem**:

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)}. \tag{2.7.5}$$

Example 2.7.10. In the setting of Example 2.7.6, suppose that a customer returns a TV for warranty repair. What is the probability it is a brand 1 TV?

Solution. We want to find $P(A_1|B)$. Using the Bayes theorem we obtain

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{\sum_{i=1}^{3} P(A_i)P(B|A_i)} = \frac{(.5)(.25)}{.205} = .61.$$

Note: $P(A_1|B) = .61$ is called the *posterior* probability while $P(A_1) = .5$ is the *prior* probability.

2.7.3 Exercises

1. Rusty's Body Shop repairs GM and Ford cars of all ages. The following table provides information about the probability that a randomly chosen car is of a particular make and age class:

Car make\ Age class	< 3 years	$3-8$ years	> 8 years
GM	.10	.20	.30
Ford	.05	.15	.20

Let A denote the event that a randomly chosen car is less than 3 years old, and B the event that a randomly chosen car is a Ford.

   (a) Determine the probabilities $P(A)$, $P(B)$ and $P(A \cap B)$.

   (b) Are the events A and B independent? Why or why not.

   (c) Calculate $P(A|B)$ and $P(B|A)$ and interpret their meaning.

2. A batch of 10 fuses has three defective ones. A sample of size two is taken at random and without replacement, and the fuses are examined. Given that there is exactly one defective fuse in the sample, what is the probability that the defective fuse was the first one selected?

3. Thirty percent of credit card holders carry no monthly balance, while 70% do. Of those card holders carrying a balance, 30% have annual income \$20,000 or less, 40% between \$20,001 - \$50,000, and 30% over \$50,000. Of those card holders carrying no balance, 20%, 30%, and 50% have annual incomes in these three respective categories.

(a) What is the probability that a randomly chosen card holder has annual income \$20,000 or less?

(b) If this card holder has an annual income that is \$20,000 or less, what is the probability that (s)he carries a balance?

4. Bean seeds from supplier A have a 50% germination rate and from supplier B have a 75% germination rate. A seed packaging company purchases 40% of their bean seeds from supplier A and 60% from supplier B.

   (a) A seed is selected at random from the mixed seeds. Find the probability that it will germinate.

   (b) If a randomly selected seed germinates, what is the probability that it came from supplier A?

5. Bowl A contains two red chips, bowl B contains two white chips, and bowl C contains one red chip and a white chip. A bowl is selected at random (with equal probabilities), and one chip is taken at random from that bowl.

   (a) Find the probability of selecting a white chip, say P(W).

   (b) If the selected chip is white, compute the conditional probability that bowl C was selected.

6. A bookstore specializes in three kinds of books: Romance, adventure and fiction. Books are either hardcover or paperback. The following table provides the probability that a typical customer will choose a book of a particular kind and cover:

Kind/Cover	Paperback	Hardcover
Romance	0.15	0.20
Adventure	0.25	0.10
Fiction	0.20	0.10

   Let A_1, A_2 denote the events that the next customer will choose paperback, hardcover, respectively. Similarly, let B_1, B_2, B_3 denote the events that the next customer will choose romance, adventure, fiction, respectively.

   (a) The probability 0.25 in the above table corresponds to (circle one)

   (i) $P(A_1 \cap B_2)$ (ii) $P(A_1 \cup B_2)$ (iii) Other (specify)

   (b) Find $P(A_1)$, $P(B_1)$.

   (c) Are events A_1 and B_1 independent? Why or why not ?

(d) Are events A_1 and A_2 independent? Why or why not ?

(e) Suppose that the next customer comes determined to buy an adventure story. Find the probability that s/he will buy paperback.

7. Suppose that Ford has 40% of the American new automobile market, and that General Motors has 35% and Chrysler has the remaining 25%. Suppose also that 30% of all new automobiles sold are trucks, while 70% are cars. The distribution of new automobiles is as in the table below:

	Car	Truck	
Ford	.28	.12	.40
General Motors	.25	.10	.35
Chrysler		.08	.25
	.70	.30	1.00

What is the probability that the next person who applies for license plates for a new automobile

(a) just purchased a Ford or a truck?

(b) just purchased a Chrysler car?

(c) just purchased a car given that he just purchased a General Motors automobile?

(d) Are the events {the automobile is a car} and {the automobile was made by General Motors} independent? (Justify your answer)

(e) A small Ford dealership received three new cars and two new trucks this week. What is the probability that each of the next three customers gets the automobile type (car or truck) he/she wants, assuming that the 3 customers act independently? (**Hint:** The complement of the event of interest is that the three customers will not all get the type of car they want. This can happen only if they all want a truck.)

8. State College Police plans to enforce speed limits by using radar traps at 4 different locations. The radar traps at each of the locations, L_1, L_2, L_3, L_4 are operated 40%, 30%, 20%, and 30% of the time. A person who is speeding on his way to Penn State has probabilities of 0.2, 0.1, 0.5, and 0.2, respectively, of passing through these locations.

(a) What is the probability the speeding person will receive a speeding ticket?

(b) If the person received a speeding ticket on his way to Penn State, what is the probability that he passed through the radar trap located at L_2?

9. Seventy percent of the light aircraft that disappear while in flight in a certain country are subsequently discovered. Of the aircraft that are discovered, 60% have an emergency locator, whereas 10% of the aircraft not discovered have such a locator. Define events A and B as follows:
   A = {light aircraft that disappears is discovered}
   B = {light aircraft that disappears has an emergency locator}
   Suppose a light aircraft has disappeared.

   (a) What is the probability that it has an emergency locator and it will not be discovered?

   (b) What is the probability that it has an emergency locator?

   (c) If it has an emergency locator, what is the probability that it will not be discovered?

10. Each morning the manager of a Barney toy fabrication plant inspects the output of an assembly line. There are two assembly lines, A and B, at the plant, and, so as not to be predictable to the line foreman, she flips a fair coin to determine which assembly line to inspect that morning. She then inspects three Barneys from the chosen assembly line. In assembly line A, each Barney is defective with probability 0.10, independently of all the other Barneys. In assembly line B, each Barney is defective with probability 0.01, independently of all the other Barneys.

    (a) Find the probability that, in a particular morning's inspection, she will find no defective Barneys.

    (b) Suppose that in a particular morning she finds no defective Barneys. What is the probability that the Barneys she inspected came from assembly line A?

2.8 Computer Activities

1. Here we will demonstrate how to use Minitab to do probability sampling from the finite sample space population $\{1, 2, 3, 4, 5\}$, when the probability for each number being selected is 0.1, 0.2, 0.4, 0.2, 0.1, respectively.

   (a) Enter numbers 1-5 in column 1.

   (b) Enter the probabilities 0.1, 0.2, 0.4, 0.2, 0.1 in column 2.

   (c) Use the sequence of commands:

      Calc¿Random Data¿Discrete¿Enter "100" into Generate; enter "C3" into Store in c

      enter "C1" into Values in; enter "C2" in Probabilities in¿OK

These commands will generate a probability sample of size 100 from the numbers 1-5, according to the probability mass function given in C2. [Note that probability sampling is with replacement.]

2. In this activity we will see numerical and graphical methods to display the results of the previous problem.

   (a) For a numerical display of the numbers use the commands:

      Stat¿Basic Statistics¿Display Descriptive Statistics¿Enter c3 in "Variables" and in "

      Click on Statistics and select "N total" "Percent", "Cumulative percent". OK, OK

   (b) For graphical view of the sample, construct the histogram with the commands:

      Graph¿Histogram¿Choose "Simple", OK¿Enter "C3" in Graph variables¿OK

   (c) To compare with the probability mass function, use the commands:

      Graph¿Scatterplot¿ select "simple"¿ OK¿Enter "C2" for Y and "C1" for X¿OK

Chapter 3

Random Variables and Their Distributions

3.1 Introduction

In Chapter 1 we briefly introduced the concept of a *random variable* as the numerical description of the outcome of a random (probabilistic) experiment which consists of the random selection of a unit from a population of interest (the *underlying population* of the random variable) and the recording of a numerical description of a characteristic(s) of the selected unit. In this chapter we will give a slightly more precise and detailed definition of a random variable. Moreover, we will give concise descriptions of the *probability distribution* of a random variable, or a (sample space) population, and will introduce several *parameters* of the probability distribution that are often used for descriptive and inferential purposes. Probability *models*, and corresponding *model parameters*, will be introduced. The above concepts will be introduced separately for the two main categories of random variables, *discrete* random variables and *continuous* random variables. This chapter focuses on univariate variables. Similar concepts for bivariate and multivariate random variables will be discussed in the next chapter.

3.2 Random Variables

In Section 2.2, it was pointed out that different variables can be associated with the same experiment, depending on the objective of the study and/or preference of the investigator;

see Example 2.2.2 and the note that follows. The following examples reinforce this point and motivate a more precise definition of *random variable*.

Example 3.2.1. Consider the experiment where two fuses are examined for the presence of a defect. Depending on the objective of the investigation the variable of interest can be the number of defective fuses among the two that are examined. Call this random variable X. Thus, the sample space corresponding to X is $\mathcal{S}_X = \{0, 1, 2\}$. Alternatively, the investigation might focus on the whether or not the number of defective items among the two that are examined is zero. Call Y the random variable that takes the value zero when there are zero defective fuses, and the value 1 when there is at least one defective fuse, among the two that are examined. Thus, the sample space that corresponds to Y is $\mathcal{S}_Y = \{0, 1\}$. Note that both X and Y are functions of the outcome in the basic, most detailed, sample space of the experiment, namely $\mathcal{S} = \{\mathcal{DD}, \mathcal{ND}, \mathcal{DN}, \mathcal{NN}\}$.

Example 3.2.2. In accelerated life testing, products are operated under harsher conditions than those encountered in real life. Consider the experiment where one such product is tested until failure, and let X denote the time to failure. The sample space of this experiment, or of X, is $\mathcal{S}_X = [0, \infty)$. However, if the study is focused on whether or not the product lasts more than, say, 1500 hours of operation, the investigator might opt to record a coarser random variable, Y, which takes the value 1 if the product lasts more than 1500 hours and the value 0 if it does not. Thus, the sample space of Y is $\mathcal{S}_Y = \{0, 1\}$. Note that Y is a function of X, which is the most basic and detailed outcome of the experiment.

Example 3.2.3. In the accelerated life testing context of Example 3.2.2, consider the experiment which records the times until failure of each of a sample of 10 products. Thus, the outcome of the experiment consists of the life times of the 10 products, which we denote by $X_1, X_2, \ldots, X_{10}$. However, if the study is focused on the (population) average life time, then an outcome variable that the investigator might opt to record is the (sample) average

$$\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i,$$

which is a function of the basic outcome $X_1, X_2, \ldots, X_{10}$ of the experiment. Alternatively, if the study is focused on the (population) proportion of products that last more than 1500 hours of operation, the investigator might opt to record, as outcome variable, the (sample) proportion

$$\widehat{p} = \frac{1}{10}\sum_{i=1}^{10} Y_i,$$

where Y_i takes the value 1 if the ith product lasts more than 1500 hours of operation (i.e. if $X_i > 1500$), and the value 0 otherwise. Note that the variable $\widehat{p}$ is a function of $Y_1, \ldots, Y_{10}$, and, therefore, also a function of the basic outcome $X_1, X_2, \ldots, X_{10}$ of the experiment.

As illustrated by the above examples, in any given experiment we can define several random variables, all of which are functions of the outcome in the basic, most detailed, sample space. This leads to the following most common definition of a random variable.

Definition 3.2.1. *A* **random variable** *is a rule (or function) that associates a number with each outcome of the basic, most detailed, sample space of a random experiment.*

The variables in Example 3.2.1, as well as the variable Y in Example 3.2.2, are examples of a category of random variables called *discrete* random variables.

Definition 3.2.2. *A* **discrete random variable** *is a random variable whose sample space has a finite or at most countably infinite number of values.*

Following are some additional examples of discrete random variables, which are common and important enough to have been named. They all arise in connection to certain sample inspection experiments.

Example 3.2.4. A quality engineer collects a simple random sample of size n product items, from the production line, which are then inspected for the presence or absence of a certain defect. This is called the **binomial experiment**. The random variable, X, which denotes the number of defective items found, is called a **binomial random variable**. If $n = 1$, X is also called a **Bernoulli random variable**. The sample space of X is $\mathcal{S} = \{0, 1, \ldots, n\}$.

Example 3.2.5. A batch of N items arrives at a distributor. The distributor draws a simple random sample of size n for thorough inspection. This is called the **hypergeometric experiment**. The random variable, X, which denotes the number of defective items found, is called a **hypergeometric random variable**. The sample space of X depends also on the (usually unknown) number of defective items in the batch. Whatever that might be, X cannot take a value outside the set of values $\{0, 1, \ldots, n\}$.

Example 3.2.6. A quality engineer continues to inspect product items, for the presence or absence of a certain defect, until the rth defective item is found. This is called the **negative binomial experiment**. The random variable, X, which denotes the number of items inspected, is called a **negative binomial random variable**. If $r = 1$, X is called a **geometric random variable**. The sample space of X is $\mathcal{S} = \{r, r+1, r+2, \ldots\}$. The reason why this sample space is open-ended on the right is because the probability that X will be greater than any specified number is positive (though this probability will be very small if the specified number is large).

In Example 3.2.2, we saw that the sample space of the variable $X =$ time to failure of a product is $[0, \infty)$. This is an example of a sample space which is infinite, but not *countably* infinite (i.e. it cannot be enumerated), as is that of the negative binomial variable of Example 3.2.6. [As an indication that the numbers in $[0, \infty)$ cannot be enumerated, note that it is impossible to identify the number that comes after 0. Even finite intervals, such as $[0, 1]$ contain uncountably infinite many numbers.]

Definition 3.2.3. *A random variable X is called* **continuous** *if it can take any value within a finite or infinite interval of the real line* $(-\infty, \infty)$.

Thus, the sample space of a continuous random variable has uncountably infinite values. Examples of experiments resulting in continuous variables include measurements of length, weight, strength, hardness, life time, pH or concentration of contaminants in soil or water samples.

Note that, although a continuous variable can take any possible value in an interval, its measured value cannot. This is because no measuring device has infinite precision. Thus, continuous variables do not really exist in real life; they are only ideal versions of the discretized variables which are measured. Nevertheless, the study of continuous random variables is meaningful as it provides useful, and quite accurate, approximations to probabilities pertaining to their discretized versions.

Note further that, if the underlying population of units is finite, there is a finite number of values that a variable can possibly take, regardless of whether it is thought of in its ideal continuous state, or in its discretized state. For example, an experiment investigating the relation between height, weight and cholesterol level of men 55-65 years old, records these three continuous variables for a sample of the aforementioned finite population. Even if we think of these variables as continuous (i.e. non-discretized), the number of different values that each of them can take cannot exceed the number of existing 55-65 year old men. Even in such cases the model of a continuous random variable offers both convenience and accurate approximation of probability calculations.

3.2.1 Exercises

1. The uniting of the stem of one plant with the stem or root of another is called grafting. Suppose that each graft fails independently with probability 0.3. Five grafts are scheduled to be performed next week. Let X denote the number of grafts that will fail next week.

   (a) Give the sample space of X.

   (b) The random variable X is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

2. In the grafting context of the previous exercise, suppose that grafts are done one at a time, and the process continues until the first failed graft. Let Y denote the number of successful grafts until the first failed graft.

   (a) Give the sample space of Y.

   (b) The random variable Y is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

3. Suppose that 8 of the 20 buses in a particular city have developed cracks on the underside of the main frame. Five buses are to be selected for thorough inspection. Let X denote the number of buses (among the five that are inspected) that have cracks.

   (a) Give the sample space of X.

   (b) The random variable X is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

4. A distributor receives a new shipment of 20 ipods. He draws a random sample of five ipods and thoroughly inspects the click wheel of each of them. Suppose that the new shipment of 20 ipods contains three with malfunctioning click wheel. Let X denote the number of ipods with defective click wheel in the sample of five.

   (a) Give the sample space of X.

(b) The random variable X is (choose one)

`(i) binomial     (ii) hypergeometric     (iii) negative binomial`

5. Suppose that 10% of all components manufactured at General Electric are defective. Let X denote the number of defective components among 15 randomly selected ones.

   (a) Give the sample space of X.

   (b) The random variable X is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

6. A letter sent by snail mail is delivered within 3 working days with probability 0.9. You send out 10 letters on Tuesday to invite friends for dinner. Only those who receive the invitation by Friday (i.e., within 3 working days) will come. Let X denote the number of friends who come to dinner.

   (a) Give the sample space of X.

   (b) The random variable X is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

7. In a shipment of 10 electronic components, 2 are defective. Suppose that 5 components are selected at random for inspection, and let X denote the number of defective components found.

   (a) Give the sample space of X.

   (b) The random variable X is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

8. Suppose that 30% of all drivers stop at an intersection having flashing red lights when no other cars are visible. Of 15 randomly chosen drivers coming to an intersection under these conditions, let X denote the number of those who stop.

   (a) Give the sample space of X.

   (b) The random variable X is (choose one)

   `(i) binomial     (ii) hypergeometric     (iii) negative binomial`

9. Three electrical engineers toss coins to see who pays for coffee. If all three match, they toss another round. Otherwise the 'odd person' pays for coffee. Let X denote the number of times the engineers toss coins until odd person results

(a) Give the sample space of X.

(b) The random variable X is (choose one)

`(i) binomial     (ii) hypergeometric     (iii) negative binomial`

3.3 Probability Distributions

When an experiment is performed, the value that the univariate variable of interest, X, takes will be some real number. In other words, the event that X takes a value between $-\infty$ and ∞, $[-\infty < X < \infty]$, happens with probability 1 or, in mathematical notation,

$$P([-\infty < X < \infty]) = 1.$$

We say that we know the **probability distribution** of a univariate random variable if we know the probability with which its value will fall in any given interval. That is, if we know how the total probability of 1 is *distributed* among all subintervals of the real line. A concise description of the probability distribution of a random variable, both discrete and continuous, can be achieved through its *cumulative distribution function* which is defined in the next subsection. In this section we will also give, for each of the two types of random variables, an additional method for achieving a concise description of the probability distribution. For discrete random variables this method will be called *probability mass function*, while for continuous random variables it will be called *probability density function.*

3.3.1 The cumulative distribution function

Definition 3.3.1. *Let X be a univariate, discrete or continuous, random variable associated with some experiment, and let $[X \leq x]$ denote the event that, when the experiment is performed, X takes a value that is less than or equal to x. [Here x is a generic symbol that stands for any real number.] Then the* **cumulative distribution function***, abbreviated by* **cdf***, of X is defined by*

$$F_X(x) = P([X \leq x]).$$

Thus, the cdf is a function defined for all real numbers. When no confusion is possible, it will be denoted simply as $F(x)$, i.e. the subscript X will be omitted. Also, when referring to the cdf as a function, and not in regard to its value at a particular x, we denote it by F_X, or simply by F. Finally, $P([X \leq x])$ will also be denoted by $P(X \leq x)$.

Example 3.3.1. Consider a random variable X with cumulative distribution function given by

$$\begin{aligned}
F(x) &= 0, \quad \text{for all } x \text{ that are less than } 1, \\
F(x) &= 0.4, \quad \text{for all } x \text{ such that } 1 \leq x < 2, \\
F(x) &= 0.7, \quad \text{for all } x \text{ such that } 2 \leq x < 3, \\
F(x) &= 0.9, \quad \text{for all } x \text{ such that } 3 \leq x < 4, \\
F(x) &= 1, \quad \text{for all } x \text{ that are greater than or equal to } 4.
\end{aligned}$$

Plotting $F(x)$ versus x gives the following figure.

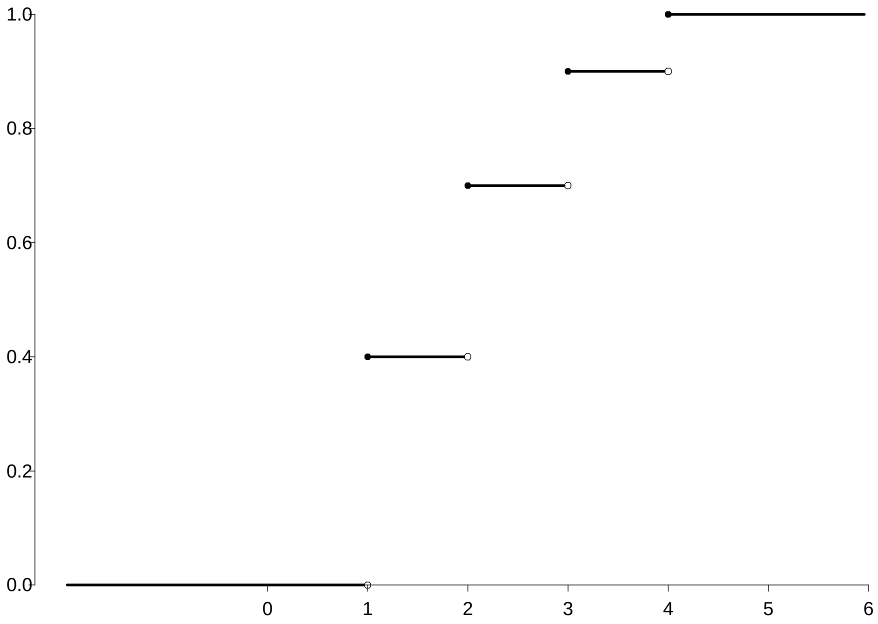


Figure 3.1: The CDF of a Discrete Distribution is a Step or Jump Function

The form of the cdf in Example 3.3.1 implies, through the properties of probability, all relevant information about probabilities relating to the random variable X, as demonstrated in the following example.

Example 3.3.2. Let X have cdf as given in Example 3.3.1. Use the form of the cdf to deduce the distribution of X.

Solution. Since the cdf given in Example 3.3.1 is a jump function, the first deduction is that X is a discrete variable; the second deduction is that the possible values that X can take are the jump points of its cdf, i.e. 1,2,3, and 4; the third deduction is that the probability with which X takes each value equals the size of the jump at that value (for example, $P(X = 1) = 0.4$). We now justify how we arrived at these deduction. First note that the first of the equations defining F implies that

$$P(X < 1) = 0,$$

which means that X cannot a value less than one. Moreover, the second of the equations defining F, in particular the relation $F(1) = 0.4$, implies that X takes the value one with probability 0.4. To see this, note first that the event $[X \leq 1]$ can be written as the union of the disjoint events $[X < 1]$ and $[X = 1]$. Using now the additivity property of probability we obtain that

$$0.4 = F(1) = P([X \leq 1]) = P([X < 1]) + P([X = 1]) = 0 + P([X = 1]) = P([X = 1]).$$

The second of the equations defining F also implies that X cannot take a value in the interval $(1, 2)$ (the notation $(1, 2)$ means that the endpoints, 1, 2, are excluded from the interval). By similar reasoning it can be deduced from the form of F that X is a discrete random variable taking values $1, 2, 3, 4$, with respective probabilities $0.4, 0.3, 0.2, 0.1$. See also Example 3.3.4 for further discussion of this example.

Generally, using the calculus of probability, which was developed in Chapter 2, it is possible to show that knowing the cdf of a random variable X amounts to knowing the probability distribution of X; see part 3 of Proposition 3.3.1 below, where some further properties of the cdf are also listed.

Proposition 3.3.1. *The cumulative distribution function, F, of any random variable X satisfies the following basic properties:*

1. *It is non-decreasing. Thus, for any two real numbers $a \leq b$, we have*

$$F(a) \leq F(b).$$

2. $F(-\infty) = 0,\ F(\infty) = 1.$

3. *If $a,\ b$ are any two real numbers such that $a < b$, then*

$$P([a < X \leq b]) = F(b) - F(a).$$

4. If X is a discrete random variable with sample space $\mathcal{S}$, then its cdf F is a jump or step function with jumps occurring only at the values x that belong in $\mathcal{S}$, while the flat regions of F correspond to regions where X takes no values. Moreover, the size of the jump at each member x of $\mathcal{S}$ equals $P(X = x)$.

Formal proof of the entire proposition will not be given. However, the first property, which can be rephrased as $P(X \leq a) \leq P(X \leq b)$ if $a \leq b$, is quite intuitive: It is more likely that X will take a value which is less than or equal to 10 than one which is less than or equal to 5, since the latter possibility is included in the first. As a concrete example, if we take $a = 1.5$ and $b = 2.5$ then, for the cdf of Example 3.3.1, we have $P(X \leq 1.5) = 0.4$ while $P(X \leq 2.5) = 0.7$. Property 2 is also quite intuitive; indeed the event $[X \leq -\infty]$ never happens and hence it has probability 0, while the event $[X \leq \infty]$ always happens and hence it has probability 1. To see property 3, note first that the event $[X \leq b]$ is the union of the disjoint events $[X \leq a]$ and $[a < X \leq b]$, and apply the additivity property of probability. (Note that property 3 implies property 1.) Finally note that the property 4 is illustrated in Example 3.3.2, in terms of the cdf of Example 3.3.1.

3.3.2 The probability mass function of a discrete distribution

A concise description of the probability distribution of discrete random variable, X, can also be achieved through its *probability mass function*, which gives the probability with which X takes each of its possible values.

Definition 3.3.2. *Let X be a discrete random variable. Then, the function*

$$p(x) = P(X = x),$$

is called the **probability mass function**, *abbreviated* **pmf**, *of the random variable X.*

Let the sample space of X be $\mathcal{S} = \{x_1, x_2, \ldots\}$. It should be clear from the above definition that $p(x) = 0$, for all x that do not belong in $\mathcal{S}$ (i.e. all x different from $x_1, x_2, \ldots$); also, the axioms of probability imply that

$$p(x_i) \geq 0, \quad \text{for all } i, \text{ and } \sum_i p(x_i) = 1. \tag{3.3.1}$$

When X can take only a small number of possible values, its pmf is easily given by listing the possible values and their probabilities. In this case, the *probability histogram* or *probability bar chart* are two ways of graphing a pmf. See Figure 3.2.

Example 3.3.3. In the context of the hypergeometric experiment of Example 3.2.5, suppose that the batch size is $N = 10$, that the batch contains 3 defective items, and that we draw a random sample of size $n = 3$ without replacement. The probabilities that the number of defective items found, X, is 0, 1, 2, and 3 can be calculated as:

$$\begin{aligned} P(X = 0) &= \frac{7}{10}\frac{6}{9}\frac{5}{8} = \frac{7}{24}, \quad P(X = 1) = \frac{3}{10}\frac{7}{9}\frac{6}{8} + \frac{7}{10}\frac{3}{9}\frac{6}{8} + \frac{7}{10}\frac{6}{9}\frac{3}{8} = \frac{21}{40}, \\ P(X = 2) &= 3 \times \frac{3 \times 2 \times 7}{720} = \frac{7}{40}, \quad P(X = 3) = \frac{3 \times 2 \times 1}{720} = \frac{1}{120}. \end{aligned}$$

Thus, the pmf of X is: . Note that property (3.3.1) holds for this pmf. The hypergeometric pmf will be discussed again in Section 3.5 below.

x	0	1	2	3
$p(x)$	0.292	0.525	0.175	0.008

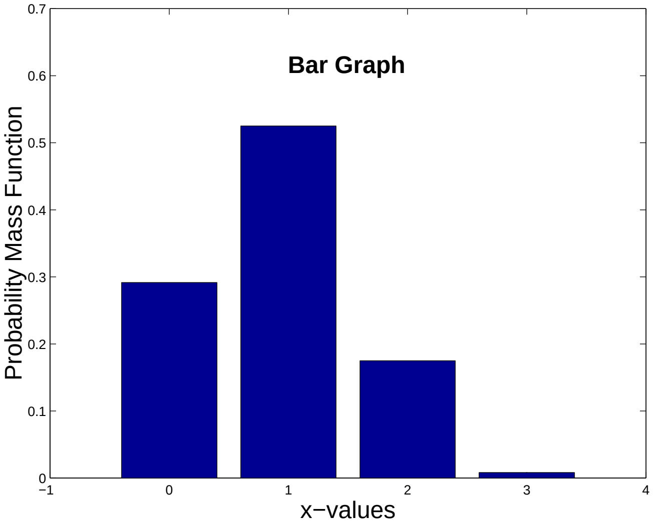


Figure 3.2: Bar Graph for the pmf of Example 3.3.3

If $\mathcal{S}$ consists of a large (or infinite) number of values, using a formula to describe the probability assignments is more convenient.

Example 3.3.4. Consider the negative binomial experiment, described in Example 3.2.6, with $r = 1$. Thus, the inspection of items continues until the first defective item is found. Here the sample space of X, the geometric random variable, is countably infinite and thus it is impossible to give the pmf of X by listing its values and corresponding probabilities as was done in Example 3.3.3. Let p denote the probability that a randomly selected product will be defective; for example p can be 0.01, meaning that 1% of all products are defective. Then, reasoning as we did in Example 2.5.4, we have that the pmf of X is given by the formula

$$p(x) = P(X = x) = (1 - p)^{x-1}p, \quad \text{for} \quad x = 1, 2, \ldots$$

In Example 2.5.4, we used $p = 0.01$, obtained numerical values for some of the above probabilities, and verified that property (3.3.1) holds for this pmf.

Part 4 of Proposition 3.3.1 indicates how the pmf can be obtained from the cdf; see also Example 3.3.1. We now illustrate how the cdf can be obtained from the pmf.

Example 3.3.5. Consider a random variable X with sample space $\mathcal{S} = \{1, 2, 3, 4\}$ and probability mass function

x	1	2	3	4
$p(x)$	.4	.3	.2	.1

The cumulative probabilities are

$$\begin{aligned} F(1) &= P(X \leq 1) = 0.4, \quad F(2) = P(X \leq 2) = 0.7 \\ F(3) &= P(X \leq 3) = 0.9, \quad F(4) = P(X \leq 4) = 1. \end{aligned}$$

Note that the jump-function nature of a cdf of a discrete random variable (see Proposition 3.3.1) imply that it is completely specified by its values for each x in $\mathcal{S}$. For example, $F(1.5) = P(X \leq 1.5) = P(X \leq 1) = F(1)$. In fact, the above cdf is the cdf of Example 3.3.1.

The above example implies that pmf and cdf provide complete, and equivalent, descriptions of the probability distribution of a discrete random variable X. This is stated precisely in the following

Proposition 3.3.2. *Let $x_1 < x_2 < \cdots$ denote the possible values of the discrete random variable X arranged in an increasing order. Then*

1. *$p(x_1) = F(x_1)$, and $p(x_i) = F(x_i) - F(x_{i-1})$, for $i = 2, 3, \cdots$.*

2. *$F(x) = \sum_{x_i \leq x} p(x_i)$.*

3. *$P(a \leq X \leq b) = \sum_{a \leq x_i \leq b} p(x_i)$.*

4. *$P(a < X \leq b) = F(b) - F(a) = \sum_{a < x_i \leq b} p(x_i)$.*

3.3.3 The density function of a continuous distribution

Though the cumulative distribution function offers a common way to describe the distribution of both discrete and continuous random variables, the alternative way offered by the probability mass function applies only to discrete random variables. The corresponding alternative for continuous random variables is offered by the *probability density function.* The reason that a continuous random variable cannot have a pmf is

$$P(X = x) = 0, \text{ for any value } x. \tag{3.3.2}$$

This is a consequence of the fact that the possible values of a continuous random variable X are uncountably many. We will demonstrate (3.3.2) using the simplest and most intuitive random variable, the *uniform in* $[0, 1]$ random variable.

Definition 3.3.3. *Consider selecting a number at random from the interval* $[0, 1]$ *in such a way that any two subintervals of* $[0, 1]$ *of equal length, such as* $[0, 0.1]$ *and* $[0.9, 1]$, *are equally likely to contain the selected number. If* X *denotes the outcome of such a random selection, then* X *called the* **uniform** *in* $[0, 1]$ *random variable; alternatively, we say that* X *has the* **uniform** *in* $[0, 1]$ *distribution; this is denoted by* $X \sim U(0, 1)$.

If $X \sim U(0, 1)$, one can argue that the following statements are true: a) $P(X < 0.5) = 0.5$, b) $P(0.4 < X < 0.5) = 0.1$, c) $P(0.49 < X < 0.5) = 0.01$. In general, it can be argued that

$$P(X \text{ in an interval of length } l) = l. \tag{3.3.3}$$

Since any single number x is an interval of length zero, relation (3.3.3) implies that $P(X = x) = 0$, demonstrating thus (3.3.2). Thus, for any $0 \leq a < b \leq 1$ we have

$$P(a < X < b) = P(a \leq X \leq b) = b - a. \tag{3.3.4}$$

Recalling that we know the probability distribution of a random variable if we know the probability with which its value will fall in any given interval, relation (3.3.3), or (3.3.4), implies that we know the distribution of a uniform in $[0, 1]$ random variable. Since $b - a$ is also the area under the constant curve at height one above the interval $[a, b]$, an alternative way of describing relation (3.3.4) is through the constant curve at height one. This is the simplest example of a *probability density function.*

Definition 3.3.4. *The* **probability density function**, *abbreviated* **pdf**, f, *of a continuous random variable* X *is a nonnegative function* ($f(x) \geq 0$, *for all* x), *with the property that* $P(a < X < b)$ *equals the area under it and above the interval* (a, b). *Thus,*

$$P(a < X < b) = \begin{cases} \textit{area under } f_X \\ \textit{between } a \textit{ and } b. \end{cases}$$

Since the area under a curve is found by integration, we have

$$P(a < X < b) = \int_a^b f_X(x)dx. \tag{3.3.5}$$

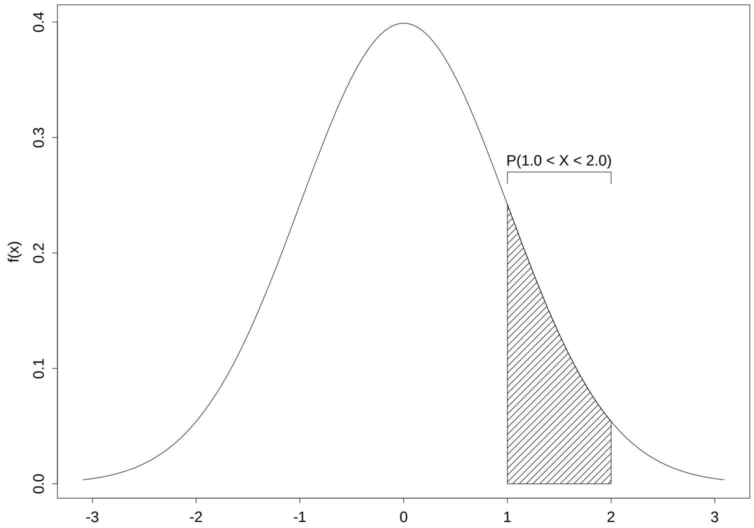


Figure 3.3: $P(a < X < b)$ Equals the Area under the Curve, Above (a, b)

It can be argued that for any continuous random variable X, $P(X = x) = 0$, for any value x. The argument is very similar to that given for the uniform random variable. In particular,

$$P(a < X < b) = P(a \leq X \leq b). \tag{3.3.6}$$

Remark 3.3.1. Relation (3.3.6) is true only for the idealized version of a continuous random variable. As we have pointed out, in real life, all continuous variables are measured on a discrete scale. The pmf of the discretized random variable which is actually measured is readily approximated from the formula

$$P(x - \Delta x \leq X \leq x + \Delta x) \approx 2f_X(x)\Delta x,$$

where Δx denotes a small number. Thus, if Y denotes the discrete measurement of the continuous random variable X, and if Y is measured to three decimal places with the usual rounding, then

$$P(Y = 0.123) = P(0.1225 < X < 0.1235) \approx f_X(0.123)(0.001).$$

Moreover, breaking up the interval $[a, b]$ into n small subintervals $[x_k - \Delta x_k, x_k + \Delta x_k]$, $k = 1, \ldots, n$, we have

$$P(a \leq Y \leq b) = \sum_{k=1}^{n} 2f_X(x_k)\Delta x_k.$$

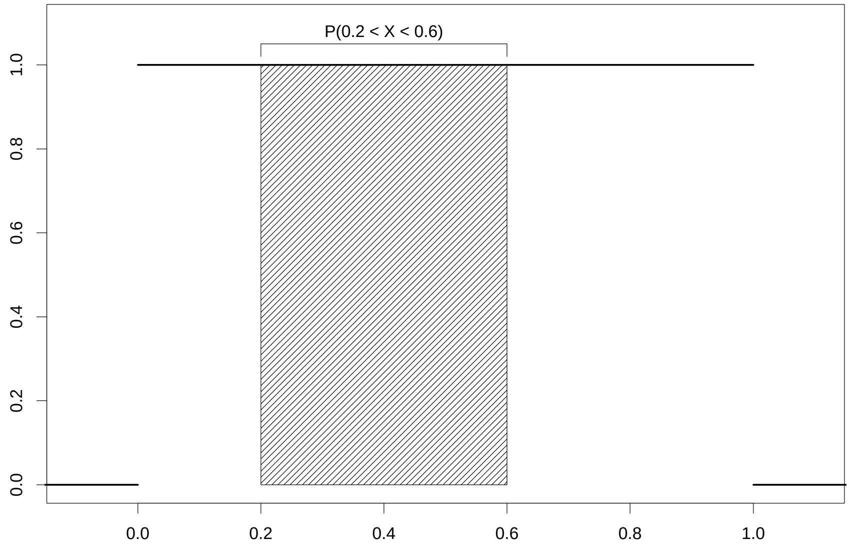


Figure 3.4: PDF of the Uniform in (0,1) Distribution

Since the summation on the left approximates the integral $\int_a^b f_X(x)dx$, the above confirms the approximation

$$P(a \leq X \leq b) \approx P(a \leq Y \leq b),$$

namely, that the distribution of the discrete random variable Y is approximated by that of its idealized continuous version.

Relation (3.3.5) implies that the cumulative distribution function can also be obtained by integrating the pdf. Thus, we have

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy. \tag{3.3.7}$$

By a theorem of calculus, we also have

$$F'(x) = \frac{d}{dx}F(x) = f(x), \tag{3.3.8}$$

which derives the pdf from the cdf. Finally, relations (3.3.5), (3.3.6) and (3.3.7) imply

$$P(a < X < b) = P(a \leq X \leq b) = F(b) - F(a). \tag{3.3.9}$$

Thus, for continuous random variables, we do not need to worry as to whether or not the inequalities are strict when using the cdf for calculating probabilities.

Example 3.3.6. Let $X \sim U(0,1)$, i.e. X has the uniform in $[0,1]$ distribution. The pdf of X is (see Figure 3.4)

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x > 1 \end{cases}$$

The cdf of X is

$$F(x) = \int_0^x f(y)dy = \int_0^x dy = x,$$

for $0 \le x \le 1$, and $F(x) = P(X \le x) = 1$, for $x \ge 1$. See Figure 3.5

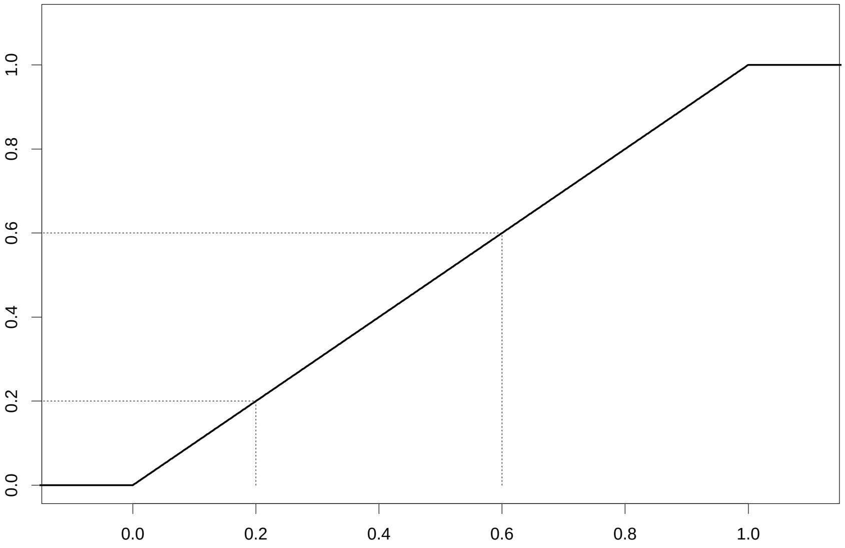


Figure 3.5: CDF of the Uniform in (0,1) Distribution

Example 3.3.7. A random variable X is said to have the uniform distribution in the interval $[A, B]$, denoted $X \sim U(A, B)$, if it has pdf

$$f(x) = \begin{cases} 0 & \text{if } x < A \\ \frac{1}{B-A} & \text{if } A \le x \le B \\ 0 & \text{if } x > B. \end{cases}$$

Find $F(x)$.

Solution: Note first that since $f(x) = 0$, for $x < A$, we also have $F(x) = 0$, for $x < A$. This and relation (3.3.7) imply that for $A \leq x \leq B$,

$$F(x) = \int_A^x \frac{1}{B-A} dy = \frac{x-A}{B-A}.$$

Finally, since $f(x) = 0$, for $x > B$, it follows that $F(x) = F(B) = 1$, for $x > B$.

Example 3.3.8. A random variable X is said to have the **exponential distribution** with parameter λ, also denoted $X \sim \exp(\lambda)$, $\lambda > 0$, if $f(x) = \lambda e^{-\lambda x}$, for $x > 0, f(x) = 0$, for $x < 0$. Find $F(x)$.

Solution:

$$F(x) = \int_{-\infty}^x f(y)dy = \int_0^x f(y)dy = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}.$$

Figure 3.6 presents plots of the pdf of the exponential distribution for different values of the parameter λ.

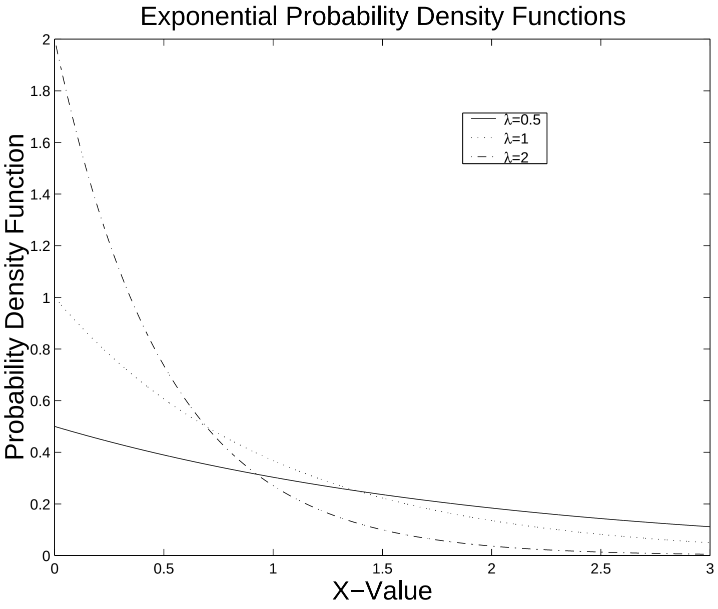


Figure 3.6: PDFs of Three Exponential Distributions

Example 3.3.9. Let T denote the life time of a randomly selected component, measured in hours. Suppose $T \sim \exp(\lambda)$ where $\lambda = 0.001$. Find the $P(900 < T < 1200)$.

Solution: Using (3.3.5),

$$P(900 < T < 1200) = \int_{900}^{1200} .001 e^{-.001x} dx$$
$$= e^{-.001(900)} - e^{-.001(1200)} = e^{-.9} - e^{-1.2} = .1054$$

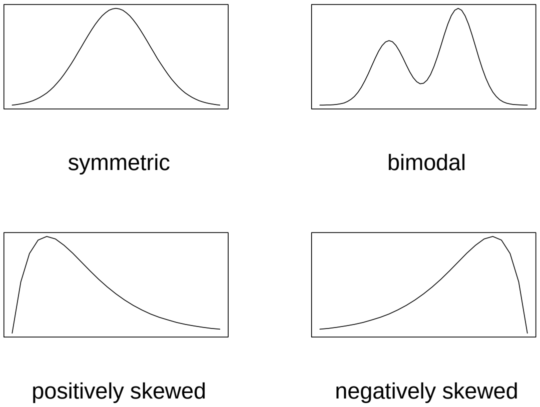


Figure 3.7: Typical Shapes of PDFs

Using the closed form expression of the cdf,

$$P(900 < T < 1200) = F_T(1200) - F_T(900)$$
$$= \left[1 - e^{-(.001)(1200)}\right] - \left[1 - e^{-(.001)(900)}\right] = .1054.$$

We close this section by giving names to some typical shapes of probability density functions, which are presented in the following figure. A **positively skewed** distribution is also called **skewed to the right**, and a **negatively skewed** distribution is also called **skewed to the left**. See Figure 3.7

3.3.4 Exercises

1. An unfair die with six sides is rolled. Let X be outcome of the die. The probability mass function is given to be P(X=i)=i/21 , i=1,2,....6.

   (a) Show that this is a legitimate pmf.

   (b) Find and plot the cumulative distribution function.

2. Put a check-mark in the table below indicating whether or not each of $p_1(x)$, $p_2(x)$, $p_3(x)$ is a legitimate probability mass function (pmf) for the random variable X.

x	0	1	2	3	Is legitimate	Is not legitimate
$p_1(x)$	0.2	0.3	0.4	0.2		
$p_2(x)$	0.3	0.3	0.5	-0.1		
$p_3(x)$	0.1	0.4	0.4	0.1		

3. A metal fabricating plant currently has five major pieces under contract each with a deadline for completion. Let X be the number of pieces completed by their deadlines. Suppose that X is a random variable with p.m.f. $p(x)$ given by

x	0	1	2	3	4	5
$p(x)$	0.05	0.10	0.15	0.25	0.35	0.10

   (a) Find and plot the cdf of X.

   (b) Use the cdf to find the probability that between one and four pieces, inclusive, are completed by deadline.

4. Let X denote the daily sales for a computer manufacturing firm. The cumulative distribution function of the random variable X is

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.2 & 0 \le x < 1 \\ 0.7 & 1 \le x < 2 \\ 0.9 & 2 \le x < 3 \\ 1 & 3 \le x \end{cases}$$

   (a) Plot the cumulative distribution function. What is the probability of two or more sales in a day?

   (b) Write down the probability mass function of X.

5. The cumulative (probability) distribution function of checkout duration in a certain supermarket is

$$F(x) = \frac{x^2}{4}\ , \quad \text{for } x \text{ between 0 and 2 .}$$

(Thus, $F(x) = 0$ for $x \le 0$ and $F(x) = 1$ for $x > 2$.)

   (a) Find the probability that the duration is between 0.5 and 1.

   (b) Find the density function $f(x)$.

6. In the context of the above Exercise 5, let Y denote the checkout duration measured in seconds.

   (a) Find the cumulative distribution function $F_Y(y)$ of Y. (Hint: $Y \le y$ holds if and only if $X \le y/60$.)

   (b) Find the density function $f_Y(y)$ of Y.

7. Let X denote the amount of time for which a statistics reference book, on two-hour reserve at the engineering library, is checked out by a randomly selected student. Suppose that X has density function

$$f(x) = \begin{cases} 0.5x & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the probability that the book is checked out between 0.5 and 1.5 hours?

8. Let X denote the resistance of a randomly chosen resistor, and suppose that its pdf is given by

$$f(x) = \begin{cases} kx & \text{if } 8 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find k.

   (b) Give a formula for the cdf of X.

   (c) Find $P(8.6 \leq X \leq 9.8)$.

   (d) Find the conditional probability that $X \leq 9.8$ given that $X \geq 8.6$

9. Let X be a continuous random variable with a standard exponential density, that is, $f(x) = e^{-x}$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$. Find the following probabilities.

   (a) $P(-1 < X \leq 1)$

   (b) $P(X \geq 1)$

   (c) $P(X > 2 \mid X > 1)$

10. Let X denote the life time of a component when measured in hours, and assume that X has an exponential distribution.

    (a) The cumulative distribution function of an exponential random variable with parameter λ is $F(x) = 1 - exp(-\lambda x)$. Give an expression (involving λ) for the probability $P(2 < X < 5)$.

    (b) Let now Y denote the same life time but in minutes instead of hours. Give an expression for the cumulative distribution function of Y. (Hint: $Y \leq y$ holds if and only if $X \leq y/60$.)

    (c) Find the probability density function of Y.

11. The time X in hours for a certain plumbing manufacturer to deliver a custom made fixture is a random variable with pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-\theta)} & \text{if } x \geq \theta \\ 0 & \text{otherwise,} \end{cases}$$

with $\lambda = 0.02$, $\theta = 48$. An architect overseeing a renovation must order a custom made piece to replace an old fixture which unexpectedly broke. The architect has determined that it will cost him 200 dollars a day for every day beyond three days that he does not have the piece.

(a) What is the probability that, if he orders the piece from that manufacturer, he will lose no money?

(b) What is the probability that he will lose no more that 400 dollars?

(c) What is the probability that he will lose between 400 and 800 dollars?

12. Plumbing suppliers typically ship packages of plumbing supplies containing many different combinations of pipes, sealants, drains, etc. Almost invariably there are one or more parts in the shipment that are not correct: the part may be defective, missing, not the one that was ordered, etc. In this question the random variable of interest is the proportion P of parts in a shipment, selected at random, that are not correct.

A family of distributions for modeling a random variable P, which is a proportion, has the probability density function

$$f_P(p) = \theta p^{\theta-1}, \quad 0 < p < 1, \quad \theta > 0.$$

(a) Find the cdf of P, in terms of the parameter θ.

(b) Assume that $\theta = 0.07$. What proportion of shipments have a proportion of incorrect parts less than 0.2?

13. The life time, X, of certain equipment in some units is believed to follow the probability density function

$$f(x) = (1/\theta^2)xe^{-x/\theta}, \quad x > 0, \quad \theta > 0.$$

(a) Find the cdf of X.

(b) What proportion of equipment have a life time less than 25 time units?

3.4 Parameters of a Univariate Distribution

The pmf and cdf provide complete descriptions of the probability distribution of a discrete random variable, and their graphs can help identify key features of its distribution regarding the *location* (by this we mean the most ”typical” value, or the ”center” of the range of values, of the random variable), variability, and shape. Similarly, the pdf and cdf of a continuous random variable provide a complete description of its distribution. In this section we will introduce certain *summary parameters* that are useful for describing/quantifying the prominent features of the distribution of a random variable. The parameters we will consider are the *mean value*, also referred to as the *average value* or *expected value*, and the *variance*, along with its close relative, the *standard deviation.* For continuous random variables, we will consider, in addition, *percentiles* such as the *median* which are also commonly used as additional parameters to describe the location, variability and shape of a continuous distribution. Such parameters of the distribution of a random variable are also referred to, for the sake of simplicity, as parameters of a random variable, or parameters of a (statistical) population.

3.4.1 Discrete random variables

Expected value

In Chapter 1, Section 1.6.2, we defined the *population average* or *population mean* for a finite population. In this subsection, we will expand on this concept and study its properties.

Definition 3.4.1. *Consider a finite population of N units and let $v_1, \ldots, v_N$ denote the values of the variable of interest for each of the N units. Then, the* **population average** *or* **population mean value** *is defined as*

$$\mu = \frac{1}{N} \sum_{i=1}^{N} v_i. \tag{3.4.1}$$

If X denotes the outcome of taking a simple random sample of size one from the above population and recording the unit value, then μ of relation (3.4.1) is also called the **expected value** *of X. The expected value of X is also denoted by $E(X)$ or μ_X.*

Example 3.4.1. Suppose that the population of interest is a batch of $N = 100$ units received by a distributor, and that ten among them have some type of defect. A defective unit is indicated by 1, while 0 denotes no defect. In this example, each v_i, $i = 1, 2, \ldots, 100$, is either 0 or 1. Since

$$\sum_{i=1}^{100} v_i = (90)(0) + (10)(1) = 10,$$

the population average is

$$\mu = \frac{1}{100} \sum_{i=1}^{100} v_i = \frac{10}{100} = p,$$

where p denotes the proportion of defective items. If X denotes the state (i.e. 0 or 1) of a unit selected by simple random sampling from this batch, then its expected value is also

$$E(X) = p.$$

In preparation for the next proposition, let $x_1, \ldots, x_m$ denote the distinct values among the values $v_1, \ldots, v_N$ of the statistical population. Also, let n_j, denote the number of times that the distinct value x_j is repeated in the statistical population; in other words, n_j is the number of population units that have value x_j. For example, in Example 3.4.1 $m = 2$, $x_1 = 0$, $x_2 = 1$, $n_1 = 90$ and $n_2 = 10$. Finally, let X denote the outcome of one simple random selection from the above statistical population; thus the sample space of X is $\mathcal{S} = \{x_1, \ldots, x_m\}$.

Proposition 3.4.1. *In the notation introduced in the preceding paragraph, the expected value of the random variable X, as well as the average of its underlying (statistical) population, are also given by*

$$\mu_X = \sum_{j=1}^{m} x_j p_j,$$

where $p_j = n_j/N$ is the proportion of population units that have the value x_j, $j = 1, 2, \ldots, m$.

The proof of this proposition follows by noting that $\sum_{i=1}^{N} v_i = \sum_{j=1}^{m} n_j x_j$. Therefore, the expression for μ (which is also μ_X) in Definition 3.4.1 can also be written as

$$\mu_X = \sum_{j=1}^{m} x_j \frac{n_j}{N},$$

which is equivalent to the expression given in the proposition, since $p_j = n_j/N$. Proposition

3.4.1 is useful in two ways. First, taking a *weighted* average of the (often much fewer) values in the sample space is both simpler and more intuitive than averaging the values of the underlying statistical population, especially when that population is a large hypothetical one. Second, this proposition affords an abstraction of the random variable in the sense that it disassociates it from its underlying population, and refers it to an equivalent experiment whose underlying population coincides with the sample space. This abstraction will be very useful in Chapter 5, where we will introduce models for probability distributions. Because of the simplicity of the formula given in Proposition 3.4.1, the general definition of expected value of a random variable, which is given in Definition **??** below, is an extension of it, rather than of the formula in Definition 3.4.1.

Example 3.4.2. Consider the population of Example 3.4.1. Using Proposition 3.4.1, we can simplify the calculation of the population mean (or of the expected value of X) as follows: Let $x_1 = 0, x_2 = 1$ denote the sample space of X (or, equivalently, the distinct values among the $N = 100$ values of the statistical population), and let $p_1 = n_1/N = 0.9$, $p_2 = n_2/N = 0.1$ denote the proportions of x_1, x_2, respectively, in the statistical population. Then, according to Proposition 3.4.1,

$$\mu = x_1p_1 + x_2p_2 = x_2p_2 = p_2 = 0.1.$$

Note that p_2 is what we called p in Example 3.4.1.

We now give the definition of expected value for an arbitrary discrete random variable, and the mean value of its underlying population.

Definition 3.4.2. *Let X be an arbitrary discrete random variable, and let $\mathcal{S}$ and $p(x) = P(X = x)$ denote its sample space and probability mass function, respectively. Then, the* **expected value***, $E(X)$ or μ_X, of X is defined as*

$$\mu_X = \sum_{x \text{ in } \mathcal{S}} xp(x).$$

The **mean value** *of an arbitrary discrete population is the same as the expected value of the random variable it underlies.*

Note that this definition generalizes Definition 3.4.1 as it allows for a random variables with an infinite sample space. We now demonstrate the calculation of the expected value with some examples.

Example 3.4.3. Roll a die and let X denote the outcome. Here $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ and $p(x) = 1/6$ for each member x of $\mathcal{S}$. Thus,

$$\mu_X = \sum_{x \ in \ \mathcal{S}} xp(x) = \sum_{i=1}^{6} \frac{i}{6} = \frac{21}{6} = 3.5.$$

Note that rolling a die can be thought of as taking a simple random sample of size one from the population $\{1, 2, 3, 4, 5, 6\}$, in which case Definition 3.4.1 can also be used.[1]

Example 3.4.4. Select a product from the production line and let X take the value 1 or 0 as the product is defective or not. Note that the random variable in this experiment is similar to that of Example 3.4.2 except for the fact that the population of all products is infinite and conceptual. Thus, the approach of Example 3.4.1 cannot be used for finding the expected value of X. Letting p denote the proportion of all defective items in the conceptual population of this experiment, we have

$$E(X) = \sum_{x \ in \ \mathcal{S}} xp(x) = 1p + 0(1 - p) = p,$$

which is the same answer as the one we obtained in Example 3.4.2, when the latter is expressed in terms of the proportion of defective products.

Example 3.4.5. Consider the experiment of Example 3.3.4, where the quality engineer keeps inspecting products until the first defective item is found, and let X denote the number of items inspected. This experiment can be thought of as random selection from the conceptual population of all such experiments. The statistical population corresponding to this hypothetical population of all such experiments is $\mathcal{S} = 1, 2, 3, \ldots$. As it follows from Example 3.3.4, the proportion of population units (of experiments) with (outcome) value $x_j = j$ is $p_j = (1-p)^j p$, where p is the probability that a randomly selected product is defective. Application of Definition 3.4.2 yields, after some calculations, that

$$\mu_X = \sum_{x \ in \ \mathcal{S}} xp(x) = \sum_{j=1}^{\infty} x_j p_j = \frac{1}{p}.$$

We finish this subsection by giving some properties of expected values, which have to do with the computation of the expected value of a function $Y = h(X)$ of a random variable X.

[1] As a demonstration of the flexibility of Proposition 3.4.1, we consider the die rolling experiment as a random selection from the conceptual population of all rolls of a die. The i-th member of this population, $i = 1, 2, 3 \ldots$, has value v_i which is either $x_1 = 1$ or $x_2 = 2$ or $x_3 = 3$ or $x_4 = 4$ or $x_5 = 5$, or $x_6 = 6$. The proportion of members having the value x_j is $p_j = 1/6$, for all $j = 1, \ldots, 6$, and thus Proposition 3.4.1 applies again to yield the same answer for μ_X.

Proposition 3.4.2. *Let X be a random variable with sample space $\mathcal{S}$ and pmf $p_X(x) = P(X = x)$. Also, let $h(x)$ be a function on $\mathcal{S}$ and let Y denote the random variable $Y = h(X)$. Then*

1. *The expected value of $Y = h(X)$ can be computed using the pmf of X as*

$$E(Y) = \sum_{x \ in \ \mathcal{S}} h(x)p_X(x).$$

2. *(MEAN VALUE OF A LINEAR TRANSFORMATION) If the function $h(x)$ is linear, i.e. $h(x) = ax + b$, then*

$$E(Y) = E(aX + b) = aE(X) + b.$$

Example 3.4.6. A bookstore purchases 3 copies of a book at \$6.00 each, and sells them for \$12.00 each. Unsold copies will be returned for \$2.00. Let X be the number of copies sold, and Y be the net revenue. Thus $Y = h(X) = 12X + 2(3 - X) - 18$. If the pmf of X is

x	0	1	2	3
p$_X$(x)	.1	.2	.2	.5

then it can be seen that the pmf of Y is

y	-12	-2	8	18
p$_Y$(y)	.1	.2	.2	.5

Thus $E(Y)$ can be computed using the definition of the expected value of a random variable (Definition 3.4.2):

$$E(Y) = \sum_{all \ y \ values} yp_Y(y) = (-12)(.1) + (-2)(.2) + (.8)(.2) + (18)(.5) = 9.$$

However, using part 1 of Proposition 3.4.2, $E(Y)$ can be computed without first computing the pmf of Y:

$$E(Y) = \sum_{all \ x \ values} h(x)p_X(x) = 9.$$

Finally, since $Y = 10X - 12$ is a linear function of X, part 2 of Proposition 3.4.2 implies that $E(Y)$ can be computed by only knowing $E(X)$. Since $E(X) = \sum_x xp_X(x) = 2.1$, we have $E(Y) = 10(2.1) - 12 = 9$, the same as before.

Variance and standard deviation

The variance of a random variable X, or of its underlying population, indicates/quantifies the extent to which the values in the statistical population differ from the expected value of X. The variance of X is denoted either by σ_X^2, or $\mathrm{Var}(X)$, or simply by σ^2 if no confusion is possible. For a finite underlying population of N units, with corresponding statistical population $v_1, \ldots, v_N$, the population variance is defined as

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \mu_X)^2,$$

where μ_X is the mean value of X. Though this is neither the most general definition, nor the definition we will be using for calculating the variance of random variables, its form demonstrates that the population variance is the average squared distance of members of the (statistical) population from the population mean. As it is an average square distance, it goes without saying that, the variance of a random variable can never be negative. As with the formulas pertaining to the population mean value, averages over the values of the statistical population can be replaced by weighted averages over the values of the sample space of X. In particular, if $x_1, \ldots, x_m$ are the distinct values in the sample space of X and $p_j = n_j/N$ is the proportion of population units that have the value x_j, $j = 1, 2, \ldots, m$, then the population variance is also given by

$$\sigma_X^2 = \sum_{j=1}^{m}(x_j - \mu_X)^2 p_j.$$

The general definition of the variance of a discrete random variable allows for countably infinite sample spaces, but is otherwise similar to the above weighted average expression.

Definition 3.4.3. *Let X be any discrete random variable with sample space $\mathcal{S}$ and pmf $p(x) = P(X = x)$. Then the* **variance** *of X, or of its underlying population, is defined by*

$$\sigma^2 = \sum_{x \ in \ \mathcal{S}} (x - \mu)^2 p(x),$$

where p_j is the proportion of population units that have the value x_j, $j = 1, 2, \ldots$.

Some simple algebra reveals the simpler-to-use formula (also called the *short-cut formula* for σ^2)

$$\sigma^2 = \sum_{x \ in \ \mathcal{S}} x^2 p(x) - \mu^2 = E(X^2) - [E(X)]^2.$$

Definition 3.4.4. *The positive square root, σ, of σ^2 is called the* **standard deviation** *of X or of its distribution.*

Example 3.4.7. Suppose that the population of interest is a batch of $N = 100$ units received by a distributor, and that 10 among them have some type of defect. A defective unit is indicated by 1, while 0 denotes no defect. Thus, the statistical population consists of v_i, $i = 1, 2, \ldots, 100$, each of which is either 0 or 1. Let X denote the outcome of a simple random selection of a unit from the batch. In Example 3.4.1 we saw that $E(X) = p$, where p denotes the proportion of defective items. To use the short-cut formula for the variance, it should be noted that here $X^2 = X$ holds by the fact that X takes only the value of 0 or of 1. Thus, $E(X^2) = E(X) = p$, so that

$$\sigma^2 = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p).$$

Example 3.4.8. Roll a die and let X denote the outcome. As in Example 3.4.3, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ and $p(x) = 1/6$ for each member x of $\mathcal{S}$, and $\mu = 3.5$. Thus,

$$\sigma^2 = E(X^2) - \mu^2 = \sum_{j=1}^{6} x_j^2 p_j - \mu^2 = \frac{91}{6} - 3.5^2 = 2.917.$$

Example 3.4.9. Consider the experiment of Example 3.4.5, where the quality engineer keeps inspecting products until the first defective item is found, and let X denote the number of items inspected. Thus, the experiment can be thought of as random selection from the conceptual population of all such experiments. As it follows from Example 3.3.4, the proportion of members having the value $x_j = j$ is $p_j = (1-p)^j p$, $j = 1, 2, \ldots$, where p is the proportion of defective product items in the conceptual population of all product. Also from Example 3.4.5 we have that $\mu_X = 1/p$. Application of the short-cut formula and some calculations yield that

$$\sigma^2 = \sum_{j=1}^{\infty} x_j^2 p_j - \mu^2 = \frac{1-p}{p^2}.$$

We conclude this subsection by providing a result about the variance of a linear transformation of a random variable, and by demonstrating its use.

Proposition 3.4.3. *(VARIANCE OF A LINEAR TRANSFORMATION) Let X be a random variable with variance σ_X^2, and let $Y = a + bX$ be a linear transformation of X. Then*

$$\sigma_Y^2 = b^2 \sigma_X^2, \quad \text{and} \quad \sigma_Y \;=\; |b| \sigma_X.$$

Example 3.4.10. Consider Example 3.4.6, where a bookstore purchases 3 copies of a book at \$6.00 each, sells each at \$12.00 each, and returns unsold copies for \$2.00. X is the number of copies sold, and $Y = 10X - 12$ is the net revenue. Using the pmf of X given in Example 3.4.6, we find that

$$\sigma_X^2 = 0.2 + 4 \times 0.2 + 9 \times 0.5 - 2.1^2 = 1.09, \ \sigma_X = 1.044.$$

Using Proposition 3.4.3 and the above information we can compute the variance and standard deviation of Y without making use of its pmf:

$$\sigma_Y^2 = 10^2 \sigma_X^2 = 109, \quad \text{and} \quad \sigma_Y = 10.44.$$

Note that the variance and standard deviation of Y is identical to those of $Y_1 = -10X$ and of $Y_2 = -10X + 55$.

3.4.2 Continuous random variables

As in the discrete case, the *average*, or *mean value*, and the *variance* of a population are concepts that are used synonymously with the *mean* or *expected value*, and the *variance* of the corresponding random variable (or, more precisely, of the distribution of the random variable), respectively. [By random variable corresponding to a population, we mean the outcome of a simple random selection from that population.] Thus, we will define directly the expected value and variance of a random variable. Moreover, we will define the *median* and other *percentiles* as additional population parameters of interest.

Expected Value and Variance

Definition 3.4.5. *The* **expected value** *and* **variance***, respectively, of a continuous random variable X with probability density function $f(x)$ are defined by*

$$\begin{aligned} E(X) &= \mu_X = \int_{-\infty}^{\infty} x f(x) dx, \\ Var(X) &= \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, \end{aligned}$$

provided the integrals exists. As in the discrete case, the **standard deviation***, σ_X, is the positive square root of the variance σ_X^2.*

The approximation of integrals by sums, as we saw in Subsection 3.3.3, helps connect the definitions of expected value for discrete and continuous variables. The next proposition asserts that the properties of mean value and variance of continuous random variables are similar to those of discrete variables.

Proposition 3.4.4. *For a continuous random variable X we have*

1. *The expected value of a function $Y = h(X)$ of X can be computed, without first finding the pdf of Y, as*

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx$$

2. *The short-cut version of the variance is, as before,*

$$\sigma_X^2 = E(X^2) - [E(X)]^2.$$

3. *The expected value, variance and standard deviation of a linear function $Y = a+bX$ of X are, as in the discrete case,*

$$E(a+bX) = a + bE(X), \quad \sigma^2_{a+bX} \quad = \quad b^2\ \sigma_X^2, \quad \textit{and} \quad \sigma_{a+bX} = |b|\ \sigma_X.$$

Example 3.4.11. Let $X \sim U(0,1)$, i.e. X has the uniform in $[0,1]$ distribution (see Example 3.3.6), and show that $E(X) = 0.5$ and $\text{Var}(X) = 1/12$.

Solution. We have $E(X) = \int_0^1 xdx = 0.5$. Moreover, $E(X^2) = \int_0^1 x^2dx = 1/3$, so that $Var(X) = 1/3 - 0.5^2 = 1/12$.

Example 3.4.12. Let $Y \sim U(A,B)$, i.e. Y has the uniform in $[A,B]$ distribution (see Example 3.3.7), and show that $E(Y) = (B+A)/2$, $Var(Y) = (B-A)^2/12$.

Solution. This computation can be done with the use of the pdf for the uniform in $[A,B]$ random variable, which was given in Example 3.3.7. But here we will use the fact that if $X \sim U(0,1)$, then

$$Y = A + (B-A)X \sim U(A,B).$$

Using this fact, the results of Example 3.4.11, and part 3 of Proposition 3.4.4, we have

$$\begin{aligned} E(Y) &= A + (B-A)E(X) = A + \frac{B-A}{2} = \frac{B+A}{2}, \\ Var(Y) &= (B-A)^2 Var(X) = \frac{(B-A)^2}{12}. \end{aligned}$$

Example 3.4.13. Let $X \sim Exp(\lambda)$, i.e. X has the exponential distribution with parameter $\lambda > 0$ (see Example 3.3.8), and show that $E(X) = 1/\lambda$, and $\sigma_X^2 = 1/\lambda^2$.

Solution. We have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \frac{1}{\lambda}, \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2\lambda e^{-\lambda x}dx = -x^2 e^{-\lambda x}\big|_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x}dx = \frac{2}{\lambda^2}. \end{aligned}$$

Thus,

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

The plots of the pdf of the three exponential distributions that were presented in Example 3.3.8, suggest that the smaller λ gets the more probability is allocated to large positive numbers. Thus, it is reasonable that both the expected value and the variance of an exponential distribution will increase as λ decreases. In particular, for $\lambda = 0.5,\ 1$, and 2, the three values of λ that were used in the pdf and cdf plots of Example 3.3.8, the corresponding mean values are

$$E_{\lambda=0.5}(X) = 2, \quad E_{\lambda=1}(X) = 1, \quad E_{\lambda=2}(X) = 0.5,$$

and the corresponding variances are

$$\text{Var}_{\lambda=0.5}(X) = 4, \quad \text{Var}_{\lambda=1}(X) = 1, \quad \text{Var}_{\lambda=2}(X) = 0.25.$$

For exponential distributions, the standard deviation equals the mean value.

Percentiles (or Quantiles)

We now proceed with the definition of some additional population parameters that are of interest in the continuous case.

Definition 3.4.6. *Let X be a continuous random variable with cdf F. Then the* **median** *of X, or of the distribution of X, is defined as the number $\tilde{\mu}_X$ with the property*

$$F(\tilde{\mu}_X) = P(X \leq \tilde{\mu}_X) = 0.5,$$

The median also has the property that it splits the area under the pdf of X in two equal parts, so the probability of X taking a value less than $\tilde{\mu}_X$ is the same as that of taking a value greater than it.

Example 3.4.14. Suppose $X \sim Exp(\lambda)$, so $f(x) = \lambda e^{-\lambda x}$, for $x > 0$. Find the median of X.

Solution. According to its definition, the median is found by solving the equation

$$F(\tilde{\mu}) = 0.5.$$

Since $F(x) = 1 - e^{-\lambda x}$ (see Example 3.3.8), the above equation becomes

$$1 - e^{-\lambda \tilde{\mu}} = 0.5,$$

or $e^{-\lambda \tilde{\mu}} = 0.5$, or $-\lambda \tilde{\mu} = ln(0.5)$, or

$$\tilde{\mu} = -\frac{ln(0.5)}{\lambda}.$$

In symmetric distributions, the population mean and median coincide but otherwise they differ. In positively skewed distributions the mean is larger than the median, but in negatively skewed distributions the opposite is true.

Definition 3.4.7. *Let α be a number between 0 and 1. The* **100(1-α)th percentile** *(or* **quantile***) of a continuous random variable X is the number, denoted by x_α, with the property*

$$F(x_\alpha) = P(X \leq x_\alpha) = 1 - \alpha.$$

Thus, the 95th percentile, denoted by $x_{0.05}$, of a random variable X separates the units of the underlying population with characteristic value in the upper 5% of values from the rest. For example, if the height of a newborn is in the 95th percentile, it means that only 5% of newborns are taller than it and 95% are shorter than it. In this terminology, the median is the 50th percentile and can also be denoted as $x_{0.5}$. The 25th percentile is also called the **lower quartile** and denoted q_1, while the 75th percentile is also called the **upper quartile** and denoted q_3. In the spirit of this terminology and notation, the median is also called the middle quartile and denoted by q_2.

Like the median, the percentiles can be found by solving the equation which defines them, i.e. by solving $F(x_\alpha) = 1 - \alpha$ for x_α. For the exponential distribution with mean value $1/\lambda$, this equation becomes $1 - exp(-\lambda x_\alpha) = 1 - \alpha$ or

$$x_\alpha = -\frac{ln(\alpha)}{\lambda}.$$

Thus, the 95th percentile of the exponential random variable T of Example 3.3.9 is $t_{0.05} = -ln(0.05)/0.001 = 2995.73$, whereas its median is $t_{0.5} = -ln(0.5)/0.001 = 693.15$. Recall that, according to the formula of Example 3.4.13, the expected value of T is $\mu_T = 1/0.001 = 1000$. Note that the exponential distribution is positively skewed and thus the expected value should be larger than the median.

Measures of Location and Spread Based on Percentiles

The percentiles, in addition to being measures of location, in the sense of identifying points of interest of a continuous distribution, can also be used to define quantifications of spread (or variability) of a distribution. Such measures of spread or variability are defined as the distance between selected percentiles can serve as a measure of variability. The most common such measure of spread is defined below.

Definition 3.4.8. *The* **interquartile range***, abbreviated by* **IQR***, is the distance between the 25th and 75th percentile (or between the upper and lower quartiles).*

Example 3.4.15. In this example, we present the mean, the 25th, 50th and 75th percentiles, and the interquartile range of the exponential distributions corresponding to the three values of λ that were used in the pdf and cdf plots of Example 3.3.8 ($\lambda = 0.5,\ 1$, and 2). Recall that in Example 3.4.13 we saw that the standard deviation coincides with the mean in the exponential distribution.

λ	0.5	1	2
$\mu = \sigma$	2	1	0.5
$x_{0.75}$	0.5754	0.2877	0.1438
$x_{0.5}$	1.3862	0.6931	0.3466
$x_{0.75}$	2.7726	1.3863	0.6932
IQR	2.1972	1.0986	0.5494

This table provides another illustration of the fact that, for the exponential distribution, the median is smaller than the mean, as it should be for all positively skewed distributions. In addition, the table demonstrates that the standard deviation and the IQR, two measures of variability of the distribution, decrease as λ increases.

Remark 3.4.1. It can be seen from the above table that the ratio IQR/σ remains constant as the value of λ changes. This is true more generally. In particular, IQR shares the property of σ that IQR_{a+bX}, the IQR of $a + bX$, equals $|b|IQR_X$. Thus, for any r.v. X, the ratio IQR_X/σ_X will not change if X gets multiplied by a constant.

In the following two sections we introduce the most common models for probability distributions of discrete and continuous random variables, respectively.

3.4.3 Exercises

1. A metal fabricating plant currently has five major pieces under contract each with a deadline for completion. Let X be the number of pieces completed by their deadlines. Suppose that X is a random variable with p.m.f. $p(x)$ given by

x	0	1	2	3	4	5
$p(x)$	0.05	0.10	0.15	0.25	0.35	0.10

   (a) Compute the expected value and variance of X.

   (b) For each contract completed before the deadline, the plant receives a bonus of \$15,000. Find the expected value and variance of the total bonus amount.

2. Let X denote the daily sales for a computer manufacturing firm. The probability mass function of the random variable X is

x	0	1	2	3
$p(x)$	0.2	0.5	0.2	0.1

   (a) Calculate the expected value and variance of X.

   (b) Suppose the firm makes \$1,200 per sale. Thus $Y = 1,200 \times X$ is the daily revenue in dollars. Find the mean value and variance of Y.

3. An unfair die with six sides is rolled. Let X be outcome of the die. The probability mass function is given to be P(X=i)=i/21 , i=1,2,....6.

   (a) Calculate $E(X)$ and $E(X^2)$.

   (b) In a win-win game, the player is offered

$$\$4,000 \times X \quad \text{or} \quad \$1000 \times X^2,$$

   where the random variable X has the distribution given above. When the choice is made a value of X is generated and the player receives the chosen price. Which choice would you recommend the player to make?

4. Let X have pmf

x	1	2	3	4
$p(x)$	0.4	0.3	0.1	0.2

(a) Calculate $E(X)$ and $E(1/X)$.

(b) In a win-win game, the player will win a monetary price, but he/she has to decide between a fixed and a random price. In particular the player is offered

$$\frac{\$1000}{E(X)} \quad \text{or} \quad \frac{\$1000}{X},$$

where the random variable X has the distribution given above. When the choice is made a value of X is generated and the player receives the chosen price. Which choice would you recommend the player to make?

5. The cumulative (probability) distribution function of X, the checkout duration in a certain supermarket measured in minutes, is

$$F(x) = \frac{x^2}{4}\ , \quad \text{for } x \text{ between 0 and 2 .}$$

(Thus, $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x > 2$.)

(a) Find the median checkout duration.

(b) Find the interquartile range.

(c) Find the expected value and the standard deviation of X.

6. In the context of the above Exercise 5, let Y denote the checkout duration measured in seconds.

(a) Find the cumulative distribution function $F_Y(y)$ of Y.

(b) Find the median and the interquartile range of Y.

(c) Find the expected value and the standard deviation of Y.

7. Let X denote the amount of time that a statistics reference book on a two-hour reserve at the engineering library is checked out by a randomly selected student. Suppose that X has density function

$$f(x) = \begin{cases} 0.5x & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the cumulative distribution function $F_X(x)$ of X.

(b) Find the 25th, 50th and 75th percentile of the checkout duration.

(c) Find the expected value and the standard deviation of X.

8. Let X denote the life time of a component when measured in hours, and assume it has the exponential distribution with parameter λ. Thus, the pdf of X is $f(x) = \lambda exp(-\lambda x)$, and its cdf is $F(x) = 1 - exp(-\lambda x)$.

   (a) Using the examples in the text, give the mean value and variance of X.

   (b) Find the expected value and variance of Y.

9. Let X denote the life time (in some units) of a randomly selected component from an assembly line, and suppose that the pdf of X is

$$f(x) = \begin{cases} \lambda e^{-\lambda(x-5)} & \text{if } x \geq 5 \\ 0 & \text{otherwise.} \end{cases},$$

   for some $\lambda > 0$. Express in terms of λ the following quantities:

   (a) The expected value and variance of X.

   (b) The 25th, the 50th and the 75th percentiles of X.

10. Plumbing suppliers typically ship packages of plumbing supplies containing many different combinations of pipes, sealants, drains, etc. Almost invariably there are one or more parts in the shipment that are not correct: the part may be defective, missing, not the one that was ordered, etc. In this question the random variable of interest is the proportion P of parts in a shipment, selected at random, that are not correct.

    A family of distributions for modeling a random variable P, which is a proportion, has the probability density function

$$f_P(p) = \theta p^{\theta-1}, \quad 0 < p < 1, \quad \theta > 0.$$

    (a) Find the expectation of P, in terms of the parameter θ.

    (b) Find the variance of P, in terms of the parameter θ.

    (c) Find the cdf of P, in terms of the parameter θ.

    (d) Find the 25th, the 50th and the 75th percentiles of P, in terms of the parameter θ.

11. The random variable Y has a probability density given by the following function:

$$f(y) = \frac{1}{y\ln(5)} \text{ if } 1 \le y \le 5$$

where ln is the natural logarithm, $f(y) = 0$ if $y < 1$ or $y > 5$.

    (a) Find the cdf of Y.

    (b) Find the mean and variance of Y.

    (c) What is the 25th, the 50th and the 75th percentiles of Y.

12. The life time, X, of certain equipment in some units is believed to follow the probability density function

$$f(x) = (1/\theta^2)xe^{-x/\theta}, \;\; x > 0, \;\; \theta > 0.$$

    (a) Find the expectation of X.

    (b) Find the variance of X.

    (c) Find the cdf of X.

    (d) Find the 25th, the 50th and the 75th percentiles of X.

3.5 Models for Discrete Random Variables

In this section we develop *models* for the probability distributions of random variables. Each model will pertain to an entire class of random variables which share a common, up to a some *parameters*, probabilistic behavior. The population parameters, such as the expected value and variance of the random variable in any given experiment, can be expressed in terms of the *model parameters* of the assumed model. In parametric statistical inference, the model parameters are the focus of inference.

The advantage of studying classes of random variables is seen by considering the simplest random variable, which is one that takes only two values and is called Bernoulli random variable, as already defined in Example 3.2.4. The two values can always be re-coded to 0 and 1, so the sample space of a Bernoulli random variable will always be taken to be $\mathcal{S} = \{0, 1\}$. The random variables Y in Examples 3.2.1 and 3.2.2 are examples of Bernoulli random variables. The random variables X in Examples 3.2.4 and 3.2.5 are also Bernoulli when $n = 1$. These examples illustrate the wide variety of Bernoulli experiments, i.e. of experiments that give rise to a Bernoulli random variable. The advantage of modeling is that the probabilistic structure of any Bernoulli random variable follows from that of the **prototypical** one. In the prototypical Bernoulli experiment, X takes the value 1 with probability p and the value 0 with probability $1-p$. Thus, the pmf of X is

x	0	1
$p(x)$	$1-p$	p

and its cdf is

x	0	1
$F(x)$	$1-p$	1

We have already seen that the expected value and variance of a Bernoulli random variable X are

$$\mu_X = p, \quad \sigma_X^2 = p(1-p).$$

The pmf and cdf of any specific Bernoulli random variable is of the above form for some value of the *model parameter* p. Thus, if X takes the value 0 or 1 as a randomly selected item is defective or non-defective, respectively, and the probability of a non-defective item is 0.9, the pmf and cdf of X follow from those of the prototypical pmf by replacing p by 0.9, i.e.

x	0	1
$p(x)$	.1	.9

In Section 3.2 we already saw three main classes of random variables, namely the *Binomial*, the *Hypergeometric*, and the *Negative Binomial* (see Examples 3.2.4, 3.2.5, and 3.2.6, respectively). In this section we will give the pmf, the expected value, and the variance, for the prototypical variable in each of the above classes of random variables. An additional class of discrete random variables, the *Poisson*, will also be introduced.

3.5.1 The Binomial Distribution

Suppose that n independent Bernoulli experiments are performed, and that each experiment results in 1 with the same probability p. The total number, X, of experiments resulting in 1 is a **binomial** random variable with parameters n and p. Thus the possible values of X are $0, 1, \cdots, n$.

Example 3.5.1. Consider the random variable X of Example 3.2.4. Thus, X denotes the number of defective products in a sample of n products selected from the production line. Here each inspected product corresponds to a Bernoulli trial, or experiment. For X to be binomial (as we called it), the model assumptions of independence and equal probability for 1 in each Bernoulli trial must hold. In such experiments (i.e. where items are inspected as they come off the production line), the model assumptions are usually realistic, but they would be violated if, due to some malfunction in the production process, there are clusters of defective products.

Example 3.5.2. Suppose that n product items are placed in accelerated life testing and let X denote the number of those that last more than 1500 hours of operation. (In Example 3.2.3, where $n = 10$, this number was expressed as $\sum_{i=1}^{10} Y_i$.) Here, each product item that is placed in accelerated life testing corresponds to a Bernoulli trial (either the product item lasts more than 1500 hours of operation or not). Then, X is a binomial random variable provided that the model assumptions of independence and equal probability for 1 in each Bernoulli trial holds. [Would you worry about the validity of the model assumptions if the life tests were conducted in two different labs? What if the product items tested came from two different production facilities?]

The pmf $p(x) = P(X = x)$ of a binomial random variable X, or the binomial distribution, with parameters n, p is denoted by $b(x; n, p)$. It can be found in the following way: By the assumption of independence, the probability that the n Bernoulli trials result in any sequence with x 1's is $p^x(1-p)^{n-x}$. The total number of such sequences is the total number of ways that x of the n trials can be selected (the x selected will get assigned 1, while the others will get assigned 0). This number, of course, is $\binom{n}{x}$. Thus,

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The the cdf $F(x) = P(X \leq x)$ of a binomial random variable X with parameters n, p is denoted by $B(x; n, p)$. There is no closed form expression for $B(x; n, p)$, but Table A.1 of the appendix gives the cdf for $n = 5, 10, 20, 25$ and selected values of p.

The mean value and variance of a binomial X with parameters n, p are

$$E(X) = np, \ \sigma_X^2 = np(1-p).$$

For $n = 1$ we have the mean and variance of the Bernoulli random variable, which were derived in Examples 3.4.4 and 3.4.7, respectively.

Example 3.5.3. Suppose 70% of all purchases in a certain store are made with credit card. Let X denote the number of credit card uses in the next 10 purchases. Find a) the expected value and variance of X, and b) the probability that $P(5 \le X \le 8)$.

Solution. It seems reasonable to assume that $X \sim \text{Bin}(10, 0.7)$. Thus,

$$E(X) = np = 10(0.7) = 7, \;\; \sigma_X^2 = 10(0.7)(0.3) = 2.1.$$

Next, using part 3 of Proposition 3.3.1 and Table A.1 we have

$$\begin{aligned} P(5 \le X \le 8) &= P(4 < X \le 8) = F(8) - F(4) \\ &= 0.851 - 0.047 = 0.804. \end{aligned}$$

Alternatively, this probability can be calculated as

$$P(5 \le X \le 8) \;=\; P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8),$$

but this seems more labor intensive.

3.5.2 The Hypergeometric Distribution

The hypergeometric model applies to situations where a simple random sample of size n is taken without replacement from a finite population of N units of which M are labeled 1 and the rest are labeled 0. The number X of units labeled 1 in the sample is a **hypergeometric** random variable with parameters n, M and N. The random variable X of Example 3.2.5, is an example of a hypergeometric random variable.

Since all $\binom{N}{n}$ groups of size n are equally likely to be selected, the pmf of a hypergeometric random variable X, or the hypergeometric distribution, is easily seen to be

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

As in the binomial case we have $P(X = x) = 0$ if $x > n$. In addition we now have that $P(X = x) = 0$ if $x > M$ (i.e. the sample cannot contain more 1's than the population) or if $n - x > N - M$ (i.e. the sample cannot contain more 0's than the population). It can be shown that the expected value and variance of a hypergeometric random variable X are

$$\mu_X = n\frac{M}{N}, \;\; \sigma_X^2 = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}.$$

Remark 3.5.1. Note that n successive selections from a population of N units, M of which are labeled 1 and the rest 0, are successive Bernoulli experiments with probability for 1 equal to $p = M/N$ for each of them. Thus, the hypergeometric random variable is the sum of n successive Bernoulli random variables. It differs from the binomial random variable in that the Bernoulli trials are not independent. Note, however, that the formula for the expected value of the hypergeometric random variable coincides with that of a binomial random variable with parameters n and $p = M/N$. Thus, the dependence of the Bernoulli trials affects only the variance formula.

Example 3.5.4. 12 refrigerators have been returned to the distributor because of a high-pitched oscillating noise. Suppose that 4 of the 12 have a defective compressor and the rest less serious problems. 6 refrigerators are selected at random for problem identification. Let X be the number of those found with defective compressor. Find the expected value and variance of X and the probability $P(X = 3)$.

Solution. Here $N = 12,\ n = 6,\ M = 4$. Thus, the possible values of X are 2, 3, 4, 5, 6.

$$P(X = 3) \quad = \quad \frac{\binom{4}{3}\binom{8}{3}}{\binom{12}{6}} = 0.3232$$

Also,

$$E(X) \quad = \quad n\frac{M}{N} = 6\frac{4}{12} = 2, \quad V(X) = 6\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\frac{12-6}{12-1} = \frac{8}{11}.$$

Example 3.5.5. (*The Capture/Recapture Method*) The capture/recapture method is sometimes used to estimate the size N of a wildlife population. Suppose that 10 animals are captured, tagged and released. On a later occasion, 20 animals are captured, and the hypergeometric model is used for the number X of them that are tagged. (The model is appropriate if all $\binom{N}{20}$ possible groups are equally likely; with appropriate care, this might be almost true.) Clearly X is more likely to take small values if N is large. The precise form of the hypergeometric model for the pmf of X can be used to estimate N from the value that X takes. This example will be discussed again in the Chapter on point estimation.

As remarked earlier, the hypergeometric random variable is the sum of n successive Bernoulli random variables with probability of 1 equal to $p = M/N$ for each of them. In this respect, it is similar to the binomial random variable. However, the independence assumption of the binomial model is not satisfied when sampling from finite populations; thus the two classes of distributions differ. It is clear, however, that if the population size N is large and the sample size n small, the dependence of the Bernoulli trials will be weak. Thus, the binomial model will be a good approximation to the hypergeometric one. This is a useful approximation because binomial probabilities are easier to work with. The approximation is also evident from the formulas for the expected value and variance. Indeed, if we set $p = M/N$, then the hypergeometric expected value is like that of the binomial, while σ_X^2 differs from that of the binomial by $\frac{N-n}{N-1}$, which is close to 1 if N is large and n is small. The factor $\frac{N-n}{N-1}$ is called **finite population correction factor**.

A rule of thumb for applying the binomial approximation to hypergeometric probabilities is as follows: Let X be a hypergeometric random variable with parameters n, M, N. Let Y be a binomial random variable with parameters n and $p = \frac{M}{N}$. Then if $\boxed{\frac{n}{N} \leq 0.05, \text{ or } 0.1}$

$$P(X = x) \simeq P(Y = x)$$

Example 3.5.6. (*Binomial Approximation to Hypergeometric Probabilities.*) A shipment of $N = 20$ electronic toys contains 5 that malfunction. If $n = 10$ of them are randomly chosen what is the probability that 2 of the 10 will be defective?

Solution. The correct model here is the hypergeometric. It gives

$$P(X = 2) = \frac{\binom{5}{2}\binom{15}{8}}{\binom{20}{10}} = 0.348.$$

Application of the binomial($n = 10, p = 0.25$) approximation gives $P(Y = 2) = 0.282$. In this case N is not very large and n is not sufficiently small compared to N. So the approximation is not good. However, if n remained 10, while $N = 100$ and $M = 25$ (so p remains 0.25), we have

$$P(X = 2) = \frac{\binom{25}{2}\binom{75}{8}}{\binom{100}{10}} = 0.292,$$

so the binomial probability of 0.282 provides a better approximation to this hypergeometric probability.

3.5.3 The Geometric and Negative Binomial Distributions

Assume, as we did for the binomial distribution, that independent Bernoulli trials are being performed, and that each trial results in 1 with the same probability p. A **geometric** experiment is one where the Bernoulli trials continue until the occurrence of the first 1. The geometric random variable X is the total number of trials up to and including the first

1. The quality inspection experiment described in Example 3.2.6 includes, as a special case, an example of a geometric experiment.

Clearly, $X = x$ means that there are $x - 1$ 0's followed by a 1. Using the independence of the Bernoulli trials we arrive at the following formula for the pmf of the geometric distribution with parameter p (see also Example 3.3.4):

$$p(x) = P(X = x) = (1 - p)^{x-1}p, \quad x = 1, 2, 3, \ldots$$

The cdf also has a closed form expression:

$$\begin{aligned} F(x) &= \sum_{y \le x} p(y) = p \sum_{y=1}^{x} (1 - p)^{y-1} \\ &= 1 - (1 - p)^x, \ x = 1, 2, 3, \cdots \end{aligned}$$

where the last equality follows from the formula for the partial sums of a geometric series: $\sum_{y=0}^{x} a^y = (1 - a^{x+1})/(1 - a)$.

It can be shown that for a geometric random variable X,

$$E(X) = \frac{1}{p}, \quad \text{and} \quad Var(X) = \frac{1 - p}{p^2}.$$

The **negative binomial** distribution arises as a generalization of the geometric distribution. The negative binomial experiment terminates at the occurrence of the r-th 1, and the negative binomial random variable X is the total number of trials performed. The quality inspection experiment described in Example 3.2.6 is an example of a geometric experiment. The pmf of the negative binomial distribution with parameters r and p is

$$p(x) = \binom{x - 1}{r - 1} p^r (1 - p)^{x-r}, \quad x = r, r + 1, r + 2, \ldots$$

To see this argue as follows: Any particular outcome sequence has r 1's and $x - r$ 0's and thus has probability $p^r(1 - p)^{x-r}$. Because the last trial is a 1, the remaining $r - 1$ 1's can be assigned to the remaining $x - 1$ trials in $\binom{x-1}{r-1}$ ways.

Remark 3.5.2. The outcome recorded in a negative binomial experiment is often $Y = X - r$, i.e. the total number of 0's. Thus, the possible values of Y are $0, 1, 2, \cdots$ and its pmf is

$$P(Y = y) = \binom{y + r - 1}{r - 1} p^r (1 - p)^x$$

3.5.4 The Poisson Distribution

The Poisson model and its applications

The Poisson distribution models the distribution of the number of occurrences, X, of some event in a given time interval (or a given area, or space). It is an appropriate model for counts of events which occur randomly in time or space. One of the earliest uses of the Poisson distribution was in modeling the number of alpha particles emitted from a radioactive source during a given period of time. The following example lists some additional applications of the Poisson model.

Example 3.5.7. Applications of the Poisson model include:

a) In the analysis of telephone systems to model the number of calls incoming into an exchange in a given time period.

b) In the insurance industry to model the number of freak accidents such as falls in the shower in a given time period.

c) In traffic engineering to model the number of vehicles that pass a marker on a roadway in a given time period.

d) In forestry to model the distribution of trees is a forest.

e) In astronomy to model the distribution of galaxies in a given region of the sky.

As seen from these applications of the Poisson model, the random phenomena that the Poisson distribution models differ from those of the previously discussed distributions in that they are not outcomes of sampling experiments from a well understood population. Consequently, the Poisson pmf is derived by arguments that are different from the ones used for deriving the pmfs of the previously discussed distributions (which were based on counting techniques and independence). Such arguments are beyond the scope of this course.

The Poisson family of distributions is described by a single parameter λ, which takes only positive values ($\lambda > 0$). To indicate that X is a Poisson random variable, i.e. that it has the Poisson distribution, with parameter λ we write $X \sim$ Poisson(λ). If $X \sim$ Poisson(λ), its pmf is

$$P(X = x) \quad = \quad e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Since $e^\lambda = \sum_{k=0}^{\infty}(\lambda^k/k!)$, the probabilities sum to 1. The parameter λ specifies the 'average' number of occurrences in the given interval (of time, area or space). In particular, it can be shown that for the Poisson(λ) random variable X,

$$\mu_X = \lambda \text{ and } \sigma_X^2 = \lambda.$$

Thus, if a random variable has the Poisson distribution, then its expected value equals its variance.

The cdf is given in Table A.2. An illustration of its use follows

Example 3.5.8. Let $X \sim$ Poisson(8), i.e. X has the Poisson distribution with parameter $\lambda = 8$. Find: a) $P(X \leq 5)$, b) $P(6 \leq X \leq 9)$, and c) $P(X \geq 10)$.

Solution. For a) we have $P(X \leq 5) = F(5) = 0.191$. For b) write

$$\begin{aligned} P(6 \leq X \leq 9) &= P(5 < X \leq 9) = P(X \leq 9) - P(X \leq 5) \\ &= F(9) - F(5) = 0.717 - 0.191. \end{aligned}$$

Finally, for c) write

$$P(X \geq 10) = 1 - P(X \leq 9) = 1 - F(9) = 1 - 0.717.$$

Requirements for the appropriateness of the Poisson model

It should be emphasized that the Poisson model does not apply to all events that occur randomly in time or space. One important requirement for the appropriateness of the Poisson model is that the *rate* with which events occur remains constant in time (area or space). For example, phone calls might be coming in at a higher rate during certain peak hours, in which case the Poisson model would not provide a good approximation to the distribution of the number of incoming calls during a period that encompasses both peak and off-peak hours. Another requirement is that the events under study occur *independently*. A final requirement is that these events are *rare*, in a technical sense of the word, i.e. that the probability of two events occurring simultaneously is negligible.

The following example, illustrates the relevance of these requirements in the context of applications of the Poisson model mentioned in Example 3.5.7.

Example 3.5.9. *a*) If the telephone exchange services a large number of customers, then the assumption of constant rate of incoming phone calls seems tenable (at least for certain intervals of time), and so is the requirement the phone calls occur independently. Finally, though phone calls are not considered rare in telephone exchange service, the assumption that two or more calls come in at exactly the same time is indeed negligible, and thus they are rare according to our technical definition. Thus, the Poisson model is expected to provide a reasonable approximation to the distribution of the number of incoming phone calls.

b) If the population is large, then the requirement of constant rate of occurrence of freak accidents seems tenable (at least for certain intervals of time), and so is the requirement the such accidents occur independently. Thus, the Poisson model is expected to provide a reasonable approximation to the distribution of the number of freak accidents during certain intervals of time.

c) If traffic is light, then individual vehicles act independently of each other, but in heavy traffic one vehicle's movement may influence another's, so the approximation might not be good.

d), *e*) Whether or not the distribution of trees or galaxies in a forest or space segment, respectively, satisfy the assumptions of the Poisson distribution can itself be a research question.

Poisson approximation to Binomial probabilities

We will now demonstrate, with a heuristic argument, that the Poisson pmf can be derived as the limit of a binomial pmf. To fix ideas, suppose that the random variable X, whose distribution is Poisson(λ), counts the events that occur in a particular interval of time (or area or space). Imagine that the interval is divided into a very large number of very small subintervals of equal length (area or volume). Because events occur at a constant rate in time (or area, or space), it follows that the probability of an event occurring in a subinterval is the same for all subintervals. Because events occur independently, occurrence in one subinterval does not affect the chances for an occurrence in another subinterval. Finally, because events are rare, in the technical sense of the word, the probability of two or more occurrences in any given subinterval is negligible. Let X_i denote the number of occurrences in the ith subinterval. Because the probability of X_i taking a value other than 0 or 1 is negligible, X_i is approximately Bernoulli. These approximately Bernoulli random variables are independent, the probability of 1 is the same for all, and thus their sum has approximately a Binomial distribution. Their sum, however, equals the Poisson random variable X, revealing a connection between the two distributions. Note that in the limit, as the length of each subinterval shrinks to zero, the probability of more than 1 occurrence in each subinterval also shrinks to zero, and thus each X_i becomes exactly Bernoulli. This demonstrates that Poisson probabilities can be derived as the limit of corresponding binomial probabilities.

The connection between the two distributions that was just established, implies that certain Binomial probabilities can be approximated by corresponding Poisson probabilities. This important result is stated as a proposition.

Proposition 3.5.1. *A binomial experiment where the number of trials, n, is large ($n \geq 100$), the probability of a 1 in each trial, p, is small ($p \leq 0.01$), and the product np is not large ($np \leq 20$), can be modeled (to a good approximation) by a Poisson distribution with $\lambda = np$. In particular, if $Y \sim$ Bin(n, p), with $n \geq 100$, $p \leq 0.01$, and $np \leq 20$, then*

$$P(Y \geq k) \simeq P(X \geq k), \quad k = 0, 1, 2, \ldots, n,$$

where $X \sim$ Poisson$(\lambda = np)$.

Example 3.5.10. Due to a serious defect a car manufacturer issues a recall of $n = 10,000$ cars. Let $p = 0.0005$ be the probability that a car has the defect, and let Y be the number of defective cars. Find: (a) $P(Y \geq 10)$, and (b) $P(Y = 0)$.

Solution. Here each car represents a Bernoulli trial, with outcome 1 if the car has the defect and 0 if it does not have the defect. Thus, Y is a Binomial random variable with $n = 10,000$, and $p = 0.0005$. Note that the three conditions on n and p, mentioned in Proposition 3.5.1, for the approximation of the Binomial probabilities by corresponding Poisson probabilities, are satisfied. Thus, let $X \sim$ Poisson$(\lambda = np = 5)$. For part (a) write

$$P(Y \geq 10) \simeq P(X \geq 10) = 1 - P(X \leq 9) = 1 - 0.968,$$

where the last equality follows from the Poisson Tables. Similarly, for part (b) write

$$P(Y = 0) \simeq P(X = 0) = e^{-5} = 0.007.$$

The Poisson process

Often we want to record the number of occurrences continuously as they increase with time. Thus we record

$$X(t) \quad = \quad \text{number of occurrences in the time interval } [0, t]. \tag{3.5.1}$$

Definition 3.5.1. *The number of occurrences as a function of time,* $X(t), \quad t \geq 0$*, is called a* **Poisson process***, if the following assumptions are satisfied.*

1. *The probability of exactly one occurrence in a short time period of length* Δt *is approximately* $\alpha \Delta t$*.*
2. *The probability of more than one occurrence in a short time period of length* Δt *is approximately 0.*
3. *The number of occurrences in* Δt *is independent of the number prior to this time.*

Note that the assumptions stated in Definition 3.5.1 restate, in a more formal way, the requirements for the appropriateness of the Poisson model. The parameter α in the first assumption specifies the 'rate' of the occurrences, i.e. the 'average' number of occurrences per unit of time; moreover, the first assumption implies that the rate is constant in time. The second assumption restates that the events are rare, in the technical sense of the word. Finally, the third assumption restates the requirement of independence of the occurrences. Because in an interval of length t_0 time units we would expect, on average, $\alpha \times t_0$ occurrences, we have the following result.

Proposition 3.5.2. *a) If* $X(t), \quad t \geq 0$ *is a Poisson process, then for each fixed* t_0*, the random variable* $X(t_0)$*, which counts the number of occurrences in* $[0, t_0]$*, has the Poisson distribution with parameter* $\lambda = \alpha \times t_0$*. Thus,*

$$P(X(t_0) = k) = e^{-\alpha t_0} \frac{(\alpha t_0)^k}{k!}, \ k = 0, 1, 2, \cdots \tag{3.5.2}$$

b) If $t_1 < t_2$ *are two positive numbers and* $X(t), \quad t \geq 0$ *is a Poisson process, then the random variable* $X(t_2) - X(t_1)$*, which counts the number of occurrences in* $[t_1, t_2]$*, has the Poisson distribution with parameter* $\lambda = \alpha \times (t_2 - t_1)$*. Thus, the pmf of* $X(t_2) - X(t_1)$ *is given by (3.5.2) with* t_0 *replaced by* $(t_2 - t_1)$*.*

Example 3.5.11. Continuous electrolytic inspection of a tin plate yields on average 0.2 imperfections per minute. Find

a) The probability of one imperfection in three minutes.

Solution. Here $\alpha = 0.2$, $t = 3$, $\lambda = \alpha t = 0.6$. Thus,

$$P(X(3) = 1) \ = \ F(1; \lambda = 0.6) - F(0; \lambda = 0.6) = .878 - .549 = .329.$$

b) The probability of at least two imperfections in five minutes.

Solution. Here $\alpha = 0.2$, $t = 5$, $\lambda = \alpha t = 1.0$. Thus,

$$P(X(5) \geq 2) \;=\; 1 - F\big(1; \lambda = 1.0\big) = 1 - .736 = .264.$$

c) The probability of at most one imperfection in 0.25 hours.

Solution. Here $\alpha = 0.2$, $t = 15$, $\lambda = \alpha t = 3.0$. Thus,

$$P(X(15) \leq 1) = F\big(1; \lambda = 3.0\big) = .199.$$

We close this section with two interesting properties of the Poisson process.

Proposition 3.5.3. *Let $X(t)$ be a Poisson process with rate α, and let T denote the time until the first occurrence. Then the pdf of T is $f_T(t) = \alpha e^{-\alpha t}$, i.e. it has the exponential distribution.*

Proof. We will first find $1 - F_T(t) = P(T > t)$. (The function $S_T(t) = 1 - F_T(t)$ is also known as the *survival function* of the random variable T.)

$$\begin{aligned} P(T > t) &= P\big(\text{No occurrences in } (0, t)\big) \\ &= P\big(X(t) = 0\big) = e^{-\alpha t}. \end{aligned}$$

Thus $F_T(x) = P(T \leq x) = 1 - e^{-\alpha x}$, and

$$f_T(x) \;=\; F'_T(x) = \alpha e^{-\alpha x}$$

so T is exponential with $\lambda = \alpha$.

Note that any *inter-arrival* time, i.e. the time between any two successive occurrences, has the same distribution as T of Proposition 3.5.3. The next proposition shows that if we condition on the number of occurrences in a given interval, the time points of occurrence of the events are uniformly distributed in that interval. This justifies the notion that Poisson events occur randomly.

Proposition 3.5.4. *Let $X(t)$ be a Poisson process with rate α, and suppose that there are n occurrences in the time period $[0, 1]$. Then, for any $0 < t < 1$, the conditional probability that there are $m \leq n$ occurrences in $[0, t]$ is $\binom{n}{m} t^m (1 - t)^{n-m}$.*

Note: If $X_1, \dots, X_n$ are iid $U(0,1)$ random variables, then the probability that exactly m of them fall in the interval $[0,t]$ is also $\binom{n}{m} t^m (1-t)^{n-m}$.

Proof of Proposition. We have

$$\begin{aligned} & P(X(t)=m|X(1)=n) \\ = & \frac{P(X(t)=m, X(1)-X(t)=n-m)}{P(X(1)=n)}. \end{aligned}$$

Since $X(t), X(1)-X(t)$ are independent, the numerator is

$$e^{-\alpha t}\frac{(\alpha t)^m}{m!} e^{-\alpha(1-t)}\frac{(\alpha(1-t))^{n-m}}{(n-m)!}$$

and the denominator is $e^{-\alpha}\alpha^n/n!$. The result of the proposition follows by division.

3.5.5 Exercises

1. Grafting, the uniting of the stem of one plant with the stem or root of another, is widely used commercially to grow the stem of one variety that produces fine fruit on the root system of another variety with a hardy root system. For example, most sweet oranges grow on trees grafted to the root of a sour orange variety. Suppose that each graft fails independently with probability 0.3. Five grafts are scheduled to be performed next week. Let X denote the number of grafts that will fail next week.

   (a) Write the formula for the pmf for X.

   (b) Find the expected value of X.

   (c) Find the variance of X.

   (d) Suppose that the cost of each failed graft is \$9.00. Find:

      i. The probability that the cost from failed grafts will exceed \$20.00.

      ii. The expected cost from failed grafts.

      iii. The variance of the cost from the failed grafts.

2. In the grafting context of the previous exercise, suppose that grafts are done one at a time. What is the expected number of successful grafts until the first failed graft?

3. Suppose that 8 of the 20 buses in a particular city have developed cracks on the underside of the main frame. Five buses are to be selected for thorough inspection.

Let X denote the number of buses (among the five that are inspected) that have cracks.

(a) Give the pmf of X.

(b) Find the expected value of X.

(c) Find the variance and the standard deviation of X.

4. A distributor receives a new shipment of 20 ipods. He draws a random sample of five ipods and thoroughly inspects the click wheel of each of them. Suppose that the new shipment of 20 ipods contains three with malfunctioning click wheel. Let X denote the number of ipods with defective click wheel in the sample of five.

   (a) Give the pmf of X.

   (b) Find the expected value of X.

   (c) Find the variance and the standard deviation of X.

5. In the context of Exercise 3.3.4,8, find the probability that out of four randomly and independently selected resistors, exactly two have resistance between 8.6 and 9.8.

6. Suppose that 10% of all components manufactured at General Electric are defective. Let X denote the number of defective components among 15 randomly selected ones.

   (a) What is the expected value of X?

   (b) What is the probability that there are exactly 3 defective components, $P(X = 3)$? Use the probability mass function to answer the question.

   (c) What is the probability that the number of defective components is between 2 and 5 inclusive, $P(2 \leq X \leq 5)$? Use the cumulative distribution function to answer the question.

   (d) Let $Y = n - X$. In words, what does Y represent?

   (e) Give the distribution of Y, including the value of the parameter(s).

   (f) Find the probability that, of the 15 components that were randomly selected, exactly 12 are **not** defective.

7. On average, 90% of letters are delivered within 3 working days. You send out 10 letters on Tuesday to invite friends for dinner. Only those who receive the invitation

by Friday (i.e., within 3 working days) will come. Let X denote the number of friends who come to dinner.

(a) What is the expected value and variance of X?

(b) Determine the probability that at least 7 friends are coming.

(c) A catering service charges a base fee of \$100 plus \$10 for each guest coming to the party. What is the expected value of the total catering cost? What is the variance of the total cost?

8. In a shipment of 10 electronic components, 2 are defective. Suppose that 5 components are selected at random for inspection, and let X denote the number of defective components found.

   (a) Give the probability mass function of X.

   (b) Use the probability mass function to compute μ_X and σ_X^2.

   (c) Compare your results in part (b) with those obtained from the formulas for the expected value and variance of the hypergeometric distribution.

9. Suppose that 30% of all drivers stop at an intersection having flashing red lights when no other cars are visible. Of 15 randomly chosen drivers coming to an intersection under these conditions, let X denote the number of those who stop.

   (a) Find the probabilities $P(X = 6)$ and $P(X \geq 6)$.

   (b) Find μ_X and σ_X^2.

10. Three electrical engineers toss coins to see who pays for coffee. If all three match, they toss another round. Otherwise the 'odd person' pays for coffee.

    (a) Find the probability that a round of tossing will result in a match (that is, either three 'heads' or three 'tails').

    (b) Find the probability that they will have to toss more than two times to determine an 'odd person'.

11. A telephone operator receives calls at a rate of 0.3 per minute. Let X denote the number of calls received in a given 3-minute period.

    (a) The distribution of the random variable X is (choose one) `(i) binomial  (ii) hypergeometric  (iii) negative binomial  (iv) Poisson`

(b) What is the mean value and variance of X?

(c) Find the probability that exactly 1 call arrive in a given 3-minute period.

(d) Find the probability that exactly 1 call will arrive in a given 9-minute period

12. Suppose that on average there are 1.5 cracks per concrete specimen for a particular type of cement mix.

    (a) Let X denote the number of cracks in a randomly chosen concrete specimen. State the distribution of X.

    (b) The random variable Y takes on the values 0, 1, 2, and 3, as X takes the values 0, 1, 2, and $\geq$3, respectively. (In other words, $Y = X$ if $X = 0$, 1, 2, or 3, and $Y = 3$ if $X > 3$.) Write the probability mass function of Y.

    (c) Calculate the mean and variance of Y.

    (d) The cost of cracks in a concrete specimen is $C = 15 \times Y$ dollars. Find the mean and variance of C.

13. In the inspection of tin plate by continuous electrolytic process, 2 imperfections are spotted on the average per minute. Let the random variable X denote the number of imperfections that will be spotted during the next 24 seconds.

    (a) Specify the distributions of X.

    (b) Find the probability that X takes the value 1 (i.e. of spotting exactly one imperfection in the next 24 seconds.)

14. It is known that 0.1% of the books bound at a certain bindery have defective bindings. Let the random variable X denote the number of books that have defective bindings out of the next 500 books that are bound.

    (a) What is the distribution of X?

    (b) What distribution would you use to approximate probabilities related to the random variable X?

    (c) Find the (approximate) probability that at most 5 of the next 500 books that are bound have defective bindings. (Hint: E(X)=5.)

15. Each morning the manager of a Barney toy fabrication plant inspects the output of the assembly line. The line produces Barneys at a constant rate of 1000 every hour. Each Barney is defective with probability 0.01, independently of all the other Barneys.

    (a) State the exact distribution of X, the number of defective items found in one hour.

    (b) Suggest an approximating distribution and use it to find the approximate proximate probability that the number of defective items find in one hour is four or fewer.

16. An engineer at a construction firm has a subcontract for the electrical work in the construction of a new office building. From past experience with this electrical subcontractor, the engineer knows that each light switch that is installed will be faulty with probability $p = .002$ independent of the other switches installed. The building will have $n = 1500$ light switches in it. Let X be the number of faulty light switches in the building.

    (a) How is X distributed? Do not forget to specify parameter values.

    (b) What is the probability of no faulty switches?

    (c) Can the exact probability of the event $X > 5$ be easily evaluated? If not suggest a way to approximate this probability and carry out your approximation

17. It has been suggested that a Poisson process can be used to model the occurrence of structural loads over time in aging concrete structures. Suppose that there is an average of two occurrences per year.

    (a) Let X denote the number of loads that will occur in an aging structure during the next quarter of a year. What is the distribution of X?

    (b) What is the probability that more than 2 loads will occur during the next quarter of a year?

18. In a shipment of 10,000 of a certain type of electronic component, 100 are defective.

    (a) Suppose that 15 components are selected at random for inspection, and let X denote the number of defective components found. State the exact distribution of X and give the formula for $P(X = k)$. (Hint: Choose one distribution from: Bernoulli, binomial, Hypergeometric, negative binomial, Poisson.)

(b) Using the table of probabilities for the appropriate distribution, approximate the probability that $X = 0$.

(c) Suppose that 400 components are selected at random for inspection, and let Y denote the number of defective components found. State the exact distribution of Y and approximate (using the appropriate table of probabilities) the probability that at least three defective components are found.

3.6 Models for Continuous Random Variables

Continuous random variables arise primarily from taking different types of measurement (duration, strength, hardness, elasticity, concentration of a particular soil contaminant etc). As such, we often have no indication of which probability model will best describe the true distribution of a particular continuous random variable. For example, there may be no a priori knowledge that the distribution of the duration (life time) of a randomly chosen electrical component is exponential, as it was assumed in Example 3.3.9. This is in stark contrast to the three probability models for inspection experiments (i.e. the Binomial, the Hypergeometric and the Negative Binomial), where the nature of the experiment determines the type of probability model. Thus, the probability models for continuous random variables serve not only as a convenient classification of experiments with similar probabilistic structure, but also as approximations to the true probability distribution of random variables.

In this section we will introduce some useful models for classes, or families, of continuous distributions. Two such models, the *Uniform* and of the *Exponential* were introduced in Section 3.4.2 for the purpose of illustrating the calculations involved in finding the mean (or expected) value and the variance of continuous random variables. For each of the probability models that we introduce below, we give the formula for the pdf, the mean value and variance. In addition, we will describe how to find the median and other percentiles, with the use of tables, when the cdf does not have a closed form expression.

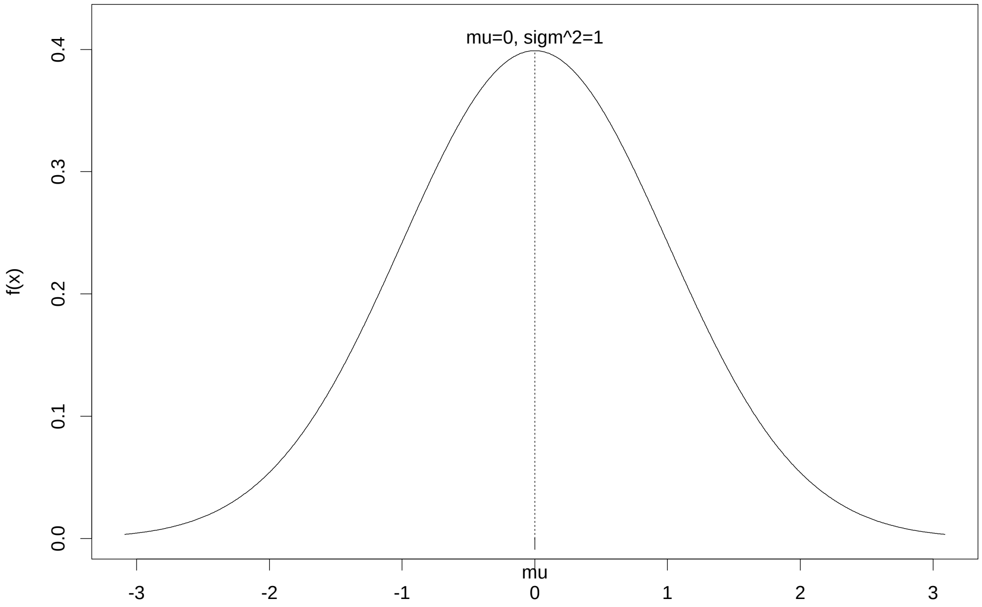


Figure 3.8: PDF of the N$(0, 1)$ Distribution

3.6.1 The Normal Distribution

A random variable X is said to be **normal**, or to have the normal distribution, with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if its pdf is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty.$$

This is a symmetric distribution with center of symmetry μ. Thus, μ is both the mean and the median of X. The parameter σ controls the spread of the distribution; in particular, σ is the standard deviation of X.

When $\mu = 0$ and $\sigma = 1$, X is said to have the **standard normal distribution** and is denoted by Z. The pdf of Z is denoted by ϕ, and is shown in Figure 3.8. Thus,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \ -\infty < z < \infty.$$

The cdf of Z is denoted by Φ. Thus

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(x) dx.$$

This integral is hard to evaluate, and it does not have a closed form expression. The same is true for the cdf of the normal distribution with any mean μ and variance σ^2. However,

$\Phi(z)$ is tabulated in Table A.3 for values of z from -3.49 to 3.49 in increments of 0.01. We will see how this table is used for calculating probabilities and for finding percentiles, not only for the standard normal random variable but also for other normal random variables. Such calculations are based on an important property of the normal distribution given in the following proposition.

Proposition 3.6.1. *If $X \sim N(\mu, \sigma^2)$, then*

$$Y = a + bX \sim N(a + b\mu, b^2\sigma^2).$$

The new element of this proposition is that a linear transformation of a normal random variable, X, is also a normal random variable. The mean value $(a + b\mu)$, and the variance $(b^2\sigma^2)$ of the transformed variable Y, are as specified in Propositions 3.4.2 and 3.4.3, so there is nothing new in that.

Finding probabilities

We first illustrate the use of Table A.3 for finding probabilities associated with the standard normal random variable.

Example 3.6.1. Let $Z \sim N(0, 1)$. Find: a) $P(-1 < Z < 1)$, b) $P(-2 < Z < 2)$, and c) $P(-3 < Z < 3)$.

Solution. In Table A.3, z-values are listed in two decimal places, with the second decimal place identified in the top row of the table. Thus, the z-value 1 is identified by 1.0 in the left column of the table and 0.00 in the top row of the table. The probability $\Phi(1) = P(Z \leq 1)$ is the number that corresponds to the row and column identified by 1.0 and 0.00, which is 0.8413. Working this way, we find:
a) $\Phi(1) - \Phi(-1) = .8413 - .1587 = .6826$,
b) $\Phi(2) - \Phi(-2) = .9772 - .0228 = .9544$,
c) $\Phi(3) - \Phi(-3) = .9987 - .0013 = .9974$

Remark 3.6.1. In words, this example means that approximately 68% of the values of a standard normal random variable fall within one standard deviation from its mean (part a), approximately 95% fall within two standard deviations and approximately 99.7% of its values fall within 3 standard deviation. This is called the 68-95-99.7% rule.

The use of Table A.3 for finding probabilities associated with any normal random variable (i.e. having a general mean $\mu \neq 0$, and a general variance $\sigma^2 \neq 1$) is made possible through the following corollary to Proposition 3.6.1.

Corollary 3.6.1. *If $X \sim N(\mu, \sigma^2)$, then*

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

To see how the corollary follows from the Proposition 3.6.1, apply first the proposition with $a = -\mu$ and $b = 1$, to obtain that if $X \sim N(\mu, \sigma^2)$, then

$$Y = X - \mu \sim N(0, \sigma^2).$$

A second application of the Proposition 3.6.1, on the normal random variable Y, with $a = 0$ and $b = 1/\sigma$, yields

$$Z = \frac{Y}{\sigma} = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

In words, Corollary 3.6.1, shows that any normal random variable, X, can be **standardized**, i.e. transformed to a standard normal random variable, Z, by subtracting from it its mean and dividing by its standard deviation. This implies that any event of the form $a \leq X \leq b$ can be expressed in terms of the standardized variable:

$$[a \leq X \leq b] = \left[\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right].$$

(The " = " sign here expresses the equivalence of the two events.) It follows that

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right), \quad (3.6.1)$$

where the last equality follows from the fact that $(X - \mu)/\sigma$ has the standard normal distribution.

This calculation is illustrated in the next example .

Example 3.6.2. Let $X \sim N(1.25, 0.46^2)$. Find a) $P(1 \leq X \leq 1.75)$, and b) $P(X > 2)$.

Solution. According to Corollary 3.6.1, we have that

$$Z = \frac{X - 1.25}{0.46} \sim N(0, 1).$$

The idea now is to express the events, whose probabilities we want to find, in terms of $Z = (X - 1.25)/0.46$, and then use Table A.3. For the event in part a) we have

$$\begin{aligned} P(1 \leq X \leq 1.75) &= P\left(\frac{1-1.25}{.46} \leq \frac{X-1.25}{.46} \leq \frac{1.75-1.25}{.46}\right) \\ &= \Phi(1.09) - \Phi(-.54) = .8621 - .2946 = .5675. \end{aligned}$$

Working similarly for the event in part b), we have

$$P(X > 2) = P\left(Z > \frac{2-1.25}{.46}\right) = 1 - \Phi(1.63) = .0516.$$

Finding percentiles

We first demonstrate the use of Table A.3 in finding percentiles of the standard normal distribution. The notation for the percentiles of a standard normal random variable is consistent with the notation of percentiles that we used in Subsection 3.4.2. In particular, we have

Notation: The $(1-\alpha)100$th percentile of Z is denoted by z_α.

For example, the 90th percentile is denoted by $z_{0.1}$, the 95th percentile by $z_{0.05}$, etc. Figure 3.9 recalls the defining property of percentiles.

To use Table A.3 for finding z_α, one first locates $1-\alpha$ in the body of Table A.3 and then reads z_α from the margins. If the exact value of $1-\alpha$ does not exist in the main body of the table, then an approximation is used as described in the following.

Example 3.6.3. Find the 95th percentile of Z.

Solution. Here $\alpha = 0.05$, so $1-\alpha = 0.95$. However, the exact number of 0.95 does not exist in the body of Table A.3. So we use the entry that is closest to, but larger than 0.95 (which is 0.9505), as well as the entry that is closest to, but smaller than 0.95 (which is 0.9495), and approximate $z_{0.05}$ by averaging the z-values that correspond to these two closest entries: $z_{0.05} \simeq (1.64 + 1.65)/2 = 1.645$.

The use of Table A.3 for finding percentiles of any normal random variable (i.e. having a general mean $\mu \neq 0$, and a general variance $\sigma^2 \neq 1$) is made possible through the following corollary to Proposition 3.6.1.

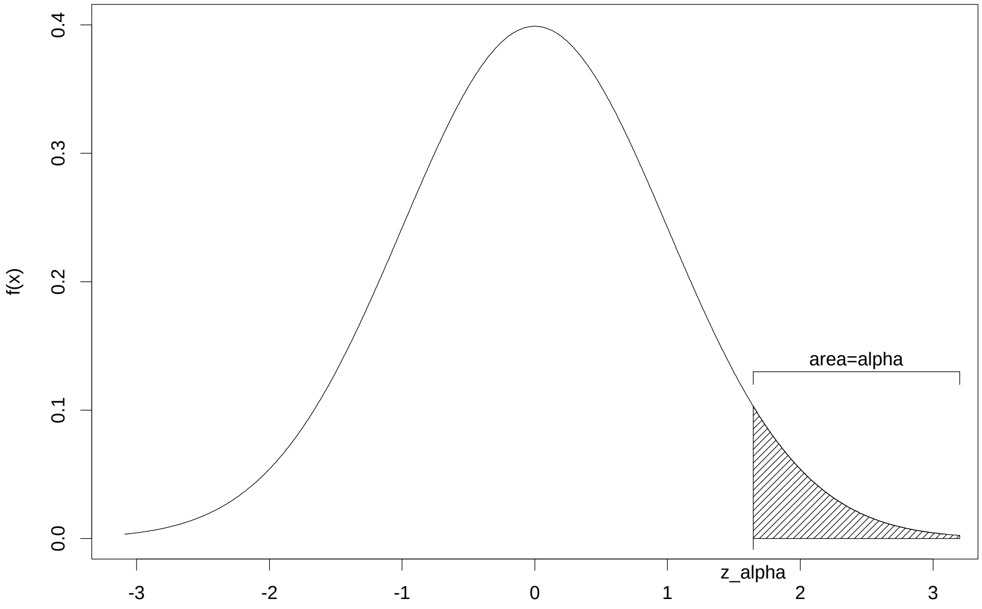


Figure 3.9: Illustration of Percentile Definition and Notation

Corollary 3.6.2. *Let $X \sim N(\mu, \sigma^2)$, and let x_α denote the $(1-\alpha)$100th percentile of X. Then,*

$$x_\alpha = \mu + \sigma z_\alpha. \tag{3.6.2}$$

To prove this corollary, i.e. to show that $\mu + \sigma z_\alpha$ is indeed the $(1-\alpha)$100th percentile of X, we must show that $P(X \leq \mu + \sigma z_\alpha) = 1 - \alpha$. But this follows by an application of relation (3.6.1) with $a = -\infty$ and $b = \mu + \sigma z_\alpha$:

$$P(X \leq \mu + \sigma z_\alpha) = \Phi(z_\alpha) - \Phi(-\infty) = 1 - \alpha - 0 = 1 - \alpha,$$

since $\Phi(-\infty) = 0$. An illustration of finding percentiles of any normal distribution through the formula (3.6.2) is given next.

Example 3.6.4. Let $X \sim N(1.25, .46^2)$. Find the 95-th percentile, $x_{.05}$, of X.

Solution. From (3.6.2)) we have

$$x_{.05} = 1.25 + .46 z_{.05} = 1.25 + (.46)(1.645) = 2.01.$$

Remark 3.6.2. The 68-95-99.7% rule applies for any normal random variable $X \sim N(\mu, \sigma^2)$. Indeed,

$$\begin{aligned}
P(\mu - 1\sigma < X < \mu + 1\sigma) &= P(-1 < Z < 1), \\
P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2), \\
P(\mu - 3\sigma < X < \mu + 3\sigma) &= P(-3 < Z < 3).
\end{aligned}$$

3.6.2 Distribution Models Used in Reliability

A random variable that records the life time, or time to failure, of a product or equipment is called a failure time, or survival time, or life time. Reliability theory deals with models for probability distributions for life times. Such models of probability distributions, which are also referred to as **failure-time distributions**, assign all probability mass on the positive half of the real line. This is because life times take only positive values. The exponential distribution, which was introduced in Section 3.4.2, is an example of a failure-time distribution.

Failure-time distributions have their own terminology. If $f(x)$, $F(x)$ denote the pdf and cdf of the time to failure, X, of some product, the probability $P(X > t)$ is called the **reliability** of the product at time t. The **reliability function**, also called **survival function**, is defined as

$$R(t) = 1 - F(t) = 1 - \int_0^t f(x)dx. \tag{3.6.3}$$

The **failure rate**, also called **hazard function**, is defined as

$$h(t) = \frac{f(t)}{R(t)}. \tag{3.6.4}$$

The failure or hazard rate provide information regarding the conditional probability of failure in the small interval $(t, t + \Delta t)$ given that the product has not failed up to time t. In particular,

$$P(t < X < t + \Delta t) \simeq h(t)\Delta t. \tag{3.6.5}$$

The failure rate is characteristic of a failure-time distribution in the sense that the pdf f can be determined from the failure rate h. It can be shown that

$$f(t) = h(t) \exp\{-\int_0^t h(x)dx\}. \tag{3.6.6}$$

Example 3.6.5. Let the failure time T have the exponential distribution with parameter value λ. Thus, $f(t) = \lambda \exp\{-\lambda t\}$, for $t > 0$, and $f(t) = 0$, for $t < 0$. Then show:

1. That the reliability function of X is: $R(t) = \exp\{-\lambda t\}$.
2. That the failure rate of X is: $h(t) = \lambda$.
3. Relation (3.6.6) for this particular case.

Solution: The first two parts follow immediately from the definitions. To see the third part, substitute the constant λ for the function h in (3.6.6) to obtain

$$f(t) = \lambda \exp\{-\int_0^t \lambda dx\} = \lambda \exp\{-\lambda t\}.$$

Remark 3.6.3. The fact that the failure rate of an exponentially distributed failure time is constant is related to the so-called **no-aging property** of the exponential distribution. See Exercise 8

Three other models of probability distributions for failure-times are given in the next three subsections.

3.6.3 The Log-normal distribution

A failure-time T is said to have the log-normal distribution if its natural logarithm is normally distributed. Thus, if $\mu_{\ln}$, $\sigma_{\ln}^2$ are the mean and variance of $\ln T$, the cdf of T is

$$F_T(t) = P(T \leq t) = P(\ln T \leq \ln t) = \Phi\left(\frac{\ln t - \mu_{\ln}}{\sigma_{\ln}}\right), \quad \text{for } t > 0, \tag{3.6.7}$$

and $F_T(t) = 0$, for $t < 0$. Differentiating $F_T(t)$ with respect to t, we obtain that the pdf of T satisfies $f_T(t) = 0$, for $t < 0$, and

$$f_T(t) = \frac{1}{t\sigma_{\ln}}\phi\left(\frac{\ln t - \mu_{\ln}}{\sigma_{\ln}}\right) = \frac{t^{-1}}{\sqrt{2\pi\sigma_{\ln}^2}} \exp\left\{-\frac{(\ln t - \mu_{\ln})^2}{2\sigma_{\ln}^2}\right\} \tag{3.6.8}$$

This is a family of positively skewed pdf's. Note that the parameters $\mu_{\ln}$, $\sigma_{\ln}$ of the log-normal distribution, or the rv T, are the mean and standard deviation of $\ln T$, not those of T (see (3.6.9)). The one that corresponds to $\mu_{\ln} = 0$ and $\sigma_{\ln} = 1$ is called the standard log-normal pdf. Its plot is shown in Figure 3.10.

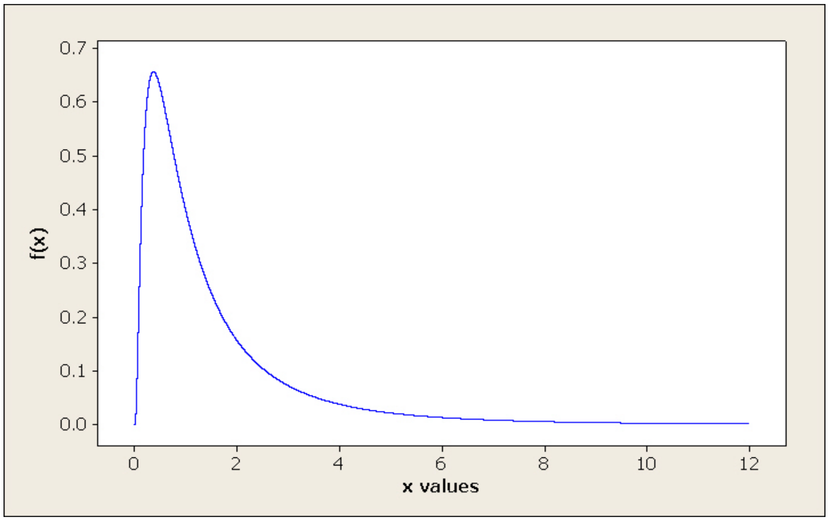


Figure 3.10: Graph of the standard log-normal pdf ($\mu_{\ln} = 0$, $\sigma_{\ln} = 1$).

If T has the log-normal distribution with pdf (3.6.8), the mean and variance of of T are

$$\mu_T = e^{\mu_{\ln} + \sigma^2_{\ln}/2}, \quad \sigma^2_T = e^{2\mu_{\ln} + \sigma^2_{\ln}} \times \left(e^{\sigma^2_{\ln}} - 1 \right). \tag{3.6.9}$$

Thus, the variance of a log-normal rv increases both with the mean and with the variance of its logarithm. For example, if $\ln T$ has the standard log-normal distribution, thus mean $\mu_{\ln} = 0$ and variance $\sigma^2_{\ln} = 1$, the variance of T is 4.67. If $\ln T$ has mean $\mu_{\ln} = 5$ and variance $\sigma^2_{\ln} = 1$, the variance of T is 102,880.54. If the variance of $\ln T$ increases to $\sigma^2_{\ln} = 4$ the variance of T increases to 64,457,184.42.

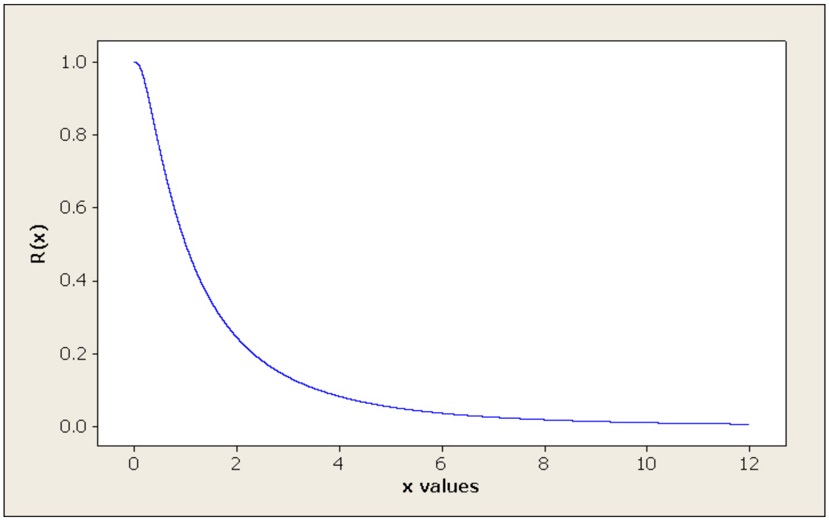


Figure 3.11: Reliability function of the log-normal distribution ($\mu_{\ln} = 0$, $\sigma_{\ln} = 1$).

The reliability and hazard functions of the standard log-normal distribution are shown in Figures 3.11 and 3.12, respectively.

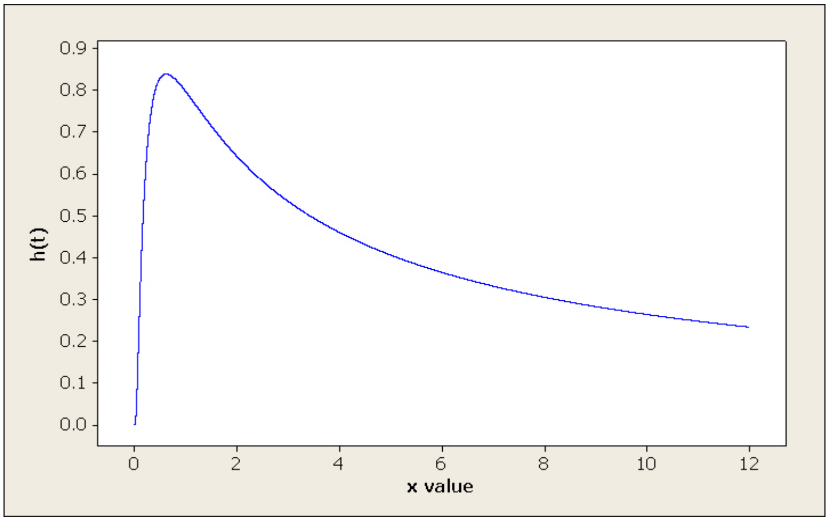


Figure 3.12: Hazard function of the log-normal distribution ($\mu_{\ln} = 0$, $\sigma_{\ln} = 1$).

The form of the hazard function suggests that the failure rate of equipment whose life time follows a log-normal distribution starts by a sharp increase during the fist, break-in, period of their use, but later it decreases. This feature has been observed empirically in the life times of many products, including that of a semiconductor laser, as well as in medical applications, including the blood pressure of humans. Contributing to the popularity of the log-normal distribution is the fact that the probabilities and percentiles are easily evaluated by reference to the normal distribution. This is illustrated in the following example.

Example 3.6.6. Suppose that the life time, T, of a product has a log-normal distribution with parameters $\mu_{\ln} = 10$ hours $\sigma_{\ln} = 1.5$ hours. Find a) $t_{0.05}$, the 95th percentile of T, and b) the probability that the lifetime of a randomly selected product exceeds 20,000 hours.

Solution: a) $t_{0.05}$ satisfies the equation $P(T > t_{0.05}) = 0.05$. Because

$$P\left(T > t_{0.05}\right) = P\left(\ln T > \ln t_{0.05}\right) = 1 - \Phi\left(\frac{\ln t_{0.05} - 10}{1.5}\right) = 0.05,$$

it follows that

$$\frac{\ln t_{0.05} - 10}{1.5} = z_{0.05} = 1.645, \quad \text{so that} \quad t_{0.05} = \exp(12.4675) = 259,756.52 \quad \text{hours}.$$

b) We have

$$P(T > 20,000) = P\left(\ln T > \ln 20,000\right) = 1 - \Phi\left(\frac{\ln 20,000 - 10}{1.5}\right) = 0.53.$$

3.6.4 The Gamma and the χ^2 distributions

The gamma family of distributions is important not only because it offers a rich variety of right skewed pdf's, but because it includes, as special cases, several other important distributions. In particular, it includes a) the exponential family, which models the time until the fist occurrence in a Poisson process, b) the **Erlang** family, which models the time until the rth occurrence in a Poisson process, and c) the family of (central) χ^2 distributions, which is useful in some statistical inference contexts.

The rv T has a gamma distribution with *shape* parameter $\alpha > 0$ and *scale* parameter $\beta > 0$ if its pdf is zero for negative values and

$$f_T(t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-t/\beta}, \quad \text{for} \quad t \geq 0,$$

where Γ is the *gamma function* defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

Using integration by parts it can be shown that

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1), \text{ and, for } r \geq 1 \text{ integer, } \Gamma(r) = (r-1)!$$

Moreover, it can be shown that $\Gamma(1/2) = \pi^{1/2}$. Figure 3.13 shows how the shape of the pdf changes for different values of α.

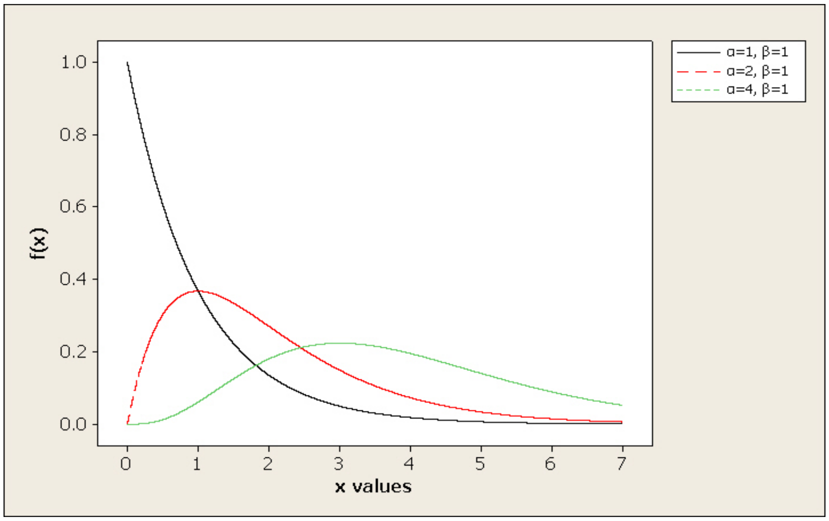


Figure 3.13: PDFs of some gamma distributions.

When $\alpha = 1$ we get the family of exponential distributions with $\lambda = 1/\beta$, and for $\alpha = r$, $r \geq 1$, we get the family of Erlang distributions. Finally, the χ^2_ν (which is read as "chi-square with ν degrees of freedom") distribution, where $\nu \geq 1$ is an integer, corresponds to $\alpha = \nu/2$ and $\beta = 2$.

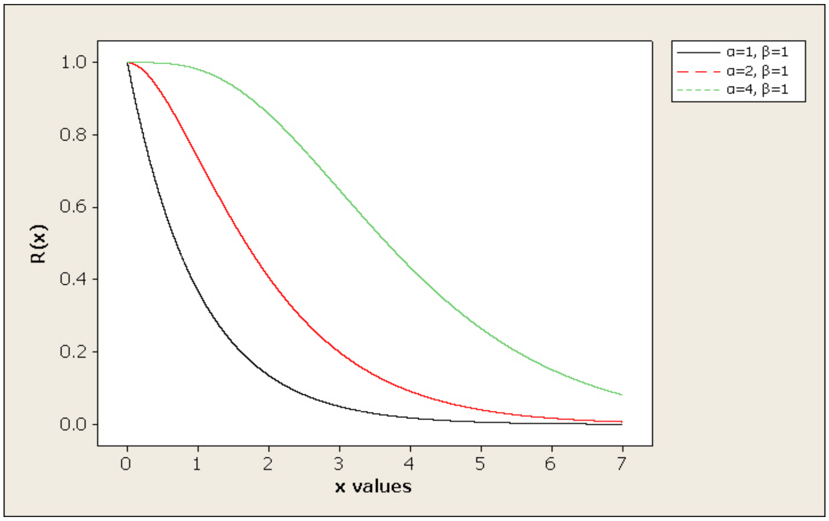


Figure 3.14: Reliability functions of some gamma distributions.

The mean and variance of a gamma distribution, or rv T, are given by

$$\mu_T = \alpha\beta, \quad \sigma_T^2 = \alpha\beta^2.$$

Figures 3.14 and 3.15 show the reliability and hazard functions for different values of the shape parameter, and $\beta = 1$.

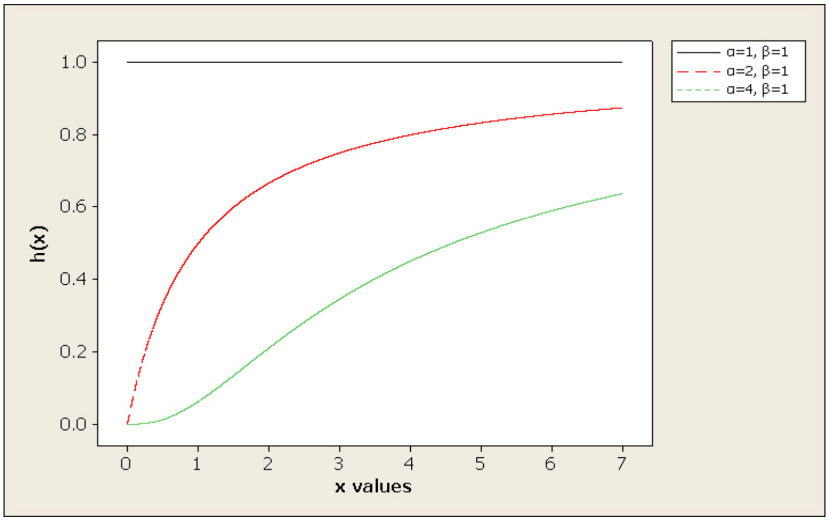


Figure 3.15: Hazard functions of some gamma distributions.

As can be seen from Figure 3.15, if the life time of a product has the gamma distribution with $\alpha > 1$, its failure rate keeps increasing with time of use which is appropriate for modeling the usual aging process (products become less reliable as they age). The case $\alpha = 1$ corresponds to the exponential distribution which, as we have seen, has a constant failure rate equal to λ, which, in Figure 3.15, is $\lambda = 1/\beta = 1$. Though not shown here, if $\alpha < 1$ the failure rate decreases sharply in the beginning and then it stabilizes.

3.6.5 The Weibull distribution

Another generalization of the exponential distribution is the Weibull distribution. A rv T is said to have a Weibull distribution with *shape* parameter $\alpha > 0$ and *scale* parameter $\beta > 0$ if its pdf is zero for $t < 0$ and

$$f_T(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} e^{-(t/\beta)^\alpha}, \quad \text{for} \quad t \geq 0.$$

When $\alpha = 1$ the Weibull pdf reduces to the exponential pdf with $\lambda = 1/\beta$. Both α and β can vary to obtain a number of different pdf's, as Figure 3.16. Though the exponential family of distributions is included in both the gamma and the Weibull, these are distinct families of distributions.

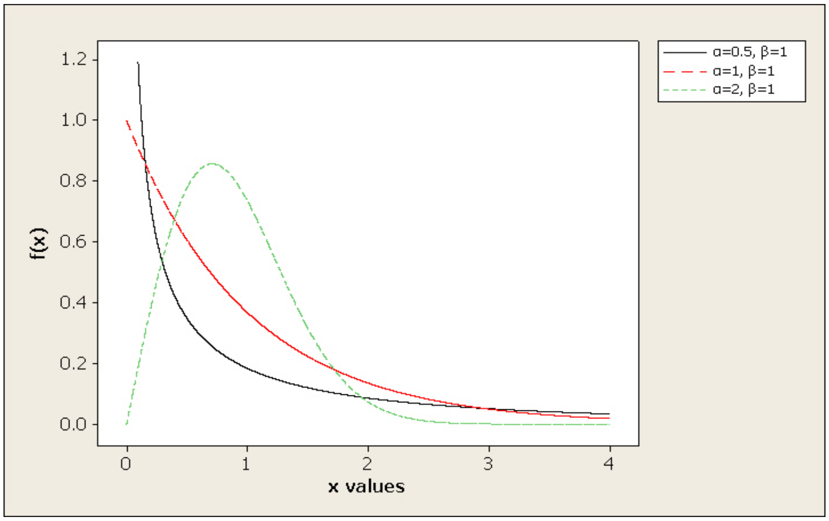


Figure 3.16: PDFs of some Weibull distributions.

The mean and variance of a Weibull distribution, or rv T, are given by

$$\mu_T = \beta\Gamma\left(1 + \frac{1}{\alpha}\right), \quad \sigma_T^2 = \beta^2 \left\{ \Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2 \right\}.$$

Moreover, the cdf and the reliability functions have closed form expressions that facilitate the calculation of probabilities and percentiles:

$$F_T(t) = 1 - e^{-(t/\beta)^\alpha}, \quad \text{and} \quad R_T(t) = e^{-(t/\beta)^\alpha}.$$

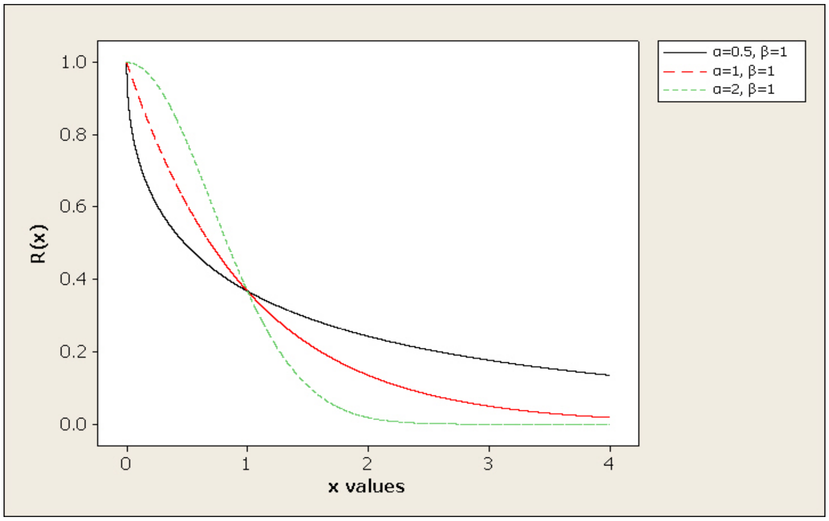


Figure 3.17: Reliability functions of some Weibull distributions.

The reliability function, and hazard function for selected values of the shape and scale parameters are shown in Figures 3.17 and 3.18, respectively.

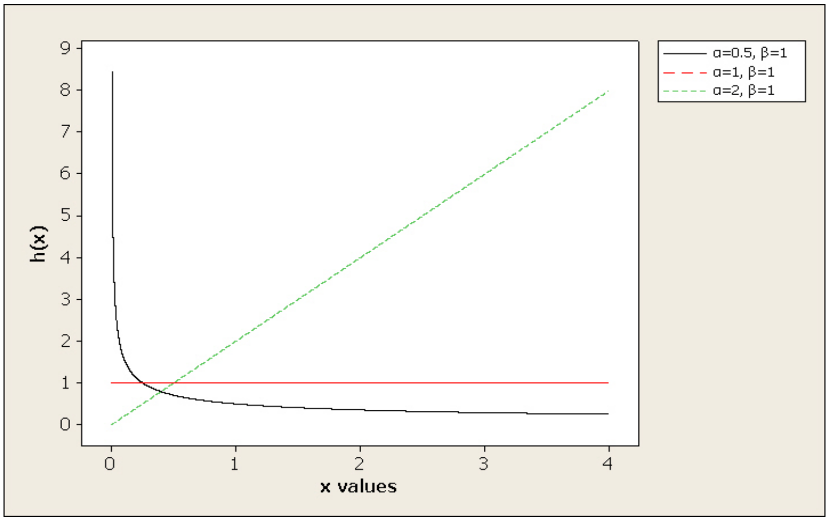


Figure 3.18: Hazard functions of some Weibull distributions.

Example 3.6.7. Suppose that T has the Weibull distribution with parameters $\alpha = 20$ and $\beta = 100$. Find a) the probability that T takes a value less than 102, and b) the 90th percentile, $t_{0.1}$, of T.

Solution: a) $P(T \leq 102) = 1 - e^{-(102/100)^{20}} = 0.77$.
b) $t_{0.05}$ satisfies $1 - e^{-(t_{0.1}/100)^{20}} = 0.9$. Solving, we obtain $t_{0.1} = 100 \times (-\ln(0.1))^{0.05} = 104.26$.

3.6.6 Exercises

1. Suppose that the yield strength (ksi) for A36 steel is normally distributed with $\mu = 43$ and $\sigma = 4.5$

   (a) What is the 25th percentile of the distribution of A36 steel strength?

   (b) What strength value separates the strongest 10% from the others?

   (c) What is the value of c such that the interval $(43 - c, 43 + c)$ includes 99% of all strength values?

   (d) What is the probability that at most 3 of 15 independently selected A36 steels have strength less than 43?

2. The expected value of the weight of frozen yogurt cups sold at the HUB is 8 oz. Suppose the weight of a cup is normally distributed with standard deviation 0.5 oz, and independent from previous cups.

   (a) What is the probability to get a cup weighing more than 8.64 oz?

   (b) What is the probability to get a cup weighing more than 8.64 oz three days in a row?

3. The resistance for resistors of a certain type is a random variable X having the normal distribution with mean 9 ohms and standard deviation 0.4 ohms. An resistor is acceptable if its resistance is between 8.6 ohms and 9.8 ohms.

   (a) What is the probability that a randomly chosen resistor is acceptable?

   (b) What is the probability that out of four randomly and independently selected resistors, two are acceptable?

4. Admission officers in Colleges A and B use the SAT scores as their admission criteria. SAT scores are normally distributed with mean 500 and standard deviation 80. College A accepts people whose scores are above 600, and College B accepts the top 1% of people in terms of their SAT scores.

   (a) What percentage of high school seniors can get into College A?

   (b) What is the minimum score needed to get accepted by College B ?

5. The finished inside diameter of a piston ring is normally distributed with a mean of 10 cm and a standard deviation of 0.03 cm.

   (a) Above what value of inside diameter will 85.08% of the piston rings fall?

   (b) What is the probability that the diameter of a randomly selected piston will be less than 10.06?

6. The actual content of apple juice in 16 oz bottles bottled by a certain firm is normally distributed with mean 16 oz and standard deviation 1 oz.

   (a) What proportion of bottles contain more than 17 oz of apple juice?

   (b) If 33% of all bottles contain less than c oz, find c.

7. A machine manufactures tires with a tread thickness that is normally distributed with mean 10 millimeters (mm) and standard deviation 2 millimeters. The tire has a 50,000 mile warranty. In order to last for 50,000 miles the manufactures guidelines specify that the tread thickness must be at least 7.9 mm. If the thickness of tread is measured to be less than 7.9 mm, then the tire is sold as an alternative brand with a warranty of less than 50,000 miles.

   (a) Find expected proportion of rejects.

   (b) The demand for the alternative brand of tires is such that 30% of the total output should be sold under the alternative brand name. How should the critical thickness, originally 7.9 mm, be set to achieve 30% rejects.

8. (**No-aging property of the exponential distribution.**) If a life time T has an exponential distribution, i.e. its pdf and cdf are $f(t) = \lambda \exp\{-\lambda t\}$, $F(t) = 1 - \exp\{-\lambda t\}$, it has the following interesting *no-aging* property:

$$P(T \le t_0 + t_1 | T \ge t_0) = 1 - \exp\{-\lambda t_1\}$$

   In words the no-aging property means that, given that a product (whose life time is T) has not failed by time t_0, then the probability that it will fail in the next time period of length t_1 is as it has never been used. Prove the no-aging property.

9. Suppose that the life time, T, of a product has a log-normal distribution with parameters $\mu_{\ln} = 10$ hours $\sigma_{\ln} = 1.5$ hours. Find

   (a) the mean and variance of T,

   (b) the probability that the life time of a randomly selected product does not exceed 1000 hours, and

   (c) the 95th percentile of T.

10. Suppose that T has the Weibull distribution with parameters $\alpha = 0.2$ and $\beta = 10$. Find

   (a) the mean and variance of T,

   (b) the probability that T takes a value between 20 and 30, and

   (c) the 95th percentile of T.

Chapter 4

Multivariate Variables and Their Distribution

4.1 Introduction

In Chapter 3, we defined a univariate random variable as a rule that assigns a number to each outcome of a random experiment. When there are several rules, each of which assigns a (different) number to each outcome of a random experiment, then we have a **multivariate** random variable. The notion of a multivariate random variable was introduced in Chapter 1, where several examples of studies with a multivariate outcome are given. Some additional examples are

1. The X, Y and Z components of wind velocity can be measured in studies of atmospheric turbulence.
2. The velocity X and stopping distance Y of an automobile can be studied in an automobile safety study.
3. The diameter at breast height (DBH) and age of a tree can be measured in a study aimed at developing a method for predicting age from diameter.

In such studies, it is not only interesting to investigate the behavior of each variable separately, but also to a) investigate and quantify the degree of relationship between them, and b) develop a method for predicting one from another. In this chapter we will introduce, among other things, the notion of *correlation*, which serves as a quantification of the relationship between two variables, and the notion of *regression function* which forms the basis for predicting one variable from another. To do that, we first need to define the *joint distribution* of multivariate random variables, discuss ways of providing concise descriptions of it, and introduce the additional concepts of *marginal distributions*, *conditional distributions* and *independent* random variables. Parameters of a joint distribution, such as the correlation and the regression function, will also be introduced. Finally, in this chapter we will present probability models for joint distributions.

4.2 Joint Distributions and the Joint CDF

We say that we know the joint distribution of a bivariate variable, (X, Y) if we know the probability for (X, Y) to take value in any given rectangle on the plane, i.e. if we know all probabilities of the form

$$P(a < X \leq b,\ c < Y \leq d), \quad \text{with} \ \ a < b,\ c < d, \tag{4.2.1}$$

where $P(a < X \leq b,\ c < Y \leq d)$ is to be interpreted as $P(a < X \leq b \text{ and } \ c < Y \leq d)$ or as $P([a < X \leq b] \ \cap \ [c < Y \leq d])$. Similarly, we say that we know the joint distribution of a multivariate variable, $X_1, X_2, \ldots, X_m$, if we know all probabilities of the form

$$P(a_1 < X_1 \leq b_1,\ a_2 < X_2 \leq b_2,\ \ldots,\ a_m < X_m \leq b_m), \quad \text{with} \ \ a_k < b_k,\ k = 1, \ldots, m.$$

As in the univariate case, which was considered in Chapter 3, two ways for providing a concise description of the joint probability distribution of a multivariate random variable will be discussed. The first way, which is discussed in this section, generalizes the notion of *cumulative distribution function* (cdf) to the multivariate case. The second way generalizes the notions of *probability mass function* (pmf) and *probability density function* to the multivariate case, and is discussed in the next two sections.

Definition 4.2.1. *The* **joint** *or* **bivariate** *cdf of two random variables, X, Y, is defined by*

$$F(x, y) = P(X \leq x, Y \leq y).$$

The **joint** *or* **multivariate** *cdf of several random variables, $X_1, X_2, \ldots, X_m$, is defined by*

$$F(x_1, x_2, \ldots, x_m) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m).$$

The joint cdf is convenient for obtaining the cdf of the individual random variables that comprise the multivariate random variable, and, more generally, for calculating the probability that any two components, (X, Y), of a multivariate random variable will lie in a rectangle. These results are given in the following proposition.

Proposition 4.2.1. *a) Let $F(x, y)$ is the bivariate cdf of (X, Y). Then, the cdf of X is given by*

$$F_X(x) = F(x, \infty). \tag{4.2.2}$$

Similarly, $F_Y(y) = F(\infty, y)$ is the formula for obtaining the cdf of Y from the bivariate cdf of (X, Y).

b) Let $F(x_1, x_2, x_3)$ be the tri-variate cdf (X_1, X_2, X_3). Then the cdf of X_1 is

$$F_{X_1}(x_1) = F(x_1, \infty, \infty). \tag{4.2.3}$$

Similarly, the cdf of X_2 is $F_{X_2}(x_2) = F(\infty, x_2, \infty)$, and the cdf of X_3 is $F_{X_3}(x_3) = F(\infty, \infty, x_3)$.

c) Let $F(x_1, x_2, x_3)$ be the tri-variate cdf (X_1, X_2, X_3). Then the bivariate cdf of (X_1, X_2) is

$$F_{X_1,X_2}(x_1, x_2) = F(x_1, x_2, \infty). \tag{4.2.4}$$

Similarly, the bivariate cdf of (X_1, X_3) is $F_{X_1,X_3}(x_1, x_3) = F(x_1, \infty, x_3)$, and the cdf of (X_2, X_3) is $F_{X_2,X_3}(x_3) = F(\infty, x_2, x_3)$.

d) Let $F(x, y)$ is the bivariate cdf of (X, Y). Then,

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1). \tag{4.2.5}$$

A proof of part a) of the proposition is offered by the following series of equalities.

$$F_X(x) = P(X \leq x) = P(X \leq x,\ Y \leq \infty) = F(x, \infty).$$

Proofs of parts b) and c) are similar. Part d) expresses the probability of relation (4.2.1) in terms of the bivariate cdf, showing that knowledge of the bivariate cdf entails knowledge of the bivariate distribution. A similar (but more complicated) relation shows that the multivariate cdf entails knowledge of the multivariate distribution.

The following example illustrate the application of the properties of the joint cdf that are listed in Proposition 4.2.1.

Example 4.2.1. *[BIVARIATE UNIFORM DISTRIBUTION.]* Let (X, Y) have a **bivariate uniform distribution** on the unit rectangle $[0, 1] \times [0, 1]$. This means that the probability that (X, Y) lies in a subspace A of the unit rectangle equals the area of A. The bivariate cdf of (X, Y) is

$$\begin{aligned} F(x,y) &= xy, \quad \text{for } 0 \le x, y \le 1, \\ F(x,y) &= x, \quad \text{for } 0 \le x \le 1,\ y \ge 1, \\ F(x,y) &= y, \quad \text{for } 0 \le y \le 1,\ x \ge 1, \\ F(x,y) &= 1, \quad \text{for } x, y \ge 1, \text{ and} \\ F(x,y) &= 0, \quad \text{if either } x \le 0 \text{ or } y \le 0. \end{aligned}$$

a) Find the cdf of the variables X and Y.
b) Find the probability that (X, Y) will take a value in the rectangle $A = (0.3, 0.6) \times (0.4, 0.7)$.

Solution. For a) use Proposition 4.2.1a) to get

$$F_X(x) = F(x, \infty) = x, \quad 0 \le x \le 1,$$

which is recognized to be the cdf of the uniform in $[0, 1]$ distribution. Similarly $F_Y(y) = F(\infty, y) = y$, $0 \le y \le 1$.
For b) use Proposition 4.2.1d) to get

$$\begin{aligned} &P(0.3 < X \le 0.6, 0.4 < Y \le 0.7) = (0.6)(0.7) \\ &-(0.3)(0.7) - (0.6)(0.4) + (0.3)(0.4) = 0.09, \end{aligned}$$

which is equal to the area of A, as it should be.

Though very convenient for calculating the probability that (X, Y) will take value in a rectangle, and also for calculating the individual probability distributions of X and Y, the cdf is not the most convenient tool for calculating directly the probability that (X, Y) will take value in regions other than rectangles, such as circles.

4.2.1 Exercises

1. Let (X, Y) have joint cdf given by

$$\begin{aligned}
F(x,y) &= 0, \quad \text{for } x \le 0,\ y \le 0, \\
F(x,y) &= x\left(1-e^{-y/2}\right), \quad \text{for } 0 \le x \le 0.5,\ 0 \le y, \\
F(x,y) &= \frac{1}{2}\left(1-e^{-y/2}\right)+(x-0.5)\left(1-e^{-y/4}\right), \quad \text{for } 0.5 \le x \le 1,\ 0 \le y, \\
F(x,y) &= \frac{1}{2}\left(1-e^{-y/2}\right)+\frac{1}{2}\left(1-e^{-y/4}\right), \quad \text{for } 1 \le x,\ 0 \le y.
\end{aligned}$$

   (a) Find the probability that (X, Y) will take a value in the rectangle $A = (0.3, 0.6) \times (0.4, 0.7)$.

   (b) Find the probability that (X, Y) will take a value in the rectangle $A = (1.3, 1.6) \times (0.4, 0.7)$.

   (c) Sketch the range of the possible values of (X, Y).

   (d) Find the cdf of the variables X and Y.

4.3 The Joint Probability Mass Function

4.3.1 Definition and Basic Properties

Definition 4.3.1. *The* **joint** *or* **bivariate** *probability mass function (pmf) of the discrete random variables X, Y is defined as*

$$p(x,y) = P(X = x, Y = y).$$

The **joint** *or* **multivariate** *pmf of the discrete random variables $X_1, X_2, \ldots, X_n$ is similarly defined as*

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Let the sample space of (X, Y) be $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots\}$. The above definition implies that $p(x, y) = 0$, for all (x, y) that do not belong in $\mathcal{S}$ (i.e. all (x, y) different from $(x_1, y_1), (x_2, y_2), \ldots\}$). Moreover, from the Axioms of probability it follows that

$$p(x_i, y_i) \ge 0, \quad \text{for all } i, \text{ and } \sum_i p(x_i, y_i) = 1. \tag{4.3.1}$$

Example 4.3.1. A robot performs two tasks, welding joints and tightening bolts. Let X be the number of defective welds, and Y be the number of improperly tightened bolts per car. Suppose that the possible values of X are 0,1,2 and those of Y are 0,1,2,3. Thus, the sample space $\mathcal{S}$ consists of 12 pairs $(0,0),(0,1),(0,2),(0,3),\ldots,(2,3)$. The joint pmf of X,Y is given by

		Y				
		0	1	2	3	
	0	.84	.03	.02	.01	
X	1	.06	.01	.008	.002	
	2	.01	.005	.004	.001	
						1.0

In agreement with (4.3.1), the sum of all probabilities equals one. Inspection of this pmf reveals that 84% of the cars have no defective welds and no improperly tightened bolts, while only one in a thousand have two defective welds and three improperly tightened bolts. In fact, the pmf provides answers to all questions regarding the probability of robot errors. For example,

$$P(\text{exactly one error}) = P(X=1,Y=0) + P(X=0,Y=1) = .06 + .03 = .09.$$

and

$$\begin{aligned} &P(\text{exactly one defective weld}) = P(X=1,Y=0) + P(X=1,Y=1) \\ &+ P(X=1,Y=2) + P(X=1,Y=3) = .06 + .01 + .008 + .002 = .08. \end{aligned}$$

Example 4.3.2. *[MULTINOMIAL RANDOM VARIABLE]* The probabilities that a certain electronic component will last less than 50 hours of continuous use, between 50 and 90 hours, or more than 90 hours, are $p_1 = 0.2$, $p_2 = 0.5$, and $p_3 = 0.3$, respectively. Consider a simple random sample of size eight of such electric components, and set X_1 for the number of these components that last less than 50 hours, X_2 for the number of these that last between 50 and 90 hours, and X_3 for the number of these that last more than 90 hours. Then (X_1, X_2, X_3) is an example of a **multinomial** random variable. The sample space $\mathcal{S}$ of (X_1, X_2, X_3) consists of all nonnegative sets of integers (x_1, x_2, x_3) that satisfy

$$x_1 + x_2 + x_3 = 8.$$

It can be shown that the joint pmf of (X_1, X_2, X_3) is given by

$$p(x_1, x_2, x_3) = \frac{8!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3},$$

for any x_1, x_2, x_3 in the sample space. For example, the probability that one of the eight components will last less than 50 hours, five will last between 50 and 90 hours, and two will last more than 90 hours is

$$p(1, 5, 2) = \frac{8!}{1!5!2!} 0.2^1 0.5^5 0.3^2 = 0.0945.$$

4.3.2 Marginal Probability Mass Functions

In Section 5.1, we learned how to obtain the cdf of the variable X from the bivariate cdf of (X, Y). This individual distribution of the variable X is called the *marginal* distribution of X, and, similarly, the individual distribution of Y is called the *marginal* distribution of Y. The pmf of the marginal distribution of X, also called the **marginal pmf** of X, is denoted by p_X, and, similarly, the marginal pmf of Y is denoted by p_Y. In this section, we will learn how to obtain these marginal pmf's from a bivariate (or multivariate) pmf. The basic technique for doing so is illustrated in the following example.

Example 4.3.3. Consider the bivariate distribution of (X, Y) given in Example 4.3.1. Find the marginal pmf of X.

Solution. Note that each of the events $[X = x]$ is the union of the disjoint events $[X = x, Y = 0]$, $[X = x, Y = 1]$, $[X = x, Y = 2]$, and $[X = x, Y = 3]$. For example,

$$[X = 0] = [X = 0, Y = 0] \cup [X = 0, Y = 1] \cup [X = 0, Y = 2] \cup [X = 0, Y = 3].$$

Thus, each marginal probability $p_X(x)$, is found by summing the probabilities in the $X = x$ row of the table giving the joint pmf of (X, Y). For example,

$$\begin{aligned} p_X(0) &= P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 0, Y = 2) + P(X = 0, Y = 3) \\ &= 0.84 + 0.03 + 0.02 + 0.01 = 0.9. \end{aligned}$$

Similarly, the probability $p_X(1) = P(X = 1)$ is found by summing the probabilities in the $X = 1$ row (which is the second row) of the joint pmf table, and the probability $p_X(2) = P(X = 2)$ is found by summing the probabilities in the $X = 2$ row (which is the third row) of the joint pmf table. Thus, the marginal pmf of X is:

x	0	1	2
$p_X(x)$	.9	.08	.02

Note that the pmf of X in the above example can be read from the vertical margin of the joint pmf table; see the table below. This justifies naming p_X the *marginal pmf* of X. Also, the marginal pmf p_Y of Y is read from the lower horizontal margin of the joint pmf.

		Y				
		0	1	2	3	
	0	.84	.03	.02	.01	.9
X	1	.06	.01	.008	.002	.08
	2	.01	.005	.004	.001	.02
		.91	.045	.032	.013	1.0

In general, we have the following proposition.

Proposition 4.3.1. *Let $x_1, x_2, \ldots$ be the possible values of X and $y_1, y_2, \ldots$ be the possible values of Y. The marginal pmf of X, respectively, Y are given by*

$$p_X(x) = \sum_i p(x, y_i), \quad p_Y(y) = \sum_i p(x_i, y),$$

where x is one of the possible values of X, and y is one of the possible values of Y.

In the case of more than two random variables, $X_1, X_2, \ldots, X_m$, the marginal pmf of each variable X_i can be obtained by summing the joint pmf over all possible values of the other random variables. For example, it can be shown that, if the joint distribution of X_1, X_2, X_3 is multinomial (see Example 4.3.2), then each X_i has a binomial distribution.

4.3.3 Conditional Probability Mass Functions

The concept of a *conditional distribution* of a discrete random variable is an extension of the concept of conditional probability of an event.

For a discrete (X, Y), if x is a possible value of X, i.e. $p_X(x) > 0$, the concept of conditional probability provides answers to questions regarding the value of Y, given that $X = x$ has been observed. For example, the conditional probability that Y takes the value y given that $X = x$ is

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x, y)}{p_X(x)},$$

where $p(x, y)$ is the joint pmf of (X, Y) and $p_X(x)$ is the marginal pmf of X. The above relation follows simply from the definition of conditional probability, but when we think of it as a function of y with x being kept fixed, we call it the *conditional pmf* of Y given the information that X has taken the value x.

Definition 4.3.2. *Let (X, Y) be discrete, and let $\mathcal{S}_Y = \{y_1, y_2, \ldots\}$ be the sample space of Y. If x is a possible value of X, i.e. $p_X(x) > 0$, where p_X is the marginal pmf of X, then*

$$p_{Y|X}(y_j|x) = \frac{p(x, y_j)}{p_X(x)}, \quad j = 1, 2, \ldots,$$

where $p(x, y)$ is the joint pmf of (X, Y), is called the **conditional pmf** *of Y given that $X = x$. An alternative notation is $p_{Y|X=x}(y_j)$*

When the joint pmf of (X, Y) is given in a table form, as in Examples 4.3.1, and 4.3.3 the conditional pmf of Y given $X = x$ is found simply by dividing the joint probabilities in the row that corresponds to the X-value x by the marginal probability that $X = x$.

Example 4.3.4. Consider the discrete (X, Y) of Example 4.3.1, and find the conditional pmf of Y given $X = 0$.

Solution. The conditional pmf of Y given that $X = 0$ is obtained by dividing each joint probability in the row that corresponds to the X-value 0, by the marginal probability that $X = 0$. This gives,

y	0	1	2	3
$p_{Y\|X}(y\|X = 0)$	.9333	.0333	.0222	.0111

For example, the values .9333 is obtained as

$$p_{Y|X}(0|X = 0) = \frac{p(0, 0)}{p_X(0)} = \frac{.84}{.9} = .9333.$$

Proposition 4.3.2.

1. *The conditional pmf is a proper pmf. Thus, for each x with $p_X(x) > 0$,*

$$p_{Y|X}(y_j|x) \geq 0, \quad \textit{for all } j = 1, 2, \ldots, \quad \textit{and} \sum_j p_{Y|X}(y_j|x) = 1.$$

2. *If we know the conditional pmf of Y given $X = x$, for all values x in the sample space of X (i.e. if we know $p_{Y|X}(y_j|x)$, for all $j = 1, 2, \ldots,$ and for all possible values x of X), and also know the marginal pmf of X, then the joint pmf of (X, Y) can be obtained as*

$$p(x, y) = p_{Y|X}(y|x) p_X(x). \tag{4.3.2}$$

3. The marginal pmf of Y can be obtained as

$$p_Y(y) = \sum_{x \ in \ \mathcal{S}_X} p_{Y|X}(y|x)p_X(x). \tag{4.3.3}$$

Relation (4.3.2), which follows directly from the definition of conditional pmf, is useful because it is often easier to specify the marginal distribution of X and the conditional distribution of Y given X, than to specify directly the joint distribution of (X, Y); see Example 4.3.5 below. Relation (4.3.3) follows by summing (4.3.2) over all possible values X; it is also a direct consequence of the Law of Total Probability.

Example 4.3.5. It is known that, with probability 0.6, a new lap-top owner will install wireless internet connection at home within a month. Let X denote the number (in hundreds) of new lap-top owners in a week from a certain region, and let Y denote the number among them who install wireless connection at home within a month. Suppose that the pmf of X is

x	0	1	2	3	4
$p_X(x)$	0.1	0.2	0.3	0.25	0.15

Find the joint distribution of (X, Y). Find the probability that $Y = 4$.

Solution. According to part 2 of Proposition 4.3.2, since the marginal distribution of X is known, the joint distribution of (X, Y) can be specified if $p_{Y|X}(y|x)$ is known for all possible values x of X. Given that $X = x$, however, Y has the binomial distribution with $n = x$ trials and probability of success $p = 0.6$, so that

$$p_{Y|X}(y|x) = \binom{x}{y} 0.6^y 0.4^{x-y}.$$

For example, if $X = 3$ then the probability that $Y = 2$ is

$$p_{Y|X}(2|3) = \binom{3}{2} 0.6^2 0.4^{3-2} = 0.4320.$$

Next, according to part 3 of Proposition 4.3.2,

$$p_Y(4) = \sum_{x=0}^{4} p_{Y|X}(4|x)p_X(x) = 0 + 0 + 0 + 0 + 0.6^4 \times 0.15 = 0.0194.$$

We conclude this subsection by pointing out that, since the conditional pmf is a proper pmf, it is possible to consider its expected value and its variance. These are called the **conditional expected value** and **conditional variance**, respectively. As an example,

Example 4.3.6. Consider the discrete (X, Y) of Example 4.3.1. Calculate the conditional expected value of Y given that $X = 0$.

Solution. Using the conditional pmf that we found in Example 4.3.4, we obtain,

$$E(Y|X = 0) = 0 \times (.9333) + 1 \times (.0333) + 2 \times (.0222) + 3 \times (.0222) = .111.$$

Compare this with the unconditional, or marginal, expected value of Y, which is $E(Y) = .148$.

4.3.4 Independence

The notion of independence of random variables is an extension of the notion of independence of events. For example, the events $A = [X = x]$ and $B = [Y = y]$ are independent if

$$P(X = x,\ Y = y) = P(X = x)P(Y = y).$$

This follows directly from the definition of independent events, since $P(X = x,\ Y = y)$ means $P([X = x] \cap [Y = y])$. If the above equality holds for all possible values, x, of X, and all possible values, y, of Y, then X, Y are called *independent*. In particular, we have the following definition.

Definition 4.3.3. *The discrete random variables X, Y are called* **independent** *if*

$$p_{X,Y}(x, y) = p_X(x)p_Y(y),\ \textit{for all } x, y,$$

where $p_{X,Y}$ is the joint pmf of (X, Y) and p_X, p_Y are the marginal pmf's of X, Y, respectively. In other words, X, Y are independent if the events $A = [X = x]$ and $B = [Y = y]$ are independent for all x, y. Similarly, the random variables $X_1, X_2, \ldots, X_n$ are called **independent** *if*

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n).$$

If $X_1, X_2, \ldots, X_n$ are independent and also have the same distribution (which is the case of a simple random sample from an infinite/hypothetical population) they are called **independent and identically distributed**, *or* **iid** *for short.*

The next proposition is a collection of some statements that are equivalent to the statement of independence of two random variables.

Proposition 4.3.3. *Each of the following statements implies, and is implied by, the independence of the random variables X, Y.*

1. $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, *for all* x, y, *where* $F_{X,Y}$ *is the joint cdf of* (X,Y) *and* F_X, F_Y *are the marginal cdfs of* X, Y, *respectively.*

2. $p_{Y|X}(y|x) = p_Y(y)$, *where* $p_{Y|X}(y|x)$ *is the conditional probability of* $[Y = y]$ *given that* $[X = x]$, *and* p_Y *is the marginal pmf of* Y.

3. $p_{Y|X}(y|x)$ *does not depend on* x. *In other words, the conditional pmf of* Y *given* $X = x$ *is the same for all values* x *that* X *might take.*

4. $p_{X|Y}(x|y) = p_X(x)$, *where* $p_{X|Y}(x|y)$ *is the conditional probability of* $[X = x]$ *given that* $[Y = y]$, *and* p_X *is the marginal pmf of* X.

5. $p_{X|Y}(x|y)$ *does not depend on* y. *In other words, the conditional pmf of* X *given* $Y = y$ *is the same for all values* y *that* Y *might take.*

6. *Any event associated with the random variable* X *is independent from any event associated with the random variable* Y, *i.e.* $[X \in A]$ *is independent from* $[Y \in B]$, *where* A *is any subset of the sample space of* X, *and* B *is any subset of the sample space of* Y.

7. *For any two functions* h *and* g, *the random variables* $h(X)$, $g(Y)$ *are independent.*

Example 4.3.7. Consider the joint distribution of (X,Y) given in Example 4.3.1. Are X, Y independent?

Solution. Here X, Y are not independent since

$$p(0,0) = .84 \neq p_X(0)p_Y(0) = (.9)(.91) = .819.$$

Example 4.3.8. Are the X, Y of Example 4.3.5 independent?

Solution. Here X, Y are not independent since

$$p_{X,Y}(3,4) = p_{Y|X}(4|3)p_X(3) = 0 \times p_X(3) = 0 \neq p_X(3)p_Y(4) = 0.25 \times 0.0194 = 0.0049.$$

Example 4.3.9. A system is made up of two components connected in parallel. Let A, B denote the two components. Thus, the system fails if both components fail. Let the random variable X take the value 1 if component A works, and the value 0 of it does not. Similarly, Y takes the value 1 if component B works, and the value 0 if it does not. From the repair history of the system it is known that the joint pmf of (X,Y) is

		Y		
		0	1	
	0	0.0098	0.9702	0.98
X				
	1	0.0002	0.0198	0.02
		0.01	0.99	1.0

Are X, Y independent?

Solution. In this example, it can be seen that $p_{X,Y}(x,y) = p_X(x)p_Y(y),$ for all x, y. Thus, X, Y are independent. Alternatively, calculation of the conditional pmf of Y given X reveals that the the conditional pmf of Y given $X = x$ does not depend on the particular value x of X, and thus, in accordance with part 3 of Proposition 4.3.3, X, Y are independent.

4.3.5 Exercises

1. In a gasoline station there are two self-service pumps and two full-service pumps. Let X denote the number of self-service pumps used at a particular time and Y the number of full-service pumps in use at that time. The joint pmf $p(x, y)$ of (X, Y) appears in the next table.

			y	
		0	1	2
	0	.10	.04	.02
x	1	.08	.20	.06
	2	.06	.14	.30

   (a) Find the probability $P(X \leq 1, Y \leq 1)$.

   (b) Compute the marginal pmf of X and Y.

   (c) Compute the mean value and variance of X and Y.

2. Suppose that X and Y are discrete random variables taking values -1, 0, and 1, and that their joint pmf is given by

			X	
	$p(x,y)$	-1	0	1
	-1	0	1/9	2/9
Y	0	1/9	1/9	1/9
	1	2/9	1/9	0

(a) Find the marginal pmf of X.

(b) Are X and Y independent? Why?

(c) Find $E(X)$.

(d) Find the probability $P(X \geq 0 \text{ and } Y \geq 0)$.

3. The following table shows the joint probability distribution of X, the amount of drug administered to a randomly selected laboratory rat, and Y, the number of tumors present on the rat.

		y			
	$p(x, y)$	0	1	2	
	0.0 mg/kg	.388	.009	.003	.400
x	1.0 mg/kg	.485	.010	.005	.500
	2.0 mg/kg	.090	.008	.002	.100
		.963	.027	.010	1.000

(a) Are X and Y independent random variables? Explain.

(b) What is the probability that a randomly selected rat has: (i) one tumor, and (ii) at least one tumor?

(c) For a randomly selected rat in the 1.0 mg/kg drug dosage group, what is the probability that it has: (i) no tumor, (ii) at least one tumor?

(d) What is the conditional pmf of the number of tumors for a randomly selected rat in the 1.0 mg/kg drug dosage group?

(e) What is the expected number of tumors for a randomly selected rat in the 1.0 mg/kg drug dosage group?

4. Let X take the value 0 if a child under 5 uses no seat belt, 1 if it uses adult seat belt, and 2 if it uses child seat. And let Y take the value 0 if a child survived a motor vehicle accident, and 1 if it did not. An extensive study undertaken by the National Highway Traffic Safety Administration resulted in the following conditional distributions of Y given $X = x$:

y	0	1
$p_{Y\|X=0}(y)$	0.69	0.31
$p_{Y\|X=1}(y)$	0.85	0.15
$p_{Y\|X=2}(y)$	0.84	0.16

while the marginal distribution of X is

x	0	1	2
$p_X(x)$	0.54	0.17	0.29

(a) Use the table of conditional distributions of Y given $X = x$ to conclude whether or not X and Y independent. Justify your answer.

(b) Tabulate the joint distribution of X and Y.

(c) Use the joint distribution of X and Y to conclude whether or not X and Y are independent. Justify your answer.

(d) Find the marginal distribution of Y.

5. A popular local restaurant offers dinner entrees in two price ranges: $7.50 meals and $10.00 meals. From past experience, the restaurant owner knows that 65% of the customers order from the $7.50 meals, while 35% order from the $10.00 meals. Also from past experience, the waiters and waitresses know that the customers always tip either $1.00, $1.50, or $2.00 per meal. The table below gives the joint probability distribution of the price of the meal ordered and the tip left by a random customer.

		Tip Left			
		$1.00	$1.50	$2.00	
Meal Price	$7.50	.455	.195	0	.65
	$10.00	.035	.210	.105	.35
					1

(a) Find the marginal mean value and variance of the variable ”price of meal”. Do the same thing for the variable ”tip left”.

(b) Is the amount of tip left independent of the price of the meal? Justify your answer.

6. Consider selecting two products from a batch of 10 products. Suppose that the batch contains 3 defective and 7 non-defective products. Let X take the value 1 or 0 as the first selection from the 10 products is defective or not. Let Y take the value 1 or 0 as the second selection (from the nine remaining products) is defective or not.

(a) Find the marginal distribution of X.

(b) Find the conditional distributions of Y given each of the possible values of X.

(c) Use your two results above to find the joint distribution of X and Y.

(d) Find the conditional variance of Y given $X = 1$.

(e) Find the marginal distribution of Y. Is it the same as that of X?

4.4 The Joint Probability Density Function

4.4.1 Definition and Basic Properties

In the univariate case, the pdf of a random variable X is a function $f(x)$ such that the area under the curve it defines equals one, and the probability that X takes a value in any given interval equal the area under the curve and above the interval. In the bivariate case, the pdf is a function of two arguments, $f(x,y)$, which defines a surface in the three dimensional space, and probabilities are represented as volumes under this surface. More precisely we have the following

Definition 4.4.1. *The* **joint** *or* **bivariate** *density function of the continuous* (X,Y) *is a non-negative function* $f(x,y)$ *such that*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ dx\ dy = 1,$$

i.e. the volume under the surface that $f(x,y)$ *defines is one, and the probability that* (X,Y) *will take a value in a region* A *of the plane is*

$$P((X,Y)\in A) = \int\int_A f(x,y)\ dx\ dy, \qquad (4.4.1)$$

or, if A *is a rectangle,* $A = \{(x,y)|a \le x \le b,\ c \le y \le d\}$,

$$P(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f(x,y)\ dy\ dx. \qquad (4.4.2)$$

Similarly, the **joint** *or* **multivariate** *probability density function of the continuous* $(X_1, X_2, \ldots, X_n)$ *is a non-negative function* $f(x_1, x_2, \ldots, x_n)$ *such that*

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,x_2,\ldots,x_n)\ dx_1\cdots dx_n = 1 \quad \textit{and}$$

$$P((X_1,X_2,\ldots,X_n)\in B) = \int\cdots\int_B f(x_1,x_2,\ldots,x_n)\ dx_1\cdots dx_n, \qquad (4.4.3)$$

where B *is a region in* n*-dimensions.*

The definition implies that the cdf, F, of (X,Y) can be obtained from the pdf by

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)\ dt\ ds. \qquad (4.4.4)$$

Conversely, the pdf of (X,Y) can be derived from the cdf by differentiating:

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y). \qquad (4.4.5)$$

Formulas analogous to (4.4.4) and (4.4.5) exist in the multivariate case.

Example 4.4.1. Using the joint cdf of the bivariate uniform distribution in the unit rectangle $[0,1] \times [0,1]$, which was given in Example 4.2.1, find the corresponding joint pdf.

Solution. Using (4.4.5) and the form of the bivariate cdf of the uniform distribution in the unit rectangle, we obtain that $f(x,y) = 0$, if (x,y) is outside the unit rectangle, and

$$f(x,y) = 1, \quad \text{for } 0 \leq x, y \leq 1.$$

Example 4.4.2. Consider the bivariate density function

$$f(x,y) = \frac{12}{7}(x^2 + xy), \quad 0 \leq x, y \leq 1.$$

Find the probability that $X > Y$.

Solution. The desired probability can be found by integrating f over the region $A = \{(x,y)|0 \leq y \leq x \leq 1\}$. Note that A is not a rectangle, so we use (4.4.1):

$$P(X > Y) = \frac{12}{7}\int_0^1 \int_0^x (x^2 + xy)\ dy\ dx = \frac{9}{14}.$$

Example 4.4.3. Consider the bivariate distribution of Example 4.4.2. Find a) the probability that $X \leq 0.6$ and $Y \leq 0.4$, and b) the joint cdf of (X, Y).

Solution. Since the region specified in part a) is a rectangle, we use (4.4.2):

$$P(X \leq 0.6,\ Y \leq 0.4) = F(0.6, 0.4) = \frac{12}{7}\int_0^{0.6} \int_0^{0.4} (x^2 + xy)\ dy\ dx = 0.0741.$$

For part b) we use (4.4.4):

$$F(x,y) = \int_0^x \int_0^y \frac{12}{7}(s^2 + st)\ dt\ ds = \frac{12}{7}\left(\frac{x^3 y}{3} + \frac{x^2 y^2}{4}\right).$$

Note that an easy differentiation verifies that $f(x,y) = \frac{\partial^2}{\partial x \partial y}F(x,y)$, is indeed the joint pdf given in Example 4.4.2.

4.4.2 Marginal Probability Density Functions

In the discrete case the marginal pmf, $p_X(x)$, of X is obtained (for any of the possible values x of X) by summing the bivariate pmf, $p_{X,Y}(x,y)$, of X, Y over all possible values of Y, while keeping x fixed; see Proposition 4.3.1. In the continuous case, the corresponding task is achieved by replacing summation by integration. In particular, we have

Proposition 4.4.1. *Let X, Y have bivariate pdf f. Then the* **marginal pdf**, *$f_X(x)$, of X is obtained (for any x) by integrating $f(x, y)$ over y, while keeping x fixed. Similarly for the marginal pdf of Y. Thus,*

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\ dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\ dx \tag{4.4.6}$$

Example 4.4.4. Find the marginal pdf of X and Y from their joint bivariate uniform distribution given in Example 4.4.1.

Solution. From (4.4.6), we have

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\ dy = \int_0^1 1dy = 1, \quad \text{for } 0 \leq x \leq 1, \text{ and } \ f_X(x) = 0, \quad \text{for } x \notin [0, 1].$$

Similarly, the marginal pdf of Y is obtained by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\ dx = \int_0^1 1dx = 1, \quad \text{for } 0 \leq y \leq 1, \text{ and } \ f_Y(y) = 0, \quad \text{for } y \notin [0, 1].$$

Thus, each of X and Y have a uniform in $[0, 1]$ distribution, in agreement with Example 4.2.1.

Example 4.4.5. Find the marginal pdf of X and Y from their joint bivariate distribution given in Example 4.4.2.

Solution. From (4.4.6), we have

$$f_X(x) = \int_0^1 \frac{12}{7}(x^2 + xy)dy = \frac{12}{7}x^2 + \frac{6}{7}x, \quad \text{for } 0 \leq x \leq 1, \text{ and } \ f_X(x) = 0, \quad \text{for } x \notin [0, 1].$$

Similarly, the marginal pdf of Y is given by

$$f_Y(y) = \int_0^1 \frac{12}{7}(x^2 + xy)dx = \frac{4}{7} + \frac{6}{7}y.$$

4.4.3 Conditional Probability Density Functions

In analogy with the definition in the discrete case, if (X, Y) are continuous with joint pdf f, and marginal pdfs f_X, f_Y, then the **conditional pdf** of Y given $X = x$ is defined to be

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_X(x)}, \tag{4.4.7}$$

if $f_X(x) > 0$.

Remark 4.4.1. *In the case of discretized measurements, $f_{Y|X=x}(y)\Delta y$ approximates $P(y \leq Y \leq y + \Delta y|x \leq X \leq x + \Delta x)$, as can be seen from*

$$\begin{aligned}
&P(y \leq Y \leq y + \Delta y|x \leq X \leq x + \Delta x) \\
= \ &\frac{P(y \leq Y \leq y + \Delta y, x \leq X \leq x + \Delta x)}{P(x \leq X \leq x + \Delta x)} \\
= \ &\frac{f(x, y) \ \Delta x \ \ \Delta y}{f_X(x) \ \Delta x} = \frac{f(x, y)}{f_X(x)} \ \Delta y.
\end{aligned}$$

Proposition 4.4.2. *1. The conditional pdf is a proper pdf. Thus, for each x with $f_X(x) > 0$,*

$$f_{Y|X=x}(y) \geq 0, \quad \text{for all } y, \quad \text{and} \quad \int_{-\infty}^{\infty} f_{Y|X=x}(y) \ dy = 1.$$

2. If we know the conditional pdf of Y given $X = x$, for all values x that satisfy $f_X(x) > 0$, and also know the marginal pdf of X, then the joint pdf of (X, Y) can be obtained as

$$f(x, y) = f_{Y|X=x}(y) f_X(x). \tag{4.4.8}$$

3. The marginal pdf of Y can be obtained as

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y) f_X(x) dx, \tag{4.4.9}$$

Relation (4.4.8) of the above proposition follows directly from the definition of conditional pdf; is useful because it is often easier to specify the marginal distribution of X and the conditional distribution of Y given X, than to specify directly the joint distribution of (X, Y); see Example 4.4.6 below. Relation (4.4.9), which follows by integrating (4.4.8) over x, is the **Law of Total Probability** for continuous variables.

Example 4.4.6. Let Y denote the age of a tree, and let X denote the tree's diameter (in cm) at breast height. Suppose it is known that, for a particular type of tree, the conditional distribution of Y given that $X = x$ is normal with mean $\mu_{Y|X=x} = 5 + 0.33x$, and standard deviation $\sigma_{Y|X=x} = 0.4$, for all x. That is,

$$Y|X = x \ \sim \ N(5 + 0.33x, 0.4^2).$$

Suppose also that in a given forested area, the diameter of the trees ranges from 10-80 cm, and the distribution of X is well approximated by the uniform on $[10, 80]$ distribution. Find the joint pdf of (X, Y).

Solution. From relation (4.4.8) we have

$$f(x,y) = f_{Y|X=x}(y)f_X(x) = \frac{1}{\sqrt{2\pi}0.4}\exp\left(-\frac{(y-5-0.33x)^2}{2\times 0.16}\right)\frac{1}{70}, \quad 10 < x < 80.$$

Example 4.4.7. For a cylinder selected at random from the manufacturing line, let X=height, Y=radius. Suppose X, Y have a joint pdf

$$f(x,y) = \begin{cases} \frac{3}{8}\frac{x}{y^2} & \text{if } 1 \le x \le 3, \ \frac{1}{2} \le y \le \frac{3}{4} \\ 0 & \text{otherwise.} \end{cases}$$

Find $f_X(x)$ and $f_{Y|X=x}(y)$.

Solution. According to the formulae,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{.5}^{.75}\left(\frac{3}{8}\frac{x}{y^2}\right)dy = \frac{x}{4}$$

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{3}{2}\frac{1}{y^2}.$$

4.4.4 Independence

Definition 4.4.2. *The continuous random variables X, Y are called* **independent** *if*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \textit{holds for all } x, y,$$

where $f_{X,Y}$ is the joint pdf of (X,Y), and f_X, f_Y are the marginal pdfs of X, Y, respectively. Similarly, the random variables $X_1, X_2, \ldots, X_n$ are called **independent** *if*

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n).$$

If $X_1, X_2, \ldots, X_n$ are independent and also have the same distribution (which is the case of a simple random sample from an infinite/hypothetical population) they are called **independent and identically distributed**, *or* **iid** *for short.*

The next proposition is a collection of some statements that are equivalent to the statement of independence of two random variables.

Proposition 4.4.3. *Each of the following statements implies, and is implied by, the independence of the continuous random variables X, Y.*

1. *$F_{X,Y}(x,y) = F_X(x)F_Y(y)$, for all x, y, where $F_{X,Y}$ is the joint cdf of (X,Y) and F_X, F_Y are the marginal cdfs of X, Y, respectively.*

2. *$f_{Y|X=x}(y) = f_Y(y)$, where $f_{Y|X=x}$ is the conditional pdf of Y given that $[X = x]$, and f_Y is the marginal pmf of Y.*

3. *$f_{Y|X=x}(y)$ does not depend on x. In other words, the conditional pdf of Y, given $X = x$, is the same for all values x that X might take.*

4. *$f_{X|Y=y}(x) = f_X(x)$, where $f_{X|Y=y}(x)$ is the conditional pdf of X given that $[Y = y]$, and f_X is the marginal pdf of X.*

5. *$f_{X|Y=y}(x)$ does not depend on y. In other words, the conditional pdf of X, given $Y = y$, is the same for all values y that Y might take.*

6. *Any event associated with the random variable X is independent from any event associated with the random variable Y, i.e. $[X \in A]$ is independent from $[Y \in B]$, where A is any subset of the sample space of X, and B is any subset of the sample space of Y.*

7. *For any two functions h and g, the random variables $h(X)$, $g(Y)$ are independent.*

Example 4.4.8. Consider the joint distribution of X=height, and Y=radius given in Example 4.4.7. Are X and Y independent?

Solution. The marginal pdf of X was derived in Example 4.4.7. That of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_1^3 \left(\frac{3}{8}\frac{x}{y^2}\right)dx = \frac{3}{2}\frac{1}{y^2}.$$

Finally, joint pdf of (X, Y) is given in Example 4.4.7. From this it can be verified that

$$f(x,y) = f_X(x)f_Y(y).$$

Thus, X and Y are independent. An alternative method of checking the independence of X and Y, is to examine the conditional pdf $f_{Y|X=x}(y)$. In Example 4.4.7 it was derived that $f_{Y|X=x}(y) = \frac{3}{2}\frac{1}{y^2}$. Since this is constant in x, we conclude that X and Y are independent, according to part 3 of Proposition 4.4.3. Finally, as an application of part 7 of Proposition 4.4.3, X and Y^2 are also independent.

4.4.5 Exercises

1. Let the rv's X and Y have the joint pdf given below:

$$f(x,y) = kxy^2, \text{ for } 0 \le x \le 2,\ 0 \le y \le 3.$$

(a) Find the constant k.

(b) Find the marginal pdf's of X and Y.

(c) Are X and Y independent?

2. Let the rv's X and Y have the joint pdf given below:

$$f(x, y) = kxy^2, \quad \text{for} \ \ 0 \leq x \leq 2, \ x \leq y \leq 3.$$

(a) Find the constant k.

(b) Find the marginal pdf's of X and Y.

(c) Are X and Y independent?

3. Let the rv's X and Y have the joint pdf given below:

$$f(x, y) = \begin{cases} 2e^{-x-y} & 0 \leq x \leq y < \infty \\ 0 & \textit{otherwise} \end{cases}$$

(a) Find $P(X + Y \leq 3)$.

(b) Find the marginal pdf's of Y and X.

(c) Are X and Y independent? Justify your answer.

4. Consider the following situation from architectural design. The width of a heating system vent is mechanically controlled, and essentially allows a flow Y between 0 and ∞. The actual flow through the vent, X, is either inward or outward but can never exceed Y in absolute value. The joint pdf of X and Y is given by

$$f(x, y) = \begin{cases} c(y^2 - x^2)e^{-y} & \text{if } -y \leq x \leq y \\ 0 & \text{otherwise.} \end{cases}$$

(a) Sketch the support of f, i.e. the set $\{(x, y) : f(x, y) > 0\}$.

(b) Find c.

(c) Find the marginal density of Y. In which family of distributions models does this distribution belong?

(d) Find the expectation of X.

(e) Are X and Y independent?

5. A type of steel has microscopic defects which are classified on continuous scale from 0 to 1, with 0 the least sever and 1 the most sever. This is called the defect index. Let X and Y be the static force at failure and the defect index for a particular type of structural

member made of this steel. For a member selected at random, these are jointly distributed random variables with joint pdf

$$f(x,y) = \begin{cases} 24x & \text{if } 0 \le y \le 1-2x \text{ and } 0 \le x \le .5 \\ 0 & \text{otherwise} \end{cases}$$

(a) Draw the support of this pdf, i.e. the region of (x, y) values where $f(x, y) > 0$.

(b) Are X and Y independent? Answer this question without first computing the marginal pdfs of X and Y. Justify your answer.

(c) Find each of the following: f_X, f_Y, $E(X)$, and $E(Y)$.

6. John and his trainer Yvonne have agreed to meet between 6 A.M. and 8 A.M. for a workout but will aim for 6 A.M. Let X = # of hours that John is late, and Y = # of hours that Yvonne is late. Suppose that the joint distribution of X and Y is

$$f(x,y) = \begin{cases} \frac{1}{4} & 0 \le x \le 2, 0 \le y \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

(a) Determine the marginal probability density function of X. Do the same for Y. [If you can, guess the marginal pdf of Y without any additional calculations.]

(b) Compute $E(X)$ and $\text{Var}(X)$. Do the same for Y. [If you can, guess the $E(Y)$ and $\text{Var}(Y)$ without any additional calculations.]

(c) Are X and Y independent? Justify your answer.

4.5 Statistics and Sampling Distributions

A function $h(X_1, \ldots, X_n)$ of random variables will be called a **statistic**. Statistics are, of course, random variables and, as such, they have a distribution. The distribution of a statistic is known as its **sampling distribution**. As in the univariate case, we will see in this section that the expected value and variance of a function of random variables (statistic) can be obtained without having to first obtain its distribution. This is a very useful/convenient method of calculating the expected values and variances since the sampling distribution of a statistic is typically difficult to obtain; this will be demonstrated in Section 5.4. Thus, in this section we will only be concerned with finding the mean and variance of the sampling distribution of a statistic.

Proposition 4.5.1. *1. Let (X, Y) be discrete with joint pmf $p_{X,Y}$. The expected value of a function, $h(X, Y)$, of (X, Y) are computed by*

$$E[h(X,Y)] = \sum_x \sum_y h(x,y) p_{X,Y}(x,y), \tag{4.5.1}$$

and the variance of $h(X,Y)$ *is computed by*

$$\sigma^2_{h(X,Y)} = E[h^2(X,Y)] - [E[h(X,Y)]]^2,$$

where, according to (4.5.1), $E[h^2(X,Y)] = \sum_x \sum_y h^2(x,y)p_{X,Y}(x,y)$.

2. *Let* (X,Y) *be continuous with joint pdf* $f_{X,Y}$. *The expected value of a function,* $h(X,Y)$, *of* (X,Y) *are computed by*

$$E[h(X,Y)] = \int_{\infty}^{\infty} \int_{\infty}^{\infty} h(x,y) f_{X,Y}(x,y)\ dx\ dy, \tag{4.5.2}$$

and the variance of $h(X,Y)$ *is computed by*

$$\sigma^2_{h(X,Y)} = E[h^2(X,Y)] - [E[h(X,Y)]]^2,$$

where, according to (4.5.2), $E[h^2(X,Y)] = \int_{\infty}^{\infty} \int_{\infty}^{\infty} h(x,y) f_{X,Y}(x,y)\ dx\ dy$.

Note that the formulae in the continuous case are similar to those in the discrete case, except for the fact that summation is replaced by integration.

The formulae in Proposition 4.5.1 extend directly to functions of more than two random variables. For example, in the discrete case, the expected value of the statistic $h(X_1, \ldots, X_n)$ is computed by

$$E[h(X_1, \ldots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} h(x_1, \ldots, x_n) p(x_1, \ldots, x_n),$$

where p denotes the joint pmf of $X_1, \ldots, X_n$, while in the continuous case, the expected value of $h(X_1, \ldots, X_n)$ is computed by

$$E[h(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Example 4.5.1. Consider the joint distribution of (X,Y) given in Example 4.3.1. Find the expected value of the total number of errors, T, that the robot makes on a car.

Solution. Here $T = h(X,Y) = X + Y$, where X is the number of defective welds, and Y be the number of improperly tightened bolts per car. Thus

$$\begin{aligned} E(T) &= \sum_x \sum_y (x+y) p(x,y) \\ &= 0(.84) + 1(.03) + 2(.02) + 3(.01) \\ &+ 1(.06) + 2(.01) + 3(.008) + 4(.002) \\ &+ 2(.01) + 3(.005) + 4(.004) + 5(.001) \\ &= .268. \end{aligned}$$

Example 4.5.2. Consider the joint distribution of X=height, Y= radius given in Example 4.4.7. Find the expected value of the volume of a cylinder.

Solution: The volume of the cylinder is given in terms of the height (X) and radius (Y) by the function $h(X,Y) = \pi Y^2 X$. Thus,

$$\begin{aligned} E[h(X,Y)] &= \int_1^3 \int_{.5}^{.75} \pi y^2 x \left(\frac{3}{8} \frac{x}{y^2} \right) dy\ dx \\ &= \frac{13}{16}\pi. \end{aligned}$$

We conclude this subsection with a result about the expected value of a product of independent random variables.

Proposition 4.5.2. *If X and Y are independent, then*

$$E(XY) = E(X)E(Y).$$

In general, if $X_1, \ldots, X_n$ are independent,

$$E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n).$$

As an application of Proposition 4.5.2, the expected value of the volume $h(X,Y) = \pi Y^2 X$, which was obtained in Example 4.5.2, can also be calculated as

$$E[h(X,Y)] = \pi E(Y^2)E(X),$$

since, as shown in Example 4.4.8, X and Y are independent, and thus, by part 5 of Proposition 4.4.3, X and Y^2 are also independent.

4.5.1 Exercises

1. Consider the information given in Exercise 4.3.5-5 regarding the price of dinner entrees and the tip left. Find the expected value and variance of the total price paid (meal plus tip) by a random customer.

2. Consider the information given in Exercise 4.3.5-1 on the joint distribution of the number, X, of self-service pumps used at a particular time, and Y, the number of full-service pumps in use at that time. Suppose that the profit to the gas station at a particular time is $8X + 10Y$ dollars. Find the expected value of the profit.

3. Suppose that X and Y are discrete random variables taking values -1, 0, and 1, and that their joint pmf is given by

			X	
$p(x, y)$		-1	0	1
	-1	0	1/9	2/9
Y	0	1/9	1/9	1/9
	1	2/9	1/9	0

Find the expected value $E(\max\{X, Y\})$, of the maximum of X and Y.

4.6 Parameters of a Multivariate Distribution

4.6.1 The Regression Function

It is often of interest to know how the expected value of one variable changes when we have observed the value that another variable has taken. This knowledge is the fundamental ingredient in predicting one variable from another. In the study where X is the velocity and Y is the stopping distance of an automobile, and in the study where X is the diameter at breast height and Y is the age of a tree, both of which were mentioned in Section 5.1, knowing how the expected value of Y changes with X, and thus being able to predict Y from X, would be of primary interest.

Definition 4.6.1. *For the bivariate random variable (X, Y), the conditional expected value of Y given that $X = x$, i.e.*

$$\mu_{Y|X}(x) = E(Y|X = x),$$

when considered as a function of x, is called the **regression function** *of Y on X.*

Example 4.6.1. Consider Example 4.3.1, regarding the errors a robot makes per car. Thus, X be the number of defective welds, and Y be the number of improperly tightened bolts per car. Calculate the regression function of Y on X.

Solution. In Example 4.3.4 we obtained the conditional pmf of Y given that $X = 0$ as

y	0	1	2	3
$p_{Y\|X}(y\|X = 0)$	0.9333	0.0333	0.0222	0.0111

and in Example 4.3.6, we computed the conditional expectation of Y given that $X = 0$ as $E(Y|X = 0) = 0.111$. Repeating these calculations, conditioning first on $X = 1$ and then on $X = 2$, we obtain

y	0	1	2	3
$p_{Y\|X}(y\|X = 1)$	0.75	0.125	0.1	0.025

so that $E(Y|X = 1) = 0.4$, and

y	0	1	2	3
$p_{Y\|X}(y\|X = 2)$	0.5	0.25	0.2	0.05

from which we obtain $E(Y|X = 2) = 0.8$. Summarizing the above calculations of the conditional expectation of Y in a table, we obtain the regression function of Y on X:

x	0	1	2
$\mu_{Y\|X}(x)$	0.111	0.4	0.8

The information that this regression function makes visually apparent, and which was not easily discernable from the joint probability mass function, is that in a car with more defective welds, you can expect to experience more improperly tightened bolts.

Example 4.6.2. Suppose (X, Y) have joint pdf

$$f(x, y) = \begin{cases} 24xy & 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the regression function of Y on X.

Solution. First we find the conditional pdf of Y given $X = x$. We have

$$f_X(x) = \int_0^{1-x} 24xy \; dy = 12x(1 - x)^2.$$

Thus,

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_X(x)} = 2\frac{y}{1-x}^{2}.$$

Using this conditional pdf, we calculate the regression function of Y on X as

$$E(Y|X = x) = \int_0^{1-x} y f_{Y|X=x}(y) \; dy = \frac{2}{3}(1 - x).$$

Proposition 4.6.1 (Law of Total Probability for Expectations). *A weighted average of the conditional expected values of Y, with weights equal to the marginal probabilities of X, gives the unconditional expected value of Y. That is, in the discrete case,*

$$E(Y) = \sum_{x \in \mathcal{S}_X} E(Y|X = x)p_X(x).$$

The corresponding result in the continuous case is obtained by replacing summation with integration, and the pmf with the pdf:

$$E(Y) = \int_{-\infty}^{\infty} E(Y|X = x)f_X(x)\ dx.$$

Example 4.6.3. Use the regression function obtained in Example 4.6.1, and the marginal pmf of X from Example 4.3.3, to verify the Law of Total Probability for Expectations.

Solution. Using the information from the aforementioned two examples, we have

$$\begin{aligned} & E(Y|X = 0)p_X(0) \ + \ E(Y|X = 1)p_X(1) \ + \ E(Y|X = 2)p_X(2) \\ = \ & 0.111 \times 0.9 \ + \ 0.4 \times 0.08 \ + \ 0.8 \times 0.02 \ = \ 0.148 \ = \ E(Y). \end{aligned}$$

4.6.2 The Covariance

When two random variables X and Y are not independent, they are dependent. We say that X, Y are **positively dependent** if "large" values of X are associated with "large" values of Y, and "small" values of X are associated with "small" values of Y. For example, the variables X=height and Y=weight of a randomly selected adult male, are positively dependent. In the opposite case we say that X, Y are **negatively dependent**. If the variables are either positively or negatively dependent, their dependence is called **monotone**.

Covariance is a parameter of the joint distribution of two variables whose value (or, rather its sign) identifies the dependence as being positive or negative.

Definition 4.6.2. *The* **covariance** *of X and Y, denoted by $Cov(X, Y)$ or σ_{XY}, is defined as*

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y,$$

where μ_X and μ_Y are the marginal expected values of X and Y, respectively.

The second equality in the above definition is a computational formula for the covariance, similar to the computational (short-cut) formula, $\sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$, for the variance. It is also worth pointing out that the covariance of a random variable with itself, i.e. $Cov(X, X)$ or σ_{XX}, is

$$\sigma_{XX} = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2,$$

which is the variance of X.

Example 4.6.4. Let X denote the deductible in car insurance, and let Y denote the deductible in home insurance, of a randomly chosen home and car owner in some community. Suppose that X, Y have the following joint pmf.

		y			
		0	100	200	
	100	.20	.10	.20	.5
x					
	250	.05	.15	.30	.5
		.25	.25	.50	1.0

where the deductible amounts are in dollars. Find $\sigma_{X,Y}$.

Solution. We will use computational formula $Cov(X, Y) = E(XY) - E(X)E(Y)$. First,

$$E(XY) = \sum_x \sum_y xyp(x, y) = 23.750.$$

Also,

$$E(X) = \sum_x xp_X(x) = 175, \ E(Y) = 125.$$

Thus, $Cov(X, Y) = 1875$.

Example 4.6.5. Suppose (X, Y) have joint pdf

$$f(x, y) = \begin{cases} 24xy & 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $Cov(X, Y)$.

Solution: Will use the computational formula $Cov(X, Y) = E(XY) - E(X)E(Y)$.

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) \ dxdy = \int_0^1 \int_0^{1-x} xy24xy \ dydx = \frac{2}{15}.$$

Next, using the marginal pdf of X, which was derived in Example 4.6.2, and noting that, by the symmetry of the joint pdf in x, y, the marginal pdf of Y has the same form, we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x 12x(1-x)^2 dx = \frac{2}{5} \\ E(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy = \int_0^1 y 12y(1-y)^2 dy = \frac{2}{5} \end{aligned}$$

Thus

$$Cov(X,Y) = \frac{2}{15} - \frac{2}{5}\frac{2}{5} = -\frac{2}{75}.$$

In order to develop an intuitive understanding as to why the sign of covariance identifies the nature of dependence, let us consider a finite underlying population of N units and let X, Y denote the characteristics of interest. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ denote the values of the bivariate characteristic of the N units. Let (X, Y) denote the bivariate characteristic of a randomly selected unit. Then (X, Y) has a discrete distribution (even when X and Y are continuous variables!) taking each of the possible values $(x_1, y_1), \ldots, (x_N, y_N)$ with probability $1/N$. In this case the covariance formula in Definition 4.6.2 can be computed as

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y), \tag{4.6.1}$$

where $\mu_X = \frac{1}{N}\sum_{i=1}^N x_i$, and $\mu_Y = \frac{1}{N}\sum_{i=1}^N y_i$ are the marginal expected values of X and Y, respectively. Suppose now that X, Y have positive dependence. Then the products

$$(x_i - \mu_X)(y_i - \mu_Y), \tag{4.6.2}$$

which appear in the summation of relation (4.6.1), will tend to be positive, and thus σ_{XY} will, in all likelihood, also be positive. As a concrete example, consider the population of men 25-30 years old residing in Centre County, PA, and let X=height, Y=weight be the characteristics of interest. These variables are clearly positively correlated, so that, if a man is taller than average (i.e. if $x_i - \mu_X > 0$), then it is more likely than not that this man is heavier than average (i.e $y_i - \mu_Y > 0$); similarly, if a man is shorter than average, then it is more likely than not that he will be less heavy than average. Thus, the products in (4.6.2) will tend to be positive. However, if the dependence is negative then these products will tend to be negative and so will σ_{XY}.

The above discussion reveals that σ_{XY} will be positive or negative according to whether the dependence of X and Y is positive or negative. Other properties of covariance are summarized in the following proposition.

Proposition 4.6.2. *1. If X, Y are independent, then $Cov(X, Y) = 0$.*

2. $Cov(X, Y) = -Cov(X, -Y) = -Cov(-X, Y)$.

3. For any real numbers b and d,

$$Cov(X + b, Y + d) = Cov(X, Y).$$

4. For any real numbers a, b, c and d,

$$Cov(aX + b, cY + d) = ac\, Cov(X, Y).$$

The first property follows from the formula for calculating the expected value of a product of independent random variables, which was given in Proposition 4.5.2. The second property means that, if the sign of one of the two variables changes, then a positive dependence becomes negative and vice-versa. The third property means that adding constants to the random variables will not change their covariance. The fourth property implies that the covariance of X and Y changes when the scale (or unit) changes. Thus, if X=height and Y=weight, changing the scale from (meters, kg) to (ft, lb), changes the value of the covariance of X and Y.

Example 4.6.6. Consider the information given in Example 4.6.4, but assume that the deductible amounts are given in cents. Thus, the new variables, (X', Y') are related to those of the aforementioned example by $(X', Y') = (100X, 100Y)$. Find $\sigma_{X'Y'}$.

Solution: According to part 4 of Proposition 4.6.2,

$$Cov(X', Y') = Cov(100X, 100Y) = 18,750,000.$$

4.6.3 Pearson's (or Linear) Correlation Coefficient

The basic role of covariance is to identify the nature of dependence. It is often of interest to quantify the dependence of two (positively or negatively) dependent variables.

The regression function, which was discussed in the Section 4.6.1, is a consequence and a manifestation of dependence since, if X and Y are independent, then the regression function $\mu_Y(x)$ is constant in x. However, it is important to note that the regression function is not designed to measure the degree of dependence of X and Y. Moreover, the value of the covariance is not a satisfactory quantification of dependence, because it is scale dependent (part 4 of Proposition 4.6.2). Clearly, we would like a measure of dependence to be unaffected by such scale changes. This leads to the definition of the *correlation coefficient* as a scale-free version of covariance.

Definition 4.6.3. *The* **Pearson's (or linear) correlation coefficient** *of X and Y, denoted by $Corr(X, Y)$ or ρ_{XY}, is defined as*

$$\rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X, σ_Y are the marginal standard deviations of X, Y, respectively.

The following proposition summarizes some properties of the correlation coefficient.

Proposition 4.6.3. *1. If a and c are either both positive or both negative, then*

$$Corr(aX + b, cY + d) = Corr(X, Y).$$

If If a and c are of opposite signs, then

$$Corr(aX + b, cY + d) = -Corr(X, Y).$$

2. For any two random variables X, Y,

$$-1 \leq \rho(X, Y) \leq 1,$$

3. if X, Y are independent then $\rho_{X,Y} = 0$.

4. $\rho_{X,Y} = 1$ or -1 if and only if $Y = aX + b$ for some numbers a, b with $a \neq 0$.

The properties listed in Proposition 4.6.3 imply that correlation is indeed a successful measure of *linear* dependence. First, it has the desirable property of being independent of scale. Second, the fact that it takes values between -1 and 1, makes it possible to develop a feeling for the degree of dependence between X and Y. Thus, if the variables are independent, their correlation coefficient is zero, while $\rho_{X,Y} = \pm 1$ happens if and only if X and Y have the strongest possible linear dependence (which is, knowing one amounts to knowing the other).

Example 4.6.7. Consider the information given in Example 4.6.4. Thus, X and Y denote the deductible in car and home insurance, of a randomly chosen home and car owner in some community, and the joint pmf of X, Y is

		y			
		0	100	200	
x	100	.20	.10	.10	.5
	250	.05	.15	.30	.5
		.25	.25	.50	1.0

where the deductible amounts are in dollars. Find $\rho_{X,Y}$. Next, express the deductible amounts in cents and find again the correlation coefficient.

Solution. In Example 4.6.4 we saw that $Cov(X, Y) = 1875$. Omitting the details of the calculations, the standard deviations of X and Y are computed to be $\sigma_X = 75$, $\sigma_Y = 82.92$. Thus, $\rho_{X,Y} = .301$. Next, if the deductible amounts are expressed in cents, then the new deductible amounts are $(X', Y') = (100X, 100Y)$. According to Proposition 4.6.3, part 1, the correlation remains unchanged.

Example 4.6.8. Consider the multinomial experiment of Example 4.3.2, but with a sample of size one. Thus, one electronic component will be tested, and if it lasts less than 50 hours, then $Y_1 = 1$, $Y_2 = 0$, and $Y_3 = 0$; if it lasts between 50 and 90 hours, then $Y_1 = 0$, $Y_2 = 1$, and $Y_3 = 0$; if it lasts more than 90 hours, then $Y_1 = 0$, $Y_2 = 0$, and $Y_3 = 1$. Find $Cov(Y_1, Y_2)$, $Cov(Y_1, Y_3)$ and $Cov(Y_2, Y_3)$.

Solution. We will use computational formula $Cov(X, Y) = E(XY) - E(X)E(Y)$. First note that

$$E(Y_1 Y_2) = 0.$$

This is due to the fact that the sample space of the bivariate random variable (Y_1, Y_2) is $\{(1, 0), (0, 1), (0, 0)\}$, so that the product $Y_1 Y_2$ is always equal to zero. Next, since the marginal distribution of each Y_i is Bernoulli, we have that

$$E(Y_i) = P(Y_i = 1).$$

Thus, according to the information given in Example 4.3.2, $E(Y_1) = 0.2$, $E(Y_2) = 0.5$, $E(Y_3) = 0.3$. It follows that $Cov(Y_1, Y_2) = -0.1$, $Cov(Y_1, Y_3) = -0.06$ and $Cov(Y_2, Y_3) = -0.15$.

Example 4.6.9. Suppose (X, Y) have the joint pdf given in Example 4.6.5, that is,

$$f(x,y) = \begin{cases} 24xy & 0 \leq x \leq 1, 0 \leq y \leq 1, x+y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $Cov(X, Y), \rho_{X,Y}$.

Solution: In Example 4.6.5 we saw that $\sigma_{XY} = -\frac{2}{75}$. Further, using the marginal pdfs, as done in the aforementioned example, we have

$$E(X^2) = \int_0^1 x^2 12x(1-x)^2 dx = \frac{1}{5},$$

so that $\sigma_X^2 = (1/5) - (4/25) = 1/25$, and $\sigma_X = 1/5$. Similarly, $\sigma_Y = 1/5$. Using these calculations, we have

$$\rho_{X,Y} = -\frac{2}{75}/\frac{1}{25} = -\frac{50}{75} = -\frac{2}{3}.$$

It should be emphasized that correlation measures only *linear dependence.* In particular, it is possible to have the strongest possible dependence, i.e. knowing one amounts to knowing the other, but if the relation between X and Y is not linear, then the correlation will not be one, as the following examples shows.

Example 4.6.10. Let X have the uniform in $(0, 1)$ distribution, and let $Y = X^2$. Thus, knowing X amounts to knowing Y and, conversely, X is given as the positive square root of Y. Then,

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(X^2) = \frac{1}{4} - \frac{1}{2}\frac{1}{3} = \frac{1}{12}$$

$$\sigma_X = \frac{1}{\sqrt{12}}, \quad \sigma_Y = \frac{2}{3\sqrt{5}},$$

so that

$$\rho_{X,Y} = \frac{3\sqrt{5}}{2\sqrt{12}} = 0.968.$$

A similar set of calculations reveals that with X as before and $Y = X^4$, $\rho_{X,Y} = 0.866$.

If the dependence is neither positive nor negative (i.e. it is not **monotone**) it is possible for two variables to have zero correlation, even though they are very closely related, as the following example shows.

Example 4.6.11. Let X have the uniform in $(-1,1)$ distribution, and let $Y = X^2$. Thus, X and Y are closely related since knowing X amounts to knowing Y (though not vice versa). However it is easily checked that $\rho_{X,Y} = 0$. Note that the dependence of X and Y is neither positive nor negative.

Two variables having zero correlation are called **uncorrelated**. Independent variables are uncorrelated, but uncorrelated variables are not necessarily independent.

4.6.4 Spearman's (or Monotone) Correlation Coefficient

As Example 4.6.10 demonstrated, linear correlation is not a good measure of the monotone dependence of two variables. *Spearman's* correlation coefficient, defined in this section, remedies this.

Definition 4.6.4. *The* **Spearman's (or monotone) correlation coefficient** *of X and Y, denoted by ρ^S_{XY}, is defined as the correlation coefficient of $U_1 = F_X(X)$ and $U_2 = F_Y(Y)$, where F_X is the cdf of X and F_Y is the cdf of Y. That is,*

$$\rho^S_{X,Y} = \rho_{U_1,U_2}.$$

Note that if X and Y are continuous random variables, then $U_1 = F_X(X)$ and $U_2 = F_Y(Y)$ are uniform random variables. The rationale behind the definition of the uniform correlation is based on the fact that if two uniform random variables are monotonically dependent (i.e. either positively or negatively dependent) then they are linearly dependent. Thus, by transforming X and Y to uniform random variables, their monotone dependence becomes linear dependence.

Example 4.6.12. Consider $Y = X^2$, where X is any random variable taking only positive values. As Example 4.6.10 demonstrated, $\rho_{X,Y}$ is not equal to 1 even though knowing one amounts to knowing the other. On the other hand, $\rho^S_{X,Y} = 1$, as we now show. Note first that,

$$F_Y(y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}).$$

Thus, $U_2 = F_Y(Y) = F_X(\sqrt{Y}) = F_X(X) = U_1$, so that $\rho_{U_1,U_2} = 1$.

4.6.5 Exercises

1. Consider the information given in Exercise 4.3.5-1 on the joint distribution of the number, X, of self-service pumps used at a particular time, and Y, the number of full-service pumps in use at that time.

   (a) Compute the regression function of Y on X.

   (b) Compute the covariance and correlation between X and Y.

2. Consider the information given in Exercise 4.3.5-3 on the joint distribution of X, the amount of drug administered to a randomly selected laboratory rat, and Y, the number of tumors present on the rat.

   (a) Find the regression function of Y on X.

   (b) You are given: $E(X) = .7$, $E(X^2) = .9$, $E(Y) = .047$, $E(Y^2) = .067$. Find $\sigma_{X,Y}$, and $\rho_{X,Y}$.

   (c) Does the 1.0 mg/kg drug dosage increase or decrease the expected number of tumors over the 0.0 mg/kg drug dosage? Justify your answer in two ways, a) using the regression function, and b) using the correlation coefficient.

3. Consider the information given in Exercise 4.3.5-4 regarding the use of children's seat belts and children's survival of motor vehicle accidents.

   (a) Find the regression function, $\mu_{Y|X}(x)$, of Y on X.

   (b) Find the covariance and the correlation of X and Y.

4. Consider the information given in Exercise 4.3.5-5 regarding the price of dinner entrees and the tip left.

   (a) Find the correlation between the price of the meal and the amount of the tip left. You may use the following information to help you answer the question: $E(\text{price}) = 8.38$, $Var(\text{price}) = 1.42$, $E(\text{tip}) = 1.31$, and $Var(\text{tip}) = 0.11$.

   (b) Find the regression function $E(\text{Tip}|\text{Meal Price})$.

5. Consider selecting two products from a batch of 10 products. Suppose that the batch contains 3 defective and 7 non-defective products. Let X take the value 1 or 0 as the first selection from the 10 products is defective or not. Let Y take the value 1 or 0 as the second selection (from the nine remaining products) is defective or not.

   (a) Find the regression function $E(Y|X = x)$, $x = 0, 1$.

   (b) Find the covariance and the correlation of X and Y.

6. Consider the context of Exercise 4.4.5-4, so that the variables X = actual flow through the vent and Y = maximum flow possible, have joint pdf

$$f(x,y) = \begin{cases} c(y^2 - x^2)e^{-y} & \text{if } -y \leq x \leq y \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find $f_{X|Y=y}$, and the regression function $E(X|Y = y)$.

   (b) Calculate $\text{Cov}(X,Y)$ and $\text{Corr}(X,Y)$.

   (c) On the basis of the regression function, comment on the appropriateness of the linear correlation coefficient as a measure of dependence.

7. Consider the context of Exercise 4.4.5-5, so that the variables X = static force at failure Y = defect index, have joint pdf

$$f(x,y) = \begin{cases} 24x & \text{if } 0 \leq y \leq 1 - 2x \text{ and } 0 \leq x \leq .5 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find $f_{Y|X=x}$, and the regression function $E(Y|X = x)$.

   (b) Find $\text{Var}(X)$, $\text{Var}(Y)$, and $\text{Cov}(X,Y)$.

   (c) Find the correlation coefficient for X and Y. Is there evidence for a linear relationship between these random variables? Plot the regression function, which you found above, and comment.

8. Consider the context of Exercise 4.4.5-6, so that the variables X = # of hours that John is late, and Y = # of hours that Yvonne is late, have joint pdf

$$f(x,y) = \begin{cases} \frac{1}{4} & 0 \leq x \leq 2, 0 \leq y \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

   Use the answers from Exercise 4.4.5-6 in order to find

   (a) $f_{Y|X=x}$ and $E(Y|X = x)$, and

   (b) $\text{Cov}(X,Y)$ and $\text{Corr}(X,Y)$

   without doing any additional calculations.

9. Let X be defined by the probability density function

$$f(x) = \begin{cases} -x & -1 < x \leq 0 \\ x & 0 < x \leq 1 \\ 0 & otherwise \end{cases}$$

   (a) Define $Y = X^2$ and find $\text{Cov}(X,Y)$.

   (b) Without doing any calculations, find $E(Y|X = x)$.

   (c) On the basis of the regression function found above, comment on the appropriateness of the linear correlation coefficient as a measure of dependence between X and Y.

4.7 Contrasts of Population Means

Contrasts are used to characterize differences among population means, and thus they are the main focus of statistical investigations in comparative studies (see Section 1.7).

The simplest contrast arises in the context of comparing two population means, μ_1, μ_2, and is

$$\theta = \mu_1 - \mu_2. \tag{4.7.1}$$

Example 4.7.1. A certain lake has been designated for pollution clean-up. To assess the effectiveness of the clean-up measures, water samples from 10 randomly selected locations are taken both before and after the clean-up. Let (X_{1i}, X_{2i}), $i = 1, \ldots, 10$, denote the before and after pollution indexes at the 10 locations. Because the locations were chosen randomly, the marginal mean value, μ_1, of X_{1i} is the average pollution index of the lake before the clean-up, for all $i = 1, \ldots, 10$. Similarly, the marginal mean value, μ_2, of X_{2i} is the average pollution index of the lake after the clean-up, for all $i = 1, \ldots, 10$. In this comparative study, the contrast (4.7.1) is of primary interest.

Remark 4.7.1. *The comparison of the average pollution before and after clean-up in the above example, could also be done with a different sampling scheme, according to which the researchers use a different set of randomly selected locations for taking water samples after the clean-up. Though this sampling scheme produces two independent samples,* $X_{1i},\ i = 1, \ldots, n_1$ *and* $X_{2j},\ j = 1, \ldots, n_2$*, it is still true that* $E(X_{1i}) = \mu_1$, $E(X_{2j}) = \mu_2$*, and the contrast (4.7.1) is still of primary interest.*

In this section we will see the relevant contrasts for comparative studies involving one factor and two factors.

4.7.1 Contrasts in One-Factor Designs

According to the terminology introduced in Section 1.7 one-factor designs arise in the context of comparing k population means, $\mu_1, \ldots, \mu_k$, where the populations correspond to the different levels of the factor. The comparison of k population means involves $k-1$ contrasts. The set of $k-1$ contrasts is not unique. For example, one set of contrasts is

$$\mu_1 - \mu_2,\ \mu_1 - \mu_3,\ \ldots,\ \mu_1 - \mu_k. \tag{4.7.2}$$

Example 4.7.2. Consider a study aimed at comparing the average life time of four types of truck tires. The investigators select a random sample of 16 trucks for the experiment. For each truck, four tires, one of each type, are selected and the location where each tire is put on is randomly chosen. Then the trucks are operated and the duration (in miles) of each tire is recorded. This results in multivariate observations $(X_{1i}, X_{2i}, X_{3i}, X_{4i})$, $i = 1, \ldots, 16$, where X_{ti} denotes the mile-life of tire type t, $t = 1, 2, 3, 4$, in truck i. Because the trucks, the tires, and the location where each tire is put on the truck are all chosen at random, the marginal mean value, μ_1, of X_{1i} is the average mile-life of tires of type 1 when used on such trucks, for all $i = 1, \ldots, 16$. Similarly, the marginal mean values, μ_t, of X_{ti}, $t = 2, 3, 4$, are the average mile-life of tires of types 2,3, and 4, respectively, when used on such trucks, for all $i = 1, \ldots, 16$. In this comparative study, the contrasts (4.7.2) are of primary interest

Remark 4.7.2. *The comparison of the average mile-life of the four types of tires can also be done with a different study design, according to which the researchers equip n_1 trucks with tires of type 1, a different sample of trucks with tires of type 2, and so forth. For the ith truck equipped with tires of type t, the average mile-life, X_{ti}, of its four tires is recorded. This study design produces four independent samples, X_{ti}, $i = 1, \ldots, n_t$, $t = 1, \ldots, 4$. Note that it is still true that $E(X_{ti}) = \mu_t$, $t = 1, \ldots, 4$, and the contrasts (4.7.2) are still of primary interest.*

The contrasts in (4.7.2) are particularly appropriate for the so-called *control versus treatment* studies, where a number of experimental methods or products are compared to the standard one. For example, suppose that the tire types 2, 3 and 4 of Example 4.7.2 are experimental versions and the purpose of the comparative study is to to see if the tires with the experimental design achieve, on average, higher average mile-life than the standard type (which is type 1). Alternative contrasts can be formulated according to the particular objectives of a comparative study. An example of a different set of contrasts for the comparison of k means is

$$\mu_1 - \overline{\mu}_{\cdot}, \; \mu_2 - \overline{\mu}_{\cdot}, \; \ldots, \; \mu_{k-1} - \overline{\mu}_{\cdot}, \tag{4.7.3}$$

where $\overline{\mu}_{\cdot} = k^{-1} \sum_{i=1}^{k} \mu_i$. The contrast $\mu_k - \overline{\mu}_{\cdot}$ could also be included in the set of contrasts (4.7.3), but typically investigators focus on linearly independent contrasts.

4.7.2 Contrasts in Two-Factor Designs

In this section we will consider the contrasts of interest in comparative studies aimed at assessing the effect of two factors on a variable of interest. For example, it may be of interest to assess the effect of temperature and humidity settings on the yield of a chemical reaction. Another example is presented in the next example.

Example 4.7.3. Consider the study aimed at comparing the average life time of four types of truck tires of Example 4.7.2, but the study uses five different types of trucks. Since the type of truck might also influence the mile-life of the tires, it is considered as an additional factor.

Remark 4.7.3. *The study of Example 4.7.3 might employ two different designs. In study design 1, the four tires of each truck consist of one tire of each type. In study design 2, the four tires of each truck are all of the same type; in this case the average life time of the four tires in each truck is recorded. Study design 1 produces multivariate observations, since the mile-lives of the four tires of each truck are dependent random variables. Study design 2 produces independent observations.*

	Column Factor			
Row Factor	1	2	3	4
1	μ_{11}	μ_{12}	μ_{13}	μ_{14}
2	μ_{21}	μ_{22}	μ_{23}	μ_{24}

Figure 4.1: Cell means in a 2 × 4 design

In the presence of two factors, each factor-level combination (e.g. each combination of a temperature setting and humidity setting) corresponds to a different population. These populations are represented in a table, where the rows correspond to the levels of one factor, which, therefore, is called the *row factor*, and the columns correspond to the levels of the other factor, the *column factor*. Thus, if the row factor has r levels, and the column factor has c levels, there are rc factor-level combinations or populations. A two-factor design with r and c levels for the row and column factors is called an $r \times c$ **design**. The factor-level combinations will be called *cells*. Figure 4.1 shows the means corresponding to the populations of the different factor-level combinations in a 2×4 design.

Which contrasts of the population means are of relevance in the comparison of the levels of a factor in a two-factor design? Answering this question is the main objective of this section. To better appreciate the answer, let us focus on a comparative study where the mile-lives of two types of tires are compared on four different types of trucks. (This is a simpler version of the comparative study of Example 4.7.3.) The type of contrast used for the comparison of the two tires depends on whether one wishes to compare the mean mile-lives in a *truck-specific* manner, or in an *overall* manner. To compare the mile-lives of the two types of tires overall trucks, the relevant contrast involves the average of the mean mile-lives over all the trucks, namely

$$\overline{\mu}_{1\cdot} - \overline{\mu}_{2\cdot},$$

where $\overline{\mu}_{1\cdot} = \sum_{j=1}^{4} \mu_{1j}/4$ and $\overline{\mu}_{2\cdot} = \sum_{j=1}^{4} \mu_{2j}/4$ are the averages, over all column levels, of the population means for each row level. These averages are illustrated in Figure 4.2. In general, the contrast for comparing the two row levels of a $2 \times c$ design is

	Column Factor				
Row Factor	1	2	3	4	Average
1	μ_{11}	μ_{12}	μ_{13}	μ_{14}	$\overline{\mu}_{1\cdot} = \frac{1}{4}\sum_{j=1}^{4} \mu_{1j}$
2	μ_{21}	μ_{22}	μ_{23}	μ_{24}	$\overline{\mu}_{2\cdot} = \frac{1}{4}\sum_{j=1}^{4} \mu_{2j}$

Figure 4.2: Cell means with their row averages in a 2 × 4 design

$\overline{\mu}_{1\cdot} - \overline{\mu}_{2\cdot}$	Contrast for comparing the average effect of two row levels

(4.7.4)

where $\overline{\mu}_{1\cdot} = \sum_{j=1}^{c} \mu_{1j}/c$ and $\overline{\mu}_{2\cdot} = \sum_{j=1}^{c} \mu_{2j}/c$.

Returning to the experiment where two types of tires are compared on four different types of trucks, the contrasts for a truck-specific comparison of the tire lives are

$$\mu_{11} - \mu_{21}, \ \ldots, \ \mu_{14} - \mu_{24}.$$

In general, the contrast for the column-specific comparison of two row levels in a $2 \times c$ design is

$\mu_{11} - \mu_{21}, \ \ldots, \ \mu_{1c} - \mu_{2c}$	Contrasts for comparing the column-specific effects of two row levels

(4.7.5)

If the row factor has r levels, with $r > 2$, any of the contrasts discussed in Section 4.7.1 can be used for the comparison of the averages, over all column levels, of the population means for each of the r row levels. For example, we can use the set of contrasts

$\overline{\mu}_{1\cdot} - \overline{\mu}_{2\cdot}, \ \overline{\mu}_{1\cdot} - \overline{\mu}_{3\cdot}, \ \ldots, \ \overline{\mu}_{1\cdot} - \overline{\mu}_{r\cdot}$	Contrasts for comparing the average effects of r row levels	(4.7.6)

The contrasts discussed in Section 4.7.1 can also be used for the comparison of the r row levels within each column. The set of all these contrasts for all columns are the relevant contrasts for the column-specific comparisons of k row levels. See Exercise 4.7.3-1.

The contrasts for comparing the column-levels, for both their average and their row-specific effects, are defined similarly.

The concept of interaction

In experimental designs that involve more than one factor, it is of interest to also examine whether or not the factors **interact** in influencing the response variable.

In words, we say that the factors **do not interact** if the relative performance of the levels of one factor remains the same regardless of the level of the other factor. For example, if an increase of the temperature by 50^oF increases the yield of a chemical reaction by 5gr, and this increase is the same at all humidity levels, we say that the factors ”temperature” and ”humidity” do not interact. Otherwise, we say that they interact.

To formalize the concept of interaction, let us focus first on a $2 \times c$ design, such as in Example 4.7.3 with two types of tires and four types of trucks, and consider the contrasts (4.7.5) for the truck-specific comparison of the tires. If these contrasts are equal, i.e. if

$\mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \cdots = \mu_{1c} - \mu_{2c},$	Equality of contrasts expressing no interaction in a $2 \times c$ design	(4.7.7)

then we say that the row and column factors **do not interact**, or there is no **interaction** between the two factors. Otherwise we say that the factors **interact**. Note that if the factors do not interact, the equality of the column-specific contrasts given in (4.7.7) implies their equality with the contrast for comparing the average effect of the two row levels given in (4.7.4).

If there is no interaction, the **profiles** of the two row levels are parallel as Figure 4.3 shows.

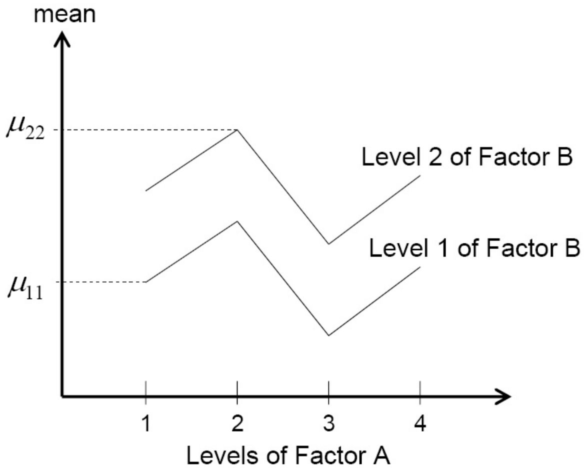


Figure 4.3: Profiles of the two levels of Factor B: No interaction

Figure 4.4 shows a case where the two factors interact.

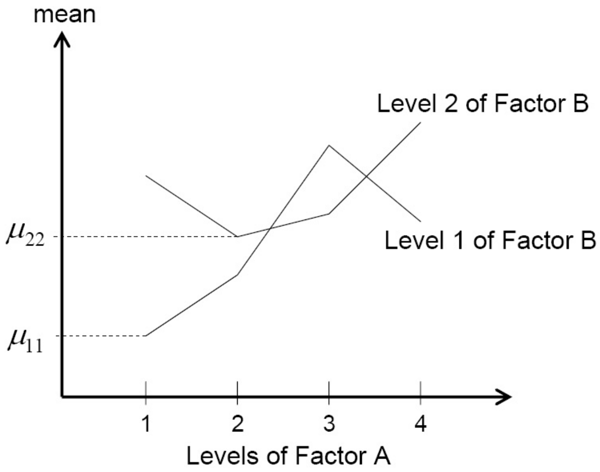


Figure 4.4: Profiles of the two levels of Factor B: no interaction

In general, we say that the two factors in a $r \times c$ design do not interact if the profiles of the r row levels are parallel. In that case, the profiles of the c column levels are also parallel.

The decomposition of means

Imagine the means of the rc populations in an $r \times c$ design to be arranged in a rectangular array, as shown in Figure 4.2 for a 2×4 design, and consider the row averages, column averages, and overall averages of these means:

$$\overline{\mu}_{i\cdot} = \frac{1}{c}\sum_{j=1}^{c}\mu_{ij}, \quad \overline{\mu}_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r}\mu_{ij}, \quad \overline{\mu}_{\cdot\cdot} = \frac{1}{r}\sum_{i=1}^{r}\overline{\mu}_{i\cdot} = \frac{1}{c}\sum_{j=1}^{c}\overline{\mu}_{\cdot j} = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\mu_{ij} \quad . \quad (4.7.8)$$

Any rectangular array of real numbers μ_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$ can be decomposed in terms of the averages in (4.7.8) as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{4.7.9}$$

where,

$$\mu = \overline{\mu}_{\cdot\cdot}, \;\; \alpha_i = \overline{\mu}_{i\cdot} - \mu, \;\; \beta_j = \overline{\mu}_{\cdot j} - \mu, \;\; \gamma_{ij} = \mu_{ij} - (\mu - \alpha_i - \beta_j)\,. \tag{4.7.10}$$

Moreover, the set of parameters α_i, β_j and γ_{ij}, as defined in (4.7.10), is the only set of parameters that satisfy the decomposition (4.7.9) and the conditions

$$\sum_{i=1}^{r} \alpha_i = 0, \;\; \sum_{j=1}^{c} \beta_j = 0, \;\; \sum_{i=1}^{r} \gamma_{ij} = 0, \text{ for all } j, \;\; \sum_{j=1}^{r} \gamma_{ij} = 0, \text{ for all } i. \tag{4.7.11}$$

The parameters α_i, $i = 1, \ldots, r$, are called **average**, or **main effects** of the row factor. The parameters β_j, $j = 1, \ldots, c$, are the average, or main, effects of the column factor, and the parameters γ_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, are the **interaction effects**.

The following example demonstrates this decomposition

Example 4.7.4. Consider the following rectangular array of numbers ($r = 2$, $c = 4$). The row averages, column averages, and total average, as well as the parameters α_i and β_j are given in the margins.

$A \setminus B$	1	2	3	4	
1	$\mu_{11} = 4$	$\mu_{12} = 8$	$\mu_{13} = 2$	$\mu_{14} = 6$	$\overline{\mu}_{1\cdot} = 5, \alpha_1 = -3$
2	$\mu_{21} = 10$	$\mu_{22} = 14$	$\mu_{23} = 8$	$\mu_{24} = 12$	$\overline{\mu}_{2\cdot} = 11, \alpha_2 = 3$
	$\overline{\mu}_{\cdot 1} = 7$	$\overline{\mu}_{\cdot 2} = 11$	$\overline{\mu}_{\cdot 3} = 5$	$\overline{\mu}_{\cdot 4} = 9$	$\overline{\mu}_{\cdot\cdot} = \mu = 8$
	$\beta_1 = -1$	$\beta_2 = 3$	$\beta_3 = -3$	$\beta_4 = 1$	

It is easily verified that $\mu_{ij} = \mu + \alpha_i + \beta_j$, for all i, j, and that $\gamma_{ij} = 0$, for all i, j.

If the interaction effects are all zero, as in the above example, the decomposition (4.7.9) becomes

$$\mu_{ij} = \mu + \alpha_i + \beta_j. \tag{4.7.12}$$

It is easily seen that when (4.7.12) holds the profiles of the levels of each factors are parallel, so the factors do not interact. Moreover, because the indices i, j in the decomposition (4.7.12) appear in separate terms, lack of interaction is also called **additivity**.

4.7.3 Exercises

1. It is known that the life time of a particular type of root system is influenced by the watering it receives and whether or not it grew at a depth of more than 4cm. An experiment is designed to study the effect of three watering regimens (row factor) and their possible interaction with the depth factor (column factor).

   (a) Make a 3×2 table showing the mean survival times of the root system for each factor level combination. (Use symbols, i.e. μ_{ij}, instead of actual numbers.)

   (b) Write the contrasts for comparing the watering regimens overall the levels of the depth factor.

   (c) Write the contrasts for the depth-specific comparison of the watering regimens.

   (d) What relation do you expect the contrasts for the depth-specific comparison of the watering regimens to satisfy if the watering and depth factors do not interact?

   (e) In the absence of interaction, what is the relation between the contrasts for the depth-specific comparison of the watering regimens and the contrasts for the average effect of the watering regimens.?

2. Show that

   (a) If $\alpha_1 = \cdots = \alpha_r$, then $\alpha_i = 0$, for all i.

   (b) If $\beta_1 = \cdots = \beta_c$, then $\beta_j = 0$, for all j.

   (c) If the condition in part 2a holds, then $\overline{\mu}_{1\cdot} = \cdots = \overline{\mu}_{r\cdot}$.

   (d) If the condition in part 2b holds, then $\overline{\mu}_{\cdot 1} = \cdots = \overline{\mu}_{\cdot c}$.

3. Consider the rectangular array of numbers given in Example 4.7.4, and change μ_{22} from 14 to 9. With this change, obtain the decomposition (4.7.9), i.e. obtain the average row and column effects and the interaction effects.

4.8 Models for Joint Distributions

4.8.1 Regression Models

From relation 4.3.2 we have that the joint pmf of X, Y can be given as the product of the conditional pmf of Y given $X = x$ and the marginal pmf of X, i.e. $p(x, y) = p_{Y|X=x}(y)p_X(x)$.

Similarly, in the continuous case we have $f(x, y) = f_{Y|X=x}(y)f_X(x)$ (see relation (4.4.8).

Regression models, specify the joint distribution of X, Y by first specifying the conditional distribution of Y given $X = x$, and then specifying the marginal distribution of X. Such models are useful when the primary objective of the study is to understand how the expected or average value of Y varies with X. A study of the speed, X, of an automobile and the stopping distance Y, or a study of the diameter at breast height, X, and age of a tree Y, or a study of the stress applied, X, and time to failure, Y, are examples of such studies. In such studies, the marginal distribution of X (which is also called **covariate** or **independent variable** or **explanatory variable**) is of little interest, and the regression model allows the investigator to focus on modeling the more interesting part of the joint distribution, namely, the conditional distribution of Y (which is called the **response variable**) given a value of X. Moreover, in some studies, the investigator can choose the values of X, in which case X is deterministic. Thus, the specific regression models discussed below, refer only to the conditional distribution of Y given $X = x$.

By specifying separately the two components of the joint distribution (i.e. the conditional and the marginal) regression models encompass very rich and flexible classes of joint distributions. In particular, they offer a way to model the joint distribution of a discrete and a continuous variable. For example, in a study focusing on the proportion of n equipment that last more than 600 hours of operation under different stress conditions, stress can be measured on a continuous scale, while the proportion (being a binomial random variable divided by n) is discrete.

In this subsection we will introduce the normal regression model, which is the only regression model that we will used in detail for data analysis in this book.

The Normal Regression Model

The normal regression model specifies that the conditional distribution of Y given $X = x$ is normal,

$$Y|X = x \sim N\left(\mu_{Y|X=x}, \sigma_\varepsilon^2\right), \tag{4.8.1}$$

where $\mu_{Y|X=x}$ is a given function of x, typically depending on unknown parameters.

Remark 4.8.1. *Note that the model (4.8.1) specifies that the variance of the conditional distribution of Y given $X = x$ does not depend of x. This is called the* **homoscedastic** *normal regression model. However, since we will not consider* **heteroscedastic** *models, by normal regression model we will mean the homoscedastic one.*

The simplest regression model is the **simple linear regression model** which specifies that

$$\mu_{Y|X}(x) = \alpha_1 + \beta_1 x, \quad \text{or} \quad \mu_{Y|X}(x) = \beta_0 + \beta_1(x - \mu_X), \tag{4.8.2}$$

where μ_X is the marginal mean value of X, and $\alpha_1,\ \beta_1$ (or $\beta_0,\ \beta_1$) are unknown model parameters. [Note that in (4.8.2), $\alpha_1 = \beta_0 - \beta_1\mu_X$.] The straight line defined by the equation (4.8.2) is called the **regression line**. The following picture illustrates the meaning of the slope of the regression line.

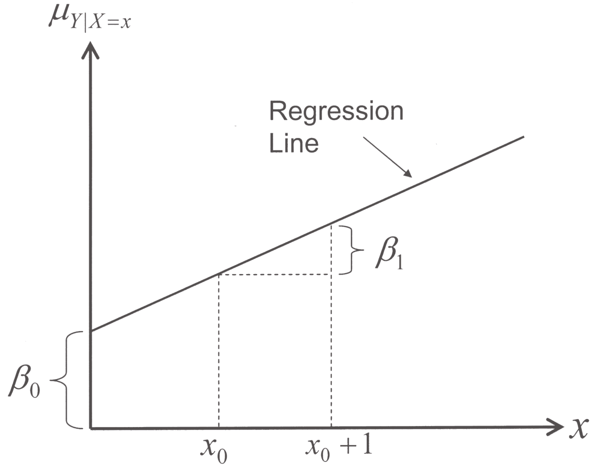


Figure 4.5: Illustration of Regression Parameters

Basically, the slope expresses the change in the average or mean value of Y when the value of X changes by one unit. Thus, if $\beta_1 > 0$ then X and Y are positively correlated, and if $\beta_1 < 0$ then X and Y are negatively correlated. If $\beta_1 = 0$ then X and Y are uncorrelated, in which case X is not of value in predicting Y. The above hint to a close connection between the slope in a simple linear regression model and correlation, and the following proposition makes the connection precise.

Proposition 4.8.1. *If the regression function of Y on X is linear (as either of the two functions given in (4.8.2)), then*

$$\beta_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} = \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}. \tag{4.8.3}$$

Moreover, (4.8.3) is true regardless of whether or not the conditional distribution of Y given $X = x$ is normal, i.e. (4.8.3) follows only from (4.8.2), regardless of whether or not (4.8.1) holds.

The normal simple linear regression model is also written as

$$Y = \alpha_1 + \beta_1 x + \varepsilon, \quad \text{or} \quad Y = \beta_0 + \beta_1(X - \mu_X) + \varepsilon, \tag{4.8.4}$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is called the **intrinsic error variable**. The intrinsic error variable expresses the conditional variability of Y around its conditional mean given $X = x$, as the following figure illustrates

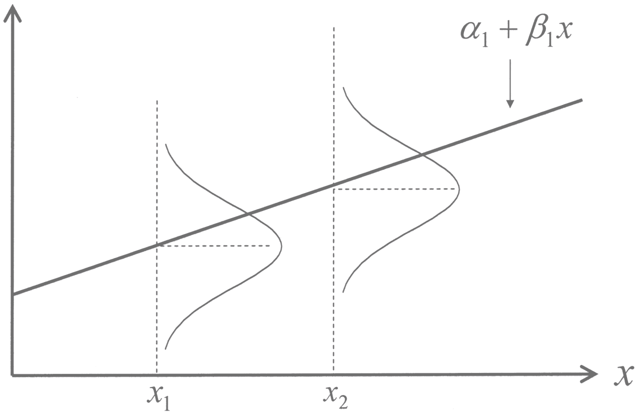


Figure 4.6: Illustration of Intrinsic Scatter in Regression

Quadratic and more complicated models are also commonly used. The advantages of such models are: a) It is typically easy to *fit* such a model to data (i.e. estimate the model parameters from the data), and b) Such models offer easy interpretation of the effect of X on the expected value of Y.

4.8.2 The Bivariate Normal Distribution

Definition 4.8.1. *If the joint distribution of* (X, Y) *is specified by the assumptions that the conditional distribution of* Y *given* $X = x$ *follows the normal linear regression model, and* X *has a normal distribution, i.e. if*

$$Y|X = x \sim N\left(\beta_0 + \beta_1(x - \mu_X), \sigma_\varepsilon^2\right), \quad \text{and} \;\; X \sim N(\mu_X, \sigma_X^2), \tag{4.8.5}$$

then (X, Y) *is said to have a* **bivariate normal distribution**.

It follows that the joint pdf of (X, Y) is

$$f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left\{-\frac{(y - \beta_0 - \beta_1(x - \mu_X))^2}{2\sigma_\varepsilon^2}\right\} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{-\frac{(x - \mu_X)^2}{2\sigma_X^2}\right\} \tag{4.8.6}$$

Using (4.8.5) and the Law of Total Probability for Expectations, we obtain

$$\begin{aligned} E(Y) &= E[E(Y|X)] = E[\beta_0 + \beta_1(X - \mu_X)] = \beta_0 \\ E(XY) &= E[E(XY|X)] = E[X(\beta_0 + \beta_1(X - \mu_X))] = \beta_0\mu_X + \beta_1\sigma_X^2 \\ E(Y^2) &= E[E(Y^2|X)] = E[\sigma_\varepsilon^2 + (\beta_0 + \beta_1(X - \mu_X))^2] = \sigma_\varepsilon^2 + \beta_0^2 + \beta_1^2\sigma_X^2. \end{aligned}$$

Thus,

$$\mu_Y = \beta_0, \; \sigma_Y^2 = \sigma_\varepsilon^2 + \beta_1^2\sigma_X^2, \; \sigma_{XY} = \beta_1\sigma_X^2, \tag{4.8.7}$$

which is in agreement with Proposition 4.8.1. It can be shown that the following is true

Proposition 4.8.2. *If (X, Y) have a bivariate normal distribution, then the marginal distribution of Y is also normal with mean μ_Y and variance σ_Y^2 given in (4.8.7).*

A bivariate normal distribution is completely specified by the mean values and variances of X and Y and the covariance of X and Y, i.e. by μ_X, μ_Y, σ_X^2, σ_Y^2 and σ_{XY}. The two variances and the covariance are typically arranged in a symmetric matrix, called the **covariance matrix**.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \tag{4.8.8}$$

In fact, using (4.8.6) and (4.8.7), the joint pdf of (X, Y) can be rewritten as

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right\} \\ &= \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}|}} \exp\left\{ -\frac{1}{2}(x-\mu_X, y-\mu_Y)\mathbf{\Sigma}^{-1} \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} \right\}, \end{aligned} \tag{4.8.9}$$

where $|\mathbf{\Sigma}|$ denotes the determinant of $\mathbf{\Sigma}$. See Exercise 4.8.4-8.

It is readily seen that if $\rho = 0$ the first expression in (4.8.9) for the joint pdf of X, Y becomes the product of the two marginal distributions. Thus we have

Proposition 4.8.3. *If (X, Y) have a bivariate normal distribution with $\rho = 0$, X and Y are independent.*

4.8.3 Multinomial Distribution

A special case of the multinomial distribution was introduced in Example 4.3.2, and used again in Example 4.6.8. Here we will give its general definition.

The multinomial distribution arises in cases where a basic experiment, which has r possible outcomes, is repeated independently n times. For example, the basic experiment can be life testing of an electric component, with $r = 3$ possible outcomes: 1, if the life time is short (less than 50 time units), 2, if the life time is medium (between 50 and 90 time units), or 3, if the life time is long (exceeds 90 time units). When this basic experiment is repeated n times, one can record the outcome of each basic experiment, resulting in n iid random variables $X_1, \ldots, X_n$. Alternatively, we may record

$$N_1, \ldots, N_r,$$

where $N_i =$ the number of times outcome i occurred.

Definition 4.8.2. *Consider n independent repetitions of an experiment which has r possible outcomes, and let $p_1, \ldots, p_r$ denote the probability of the first,...,rth outcome. Then, the random variables $N_1, \ldots, N_r$, defined above, are said to have the* **multinomial distribution** *with parameters, n, r, and $p_1, \ldots, p_r$. The joint probability mass function of the multinomial random variables $N_1, \ldots, N_r$ is*

$$P(N_1 = x_i, \ldots, N_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

if $x_1 + \cdots x_r = n$, and zero otherwise, where p_i is the probability of outcome i.

The *multinomial experiment* generalizes the binomial one. Recall that the binomial experiment consists of n independent repetitions of an experiment whose sample space has two possible outcomes. Instead of recording the individual Bernoulli outcomes, $X_1, \ldots, X_n$, the binomial experiment records $T =$ the number of times one of the outcomes occurred. To see the analogy with the multinomial experiment, note that recording T is equivalent to recording

$$N_1 = T, \quad \text{and} \quad N_2 = n - T.$$

Similarly, in the multinomial experiment one need not record N_r since

$$N_r = n - N_1 - \cdots - N_{r-1}.$$

Proposition 4.8.4. *If $N_1, \ldots, N_r$ have the multinomial distribution with parameters, n, r, and $p_1, \ldots, p_r$, the marginal distribution of each N_i is binomial with probability of 1 equal to p_i, i.e. $N_i \sim Bin(n, p_i)$. Thus,*

$$\boxed{E(N_i) = np_i \quad and \quad Var(N_i) = np_i(1 - p_i)}.$$

Moreover, it can be shown that, for $i \neq j$,

$$\boxed{Cov(N_i, N_j) = -np_ip_j}.$$

Example 4.8.1. Suppose that 60% of the supply of raw material kits used in a chemical reaction can be classified as recent, 30% as moderately aged, 8% as aged, and 2% unusable. If 16 kits are randomly chosen, find the probability that:

a) Exactly one of the 16 planned chemical reactions will not be performed due to unusable raw materials.

b) 10 chemical reactions will be performed with recent materials, 4 with moderately aged, and 2 with aged materials.

c) Let N_1, N_2, N_3, N_4 denote the number of chemical reactions performed with recent, moderately aged, aged, and unusable materials. Find Cov(N_1, N_2). How do you explain, intuitively, the negative sign of the covariance?

Solution: a) We want the probability $P(N_4 = 1)$. According to Proposition 4.8.4, $N_4 \sim \text{Bin}(16, 0.02)$. Thus,

$$P(N_4 = 1) = 16(0.02)(0.98)^{15} = 0.2363.$$

b) $P(N_1 = 10, N_2 = 4, N_3 = 2, N_4 = 0) = \dfrac{16!}{10!4!2!}0.6^{10}0.3^{4}0.08^{2} = 0.0377.$

c) According to Proposition 4.8.4, $\text{Cov}(N_1, N_2) = -16(0.6)(0.3) = -2.88$. The negative sign is explained by the fact that, intuitively, the larger N_1 is the smaller N_2 is expected to be.

4.8.4 Exercises

1. Suppose that the regression function of Y on X is

   $$\mu_{Y|X}(x) = 9.3 + 1.5x.$$

   Moreover, assume that the conditional distribution of Y given $X = x$ is normal with mean $\mu_{Y|X}(x)$ and variance $\sigma^2 = 16$.

   (a) Find the 95th percentile of the conditional distribution of Y when $X = 24$.

   (b) Let Y_1 denote an observation made at $X = x_1$, and Y_2 denote an observation made at $X = x_2$ where $x_2 = x_1 + 1$. (To fix ideas take $x_1 = 24$.) Compute the probability that Y_1 is larger than Y_2.

2. Consider the information given in the above Exercise 1. Suppose further that the marginal mean and variance of X are $E(X) = 24$ and $\sigma_X^2 = 9$. Find

   (a) the marginal mean and variance of Y, and

   (b) $\sigma_{X,Y}$ and $\rho_{X,Y}$.

3. Consider the information given in the above Exercise 1. Suppose further that the marginal distribution of X is normal with $\mu_X = 24$ and $\sigma_X^2 = 9$. Give the joint pdf of X, Y.

4. (**The Exponential Regression Model.**) The exponential regression model is common in reliability studies investigating how the expected life time of a product changes with some operational stress variable X. The exponential regression model assumes that the life time has an exponential distribution, and models the dependence of the expected life time on the operational stress by expressing the parameter λ of the exponential distribution as a function of the stress variable X. An example of such a regression model is

   $$\log \lambda = \alpha + \beta x.$$

Suppose that the above exponential regression model holds in the aforementioned reliability study with $\alpha = 4.2$ and $\beta = 3.1$. Suppose further, that the stress X a randomly selected product is exposed is uniformly distributed in the interval $[2, 6]$.

(a) Find the expected life time of a randomly selected product.

(b) Give the joint pdf of X, Y.

5. (**The Binomial Regression Model.**) The binomial regression model is common in reliability studies investigating how the probability of failure of a product (or probability of a product lasting more than, e.g. 600 hours of operation) changes with some stress variable X. Binomial regression models this change by specifying a function that captures the dependence of p, the probability of failure, on the stress variable X. An example, such a regression model is

$$\log \frac{p}{1-p} = \alpha + \beta x.$$

Suppose that the above binomial regression model holds in the aforementioned reliability study with $\alpha = 2.7$ and $\beta = 3.5$. Suppose further, that the stress X a randomly selected product is exposed is a discrete random variable with pmf $p_X(x) = 1/3$, for $x = 2, 4, 6$. Find the probability that a randomly selected product will last more that 600 hours.

6. An extensive study undertaken by the National Highway Traffic Safety Administration reported that 17% of children under 5 use no seat belt, 29% use adult seat belt, and 54% use child seat. In a sample of 15 children under five what is the probability that:

(a) Exactly 10 children use child seat?

(b) Exactly 10 children use child seat and 5 use adult seat?

(c) Exactly 8 children use child seat, 5 use adult seat and 2 use not seat belt?

7. Consider the information given in the above Exercise 6. Set N_1, N_2, N_3 for the number of children using no seat belt, adult seat belt, and child seat, respectively. Find

(a) $\text{Cov}(N_1, N_2)$.

(b) $\text{Cov}(N_1, N_2 + N_3)$.

8. (a) Use (4.8.6) and (4.8.7) to show that $\sigma_\varepsilon^2 = \sigma_Y^2 - \beta_1^2\sigma_X^2$, and $\rho = \beta_1\sigma_X/\sigma_Y$.

(b) Use the results in part a) to show that $\sigma_X\sigma_Y\sqrt{1-\rho^2} = \sigma_\varepsilon\sigma_X$.

(c) Use the results in part a) and b) to show that the combined exponential exponents in (4.8.6) equal the exponent in the first expression in (4.8.9).

(d) Use matrix operations to show equality of the two expressions in (4.8.9).

Chapter 5

Distribution of Sums and the Central Limit Theorem

5.1 Introduction

The distribution of a statistic was called its *sampling* distribution in Section 4.5. There we also saw how to calculate the expected value and variance of statistics. A class of statistics, that are of special interest, consists of sums of random variables, or, more generally, of *linear combinations* of random variables. The function $h(X_1, \ldots, X_n)$ is a **linear combination** of $X_1, \ldots, X_n$ if

$$h(X_1, \ldots, X_n) = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n,$$

where $a_1, \ldots, a_n$ are given constant numbers. For example, the sum, $T = \sum_i X_i$ of $X_1, \ldots, X_n$ is a linear combination with all $a_i = 1$, and the sample mean, $\overline{X} = \frac{1}{n} T$, is a linear combination with all $a_i = \dfrac{1}{n}$.

In this chapter we will apply the results of Section 4.5 for finding the expected value and variance of a linear combination, as well as the covariance of two linear combinations. Moreover, we will see that, though the expected value and variance of a sum, or average is easily obtained, obtaining its exact sampling distribution is much more complicated. A general result that offers an approximation to the sampling distribution of a sum, or average, will be given.

5.2 Expected Value of Sums

In this section we will see that the expected value of sums and differences of random variables follow easily from the individual (marginal) mean values of the random variables.

Proposition 5.2.1. *Let $(X_1, \ldots, X_n)$ have any joint distribution (i.e. they may be discrete or continuous, independent or dependent), and set $E(X_i) = \mu_i$. Then*

$$E(a_1 X_1 + \cdots + a_n X_n) = a_1 \mu_1 + \ldots + a_n \mu_n.$$

In other words, the expected value of a linear combination of random variables is the same linear combination of their expected values.

Corollary 5.2.1. *Let (X_1, X_2) have any joint distribution. Then*

$$E(X_1 - X_2) = \mu_1 - \mu_2, \quad \textit{and} \quad E(X_1 + X_2) = \mu_1 + \mu_2.$$

Corollary 5.2.2. *Let $(X_1, \ldots, X_n)$ have any joint distribution. If $E(X_1) = \cdots = E(X_n) = \mu$, then*

$$E(\overline{X}) = \mu, \quad \textit{and} \quad E(T) = n\mu,$$

where $T = \sum_i X_i$ and $\overline{X} = \frac{1}{n} T$.

An additional corollary of Proposition 5.2.1, whose proof uses also Proposition 4.5.2, is

Corollary 5.2.3. *Let $X_1, \ldots, X_n$ be a simple random sample from a population having variance σ^2. Then the expected value of the sample variance S^2 equals σ^2. That is*

$$E\left(S^2\right) = \sigma^2, \quad \textit{where} \quad S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right].$$

Example 5.2.1. Consider the setting of Example 4.5.1. Thus, X is the number of defective welds, and Y is the number of improperly tightened bolts per car. Find the expected value of the total number of robot errors per car.

Solution. The total number of errors is $T = X + Y$. Using the marginal pmfs of X and Y that are calculated in Example 4.3.3, and Corollary 5.2.1, we have

$$E(T) = E(X + Y) = E(X) + E(Y) = .12 + .148 = .268.$$

Note that the expected value of the total number of errors was obtained Example 4.5.1 in a different way. The present way is simpler.

Example 5.2.2. Let $X \sim \text{Bin}(n, p)$. Thus, X counts the number of 1's in n Bernoulli experiments, with the probability of 1 in each experiment equal to p. If X_i denotes the result of the ith Bernoulli experiment, then $X = \sum_i X_i$. It follows that

$$E(X) = \sum_{i=1}^{n} E(X_i) = np,$$

since p is the expected value of each X_i. This is an easier way of obtaining the expected value of a binomial random variable.

Example 5.2.3. Let X have the negative binomial distribution with parameters r and p. Thus, X counts the total number of Bernoulli trial until the rth 1. Let X_1 denote the number of trials up to and including the first 1, X_2 denote the number of trials after the first 1 up to and including the second 1, ... X_r denote the number of trials from the $(r-1)$st 1 up to and including the rth 1. Then each X_i has a geometric distribution and $X = \sum_i X_i$. It follows that

$$E(X) = \sum_{i=1}^{r} E(X_i) = r\frac{1}{p},$$

since $1/p$ is the expected value of each X_i. This is an easier way of obtaining the expected value of a negative binomial random variable.

5.2.1 Exercises

1. Suppose your waiting time for the bus in the morning has mean 3 minutes and variance 1.12 minutes^2, while the waiting time in the evening has mean 6 minutes and variance 4 minutes^2. In a typical week, you take the bus 5 times in the morning and 3 times in the evening. Calculate the expected value of the total waiting time in a typical week. (Hint: Let X_i denote the waiting time in the ith morning of the week, $i = 1, \ldots, 5$, and let Y_j denote the waiting time in the jth evening of the week. Express the total waiting as a linear combination of the random variables $X_1, \ldots, X_5, Y_1, Y_2, Y_3$.)

2. Two towers are constructed, each by stacking 30 segments of concrete vertically. The height (in inches) of a randomly selected segment is uniformly distributed in the interval (35.5,36.5).

   (a) Find the mean value of the height of a randomly selected segment. (Hint: What is the mean and variance of a uniform in (35.5,36.5) random variable?)

(b) Let $X_1, \ldots, X_{30}$ denote the heights of the segments used in tower 1, and $Y_1, \ldots, Y_{30}$ denote the heights of the segments used in tower 2. Express the difference of the heights of the two towers in terms of these Xs and Ys. (Hint: Height of tower 1 is the sum of the Xs.)

(c) Find the mean value of the difference of the heights of the two towers.

3. Using the information on the joint distribution of meal price and tip, given in Exercise 4.3.5,5, find the mean value of the total cost of the meal (entree plus tip) for a randomly selected customer.

5.3 Variance and Covariance of Sums

In the previous subsection we saw that the expected value of a linear combination of random variables is the same linear combination of the expected values, regardless of whether or not the random variables are independent. Here we will see that an analogous (but different!) result holds for the variance of a linear combination of independent random variables. In the general case, the variance of a linear combination involves the pairwise covariances.

Proposition 5.3.1. *Let the variables $X_1, \ldots, X_n$ have variances $\sigma_{X_i} = \sigma_i^2$ and covariances $\sigma_{X_i,X_j} = \sigma_{ij}$. (Recall that $\sigma_i^2 = \sigma_{ii}$.) Then*

1. *If $X_1, \ldots, X_n$ are independent, thus all $\sigma_{ij} = 0$,*

$$Var(a_1X_1 + \ldots + a_nX_n) = a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2$$

2. *Without independence,*

$$Var(a_1X_1 + \ldots + a_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\ \sigma_{ij}.$$

Corollary 5.3.1. 1. *If X_1, X_2 independent,*

$$\begin{aligned} Var(X_1 - X_2) &= \sigma_1^2 + \sigma_2^2 \\ Var(X_1 + X_2) &= \sigma_1^2 + \sigma_2^2. \end{aligned}$$

2. *Without independence,*

$$\begin{aligned} Var(X_1 - X_2) &= \sigma_1^2 + \sigma_2^2 - 2Cov(X_1, X_2) \\ Var(X_1 + X_2) &= \sigma_1^2 + \sigma_2^2 + 2Cov(X_1, X_2). \end{aligned}$$

Corollary 5.3.2. *If $X_1, \ldots, X_n$ are independent and $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$, then*

$$Var(\overline{X}) = \frac{\sigma^2}{n} \quad \text{and} \quad Var(T) = n\sigma^2,$$

where $T = \sum_i X_i$ and $\overline{X} = T/n$.

Example 5.3.1. Let $X \sim \text{Bin}(n,p)$. Thus, X counts the number of 1's in n independent Bernoulli experiments, with the probability of 1 in each experiment equal to p. If X_i denotes the result of the ith Bernoulli experiment, then $X = \sum_i X_i$. Because the X_i's are independent, and each has variance $\sigma_{X_i}^2 = p(1-p)$, it follows that the variance of the binomial random variable X is

$$\sigma_X^2 = \sum_{i=1}^{n} \sigma_{X_i}^2 = np(1-p).$$

This is an easier way of obtaining the variance of a binomial random variable.

Example 5.3.2. Let X have the negative binomial distribution with parameters r and p. Thus, X counts the total number of Bernoulli trial until the rth 1. Let X_1 denote the number of trials up to and including the first 1, X_2 denote the number of trials after the first 1 up to and including the second 1, ... X_r denote the number of trials from the $(r-1)$st 1 up to and including the rth 1. Then the X_i's are independent, each X_i has a geometric distribution and $X = \sum_i X_i$. It follows that the variance of the negative binomial random variable X is

$$\sigma_X^2 = \sum_{i=1}^{n} \sigma_{X_i}^2 = r\frac{1-p}{p^2},$$

since $(1-p)/p^2$ is the variance of each X_i. This is an easier way of obtaining the variance of a binomial random variable.

Proposition 5.3.2. *Let the variables $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ have variances denoted by $\sigma_{X_i}^2$ and $\sigma_{Y_j}^2$, and covariances between the X and Y variables be denoted by σ_{X_i, Y_j}. Then,*

$$Cov(a_1X_1 + \cdots + a_mX_m, b_1Y_1 + \cdots + b_nY_n) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j Cov(X_i, Y_j).$$

Example 5.3.3. Consider the multinomial experiment of Example 4.3.2. Thus, $n = 8$ products are tested, X_1 denotes the number of those that last less than 50 hours, X_2 denotes the number that last between 50 and 90 hours, and $X_3 = 8 - X_1 - X_2$ denotes the number that last more than 90. Find the covariance of X_1 and X_2.

Solution. For each of the eight products, i.e. for each $i = 1, \ldots, 8$, define triples of variables (Y_{i1}, Y_{i2}, Y_{i3}), as in Example 4.6.8. Thus, if the ith product lasts less than 50 hours, then $Y_{i1} = 1$ and $Y_{i2} = Y_{i3} = 0$; if it lasts between 50 and 90 hours then $Y_{i2} = 1$ and $Y_{i1} = Y_{i3} = 0$; if it lasts more than 90 hours then $Y_{i3} = 1$ and $Y_{i1} = Y_{i2} = 0$. Thus, X_1, the number of products that last less than 50 hours, is given as

$$X_1 = \sum_{i=1}^{8} Y_{i1}.$$

Similarly,

$$X_3 = \sum_{i=1}^{8} Y_{i1}, \quad \text{and} \quad X_3 = \sum_{i=1}^{8} Y_{i3}.$$

It follows that

$$\begin{aligned} Cov(X_1, X_2) &= Cov(\sum_{i=1}^{8} Y_{i1}, \sum_{i=1}^{8} Y_{i2}) \\ &= \sum_{i=1}^{8} \sum_{j=1}^{8} Cov(Y_{i1}, Y_{j2}). \end{aligned}$$

Assuming that the life times of different products are independent, we have that, if $i \neq j$, then $Cov(Y_{i1}, Y_{i2}) = 0$. Thus,

$$Cov(X_1, X_2) = \sum_{i=1}^{8} Cov(Y_{i1}, Y_{i2}) = \sum_{i=1}^{8} (-0.1) = -0.8,$$

where -0.1 is the covariance of Y_{i1} and Y_{i2}, as derived in Example 4.6.8. The covariance of X_1 and X_3 is similarly found to be $8 \times (-0.06) = -0.48$ and the covariance of X_2 and X_3 is similarly found to be $8 \times (-0.15) = -1.2$.

5.3.1 Exercises

1. Consider selecting two products from a batch of 10 products. Suppose that the batch contains 3 defective and 7 non-defective products. Let X take the value 1 or 0 as the first selection from the 10 products is defective or not. Let Y take the value 1 or 0 as the second selection (from the nine remaining products) is defective or not. Find the mean value and variance of the total number of defective items picked in the two selections.

2. Suppose your waiting time for the bus in the morning has mean 3 minutes and variance 1.12 minutes^2, while the waiting time in the evening has mean 6 minutes and variance 4 minutes^2. In a typical week, you take the bus 5 times in the morning and 3 times in

the evening. Calculate the variance of the total waiting time in a typical week. State your assumptions. (Hint: Let X_i denote the waiting time in the ith morning of the week, $i = 1, \ldots, 5$, and let Y_j denote the waiting time in the jth evening of the week. Express the total waiting as a linear combination of the random variables $X_1, \ldots, X_5, Y_1, Y_2, Y_3$.)

3. Two towers are constructed, each by stacking 30 segments of concrete vertically. The height (in inches) of a randomly selected segment is uniformly distributed in the interval (35.5,36.5).

   (a) Find the variance of the height of a randomly selected segment. (Hint: What is the mean and variance of a uniform in (35.5,36.5) random variable?)

   (b) Let $X_1, \ldots, X_{30}$ denote the heights of the segments used in tower 1, and $Y_1, \ldots, Y_{30}$ denote the heights of the segments used in tower 2. Express the difference of the heights of the two towers in terms of these Xs and Ys. (Hint: Height of tower 1 is the sum of the Xs.)

   (c) Find the variance of the difference of the heights of the two towers.

4. Using the information on the joint distribution of meal price and tip, given in Exercise 4.3.5,5, find the variance of the total cost of the meal (entree plus tip) for a randomly selected customer.

5.4 The Central Limit Theorem

In Section 5.2 we saw that the expected value of a sum of random variables is the corresponding sum of their expected values, and in Section 5.3 we saw that the variance of a sum of *independent* random variables is also a sum of their variances.

Though the expected value and the variance of sums follow by fairly straight-forward formulas, finding the distribution of a sum of random variables is, in general, a difficult task. One exception to this rule is in the case of a sum of normal random variables. The result is given in the following proposition.

Proposition 5.4.1. *Let $X_1, X_2, \ldots, X_n$ be independent and normally distributed random variables, $X_i \sim N(\mu_i, \sigma_i^2)$, and let $Y = a_1X_1 + \cdots + a_nX_n$ be a linear combination of the X_i's. Then Y also has a normal distribution, $Y \sim N(\mu_Y, \sigma_Y^2)$, with*

$$\mu_Y = a_1\mu_1 + \cdots + a_n\mu_n, \quad \sigma_Y^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2.$$

Corollary 5.4.1. *Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, and set $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$, as usual. Then $\overline{X} \sim N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$, where*

$$\mu_{\overline{X}} = \mu, \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}.$$

Remark 5.4.1. *Note that the only new item of Proposition 5.4.1 and its corollary is that linear combinations of independent normal random variables, and in particular $\overline{X}$, have a normal distribution. The expected value and variance of the linear combination are in accordance to the formulas in Sections 5.2 and 5.3.*

Example 5.4.1. The mileage of a randomly selected car of brand 1 is $X_1 \sim N(22, (1.2)^2)$, and that of brand 2 is $X_2 \sim N(26, (1.5)^2)$. Two cars, one from each brand, are selected at random and their mileage is tested. Find the probabilities that a) X_2 exceeds X_1, and b) X_2 exceeds X_1 by 5.

Solution. The answer to both parts uses $X_2 - X_1 \sim N(4, (1.2)^2 + (1.5)^2)$, which follows from Proposition 5.4.1. For a) write

$$P(X_2 - X_1 > 0) = P\left(Z > \frac{-4}{\sqrt{(1.2)^2 + (1.5)^2}}\right) = P(Z > -2.08) = 0.9812.$$

For b) write

$$P(X_2 - X_1 > 5) = P\left(Z > \frac{5-4}{\sqrt{(1.2)^2 + (1.5)^2}}\right) = P(Z > 0.52) = 0.3015.$$

As already mentioned, in general the task of finding the distribution of a sum of variables can be difficult. This is demonstrated by the following example.

Example 5.4.2. Let X_1, X_2 be iid random variables with common pmf

x	40	45	50
$p(x)$	.2	.3	.5

Find the pmf of $\overline{X} = \frac{1}{2}(X_1 + X_2)$.

Solution. To do so we first find the possible values of $\overline{X}$. These are

$$\overline{x} \ : \ 40, \quad 42.5, \quad 45, \quad 47.5, \quad 50$$

Then we need to determine the probability for each value. For example

$$\begin{aligned} & P(\overline{X}=45) = P(X_1=40, X_2=50) + P(X_1=45, X_2=45) + P(X_1=50, X_2=40) \\ = \ & P(X_1=40)P(X_2=50) + P(X_1=45)P(X_2=45) + P(X_1=50)P(X_2=40) \\ = \ & (0.2)(0.5) + (0.3)(0.3) + (0.5)(0.2) = 0.29 \end{aligned}$$

Working similarly for the other values we obtain the following pmf for $\overline{X} = \frac{1}{2}(X_1 + X_2)$:

$\overline{x}$	40	42.5	45	47.5	50
$p_{\overline{X}}(\overline{x})$	0.4	0.12	0.29	0.30	0.25

From this example, it can be appreciated that finding the pmf of $\overline{X} = n^{-1}\sum_i^n X_i$ gets harder as the number, n, of the random variables increases, even when the pmf of each X_i is quite simple (and even though $E(\overline{X})$, $Var(\overline{X})$ are known).

Thus, when the number of random variables is large, we resort to an approximation of the distribution of $\overline{X}$. This is achieved by the *Central Limit Theorem.* The importance to statistics of being able to approximate the distribution of sums cannot be overstated. For this reason, the Central Limit Theorem is considered the most important theorem in probability and statistics.

Theorem 5.4.1 (THE CENTRAL LIMIT THEOREM)**.** *Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2. Then for large enough n,*

1. *$\overline{X}$ has approximately a normal distribution with mean μ and variance σ^2/n, i.e.*

$$\overline{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

2. *$T = X_1 + \ldots + X_n$ has approximately a normal distribution with mean $n\mu$ and variance $n\sigma^2$, i.e*

$$T = X_1 + \ldots + X_n \dot{\sim} N\left(n\mu, n\sigma^2\right).$$

The quality of the approximation increases with the sample size n. For our purposes, we will consider that the approximation is good enough if $n \geq 30$.

Remark 5.4.2. *(1) The CLT does not require the X_i to be continuous.*

(2) The CLT can be stated for more general linear combinations of independent random variables, and also for certain dependence random variables. The present statement suffices for the range of applications of this book.

(3) The CLT asserts that the result of Proposition 5.4.1 holds approximately even if the the X_i's do not have a normal distribution.

(3) The CLT explains the central role of the normal distribution in probability and statistics.

Example 5.4.3. The amount of impurity in a randomly selected batch of chemicals has μ =4.0g and σ =1.5g. In a random sample of 50 batches, what is the (approximate) probability that average amount, $\overline{X}$, of impurity is between 3.5 and 3.8g?

Solution. By the CLT, $\overline{X} \overset{\cdot}{\sim} N(4.0,\ 1.5^2/50) = N(4.0, .2121^2)$, and since $n \geq 30$, this is a good enough approximation to the distribution of $\overline{X}$. Thus,

$$P(3.5 < \overline{X} < 3.8) \simeq P\left(\frac{3.5 - 4.0}{.2121} < Z < \frac{3.8 - 4.0}{.2121}\right) = \Phi(-0.94) - \Phi(-2.36) = 0.1645.$$

Example 5.4.4. Suppose that the waiting time for a bus, in minutes, has the uniform in $[0, 10]$ distribution. In five months a person catches the bus 120 times. Find the 95th percentile, $t_{0.05}$, of the person's total waiting time T.

Solution. Write $T = X_1 + \ldots + X_{120}$, where X_i=waiting time for catching the bus the ith time. It is given that $X_i \sim U(0, 10)$, so that

$$E(X_i) = 5, \text{ and } Var(X_i) = 10^2 \frac{1}{12} = \frac{100}{12}.$$

Since $n = 120 \geq 30$, the CLT provides a good enough approximation to the distribution of T. Thus,

$$T \overset{\cdot}{\sim} N\left(120 \times 5,\ 120\frac{100}{12}\right) = N(600, 1000).$$

Therefore, the 95th percentile of T can be approximated by the 95th percentile of the $N(600, 1000)$ distribution:

$$t_{0.05} \simeq 600 + z_{.05}\sqrt{1000} = 600 + 1.645\sqrt{1000} = 652.02.$$

We conclude the section on the normal distribution by describing in detail the normal approximation to binomial probabilities.

Approximation of binomial probabilities

Consider n replications of a success/failure experiment with probability of success being p. Let

$$X_i = \begin{cases} 1 & \text{if } i\text{th replication results in success} \\ 0 & \text{if } i\text{th replication results in failure} \end{cases}$$

Thus $X_1, \ldots, X_n$ is a simple random sample from the Bernoulli distribution, so that $E(X_i) = p$, $Var(X_i) = p(1-p)$. Also the binomial random variable T that counts the number of 1's in the n repetitions is

$$T = X_1 + \cdots + X_n.$$

Thus, by the CLT T has approximately a normal distribution. The general condition $n \geq 30$ for achieving acceptable quality in the approximation, can be specialized for the binomial distribution to

$$\boxed{np \geq 5 \text{ and } n(1-p) \geq 5 \quad \begin{array}{l}\text{sample size requirement for approximating} \\ \text{binomial probabilities by normal probabilities}\end{array}}.$$

Moreover, due to the discreteness of the binomial distribution, the approximation is improved by the so-called **continuity correction**, namely, if $X \sim B(n,p)$ and the condition $np \geq 5$ and $n(1-p) \geq 5$ holds, then

$$P(X \leq x) \simeq P(Y \leq x + 0.5) = \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right), \tag{5.4.1}$$

where Y is a random variable having the normal distribution with mean and variance equal to the mean and variance of the Binomial random variable X, i.e. $Y \sim N\left(np, np(1-p)\right)$.

Remark 5.4.3. *The normal approximation works in situations where the Poisson approximation does not work. For example p does not have to be* ≤ 0.01.

Example 5.4.5. Suppose that 10% of a certain type of components last more than 600 hours of operation. For $n = 200$ components, let X denote the number of those that last more than 600 hours. Find (approximately): a) $P(X \leq 30)$, b) $(15 \leq X \leq 25)$ and c) $P(X = 25)$.

Solution. Here X has a binomial distribution with $n = 200$ and $p = 0.1$. Thus, $np = 20$, $n(1-p) = 180$, so the sample size conditions for approximating binomial probabilities with normal probabilities are met. Moreover, the mean and variance of X are $np = 20$ and $np(1-p) = 18$. Thus, using also the continuity correction, we have

$$P(X \le 30) \simeq \Phi\left(\frac{30.5-20}{\sqrt{18}}\right) = \Phi(2.47) = 0.9932,$$

$$\begin{aligned} P(15 \le X \le 25) &= P(X \le 25) - P(X \le 14) \simeq \Phi\left(\frac{25.5-20}{\sqrt{18}}\right) - \Phi\left(\frac{14.5-20}{\sqrt{18}}\right) \\ &= 0.9032 - 0.0968 = 0.8064, \end{aligned}$$

$$\begin{aligned} P(X = 25) &= P(X \le 25) - P(X \le 24) \simeq \Phi\left(\frac{25.5-20}{\sqrt{18}}\right) - \left(\frac{24.5-20}{\sqrt{18}}\right) \\ &= 0.9032 - 0.8554 = 0.0478, \end{aligned}$$

for a), b) and c), respectively.

5.4.1 Exercises

1. Three friends on a camping trip each bring one flashlight with a fresh battery but they decide to use one flashlight at a time. Let X_1, X_2, and X_3 denote the lives of the batteries in each of the 3 flashlights, respectively. Suppose that they are independent normal rv's with expected values $\mu_1 = 6$, $\mu_2 = 7$, and $\mu_3 = 8$ hours and variances $\sigma_1^2 = 2$, $\sigma_2^2 = 3$, and $\sigma_3^2 = 4$, respectively.

   (a) Calculate the probability that the flashlights will last a total of less than 25 hours.

   (b) Suppose that the three friends have five camping trips that year and each time they start with the same types of fresh batteries as above. Find the probability that the batteries last more than 25 hours exactly three of the five times.

2. Using the information on the joint distribution of meal price and tip, given in Exercise 4.3.5,5, answer the following question: If a waitress at the restaurant is able to serve 70 customers in an evening, what is the probability that her tips for the night exceed \$100 (assuming that her customers act independently)?

3. Suppose the stress strengths of two types of material follow the gamma distribution with parameters $\alpha_1 = 2$, $\beta_1 = 2$ for type one and $\alpha_2 = 1$, $\beta_2 = 3$ for type two. Let $\bar{X}_1$, $\bar{X}_2$ be the averages corresponding to sample sizes $n_1 = 36$ and $n_2 = 42$ for the two types of material.

   (a) Specify the (approximate) distributions of $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_1 - \overline{X}_2$. Justify your answers.

(b) Find the (approximate) probability that $\overline{X}_1$ will be larger than $\overline{X}_2$.

4. Two towers are constructed, each by stacking 30 segments of concrete vertically. The height (in inches) of a randomly selected segment is uniformly distributed in the interval (35.5,36.5). A roadway can be laid across the two towers provided the heights of the two towers are within 4 inches of each other. Find the probability that the roadway can be laid. Be careful to justify the steps in your argument, and state whether the probability is exact or approximate.

5. An optical company uses a vacuum deposition method to apply a protective coating to certain lenses. The coating is built up one layer at a time. The thickness of a given layer is a random variable with mean $\mu = .5$ microns and standard deviation $\sigma = .2$ microns. The thickness of each layer is independent of the the others and all layers have the same thickness distribution. In all, thirty-six (36) layers are applied. The company has determined that a minimum thickness of 16 microns for the entire coating is necessary to meet all warranties. Consequently, each lens is tested and additional layers are applied if the lens does not have at least a 16 micron thick coat.

   (a) What is the approximate distribution of the coating thickness? Cite the appropriate theorem to justify your answer.

   (b) The company has determined that a minimum thickness of 16 microns for the entire coating is necessary to meet all warranties. Consequently, each lens is tested and additional layers are applied if the lens does not have at least a 16 micron thick coat. What proportion of lenses must have additional layers applied?

6. Flights from State College to Pittsburgh take on average 55 minutes, with a standard deviation of 7 minutes.

   (a) Let $X_1, \ldots, X_{49}$ denote the duration of the next 49 flights you take to Pittsburgh. What is the (approximate) probability that the average duration $\overline{X}$ is between 53 and 57 minutes?

   (b) What is the probability that the average duration of the next 100 flights is between 53 and 57 minutes?

   (c) Find a number c such that $P(\overline{X} > c) = 0.05$, when $\overline{X}$ is the average flight time of 49 flights.

7. A quality control inspector will accept a batch of 100 presumably 0.5cm diameter steel rods only if their sample mean diameter falls between 0.495cm and 0.505cm. It is known that the standard deviation for the diameter of a randomly selected rod is 0.03cm.

(a) Let μ denote the actual mean diameter of a randomly selected rod. What is the expected value of the sample mean diameter of the 100 steel rods?

(b) Using the information given above regarding the standard deviation for the diameter of a randomly selected rod, what is the variance of the sample mean diameter of the 100 steel rods?

(c) If the actual mean diameter of a randomly selected rod is $\mu = 0.503$cm, what is the probability the inspector will accept the batch?

8. Seventy identical automobiles will be driven 300km. Of them, 35 will use Brand A gasoline, and the remaining 35 will use Brand B. For this type of car, Brand A gasoline delivers a km/liter fuel economy with mean 50 km/L and standard deviation of 4 km/L, while Brand B delivers a mean of 52 km/L with standard deviation also 4 km/L. Let $\overline{X}$, $\overline{Y}$ denote the sample average km/L achieved by cars using Brands A, B, respectively.

   (a) What is the approximate distributions of $\overline{X}$ and $\overline{Y}$?

   (b) What is the approximate distribution of $\overline{X} - \overline{Y}$?

   (c) Find the (approximate) probability that $\overline{X}$ comes out larger than $\overline{Y}$.

9. Hardness of pins of a certain type is known to have a mean of 50 and standard deviation of 3.

   (a) A sample of size 9 is drawn. Do you need the normality assumption in order to compute the probability that the sample mean hardness is greater than 51 (i.e. $\bar{X} > 51$)? Justify your answer.

   (b) Find the approximate probability that sample mean hardness is greater than 51 (i.e. $\bar{X} > 51$) when a sample of size 36 is chosen.

10. Two airplanes are traveling parallel to each other in the same direction. At Noon, plane A is 10 km ahead of plane B. Let X_1 and X_2 be the speeds of planes A and B, respectively, which are independently Normally distributed with means 510 km/hr and 495 km/hr, respectively, and standard deviations 8 km/hr and 10 km/hr, respectively. At 3:00 PM, the distance (in km) that plane A is ahead of plane B is given by the expression $3X_1 - 3X_2 + 10$.

    (a) At 3:00 PM, how large do you expect this distance to be?

    (b) At 3:00 PM, what is the variance of this distance?

    (c) What is the distribution of this distance?

    (d) Find the probability that plane B has not yet caught up to plane A at 3:00 PM (i.e., that the distance is still positive).

11. A random sample of size 25 is taken from a normal population having a mean of 80 and a standard deviation of 5. A second random sample of size 36 is taken from a different normal population having a mean of 75 and a standard deviation of 3. Find the probability that the sample mean computed form the 25 measurements will exceed the sample mean computed from the 36 measurements by at least 3.4 but less than 5.9.

12. Five identical automobiles will be driven for a fixed (same for all cars) distance using one of two types of gasoline – two using Brand A and three using Brand B. Both gasolines are known to deliver mileage which is normally distributed. The mileage of Brand A has a mean of 20 mpg and a standard deviation of 2 mpg, while the mileage Brand B has a mean of 21 mpg and also a standard deviation of 2 mpg. Find the probability that the average mileage for two cars using Brand A will be higher than the average mileage for three cars using Brand B.

13. Let X_1, X_2, and X_3 be three independent normal random variables with expected values $\mu_1 = 90$, $\mu_2 = 100$, and $\mu_3 = 110$ and variances $\sigma_1^2 = 10$, $\sigma_2^2 = 12$, and $\sigma_3^2 = 14$, respectively.

    (a) Specify the distribution of $X_1 + X_2 + X_3$.

    (b) Calculate $P(X_1 + X_2 + X_3 \leq 306)$.

14. Suppose that only 60% of all drivers in a certain state wear seat belts . A random sample of 500 drivers is selected, and let X denote the number of drivers that wear seatbelt.

    (a) State the exact distribution of X. (Choose from: Binomial, hypergeometric, negative binomial, and Poisson.)

    (b) Find the (approximate) probability that X is between 270 and 320 (inclusive).

15. A machine manufactures tires with a tread thickness that is normally distributed with mean 10 millimeters (mm) and standard deviation 2 millimeters. The tire has a 50,000 mile warranty. In order to last for 50,000 miles the manufactures guidelines specify that the tread thickness must be at least 7.9 mm. If the thickness of tread is measured to be less than 7.9 mm, then the tire is sold as an alternative brand with a warranty of less than 50,000 miles. Compute the probability that, in a batch of 100 tires, there are no more that 10 rejects.

16. Each morning the manager of a Barney toy fabrication plant inspects the output of an assembly line. There are two assembly lines, A and B, at the plant, and, so as not to be predictable to the line foreman, she flips a fair coin to determine which assembly line to inspect that morning. Assembly line A produces Barneys at a constant rate of 200 every hour. Each Barney is defective with probability 0.10, independently of all the other

Barneys. Assembly line B produces Barneys at a constant rate of 1000 every hour. Each Barney is defective with probability 0.01, independently of all the other Barneys.

(a) Find the probability that, if the manager inspects the output of line B for one hour, she will find 6 or fewer defective.

(b) Find the probability that, if the manager inspects the output of line A for one hour, she will find 6 or fewer defective.

(c) (*Review part.*) Suppose you learn that she finds 6 or fewer defective in a one hour inspection, but you are not told which line, A or B, she inspected. Find the probability that she inspected line A that morning.

Chapter 6

Empirical Statistics: Data Graphics and Estimators

Empirical, or descriptive, statistics encompasses both graphical visualization methods and numerical summaries of the data. The data displays help identify key features of the data, such as symmetry of skewness, and provide valuable insight about corresponding population features. The numerical summaries, such as measures of location and spread, serve as *estimators* of corresponding population parameters.

In descriptive statistics, the feature identification and parameter estimation, are done with no or minimal assumptions on the underlying population, and thus they are part of nonparametric statistics. Here we will present some of the simplest, and most widely used, methods for univariate, bivariate, and tri-variate data.

In all that follows, the sample size of a single sample will usually be denoted by n. If two samples are simultaneously under consideration the sample sizes will be denoted by m and n or n_1 and n_2. The later convention, which uses subscripts, extends readily to more than two samples. Thus, if $k > 2$ samples are simultaneously under consideration the sample sizes will be denoted by $n_1, \ldots, n_k$.

The variable of interest (e.g. cement hardness) will typically be denoted by X (though Y or another letter might also be used), and a single data set consisting of n measurements on the variable X, will be denoted by $x_1, \ldots, x_n$. The numbering of the measurements or observations does not indicate ordering by their magnitude. Typically, the first observation to be obtained will be indexed by 1, the second by 2, and so on. If two samples are simultaneously under consideration they will be denoted by $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, or $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_2}$. The later notation extends readily to $k > 2$ samples as $x_{i1}, \ldots, x_{in_i}$, for $i = 1, \ldots, k$. Finally, a bivariate or tri-variate measurement will be denoted by $\mathbf{X}$ and a bivariate sample will be denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

6.1 Estimators for Univariate Data

In this section we present sample versions of the population parameters we discussed in Chapter 3. These population parameters can be classified as *location parameters*, e.g. the mean and, for continuous distributions, the median and various percentiles, and *scale (or spread) parameters*, e.g. the variance and, for continuous distributions, the interquartile range (IQR). Based on a simple random sample $X_1, \ldots, X_n$ from a population of interest, we will discuss empirical statistics that serve as *estimators* of the location and scale population parameters.

6.1.1 Sample Mean, Proportion and Variance: A Review

Let $X_1, \ldots, X_n$ be a simple random sample, from a population with mean μ and variance σ^2. In Chapter 1 we introduced the sample mean or average, and the sample variance,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

as estimators of the corresponding population parameters. If the X_is are transformed to $Y_i = a + bX_i$, then

$$\overline{Y} = a + b\overline{X}, \quad \text{and} \quad S_Y^2 = b^2 S^2.$$

We also saw the sample proportion, $\widehat{p}$, as a special case of the sample mean when the X_is are Bernoulli(p) random variables (so that $\mu = p$). Moreover, in Section 4.5, we saw that

$$E\left(\overline{X}\right) = \mu, \quad E\left(\widehat{p}\right) = p, \quad \text{and} \quad E(S^2) = \sigma^2,$$

(see Corollaries 5.2.2 and 5.2.3). This property reinforces the notion that it is reasonable to use $\overline{X}$, $\widehat{p}$, and S^2 as estimators of μ, p, and σ^2. This will be discussed further in Chapter 7.

Consistency of Averages

The limiting relative frequency definition of probability suggests that $\widehat{p}$ becomes a more accurate estimator of p as the sample size increases. In other words, the **error of estimation** $|\widehat{p} - p|$ decreases as the sample size increases. This basic property, which is desirable for all estimators, is called **consistency**. The following proposition suggests that a similar property is true for the sample mean.

Proposition 6.1.1. *As the size, n, of a simple random sample, $X_1, \ldots, X_n$, from a population with mean μ and finite variance σ^2, increases to infinity, the estimation error*

$$|\overline{X} - \mu|$$

decreases to zero.

Remark 6.1.1. *The above proposition extends a more general result which asserts that the average $\overline{h(X)} = n^{-1} \sum_{i=1}^{n} h(X_i)$ is an estimator of $\mu_{h(X)} = E(h(X_i))$, and the estimation error $|\overline{h(X)} - \mu_{h(X)}|$ decreases to zero as the sample size increases to infinity, for any function h which satisfies $E(h(X)^2) < \infty$. In particular, the estimation of σ^2 by S^2 becomes increasingly accurate as the sample size increases.*

A formal proof of this proposition is beyond the scope of this book. The following example, offers a demonstration of the consistency of the sample mean in the simple setting of rolling a die.

Example 6.1.1. Consider the conceptual population of all throws of a die, and let X denote a random selection from that population (i.e. X denotes the outcome of a roll of a die). The population mean μ is

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5. \tag{6.1.1}$$

Let now $X_1, \ldots, X_{100}$ denote a simple random sample of size 100 from the conceptual population of all throws of a die (i.e. $X_1, \ldots, X_{100}$ denote the outcomes of 100 throws of a die), and suppose that the outcomes of these 100 throws are summarized in the following frequency table. Thus, the 1 occurred in 17 of the 100 throws, 2 occurred in 16 throws etc. Note that, according to the limiting relative interpretation of probability, each relative frequency, $\hat{p}_i = f_i/100$, estimates the probability that i will occur in a throw of a die (which is $1/6 = 0.1666$, for all $i = 1, \ldots, 6$).

Frequency Table of 100 Rolls of a Die

i	f_i	$\hat{p}_i = \frac{f_i}{100}$	$p_i = P(X = i)$
1	17	.17	.1666
2	16	.16	.1666
3	19	.19	.1666
4	14	.14	.1666
5	15	.15	.1666
6	19	.19	.1666
	n=100	1.0	1.0

To see the relation between $E(X)$ and $\overline{X}$ write

$$\begin{aligned} \overline{X} &= \frac{1}{100}\sum_{i=1}^{100} X_i = \frac{1}{100}(1 \cdot f_1 + 2 \cdot f_2 + \cdots + 6 \cdot f_6) \\ &= 1 \cdot \hat{p}_1 + 2 \cdot \hat{p}_2 + \cdots + 6 \cdot \hat{p}_6. \end{aligned}$$

Comparing this expression of $\overline{X}$ with the expression in (6.1.1), it is seen that $\overline{X}$ is a reasonable estimator of μ since each $\hat{p}_i$ is a reasonable estimator of 0.1666.

6.1.2 Sample Percentiles

Let $X_1, \ldots, X_n$ be a simple random sample from a continuous population distribution. Roughly speaking, the $(1-\alpha)100$th **sample percentile** divides the sample in two parts, the part having the $(1-\alpha)100\%$ smaller values, and the part having the $\alpha 100\%$ larger values. For example, the 90th sample percentile (note that $90 = (1 - 0.1)100$) separates the upper (largest) 10% from the lower 90% values in the data set. The 50th sample percentile is also called the **sample median**; it is the value which separates the upper or largest 50% from the lower or smallest 50% of the data. The 25th, the 50th and the 75th sample percentiles are also called **sample quartiles**, as they divide the sample in four equal parts. We also refer to the 25th and the 75th sample percentiles as the **lower sample quartile** (**q1**) and **upper sample quartile** (**q3**), respectively.

Sample percentiles serve as estimators of corresponding population percentiles. It can be argued that they are, in fact, consistent estimators, but the argument will not be presented here.

For a precise definition of sample percentiles we need to introduce notation for the **ordered** sample values, or **ordered statistics**: The sample values $X_1, \ldots, X_n$ arranged in increasing order are denoted

$$\boxed{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}. \tag{6.1.2}$$

Thus, $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$.

We begin by identifying each $X_{(i)}$ as an estimator of a population percentile. Following that, we give precise (computational) definitions of the sample median and the upper and lower quartiles.

Definition 6.1.1. *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the ordered sample values in a sample of size n. Then $X_{(i)}$, the ith smallest sample value, is taken to be the $\left[100\dfrac{i-0.5}{n}\right]$-th* **sample percentile**. *Sample percentiles are estimators of corresponding population percentiles.*

Note that, depending on the sample size n, the above definition may not identify the sample median. For example, if $n = 5$ then $X_{(3)}$, the 3rd smallest value, is the $100\frac{2.5}{5} =$ 50th sample percentile or median. But if $n = 4$ then $X_{(2)}$ is the $100\frac{1.5}{4} = 37.5$th sample percentile, while $X_{(3)}$ is the $100\frac{2.5}{4} = 62.5$th sample percentile, so that none of the ordered values is the median. In general, if the sample size is even, none of the order statistics is the 50th percentile. Similarly, depending on the sample size, Definition 6.1.1 may not identify the upper and lower quartiles. (Only for sample sizes of the form $6 +$ (a multiple of 4) we have that the 25th and 75th percentiles are among the order statistics.) Thus, we need computational definitions for the sample median and quartiles.

Definition 6.1.2. *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the ordered sample values in a sample of size n. The* **sample median** *is defined as*

$$\widetilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \textit{if } n \textit{ is odd} \\[2ex] \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & \textit{if } n \textit{ is even} \end{cases}$$

Thus, when the sample size is even, the sample median is defined by interpolating between the nearest sample percentiles.

The next example, and the remark that follows, illustrate the similarities and differences between the sample mean and the sample median.

Example 6.1.2. Let the sample values, of a sample of size $n = 5$, be $X_1 = 2.3,\ X_2 = 3.2,\ X_3 = 1.8,\ X_4 = 2.5,\ X_5 = 2.7$. Find the sample mean and the sample median.

Solution. The sample mean is

$$\overline{X} = \frac{2.3 + 3.2 + 1.8 + 2.5 + 2.7}{5} = 2.5.$$

To find the median, we first order the values from smallest to largest:

$$X_{(1)} = 1.8, \quad X_{(2)} = 2.3, \quad X_{(3)} = 2.5, \quad X_{(4)} = 2.7, \quad X_{(5)} = 3.2.$$

The sample size here is odd, and $\dfrac{n+1}{2} = 3$. Thus the median is

$$\widetilde{X} = X_{\left(\frac{n+1}{2}\right)} = X_{(3)} = 2.5.$$

Remark 6.1.2. *If, in the sample used in Example 6.1.2, the largest observation had been $X_{(5)} = 4.2$ instead of 3.2, we would have*

$$\overline{X} = 2.7, \quad \widetilde{X} = 2.5.$$

*Thus the value of $\overline{X}$ is affected by extreme observations (**outliers**), where as the median is not. In general, if the histogram (or stem-and-leaf display) of the data displays positive skewness, the sample mean will be larger than the sample median, and if it displays negative skewness the sample mean will be smaller than the sample median. This is analogous to the relation of the population mean and median when the pdf is positively or negatively skewed.*

Definition 6.1.3. *The* **sample lower quartile** *or q_1 is defined as*

$$\boxed{q_1 = \textit{ Median of smallest half of the values}}$$

where, if n is even the smallest half of the values consists of the smallest $n/2$ values, and if n is odd the smallest half consists of the smallest $(n+1)/2$ values. Similarly, the **sample upper quartile** *or q_3 is defined as*

$$\boxed{q_3 = \textit{ Median of largest half of the values}}$$

where, if n is even the largest half of the values consists of the largest $n/2$ values, and if n is odd the largest half consists of the largest $(n+1)/2$ values. The sample quartiles are estimators of the corresponding population quartiles.

For sample sizes that are not of the form 6 + (a multiple of 4), the above definition uses interpolation to define the sample lower and upper quartiles. This interpolation, though cruder than that employed in most software packages, is convenient for hand calculations.

Definition 6.1.4. *The* **sample interquartile range***, or simply* **sample IQR***, defined as*

$$IQR = q_3 - q_1$$

is an estimator of the population IQR, which is a measure of variability.

Example 6.1.3. Suppose that the sample values of a sample of size $n = 8$ are 9.39, 7.04, 7.17, 13.28, 7.46, 21.06, 15.19, 7.50. Find the lower and upper quartiles.

Solution. Since n is even, q_1 is the median of the

$$\text{Smallest } 4(= n/2) \text{ values: } 7.04\ 7.17\ 7.46\ 7.50$$

and q_3 is the median of the

$$\text{Largest } 4(= n/2) \text{ values: } 9.39\ 13.28\ 15.19\ 21.06.$$

Thus $q_1 = \dfrac{7.17 + 7.46}{2} = 7.31$, and $q_3 = \dfrac{13.28 + 15.19}{2} = 14.23$.
If there was an additional observation of 8.20, so $n = 9$, then (check)

$$q_1 = 7.46, \qquad q_3 = 13.28.$$

6.2 Estimators for Bivariate Data

6.2.1 Pearson's Linear Correlation Coefficient

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a bivariate population. The **sample covariance**, which estimates the population covariance, is defined as

$$\begin{aligned} \widehat{\sigma}_{X,Y} = \widehat{\mathrm{Cov}}(X,Y) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\right], \end{aligned}$$

where the second expression is the recommended computational formula. With this definition the sample version of **Pearson's linear correlation coefficient** is defined as

$$\widehat{\rho}_{X,Y} = \frac{\widehat{\sigma}_{X,Y}}{S_X S_Y},$$

where S_X and S_Y are the sample standard deviations of the X- and Y-samples.

Example 6.2.1. Consider the $n = 10$ pairs of (X, Y)-values:

```
X-values:  4.9681 2.1757 3.47928 2.2873 1.7415  4.0740 2.3046
Y-values: 95.2380 0.5249 21.4913 0.7289 0.1404 75.8636 0.7781
```

```
X-values:  3.6008 2.2666  0.7241
Y-values: 28.2765 0.6569 -0.0068
```

With this data, $\sum X_i = 27.622$, $\sum Y_i = 223.692$, $\sum X_i Y_i = 965.142$. Thus,

$$\widehat{\sigma}_{X,Y} = \frac{1}{9}\left[965.142 - \frac{1}{10}(27.622)(223.692)\right] = 38.5844,$$

and

$$\rho_{X,Y} = \frac{38.5844}{(1.2483)(35.0952)} = 0.881.$$

6.2.2 Spearman's Rank Correlation Coefficient

The sample version of Spearman's correlation coefficient is also called **Spearman's rank correlation coefficient**. In loose terms, it is defined as the Pearson's correlation coefficient computed by replacing the observations by their *ranks*, or *mid-ranks*.

Definition 6.2.1. *Given a sample $X_1, \ldots, X_n$, such that no two observations have the same value, arrange them from smallest to largest (in increasing order). Then the* **rank**, *R_i, of observation X_i, is defined to be the position of X_i in this ordering of the observations. In other words, the rank of X_i equals the number of observations that are less than it plus one, or the number of observations that are less than or equal to it.*

Thus, the smallest observation has rank 1 while the largest has rank n.

If some observations share the same value, then they cannot be assigned ranks as described above because they cannot be arranged from smallest to largest in a unique way.

Example 6.2.2. Suppose that $X_1 = 2.0$, $X_2 = 1.9$, $X_3 = 1.8$, $X_4 = 2.3$, $X_5 = 2.1$, $X_6 = 1.9$, $X_7 = 2.5$. Because X_2 and X_6 are tied, the observations can be arranged in increasing order either as

$$X_3 \quad X_2 \quad X_6 \quad X_1 \quad X_5 \quad X_4 \quad X_7, \quad \text{or as}$$

$$X_3 \quad X_6 \quad X_2 \quad X_1 \quad X_5 \quad X_4 \quad X_7$$

In either way of ordering of the observations in the above example, the pair of observations X_2 and X_6 occupies the pair of ranks 2 and 3, but it is not clear which observation has which rank. The solution is to assign **mid-ranks** to tied observations, which is the average of the ranks they would have received if they were all different, or, in other words, the average of the ranks that they occupy. In the context of the above example, X_2 and X_6 occupy the pair of ranks 2 and 3, and thus the are both assigned the mid-rank of 2.5. Another example of computing mid-ranks follows.

Sample:	2.1	1.9	1.9	2.0	2.1	2.1	1.8
Mid-ranks:	6	2.5	2.5	4	6	6	1

With the above definition of ranks and mid-ranks, the computation of Spearman's rank correlation coefficient proceeds as follows:

- Rank the observations in the X-sample from smallest to largest. Let

  $$R_1^X, R_2^X, \ldots, R_n^X$$

  denote the (mid-)ranks of $X_1, X_2, \ldots, X_n$.

- Rank the observations in the Y-sample from smallest to largest. Let

  $$R_1^Y, R_2^Y, \ldots, R_n^Y$$

  denote the (mid-)ranks of $Y_1, Y_2, \ldots, Y_n$.

- Compute Pearson's linear correlation coefficient on the pairs of (mid-)ranks

  $$(R_1^X, R_1^Y), \ldots, (R_n^X, R_n^Y).$$

  This is Spearman's rank correlation coefficient.

Example 6.2.3. Consider the (X, Y)-values given in Example 6.2.1. Then,

```
Ranks of X-values: 10 3 7 5 2 9 6 1 8 4
Ranks of Y-values: 10 3 7 5 2 9 6 1 8 4
```

Pearson's correlation coefficient on the ranks, which is Spearman's rank correlation, is 1.

6.2.3 The Regression Function

Let (X, Y) be a bivariate random variable and suppose that the regression function, $\mu_{Y|X=x}$ or $E(Y|X = x)$, of Y on X is of the form

$$E(Y|X = x) = \alpha_1 + \beta_1 x.$$

Then, it can be shown (see also Proposition 4.8.1) that

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \text{and} \quad \alpha_1 = E(Y) - \beta_1 E(X). \tag{6.2.1}$$

Thus, if we have a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the underlying population, the intercept and slope of the regression line can be estimated by

$$\widehat{\beta}_1 = \frac{\widehat{\text{Cov}}(X, Y)}{S_X^2}, \quad \text{and} \quad \widehat{\alpha}_1 = \overline{Y} - \widehat{\beta}_1 \overline{X}, \tag{6.2.2}$$

where $\widehat{\text{Cov}}(X, Y)$ is the sample covariance, and S_X^2 is the sample variance of the X_is. The estimated regression line is

$$\widehat{\mu}_{Y|X=x} = \widehat{\alpha}_1 + \beta_1 x \ . \tag{6.2.3}$$

Remark 6.2.1. *The estimators of the regression coefficients in (6.2.2) can also be derived with the method of least squares (see Section 7.3) and thus are commonly referred to as the least squares estimators (LSE).*

Example 6.2.4. To calibrate a method for measuring lead concentration in water, the method was applied to twelve water with known lead content. The concentration measurements (y) and the known concentration (x) is given below.

x	5.95	2.06	1.02	4.05	3.07	8.45	2.93	9.33	7.24	6.91	9.92	2.86
y	6.33	2.83	1.65	4.37	3.64	8.99	3.16	9.54	7.11	7.10	8.84	3.56

Assuming that the regression function of Y on X is linear, estimate the regression line.

Solution: With this data,

$$\widehat{\text{Cov}}(X, Y) = 8.17, \quad S_X^2 = 9.19, \quad \overline{X} = 5.32, \quad \text{and} \quad \overline{Y} = 5.6.$$

Thus,

$$\widehat{\beta}_1 = \frac{8.17}{9.19} = 0.89, \quad \widehat{\alpha}_1 = 5.6 - 0.89 \times 5.32 = 0.87,$$

so the estimated regression line is

$$\widehat{\mu}_{Y|X=x} = 0.87 + 0.89x.$$

The figure below shows the scatter plot of the data superimposed on the estimated regression line.

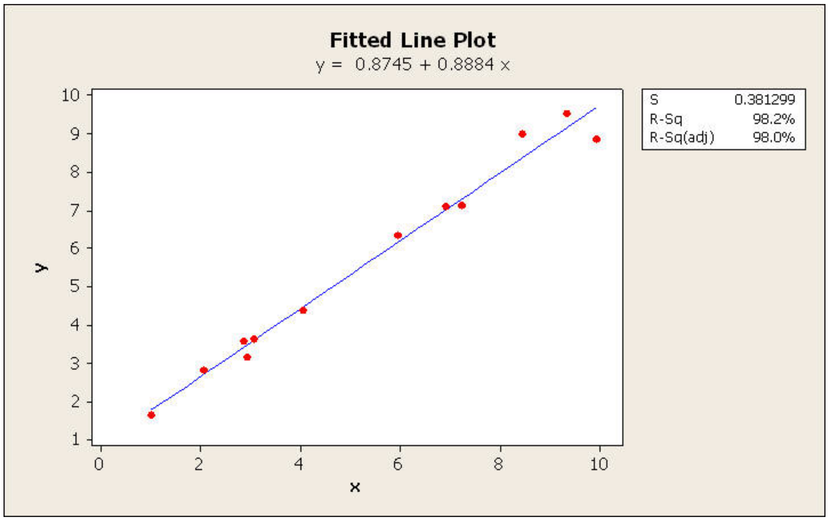


Figure 6.1: Calibration Data and Fitted Regression Line

6.3 Box Plots

6.3.1 One Sample Box Plots

A **box plot** presents a simple but effective visual description of the main features, including symmetry or skewness, of a data set $X_1, \ldots, X_n$. This is achieved, essentially, by presenting the five numbers: the smallest observation ($X_{(1)}$), the sample lower quartile (q_1), the sample median ($\widetilde{X}$), the sample upper quartile (q_3), and the largest observation ($X_{(n)}$).

A box plot displays the central 50% of the data with a box, the left (or lower) edge of which is at q_1 and the right (or upper) edge at q_3. A line inside the box represents the median. The lower 25% and upper 25% of the data are represented by lines (or *whiskers*) which extend from each edge of the box. The lower whisker extends from q_1 until the smallest observation within 1.5 interquartile ranges from q_1. The upper whisker extends from q_3 until the largest observation within 1.5 interquartile ranges from q_3.

Observations further from the box than the whisker ends (i.e. smaller than $q_1 - 1.5 \times \text{IQR}$ or larger than $q_3 + 1.5 \times \text{IQR}$) are called **outliers**, and are plotted individually.

Example 6.3.1. The following are 14 Ozone measurements (Dobson units) taken in 2002 from the lower stratosphere, between 9 and 12 miles (15 and 20 km) altitude are: 315 249 234 269 276 252 230 297 256 286 390 280 441 213. For this data, $X_{(1)} = 213$, $q_1 = 249$, $\widetilde{X} = 272.5$, $q_3 = 297$, $X_{(14)} = 441$. The interquartile range is $\text{IQR} = 297 - 249 = 48$, and

$q_3 + 1.5 \times \text{IQR} = 369$. Thus, the 390 and 441 are outliers. The boxplot of this data is shown below.

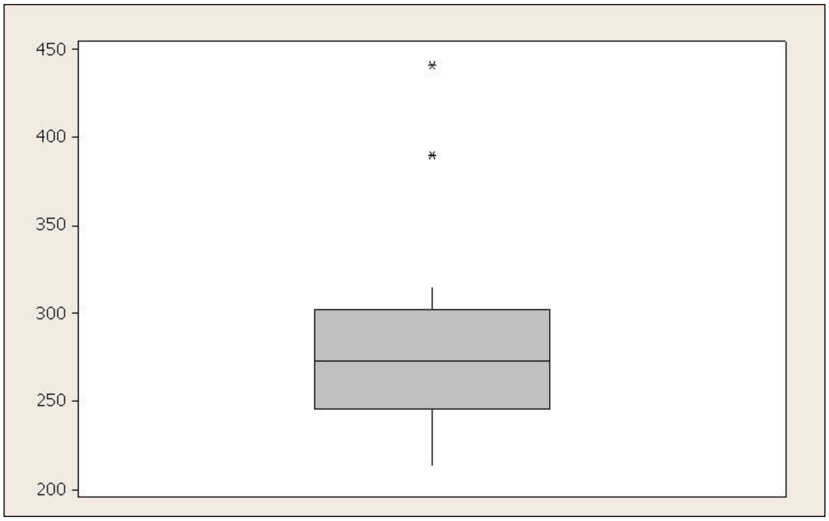


Figure 6.2: Boxplot of Ozone Data

6.3.2 Comparative Box Plots

Figures showing boxplots of different data sets side-by-side are useful for comparing the main features of samples from different populations.

Example 6.3.2. The ignition times of two types of material used for children's clothing are (to the nearest hundredth of a second). The $m = 25$ observations $(x_1, \ldots, x_{25})$ from material type A, and $n = 28$ observations $(y_1, \ldots, y_{28})$ from material type B are

x-values :	7.44	8.05	2.64	6.31	6.55	2.25	3.14	7.74	3.40	2.99	4.61
	3.01	4.27	4.96	5.82	4.25	3.35	3.42	4.89	4.41	2.42	8.61
	8.35	2.87	4.22								
y-values :	8.25	10.95	1.47	1.93	7.31	6.18	6.21	3.35	4.20	4.01	4.97
	7.53	8.84	4.58	7.20	7.30	9.12	8.08	8.40	8.88	8.97	7.54
	7.45	10.86	2.10	10.49	10.12	2.79					

A comparative boxplot for these two samples is shown below.

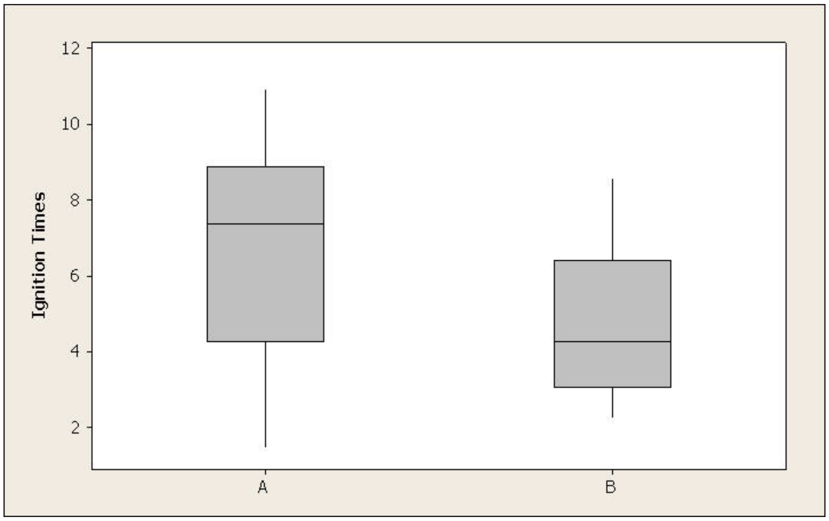


Figure 6.3: Comparative Boxplot of Ignition Times

6.4 Stem-and-Leaf Diagrams

Stem-and-leaf diagrams are useful for organizing and displaying univariate data. The basic steps for presenting a data set $x_1, \ldots, x_n$ by a stem-and-leaf diagram are:

1. Break each observation x_i in two parts

   (a) The **stem**, which consists of the beginning digit(s) of x_i, and

   (b) the **leaf**, which consists of the remaining digit(s) of x_i.

2. Form a vertical column listing the stem values.

3. Write each leaf value in a horizontal column on the right of the corresponding stem.

4. Display counts on the left margin.

Example 6.4.1. The following figure presents a stem-and-leaf display of $n = 40$ solar intensity measurements (watts/m^2) on different days at a location in southern Australia. (The decimals digits are not shown.)

4	64	3	3	6	7			
8	65	0	2	2	8			
11	66	0	1	9				
18	67	0	1	4	7	7	9	9
(4)	68	5	7	7	9			
18	69	0	0	2	3			
14	70	0	1	2	4	4	5	
8	71	0	1	3	6	6	6	
2	72	0	8					

Stem: Hundreds and tens digits; Leaf: Ones digits

The numbers in the left column help obtain certain counts, as follows:
(a) The number in parenthesis is the *size* or *depth* of the *central* stem, i.e. the one which contains the middle value(s); thus, 4 is the size of the stem 68 which contains the middle values 687 and 687.
(b) The sum of the number in parenthesis and the ones above and below it equal the sample size n; thus, $18 + 4 + 18 = 40$.
(c) the depth of each stem can be obtained as difference of the corresponding number in the left column minus the number above it (for stems above the central stem) or below it (for stems below the central stem); thus, $11 - 8 = 3$ which is the depth of the stem 66, and $14 - 8 = 6$ which is the depth of the stem 70.

As seen from this example, stem-and-leaf diagrams provide a basis for evaluating how the data values are *distributed* within the *sample range* (i.e. from 643 to 728). Some of the noteworthy aspects of the distribution of the data include: the shape of the distribution, the typical value (or center) of the data, the variability (or spread) of the data, the number and location of peaks or *modes*, the presence of gaps in the data, and the presence of observations that are unusually far from the bulk of the data (*outliers*). The amazing thing about stem-and-leaf diagrams is not only that they readily provide so much insight about the data, but that they do so while keeping most of the original information. The only information lost is the order in which the observations were obtained, making them unsuitable for detecting a possible time pattern in the data.

At a first glance, the steam-and-leaf diagram diagram in Example 6.4.1 seems to indicate bimodality. But the peaks are not very high and may be due to the randomness of the sample. Thus, the diagram is not inconsistent with a shape that is roughly flat (uniform distribution).

Choice of Stems in a Stem-and-Leaf Diagram

The number of stems used in a steam-and-leaf diagram plays an important role in our ability to see interesting features of the data. For example, if we had chosen a one-digit stem (only hundreds) to display solar intensity data, the diagram would be too crude or clumpy. On the other hand, in a display with three-digit stems and only the decimals as stems, the information about the shape of the distribution would be compromised my the excessive detail.

Occasionally we need to split the stems to get a better display, as in the following example.

Example 6.4.2. The following figure presents a stem-and-leaf display of US beer production (in millions of barrels) for a different quarter during the period 1975-1982. In this figure, stem 4 has been split into 4L for the 'low' leafs (0, 1, 2, 3, 4), while 4H hosts the 'high' leafs.

1	5H	5										
7	5L	0	2	2	3	3	4					
14	4H	6	6	7	8	8	9	9				
(11)	4L	1	1	1	2	2	4	4	4	4	4	4
6	3H	5	6	6	6	9	9					

More detailed splits of stems can also be used in order to achieve a better visual impression of the distribution of the data.

6.5 Histograms: Univariate Data

Histograms offer a different way for organizing and displaying data. A histogram does not retain as much information on the original data as a stem-and-leaf diagram, in the sense that the actual values of the data are not displayed. On the other hand, histograms offer more flexibility in selecting the *classes* (which are the analogue of the stems), and can be used for bivariate data as well. The flexibility in selection the classes makes them suitable as estimators of the underlying probability density function.

Basic steps for presenting a data set $x_1, \ldots, x_n$ as a histogram:

1. Divide the range of the data values into consecutive intervals.

2. Define the end points of the intervals with one more decimal than used for the data set.

3. Count the number of data points that fall in each interval. These are the **frequencies**.

4. Construct a table (the **frequency distribution table**) showing the intervals and their frequencies. The table can be enhanced by showing the **relative frequencies** and the **cumulative relative frequencies**.

5. Construct the histograms by drawing a box, or vertical bar, above each class interval whose height is $f_i/(\text{class width})$.

Example 6.5.1. The following is a frequency table of lifetimes (measured in thousands of cycles) of 100 strips of aluminum subjected to a stress of 20,000 psi.

	Class	nf_i	f_i	F_i
		(Freq.)	(Rel. Freq.)	(Cum. Rel. Freq.)
1.	350-<550	1	0.01	0.01
2.	550-<750	4	0.04	0.05
3.	750-<950	9	0.09	0.14
4.	950-<1150	15	0.15	0.29
5.	1150-<1350	18	0.18	0.47
6.	1350-<1550	19	0.19	0.66
7.	1550-<1750	14	0.14	0.80
8.	1750-<1950	10	0.10	0.90
9.	1950-<2150	7	0.07	0.97
10.	2150-<2350	3	0.03	1.000
		n=100	1.000	

The histogram for the above frequency distribution is given below.

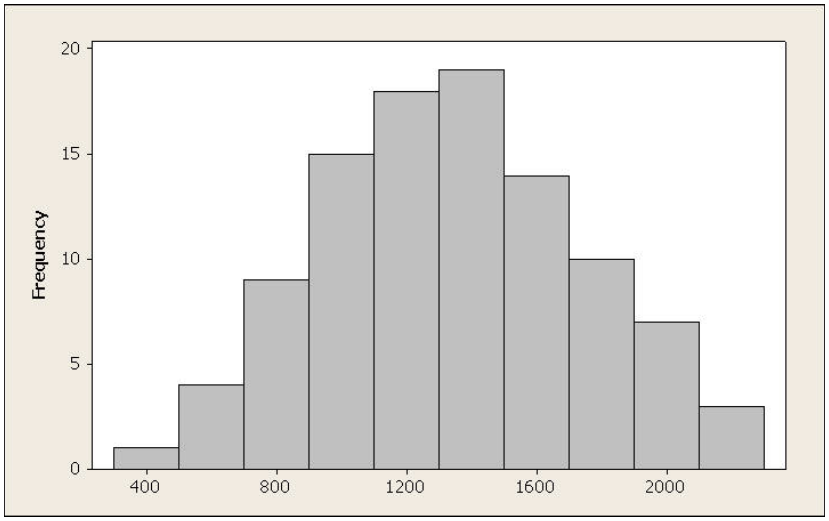


Figure 6.4: Histogram of 100 Lifetimes of Aluminum Strips

6.5.1 Plotting the Cumulative Relative Frequencies

A plot of the cumulative relative frequencies gives a different perspective on the distribution of the data. The cumulative relative frequencies approximate (serve as estimators of) the cumulative distribution function. The following figure shows such a plot for the frequency table given in Example 6.5.1.

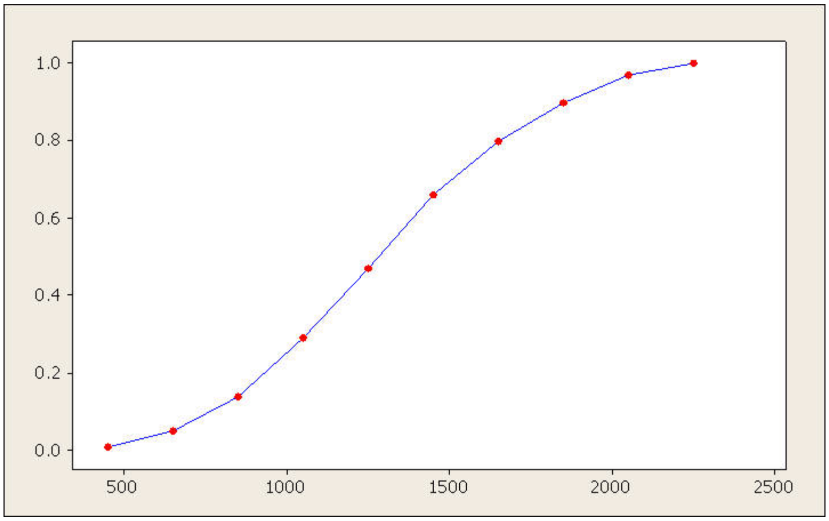


Figure 6.5: Cumulative Relative Frequencies of 100 Lifetimes of Aluminum Strips

6.5.2 Relative Frequency Histograms as Empirical Estimators of the PDF

Taking the heights of the boxes in a histogram equal to the relative frequencies, results in the *relative frequency histogram.* Since changing from frequencies to relative frequencies involves a change in the scale of the y-axis, the shape of the relative frequency histogram remains the same.

According to the relative frequency interpretation of probability, the relative frequency of observations falling in a given class interval approximates the probability of an observation falling in that interval. This is true even if the class intervals become suitably shorter as the sample size increases. Since probabilities are areas under the probability density function, it follows that the outline of the histogram will approximate the underlying pdf, and the approximation becomes increasingly better as the sample size increases. This is illustrated in the following figures which show histograms from 150 and of 5,000 observations from a gamma distribution with shape parameter 2 and scale parameter 1, with the pdf of the fitted gamma model superimposed.

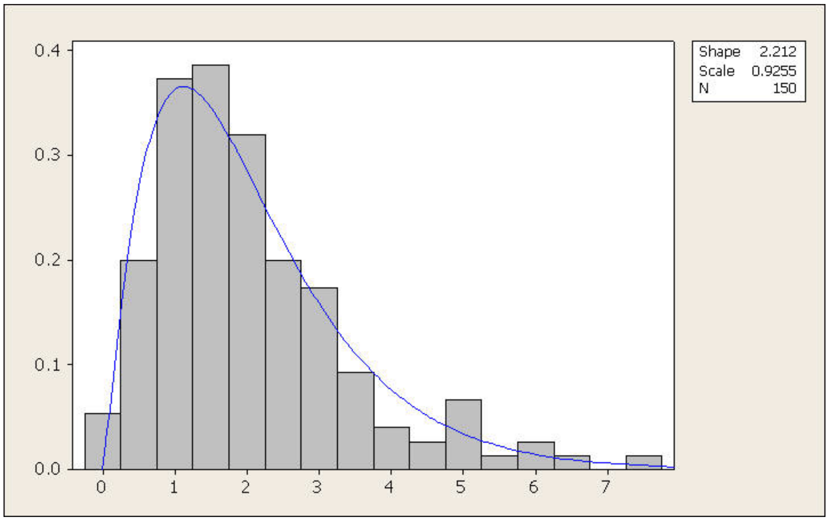


Figure 6.6: Histogram of 150 Random Data Drawn from Gamma(2,1), with the PDF of the Best Fitting Gamma distribution superimposed.

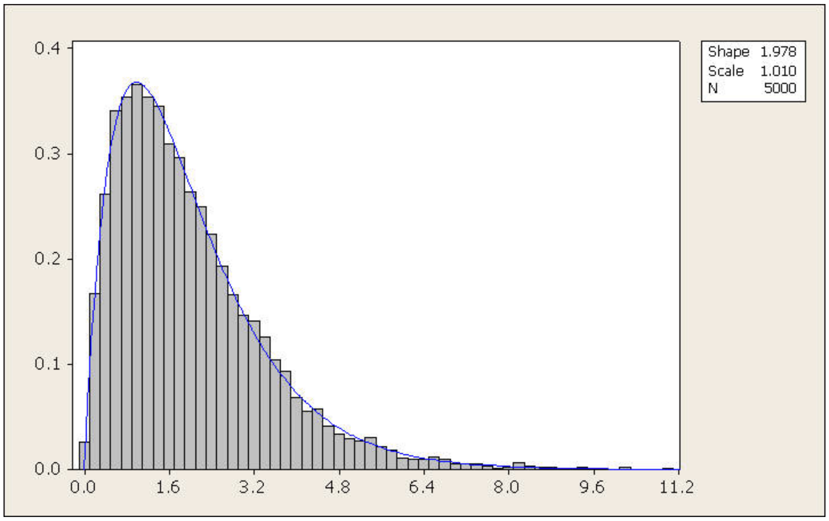


Figure 6.7: Histogram of 5,000 Random Data Drawn from Gamma(2,1), with the PDF of the Best Fitting Gamma Distribution Superimposed.

The box in the top right corner of the figures also shows the empirical values (estimators) of the shape and scale parameters. Note that the estimators based on 5,000 observations are much closer to the true values of the parameters.

6.5.3 Bivariate Data

A useful plot for bivariate data is the so-called **marginal plot**. The plot displays a scatter plot of the data points (X_i, Y_i), $i = 1, \ldots, n$, with histograms of the marginal data distributions on the margins.

Example 6.5.2. Figure 6.8 shows the marginal plot for 50 data points (X_i, Y_i), where the X_is are computer generated from the exponential distribution with mean 1 (Exp(1)) and the Y_is are generated independently from the $N(0, 1)$ distribution.

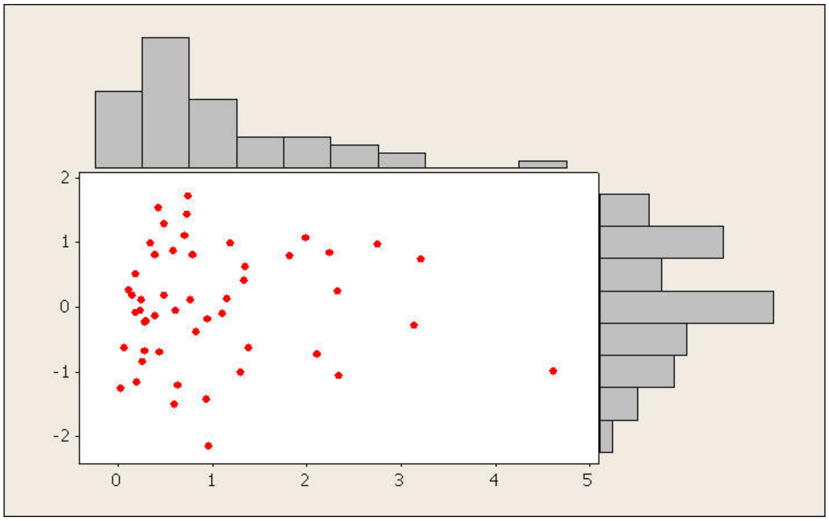


Figure 6.8: Example of a Marginal Plot.

A different version of the marginal plot used box plots in the margins.

6.6 3-D Scatter Plots

For 3-dimensional data, a *3-D scatter plot*, also known as *3-D dot plot*, is helpful both for visualizing how the data points are distributed in space, and for discerning relationships between the variables.

Once the data are entered in columns labeled X, Y, and Z, a 3-D scatter plot can be constructed with Minitab using the following commands.

```
  Graph>3D Scatterplot>"Simple">OK>Enter the Z, Y and X variables>OK
```

The graph can be rotated about each of the axes to achieve a better visual understanding of the data distribution and the relation between the variables.

Example 6.6.1. The electricity consumed in an industrial plant in 30 consecutive 30-day periods, together with the average temperature and amount (in tons) of production is given in the following table.

Temp	46	45	57	58	63	70	84	45	83	82	35	51	78	66	71
Prod	95	91	93	102	94	86	106	95	93	98	90	96	89	110	109
Elec	330	329	363	375	363	381	358	362	407	388	350	353	376	406	382

Temp	33	36	79	81	39	48	62	46	71	47	42	68	81	84	64
Prod	85	104	97	88	90	92	90	93	109	103	110	94	95	107	95
Elec	337	398	380	402	328	351	336	354	372	354	340	391	386	383	376

Entering the data in Minitab and using the above commands we get the plot

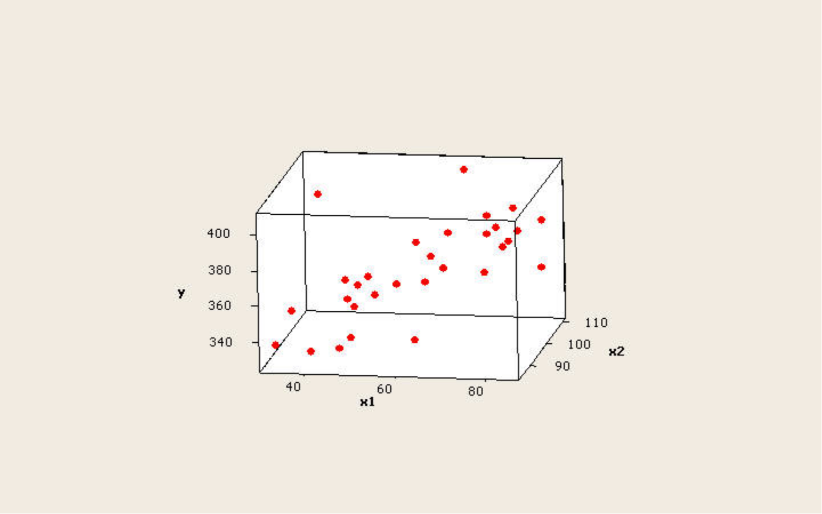


Figure 6.9: 3D Scatterplot of the Electricity Data.

6.7 Bar Graphs and Pie Charts

Bar graphs and *pie charts* are used to depict the breakdown of a certain population into categories. Thus, they apply to qualitative or categorical data. For example, the ethnic, or religion composition, the education or income level of a certain population. Bar graphs resemble histograms in the sense of using bars to represent proportions, or counts, of the different categories. In a pie chart, the population is represented by a circle (pie) which is sliced up in pieces whose size represents proportions.

Example 6.7.1. The site http://wardsauto.com/keydata/USSalesSummary0702.xls gives information on the light vehicle market share of car companies. The bar graph and pie chart below depict this data set.

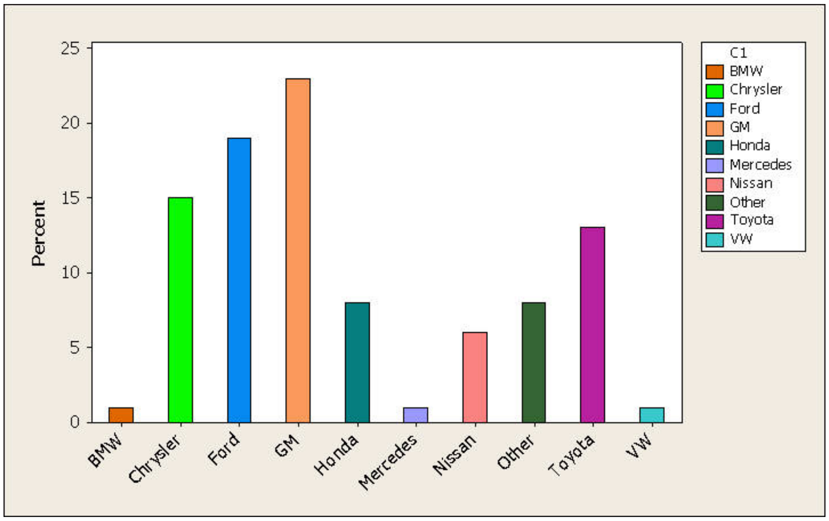


Figure 6.10: Bar Graph of Light Vehicle Market Share Data.

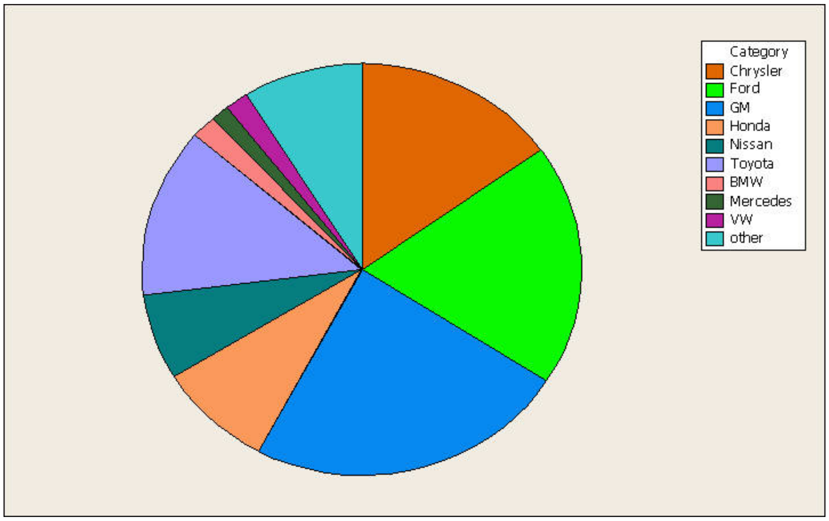


Figure 6.11: Pie Chart of Light Vehicle Market Share Data.

6.8 The Probability Plot

As already mentioned, most experiments resulting in the measurement of a continuous random variable provide little insight as to which probability model best describes the distribution of the measurements. Thus, several procedures have been devised to test the goodness-of-fit of a particular model to a random sample obtained from some population. Here we discuss a very simple graphical procedure, called the **probability plot** (PP).

The basic idea of the PP is to compare the sample percentiles (i.e. the ordered sample values) with the corresponding percentiles of the assumed model distribution. Since sample percentiles estimate corresponding population percentiles, if the population distribution equals the assumed, then a plot of the sample percentiles versus the percentiles of the assumed distribution should fall (approximately) on a straight line of 45^o that passes through the origin.

More precisely, if $X_1, \ldots, X_n$ denotes a simple random sample from some population, the ith order statistic, $X_{(i)}$, is the

$$100(1-\alpha_i)\text{th sample percentile, where } \alpha_i = 1 - \frac{i-0.5}{n}, \quad i = 1, \ldots, n\ . \tag{6.8.1}$$

See Definition 6.1.1. The normal probability plot tests the assumption that the population distribution equals a particular model distribution by plotting the pairs

$$\left(X_{(i)}, Y_{\alpha_i}\right), \quad i = 1, \ldots, n, \tag{6.8.2}$$

where the α_i are given in (6.8.1) and Y_{α_i} are the percentiles of the assumed model distribution.

Example 6.8.1. 1. To test the assumption that the population distribution is normal with a specified mean μ_0 and a specified variance σ_0^2, the PP plots the pairs

$$\left(X_{(i)}, \mu_0 + \sigma_0 z_{\alpha_i}\right), \quad i = 1, \ldots, n.$$

Using a computer generated random sample of size $n = 50$ from $N(0,1)$, Figure 6.12 shows the PP for testing the assumption that the data indeed came from a

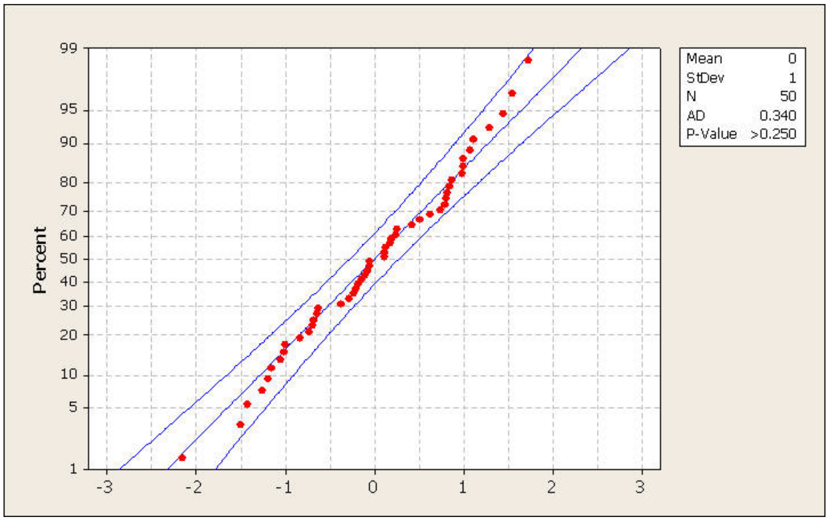


Figure 6.12: Probability Plot for Testing the Normal Model with $\mu = 0$ and $\sigma = 1$.

$N(0,1)$ distribution. The bands around the straight line are 95% *confidence bands*, indicating the degree a deviation from the straight line that each point is expected to have in 95% of the random samples of size 50. The *p-value* mentioned in the upper right-hand insert offers a more objective criterion of the goodness of fit of the assumed model distribution: As a rule of thump, if the p-value is larger than 0.1 there is no reason the suspect the validity of the model assumption.

2. In order to test the assumption that the population distribution is exponential with specified parameter value λ_0, or mean $\mu_0 = 1/\lambda_0$, the probability plot plots the pairs

$$\left(X_{(i)}, -\ln(\alpha_i)\mu_0\right), \quad i = 1, \ldots, n.$$

Using a computer generated random sample of size $n = 50$ from the exponential distribution with parameter value $\lambda = (\mu =)1$, Figure 6.13 shows the probability plot for testing the assumption that the data came from an Exp(1) distribution.

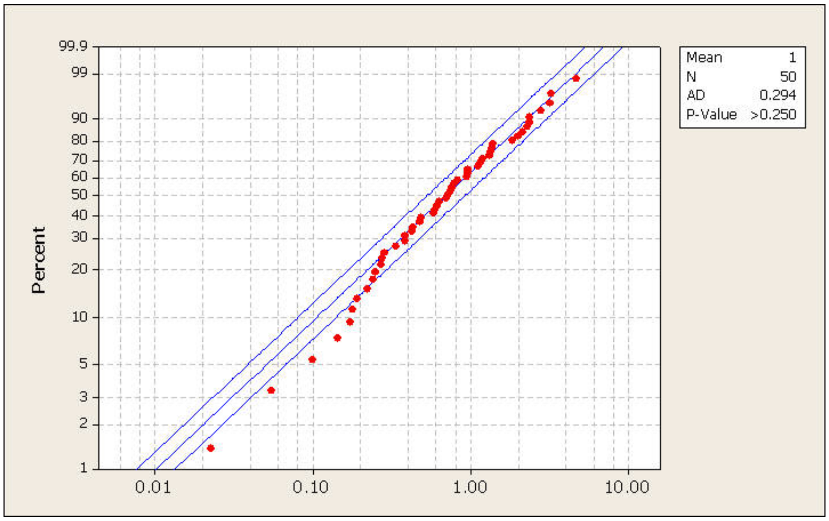


Figure 6.13: Probability Plot for Testing the Exponential Model with $\lambda = 1$.

Though some points fall outside the 95% confidence bands, the p-value suggests that, overall, the deviation of the points from the straight line is well within the expected, so there is no reason to suspect the assumption that the data came from the Exp(1) distribution.

Because in most cases, the distributional assumption made involves an entire family of distributions (i.e. it is assumed that the population distribution belongs in a parametric family of distributions, without specifying a particular member of the family), the model parameters are first estimated and the percentiles of the best-fitting model distribution are used in the probability plot.

Example 6.8.2. 1. In order to test the assumption that the population distribution is a member of the normal family, the mean and standard deviation are estimated by $\overline{X}$ and S, respectively, and the probability plot plots the pairs

$$\left(X_{(i)}, \overline{X} + S z_{\alpha_i}\right), \quad i = 1, \ldots, n.$$

2. In order to test the assumption that the population distribution is a member of the exponential, μ is estimated by $\overline{X}$ and the probability plot plots the pairs

$$\left(X_{(i)}, -\ln(\alpha_i)\overline{X}\right), \quad i = 1, \ldots, n.$$

Using the data from the Exp(1) distribution that were generated in Example 6.8.1, Figure 6.14 shows the probability plot when the mean μ is estimated from the data.

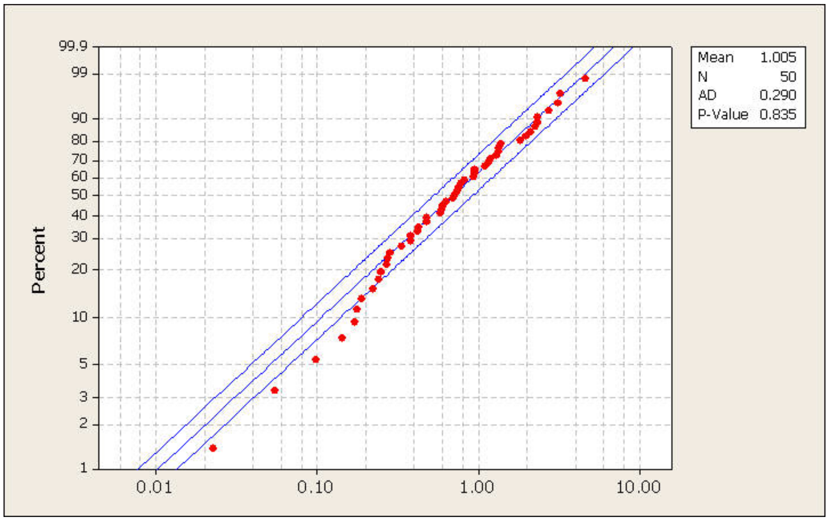


Figure 6.14: Probability Plot for Testing the Exponential Model with Unspecified λ.

In the final figure, Figure 6.15 , the probability plot is used to test the goodness-of-fit of the normal model to the data generated from the Exp(1) distribution that were in Example 6.8.1,

As seen the points plotted display a rather serious deviation from the straight line, and the p-value is quite small. Thus, we would reject the assumption that the data were generated from a normal distribution.

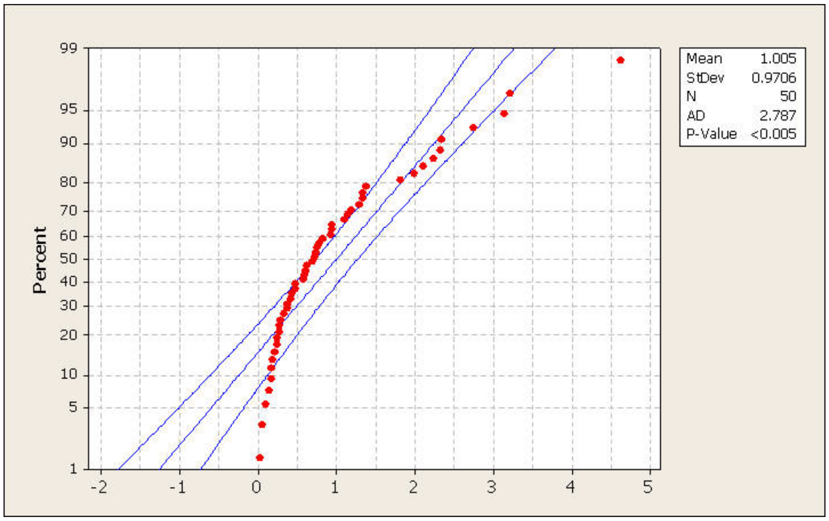


Figure 6.15: Probability Plot for Testing the Normal Model with Unspecified μ and σ.

6.9 Exercises

1. A robot's reaction time was measured for a sample of 22 simulated malfunctions. The measurements, in nanoseconds, are

   28.2 27.9 30.1 28.5 29.4 29.8 30.3 30.6 28.9 27.7 28.4 30.2 31.1

   29.7 29.0 29.5 31.4 28.7 27.0 27.6 29.9 28.0

   Do a histogram of this data set. Construct the corresponding frequency table, displaying also the relative frequencies and the cumulative relative frequencies.

2. The reaction time of a robot of a different design than that in Exercise 1 was measured for a similar sample of 22 simulated malfunctions. The measurements, in nanoseconds, are

   27.67 27.68 27.77 27.85 27.88 27.98 28.04 28.06 28.24 28.50 28.60

   28.68 29.24 29.36 29.48 29.50 29.53 29.56 29.57 29.79 29.88 30.93

   (a) Estimate the population median, the 25th and the 75th percentiles.

   (b) Estimate the population interquartile range.

   (c) What percentile is the 19th ordered value?

   (d) How would you approximate the 90th percentile?

3. The following data show the starting salaries, in $1000 per year, for a sample of 15 senior engineers:

```
152 169 178 179 185 188 195 196 198 203 204 209 210 212 214
```

(a) Do a stem and leaf diagram of the salary data.

(b) Assuming that the 15 senior engineers represent a simple random sample from the population of senior engineers, estimate the population median.

(c) Under the same assumption done above, estimate the 25th and 75th population percentiles.

4. The sample variance of the salary data in Exercise 3 is $S^2 = 312.31$. Give the variance for the data on second-year salaries for the same group of engineers if

   (a) if each engineer gets a \$5000 raise, and

   (b) if each engineer gets a 5% raise.

5. Consider the data in Exercise 1 and do a cumulative relative frequency plot using the frequency table constructed in that exercise.

6. Do a boxplot for the data in Exercise 3, and comment on what data features are immediately discernible from this plot, but were not from the stem-and-leaf diagram.

7. Construct a comparative boxplot using the reaction times for the two designs of robot, given in Exercises 1 and 2. Comment on any differences in the two data distributions. Which robot is preferable?

8. A random sample of 10 cars of type A that were test driven yielded the following gas mileage on the highway: 29.1, 29.6, 30, 30.5, 30.8. A random sample of 14 cars of type B yielded the following gas mileage when test driven under similar conditions: 21, 26, 30, 35, 38.

   (a) For each type of car, estimate the population mean gas mileage.

   (b) For each type of car, estimate the population variance of gas mileage.

   (c) Do a comparative boxplot.

   (d) On the basis of the above analysis, rank the two types of car in terms of quality. Justify your answer.

9. An experiment examined the effect of temperature on the strength of new concrete. After curing for several days at 20^oC, specimens were exposed to temperatures of -8^oC, or 15^oC for 28 days, at which time their strengths were determined. The results are listed in the table below.

Temperature (^oC)	28 Day Strength (MPa)								
-8	56.53	60.08	59.46	65.33	65.13	62.53	61.33	66.13	61.53
15	71.53	62.99	67.33	65.13	66.33	65.13	77.53	69.33	61.13

Do a comparative box plot of the data, and comment on any differences the two data sets display.

10. Consider the data in Exercise 9.

    (a) For each of the two temperatures, estimate the probability that the concrete strength will exceed 63.

    (b) For each of the two temperatures, what population percentile does 63 estimate?

11. The diameter of a ball bearing was measured by 8 inspectors, each using two different kinds of calipers. The results are

Inspector	1	2	3	4	5	6	7	8
Caliper1	0.265	0.255	0.266	0.267	0.269	0.258	0.254	0.267
Caliper2	0.264	0.255	0.265	0.266	0.267	0.259	0.252	0.265

    (a) Do a scatter plot of the data.

    (b) Compute Pearson's correlation coefficient.

    (c) Compute Spearman's correlation coefficient.

12. The article "Toxicity Assessment of Wastewaters, River Waters, and Sediments in Austria Using Cost-Effective Microbiotests", by M. Latif and E. Licek (2004, Environmental Toxicology, Vol 19, No 4, 302-308) reports data on conductivity (μS/cm) measurements of surface water (X), and water in the sediment at the bank of a river (Y), taken at 10 points during winter. The data are

```
X : 220 800 277 223 226 240 752 317 333 340
Y : 386 889 358 362 411 390 927 612 554 532
```

Assume that the regression function of Y on X is linear.

    (a) Estimate the slope and intercept of the regression line.

    (b) Use the estimated regression line to estimate the average sediment conductivity when the surface conductivity is 300.

(c) By how much do you expect the average sediment conductivity to change when the surface conductivity increases by 50?

(d) Assume that the conditional distribution of the response, Y, given that $X = x$ is normal with mean $\mu_{Y|X}(x)$ and variance σ^2. Estimate the probability that a response at $X = 250$ will be larger than a response at $X = 300$.

13. The article ”Effects of Bike Lanes on Driver and Bicyclist Behavior” (Transportation Eng. J., 1977: 243-256) reports data from a study on X = distance between a cyclist and the roadway center line, and Y = the separation distance between the cyclist and a passing car (both determined by photography). The data from ten streets with bike lanes are:

X :	12.8	12.9	12.9	13.6	14.5	14.6	15.1	17.5	19.5	20.8
Y :	5.5	6.2	6.3	7.0	7.8	8.3	7.1	10.0	10.8	11.0

Assume that the regression function of Y on X is linear.

(a) Estimate the slope and intercept of the regression line.

(b) Assume that the unit was changed from feet to inches. How would that affect your estimates of the slope and intercept?

(c) What is the average distance between the cyclist and a passing car if $X = 14$?

(d) By how much do you expect the average average distance between the cyclist and a passing car to change when X increases by 4 feet?

(e) Would you use the estimated regression line to estimate the average distance between the cyclist and a passing car if $X = 10$?

14. Use a statistical software package to generate 50 (X, Y)-values according to the following regression model. X is normal with mean 5 and variance 1. Y is generated as $Y = 15 - 2 \times X + E$, where E is normal with mean 0 and variance 1.

(a) Fit the regression line with the scatter plot. Give the estimates of the coefficients of the regression line. Give the correlation coefficient.

(b) Repeat the above with E normal with mean zero and variance 9. Compare the correlation coefficients.

(c) Do a scatter plot of the two data sets and use it to explain the difference in the correlation coefficients.

Chapter 7

Point Estimation: Fitting Models to Data

7.1 Introduction

A common objective of data collection is to learn about the distribution (or aspects of it), of some characteristic of the units of a population of interest. In technical parlance, the objective is to *estimate* a distribution (or aspects of it). This is done by a suitable extrapolation of sample information to the population.

One approach for achieving this, called **nonparametric**, or **NP**, for short, is to use the sample characteristics as estimates of the corresponding population characteristics. For example, estimation of a population proportion, mean, variance, percentiles, IQR, correlation, and the slope of the regression line, by the corresponding sample quantities, all of which were discussed in Chapters 1 and 6, are examples of nonparametric estimation.

This chapter will focus on another main approach for extrapolating sample information to the population, called the **parametric** approach. This approach starts with the assumption that the distribution of the population of interest belongs in a specific *parametric* family of distribution models. Many such models, including all we have seen, depend on a small number of parameters. For example, Poisson models are identified by the single parameter λ, and normal models are identified by two parameters, μ and σ^2. Under this assumption (i.e. that there is a member of the assumed parametric family of distributions that equals the population distribution of interest), the objective becomes that of estimating the model parameters, in order to identify which member of the parametric family of distributions best *fits* the data. In this chapter, we will learn how to *fit* a particular family of distribution models to the data, i.e. identify the member of the parametric family that best fits the data. Three methods of fitting models to data that we will discuss are: a) the *method of moments*, which derives its name because it identifies the model parameters that correspond (in some sense) to the nonparametric estimation of selected moments, b) the *method of maximum likelihood*, and c) the *method of least squares* which is most commonly used for fitting regression models.

Estimators obtained from the nonparametric and parametric approaches will occasionally differ. For example, under the assumption that the population distribution is normal, estimators of population percentiles and proportions depend only on the sample mean and sample variance, and thus differ from the sample percentiles and proportions; the assumption of a uniform distribution yields an estimator of the population mean value which is different from the sample mean; the assumption of Poisson distribution yields an estimator of the population variance which is different from the sample variance. Moreover, even the three different methods for fitting parametric models, will, occasionally, produce different estimates of the model parameters. Thus, another learning objective of this chapter is to develop criteria for selecting the best among different estimators of the same quantity, or parameter.

In all that follows, the Greek letter θ serves as a generic notation for any model or population parameter(s) that we are interested in estimating. Thus, if we are interested in the population mean value, then $\theta = \mu$, and, if we are interested in the population mean value and variance then, $\theta = (\mu, \sigma^2)$. If θ denotes a population parameter, then, by **true value of** θ we mean the (unknown to us) population value of θ. For example, if θ denotes the population mean or variance, then the true value of θ is the (unknown to us) value of the population mean or variance. If θ denotes a model parameter then, by **true value of** θ we mean the (unknown to us) value of θ that corresponds to the **best-approximating** model, which is the member of the parametric family that best approximates the population distribution. For example, if we use the exponential family to model the distribution of life times of a certain component, then the unknown to us value of $\lambda = 1/\mu$, where μ is the population mean, is the true value of $\theta = \lambda$.

7.2 Estimators and Estimates

7.2.1 Overview

An **estimator**, $\hat{\theta}$, of the true value of θ is a function of the random sample $X_1, \ldots, X_n$,

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n).$$

Here we refer to the random sample in the planning stage of the experiment, when $X_1, \ldots, X_n$ are random variables (hence the capital letters). Being a function of random variables, i.e. a statistic, an estimator is a random variable, and thus it has a sampling distribution. If θ is a population parameter, or, if it is a model parameter and the assumed parametric model is correct, the sampling distribution of an estimator $\hat{\theta}$ depends on the true value, θ_0, of θ. This dependence of the sampling distribution of $\hat{\theta}$ on the true value of θ is indicated by writing the expected value and variance ($E(\hat{\theta})$ and $\text{Var}(\hat{\theta})$) of $\hat{\theta}$ as

$$E_{\theta=\theta_0}\left(\hat{\theta}\right), \quad \text{and} \quad \text{Var}_{\theta=\theta_0}(\hat{\theta})$$

and read "the expected value of $\hat{\theta}$ when the true value of θ is θ_0", and "the variance of $\hat{\theta}$ when the true value of θ is θ_0".

Example 7.2.1. a) Car manufacturers often advertise damage results from low impact crash experiments. In an experiment crashing $n = 20$ randomly selected cars of a certain type against a wall at 5 mph, let X denote the number of cars that sustain no visible damage. The parameter of interest here is $\theta = p$, the probability that a randomly selected car will sustain no visible damage in such low impact crash. Note that p is both a population parameter and a model parameter, since, under the reasonable assumption that crash results of different cars are independent, $X \sim Bin(20, p)$. As estimator of p we use the proportion, $\hat{p} = X/20$, of cars that sustain no visible damage when crashed at 5 mph. For example, if $X = 12$ of the 20 cars in the experiment sustain no visible damage, the estimate of p is $\hat{p} = 12/20$, and the best fitting model is $\text{Bin}(20, 0.6)$. Because $X = 20\hat{p} \sim Bin(20, p)$, we see that the distribution of $\hat{p}$ depends on the true value, p_0, of p. For example, if the true value is $p_0 = 0.7$, then $n\hat{p} = X \sim Bin(n, 0.7)$. This dependence of the distribution of X on the true value p is often indicated in the formula for the expected value and variance of a binomial random variable, which is $E(X) = 20p$ and $\text{Var}(X) = 20p(1-p)$, by writing $E_p(X) = 20p$ and $\text{Var}_p(X) = 20p(1-p)$. In particular, if the true value of p is $p_0 = 0.7$, we write

$$E_{p=0.7}(X) = 20 \times 0.7 = 14, \quad \text{and} \quad E_{p=0.7}(\hat{p}) = 0.7,$$

and read "when the true value of p is 0.7, the expected value of X is 14 and the expected value of $\hat{p}$ is 0.7". Similarly, we write

$$\text{Var}_{p=0.7}(X) = 20 \times 0.7 \times 0.3 = 4.2, \quad \text{and} \quad \text{Var}_{p=0.7}(\hat{p}) = 0.7 \times 0.3/20 = 0.0105,$$

and read "when the true value of p is 0.7, the variance of X is 4.2 and the variance of $\hat{p}$ is 0.0105".

Example 7.2.2. The response time, X, of a robot to a type of malfunction in a certain production process (e.g. car manufacturing) is often the variable of interest. Let $X_1, \ldots, X_{36}$ denote 36 response times that are to be measured. Here it is not clear what the population distribution (i.e. the distribution of each X_i) might be, but we might be interested in estimating the population mean, μ (which is also the expected value of each X_i). As estimator of μ we use the average, $\overline{X}$. Since we do not know the distribution of each X_i, we also do not know the exact distribution of their average, $\overline{X}$ (though the CLT can be used to approximate its distribution). Nevertheless, we know that the expected value of $\overline{X}$ is the same as the population mean (i.e. the same as the expected value of each X_i). We indicate this by writing $E_\mu(\overline{X}) = \mu$. In particular, if the true value of μ is 8.5, we write

$$E_{\mu=8.5}\left(\overline{X}\right) = 8.5,$$

and read "when the true value of μ is 8.5, the expected value of $\overline{X}$ is 8.5". Similarly, the variance of $\overline{X}$ equals the population variance (i.e. the variance of each X_i) divided by the sample size, which here is $n = 36$. We indicate this by writing $\text{Var}_{\sigma^2}(\overline{X}) = \sigma^2/36$. In particular, if the true value of σ^2 is 18, we write

$$\text{Var}_{\sigma^2=18}(\overline{X}) = \frac{18}{36} = 0.5,$$

and read "when the true value of σ^2 is 18, the variance of $\overline{X}$ is 0.5".

Let $X_1 = x_1, \ldots, X_n = x_n$ be the observed values when the experiment has been carried out. The value, $\hat{\theta}(x_1, \ldots, x_n)$, of the estimator evaluated at the observed sample values will be called a **point estimate** or simply an **estimate**. Point estimates will also be denoted by $\hat{\theta}$. Thus, an estimator is a random variable, while a (point) estimate is a specific value that the estimator takes.

Example 7.2.3. a) In Example 7.2.1, the proportion $\widehat{p} = X/20$ of cars that will sustain no visible damage among the 20 that are to be crashed is an estimator of p. If the experiment yields that $X = 12$ sustain no visible damage, then $\widehat{p} = 12/20$ is the point

estimate of p.
b) In Example 7.2.2, $\widehat{\mu} = \overline{X} = (X_1 + \cdots + X_{36})/36$ is an estimator of μ. If the measured response times $X_1 = x_1, \ldots, X_{36} = x_{36}$ yield an average of $(x_1 + \cdots + x_{36})/36 = 9.3$, then $\widehat{\mu} = 9.3$ is a point estimate of μ.

7.2.2 Biased and Unbiased Estimators

Being a random variable, $\widehat{\theta}$ serves only as an approximation to the true value of θ. With some samples, $\widehat{\theta}$ will overestimate the true θ, whereas for others it will underestimate it. Therefore, properties of estimators and criteria for deciding among competing estimators are most meaningfully stated in terms of their distribution, or the distribution of

$$\hat{\theta} - \theta = \textbf{error of estimation}. \tag{7.2.1}$$

The first criterion we will discuss is that of *unbiasedness.*

Definition 7.2.1. *The estimator $\hat{\theta}$ of θ is called* **unbiased** *for θ if*

$$E(\hat{\theta}) = \theta.$$

(More exactly, $\hat{\theta}$ is unbiased for θ if $E_\theta(\hat{\theta}) = \theta$.) The difference

$$E(\hat{\theta}) - \theta$$

is called the **bias** *of $\hat{\theta}$ and is denoted by bias$(\hat{\theta})$. (More exactly, bias$_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$.)*

Example 7.2.4. The estimators $\widehat{p}$, $\overline{X}$, and S^2 are unbiased estimators of p, μ and σ^2. That is,

$$E_p(\widehat{p}) = p, \quad E_\mu(\overline{X}) = \mu, \quad E_{\sigma^2}(S^2) = \sigma^2.$$

Example 7.2.5. *The least squares estimators of the regression parameters, given in (6.2.2), are unbiased. Thus,*

$$E_{\beta_1}\left(\widehat{\beta}_1\right) = \beta_1, \quad \textit{and} \quad E_{\alpha_1}\left(\widehat{\alpha}_1\right) = \alpha_1.$$

Unbiased estimators have zero bias. This means that, though with any given sample $\hat{\theta}$ may underestimate or overestimate the true value of θ, the estimation error $\hat{\theta} - \theta$ averages to zero. Thus, when using unbiased estimators, there is no tendency to overestimate or underestimate the true value of θ.

7.2.3 The Standard Error and the Mean Square Error

Though unbiasedness is a desirable property, many commonly used estimators do have a small, but non-zero, bias. Moreover, there are cases where there is more than one unbiased estimator. The following examples illustrate both possibilities.

Example 7.2.6. a) Let $X_1, \ldots, X_n$ be a random sample from some population having a continuous symmetric distribution. Thus, the mean and the median coincide. Then the sample mean $\overline{X}$, the sample median $\widetilde{X}$, any trimmed mean, as well as hybrid estimators such as

$$\text{median}\left\{\frac{X_i + X_j}{2}; i, j = 1, \ldots, n, i \neq j\right\},$$

are unbiased estimators for μ.
b) The sample standard deviation S is a biased estimator of the population standard deviation σ.
c) Let $X_1, \ldots, X_n$ be the life times of a random sample of n valves, and assume that each life time has the exponential distribution with parameter λ (so $\mu = 1/\lambda$). Then $1/\overline{X}$ is a biased estimator of $\lambda = 1/\mu$, and $\exp\{-500/\overline{X}\}$ is a biased estimator of $\exp\{-\lambda 500\} = P(X > 500)$, the probability that the life of a randomly chosen valve exceeds 500 time units of operation.
d) Let $X_1, \ldots, X_n$ be a sample from a normal distribution with parameters μ, σ^2. Then,

$$\Phi\left(\frac{17.8 - \overline{X}}{S}\right) - \Phi\left(\frac{14.5 - \overline{X}}{S}\right)$$

is a biased estimator of $P(14.5 < X < 17.8)$, where $X \sim N(\mu, \sigma^2)$, and

$$x_\alpha = \overline{X} + S z_\alpha$$

is a biased estimator of the $(1 - \alpha)$100th percentile of X.

In this section we will describe a criterion, based on the concept of *mean square error*, for comparing the performance, or quality, of estimators that are not necessarily unbiased. As a first step, we will discuss a criterion for choosing between two unbiased estimators. This criterion is based on the *standard error*, a term synonymous with standard deviation.

Definition 7.2.2. *The* **standard error** *of an estimator $\widehat{\theta}$ is its standard deviation $\sigma_{\widehat{\theta}} = \sqrt{\sigma^2_{\widehat{\theta}}}$, and an estimator/estimate of the standard error, $\widehat{\sigma}_{\widehat{\theta}} = \sqrt{\widehat{\sigma}^2_{\widehat{\theta}}}$, is called the* **estimated standard error**.

Example 7.2.7. a) The standard error, and the estimated standard error, of the estimator $\widehat{p}$ discussed in Example 7.2.1 are

$$\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}, \quad \text{and} \quad \widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

With the given information that 12 of 20 cars sustain no visible damage, we have $\widehat{p} = 12/20 = 0.6$, so that the estimated standard error is

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{0.6 \times 0.4}{20}} = 0.11.$$

b) The standard error, and the estimated standard error of $\overline{X}$ are

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}, \quad \text{and} \quad \widehat{\sigma}_{\overline{X}} = \frac{S}{\sqrt{n}},$$

where S is the sample standard deviation. If the sample standard deviation of the $n = 36$ robot reaction times mentioned in Example 7.2.2 is $S = 1.3$, the estimated standard error of $\overline{X}$ in that example is

$$\widehat{\sigma}_{\overline{X}} = \frac{1.3}{\sqrt{36}} = 0.22.$$

A comparison of unbiased estimators is possible on the basis of their standard errors: Among two unbiased estimators of a parameter θ, choose the one with the smaller standard error. This is the **standard error selection criterion for unbiased estimators**. The rationale behind this criterion is that a smaller standard error implies that the distribution is more concentrated about the true value of θ, and thus is preferable. For example, in Figure 7.1 estimator $\widehat{\theta}_1$ is preferable to $\widehat{\theta}_2$, though both are unbiased.

The standard error selection criterion for unbiased estimators will be used in Section **??** to determine whether or not stratified random sampling yields a better estimator of the population mean than simple random sampling.

Depending on the assumed parametric model, and the model parameter θ, it may be possible to find an unbiased estimator of θ that has the smallest variance among all unbiased estimators of θ (under the assumed parametric model). When such estimators exist they are called **minimum variance unbiased estimators** (MVUE). The following proposition gives some examples of MVUE.

Proposition 7.2.1. *a) If $X_1, \ldots, X_n$ is a sample from a normal distribution then $\overline{X}$ is MVUE for μ, and S^2 is MVUE for σ^2.*
b) If $X \sim Bin(n,p)$, then $\hat{p} = X/n$ is MVUE for p.

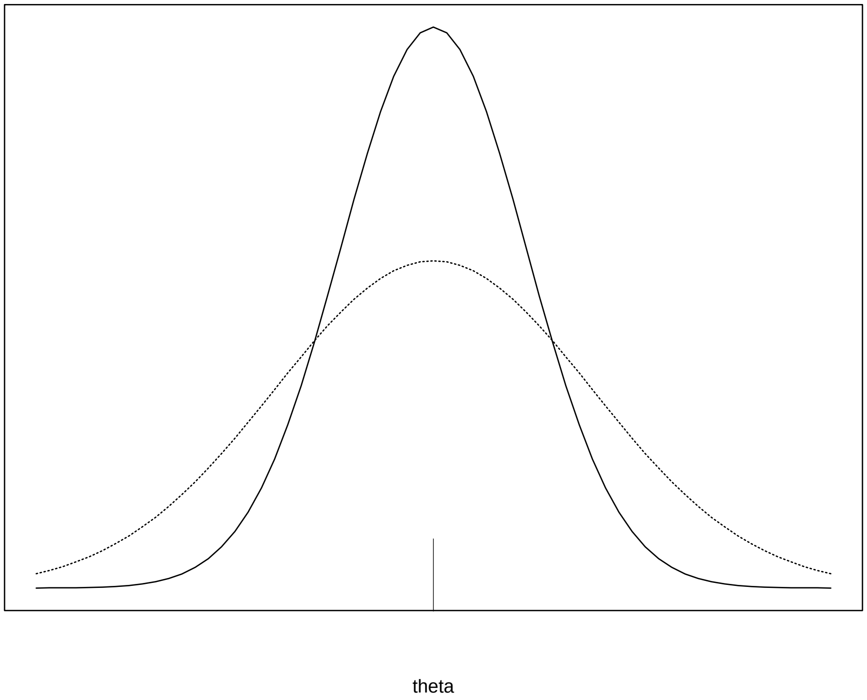


Figure 7.1: Solid Line: PDF of $\widehat{\theta}_1$. Dotted Line: PDF of $\widehat{\theta}_2$

REMARK 1: The fact that $\overline{X}$ is MVUE for μ when sampling under a normal distribution does not imply that $\overline{X}$ is always the preferred estimator for μ. Thus, if the population distribution is *logistic* or *Cauchy* then the sample median, $\tilde{X}$, is better than $\overline{X}$. Estimators such as the various trimmed means and the hybrid estimator discussed in Example 7.2.6 perform well over a wide range of underlying population distributions.

We now proceed with the definition of *mean square error* and the corresponding more general selection criterion.

Definition 7.2.3. *The* **mean square error** *(MSE) of an estimator $\hat{\theta}$ for the parameter θ is defined to be*

$$MSE\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2.$$

(More exactly, $MSE_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2$.) The **MSE selection criterion** *says that among two estimators, the one with smaller MSE is preferred.*

The mean square error of an unbiased estimator is the variance of that estimator. Thus, the MSE selection criterion reduces to the standard error selection criterion when the estimators are unbiased. The next proposition reiterates this, and shows how the MSE criterion incorporates both the standard error and the bias in order to compare estimators that are not necessarily unbiased.

Proposition 7.2.2. *If $\hat{\theta}$ is unbiased for θ then*

$$MSE\left(\hat{\theta}\right) = \sigma^2_{\hat{\theta}}.$$

In general,

$$MSE\left(\hat{\theta}\right) = \sigma^2_{\hat{\theta}} + \left[bias\left(\hat{\theta}\right)\right]^2.$$

In the section on *maximum likelihood estimation*, the MSE criterion suggests that, with large enough sample sizes, the biased estimators of probabilities and percentiles, which are given in parts c) and d) of Example 7.2.6, are to be preferred to the corresponding nonparametric estimators, i.e. sample proportions and sample percentiles, provided that the assumption of exponential distribution made in part c) and of normal distribution in part d) hold true.

7.2.4 Exercises

1. The financial manager of a large department store chain selected a random sample of 200 of its credit card customers and found that 136 had incurred an interest charge during the previous year because of an unpaid balance.

   (a) What do you think is the parameter of interest in this study?

   (b) Calculate a point estimate for the parameter of interest listed in part (a).

   (c) Give the standard error of the estimator in (b).

   (d) Is the estimator unbiased?

2. Suppose a person is selected at random and is given a hand-writing-recognition-equipped Apple Newton computer to try out. The Apple Newton is fresh from the manufacturer, and the person is asked to write "I think Apple Newton's are neat". But often the Apple Newton mis-interprets the words and types back something quite different (we needn't go into all the possibilities here).

   A computer magazine conducts a study to estimate the probability, p, that, for a randomly selected owner of a hand-writing-recognition-equipped Apple Newton computer, the phrase "I think Apple Newton's are neat" will be mis-interpreted. Suppose that they will observe X mistakes in n tries (n owners).

(a) What is your estimator for p? What properties does this estimator have? Are there any general principles that support the use of this estimator?

(b) What is your estimate $\hat{p}$ if the magazine finds 33 mistakes in 50 tries?

(c) Now suppose that the magazine's pollster recommends that they also estimate the standard error of the estimator $\hat{p}$. For the given data, what is your estimate of the standard error?

3. A food processing company is considering the marketing of a new product. Among 40 randomly chosen consumers 9 said that they would purchase the new product and give it a try. Estimate the true proportion of potential buyers, and state its standard error. Is the estimator unbiased?

4. To estimate the parameter θ we must choose between 4 different estimators:

$\hat{\theta}_1$ has $E(\hat{\theta}_1) = \theta$ and $Var(\hat{\theta}_1) = \frac{\sigma^2}{\sqrt{n}}$
$\hat{\theta}_2$ has $E(\hat{\theta}_2) = \left(\frac{n}{n-1}\right)\theta$ and $Var(\hat{\theta}_2) = \frac{\sigma^2}{n}$
$\hat{\theta}_3$ has $E(\hat{\theta}_3) = \left(\frac{n}{n-1}\right)\theta$ and $Var(\hat{\theta}_3) = \frac{\sigma^2}{\sqrt{n}}$
$\hat{\theta}_4$ has $E(\hat{\theta}_4) = \theta$ and $Var(\hat{\theta}_4) = \frac{\sigma^2}{n}$

(a) Which estimators are unbiased?

(b) Among the unbiased estimators, which one would you choose to estimate θ? Why?

5. To estimate the proportion p_1 of males that support Amendment Q to the Constitution, we take a random sample of m males and record the number X in support. To estimate the corresponding proportion p_2 for females, we take a random sample of n females and record the number Y in support.

(a) Write the estimators $\widehat{p_1}$, $\widehat{p_2}$ of p_1, p_2.

(b) Show that $\widehat{p_1} - \widehat{p_2}$ is an unbiased estimator of $p_1 - p_2$.

(c) What is the standard error of the estimator in b)?

(d) Suppose that $m = 100$, $n = 200$ and the observed values of X and Y are 60 and 150, respectively. Use the estimator in part b) to obtain an estimate of $p_1 - p_2$.

(e) Use the result of part c) and the data given in part d) to estimate the standard error of the estimator.

6. Let $X_1, \ldots, X_{10}$ be a random sample from a population with mean μ and variance σ^2. In addition, let $Y_1, \ldots, Y_{10}$ be a random sample from another population with mean also equal to μ and variance $4\sigma^2$.

(a) Show that for any δ, $0 \leq \delta \leq 1$, $\widehat{\mu} = \delta\overline{X} + (1-\delta)\overline{Y}$ is unbiased for μ.

(b) Obtain an expression for the variance of $\widehat{\mu}$.

(c) You are asked to select one of the two estimators: $\widehat{\mu}$ given above with $\delta = 0.5$ or $\overline{X}$. Which would you select and why?

7. Use Example 7.2.5 and properties of expectation to show that the estimator of the regression line, $\widehat{\mu}_{Y|X=x} = \widehat{\alpha}_1 + \beta_1 x$, given in (6.2.3), is unbiased.

7.3 Methods for Fitting Models to Data

In this section we will present three methods for fitting parametric models to data. As mentioned in Section 7.1, the parametric approach to extrapolating sample information to the population is to assume that the population distribution is a member of a particular parametric family of distribution models, and then fit the assumed family to the data. Fitting a model to data, amounts to estimating the model parameters. The estimated model parameters identify the **fitted model**, i.e. the member of the parametric family that best fits the data, as the one that corresponds to the estimated parameters. The fitting methods we will present correspond to different criteria for determining the best fitting model.

7.3.1 Overview and the Method of Moments

Under the assumption that the population distribution belongs in a parametric family (such as the exponential, or the uniform, or the normal, or the lognormal etc), then population parameters such as the mean and variance are expressed in terms of the model parameters. Conversely, the model parameters can be expressed in terms of population parameters. For example, if we assume that the population distribution belongs in the exponential family (so it has pdf of the form $f_\lambda(x) = \lambda \exp(-\lambda x)$, the model parameter λ is expressed in terms of the population mean as $\lambda = \mu^{-1}$; if we assume that the population distribution belongs in the uniform family (so its pdf is constant an interval $[\alpha, \beta]$), then, solving the equations $\mu = (\alpha + \beta)/2$, $\sigma^2 = (\beta - \alpha)^2/12$ (which express the population parameters μ, σ^2 in terms of the model parameters) for α, β, we can express the model parameters in terms of μ, σ^2 as $\alpha = \mu - \sqrt{3}\sigma$, $\beta = \mu + \sqrt{3}\sigma$. The idea behind the **method of moments** is that, since the population parameters can be estimated nonparametrically, then, expressing the model parameters in terms of the population parameters leads to estimators of the model parameters. This is approach to estimating model parameters is illustrated in the examples that follow.

Remark 7.3.1. *The method of moments derives its name from the fact that the expected value of the kth power of a random variable is called its kth* **moment***, or the kth moment of the underlying population. Thus, moments are also population parameters; the kth moment is denoted by μ_k. According to this terminology, the population mean is the first moment, and is also denoted by μ_1, while the population variance can be expressed in terms of the fist two moments as $\sigma^2 = \mu_2 - \mu_1^2$. If $X_1, \ldots, X_n$ is a sample from a population with kth moment μ_k (so that $E(X_i^k) = \mu_k$, for all $i = 1, \ldots, n$), then the nonparametric estimator of μ_k is the kth* **sample moment***:*

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^n.$$

Under a parametric assumption, the population moments can be expressed in terms of the model parameters. If the parametric family does not involve more than two parameters, then these parameters can be expressed in terms of the first two moments (or, equivalently, in terms of the population mean and variance), and be estimated by substituting nonparametric estimators of the moments in that expression. If the parametric family involves more than two parameters, then more than two moments are used in the method of moments estimation of the model parameters. In this book we will not consider distribution models with more than two model parameters, so in our applications of the method of moments, we will use only the sample mean and sample variance.

Example 7.3.1. Consider the low impact car crash experiment described in Example 7.2.1. Thus, 20 cars are crashed at 5 mph, and X denote the number among them that sustain no visible damage. Under the reasonable assumption that crash results of different cars are independent, $X \sim Bin(20, p)$. Here the model parameter p coincides with the population proportion p of cars that sustain no visible damage when crashed at 5 mph. Thus the method of moments estimator of p is $\hat{p} = X/n$, which is the same as the nonparametric estimator of p. The best fitting model is the binomial distribution that corresponds to the estimated value of p. For example, if $X = 12$ of the 20 cars in the experiment sustain no visible damage, the estimate of p is $\hat{p} = 12/20$, and the best fitting model is the $\text{Bin}(20, 0.6)$.

Example 7.3.2. Consider the robot reaction time experiment described in Example 7.2.2. Thus, $X_1, \ldots, X_{36}$ denote 36 response times that are to be measured. Though it is not clear what the population distribution (i.e. the distribution of each X_i) is, it might be assumed (at least tentatively) that this distribution is a member of the normal family of distributions. The model parameters that identify a normal distribution are its mean and variance. Thus, under the assumption that the population distribution belongs in the normal family, the model parameters coincide with the population parameters μ and σ^2. It follows, that the method of moments estimators of the model parameters are simply $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma}^2 = S^2$, so that the best fitting model is the normal distribution with mean and variance equal to the sample mean and sample variance, respectively. For example, if the sample mean of the 36 response times is $\overline{X} = 9.3$, and the sample variance is $S^2 = 4.9$, the best fitting model is the $N(9.3, 4.9)$.

Example 7.3.3. The lifetime of electric components is often the variable of interest in reliability studies. Let $T_1, \ldots, T_{25}$ denote the life times, in hours, of a random sample of 25 components. Here it is not clear what the population distribution might be, but it might be assumed (at least tentatively) that it is a member of the exponential family of models. Thus, each T_i has pdf $f_\lambda(t) = \lambda \exp(-\lambda t)$, for $t \geq 0$ ($f_\lambda(t) = 0$, for $t < 0$), for some $\lambda > 0$. Here the single model parameter λ identifies the exponential distribution. Since, under the assumption that the population distribution belongs in the exponential family, the model parameter is given in terms of the population mean by $\lambda = \mu^{-1}$, it follows that the method of moments estimator of λ is $\hat{\lambda} = 1/\overline{X}$. Thus, the best fitting exponential model is the exponential distribution with model parameter equal to $\hat{\lambda}$. For example, if the average of the 25 life times is 113.5 hours, the best fitting model is the exponential distribution with $\lambda = 113.5^{-1}$.

Example 7.3.4. Suppose, as in the previous example, that interest lies in the distribution of the life time of some type of electric component, and let $T_1, \ldots, T_{25}$ denote the life times, in hours, of a random sample of 25 such components. If the assumption that the population distribution belongs in the exponential family does not appear credible, it might be assumed that it is a member of the gamma family of distribution models. This is a richer family of models and it includes the models of the exponential distribution. The gamma distribution is identified by two parameters, α and β, and its pdf is of the form

$$f_{\alpha,\beta}(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \ x > 0.$$

The model parameters α and β are related to the model mean and variance through

$$\alpha = \frac{\mu^2}{\sigma^2}, \quad \beta = \frac{\sigma^2}{\mu}.$$

Thus, the method of moments estimators of the model parameters α and β are

$$\widehat{\alpha} = \frac{\overline{X}^2}{S^2},\ \widehat{\beta} = \frac{S^2}{\overline{X}},$$

where $\overline{X}$, S^2 denote the sample mean and sample variance of the 25 life times. For example, if the average of the 25 life times is 113.5 hours, and their sample variance is 1205.55 hours^2, the best fitting model is the gamma distribution with $\alpha = 113.5^2/1205.55 = 10.69$, and $\beta = 1205.55/113.5 = 10.62$.

We finish this subsection with a discussion that applies to the parametric approach in general, not just to the method of moments.

Having a fitted distribution model means that the entire distribution has been estimated. In particular, the fitted model provides immediate and direct estimation of any percentile and probability that might be of interest.

Example 7.3.5. For example, the fitted model of Example 7.3.2 allows direct and immediate estimation of the probability that the reaction time of the robot will exceed 50 nanoseconds, and of the 95th percentile of the robot reaction time, by the corresponding probability and percentile of the fitted normal distribution:

$$\widehat{P}(X > 50) = 1 - \Phi\left(\frac{50 - \overline{X}}{S}\right), \quad \text{and} \quad \hat{x}_{0.05} = \overline{X} + z_{0.05}S, \qquad (7.3.1)$$

where a hat place above a quantity denotes an estimator of the quantity, $\overline{X}$ is the sample mean and S is the sample standard deviation obtained from the 36 measured response times.

Of course, the nonparametric estimates of proportions and percentiles that were discussed in Section **??**, can still be used. For example, the probability $P(X > 50)$ in Example 7.3.5, can be estimated as the proportion of the 36 robot reaction times that exceed 50 nanoseconds, and $\hat{x}_{0.05}$ can be estimated by the 95th sample percentile. If the assumption of normality is correct, i.e. if the population distribution of the robot reaction time is normal, then the estimates in (7.3.1) are to be preferred to the nonparametric ones, according to criteria that will be discussed in this chapter.

Keep in mind that, if the parametric assumption is not correct, i.e. if the population distribution is not a member of the assumed parametric family of models, then the fitted model is an estimator of the **best-approximating** model, which is the member of the parametric family that best approximates the population distribution. Consequently, estimators of probabilities and percentiles based on the fitted model (as was done in Example 7.3.5) will not have desirable properties. The following example illustrates this point.

Example 7.3.6. To demonstrate the effect that an incorrect modeling assumption can have on the estimation of probabilities and percentiles, suppose, as in Example 7.3.4, that interest lies in the distribution of the life time of some type of electric component, and the life times of a random sample of 25 such components is observed. With the information given in that example, the best fitting gamma model is the one with $\alpha = 113.5^2/1205.55 = 10.69$, and $\beta = 1205.55/113.5 = 10.62$. Thus, using the assumption of a gamma distribution model, the probability that a randomly selected component will last more than 140 hours, and the 95th percentile of life times are estimated as those of the best fitting gamma model. Using a statistical software package we obtain the estimates

$$\widehat{P}(X > 140) = 0.21, \quad \text{and} \quad \hat{x}_{0.05} = 176.02.$$

However, if the assumption of an exponential model distribution is made, as in Example 7.3.3, the best fitting exponential model gives estimates

$$\widehat{P}(X > 140) = 0.29, \quad \text{and} \quad \hat{x}_{0.05} = 340.02.$$

Fortunately, there are diagnostic tests that can help decide whether a parametric assumption is not correct, or which of two parametric families provides a better fit to the data. A way of gaining confidence that the population distribution belongs in the assumed parametric family (ideal case), or at least it is well approximated by the best approximating model, is the probability plot, which we saw in Chapter 6.

7.3.2 *The Method of Maximum Likelihood (ML)

The method of ML estimates θ by addressing the question "what value of the parameter is most likely to have generated the data?"

The answer to this question, which is the ML estimator (MLE), is obtained by maximizing (with respect to the parameter) the so-called **likelihood function** which is simply the joint p.d.f. (or p.m.f.) of $X_1, \ldots, X_n$ evaluated at the sample points:

$$\text{lik}(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Typically, it is more convenient to maximize the logarithm of the likelihood function, which is called the **log-likelihood function**. Since the logarithm is a monotone function, this is equivalent to maximizing the likelihood function.

Example 7.3.7. A Bernoulli experiment with outcomes 0 or 1 is repeated independently 20 times. If we observe 5 1s, find the MLE of the probability of 1, p.

Solution. Here the observed random variable X has a binomial$(n = 20, p)$ distribution. Thus,

$$P(X = 5) = \binom{20}{5} p^5 (1-p)^{15}.$$

The value of the parameter which is most likely to have generated the data $X = 5$ is the one that maximizes this probability, which, in this case, is the likelihood function. The log-likelihood is

$$\ln P(X = 5) = \ln \binom{20}{5} + 5\ln(p) + 15\ln(1-p).$$

Setting the first derivative of it to zero yields the MLE $\widehat{p} = \dfrac{5}{20}$. In general, the MLE of the binomial probability p is $\widehat{p} = \frac{X}{n}$.

Example 7.3.8. Let $X_1 = x_1, \ldots, X_n = x_n$ be a sample from a population having the exponential distribution, i.e. $f(x|\lambda) = \lambda \exp(\lambda)$. Find the MLE of λ.

Solution, The likelihood function here is

$$\text{lik}(\lambda) = \lambda e^{-\lambda x_1} \ldots \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum x_i},$$

and the first derivative of the log-likelihood function is

$$\frac{\partial}{\partial \lambda}\left[n\ln(\lambda) - \lambda \sum X_i \right] = \frac{n}{\lambda} - \sum X_i.$$

Setting this to zero yields $\widehat{\lambda} = \dfrac{n}{\sum X_i} = \dfrac{1}{\overline{X}}$ as the MLE of λ.

Example 7.3.9. The lifetime of a certain component is assumed to have the $\exp(\lambda)$ distribution. A n sample of n components is tested. Due to time constraints, the test is terminated at time L. So instead of observing the life times $X_1, \ldots, X_n$ we observe $Y_1, \ldots, Y_n$ where

$$Y_i = \begin{cases} X_i \text{ if } X_i \le L \\ L \text{ if } X_i > L \end{cases}$$

We want to estimate λ.

Solution. For simplicity, set $Y_1 = X_1, \ldots,\ Y_k = X_k,\ Y_{k+1} = L, \ldots, Y_n = L$. The likelihood function is

$$\lambda e^{-\lambda Y_1} \ldots \lambda e^{-\lambda Y_k} \cdot e^{-\lambda Y_{k+1}} \ldots e^{-\lambda Y_n}$$

and the log-likelihood is

$$\begin{aligned} & \left(\ln(\lambda) - \lambda Y_1\right) + \ldots \\ + & \left(\ln(\lambda) - \lambda Y_k\right) - \lambda L - \ldots - \lambda L \\ = & \; k\ln(\lambda) - \lambda \sum_{i=1}^{k} Y_i - (n-k)\lambda L. \end{aligned}$$

Setting the derivative w.r.t. λ to zero and solving gives the MLE

$$\begin{aligned} \hat{\lambda} &= \frac{k}{\sum_{i=1}^{k} Y_i + (n-k)L} \\ &= \frac{k}{\sum_{i=1}^{n} Y_i} = \frac{k}{n\overline{Y}} \end{aligned}$$

Example 7.3.10. Let $X_1 = x_1, \ldots, X_n = x_n$ be a sample from a population having the uniform distribution on $(0, \theta)$. Find the MLE of θ.

Solution. Here $f(x|\theta) = \dfrac{1}{\theta}$, if $0 < x < \theta$, and 0 otherwise. Thus the likelihood function is $\dfrac{1}{\theta^n}$ provided, $0 < X_i < \theta$ for all i, and is 0 otherwise. This is maximized by taking θ as small as possible. However if θ smaller than $\max(X_1, \ldots, X_n)$ then likelihood function is zero. Thus the MLE is $\hat{\theta} = \max(X_1, \ldots, X_n)$.

Theorem 7.3.1. *(Optimality of MLEs) Under smoothness conditions on $f(x|\theta)$, when n is large, the MLE $\hat{\theta}$ has sampling distribution which is approximately normal with mean value equal to (or approximately equal to) the true value of θ, and variance nearly as small as that of any other estimator. Thus, the MLE $\hat{\theta}$ is approximately a MVUE of θ.*

Remark 7.3.2. Among the conditions needed for the validity of Theorem 7.3.1 is that the set of x-values for which $f(x|\theta) > 0$ should not depend on θ. Thus, the sampling distribution of the MLE $\widehat{\theta} = \max(X_1, \ldots, X_n)$ of Example 7.3.10 is not approximately normal even for large n. However, application of the MSE criterion yields that the biased estimator $\max(X_1, \ldots, X_n)$ should be preferred over the unbiased estimator $2\overline{X}$ if n is sufficiently large. Moreover, in this case, it is not difficult to remove the bias of $\widehat{\theta} = \max(X_1, \ldots, X_n)$.

Theorem 7.3.2. *(Invariance of MLEs) If $\widehat{\theta}$ is the MLE of θ and we are interested in estimating a function, $\vartheta = g(\theta)$, of θ then*

$$\widehat{\vartheta} = g(\widehat{\theta})$$

is MLE of ϑ. Thus, $\widehat{\vartheta}$ has the stated optimality of MLEs stated in Theorem 7.3.1.

According to Theorem 7.3.2, the estimators given in Example 7.2.6c),d) are optimal. The following examples revisit some of them.

Example 7.3.11. Consider the setting of Example 7.3.8, but suppose we are interested in the mean lifetime. For the exponential distribution $\mu = \frac{1}{\lambda}$. (So here $\theta = \lambda$, $\vartheta = \mu$ and $\vartheta = g(\theta) = \frac{1}{\theta}$.) Thus $\widehat{\mu} = 1/\widehat{\lambda}$ is the MLE of μ.

Example 7.3.12. Let $X_1, \ldots, X_n$ be a sample from $N(\mu, \sigma^2)$. Estimate:
a) $P(X \leq 400)$, and b) $x_{.1}$.

Solution. a) $\vartheta = P(X \leq 400) = \Phi\left(\frac{400 - \mu}{\sigma}\right) = g(\mu, \sigma^2)$. Thus

$$\widehat{\vartheta} = g(\widehat{\mu}, S^2) = \Phi\left(\frac{400 - \overline{X}}{S}\right).$$

b) $\widehat{\vartheta} = \widehat{x}_{.1} = \widehat{\mu} + \widehat{\sigma} z_{.1} = g(\widehat{\mu}, S^2) = \overline{X} + S z_{.1}$

Remark 7.3.3. As remarked also in the examples of Section 7.3.1, Example 7.3.12 shows that the estimator we choose depends on what assumptions we are willing to make. If we do not assume normality (or any other distribution)

a) $P(X \leq 400)$ would be estimated by $\widehat{p}$=the proportion of X_is that are ≤ 400,

b) $x_{.1}$ would be estimated by the sample 90th percentile.

If the normality assumption is correct, the MLEs of Example 7.3.12 are to be preferred by Theorem 1.

7.3.3 The Method of Least Squares

In this subsection we will present the method of *least squares*, which is the most common method for fitting regression models. We will, describe this method for the case of fitting the simple linear regression model, namely the model which assumes that the regression function of Y on X is

$$\mu_{Y|X}(x) = E(Y|X = x) = \alpha_1 + \beta_1 x \tag{7.3.2}$$

(see relation (4.8.2)). The estimators we will obtain, called the **least squares estimators**, are the same as the empirical estimators that were derived in Section 6.2.3.

Estimation of the Intercept and the Slope

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote a simple random sample from a bivariate population of covariate and response variables (X, Y), and assume that (7.3.2) holds. To explain the

LS method consider the problem of deciding which of two lines fit the data better. For example, which of the two lines shown in Figure 7.2 fits the data better?

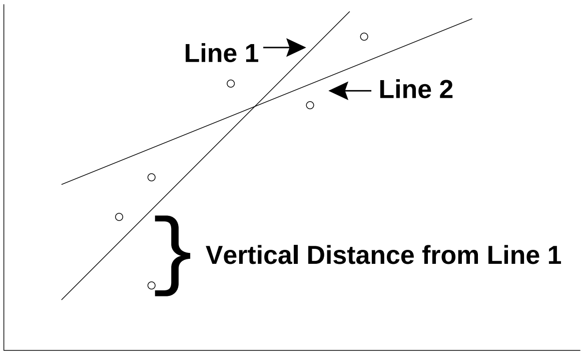


Figure 7.2: Two Lines Through a Data Set. Illustration of Vertical Distance

To answer the question of which of two lines fit the data better, one must first adopt a principle, on the basis of which, to judge the quality of a fit. The principle we will use is the principle of *least squares*. According to this principle, the quality of the fit of a line to data $(X_1, Y_1), \ldots, (X_n, Y_n)$ is judged by the sum of the squared vertical distances of each point (X_i, Y_i) from the line. The vertical distance of a point from a line is illustrated in Figure 7.2. The line for which this sum of squared vertical distances is smaller, is said to provide a better fit to the data.

The least squares estimates of the *intercept*, α_1, and the *slope* β_1, are the intercept and the slope, respectively, of the **best fitting line**, i.e. of the line with the smallest sum of vertical square distances. The best fitting line is also called the **estimated regression line**.

Since the vertical distance of the point (X_i, Y_i) from a line $a + bx$ is $Y_i - (a + bX_i)$, the method of least squares finds the values $\widehat{\alpha}_1$, $\widehat{\beta}_1$ which minimize

$$\sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

with respect to a, b. This minimization problem has a simple closed-form solution:

$$\widehat{\beta}_1 = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2}, \quad \widehat{\alpha}_1 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

Thus, the **estimated regression line** is

$$\widehat{\mu}_{Y|X}(x) = \widehat{\alpha}_1 + \widehat{\beta}_1 x.$$

Remark 7.3.4. Note that $\widehat{\beta}_1$ can also be written as

$$\widehat{\beta}_1 = \frac{\widehat{\sigma}_{X,Y}}{S_X^2} = \rho_{X,Y}\frac{S_X}{S_Y},$$

where $\widehat{\sigma}_{X,Y}$ is the sample covariance, $\rho_{X,Y}$ is Pearson's correlation coefficient, and S_X^2, S_Y^2 are the sample variances of the X_is and Y_is.

Example 7.3.13. Suppose that $n = 10$ data points on X=stress applied and Y=time to failure yield summary statistics $\sum X_i = 200$, $\sum X_i^2 = 5412.5$, $\sum Y_i = 484$, $\sum X_iY_i = 8407.5$. Thus the best fitting line has slope and intercept of

$$\begin{aligned}\widehat{\beta}_1 &= \frac{10(8407.5) - (200)(484)}{10(5412.5) - (200)^2} = -.900885,\\ \widehat{\alpha}_1 &= \frac{1}{10}(484) - (-.900885)\frac{200}{10} = 66.4177,\end{aligned}$$

respectively.

Estimation of the Intrinsic Variability

The simple linear regression model can also be written as

$$Y_i = \alpha_1 + \beta_1 X_i + \varepsilon_i, \tag{7.3.3}$$

where the ε_i were called the intrinsic error variables in Subsection 4.8.1, and are assumed to be iid. Note that (7.3.3) is similar to (4.8.4) except for the different definition of the intercept, and the assumption that the intrinsic error variables have a normal distribution.

The idea for estimating the conditional variance, σ^2, of the response variable Y given $X = x$, which is the same as the variance of each intrinsic error variable ε_i, is that, if the true values of α_1 and β_1 were know, then the intrinsic error variables can be obtained as

$$\varepsilon_i = Y_i - \alpha_1 - \beta_1 X_i,$$

and σ^2 would be estimated by the sample variance of the ε_i, $i = 1, \ldots, n$. Of course, α_1, β_1 are not known. Thus we use the **residuals** or **estimated errors**

$$\widehat{\varepsilon}_1 = Y_1 - \widehat{\alpha}_1 - \widehat{\beta}_1 X_1 = Y_1 - \widehat{Y}_1, \ldots, \widehat{\varepsilon}_n = Y_n - \widehat{\alpha}_1 - \widehat{\beta}_1 X_n = Y_1 - \widehat{Y}_n, \tag{7.3.4}$$

where $\widehat{Y}_i = \widehat{\alpha}_1 - \widehat{\beta}_1 X_i$ are called the **fitted** or **predicted** values, and estimate σ^2 by

$$\widehat{\sigma}^2 = S^2 = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2. \tag{7.3.5}$$

The division by $n-2$ is justified by the fact that two parameters are estimated for the computation of S^2. The quantity $\sum_{i=1}^n \widehat{\varepsilon}_i^2$ in (7.3.5), is called the **error sum of squares**, and is denoted by **SSE**. A computational formula it is

$$SSE = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} Y_i^2 - \widehat{\alpha}_1 \sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \sum_{i=1}^{n} X_i Y_i. \tag{7.3.6}$$

The fitted values and the residuals are illustrated in Figure 7.3.

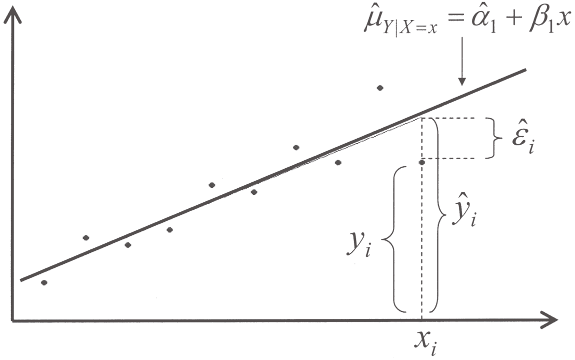


Figure 7.3: Illustration of Fitted Values and Residuals

Remark 7.3.5. *The intrinsic variance σ^2 measures the variability in Y that cannot be attributed to (or explained by) the regression model, and thus it is called the variability of Y which is* **unexplained** *by the model. For example if there was no intrinsic error (so $\sigma^2 = 0$), the model is $Y_i = \alpha_1 + \beta_1 X_i$, so that the data will fall exactly on a straight line, as in Figure 7.4.*

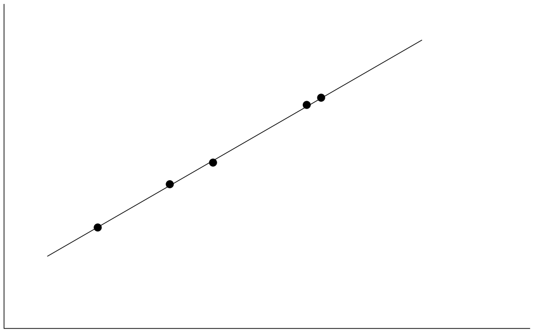


Figure 7.4: No Intrinsic Variability. The Variability in Y is Explained by that of X

Here there is no unexplained variation in Y (or no intrinsic variation), and the variability in the Y values is only due to the regression model, i.e. entirely due to its dependence on

X. On the other extreme, if $\beta_1 = 0$ (i.e. if the X-variable has no predictive value for the Y-variable), as in the model $Y_i = \alpha_1 + \varepsilon_i$, then all the variability in Y is due to the error variable and is not explained by the regression model.

Example 7.3.14. Consider the following data on Y=propagation of an ultrasonic stress wave through a substance, and X=tensile strength of substance.

X	:	12	30	36	40	45	57	62	67	71	78	93	94	100	105
Y	:	3.3	3.2	3.4	3.0	2.8	2.9	2.7	2.6	2.5	2.6	2.2	2.0	2.3	2.1

Fit the simple linear regression model.

Solution. Here $n = 14$. $\sum_i X_i = 890, \sum_i X_i^2 = 67,182, \sum_i Y_i = 37.6, \sum_i Y_i^2 = 103.54$ and $\sum_i X_iY_i = 2234.30$. From this we get $\widehat{\beta}_1 = -0.0147209, \widehat{\beta}_0 = 3.6209072$, and

$$\widehat{\sigma}^2 = S^2 = \frac{103.54 - \widehat{\beta}_0(37.6) - \widehat{\beta}_1(2234.30)}{12} = \frac{0.2624532}{12} = 0.02187.$$

(In this example it is important to allow several decimal places in the calculations. For example, when $\widehat{\beta}_0, \widehat{\beta}_1$ are rounded to three decimal places, the numerator of $\widehat{\sigma}^2$ becomes 0.905). As illustration, we compute the fitted value and residual at the first X-value $X_1 = 12$

$$\widehat{Y}_1 = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 = \widehat{\beta}_0 + \widehat{\beta}_1(12) = 3.444, \quad \widehat{\varepsilon}_1 = Y_1 - \widehat{Y}_1 = -0.144.$$

Example 7.3.15. Consider the information given in Example 7.3.14. Let Y_1 denote an observation made at $X_1 = 30$, and Y_2 denote an observation made at $X_2 = 35$. Estimate the $E(Y_1 - Y_2)$, the expected difference between Y_1 and Y_2.

Solution. According to the regression model,

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1, \;\; Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2, \text{ so that}$$

$$Y_1 - Y_2 = \beta_1(X_1 - X_2) + \varepsilon_1 - \varepsilon_2 = -5\beta_1 + \varepsilon_1 - \varepsilon_2.$$

It follows that the expected value of the difference $Y_1 - Y_2$ is

$$E(Y_1 - Y_2) = -5\beta_1 + E(\varepsilon_1 - \varepsilon_2) = -5\beta_1$$

Thus, $E(Y_1 - Y_2)$ can be estimated as $\widehat{E}(Y_1 - Y_2) = -5\widehat{\beta}_1 = 5(0.0147209)$.

7.3.4 Exercises

1. Use Minitab to generate a simple random sample of 50 observations from a normal population with $\mu = 11.7$ and $\sigma = 4.6$.

   (a) Give the true (population) values of $x_{0.85}$, $x_{0.75}$, $x_{0.45}$, $x_{0.05}$, and of $P(12 < X \leq 16)$.

   (b) Give the empirical (nonparametric) estimates of the above population quantities. (Note: If the 50 observations are in column C1, the sequence of commands Calc> Calculator, Store results in "C2", Expression: "12<C1≤16" > OK, and then Calc> Column statistics, Click "Sum", Input variable C2 > OK, gives the count of observations that greater than 12 and less than or equal to 16.)

   (c) Assume that the underlying population is normal (which is correct) and give the parametric estimates of the above population quantities. Compare how close the two types of estimates are.

2. Repeat Exercise 1 using 5,000 observations. Are the two types of estimators closer than when 50 observations were used?

3. Use Minitab to generate a simple random sample of 5,000 observations from a gamma population with shape parameter 2 and scale parameter 1 (Gamma(2,1)).

   (a) Use Minitab to find $P(0.8 < X \leq 1.2)$, if $X \sim$ Gamma(2,1).

   (b) Compare the empirical (nonparametric) and parametric (assuming the correct, i.e. gamma model) estimators of $P(0.8 < X \leq 1.2)$.

   (c) Use Minitab to superimpose the fit of the Weibull model on the histogram of the data. On the basis of the fitted Weibull pdf and the histogram, do you think that the parametric estimator of $P(0.8 < X \leq 1.2)$, when the Weibull model is (wrongly) assumed will overestimate or underestimate the true value of $P(0.8 < X \leq 1.2)$?

   (d) Repeat the above part using the lognormal model instead of Weibull.

4. For a sample of 6 jugs of 2 % lowfat milk produced by "Happy Cow Dairy" the fat content X_i has been determined as:

   2.08 2.10 1.81 1.98 1.91 2.06

   (a) Making no assumptions on the distribution of the fat content, estimate the proportion of milk jugs having a fat content of 2.05 % or more.

   (b) Making no assumptions on the distribution of the fat content, estimate the mean fat content of "Happy Cow Dairy" 2 % lowfat milk.

(c) Making no assumptions on the distribution of the fat content,

(i) your estimator in part 4a is (circle all correct statements):

`Unbiased   Maximum likelihood estimator   Moments estimator`

(ii) your estimator in part 4b is (circle all correct statements):

`Unbiased   Maximum likelihood estimator   Moments estimator`

(d) Assuming normality, the maximum likelihood estimators of the mean and variance are given by $\overline{X}$ and $[(n-1)/n]s^2$, respectively. Calculate the maximum likelihood estimator of the proportion of milk jugs having a fat content of 2.05 % or more. [Hints: a) $P(X > x) = 1 - \Phi((x-\mu)/\sigma)$, b) For the given data, $\sum(x_i - \overline{x})^2 = 0.064$.]

5. Fifty newly manufactured items are examined and the number of scratches per item are recorded. The resulting frequencies of the number of scratches is:

Number of scratches per item	0	1	2	3	4
Observed frequency	4	12	11	14	9

Let X be the number of scratches on a randomly selected item, and suppose that X has a Poisson distribution with parameter λ.

(a) Find the moments estimator of λ, and compute it using the above data. [Hint: The mean of a Poisson distribution with parameter λ is λ.]

(b) Is the moments estimator unbiased?

6. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f(x, \theta) = \theta x^{\theta-1}, 0 < x < 1$, where $\theta > 0$.

(a) It is given that $\mu = \theta/(1+\theta)$. Find the moments estimator of θ.

(b) Suppose 5 data points from this distribution are selected with $x_1 = 0.5, x_2 = 0.3, x_3 = 0.7, x_4 = 0.4$, and $x_5 = 0.8$. Evaluate the moments estimate of θ.

7. Let $X_1, \ldots, X_n$ denote the life times of n components selected randomly from an assembly line.

(a) We are interested in estimating the probability that a randomly chosen component has life time greater than 8 time units. Find an unbiased estimator of this probability.

(b) Give the standard error of the estimator in (a).

(c) It is believed that the life times have pdf $f(x) = \lambda e^{-\lambda(x-5)}$ for $x \geq 5$ and zero otherwise. Find the method of moments estimator of λ.

(d) Using the above model, the probability that a life time exceeds x is $exp(-\lambda(x-5))$. Give a model-based estimator of the probability that a life time will exceed 8.

(e) Give a model-based estimator of the 95-th percentile of this distribution.

8. The life times of 13 components selected randomly from an assembly line are

   5.60, 6.06, 6.77, 9.21, 5.41, 7.19, 6.31, 8.23, 6.16, 7.76, 5.59, 10.32, 9.01

   Given numerical answers to the questions of Exercise 7.

9. Plumbing suppliers typically ship packages of plumbing supplies containing many different combinations of pipes, sealants, drains, etc. Almost invariably there are one or more parts in the shipment that are not correct: the part may be defective, missing, not the one that was ordered, etc. In this question the random variable of interest is the proportion P of parts in a shipment, selected at random, that are not correct. A family of distributions for modeling a random variable P that is a proportion (thus $0 < P < 1$), has pdf

$$f_P(p) = \theta p^{\theta-1}, \quad \theta > 0.$$

   (a) Find the expected value of P.

   (b) Suppose that the proportion of not correct items in 5 shipments is 0.05, 0.31, 0.17, 0.23, and 0.08. What would be your estimate of θ, using the method of moments?

10. The life time of certain equipment is believed to follow the probability density function

$$f(x) = (1/\theta^2)xe^{-x/\theta}, \quad x > 0, \quad \theta > 0.$$

   (a) Estimate $E(X)$ using the data: 3, 1, 5, 4, 2, 3, 3, 3, 4, 3.

   (b) Find the method of moments estimator for θ, and evaluate it with the above data.

   (c) Is the moments estimator for θ unbiased?

11. A company manufacturing bike helmets wants to estimate the proportion p of helmets with a certain type of flaw. They decide to keep inspecting helmets until they find $r = 5$ flawed ones. Let X denote the number of helmets that were not flawed among those examined. Thus, if the 52-nd helmet examined is the 5-th flawed one, then $X = 47$.

   (a) What is the distribution of X?

   (b) Find the MLE of p.

   (c) Find the method of moments estimator of p.

   (d) If $X = 47$, give a numerical value to your estimators in a), b).

Chapter 8

Confidence and Prediction Intervals

8.1 Introduction to Confidence Intervals

By virtue of the Central Limit Theorem, if the sample size n is large enough, many estimators, $\hat{\theta}$, are approximately normally distributed. Moreover, such estimators are typically unbiased (or nearly unbiased), and their estimated standard error, $\widehat{\sigma}_{\hat{\theta}}$, typically provides a reliable estimate of $\sigma_{\hat{\theta}}$. Thus, if n is large enough, many estimators $\hat{\theta}$ satisfy

$$\hat{\theta} \stackrel{\cdot}{\sim} N\left(\theta, \widehat{\sigma}_{\hat{\theta}}^2\right), \tag{8.1.1}$$

where θ is the true value of the parameter. For example, this is the case for the nonparametric estimators discussed in Chapter 6, the moment estimators and many maximum likelihood estimators, as well as the least squares estimators which are described in Section 7.3. For such estimators, it is customary to report point estimates together with their estimated standard errors. The estimated standard error helps assess the size of the estimation error through the 68-95-99.7% rule of the normal distribution. For example, (8.1.1) implies that

$$\left|\hat{\theta} - \theta\right| \le 2\widehat{\sigma}_{\hat{\theta}}$$

holds approximately 95% of the time. Alternatively, this is can be written as

$$\hat{\theta} - 2\widehat{\sigma}_{\hat{\theta}} \le \theta \le \hat{\theta} + 2\widehat{\sigma}_{\hat{\theta}}$$

which gives an interval of plausible values for the true value of θ, with degree of plausibility approximately 95%. Such intervals are called **confidence intervals**. The abbreviation **CI** will be used for ”confidence interval”. Note that if $\hat{\theta}$ is unbiased and we believe that the normal approximation to its distribution is quite accurate, the 95% CI uses $z_{0.025} = 1.96$ instead of 2, i.e.

$$\hat{\theta} - 1.96\widehat{\sigma}_{\hat{\theta}} \leq \theta \leq \hat{\theta} + 1.96\widehat{\sigma}_{\hat{\theta}}.$$

The general technique for constructing $(1-\alpha)100\%$ confidence intervals for a parameter θ based on an approximately unbiased and normally distributed estimator, $\hat{\theta}$, consists of two steps:

a) Obtain an *error bound*, which holds with probability $1-\alpha$. This error bound is of the form

$$\left|\hat{\theta} - \theta\right| \leq z_{\alpha/2}\widehat{\sigma}_{\hat{\theta}}$$

b) Convert the error bound into an interval of plausible values for θ of the form

$$\hat{\theta} - z_{\alpha/2}\widehat{\sigma}_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\widehat{\sigma}_{\hat{\theta}},$$

or, in short-hand notation,

$$\hat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\hat{\theta}}.$$

The degree of plausibility, or **confidence level**, of the interval will be $(1-\alpha)100\%$.

In this chapter we will discuss the interpretation of CIs, present nonparametric confidence intervals for population means and proportions, as well as an alternative, normal based CI for the population mean. The issue of *precision* in estimation will be considered, and we will see how the sample size can be manipulated in order to increase the precision of the estimation of a population mean and proportion. Furthermore, we will discuss the related issue of constructing *prediction* intervals under the normality assumption, and we will present CIs for other parameters such as percentiles and the variance.

8.2 Confidence Interval Semantics

There are two ways to view any CI, depending on whether $\widehat{\theta}$ is viewed as an estimator (i.e. a random variable) or as an estimate. When $\widehat{\theta}$ is viewed as a random variable the CI is a random interval. Hence we can say that

$$\text{the true value of } \theta \text{ belongs in } \widehat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$$

holds with probability $1-\alpha$. But when data are obtained and used to compute the point estimate $\widehat{\theta}$, the CI is a fixed interval which either contains the true (and unknown to us) value of θ or it does not. For example, as we will see in the next subsection, the estimate $\widehat{p}=0.6$, based on $n=20$ Bernoulli trials, leads to the fixed 95% confidence interval of

$$(0.38, 0.82)$$

for p. This either contains the true value of p or it does not.

So how is a computed $(1-\alpha)100\%$ CI to be interpreted? To answer this question, think of the process of constructing a $(1-\alpha)100\%$ CI as performing a Bernoulli trial where the outcome is "success" if the true value of θ belongs in the CI. Thus, the probability of "success" is $1-\alpha$, but we are not able to observe the outcome of the trial. Not knowing the outcome, the only thing we can say is that our degree of confidence that the outcome was "success" is measured by $(1-\alpha)100\%$.

8.3 Nonparametric Confidence Intervals

In this section we will construct confidence intervals for a population mean, proportion, and median. One of the methods for constructing CIs for a median extends to constructing CIs for other percentiles. The procedures for constructing CIs for a mean rely on the normal approximation to the distribution of the sample mean. The other procedures can also be used with small sample sizes through software packages, but here will only describe procedures that use the normal approximation. Thus, the actual confidence level of these CIs will be approximately equal to the specified $(1-\alpha)100\%$. The larger the sample size, the better the approximation. Except for requiring a large sample size, no other requirement is needed for the validity of the present procedures, and thus are called nonparametric (NP).

8.3.1 Nonparametric CIs for Means

Let $X_1, \ldots, X_n$ denote a simple random sample from a population with mean value μ and variance σ^2. If the sample size is large ($n>30$) the Central Limit Theorem asserts that

$$\overline{X} \overset{\cdot}{\sim} N\left(\mu, \sigma^2/n\right). \tag{8.3.1}$$

Moreover, if n is large enough, the sample standard deviation, S, is a good estimator of σ. Thus,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{\cdot}{\sim} N(0,1). \tag{8.3.2}$$

Relation (8.3.2) implies that the bound on the error of estimation

$$\left|\overline{X}-\mu\right| \leq z_{\alpha/2}\frac{S}{\sqrt{n}} \tag{8.3.3}$$

holds with approximate probability $1-\alpha$, and this leads to the (approximate) $(1-\alpha)100\%$ CI

$$\boxed{\left(\overline{X}-z_{\alpha/2}\tfrac{S}{\sqrt{n}},\ \overline{X}+z_{\alpha/2}\tfrac{S}{\sqrt{n}}\right), \quad \text{or} \quad \overline{X}\pm z_{\alpha/2}\tfrac{S}{\sqrt{n}}, \quad \text{NP CI for } \mu} \tag{8.3.4}$$

for μ. (In the rare cases where the true value of σ is known, we can use it instead of its estimate S.) This is a nonparametric CI for the population mean, because it does not use any distributional assumptions (although it does assume, as the Central Limit Theorem does, that the population mean and variance exist, and are finite).

Example 8.3.1. In Examples 7.3.2 and 7.2.7, a sample of size $n = 36$ measured reaction times yielded a sample average of $\overline{X} = 5.4$, and a sample standard deviation of $S = 1.3$. Since $n = 36$ is large enough to apply the Central Limit Theorem, the available information yields an approximate 68% CI for the true value of μ (the mean response time of the robot in the conceptual population of all malfunctions of that type), of

$$5.4-\frac{1.3}{\sqrt{36}} \leq \mu \leq 5.4+\frac{1.3}{\sqrt{36}}, \quad \text{or} \quad 5.18 \leq \mu \leq 5.62.$$

Example 8.3.2. A random sample of $n = 56$ cotton samples gave average percent elongation of 8.17 and a sample standard deviation of $S = 1.42$. Calculate a 95% CI for μ, the true average percent elongation.

Solution. The sample size of $n = 56$ is large enough for the application of the Central Limit Theorem, which asserts that the distribution of the sample mean, $\overline{X}$, is well approximated by the normal distribution. Thus, the error bound (8.3.3) and the resulting nonparametric CI (8.3.4) for μ can be used. The information given yields the CI

$$\bar{X}\pm z_{\alpha/2}\frac{S}{\sqrt{n}} = 8.17 \pm 1.96\frac{1.42}{\sqrt{56}} = 8.17 \pm .37 = (7.8,\ 8.54).$$

8.3.2 Nonparametric CIs for Proportions

A special case of the NP CI (8.3.4) arises when $X_1, \ldots, X_n$ is a sample from a Bernoulli population. Thus, each X_i takes the value 1 or 0, has mean $\mu = p$ (the probability of

1), and variance is $\sigma^2 = p(1-p)$. When sampling from a Bernoulli population, we are typically given only the binomial random variable, $T = X_1 + \cdots + X_n$, or the observed proportion of 1s (which is the sample average), $\overline{X} = \widehat{p}$. By the Central Limit Theorem, the normal distribution provides an adequate approximation to the distribution of T, or that of $\widehat{p}$, whenever $n \geq 5$ and $n(1-p) \geq 5$, and thus (8.3.1) is replaced by

$$\widehat{p} \overset{\cdot}{\sim} N\left(p, p(1-p)/n\right).$$

The above approximate distribution of $\widehat{p}$ leads to the (approximate) $(1-\alpha)100\%$ CI

$$\boxed{\left(\widehat{p} - z_{\alpha/2}\sqrt{\tfrac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{\alpha/2}\sqrt{\tfrac{\widehat{p}(1-\widehat{p})}{n}}\right).\ \text{NP CI for } p} \qquad (8.3.5)$$

Note that, because p is unknown, the condition $np \geq 5$ and $n(1-p) \geq 5$ for applying the CLT is replaced in practice by $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, i.e. at least five 1s and at least five 0s.

Example 8.3.3. In the car crash experiment of Example 7.3.1, it was given that 12 of 20 cars sustained no visible damage. Thus, the number of those that did not sustain visible damage and the number of those that did, exceeds five. With this information, we have $\hat{p} = 12/20 = 0.6$, so that, an approximate 95% CI, of the form given in (8.3.5), for the true value of p (the population proportion of cars that sustain no visible damage) is

$$0.6 - 1.96\sqrt{\frac{0.6 \times 0.4}{20}} \leq p \leq 0.6 + 1.96\sqrt{\frac{0.6 \times 0.4}{20}},$$

or $0.6 - 0.2191 \leq p \leq 0.6 + 0.2191$ or $0.38 \leq p \leq 0.82$.

Example 8.3.4. The point estimate for the probability that a certain component functions properly for at least 1000 hours, based on a sample of $n = 100$ of such components, is $\hat{p} = 0.91$. Give a 95% CI for p.

Solution. The sample size conditions for applying the aforementioned CI for a binomial parameter p hold here. Thus the desired CI is:

$$\hat{p} \pm z_{.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .91 \pm 1.96\sqrt{\frac{.91(.09)}{100}} = .91 \pm .059.$$

8.3.3 Nonparametric CIs for the Median and Percentiles

Let $X_1, \ldots, X_n$ denote a sample from a population having a continuous distribution, and let x_p denote the $(1-p)100$th percentile. The basic idea for constructing a nonparametric

CI for x_p is to associate a Bernoulli trial Y_i with each observation X_i:

$$Y_i = \begin{cases} 1 & \text{if } X_i > x_p \\ 0 & \text{if } X_i < x_p \end{cases}$$

Thus, the probability of a 1 (or success) in each Bernoulli trial is p. Let $Y = \sum_i Y_i$ be the Binomial(n, p) random variable. Let also $X_{(1)} < \cdots < X_{(n)}$ denote the ordered sample values. Then the events

$$X_{(k)} < x_p < X_{(k+1)}, \quad X_{(k)} < x_p, \quad x_p < X_{(k+1)},$$

are equivalent to

$$Y = n - k, \quad Y \leq n - k, \quad Y \geq n - k,$$

respectively.

Nonparametric $(1 - \alpha)100\%$ CIs for x_p will be of the form

$$\boxed{X_{(a)} < x_p < X_{(b)}, \quad \text{NP CI for } x_p} \tag{8.3.6}$$

where the indices a, b are found from the requirements that

$$\begin{aligned} P\left(x_p < X_{(a)}\right) &= \alpha/2 \\ P\left(X_{(b)} < x_p\right) &= \alpha/2. \end{aligned}$$

These requirements can equivalently be expressed in terms of Y as

$$P\left(Y \geq n - a + 1\right) = \alpha/2 \tag{8.3.7}$$

$$P\left(Y \leq n - b\right) = \alpha/2, \tag{8.3.8}$$

and thus, a and b can be found using either the binomial tables (since $Y \sim$ Binomial(n, p)) or the normal approximation to the binomial. For example, using the normal approximation with continuity correction, a and b are found from:

$$\frac{n - a - np + 0.5}{\sqrt{np(1-p)}} = z_{\alpha/2}, \quad \frac{n - b - np + 0.5}{\sqrt{np(1-p)}} = -z_{\alpha/2}.$$

The special case of $p = 0.5$, which corresponds to the median, deserves separate consideration. In particular, the $(1 - \alpha)100\%$ CIs for the median $\widetilde{\mu}$ will be of the form

$$\boxed{X_{(a)} < \widetilde{\mu} < X_{(n-a+1)}, \quad \text{NP CI for } \widetilde{\mu}} \tag{8.3.9}$$

so that only a must be found. To see that this is so, note that if $p = 0.5$ then Y, the number of 1s, has the same distribution as $n - Y$, the number of 0s (both are Binomial$(n, 0.5)$). Thus,

$$\begin{aligned} P\left(Y \geq n - a + 1\right) &= P\left(n - Y \geq n - a + 1\right) \\ &= P\left(Y \leq a - 1\right), \end{aligned}$$

which implies that if a, b satisfy (8.3.7), (8.3.8), respectively, then they are related by $n - b = a - 1$ or $b = n - a + 1$. The a in relation (8.3.9) can be found from the requirement that

$$P\left(Y \leq a - 1\right) = \alpha/2. \tag{8.3.10}$$

Example 8.3.5. Let $X_1, \ldots, X_{25}$ be a sample from a continuous population. Find the confidence level of the following CI for the median:

$$\left(X_{(8)}, X_{(18)}\right).$$

Solution. First note that the CI $(X_{(8)}, X_{(18)})$ is of the form (8.3.9) with $a = 8$. According to the formula (8.3.10), and the binomial tables,

$$\alpha = 2P\left(Y \leq 7\right) = 2(0.022) = 0.044.$$

Thus, the confidence level of the CI $\left(X_{(8)}, X_{(18)}\right)$, is $(1 - \alpha)100\% = 95.6\%$.

8.3.4 Exercises

1. For a random sample of 50 measurements of the breaking strength of cotton threads, $\bar{X} = 210$ grams and $S = 18$ grams.

   (a) Obtain an 80% CI for the true mean breaking strength. What assumptions, if any, are needed for the validity of the confidence interval?

   (b) Would the 90% CI be wider the 80% CI?

   (c) A classmate offers the following interpretation of the CI you obtained in a): We are confident that 80% of all breaking strength measurements of cotton threads will be within the calculated CI. Is this interpretation correct?

2. To determine the probability that a certain component lasts (i.e. operates properly) for more than 350 hours of operation, a random sample of 37 components was put to test. Of these 24 lasted longer than 350 hours.

(a) Do a 95% CI for the probability p that a randomly selected component lasts more than 350 hours.

(b) A system consists of two such components connected in line. Thus, the system operates if and only if both components operate properly. Do a 95% C.I. for the probability that the system lasts more than 350 hours. You can assume that the life times of the two components in the system are independent. (Hint: Express the probability that the system lasts more than 350 hours in terms of p.)

3. A Reader's Digest/Gallup Survey on the drinking habits of Americans estimated the proportion of adults across the country who drink beer, wine, or hard liquor, at least occasionally. Of the 1516 adults interviewed, 985 said they drank.

   (a) Find a 95% confidence interval for the proportion, p, of all Americans who drink beer, wine, or hard liquor, at least occasionally.

   (b) An interpretation of the CI obtained in (a) is that the probability is 0.95 that the true proportion of adults who drink lies in the interval you obtained. True or false?

4. A random sample of 100 voters is selected and 65 of them are found to favor Candidate C. We want to find a CI for the proportion of the voting population favoring Candidate C.

   (a) State the point estimator (in symbol) and the point estimate.

   (b) Check the appropriate conditions to see whether one can use the NP CI for the parameter in (a).

   (c) Find a 95% NP CI for the proportion of the voting population favoring Candidate C.

8.4 Confidence Intervals Based on the t-Distribution

When sampling from normal populations, an estimator $\widehat{\theta}$ of some parameter θ often satisfies, for all sample sizes n,

$$\frac{\widehat{\theta}-\theta}{\widehat{\sigma}_{\hat{\theta}}} \sim t_{\nu}, \quad \text{where } \widehat{\sigma}_{\hat{\theta}} \text{ is the estimated s.e.,} \tag{8.4.1}$$

and t_ν stands for "t-distribution with ν degrees of freedom". t-distributions are symmetric. Relation (8.4.1) leads to the following bound on the error of estimation of θ,

$$\left|\widehat{\theta}-\theta\right| \leq t_{\nu,\alpha/2}\widehat{\sigma}_{\hat{\theta}}, \tag{8.4.2}$$

which holds with probability $1 - \alpha$.

The notation $t_{\nu,\alpha/2}$ is similar to the z_α notation, i.e. it corresponds to the $100(1-\alpha/2)$th percentile of the t-distribution with $n-1$ degrees of freedom. See Figure 8.1.

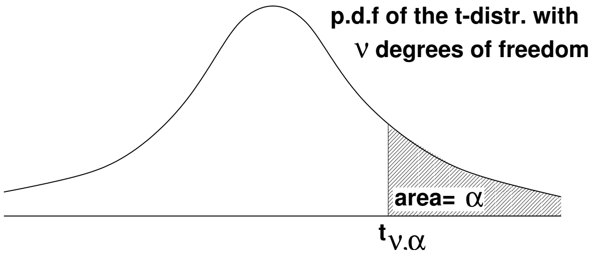


Figure 8.1: PDF and Percentile of a t Distribution.

Note that, as the degrees of freedom ν gets large, $t_{\nu,\alpha/2}$ approaches $z_{\alpha/2}$; for example, the 95th percentile for the t-distributions with $\nu = 9, 19, 60$ and 120, are 1.833, 1.729, 1.671, 1.658, respectively, while $z_{0.05} = 1.645$. Selective percentiles of t-distributions are given in the t-table.

The error bound (8.4.2) leads to the following $100(1-\alpha)\%$ CI for θ:

$$\left(\widehat{\theta} - t_{\nu,\alpha/2}\widehat{\sigma}_{\hat{\theta}},\ \widehat{\theta} + t_{n-1,\alpha/2}\widehat{\sigma}_{\hat{\theta}}\right). \tag{8.4.3}$$

We note that, if the assumption that the sampled population is normal is correct, the confidence of the above CI is exactly $100(1-\alpha)\%$ for all sample sizes n.

In the next subsections we will construct such intervals for a normal mean, the slope of a regression line, and a regression line evaluated at some value x of the covariate.

8.4.1 Confidence Intervals for a Normal Mean

Let $X_1, \ldots, X_n$ be a simple random sample from a population having a normal distribution, with mean μ and variance σ^2, both unknown.

In Chapter 4 we saw that, by subtracting the population mean from the sample mean, $\overline{X}$, and dividing the difference by the estimated standard error, $S/\sqrt{n}$, we obtain a random variable which has a t-distribution with $n-1$ degrees of freedom. That is, we have

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \tag{8.4.4}$$

Note that (8.4.4) gives the exact distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$, whereas (8.3.2) gives the approximate distribution when n is sufficiently large. Using relation (8.4.4) we obtain an alternative bound on the error of estimation of μ, namely

$$\left|\overline{X} - \mu\right| \le t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \tag{8.4.5}$$

which holds with probability $1 - \alpha$. We note again that the probability for this error bound is exact, whereas the error bound (8.3.3) holds with probability approximately $1 - \alpha$, provided that n is sufficiently large. Of course, (8.4.5) requires the normality assumption, whereas (8.3.3) does not.

The error bound (8.4.5) leads to the following $100(1 - \alpha)\%$ CI for the normal mean:

$$\boxed{\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right). \quad \begin{array}{l}\text{Normality-based, } t\text{-Confidence}\\ \text{Interval for } \mu.\end{array}} \tag{8.4.6}$$

Example 8.4.1. The mean weight loss of $n = 16$ grinding balls after a certain length of time in mill slurry is $3.42g$ with $S = 0.68g$. Construct a 99% CI for the true mean weight loss.

Solution. Here $\alpha = .01$ and $t_{n-1,\alpha/2} = t_{15,0.005} = 2.947$. Thus a 99% CI for μ is

$$\overline{X} \pm t_{n-1,\alpha/2}(S/\sqrt{n}) = 3.42 \pm 2.947(0.68/\sqrt{16}), \quad \text{or} \quad 2.92 < \mu < 3.92.$$

8.4.2 CIs for the Slope of Linear Regression

Consider a population of (X, Y) values such that the regression function of Y on X is

$$\mu_{Y|X}(x) = E(Y|X = x) = \alpha_1 + \beta_1 x,$$

and the conditional variance of Y given $X = x$,

$$\mathrm{Var}(Y_i|X_i = x) = \sigma^2,$$

is the same for all x. Under these conditions, estimation of

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$, be a simple random sample from this population. Estimation of α_1 and β_1, and, consequently, for $\mu_{Y|X}(x)$, was considered in Chapters 6 and 7. All methods of estimation, the empirical estimators of Section 6.2.3 (which can be thought of as moments estimators), the method of least squares (Subsection 7.3.3), and the method of maximum likelihood under the normality assumption, give the same estimators, which we called the least squares estimators (LSE). Restated here for easy reference, the LSE are

$$\widehat{\beta}_1 = \frac{n\sum X_iY_i - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2}, \quad \widehat{\alpha}_1 = \overline{Y} - \widehat{\beta}_1\overline{X}, \quad \widehat{\mu}_{Y|X}(x) = \widehat{\alpha}_1 + \widehat{\beta}_1 x. \qquad (8.4.7)$$

In Example 7.2.5 we saw that $\widehat{\alpha}_1$ and $\widehat{\beta}_1$ are unbiased estimators for α_1 and β_1, respectively; consequently, $\widehat{\mu}_{Y|X}(x)$ is unbiased for $\mu_{Y|X}(x)$ (see Exercise 7.2.4,7).

The construction of confidence intervals for β_1 is based on the following proposition which gives the standard error and the estimated standard error of $\widehat{\beta}_1$, and, under the additional assumption of normality, the distribution of $\widehat{\beta}_1$.

Proposition 8.4.1. *Assume that the conditional variance of each Y_i, given that $X_i = x$, is σ^2, let $\hat{\beta}_1$ be the LSE of the slope parameter β_1, and let*

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\alpha}_1 - \widehat{\beta}_1 X_i)^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n} Y_i^2 - \widehat{\alpha}_1\sum_{i=1}^{n} Y_i - \widehat{\beta}_1\sum_{i=1}^{n} X_iY_i\right]$$

be the estimator of σ^2, introduced in Subsection 7.3.3. Then,

1. *The standard error, and the estimated standard error of $\widehat{\beta}_1$ are*

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}}, \quad \textit{and} \quad \widehat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{S^2}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}},$$

   respectively.

2. *Assume now, in addition, that the conditional distribution of Y given $X = x$ is normal. Thus,*

$$Y|X = x \sim N(\alpha_1 + \beta_1 x, \sigma^2).$$

   Then, $\widehat{\beta}_1$ has a normal distribution,

$$\widehat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right), \quad \textit{or} \quad \frac{\widehat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1).$$

3. *Under the above assumption of normality,*

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}.$$

Remark 8.4.1. *The estimated standard error, $\widehat{\sigma}_{\hat{\beta}_1}$, of $\widehat{\beta}_1$ is also denoted by $S_{\hat{\beta}_1}$.*

Part 3 of Proposition 8.4.1, can be used for constructing bounds for the estimation error, $\widehat{\beta}_1 - \beta_1$, and CIs for β_1, as in relations (8.4.2) and (8.4.3). Thus, the bound

$$\left|\widehat{\beta}_1 - \beta_1\right| \leq t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\beta}_1}. \tag{8.4.8}$$

on the error of estimation of β_1, holds with probability exactly $1-\alpha$, for all n. This bound leads to an $100(1-\alpha)\%$ CI for the slope β_1 of the true regression slope of the form

$$\boxed{\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\widehat{\sigma}_{\widehat{\beta}_1} \quad \begin{array}{l}\text{Normality-based, } t\text{-Confidence}\\ \text{Interval for } \beta_1.\end{array}} \tag{8.4.9}$$

Example 8.4.2. Consider the setting and data of Example 7.3.14. Thus there are $n = 14$ observations, $\sum_i X_i = 890$, $\sum_i X_i^2 = 67,182$, $\sum_i Y_i = 37.6$, $\sum_i Y_i^2 = 103.54$ and $\sum_i X_iY_i = 2234.30$. Also, we saw that $\widehat{\beta}_1 = -0.0147209$, $\widehat{\alpha}_1 = 3.6209072$, and the point estimate of σ^2 is

$$S^2 = \frac{103.54 - \widehat{\alpha}_1(37.6) - \widehat{\beta}_1(2234.30)}{12} = \frac{.2624532}{12} = 0.02187.$$

a) Using this information construct a 95% CI for β_1.

b) Let Y_1 denote an observation made at $X_1 = 30$, and Y_2 denote an observation made at $X_2 = 35$. Construct a 95% CI for $E(Y_1 - Y_2)$, the expected difference between Y_1 and Y_2.

Solution. a) With the given information, the estimated standard error of $\hat{\beta}_1$ is

$$\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{S^2}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2}} = \sqrt{\frac{0.02187}{67,182 - \frac{1}{14}890^2}} = 0.001414.$$

Thus, the 95% CI for β_1 is

$$\begin{aligned} & -0.0147209 \pm t_{0.025,12}0.001414 = -0.0147209 \pm 2.179 \times 0.001414 \\ = \quad & -0.0147209 \pm 0.00308 = (-0.0178, -0.01164). \end{aligned}$$

b) Arguing as in Example 7.3.15, $E(Y_1 - Y_2) = -5\beta_1$, so that a point estimate of $E(Y_1 - Y_2)$ is $E(\widehat{Y_1 - Y_2}) = -5\widehat{\beta}_1 = 5(0.0147209)$. A 95% CI for this is

$$-5\widehat{\beta}_1 \pm 5t_{\alpha/2,n-2}\widehat{\sigma}_{\widehat{\beta}_1} = 5(0.0147209) \pm 5 \times 2.179 \times 0.001414.$$

8.4.3 CIs for the Regression Line

The construction of confidence intervals for $\mu_{Y|X=x} = \alpha_1 + \beta_1 x$ is based on the following proposition which gives the standard error and the estimated standard error of $\widehat{\mu}_{Y|X=x}$, and, under the additional assumption of normality, the distribution of $\widehat{\mu}_{Y|X=x}$.

Proposition 8.4.2. *Assume that the conditional variance of each Y_i, given that $X_i = x$, is σ^2, let $\widehat{\mu}_{Y|X=x}$ be the estimator of the regression line $\mu_{Y|X=x}$, and let*

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\alpha}_1 - \widehat{\beta}_1 X_i)^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n} Y_i^2 - \widehat{\alpha}_1 \sum_{i=1}^{n} Y_i - \widehat{\beta}_1 \sum_{i=1}^{n} X_i Y_i\right]$$

be the estimator of σ^2, introduced in Subsection 7.3.3. Then,

1. *The standard error and the estimated standard error of $\widehat{\mu}_{Y|X=x}$ are*

$$\sigma_{\widehat{\mu}_{Y|X=x}} = \sigma\sqrt{\frac{1}{n} + \frac{n(x-\overline{X})^2}{n\sum X_i^2 - (\sum X_i)^2}}, \quad \widehat{\sigma}_{\widehat{\mu}_{Y|X=x}} = S\sqrt{\frac{1}{n} + \frac{n(x-\overline{X})^2}{n\sum X_i^2 - (\sum X_i)^2}}.$$

2. *Under the normality assumption,*

$$\widehat{\mu}_{Y|X=x} = \widehat{\alpha}_1 + \widehat{\beta}_1 x \sim N(\mu_{Y|X=x}, \sigma^2_{\widehat{\mu}_{Y|X=x}}).$$

3. *Under the normality assumption,*

$$T = \frac{\widehat{\mu}_{Y|X=x} - \mu_{Y|X=x}}{\widehat{\sigma}_{\widehat{\mu}_{Y|X=x}}} \sim t_{n-2}.$$

Remark 8.4.2. *The estimated standard error, $\widehat{\sigma}_{\widehat{\mu}_{Y|X=x}}$, of $\widehat{\mu}_{Y|X=x}$ is also denoted by $S_{\widehat{\mu}_{Y|X=x}}$.*

Remark 8.4.3. *One feature of the standard error $\sigma_{\hat{\mu}_{Y|X=x}}$, and of $S_{\hat{\mu}_{Y|X=x}}$, of $\hat{\mu}_{Y|X=x}$ is worth pointing out: the numerator of the second term has $(x - \bar{X})^2$ meaning that the farther away x is from $\bar{X}$, the larger the standard error.*

Part 3 of Proposition 8.4.2, can be used for constructing bounds for the estimation error, $\widehat{\mu}_{Y|X=x} - \mu_{Y|X=x}$, and CIs for $\mu_{Y|X=x}$. In particular, a $100(1-\alpha)\%$ CI for the regression line is

$$\boxed{\widehat{\mu}_{Y|X=x} \pm t_{\alpha/2,n-2} S_{\widehat{\mu}_{Y|X=x}} \quad \begin{array}{l}\text{Normality-based } t\text{-Confidence}\\ \text{Interval for } \mu_{Y|X=x}.\end{array}} \tag{8.4.10}$$

Remark 8.4.3 regarding the term $(x - \overline{X})^2$ in the expression for $S_{\widehat{\mu}_{Y|X=x}}$ implies that confidence intervals for $\mu_{Y|X=x}$ get wider as x get farther away from $\overline{X}$.

Because of this, it is not recommended to make a confidence interval for the expected response at $x < X_{(1)}$ (i.e. x less than the smallest X-value in the data) or at $x > X_{(n)}$.

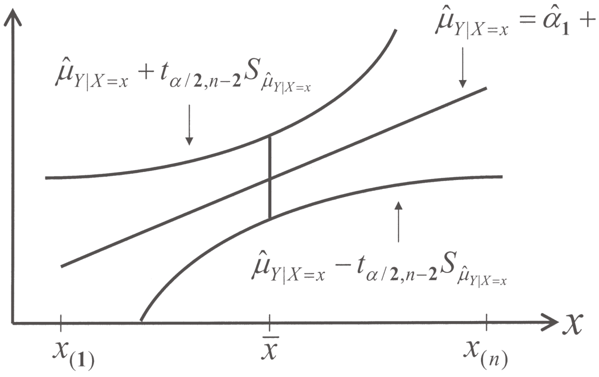


Figure 8.2: Confidence Intervals for $\mu_{Y|X=x}$ Get Wider Away from $\overline{X}$

Example 8.4.3. Let $n = 11, \sum X_i = 292.90, \sum Y_i = 69.03, \sum X_i^2 = 8141.75, \sum X_i Y_i = 1890.200, \sum Y_i^2 = 442.1903, \widehat{\mu}_{Y|X} = 2.22494 + .152119X$, and $S = 0.3444$. Thus,

$$S_{\widehat{\mu}_{Y|X=x}} = 0.3444 \sqrt{\frac{1}{11} + \frac{11(x - 26.627)^2}{11(8141.75) - (292.9)^2}}.$$

It follows that the estimated standard error of $\widehat{\mu}_{Y|X=26.627}$ is $S_{\widehat{\mu}_{Y|X=26.627}} = 0.1038$, while the estimated standard error of $\hat{\mu}_{Y|X=25}$ is $S_{\widehat{\mu}_{Y|X=25}} = 0.1082$. Note that $S_{\widehat{\mu}_{Y|X=25}} > S_{\widehat{\mu}_{Y|X=26.627}}$ which is expected since $\bar{X} = 26.627$. As a consequence, the 95% CI for $\mu_{Y|X=25}$ is

$$\widehat{\mu}_{Y|X=25} \pm t_{.025,9} 0.1082 = 6.028 \pm 0.245$$

which is wider than the 95% CI for $\mu_{Y|X=26.627}$:

$$\widehat{\mu}_{Y|X=26.627} \pm t_{.025,9} 0.1038 = 6.275 \pm 0.235.$$

8.4.4 Exercises

1. A sample of 15 dowel joints using 1.25in rails, yielded a sample mean and sample standard deviation of flat-wise bonding strength of 1215.6 lb/in and 137.4 lb/in, respectively. Assume that the strength measurements are normally distributed.

   (a) Obtain a 99% confidence interval for the mean flat-wise bonding strength of this type of dowel joint.

   (b) Would a 95% CI be wider than the 99% CI constructed above?

   (c) The true mean flat-wise bonding strength is between the two numbers you obtained in part (a) with probability 0.99. True of false?

(d) Ninety nine percent of all dowel joint have flat-wise bonding strength between the two numbers you obtained in part (a). True or false?

2. A random sample of 25 engineering seniors is selected. The average amount of study time for this group turned out 15 hours/week with a sample standard deviation of 4 hours/week. Assume that the amount of study time of a randomly selected engineering student is normally distributed.

   (a) Construct a 98% CI for the true population mean.

   (b) Supposing that population standard deviation is known to be 4 hours/week, repeat part a.

3. Analysis of the venom of seven eight-day-old worker bees yielded the following observations on histamine content in nanograms: 649, 832, 418, 530, 384, 899, 755.

   (a) Compute a 90% confidence interval for the true average histamine content for all worker bees of this age. State the formula you are using.

   (b) Were any assumptions necessary to justify your computations? If so, list them.

   (c) Explain what 90% confidence means.

4. A question relating to a study of the echolocation system for bats is how far apart are the bat and an insect when the bat first senses the insect. The technical problems for measuring this are complex and so only 11 data points, $Y_1, \ldots, Y_{11}$, were obtained. It is given that $\sum_{i=1}^{11} Y_i = 532$ and $\sum_{i=1}^{11} Y_i^2 = 29,000$. Construct a 95% confidence interval for the mean distance μ. What assumption did you use?

5. A tire company wants to get a 99% CI for the true average lifetime, (μ), of a brand of tires. 10 tires are selected and tested on a test wheel that simulated normal road condition. The sample average lifetime is 41 (in thousands of miles), and the sample standard deviation is 3.59.

   (a) State the assumptions you make in order to construct a 99% confidence interval for μ.

   (b) Find a 99% confidence interval for the true average lifetime of the brand of tires, μ.

6. From twelve experiments it is found that the average amount of DDT causing 50% fatality in rainbow trout is 9.10 ppm with a standard deviation of 3.16 ppm.

   (a) Obtain an 90% confidence interval for the true mean amount of DDT causing 50% fatality in rainbow trout.

   (b) What assumptions, if any, are needed for the validity of the confidence interval?

7. The article ”Toxicity Assessment of Wastewaters, River Waters, and Sediments in Austria Using Cost-Effective Microbiotests”, by M. Latif and E. Licek (2004, Environmental Toxicology, Vol 19, No 4, 302-308) reports data on conductivity (μS/cm) measurements of surface water (X), and water in the sediment at the bank of a river (Y), taken at 10 points during winter. The data are

   ```
   X : 220 800 277 223 226 240 752 317 333 340
   Y : 386 889 358 362 411 390 927 612 554 532
   ```

   Assume that the regression function of Y on X is linear, and the conditional variance of Y given a value x of X is σ^2, the same for all x.

   (a) Construct a 95% CI for the true slope of the regression line.

   (b) Construct a 90% confidence interval for the expected sediment conductivity when the surface conductivity is 300.

   (c) A 90% prediction interval for the next sediment conductivity measurement at surface conductivity 300 will be shorter than the interval you constructed in part b). TRUE FALSE

   (d) Construct a confidence interval for the expected change in the average sediment conductivity when the surface conductivity increases by 50.

8. The article ”Effects of Bike Lanes on Driver and Bicyclist Behavior” (Transportation Eng. J., 1977: 243-256) reports data from a study on X = distance between a cyclist and the roadway center line, and Y = the separation distance between the cyclist and a passing car (both determined by photography). The data from ten streets with bike lanes are:

   ```
   X : 12.8 12.9 12.9 13.6 14.5 14.6 15.1 17.5 19.5 20.8
   Y :  5.5  6.2  6.3  7.0  7.8  8.3  7.1 10.0 10.8 11.0
   ```

   Assume that the regression function of Y on X is linear, and the conditional variance of Y given a value x of X is σ^2, the same for all x.

   (a) Obtain an estimate of σ, and the estimated standard error of the estimator of the regression slope.

   (b) Assume that the conditional distribution of Y when a value of X is given, is normal. Let Y_1 be the separation distance between a cyclist and a passing car when the cyclist's distance from the roadway center line is 14 feet, and let Y_2 be the same distance for a cyclist whose distance from the roadway center line is 16 feet. Estimate the probability that Y_1 will be greater than Y_2.

(c) Construct a 95% CI for the average separation distance between a cyclist and a passing car when the cyclist's distance from the roadway center line is 14 feet.

(d) Will the 95% CI for the average separation distance between a cyclist and a passing car when the cyclist's distance from the roadway center line is 19 feet be shorter than the CI of part (c).

(e) Construct a 90% prediction interval for the separation distance between the next cyclist whose distance from the roadway center line is 15 feet and a passing car.

8.5 The Issue of Precision

Precision in the estimation of a parameter θ is quantified by the size of the bound of the error of estimation $|\hat{\theta} - \theta|$. Equivalently, it can be quantified by the length of the corresponding CI, which is twice the size of the error bound. A shorter error bound, or shorter CI, implies more precise estimation.

The bounds we have seen on the error of estimation of a population mean μ and population proportion p, are of the form

$$\begin{aligned}
\left|\overline{X} - \mu\right| &\leq z_{\alpha/2}\frac{S}{\sqrt{n}} \quad \text{(nonparametric case, } n > 30\text{)} \\
\left|\overline{X} - \mu\right| &\leq t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \quad \text{(normal case, any } n\text{)} \\
|\hat{p} - p| &\leq z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (n\hat{p} \geq 5, n(1-\hat{p}) \geq 5).
\end{aligned}$$

The first and the third of these bounds hold with probability approximately $1-\alpha$, while the second holds with probability exactly $1-\alpha$, if the underlying population has a normal distribution. In the rare cases when σ can be considered know, the S of the first error bound is replaced by σ.

The above expressions suggest that the size of the error bound (or length of CI) depends on the sample size n. In particular, a larger sample size yields a smaller error bound, and thus more precise estimation. Shortly we will see how to choose n in order to achieve a prescribed degree of precision.

It can also be seen that the probability with which the error bound holds, i.e. $1-\alpha$, also affects the size of the error bound, or length of the CI. For example, a 90% CI (($\alpha = .1$, so $\alpha/2 = .05$)), is narrower than a 95% CI (($\alpha = .05$, so $\alpha/2 = .025$)), which is less narrower than a 99% CI (($\alpha = .01$, so $\alpha/2 = .005$)). This is so because

$$z_{.05} = 1.645 < z_{.025} = 1.96 < z_{.005} = 2.575,$$

and similar inequalities for the t-critical values. The increase of the length of the CI with the level of confidence is to be expected. Indeed, we are more confident that the wider CI will contain the true value of the parameter. However, we rarely want to reduce the length of the CI by decreasing the level of confidence.

We will now deal with the main learning objective of this section, which is how to choose n in order to achieve a prescribed degree of precision in the estimation of μ and of p. As we will see, the main obstacle to getting a completely satisfactory answer to this question lies in the fact that the estimated standard error, which enters all expressions of error bounds, is unknown prior to the data collection. For the two estimation problems, we will discuss separately ways to bypass this difficulty.

8.5.1 Sample size determination for μ

Consider first the task of choosing n for precise estimation of μ, in the rare case that σ is known. In that case, if normality is assumed, or if we know that the needed sample size will be > 30, the required n for achieving a prescribed length L of the $(1-\alpha)100\%$ CI, is found equating the expression for the length of the CI to L, and solving the resulting equation for n. That is, we solve

$$2z_{.025}\frac{\sigma}{\sqrt{n}} = L,$$

for n. The solution is

$$n = \left(2z_{\alpha/2}\frac{\sigma}{L}\right)^2. \tag{8.5.1}$$

More likely than not, the solution will not be an integer, in which case the recommended procedure is to round up. The practice of rounding up guarantees that the prescribed objective will be more than met.

Example 8.5.1. The time to response (in milliseconds) to an editing command with a new operating system is normally distributed with an unknown mean μ and $\sigma = 25$. We want a 95% CI for μ of length $L = 10$ milliseconds. What sample size n should be used?

Solution. For 95% CI, $\alpha = .05, \alpha/2 = .025$ and $z_{.025} = 1.96$. Thus, from formula (8.5.1) we obtain

$$n = \left(2 \cdot (1.96)\frac{25}{10}\right)^2 = 96.04,$$

which is rounded up to $n = 97$.

Typically however, σ is unknown, and thus, sample size determinations must rely on some preliminary approximation, S_{prl}, to it. Two commonly used methods for obtaining this approximation are:

a) If the range of population values is known, then σ can be approximated by dividing the range by 3.5 or 4. That is, use

$$S_{prl} = \frac{\text{range}}{3.5}, \quad \text{or} \quad S_{prl} = \frac{\text{range}}{4},$$

instead of σ in the formula (8.5.1). This approximation is obtained by considering the standard deviation of a uniform in (a, b) random variable, which is $\sigma = (b - a)/\sqrt{12} = (b - a)/3.464$.

b) Alternatively, the approximation can be based on the sample standard deviation, S_{prl}, of a preliminary sample.

The reason why this is not a very satisfactory solution is because the standard deviation of the final sample, upon which the CI will be calculated, will be different from that of the preliminary approximation of it, regardless of how this approximation was obtained. Thus the prescribed precision objective might not be met, and some trial-and-error iteration might be involved. The trial-and-error process gets slightly more complicated with the t-distribution CIs, because the t-percentiles change with the sample size.

8.5.2 Sample size determination for p

Consider next the selection of sample size for meeting a prescribed level of precision in the estimation of the binomial parameter p.

As noted previously, the required n for achieving a prescribed length L of the $(1-\alpha)100\%$ CI, is found by equating the expression for the length of the CI for p to L, and solving the resulting equation for n. That is, we solve

$$2z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = L$$

for n. The solution is

$$n = \frac{4z_{\alpha/2}^2 \widehat{p}(1-\widehat{p})}{L^2}. \tag{8.5.2}$$

The problem is that the value of $\widehat{p}$ is not known before the sample is collected, and thus sample size determinations must rely on some preliminary approximation, $\widehat{p}_{prl}$, to it. Two commonly used methods for obtaining this approximation are:

a) When preliminary information about p exists. This preliminary information may come either from a small pilot sample or from expert opinion. In either case, the preliminary $\widehat{p}_{prl}$ is entered in (8.5.2) instead of $\widehat{p}$ for sample size calculation. Thus,

$$n = \frac{4z_{\alpha/2}^2 \hat{p}_{prl}(1-\hat{p}_{prl})}{L^2}, \tag{8.5.3}$$

and we round up.

b) When no preliminary information about p exists. In this case, we replace $\widehat{p}(1-\widehat{p})$ in (8.5.2) by 0.25. The rationale for doing so is seen by noting that $\widehat{p}(1-\widehat{p}) \leq 0.25$; thus, by using the larger 0.25 the calculated sample size will be at least as large as needed for meeting the precision specification. This gives

$$n = \frac{z_{\alpha/2}^2}{L^2}. \tag{8.5.4}$$

Example 8.5.2. A preliminary sample gave $\hat{p}_{prl} = 0.9$. How large should n be to estimate the probability of interest *to within* 0.01 with 95% confidence?

Solution. "To within 0.01" is another way of saying that the 95% bound on the error of estimation should be 0.01, or the desired CI should have a width of 0.02. Since we have preliminary information, we use (8.5.3):

$$n = \frac{4(1.96)^2(.91)(.09)}{(.02)^2} = 3146.27.$$

This is rounded up to 3147.

Example 8.5.3. A new method of pre-coating fittings used in oil, brake and other fluid systems in heavy-duty trucks is being studied. How large n is needed to estimate the proportion of fittings that leak to within .02 with 90% confidence? (No prior info available).

Solution. Here we have no preliminary information about p. Thus, we apply the formula (8.5.4) and we obtain

$$n = z_{\alpha/2}^2/L^2 = (1.645)^2/(.04)^2 = 1691.26.$$

This is rounded up to 1692.

8.5.3 Exercises

1. How many test specimens do we need to obtain a 98% CI with length 0.4 for the average shrinkage percentage of plastic clay if it is known that the shrinkage percentage is normally distributed with $\sigma = 1.2$.

2. Assuming that the true value of σ is 150.0, find the sample size needed to estimate the mean to within 30 lb/in (that is, to give the confidence interval length 60 lb/in) with 90% confidence.

3. With the information given in Exercise 8.3.4,1, what sample size is needed for obtain an 80% confidence interval of length 5?

4. A food processing company is considering the marketing of a new product. Among 40 randomly chosen consumers 9 said that they would purchase the new product and give it a try. Construct a 90% confidence interval for the the true proportion of potential buyers.

5. Use the information in Exercise 4 to determine the sample size that is needed for a 90% confidence interval of length 0.1. What would your answer be if no information were available?

6. (a) A survey of potential voters is being planned by the governor's staff to determine the proportion of the voting population favoring Referendum C to within 2 percentage points (0.02) with 95% confidence. What sample size is necessary to assure this objective?

   (b) Independently from the governor's office, a local television station conducted a survey regarding the same issue. A random sample of 100 potential voters was selected and 65 favored the referendum. Compute a 95% confidence interval for the proportion of potential voters favoring Referendum C.

   (c) The governor's office hears about the TV station survey and wants to make use of the information in their sample size calculation. Give the required sample size to achieve the objective stated in part (a) using this new information.

7. a) With the information given in Exercise 8.3.4,4 determine the needed sample size needed for the 95% CI to have length at most 0.05
   b) What sample size is necessary if the 94% confidence interval is to have length at most 0.05 irrespective of the true value of the proportion?

8. A study of the electro-mechanical protection devises used in electrical power systems showed that of 193 devices that failed when tested 75 where due to mechanical parts failures.

(a) Find a 95% confidence interval for p the proportion of failures due to mechanical causes.

(b) How large a sample is required to estimate p within 0.03, i.e. to have a confidence interval of length 0.06, with 95% confidence:
(i) When you have no idea of the approximate value of p;
(ii) When a preliminary investigation leads you to believe that $p \approx 0.39$

9. a) Using the information given in Exercise 8.3.4,3, obtain a sample size that will ensure the length of a 95% confidence interval for p is 0.02.

   b) Suppose now that no prior estimate of p exists. Obtain a sample size that will ensure the length of a 95% confidence interval for p is 0.02.

10. a) With the information given in Exercise 8.3.4,2, what sample size is needed to yield a CI for p whose length is at most 0.05?

    b) Assuming no prior information on p (i.e. assuming the experiment which yielded 24 of 37 components lasting more than 350 hours had not been conducted), determine the sample required to yield a 95% CI for p whose length is at most 0.1.

8.6 Prediction Intervals

The meaning of the word *prediction* is related to, but distinct from, the word *estimation*. The latter is used when we are interested in learning the value of a population or model parameter, while the former is used when we want to learn about the value that a future observation might take. In this section we will discuss prediction of a (future) observation, both in a univariate context, and in a bivariate context when a linear regression model is appropriate.

8.6.1 Prediction Based on a Univariate Sample

For a concrete example, suppose you contemplate eating a hot dog and you wonder about the amount of fat in the hot dog which *you will eat*. This is different from the question ”what is the expected (mean) amount of fat in hot dogs?” To further emphasize the difference between the two, suppose that the amount of fat in a randomly selected hot dog is known to be $N(20, 9)$. Thus there are no unknown parameters to be estimated. In particular we know that the expected amount of fat in hot dogs is 20 gr. Still the amount of fat in the hot dog which you will eat is unknown, simply because it is a random variable.

How do we predict it? According to well-accepted criteria, the best point-predictor of a normal random variable with mean μ is μ. A $(1-\alpha)100\%$ **prediction interval**, or **PI**, is an interval that contains the random variable (which is being predicted) with probability $1-\alpha$. If both μ and σ are known, then a $(1-\alpha)100\%$ PI is

$$\mu \pm z_{\alpha/2}\sigma.$$

In our particular example, $X \sim N(20, 9)$, so the best point predictor of X is 20 and a 95% PI is $20 \pm (1.96)3 = (14.12, 25.88)$.

When μ, σ are unknown (as is typically the case) we use a sample $X_1, \ldots, X_n$ to estimate μ, σ by $\overline{X},\ S$, respectively. Then, as best point predictor of a future observation, we use $\overline{X}$. But now, the prediction interval (always assuming normality) must take into account the variability of $\overline{X},\ S$ as estimators of $\mu,\ \sigma$. Doing so yields the following $(1-\alpha)100\%$ PI for the next observation X:

$$\boxed{\left(\overline{X} - t_{\alpha/2,n-1}S\sqrt{1+\tfrac{1}{n}}, \overline{X} + t_{\alpha/2,n-1}S\sqrt{1+\tfrac{1}{n}}\right). \quad \text{PI for a future observation}} \qquad (8.6.1)$$

In the above formula, the variability of $\overline{X}$ is accounted for by the $\dfrac{1}{n}$, and the variability of S is accounted for by the use of the t-percentiles.

Example 8.6.1. The fat content measurements from a sample of size $n = 10$ hot dogs, gave sample mean and sample standard deviation of $\overline{X} = 21.9$, and $S = 4.134$. Give a 95% PI for the fat content of the next hot dog to be sampled.

Solution. Using the given information in the formula (8.6.1), we obtain the PI

$$\overline{X} \pm t_{.025,9}\, S\sqrt{1+\frac{1}{n}} = (12.09, 31.71).$$

8.6.2 Prediction Intervals Based on Simple Linear Regression

Often we are interested in a confidence statement regarding a future observation of the response variable at covariate value $X = x$. This is quite different from a confidence statement regarding the expected response $\mu_{Y|X=x}$ at $X = x$ because a future observation is a random variable. As we saw above, confidence statements regarding a random variable are called prediction intervals (PIs) and incorporate the additional variation of the random variable that is to be predicted. A $100(1-\alpha)\%$ PI for a future observation at $X = x$ is

$$\boxed{\widehat{\mu}_{Y|X=x} \pm t_{\alpha/2,n-2} S\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\sum X_i^2 - (\sum X_i)^2}} \quad \begin{array}{l}\text{PI for a future}\\ \text{observation at } X = x\end{array}} . \qquad (8.6.2)$$

Example 8.6.2. Consider the information given in Example 8.4.3. Thus, $n = 11, \sum X_i = 292.90, \sum Y_i = 69.03, \sum X_i^2 = 8141.75, \sum X_i Y_i = 1890.200, \sum Y_i^2 = 442.1903, \widehat{\mu}_{Y|X} = 2.22494 + .152119X$, and $S = 0.3444$. Construct a 95% PI for a future observation, made at $X = 25$.

Solution. With the given information, the 95% PI at $X = 25$ is

$$\widehat{\mu}_{Y|X=25} \pm t_{.025,9}(0.344)\sqrt{1+\frac{1}{11}+\frac{11(1.627)^2}{11\sum X_i^2 - (\sum X_i)^2}} = 6.028 \pm 0.8165.$$

Recall that the 95% CI for $\mu_{Y|X=25}$ was found to be

$$\widehat{\mu}_{Y|X=25} \pm t_{.025,9} 0.1082 = 6.028 \pm 0.245.$$

This demonstrates that PIs are wider than CIs.

8.6.3 Exercises

1. A random sample of 16 chocolate chip cookies made by a machine has an average weight of 3.1 oz and a standard deviation of 0.3 oz. Assume that the weight of the cookies is normally distributed.

   (a) Compute an interval which contains the true weight of the next chocolate chip cookie one is about to eat with confidence 90%.

   (b) Which interval (confidence or prediction) will be wider?

2. Using the context and information in Exercise 8.4.4,1, obtain a 85% prediction interval for the flat-wise bonding strength of the next dowel joint.

3. Using the context and information given in Exercise 8.4.4,7, answer the following questions:

   (a) Will a 90% prediction interval for the next sediment conductivity measurement at surface conductivity 300 be wider than the confidence interval you constructed in part (b) of the aforementioned exercise?

   (b) Will a 90% prediction interval for the next sediment conductivity measurement at surface conductivity $\overline{X} = 372.8$ be wider than the prediction interval at 300?

(c) Construct the 90% prediction interval at $\overline{X} = 372.8$.

4. Using the context and information given in Exercise 8.4.4,8, construct a 90% prediction interval for the separation distance between the next cyclist whose distance from the roadway center line is 15 feet and a passing car.

8.7 *CIs for a Normal Variance

Let $X_1, \ldots, X_n$ be a random sample from a population whose distribution belongs in the normal family. In Chapter 5, we saw that, for normal samples, the sampling distribution of S^2 is a multiple of χ^2_{n-1} random variable. Namely,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

This fact implies that

$$\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}$$

will be true $(1-\alpha)100\%$ of the time, where $\chi^2_{\alpha/2,n-1}$, $\chi^2_{1-\alpha/2,n-1}$ denote percentiles of the χ^2_{n-1} distribution as shown in the figure below.

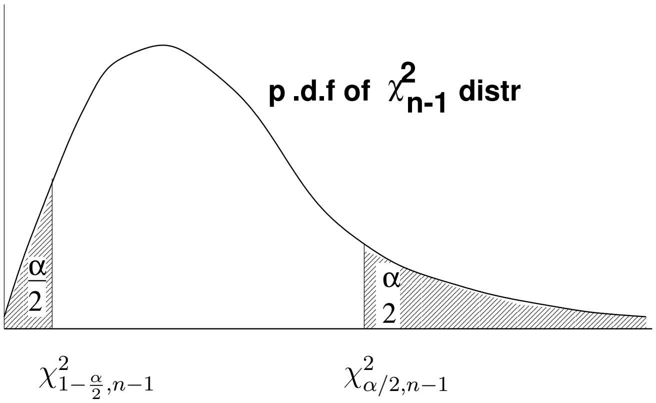


Figure 8.7: PDF and Percentile of a Chi-Square Distribution.

Note that the bounds on the error of estimation of σ^2 by S^2 are given in terms of the ratio S^2/σ^2. After some algebraic manipulations, we obtain that

$$\boxed{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \quad \text{Normality-based CI for } \sigma^2} \tag{8.7.1}$$

is a $(1-\alpha)100\%$ CI for σ^2.

Selective percentiles of χ^2 distributions are given in chi-square table.

Example 8.7.1. An optical firm purchases glass to be ground into lenses. As it is important that the various pieces of glass have nearly the same index of refraction, interest lies in the variability. A srs of size $n = 20$ measurements, yields $S^2 = (1.2)10^{-4}$. Find a 95% CI for σ.

Solution. Here $n - 1 = 19$, $\chi^2_{.975,19} = 8.906$, and $\chi^2_{.025,19} = 34.852$. Thus, according to (8.7.1),

$$\frac{(19)(1.2 \times 10^{-4})}{32.852} < \sigma^2 < \frac{(19)(1.2 \times 10^{-4})}{8.906}.$$

It follows that a 95% CI for σ is

$$\sqrt{\frac{(19)(1.2 \times 10^{-4})}{32.852}} < \sqrt{\sigma^2} < \sqrt{\frac{(19)(1.2 \times 10^{-4})}{8.906}},$$

or $.0083 < \sigma < .0160$.

8.7.1 Exercises

1. An important quality characteristic of the lapping process which is used to grind certain silicon wafers to the proper thickness is the population standard deviation, σ, of the thickness of dice cut from the wafers. If the thickness of 15 dice cut from such wafers have sample standard deviation of 0.64 mil, construct a 95% confidence interval for σ. What assumption did you use?

2. With the information given in Exercise 8.6.3,1, find a 95% confidence interval for σ, the true standard deviation of the weight of chocolate chip cookies made by the machine.

Chapter 9

Testing of Hypotheses

9.1 Introduction

In this chapter we will consider the problem of testing whether a given hypothesis regarding the true value of a parameter θ is supported by the data.

We have seen that confidence intervals provide a set of plausible (i.e. compatible with the data) values for the true value of θ. Thus, confidence intervals can be (and are) used to conduct hypothesis testing. For example, consider testing a hypothesis of the form

$$H_0 : \theta = \theta_0, \tag{9.1.1}$$

where θ_0 is a specified value. A sensible way of testing this hypothesis is to construct a CI for θ and check if the specified value θ_0 is one of the plausible values for the true θ. If θ_0 does not belong in the CI then the hypothesis H_0 is refuted, or *rejected*, and if θ_0 belongs in the CI then H_0 is not rejected. As an illustration, suppose we are interested in testing whether or not the mean value of a population equals 9.8 (so $\theta = \mu$ and $\theta_0 = 9.8$), which is written as

$$H_0 : \mu = 9.8.$$

Suppose, further, that the data yield a 95% CI for the true μ of $(9.3, 9.9)$. Since 9.8 belongs in the 95% CI, we say that H_0 is not rejected (it is not refuted by the data) at *level of significance* $\alpha = 0.05$.

Even though there is a close connection between CIs and hypothesis testing, there are a number of specific questions and issues that arise in hypothesis testing and those deserve separate treatment. These issues are:

1. **The null hypothesis and the alternative hypothesis.** In every hypothesis testing situation, there is a null hypothesis and an alternative hypothesis. Typically, the statement of the alternative hypothesis is the complement of the statement of the null hypothesis. For example, the alternative to the null hypothesis $H_0 : \theta = \theta_0$ is

   $$H_a : \theta \neq \theta_0.$$

   This alternative hypothesis is called *two-sided.* Other common null hypotheses are of the form

   $$H_0 : \theta \leq \theta_0, \text{ or } H_0 : \theta \geq \theta_0, \tag{9.1.2}$$

   with corresponding alternative hypotheses

   $$H_a : \theta > \theta_0, \text{ or } H_a : \theta < \theta_0. \tag{9.1.3}$$

   The alternative hypotheses in (9.1.3) are called *one-sided.* Testing procedures, however, do not treat the null and the alternative hypotheses equally. Basically, test procedures treat the null hypothesis in a manner similar to manner in which the presumption of innocence is treated in a court of law. One of the learning objectives of this chapter is to learn which of the two complementary statements should be designated as the null hypothesis.

2. **Rejection rules.** The intuitive, confidence-interval-based, procedure for rejecting the null hypothesis in relation (9.1.1), is not suitable for testing the one-sided null hypotheses in (9.1.2). While it is possible to define *one-sided* CIs and base test procedures for one-sided hypotheses on them, it is more common to present rules for rejecting a null hypothesis without making explicit reference to a CI.

3. **Sample size determination.** This issue involves considerations that are quite distinct from the considerations that determine sample size in the construction of CIs.

4. **Reporting the outcome.** Though the basic outcome of a test procedure is to reject or not to reject a null hypothesis, one can be more informative by also reporting the so-called *p-value.*

In this chapter we will learn how to deal with these issues for testing hypotheses about a population mean and proportion, and about regression parameters in the simple linear regression model.

9.2 The Null and Alternative Hypotheses

The hypothesis testing problems we will consider take the form of deciding between two competing hypotheses, the **null hypothesis**, denoted by H_0, and the **alternative hypothesis**, denoted by H_a. Proper designation of H_0 and H_a is very important, because the test procedures we will describe do not treat the two hypotheses symmetrically. In particular, test procedures are designed to favor the null hypothesis, so that, H_0 will not be rejected unless the data present an overwhelming evidence against it. To draw an analogy, test procedures treat a null hypothesis like the presumption of innocence is treated in a court of law, where the accused is presumed innocent unless proven guilty.

An immediate implication of this, is that a test procedure provides (statistical) proof that the alternative hypothesis is true, whenever the null hypothesis is rejected. Test procedures, however, are not meant to provide (statistical) proof that a null hypothesis is true. Thus, even if the null hypothesis is not rejected, one cannot claim that it is true.

To demonstrate the fact that test procedures do not provide (statistical) proof that the null hypothesis is true, consider the test procedure which rejects the null hypothesis $H_0 : \theta = \theta_0$ if θ_0 does not belong in the CI for θ. Since a CI contains a range values, all of which are plausible (given the data set) values for the true parameter θ, it is evident that, by not rejecting $H_0 : \theta = \theta_0$, we have not proved that H_0 is true. For example, if $H_0 : \mu = 9.8$ and the 95% CI for μ is $(9.3, 9.9)$, then $H_0 : \mu = 9.8$ is not rejected at level $\alpha = 0.05$; on the other hand $H_0 : \mu = 9.4$ would not have been rejected either, so that, by not rejecting $H_0 : \mu = 9.8$, we have not proved that $H_0 : \mu = 9.8$ is true.

When a null hypothesis is not rejected, the only thing that can be said is that the data does not provide enough evidence against it. (Similarly if an accused is acquitted his/her innocence has not been established.) On the other hand, if the null hypothesis is rejected, one can claim that the alternative has been proved (in the statistical sense) at level of significance α. The level of significance quantifies the *reasonable doubt* we are willing to accept when rejecting a null hypothesis.

The above discussion leads to the following **rule for designating H_0 and H_a**: *Alternative is the hypothesis the investigator wants to claim as true, or wants evidence for.*

Example 9.2.1. Consider the following two scenaria.

a) A trucking firm suspects the claim, made by a tire manufacturer, that certain tires last, on average, at least 28,000 miles. The firm intends to initiate a study, involving data collection and hypothesis testing, to confirm this suspicion.

b) A tire manufacturing firm wants to claim that certain tires last, on average, at least 28,000 miles. The firm intends to initiate a study, involving data collection and hypothesis testing, to support the validity of the claim.

How should the null and alternative hypotheses be specified in each case?

Solution. a) The trucking firm wants evidence that the claim is wrong, i.e. that $\mu < 28,000$. The statement, in support of which, the investigator (the trucking firm in this case) wants evidence, is designated as H_a. The complementary statement is designated as H_0. Thus the hypotheses to be tested are $H_0 : \mu \geq 28,000$ vs $H_a : \mu < 28,000$.

b) Here the manufacturing firm wants evidence in support of the claim that is about to be made. Reasoning as in part a), the hypotheses to be tested are $H_0 : \mu \leq 28,000$ vs $H_a : \mu > 28,000$.

CONVENTION: For reasons that have to do with convenience and simplicity in notation, the null hypothesis will be stated as equality. Thus, the null hypothesis will be stated as

$$H_0 : \theta = \theta_0, \tag{9.2.1}$$

regardless of whether the alternative hypothesis is of the form

$$H_a : \theta < \theta_0, \quad \text{or} \quad H_a : \theta > \theta_0, \quad \text{or} \quad H_a : \theta \neq \theta_0.$$

9.2.1 Exercises

1. In 10-mph crash tests, 25% of a certain type of automobiles sustain no visible damage. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the probability that a car with the modified bumper design sustains no visible damage in a 10-mph crash test. Because of cost considerations, the new design will not be implemented unless there is significant evidence that the new bumper design improves the crash test results. Formulate this decision problem as a hypothesis testing problem by

   (a) stating the null and alternative hypotheses, and

   (b) stating what action should be taken if the null hypothesis is rejected.

2. An appliance manufacturer is considering the purchase of a new machine for cutting sheet metal parts. If μ_0 is the average number of metal parts cut per hour by her old machine

and μ is the corresponding average for the new machine, the manufacturer wants to test the null hypothesis $H_0 : \mu = \mu_0$ against a suitable alternative. What should H_a be if

(a) She does not want to buy the new machine unless there is evidence it is more productive than the old one.

(b) She wants to buy the new machine (which has some other nice features) unless there is evidence it is less productive than the old one.

(c) For each of the two cases above state whether she should buy the new machine if the null hypothesis gets rejected.

9.3 Setting up a Test Procedure

9.3.1 Test Statistics and Rejection Regions

A test procedure is specified in terms of a **test statistic**, and a **rejection rule** (RR).

For testing a null hypothesis H_0 about a parameter θ, the test statistic can be (and typically is) based on a point estimator $\widehat{\theta}$ of θ. (Other types of test statistics will be seen in Section **??**.)

The rejection rule prescribes when H_0 is to be rejected. Basically, H_0 is rejected when the test statistic takes a value which is deemed very unlikely if H_0 were true. Clearly, the values deemed very unlikely under H_0 are determined on the basis of the *null distribution* of the test statistic, i.e. its distribution under the null hypothesis. The set, or region, of unlikely values of the test statistic is called the *rejection region.* Thus a rejection rule is specified in terms of a rejection region: if the value of the test statistic falls in the rejection region, the null hypothesis is rejected.

Example 9.3.1. Consider the two hypothesis testing problems given in Example 9.2.1. To test either of these, a random sample of n is chosen and their lifetimes recorded. Let $\overline{X}$ denote the average lifetime of the n tires. Since $\overline{X}$ is an estimator of μ, it can serve as the test statistic for both problems. Consider the first testing problem, which can be written as

$$H_0 : \mu = 28,000 \;\text{ vs }\; H_a : \mu < 28,000$$

according to the convention for stating null hypotheses given in (9.2.1). Under the null hypothesis (which really is $H_0 : \mu \geq 28,000$) the test statistic is unlikely to take values which are much smaller than 28,000. Thus, the rejection region is of the form $\overline{X} < C$, for some constant. For example, the rejection region can be $\overline{X} < 27,000$.

Using the same reasoning, the rejection region for the second testing problem, which can be written as

$$H_0 : \mu = 28,000 \;\text{ vs }\; H_a : \mu > 28,000$$

according to the convention, is of the form $\overline{X} > C$, for some constant.

Example 9.3.2. A coal mining company suspects that certain detonators used with explosives do not meet the requirement that at least 90% will ignite. Using the convention (9.2.1), the relevant hypothesis testing problem for confirming this suspicion is

$$H_0 : p = 0.9, \;\text{ vs }\; H_a : p < 0.9,$$

where p is the population proportion of detonators that ignite. The data consist of the number, X, among n randomly selected detonators, which ignite. In this case the test statistic can be either X, or the estimator $\hat{p} = \frac{X}{n}$ of p. If the null hypothesis (which really is $H_0 : p \geq 0.9$) is true, the test statistic is unlikely to take values which are much smaller than 0.9. Thus, the rejection region is of the form $\widehat{p} < C$, for some constant. For example, the RR can be $\widehat{p} < 0.8$.

9.3.2 Type I and Type II Errors

Because of sampling variability, it is possible that the test statistic will take a value in the rejection region when the null hypothesis is true. Similarly, it is possible that test statistic will not take a value in the rejection region when the alternative hypothesis is true. For example, in the testing problem

$$H_0 : \mu = 28,000 \;\text{ vs }\; H_a : \mu < 28,000$$

with rejection region $\overline{X} < 27,000$, as mentioned in Example 9.3.1, it is possible that $\overline{X} < 27,000$ (which would lead the investigator to reject H_0) even though the true value of the mean is $\mu = 28,000$, and, similarly, it is possible that $\overline{X} > 27,000$ (which would lead the investigator not to reject H_0) even though the true value of the mean is $\mu = 26,000$. **Type I error** is committed when the null hypothesis is rejected when in fact it is true. **Type II error** is committed when the null hypothesis is not rejected when in fact it is false. These two types of error are illustrated in the following table.

		Truth	
		H_0	H_a
Outcome of test	H_0	Correct decision	Type II
	H_a	Type I	Correct decision

In order to decide which of the possible rejection rules to adopt, we need to consider two basic properties of a rejection rule. These are the *probability of rejecting H_0, when H_0 is in fact true*, or the *probability of type I error*, and the *probability of not rejecting H_0, when H_0 is in fact false*, or the *probability of type II error.* The next example illustrates the calculation of these probabilities in the context of testing a hypothesis for a binomial proportion.

Example 9.3.3. Consider the testing problem of Example 9.3.2, namely

$$H_0 : p = 0.9 \text{ vs } H_a : p < 0.9,$$

suppose that $n = 20$ detonators are to be tested, and let X denote the number of those that function correctly. For the two RR

$$\text{Rule 1: } X \leq 16, \quad \text{Rule 2: } X \leq 17,$$

calculate the probability of type I error, and also the probability of type II error when the true value of p is 0.8.

Solution. Note first that the RR can equivalently be stated in terms of $\hat{p}$. For example, $\hat{p} \leq 16/20$ is equivalent to Rule 1. However, it is slightly more convenient to state the RR in terms of X, as done above.

The probability of committing type I error with Rule 1 is

$$\begin{aligned} P(\text{type I error}) &= P(H_0 \text{ is rejected when it is true}) \\ &= P(X \leq 16 \mid p = 0.9, n = 20) = 0.133. \end{aligned}$$

(The probability of 0.133 is found from the binomial tables.) Thus, there is a 13.3% chance that H_0 will be rejected even though it is true. Now suppose that the true value of p is 0.8, so H_a is true. Then the probability of type II error is

$$\begin{aligned} P(\text{type II error when } p = 0.8) &= P(H_0 \text{ is not rejected when } p = 0.8) \\ &= P(X > 16 \mid p = 0.8, n = 20) = 1 - 0.589 = 0.411. \end{aligned}$$

Thus, there is a 41.1% that H_0 will not be rejected even though it is false (true value of $p = 0.8$).

Using the same calculations for Rule 2 we obtain

$$\begin{aligned} P(\text{type I error}) &= P(X \leq 17 \mid p = 0.9, n = 20) = 0.323, \text{ and} \\ P(\text{type II error when } p = 0.8) &= P(X > 17 \mid p = 0.8, n = 20) = 1 - 0.794 = 0.206. \end{aligned}$$

The calculations in the above example illustrate the very important fact that it is not possible to control, or reduce, the probabilities of both types of errors with the same sample size. For example, changing the RR from $X \leq 16$ to $X \leq 17$ in Example 9.3.3, the probability of type I error increased from 0.133 to 0.323, while the probability of type II error decreased from 0.411 to 0.206. In general, the inability to reduce the probability of both types of errors is due to the fact that the two events involved are complementary. Thus, if we shrink the rejection region (thereby decreasing the probability of type I error), we expand the complement of the rejection region (thereby increasing the probability of type II error).

Remark: Though the events involved in the calculation of the probabilities of type I and type II errors are complementary, the two probabilities do not sum to one. This is because they are evaluated at different values of the parameter. For example, for Rule 1 of Example 9.3.3, the events involved are $X \leq 16$ and $X > 16$, for the type I and type II error probabilities, respectively. However, for the calculation of the probability of type I error it is assumed that $X \sim Bin(n = 20, p = 0.9)$, whereas for the calculation of the probability of type II error it is assumed that $X \sim Bin(n = 20, p = 0.8)$.

We close this subsection with one more definition.

Definition 9.3.1. *The probability of committing a type II error (i.e. of not rejecting H_0), when the true value, θ_a, of the parameter θ belongs in the domain of the alternative hypothesis H_a, is denoted by $\beta(\theta_a)$. That is,*

$$\beta(\theta_a) = P(\textit{type II error when } \theta = \theta_a), \tag{9.3.1}$$

where θ_a is a value in the domain of H_a. One minus the probability of type II error, evaluated at θ_a, i.e. the probability of rejecting H_0 when it is false, is called the **power** *of the test procedure at θ_a. That is*

$$\textit{Power at } \theta_a = 1 - \beta(\theta_a). \tag{9.3.2}$$

The power of a test procedure expresses the ability the procedure to reject the null hypothesis when in it not true.

Example 9.3.4. The power at $p = 0.8$ of the test procedure that correspond to Rule 1 of Example 9.3.3 is

$$\text{Power at } 0.8 = 1 - P(\text{type II error when } p = 0.8) = 1 - 0.411 = 0.589.$$

9.3.3 The Philosophy for Choosing a Rejection Rule

Because we cannot control the probability of both types of error, a rejection rule, and thus a test procedure, is adopted on the basis of the philosophy that **a type I error is far more serious than a type II error**. According to this philosophy, one specifies a level for the probability of type I error that one is willing to tolerate, and adopt the RR that meets the specified tolerance level.

Definition 9.3.2. *The level at which one decides to control the probability of type I error is is called* **level of significance***, and is denoted by* α*.*

Usually α is taken to be $0.1,\ 0.05$ or 0.01.

As the calculations in Example 9.3.3 suggest, a decision on the level of significance, α, leads to the appropriate RR, and, thereby, to the identification of the testing procedure.

Example 9.3.5. A decision to use a level of significance of $\alpha = 0.133$ in the testing problem of Example 9.3.3, leads to the RR is $X \leq 16$. Indeed, there is no other rejection region of the form $X \leq C$ that achieves probability of type I error equal to 0.133.

Note that the discreteness of the binomial distributions does not allow the usual choices of α. For example, no RR of the form $X \leq C$ can have a level of significance of $\alpha = 0.1$, for the testing problem $H_0 : p = 0.9$ vs $H_a : p < 0.9$, based on a sample of size 20.

The next two sections apply the ideas of this section for the construction of test procedures for a population mean, proportion, and median.

9.3.4 Exercises

1. a) When a null hypothesis is rejected, there is risk of committing which type of error?

   b) When a null hypothesis is not rejected, there is risk of committing which type of error?

2. In 10-mph crash tests, 25% of a certain type of automobiles sustain no visible damage. A modified bumper design has been proposed in an effort to increase this percentage. Let p denote the proportion of all cars with this new bumper that sustain no visible damage in 10-mph crash tests. The hypotheses to be tested are $H_0 : p = 0.25$ vs $H_0 : p > 0.25$. The test will based on an experiment involving $n = 20$ independent crashes of car prototypes with the new bumper design. Let X denote the number of crashes resulting in no visible damage, and consider the test procedure that rejects the null hypothesis if $X \geq 8$.
   a) Use the binomial table to find the probability of type I error.
   b) Use the binomial table to find $\beta(0.3)$, i.e. the probability of type II error when the true value of p is 0.3.

9.4 Nonparametric Tests

In this section we will develop test procedures, for testing hypotheses concerning population means, proportions, and medians. One of the procedures for testing for a median extends to testing hypotheses for percentiles. The procedures for testing hypotheses regarding the mean rely on the normal approximation to the distribution of the sample mean. The other procedures can also be used with small sample sizes through software packages, but here will only describe procedures that use the normal approximation. Except for requiring a large sample size, no other requirement is needed for the validity of the present procedures and thus are called *nonparametric tests.*

Because the rules for rejecting a null hypothesis will be derived using the normal approximation, the actual level of significance of these procedures will be approximately equal to the specified level of α. The larger the sample size, the better the approximation.

9.4.1 Nonparametric Tests for a Population Mean

As mentioned in the previous section, specification of the α identifies the RR, and thus the test procedure. The details of doing so, when using the normal approximation, are demonstrated in the following example.

Example 9.4.1. Suppose that a trucking firm suspects the claim that certain tires last at least 28,000 miles. To gain evidence in support of this suspicion, a study is initiated to test

$$H_0 : \mu = 28,000 \text{ vs } H_a : \mu < 28,000.$$

For this purpose, the life-times of a random sample of $n = 40$ tires will be measured.

a) What should the RR be, so that the test procedure will have level of significance of $\alpha = 0.01$?

b) Suppose that the measured life times yield sample mean and sample standard deviation of $\overline{X} = 27,463$ and $S = 1,348$, respectively. With this data, and the RR determined in part a), should H_0 be rejected?

Solution. a) According to the rationale that was used in the previous section, $H_0 : \mu = 28,000$ should be rejected in favor of $H_a : \mu < 28,000$, if $\overline{X}$ is small enough. Thus, the rejection rule for this testing problem should be of the form

$$\overline{X} < C. \tag{9.4.1}$$

The constant C will be determined from the requirement that the level of significance of the test procedure be $\alpha = 0.01$, i.e. the requirement that the probability of rejecting H_0, when H_0 is true, be equal to $\alpha = 0.01$. This requirement is expressed as

$$P(\text{type I error}) = P(\overline{X} < C|\mu = 28,000) = 0.01. \tag{9.4.2}$$

Thus, we have to solve the above equation for C. The exact distribution of $\overline{X}$ when $\mu = 28,000$ is not known. However, since the sample size is large enough, according to the CLT, it can be approximated by a normal distribution with mean 28,000 and variance σ^2/n. Moreover, for large sample sizes, S is a reliable estimator of σ. Thus, under the null hypothesis, i.e. assuming that $\mu = 28,000$,

$$\overline{X} \dot{\sim} N\left(28,000, \frac{S^2}{40}\right).$$

If follows, that the approximate version of the equation (9.4.2) is

$$P\left(Z_{H_0} < \frac{C - 28,000}{S/\sqrt{40}}\right) = \Phi\left(\frac{C - 28,000}{S/\sqrt{40}}\right) = 0.01, \tag{9.4.3}$$

where

$$Z_{H_0} = \frac{\overline{X} - 28,000}{S/\sqrt{40}}. \tag{9.4.4}$$

Solving equation (9.4.3) for C yields

$$C = 28,000 - z_{0.01}\frac{S}{\sqrt{40}}.$$

Thus, the (unique) test procedure with rejection rule of the form (9.4.1) and level of significance $\alpha = 0.01$ is

$$\text{REJECT } H_0 \text{ IF: } \overline{X} < 28,000 - z_{0.01}\frac{S}{\sqrt{40}}.$$

Note that the rejection rule can be equivalently expressed in terms of the test statistic Z_{H_0} given in (9.4.4) as

$$\text{REJECT } H_0 \text{ IF: } Z_{H_0} < -z_{0.01} = -2.33. \tag{9.4.5}$$

b) With the given data, the test statistic Z takes the value $Z = -2.52$. Since $-2.52 < -2.33$, H_0 is rejected according to the rejection rule (9.4.5).

Rejection rules for testing a null hypothesis

$$H_0: \ \mu = \mu_0, \tag{9.4.6}$$

where μ_0 is a given value, against other alternative hypotheses can be derived in a manner similar to that used in the above example. To present these test procedures, let $X_1, \ldots, X_n$ be a random sample from any population with finite variance, σ^2, and such that $n > 30$. Also, set

$$Z_{H_0} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \tag{9.4.7}$$

for the test statistic. Then, the rejection rules for testing the null hypothesis in (9.4.6), at level of significance α, against the different alternatives are:

H_a	RR at level α
$\mu > \mu_0$	$Z_{H_0} \geq z_\alpha$
$\mu < \mu_0$	$Z_{H_0} \leq -z_\alpha$
$\mu \neq \mu_0$	$\lvert Z_{H_0} \rvert \geq z_{\alpha/2}$

In the rare cases where σ is known, we can use it instead of S in the definition of the test statistic in (9.4.7).

Example 9.4.2. A tire company wants to change the tire design. Economically the modification can be justified if the average lifetime with the new design exceeds 20,000 miles. A random sample of $n = 16$ new tires is tested. Assume that the distribution of the lifetimes is $N(\mu, 1500^2)$. The 16 tires yield $\overline{X} = 20,758$. Should the new design be adopted? Test at level of significance $\alpha = 0.01$.

Solution. The purpose of this study is to gain evidence in support of the economic feasibility of the new design. Thus, $H_0 : \mu = 20,000$ is to be tested against $H_a : \mu >$

20, 000. Because the standard deviation, σ, is given, we use it in the formula for the test statistic. With this substitution, the test statistic is

$$Z_{H_0} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{20,758 - 20,000}{1500/\sqrt{16}} = 2.02.$$

The rejection region, at level of significance $\alpha = 0.01$, is $Z_{H_0} > z_{.01} = 2.33$. Since $2.02 \not> 2.33, H_0$ is not rejected, and thus the new design is not adopted.

REMARK: The fact that, in the above example, $\overline{X} = 20,758 > 20,000$ but the testing procedure decides against $H_a : \mu > 20,000$, serves to highlight the fact that test procedures favor the null hypothesis.

9.4.2 Nonparametric Tests for a Population Proportion

In Example 9.3.3 we saw how to specify the RR for testing $H_0 : p = p_0$ vs $H_a : \ p < p_0$, with the use of binomial tables. A similar approach can be used for specifying the RR for testing against other alternative hypotheses. In this, section we will concentrate on specifying the RR for testing hypotheses for a binomial proportion using the normal approximation to the binomial probabilities.

A sample proportion, $\widehat{p}$, is also a sample mean, since it is the average of the Bernoulli random variables $X_1, \ldots, X_n$, where X_i takes the value 1 or 0, as the ith trial is a success or a failure. Moreover, p is a population mean, since it is the expected value of each Bernoulli random variable X_i. Thus, the RRs for testing for a population proportion based on the normal approximation to the binomial distribution, can be considered a special case of the RRs for testing for a population mean, which we saw in the previous subsection. The main difference is that, in this case, the population variance, σ^2, is related to the population mean, p, by

$$\sigma^2 = p(1-p).$$

Since the rejection region of a test procedure is determined from the null distribution of the test statistic, and since a null hypothesis specifies that the true value of p is p_0, it also specifies that the true value of $\sigma = \sqrt{p_0(1-p_0)}$. This means that we can use it instead of S in the formula for the test statistic (9.4.7). Thus, the test statistic we will use is

$$Z_{H_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}. \tag{9.4.8}$$

Provided that $np_0 \geq 5$ and $n(1 - p_0) \geq 5$, the null distribution of this test statistic can be approximated by $N(0, 1)$. Thus, provided the aforementioned conditions on the sample size hold, RRs for the different alternative hypotheses, having a level of significance approximately equal to α, are:

H_a	RR at level α
$p > p_0$	$Z_{H_0} \geq z_\alpha$
$p < p_0$	$Z_{H_0} \leq -z_\alpha$
$p \neq p_0$	$\lvert Z_{H_0}\rvert \geq z_{\alpha/2}$

Example 9.4.3. It is thought that more than 70% of all faults in transmission lines are caused by lightning. To gain evidence in support of this contention a random sample of 200 faults from a large data base yields that 151 of 200 are due to lightning. Test $H_0 : p = 0.7$ vs $H_a : p > 0.7$ at level of significance $\alpha = 0.01$.

Solution. Since $200(0.7) \geq 5$ and $200(0.3) \geq 5$, the null distribution of the test statistic Z, given in (9.4.8), can be approximated by $N(0, 1)$. From the data given we have $\hat{p} = 151/200 = 0.755$, and thus the test statistic takes a value of

$$Z_{H_0} = \frac{\hat{p} - 0.7}{\sqrt{(0.7)(0.3)/200}} = 1.697.$$

The RR, at level of significance $\alpha = 0.01$, is $Z_{H_0} > z_{0.01} = 2.33$. Here $1.697 \not> 2.33$ so H_0 is not rejected.

9.4.3 The Sign Test for the Median and Percentiles

When the sample size is small and the population non-normal we have no procedure for testing $H_0 : \mu = \mu_0$. The *sign test* fills this gap, but it tests for the median instead of the mean. Note that when the distribution is symmetric $\mu = \widetilde{\mu}$; when the distribution is not symmetric the median might be as meaningful as the mean. The test is based on the idea that when $H_0 : \tilde{\mu} = \tilde{\mu}_0$ is true about half of the observations should be greater than $\tilde{\mu}_0$. In particular, we use

$$\text{test statistic} = Y = \text{number of observations} > \tilde{\mu}_0.$$

When $H_0 : \tilde{\mu} = \tilde{\mu}_0$ is true then Y has a binomial distribution with parameters n and $p = .5$. Thus the RR are:

H_a	R.R. at level α
$\tilde{\mu} > \tilde{\mu}_0$	$Y \geq c_1$
$\tilde{\mu} < \tilde{\mu}_0$	$Y \leq c_2$
$\tilde{\mu} \neq \tilde{\mu}_0$	$Y \leq c$ or $Y \geq n - c$

where c_1, c_2 and c are found from the binomial $bin(n, .5)$ table.

The procedure that uses normal approximation **requires** $n \geq 10$. It uses test statistic

$$\text{test statistic} = \frac{Y - .5n}{.5\sqrt{n}} = Z_{H_0}$$

and RR

H_a	R.R. at level α
$\tilde{\mu} > \tilde{\mu}_0$	$Z_{H_0} > z_\alpha$
$\tilde{\mu} < \tilde{\mu}_0$	$Z_{H_0} < -z_\alpha$
$\tilde{\mu} \neq \tilde{\mu}_0$	$\lvert Z_{H_0} \rvert > z_{\alpha/2}$

Example 9.4.4. *It is claimed that the median increase in home owners taxes in a certain county is \$300. A r.sample of 20 homeowners gives the following tax-increase data*

342, 176, 517, 296, 312, 143, 279,
195, 241, 211, 285, 329, 137, 188,
260, 233, 357, 412, 228, 209

Does the data set indicate that the claim is not true at $\alpha = .05$?

Solution: $H_0 : \tilde{\mu} = 300$, $H_0 : \tilde{\mu} \neq 300$. *For this data*

$$Y = \#\ \textit{of observations that are} > 300 = 6.$$

Thus the test statistic is

$$Z_{H_0} = \frac{Y - .5n}{.5\sqrt{n}} = \frac{6 - (.5)(20)}{.5\sqrt{20}} = -1.79.$$

The p-value for the two-sided alternative is $2[1 - \Phi(|Z_{H_0}|)] = 2(.0367) = .0734$ *which is larger than* $\alpha = .05$ *and thus* H_0 *cannot be rejected.*

9.4.4 The Signed-Rank Test for the Median of a Symmetric Population

Let $X_1, \ldots, X_n$ be a sample from a continuous population, and consider testing the hypothesis $H_0 : \widetilde{\mu} = \mu_0$ about the median. The sign test is based only on the signs of the difference $X_i - \mu_0$, ignoring their magnitudes. Thus, though it is useful in that it does not require symmetry or any other distributional assumption, when symmetry can be assumed other test procedures exist which have improved power.

Here we will describe the *signed-rank test* which is valid under the assumption that the population distribution is symmetric so that it can be used for hypotheses either about the mean or the median. When the population distribution is normal, the signed-rank test is somewhat less powerful than the t-test, but it can be much more powerful for other symmetric distributions. These include the logistic, Laplace, t, uniform, Cauchy, which are also unimodal.

The signed-rank test statistic for testing $H_0 : \tilde{\mu} = \mu_0$, or $H_0 : \mu = \mu_0$ based on a sample $X_1, \ldots, X_n$ from a symmetric population is calculated as follows:

1. Rank the absolute differences $|X_1 - \mu_0|, \ldots, |X_n - \mu_0|$ from smallest to largest. Let R_i denote the rank of $|X_i - \mu_0|$.

2. Assign to R_i the sign of $X_i - \mu_0$, forming thus signed ranks.

3. Let S_+ be the sum of the ranks R_i with positive sign, i.e. the sum of the positive signed ranks.

To appreciate the relevance of S_+ for testing $H_0 : \widetilde{\mu} = \mu_0$, note that if the true population mean or median is larger than μ_0, then more differences $X_i - \mu_0$ will tend to be positive and also the positive differences will tend to be larger than the absolute value of the negative differences. Thus, S_+ will tend to take larger values if the alternative $H_a : \widetilde{\mu} > \mu_0$ is true. Similarly, S_+ will tend to take smaller values if the alternative $H_a : \widetilde{\mu} < \mu_0$ is true.

The exact null distribution of S_+ does not depend on the distribution of the X_i (provided it is continuous and symmetric). This exact null distribution is known. Thus the signed-rank test can be carried out even with small sample sizes, with the use of tables, or a software package. Here we will only present the test procedure based on the standardized test statistic.

It can be shown that, if H_0 holds,

$$\mu_{S_+} = \frac{n(n+1)}{4}, \quad \sigma^2_{S_+} = \frac{n(n+1)(2n+1)}{24}.$$

When $n > 10$, the null distribution of S_+ is approximately normal with the above mean and variance. Thus, test procedures can be based on

$$Z_{H_0} = \frac{S_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}, \tag{9.4.9}$$

and the usual rejection regions of a Z-test. Namely,

H_a	RR at level α
$\widetilde{\mu} > \mu_0$	$Z_{H_0} > z_\alpha$
$\widetilde{\mu} < \mu_0$	$Z_{H_0} < -z_\alpha$
$\widetilde{\mu} \neq \mu_0$	$\lvert Z_{H_0}\rvert > z_{\alpha/2}$

Example 9.4.5. As part of quality control, data are periodically collected from a fill and seal machine to test if the average fill weight deviates from the target value of 965 grams. It is believed that the distribution of the fill weights is symmetric about its true value. Use the signed-rank test to test $H_0 : \widetilde{\mu} = 965$ vs $H_a : \widetilde{\mu} \neq 965$.

Solution: The sample of size $n = 12$ is large enough to apply the normal approximation. Here $n(n+1)/4 = 39$, $n(n+1)(2n+1)/24 = 162.5$.

$X_i - 965$	4.5	8.7	5.9	2.6	-4.6	9.6	-13.0	1.0	-11.2	0.2	-0.3	11.0
Signed Rank	5	8	7	4	-6	9	-12	3	-11	1	-2	10

The sum of the positively signed ranks is $5 + 8 + 7 + 4 + 9 + 3 + 1 + 10 = 47$. Thus,

$$Z_{H_0} = \frac{47 - 39}{\sqrt{162.5}} = 0.63.$$

Thus H_0 is not rejected at any of the common levels of significance.

9.4.5 Exercises

1. In the context of Exercise 9.2.1,2, suppose that the the old machine achieves $\mu_0 = 9.5$ cuts per hour. The manufacturer decides to test $H_0 : \mu = 9.5$ against $H_a : \mu \geq 9.5$ at the $\alpha = 0.05$ level. She gets the machine manufacturer to lend her a new machine and she measures the number of cuts made by the machine in fifty (50) randomly selected one hour time periods. The summary statistics for this random sample are $\overline{X} = 9.8$ and $S = 1.095$. Would you reject H_0?

2. Consider the car crash experiment of Exercise 2 above, where $H_0 : p = 0.25$ is tested against $H_0 : p > 0.25$, but suppose that $n = 50$ cars are crashed, and the null hypothesis is rejected if the number of crashes, X, resulting in no visible damage is $X \geq 15$.
   a) Use the normal approximation to the binomial probabilities to find the probability of type I error.
   b) Use the normal approximation to the binomial probabilities to find the probability of type II error when the true value of p is 0.3.

3. Scientists have labeled benzene as a possible cancer-causing agent. Studies have shown that people who work with benzene more than 5 years have 20 times the incidence of leukemia than the general population. As a result, the federal government has lowered the maximum allowable level of benzene in the workplace from 10 parts per million (ppm) to 1 ppm. Suppose a steel manufacturing plant, which exposes its workers to benzene daily, is under investigation by the Occupational Safety and Health Administration (OSHA). Thirty six air samples, collected over a period of 1.5 months and examined for benzene content, yielded the following summary statistics: $\overline{Y} = 2.1$ ppm, $S = 4.1$ ppm.

   (a) Is the steel manufacturing plant in violation of the new government standards? State the appropriate hypothesis.

   (b) Give the test statistic for testing the hypothesis in (a), and specify the rejection region for a test at level α. What assumptions are required for the procedure to be valid?

   (c) Carry out the test at $\alpha = .05$.

4. A manufacturer of automatic washers provides a particular model in one of three colors, white, almond, or blue. Of the first 1000 washers sold, it is noted that 400 of the washers were of color A. The manufacturer wants to know if this is evidence that more than a third of consumers prefer white washers.

   (a) Set up the appropriate null and alternative hypothesizes for the manufacturer.

   (b) Carry out a test of your null hypothesis versus your alternative hypothesis at the $\alpha = 0.05$ level. Is there evidence that more than one third of consumers prefer white washers?

5. A food processing company is considering the marketing of a new product. The marketing would be profitable if at least 20% of the consumers would be willing to try this new product. Among 42 randomly chosen consumers 9 said that they would purchase the new product and give it a try.

   (a) Set up the appropriate null and alternative hypothesizes.

(b) Carry out a test of your null hypothesis versus your alternative hypothesis at the $\alpha = 0.01$ level. Is there evidence that more than 20% of consumers are willing to try the new product?

6. The output, in hundreds of pounds, of a production facility in 16 randomly selected one-hour periods is

21.72	88.98	763.32	752.74	656.30	810.70	103.81	583.91
268.97	561.17	252.18	82.00	579.27	68.32	386.56	242.41

Is there sufficient evidence to reject the hypothesis that the median production is 200?

(a) State the null and alternative hypotheses.

(b) Is a sign test or signed rank test appropriate in this case?

(c) State any assumptions needed for conducting the test you specified in part 2.

(d) Carry out the test you specified in part 2, at level $\alpha = 0.05$, and state your conclusion.

9.5 Tests Based on the t-Distribution

As mentioned in Section 8.4, when sampling from normal populations, an estimator $\widehat{\theta}$ of some parameter θ often satisfies, for all sample sizes n,

$$\frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\hat{\theta}}} \sim t_\nu, \quad \text{where } \widehat{\sigma}_{\hat{\theta}} \text{ is the estimated s.e.,} \tag{9.5.1}$$

and t_ν stands for "t-distribution with ν degrees of freedom". In such cases, the TS for testing $H_0 : \theta = \theta_0$ is

$$T_{H_0} = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\hat{\theta}}} \sim t_\nu.$$

According to (9.5.1), if the null hypothesis is true, T_{H_0} has a t_ν distribution. That is

$$T_{H_0} = \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}_{\hat{\theta}}} \sim t_\nu, \quad \text{provided } H_0 \text{ holds.} \tag{9.5.2}$$

This leads to the following test procedures

H_a	RR at level α
$\theta > \theta_0$	$T_{H_0} > t_{\alpha,\nu}$
$\theta < \theta_0$	$T_{H_0} < -t_{\alpha,\nu}$
$\theta \neq \theta_0$	$\lvert T_{H_0}\rvert > t_{\alpha/2,\nu}$

We note that, if the assumption that the sampled population is normal is correct, the level of significance of the above test procedures is exactly α for all sample sizes n.

In the next subsections we will discuss such tests for the population mean, the regression slope, and the regression line.

9.5.1 Tests About a Normal Mean

Here we will consider testing hypotheses for a population mean, when we know that the population distribution is $N(\mu, \sigma^2)$, where σ^2 is unknown (as, of course, is μ). Thus, we assume that $X_1, \ldots, X_n$ is a sample from such a normal population. To test $H_0: \ \mu = \mu_0$ against the various alternative hypotheses, we will still use (9.4.7) as test statistic, but now we will name it T_{H_0}, instead of Z_{H_0}, because its (exact) null distribution is t_{n-1}. That is, the test statistic we will use, and its null distribution are

$$T_{H_0} = \frac{\overline{X} - \mu_0}{S\sqrt{n}} \sim t_{n-1} \ . \tag{9.5.3}$$

The RRs for the various alternative hypotheses are

H_a	RR at level α
$\mu > \mu_0$	$T_{H_0} > t_{\alpha,n-1}$
$\mu < \mu_0$	$T_{H_0} < -t_{\alpha,n-1}$
$\mu \neq \mu_0$	$\lvert T_{H_0}\rvert > t_{\alpha/2,n-1}$

Example 9.5.1. The maximum acceptable level of exposure to microwave radiation in US is an average of 10 microwatts per $(\text{cm})^2$. It is suspected that a large television transmitter may be pushing the average level of radiation above the safe limit. A random sample of $n = 25$ gives $\overline{X} = 10.3$, and $S = 2.0$. To gain evidence in support of the suspicion we test $H_0: \mu = 10$ vs $H_a: \mu > 10$ at $\alpha = 0.1$.

Solution. Test statistic is $T_{H_0} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = .75$ and the RR is $T_{H_0} > t_{.1,24} = 1.318$. Here $0.75 \not> 1.318$, so H_0 is not rejected.

9.5.2 Tests about the Regression Slope

According to part 3 of Proposition 8.4.1, if the null hypothesis $H_0: \beta_1 = \beta_{1,0}$ is true, then

$$T_{H_0} = \frac{\hat{\beta}_1 - \beta_{1,0}}{S_{\hat{\beta}_1}}, \tag{9.5.4}$$

where $\widehat{\beta}_1$ was last given in (8.4.7), and $S_{\hat{\beta}_1} = \widehat{\sigma}_{\hat{\beta}_1}$ is given in the aforementioned proposition, has the t-distribution with $n-2$ df. We will use T_{H_0} as the test statistic for testing hypotheses about the slope. The rejection rules are summarized below:

H_a	R.R. at level α
$\beta_1 > \beta_{1,0}$	$T_{H_0} > t_{\alpha,n-2}$
$\beta_1 < \beta_{1,0}$	$T_{H_0} < t_{\alpha,n-2}$
$\beta_1 \neq \beta_{1,0}$	$\lvert T_{H_0}\rvert > t_{\alpha/2,n-2}$.

Testing the hypothesis $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ is the most common testing problem in the context of the simple liner regression model,and it called the **model utility test**. The terminology is justified by the fact that, if the null hypothesis $H_0 : \beta_1 = 0$ is true, then the regression model has no utility, in the sense that the X variable has no predictive value for the Y variable. Clearly, the model utility test is carried out as described above with 0 replacing $\beta_{1,0}$ (so the R.R. at level α is $|T_{H_0}| > t_{\alpha/2,n-2}$, where $T_{H_0} = \dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$).

9.5.3 Tests about the Regression Line

According to part 3 of Proposition 8.4.2, if the null hypothesis $H_0 : \mu_{Y|X=x} = \mu_{Y|X=x,0}$, where $\mu_{Y|X=x,0}$ is a specified value, is true, then

$$T_{H_0} = \frac{\widehat{\mu}_{Y|X=x} - \mu_{Y|X=x,0}}{S_{\hat{\mu}_{Y|X=x}}}, \tag{9.5.5}$$

where $\widehat{\mu}_{Y|X=x}$ was last given in (8.4.7), and $S_{\hat{\mu}_{Y|X=x}} = \widehat{\sigma}_{\hat{\mu}_{Y|X=x}}$ is given in the aforementioned proposition, has the t-distribution with $n-2$ df. We will use T_{H_0} as the test statistic for testing hypotheses about the regression line at $X = x$. The rejection rules are summarized below:

H_a	R.R. at level α
$\mu_{Y|X=x} > \mu_{Y|X=x,0}$	$T_{H_0} > t_{\alpha,n-2}$
$\mu_{Y|X=x} < \mu_{Y|X=x,0}$	$T_{H_0} < t_{\alpha,n-2}$
$\mu_{Y|X=x} \neq \mu_{Y|X=x,0}$	$\lvert T_{H_0}\rvert > t_{\alpha/2,n-2}$.

9.5.4 Exercises

1. A tire company wants to claim that the average lifetime, (μ), of a brand of tires exceeds 35 thousand miles. A simple random sample of 10 tires are tested on a test wheel that

simulates normal road condition. The sample average lifetime is 41 (in thousands of miles), and the sample standard deviation is $S = 3.59$.

(a) State the null and alternative hypotheses.

(b) Is a Z-test or a t-test appropriate in this case?

(c) State any assumptions that are needed for the validity of the test you selected in b).

(d) Carry out the test you specified in part b), and state your conclusion.

2. To investigate the corrosion-resistance properties of a certain type of steel conduit, 16 specimens are buried in soil for a 2-year period. The maximum penetration (in mils) for each specimen is then measured, yielding a sample average penetration of $\overline{X} = 52.7$ and a sample standard deviation of $S = 4.8$. The conduits will be used unless it can be demonstrated conclusively that the (population) mean penetration exceeds 50 mils.

    (a) State the null and alternative hypotheses.

    (b) Is a Z-test or a t-test appropriate in this case?

    (c) State any assumptions needed for conducting the test you specified in part 2.

    (d) Carry out the test you specified in part 2, at level $\alpha = 0.1$, and state your conclusion.

3. An experiment examined the effect of temperature on the strength of new concrete. After curing for several days at 20^oC, specimens were exposed to temperatures of -10^oC, -5^oC, 0^oC, 10^oC, or 20^oC for 28 days, at which time their strengths were determined. The results are listed in the table below.

Exposure Temperature	28 Day Strength		
(^oC)	(MPa)		
-10	56.53	65.33	65.13
-5	62.53	61.33	66.13
0	61.53	66.33	65.13
10	71.53	67.33	65.13
20	77.53	69.33	61.13

a) Fit a simple linear model for estimating the regression function of Y =28 day strength, on X =exposure temperature. Give the estimated regression line.

b) Because concrete is used in structures located in cold waters, there is concern that the decrease of temperature would weaken the concrete. What are the null and alternative hypotheses that should be tested to gain evidence in support of this concern?

c) Carry out the test of part (b) using $\alpha = .05$. Is there evidence that decreasing the exposure temperature weaken the concrete?

4. Consider the 10 data points on on conductivity (μS/cm) measurements of surface water (X), and water in the sediment at the bank of a river (Y), given in Exercise 8.4.4,7. Assume that the regression function of Y on X is linear.
   The summary statistics are $\sum x_i = 3728$, $\sum y_i = 5421$, $\sum x_i^2 = 1,816,016$, $\sum y_i^2 = 3,343,359$, $\sum x_i y_i = 2,418,968$.
   a) Conduct the model utility test, i.e. test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$, at level $\alpha = 0.05$.
   b) Test, at $\alpha = 0.05$ the null hypothesis $H_0 : \mu_{Y|X=300} = 515$ vs $H_a : \mu_{Y|X=300} > 515$.

9.6 P-Values

Reporting the outcome of a test procedure as a rejection or not of a null hypothesis H_0, does not convey the full information contained in the data regarding the evidence against H_0. This is because such an outcome depends on the chosen α-level. For example, in Example 9.4.3 H_0 was not rejected at $\alpha = 0.01$. The value of the test statistic, however, was $Z = 1.697$, and thus, had we chosen $\alpha = 0.05$, H_0 would have been rejected.

Instead of reporting only whether or not H_0 is rejected at the chosen α-level, we can be more informative by reporting the so-called *p-value.*

Definition 9.6.1. P-value *is the smallest level of significance at which H_0 would be rejected for a given data set.*

The p-value captures the strength of the evidence against H_0: *The smaller the p-value, the stronger the evidence against H_0.* The p-value determines the outcome of the test:

If p-value$\leq \alpha \Rightarrow$ reject H_0 at level α

9.6.1 P-value for a Z-test

. Let Z_{H_0} denote the test statistic of any one of the Z-tests. Then the p-value is given by

$$P - \text{value} = \begin{cases} 1 - \Phi(Z_{H_0}) & \text{for upper-tailed test} \\ \Phi(Z_{H_0}) & \text{for lower-tailed test} \\ 2\big[1 - \Phi(|Z_{H_0}|)\big] & \text{for two-tailed test} \end{cases}$$

Example 9.6.1. (Example 9.4.3, continued) The test statistic for $H_0 : p = 0.7$ vs $H_a : p > 0.7$ is $Z_{H_0} = 1.697$. Thus the p-value is $1 - \Phi(1.697) \simeq 1 - \Phi(1.7) = 0.0446$

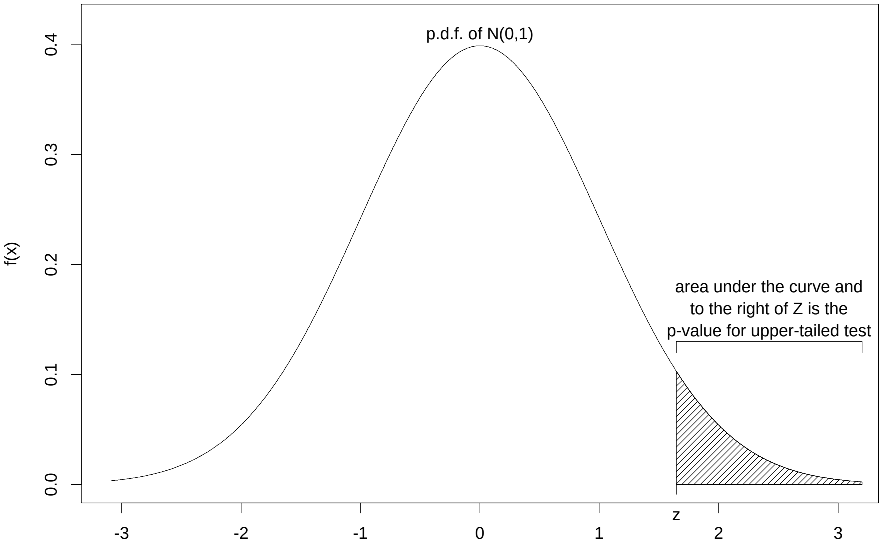


Figure 9.1: Illustration of the p-Value for a Right Tailed Z-Test

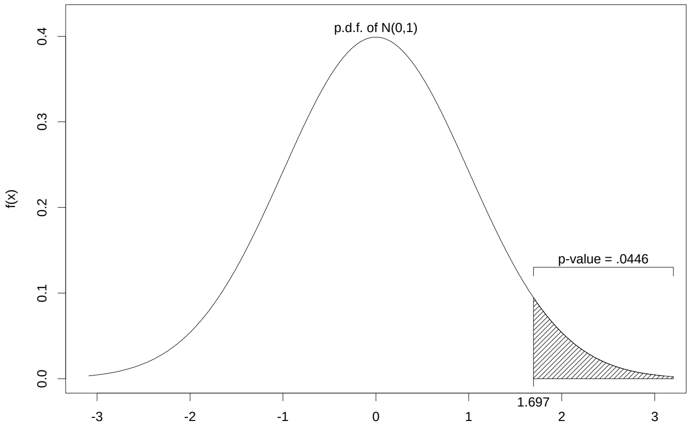


Figure 9.2: Right Tail p-Value Corresponding to $Z_{H_0} = 1.697$.

9.6.2 P-value for a t-test

. This is determined in exactly the same way except that you use test statistic T_{H_0} instead of Z_{H_0} and the t-table instead of the Z-table.

Example 9.6.2. (Example 9.5.1 continues) The test statistic for $H_0 : \mu = 10$ vs $H_a : \mu > 10$ (right tailed test) is $T_{H_0} = 0.75$. As the figure below illustrates, the p-value in this case is the probability with which, a random variable having the t_{24} distribution, takes a value larger than 0.75.

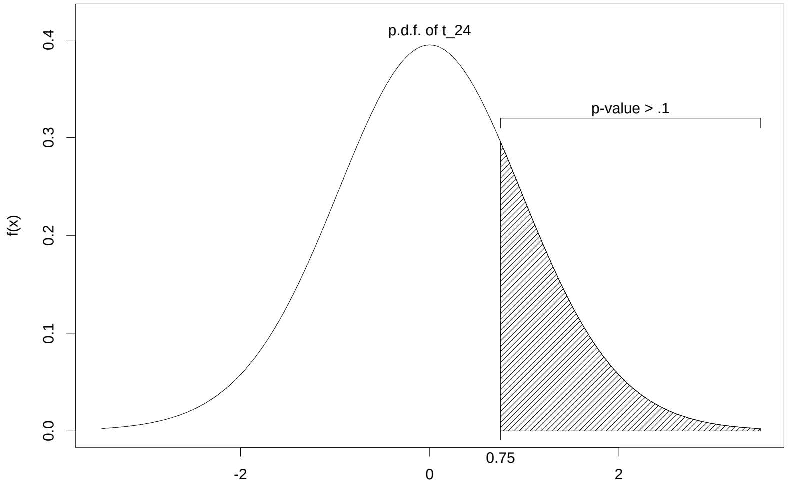


Figure 9.3: Right Tail p-Value Corresponding to $T_{H_0} = 0.75$.

The value 0.75 does not appear in the t-table (these tables are not detailed enough), but we can say that the p-value is larger than 0.1. Thus the null hypothesis would not be rejected at any of the common levels of significance.

9.6.3 Exercises

1. Consider the testing problem and the data given in Exercise 9.4.5,3, and give the (approximate) p-value of the test.

2. Consider the testing problem and the data given in Exercise 9.4.5,4, and give the (approximate) p-value of the test.

9.7 Precision in Hypothesis Testing

Precision in hypothesis testing is quantified by the power of the test procedure. Recall that the power of a test procedure is the probability of rejecting the null hypothesis when the alternative is true, or one minus the probability of type II error. Thus, the desired level of power, at a given alternative value of the parameter, is specified, and the sample size, which is needed for achieving the desired level of power, is determined.

Precision considerations in testing arise naturally in cases like that of Example 9.4.2. There we saw that $\overline{X} = 20,758$, but the null hypothesis $H_0 : \mu = 20,000$ was not rejected, even though the point estimate suggests that the alternative, $H_a : \mu > 20,000$, is true. While this is a manifestation of the fact that test procedures favor the null hypothesis, situations like this raise questions regarding the performance characteristics of the test procedure. For example, in the context of the aforementioned example, it would be of interest to know what is the the power of the procedure when the true value of the population mean is $21,000$.

In this section we will demonstrate the calculation of the power, at a given alternative value of the parameter. It will be seen that the power increases with the sample size. (Equivalently, the probability of type II error decreases as the sample size increases.) The formulas for the power of a test procedure can then be used to determine the sample size which is needed for achieving a prescribed level of precision. Such sample size determinations will be demonstrated for testing for a normal mean and for a proportion.

9.7.1 Precision in testing for a normal mean

We will only discuss the case where the variance is assumed known, as this allows a clean solution to the problem of sample size determination. The case where σ is unknown will be discussed briefly at the end of this subsection.

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, where σ^2 is known, and suppose we want to test $H_0 : \mu = \mu_0$ vs some alternative hypothesis. The null distribution of the Z_{H_0} test statistic, given in relation (9.4.7) but with the known true value of σ replacing S, is exactly $N(0, 1)$. That is,

$$Z_{H_0} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1). \tag{9.7.1}$$

The RRs, for the different alternative hypotheses, are as given in Subsection 9.4.1, but with Z_{H_0} given in (9.7.1). Determining the sample size necessary for achieving a prescribed level of precision for these test procedures, requires formulas for the power of these procedures, or RRs. The following example demonstrates the calculations leading to such a formula for one of the RRs.

Example 9.7.1. Consider the setting of Example 9.4.2. Thus, a tire company wants to change the tire design, but the modification can be justified economically only if the average lifetime of tires with the new design exceeds 20,000 miles. It is known that the population distribution of lifetimes is normal with $\sigma = 1,500$. The lifetimes of a random sample of $n = 16$ new tires are to be measured. The project manager, who is keen on implementing the new design, wants to know the probability of rejecting the null hypothesis $H_0 : \mu = 20,000$, at level of significance $\alpha = 0.01$, when the mean lifetime is $\mu = 21,000$. Calculate this probability under the assumption that the lifetimes are normally distributed with $\sigma = 1500$.

Solution. In this case, the alternative hypothesis is $H_a : \mu > 20,000$, so that the appropriate rejection region is $Z_{H_0} > z_\alpha$, where Z_{H_0} is the test statistic given in (9.7.1). Thus,

$$\begin{aligned}
\beta(21,000) &= P(\text{type II error} \mid \mu = 21,000) \;=\; P(Z_{H_0} < z_\alpha | \mu = 21,000) \\
&= P\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \;\middle|\; \mu = 21,000\right) \\
&= P\left(\overline{X} < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \;\middle|\; \mu = 21,000\right) \\
&= P\left(\frac{\overline{X} - 21,000}{\sigma/\sqrt{n}} < \frac{\mu_0 - 21,000}{\sigma/\sqrt{n}} + z_\alpha \;\middle|\mu = 21,000\right) \\
&= \Phi\left(\frac{\mu_0 - 21,000}{\sigma/\sqrt{n}} + z_\alpha\right). \qquad (9.7.2)
\end{aligned}$$

Here, $\mu_0 = 20,000$, $\sigma = 1,500$, and $\alpha - 0.01$, so that $z_\alpha = 2.33$. Thus, we obtain that

$$\beta(21,000) \;=\; \Phi(-0.34) = 0.3669,$$

so that, the power of the test procedure at $\mu = 21,000$ is $1 - 0.3669 = 0.6331$.

The formulas for the probability of type II error, when testing $H_0 : \mu = \mu_0$ against the different alternative hypotheses, are:

H_a	$\beta(\mu_a)$ for a Level α test
$\mu > \mu_0$	$\Phi\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_\alpha\right)$
$\mu < \mu_0$	$1 - \Phi\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_\alpha\right)$
$\mu \neq \mu_0$	$\Phi\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_{\alpha/2}\right) - \Phi\left(\frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_{\alpha/2}\right)$

Note that the first of the above formulas was essentially derived in Example 9.7.1. (This is seen by replacing 21,000 by μ_a in the relation (9.7.2).) The other formulas can be derived by similar calculations.

Having formulas for the probability of type II error, means that we can determine the sample size needed for achieving a prescribed level of precision in testing. The way of doing so is demonstrated in the following example.

Example 9.7.2. In the setting of Example 9.7.1, so that the population distribution of life times of tires is normal with $\sigma = 1,500$, find the sample size needed to achieve a probability of type II error, at $\mu_a = 21,000$, of 0.1.

Solution. The alternative hypothesis here is $H_a : \mu > 20,000$, and we want to find the sample size for which $\beta(21,000) = 0.1$. According to the formula for the probability of type II error, for the procedure that tests against an alternative of the form $H_a : \mu > \mu_0$, we have to solve for n the equation

$$\Phi\left(\frac{\mu_0 - 21,000}{1,500/\sqrt{n}} + z_\alpha\right) = 0.1 \ .$$

Recall that the z value that satisfies $\Phi(z) = 0.1$, is denoted by $z_{0.9}$. Thus, the above equation is equivalent to

$$\frac{\mu_0 - 21,000}{1,500/\sqrt{n}} + z_\alpha = z_{0.9}.$$

Since, by the symmetry (about zero) of the standard normal distribution, we have $z_{0.9} = -z_{0.1}$, the solution to the above equation is

$$n = \left[\frac{1,500(z_\alpha + z_{0.1})}{\mu_0 - 21,000}\right]^2 = \left[\frac{1500(2.33 + 1.28)}{20,000 - 21,000}\right]^2 = (5.42)^2 = 29.32,$$

which is rounded up (as usual) to $n = 30$.

The formulas for the determination of the sample size needed to achieve a prescribed level, β, of probability of type II error, when testing $H_0 : \mu = \mu_0$ against the different alternative hypotheses, at level α, are:

H_a	Sample Size Needed for $\beta(\mu_a)$ to Equal β
$\mu > \mu_0$	$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu_a}\right]^2$
$\mu < \mu_0$	$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu_a}\right]^2$
$\mu \neq \mu_0$	$n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu_a}\right]^2$

9.7.2 Precision in testing for a population proportion

The issue of precision arises as naturally in testing problems for a population proportion, as it does for in testing for a population mean. In Example 9.4.3, a sample of size $n = 200$ yielded $\hat{p} = 0.755$. This, however, was not enough to reject the null hypothesis $H_0 : p = 0.7$ versus $H_a : p > 0.7$, at $\alpha = 0.01$. The failure to reject the null hypothesis, when the estimate suggests that the alternative is true, raises questions regarding the power of the test procedure.

Here we will present formulas for the probability of type II error of the large sample test procedures for testing $H_0 : p = p_0$ vs some alternative. The formulas for the probability of type II error, will lead to formulas for the sample size needed to achieve a specified degree of precision.

As seen in Section 8.3, when $np_0 \geq 5$ and $n(1 - p_0) \geq 5$, the test statistic, for testing $H_0 : p = p_0$ vs the different alternatives, and its null distribution, is

$$Z_{H_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\cdot}{\sim} N(0,1).$$

The RRs, for the different alternative hypotheses, are as given in Subsection 9.4.1, but with Z_{H_0} given above.

The formulas for the probability of type II error, when testing $H_0 : p = p_0$ against the different alternative hypotheses, are:

H_a	$\beta(p_a)$ for a Level α test
$p > p_0$	$\Phi\left(\frac{p_0 - p_a + z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_a(1-p_a)/n}}\right)$
$p < p_0$	$1 - \Phi\left(\frac{p_0 - p_a - z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_a(1-p_a)/n}}\right)$
$p \neq p_0$	$\Phi\left(\frac{p_0 - p_a + z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_a(1-p_a)/n}}\right) - \Phi\left(\frac{p_0 - p_a - z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p_a(1-p_a)/n}}\right)$

A brief derivation of the first of the above formulas for $\beta(p_a)$, i.e. when the alternative is $H_a : p > p_0$, is given below. Thus, in the following derivation p_a is a value greater than

p_0.

$$\begin{aligned}
\beta(p_a) &= P(\text{type II error when } p = p_a) = P(Z < z_\alpha \mid p = p_a) \\
&= P\bigg(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < z_\alpha \,\bigg|\, p = p_a \bigg) \\
&= P\bigg(\hat{p} < p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \,\bigg|\, p = p_a \bigg) \\
&= \Phi\bigg(\frac{p_0 - p_a + z_\alpha \sqrt{p_0(1-p_0)/n}}{\sqrt{p_a(1-p_a)/n}} \bigg)
\end{aligned}$$

The formulas for the determination of the sample size needed to achieve a prescribed level, β, of probability of type II error, when testing $H_0 : p = p_0$ against the different alternative hypotheses, at level α, are:

H_a	Sample Size Needed for $\beta(p_a)$ to Equal β
$\mu > \mu_0$	$n = \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_a(1-p_a)}}{p_0 - p_a} \right]^2$
$\mu < \mu_0$	$n = \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_a(1-p_a)}}{p_0 - p_a} \right]^2$
$\mu \neq \mu_0$	$n = \left[\frac{z_{\alpha/2} \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_a(1-p_a)}}{p_0 - p_a} \right]^2$

Example 9.7.3. (Example 9.4.3, continued) In the setting of Example 9.4.3 find $\beta(0.8)$.

Solution.

$$\begin{aligned}
\beta(0.8) &= \Phi\bigg(\frac{0.7 - 0.8 + 2.33\sqrt{(0.7)(0.3)/200}}{\sqrt{(0.8)(0.2)/200}} \bigg) \\
&= \Phi(-0.866) = 0.1936
\end{aligned}$$

9.7.3 Exercises

1. In the context of Exercise 9.4.5,3, assume that the true value of σ is known to be $\sigma = 4.1$ ppm.

   (a) Find the probability of type II error when the true ppm concentration is 2 ppm.

   (b) OSHA would like the probability of type II error not to exceed 10% when the true concentration is 2 ppm. What should be changed in their sampling procedure, and by how much, in order to achieve this?

2. An acid bath is used to clean impurities from the surfaces of metal bars used in laboratory experiments. The laboratory has instructed the chemical company, which provides batches of acid to the lab in large lots, that the solution must not be too strong or the bars will suffer damage to their plated surfaces. The nominal strength of the acid should be 8.5 on an acidity index, and it is important that the average index should not exceed 8.65. The distribution of acid index for this type of solution is known to be normal with standard deviation 0.2 units. The chemical company will take a sample of several batches of acid from a lot and test $H_0 : \mu = 8.5$ against $H_a : \mu > 8.5$. If the null hypothesis is rejected they will not deliver the lot. On the other hand, if the test fails to reject the null hypothesis the lot is delivered. The chemical company has determined that a type I error, α of 0.05 is acceptable to them, while the lab had determined that a type II error at $\mu = 8.65$, $\beta(8.65)$, of 0.05 is acceptable to them. How many batches must be sampled to achieve these errors?

3. In the context of Exercise 9.4.5,5, and find the probability of type II error at $p_a = 0.22$.

9.8 Review Exercises

1. A study examined the effect of varying the water/cement ratio (W/C R) on the strength of concrete that has been aged 28 days. The data were entered into Minitab and the following output was obtained.

   The regression equation is: Strength = 2.56 - 1.06 W/C R

   ```
   Predictor     Coef     SE Coef      T
   Constant     2.5618    0.1406     18.23
   Water/Ce    -1.0551    0.09562   -11.03

              Analysis of Variance Table
   Source            DF       SS       MS    F
   Regression         1    0.26042
   Residual Error
   Total              5    0.26898

              S=0.04626, R-sq=96.82%
   ```

   (a) What is the sample size of this data set?

   (b) What proportion of the total variability in cement strength is explained by the regression model?

   (c) What is the estimate of the standard deviation of the intrinsic error?

(d) (2 points) Give the estimated mean cement strength at water/cement ratio of $X = 1.4$.

(e) Fill in the missing entries in the ANOVA table.

(f) State the null and the alternative hypothesis for the model utility test.

(g) According to the ANOVA table, would you reject the null hypothesis of the model utility test? Justify.

(h) The question of interest is whether increasing the water content weakens the cement. State an appropriate null and alternative hypothesis.

(i) Test the hypothesis you specified at $\alpha = 0.05$.

(j) Two concrete pieces are made at water/cement ratios 1.35, and 1.55, respectively.

   (i) Estimate the mean difference of the strength at water/cement ratios 1.35, and 1.55.

   (ii) Make a 95% Confidence Interval for the above mean difference.

2. Regression analysis was used to develop an equation for predicting concrete strength from its modulus of elasticity. Assuming that the regression function is linear, a statistical software gave the following output.

The regression equation is: Strength = 3.29 + 0.107 MoE

Predictor	Coef	SE Coef	T
Constant	3.2925	0.60080	5.48
MoE	0.10748	0.01280	8.40

Analysis of Variance Table

Source	DF	SS	MS	F
Regression	1	52.870		
Residual Error				
Total	26	71.605		

S = 0.8657, R-Sq = 73.8%, R-Sq(adj) = 72.8%

(a) What is the sample size of this data set?

(b) What percent of the total variability is accounted for by the regression model?

(c) Fill in the missing entries in the ANOVA table.

(d) Use the information given above the ANOVA table to construct a 95% CI for the slope of the true regression line.

(e) Use the information given above the ANOVA table to test the hypothesis $H_0 : \beta_1 = 0.1$ versus the alternative $H_a : \beta_1 < 0.1$, where β_1 is the slope of the true regression line, at $\alpha = 0.05$.

(f) Use the information given in the ANOVA table to conduct the model utility test. Make sure you state what hypotheses are tested in the model utility test.

(g) It is given that

$$\overline{X} = 55 \quad \text{and} \quad \sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2 = \sum_i (X_i - \overline{X})^2 = 4,574.2$$

Construct a 95% CI for the expected strength at MoE=65. (The range of MoE values in the data set is from 30 to 80.)

(h) Use the information given in part g) to construct a 95% prediction interval for the next strength measurement at MoE=65.

Chapter 10

Comparing Two Populations

10.1 Introduction

In this chapter we use confidence intervals and hypothesis testing for comparing two population means and two population proportions. For example, we may want to compare the hardness of two cement mixtures, the strength of two types of steel, the drying time of two types of paint, the effect of two different temperatures on the yield of a chemical reaction, the proportions of warrantee claims for two different types of products, the proportion of defective products coming out of two different assembly lines, and so forth.

Studies aimed at comparing two populations are the simplest kind of comparative studies mentioned in Section 1.7 of Chapter 1. In the jargon introduced there, the two populations are also called treatments, or the two levels of a factor. For example, the two different temperatures for performing a chemical reaction are called the two treatments, or the two levels of the factor "temperature". Scientists are often interested in comparing two different treatments to determine if either produces a significantly more desirable effect on the response. Recall, however, that the experimenter can be confident that an observed statistically significant difference in the responses is due to a difference in the treatments (that is, that a cause-and-effect relationship has been established) only if the allocation of experimental units to the treatments is done in a randomized fashion controlled by the experimenter, i.e. only when a statistical experiment is performed.

Let μ_1, σ_1^2 denote the mean and variance of population 1, and μ_2, σ_2^2 denote the mean and variance of population 2. If the two populations are Bernoulli, then $\mu_1 = p_1$, where p_1 is the probability of "success" in a random selection from population 1, $\sigma_1^2 = p_1(1 - p_1)$, and similarly, $\mu_2 = p_2$, $\sigma_2^2 = p_2(1-p_2)$. The comparison of the two populations will be based on a simple random sample from each of the two populations. Let

$$X_{i1}, X_{i2}, \ldots, X_{in_i},\ i = 1, 2, \tag{10.1.1}$$

denote the two samples. Thus, the random sample from population 1 has size n_1, with observations denoted by $X_{11}, \ldots, X_{1n_1}$, and the random sample from population 2 has size n_2 with observations denoted by $X_{21}, \ldots, X_{2n_2}$. Moreover, let

$$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{in_i},\ S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2, \tag{10.1.2}$$

be the sample mean and sample variance from the ith sample, $i = 1, 2$. When the populations are Bernoulli

$$\overline{X}_i = \widehat{p}_i, \text{ and } S_i^2 = \frac{n_i}{n_i - 1}\widehat{p}_i(1-\widehat{p}_i),$$

and typically only the proportion, $\widehat{p}_i$, or the number, $T_i = n_i\widehat{p}_i$, of successes in the sample from population i, $i = 1, 2$, are given. In previous chapters we saw how to construct confidence intervals for each μ_i, or p_i, and how to test hypotheses regarding their value. These one-sample procedures, however, are not suitable for the present task of comparing the two population parameters.

One approach for comparing the two population means is based on the difference, or **contrast**

$$\overline{X}_1 - \overline{X}_2. \tag{10.1.3}$$

For the comparison of two population proportions, this approach is based on the contrast

$$\widehat{p}_1 - \widehat{p}_2. \tag{10.1.4}$$

An alternative approach for the comparison of two populations, which uses the ranks of the observations, will also be described.

With the exception of Section 10.4, we will make the further assumption that the two samples in (10.1.1) are independent (have been collected independently).

We close the introduction with a proposition that summarizes the basic facts on which the contrast based approach to confidence intervals and hypothesis testing is based.

Proposition 10.1.1. *The estimator $\overline{X}_1 - \overline{X}_2$ of $\mu_1 - \mu_2$ satisfies*

1. $\overline{X}_1 - \overline{X}_2$ is unbiased for $\mu_1 - \mu_2$, i.e.

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2.$$

In particular, if the X_{1j}'s and the X_{2j}'s are Bernoulli random variables, then

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2.$$

2. The variance of $\overline{X}_1 - \overline{X}_2$ is

$$\sigma^2_{\overline{X}_2 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If the X_{1j}'s and the X_{2j}'s are Bernoulli random variables, then

$$\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

3. If both populations are normal,

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma^2_{\overline{X}_1 - \overline{X}_2}). \tag{10.1.5}$$

4. For non-normal populations, (10.1.5) is approximately true for large samples ($n_1 \geq 30$, $n_2 \geq 30$) by the CLT. If the X_{1i}'s and the X_{2i}'s are Bernoulli random variables, so $T_1 = \sum_{j=1}^{n_1} X_{1j}$ and $T_2 = \sum_{j=1}^{n_2} X_{2j}$ are binomial random variables,

$$\widehat{p}_1 - \widehat{p}_2 = \frac{T_1}{n_1} - \frac{T_2}{n_2} \dot{\sim} N\left(p_1 - p_2, \sigma^2_{\widehat{p}_1 - \widehat{p}_2}\right),$$

if $n_1 p_1 \geq 10$, $n_1(1-p_1) \geq 10$, $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$ are all true

Corollary 10.1.1. *The estimated standard error of $\overline{X}_1 - \overline{X}_2$ is*

$$\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \tag{10.1.6}$$

where S_1^2, S_2^2 are the sample variances from populations 1, 2, respectively. When the sample sizes are large enough (≥ 30), and so the estimated standard error of $\overline{X}_1 - \overline{X}_2$ is a good approximation to its true standard error,

$$\overline{X}_1 - \overline{X}_2 \dot{\sim} N\left(\mu_1 - \mu_2, \widehat{\sigma}^2_{\overline{X}_1 - \overline{X}_2}\right). \tag{10.1.7}$$

The estimated standard error of $\widehat{p}_1 - \widehat{p}_2$ is

$$\widehat{\sigma}_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}. \tag{10.1.8}$$

When the sample sizes are large enough ($n_1\widehat{p}_1 \geq 10$, $n_1(1-\widehat{p}_1) \geq 10$, and $n_2\widehat{p}_2 \geq 10$, $n_2(1-\widehat{p}_2) \geq 10$, for our purposes), and so the estimated standard error of $\widehat{p}_1 - \widehat{p}_2$ is a good approximation to its true standard error,

$$\widehat{p}_1 - \widehat{p}_2 \dot{\sim} N\left(p_1 - p_2, \widehat{\sigma}^2_{\widehat{p}_1 - \widehat{p}_2}\right), \tag{10.1.9}$$

10.2 Nonparametric Methods

In this section we will describe two nonparametric procedures for comparing two populations. They are called nonparametric because they do not require the assumption of normality, or any other distributional assumption, and are valid for all ordinal data, i.e. both discrete and continuous.

10.2.1 The Nonparametric Contrast-Based Procedure

In this subsection we will learn how to use the sample contrasts (10.1.3) and (10.1.4) for constructing confidence intervals for the corresponding contrasts between population parameters, i.e. $\mu_1 - \mu_2$ and $p_1 - p_2$, and for testing hypotheses regarding these contrasts. This approach uses the Central Limit Theorem to approximate the distribution of the sample contrasts in a nonparametric way. These facts will now be used for the construction of confidence intervals and for testing hypotheses.

Confidence Intervals for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and for $\mathbf{p}_1 - \mathbf{p}_2$

Relation (10.1.7) yields the following $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\widehat{\sigma}_{\overline{X}_1-\overline{X}_2} \boxed{\begin{array}{l}(1-\alpha)100\% \text{ confidence}\\ \text{interval for } \ \mu_1-\mu_2\end{array}}, \tag{10.2.1}$$

where $\widehat{\sigma}_{\overline{X}_1-\overline{X}_2}$ is given in (10.1.6).

Similarly, relation (10.1.9) yields the following $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$.

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{p}_1-\widehat{p}_2} \boxed{\begin{array}{l}(1-\alpha)100\% \text{ confidence}\\ \text{interval for } \ p_1-p_2\end{array}}, \tag{10.2.2}$$

where $\widehat{\sigma}_{\widehat{p}_1-\widehat{p}_2}$ is given in (10.1.8).

Example 10.2.1. A random sample of $n_1 = 32$ specimens of cold-rolled steel give average strength $\overline{X}_1 = 29.8$ ksi, and sample standard deviation of $S_1 = 4$. A random sample of $n_2 = 35$ specimens of two-sided galvanized steel give average strength $\overline{X}_2 = 34.7$ ksi, and $S_2 = 6.74$. Find a 99% CI for $\mu_1 - \mu_2$.

Solution. Here $\alpha = .01$, $z_{\alpha/2} = z_{.005} = 2.58$, and $\widehat{\sigma}_{\overline{X}_1-\overline{X}_2} = 1.34$. Thus the 99% CI is

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\widehat{\sigma}_{\overline{X}_1-\overline{X}_2} = -4.90 \pm 2.58 \cdot 1.34 = (-8.36, -1.44).$$

Example 10.2.2. A certain plant manufactures tractors in each of two different assembly lines labeled L_1 and L_2. It is suspected that assembly line L_1 produces a higher proportion of tractors requiring extensive adjustments. A random sample of 200 tractors produced by L_1 yield 16 tractor requiring extensive adjustment, while a random sample of 400 tractors produced by L_2 yield 14 requiring extensive adjustments. Do a 99% CI for $p_1 - p_2$, the difference of the two proportions.

Solution. Here $\widehat{p}_1 = 16/200 = 0.08$ and $\widehat{p}_2 = 14/400 = 0.035$. Moreover, $\alpha = .01$, so that $z_{\alpha/2} = z_{.005} = 2.575$. Thus the 99% C.I. is

$$0.08 - .035 \pm 2.575\sqrt{\frac{(.08)(.92)}{200} + \frac{(.05)(.95)}{400}} = (-0.012,\ 0.102).$$

Testing Hypotheses About $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

In comparing two population means, the null hypothesis of interest can be put in the form

$$H_0 : \mu_1 - \mu_2 = \Delta_0, \tag{10.2.3}$$

where Δ_0 is a constant which is specified in the context a particular application. Note that if $\Delta_0 = 0$, the null hypothesis specifies that $\mu_1 = \mu_2$. The test statistic for testing this null hypothesis is

$$Z_{H_0} = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}}, \tag{10.2.4}$$

where $\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}$ is given in (10.1.6). If the null hypothesis is true, i.e. if $\mu_1 - \mu_2 = \Delta_0$, then, according to (10.1.7)

$$Z_{H_0} \overset{\cdot}{\sim} N(0, 1), \tag{10.2.5}$$

which justifies labeling the test statistic by Z_{H_0}. If the null hypothesis is not true, then Z_{H_0} no longer has the standard normal distribution.

The null hypothesis in (10.2.3) is tested against one of the alternatives

$$H_a : \mu_1 - \mu_2 > \Delta_0, \quad \text{or} \quad H_a : \mu_1 - \mu_2 < \Delta_0, \quad \text{or} \quad H_a : \mu_1 - \mu_2 \neq \Delta_0.$$

To motivate the rejection rules listed below, we note that if the true value of $\mu_1 - \mu_2$ is larger than the Δ_0 specified by the null hypothesis (so $H_a : \mu_1 - \mu_2 > \Delta_0$ is true), then the test statistic Z_{H_0} in (10.2.4) tends to take larger values than a standard normal random variable. Analogously, if the true value of $\mu_1 - \mu_2$ is smaller than the Δ_0 specified by the null hypothesis (so $H_a : \mu_1 - \mu_2 < \Delta_0$ is true), then Z_{H_0} tends to take smaller values than a standard normal random variable. To see this suppose that $\mu_1 - \mu_2 = \Delta_1$, where Δ_1 is a value greater than Δ_0 (so $H_a : \mu_1 - \mu_2 > \Delta_0$ is true). To fix ideas, suppose that $\Delta_0 = 0$ (so the null hypothesis specifies that $\mu_1 = \mu_2$), but the true value of the difference of the means is 1 (so $\Delta_1 = 1$). In this case, according to (10.1.7),

$$Z_{H_a} = \frac{\overline{X}_1 - \overline{X}_2 - 1}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \mathrel{\dot\sim} N(0,1), \quad \text{while} \quad Z_{H_0} = \frac{\overline{X}_1 - \overline{X}_2}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}}.$$

Thus,

$$Z_{H_0} = Z_{H_a} + \frac{1}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}},$$

which shows that the test statistic is larger than the standard normal variable Z_{H_a}. It can similarly be demonstrated that, if $H_a : \mu_1 - \mu_2 < \Delta_0$ is true, then Z_{H_0} will take values smaller than a standard normal random variable. This leads to the following rejection rules:

$$\begin{array}{cc}
H_a & \text{RR at level } \alpha \\
\hline
\mu_1 - \mu_2 > \Delta_0 & Z_{H_0} > z_\alpha \\
\mu_1 - \mu_2 < \Delta_0 & Z_{H_0} < -z_\alpha \\
\mu_1 - \mu_2 \neq \Delta_0 & |Z_{H_0}| > z_{\alpha/2}
\end{array} \tag{10.2.6}$$

The p-value is computed the same way as for the one-sample Z-test. The calculation of the probability of Type II error when the true value of the contrast is $\mu_1 - \mu_2 = \Delta_1$, $\beta(\Delta_1$, is done according to the following formulas for the different alternative hypotheses

Formulas for Calculating the Probability of Type II Error

$$\begin{array}{cc}
H_a & \beta(\Delta_1) \\
\hline
\mu_1 - \mu_2 > \Delta_0 & \Phi\left(z_\alpha - \frac{\Delta_1 - \Delta_0}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \right) \\
\mu_1 - \mu_2 < \Delta_0 & \Phi\left(-z_\alpha - \frac{\Delta_1 - \Delta_0}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \right) \\
\mu_1 - \mu_2 \neq \Delta_0 & \Phi\left(z_{\alpha/2} - \frac{\Delta_1 - \Delta_0}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta_1 - \Delta_0}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \right)
\end{array} \tag{10.2.7}$$

where $\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}$ is given in (10.1.6).

Example 10.2.3. Consider the data from the experiment of Example 10.2.1, which compares the strengthes of cold-rolled steel and two-sided galvanized steel. Are the strengths of the two types of steel different at $\alpha = .01$? What is the probability of type II error when $\mu_1 - \mu_2 = 5$?

Solution. Here we want to test $H_0 : \mu_1 - \mu_2 = 0$ vs $H_a : \mu_1 - \mu_2 \neq 0$ (i.e. $\Delta_0 = 0$). Test statistic takes the value

$$Z_{H_0} = \frac{\overline{X}_1 - \overline{X}_2}{\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}} = \frac{-4.90}{1.34} = -3.66.$$

Here $z_{\alpha/2} = z_{.005} = 2.58$. Since $|-3.66| > 2.58$, H_0 is rejected. To compute the probability of Type II error, we use the formula (10.2.7) with $\Delta_0 = 0$, $\Delta_1 = 5$ and $\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2} = 1.34$. Thus,

$$\beta(5) = \Phi(-1.15) - \Phi(-6.31) = .1251.$$

Remark 10.2.1. In view of the fact that the confidence interval in Example 10.2.1 does not include 0, the outcome of the test procedure in Example 10.2.3 is to be expected. Indeed, when testing against a two-sided alternative, the null hypothesis will be rejected at level α, when ever Δ_0 is not included in the $(1 - \alpha)100\%$ confidence interval.

Testing Hypotheses About $\mathbf{p}_1 - \mathbf{p}_2$

In comparing two population proportions, the null hypothesis of interest can be put in the form

$$H_0 : p_1 - p_2 = \Delta_0, \tag{10.2.8}$$

where Δ_0 is a constant which is specified in the context a particular application. If $\Delta_0 \neq 0$, the test statistic is as given in (10.2.4) with $\widehat{p}_1 - \widehat{p}_2$ replacing $\overline{X}_1 - \overline{X}_2$, and $\widehat{\sigma}_{\widehat{p}_1 - \widehat{p}_2}$ replacing $\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}$, i.e.

$$Z_{H_0} = \frac{\widehat{p}_1 - \widehat{p}_2 - \Delta_0}{\widehat{\sigma}_{\widehat{p}_1 - \widehat{p}_2}}.$$

The regions for rejecting the null hypothesis in (10.2.8) in favor of one of the alternatives

$$H_a : p_1 - p_2 > \Delta_0, \quad \text{or} \quad H_a : p_1 - p_2 < \Delta_0, \quad \text{or} \quad H_a : p_1 - p_2 \neq \Delta_0$$

are as given in (10.2.6) with $p_1 - p_2$ replacing $\mu_1 - \mu_2$ in the specification of H_a. Finally the probability of Type II error is calculated as in (10.2.7) with $\widehat{\sigma}_{\widehat{p}_1 - \widehat{p}_2}$ replacing $\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}$.

The only new item in the comparison of two proportions involves testing the hypothesis (10.2.8) with $\Delta_0 = 0$, i.e. $H_0 : p_1 = p_2$. While this hypothesis can be tested as described above for the case of the general Δ_0, it is customary to use a different estimate of $\sigma_{\widehat{p}_1 - \widehat{p}_2}$ in this case. In particular, it is customary to estimate the standard error of $\widehat{p}_1 - \widehat{p}_2$ by

$$\widetilde{\sigma}_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{\widehat{p}(1 - \widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \boxed{\begin{array}{l}\text{Estimated standard error} \\ \text{for } \widehat{p}_1 - \widehat{p}_2 \text{ when } p_1 = p_2\end{array}} \tag{10.2.9}$$

where

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2},$$

is the *pooled* estimator of the common value of p_1 and p_2, assuming, of course, that $p_1 = p_2$. Note that (10.2.9) can be obtained from the formula for the estimated standard error given in (10.1.8) by substituting $\widehat{p}$ in place of $\widehat{p}_1$ and $\widehat{p}_2$ ($\widetilde{\sigma}_{\widehat{p}_1-\widehat{p}_2}$ is a "correct" estimator of the standard error of $\widehat{p}_1 - \widehat{p}_2$, only under the null hypothesis.) Thus the test statistic for testing $H_0 : p_1 = p_2$ is

$$Z_{H_0} = \frac{\widehat{p}_1 - \widehat{p}_2}{\widetilde{\sigma}_{\widehat{p}_1-\widehat{p}_2}},$$

and the rejection regions remain the same as described above, i.e. are as given in (10.2.6) with $p_1 - p_2$ replacing $\mu_1 - \mu_2$ in the specification of H_a.

Example 10.2.4. Consider the manufacturing of tractors using two different assembly lines, as described in Example 10.2.2. Let p_1 denote the proportion of tractors coming out of assembly line L_1 that require adjustments, and let p_2 be the corresponding proportion for assembly line L_2. Test the null hypothesis $H_0 : p_1 = p_2$ against the alternative $H_a : p_1 > p_2$, at level of significance at $\alpha = .01$, and calculate the p-value.

Solution. Here $n_1 = 200$, $n_2 = 400$, $\widehat{p}_1 = 0.08$, and $\widehat{p}_2 = 0.035$. Thus, the pooled estimate of the (assumed, under H_0) common value of the two probabilities is $\widehat{p} = .05$. The test statistic is

$$Z_{H_0} = \frac{.08 - .035}{\sqrt{(.05)(.95)\left(\frac{1}{200} + \frac{1}{400}\right)}} = 2.38.$$

Since $2.38 > z_{.01} = 2.33$, H_0 is rejected. The p-value is

$$p\text{-value} = 1 - \Phi(2.38) = 1 - .9913 = .0087.$$

10.2.2 The Rank-Sum Test Procedure

The contrast-based procedure described in the previous subsection can only be applied if we have large sample sizes. In this subsection we will describe a popular test procedure, called the *Mann-Whitney-Wilcoxon rank-sum test* (or *rank-sum test* for short), which can be used with both small and large sample sizes. If the two statistical populations are continuous, the null distribution of the test statistic is known even with very small sample sizes, while for discrete populations, the null distribution of the test statistic can be well approximated with much smaller sample sizes than what the nonparametric contrast-based procedure requires. However, the popularity of the rank-sum test is also due to its desirable power properties (i.e. low probability of type II error), especially if the two population distributions are heavy tailed, or skewed.

Let μ_1, μ_2 denote the two population means, and F_1, F_2 denote the two population cumulative distribution functions. We will concentrate on testing the hypothesis

$$H_0^F : F_1 = F_2. \tag{10.2.10}$$

Note that if H_0^F is true, then so is the hypothesis of equality of the two population means, $H_0 : \mu_1 - \mu_2 = 0$.

As the name of the procedure suggests, the test is based on the *ranks* of the observations, or the *mid-ranks* if some observations are *tied*, i.e. if some observations share the same value. The ranks and mid-ranks of a set of observations were defined in Chapter 6.

Implementation of the rank sum test procedure begins by combining the observations, $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$, from the two samples into an overall set of $n_1 + n_2$ observations. Let R_{ij} denote the (mid-)rank of observation X_{ij} in this combined sample, and set

$$\overline{R}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} R_{1j}, \quad \overline{R}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} R_{2j}.$$

Then, the standardized version of the rank-sum test statistic is

$$Z_{H_0} = \frac{\overline{R}_1 - \overline{R}_2}{\sqrt{S_R^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{10.2.11}$$

where

$$S_R^2 = \frac{1}{N-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(R_{ij} - \frac{N+1}{2}\right)^2, \quad \text{where } N = n_1 + n_2. \tag{10.2.12}$$

If the null hypothesis 10.2.10 is true, then

$$Z_{H_0} \dot{\sim} N(0,1),$$

and the approximation is good for $n_1,\ n_2 > 8$. Thus, the regions for rejecting H_0 against the different alternatives are:

H_a	Rejection region at level α
$\mu_1 - \mu_2 > 0$	$Z_{H_0} \geq z_\alpha$
$\mu_1 - \mu_2 < 0$	$Z_{H_0} \leq -z_\alpha$
$\mu_1 - \mu_2 \neq 0$	$\lvert Z_{H_0}\rvert \geq z_{\alpha/2}$

Remark 10.2.2. 1. In the context of rank tests, the alternative $\mu_1 - \mu_2 > 0$ should be interpreted more generally as

$$P(X_{1j} > X_{2j}) > 0.5,$$

where X_{ij} denotes a random observation from population i. Similar interpretations apply to the other alternative hypotheses.

2. For $n1,\ n_2 \leq 8$ the rank-sum statistic is simply the sum of the ranks of the first sample, i.e.

$$W = \sum_{j=1}^{n_1} R_{1j}.$$

For continuous populations (so no ties), the distribution of W is the same no matter what the (continuous) population distribution is. Moreover, this distribution has been tabulated, and so testing can be carried out even for very small sample sizes, using either tables or a software package.

3. In the case of no ties, an alternative form of the standardized MWW statistic Z_{H_0} is

$$Z_{H_0} = \sqrt{\frac{12}{n_1 n_2 (N+1)}} \left(W - n_1 \frac{N+1}{2} \right).$$

In the case of ties, an alternative form of the standardized MWW statistic is

$$Z_{H_0} = \left[\frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2 \sum_k d_k (d_k^2 - 1)}{12 N (N-1)} \right]^{-1/2} \left(W - n_1 \frac{N+1}{2} \right),$$

where the summation is over all groups of tied observations in the combined sample, and d_k is the number of tied observations at the kth group.

4. Two valid variations of the statistic Z_{H_0} (which, however, might yield different results) are: a) use the two-sample t-statistic evaluated on the ranks (Section 10.3), and b) use the two-sample Z-statistic evaluated on the ranks. The first consists of replacing S_R^2 in the formula (10.2.11) by the pooled sample variance, evaluated on the ranks, i.e.

$$\frac{1}{N-2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left(R_{ij} - \overline{R}_i \right)^2,$$

while the second consists of replacing the denominator in (10.2.11) by

$$\sqrt{\frac{S_{1,R}^2}{n_1} + \frac{S_{2,R}^2}{n_2}},$$

where $S_{1,R}^2$ is the sample variance of $R_{11}, \ldots, R_{1n_1}$, and $S_{2,R}^2$ is the sample variances of $R_{21}, \ldots, R_{2n_2}$.

Example 10.2.5. The sputum histamine levels from a sample of size 9 allergic individuals and 13 non-allergic individuals are

Allergic:	67.7, 39.6, 1651.0, 100.0, 65.9, 1112.0, 31.0, 102.4, 64.7
Non-Allergic:	34.3, 27.3, 35.4, 48.1, 5.2, 29.1, 4.7, 41.7, 48.0, 6.6, 18.9, 32.4, 45.5

Is there a difference between the two populations?

Solution. Here sample 1 is the sample of allergic individuals. The ranks of the observations of sample 1 are $R_{11} = 18$, $R_{12} = 11$, $R_{13} = 22$, $R_{14} = 19$, $R_{15} = 17$, $R_{16} = 21$, $R_{17} = 7$, $R_{18} = 20$, $R_{19} = 16$. Thus $W = \sum_j R_{1j} = 151$. Because both sample sizes are larger than 8, we use the normal approximation to conduct the test. Here

$$Z_{H_0} = \frac{151 - 9(23)/2}{\sqrt{9(13)(23)/12}} = 3.17,$$

which yields p-value=$2[1 - \Phi(3.17)] = .0016$. Thus H_0 is rejected in favor of $H_a : \mu_1 \neq \mu_2$ even at $\alpha = .01$.

10.2.3 Exercises

1. An article in the journal of *Engineering Fracture Mechanics, 1997, Vol. 56, No.1, 65-76* reports on a study regarding the effect of thickness in fatigue crack growth in aluminum alloy 2024-T351. Two groups of specimens were created, one with thickness of 3mm and the other with a thickness of 15mm. Each specimen had an initial crack length of 15mm. The same cyclic loading was applied to all specimens, and the number of cycles it took to reach a final crack length of 25mm was recorded. Suppose that for the group having thickness 3mm, a sample of size 36 gave $\overline{X}_1 = 160,592$ and $S_1 = 3,954$, and for the group having thickness 15mm, a sample of size 42 gave $\overline{X}_2 = 159,778$ and $S_2 = 15,533$. The scientific question is whether or not thickness affects fatigue crack growth.

   (a) State the null and alternative hypotheses.

   (b) State the test statistic and the rejection region, with level of significance α.

   (c) What assumptions, if any, are needed for the validity of the test procedure you specified in part b)?

   (d) Carry out the test at level $\alpha = 0.05$, and state the p-value.

   (e) Construct a 95% CI for the difference in the two means. Explain how the testing problem in a) can be conducted in terms of the CI, and check if the test result remains the same.

2. To compare the corrosion-resistance properties of two types of material used in underground pipe lines, specimens of both types are buried in soil for a 2-year period and the maximum penetration (in mils) for each specimen is measured. A sample of size 42 specimens of material type A yielded $\overline{X}_1 = 0.49$ and $S_1 = 0.19$, and a sample of size 42 specimens of material type B gave $\overline{X}_2 = 0.36$ and $S_2 = 0.16$. The scientific question is whether the two types of material have the same corrosion resistance.

   (a) State the null and alternative hypotheses.

   (b) State the test statistic and the rejection region, with level of significance α.

   (c) What assumptions, if any, are needed for the validity of the test procedure you specified in part b)?

   (d) Carry out the test at level $\alpha = 0.05$, and state the p-value.

   (e) Construct a 95% CI for the difference in the two means. Explain how the testing problem in a) can be conducted in terms of the CI, and check if the test result remains the same.

3. The lifetimes of a random sample of 49 batteries of brand A gave $\overline{X}_1 = 4,250$ hours, and $S_1 = 220$ hours. The lifetimes of a random sample of 49 batteries of brand B gave $\overline{X}_2 = 4,040$ hours, and $S_2 = 190$ hours. Find the p-value for testing $H_0 : \mu_1 - \mu_2 = 100$ versus $H_0 : \mu_1 - \mu_2 > 100$, and use the p-value to determine if H_0 is rejected at level $\alpha = 0.05$.

4. The article *Improving the thermal fatigue resistance of brake discs* in the journal *Materials, Experimentation and Design in Fatigue, 1981, 60-71* reports on a study using high temperature strain gages to measure the total strain amplitude of different types of cast iron for use in disc brakes. The results for spheroidal graphite (SG) and compacted graphite (CG), multiplied by 10,000, are:

   ```
   SG: 105  77  52  27  22  17  12  14  65
   CG:  90  50  30  20  14  10  60  24  76
   ```

   The scientific question is whether the total amplitude strain properties of the different types of cast iron differ.

   (a) State the null and the alternative hypothesis.

   (b) Is the nonparametric contrast-based procedure appropriate for this data set? Justify your answer.

   (c) Conduct the rank sum test, at $\alpha = 0.05$.

5. Carbon fiber reinforced polymer matrix is used to reinforce concrete columns. Impacted columns are wrapped with recycled crumb rubber. A study considered the effect that

wrapping with two types of recycled crumb rubber has on the bond strength. The two types of rubber were GF80 and GF170, using a concrete to rubber crumb ratio of 9:1. A pull off tester was used to measure the bond strength. The data are:

```
GF80 : 5.03 5.36 5.39 3.90
GF170: 4.36 3.66 5.77 5.38
```

(a) State the null and the alternative hypothesis.

(b) Is the nonparametric contrast-based procedure appropriate for this data set? Justify your answer.

(c) State a rank-sum test statistic and the rejection region.

(d) Conduct the test at level $\alpha = 0.05$.

6. An article in the journal *Knee Surgery, Sports Traumatology, Arthroscopy* (2005, Vol. 13, 273-279) reported results of arthroscopic meniscal repair with an absorbable screw. For tears greater than 25 millimeters, 14 of 18 repairs were successful, while for tears less than 25 millimeters, 22 of 30 were successful.

   (a) Is there evidence that the success rate for the two types of tears are different? Use $\alpha = 0.1$.

   (b) Report the p-value for the above testing problem.

   (c) Construct a 90% confidence interval for the difference of the two population proportions.

7. A decision is to be made whether to site an incinerator for the county in Placerville or Centreville. The county commissioners favor Centreville. However, the residents of Centreville claim that a much larger percentage of residents in Centreville are strongly opposed to the incinerator being situated in their town than the corresponding percentage of residents of Placerville. The county commissioners agree to place the incinerator in Placerville, if a survey shows almost conclusive evidence in support of the Centreville residents' contention. Suppose 330 of the 890 residents of Centreville strongly oppose the incinerator being situated in Centreville, and that 170 of the 550 residents of Placerville strongly oppose the incinerator being situated in Placerville. Does this survey provide the evidence that the Centreville residents hoped for?

   (a) State the null and the alternative hypotheses.

   (b) State the test statistic and give the rejection region for the hypothesis you stated in part a).

   (c) Conduct the test at level $\alpha = 0.1$, and state the p-value.

(d) Construct a 90% confidence interval for the difference of the two population proportions.

8. A tracking device, used to enable a robot to home in on a beacon which produces a audio signal, is fine-tuned if the probability of correct identification of the direction of the beacon is the same for each side (left and right) of the tracking device. Out of 100 signals from the right, the device identifies the direction correctly 85 times. Out of 100 signals from the left, the device identifies the direction correctly 87 times.

   (a) State the null and the alternative hypotheses.

   (b) State the test statistic and give the rejection region for the hypothesis you stated in part a).

   (c) Conduct the test at level $\alpha = 0.01$, and state the p-value.

   (d) Construct a 99% confidence interval for the difference of the two population proportions. Carry out the test of the hypotheses you specified in part a) using the confidence interval.

9. Suppose that in 85 10-mph crash tests for type A car, 19 sustained no visible damage. In 85 10-mph crash tests for type B car, 22 sustained no visible damage. Is this evidence sufficient to claim that type B cars do better in 10-mph crash tests than type A cars? State the null and the alternative hypothesis, and test at $\alpha = 0.05$.

10. With the information given in Exercise 3 above, find the probability of type II error at $\Delta_1 = 200$.

11. with the information given in Exercise 2 above, find the probability of type II error at $\Delta_1 = 0.1$.

10.3 Contrast-Based Procedures Under Normality

When the two population distributions can be assumed normal, the contrast-based procedure is recommended. If the two sample sizes are large, then the nonparametric contrast based procedure discussed in Subsection 10.2.1 for constructing confidence intervals and conducting hypothesis testing remain the same even if the two population distributions can be assumed normal. If either of the two sample sizes is small, the approximate distribution of the test statistic is not normal, as it is in the large-sample case. We will discuss two different approximating distributions, one when the two population variances can be assumed equal, and a different one when this is not the case. When the two population variances can be assumed equal, the denominator of the contrast-based test statistic will also be different.

The small sample test procedures that will be discussed in this section are alternative procedures to the nonparametric rank-sum test and are valid only if the two population distributions are normal.

10.3.1 Variances Unknown but Equal: The Two-Sample t-test

In addition to the normality assumption, in this subsection we will assume that the two population variances are equal. We let σ^2 denote the common, but unknown, value of the two variances, i.e.

$$\sigma_1^2 = \sigma_2^2 = \sigma^2. \tag{10.3.1}$$

Under this assumption, the variance of the contrast $\overline{X}_1 - \overline{X}_2$ is

$$\sigma^2_{\overline{X}_1-\overline{X}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Let S_i^2 be the sample variance from the ith sample, $i = 1, 2$. Because of (10.3.1), S_1^2 and S_2^2 are both unbiased estimators of σ^2. However, the estimator of choice combines, or *pools*, the two sample variances giving more weight to the sample variance from the sample with the larger sample size. The **pooled** estimator of σ^2 is

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Using it, the contrast-based test statistic for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ becomes

$$T_{H_0} = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

If $H_0 : \mu_1 - \mu_2 = \Delta_0$ is true then

$$T_{H_0} \sim t_{n_1+n_2-2}.$$

This leads to the following rejection regions:

H_a	R.R. at level α
$\mu_1 - \mu_2 > \Delta_0$	$T_{H_0} > t_{\alpha,n_1+n_2-2}$
$\mu_1 - \mu_2 < \Delta_0$	$T_{H_0} < -t_{\alpha,m+n-2}$
$\mu_1 - \mu_2 \neq \Delta_0$	$\|T_{H_0}\| > t_{\alpha/2,n_1+n_2-2}$

(10.3.2)

A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2,n_1+n_2-2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Example 10.3.1. To examine if two catalysts are equivalent in terms of the mean yield of a chemical process, n_1=8 chemical process are performed with catalyst A, and n_2=8 are performed with catalyst B. From catalyst A we obtain $\overline{X}_1 = 92.255$, $S_1 = 2.39$. From catalyst B we obtain $\overline{X}_2 = 92.733$, $S_2 = 2.98$. Test $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$ at $\alpha = .05$, and construct a 95% CI for the contrast $\mu_1 - \mu_2$.

Solution. Since the sample sizes are small and the population variances unknown, we assume that the populations are normal and also that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The pooled estimator of σ is

$$S_p = \sqrt{\frac{(8-1)S_1^2 + (8-1)S_2^2}{8+8-2}} = \sqrt{7.30} = 2.7.$$

The test statistics is

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{-.0508}{1.351} = -.0376.$$

Since $t_{.025,14} = 2.145$, H_0 is not rejected. Next, the requested 95% CI for $\mu_1 - \mu_2$ is

$$\begin{aligned} & \overline{X}_1 - \overline{X}_2 \pm t_{.025,14}\sqrt{S_p^2\left(\frac{1}{8} + \frac{1}{8}\right)} \\ = \quad & -.508 \pm 2.145 \times 1.351 = (-3.406,\ 2.390). \end{aligned}$$

10.3.2 Variances Unequal: The Smith-Satterthwaite t-test

When the assumption $\sigma_1^2 = \sigma_2^2$ appears to be violated, the denominator test statistic (and indeed the entire test statistic) remains the same as in Subsection 10.2.1, but its critical values are found from a t-distribution with degrees of freedom calculated on the basis of the sample variances and the sample sizes. This approximation of the null distribution of the test statistic, when the sample sizes are small, is due to Smith and Satterthwaite, and thus the test procedure described in the subsection is called the *Smith-Satterthwaite test.* Let

$$T'_{H_0} = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Smith and Satterthwaite proposed that, when n_1 and n_2 are small and the populations distributions are normal, then, if $H_0 : \mu_1 - \mu_2 = \Delta_0$ is true

$$T'_{H_0} \overset{\cdot}{\sim} t_\nu, \quad \text{where} \quad \nu = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}\right],$$

where brackets around a number x, $[x]$, denote the integer part of x (i.e. x rounded down to its nearest integer). The rejection regions are as in (10.3.2) with T'_{H_0} replacing T_{H_0} and ν replacing $n_1 + n_2 - 2$.

Example 10.3.2. A manufacturer of video display units is testing two micro circuit designs to determine whether they produce equivalent current flow. Using Design 1, $n_1 = 15$ current flow measurements give $\overline{X}_1 = 24.2$ and $S_1^2 = 10$. Using Design 2, $n_2 = 20$ measurements give $\overline{X}_2 = 23.9$ and $S_2^2 = 20$. Test $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$ at $\alpha = 0.1$.

Solution. Here

$$T'_{H_0} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = 0.18,$$

and $\nu = [14.17] = 14$. Thus, the value of the test statistic, 0.18, is compared to $t_{0.05,14} = 1.761$. Since T'_{H_0} is not greater that 1.761, the null hypothesis is not rejected.

10.3.3 Exercises

1. A fill and seal machine processes a particular product on two lines. As part of quality control, data are periodically collected to test if the fill weight is the same in the two lines. Suppose that a particular data collection of 12 fill weights from each line yield sample mean and sample variance of $\overline{X}_1 = 966.75$, $S_1^2 = 29.30$ from the first line, and $\overline{X}_2 = 962.33$, $S_2^2 = 26.24$ from the second line.

   (a) State the null and alternative hypotheses.

   (b) Assume that the two population variances are the equal. State the test statistic and the rejection region, with level of significance α.

   (c) What assumptions, if any, are needed for the validity of the test procedure you specified in part b)?

   (d) Carry out the test at level $\alpha = 0.05$, and state the p-value.

   (e) Construct a 95% CI for the difference in the two means. Explain how the testing problem in a) can be conducted in terms of the CI, and check if the test result remains the same.

2. Consider the testing problem of Exercise 1, and do parts b)-e) without the assumption that the two population variances are equal.

3. The article "Compression of Single-Wall Corrugated Shipping Containers Using Fixed and Floating Test Platens" in the *Journal of Testing and Evaluation, 1992, Vol. 20, No. 4, 318-320*, considered the compression strength of five different types of corrugated fiberboard containers in order to determine the effect of the platen on the results. For one type of corrugated containers, the floating platen strength measurements of a random sample of $n_1 = 10$ such containers gave $\overline{X}_1 = 757$ and $S_1 = 41$. For the same type of

corrugated containers, the fixed platen strength of an independent sample of $n_2 = 10$ such containers gave $\overline{X}_2 = 807$ and $S_2 = 27$.

(a) State the null and the alternative hypotheses.

(b) Give the two-sample test statistic assuming that the population variances are equal. State the rejection region, with level of significance α.

(c) What assumptions are needed for the validity of the test procedure in part b)?

(d) Carry out the test at level $\alpha = 0.1$, and state the p-value.

(e) Construct a 90% CI for the difference in the two means. Explain how the testing problem in a) can be conducted in terms of the CI, and check if the test result remains the same.

4. Consider the testing problem of Exercise 3, and do parts b)-e) without the assumption that the two population variances are equal.

5. 7075-T6 wrought aluminum alloy is commonly used in applications such as ski poles, aircraft structures, and other highly stressed structural applications, where very high strength and good corrosion resistance are needed. An experiment undertaken by a laboratory compared the ultimate tensile strengths (UTS) of notched and holed specimens of 7075-T6 wrought aluminum to determine which has higher sensitivity to stress concentrations. The experimental data are:

```
Notched UTS: 403 400 411 398 407 403 392 408 402 405 397 407 403
             396 404
Holed UTS  : 537 545 529 536 539 553 532 537 549 528 538 531 536
             544 543
```

(a) State the null and the alternative hypothesis.

(b) Assuming normality, conduct the two-sample t-test, at level $\alpha - 0.05$, assuming equal population variances. State the p-value.

(c) Use probability plots to assess the validity of the normality assumption.

(d) Construct a 90% CI for the difference in the two means. Explain how the testing problem in a) can be conducted in terms of the CI, and check if the test result remains the same.

6. Consider the testing problem of Exercise 5, and do parts b)-e) without the assumption that the two population variances are equal.

7. Consider the setting and information in Exercise 10.2.3,5.

(a) Assume normality, and conduct the two-sample t-test, at level $\alpha = 0.05$, assuming equal population variances.

(b) Construct a 95% CI for the difference of the population means.

10.4 Paired Data

Paired data arise when a random sample of n individuals (subjects or objects) receives each of the two treatments that are to be compared. Paired data are not independent, and thus the preceding procedures cannot be used.

The following examples highlight two contexts where such data arise.

Example 10.4.1. Two different types of materials for making soles for children's shoes are to be compared for durability. One way of designing this comparative experiment is to make n pairs of shoes where one (either the left or the right) is randomly selected to be made with material A, and the other with material B. Then a random sample of n children is selected and fitted with such pair of shoes. After a certain amount of time, the shoes are evaluated for ware and tear. In this example, the sample of n children are the subjects, and the two treatments are the two types of material. For each subject there will be two measurements, the quantification of wear and tear in the shoe made with material A, and the corresponding quantification in the shoe made with material B. This results in paired data. Another way of designing the comparative study is to select a random sample of n_1 children who are to be fitted with shoes made with material A, and a random sample of n_2 children who are to be fitted with shoes made with material B. This will result in two independent samples, one from each population.

Example 10.4.2. Two different methods for determining the percentage of iron in ore samples are to be compared. One way of designing this comparative study is to obtain n ore samples and subject each of them to the two different methods for determining the iron content. In this example, the n ore samples are the objects, and the two methods are the treatments. For each ore sample there will be two measurements, resulting in paired data. Another way of designing the comparative study is to obtain n_1 ore samples to be evaluated with method 1, and, independently, obtain a different set of n_2 ore samples to be evaluated with methods 2. This will result in two independent samples, one from each population.

From the above two examples it follows that the design that results in paired data eliminates a lot of uncontrolled variability. In Example 10.4.1, the uncontrolled variability is caused by the children having different weights, walking in different terrains, etc. Fitting the children with pairs of shoes, one of which is made with material A and the other with material B, assures that both materials are subject to the same weight and terrain, eliminating thus this source of uncontrolled variability. Similarly, in Example 10.4.2, though the ore samples came from the same area, individual ore samples might differ in iron content, due to natural variability. Subjecting the same ore samples to each method eliminates this uncontrolled variability. Elimination of uncontrolled variability means that we can have a more accurate comparison with smaller sample sizes. Thus, comparative studies should be designed to yield paired data whenever possible.

Let X_{1i} denote the observation on individual i receiving treatment 1, and X_{2i} denote the observation on individual i receiving treatment 2. Thus, individual i contributes the pair of observations (X_{i1}, X_{i2}) to the data set, which can be put in the form

$$(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n}).$$

Because they are associated with the same individual, X_{1i} and X_{i2} are not independent. Thus, the preceding procedures, which assume that the two samples are independent, cannot be used. The adaptation of these procedures to paired data is discussed in the next subsections.

10.4.1 The Contrast-Based Procedure for Paired Data

The fact that X_{1i} and X_{i2} are not independent means that the sample averages $\overline{X}_1$ and $\overline{X}_2$ are not independent. Therefore, the formula in (10.1.6) for the standard error of the sample contrast $\overline{X}_1 - \overline{X}_2$ does not apply. (The standard error involves also the covariance of $\overline{X}_1$ and $\overline{X}_2$; see Corollary 5.3.1). We will now describe a way of estimating the standard error of the sample contrast without estimating the covariance of $\overline{X}_1$ and $\overline{X}_2$.

Let $D_i = X_{1i} - X_{2i}$ denote the difference of the two observations on the ith individual, $i = 1, \ldots, n$, and let

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, \; S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2$$

denote the sample average and sample variance of these differences. The basic idea is that, since

$$\overline{D} = \overline{X}_1 - \overline{X}_2,$$

the standard error of $\overline{X}_1 - \overline{X}_2$ equals the standard error of $\overline{D}$. Thus,

$$\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2} = \widehat{\sigma}_{\overline{D}} = \frac{S_D}{\sqrt{n}}. \tag{10.4.1}$$

Moreover,

$$\mu_D = E(D_i) = E(X_i) - E(Y_i) = \mu_1 - \mu_2,$$

which means that $H_0 : \mu_1 - \mu_2 = 0$ is true if and only if $\tilde{H}_0 : \mu_D = 0$. But $\tilde{H}_0 : \mu_D = 0$ can be tested with the procedures for one sample we saw in Chapter 9. In particular we use the **paired t test** statistic

$$T_{H_0} = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t_{n-1}, \;\; [\text{under} H_0], \tag{10.4.2}$$

with corresponding rejection regions

H_a	RR at level α
$\mu_1 - \mu_2 > 0$	$T_{H_0} > t_{\alpha,n-1}$
$\mu_1 - \mu_2 < 0$	$T_{H_0} < -t_{\alpha,n-1}$
$\mu_1 - \mu_2 \neq 0$	$\lvert T_{H_0}\rvert > t_{\alpha/2,n-1}$.

If $n < 30$, the above paired t test is valid under the assumption that the D_i's are normal. If $n \geq 30$ the normality assumption is not needed. In this case the **paired Z test** can also be used. The paired Z test uses the statistic Z_{H_0}, which equals the paired t statistic in (10.4.2), and rejection regions as above but with the standard normal percentiles replacing the t percentiles.

A $(1-\alpha)100\%$CI for the difference of the two means based on paired data is

$$\overline{D} - t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \overline{D} + t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}. \tag{10.4.3}$$

If $n < 30$, the validity of this CI requires the assumption that the D_i's are normal. If $n \geq 30$ the normality assumption is not needed. In this case the paired Z interval can also be used. This has the form of the interval in (10.4.3) but with the standard normal percentiles replacing the t percentiles.

Example 10.4.3. Consider the study comparing two methods for determining the iron content in ore samples described in Example 10.4.2. A total of 12 ore samples are analyzed by both methods producing the paired data

Ore Sample	Method A	Method B	D
1	38.25	38.27	-0.02
2	31.68	31.71	-0.03
⋮	⋮	⋮	⋮
12	30.76	30.79	-0.03

yield $\overline{D} = -0.0167$ and $S_D = 0.02645$. We want to see if there is evidence (at $\alpha = .05$) that method B gives a higher average percentage than method A.

Solution. Here $H_0 : \mu_1 - \mu_2 = 0,\ H_a : \mu_1 - \mu_2 < 0$, and because the sample size is small, we must assume normality (or use the nonparametric small sample procedures of the next subsection). Assuming normality, the test statistic is

$$T_{H_0} = \frac{\overline{D}}{S_D/\sqrt{n}} = \frac{-0.0167}{.02645/\sqrt{12}} = -2.1865.$$

Since $T_{H_0} < -t_{.05,11} = -1.796$, H_0 is rejected.

10.4.2 The Signed-Rank Test for Paired Data

Under the null hypothesis $H_0^F : F_1 = F_2$, the distribution of the differences $D_i = X_{1i} - X_{2i}$ is symmetric about zero. Thus, for continuous data, the signed-rank procedure described in Subsection 9.4.4 can be applied on the differences for testing $H_0 : \widetilde{\mu}_D = 0$. Rejection of this hypothesis implies that $H_0^F : F_1 = F_2$ must also be rejected. The sign test of Subsection 9.4.3 can also be applied, but the signed-rank test, which uses the symmetry of the distribution of the D_i, is more powerful and thus it is preferable to the sign test.

For convenience we repeat here the steps for carrying out the signed rank test. First, compute the test statistic using the following steps.

1. Rank the absolute differences $|D_1|, \ldots, |D_n|$ from smallest to largest. Let R_i denote the rank of $|D_i|$.
2. Assign to R_i the sign of D_i, forming thus signed ranks.
3. Let S_+ be the sum of the ranks R_i with positive sign, i.e. the sum of the positive signed ranks.

Then, form the standardized test statistic using the following.

- If H_0 holds, $\mu_{S_+} = \frac{n(n+1)}{4}$, $\sigma^2_{S_+} = \frac{n(n+1)(2n+1)}{24}$.
- If H_0 holds, and $n > 10$, $\quad S_+ \overset{\cdot}{\sim} N(\mu_{S_+}, \sigma^2_{S_+})$.
- The TS for testing $H_0 : \mu_D = 0$ is
$$Z_{H_0} = \left(S_+ - \frac{n(n+1)}{4}\right) / \sqrt{\frac{n(n+1)(2n+1)}{24}},$$
- The RRs are the usual RRs of a Z-test.

Example 10.4.4. It is suspected that Lab A tends to underestimate the concentration of mercury in water samples. 12 water samples were analyzed for mercury content by Labs A and B. The 12 differences, D_i, and the ranks of their absolute values are given in the table below. Test $H_0 : \mu_1 - \mu_2 = 0$, $H_a : \mu_1 - \mu_2 < 0$ at $\alpha = 0.05$.

D_i	-0.0206	-0.0350	-0.0161	-0.0017	0.0064	-0.0219
R_i	5	10	4	1	2	6
D_i	-0.0250	-0.0279	-0.0232	-0.0655	0.0461	-0.0159
R_i	8	9	7	12	11	3

Solution: Here $S_+ = 2 + 11 = 13$. Thus

$$Z_{H_0} = \frac{13 - 39}{\sqrt{162.5}} = -2.04.$$

The p-value equals $\Phi(-2.04) = 0.02$ and thus, H_0^F is rejected. Note that the exact p-value, as obtained from Minitab, is 0.0193, indicating that the normal approximation is good.

10.4.3 Exercises

1. A study sponsored by the National Highway Institute considered considered the percent of soil passing a 3/8 inch sieve for soil taken from two separate locations. It is known that the percent of soil passing the sieve is affected by weather conditions. One measurement on each of the two types of soil was made on each of 32 different days. The data are:

Day :	1	2	3	4	5	6	7	8	9	10	11	12
Soil 1:	34.6	28.0	29.4	35.8	45.9	31.4	36.7	36.3	26.3	29.1	38.5	31.5
Soil 2:	47.1	30.9	35.1	49.2	51.8	32.1	49.3	33.4	26.4	27.2	29.3	37.2

Day :	13	14	15	16	17	18	19	20	21	22	23	24
Soil 1:	24.1	43.2	35.0	42.8	42.0	30.2	21.6	25.1	30.4	33.4	33.4	26.5
Soil 2:	29.1	39.8	38.0	50.4	53.5	35.7	24.6	21.2	34.2	38.0	28.7	22.1

Day :	25	26	27	28	29	39	31	32
Soil 1:	41.5	18.7	28.2	60.0	35.3	34.3	35.5	31.9
Soil 2:	36.6	26.7	27.1	47.4	45.9	34.3	57.6	27.5

   Test the hypothesis that the population mean percent of soil passing through the sieve is the same for the two locations against the two-sided alternative.

   (a) State the test statistic and the rejection region at level α.

   (b) What assumptions are needed for the validity of the test procedure in part a)?

   (c) Conduct the test at level $\alpha = 0.01$.

   (d) Construct a 99% CI for the difference of the population means.

2. Two brands of motorcycle tires are to be compared for durability. Eight motorcycles are selected at random and one tire from each brand is randomly assigned (front or back) on each motorcycle. The motorcycles are then run until the tires wear out. The data, in kilometers, are:

Motorcycle:	1	2	3	4	5	6	7	8
Brand 1:	36,925	45,300	36,240	32,100	37,210	48,360	38,200	33,500
Brand 2:	34,318	42,280	35,500	31,950	38,015	47,800	37,810	33,215

   (a) State the null and alternative hypotheses.

(b) Conduct, at level $\alpha = 0.05$, the contrast-based test.

(c) What assumptions are needed for the validity of the test procedure in part b)?

(d) Conduct, at level $\alpha = 0.05$, the signed-rank test.

(e) What assumptions are needed for the validity of the test procedure in part d)?

3. With the information given in Exercise 2 above, construct a 95% CI for the difference of the population means. Next, using Minitab, construct the 95% CI that corresponds to the signed-rank test.

4. Changes in the turbidity of a body of water is used by Environmental or Soils Engineer as an indication that surrounding land may be unstable causing sediments to be pulled into the water. The Wagner turbidity test was performed on water samples from ten locations around a lake, before and after a land stabilization project that included the planting of trees:

Location:	1	2	3	4	5	6	7	8	9	10
Before :	19.6	21.2	24.8	20.0	18.8	20.9	21.6	24.5	29.5	30.9
After :	18.6	19.6	22.0	19.1	17.4	19.7	19.8	22.2	27.7	28.6

The scientific question is whether the land stabilizing measures have reduced the turbidity.

(a) State the null and the alternative hypothesis.

(b) State a test statistic and the rejection region.

(c) What assumptions, if any are needed for the validity of the test procedure you suggested?

(d) Conduct the test at level $\alpha = 0.05$.

(e) Conduct the test assuming that the two samples are independent and compare the results with the those of the test in part d).

5. With the information given in Exercise 4 above, conduct the signed-rank test at level $\alpha = 0.05$. Next, assume that the two samples are independent and conduct the rank sum test at level $\alpha = 0.05$. Compare the results from the two tests.

10.5 Review Exercises

1. A study was undertaken to determine the truth of the claim that more Italians than British prefer white champagne to pink champagne at weddings. In a random sample of 100 Italians, 32 preferred white champagne, while in a random sample of 200 British, 50 preferred white champagne. Can we conclude that a higher proportion of Italians rather than Americans prefer white champagne at weddings? Use a significance level of 0.06.

(a) State the null and alternative hypotheses.

(b) Write down your test statistic, and compute its p-value.

(c) Use the p-value to conduct the hypothesis testing at level of significance $\alpha = 0.05$.

2. A study aimed at determining the extend to which plasma ascorbic acid is elevated in smokers, took measurements from a random sample five smokers and five nonsmokers males, all in the 25-30 age group. The data are:

Smokers	1.48	1.71	1.98	1.68	1.18	$s_1 = 0.297$
Nonsmokers	0.42	0.68	0.51	0.73	0.91	$s_2 = 0.192$

(a) The appropriate CI for the difference, $\mu_1 - \mu_2$, between the true average ascorbic acid values of smokers and nonsmokers is of the form (choose one):

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2}, \quad \text{or} \quad \overline{X}_1 - \overline{X}_2 \pm t_{\alpha/2,\nu}\widehat{\sigma}_{\overline{X}_1 - \overline{X}_2},$$

where ν denotes the appropriate degrees of freedom.

(b) State the basic assumption needed for the validity of the type of CI you chose above.

(c) Assuming, in addition, that the two population variances are equal, give a 95% CI of the type you chose above.

(d) Give a 95% CI of the type you chose above without the assumption that the two population variances are equal.

3. Two different analytical tests can be used to determine the impurity levels in steel alloys. The first test is known to perform very well, but the second test is cheaper than the first. A specialty steel manufacturer will adopt the second method unless there is evidence that it gives significantly different answers than the first. Eight steel specimens are cut in half and one half is randomly assigned to one test and the other half to the other test. The results are shown in the following table.

	Specimen							
Test	1	2	3	4	5	6	7	8
1	1.2	1.3	1.5	1.4	1.7	1.8	1.4	1.3
2	1.4	1.7	1.5	1.3	2.0	2.1	1.7	1.6

(a) State the appropriate null and alternative hypotheses.

(b) Which test procedure is correct for testing the above hypothesis? (Choose one.)

`two-sample z test; two-sample t test; paired z test; paired t test.`

(c) State any assumptions you must make to ensure the validity of the test you chose.

(d) Carry out the test using $\alpha = 0.05$, and state your conclusions.

(e) Construct the 95% CI which corresponds to the test you chose.

4. The Field Research Group of an Indianapolis based company is charged with not allowing any cost-reduced product to be put into the field before it has been shown to be at least as reliable as the product that it is replacing. A field test of AutoCaller 2.0 is conducted in Illinois. Two samples of 1000 each of the old AutoCaller 1.0 and AutoCaller 2.0 are put into service in randomly selected Illinois customers' homes. After 6 months, 52 of the old units and 66 of the new units have required some sort of repair.

   (a) State appropriate null and alternative hypotheses.

   (b) Carry out the test at level $\alpha = 0.05$. What do you conclude about marketing the new device?

5. An article in *Technometrics, Vol. 17, 161-166* reports on the results of a cloud seeding experiment. The question of interest is whether cloud seeding with silver nitrate increases rainfall. Out of 52 clouds, 26 were randomly selected for seeding, with the remaining 26 serving as controls. The rainfall measurements, in acre-feet, were

```
Control: 1202.6, 830.1, 372.4, 345.5, 321.2, 244.3, 163.0, 147.8,
      95.0, 87.0, 81.2, 68.5, 47.3, 41.1, 36.6, 29.0, 28.6, 26.3,
      26.1, 24.4, 21.7, 17.3, 11.5, 4.9, 4.9, 1.0
Seeded : 2745.6, 1697.8, 1656.0, 978.0, 703.4, 489.1, 430.0, 334.1,
      302.8, 274.7, 274.7, 255.0, 242.5, 200.7, 198.6, 129.6, 119.0,
      118.3, 115.3, 92.4, 40.6, 32.7, 31.4, 17.5, 7.7, 4.1
```

   (a) State the null and the alternative hypothesis.

   (b) Give an appropriate test statistic based on the actual observations, and the corresponding rejection region.

   (c) Is normality needed for the validity of the test procedure you described in part b)? If so, construct normal probability plots to assess if the normality assumption is supported by the data.

   (d) Carry out the test at $\alpha = 0.05$, and state the p-value.

6. Consider the context and information given in Exercise 5 above.

   (a) State the null and the alternative hypothesis.

   (b) Give an appropriate test statistic based on the ranks of the observations, and the corresponding rejection region.

   (c) Carry out the test at $\alpha = 0.05$, and state the p-value.

Chapter 11

Comparing k(> 2) Populations

11.1 Introduction

In this chapter we will consider the comparison of k population means or proportions. Scientists are often interested in comparing several different treatments to determine if one produces a significantly more desirable effect on the response than the others. For example, we may be of interest to compare the mean yield of a chemical process under $k = 4$ different temperature settings. Here, temperature is the *factor* and temperature settings are also called *factor levels*. Each factor level corresponds to a different population or treatment. We emphasize again that a cause-and-effect relationship can been established only if the allocation of experimental units to the treatments is done in a randomized fashion controlled by the experimenter, i.e. only when a statistical experiment is performed. Similarly, it may be of interest to compare $k > 2$ materials in terms of the proportions that remain intact when subjected to a certain extreme temperature.

Though still a simple kind of a comparative study, the comparison of $k > 2$ populations serves to introduce three important methodologies that will be used also in more complicated comparative studies. First, we will see how the hypothesis of equality of the k population means, or proportions, can be formulated in terms of contrasts, and how this formulation leads to a general chi-square test statistic. Second, the analysis of variance approach will be introduced here as an alternative testing method which is recommended for normal *homoscedastic* (i.e. having the same variance) populations. Finally, we will introduce the related concepts of *simultaneous* confidence intervals and *multiple comparisons*, which generalize the idea of a confidence interval. We begin by formulating a *statistical model*, which is commonly used for the problem of comparing $k > 2$ means of continuous populations.

11.2 The Statistical Model and Hypothesis

Let μ_i, σ_i^2, denote the mean and variance of population (or factor level) i, $i = 1, \ldots, k$. The comparison of the k populations will be based on a simple random sample from each populations. Let

$$X_{i1}, X_{i2}, \ldots, X_{in_i}, \ i = 1, \ldots, k, \tag{11.2.1}$$

denote the k samples. Thus, the random sample from population 1 has size n_1, with observations denoted by $X_{11}, \ldots, X_{1n_1}$, the random sample from population 2 has size n_2 with observations denoted by $X_{21}, \ldots, X_{2n_2}$, and so forth. This is the simplest comparative study for which it is common to write a **statistical model** for the generation of the observations X_{ij}. This modeling begins by writing each observation as the sum of its mean, μ_i, plus an **error term**, ϵ_{ij},

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad \text{where} \ \ \epsilon_{ij} = X_{ij} - \mu_i, \tag{11.2.2}$$

and further decomposing μ_i into an **overall mean**, μ, plus a **treatment effect**, α_i. The end result is

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \text{where} \ \ \mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i, \ \ \text{and} \ \ \alpha_i = \mu_i - \mu. \tag{11.2.3}$$

From their definitions it follows that the error term has mean value zero, and the treatment effects sum to zero. That is,

$$E\left(\epsilon_{ij}\right) = 0, \quad \text{and} \ \ \sum_{i=1}^{k}\alpha_i = 0. \tag{11.2.4}$$

Of interest here is the testing of

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \ \ \text{or, equivalently,} \ \ H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0, \tag{11.2.5}$$

against the alternative

$$H_a : H_0 \ \ \text{is false.} \tag{11.2.6}$$

Remark 11.2.1. Procedures for testing against one-sided alternatives, such as $H_a : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$, with at least one inequality strict, exist but are beyond the scope of this book.

Let

$$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{in_i}, \; S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i - 1} \left(X_{ij} - \overline{X}_i \right)^2, \tag{11.2.7}$$

be the sample mean and sample variance from the ith sample, $i = 1, \ldots, k$.

We will discuss four approaches to testing (11.2.5) against (11.2.6). One approach generalizes the nonparametric contrast-based approach which was discussed in the two-sample context. It is based on the set of $k - 1$ **contrasts**

$$\overline{X}_1 - \overline{X}_2, \; \overline{X}_1 - \overline{X}_3, \; \ldots, \; \overline{X}_1 - \overline{X}_k. \tag{11.2.8}$$

The contrast-based approach applies also to the comparison of k proportions. Another approach generalizes the rank-sum test. The analysis of variance approach, suitable for normal homoscedastic populations, will also be described. Finally, the (seemingly) simplest approach of all, namely performing all pairwise tests, will be presented through the concepts of multiple comparisons and simultaneous confidence intervals.

We close this section by introducing the **dot-notation**, which is common in factorial experiments. According to the dot-notation, a $\cdot$ replacing an index means summation over that index, while an over-line continues to denote averaging. This notation is useful when observations are labeled by more than one index, and there is need to sum or average over some of these indices. According to this notation,

$$\begin{aligned} X_{i\cdot} &= \sum_{j=1}^{n_i} X_{ij}, \; X_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} && (11.2.9) \\ \overline{X}_{i\cdot} &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} = \overline{X}_i, \quad \text{is the } i\text{-th sample mean, while} \\ \overline{X}_{\cdot\cdot} &= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k} \frac{n_i}{N} \overline{X}_i, \quad \text{is the overall sample average,} && (11.2.10) \end{aligned}$$

where $N = \sum_{i=1}^{k} n_i$ is the total sample size. In this chapter, we will use $\overline{X}_i$ and $\overline{X}_{i\cdot}$ interchangeably.

11.3 Nonparametric Contrast-Based Method

In this section we will describe the extension of the nonparametric contrast-based method and the rank-sum procedure to the comparison of more than two populations. These procedures are called nonparametric because they do not require the assumption of normality, or any other distributional assumption, and are valid for all ordinal data, homoscedastic or not.

11.3.1 Unequal Variances: The χ^2 Test and p-Value

In this subsection we will learn how to use the sample contrasts (11.2.8) for testing (11.2.5) against (11.2.6). The relevance of the sample contrasts in testing this hypothesis can be seen from the fact that the population contrasts, which the sample contrasts estimate, i.e.

$$\mu_1 - \mu_2,\ \mu_1 - \mu_3,\ \ldots,\ \mu_1 - \mu_k,$$

are all zero if and only if the null hypothesis is true. Implementation of the contrast-based approach requires some basic matrix operations.

A $k \times k$ matrix is called a *diagonal* matrix, if its off-diagonal elements are all zero. For any matrix A, A' denotes the *transposed* matrix, i.e. the matrix whose columns are the rows of A. Similarly, if $\mathbf{a}$ is a row vector, $\mathbf{a}'$ transposes if to a column vector and vice-versa.

A matrix is called a *contrast matrix* if the elements of each one of its rows sum to zero. For example, a $(k-1) \times k$ contrast matrix is

$$\mathbf{C}_k = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix}_{(k-1)\times k} \qquad (11.3.1)$$

As its name indicates, contrast matrices are used to generate sets of contrasts. Thus, if $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_k)'$, is the column vector of group sample means, then the multiplication

$\mathbf{C}_k\overline{\mathbf{X}}$ gives the set of contrasts (11.2.8):

$$\mathbf{C}_k\overline{\mathbf{X}} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix}_{(k-1)\times k} \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \overline{X}_3 \\ \vdots \\ \overline{X}_k \end{pmatrix}_{k\times 1} = \begin{pmatrix} \overline{X}_1 - \overline{X}_2 \\ \overline{X}_1 - \overline{X}_3 \\ \overline{X}_1 - \overline{X}_4 \\ \vdots \\ \overline{X}_1 - \overline{X}_k \end{pmatrix}_{(k-1)\times 1}$$

Let $\mathbf{V}$ be the $k \times k$ diagonal matrix whose diagonal elements are $\text{Var}(\overline{X}_i) = \sigma_i^2/n_i$. Also let $\widehat{\mathbf{V}}$ be the $k \times k$ diagonal matrix whose diagonal elements are the estimates, S_i^2/n_i, of $\text{Var}(\overline{X}_i) = \sigma_i^2/n_i$. In notation,

$$\mathbf{V} = \text{diag}\left(\frac{\sigma_1^2}{n_1}, \ldots, \frac{\sigma_k^2}{n_k}\right), \quad \text{and} \quad \widehat{\mathbf{V}} = \text{diag}\left(\frac{S_1^2}{n_1}, \ldots, \frac{S_k^2}{n_k}\right). \tag{11.3.2}$$

The test statistic for testing the null hypothesis in (11.2.5), i.e. $H_0 : \mu_1 = \cdots = \mu_k$, is

$$Q_k = (\mathbf{C}_k\overline{\mathbf{X}})'(\mathbf{C}_k\widehat{\mathbf{V}}\mathbf{C}_k')^{-1}(\mathbf{C}_k\overline{\mathbf{X}}). \boxed{\begin{array}{c}\text{Contrast-based test statistic} \\ \text{for testing } H_0 : \mu_1 = \cdots = \mu_k\end{array}}. \tag{11.3.3}$$

If H_0 holds and $n_i \geq 30$, for all i, Q_k has approximately a χ^2 distribution with $k-1$ degrees of freedom:

$$Q_k \mathrel{\dot{\sim}} \chi^2_{k-1}.$$

Thus, the test procedure rejects the null hypothesis if

$$Q_k > \chi^2_{k-1}(\alpha). \boxed{\begin{array}{c}\text{Region for rejecting } H_0 : \mu_1 = \cdots = \mu_k \\ \text{at level of significance } \alpha\end{array}} \tag{11.3.4}$$

Remark 11.3.1. *The set of population contrasts $\mu_1 - \mu_2, \ldots, \mu_1 - \mu_k$ are not the only ones that can characterize the null hypothesis $H_0 : \mu_1 = \ldots = \mu_k$. For example, if $\overline{\mu} = k^{-1}\sum_{i=1}^k \mu_i$, then H_0 is true if and only if the contrasts $\mu_1 - \overline{\mu}, \mu_2 - \overline{\mu}, \ldots, \mu_{k-1} - \overline{\mu}$ are all zero. It can be shown that the value of the test statistic Q_k does not depend on which set of contrasts is used to represent the null hypothesis.*

To define the p-value of the above χ^2-test, let Ψ_ν be the cumulative distribution function of the chi-square distribution with ν degrees of freedom. That is, if $X \sim \chi^2_\nu$, then $P(X \leq x) = \Psi_\nu(x)$. Then, the p-value of the chi square test is

$$p\text{-value} = 1 - \Psi_{k-1}(Q_k). \boxed{p\text{-value of the chi-square test}} \tag{11.3.5}$$

The procedure (11.3.4) generalizes the nonparametric (large-sample) Z-test for testing the equality of two population means versus the two-sided alternative, as the following example shows.

Example 11.3.1 (Application to the comparison of 2 population means). For the comparison of two populations, $\mathbf{C}_2$, the $\widehat{\mathbf{V}}$ of (11.3.2), and $(\mathbf{C}_2\widehat{\mathbf{V}}\mathbf{C}_2')^{-1}$ become

$$\mathbf{C}_2 = (1, -1),\ \widehat{\mathbf{V}} = \begin{pmatrix} \frac{S_1^2}{n_1} & 0 \\ 0 & \frac{S_2^2}{n_2} \end{pmatrix},\ (\mathbf{C}_2\widehat{\mathbf{V}}\mathbf{C}_2')^{-1} = 1/\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right).$$

Thus, since $\mathbf{C}_2\overline{\mathbf{X}} = \overline{X}_1 - \overline{X}_2$, it follows that the test statistic is

$$Q_2 = \frac{(\overline{X}_1 - \overline{X}_2)^2}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

and the null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected in favor of the alternative $H_a : \mu_1 \neq \mu_2$ if

$$Q_2 > \chi_1^2(\alpha). \tag{11.3.6}$$

Note that if Z_{H_0} is the test statistic (10.2.4) with $\Delta_0 = 0$, then

$$Q_2 = Z_{H_0}^2, \quad \text{and} \quad \chi_1^2(\alpha) = z_{\alpha/2}^2,$$

where the last equality above can be verified from the normal and χ^2 tables. Thus, the rejection region, given in (10.2.6) for testing $H_0 : \mu_1 = \mu_2$ against $H_0 : \mu_1 \neq \mu_2$, is the same as the rejection region in (11.3.6).

Example 11.3.2. The American Society for Testing Materials had a standard flammability test performed on five pieces from each of three types of fabric used in children's clothing. The response variable is the length of the burn mark made when the fabric piece is exposed to a flame in a specified way.

	Observations						$\overline{X}_i$	S_i^2
Material 1	1.56	2.12	1.90	2.07	1.92	1.94	1.918	0.0386
Material 2	1.72	1.69	1.87	1.78	1.79	1.91	1.793	0.0071
Material 3	1.62	2.08	2.07	1.99	1.95	1.93	1.94	0.0283

Test $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a : H_0$ is false, at $\alpha = 0.1$, and find the p-value of the statistic.

Solution. Here the sample sizes are not large enough, so ideally we should apply the rank based procedure, which is described in the next subsection. But for the sake of illustration, using

$$\mathbf{C}_3 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \ \widehat{\mathbf{V}} = \frac{1}{6} \begin{pmatrix} 0.0386 & 0 & 0 \\ 0 & 0.0071 & 0 \\ 0 & 0 & 0.0283 \end{pmatrix}, \text{ and } \overline{\mathbf{X}} = \begin{pmatrix} 1.918 \\ 1.793 \\ 1.94 \end{pmatrix},$$

the test statistic (11.3.3) for $k = 3$ takes the value

$$Q_3 = 4.9, \text{ which is larger than } \chi^2_2(0.1) = 4.605,$$

and thus the null hypothesis is rejected at $\alpha = 0.1$. Using a software package we find that the exact p-value is 0.086.

11.3.2 Equal Variances and the ANOVA Table

If the assumption that the population variances are equal ($\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$) appears tenable, it makes sense to use the pooled variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + \cdots + (n_k - 1)S_k^2}{n_1 + \cdots + n_k - k} \tag{11.3.7}$$

as the estimator of the common variance σ^2. In this case, the matrix $\widehat{\mathbf{V}}$ given in (11.3.2) becomes

$$\widehat{\mathbf{V}} = \text{diag}\left(\frac{S_p^2}{n_1}, \ldots, \frac{S_p^2}{n_k}\right). \tag{11.3.8}$$

With this substitution, the statistic Q_k of relation (11.3.3) has the simpler form:

$$Q_k = \frac{\sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}_{\cdot\cdot}\right)^2}{S_p^2}, \quad \boxed{\begin{array}{l}\text{Contrast-based test statistic for testing} \\ H_0 : \mu_1 = \cdots = \mu_k \text{ under homoscedasticity}\end{array}}, \tag{11.3.9}$$

where S_p^2 is the pooled sample variance given in (11.3.7), and $\overline{X}_{\cdot\cdot} = \sum_{i=1}^{k}(n_i/N)\overline{X}_i$ is the overall sample average defined in (11.2.10). As before, if the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ holds, $Q_k \overset{\cdot}{\sim} \chi^2_{k-1}$. Thus, the test procedure rejects the null hypothesis if

$$Q_k > \chi^2_{k-1}(\alpha). \quad \boxed{\begin{array}{l}\text{Region for rejecting } H_0 : \mu_1 = \cdots = \mu_k \\ \text{at level of significance } \alpha\end{array}}$$

Remark 11.3.2. *The assumption of equal population variances should only be made with care. Use of the simpler statistic (11.3.9) when the sample variances contradict the assumption of equal population variances can produce misleading results. Thus, with the data in Example 11.3.2, the statistic (11.3.9) is* $Q_3 = 3.058$*, yielding a p-value of 0.22, so* H_0 *is not rejected at* $\alpha = 0.1$*. In this example, the assumption of equal variances does not appear tenable since* $S_1^2/S_2^2 = 5.4$*.*

The ANOVA Table

Most software packages compute the statistic (11.3.9) through the analysis of variance (ANOVA) table, the construction of which we now describe.

First compute the **Treatment Sum of Squares** (SSTr) and the **Error Sum of Squares** (SSE) as follows:

$$\text{SSTr} = \sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 \quad \boxed{\text{Treatment Sum of Squares}} \tag{11.3.10}$$

$$\text{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2 \quad \boxed{\text{Error Sum of Squares}} \tag{11.3.11}$$

SSTr equals the numerator of Q_k in (11.3.9). Next, form the **Treatment Mean sum of Squares** (MSTr) and **Error Mean sum of Squares** (MSE) by dividing SSTr and SSE by their **degrees of freedom**, which are $k-1$ and $N-k$, respectively, where $N = n_1 + \cdots + n_k$:

$$\text{MSTr} = \frac{\text{SSTr}}{k-1}, \quad \text{MSE} = \frac{\text{SSE}}{N-k}. \tag{11.3.12}$$

MSE equals the pooled variance S_p^2 given in (11.3.7).

Remark 11.3.3. 1. The numerator of the sample variance of the combined data is called the **Total Sum of Squares** (SST): $\text{SST} = \sum_i \sum_j (X_{ij} - \overline{X}_{\cdot\cdot})^2$.

2. The identity TSS = TrSS + ESS is responsible for the name Analysis of Variance.

The ANOVA table is essentially a summary of these calculations:

Source	df	SS	MS	F
Treatment	$k-1$	SSTr	MST$=\dfrac{\text{SSTr}}{k-1}$	$F=\dfrac{\text{MST}}{\text{MSE}}$
Error	$N-k$	SSE	MSE$=\dfrac{\text{SSE}}{N-k}$	
Total	$N-1$	SST		

The Q_k statistic is obtained from the ANOVA table as follows:

$$Q_k = (k-1)F. \tag{11.3.13}$$

Example 11.3.3. The following data resulted from comparing the degree of soiling in fabric treated with three different mixtures of methacrylic acid.

Mix 1	.56	1.12	.90	1.07	.94
		$X_{1\cdot} = 4.59$, $\overline{X}_{1\cdot} = .918$			
Mix 2	.72	.69	.87	.78	.91
		$X_{2\cdot} = 3.97$, $\overline{X}_{2\cdot} = .794$			
Mix 3	.62	1.08	1.07	.99	.93
		$X_{3\cdot} = 4.69$, $\overline{X}_{3\cdot} = .938$			

Test $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a : H_0$ is false, at $\alpha = 0.1$.

Solution. For the sake of illustration, we use the present nonparametric procedure even though the sample sizes in this application are small. A computer printout gives the following ANOVA table.

Source	df	SS	MS	F
Treatment	$k-1=2$	.0608	.0304	.99
Error	$N-k=12$	.3701	.0308	
Total	$N-1=14$	.4309		

Thus, $Q_3 = (3-1) \times (0.99) = 1.98$. Since $\chi^2_2(0.1) = 4.605$, H_0 is not rejected.

11.3.3 Comparing k (>2) Proportions

When we want to compare different types of cars in terms of the proportion that sustain no damage in 5 miles/hour crash tests, or when we want to compare different age groups of the voting population in terms of the proportion of approval of a certain political reform, or when we want to compare the proportion of defective products of a given production process in different weeks, we end up wanting to test

$$H_0 : p_1 = p_2 = \cdots = p_k = p \text{ versus } H_a : \; H_0 \text{ is not true.}$$

Since the probability, p, of a 1, or of a ”success”, in a Bernoulli experiment is also the mean value of the Bernoulli random variable, this testing problem is a special case of testing for the equality of k means. The only difference is that the variance of a Bernoulli random variable is related to its mean: $\sigma^2 = p(1-p)$. Because the null hypothesis specifies that the probabilities are the same in all k experiments, it follows that, under the null hypothesis, we have homoscedasticity:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = p(1-p).$$

Thus, we only need the simper form of the Q_k statistic, namely the form given in (11.3.9). Substituting the sample means $\overline{X}_i$ by sample proportions $\widehat{p}_i$, the overall sample mean $\overline{X}_{\cdot\cdot}$ by the overall, or pooled, sample proportion

$$\widehat{p} = \sum_{i=1}^{k} \frac{n_i}{N} \widehat{p}_i, \tag{11.3.14}$$

and the pooled sample variance S_p^2 by $\widehat{p}(1-\widehat{p})$, the statistic Q_k takes the form

$$Q_k = \sum_{i=1}^{k} \frac{n_i \left(\widehat{p}_i - \widehat{p}\right)^2}{\widehat{p}(1-\widehat{p})}. \tag{11.3.15}$$

Thus, the test procedure rejects the null hypothesis if

$$Q_k > \chi^2_{k-1}(\alpha). \quad \boxed{\begin{array}{l}\text{Region for rejecting } H_0 : p_1 = \cdots = p_k \\ \text{at level of significance } \alpha.\end{array}}$$

Even though the expression of Q_k in (11.3.15) is quite simple, an equivalent form of it is more common. To describe this equivalent form, we need the following additional notation:

$$O_{1i} = n_i \widehat{p}_i, \qquad O_{2i} = n_i \left(1 - \widehat{p}_i\right) \tag{11.3.16}$$

$$E_{1i} = n_i \widehat{p}, \qquad E_{2i} = n_i \left(1 - \widehat{p}\right), \tag{11.3.17}$$

where $\widehat{p}$ is the overall sample proportion defined in (11.3.14). Thus, O_{1i} is the observed number of successes (or 1s) in the ith group, O_{2i} is the observed number of failures (or

0s) in the ith group, and E_{1i}, E_{2i} are the corresponding expected numbers under the null hypothesis that the probability of 1 is the same in all groups. With this notation, the alternative form for Q_k, called the *contingency table* form, is

$$Q_k = \sum_{i=1}^{k} \sum_{\ell=1}^{2} \frac{\left(O_{\ell i} - E_{\ell i}\right)^2}{E_{\ell i}}. \tag{11.3.18}$$

The equivalence of the two expressions, (11.3.15) and (11.3.18), for Q_k is easily seen by noting the following easy algebraic identities:

$$\begin{aligned} \left(O_{1i} - E_{1i}\right)^2 &= \left(O_{2i} - E_{2i}\right)^2 = n_i^2 \left(\widehat{p}_i - \widehat{p}\right)^2 \\ \frac{1}{E_{1i}} + \frac{1}{E_{2i}} &= \frac{E_{1i} + E_{2i}}{E_{1i}E_{2i}} = \frac{1}{n_i \widehat{p}\left(1 - \widehat{p}\right)}. \end{aligned}$$

Thus, for each i,

$$\sum_{\ell=1}^{2} \frac{\left(O_{\ell i} - E_{\ell i}\right)^2}{E_{\ell i}} = n_i^2 \left(\widehat{p}_i - \widehat{p}\right)^2 \left(\frac{1}{E_{1i}} + \frac{1}{E_{2i}}\right) = \frac{n_i \left(\widehat{p}_i - \widehat{p}\right)^2}{\widehat{p}(1 - \widehat{p})},$$

which shows the equivalence of the expressions (11.3.15) and (11.3.18).

Example 11.3.4. A commercial airline is considering four different designs of the control panel for the new generation of airplanes. To see if the designs have an effect on the pilot's response time to emergency displays, emergency conditions were simulated and the response times of pilots were recorded. The sample sizes, n_i, and number of times, O_{1i}, that the response times were below 3 seconds for the four designs are as follows: $n_1 = 45$, $O_{11} = 29$; $n_2 = 50$, $O_{12} = 42$; $n_3 = 55$, $O_{13} = 28$; $n_4 = 50$, $O_{14} = 24$. Perform the test at level of significance $\alpha = 0.05$.

Solution. Here the null hypothesis is $H_0 : p_1 = p_2 = p_3 = p_4$ and the alternative is $H_a : H_0$ *is not true.* The level of significance is $\alpha = 0.05$, so that, the rejection region is $Q_4 > \chi^2_3(0.05) = 7.815$. With the data given,

$$\widehat{p} = \frac{O_{11} + O_{12} + O_{13} + O_{14}}{n_1 + n_2 + n_3 + n_4} = \frac{123}{200} = 0.615,$$

so the common denominator of the terms in the expression (11.3.15) for Q_4 is $\widehat{p}(1 - \widehat{p}) = 0.2368$. Thus, the value of Q_4 is

$$\begin{aligned} Q_4 &= \frac{45(0.6444 - 0.615)^2}{0.2368} + \frac{50(0.84 - 0.615)^2}{0.2368} \\ &\quad + \frac{55(0.5091 - 0.615)^2}{0.2368} + \frac{50(0.48 - 0.615)^2}{0.2368} \\ &= 0.1643 + 10.6894 + 2.6048 + 3.8482 = 17.307. \end{aligned}$$

Since $17.307 > 7.815$, the null hypothesis is rejected.

11.3.4 Exercises

1. A study conducted at the Delphi Energy and Engine Management Systems considered the effect of blow-off pressure, during the manufacture of spark plugs, on the spark plug resistance. The resistance measurements of 150 spark plugs manufactured at each of three different blow-off pressures yield the following summary statistics:

	N	Sample mean	Sample St.Dev.
10psi:	150	5.365	2.241
12psi:	150	5.415	1.438
15psi:	150	5.883	1.065

   (a) Use the contrast based method with unequal variances to test the hypothesis that the three corrugated containers have, on average, the same strength. Use $\alpha = 0.05$.

   (b) Use the contrast based method, assuming equal variances, to test the hypothesis that the three corrugated containers have, on average, the same strength. Use $\alpha = 0.05$.

   (c) Compare the p-values of the tests in parts a) and b).

   (d) Perform the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly at $\alpha = 0.05$.

2. The article "Compression of Single-Wall Corrugated Shipping Containers Using Fixed and Floating Test Platens" in the *Journal of Testing and Evaluation, 1992, Vol. 20, No. 4, 318-320*, considered the compression strength of five different types of corrugated fiberboard containers in order to determine the effect of the platen on the results. Suppose that the fixed platen strength measurements for three types of corrugated containers yielded the following summary statistics: For type A, a sample of size 36 yielded $\overline{X}_1 = 754$ and $S_1 = 16$; for type B, a sample of size 49 yielded $\overline{X}_2 = 769$ and $S_2 = 27$; for type C, a sample of size 42 yielded $\overline{X}_3 = 776$ and $S_3 = 38$.

   (a) Use the contrast based method with unequal variances to test the hypothesis that the three corrugated containers have, on average, the same strength. Use $\alpha = 0.05$.

   (b) Use the contrast based method, assuming equal variances, to test the hypothesis that the three corrugated containers have, on average, the same strength. Use $\alpha = 0.05$.

   (c) Compare the p-values of the tests in parts a) and b).

   (d) Perform the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly at $\alpha = 0.05$.

3. The flame resistance of three materials used in children's pajamas was tested by subjecting specimens of the materials to high temperatures. Out of 111 specimens of material A, 37 ignited. Out of 85 specimens of material B, 28 ignited. Out of 100 specimens of material C, 21 ignited.

    (a) Test the hypothesis that the probability of ignition is the same for all three materials, versus the alternative that this hypothesis is false.

    (b) Perform the Bonferroni multiple comparisons method to identify which population proportions differ significantly at $\alpha = 0.05$.

4. A certain type of John Deere tractor is assembled in five different locations. To see if the proportion of tractors that require warranty repair work, is the same for all locations, a random sample of 50 tractors from each location is selected and followed up for the duration of the warranty period. The numbers requiring warranty repair work are: 18 for type A, 8 for type B, 21 for type C, 16 for type D, and 13 for type E.

    (a) Test the hypothesis that the five population proportions are the same $\alpha = 0.05$.

    (b) Perform the Bonferroni multiple comparisons method to identify which population proportions differ significantly at $\alpha = 0.05$.

5. Wind-born debris (from roofs, passing trucks, insects or birds) can wreak havoc on architectural glass on upper-stories of a building. The paper "Impact resistance of laminated glass using "sacrificial ply" design concept" in the *Journal of Architectural Engineering*, 2000, 24-34, reports the results of an experiment, where 10 configurations of glass were subjected to a 2 gram steel ball projectile traveling under 5 impact velocity ranges. Here we report the results for configurations 1, 2, 3 and 5. Out of 105 inner glass ply breaks (IPBs) of configuration 1, 91 were at impact velocity of 139 ft/s or less. For configurations 2, 3 and 5 the results were 128 out of 148, 46 out of 87, and 62 out of 93.

    (a) Test the hypothesis that the five population proportions are the same $\alpha = 0.05$.

    (b) Perform the Bonferroni multiple comparisons method to identify which population proportions differ significantly at $\alpha = 0.05$.

11.4 Nonparametric Rank-Based Methods

In this subsection we will describe two test procedure which are based on ranks and can be used with both small and large sample sizes. One of them is the *Kruskal-Wallis* test, while the other uses contrasts on the average group ranks. The popularity of these procedures is also due to their desirable power properties (i.e. low probability of type II error), especially if the two population distributions are heavy tailed, or skewed.

Let $\mu_1, \ldots, \mu_k$ denote the k population means, and $F_1, \ldots, F_k$ denote the cumulative distribution functions of the k populations. We will concentrate on testing the hypothesis

$$H_0^F : F_1 = \cdots = F_k. \tag{11.4.1}$$

Note that if H_0^F is true, then so is the hypothesis of equality of the two population means, $H_0 : \mu_1 = \cdots = \mu_k$.

The calculation of the rank test statistics begins by combining the observations, $X_{i1}, \ldots, X_{in_i}$, $i = 1, \ldots, k$, from the k samples into an overall set of $N = n_1 + \cdots + n_k$ observations, and arrange them from smallest to largest. The assignment of ranks, or mid-ranks, to a set of observations is carried out as discussed in subsection 10.2.2. Let R_{ij} denote the rank, or mid-rank of observation X_{ij}, and set

$$\overline{R}_i = n_i^{-1} \sum_j R_{ij}, \ S_{R,i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(R_{ij} - \overline{R}_i\right)^2, \tag{11.4.2}$$

for the average rank in group i, and the sample variance of the ranks in group i.

11.4.1 The Kruskal-Wallis Test

This test applies if there are no ties (continuous populations). In that case, the null distribution of the Kruskal-Wallis test statistic is known even with very small sample sizes. However, this exact null distribution depends on the group sample sizes, so it requires extensive tables for its presentation. Since a fairly accurate approximation of this exact distribution is possible even with small (≥ 8) sample sizes, tables are not given.

With the notation introduced in (11.4.2), the Kruskal-Wallis test statistic is

$$KW_k = \frac{1}{S_{KW}^2} \sum_{i=1}^{k} n_i \left(\overline{R}_i - \frac{N+1}{2}\right)^2, \quad \boxed{\begin{array}{c}\text{General Form of the}\\ \text{Kruskal-Wallis test statistic}\end{array}} \tag{11.4.3}$$

where

$$S_{KW}^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(R_{ij} - \frac{N+1}{2}\right)^2, \tag{11.4.4}$$

is the sample variance of the collection of all ranks.

If there are no tied observations, the Kruskal-Wallis test statistic can also be rewritten as

$$KW_k = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2 . \quad \boxed{\begin{array}{c}\text{Kruskal-Wallis test statistic}\\ \text{when there are no ties}\end{array}} \tag{11.4.5}$$

Remark 11.4.1. *The equivalence of the formulas (11.4.3) and (11.4.5) in the case of no ties follows from the identity*

$$\sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 = \frac{(N-1)N(N+1)}{12}.$$

Even with small group sample sizes, the null distribution of the test statistic KW_k is approximately chi-square with $k-1$ degrees of freedom. Thus, the test procedure rejects $H_0^F : F_1 = \cdots = F_k$ is

$$KW_k > \chi^2_{k-1}(\alpha). \quad \boxed{\begin{array}{c}\text{Rejection region for the Kruskal-Wallis}\\ \text{test at level of significance } \alpha\end{array}} \tag{11.4.6}$$

The p-value of the Kruskal-Wallis statistic is computed as in (11.3.5) with Q_k replaced by KW_k.

Example 11.4.1. Consider the setting of Example 11.3.2, and use the Kruskal-Wallis statistic to test the hypothesis that the flammability of the three materials is the same. Use $\alpha = 0.1$, and find the p-value.

Solution: The ranks, rank averages and rank variances of the flammability data are

	Ranks						$\overline{R}_i$	$S^2_{R,i}$
Material 1	1	18	8	15.5	10	12	10.75	35.98
Material 2	4	3	7	5	6	9	5.667	4.67
Material 3	2	17	15.5	14	13	11	12.08	28.64

Here $N = 18$, so $(N+1)/2 = 9.5$. The main component of the Kruskal-Wallis statistic is

$$\sum_{i=1}^{k} n_i \left(\overline{R}_i - \frac{N+1}{2} \right)^2 = 6(10.75 - 9.5)^2 + 6(5.667 - 9.5)^2 + 6(12.08 - 9.5)^2 = 137.465.$$

Moreover, the sample variance of the combined set of ranks is $S^2_{KW} = 28.47$. Thus, the Kruskal-Wallis statistic is

$$KW_3 = 137.465/28.47 = 4.83.$$

Since 4.83 is larger than $\chi^2_2(0.1) = 4.605$, the null hypothesis is rejected at $\alpha = 0.1$. Using the formula in (11.3.5), with Q_k replaced by KW_3, and a software package we find that the exact p-value is 0.089. We close this example by noting that, in this data set, the observation 2.07 in group 1 is tied with an observation in group 3. Thus, the expression in (11.4.5) is not valid. On the other hand, there is only one tie, so the difference between the two expressions should not be great. Indeed, $12/[N(N+1)] = 0.03509$, which is very close to $1/S^2_{KW} = 0.03512$. Thus, expression (11.4.5) also yields a value of 4.83.

11.4.2 The Rank-Contrasts Method

The contrast-based rank test statistics apply to data with or without ties. Their exact distribution is not known, but it can also be approximated fairly accurately even with small (≥ 8) group sample sizes.

The contrast-based rank statistic can be computed by using the ranks instead of the observations in the contrast-based statistic Q_k in (11.3.3), i.e.

$$QR_k = (\mathbf{C}_k\overline{\mathbf{R}})'(\mathbf{C}_k\widehat{\mathbf{V}}(\mathbf{R})\mathbf{C}_k')^{-1}(\mathbf{C}_k\overline{\mathbf{R}}), \quad \boxed{\begin{array}{c}\text{The rank-contrasts statistic}\\ \text{for data with or without ties}\end{array}} \tag{11.4.7}$$

where $\overline{\mathbf{R}} = \left(\overline{\mathbf{R}}_1, \ldots, \overline{\mathbf{R}}_k\right)'$, is the column vector of group rank averages, and $\widehat{\mathbf{V}}(\mathbf{R})$ uses the rank sample variances $S_{R,i}$ instead of the sample variances S_i^2, $i = 1, \ldots, k$; thus

$$\widehat{\mathbf{V}}(\mathbf{R}) = \text{diag}\left(\frac{S^2_{R,1}}{n_1}, \ldots, \frac{S^2_{R,k}}{n_k}\right).$$

If the group rank variances are not too dissimilar (as they would be under H_0^F) then it makes sense to use the pooled variance,

$$S^2_{R,p} = \frac{(n_1-1)S^2_{R,1} + \cdots + (n_k-1)S^2_{R,k}}{N-k}. \tag{11.4.8}$$

In this case,

$$\widehat{\mathbf{V}}(\mathbf{R}) = \text{diag}\left(\frac{S^2_{R,p}}{n_1}, \ldots, \frac{S^2_{R,p}}{n_k}\right),$$

and, similar to Section 11.3.2, the statistic QR_k takes the simpler form

$$QR_k = \frac{1}{S^2_{R,p}}\sum_{i=1}^{k} n_i\left(\overline{R}_i - \frac{N+1}{2}\right)^2. \quad \boxed{\begin{array}{c}\text{The rank-contrasts statistic with}\\ \text{the pooled group rank variace}\end{array}} \tag{11.4.9}$$

Remark 11.4.2. *The above form of the rank-contrasts statistic makes transparent its difference from the Kruskal-Wallis statistic. Thus, the only difference between the statistics in (11.4.3) and in (11.4.9) is in the way the common (under H_0^F) group rank variance is estimated. Under the alternative hypothesis, however, the group rank variances are not the same, and the two expressions estimate different quantities. It can be shown that under location shift alternatives, $S_{R,p}^2$ tends to be smaller than S_{KW}^2, which implies that the rank-contrasts statistic in (11.4.9) is more powerful than the Kruskal-Wallis statistic. Moreover, when the variances are unequal, it is known that the weighted statistic (11.4.7) is more powerful than the unweighted one in (11.4.9). Finally, it should be pointed out that scale alternatives also induce unequal group rank variances, making the weighted statistic (11.4.7) less sensitive to them.*

The null distribution of the test statistic QR_k is approximately chi-square with $k-1$ degrees of freedom. Thus, the test procedure rejects $H_0^F : F_1 = \cdots = F_k$ if

$$QR_k > \chi^2_{k-1}(\alpha). \quad \boxed{\begin{array}{c}\text{Rejection region for the contrast-based}\\ \text{rank test at level of significance } \alpha\end{array}} \tag{11.4.10}$$

The p-value of the rank-contrasts statistic is computed as in (11.3.5) with Q_k replaced by QR_k.

Example 11.4.2. Consider the setting of Example 11.3.2, and use the rank-contrasts statistic to test the hypothesis that the flammability of the three materials is the same. Use $\alpha = 0.1$, and find the p-value.

Solution: The ranks, rank averages and rank variances of the flammability data have already been given in Example 11.4.1. From the same example we have

$$\sum_{i=1}^{k} n_i \left(\overline{R}_i - \frac{N+1}{2}\right)^2 = 6(10.75-9.5)^2 + 6(5.667-9.5)^2 + 6(12.08-9.5)^2 = 137.465.$$

Next, the pooled group rank variance of relation (11.4.8) is

$$S_{R,p}^2 = \frac{5 \times 35.98 + 5 \times 4.67 + 5 \times 28.64}{18-3} = 23.0967.$$

Thus, the rank-contrasts statistic that uses the pooled rank variance (the unweighted one in (11.4.9)) is

$$QR_3 = 137.465/23.0967 = 5.95$$

Since 5.95 is larger than $\chi^2_2(0.1) = 4.605$, the null hypothesis is rejected at $\alpha = 0.1$. Using the formula in (11.3.5) with Q_k replaced by QR_3, and a software package, we find that the exact p-value is 0.051. Omitting the detailed calculations, the weighted statistic (11.4.7) is $QR_3 = 10.04$, which corresponds to a p-value of 0.0066. Note that this p-value make the weighted statistic significant even at $\alpha = 0.01$.

11.4.3 Exercises

1. Porous carbon materials are used commercially in several industrial applications, including gas separation, membrane separation, and fuel cell applications (*J. Phys. Chem. B*, 1997, *101*, 3988; *Microporous Mater*, 1994, *4*, 407; *Nature*, 2001, *412*, 169). For the purpose of gas separation, the pore size distribution is important. An experiment conducted to compare the pore size distribution when the carbon is made at four different temperatures (300^{o}C, 400^{o}C, 500^{o}C, and 600^{o}C) yielded the following results:

300 :	7.75	7.0	6.9	7.5	8.0
400 :	6.9	7.6	7.0	7.0	7.7
500 :	6.7	6.9	6.9	6.2	6.6
600 :	6.4	6.2	6.0	6.6	6.0

   Of interest is to test the hypothesis that the population mean of pore size is the same for all four temperatures, i.e. $H_0 : \mu_1 = \cdots = \mu_4$ vs $H_a : H_0$ is not true.

   (a) What assumptions, if any, are needed for the validity of the Kruskal-Wallis test?

   (b) What assumptions, if any, are needed for the validity of the rank-contrast test?

   (c) Conduct the rank-contrast test, at $\alpha = 0.05$, assuming equal variances.

2. Records of teaching an honors statistics class in Experimental Design for engineers, indicated that different Professors had adopted one of three different teaching methods: A) Use of a textbook as the main source of teaching material, B) Use of a textbook combined with computer activities, and C) Use of specially designed instructional notes together with computer activities. It was decided to compare the three teaching methods by randomly dividing 24 students in three groups of eight, to receive each of the three teaching methods. A common exam was administered in the end. The scores for the three groups were

A:	71	76	68	72	69	91	77	81
B:	73	78	75	80	84	93	88	79
C:	92	82	85	86	98	95	73	87

Of interest is to test the hypothesis that the three methods are equally effective, i.e. $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a : H_0$ is not true.

(a) What assumptions, if any, are needed for the validity of the rank based procedures (Kruskal-Wallis, or rank-contrasts)?

(b) Conduct the rank-contrasts with the pooled group rank variance at level $\alpha = 0.05$.

(c) Use Minitab to conduct the Kruskal-Wallis test. Compare the p-value with that obtained in part b).

3. The article *Improving the thermal fatigue resistance of brake discs* in the journal *Materials, Experimentation and Design in Fatigue, 1981, 60-71* reports on a study using high temperature strain gages to measure the total strain amplitude of three different types of cast iron (gray (G), spheroidal graphite (SG) and compacted graphite (CG)) for use in disc brakes. The results (multiplied by 10,000) are:

```
SG: 105  77  52  27  22  17  12  14  65
CG:  90  50  30  20  14  10  60  24  76
G :  40  20  15  25   8  10  30  19  28
```

The scientific question is whether the total amplitude strain properties of the different types of cast iron differ. Conduct the rank-contrast test, at $\alpha = 0.05$, with the assumption of equal variances.

4. The article ”Flexural fatigue behavior of threaded connections for large diameter pipes” in the journal *Experimental Mechanics*, 2002, *42* 1-7, reports on a study where fatigue tests were performed by subjecting the threaded connection of large diameter pipes to constant amplitude stress of either 10 ksi, 15 ksi, 18 ksi and 22 ksi. The measured fatigue lives, in number of cycles to failure are (the number of cycles have been divided by 10,000):

```
10.0 ksi: 485.22 195.40 411.11
12.5 ksi:  55.34  37.20
15.0 ksi:  40.59  76.45  22.56
18.0 ksi:  92.70  33.31  16.51
22.0 ksi:   7.45   4.50   7.06   7.21
```

(a) Are the conditions, need for a valid application of the contrast based method (on the original data) with unequal population variances, met? Justify your answer.

(b) Conduct the rank-contrast test, at $\alpha = 0.05$, without the assumption of equal variances.

(c) Use Minitab to conduct the Kruskal-Wallis test. Compare the p-value with that obtained in part b).

11.5 The F-Test Under the Normality Assumption

Under the assumption that the k population distributions are all normal with the same variance, it is possible to test the hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ even with very small group sample sizes. Before we describe this exact testing procedure, we highlight the two assumptions that must hold for its validity.

Homoscedasticity Assumption: The k populations variances are equal: $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$.

Normality Assumption: The k population distributions are normal:

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad \text{for all } j = 1, \ldots, n_i.\ i = 1, \ldots, k.$$

If the above two assumptions hold, the exact distribution of the F statistic given in the ANOVA table (see Section 11.3.2) is known to be F with numerator degrees of freedom $k-1$ and denominator degrees of freedom $N-k$, where $N = n_1 + \cdots + n_k$. That is,

$$F = \frac{MST}{MSE} \sim F_{k-1,N-k}. \quad \boxed{\begin{array}{l}\text{Exact distribution of the ANOVA}\\ \text{F-statistic for testing } H_0 : \mu_1 = \cdots = \mu_k\end{array}} \tag{11.5.1}$$

Thus, the rejection region is

$$F > F_{\alpha,k-1,N-k}. \quad \boxed{\begin{array}{l}\text{Region for rejecting } H_0 : \mu_1 = \cdots = \mu_k\\ \text{at level of significance } \alpha\end{array}} \tag{11.5.2}$$

Example 11.5.1. Consider again the experiment and the data of Example 11.3.3. Assuming that the three populations are normal and homoscedastic, test $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_a : H_0$ is false, at $\alpha = 0.1$.

Solution. Recall that the ANOVA table in this case is

Source	df	SS	MS	F
Treatment	$k-1=2$	.0608	.0304	.99
Error	$k(n-1)=12$	.3701	.0308	
Total	$nk-1=14$	.4309		

The rejection rule specifies that H_0 be rejected if $F > F_{0.1,2,12}$. Since $F_{0.1,2,12} = 2.81$, H_0 is not rejected.

11.5.1 Exercises

1. Consider the setting and data of Exercise 11.4.3,1.
   (a) What assumptions, if any, are needed for the validity of the ANOVA F-test?
   (b) Conduct the F-test at level $\alpha = 0.05$.
2. Consider the setting and data of Exercise 11.4.3,2.
   (a) What assumptions, if any, are needed for the validity of the ANOVA F-test?
   (b) Conduct the F-test at level $\alpha = 0.05$.
3. (a) Consider the setting and data of Exercise 11.4.3,3. Do the assumptions of the ANOVA F-test appear to be satisfied for this data set? Justify your answer.
   (b) Consider the setting and data of Exercise 11.4.3,4. Do the assumptions of the ANOVA F-test appear to be satisfied for this data set? Justify your answer.

11.6 Multiple Comparisons and Simultaneous CIs

Because the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is tested against the alternative $H_a : H_0$ is false, it follows that when H_0 is rejected it is not clear which μ_i's are significantly different. It would seem that this question can be addressed quite simply by making confidence intervals for all pairwise differences $\mu_i - \mu_j$, $i \neq j$, and checking whether or not each interval contains zero. If a confidence interval does not contain zero, then the corresponding means are declared significantly different. With some fine-tuning, this approach leads to a correct *multiple comparisons* method.

The reason why the above simple procedure needs fine-tuning has to do with the **overall**, or **experiment-wise error rate**, which is the probability of at least one pair of means being declared different when all means are equal.

To appreciate the experiment-wise error rate, suppose that the problem involves the comparison of $k = 5$ population means. If the null hypothesis is rejected, then, to determine which pairs of means are significantly different, confidence intervals for all 10 pairwise differences,

$$\mu_1 - \mu_2, \ldots, \mu_1 - \mu_5, \mu_2 - \mu_3, \ldots, \mu_2 - \mu_5, \ldots, \mu_4 - \mu_5, \tag{11.6.1}$$

must be made. Assume for the moment that the confidence intervals for the 10 contrasts in (11.6.1) are independent. (They are not independent because the confidence intervals for, say, $\mu_1 - \mu_2$ and $\mu_1 - \mu_3$ both involve the sample $X_{11}, \ldots, X_{1n_1}$ from population 1.) In that case, the null probability that all 10 confidence intervals contain zero is $(1-\alpha)^{10}$. Thus, the experiment-wise error rate is $1-(1-\alpha)^{10}$. If $\alpha = 0.05$, then the experiment-wise error rate is

$$1 - (1 - 0.05)^{10} = 0.401. \tag{11.6.2}$$

It turns out that, in spite of the dependence of the confidence intervals, the above calculation gives an fairly close approximation to the true experiment-wise error rate. Thus, the chances are approximately 40% that at least one of comparisons will be declared significantly different when using traditional 95% confidence intervals. Confidence intervals which control the experiment-wise error rate at a desired level α will be called $(1-\alpha)100\%$ **simultaneous confidence intervals**.

We will see two methods of fine-tuning the naive procedure of using traditional confidence intervals in order to bring to experiment-wise error rate to a desired level α. One method is based on *Bonferroni's inequality*, which gives an upper bound on the experiment-wise error rate. The other is *Tukey*'s procedure, which gives the exact experiment-wise error rate in the case of sampling normal homoscedastic populations, but can also be used as a good approximation with large samples from any distribution. Moreover, Tukey's method can be applied on the ranks, with smaller sample sizes, when sampling from skewed homoscedastic distributions.

11.6.1 Bonferroni Multiple Comparisons and Simultaneous CIs

The idea behind Bonferroni's intervals is to adjust the level of the traditional confidence intervals in order to achieve the desired experiment-wise error rate. As mentioned in connection to the calculation in (11.6.2) above, due to the dependence among the confidence intervals, it is not possible to know the exact experiment-wise error rate when performing a total of m confidence intervals. However, **Bonferroni's inequality** asserts that, when each of m confidence intervals are performed at level α, then the probability that at least one does not contain the true value of the parameter, i.e. the experiment-wise error rate, is no greater than $m\alpha$. Similarly, if each of m pairwise tests are performed at level α, the experiment-wise level of significance (i.e. the probability of rejecting at least one of the m null hypotheses when all are true), is no greater than $m\alpha$.

Referring to the calculation in (11.6.2), if each confidence interval is constructed at level 0.05/10, and if the confidence intervals were constructed independently, the experiment-wise error rate is

$$1-(1-\frac{0.05}{10})^{10}=1-(1-0.005)^{10}=0.0489,$$

which, indeed, is no larger than 0.05.

The above discussion leads to the following procedure for constructing Bonferroni simultaneous CIs and multiple comparisons

Proposition 11.6.1.

- **$(1-\alpha)100\%$ Bonferroni simultaneous CIs for m contrasts:** *For each contrast construct a $(1-\alpha/m)100\%$ CI. This set of m CIs are the $(1-\alpha)100\%$ Bonferroni simultaneous CIs for the m contrasts.*
- **Bonferroni multiple comparisons at level α:** *If any of the m $(1-\alpha)100\%$ Bonferroni simultaneous CIs does not contain zero, the corresponding contrast is declared significantly different from zero at experiment-wise level α. Equivalently, if any of the m pairwise tests, conducted at level α/m rejects the null hypothesis that the corresponding contrast is zero, that contrast is declared significantly different from zero at experiment-wise level α.*

Example 11.6.1. In the context of Example 11.3.4, test the null hypothesis that the panel design has no effect on the pilot reaction time, i.e. $H_0 : p_1 = p_2 = p_3 = p_4$ vs H_a : H_0 is not true, at level $\alpha = 0.05$ using Bonferroni multiple comparisons.

Solution: In order to test the null hypothesis with multiple comparisons, and determine which panel designs have a significantly different effect on the reaction time, we construct 95% Bonferroni simultaneous CIs for the contrasts

$$p_1-p_2,\ p_1-p_3,\ p_1-p_4,\ p_2-p_3,\ p_2-p_4,\ p_3-p_4.$$

If one of these CIs does not contain zero, the null hypothesis is rejected at level 0.05. The CIs that do not contain zero correspond to contrasts that are significantly different from zero. Because there are $m = 6$ contrasts, we construct $(1-0.05/6)100\% = 99.17\%$ CIs for each of the above contrasts. Recall that the data for the three panel designs are: $n_1 = 45$, $O_{11} = 29$; $n_2 = 50$, $O_{12} = 42$; $n_3 = 55$, $O_{13} = 28$; $n_4 = 50$, $O_{14} = 24$. The resulting CIs are:

Contrast	99.17% CI	Contains zero?
$p_1 - p_2$	(-0.428, 0.0373)	Yes
$p_1 - p_3$	(-0.124, 0.394)	Yes
$p_1 - p_4$	(-0.101, 0.429)	Yes
$p_2 - p_3$	(0.106, 0.555)	No
$p_2 - p_4$	(0.129, 0.591)	No
$p_3 - p_4$	(-0.229, 0.287)	Yes

Thus, p_2 is significantly different, at experiment-wise level $\alpha = 0.05$, from p_3 and p_4. All other contrasts are not significantly different from zero.

Example 11.6.2. Consider the data in Exercise 11.4.3,2 on the exams scores for students exposed to three different teaching methods. Use the Bonferroni multiple comparisons procedure, based on the rank-sum test, to identify the methods achieving significantly, at $\alpha = 0.05$, better score results.

Solution: In this example, we are interested only in multiple comparisons, not in simultaneous CIs. Because the desired experiment-wise error rate is 0.05, we will conduct each of the $m = 3$ pair-wise comparisons (A vs B, A vs C, and B vs C) at level $0.05/3 = 0.0167$. If the p-value of one of these comparisons is smaller than 0.0167, the corresponding methods are declared different at level $\alpha = 0.05$. The results from the tree rank-sum tests are summarized in the following table:

Comparison	p-value	Less than 0.0167?
A vs B	0.104	No
A vs C	0.0136	Yes
B vs C	0.1415	No

Thus, methods A and C are significantly different at $\alpha = 0.05$, but methods A and B, as well as methods B and C are not significantly different.

11.6.2 Tukey's Multiple Comparisons and Simultaneous CIs

These intervals are appropriate under normality and homoscedasticity, i.e. under the assumptions of Section 11.5. If group sample sizes are all large (≥ 30), they are approximately valid without the normality assumption, though the homoscedasticity assumption is still needed.

Tukey's intervals are based on the so-called studentized range distribution which is characterized by a numerator degrees of freedom, and a denominator degrees of freedom (denoted by ν). The numerator degrees of freedom equals the number of means, k, that are being compared. The denominator degrees of freedom equals $N - k$, where $N = n_1 + \cdots + n_k$ (which equals the degrees of freedom corresponding to the pooled sample variance, S_p^2, or, equivalently, to the SSE in the ANOVA table). Critical points of the studentized range distribution are given in Table A.7. Tukey's simultaneous CIs and multiple comparisons are as follows.

Proposition 11.6.2. *Let S_p^2 be the pooled sample variance given in (11.3.7) (so that $S_p^2 =$ MSE of the ANOVA table), and let $Q_{\alpha,k,N-k}$ denote the upper-tail α critical value of the studentized range distribution with k and $\nu = N - k$ degrees of freedom. Then*

- $(1-\alpha)100\%$ **Tukey's simultaneous CIs for all contrasts $\mu_i - \mu_j$, $i \neq j$:**

$$\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - Q_{\alpha,k,N-k}\sqrt{\frac{S_p^2}{2}(\frac{1}{n_i} + \frac{1}{n_j})} \leq \mu_i - \mu_j \leq \overline{X}_{i\cdot} - \overline{X}_{j\cdot} + Q_{\alpha,k,N-k}\sqrt{\frac{S_p^2}{2}(\frac{1}{n_i} + \frac{1}{n_j})}$$

- **Tukey's multiple comparisons at level α:** *If for a pair (i,j), $i \neq j$ the interval*

$$\overline{X}_{i\cdot} - \overline{X}_{j\cdot} \pm Q_{\alpha,k,N-k}\sqrt{\frac{S_p^2}{2}(\frac{1}{n_i} + \frac{1}{n_j})}$$

  does not contain zero, it is concluded that μ_i and μ_j differ significantly at level α.

The following steps for carrying out Tukey's procedure lead to an organized way of presenting the results from all pairwise comparisons.

1. Select α and find $Q_{\alpha,k,N-k}$ from Table A.7.
2. Calculate $w = Q_{\alpha,k,N-k}\sqrt{\frac{S_p^2}{2}(\frac{1}{n_i} + \frac{1}{n_j})}$.
3. List the sample means in increasing order and underline each pair that differs by less than w. Pairs that are not underlined indicate that the corresponding population means differ significantly at level α.

Example 11.6.3. Four different concentrations of ethanol are compared, at level $\alpha = 0.01$, for their effect on sleep time. Each concentration was given to a sample of 5 rats and the REM sleep time for each rat was recorded. The resulting four sample means are $\overline{X}_{1\cdot} = 79.28$, $\overline{X}_{2\cdot} = 61.54$, $\overline{X}_{3\cdot} = 47.92$, $\overline{X}_{4\cdot} = 32.76$. The remaining details of the data set are summarized in the ANOVA table:

Source	df	SS	MS	F
Treatment	3	5882.3575	1960.78583	21.09
Error	16	1487.4000	92.9625	
Total	19	7369.7575		

Assuming that the four populations are normal, with the same variance, we can use the F-test. Here, $F_{0.01,3,16} = 4.20$ and since $21.09 > 4.20$, $H_0 : \mu_1 = \cdots = \mu_4$ is rejected at level $\alpha = 0.01$.

To identify which of the means differ we apply Tukey's procedure. Since $S_p^2 = \text{MSE} = 92.9625$ and $Q_{.01,4,16} = 5.19$, we have $w = 5.19\sqrt{92.96/5} = 22.38$. The three steps outlined above yield the following summary of the multiple comparisons results:

$\overline{X}_{4\cdot}$	$\overline{X}_{3\cdot}$	$\overline{X}_{2\cdot}$	$\overline{X}_{1\cdot}$
32.76	47.92	61.54	79.28

Thus, $(\mu_1, \mu_3), (\mu_1, \mu_4)$, and (μ_2, μ_4) are significantly different at $\alpha = .01$.

11.6.3 Tukey's Multiple Comparisons on the Ranks

For non-normal data, it is preferable to use Tukey's multiple comparisons procedure on the (mid-)ranks, rather than on the original observations. To apply this procedure, rank the combined data as described in Section 11.4, and apply Tukey's multiple comparisons procedure on these ranks as described in the steps following Proposition 11.6.2 (not forgetting to substitute the pooled rank variance $S_{R,p}^2$, given in (11.4.8), for S_p^2). Note that now the simultaneous CIs are not relevant to the contrasts $\mu_i - \mu_j$; only the multiple comparisons are relevant.

Example 11.6.4. Consider the setting of Example 11.3.2, and use Tukey's multiple comparisons method on the ranks to identify which materials are significantly different in terms of flammability at level $\alpha = 0.1$.

Solution: The ranks, rank averages and rank variances of the flammability data for the $k = 3$ materials are given in Example 11.4.1. From that we get that the pooled rank variance (11.4.8) is $S_{R,p}^2 = 23.1$. Next, $Q_{0.1,3,15} = 3.14$, so that $w = 3.14\sqrt{23.1/6} = 6.16$

Since $S_p^2 = \text{MSE} = 92.9625$ and $Q_{.01,4,16} = 5.19$, we have $w = 5.19\sqrt{92.96/5} = 22.38$. The three steps outlined above yield the following summary of the multiple comparisons results:

$\overline{R}_3$	$\overline{R}_1$	$\overline{R}_2$
12.08	10.75	5.667

Thus, only materials 2 and 3 are significantly different at $\alpha = 0.1$.

11.6.4 Exercises

1. For the data in Exercise 11.3.4,1 use the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly, at $\alpha = 0.05$, by performing all pair-wise nonparametric contrast-based tests.

2. For the data in Exercise 11.3.4,2 use the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly, at $\alpha = 0.05$, by performing all pair-wise nonparametric contrast-based tests.

3. For the data in Exercise 11.3.4,3 use the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly, at $\alpha = 0.05$, by performing all pair-wise nonparametric contrast-based tests for proportions.

4. For the data in Exercise 11.3.4,4 use the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly, at $\alpha = 0.05$, by performing all pair-wise nonparametric contrast-based tests for proportions.

5. For the data in Exercise 11.3.4,5 use the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly, at $\alpha = 0.05$, by performing all pair-wise nonparametric contrast-based tests for proportions.

6. For the data in Exercise 11.4.3,3 use the Bonferroni multiple comparisons method to identify the groups whose population means differ significantly, at $\alpha = 0.05$, by performing all pair-wise rank sum tests.

7. For the data in Exercise 11.4.3,4 apply Bonferroni's multiple comparisons procedure, at experiment-wise error rate of 0.05, using the rank-sum test.

8. Consider the setting and data of Exercise 11.4.3,1.

   (a) Perform Tukey's multiple comparison method to determine the temperatures with (statistically) significantly different, at level $\alpha = 0.05$, mean pore sizes.

   (b) Perform Tukey's multiple comparison procedure on the ranks to determine the temperatures with significantly different, at level $\alpha = 0.05$, mean pore sizes.

9. Consider the setting and data of Exercise 11.4.3,2.

   (a) Perform Tukey's multiple comparison method to determine which teaching methods are (statistically) significantly different at level $\alpha = 0.05$.

   (b) Perform Tukey's multiple comparison procedure on the ranks to determine which teaching methods are (statistically) significantly different at level $\alpha = 0.05$.

11.7 Review Exercises

1. As part of the investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three different positions on the roof. The forces required to shear each of these bolts (coded values) are as follows

   Position 1: 90 82 79 93 83
   Position 2: 105 93 98 104 95
   Position 3: 83 89 98 94 86

   The question of interest is whether or not the three positions require the same force. The data are entered into Minitab columns C1, C2, and C3, for Positions 1, 2, and 3, respectively, and the ANOVA table method of analysis was applied.

   (a) Fill in the missing entries in the following ANOVA Table.

   ```
   Source  DF     SS       MS        F         P
   Factor   2   478.5   (    )  (    )   0.009
   Error   (  ) (      )  (    )
   Total   14   875.7
   ```

   (b) On the basis of the ANOVA Table, would you reject the null hypothesis at level $\alpha = 0.05$? Justify your answer.

   (c) An engineer who does not understand the ANOVA methodology wants to conduct three tests, each at level $\alpha = 0.05$, based on the individual contrasts $\overline{X}_1 - \overline{X}_2$, $\overline{X}_1 - \overline{X}_3$, $\overline{X}_2 - \overline{X}_3$. Is this a good idea? Justify your answer.

   (d) A simple modification of the above naive idea of the statistically uneducated engineer can turn it into a proper statistical procedure for multiple comparisons. Give the name of this procedure and perform this multiple comparisons procedure.

   (e) The following Minitab output gives Tukey's 95% simultaneous CIs. Use these CIs to perform Tukey's multiple comparisons at experiment-wise error rate of 0.05.

```
C1 subtracted from:
      Lower   Center   Upper   ------+---------+---------+---------+---
C2    3.900   13.600  23.300                  (-------*-------)
C3   -5.100    4.600  14.300           (-------*-------)
                               ------+---------+---------+---------+---
                                   -12         0        12        24
C2 subtracted from:
       Lower   Center   Upper  ------+---------+---------+---------+---
C3   -18.700  -9.000   0.700   (--------*-------)
                               ------+---------+---------+---------+---
                                  -12         0        12        24
```

2. Consider the setting and information provided in Exercise 1.

   (a) Perform the Kruskal-Wallis test at $\alpha = 0.05$.

   (b) Perform Bonferroni's multiple comparisons procedure, at experiment-wise error rate of 0.05, using the rank-sum test procedure.

3. In an industrial experiment an engineer is interested in how the mean absorption of moisture in concrete varies among five concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate. The data are recorded in the following table.

		1	2	3	4	5	6	$x_{i\cdot}$	$x_{i\cdot}^2$	$\bar{x}_{i\cdot}$
Aggregate	1	551	457	450	731	499	632	3320	11022400	553.33
Aggregate	2	595	580	508	583	633	517	3416	11669056	569.33
Aggregate	3	639	615	511	573	648	677	3663	13417569	610.50
Aggregate	4	417	449	517	438	415	555	2791	7789681	465.17
Aggregate	5	563	631	522	613	656	679	3664	13424896	610.67
							Total	16854	46301202	

   (a) State the null and the alternative hypotheses.

   (b) Fill the empty spots in the following ANOVA table.

```
Source  DF     SS       MS       F
Factor ( ) (      )  (    )  (    )
Error  ( ) (      )  (    )
Total   29  209.377
```

   (c) Conduct the F test at level 0.05 to test H_0 vs H_1.

   (d) Use Tukey's procedure to investigate differences in true average absorption of moisture in concrete aggregates. Report your conclusions in the standard format.

4. Consider the setting of the above Exercise 3.

   (a) Using the data for aggregates 1, 4, and 5, perform the rank-contrasts test procedure, at level 0.05, without pooling the variances.

   (b) Using the data for aggregates 1, 4, and 5, perform the rank-contrasts test procedure, at level 0.05, with the variances pooled.

   (c) Using the data for aggregates 1, 4, and 5, perform Bonferroni's multiple comparisons procedure, at experiment-wise error rate of 0.05, using the rank-sum statistic.

5. As part of a study on the rate of combustion of artificial graphite in humid air flow, researchers conducted an experiment to investigate oxygen diffusivity through a water vapor mixture. An experiment was conducted with mole fraction of water at three levels. The data are shown in the following table

Mole Fraction of H_2O		
.0022	.017	.08
1.68	1.69	1.72
1.98	1.99	2.02
2.30	2.31	2.35
2.64	2.65	2.70
3.00	3.01	3.06
3.38	3.39	3.45
3.78	3.79	3.85
4.19	4.21	4.27
4.63	4.64	4.71

Does variation in mole fraction of water have any effect on true average diffusivity? Use the ANOVA table method to test, $\alpha = 0.05$, if the variation in mole fraction of water has any effect on true average diffusivity. What assumptions are needed for the validity of this test procedure?

Chapter 12

Randomized Block Designs

12.1 Introduction

A randomized block design is a generalization of the paired data design, which we saw in Section 10.4. It is an alternative design used to compare $k \geq 2$ populations, which correspond to different treatments, methods, or product types. It differs from the design of Chapter 11 in that the samples from the different treatments are not independent. Thus, the difference between the randomized block design and the design used in Chapter 11 is analogous to the difference between the paired data design and the two-sample design that uses two independent samples; see Chapter 10.

A randomized block design arises when a random sample of n individuals (subjects or objects) receives each of the k treatments that are to be compared. Because all observations are based on the same subjects or objects, it follows that the k samples are not independent, and thus the methods discussed in Chapter 11 cannot be used. The subjects or objects used are called **blocks**. A block design is called **randomized** if the order in which the k treatments are applied is randomized for each block. The term **randomized complete block design** is also used to emphasize the fact that each block receives all k treatments.

The following examples highlight two contexts where such data arise.

Example 12.1.1. Four different types of truck tires, A, B, C and D, are to be compared for durability. One way of designing this comparative experiment is to select a random sample of n trucks and fit each of them with one tire of each type. The locations (front left, front right, rear left, and rear right) where each of the tire types are fitted are selected random for each truck. After a pre-specified number of miles on the road, the tires are evaluated for ware and tear. In this example, the sample of n trucks are the blocks, and the four populations to be compared correspond to the four tire types. From each block four measurements are made, which are quantifications of wear and tear of each tire type. Because of the specific way that a truck affects the wear and tear of its tires (load, road conditions etc), the four measurements from each block cannot be assumed independent. However, measurements from different trucks can be assumed independent. Another design for this comparative study is to use tire type A on a random sample of n_1 trucks, fit a different sample of n_2 trucks with tire type B, a different sample of n_3 trucks for tire type C, and a different sample of n_4 trucks from tire type D, and from each truck record the average wear and tear of its four tires. This design gives four independent samples, and the methods of Chapter 11 can be used.

Example 12.1.2. Three different methods for determining the percentage of iron in ore samples are to be compared. A randomized block design for this comparative study consists of obtaining n ore samples and subject each of them to the three different methods for determining the iron content. The order in which the three methods are applied is randomized for each ore sample. In this example, the n ore samples are the blocks, and the populations that are compared correspond to the three different methods. For each ore sample there will be three measurements which are dependent, because they depend on the ore sample's true iron content. (Note that because the ore samples are selected at random, their true iron content is a random variable.) The design of Chapter 11 uses different ore samples for each method, resulting in three independent samples.

From the above two examples it follows that the a randomized block design eliminates a lot of uncontrolled variability in the measurements. In Example 12.1.1, the randomized block design eliminates the uncontrolled variability caused by the trucks having different loads, traveling different routs at different speeds. Similarly, in Example 12.1.2, the randomized block design eliminates the uncontrolled variability caused by the different iron content of the various ore samples.

A randomized block design can also be viewed as a two-factor design, with the blocks serving as the levels of the "block" factor. Because of the within each block randomization, the assumption of additivity (see Section 4.7.2) appears tenable. Moreover, the methods of analysis presented in the following sections apply to any two-factor design (i.e. not necessarily block design) with one observation per cell, provided the assumption of additivity holds; see Remark 12.2.1.

12.2 The Statistical Model and Hypothesis

Let X_{ij}, $j = 1, \ldots, k$, denote the k observations from block i, where the index j enumerates the treatments. Thus, the data set for the randomized block design consists of the k-tuples

$$(X_{i1}, X_{i2}, X_{i3}, X_{i4}), \quad i = 1, \ldots, n, \tag{12.2.1}$$

which are assumed to be independent. Figure 12.1 shows the data for a randomized block design with four treatments.

	Treatments			
Blocks	1	2	3	4
1	X_{11}	X_{12}	X_{13}	X_{14}
2	X_{21}	X_{22}	X_{23}	X_{24}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}

Figure 12.1: Data display for a randomized block design

The observations for each treatment (each column in Figure 12.1) are also assumed to have the same distribution which depends on the treatment; that is, for each j, $j = 1, \ldots, k$,

$$X_{1j}, X_{2j}, \ldots, X_{nj}, \quad \text{are iid} \quad F_j. \tag{12.2.2}$$

The main modeling assumption is that the different treatments (or different F_j) have the same variance. Thus, it is assumed that for all i and all j,

$$\text{Var}\,(X_{ij}) = \sigma^2 \boxed{\begin{array}{c}\text{Assumption of equal}\\ \text{variances}\end{array}}. \tag{12.2.3}$$

The observations within each block can be dependent, but the test procedures we will describe use the additional modeling assumption that all pairs of observations have the same covariance. (Randomization within each block makes this assumption tenable.) That is, we assume that for any two pairs of treatments (j_1, j_2), (j_3, j_4), and blocks i_1, i_2

not necessarily different,

$$\mathrm{Cov}\left(X_{i_1j_1}, X_{i_1j_2}\right) = \mathrm{Cov}\left(X_{i_2j_3}, X_{i_2j_4}\right) = \sigma_a^2 \quad \boxed{\begin{array}{c}\text{Assumption of equal}\\ \text{covariances}\end{array}}. \tag{12.2.4}$$

For example, any two observations from block 1 have the same covariance, which is the same with the covariance of any two observations from block 2.

Assumptions (12.2.4) and (12.2.3) imply that any two observations within the same block are equally correlated. These assumptions are implied from the following model

$$X_{ij} = \mu + a_i + \beta_j + \epsilon_{ij}, \;\; E(a_i) = 0, \;\; \sum_j \beta_j = 0, E(\epsilon_{ij}) = 0, \tag{12.2.5}$$

where μ is the overall mean, a_i denotes the **random effect** of block i, β_j denotes the fixed effect of treatment j, and ϵ_{ij} is the intrinsic error associated with observation X_{ij}. The random block effects a_i are iid with variance σ_a^2, the intrinsic errors are iid with variance σ_ϵ^2, and the random block effects are independent from the intrinsic errors.

Remark 12.2.1. *The comparison of the treatment effects presented in the next sections remains the same even if the block effects are fixed instead of random; that is, even under the model*

$$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \;\; \sum_i \alpha_i = 0, \;\; \sum_j \beta_j = 0, E(\epsilon_{ij}) = 0, \tag{12.2.6}$$

which is the model for a two-factor design without interaction (see Section 4.7.2). This will be discussed further in Chapter **??**.

As mentioned in the introductory section, the objective in a randomized block design is to test equality/equivalence of the k treatments. If

$$\mu_j = E\left(X_{ij}\right), \;\; \text{so that, according to (12.2.5),} \;\; \mu_j = \mu + \beta_j, \tag{12.2.7}$$

Interest lies in testing the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k, \;\; \text{or} \;\; \beta_1 = \cdots = \beta_k = 0. \tag{12.2.8}$$

In terms of contrasts, this hypothesis can be written as

$$H_0 : \mu_1 - \mu_2 = \mu_1 - \mu_3 = \cdots = \mu_1 - \mu_k = 0, \tag{12.2.9}$$

or, using the contrast matrix $\mathbf{C}_k$ defined in (11.3.1), as

$$H_0 : \mathbf{C}_k \boldsymbol{\mu} = \mathbf{0}, \tag{12.2.10}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$. The next sections provide procedures for doing so.

12.3 The Large-Sample Contrast-Based Procedure

In this section we will describe a contrast-based test procedure which requires a sample size of $n \geq 30$. This procedure uses only the modeling assumptions (12.2.4) and (12.2.3). A procedure that holds for any sample size, but requires in addition the normality assumption, is presented in Section 12.4. Let $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_k)'$ be the column vector of group sample means (thus, $\overline{X}_j = n^{-1} \sum_{i=1}^{n} X_{ij}$), so that the sample contrasts

$$\mathbf{C}_k \overline{\mathbf{X}} = (\overline{X}_1 - \overline{X}_2, \overline{X}_1 - \overline{X}_3, \ldots, \overline{X}_1 - \overline{X}_k)' \tag{12.3.1}$$

are estimators of the contrasts in (12.2.9). Recall that, for each j, $X_{1j}, X_{2j}, \ldots, X_{nj}$ are iid with variance σ^2 (see (12.2.2) and (12.2.3)), that observations in different blocks are independent (see (12.2.1)), and the covariance of any two observations within the same block is σ_a^2. From these facts it follows that

$$\text{Var}\left(\overline{X}_j\right) = \frac{\sigma^2}{n}, \quad \text{and} \quad \text{Cov}\left(\overline{X}_{j_1}, \overline{X}_{j_2}\right) = \frac{\sigma_a^2}{n}, \quad \text{for all} \quad j, \quad \text{and} \quad j_1 \neq j_2.$$

These relations imply that, for $j_1 \neq j_2$,

$$\text{Var}\left(\overline{X}_1 - \overline{X}_j\right) = 2\frac{\sigma_\epsilon^2}{n}, \quad \text{and} \quad \text{Cov}\left(\overline{X}_1 - \overline{X}_{j_1}, \overline{X}_1 - \overline{X}_{j_2}\right) = \frac{\sigma_\epsilon^2}{n}, \tag{12.3.2}$$

where $\sigma_\epsilon^2 = \sigma^2 - \sigma_a^2$ is the variance of the intrinsic error in model (12.2.5). An unbiased and consistent estimator of σ_ϵ^2 is

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{n-1} \frac{1}{k-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(X_{ij} - \overline{X}_{\cdot j} - \overline{X}_{i\cdot} + \overline{X}_{\cdot\cdot}\right)^2, \tag{12.3.3}$$

where according to the dot notation introduced in (11.2.9), (11.2.10), $\overline{X}_{\cdot j}$ is the sample mean from the jth treatment (which we also denote by $\overline{X}_j$), $\overline{X}_{i\cdot} = k^{-1} \sum_{j=1}^{k} X_{ij}$, and $\overline{X}_{\cdot\cdot}$ is the average of all kn observations. Relations (12.3.2) and (12.3.3) imply that the estimated variance-covariance matrix of the vector of estimated contrasts (12.3.1) is

$$\widehat{\mathbf{\Sigma}}_k = \frac{1}{n} \widehat{\sigma}_\epsilon^2 \begin{pmatrix} 2 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 2 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & 2 \end{pmatrix}_{(k-1)\times(k-1)} ,$$

and its inverse is

$$\widehat{\mathbf{\Sigma}}_k^{-1} = \frac{n}{k\widehat{\sigma}_\epsilon^2}\begin{pmatrix} k-1 & -1 & -1 & -1 & \cdots & -1 \\ -1 & k-1 & -1 & -1 & \cdots & -1 \\ -1 & -1 & k-1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & -1 & \cdots & k-1 \end{pmatrix}_{(k-1)\times(k-1)}.$$

Then, the statistic for testing H_0 in (12.2.8) is

$$\begin{aligned} Q_k &= \left(\mathbf{C}_k\overline{\mathbf{X}}\right)'\widehat{\mathbf{\Sigma}}_k^{-1}\left(\mathbf{C}_k\overline{\mathbf{X}}\right) = \frac{n}{k\widehat{\sigma}_\epsilon^2}k\sum_{j=1}^{k}\left(\overline{X}_j - \overline{X}_{\cdot\cdot}\right)^2 \\ &= \frac{\sum_{j=1}^{k} n\left(\overline{X}_j - \overline{X}_{\cdot\cdot}\right)^2}{\widehat{\sigma}_\epsilon^2}. \qquad (12.3.4) \end{aligned}$$

Under the null hypothesis in (12.2.8), $Q_k \overset{\cdot}{\sim} \chi^2_{k-1}$, so that H_0 is rejected at level α if

$$Q_k > \chi^2_{k-1,\alpha}, \quad \boxed{\begin{array}{c}\text{Large sample region for rejecting} \\ H_0 : \mu_1 = \cdots = \mu_k \text{ at level } \alpha\end{array}}. \qquad (12.3.5)$$

Example 12.3.1. Each of a random sample of 36 Napa Valley visitors tested and rated four wine varieties on a scale of 1-10. For impartiality purposes, the wines were identified only by numbers 1-4. The order in which each of the four wines were presented was randomized for each visitor. The summary statistics from the resulting data set are $\overline{X}_1 = 8.97$, $\overline{X}_2 = 9.05$, $\overline{X}_3 = 8.36$, $\overline{X}_4 = 8.31$, $\overline{X}_{\cdot\cdot} = 8.67$, and $\sigma_\epsilon^2 = 0.36$. Is there a significant difference in the rating of the four wines? Test at $\alpha = 0.05$

Solution: Plugging the given summary statistics into the formula (12.3.4), we have

$$\begin{aligned} Q_4 &= \frac{36\left[(8.97-8.67)^2 + (9.05-8.67)^2 + (8.36-8.67)^2 + (8.31-8.67)^2\right]}{0.36} \\ &= \frac{36 \times 0.4601}{0.36} = 46.01. \end{aligned}$$

Since $\chi^2_{3,0.05} = 7.815$, we conclude that the difference in the ratings is significant at $\alpha = 0.05$. (The p-value here is 0 to the first three decimal places, so the difference is significant at any of the common α levels.)

12.4 ANOVA and the F-Statistic

In this section we assume that the random effects and the error terms in the model (12.2.5) have the normal distribution. (In the case of the fixed effects model (12.2.6) assume that the error terms have a normal distribution.) Under this assumption, the exact distribution of a multiple of the statistic Q_k is known. In particular,

$$F_k = \frac{1}{k-1} Q_k \sim F_{k-1,(k-1)(n-1)}, \quad \boxed{\begin{array}{c}\text{Exact distribution of the ANOVA}\\ \text{F-statistic under normality}\end{array}} \tag{12.4.1}$$

where $F_{k-1,n-1}$ denotes the F-distribution with $k-1$ and $n-1$ degrees of freedom. Thus, the rejection region is

$$F_k > F_{\alpha,k-1,(k-1)(n-1)}. \quad \boxed{\begin{array}{c}\text{The ANOVA region for rejecting}\\ H_0 : \mu_1 = \cdots = \mu_k \text{ at level } \alpha\end{array}} \tag{12.4.2}$$

Software packages present the F_k statistic in the context of the ANOVA table which is based on the decomposition of the total sum of squares into block, treatment and error sum of squares:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \overline{X}_{\cdot\cdot}\right)^2 \;&=\; \sum_{i=1}^{n} k\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2 + \sum_{j=1}^{k} n\left(\overline{X}_{j} - \overline{X}_{\cdot\cdot}\right)^2 \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \overline{X}_{\cdot j} - \overline{X}_{i\cdot} + \overline{X}_{\cdot\cdot}\right)^2,\end{aligned} \tag{12.4.3}$$

or, symbolically,

$$SS_{\text{Total}} = SS_{\text{Blocks}} + SS_{\text{Treatments}} + SS_{\text{Error}}.$$

Below we will use the abbreviated notation SST, SS_{Bl}, SS_{Tr} and SSE for the total, block, treatment and error sum of squares, respectively.

The basic structure of the ANOVA table for a randomized block design is the same as that of Section 11.3.2, except that there is one more row corresponding to the block effects. In particular the ANOVA table now takes the form

Source	df	SS	MS	F
Blocks	$n-1$	SS_{Bl}	$MS_{\text{Bl}} = \dfrac{SS_{\text{Bl}}}{n-1}$	$F_{\text{Bl}} = \dfrac{MS_{\text{Bl}}}{MSE}$
Treatment	$k-1$	SS_{Tr}	$MS_{\text{Tr}} = \dfrac{SS_{\text{Tr}}}{k-1}$	$F_k = \dfrac{MS_{\text{Tr}}}{MSE}$
Error	$(n-1)(k-1)$	SSE	$MSE = \dfrac{\text{SSE}}{(n-1)(k-1)}$	
Total	$nk-1$	SST		

Remark 12.4.1. *The F-statistic F_{Bl} can be used to test the null hypothesis of no block effect (or the hypothesis of no row-factor effect in the additive fixed effects design (12.2.6)). This null hypothesis is rejected at level α if $F_{Bl} > F_{\alpha,n-1,(n-1)(k-1)}$.*

Example 12.4.1. The ANOVA table for the wine tasting data of Example 12.3.1 is

Source	df	SS	MS	F	
Blocks	35	11.1146	0.3176	0.89	0.642
Treatments	3	16.3824	5.4608	15.33	0.000
Error	105	37.4023	0.3562		
Total	143	64.8993			

Treatments here are the four wine varieties, so $F_4 = 15.33$ with p-value indicating that the four varieties are significantly different at any of the common α levels. Note that in Example 12.3.1 we found $Q_4 = 46.01$ so that $Q_4/3 = 15.34 = F_4$, up to round-off error. Finally, from the ANOVA table we see that there was not a significant difference among the 36 wine tasters in the way they rated the four wine varieties.

12.5 The Rank-Contrasts Method

Let $F_1, \ldots, F_k$ denote the cumulative distribution functions of the k populations. (More precisely they are the marginal cdf's of the joint distribution of the random vector $X_{i1}, \ldots, X_{ik}$.) In this section we will use the ranks of the combined observations for testing the hypothesis

$$H_0^F : F_1 = \cdots = F_k. \tag{12.5.1}$$

The test procedure is approximately valid for $n \geq 8$. Note that if H_0^F is true, then so is the hypothesis of equality of the two population means, $H_0 : \mu_1 = \cdots = \mu_k$.

Consider ranking the combined set of observations from smallest to largest as was done in Section 11.4. In particular, the data from the k samples, $X_{i1}, \ldots, X_{in}$, $i = 1, \ldots, k$, are combined into an overall set of $n \times k$ observations, and ranks, or mid-ranks, are assigned to this set of observations as discussed in subsection 10.2.2. Let R_{ij} denote the (mid-)rank of observation X_{ij}. The rank-contrasts method, replaces the observations, X_{ij} by their ranks, R_{ij}, in the statistic Q_k of (12.3.4). Thus, if

$$\widehat{\sigma}^2_{R,\epsilon} = \frac{1}{n-1}\frac{1}{k-1}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(R_{ij} - \overline{R}_{\cdot j} - \overline{R}_{i\cdot} + \overline{R}_{\cdot\cdot}\right)^2, \tag{12.5.2}$$

where

$$\overline{R}_i = \overline{R}_{i\cdot} = \frac{1}{n}\sum_j R_{ij}, \ \overline{R}_{\cdot j} = \frac{1}{k}\sum_i R_{ij}, \quad \text{and} \ \ \overline{R}_{\cdot\cdot} = \frac{1}{nk}\sum_i\sum_j R_{ij},$$

the rank-contrasts test statistic for testing the hypothesis H_0^F in (12.5.1) is

$$QR_k = \frac{\sum_{j=1}^k n\left(\overline{R}_j - \overline{R}_{\cdot\cdot}\right)^2}{\widehat{\sigma}^2_{R,\epsilon}}. \tag{12.5.3}$$

Under the null hypothesis H_0^F, $QR_k \mathrel{\dot\sim} \chi^2_{k-1}$, so that H_0^F is rejected at level α if

$$QR_k > \chi^2_{k-1,\alpha}, \quad \boxed{\begin{array}{c}\text{Large sample region for rejecting}\\ H_0^F : F_1 = \cdots = F_k \text{ at level } \alpha\end{array}}. \tag{12.5.4}$$

Example 12.5.1. For the wine tasting data set of Example 12.3.1, the summary statistics on the ranks are $\overline{R}_1 = 91.03$, $\overline{R}_2 = 95.17$, $\overline{R}_3 = 52.92$, $\overline{R}_4 = 50.89$, $\overline{X}_{\cdot\cdot} = 72.50$, and $\sigma^2_\epsilon = 1386.60$. Calculate the rank-contrasts statistic and test the hypothesis of no difference in the ratings of the four wines at $\alpha = 0.05$

Solution: Plugging the given summary statistics into the formula (12.3.4), we have

$$\begin{aligned} Q_4 &= \frac{36\left[(91.03-72.5)^2 + (95.17-72.5)^2 + (52.92-72.5)^2 + (50.89-72.5)^2\right]}{1386.60} \\ &= 8.9146 + 13.3430 + 9.9535 + 12.1244 = 44.34. \end{aligned}$$

Since $\chi^2_{3,0.05} = 7.815$, we conclude that the difference in the ratings is significant at $\alpha = 0.05$. (The p-value here is 0 to the first three decimal places, so the difference is significant at any of the common α levels.)

Remark 12.5.1. *The original, and still more common, rank based test procedure for randomized block designs is Friedman's test. This test, however, uses the within blocks ranks and is less powerful (less effective in detecting alternatives) than the rank-contrasts method.*

12.6 Multiple Comparisons

As it was discussed in Section 11.6, when the null hypothesis of equality of the k means is rejected, the further question arises as to which of the pairwise contrasts is (statistically significantly) different from zero. In this section we present extensions, to the present randomized block design case, of the three multiple comparison procedures procedures discussed in Section 11.6.

12.6.1 Bonferroni Multiple Comparisons and Simultaneous CIs

Bonferroni's is the easiest one to generalize. In fact, the procedure for constructing Bonferroni simultaneous CIs and multiple comparisons remains as described in Proposition 11.6.1. The only difference now is that the simultaneous CIs for the pairwise differences of means take the form of the paired t CI, or paired Z CI (see (10.4.3)); if a software package is at hand, the CI corresponding to the signed rank test can also be used. Similarly if the multiple comparisons are to be done with pairwise testing, the only difference is that one uses either the paired Z test, or the paired t test, or the signed rank test in the pairwise comparisons.

Example 12.6.1. In the context of Example 12.3.1 perform the Bonferroni multiple comparisons procedure, using a rank method for the pairwise comparisons, to identify which pairs of wine types differ significantly in terms of their ratings. Perform the multiple comparison at experiment-wise level of significance $\alpha = 0.05$.

Solution: In this example, we are interested only in multiple comparisons, not in simultaneous CIs. Because the desired experiment-wise error rate is 0.05, we will conduct each of the $m = 6$ pair-wise comparisons (1 vs 2, 1 vs 3, 1 vs 4, 2 vs 3, 2 vs 4, and 3 vs 4) at level $0.05/6 = 0.00833$. If the p-value of any one of these comparisons is smaller than 0.00833, the corresponding methods are declared different at level $\alpha = 0.05$. The results from the six signed-rank tests are summarized in the following table:

Comparison	p-value	Less than 0.0083?
1 vs 2	0.621	No
1 vs 3	0.000	Yes
1 vs 4	0.000	Yes
2 vs 3	0.000	Yes
2 vs 4	0.000	Yes
3 vs 4	0.621	No

Thus, the ratings of wines 1 and 2 are both significantly different from the ratings of wines 3 and 4, but the ratings of wines 1 and 2, as well as those of wines 3 and 4 are not significantly different at $\alpha = 0.05$.

Example 12.6.2. In the context of Example 12.3.1 construct 95% Bonferroni simultaneous paired Z CIs for all 6 pairs, and perform multiple comparisons, at experiment-wise level of significance $\alpha = 0.05$, based on them.

Solution: Because there are $m = 6$ simultaneous CIs to be constructed, the Bonferroni method constructs each CI at confidence level of $(1 - 0.05/6)100\% = 99.167\%$. The six simultaneous CIs are displayed in the following table:

Comparison	99.17% CI	Includes 0?
$\mu_1 - \mu_2$	(-0.450, 0.296)	Yes
$\mu_1 - \mu_3$	(0.255, 0.952)	No
$\mu_1 - \mu_4$	(0.302, 1.022)	No
$\mu_2 - \mu_3$	(0.327, 1.033)	No
$\mu_2 - \mu_4$	(0.348, 1.130)	No
$\mu_3 - \mu_4$	(-0.341, 0.458)	Yes

To perform multiple comparisons using the simultaneous CIs, we check which of them includes zero. If an interval does not include zero, the corresponding comparison is declared significant at experiment-wise level of 0.05. The results, listed in the last column of the above table, coincide with the results of Example 12.6.1.

Remark 12.6.1. *The simultaneous CIs constructed in Example 12.6.2 follow the formula (10.4.3) with the standard normal percentiles replacing the t percentiles; namely they are of the form*

$$\overline{D}_{j_1j_2} - z_{\alpha/2}\frac{S_{j_1j_2}}{\sqrt{n}} \leq \mu_{j_1} - \mu_{j_2} \leq \overline{D}_{j_1j_2} + z_{\alpha/2}\frac{S_{j_1j_2}}{\sqrt{n}},$$

where $\overline{D}_{j_1j_2} = \overline{X}_{j_1} - \overline{X}_{j_2}$, and $S_{j_1j_2}$ is the sample standard deviation of the differences $X_{ij_1} - X_{ij_2}$, $i = 1, \ldots, n$. Alternatively, since under the model (12.2.5) these differences have the common variance σ_ϵ^2, the estimator $\widehat{\sigma}_\epsilon$, given in (12.3.3), can be used in all CIs instead of $S_{j_1j_2}$.

12.6.2 Tukey's Multiple Comparisons and Simultaneous CIs

Tukey's method is appropriate under the normality assumption of Section 12.4. If the number of blocks is large (≥ 30), they are approximately under the more general context of Section 12.3.

The procedure for simultaneous confidence intervals of all pairwise contrasts, and for multiple comparisons is similar to that described in Section 11.6.2. The only difference is that the denominator degrees of freedom of the studentized range distribution is now $(k - 1)(n - 1)$, and S_p^2 is replaced by $\widehat{\sigma}_\epsilon^2$ in (12.3.3). More precisely, the procedures are as follows.

- $(1-\alpha)100\%$ **Tukey's simultaneous CIs for all contrasts $\mu_i - \mu_j$, $i \neq j$:**

$$\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - Q_{\alpha,k,(k-1)(n-1)}\sqrt{\frac{\widehat{\sigma}_\epsilon^2}{n}} \le \mu_i - \mu_j \le \overline{X}_{i\cdot} - \overline{X}_{j\cdot} + Q_{\alpha,k,(k-1)(n-1)}\sqrt{\frac{\widehat{\sigma}_\epsilon^2}{n}}$$

- **Tukey's multiple comparisons at level α:** If for a pair (i,j), $i \neq j$ the interval

$$\overline{X}_{i\cdot} - \overline{X}_{j\cdot} \pm Q_{\alpha,k,(k-1)(n-1)}\sqrt{\frac{\widehat{\sigma}_\epsilon^2}{n}}$$

does not contain zero, it can be concluded that μ_i and μ_j differ significantly at level α.

The following steps for carrying out Tukey's procedure lead to an organized way of presenting the results from all pairwise comparisons.

1. Select α and find $Q_{\alpha,k,(k-1)(n-1)}$ from Table A.7.
2. Calculate $w = Q_{\alpha,k,(k-1)(n-1)}\sqrt{\frac{\widehat{\sigma}_\epsilon^2}{n}}$.
3. List the sample means in increasing order and underline each pair that differs by less than w. Pairs that are not underlined indicate that the corresponding population means differ significantly at level α.

Example 12.6.3. In the context of Example 12.3.1 use Tukey's method to construct 95% simultaneous CIs for all 6 pairs, and perform multiple comparisons, at experiment-wise level of significance $\alpha = 0.05$, based on them.

Solution: Following the steps outlined above we have the following 95% simultaneous CIs for the six pairwise differences of means.

Comparison	95% Tukey's SCI	Includes 0?
$\mu_1 - \mu_2$	(-0.444, 0.290)	Yes
$\mu_1 - \mu_3$	(0.236, 0.970)	No
$\mu_1 - \mu_4$	(0.295, 1.029)	No
$\mu_2 - \mu_3$	(0.313, 1.047)	No
$\mu_2 - \mu_4$	(0.372, 1.106)	No
$\mu_3 - \mu_4$	(-0.308, 0.426)	Yes

Some of these intervals are shorter than the corresponding Bonferroni intervals of Example 12.6.2 and some are longer. Overall, however, there is close agreement. The small discrepancy is due, in part, to the fact that Tukey's intervals use the estimator $\widehat{\sigma}_\epsilon^2$ of the common, under the model (12.2.5), variance of the differences $X_{ij_1} - X_{ij_2}$, while those of Example 12.6.2 do not. See also Remark 12.6.1. The multiple comparison results that the Tukey intervals entail are in agreement with those of Examples 12.6.1 and 12.6.2.

12.6.3 Tukey's Multiple Comparisons on the Ranks

For non-normal data, it is preferable to use Tukey's multiple comparisons procedure on the (mid-)ranks, rather than on the original observations. To apply this procedure, rank the combined data as described in Section 11.4, replace the data by their ranks and apply Tukey's procedure. The application of Tukey's method on the (mid-)ranks of the combined data is approximately valid for $n \geq 8$. Note that now the simultaneous CIs are not relevant to the contrasts $\mu_i - \mu_j$; only the multiple comparisons are relevant.

Example 12.6.4. In the context of Example 12.3.1 perform multiple comparisons, at experiment-wise level of significance $\alpha = 0.05$, using Tukey's method on the ranks.

Solution: Replacing the data by their (mid-)ranks, and applying Tukey's method on the ranks gives the following 95% simultaneous CIs and multiple comparisons.

Comparison	95% Tukey's SCI	Includes 0?
F_1 vs F_2	(-27.04, 18.76)	Yes
F_1 vs F_3	(15.21, 61.01)	No
F_1 vs F_4	(17.24, 63.04)	No
F_2 vs F_3	(19.35, 65.15)	No
F_2 vs F_4	(21.38, 67.18)	No
F_3 vs F_4	(-20.87, 24.93)	Yes

Note that the CIs based on the ranks have no relevance (are not CIs for the mean differences). However, the multiple comparisons results that these intervals entail are in agreement with those of Examples 12.6.1, 12.6.2, and 12.6.3.

12.7 Exercises

1. Consider the rate of combustion in humid air flow study of Exercise 11.7,5. An engineer with statistical training observes that there is a certain pattern in the data and inquires about other experimental conditions. It turned out that different rows corresponded to

different temperatures, incrementing by 100^oK. For example, the temperature in row 1 was 1,000^oK, in row 2 was 1,100^oK etc.

(a) Explain why, given this additional information, the one-way methodology of Chapter 11 is not recommended.

(b) Complete the following ANOVA table that corresponds to the new model.

```
ANALYSIS OF VARIANCE TABLE
SOURCE          DF        SS           MS          F

Mole Fraction (   )     0.0191    (        )  (       )
Temperature   (   )  (       )    (        )  (       )
ERROR         (   )     0.0012    (        )
TOTAL         (   )  (       )
```

(c) Test again if the variation in mole fraction of water has any effect on true average diffusivity. Use $\alpha = 0.05$.

(d) Construct Tukey's 97% simultaneous CIs and give the multiple comparisons results that these confidence intervals entail.

2. A commercial airline is considering four different designs of the control panel for the new generation of airplanes. To see if the designs have an effect on the pilot's response time to emergency displays, emergency conditions were simulated and the response times, in seconds, of 8 pilots were recorded. The same 8 pilots were used for all four designs. The order in which the designs were evaluated was randomized for each pilot.

	Pilot							
Design	1	2	3	4	5	6	7	8
1	4.02	1.74	2.61	0.42	1.85	2.19	4.21	2.54
2	2.58	2.28	1.85	0.37	0.86	1.75	2.78	0.40
3	3.25	2.71	2.95	0.66	1.14	3.04	4.95	0.81
4	2.01	1.45	1.80	0.30	0.58	1.07	2.17	0.89

(a) Which procedure would you use for testing the hypothesis that there is difference in the pilot's response time with the above data?

(b) Carry out the procedure you specified in part a) at level of significance $\alpha = 0.01$.

(c) Perform multiple comparisons, at experiment-wise level 0.05, using Tukey's method on the ranks.

3. An experiment was performed to determine the effect of four different chemicals on the strength of a fabric. Five fabric samples were selected and each chemical was tested once in random order on each fabric sample. The data are shown below:

	Fabric Sample				
Chemical Type	1	2	3	4	5
1	1.3	1.6	0.5	1.2	1.1
2	2.2	2.4	0.4	2.0	1.8
3	1.8	1.7	0.6	1.5	1.3
4	2.5	2.8	1.0	2.1	1.9

(a) Complete the ANOVA table, and carry out the test at $\alpha = 0.05$.

(b) Construct Tukey's 95% simultaneous CIs and perform the corresponding multiple comparisons.

(c) Construct Bonferroni's 95% simultaneous CIs and perform the corresponding multiple comparisons.

4. A service center for electronic equipment is interested in investigating possible differences in service times of the three types disk drives that it regularly services. Each of the three technicians employed was randomly assigned to one repair of each type of drive and the repair times were recorded. The results are shown in the table below:

	Drive		
Technician	1	2	3
1	44.8	47.8	73.4
2	33.4	61.2	71.2
3	45.2	60.8	64.6

(a) Complete the ANOVA table.

(b) State the two hypotheses that can be tested from the ANOVA table.

(c) What assumptions are needed for the validity of the ANOVA table tests?

(d) Test each of the two hypotheses using $\alpha = 0.05$, and state your conclusions.

5. A study was conducted to see whether three cars, A, B, and C, having very different wheel bases and turning radii, took the same time to parallel park. A random sample of seven drivers was obtained and the time required for each of them to parallel park each of the three cars was measured. The results are listed in the table below.

	Driver						
Car	1	2	3	4	5	6	7
A	19.0	21.8	16.8	24.2	22.0	34.7	23.8
B	17.8	20.2	16.2	41.4	21.4	28.4	22.7
C	21.3	22.5	17.6	38.1	25.8	39.4	23.9

Is there evidence that the time required to parallel park the three types of car are different?

(a) Test at $\alpha = 0.05$ using the ANOVA table method.

(b) Test at $\alpha = 0.05$ using the rank-contrasts method.

(c) Compare the two results. State which of the two results is more credible, justifying your answer.

(d) Perform Bonferroni's multiple comparisons at experiment-wise error rate of 0.05, using the signed rank test.

(e) Perform Tukey's multiple comparisons on the ranks at experiment-wise error rate of 0.05

Table A.1. Cumulative Binomial Probabilities

		p								
n	x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	0	0.591	0.328	0.168	0.078	0.031	0.010	0.002	0.000	0.000
	1	0.919	0.737	0.528	0.337	0.188	0.087	0.031	0.007	0.000
	2	0.991	0.942	0.837	0.683	0.500	0.317	0.163	0.058	0.009
	3	0.995	0.993	0.969	0.913	0.813	0.663	0.472	0.263	0.082
	4	1.000	1.000	0.998	0.990	0.699	0.922	0.832	0.672	0.410
10	0	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000
	1	0.736	0.376	0.149	0.046	0.011	0.002	0.000	0.000	0.000
	2	0.930	0.678	0.383	0.167	0.055	0.012	0.002	0.000	0.000
	3	0.987	0.879	0.650	0.382	0.172	0.055	0.011	0.001	0.000
	4	0.988	0.967	0.850	0.633	0.377	0.166	0.047	0.006	0.000
	5	1.000	0.994	0.953	0.834	0.623	0.367	0.150	0.033	0.002
	6	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013
	7	1.000	1.000	0.998	0.988	0.945	0.833	0.617	0.322	0.070
	8	1.000	1.000	1.000	0.998	0.989	0.954	0.851	0.624	0.264
	9	1.000	1.000	1.000	1.000	0.999	0.994	0.972	0.893	0.651
15	0	0.206	0.035	0.005	0.001	0.000	0.000	0.000	0.000	0.000
	1	0.549	0.167	0.035	0.005	0.001	0.000	0.000	0.000	0.000
	2	0.816	0.398	0.127	0.027	0.004	0.000	0.000	0.000	0.000
	3	0.944	0.648	0.297	0.091	0.018	0.002	0.000	0.000	0.000
	4	0.987	0.836	0.516	0.217	0.059	0.009	0.001	0.000	0.000
	5	0.998	0.939	0.722	0.403	0.151	0.034	0.004	0.000	0.000
	6	1.000	0.982	0.869	0.610	0.304	0.095	0.015	0.001	0.000
	7	1.000	0.996	0.950	0.787	0.500	0.213	0.050	0.004	0.000
	8	1.000	0.999	0.985	0.905	0.696	0.390	0.131	0.018	0.000
	9	1.000	1.000	0.996	0.966	0.849	0.597	0.278	0.061	0.002
	10	1.000	1.000	0.999	0.991	0.941	0.783	0.485	0.164	0.013
	11	1.000	1.000	1.000	0.998	0.982	0.909	0.703	0.352	0.056
	12	1.000	1.000	1.000	1.000	0.996	0.973	0.873	0.602	0.184
	13	1.000	1.000	1.000	1.000	1.000	0.995	0.965	0.833	0.451
	14	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.965	0.794

Table A.1. Cumulative Binomial Probabilities

		p								
n	x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0	0.122	0.012	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.392	0.069	0.008	0.001	0.000	0.000	0.000	0.000	0.000
	2	0.677	0.206	0.035	0.004	0.000	0.000	0.000	0.000	0.000
	3	0.867	0.411	0.107	0.016	0.001	0.000	0.000	0.000	0.000
	4	0.957	0.630	0.238	0.051	0.006	0.000	0.000	0.000	0.000
	5	0.989	0.804	0.416	0.126	0.021	0.002	0.000	0.000	0.000
	6	0.998	0.913	0.608	0.250	0.058	0.006	0.000	0.000	0.000
	7	1.000	0.968	0.772	0.416	0.132	0.021	0.001	0.000	0.000
	8	1.000	0.990	0.887	0.596	0.252	0.057	0.005	0.000	0.000
	9	1.000	0.997	0.952	0.755	0.412	0.128	0.017	0.001	0.000
	10	1.000	0.999	0.983	0.873	0.588	0.245	0.048	0.003	0.000
	11	1.000	1.000	0.995	0.944	0.748	0.404	0.113	0.010	0.000
	12	1.000	1.000	0.999	0.979	0.868	0.584	0.228	0.032	0.000
	13	1.000	1.000	1.000	0.994	0.942	0.750	0.392	0.087	0.002
	14	1.000	1.000	1.000	0.998	0.979	0.874	0.584	0.196	0.011
	15	1.000	1.000	1.000	1.000	0.994	0.949	0.762	0.370	0.043
	16	1.000	1.000	1.000	1.000	0.999	0.984	0.893	0.589	0.133
	17	1.000	1.000	1.000	1.000	1.000	0.996	0.965	0.794	0.323
	18	1.000	1.000	1.000	1.000	1.000	1.000	0.992	0.931	0.608
	19	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.988	0.878

Table A.2. Cumulative Poisson Probabilities

	λ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	0.368
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920
3		1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981
4				1.000	1.000	1.000	0.999	0.999	0.998	0.996
5							1.000	1.000	1.000	0.999

	λ									
x	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
0	0.301	0.247	0.202	0.165	0.135	0.111	0.091	0.074	0.061	0.050
1	0.663	0.592	0.525	0.463	0.406	0.355	0.308	0.267	0.231	0.199
2	0.879	0.833	0.783	0.731	0.677	0.623	0.570	0.518	0.469	0.423
3	0.966	0.946	0.921	0.891	0.857	0.819	0.779	0.736	0.692	0.647
4	0.992	0.986	0.976	0.964	0.947	0.928	0.904	0.877	0.848	0.815
5	0.998	0.997	0.994	0.990	0.983	0.975	0.964	0.951	0.935	0.961
6	1.000	0.999	0.999	0.997	0.995	0.993	0.988	0.983	0.976	0.966
7		1.000	1.000	0.999	0.999	0.998	0.997	0.995	0.992	0.988
8				1.000	1.000	1.000	0.999	0.999	0.998	0.996
9							1.000	1.000	0.999	0.999

	λ									
x	3.2	3.4	3.6	3.8	4.0	4.2	4.4	4.6	4.8	5.0
0	0.041	0.033	0.027	0.022	0.018	0.015	0.012	0.010	0.008	0.007
1	0.171	0.147	0.126	0.107	0.092	0.078	0.066	0.056	0.048	0.040
2	0.380	0.340	0.303	0.269	0.238	0.210	0.185	0.163	0.143	0.125
3	0.603	0.558	0.515	0.473	0.433	0.395	0.359	0.326	0.294	0.265
4	0.781	0.744	0.706	0.668	0.629	0.590	0.551	0.513	0.476	0.440
5	0.895	0.871	0.844	0.816	0.785	0.753	0.720	0.686	0.651	0.616
6	0.955	0.942	0.927	0.909	0.889	0.867	0.844	0.818	0.791	0.762
7	0.983	0.977	0.969	0.960	0.949	0.936	0.921	0.905	0.887	0.867
8	0.994	0.992	0.998	0.984	0.979	0.972	0.964	0.955	0.944	0.932
9	0.998	0.997	0.996	0.994	0.992	0.989	0.985	0.980	0.975	0.968
10	1.000	0.999	0.999	0.998	0.997	0.996	0.994	0.992	0.990	0.986
11		1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996	0.995
12				1.000	1.000	1.000	0.999	0.999	0.999	0.998
13							1.000	1.000	1.000	0.999

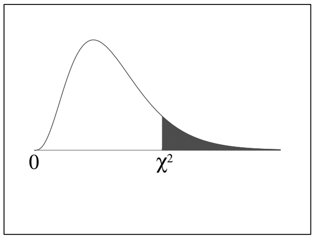
Table A.3. The Cumulative Distribution Function for the Standard Normal Distribution: Values of $\Phi(z)$ for nonnegative z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table A.4. Percentiles of the t-Distribution

df	90%	95%	97.5%	99%	99.5%	99.9%
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.183	4.541	5.841	10.215
4	1.533	2.132	2.777	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.708	5.208
7	1.415	1.895	2.365	2.998	3.500	4.785
8	1.397	1.860	2.306	2.897	3.355	4.501
9	1.383	1.833	2.262	2.822	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.625	2.977	3.787
15	1.341	1.753	2.132	2.603	2.947	3.733
16	1.337	1.746	2.120	2.584	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.879	3.611
19	1.328	1.729	2.093	2.540	2.861	3.580
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.788	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.705	3.307
80	1.292	1.664	1.990	2.374	2.639	3.195
∞	1.282	1.645	1.960	2.326	2.576	3.090

Table A.5. Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table A.6. Percentiles of the F-Distribution (ν_1 = Numerator df; ν_2 = Denominator df)

		ν_1										
ν_2	α	1	2	3	4	5	6	7	8	12	24	1000
1	0.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	60.71	62.00	63.30
	0.05	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.2
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.41	9.45	9.49
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.41	19.45	19.49
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.22	5.18	5.13
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.74	8.64	8.53
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.90	3.83	3.76
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.91	5.77	5.63
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.27	3.19	3.11
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.68	4.53	4.37
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.90	2.82	2.72
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.00	3.84	3.67
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.67	2.58	2.47
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.57	3.41	3.23
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.50	2.40	2.30
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.28	3.12	2.93
10	0.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.28	2.18	2.06
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.91	2.74	2.54
12	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.15	2.04	1.91
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.69	2.51	2.30
14	0.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.05	1.94	1.80
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.53	2.35	2.14
16	0.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	1.99	1.87	1.72
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.42	2.24	2.02
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.89	1.77	1.61
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.28	2.08	1.85
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.77	1.64	1.46
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.09	1.89	1.63
50	0.10	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.68	1.54	1.33
	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	1.95	1.74	1.45
100	0.10	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.61	1.46	1.22
	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.85	1.63	1.30
1000	0.10	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.55	1.39	1.08
	0.05	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.76	1.53	1.11

Table A.7. Percentiles of the Studentized Range Distribution

$Q_{\alpha,k,\nu}$ for $\alpha = 0.10$ and $\alpha = 0.05$

		k									
ν	α	2	3	4	5	6	7	8	9	10	11
5	.10	2.85	3.72	4.26	4.66	4.98	5.24	5.46	5.65	5.82	5.96
	.05	3.63	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
6	.10	2.75	3.56	4.06	4.43	4.73	4.97	5.17	5.34	5.50	5.64
	.05	3.46	4.34	4.90	5.30	5.63	5.89	6.12	6.32	6.49	6.65
7	.10	2.68	3.45	3.93	4.28	4.55	4.78	4.97	5.14	5.28	5.41
	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.81	6.00	6.16	6.30
8	.10	2.63	3.37	3.83	4.17	4.43	4.65	4.83	4.99	5.13	5.25
	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
10	.10	2.56	3.27	3.70	4.02	4.26	4.46	4.64	4.78	4.91	5.03
	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
12	.10	2.52	3.20	3.62	3.92	4.16	4.35	4.51	4.65	4.78	4.89
	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.26	5.39	5.51
13	.10	2.50	3.18	3.59	3.88	4.12	4.30	4.46	4.60	4.72	4.83
	.05	3.05	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
14	.10	2.49	3.16	3.56	3.85	4.08	4.27	4.42	4.56	4.68	4.79
	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
16	.10	2.47	3.12	3.52	3.80	4.03	4.21	4.36	4.49	4.61	4.71
	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
18	.10	2.45	3.10	3.49	3.77	3.98	4.16	4.31	4.44	4.55	4.65
	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.95	5.07	5.17
20	.10	2.44	3.08	3.46	3.74	3.95	4.12	4.27	4.40	4.51	4.61
	.05	2.95	3.58	3.96	4.23	4.44	4.62	4.77	4.89	5.01	5.11
25	.10	2.42	3.04	3.42	3.68	3.89	4.06	4.20	4.32	4.43	4.53
	.05	2.91	3.52	3.89	4.15	4.36	4.53	4.67	4.79	4.90	4.99
30	.10	2.40	3.02	3.39	3.65	3.85	4.02	4.15	4.27	4.38	4.47
	.05	2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.82	4.92
40	.10	2.38	2.99	3.35	3.60	3.80	3.96	4.10	4.21	4.32	4.41
	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
60	.10	2.36	2.96	3.31	3.56	3.75	3.91	4.04	4.15	4.25	4.34
	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
80	.10	2.35	2.94	3.29	3.54	3.73	3.88	4.01	4.12	4.22	4.31
	.05	2.81	3.38	3.71	3.95	4.13	4.28	4.40	4.51	4.60	4.69
∞	.10	2.33	2.90	3.24	3.48	3.66	3.81	3.93	4.04	4.13	4.21
	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55

Index