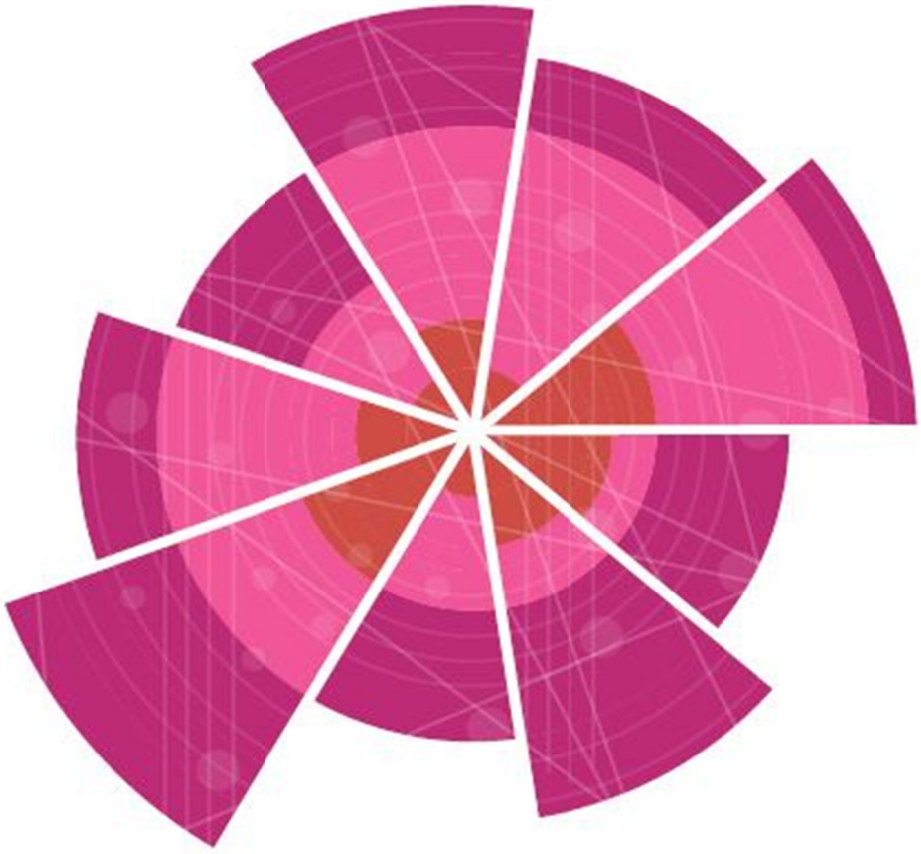
STAT 221/231
Course Notes
Winter 2022 Edition

Department of Statistics and
Actuarial Science
University of Waterloo



STATISTICS 221/231 COURSE NOTES

Department of Statistics and Actuarial Science, University of Waterloo

Winter 2022 Edition

Contents

Preface

These notes are a work-in-progress with contributions from those students taking the courses and the instructors teaching them. An original version of these notes was prepared by Jerry Lawless. Additions and revisions were made by Cyntha Struthers, Don McLeish, Jock MacKay, and others. Richard Cook supplied the example in Chapter 8. In order to provide improved versions of the notes for students in subsequent terms, please email typos and errors, or sections that are confusing, or additional comments/suggestions to castruth@uwaterloo.ca.

Specific topics in these notes also have associated video files or Powerpoint shows that can be accessed at www.watstat.ca.

1. INTRODUCTION TO STATISTICAL SCIENCES

1.1 Empirical Studies and Statistical Sciences

An *empirical study* is one in which knowledge is gained by observation or by experiment. Empirical studies may be conducted to further knowledge, improve systems, or determine public policy. For example, in disciplines such as insurance or finance, decisions must be made about what premium to charge for an insurance policy or whether to buy or sell a stock on the basis of available data. In medical research, decisions must be made about the safety and efficacy of new treatments for diseases such as cancer based on clinical trials. Government scientists collect data on fish stocks in order to provide information to policy makers who must set quotas or limits on commercial fishing.

Empirical studies deal with *populations* and *processes* which are collections of individual *units*. In order to study a population a *sample* of units is carefully selected from that population. To study a process a sample of units generated by the process is examined. Since only a sample from the population or process is observed and not all of the units are the same, there will be uncertainty in the conclusions drawn from such a study. For example, researchers at a pharmaceutical company may conduct a study to assess the effect of a new drug for controlling hypertension (high blood pressure) in a person (the unit). For cost and ethical reasons, they can only involve a relatively small sample of people in the study. Since people have varying degrees of hypertension, they react differently to the drug, and they have different side effects, there will be uncertainly in the conclusions drawn from the study. In another example, a financial engineer may collect data on stock values during a previous time period to try and predict their values in a future time period. In this case a unit is a stock. These predictions would involve uncertainty due to the variability present in such data. Finally a commercial website may conduct a study to examine changes in the number of website hits before and after an advertising campaign. Data on the number of hits per hour might be collected over a fixed period of time before and after the campaign. In this case the unit would be an hour on the website. Since only a sample of the process is collected, the conclusions about any changes would also involve uncertainty.

Statistical Sciences are concerned with all aspects of empirical studies including formu-

lating the problem, planning the experiment, collecting the data, analyzing the data, and making conclusions. In particular, Statistical Sciences deal with the study of variability in populations and processes, and with informative and cost-effective ways to collect and analyze data about such populations and processes.

Statistical data analysis occurs in a huge number of areas. For example, statistical algorithms are the basis for software involved in the automated recognition of handwritten or spoken text; statistical methods are commonly used in law cases, for example in DNA profiling; statistical process control is used to increase the quality and productivity of manufacturing and service processes; individuals are selected for direct mail marketing campaigns through a statistical analysis of their characteristics. With modern information technology, massive amounts of data are routinely collected and stored. But data do not equal information, and it is the purpose of Statistical Sciences to provide and analyze data so that the maximum amount of information or knowledge may be obtained[1]. Poor or improperly analyzed data may be useless or misleading. The same could be said about poorly collected data.

Probability models are used to represent many phenomena, populations, or processes and to deal with problems that involve variability. You studied these models in your probability course and you have seen how they can be used to describe variability. This course will focus on the collection, analysis and interpretation of data and the probability models you studied previously will be used extensively. The most important material from your probability course is the material dealing with random variables, including distributions such as the Binomial, Poisson, Multinomial, Normal or Gaussian, Uniform and Exponential. It is important to review this material on our own.

Statistical Sciences is a large discipline and this course is only an introduction. The broad objective of this course is to discuss all aspects of: problem formulation, planning of an empirical study, formal and informal analysis of data, and the conclusions and limitations of such an analysis. We must remember that data are collected and models are constructed for a specific reason. In any given application we should keep the big picture in mind (e.g. Why are we studying this? What else do we know about it?) even when considering one specific aspect of a problem.

Here is a quote[2] from Hal Varien, Google's chief economist.

"The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary(sic) scarce factor is the ability to understand that data and extract value from it.

[1] A brilliant example of how to create information through data visualization is found in the video by Hans Rosling at: http://www.youtube.com/watch?v=jbkSRLYSojo

[2] For the complete article see "How the web challenges managers" Hal Varian, *The McKinsey Quarterly*, January 2009

I think statisticians are part of it, but it's just a part. You also want to be able to visualize the data, communicate the data, and utilize it effectively. But I do think those skills - of being able to access, understand, and communicate the insights you get from data analysis - are going to be extremely important. Managers need to be able to access and understand the data themselves."

1.2 Data Collection

A *population* is a collection of *units*. Examples are: the population of all students taking STAT 231 this term (the unit is a student); the population of all persons aged 18-25 living in Ontario on January 1, 2020 (the unit is a person aged 18-25); and the population of all car insurance policies issued by a particular insurance company in the year 2019 (the unit is a car insurance policy). A *process* is a system by which units are produced. Students taking STAT 231 now and into the future could be considered as units in a process. Car insurance claims for a particular insurance company now and into the future could also be considered as units in a process. A key feature of processes is that they usually occur over time whereas populations are often static (defined at one moment in time).

Definition 1 *A variate is a characteristic of a unit.*

Variates can be of different types. If the units are people then variates such as their height, weight, age, and time until recurrence of disease after medical treatment are all examples of *continuous* variates. The lifetime of an electrical component (the unit) is also a continuous variate.

Suppose the unit of interest is a production run of smartphones made by a particular company then the number of defective smartphones produced in a run is an example of a *discrete* variate. If the units are trees in a particular forest then the number of aphids on a tree is also a discrete variate. The number of deaths in a year on a particular section of dangerous highway is another example of a discrete variate. What is the unit in this example?

Variates such as hair colour, university program or marital status of a person (the unit) are examples of *categorical* variates since these variates do not take on numerical values. Another example of a categorical variate would be the presence or absence of a disease in a unit. Sometimes, to facilitate the analysis of the data, we might redefine the variate of interest to be 1 if the disease is present and 0 if the disease is absent. We would now call the variate a discrete variate. Since the variate only takes on values 0 or 1 such a variate is often referred to as a *binary* variate.

If a categorical variate has a natural ordering then it is called an *ordinal* variate. For example, the size of a unit is an ordinal variate if the categories for size are: large, medium, small. Another example of an ordinal variate would be the opinion of a person (the unit) on a given statement in a poll for which the categories might be: strongly agree, agree, neutral, disagree, strongly disagree.

Variates can also be *complex*. The open ended response by a person (the unit) to a question on a survey is an example of a complex variate. If the units are cities and an aerial image is associated with each city (the unit) then the image is also a complex variate.

The values of a variate typically vary across the units in a population or process. This variability generates uncertainty and makes it necessary to study populations and processes by collecting data about them. By data, we mean the values of the variates for a sample of units drawn from a population or process. It is important to identify the types of variates in an empirical study since this identification will help us in choosing statistical models for the data which will aid us in the analysis of the data.

Definition 2 *An attribute of a population or process is a function of the variates over the population or process.*

Questions about a population or process are stated in terms of attributes. Examples of attributes are the *average* drop in blood pressure due to a treatment for individuals with hypertension, the *variability* in heights of children aged 10, or the *proportion* of a population having a certain characteristic.

In planning to collect data about a population or process, we must carefully specify what the objectives are. Then, we must consider feasible methods for collecting data as well as the extent it will be possible to answer questions of interest. This sounds simple but is usually difficult to do well, especially since resources are always limited.

There are several ways in which we can obtain data. One way is purely according to what is available: that is, data are provided by some existing source. Huge amounts of data collected by many technological systems are of this type, for example, data on credit card usage or on purchases made by customers in a supermarket. Sometimes it is not clear what available data represent and they may be unsuitable for serious analysis. For example, people who voluntarily provide data in a web survey may not be representative of the population at large. Alternatively, we may plan and execute a sampling plan to collect new data. Statistical Sciences stress the importance of obtaining data that will be objective and provide maximal information at a reasonable cost.

Recall that an empirical study is one in which we learn by observation or experiment. Most often this is done by collecting data. The empirical studies we will consider will usually be one of the following types:

(i) **Sample surveys**: The object of many empirical studies is to learn about a finite population (e.g. all persons over 19 in Ontario as of September 1 in a given year). In this case information about the population may be obtained by selecting a "representative" sample of units from the population and determining the variates of interest for each unit in the sample. Obtaining such a sample can be challenging and expensive. In a survey sample the variates of interest are most often collected using a questionnaire. Sample surveys are widely used in government statistical studies, economics, marketing, public opinion polls, sociology, quality assurance and other areas.

(ii) **Observational studies**: An observational study is one in which data are collected about a population or process without any attempt to change the value of one or more variates for the sampled units. For example, in studying risk factors associated with a disease such as lung cancer, we might investigate all cases of the disease at a particular hospital (or perhaps a sample of them) that occur over a given time period. We would also examine a sample of individuals who did not have the disease. A distinction between a sample survey and an observational study is that for observational studies the population of interest is usually infinite or conceptual. For example, in investigating risk factors for a disease, we prefer to think of the population of interest as a conceptual one consisting of persons at risk from the disease recently or in the future.

(iii) **Experimental studies:** An experimental study is one in which the experimenter (that is, the person conducting the study) intervenes and changes or sets the values of one or more variates for the units in the sample. For example, in an engineering experiment to quantify the effect of temperature on the performance of a certain type of computer chip, the experimenter might decide to run a study with 40 chips, ten of which are operated at each of four temperatures 10, 20, 30, and 40 degrees Celsius. Since the experimenter decides the temperature level for each chip in the sample, this is an experiment.

These three types of empirical studies are not mutually exclusive, and many studies involve aspects of all of them. Here are some slightly more detailed examples.

Example 1.2.1 A sample survey about smoking

Suppose we wish to study the smoking behaviour of Ontario residents aged $14 - 20$ years. (Think about reasons why such studies are considered important.) Of course, the population of Ontario residents aged $14 - 20$ years and their smoking habits both change over time, so we will content ourselves with a snapshot of the population at some point in time (e.g. the second week of September in a given year). Since we cannot afford to contact all persons in the population, we decide to select a sample of persons from the population of interest. (Think about how we might do this - it is quite difficult!) We decide to measure the following variates on each person in the sample: age, sex, place of residence, occupation, current smoking status, length of time smoked, etc.

Note that we have to decide how we are going to obtain our sample and how large it should be. The former question is very important if we want to ensure that our sample provides a good picture of the overall population. The amount of time and money available to carry out the study heavily influences how we will proceed.

Example 1.2.2 A study of a manufacturing process

When a manufacturer produces a product in packages stated to weigh or contain a certain amount, they are generally required by law to provide at least the stated amount in

each package. Since there is always some inherent variation in the amount of product which the manufacturing process deposits in each package, the manufacturer has to understand this variation and set up the process so that no packages or only a very small fraction of packages contain less than the required amount.

Consider, for example, soft drinks sold in nominal 355 ml cans. Because of inherent variation in the filling process, the amount of liquid y that goes into a can varies over a small range. Note that the manufacturer would like the variability in y to be as small as possible, and for cans to contain at least 355 ml. Suppose that the manufacturer has just added a new filling machine to increase the plant's capacity. The process engineer wants to compare the new machine with an old one. Here the population of interest is the cans filled in the future by both machines. The process engineer decides to do this by sampling some filled cans from each machine and accurately measuring the amount of liquid y in each can. This is an observational study.

How exactly should the sample be chosen? The machines may *drift* over time (that is, the average of the y values or the variability in the y values may vary systematically up or down over time) so we should select cans over time from each machine. We have to decide how many, over what time period, and when to collect the cans from each machine.

Example 1.2.3 A clinical trial in medicine

In studies of the treatment of disease, it is common to compare alternative treatments in experiments called clinical trials. Consider, for example, a population of persons who are at high risk of a stroke. Some years ago it was established in clinical trials that small daily doses of aspirin (which acts as a blood thinner) could lower the risk of stroke. This was done by giving some high risk subjects daily doses of aspirin (call this Treatment 1) and others a daily dose of a placebo (an inactive compound) given in the same form as the aspirin (call this Treatment 2). The two treatment groups were then followed for a period of time, and the number of strokes in each group was observed. Note that this is an experimental study because the researchers decided which subjects in the sample received Treatment 1 and which subjects received Treatment 2.

This sounds like a simple plan to implement but there are several important points. For example, patients should be assigned to receive Treatment 1 or Treatment 2 in some random fashion to avoid unconscious bias (e.g. doctors might otherwise tend to put persons at higher risk of stroke in the aspirin group) and to balance other factors (e.g. age, sex, severity of condition) across the two groups. It is also best not to let the patients or their doctors know which treatment they are receiving. This type of study is called a double-blind study. Many other questions must also be addressed. For example, what variates should we measure other than the occurrence of a stroke? What should we do about patients who are forced to drop out of the study because of adverse side effects? Is it possible that the aspirin treatment works for certain types of patients but not others? How long should the study go on? How many persons should be included?

As an example of a statistical setting where the data are not obtained by a sample survey, an experimental study, or even an observational study, consider the following.

Example 1.2.4 Direct marketing campaigns

Nearly every major retailer has a predictive analytics department devoted to understanding not just consumers' shopping habits but also their personal habits, so they can market to them more efficiently. The retail chain Target has been particularly good at this. Since Target sells everything from food to toys to lawn furniture to electronics, one of its primary goals is to try and convince customers that Target is the only store they need. Once consumers adopt certain shopping habits, however, it is very difficult to change them even with the most ingenious ad campaigns. One group that is more open to changes in their buying habits is new parents. Because birth records are usually public, new parents are bombarded with offers and advertisements from all sorts of companies as soon as the baby arrives. Target hypothesized that if they could identify women earlier in their pregnancy and send them specially designed ads, then there was a good chance of getting them to shop at Target for years. How did Target determine if a woman was pregnant?

Target has collected large amounts of data on their customers for decades. Every person who makes a credit card purchase, fills out a survey, mails in a refund on a purchase, calls the customer help line, or visits the website is assigned a guest ID. Linked to the guest ID is information on credit card purchases as well as demographic information like age, marital status, number of children, address, estimated salary, types of credit cards and websites visited. Target also buys data about ethnicity, job history, magazines read, college attended, topics discussed online, etc. The data scientist working for Target was able to identify a large number of variates that, when analyzed together, allowed them to assign each shopper a "pregnancy prediction" score. Based on these scores Target could then select which women to send the specially designed ads. For more information on how Target used these scores to increase sales see www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

1.3 Data Summaries

When we study a population or process we collect data. We cannot answer the questions of interest without summarizing the data. Summaries are especially important when we report the conclusions of the study. Summaries must be clear and informative with respect to the questions of interest and, since they are summaries, we need to make sure that they are not misleading. There are two classes of summaries: numerical and graphical.

We represent variates by letters such as x, y, z. For example, we might define a variate y as the size in dollars of an insurance claim or the first language that a person learned to speak.

Suppose that data on a variate y is collected for n units in a population or process. By convention, we label the units as $1, 2, \ldots, n$ and denote their respective y values as $y_1, y_2, \ldots, y_n$. We might also collect data on a second variate x for each unit, and we would

denote the values as $x_1, x_2, \ldots, x_n$. We refer to n as the *sample size* and to $\{x_1, x_2, \ldots, x_n\}$, $\{y_1, y_2, \ldots, y_n\}$ or $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$ as data sets. Most data sets contain the values for many variates.

Numerical Summaries

We now describe some numerical summaries which are useful for describing features of a single variate in a data set. These summaries fall generally into three categories: measures of location (mean, median, and mode), measures of variability or dispersion (variance, range, and interquartile range), and measures of shape (skewness and kurtosis). These summaries are used when the variate is either discrete or continuous.

Measures of location

- The *sample mean* $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ (also called the sample average).
- The *sample median* $\hat{m}$ or the middle value when n is odd and the sample is ordered from smallest to largest, and the average of the two middle values when n is even.
- The *sample mode*, or the value of y which appears in the sample with the highest frequency (not necessarily unique).

The sample mean, median and mode describe the "center" of the distribution of variate values in a data set. The units for mean, median and mode (e.g. centimeters, degrees Celsius, etc.) are the same as for the original variate.

Since the median is less affected by a few extreme observations (see Problem 1), it is a more robust measure of location.

Measures of dispersion or variability

- The *sample variance*:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 \right]$$

  and the *sample standard deviation*: $s = \sqrt{s^2}$.

- The *range* $= y_{(n)} - y_{(1)}$ where $y_{(n)} = \max(y_1, y_2, \ldots, y_n)$ and $y_{(1)} = \min(y_1, y_2, \ldots, y_n)$.
- The *interquartile range* IQR (see Definition 5).

The sample variance and sample standard deviation measure the variability or spread of the variate values in a data set. The units for standard deviation, range, and interquartile range (e.g. centimeters, degrees Celsius, etc.) are the same as for the original variate.

Since the interquartile range is less affected by a few extreme observations (see Problem 2), it is a more robust measure of variability.

Measures of shape

- The *sample skewness*

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{3/2}}$$

- The *sample kurtosis*

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{2}}$$

Measures of shape generally indicate how the data, in terms of a relative frequency histogram, differ from the Normal bell-shaped curve, for example whether one "tail" of the relative frequency histogram is substantially larger than the other so the histogram is asymmetric, or whether both tails of the relative frequency histogram are large so the data are more prone to extreme values than data from a Normal distribution.

Sample skewness and sample kurtosis have no units.

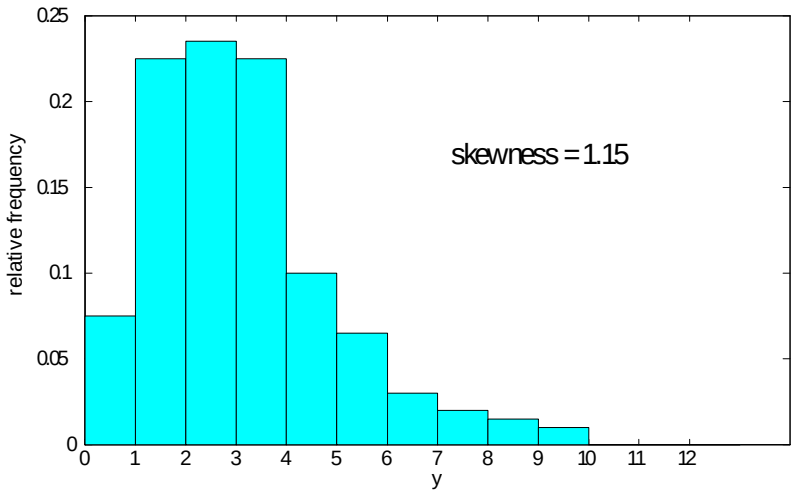


Figure 1.1: **Relative frequency histogram for data with positive skewness**

The sample skewness is a measure of the (lack of) symmetry in the data. When the relative frequency histogram of the data is approximately symmetric then there is an approximately equal balance between the positive and negative values in the sum $\sum_{i=1}^{n}(y_i - \bar{y})^3$ and this results in a value for the sample skewness that is approximately zero.

If the relative frequency histogram of the data has a long right tail (see Figure 1.1), then the positive values of $(y_i - \bar{y})^3$ dominate the negative values in the sum and the value of

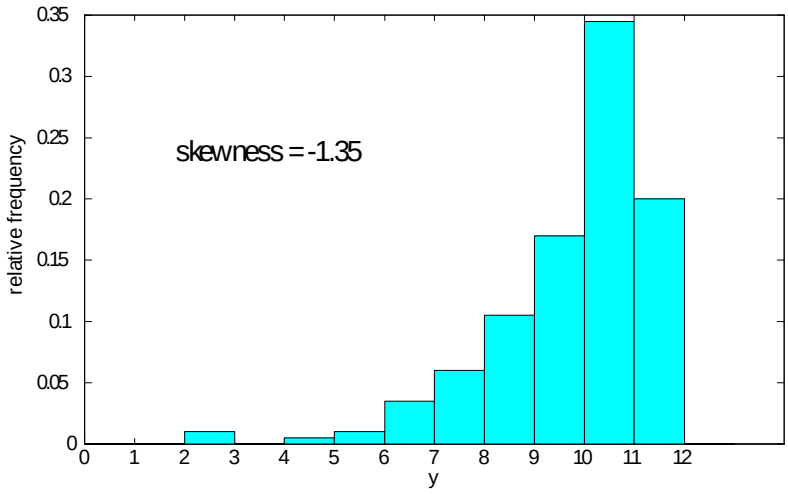


Figure 1.2: **Relative frequency histogram for data with negative skewness**

the skewness will be positive. Similarly if the relative frequency histogram of the data had a long left tail (see Figure 1.2) then the negative values of $(y_i - \bar{y})^3$ dominate the positive values in the sum and the value of the skewness will be negative.

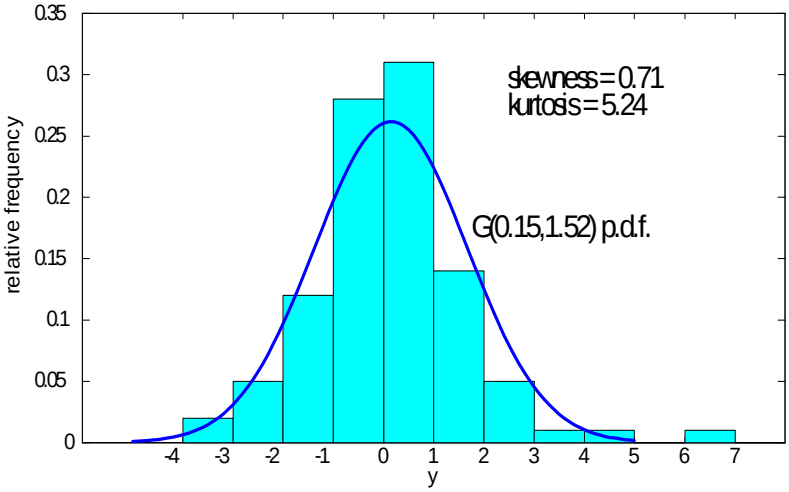


Figure 1.3: **Relative frequency histogram for data with kurtosis > 3**

The sample kurtosis measures the heaviness of the tails and the peakedness of the data relative to data that are Normally distributed. Since the term $(y_i - \bar{y})^4$ is always positive, the kurtosis is always positive. If the sample kurtosis is greater than 3 then this indicates heavier tails (and a more peaked center) than data that are Normally distributed. For data that arise from a model with no tails, for example the Uniform distribution, the sample

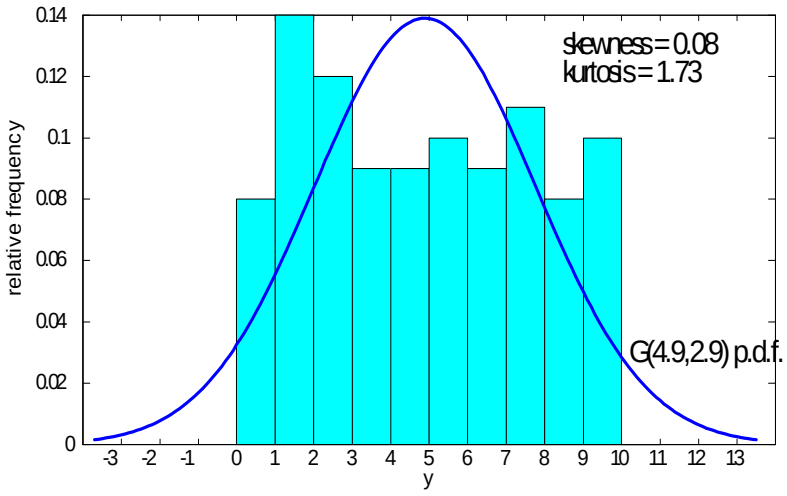


Figure 1.4: **Relative frequency histogram for data with kurtosis** < 3

kurtosis will be less than 3. See Figures 1.3 and 1.4. Typical financial data such as the S&P500 index have kurtosis values greater than three, because the extreme returns (both large and small) are more frequent than one would expect for Normally distributed data.

Another way to numerically summarize data is to use sample percentiles or quantiles.

Sample Quantiles and Percentiles

For $0 < p < 1$, the pth quantile (also called the $100p$th percentile) is a value such that approximately a fraction p of the y values in the data set are less than $q(p)$ and approximately $1-p$ are greater than $q(p)$. Depending on the size of the data set, quantiles are not uniquely defined for all values of p. There are different conventions for defining quantiles in these cases. If the sample size is large, the differences in the quantiles based on the various definitions are small. We will use the following definition to determine quantiles.

Definition 3 *Let* $\left\{y_{(1)}, y_{(2)}, \ldots, y_{(n)}\right\}$ *where* $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ *be the* order statistic *for the data set* $\{y_1, y_2, \ldots, y_n\}$. *For* $0 < p < 1$, *the* p*th (sample) quantile (also called the* $100p$*th (sample) percentile), is a value, call it* $q(p)$, *determined as follows:*

- *Let* $k = (n+1)p$ *where* n *is the sample size.*
- *If* k *is an integer and* $1 \leq k \leq n$, *then* $q(p) = y_{(k)}$.
- *If* k *is not an integer but* $1 < k < n$ *then determine the closest integer* j *such that* $j < k < j+1$ *and then* $q(p) = \frac{1}{2}\left[y_{(j)} + y_{(j+1)}\right]$.

The quantiles $q(0.25)$, $q(0.5)$ and $q(0.75)$ are often used to summarize a data set and are given special names.

Definition 4 *The quantiles* $q(0.25)$, $q(0.5)$ *and* $q(0.75)$ *are called the* lower or first quartile, *the* median, *and the* upper or third quartile *respectively.*

Example 1.3.1

Consider the data set of 12 observations which has already been ordered from smallest to largest:

$$1.2 \quad 6.6 \quad 6.8 \quad 7.6 \quad 7.9 \quad 9.1 \quad 10.9 \quad 11.5 \quad 12.2 \quad 12.7 \quad 13.1 \quad 14.3$$

For $p = 0.25$, $k = (12+1)(0.25) = 3.25$ so

$$\text{lower quartile } = q(0.25) = \frac{1}{2}\left(y_{(3)} + y_{(4)}\right) = \frac{1}{2}(6.8 + 7.6) = 7.2$$

For $p = 0.5$, $k = (12+1)(0.5) = 6.5$ so

$$\text{median} = \hat{m} = q(0.5) = \frac{1}{2}\left(y_{(6)} + y_{(7)}\right) = \frac{1}{2}(9.1 + 10.9) = 10$$

For $p = 0.75$, $k = (12+1)(0.75) = 9.75$ so

$$\text{upper quartile } = q(0.75) = \frac{1}{2}\left(y_{(9)} + y_{(10)}\right) = \frac{1}{2}(12.2 + 12.7) = 12.45$$

Also for $p = 0.1$, $k = (12+1)(0.1) = 1.3$ so

$$q(0.1) = \frac{1}{2}\left(y_{(1)} + y_{(2)}\right) = \frac{1}{2}(1.2 + 6.6) = 3.9$$

A way to quantify the variability of the variate values in a data set is to use the interquartile range (IQR) which is the difference between the lower and upper quartiles.

Definition 5 *The* interquartile range *is* $IQR = q(0.75) - q(0.25)$.

The five number summary provides a concise numerical summary of a data set which provides information about the location (through the median), the spread (through the lower and upper quartiles) and the range (through the minimum and maximum values).

Definition 6 *The* five number summary *of a data set consists of the smallest observation, the lower quartile, the median, the upper quartile and the largest value, that is, the five values:* $y_{(1)}$, $q(0.25)$, $q(0.5)$, $q(0.75)$, $y_{(n)}$.

Example 1.3.2 Comparison of body mass index

In a study of obesity in New Zealand, a sample of 150 men and 150 women were selected from workers aged 18 to 60. The height in meters and weight in kilograms was measured for each subject (unit). These variates are both continuous variates. Height and weight were recorded to 2 decimal places. This means that there are a finite number of possible values for the recorded height and the recorded weight. This does not imply that height and weight are discrete variates. The accuracy of the measuring device does not change the type of the variate. The type of a variate is important because we use this information in choosing a probability model to analyse the data as discussed in Section 1.4. Variates such as height and weight are typically modelled using a continuous distribution such as the Gaussian distribution.

For each subject the body mass index (BMI) was also calculated using

$$\text{BMI} = \frac{\text{weight}(kg)}{[\text{height}(m)]^2}$$

BMI is a continuous variate. Often the value of BMI is used to classify a subject as being "overweight", "normal weight", "underweight", etc. One possible classification is given in Table 1.1.

Underweight	BMI < 18.5
Normal Weight	$18.5 \leq$ BMI < 25.0
Overweight	$25.0 \leq$ BMI < 30.0
Moderately Obese	$30.0 \leq$ BMI < 35.0
Severely Obese	$35.0 \leq$ BMI

Table 1.1: BMI classification

Suppose Table 1.1 was used to determine the BMI class for each subject and we called this new variate "BMI class". BMI class is an example of an ordinal variate.

The data are available in the file *bmidata.txt* posted on the course website. To analyse the data, it is convenient to record the data in row-column format (see Table 1.2). The first row of the file gives the variate names, in this case, subject number, gender (M = male or F = female), height, weight and BMI. Each subsequent row gives the variate values for a particular subject.

subject	gender	height	weight	BMI
1	M	1.76	63.81	20.6
2	M	1.77	89.60	28.6
3	M	1.91	88.65	24.3
4	M	1.80	74.84	23.1

Table 1.2: First 5 rows of the file bmidata.txt

Numerical summaries for the variate BMI for each gender are given in Table 1.3. We see that there are only small differences in the sample median and the sample mean for females and males. For the sample standard deviation, IQR and the range we notice that the values are all larger for the females. In other words, there is more variability in the BMI measurements for females than for males in this sample.

Gender	$y_{(1)}$	$q\,(0.25)$	$q\,(0.5)$	$q\,(0.75)$	$y_{(n)}$	IQR	range	$\bar{y}$	s	g_1	g_2
Female	16.4	23.4	26.8	29.75	38.8	6.35	22.4	26.92	4.60	0.30	2.79
Male	18.3	24.6	26.75	29.15	37.5	4.55	19.2	27.08	3.56	0.41	3.03

Table 1.3 Numerical Summaries for BMI by Gender

We can also construct a *relative frequency table* that gives the proportion of subjects that fall within each BMI class by gender (see Table 1.4). From the table we can see that the reason that the variability in the BMI variate for females is larger than for males is because there is a larger proportion of females in the two extreme classes "underweight" and "severely obese" as compared to the males.

BMI Class	Males	Females
Underweight	0.01	0.02
Normal Weight	0.28	0.33
Overweight	0.50	0.42
Moderately Obese	0.19	0.17
Severely Obese	0.02	0.06
Total	1.00	1.00

Table 1.4: BMI Class Relative Frequency Table by Sex

Sample correlation

So far we have looked only at numerical summaries of a data set $\{y_1, y_2, \ldots, y_n\}$. Often we have bivariate data of the form $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. A numerical summary of such data is the sample correlation.

Definition 7 *The sample correlation, denoted by* r*, for data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ *is*

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

where

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ S_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \\ S_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \end{aligned}$$

The sample correlation, which takes on values between -1 and 1, is a measure of the linear relationship between the two variates x and y. If the value of r is close to 1 then we say that there is a strong positive linear relationship between the two variates while if the value of r is close to -1 then we say that there is a strong negative linear relationship between the two variates. If the value of r is close to 0 then we say that there is no linear relationship between the two variates.

Example 1.3.2 Continued

If we let $x =$ height and $y =$ weight then the sample correlation for the males is $r = 0.55$ which indicates that there is a positive linear relationship between height and weight which is what we would expect. For the females $r = 0.31$ which also indicates that there is a positive linear relationship between height and weight but the relationship is not as strong as for males.

Relative risk

Recall that values for a categorical variate are category names that do not necessarily have any ordering. If two variates of interest in a study are categorical variates then the sample correlation cannot be used as a measure of the relationship between the two variates.

Example 1.3.3 Physicians' Health Study

During the 1980's in the United States a very large study called the Physicians' Health Study was conducted to study the relationship between taking daily aspirin and the occurrence of coronary heart disease (CHD). For each physician (unit) in the study two categorical variates were collected: (1) whether the physician was assigned to the daily aspirin group or the placebo group and (2) whether or not the physician experienced CHD during the study. The data can be summarized by giving the observed frequency for each of the four possible outcomes as shown in Table 1.5.

	CHD	No CHD	Total
Placebo	189	10845	11034
Daily Aspirin	104	10933	11037
Total	293	21778	22071

Table 1.5: Physicians' Health Study

To summarize the relationship between two categorical variates consider a generalized version of Table 1.5 given by

	A	$\bar{A}$	Total
B	y_{11}	y_{12}	$y_{11} + y_{12}$
$\bar{B}$	y_{21}	y_{22}	$y_{21} + y_{22}$
Total	$y_{11} + y_{21}$	$y_{12} + y_{22}$	n

Table 1.6: General two-way table

Recall that events A and B are independent events if $P(A \cap B) = P(A)\,P(B)$ or equivalently $P(A) = P(A|B) = P\left(A|\bar{B}\right)$. If A and B are independent events then

$$\frac{P(A|B)}{P\left(A|\bar{B}\right)} = 1$$

and otherwise the ratio is not equal to one. In the Physicians' Health Study if we let $A =$ takes daily aspirin and $B =$ experienced CHD then we can estimate this ratio using the ratio of the sample proportions.

Definition 8 *For categorical data in the form of Table 1.6 the* relative risk *of event A in group B as compared to group $\bar{B}$ is*

$$\textit{relative risk} = \frac{y_{11}/(y_{11}+y_{12})}{y_{21}/(y_{21}+y_{22})}$$

Example 1.3.3 Revisited

For the Physicians' Health Study the relative risk of CHD in the placebo group as compared to the aspirin group is

$$\begin{aligned}\text{relative risk} &= \frac{189/(189+10845)}{104/(104+10933)} \\ &= 1.82\end{aligned}$$

The data suggest that the group taking the placebo are nearly twice as likely to experience CHD as compared to the group taking the daily aspirin. Can we conclude that daily aspirin reduces the occurrence of CHD? The topic of causation will be discussed in more detail in Chapter 8.

In Chapter 7 we consider methods for analyzing data which can be summarized in a two way table like Table 1.6.

Graphical Summaries

Graphical summaries or data visualizations are important tools for seeing patterns in data and for communicating results. Although the graphical summaries we present here are quite simple, they provide the building blocks for more advanced visualizations used in data science and data mining.

We consider graphical summaries for both univariate data sets $\{y_1, y_2, \ldots, y_n\}$ and bivariate data sets $\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$.

Frequency histograms

Consider measurements $\{y_1, y_2, \ldots, y_n\}$ on a variate y. Partition the range of y into k non-overlapping intervals $I_j = [a_{j-1}, a_j)$, $j = 1, 2, \ldots, k$ and then calculate

$$f_j = \text{number of values from } \{y_1, y_2, \ldots, y_n\} \text{ that are in } I_j$$

for $j = 1, 2, \ldots, k$. The f_j are called the observed *frequencies* for $I_1, I_2, \ldots, I_k$; note that $\sum_{j=1}^{k} f_j = n$.

A *histogram* is a graph in which a rectangle is constructed above each interval $I_1, I_2, \ldots, I_k$. The height of the rectangle for interval I_j is chosen so that the area of the rectangle is proportional to f_j. Two main types of frequency histograms are:

(a) a "standard" frequency histogram where the intervals I_j are of equal length. The height of the rectangle for I_j is the frequency f_j or *relative frequency* f_j/n.

(b) a "relative" frequency histogram, where the intervals $I_j = [a_{j-1}, a_j)$ may or may not be of equal length. The height of the rectangle for I_j is set equal to

$$\frac{f_j/n}{a_j - a_{j-1}}$$

so that the area of the jth rectangle equals f_j/n. With this choice of height we have

$$\sum_{j=1}^{k} (a_j - a_{j-1}) \frac{f_j/n}{(a_j - a_{j-1})} = \frac{1}{n} \sum_{j=1}^{k} f_j = \frac{n}{n} = 1$$

so the total area of the rectangles is equal to one.

If intervals of equal length are used then a standard frequency histogram and a relative frequency histogram look identical except for the labeling of the vertical axis. As just shown, the sum of the areas of the rectangles for a relative frequency histogram equals one. Recall that the area under a probability density function for a continuous random variable equals one. Therefore if we wish to superimpose a probability density function on a histogram to see how well the model fits the data we must use a relative frequency histogram. If we wish to compare two data sets which have different sample sizes then a relative frequency histogram must always be used. The vertical axis is labelled "density" to emphasize that such a histogram is being used.

To construct a frequency histogram, the number and location of the intervals must be chosen. The intervals are typically selected so that there are ten to fifteen intervals and each interval contains at least one y value from the sample (that is, each $f_j \geq 1$). If a software package is used to produce the frequency histogram then the intervals are usually chosen automatically. An option for user specified intervals is also usually provided.

Example 1.3.2 Continued

Figures 1.5 and 1.6, give the relative frequency histograms for BMI for males and females separately. We often say that histograms show the *distribution* of the data. The *shapes* of the two relative frequency histograms are somewhat bell-shaped. In each case the skewness is positive but close to zero while the kurtosis is close to three.

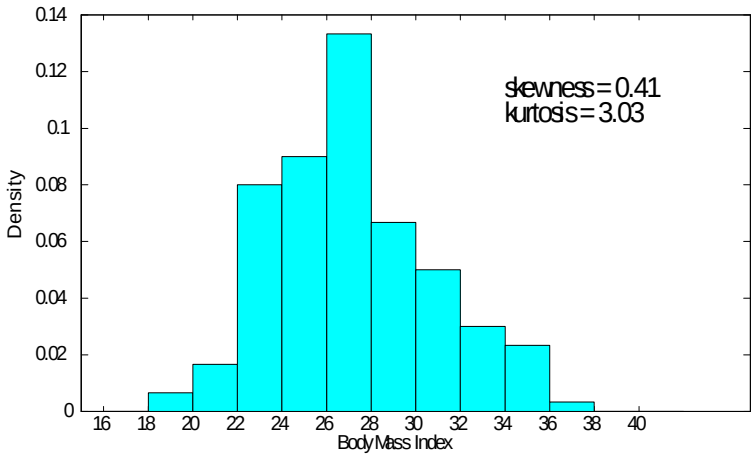


Figure 1.5: **Relative frequency histogram for male BMI data**

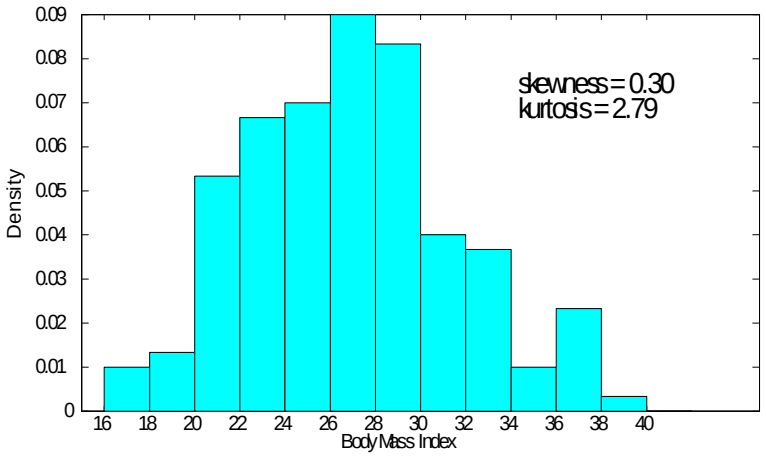


Figure 1.6: **Relative frequency histogram for female BMI data**

Example 1.3.4 Lifetimes of brake pads

A frequency histogram can have many different shapes. Figure 1.7 shows a relative frequency histogram of the lifetimes (in terms of number of thousand km driven) for the front brake pads on 200 new mid-size cars of the same type. The variate lifetime is a continuous variate.

The data are available in the file *brakepaddata.txt* posted on the course website. The relative frequency histogram for brake pad lifetimes shown in Figure 1.7 is not symmetric and has a long right tail which is consistent with a sample skewness of 1.28. Since the distribution is not symmetric the sample kurtosis is not of interest. The sample mean $\bar{y} = 49.03$ thousand km is larger than the sample median $\hat{m} = 40.35$ thousand km which is what we would expect for unimodal data with a long right tail. The sample standard deviation is $s = 36.65$ thousand km. The large variability in lifetimes is due to the wide variety of driving conditions which different cars are exposed to, as well as to variability in how soon car owners decide to replace their brake pads.

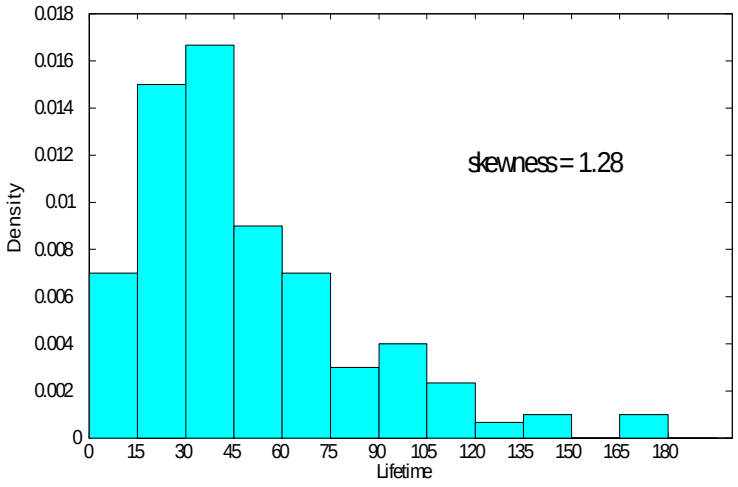


Figure 1.7: **Relative frequency histogram of brake pad lifetime data**

Bar Graphs

For categorical data, a bar graph or bar chart is a useful graphical summary. A bar graph has a bar for each of the possible values of the categorical variate with height equal to the frequency or relative frequency of that category. Usually the order of the different possible categories is not important. The width of the bar is also not important. Gaps are left between the bars to emphasize that the data are categorical.

Example 1.3.5 Global market share of browsers

The bar chart in Figure 1.8 shows the global market share of browsers in June 2017 according to StatCounter, a web analytics company. What data might StatCounter have collected to create this graphical summary? See gs.stat.counter.com for details.

Figure 1.9 illustrates how a bar graph can be used to compare the global market share of browsers in June 2015, June 2016, and June 2017.

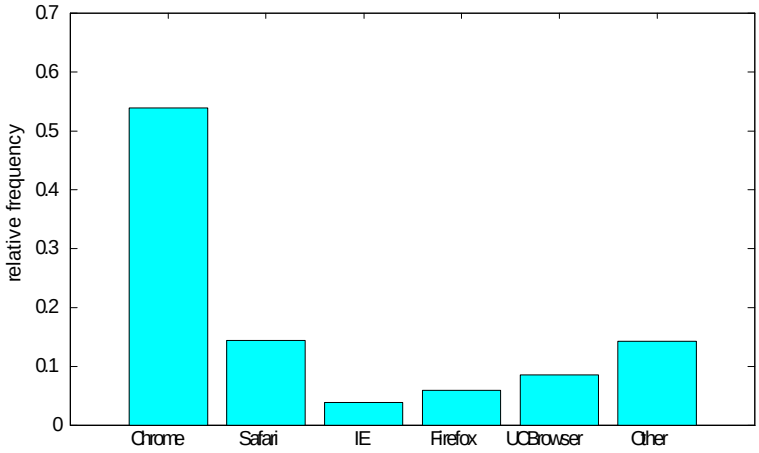


Figure 1.8: **Global market share of browsers June 2017**

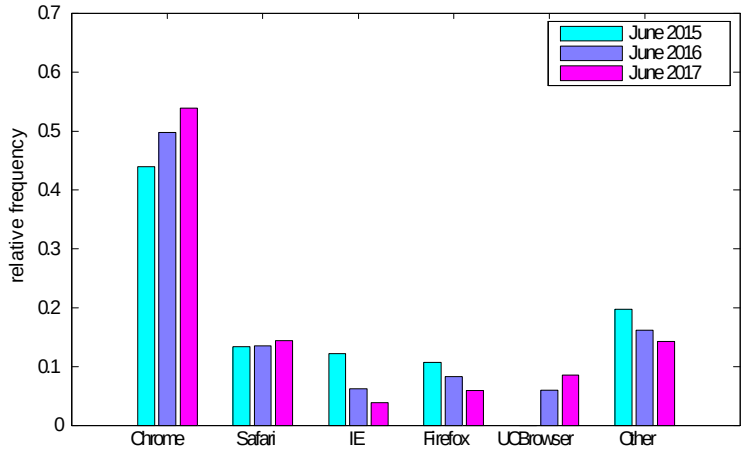


Figure 1.9: **Global market share of browsers June 2015-2017**

Pie charts, which are another way to display categorical data, are often used in the media. Pie charts are used very infrequently by statisticians since the human eye is not good at judging how much area is taken up by a wedge.

Bar graphs and pie charts are often used incorrectly in the media. See Chapter 1, Problems 21-25.

Empirical cumulative distribution function

Consider the following data set of 10 observations

$$3.1 \quad 0.6 \quad 1.6 \quad 1.8 \quad 0.3 \quad 3.8 \quad 1.0 \quad 0.8 \quad 2.9 \quad 1.7$$

Order the observations from smallest to largest to obtain the order statistic

$$0.3 \quad 0.6 \quad 0.8 \quad 1.0 \quad 1.6 \quad 1.7 \quad 1.8 \quad 2.9 \quad 3.1 \quad 3.8$$

Suppose we assume that these observations come from an unknown cumulative distribution $F(y) = P(Y \leq y)$. If we wanted to estimate $F(1.5) = P(Y \leq 1.5)$ then intuitively it seems reasonable to estimate this probability by determining the proportion of observations which are less than or equal to 1.5. Since there are four such values (0.3, 0.6, 0.8 and 1.0), we estimate $F(1.5)$ by $\hat{F}(1.5) = \frac{4}{10} = 0.4$. Since there are no observations between 1.0 and 1.6 then for any $y \in [1.0, 1.6)$ we would estimate $F(y) = P(Y \leq y)$ using $\hat{F}(y) = 0.4$.

We can estimate $F(y) = P(Y \leq y)$ in a similar way for any value of y. This leads us to the following definition:

Definition 9 *For a data set* $\{y_1, y_2, \ldots, y_n\}$*, the* empirical cumulative distribution function *or* e.c.d.f. *is defined by*

$$\hat{F}(y) = \frac{\textit{number of values in the set } \{y_1, y_2, \ldots, y_n\} \textit{ which are} \leq y}{n} \quad \textit{for all } y \in \Re$$

The empirical cumulative distribution function is an estimate, based on the data, of the population cumulative distribution function.

A graph of $\hat{F}(y)$ gives us a graphical summary of the data set $\{y_1, y_2, \ldots, y_n\}$.

For the data set of 10 observations, the graph of $\hat{F}(y)$ is given in Figure 1.10. The vertical lines are added to make the graph look visually more like a cumulative distribution function. We note that $\hat{F}(y)$ jumps a height of 0.1 at each of the unique values in the ordered data set.

More generally, for an ordered data set $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ of unique observations, $\hat{F}(y_{(j)}) = j/n$ and the jumps are all of size $1/n$.

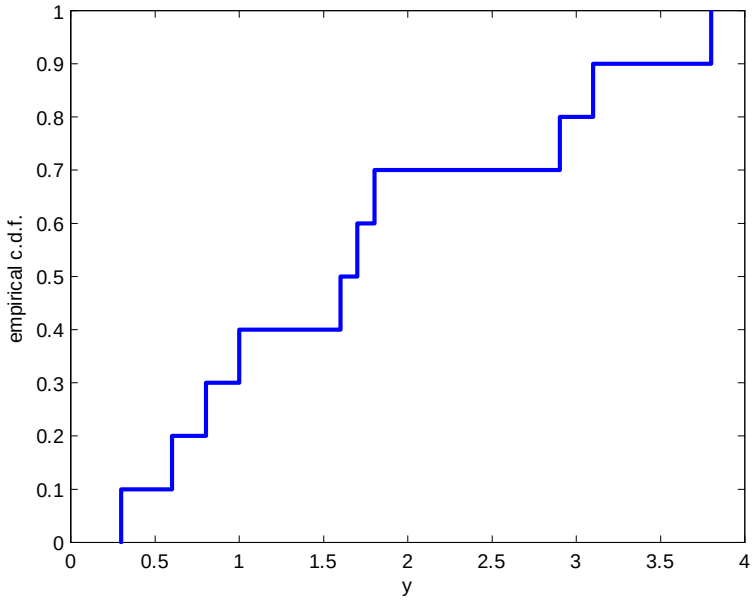


Figure 1.10: Empirical cumulative distribution function for 10 observations

Example 1.3.6

Figure 1.11 shows a graph of the empirical cumulative distribution function for 100 observations which were randomly generated from an Exponential model and then rounded to one decimal place. The observations are not all unique. In particular, at $y = 4.3$ there is a jump of height 0.04 which would indicate that there are 4 observations equal to 4.3.

The plot of the empirical cumulative distribution function does not show the shape of the distribution quite as clearly as a plot of the relative frequency histogram does. It requires more effort to determine if the distribution of the data is symmetric or skewed. We can see from Figure 1.11 that for this data set the values of $q(p)$ are changing more rapidly for $p \geq 0.8$. This means that the distribution of the data is not symmetric and has a long right tail.

Often when large data sets are reported in the media or research journals the individual observations are not reported. Sometimes only a graph like the empirical cumulative distribution function is given. What information can we obtain from the graph of the empirical cumulative distribution function? In addition to the information about the shape mentioned above, the graph allows us to determine the pth quantile or $100p$th percentile $q(p)$. For example, from Figure 1.11 we can determine, using the red dashed lines, that the lower quartile $= q(0.25) = 0.9$, the median $= q(0.5) = 2.6$, and the upper quartile $= q(0.75) = 5.3$. These are not exactly the same values that would be obtained if we had all the data and used Definition 3, however the values would be very close. From $q(0.75)$ and $q(0.25)$ we can determine that the $IQR = q(0.75) - q(0.25) = 5.3 - 0.9 = 4.4$. Finally we can also see that $y_{(1)} = 0.0$ and $y_{(100)} = 16.1$ and therefore the range $= 16.1 - 0.0 = 16.1$.

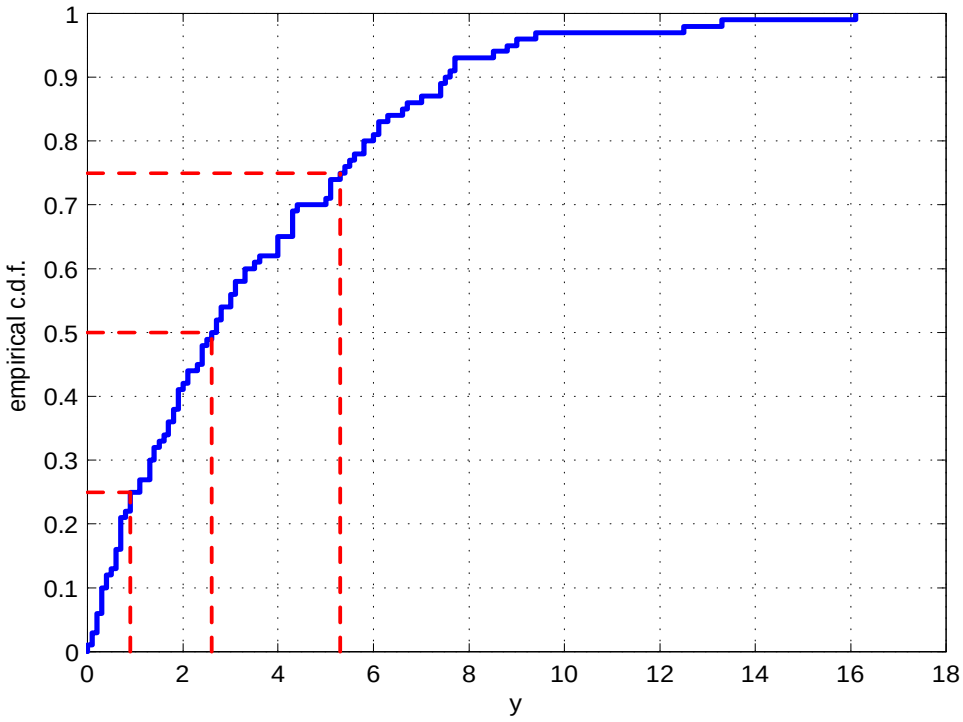


Figure 1.11: **Empirical cumulative distribution function of** 100 **observations**

The empirical cumulative distribution function can also be used to compare two data sets by graphing their empirical cumulative distribution functions on the same graph as shown in the next example.

Example 1.3.2 Continued

Figure 1.12 shows the empirical cumulative distribution function for male and female heights on the same plot for the data in the file *bmidata.txt* posted on the course website. As you might expect we see that the distribution of male heights is similar to the distribution of female heights but shifted to the right reflecting the fact that males are generally taller than females.

We can also determine from Figure 1.12 that the median height for females is 1.60 and for males the median height is 1.73. The symmetry of the two curves about their respective medians indicates that the distribution of heights is reasonably symmetric for both males and females.

For females $q(0.25) = 1.57$, $q(0.75) = 1.67$, $IQR = 1.67 - 1.57 = 0.1$, and range $= 1.79 - 1.41 = 0.38$. For males $q(0.25) = 1.71$, $q(0.75) = 1.79$, $IQR = 1.79 - 1.71 = 0.08$, and range $= 1.93 - 1.56 = 0.37$. The IQR and range for females are very similar to the IQR and range for males.

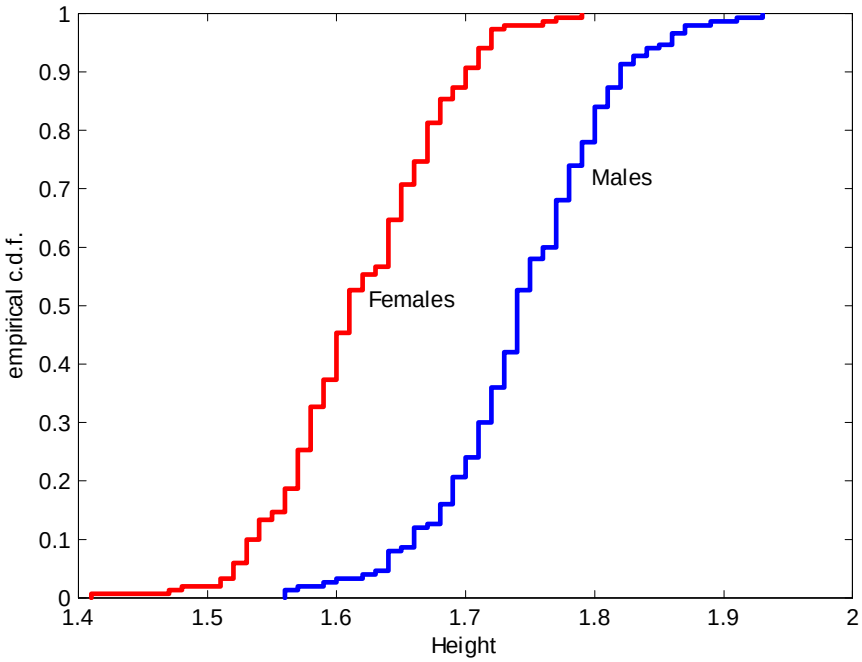


Figure 1.12: **Empirical cumulative distribution function of heights for males and for females**

Boxplots

In many situations, we want to compare the values of a variate for two or more groups. For example, in Example 1.3.2 Continued we compared the heights for males versus females by plotting side-by-side empirical distribution functions. When the number of groups is large or the sample sizes within groups are small, side-by-side *boxplots* (also called *box and whisker plots*) are a convenient way to display the data.

A boxplot gives a graphical summary about the shape of the distribution of the data and is usually displayed vertically. The line inside the box corresponds to the sample median $q(0.5)$. The top edge of the box corresponds to the upper quartile $q(0.75)$ and the lower edge of the box corresponds to the lower quartile $q(0.25)$. The so-called whiskers extend down and up from the box to a horizontal line. The lower line is placed at the smallest observed data value that is larger than the value $q(0.25) - 1.5 \times IQR$ where $IQR = q(0.75) - q(0.25)$ is the interquartile range. The upper line is placed at the largest observed data value that is smaller than the value $q(0.75) + 1.5 \times IQR$. Values beyond the whiskers (often called outliers) are plotted with special symbols.

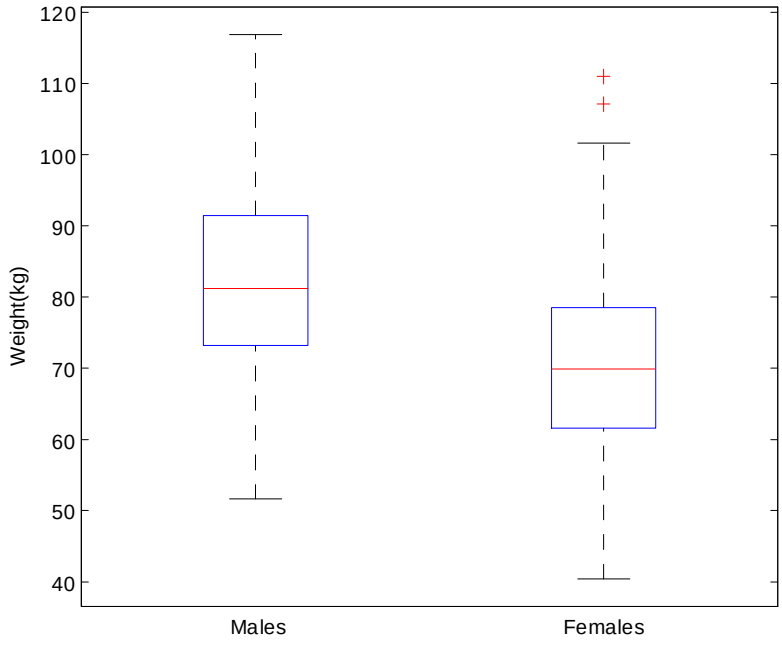


Figure 1.13: **Boxplots of weights for males and females**

Figure 1.13 displays side-by-side boxplots of male and female weights from Example 1.3.2. As mentioned previously, when large data sets are reported in the media or research journals the individual observations are not reported. What information can we obtain from these two boxplots?

The shape and spread of the two boxplots are very similar. For the males and the females, the center line in the box, which corresponds to the sample median, divides both the box and the whiskers approximately in half. This indicates that the distributions of both data sets are roughly symmetric about the median. For the females there are two large outliers.

For the boxplots we can determine that the sample median weight for females is approximately 70 and for males the sample median weight is approximately 81. For females $q(0.25) = 62$, $q(0.75) = 79$, $IQR = 79 - 62 = 17$, and range $= 111 - 40 = 71$. For males $q(0.25) = 73$, $q(0.75) = 91$, $IQR = 91 - 73 = 18$, and range $= 117 - 52 = 65$. The IQR and range for females are very similar to the IQR and range for males.

Since the boxplot for the males is shifted up relative to the boxplot for females this implies that males generally weigh more than females.

Boxplots are particularly useful for comparing more than two groups.

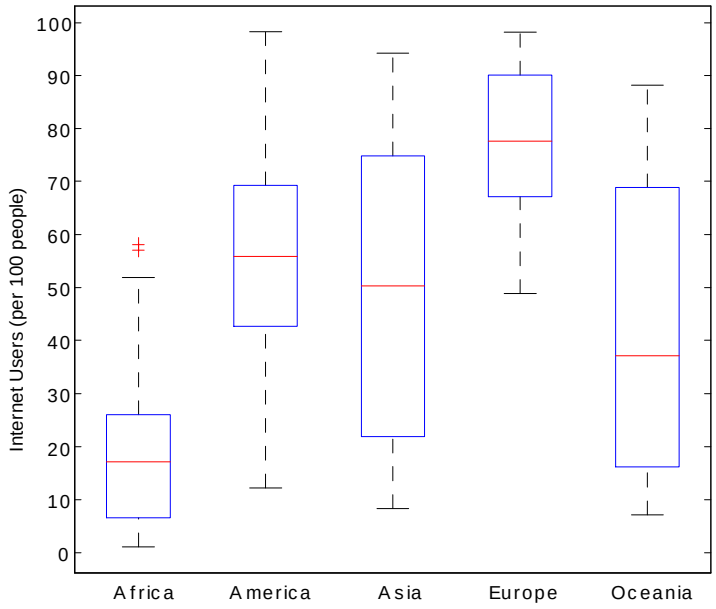


Figure 1.14: **Boxplots of internet use by different continents**

Figure 1.14 shows a comparison of internet users (per 100 people) in 2015 for countries in the world classified by continent (worldbank.org). The side-by-side boxplots make it easy to see the differences and similarities between the countries in different continents. In this example a unit is a country. The variate of interest which is measured for each country is the number of internet users per 100 people. What type of variate is this? Why is the total number of internet users not used?

For which continent is the sample median number of internet users per 100 people the smallest? For which continent is the sample median number of internet users per 100 people the largest? For which continent is the IQR the smallest? For which continent is the IQR the largest? For which continent is the range the smallest? For which continent is the range the largest? For which continent is the variability the smallest? For which continent is the variability the largest? For which continent is the distribution of the data most symmetric? For which continent is the distribution of the data most asymmetric?

The graphical summaries discussed to this point deal with a single variate. If we have data on two variates x and y for each unit in the sample then the data set is represented as $\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$. We are often interested in examining the relationships between the two variates.

Scatterplots

A *scatterplot*, which is a plot of the points (x_i, y_i), $i = 1, 2, \ldots, n$, can be used to see whether two variates are related in some way.

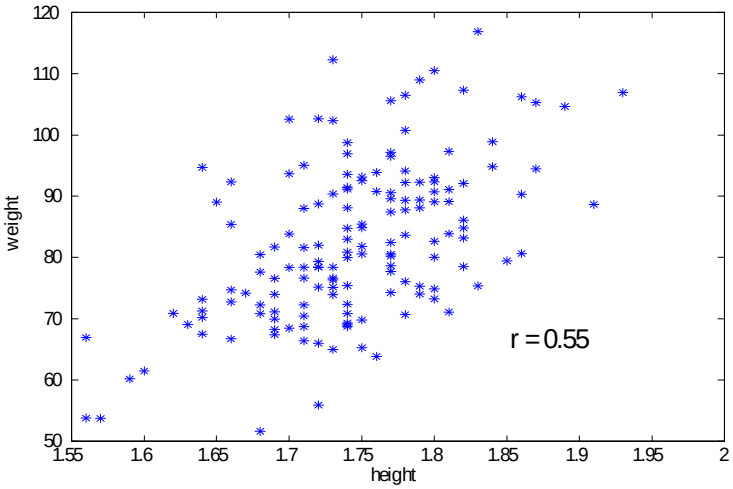


Figure 1.15: **Scatterplot of weight versus height for males**

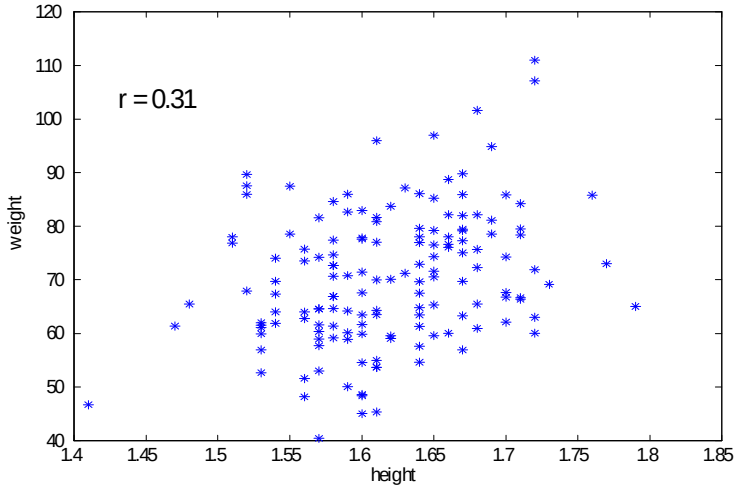


Figure 1.16: **Scatterplot of weight versus height for females**

Figures 1.15 and 1.16 give the scatterplots of y = height versus x = weight for males and females respectively for the data in Example 1.3.2. As expected, there is a tendency for weight to increase as height increases for both sexes. What might be surprising is the variability in weights for a given height.

Run charts

A *run chart* is another type of two dimensional plot which is used when we are interested in a graphical summary which illustrates how a single variate is changing over time.

In Figure 1.17 the run chart shows the closing value of the Canadian dollar in Chinese yuan for the 67 business days between May 1 and August 1, 2017. For example on August 1, 2017 the Canadian dollar was worth 5.3543 Chinese yuan. The data are from google.com/finance. In a run chart consecutive points are joined with straight lines.

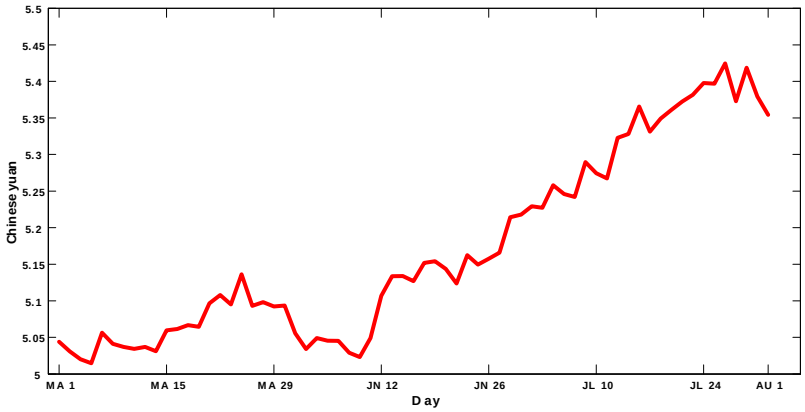


Figure 1.17: **Value of Canadian dollar in Chinese yuan May-July 2017**

In Figure 1.18 the market share for the browsers Chrome, Safari and Internet Explorer is graphed versus the months between June 2016 and July 2017 (gs.stat.counter.com).

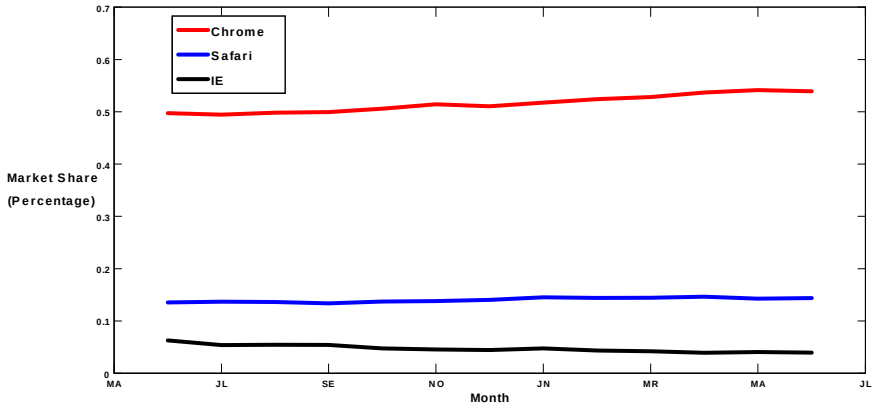


Figure 1.18: **Market share for browsers June 2016 to July 2017**

Note that for these data sets the sample correlation coefficient is not meaningful. Why?

1.4 Probability Distributions and Statistical Models

In your previous probability course you were introduced to the following statistical models and their physical setups: Binomial, Poisson, Uniform, Exponential, and Gaussian (Normal). In this course we use statistical models to describe processes such as the daily closing value of a stock or the occurrence and size of claims over time in a portfolio of insurance policies. For populations, we use a statistical model to describe the selection of the units and the measurement of the variates. The model depends on the distribution of variate values in the population (the population histogram is a graphical summary of this distribution) and the selection procedure. We exploit this connection when we want to estimate attributes of the population and quantify the uncertainty in our conclusions. We use the models in several ways:

- questions are often formulated in terms of parameters of the model
- the variate values vary so random variables can describe this variation
- empirical studies usually lead to inferences that involve some degree of uncertainty, and probability is used to quantify this uncertainty
- procedures for making decisions are often formulated in terms of models
- models allow us to characterize processes and to simulate them via computer experiments

Data summaries and properties of probability models

If we model the selection of a data set $\{y_1, y_2, \ldots, y_n\}$ as n independent realizations of a random variable Y, we can draw strong parallels between summaries of the data set described in Section 1.3 and properties of the corresponding probability model for Y. For example,

- The sample mean $\bar{y}$ corresponds to the population mean $E(Y) = \mu$.
- The sample standard deviation s corresponds to σ, the population standard deviation of Y, where $\sigma^2 = E[(Y - \mu)^2]$.
- The sample median $\hat{m}$ corresponds to the population median m. For continuous distributions the population median is the solution m of the equation $F(m) = 0.5$ where $F(y) = P(Y \leq y)$ is the cumulative distribution function of Y. For discrete distributions, it is a point m chosen such that $P(Y \leq m) \geq 0.5$ and $P(Y \geq m) \geq 0.5$.
- The relative frequency histogram corresponds to the probability histogram of Y for discrete distributions and the probability density function of Y for continuous distributions.

Example 1.4.1 A Binomial distribution example

Consider again the survey of smoking habits of teenagers described in Example 1.2.1. To select a sample of 500 units (young adults aged $14-20$), suppose we had a list of most of the units in the population of interest (young adults aged $14-20$ living in Ontario at the time of the study). Getting such a list would be expensive and time consuming so the actual selection procedure is likely to be very different. We select a sample of 500 units from the list at random and count the number of smokers in the sample. We model this selection process using a Binomial random variable Y with probability function (p.f.)

$$\begin{aligned} f(y;\theta) &= P(Y=y;\theta) \\ &= \binom{500}{y}\theta^y(1-\theta)^{500-y} \quad \text{for } y=0,1,\ldots,500 \text{ and } 0\le\theta\le 1 \end{aligned}$$

(Note that the sampling would be done without replacement so we are assuming that the number sampled is small relative to the total number in the population.) The parameter θ in the probability function represents the unknown proportion of smokers in the population of young adults aged $14-20$ living in Ontario at the time of the study, which is one attribute of interest in the study.

Note that we use the notation $P(Y=y;\theta)$ and $f(y;\theta)$ to emphasize the importance of the parameter θ in the model.

Example 1.4.2 An Exponential distribution example

In Example 1.3.4, we examined the lifetime (in 1000 km) of a sample of 200 front brake pads taken from the population of all cars of a particular model produced in a given time period. We can model the lifetime of a single brake pad by a continuous random variable Y with Exponential probability density function (p.d.f.)

$$f(y;\theta)=\frac{1}{\theta}e^{-y/\theta} \quad \text{for } y>0$$

The parameter $\theta>0$ represents the mean lifetime of the brake pads in the population since, in the model, the expected value of Y is $E(Y)=\theta$.

To model the sampling procedure, we assume that the data $\{y_1,y_2,\ldots,y_{200}\}$ represent 200 independent realizations of the random variable Y. That is, we let $Y_i=$ the lifetime for the ith brake pad in the sample, $i=1,2,\ldots,200$, and we assume that $Y_1,Y_2,\ldots,Y_{200}$ are independent Exponential random variables each having the same mean θ.

We can use the model and the data to estimate θ and other attributes of interest such as the proportion of brake pads that fail in the first $100,000$ km of use. In terms of the model, we can represent this proportion by

$$P(Y\le 100;\theta)=\int_0^{100} f(y;\theta)dy = 1-e^{-100/\theta}$$

Example 1.4.3 A Gaussian distribution example

Earlier, we described an experiment where the goal was to see if there is a relationship between operating performance y of a computer chip and ambient temperature x. In the experiment, there were four groups of 10 chips and each group operated at a different temperature $x = 10, 20, 30, 40$. The data are $\{(x_1, y_1), (x_2, y_2), \ldots, (x_{40}, y_{40})\}$. A model for $Y_1, Y_2, \ldots, Y_{40}$ should depend on the temperatures x_i and one possibility is to assume $Y_i \sim G(\beta_0 + \beta_1 x_i, \sigma)$, $i = 1, 2, \ldots, 40$ independently. In this model, the mean of Y is a linear function of the temperature x_i. The parameter σ allows for variability in performance among chips operating at the same temperature. We will consider such models in detail in Chapter 6.

Response versus explanatory variates

Suppose we wanted to study the relationship between second hand smoke and asthma among children aged 10 and under. The two variates of interest could be defined as:

x = whether the child lives in a household where adults smoke,

Y = whether the child suffers from asthma.

In this study there is a natural division of the variates into two classes: response variate and explanatory variate. In this example Y, the asthma status, is the response variate (often coded as $Y = 1$ if child suffers from asthma, $Y = 0$ otherwise) and x, whether the child lives in a household where adults smoke, is the explanatory variate (also often coded as $x = 1$ if child lives in household where adults smoke and $x = 0$ otherwise). The explanatory variate x is in the study to investigate whether the distribution of the response variate Y is different for different observed values of x.

Similarly in an observational study of 1718 men aged $40 - 55$, the men were classified according to whether they were heavy coffee drinkers (more than 100 cups/month) or not (less than 100 cups/month) and whether they suffered from CHD (coronary heart disease) or not. In this study there are also two categorical variates. One variate is the amount of coffee consumption while the other variate is whether or not the subject had experienced CHD or not. The question of interest is whether there is a relationship between coffee consumption and CHD. Unlike Example 1.4.3, neither variate is under the control of the researchers. We might be interested in whether coffee consumption can be used to "explain" CHD. In this case we would call coffee consumption an explanatory variate while CHD would be the response variate. However if we were interested in whether CHD can be used to explain coffee consumption (a somewhat unlikely proposition to be sure) then CHD would be the explanatory variate and coffee habits would be the response variate.

In some cases it is not clear which is the explanatory variate and which is the response variate. For example, suppose we were interested in studying the relationship between weight and height in the current population of females aged $16 - 25$ living in Waterloo. We might choose the response variate Y to be the weight and the explantory variate to be the height x of a randomly selected female. If we think that the average (expected) weight of a female depends linearly on her height x, then $E(Y|x) = \alpha + \beta x$ and we might choose to

model the conditional distribution of Y given x, as $G(\alpha + \beta x, \sigma)$. If, on the other hand, we are interested in the distribution of heights for a given weight then we would choose the response variate to be the height X and the explanatory variate to be the weight y.

Models for describing the relationships among two or more variates are considered in more detail in Chapters 6 and 7.

1.5 Data Analysis and Statistical Inference

Whether we are collecting data to increase our knowledge or to serve as a basis for making decisions, proper analysis of the data is crucial. We distinguish between two broad aspects of the analysis and interpretation of data. The first is what we refer to as *descriptive statistics*. This is the portrayal of the data, or parts of it, in numerical and graphical ways so as to show features of interest. (On a historical note, the word "statistics" in its original usage referred to numbers generated from data; today the word is used both in this sense and to denote the discipline of Statistics.) We have considered a few methods of descriptive statistics in Section 1.3. The terms data mining and knowledge discovery in data bases (KDD) refer to exploratory data analysis where the emphasis is on descriptive statistics. This is often carried out on very large data bases. The goal, often vaguely specified, is to find interesting patterns and relationships

A second aspect of a statistical analysis of data is what we refer to as *statistical inference*. That is, we use the data obtained in the study of a process or population to draw general conclusions about the process or population itself. This is a form of inductive inference, in which we reason from the specific (the observed data on a sample of units) to the general (the target population or process). This may be contrasted with deductive inference (as in logic and mathematics) in which we use general results (e.g. axioms) to prove specific things (e.g. theorems).

This course introduces some basic methods of statistical inference. Three main types of statistical methods will be discussed, loosely referred to as *estimation, hypothesis tests*, and *prediction*. Methods of estimation are used when we are interested in estimating one or more attributes of a process or population based on observed data. For example, we may wish to estimate the proportion of Ontario residents aged $14-20$ who smoke, or to estimate the distribution of survival times for certain types of AIDS patients. Another type of estimation problem is that of "fitting" or selecting a probability model for a process. Methods of estimation are discussed in all chapters of these Course Notes.

Hypothesis tests involve using the data to assess the truth of some question or hypothesis about the population or process. For example, we may hypothesize that in the $14-20$ age group a higher proportion of females than males smoke, or that the use of a new treatment will increase the average survival time of AIDS patients by at least 50 percent. Tests of hypotheses will be discussed in more detail in Chapter 5.

Prediction methods are used when we use the observed data to predict a future value

for a variate of a unit to be selected from the process or population. For example, based on the results of a clinical trial such as Example 1.2.3, we may wish to predict how much an individual's blood pressure would drop for a given dosage of a new drug, or, given the past performance of a stock and other data, to predict the value of the stock at some point in the future. Examples of prediction methods are given in Sections 4.7 and 6.2.

Statistical analysis involves the use of both descriptive statistics and formal methods of estimation, prediction and hypothesis testing. As brief illustrations, we return to the first two examples of section 1.2.

Example 1.5.1 Smoking behaviour survey

Suppose in Example 1.2.1, we sampled 250 males and 250 females aged $14-20$ as described in Example 1.4.1. Here we focus only on the sex of each person in the sample, and whether or not they smoked. The data are summarized in the following two-way table:

	Smokers	Non-smokers	Total
Female	82	168	250
Male	71	179	250
Total	153	347	500

Suppose we are interested in the question "Is the smoking rate among females higher than the rate among males?" From the data, we see that the sample proportion of females who smoke is $82/250 = 0.328$ or 32.8% and the sample proportion of males who smoke is $71/250 = 0.284$ or 28.4%. In the sample, the smoking rate for females is higher. But what can we say about the whole population? To proceed, we formulate the hypothesis that there is no difference in the population rates. Then assuming the hypothesis is true, we construct two Binomial models as in Example 1.4.1 each with a common parameter θ. We can estimate θ using the combined data so that $\hat{\theta} = 153/500 = 0.306$ or 30.6%. Then using the model and the estimate, we can calculate the probability of such a large difference in the observed rates. Such a large difference occurs about 20% of the time (if we selected samples over and over and the hypothesis of no difference is true) so such a large difference in observed rates happens fairly often and therefore, based on the observed data, there is no evidence of a difference in the population smoking rates. In Chapter 7 we discuss a formal method for testing the hypothesis of no difference in rates between females and males.

Example 1.5.2 Can filler study

Recall Example 1.2.2 where the purpose of the study was to compare the performance of the two machines in the future. A study was conducted in which one can from the new machine and one can from the old machine were selected each hour over a period of 40 hours. The volume in milliliters of each selected can was measured. Volume is a continuous variate. The data are available in the file *canfillingdata.txt* posted on the course website.

First we examine if the behaviour of the two machines is stable over time. In Figures 1.19 and 1.20, a run chart of the volumes over time for each machine is given. There is no

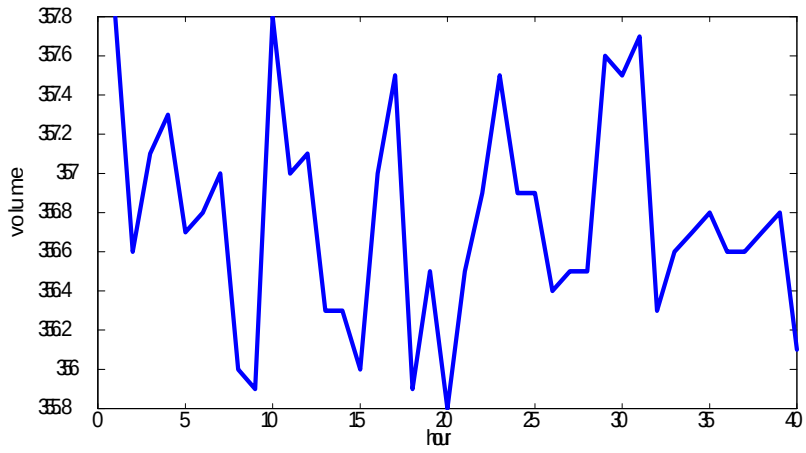


Figure 1.19: **Run chart of the volume (ml) for the new machine over time**

indication of a systematic pattern for either machine so we have some confidence that the data can be used to predict the performance of the machines in the near future.

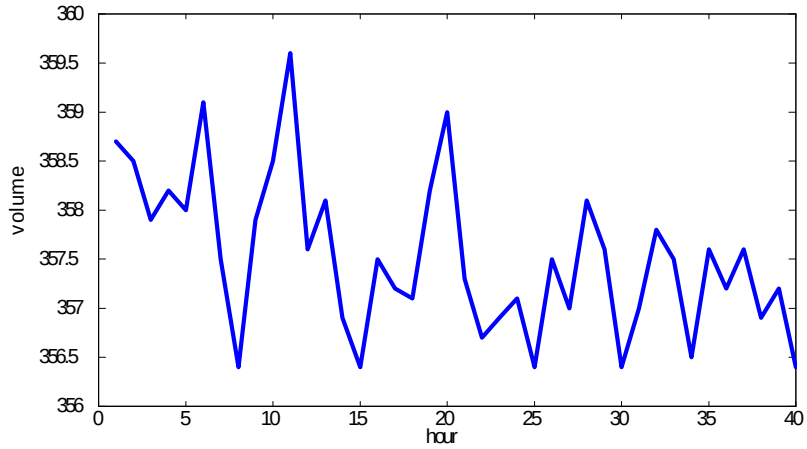


Figure 1.20: **Run chart of the volume for old machine over time**

The sample mean and standard deviation for the new machine are 356.8 and 0.54 ml respectively and, for the old machine, are 357.5 and 0.80. Figures 1.21 and 1.22 show the relative frequency histograms of the volumes for the new machine and the old machine respectively. To see how well a Gaussian model might fit these data we superimpose Gaussian probability density functions with the mean equal to the sample mean and the standard deviation equal to the sample standard deviation on each histogram. The agreement is reasonably good given that the sample size for both data sets is only 40.

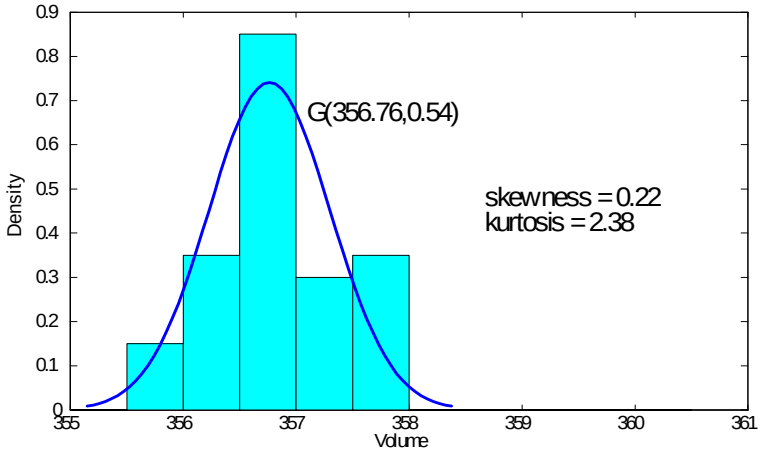


Figure 1.21: **Relative frequency histogram of volumes (ml) for the new machine**

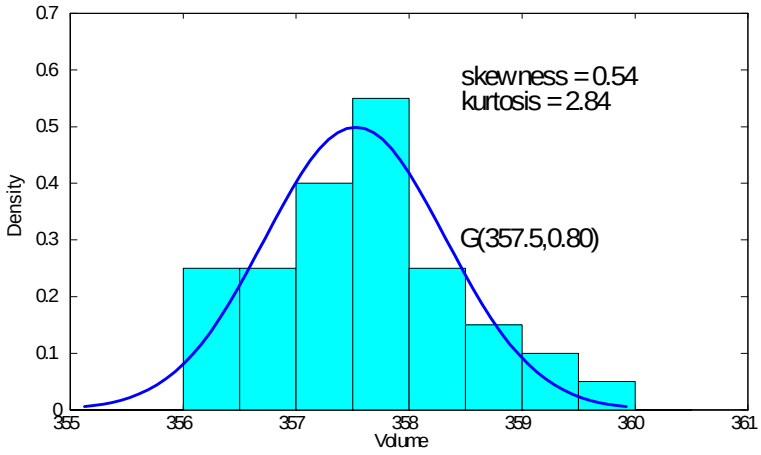


Figure 1.22: **Relative frequency histogram of volumes (ml) for the old machine**

We can use the Gaussian model to estimate the long term proportion of cans that fall below the required volume of 355 ml. For the new machine, $Y \sim G(356.8, 0.54)$ and $P(Y \leq 355) = 0.0005$ so about 5 in $10,000$ cans will be under-filled. For the old machine, $Y \sim G(357.5, 0.80)$ and $P(Y \leq 355) = 0.0008$ so about 8 in $10,000$ cans will be under-filled. Of course these estimates are subject to a high degree of uncertainty because they are based on small sample sizes.

We can see that the new machine is superior because of its smaller sample mean which translates into less overfill and hence less cost to the manufacturer. It is possible to adjust the mean of the new machine to a lower value because of its smaller standard deviation.

1.6 Statistical Software and R

Statistical software is essential for data manipulation and analysis. It is also used to deal with numerical calculations, to produce graphics, and to simulate probability models. There are many statistical software systems; some of the most comprehensive and popular are SAS, S-Plus, SPSS, Strata, Systat Minitab and R. Spreadsheet software such as EXCEL is also useful.

We will use the R software system since it has the lowest cost (free!) and the greatest functionality. It is an open source package that has extensive statistical capabilities and very good graphics procedures. R is the most common programming language among data scientists according to the 2016 O'Reilly Data Science Salary Survey (www.oreilly.com/ideas).

Information about how to use R is available in the document *Introduction to* R *and* RStudio which is posted on the course website.

1.7 Chapter 1 Problems

1. The sample mean $\bar{y}$ and the sample median $\hat{m}$ are two ways to measure the location of a data set $\{y_1, y_2, \ldots, y_n\}$.

   (a) Prove the identity
   $$\sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

   (b) Suppose the data are transformed using $u_i = a + by_i$, $i = 1, 2, \ldots, n$ where a and b are constants with $b \neq 0$. How are the sample mean and sample median of the data set $\{u_1, u_2, \ldots, u_n\}$ related to $\bar{y}$ and $\hat{m}$?

   (c) Suppose the data are transformed using $v_i = y_i^2$, $i = 1, 2, \ldots, n$. How are the sample mean and sample median of $v_1, v_2, \ldots, v_n$ related to $\bar{y}$ and $\hat{m}$?

   (d) Suppose another observation y_0 is added to the data set. Determine the mean of the augmented data set in terms of $\bar{y}$ and y_0. What happens to the sample mean as the magnitude of y_0 increases?

   (e) Suppose another observation y_0 is added to the data set. Determine the median of the augmented data set. What happens to the sample median as the magnitude of y_0 increases?

   (f) Use (d) and (e) to explain why the sample median income of a country might be a more appropriate summary than the sample mean income.

   (g) Show that $V(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2$ is minimized when $\mu = \bar{y}$.

   (h) **Challenge Problem:** Show that $W(\mu) = \sum_{i=1}^{n}|y_i - \mu|$ is minimized when $\mu = \hat{m}$.

2. The sample standard deviation s, the interquartile range $IQR = q(0.75) - q(0.25)$, and the range $= y_{(n)} - y_{(1)}$ are three different measures of the variability of a data set $\{y_1, y_2, \ldots, y_n\}$.

   (a) Prove the identity
   $$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i(y_i - \bar{y}) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

   (b) Suppose the data are transformed using $u_i = a + by_i$, $i = 1, 2, \ldots, n$ where a and b are constants and $b \neq 0$. How are the sample standard deviation, IQR, and range of the transformed data set $\{u_1, u_2, \ldots, u_n\}$ related to the sample standard deviation, IQR, and range of the original data set $\{y_1, y_2, \ldots, y_n\}$?

   (c) Suppose another observation y_0 is added to the data set. Use the result in (a) to write the sample standard deviation of the augmented data set in terms of y_0, s, and $\bar{y}$. What happens to the sample variance of the augmented data set as the magnitude of y_0 increases?

(d) If another observation y_0 is added to the data set, what happens to the IQR of the augmented data set as the magnitude of y_0 increases?

(e) If another observation y_0 is added to the data set, what happens to the range of the augmented data set as the magnitude of y_0 increases?

3. The sample skewness and kurtosis are two different measures of the shape of a data set $\{y_1, y_2, \ldots, y_n\}$. Let g_1 be the sample skewness and let g_2 be the sample kurtosis of the data set. Suppose we transform the data so that $u_i = a + by_i$, $i = 1, 2, \ldots, n$ where a and b are constants and $b \neq 0$. How are the sample skewness and sample kurtosis of the data set $\{u_1, u_2, \ldots, u_n\}$ related to g_1 and g_2?

4. Suppose the data $\{c_1, c_2, \ldots, c_{24}\}$ represents the costs of production for a firm every month from January 2018 to December 2019. For this data set the sample mean was \$2500, the sample standard deviation was \$5500, the sample median was \$2600, the sample skewness was 1.2, the sample kurtosis was 3.9, and the range was \$7500. The relationship between cost and revenue is given by $r_i = -7c_i + 1000$, $i = 1, 2, \ldots, 24$. Find the sample mean, standard deviation, median, skewness, kurtosis and range of the revenues.

5. Mass production of complicated assemblies such as automobiles depend on the ability to manufacture components to very tight specifications. The component manufacturer tracks performance by measuring a sample of parts and comparing the measurements to the specification. Suppose the specification for the diameter of a piston is a nominal value ± 10 microns (10^{-6} meters). The data below are the diameters of 50 pistons collected from the more than 10,000 pistons produced in one day. (The measurements are the diameters minus the nominal value in microns.) The data are available in the file *diameterdata.txt* posted on the course website.

−12.8	−7.3	−3.9	−3.4	−2.9	−2.7	−2.5	−2.3	−1.0	−0.9
−0.8	−0.7	−0.6	−0.4	−0.4	−0.2	0.0	0.5	0.6	0.7
1.2	1.8	1.8	2.0	2.1	2.5	2.6	2.6	2.7	2.8
3.3	3.4	3.5	3.8	4.3	4.6	4.7	5.1	5.4	5.7
5.8	6.6	6.6	7.0	7.2	7.9	8.5	8.6	8.7	8.9

$$\sum_{i=1}^{50} y_i = 100.7 \qquad \sum_{i=1}^{50} y_i^2 = 1110.79$$

(a) Plot a relative frequency histogram of the data. Is the process producing pistons within the specifications.

(b) Calculate the sample mean $\bar{y}$ and the sample median of the diameters.

(c) Calculate the sample standard deviation s and the IQR.

(d) Give the five number summary for these data.

(e) Such data are often summarized using a single performance index called Ppk defined as

$$Ppk = \min\left(\frac{U-\bar{y}}{3s},\ \frac{\bar{y}-L}{3s}\right)$$

where $(L, U) = (-10, 10)$ are the lower and upper specification limits. Calculate Ppk for these data.

(f) Explain why larger values of Ppk (i.e. greater than 1) are desirable.

(g) Suppose we fit a Gaussian model to the data with mean and standard deviation equal to the corresponding sample quantities, that is, with $\mu = \bar{y}$ and $\sigma = s$. Use the fitted model to estimate the proportion of diameters (in the process) that are out of specification.

6. In the above problem, we saw how to estimate the performance measure Ppk based on a sample of 50 pistons, a very small proportion of one day's production. To get an idea of how reliable this estimate is, we can model the process output by a Gaussian random variable Y with mean and standard deviation equal to the corresponding sample quantities. The following R code generates 50 observations and calculates Ppk. This is done 1000 times using a loop statement.

```
#Import dataset diameterdata.txt from the course website using RStudio
avgx<-mean(diameterdata$diameter)      #sample mean
sdx<-sd(diameterdata$diameter)         #sample standard deviation
temp<-rep(0,1000)  #Store the 1000 generated Ppk values in vector temp
for (i in 1:1000)  {  #Begin loop
y<-rnorm(50, avgx, sdx)   #Generate 50 new observations from a
#                         G(avgx,sdx) distribution
avg<-mean(y)       #sample mean of new data
s<-sd(y)           #sample std of new data
ppk<-min((10-avg)/(3*s),(avg+10)/(3*s))     #Ppk for new data
temp[i]<-ppk       #Store value of Ppk for this iteration
}
hist(temp)         #Plot histogram of 1000 Ppk values
mean(temp)         #average of the 1000 Ppk values
sd(temp)           #standard deviation of the 1000 Ppk values
```

(a) Compare the Ppk from the original data with the average Ppk value from the 1000 iterations. Mark the original Ppk value on the histogram of generated Ppk values. What do you notice? What would you conclude about how good the original estimate of Ppk was?

(b) Repeat the above exercise but this time use a sample of 300 pistons rather than 50 pistons. What conclusion would you make about using a sample of 300 versus 50 pistons?

7. Graph the empirical cumulative distribution function and boxplot for the data

   7.6 4.3 5.2 4.5 1.1 8.5 14.0 6.3 3.9 7.2

   without using statistical software.

8. Datasets 1 and 2 are data sets for continuous variates. The empirical cumulative distribution functions for these data sets are plotted in Figure 1.23. Answer the following questions based on this plot.

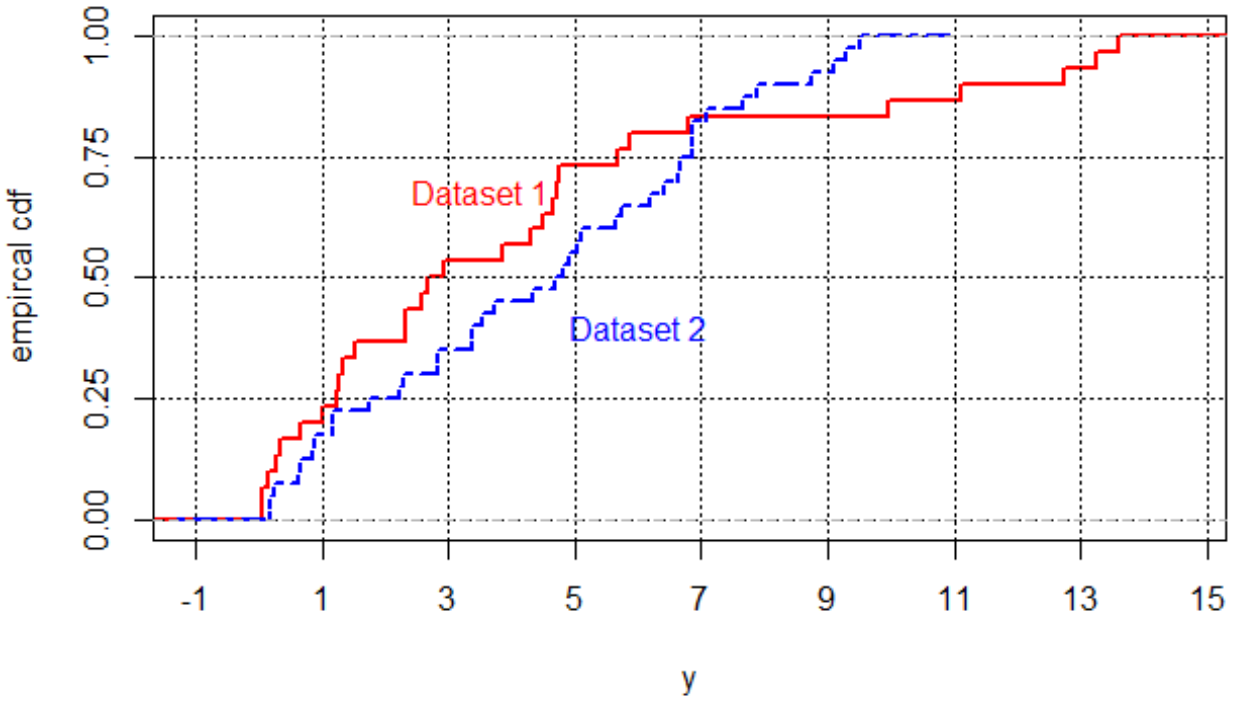


Figure 1.23: Plot for Problem 8

   (a) For each data set determine the approximate values of: $y_{(1)}$ (the minimum observation), $y_{(n)}$ (the maximum observation), the range, $q\,(0.25)$ (the lower quartile), $q\,(0.75)$ (the upper quartile), the IQR (the interquartile range), $\hat{m}$ (the median).

   (b) Which data set has more observations?

   (c) For each data set determine whether the corresponding relative frequency histogram would be reasonably symmetric, would not be reasonably symmetric but

would have a long left tail, or would not be reasonably symmetric but would have a long right tail.

(d) For each data set indicate whether the sample skewness is negative, approximately zero, or positive.

(e) Which data set would have a larger sample standard deviation?

(f) Determine the approximate values of $\hat{F}_1(9)$ where $\hat{F}_1$ is the empirical cumulative distribution function for Dataset 1, and $\hat{F}_2(3)$ where $\hat{F}_2$ is the empirical cumulative distribution function for Dataset 2.

9. Answer the following questions based on the side-by-side boxplots given in Figure 1.24.

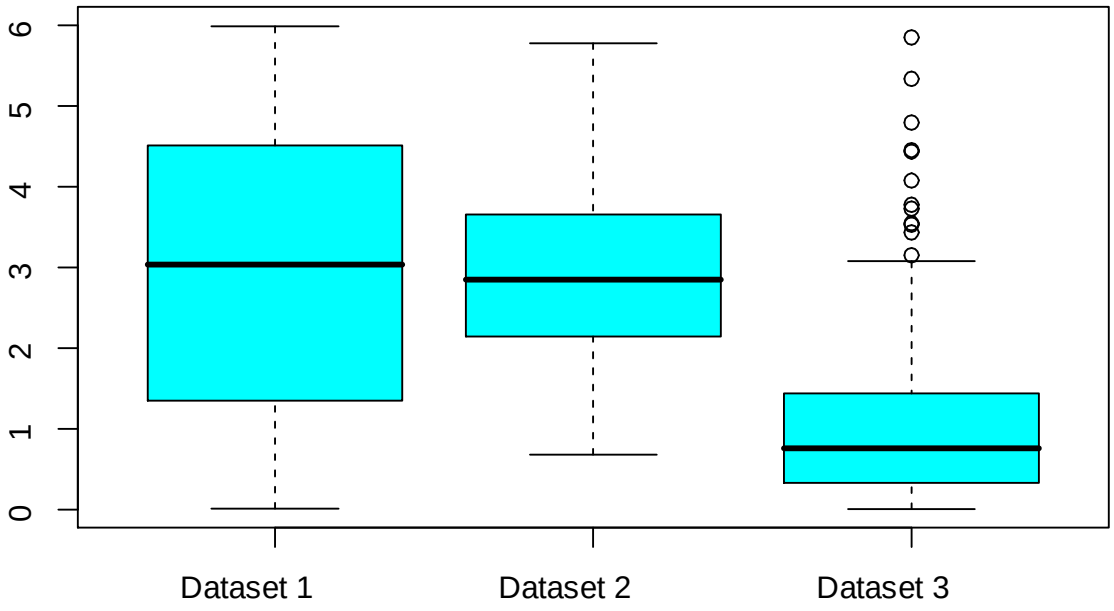


Figure 1.24: Boxplots for Problem 9

(a) For each data set determine the approximate values of: $y_{(1)}$ (the minimum observation), $y_{(n)}$ (the maximum observation), the range, $q\,(0.25)$ (the lower quartile), $q\,(0.75)$ (the upper quartile), the IQR (the interquartile range), $\hat{m}$ (the median).

(b) For each data set determine whether the corresponding relative frequency histogram would be reasonably symmetric, would not be reasonably symmetric but would have a long left tail, or would not be reasonably symmetric but would have a long right tail.

(c) For each data set indicate whether the sample skewness is negative, approximately zero, or positive.

(d) For which dataset would the corresponding relative frequency histogram look most bell-shaped? uniform?

(e) Would the sample kurtosis of Dataset 1 be larger than 3, less than 3, or approximately equal to 3?

(f) Would the sample kurtosis of Dataset 2 be larger than 3, less than 3, or approximately equal to 3?

(g) Would the sample standard deviation of Dataset 1 be larger or smaller than Dataset 2?

10. Run the following code on the can filling data and compare with the summaries given in Example 1.5.2.

```
#Import dataset canfillingdata.txt from the course website using RStudio
v1<-canfillingdata$volume[canfillingdata$machine==1]  # New Machine
v2<-canfillingdata$volume[canfillingdata$machine==2]  # Old Machine
skewness<-function(x) {(sum((x-mean(x))^3)/length(x))/
(sum((x-mean(x))^2)/length(x))^(3/2)}
kurtosis<- function(x) {(sum((x-mean(x))^4)/length(x))/
(sum((x-mean(x))^2)/length(x))^2}
# Numerical summaries by machine
summary(v1)  # Gives the 5 number summary and the sample mean
c(sd(v1),skewness(v1),kurtosis(v1))
# Note:  R defines quantiles slightly different than Def'n 3
summary(v2)
c(sd(v2),skewness(v2),kurtosis(v2))
# Plot run charts by machine, one above of the other,
# type="l" joins the points on the plots
par(mfrow=c(2,1))     # Creates 2 plotting areas, one above the other
plot(1:40,v1,xlab="Hour",ylab="Volume",main="New Machine",
ylim=c(355,360),type="l")
plot(1:40,v2,xlab="Hour",ylab="Volume",main="Old Machine",
ylim=c(355,360),type="l")
```

```
# Plot side by side relative frequency histograms with same intervals
par(mfrow=c(1,2)) # Creates 2 plotting areas side by side
# Plot relative frequency histogram for New Machine
library(MASS)     # truehist is in the MASS library
truehist(v1,h=0.5,xlim=c(355,361),xlab="Volume",ylab="Density",main="New
Machine")
# Superimpose Gaussian pdf onto histogram
curve(dnorm(x,mean(v1),sd(v1)),add=TRUE,from=355,to=359,lwd=2)
# Plot relative frequency histogram for Old Machine
truehist(v2,h=0.5,xlim=c(355,361),xlab="Volume",ylab="Density",main="Old
Machine")
# Superimpose Gaussian pdf onto histogram
curve(dnorm(x,mean(v2),sd(v2)),add=TRUE,from=355,to=361,lwd=2)
par(mfrow=c(1,1)) # Change back to one plotting area
# Plot side by side boxplots
boxplot(v1,v2,names=c("New Machine","Old Machine"))
# Plot empirical cdf's on same graph
plot(ecdf(v1),verticals=TRUE,do.points=FALSE,col="red",xlab="Volume",
ylab="e.c.d.f.",main="Empirical c.d.f.'s")
legend(356,0.8,c("New Machine (Red)","Old Machine (Blue)"))
plot(ecdf(v2),verticals=TRUE,do.points=FALSE,add=TRUE,col="blue")
```

11. The data below show the lengths in centimeters of 43 male coyotes and 40 female coyotes captured in Nova Scotia. (Based on Table 2.3.2 in Wild and Seber 1999.)

Females

71.0	73.7	80.0	81.3	83.5	84.0	84.0	84.5	85.0	85.0	86.0	86.4
86.5	86.5	88.0	87.0	88.0	88.0	88.5	89.5	90.0	90.0	90.2	91.0
91.4	91.5	91.7	92.0	93.0	93.0	93.5	93.5	93.5	96.0	97.0	97.0
97.8	98.0	101.6	102.5								

Males

78.0	80.0	80.0	81.3	83.8	84.5	85.0	86.0	86.4	86.5	87.0	88.0
88.0	88.9	88.9	90.0	90.5	91.0	91.0	91.0	91.4	92.0	92.5	93.0
93.5	95.0	95.0	95.0	94.0	95.5	96.0	96.0	96.0	96.0	97.0	98.5
100.0	100.5	101.0	101.6	103.0	104.1	105.0					

The data are available in the file *coyotedata.txt* posted on the course website. Modify the R code given in Problem 10 to answer the following questions.

(a) For each dataset determine $y_{(1)}$ (the minimum observation), $y_{(n)}$ (the maximum observation), the range, $q(0.25)$ (the lower quartile), $q(0.75)$ (the upper quartile), the IQR (the interquartile range), $\hat{m}$ (the median), $\bar{y}$ (the sample mean), s (the sample standard deviation), g_1 (the sample skewness), and g_2 (the sample kurtosis).

(b) Plot relative frequency histograms of the lengths for females and males on separate graphs. Use the same intervals so the histograms can be compared. On each relative frequency histogram overlay the graph of the $G(\bar{y}, s)$ probability density function where $\bar{y}$ and s are the sample mean and sample standard deviation respectively for that data set.

(c) Use the graphical and numerical summaries above to decide how well a Gaussian model fits each data set. Compare what you observed for your data set and what you would expect to observe if the data were generated from a Gaussian model.

(d) Plot side by side boxplots of the lengths for females and males. What similarities and differences do you notice for the female and male lengths?

(e) Plot the empirical distribution function of the lengths for females and males separately on the same graph. What similarities and differences do you notice for the female and male lengths?

12. Prove the identities

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

13. Does the value of an actor influence the amount grossed by a movie? The "value of an actor" will be measured by the average amount the actors' movies have made. The "amount grossed by a movie" is measured by taking the highest grossing movie, in which that actor played a major part. For example, Tom Hanks, whose value is 103.2 had his best results with Toy Story 3 (gross 415.0). All numbers are corrected to 2012 dollar amounts and have units "millions of U.S. dollars". Twenty actors were selected by taking the first twenty alphabetically listed by name on the website (http://boxofficemojo.com/people/). For each of the twenty actors, the value of the actor (x) and their highest grossing movie (y) were determined. The data are given below as well as in the file *actordata.txt* posted on the course website.

Actor	1	2	3	4	5	6	7	8	9	10
Value (x)	67	49.6	37.7	47.3	47.3	32.9	36.5	92.8	17.6	14.4
Gross (y)	177.2	201.6	183.4	55.1	154.7	182.8	277.5	415	90.8	83.9

Actor	11	12	13	14	15	16	17	18	19	20
Value (x)	51.1	54	30.5	42.1	23.6	62.4	32.9	26.9	43.7	50.3
Gross (y)	158.7	242.8	37.1	220	146.3	168.4	173.8	58.4	199	533

$$\begin{aligned} \sum_{i=1}^{20} x_i &= 860.6 \quad \sum_{i=1}^{20} x_i^2 = 43315.04 \\ \sum_{i=1}^{20} y_i &= 3759.5 \quad \sum_{i=1}^{20} y_i^2 = 971560.19 \\ \sum_{i=1}^{20} x_i y_i &= 184540.93 \end{aligned}$$

(a) What are the two variates in this data set? Choose one variate to be an explanatory variate and the other to be a response variate. Justify your choice.

(b) Plot a scatterplot of the data.

(c) Calculate the sample correlation for the data (x_i, y_i), $i = 1, 2, \ldots, 20$. Is there a strong positive or negative linear relationship between the two variates?

(d) Is it reasonable to conclude that the explanatory variate in this problem causes the response variate? Explain.

(e) Here is R code to plot the scatterplot (in blue) and calculate the sample correlation:

```
#Import dataset actordata.txt from the course website using RStudio
attach(actordata)
cor(Value,Gross)   # Calculates sample correlation
# plot scatterplot
plot(Value,Gross,main ="Actor Data",col="blue",pch=19,las=1)
# round correlation to 4 decimal places and convert to character
crt<-as.character(round(cor(Value,Gross),4))
txt<-paste("Sample Correlation = ",crt)  # create text
text(30,500,txt)      # add text to plot
```

14. In this course we mainly focus on methods for analyzing univariate and bivariate datasets. In the real world multivariate data sets are much more common. Learning how to analyse univariate and bivariate datasets gives us the basic tools for analyzing multivariate data sets. This problem looks at simple numerical and graphical summaries for a multivariate dataset.

Computers and smartphones are just two of the many devices that use integrated circuits. A silicon wafer is a thin slice of semiconductor material, such as a silicon crystal, used in the fabrication of integrated circuits. The thickness of such wafers is very important since thinner wafers are less costly. However the wafers cannot be too thin since then they can crack more easily.

To gain information about wafer thicknesses at a particular semiconductor fabrication plant or fab, the thickness of a single wafer, is measured at 9 locations as shown in Figure 1.25. A single wafer is removed from a tray of wafers always at the same posi-

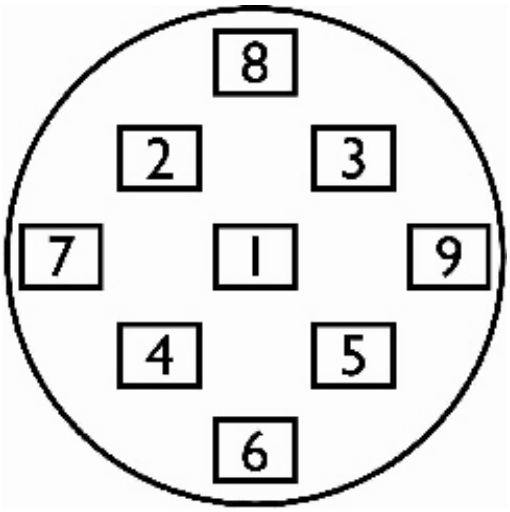


Figure 1.25: **Locations at which wafer thickness is measured**

tion for each batch of wafers. The data for 182 consecutive batches are available in the file *waferdata.txt* posted on the course website. The data have been approximately centered and scaled. These data could be used to study the relationships between the thicknesses at different locations.

Run the following R code:

```
#Import dataset waferdata.txt in folder S231Datasets using RStudio
#Install the packages moments and car if necessary
install.packages("moments")
install.packages("car")
library(moments)
library(car)
attach(waferdata)
apply(waferdata,2,mean)      # sample means for each location
apply(waferdata,2,sd)        # sample std deviations for each location
apply(waferdata,2,fivenum)   # five number summaries for each location
apply(waferdata,2,skewness)  # sample skewness for each location
```

```
apply(waferdata,2,kurtosis)  # sample kurtosis for each location
cor(waferdata)               # all sample correlations
scatterplotMatrix(waferdata[,1:5],smooth=F,regLine=F,
var.labels=colnames(waferdata[,1:5]),cex.labels=1,
diagonal=list(method="histogram"),plot.points=T)
scatterplotMatrix(waferdata[,c(1,6:9)],smooth=F,regLine=F,
var.labels=colnames(waferdata[,c(1,6:9)]),cex.labels=1,
diagonal=list(method="histogram"),plot.points=T)
```

(a) Comment on any similarities or differences in the numerical summaries for each of the 9 locations.

(b) What do you notice about the sample correlations between location 1 with locations $2, 3, 4, 5$ as compared to the sample correlations between location 1 and locations $6, 7, 8, 9$? Does what you observe make sense?

(c) Compare the variability in the points for the different scatterplots. What do you notice?

15. Two hundred volunteers participated in an experiment to examine the effectiveness of vitamin C in preventing colds. One hundred were selected at random to receive daily doses of vitamin C and the others received a placebo. (None of the volunteers knew which group they were in.) During the study period, twenty of those taking vitamin C and thirty of those receiving the placebo caught colds.

(a) Create a two-way table for these data.

(b) Calculate the relative risk of a cold in the vitamin C group as compared to the placebo group.

(c) What do these data suggest? Can you conclude that vitamin C reduces the chances of catching a cold?*****

Problems 16 to 20 are based on material covered in STAT 220/230/240. This material will be used frequently in these Course Notes. You may wish to review the relevant material from STAT 220/230/240 before attempting these problems.

16. In a very large population a proportion θ of people have blood type A. Suppose n people are selected at random. Define the random variable Y = number of people with blood type A in sample of size n.

(a) Assume $Y \sim \text{Binomial}(n, \theta)$. What are the assumptions for a Binomial model? Explain, with reasons, whether or not these assumptions might hold in this context.

(b) What is the probability function for Y? What are $E(Y)$ and $Var(Y)$?

(c) Suppose $n = 50$. What is the probability of observing 20 people with blood type A as a function of θ?

(d) If for $n = 50$ we observed $y = 20$ people with blood type A what is a reasonable estimate of θ based on this information? Estimate the probability that in a sample of $n = 10$ there will be at least one person with blood type A.

(e) More generally, suppose in a given experiment the random variable of interest Y has a Binomial(n, θ) distribution. If the experiment is conducted and y successes are observed what is a good estimate of θ based on this information?

(f) Let $Y \sim$ Binomial(n, θ). Find $E\left(\frac{Y}{n}\right)$ and $Var\left(\frac{Y}{n}\right)$. What happens to $Var\left(\frac{Y}{n}\right)$ as $n \to \infty$? What does this imply about how far $\frac{Y}{n}$ is from θ for large n? Approximate

$$P\left(\frac{Y}{n} - 1.96\sqrt{\frac{\theta(1-\theta)}{n}} \leq \theta \leq \frac{Y}{n} + 1.96\sqrt{\frac{\theta(1-\theta)}{n}}\right)$$

You may ignore the continuity correction.

(g) There are actually 4 blood types: A, B, AB, O. In a sample of size n let

Y_1 = number with type A, Y_2 = number with type B,
Y_3 = number with type AB, and Y_4 = number with type O.

Let θ_1 = proportion of type A, θ_2 = proportion of type B,
θ_3 = proportion of type AB, and θ_4 = proportion of type O in the population.

What is the joint probability function of Y_1, Y_2, Y_3, Y_4?

(h) Suppose in a sample of n people the observed data were y_1, y_2, y_3, y_4. What are reasonable estimates of $\theta_1, \theta_2, \theta_3, \theta_4$ based on these data?

17. Visits to a particular website occur at random at the average rate of θ visits per second. Suppose it is reasonable to use a Poisson process to model this process. Define the random variable Y = number of visits to the website in one second.

(a) Give the probability function for Y, $E(Y)$ and $Var(Y)$.

(b) How well do you think the Poisson process assumptions might hold in this case?

(c) Suppose over a 10 second period the number of visits to the website in each second were

1 4 5 1 0 2 5 4 3 2

(i) What is the probability of observing these data as a function of θ?

(ii) What is a reasonable estimate of θ based on these data?

(iii) Based on these data, estimate the probability that there is at least one visit to the website in a one second interval.

(d) Suppose $Y_i \sim \text{Poisson}(\theta)$, $i = 1, 2, \ldots, n$ independently.

(i) Find $E(\overline{Y})$ and $Var(\overline{Y})$. What happens to $Var(\overline{Y})$ as $n \to \infty$? What does this imply about how far $\overline{Y}$ is from θ for large n?

(ii) Approximate $P\left(\overline{Y} - 1.96\sqrt{\theta/n} \leq \theta \leq \overline{Y} + 1.96\sqrt{\theta/n}\right)$. You may ignore the continuity correction.

18. Suppose it is reasonable to model the IQ's of UWaterloo Math students using a $G(\mu, \sigma)$ distribution. Define the random variable Y = IQ of a randomly chosen UWaterloo Math student.

(a) Give the probability density function of Y, $E(Y)$ and $Var(Y)$.

(b) Show that the pth quantile of the $G(\mu, \sigma)$ distribution is $Q(p) = \mu + \sigma\Phi^{-1}(p)$ where Φ^{-1} is the inverse cumulative distribution function of a $G(0, 1)$ random variable. Thus show that the median of the $G(\mu, \sigma)$ distribution is $m = \mu$ and the IQR of the $G(\mu, \sigma)$ distribution is $\sigma\left[\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\right] = 1.34898\sigma \approx 1.35\sigma$.

(c) Suppose that the IQ's for a random sample of 16 students were:

$$127\ 108\ 127\ 136\ 125\ 130\ 127\ 117\ 123\ 112\ 129\ 109\ 109\ 112\ 91\ 134$$

$$\sum_{i=1}^{16} y_i = 1916,\ \sum_{i=1}^{16} y_i^2 = 231618$$

(i) What is a reasonable estimate of μ based on these data?

(ii) What is a reasonable estimate of σ^2 based on these data?

(iii) Based on theses data, estimate the probability that a randomly chosen UWaterloo Math student has an IQ greater than 120.

(d) Suppose $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$ independently.

(i) What is the distribution of

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Find $E(\overline{Y})$, and $Var(\overline{Y})$. What happens to $Var(\overline{Y})$ as $n \to \infty$? What does this imply about how far $\overline{Y}$ is from μ for large n?

(ii) Find $P\left(\overline{Y} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{Y} + 1.96\sigma/\sqrt{n}\right)$.

(iii) Find the smallest value of n such that $P\left(\left|\overline{Y} - \mu\right| \leq 1.0\right) \geq 0.95$ if $\sigma = 12$.

19. Suppose it is reasonable to model the battery life of a certain type of laptop using the Exponential(θ) distribution. Define the random variable Y = battery life of a randomly chosen laptop.

    (a) Give the probability density function of Y, $E(Y)$ and $Var(Y)$.

    (b) Show that the pth quantile of the Exponential(θ) distribution is $Q(p) = -\theta \log(1-p)$. Thus show that the median of the Exponential(θ) distribution is $m = \theta \log(2)$ and the IQR of the Exponential(θ) distribution is $\theta \log(3)$.

    (c) Suppose the lifetimes (in minutes) of a random sample of twenty laptop batteries were:

    $$\begin{array}{cccccccccc} 48.0 & 1047.2 & 802.3 & 165.6 & 76.7 & 64.2 & 158.6 & 338.3 & 200.6 & 362.8 \\ 119.5 & 55.9 & 411.3 & 706.9 & 16.2 & 1277.6 & 49.4 & 22.6 & 1078.4 & 440.7 \end{array}$$

    $$\sum_{i=1}^{20} y_i = 7442.8$$

    (i) What is a reasonable estimate of θ based on these data?

    (ii) Based on these data, estimate the probability that lifetime of a randomly chosen laptop is longer than 100 hours.

    (d) Suppose $Y_i \sim$ Exponential(θ), $i = 1, 2, \ldots, n$ independently.

    (i) Find $E(\overline{Y})$ and $Var(\overline{Y})$. What happens to $Var(\overline{Y})$ as $n \to \infty$? What does this imply about how far $\overline{Y}$ is from θ for large n?

    (ii) Approximate $P\left(\overline{Y} - 1.6449\theta/\sqrt{n} \leq \theta \leq \overline{Y} + 1.6449\theta/\sqrt{n}\right)$.

20. Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$, $i = 1, 2, \ldots, n$.

    (a) Find $E\left(Y_i^2\right)$.
    Hint: Rearrange the equation $Var(Y) = E\left(Y^2\right) - [E(Y)]^2$.

    (b) Find $E(\overline{Y})$, $Var(\overline{Y})$ and $E\left[(\overline{Y})^2\right]$.

    (c) Use (a) and (b) to show that $E\left(S^2\right) = \sigma^2$ where

    $$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \\ &= \frac{1}{n-1} \left[\sum_{i=1}^{n} Y_i^2 - n\left(\overline{Y}\right)^2\right] \end{aligned}$$

21. Is the graph in Figure 1.26 effective in conveying information about the snacking behaviour of students at Ridgemont High School?

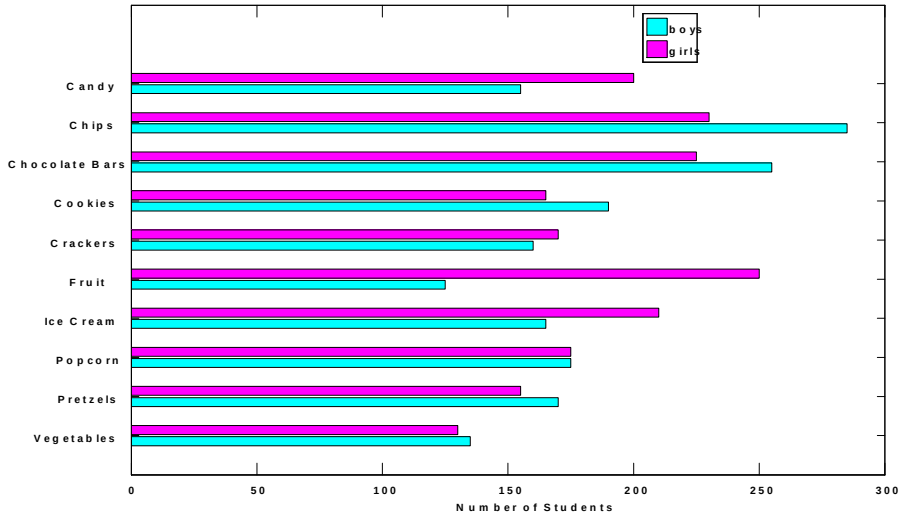


Figure 1.26: **Preferred snacks of students at Ridgemont High School**

22. The pie chart in Figure 1.27, from Fox News, shows the support for various Republican Presidential candidates in 2012. What do you notice about this pie chart?

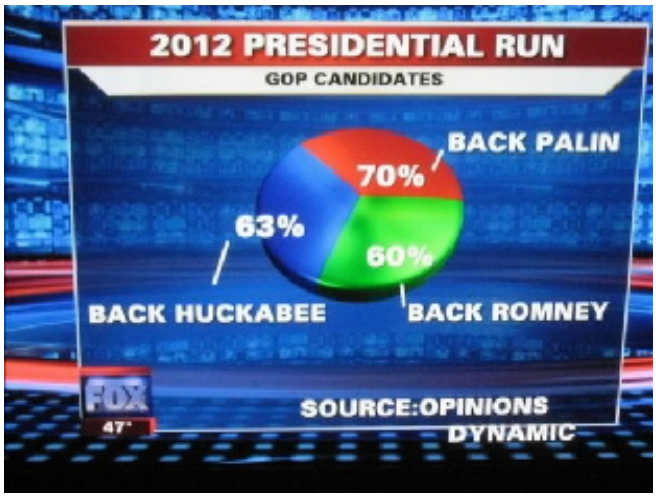


Figure 1.27: **Pie chart for support for Republican Presidental candidates**

23. The graphs in Figures 1.28 and 1.29 are two more classic Fox News graphs. What do you notice? What political message do you think they were trying to convey to their audience?

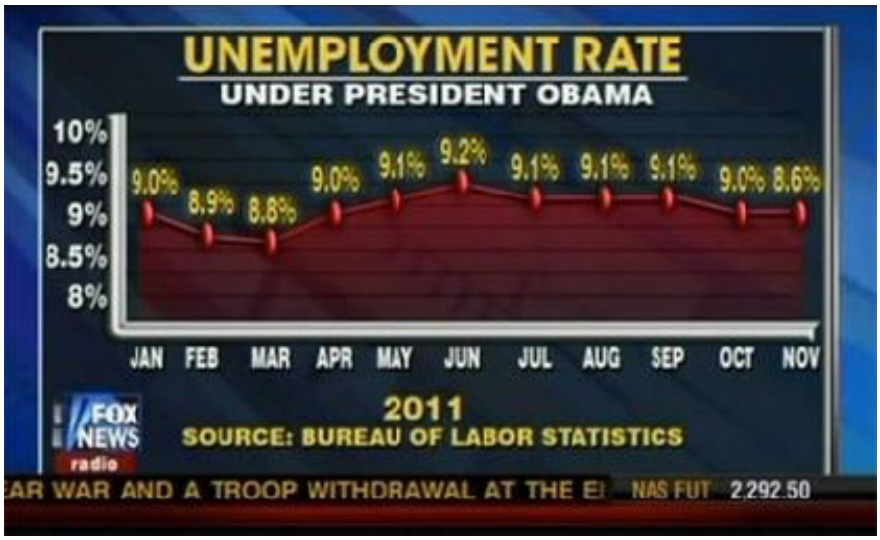


Figure 1.28: **Unemployment rate under President Obama**

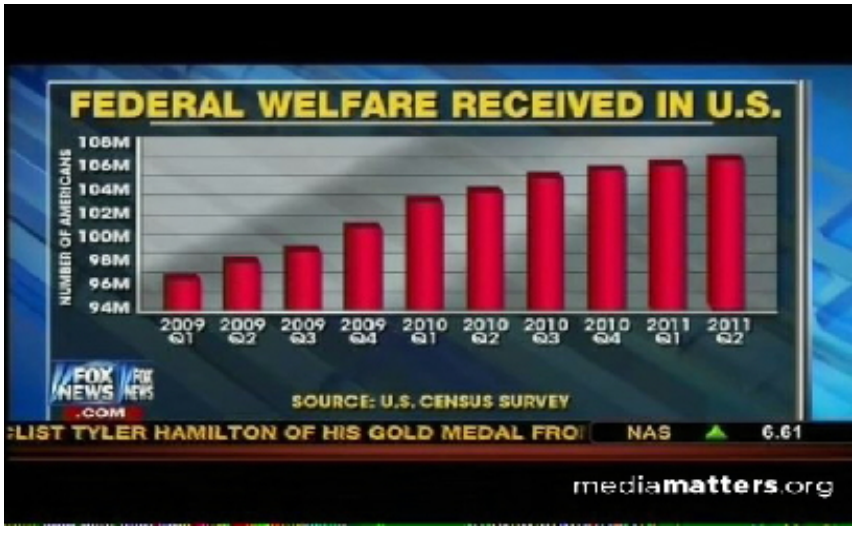


Figure 1.29: **Federal welfare in the US**

24. Information about the mortality from malignant neoplasms (cancer) for females living in Ontario is given in figures 1.30 and 1.31 for the years 1970 and 2000 respectively. The same information displayed in these two pie charts is also displayed in the bar graph in Figure 1.32. Which display seems to carry the most information?

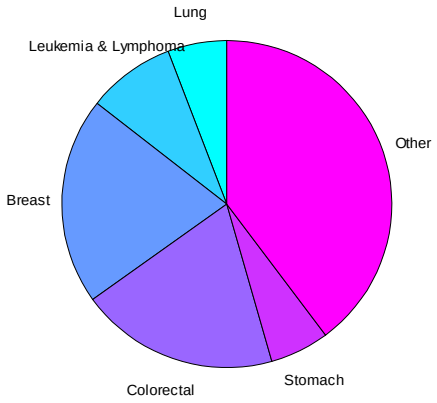


Figure 1.30: **Mortality from malignant neoplasms for females in Ontario 1970**

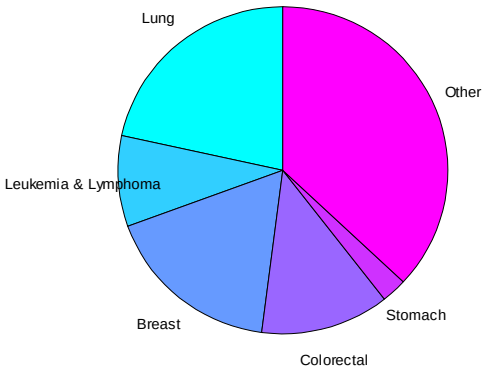


Figure 1.31: **Mortality from malignant neoplasms for females in Ontario in 2000**

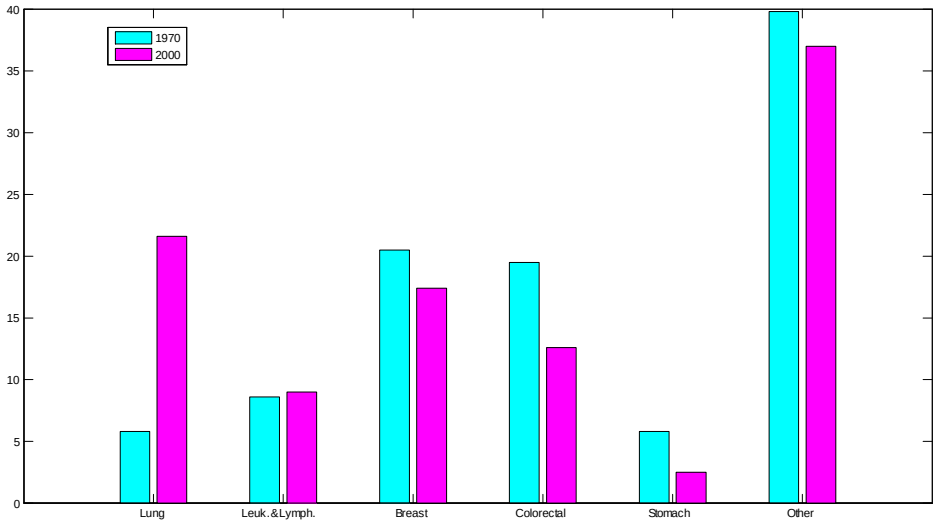


Figure 1.32: **Mortality from malignant neoplasms for females living in Ontario, 1970 and 2000**

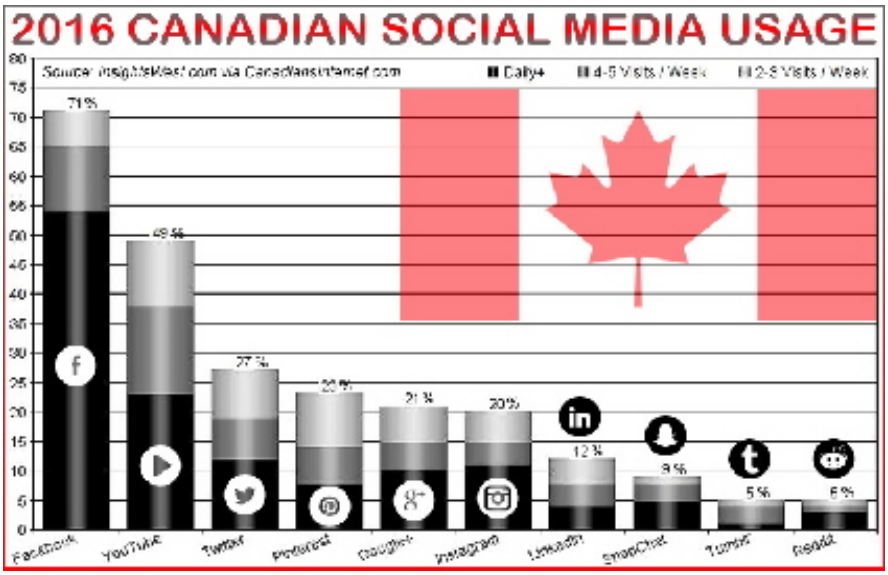
25. The following article and graphical summary appeared on canadiansinternet.com on May 24, 2016.

    Canadians love using social media in 2016 and Facebook continues to be the social network they use most, a new survey from InsightsWest has determined. Facebook, YouTube and Instagram use is still growing at a healthy pace overall in Canada, while Twitter, Google+, Pinterest, LinkedIn, Tumblr and Reddit usage has slowed down a bit. A respectable 18% of social interactions are with businesses. In spite of social media's popularity in Canada, websites are the most common way for Canadians to interact with businesses online.

    Canadian millennials use social media differently than other age groups. YouTube, Instagram, Twitter and Snapchat are growing in usage among millennials in this country. Conversely, the older our residents are, the less likely they are to have tried each social media network. In all age groups combined, only 15% said they had never tried YouTube and just 16% said they'd never used Facebook.

    Social media usage among women is growing steadily across all networks. Growth among Canadian men is slower by comparison. The ladies are using visual social networks more, with Instagram and Pinterest seeing more growth by comparison to men. LinkedIn growth among Canadian males is almost double the growth of women using the network.

    The popularity of social networks is based on more than the number of members they have. The following statistics show how many of the Canadians surveyed visit each social network at least twice per week. Comment on effective you think the graphical



summary is.

26. A study led by Beth Israel Medical Center in New York City has found that live music can be beneficial to premature babies. In the study, music therapists helped parents transform their favorite tunes into lullabies. The researchers concluded that live music, played or sung, helped to slow infants' heartbeats, calmed their breathing, improved their sucking behaviors (important for feeding), aided their sleep and promoted their states of quiet alertness. Doctors and researchers say that by reducing stress and stabilizing vital signs, music can allow infants to devote more energy to normal development.

    The two-year study was conducted between January 2011 and December 2012 in 11 hospitals in New York state. Only hospitals which received approval from their hospital's institutional review boards were included in the study. The study involved 272 premature infants aged ≤ 32 weeks with respiratory distress syndrome, clinical sepsis (a life-threatening condition that arises when the body's response to infection causes injury to its own tissues and organs), and/or SGA (small for gestational age). Over a two week period the babies experienced 4 different musical "treatments". Two of the treatments involved musical instruments, one involved singing, and the control treatment was no music at all. The instruments and singing were intended to approximate womb sounds.

    The first musical instrument was the Remo ocean disc which is a musical instrument that is round and is filled with tiny metal balls. When the disc is rotated, the metal balls move slowly to create a sound effect that is quiet and meant to simulate the fluid sounds of the womb. The second musical instrument was a gato box which is a small rectangular tuned musical instrument that is used to simulate a heartbeat sound that the baby would hear in the womb. The singing treatment consisted of live singing of a

lullaby chosen by a parent. If a parent did not chose a song then "Twinkle, Twinkle, Little Star" was used.

Each of the four treatments was given 2 times per week over the course of the two week study period. The presentation of the treatments was varied by day of the week within each week and by the time of day and randomized (either morning or afternoon) across the 2 weeks. For each treatment the baby's heart rate (beats per minute), respiratory rate (number of breaths per minute), oxygen saturation (amount of oxygen in the blood), sucking pattern (active/medium/slow/none), and activity level (active/quiet/irritable/sleeping) were recorded.

Researchers found that the gato box, the ocean disc and singing all slowed a baby's heart rate, though singing seemed to be most effective. Singing also increased the time babies stayed quietly alert. Sucking behavior improved most with the gato box. The breathing rate slowed the most and sleeping was the best with the ocean disc. Babies hearing songs their parents chose had better a better sucking pattern than those who heard "Twinkle, Twinkle, Little Star." But the "Twinkle" babies had slightly more oxygen saturation in their blood. Dr. Loewy, who trains therapists worldwide, said it did not matter whether parents or music therapists sang, or whether babies were in incubators or held.

Dr. Lance A. Parton, associate director of the regional neonatal intensive care unit at Westchester Medical Center's Maria Fareri Children's Hospital, which participated in the research, said it would be useful to see if music could help the sickest and most premature babies, who were not in the study. "Live music is optimal because it's in the moment and can adapt to changing conditions," said Dr. Standley a professor of medical music therapy at Florida State University. "If the baby appears to be falling asleep, you can sing quieter. Recorded music can't do that. But there are so many premature babies and so few trained live producers of music therapy that it's important to know what recorded music can do."

(a) Is this a sample survey, an observational study or an experimental study? Explain why.

(b) What are the units of interest in this study? Based on the given information, what population or collection of units are the researchers interested in?

(c) What variates were collected in this study? Specify the type of variate for each.

(d) What are the attributes of interest in this study?

27. Many people do not realize how important statistics is in our everyday life. We are surrounded by examples. Here is a wonderful example given by John Sall, co-founder and Executive VP of the statistical software company SAS, on the occasion of the International Year of Statistics in 2013.

"You brush your teeth. The fluoride in the toothpaste was studied by scientists using statistical methods to carefully assure the safety and effectiveness of the ingredient

and the proper concentration. The toothpaste was formulated through a series of designed experiments that determined the optimal formulation through statistical modeling. The toothpaste production was monitored by statistical process control to ensure quality and consistency, and to reduce variability.
The attributes of the product were studied in consumer trials using statistical methods. The pricing, packaging and marketing were determined through studies that used statistical methods to determine the best marketing decisions. Even the location of the toothpaste on the supermarket shelf was the result of statistically based studies. The advertising was monitored using statistical methods. Your purchase transaction became data that was analyzed statistically. The credit card used for the purchase was scrutinized by a statistical model to make sure that it wasn't fraudulent.
So statistics is important to the whole process of not just toothpaste, but every product we consume, every service we use, every activity we choose. Yet we don't need to be aware of it, since it is just an embedded part of the process. Statistics is useful everywhere you look."

Think of an example in your everyday life in which statistics played an important role.

2. STATISTICAL MODELS AND MAXIMUM LIKELIHOOD ESTIMATION

2.1 Choosing a Statistical Model

A statistical model is a mathematical model that incorporates probability[3] in some way. As described in Chapter 1, our interest here is in studying variability and uncertainty in populations and processes and drawing inferences, where warranted, in the presence of this uncertainty. This will be done by using random variables to model characteristics of randomly selected units in the population or process. It is very important to be clear about what the "target" population or process is, and exactly how the variates being considered are defined and measured. These issues are discussed in Chapter 3.

An important step in statistics is the choice of a statistical model[4] to suit a given application. The choice of a model is usually determined by some combination of the following three factors:

1. Background knowledge or assumptions about the population or process which lead to certain distributions.

2. Past experience with data sets from the population or process, which has shown that certain distributions are suitable.

3. A current data set, against which models can be assessed.

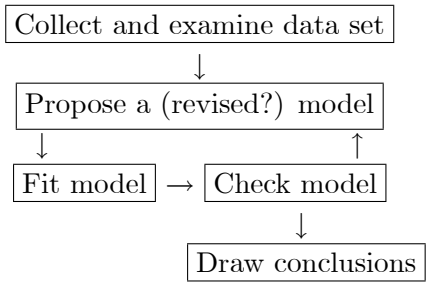
[3] The material in this section is largely a review of material you have seen in a previous probability course. This material is available in the STAT 220/230 Notes which are posted on the course website.

[4] The University of Wisconsin-Madison statistician George E.P. Box (18 October 1919 – 28 March 2013) said of statistical models that "All models are wrong but some are useful" which is to say that although models rarely fit very large amounts of data perfectly, they do assist in describing and drawing inferences from real data.

In probability, there is a large emphasis on factor 1 above, and there are many "families" of probability distributions that describe certain types of situations. For example, the Binomial distribution was derived as a model for outcomes in repeated independent trials with two possible outcomes on each trial while the Poisson distribution was derived as a model for the random occurrence of events in time or space. The Gaussian or Normal distribution, on the other hand, is often used to represent the distributions of continuous measurements such as the heights or weights of individuals. This choice is based largely on past experience that such models are suitable and on mathematical convenience.

In choosing a model we usually consider families of probability distributions. To be specific, we suppose that for a random variable Y we have a family of probability functions/probability density functions, $f(y;\theta)$ indexed by the parameter θ (which may be a vector of values). In order to apply the model to a specific problem we need to choose a value for θ. The process of selecting a value for θ based on the observed data is referred to as "estimating" the value of θ or "fitting" the model. The next section describes the method of maximum likelihood which is the most widely used method for estimating θ.

Most applications require a sequence of steps in the formulation (the word "specification" is also used) of a model. In particular, we often start with some family of models in mind, but find after examining the data set and fitting the model that it is unsuitable in certain respects. (Methods for checking the suitability of a model will be discussed in Section 2.4.) We then try other models, and perhaps look at more data, in order to work towards a satisfactory model. This is usually an iterative process, which is sometimes represented by diagrams such as:



Statistics devotes considerable effort to the steps of this process. We will focus on settings in which the models are not too complicated, so that model formulation problems are minimized. There are several distributions that you should review before continuing since they will appear in many examples. See the STAT 220/230/240 Course Notes available on the course website. You should also consult the Table of Distributions given in Chapter 10 for a condensed table of properties of these distributions including their means, variances and moment generating functions .

Table 2.1
Properties of discrete versus continuous random variables

Property	Discrete	Continuous
cumulative distribution function	$F(x) = P(X \leq x) = \sum_{t \leq x} P(X = t)$ F is a right continuous step function for all $x \in \Re$	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$ F is a continuous function for all $x \in \Re$
probability (density) function	$f(x) = P(X = x)$	$f(x) = \frac{d}{dx} F(x) \neq P(X = x) = 0$
Probability of an event	$P(X \in A) = \sum_{x \in A} P(X = x)$ $= \sum_{x \in A} f(x)$	$P(a < X \leq b) = F(b) - F(a)$ $= \int_{a}^{b} f(x)\, dx$
Total probability	$\sum_{all\ x} P(X = x) = \sum_{all\ x} f(x) = 1$	$\int_{-\infty}^{\infty} f(x)\, dx = 1$
Expectation	$E[g(X)] = \sum_{all\ x} g(x) f(x)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$

Binomial Distribution

The discrete random variable (r.v.) Y has a Binomial distribution if its probability function is of the form

$$P(Y = y; \theta) = f(y; \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \quad \text{for } y = 0, 1, \ldots, n$$

where θ is a parameter with $0 < \theta < 1$. For convenience we write $Y \sim \text{Binomial}(n, \theta)$. Recall that $E(Y) = n\theta$ and $Var(Y) = n\theta(1 - \theta)$.

Poisson distribution

The discrete random variable Y has a Poisson distribution if its probability function is of the form

$$f(y; \theta) = \frac{\theta^y e^{-\theta}}{y!} \quad \text{for } y = 0, 1, 2, \ldots$$

where θ is a parameter with $\theta \geq 0$. We write $Y \sim \text{Poisson}(\theta)$. Recall that $E(Y) = \theta$ and $Var(Y) = \theta$.

Exponential distribution

The continuous random variable Y has an Exponential distribution if its probability density function is of the form

$$f(y;\theta) = \frac{1}{\theta}e^{-y/\theta} \quad \text{for } y \geq 0$$

where θ is parameter with $\theta > 0$. We write $Y \sim \text{Exponential}(\theta)$. Recall that $E(Y) = \theta$ and $Var(Y) = \theta^2$.

Gaussian (Normal) distribution

The continuous random variable Y has a Gaussian or Normal distribution if its probability density function is of the form

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right] \quad \text{for } y \in \Re$$

where μ and σ are parameters, with $\mu \in \Re$ and $\sigma > 0$. Recall that $E(Y) = \mu$, $Var(Y) = \sigma^2$, and the standard deviation of Y is $sd(Y) = \sigma$. We write either $Y \sim G(\mu,\sigma)$ or $Y \sim N(\mu,\sigma^2)$. Note that in the former case, $G(\mu,\sigma)$, the second parameter is the standard deviation σ whereas in the latter, $N(\mu,\sigma^2)$, the second parameter is the variance σ^2. Most software syntax including R requires that you input the standard deviation for the parameter. As seen in examples in Chapter 1, the Gaussian distribution provides a suitable model for the distribution of measurements on characteristics like the height or weight of individuals in certain populations, but is also used in many other settings. It is particularly useful in finance where it is the most commonly used model for asset prices, exchange rates, interest rates, etc.

Multinomial distribution

The Multinomial distribution is a multivariate distribution in which the discrete random variable's $Y_1, Y_2, \ldots, Y_k$ $(k \geq 2)$ have the joint probability function

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k;\boldsymbol{\theta}) &= f(y_1, y_2, \ldots, y_k;\boldsymbol{\theta}) \\ &= \frac{n!}{y_1!y_2!\ldots y_k!}\theta_1^{y_1}\theta_2^{y_2\cdots}\theta_k^{y_k} \end{aligned} \tag{2.1}$$

where $y_i = 0, 1, \ldots$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} y_i = n$. The elements of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ satisfy $0 \leq \theta_i \leq 1$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} \theta_i = 1$. This distribution is a generalization of the Binomial distribution. It arises when there are repeated independent trials, where each trial has k possible outcomes (call them outcomes $1, 2, \ldots, k$), and the probability outcome i occurs is θ_i. If Y_i, $i = 1, 2, \ldots, k$ is the number of times that outcome i occurs in a sequence of n independent trials, then $(Y_1, Y_2, \ldots, Y_k)$ have the joint probability function given in (2.1). We write $(Y_1, Y_2, \ldots, Y_k) \sim \text{Multinomial}(n;\boldsymbol{\theta})$.

Since $\sum_{i=1}^{k} Y_i = n$ we can rewrite $f(y_1, y_2, \ldots, y_k; \boldsymbol{\theta})$ using only $k-1$ variables, say $y_1, y_2, \ldots, y_{k-1}$ by replacing y_k with $n - y_1 - \cdots - y_{k-1}$. We see that the Multinomial distribution with $k = 2$ is just the Binomial distribution, where the two possible outcomes are S (Success) and F (Failure).

We now turn to the problem of fitting a model. This requires estimating or assigning numerical values to the parameters in the model, for example, θ in an Exponential model or μ and σ in the Gaussian model.

2.2 Maximum Likelihood Estimation

Suppose a probability distribution that serves as a model for some random process depends on an unknown parameter θ (possibly a vector). In order to use the model we have to "estimate" or specify a value for θ. To do this we usually rely on some data set that has been collected for the random variable in question. It is important that a data set be collected carefully, and we consider this issue in Chapter 3. For example, suppose that the random variable Y represents the weight of a randomly chosen female in some population, and that we consider a Gaussian model, $Y \sim G(\mu, \sigma)$. Since $E(Y) = \mu$, we might decide to randomly select, say, 50 females from the population, measure their weights $y_1, y_2, \ldots, y_{50}$, and use the average,

$$\hat{\mu} = \overline{y} = \frac{1}{50} \sum_{i=1}^{50} y_i \tag{2.2}$$

to estimate μ. This seems sensible (why?) and similar ideas can be developed for other parameters; in particular, note that σ must also be estimated, and you might think about how you could use $y_1, y_2, \ldots, y_{50}$ to do this. (**Hint:** what does σ or σ^2 represent in the Gaussian model?) Note that although we are estimating the parameter μ we did not write $\mu = \bar{y}$. We introduced a special notation $\hat{\mu}$. This serves a dual purpose, both to remind you that $\bar{y}$ is not exactly equal to the unknown value of the parameter μ, but also to indicate that $\hat{\mu}$ is a quantity derived from the data y_i, $i = 1, 2, \ldots, 50$ and *depends on the sample.* A different draw of the sample y_i, $i = 1, 2, \ldots, 50$ will result in a different value for $\hat{\mu}$.

Definition 10 *A point estimate of a parameter is the value of a function of the observed data $y_1, y_2, \ldots, y_n$ and other known quantities such as the sample size n. We use $\hat{\theta}$ to denote an estimate of the parameter θ.*

Note that $\hat{\theta} = \hat{\theta}(y_1, y_2, \ldots, y_n) = \hat{\theta}(\mathbf{y})$ depends on the sample $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ drawn. A function of the data which does not involve any unknown quantities such as unknown parameters is called a statistic. The numerical summaries discussed in Chapter 1 are all examples of statistics. A point estimate is also a statistic.

Instead of ad hoc approaches to estimation as in (2.2), it is desirable to have a general method for estimating parameters. The method of *maximum likelihood* is a very general method, which we now describe.

Let the discrete (vector) random variable $\mathbf{Y}$ represent potential data that will be used to estimate θ, and let $\mathbf{y}$ represent the actual observed data that are obtained in a specific application. Note that to apply the method of maximum likelihood, we must know (or make assumptions about) how the data $\mathbf{y}$ were collected. It is usually assumed here that the data set consists of measurements on a random sample of units from a population or process.

Definition 11 *The* likelihood function *for θ is defined as*

$$L(\theta) = L(\theta; \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}; \theta) \quad \text{for } \theta \in \Omega$$

where the parameter space Ω *is the set of possible values for θ.*

Note that the likelihood function is a function of the parameter θ and the given data $\mathbf{y}$. For convenience we usually write just $L(\theta)$. Also, *the likelihood function is the probability that we observe the data* $\mathbf{y}$, *considered as a function of the parameter θ.* Obviously values of the parameter that make the observed data $\mathbf{y}$ more probable would seem more credible or likely than those that make the data less probable. Therefore values of θ for which $L(\theta)$ is large are more consistent with the observed data $\mathbf{y}$. This seems like a "sensible" approach, and it turns out to have very good properties.

Definition 12 *The value of θ which maximizes $L(\theta)$ for given data $\mathbf{y}$ is called the* maximum likelihood estimate [5] *(m.l. estimate) of θ. It is the value of θ which maximizes the probability of observing the data $\mathbf{y}$. This value is denoted $\hat{\theta}$.*

We are surrounded by polls. They guide the policies of political leaders, the products that are developed by manufacturers, and increasingly the content of the media. The following is an example of a public opinion poll.

Example 2.2.1 Nanos Research poll

Between February 22nd and 24th, 2016 Nanos Research (a Canadian public opinion and research company) conducted a survey of Canadian adults, 18 years or older, to determine support for the legalization of marijuana. The 1000 participants were recruited using live agents and random digit dialing (land- and cell-lines) across Canada with a maximum of five call backs. Respondents were asked "Do you support, somewhat support, somewhat oppose or oppose legalizing the recreational use of marijuana?". The data are summarized in a bar graph in Figure 2.1. Thirty-nine percent of respondents indicated that they supported the recreational use of marijuana while 29% indicated that they somewhat supported the recreational use of marijuana. Nanos reported that the margin of error for a random survey

[5] We distinguish between the random variable, the *maximum likelihood estimator*, which is the function of the potential data, and its numerical value for the given data, referred to as the *maximum likelihood estimate*.

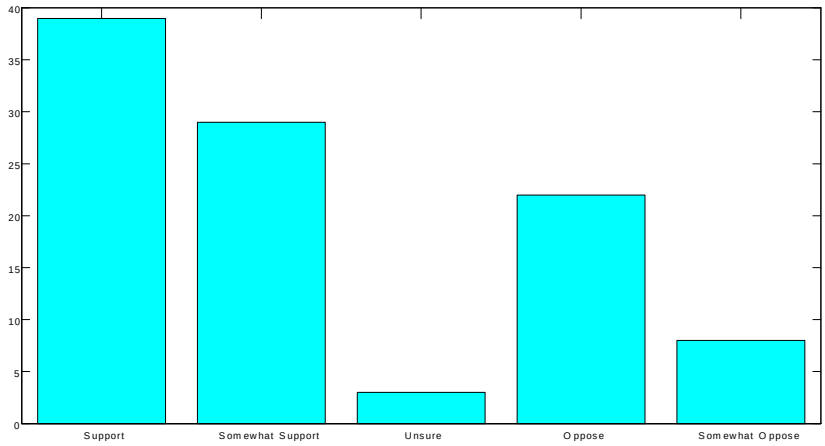


Figure 2.1: **Nanos Research poll on use of marijuana**

of 1000 Canadians is ±3.1 percentage points, 19 times out of 20. How do we interpret this statement?

Suppose that the random variable Y represents the number of units in a sample of n units drawn from a very large population who have a certain characteristic of interest. Suppose we assume that Y is closely modelled by a Binomial distribution with probability function

$$P(Y = y; \theta) = f(y; \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } y = 0, 1, \ldots, n \text{ and } 0 \leq \theta \leq 1$$

where θ represents the proportion of the large population that have the characteristic. Suppose that y units in the sample of size n have the characteristic. The likelihood function for θ based on these data is

$$\begin{aligned} L(\theta) &= P\left(y \text{ units have characteristic } ; \theta\right) \\ &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } 0 \leq \theta \leq 1 \end{aligned} \tag{2.3}$$

If $y \neq 0$ and $y \neq n$ then it can be shown that (2.3) attains its maximum value at $\theta = \hat{\theta} = y/n$ by solving $\frac{dL(\theta)}{d\theta} = 0$. The estimate $\hat{\theta} = y/n$ is called the sample proportion.

For the Nanos poll, suppose we are interested in $\theta =$ proportion of Canadian adults who support or somewhat support the recreational use of marijuana. From this poll we have $y = 680$ people out of $n = 1000$ people support or somewhat support the recreational use of marijuana so the likelihood function for θ is

$$L\left(\theta\right) = \binom{1000}{680} \theta^{680} (1-\theta)^{320} \quad \text{for } 0 \leq \theta \leq 1 \tag{2.4}$$

The maximum likelihood estimate of θ for theses data is $\hat{\theta} = y/n = 680/1000 = 0.68$ or 68% which can also easily be seen from the graph of the likelihood function (2.4) given in Figure 2.2. The interval suggested by the pollsters was $68 \pm 3.1\%$ or $[64.9, 71.1]$. Looking at Figure 2.2 we see that the interval $[0.649, 0.711]$ is a reasonable interval for the parameter θ since it seems to contain most of the values of θ with large values of the likelihood $L(\theta)$. We will return to the construction of such interval estimates in Chapter 4.

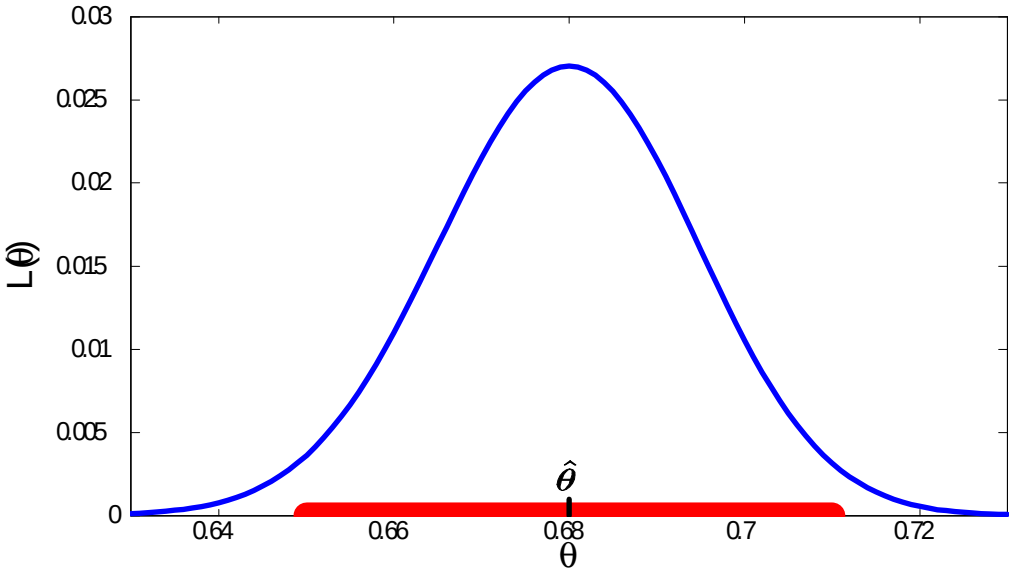


Figure 2.2: **Likelihood function for θ for the Nanos poll**

The shape of the likelihood function and the value of θ at which it is maximized are not affected if $L(\theta)$ is multiplied by a constant. Indeed it is not the absolute value of the likelihood function that is important but the relative values at two different values of the parameter, e.g. $L(\theta_1)/L(\theta_2)$. You might think of this ratio as how much more or less consistent the data are with the parameter θ_1 versus θ_2. The ratio $L(\theta_1)/L(\theta_2)$ is also unaffected if $L(\theta)$ is multiplied by a constant. In view of this the likelihood may be defined as $P(\mathbf{Y} = \mathbf{y}; \theta)$ or as any constant multiple of it, so, for example, we could drop the term $\binom{n}{y}$ in (2.3) and define $L(\theta) = \theta^y(1-\theta)^{n-y}$. This function and (2.3) are maximized by the same value $\theta = \hat{\theta} = y/n$ and have the same shape. Indeed we might rescale the likelihood function by dividing through by its maximum value $L(\hat{\theta})$ so that the new function has a maximum value equal to one.

Definition 13 *The **relative likelihood function** is defined as*

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad \text{for } \theta \in \Omega$$

Note that $0 \leq R(\theta) \leq 1$ *for all* $\theta \in \Omega$.

Sometimes it is easier to work with the *log* ($log = ln$) of the likelihood function.

Definition 14 *The **log likelihood function** is defined as*

$$l(\theta) = \ln L(\theta) = \log L(\theta) \quad \textit{for } \theta \in \Omega$$

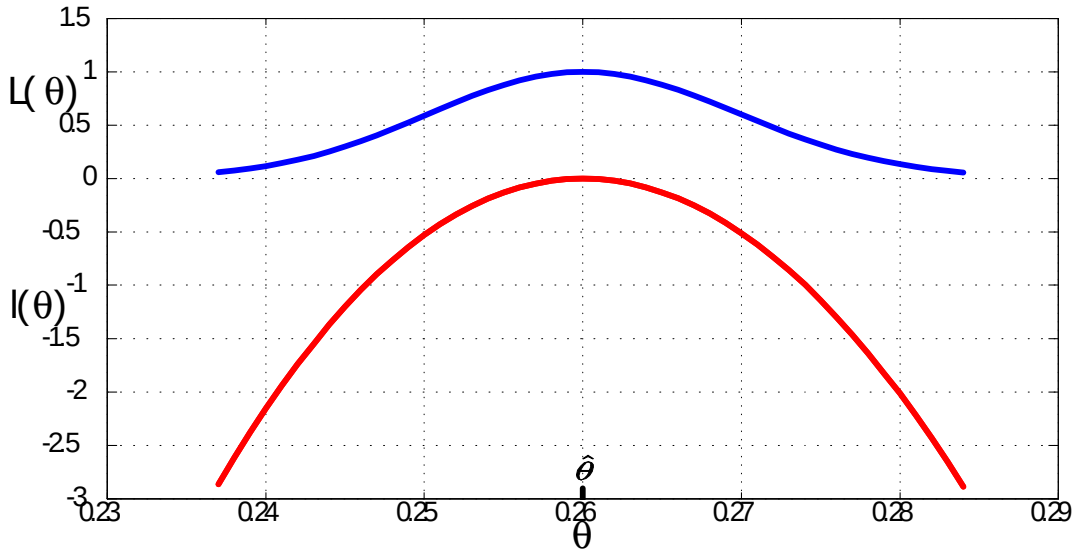


Figure 2.3: **The functions $L(\theta)$ (upper graph) and $l(\theta)$ (lower graph) are both maximized at the same value $\theta = \hat{\theta}$**

Figure 2.3 displays the graph of a likelihood function $L(\theta)$, rescaled to have a maximum value of one at $\theta = \hat{\theta}$, and the corresponding log likelihood function $l(\theta) = \log L(\theta)$ with a maximum value of $\log(1) = 0$. We see that $l(\theta)$, the lower of the two curves, is a monotone function of $L(\theta)$ so that the two functions increase together and decrease together. Both functions have a maximum at the same value $\theta = \hat{\theta}$.

Because functions are often (but not always!) maximized by setting their derivatives equal to zero, we can usually obtain $\hat{\theta}$ by solving the equation

$$\frac{d}{d\theta} l(\theta) = 0$$

For example, from $L(\theta) = \theta^y (1-\theta)^{n-y}$ we get $l(\theta) = y \log(\theta) + (n-y)\log(1-\theta)$ and

$$\frac{d}{d\theta} l(\theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta} = \frac{y - n\theta}{\theta(1-\theta)} \quad \text{for } 0 < \theta < 1$$

Solving $dl/d\theta = 0$ gives $\theta = y/n$. The First Derivative Test can be used to verify that this corresponds to a maximum value so the maximum likelihood estimate of θ is $\hat{\theta} = y/n$. This derivation holds if $y \neq 0$ and $y \neq n$. See Problem 2 for the derivation of $\hat{\theta}$ if $y = 0$ or $y = n$.

Likelihood function for a random sample

In many applications the data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ are *independent and identically distributed* (i.i.d.) random variables each with probability function $f(y;\theta)$, $\theta \in \Omega$. We refer to $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ as a random sample from the distribution $f(y;\theta)$. In this case the observed data are $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and

$$L(\theta) = L(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\theta) \quad \text{for } \theta \in \Omega$$

Recall that if $Y_1, Y_2, \ldots, Y_n$ are independent random variables then their joint probability function is the product of their individual probability functions.

Example 2.2.2 Likelihood function for Poisson distribution

Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from a Poisson(θ) distribution. The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n} P(Y_i = y_i;\theta) \quad \text{for } \theta \in \Omega \\ &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \left(\prod_{i=1}^{n} \frac{1}{y_i!}\right) \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \quad \text{for } \theta \geq 0 \end{aligned}$$

or more simply

$$L(\theta) = \theta^{n\bar{y}} e^{-n\theta} \quad \text{for } \theta \geq 0$$

The log likelihood is

$$l(\theta) = n(\bar{y}\log\theta - \theta) \quad \text{for } \theta > 0$$

with derivative

$$\frac{d}{d\theta} l(\theta) = n\left(\frac{\bar{y}}{\theta} - 1\right) = \frac{n}{\theta}(\bar{y} - \theta) \quad \text{for } \theta > 0$$

The First Derivative Test can be used to verify that the value $\theta = \bar{y}$ maximizes $l(\theta)$ and so $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate of θ.

Combining likelihoods based on independent experiments

If we have two data sets $\mathbf{y}_1$ and $\mathbf{y}_2$ from two independent studies for estimating θ, then since the corresponding random variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent we have

$$P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2;\theta) = P(\mathbf{Y}_1 = \mathbf{y}_1;\theta) P(\mathbf{Y}_2 = \mathbf{y}_2;\theta)$$

and we obtain the "combined" likelihood function $L(\theta)$ based on $\mathbf{y}_1$ and $\mathbf{y}_2$ together as

$$L(\theta) = L_1(\theta) \times L_2(\theta) \quad \text{for } \theta \in \Omega$$

where $L_j(\theta) = P(\mathbf{Y}_j = \mathbf{y}_j;\theta)$, $j = 1, 2$. This idea, of course, can be extended to more than two independent studies.

Example 2.2.3

In 2011, Harris/Decima (a research polling company) conducted a poll of the Canadian adult population in which they asked 2000 respondents whether they agreed with the statement: "University and college teachers earn too much". In 2011, $y_2 = 540$ people agreed with the statement. In a previous poll of 1500 people conducted by Harris/Decima in 2010, $y_1 = 390$ people agreed with the same statement. If we assume that $\theta =$ the proportion of the Canadian adult population that agree with the statement is the same in both years then θ may be estimated using the data from these two independent polls. The combined likelihood function would be

$$\begin{aligned} L(\theta) &= P(Y_1 = y_1, Y_2 = y_2; \theta) \\ &= P(Y_1 = y_1; \theta) P(Y_2 = y_2; \theta) \\ &= \binom{1500}{390} \theta^{390} (1-\theta)^{1110} \binom{2000}{540} \theta^{540} (1-\theta)^{1460} \\ &= \binom{1500}{390} \binom{2000}{540} \theta^{930} (1-\theta)^{2570} \quad \text{for } 0 \le \theta \le 1 \end{aligned}$$

or, ignoring the constants with respect to θ, we have

$$L(\theta) = \theta^{930} (1-\theta)^{2570} \quad \text{for } 0 \le \theta \le 1$$

The maximum likelihood estimate of θ based on the two independent experiments is $\hat{\theta} = 930/3500 = 0.2657$.

Sometimes the likelihood function for a given set of data can be constructed in more than one way as the following example illustrates.

Example 2.2.4

Suppose that the random variable Y represents the number of persons infected with the human immunodeficiency virus (HIV) in a randomly selected group of n persons. We assume the data are reasonably modeled by $Y \sim \text{Binomial}(n, \theta)$ with probability function

$$P(Y = y; \theta) = f(y; \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } y = 0, 1, \ldots, n$$

where θ represents the proportion of the population that are infected. In this case, if we select a random sample of n persons and test them for HIV, we have $\mathbf{Y} = Y$, and $\mathbf{y} = y$ as the observed number infected. Thus

$$L(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad \text{for } 0 \le \theta \le 1$$

or more simply

$$L(\theta) = \theta^y (1-\theta)^{n-y} \quad \text{for } 0 \le \theta \le 1 \tag{2.5}$$

and again $L(\theta)$ is maximized by the value $\hat{\theta} = y/n$.

For this random sample of n persons who are tested for HIV, we could also define the indicator random variable

$$Y_i = I\,(\text{person } i \text{ tests positive for HIV})$$

for $i = 1, 2, \ldots, n$. (Note: $I(A)$ is the indicator function; it equals 1 if A is true and 0 if A is false.) Now $Y_i \sim \text{Binomial}(1; \theta)$ with probability function

$$f(y_i; \theta) = \theta^{y_i}(1-\theta)^{1-y_i} \quad \text{for } y_i = 0, 1 \text{ and } 0 \leq \theta \leq 1$$

The likelihood function for the observed random sample $y_1, y_2, \ldots, y_n$ is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i; \theta) \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{\sum_{i=1}^{n}(1-y_i)} \\ &= \theta^{y}(1-\theta)^{n-y} \quad \text{for } 0 \leq \theta \leq 1 \end{aligned}$$

where $y = \sum_{i=1}^{n} y_i$. This is the same likelihood function as (2.5). The reason for this is because the random variable $\sum_{i=1}^{n} Y_i$ has a Binomial(n, θ) distribution.

In many applications we encounter likelihood functions which cannot be maximized mathematically and we need to resort to numerical methods. The following example provides an illustration.

Example 2.2.5 Coliform bacteria in water

The number of coliform bacteria Y in a random sample of water of volume v milliliters is assumed to have a Poisson distribution:

$$P(Y = y; \theta) = f(y; \theta) = \frac{(\theta v)^y}{y!} e^{-\theta v} \quad \text{for } y = 0, 1, \ldots; \quad \theta \geq 0 \tag{2.6}$$

where θ is the average number of bacteria per milliliter of water.

There is an inexpensive test which can detect the presence (but not the number) of bacteria in a water sample. In this case we do not observe Y, but rather the "presence" indicator $I(Y > 0)$, or

$$Z = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{if } Y = 0 \end{cases}$$

From (2.6) we have

$$P(Z = 0; \theta) = P\,(Y = 0; \theta) = e^{-\theta v}$$

and so

$$P(Z = 1; \theta) = 1 - P(Z = 0; \theta) = 1 - e^{-\theta v}$$

Suppose that n water samples, of volumes $v_1, v_2, \ldots, v_n$, are selected. Let $z_1, z_2, \ldots, z_n$ be the observed values of the presence indicators. Note that these observed values will be either 0 or 1.

The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(Z_i = z_i; \theta) \\ &= \prod_{i=1}^{n} (1 - e^{-\theta v_i})^{z_i} (e^{-\theta v_i})^{1-z_i} \quad \text{for } \theta \geq 0 \end{aligned}$$

and the log likelihood function is

$$l(\theta) = \sum_{i=1}^{n} [z_i \log(1 - e^{-\theta v_i}) - (1 - z_i)\theta v_i] \quad \text{for } \theta > 0$$

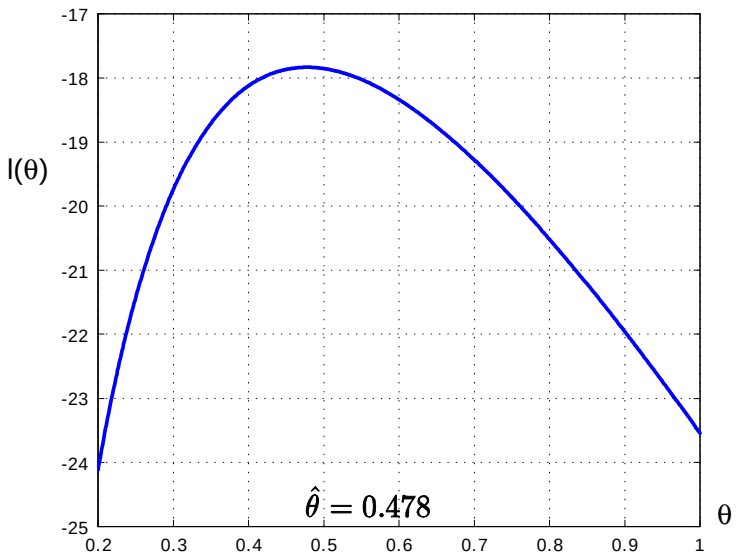


Figure 2.4: **The log likelihood function $l(\theta)$ for Example 2.2.5**

We cannot maximize $l(\theta)$ mathematically by solving $dl/d\theta = 0$, so we use *numerical methods*.

Suppose, for example, that $n = 40$ samples gave data as follows:

v_i (ml)	8	4	2	1
number of samples	10	10	10	10
number with $z_i = 1$	10	8	7	3

This gives

$$l(\theta) = 10\log(1 - e^{-8\theta}) + 8\log(1 - e^{-4\theta}) + 7\log(1 - e^{-2\theta}) + 3\log(1 - e^{-\theta}) - 21\theta \quad \text{for } \theta \geq 0$$

Either by maximizing $l(\theta)$ numerically for $\theta \geq 0$, or by solving $dl/d\theta = 0$ numerically, we find the maximum likelihood estimate of θ to be $\hat{\theta} = 0.478$. A simple way to maximize $l(\theta)$ is to plot it, as shown in Figure 2.4; the maximum likelihood estimate can then be found by inspection or, for more accuracy, by using a method like Newton's method.

A few remarks about numerical methods are in order. Aside from a few simple models, it is not possible to maximize likelihood functions explicitly. However, software exists which implements powerful numerical methods which can easily maximize (or minimize) functions of one or more variables. Multi-purpose optimizers can be found in many software packages; in R the function `nlm()` is powerful and easy to use. In addition, statistical software packages contain special functions for fitting and analyzing a large number of statistical models. The R package `MASS` (which can be accessed by the command `library(MASS)`) has a function `fitdistr` that will fit many common models.

2.3 Likelihood Functions for Continuous Distributions

Recall that we defined likelihoods for discrete random variables as the probability of observing the data $\mathbf{y}$ or

$$L(\theta) = L(\theta; \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}; \theta) \quad \text{for } \theta \in \Omega$$

For a continuous continuous random variable, $P(\mathbf{Y} = \mathbf{y}; \theta)$ is unsuitable as a definition of the likelihood since this probability always equals zero.

Suppose Y is a continuous random variable with probability density function $f(y; \theta)$. For continuous data we usually observe only the value of Y rounded to some degree of precision, for example, data on waiting times is rounded to the closest second or data on heights is rounded to the closest centimeter. The actual observation is really a discrete random variable. For example, suppose we observe Y correct to one decimal place. Then

$$P(\text{we observe } 1.1\ ; \theta) = \int_{1.05}^{1.15} f(y; \theta)dy \approx (0.1)f(1.1; \theta)$$

assuming the function $f(y;\theta)$ is reasonably smooth over the interval. More generally, suppose $y_1, y_2, \ldots, y_n$ are the observations from a random sample from the distribution with probability density function $f(y;\theta)$ which have been rounded to the nearest Δ which is assumed to be small. Then

$$P(\mathbf{Y} = \mathbf{y};\theta) \approx \prod_{i=1}^{n} \Delta f(y_i;\theta) = \Delta^n \prod_{i=1}^{n} f(y_i;\theta)$$

If we assume that the precision Δ does not depend on the unknown parameter θ, then the term Δ^n can be ignored. This argument leads us to adopt the following definition of the likelihood function for a random sample from a continuous distribution.

Definition 15 *If $y_1, y_2, \ldots, y_n$ are the observed values of a random sample from a distribution with probability density function $f(y;\theta)$, then the likelihood function is defined as*

$$L(\theta) = L(\theta;\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\theta) \quad \textit{for } \theta \in \Omega$$

Example 2.3.1 Likelihood function for Exponential distribution

Suppose that the random variable Y represents the lifetime of a randomly selected light bulb in a large population of bulbs, and that $Y \sim \text{Exponential}(\theta)$ is a reasonable model for such a lifetime.

If a random sample of light bulbs is tested and the lifetimes $y_1, y_2, \ldots, y_n$ are observed, then the likelihood function for θ is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} \\ &= \frac{1}{\theta^n} \exp\left(-\sum_{i=1}^{n} y_i/\theta\right) \\ &= \theta^{-n} e^{-n\bar{y}/\theta} \quad \text{for } \theta > 0 \end{aligned}$$

The log likelihood function is

$$l(\theta) = -n\left(\log\theta + \frac{\bar{y}}{\theta}\right) \quad \text{for } \theta > 0$$

with derivative

$$\begin{aligned} \frac{d}{d\theta} l(\theta) &= -n\left(\frac{1}{\theta} - \frac{\bar{y}}{\theta^2}\right) \\ &= \frac{n}{\theta^2}(\bar{y} - \theta) \end{aligned}$$

Now $\frac{d}{d\theta} l(\theta) = 0$ for $\theta = \bar{y}$. The First Derivative Test can be used to verify that the value $\theta = \bar{y}$ maximizes $l(\theta)$ and so $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate of θ.

Table 2.2
Summary of Maximum Likelihood Method for Named Distributions

Named Distribution	Observed Data	Maximum Likelihood Estimate	Maximum Likelihood Estimator	Relative Likelihood Function
Binomial(n, θ)	y	$\hat{\theta} = \frac{y}{n}$	$\tilde{\theta} = \frac{Y}{n}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{y} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y}$ $0 < \theta < 1$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \bar{y}$	$\tilde{\theta} = \overline{Y}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{n\hat{\theta}} e^{n\left(\hat{\theta}-\theta\right)}$ $\theta > 0$
Geometric(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \frac{1}{1+\bar{y}}$	$\tilde{\theta} = \frac{1}{1+\overline{Y}}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{n} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$ $0 < \theta < 1$
Negative Binomial(k, θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \frac{k}{k+\bar{y}}$	$\tilde{\theta} = \frac{k}{k+\overline{Y}}$	$R(\theta) = \left(\frac{\theta}{\hat{\theta}}\right)^{nk} \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n\bar{y}}$ $0 < \theta < 1$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\hat{\theta} = \bar{y}$	$\tilde{\theta} = \overline{Y}$	$R(\theta) = \left(\frac{\hat{\theta}}{\theta}\right)^{n} e^{n\left(1-\hat{\theta}/\theta\right)}$ $\theta > 0$

Example 2.3.2 Likelihood function for Gaussian distribution

As an example involving more than one parameter, suppose that $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(\mu, \sigma)$ distribution The likelihood function for $\boldsymbol{\theta} = (\mu, \sigma)$ is

$$\begin{aligned} L(\boldsymbol{\theta}) = L(\mu, \sigma) &= \prod_{i=1}^{n} f(y_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\ &= (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \quad \text{for } \mu \in \Re \text{ and } \sigma > 0 \end{aligned}$$

or more simply (ignoring constants with respect to μ and σ)

$$L(\boldsymbol{\theta}) = L(\mu, \sigma) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \quad \text{for } \mu \in \Re \text{ and } \ \sigma > 0$$

Since

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\bar{y} = \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$$

and

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \mu)^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \mu)\sum_{i=1}^{n}(y_i - \bar{y}) + \sum_{i=1}^{n}(\bar{y} - \mu)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \end{aligned}$$

we can write the likelihood function as

$$L(\mu, \sigma) = \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2\right\} \exp\left[-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right]$$

The log likelihood function for $\boldsymbol{\theta} = (\mu, \sigma)$ is

$$l(\boldsymbol{\theta}) = l(\mu, \sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{n(\bar{y} - \mu)^2}{2\sigma^2} \quad \text{for } \mu \in \Re \text{ and } \sigma > 0$$

To maximize $l(\mu, \sigma)$ with respect to both parameters μ and σ we solve [6] the two equations[7]

$$\frac{\partial l}{\partial \mu} = \frac{n}{\sigma^2}(\bar{y} - \mu) = 0 \ \text{ and } \ \frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0$$

simultaneously. We find that the maximum likelihood estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})$, where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \quad \text{and} \quad \hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}$$

[6] To maximize a function of two variables, set the derivative with respect to each variable equal to zero. Of course finding values at which the derivatives are zero does not prove this is a maximum. Showing it is a maximum is another exercise in calculus.

[7] $\frac{\partial}{\partial \mu}$ is the derivative with respect to μ holding the parameter σ constant. Similarly $\frac{\partial}{\partial \sigma}$ is the derivative with respect to σ holding μ constant.

2.4 Likelihood Functions For Multinomial Models

Multinomial models are used in many statistical applications. From Section 2.1, the Multinomial joint probability function is

$$f(y_1, y_2, \ldots, y_k; \boldsymbol{\theta}) = \frac{n!}{y_1! y_2! \cdots y_k!} \prod_{i=1}^{k} \theta_i^{y_i} \quad \text{for } y_i = 0, 1, \ldots \text{ where } \sum_{i=1}^{k} y_i = n$$

The likelihood function for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ based on data $y_1, y_2, \ldots, y_k$ is given by

$$L(\boldsymbol{\theta}) = L(\theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \prod_{i=1}^{k} \theta_i^{y_i}$$

or more simply

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{k} \theta_i^{y_i}$$

The log likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} y_i \log \theta_i$$

If y_i represents the number of times outcome i occurred in n "trials", $i = 1, 2, \ldots, k$, then it can be shown that

$$\hat{\theta}_i = \frac{y_i}{n} \quad \text{for } i = 1, 2, \ldots, k$$

are the maximum likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_k$.[8]

Example 2.4.1 A, B, AB, O blood types

Each person is one of four blood types, labelled A, B, AB and O. (Which type a person is has important consequences, for example in determining to whom they can donate a blood transfusion.) Let θ_1, θ_2, θ_3, θ_4 be the fraction of a population that has types A, B, AB, O, respectively. Now suppose that in a random sample of 400 persons whose blood was tested, the numbers who were types A, B, AB, O, were $y_1 = 172$, $y_2 = 38$, $y_3 = 14$ and $y_4 = 176$ respectively. (Note that $y_1 + y_2 + y_3 + y_4 = 400$.) Let the random variables Y_1, Y_2, Y_3, Y_4 represent the number of type A, B, AB, O persons respectively that are in a random sample of size $n = 400$. Then Y_1, Y_2, Y_3, Y_4 follow a Multinomial$(400; \theta_1, \theta_2, \theta_3, \theta_4)$.

The maximum likelihood estimates from the observed data are therefore

$$\hat{\theta}_1 = \frac{172}{400} = 0.43, \quad \hat{\theta}_2 = \frac{38}{400} = 0.095, \quad \hat{\theta}_3 = \frac{14}{400} = 0.035, \quad \hat{\theta}_4 = \frac{176}{400} = 0.44$$

(as a check, note that $\sum_{i=1}^{4} \hat{\theta}_i = 1$). These give estimates of the population fractions θ_1, θ_2, θ_3, θ_4. (Note: studies involving much larger numbers of people put the values of the θ_i's for Caucasians at close to $\theta_1 = 0.448$, $\theta_2 = 0.083$, $\theta_3 = 0.034$, $\theta_4 = 0.436$.)

[8] $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{k} y_i \log \theta_i$ is a little tricky to maximize because the θ_i's satisfy a linear constraint, $\sum_{i=1}^{k} \theta_i = 1$. The Lagrange multiplier method (Multivariate Calculus) for constrained optimization allows us to find the solution $\hat{\theta}_i = y_i/n$, $i = 1, 2, \ldots, k$.

In some problems the Multinomial parameters $\theta_1, \theta_2, \ldots, \theta_k$ may be functions of fewer than $k-1$ parameters. The following is an example.

Example 2.4.2 MM, MN, NN blood types

Another way of classifying a person's blood is through their "M-N" type. Each person is one of three types, labelled MM, MN and NN and we can let θ_1, θ_2, θ_3 be the fraction of the population that is each of the three types. In a sample of size n we let Y_1 = number of MM types observed, Y_2 = number of MN types observed and Y_3 = number of NN types observed. The joint probability function of Y_1, Y_2, Y_3 is

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \theta_1, \theta_2, \theta_3) = \frac{n!}{y_1!y_2!y_3!}\theta_1^{y_1}\theta_2^{y_2}\theta_3^{y_3}$$

According to a model in genetics, the θ_i's can be expressed in terms of a single parameter α for human populations:

$$\theta_1 = \alpha^2,\ \theta_2 = 2\alpha(1-\alpha),\ \theta_3 = (1-\alpha)^2$$

where α is a parameter with $0 \le \alpha \le 1$. In this case

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3; \alpha) = \frac{n!}{y_1!y_2!y_3!}[\alpha^2]^{y_1}[2\alpha(1-\alpha)]^{y_2}[(1-\alpha)^2]^{y_3}$$

If the observed data are y_1, y_2, y_3 then the likelihood function for α is

$$\begin{aligned} L(\alpha) &= \frac{n!}{y_1!y_2!y_3!}[\alpha^2]^{y_1}[2\alpha(1-\alpha)]^{y_2}[(1-\alpha)^2]^{y_3} \\ &= \frac{n!}{y_1!y_2!y_3!}2^{y_2}\alpha^{2y_1+y_2}(1-\alpha)^{y_2+2y_3} \quad \text{for } 0 \le \alpha \le 1 \end{aligned}$$

or more simply

$$L(\alpha) = \alpha^{2y_1+y_2}(1-\alpha)^{y_2+2y_3} \quad \text{for } 0 \le \alpha \le 1$$

The log likelihood function is

$$l(\alpha) = (2y_1 + y_2)\log\alpha + (y_2 + 2y_3)\log(1-\alpha) \quad \text{for } 0 < \alpha < 1$$

with

$$\frac{dl}{d\alpha} = \frac{2y_1 + y_2}{\alpha} - \frac{y_2 + 2y_3}{1-\alpha}$$

and

$$\frac{dl}{d\alpha} = 0 \ \text{ if } \ \alpha = \frac{2y_1 + y_2}{2y_1 + 2y_2 + 2y_3} = \frac{2y_1 + y_2}{2n}$$

so

$$\hat{\alpha} = \frac{2y_1 + y_2}{2n}$$

is the maximum likelihood estimate of α.

2.5 Invariance Property of Maximum Likelihood Estimate

Many statistical problems involve the estimation of attributes of a population or process. These attributes can often be represented as an unknown parameter or parameters in a statistical model. The method of maximum likelihood gives us a general method for estimating these unknown parameters. Sometimes the attribute of interest is a function of the unknown parameters. Fortunately the method of maximum likelihood allows us to estimate functions of unknown parameters with very little extra work. This property is called the invariance property of maximum likelihood estimates and can be stated as follows:

Theorem 16 *If* $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ *is the maximum likelihood estimate of* $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ *then* $g(\hat{\boldsymbol{\theta}})$ *is the maximum likelihood estimate of* $g(\boldsymbol{\theta})$.

Example 2.5.1

Suppose we want to estimate attributes associated with BMI for some population of individuals (for example, Canadian males age 21-35). If the distribution of BMI values in the population is well described by a Gaussian model, $Y \sim G(\mu, \sigma)$, then by estimating μ and σ we can estimate any attribute associated with the BMI distribution. For example:

(i) The mean BMI in the population corresponds to $\mu = E(Y)$ for the Gaussian distribution.

(ii) The median BMI in the population corresponds to the median of the Gaussian distribution which equals μ since the Gaussian distribution is symmetric about its mean.

(iii) For the BMI population, the 0.1 (population) quantile, $Q(0.1) = \mu - 1.28\sigma$. (To see this, note that $P(Y \leq \mu - 1.28\sigma) = P(Z \leq -1.28) = 0.1$, where $Z = (Y - \mu)/\sigma$ has a $G(0, 1)$ distribution.)

(iv) The fraction of the population with BMI over 35.0 given by

$$p = 1 - \Phi\left(\frac{35.0 - \mu}{\sigma}\right)$$

where Φ is the cumulative distribution function for a $G(0, 1)$ random variable.

Suppose a random sample of 150 males gave observations $y_1, y_2, \ldots, y_{150}$ and that the maximum likelihood estimates based on the results derived in Example 2.3.2 were

$$\hat{\mu} = \bar{y} = 27.1 \text{ and } \hat{\sigma} = \left[\frac{1}{150}\sum_{i=1}^{150}(y_i - \bar{y})^2\right]^{1/2} = 3.56$$

The estimates of the attributes in $(i) - (iv)$ would be:

(i) and (ii) $\hat{\mu} = \hat{m} = 27.1$

(iii) $\hat{Q}(0.1) = \hat{\mu} - 1.28\hat{\sigma} = 27.1 - 1.28(3.56) = 22.54$ and

(iv) $\hat{p} = 1 - \Phi\left(\frac{35.0 - \hat{\mu}}{\hat{\sigma}}\right) = 1 - \Phi(2.22) = 1 - 0.98679 = 0.01321.$

Note that (iii) and (iv) follow from the invariance property of maximum likelihood estimates.

2.6 Checking the Model

The models used in this course are probability distributions for random variables that represent variates in a population or process. A typical model has probability density function $f(y;\theta)$ if the variate Y is continuous, or probability function $f(y;\theta)$ if Y is discrete, where θ is (possibly) a vector of parameter values. If a family of models is to be used for some purpose then it is important to check that the model adequately represents the variability in Y. This can be done by comparing the model with random samples $y_1, y_2, \ldots, y_n$ of y-values from the population or process.

Comparing Observed and Expected Frequencies

One method for checking how well the model fits the data is to compare observed frequencies with the expected frequencies calculated using the assumed model. This method is particularly useful for data that have arisen from a discrete probability model. The two examples below illustrate the method.

Example 2.6.1 Rutherford and Geiger study of alpha-particles and the Poisson model

In 1910 the physicists Ernest Rutherford and Hans Geiger conducted an experiment in which they recorded the number of alpha particles emitted from a polonium source (as detected by a Geiger counter) during 2608 time intervals each of length 1/8 minute. The number of particles j detected in the time interval and the frequency f_j of that number of particles is given in Table 2.3.

We can see whether a Poisson model fit these data by comparing the observed frequencies with the expected frequencies calculated assuming a Poisson model. To calculate these expected frequencies we need to specify the mean θ of the Poisson model. We estimate θ using the sample mean for the data which is

$$\begin{aligned} \hat{\theta} &= \frac{1}{2608}\sum_{j=0}^{14} j f_j \\ &= \frac{1}{2608}(10097) \\ &= 3.8715 \end{aligned}$$

The expected number of intervals in which j particles is observed is

$$e_j = (2608)\,\frac{(3.8715)^j\, e^{-3.8715}}{j!} \quad \text{for } j = 0, 1, \ldots$$

The expected frequencies are also given in Table 2.3.

Since the observed and expected frequencies are reasonably close, the Poisson model seems to fit these data well. Of course, we have not specified how close the expected and observed frequencies need to be in order to conclude that the model is reasonable. We will look at a formal method for doing this in Chapter 7.

Number of α-particles detected: j	Observed Frequency: f_j	Expected Frequency: e_j
0	57	54.3
1	203	210.3
2	383	407.1
3	525	525.3
4	532	508.4
5	408	393.7
6	273	254.0
7	139	140.5
8	45	68.0
9	27	29.2
10	10	11.3
11	4	4.0
12	0	1.3
13	1	0.4
14	1	0.1
Total	2608	2607.9

Table 2.3 Frequency table for Rutherford/Geiger data

This comparison of observed and expected frequencies to check the fit of a model can also be used for data that have arisen from a continuous model. The following is an example.

Example 2.6.2 Lifetimes of brake pads and the Exponential model

Suppose we want to check whether an Exponential model is reasonable for modeling the data in Example 1.3.4 on lifetimes of brake pads. To do this we need to estimate the mean θ of the Exponential distribution. We use the sample mean $\bar{y} = 49.0275$ to estimate θ.

Since the lifetime Y is a continuous random variable taking on all real values greater than zero the intervals for the observed and expected frequencies are not obvious as they were in the discrete case. For the lifetime of brake pads data we choose the same intervals which were used to produce the relative frequency histogram in Example 1.3.4 except we have collapsed the last four intervals into one interval $[120, +\infty)$. The intervals are given in Table 2.4.

The expected frequency in the interval $[a_{j-1}, a_j)$ is calculated using

$$\begin{aligned} e_j &= 200 \int_{a_{j-1}}^{a_j} \frac{1}{49.0275} e^{-y/49.0275} dy \\ &= 200 \left(e^{-a_{j-1}/49.0275} - e^{-a_j/49.0275} \right) \end{aligned}$$

The expected frequencies are also given in Table 2.4. We notice that the observed and

expected frequencies are not close in this case and therefore the Exponential model does not seem to be a good model for these data.

Interval	Observed Frequency: f_j	Expected Frequency: e_j
$[0, 15)$	21	52.72
$[15, 30)$	45	38.82
$[30, 45)$	50	28.59
$[45, 60)$	27	21.05
$[60, 75)$	21	15.50
$[75, 90)$	9	11.42
$[90, 105)$	12	8.41
$[105, 120)$	7	6.19
$[120, +\infty)$	8	17.3
Total	200	200

Table 2.4: Frequency table for brake pad data

The drawback of this method for continuous data is that the intervals must be selected and this adds a degree of arbitrariness to the method. The following graphical methods provide better techniques for checking the fit of the model for continuous data.

Graphical Checks of Models

We may also use graphical techniques for checking the fit of a model. These methods are particularly useful for continuous data.

Relative frequency histograms and probability density functions

The first graphical method is to superimpose the probability density function of the proposed model on the relative frequency histogram of the data. Figure 2.5 gives the relative frequency histogram of the female BMI data with a superimposed Gaussian probability density function. Since the mean μ is unknown we estimate using the sample mean $\bar{y} = 26.9$ and since the standard deviation σ is unknown we estimate it using the sample standard deviation $s = 4.60$.

Figure 2.6 gives the relative frequency histogram of the male BMI data with a superimposed Gaussian probability density function. Since the mean μ is unknown we estimate using the sample mean $\bar{y} = 27.08$ and since the standard deviation σ is unknown we estimate it using the sample standard deviation $s = 3.56$. In both figures the relative frequency histograms are in reasonable agreement with the superimposed Gaussian probability density functions.

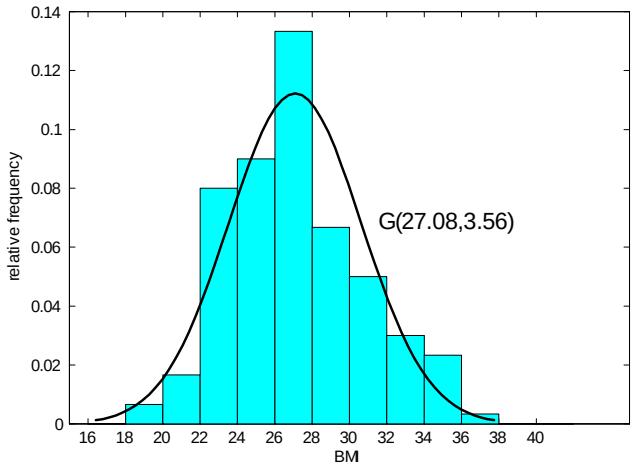


Figure 2.5: **Relative frequency histogram of female BMI data with Gaussian p.d.f.**

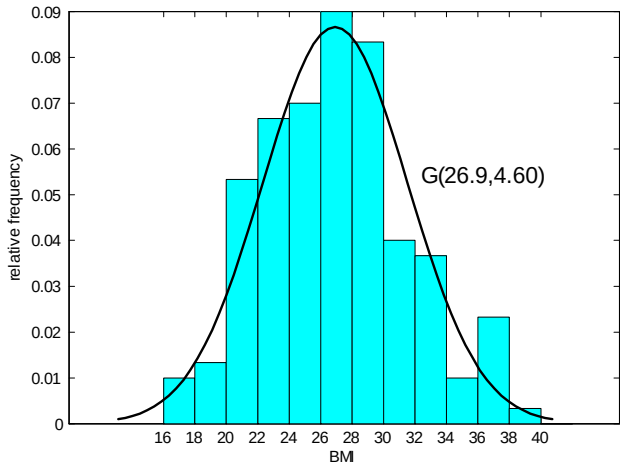


Figure 2.6: **Relative frequency histogram of male BMI data with Gaussian p.d.f.**

If we observe an obvious systematic departure between the relative frequency histogram and the superimposed probability model, the nature of the systematic departure may suggest a better model for the data. For example Figure 2.7 suggests that a more appropriate model would be a model with a longer right tail than the Gauss distribution.

The drawback of this technique is that the intervals for the relative frequency histogram must be chosen.

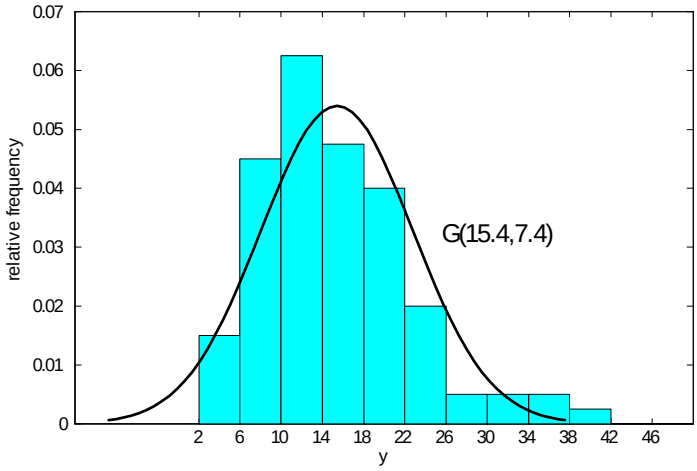


Figure 2.7: **Example of systematic departure from Gaussian model**

Empirical cumulative distribution functions and cumulative distribution functions

A second graphical method which can be used to check the fit of a model is to plot the empirical cumulative distribution function $\hat{F}(y)$ which was defined in Chapter 1 and then superimpose on this a plot of the cumulative distribution function, $P(Y \leq y; \theta) = F(y; \theta)$ for the proposed model. If the graphs of the two functions differ a great deal, this would suggest that the proposed model is a poor fit to the data. Systematic departures may also suggest a better model for the data.

Example 2.6.3 Checking an Exponential(θ) model

Figure 2.8 is a graph of the empirical cumulative distribution function $\hat{F}(y)$ for the data in Example 1.6.3 with an Exponential cumulative distribution function superimposed. The unknown mean θ is estimated using the sample mean $\bar{y} = 3.5$. Not surprisingly (since the data were randomly generated from an Exponential distribution) the agreement between the two curves is very good.

Figure 2.9 is a graph of the empirical cumulative distribution function $\hat{F}(y)$ for the data in Figure 2.7 with an Exponential cumulative distribution function superimposed. The unknown mean θ is estimated using the sample mean $\bar{y} = 15.4$. In this case the agreement between the two curves is very poor. The disagreement between the curves suggests that the proposed Exponential model disagrees with the observed distribution in both tails of the distribution.

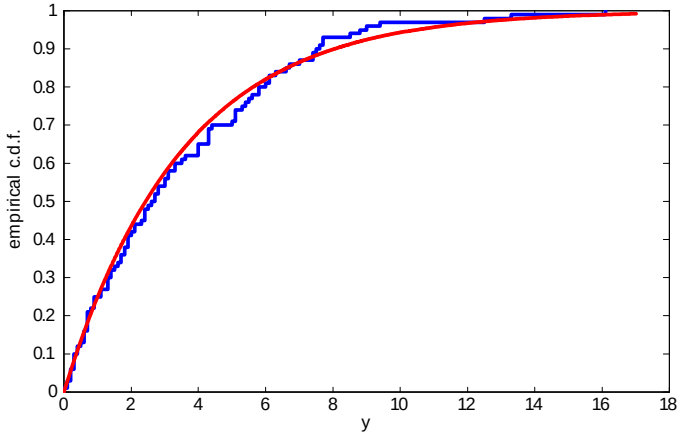


Figure 2.8: **Empirical c.d.f. and Exponential c.d.f.**

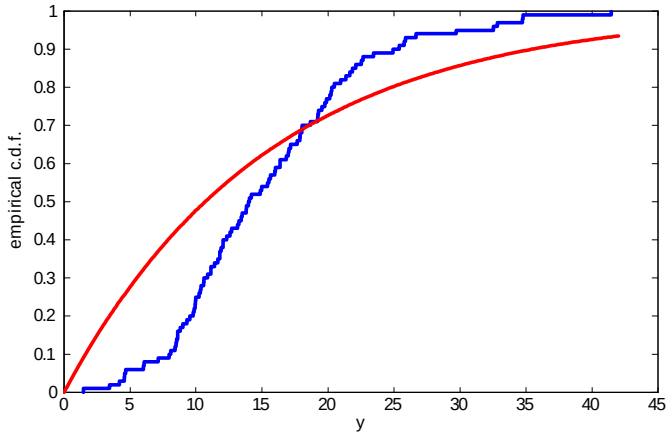


Figure 2.9: **Empirical c.d.f. and Exponential c.d.f.**

Example 2.6.4 Old Faithful data

Consider data for the time in minutes between 300 eruptions of the geyser *Old Faithful* in Yellowstone National Park, between the first and the fifteenth of August 1985. The data are available in the file *oldfaithfuldata.txt* posted on the course website. The empirical cumulative distribution function for the data are plotted in Figure 2.10. One might hypothesize that the distribution of times between consecutive eruptions follows a Gaussian distribution. To see how well a Gaussian model fits the data we could superimpose a Gaussian cumulative distribution function on the plot of the empirical cumulative distribution function. To do this we need to estimate the parameters μ and σ of the Gaussian model since they are unknown. We estimate the mean μ using the sample mean $\bar{y} = 72.3$ and the standard deviation σ using the sample standard deviation $s = 13.9$. In Figure 2.10

the cumulative distribution function of a $G\,(72.3, 13.9)$ random variable is superimposed on the empirical cumulative distribution function for the data. There is poor agreement between the two curves suggesting a Gaussian model is not suitable for these data.

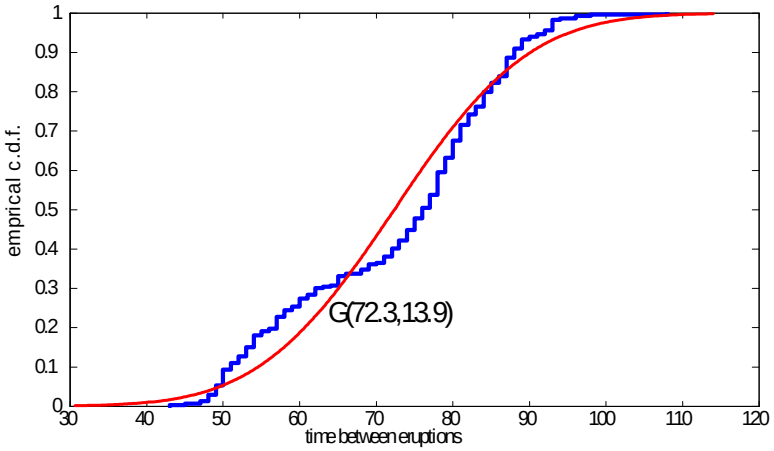


Figure 2.10: **Empirical c.d.f. of times between eruptions of Old Faithful and Gaussian c.d.f.**

The relative frequency histogram in Figure 2.11 seems to indicate that the distribution of the times appears to have two modes. The plot of the empirical cumulative distribution function does not show the shape of the distribution as clearly as the histogram.

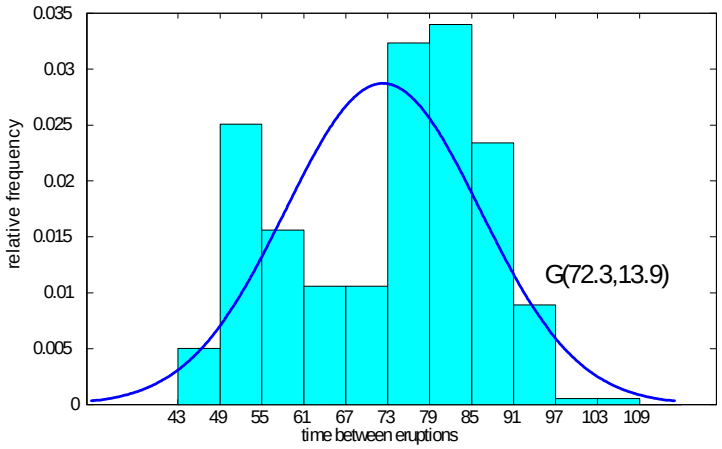


Figure 2.11: **Relative frequency histogram for times between eruptions of Old Faithful and Gaussian p.d.f.**

Example 2.6.5 Heights of females

For the data on female heights in Chapter 1 and using the results from Example 2.3.2 we obtain $\hat{\mu} = 1.62$, $\hat{\sigma} = 0.064$ as the maximum likelihood estimates of μ and σ. Figure 2.12 shows a plot of the empirical cumulative distribution function with the $G(1.62, 0.064)$ cumulative distribution function superimposed.

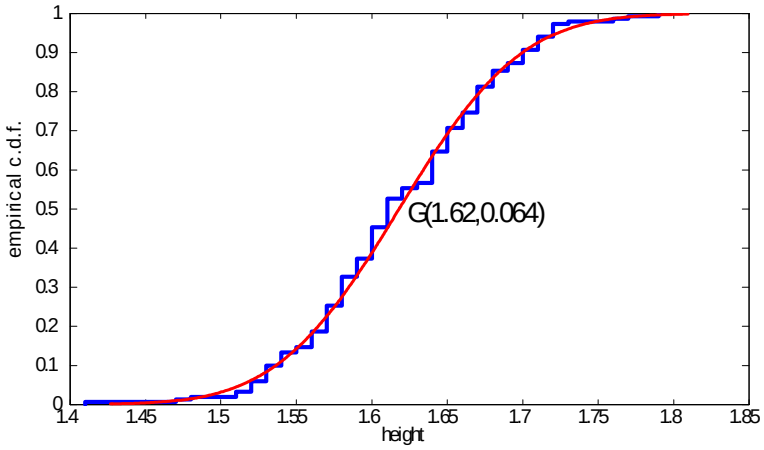


Figure 2.12: **Empirical c.d.f. of female heights and Gaussian c.d.f.**

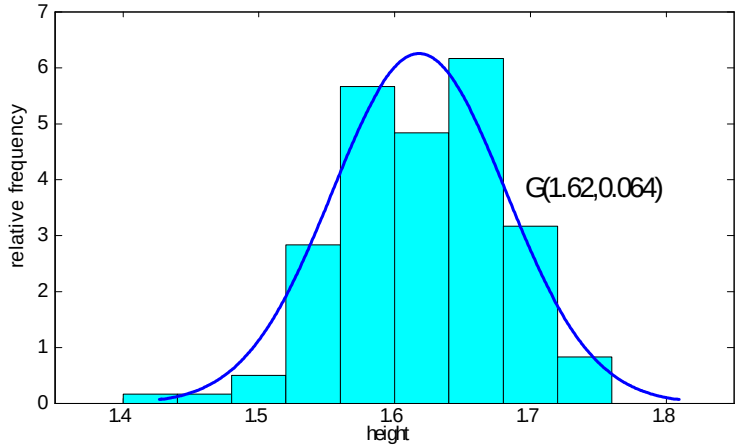


Figure 2.13: **Relative frequency histogram of female heights and Gaussian p.d.f.**

Figure 2.13 shows a relative frequency histogram for these data with the $G(1.62, 0.0637)$ probability density function superimposed. The two types of plots give complementary but consistent pictures. An advantage of the distribution function comparison is that the exact heights in the sample are used, whereas in the histogram plot the data are grouped into intervals to form the histogram. However, the histogram and probability density function

show the distribution of heights more clearly. Both graphs indicate that a Gaussian model seems reasonable for these data.

Qqplots for checking Gaussian model

Since the Gaussian model is used frequently for modeling data we look at one more graphical technique called a (Gaussian) qqplot for checking how well a Gaussian model fits a set of data. The idea behind this method is that we expect the empirical cumulative distribution function and the cumulative distribution for a Gaussian random variable to agree if a Gaussian model is appropriate for the data as we saw in Figure 2.12. Deciding if two curves are in agreement is usually more difficult than deciding if a set of points lie along a straight line. A qqplot is a graph for which the expected plot would reasonably be a straight line plot if the Gaussian model is a good fit.

Suppose for the moment that we want to check if a $G(\mu, \sigma)$ model fits the set of data $\{y_1, y_2, \ldots, y_n\}$ where μ and σ are known. As usual we let $\{y_{(1)}, y_{(2)}, \ldots, y_{(n)}\}$ represent the order statistic or the data ordered from smallest to largest. Let $Q(p)$ be the pth (theoretical) quantile for the $G(\mu, \sigma)$ distribution, that is, $Q(p)$ satisfies $P(Y \leq Q(p)) = p$ where $Y \sim G(\mu, \sigma)$. Recall also that $q(p)$ is the pth sample quantile defined in Chapter 1. If the Gaussian model is appropriate then for a reasonable size data set, we would expect $Q(0.5) = \text{median} = \mu$ to be close in value to the sample quantile $q(0.5) = \text{sample median}$, $Q(0.25)$ to be close in value to the lower quartile $q(0.25)$, $Q(0.75)$ to be close in value to the upper quartile $q(0.75)$, and so on. More generally we would expect $Q\left(\frac{i}{n+1}\right)$ to be close in value to the sample quantile $q\left(\frac{i}{n+1}\right)$ (see Definition 3) for $i = 1, 2, \ldots, n$. (Note that we use $\frac{i}{n+1}$ rather than $\frac{i}{n}$ since $Q(1) = \infty$.) For a reasonably large data set we also have $q\left(\frac{i}{n+1}\right) \approx y_{(i)}$, $i = 1, 2, \ldots, n$. Therefore if the Gaussian model fits the data we expect $Q\left(\frac{i}{n+1}\right)$ to be close in value to $q\left(\frac{i}{n+1}\right)$, $i = 1, 2, \ldots, n$. If we plot the points $\left(Q\left(\frac{i}{n+1}\right), q\left(\frac{i}{n+1}\right)\right)$, $i = 1, 2, \ldots, n$ then we should see a set of points that lie reasonably along a straight line.

But what if μ and σ are unknown? Let $Q_z(p)$ be the pth quantile for the $G(0, 1)$ distribution. We know that if $Y \sim G(\mu, \sigma)$ then $\frac{Y-\mu}{\sigma} \sim G(0, 1)$ and therefore $Q(p) = \mu + \sigma Q_z(p)$. Therefore if we plot the points $\left(Q_Z\left(\frac{i}{n+1}\right), q\left(\frac{i}{n+1}\right)\right)$, $i = 1, 2, \ldots, n$ we should still see a set of points that lie reasonably along a straight line if a Gaussian model is reasonable model for the data. Such a plot is called a (Normal) qqplot. The advantage of a qqplot is that the unknown parameters μ and σ do not need to be estimated.

Qqplots exist for other models but we only use Gaussian qqplots.

Since reading qqplots requires some experience, it is a good idea to generate many plots where we know the correct answer. This can be done by generating data from a known distribution and then plotting a qqplot. See Chapter 2, Problems 21 and22. A qqplot of 100 observations randomly generated from a $G(-2, 3)$ distribution is given in Figure

2.14. The theoretical quantiles are plotted on the horizontal axis and the empirical or sample quantiles are plotted on the vertical axis. The line in the qqplot is the line joining the lower and upper quartiles of the empirical and Gaussian distributions, that is, the line joining $(Q_Z(0.25), q(0.25))$ and $(Q_Z(0.75), q(0.75))$ where $Q_Z(0.25) = -0.674$ and $Q_Z(0.75) = 0.674$.We do not expect the points to lie exactly along a straight line since the

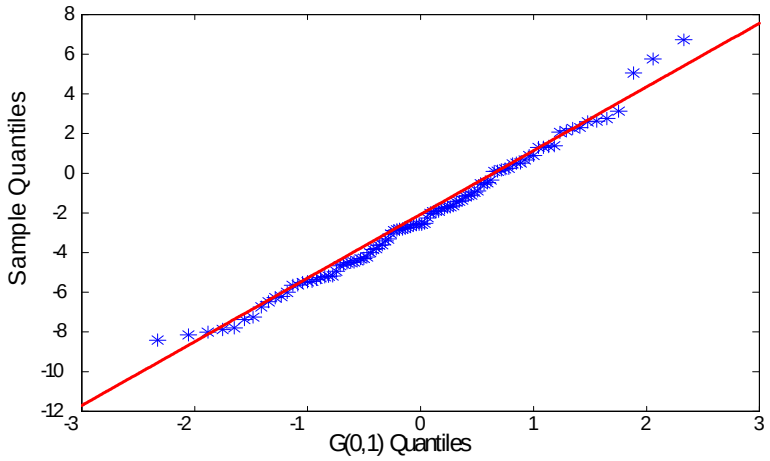


Figure 2.14: **Qqplot of a random sample of** 100 **observations from a** $G(-2, 3)$ **distribution**

sample quantiles are based on the observed data which in general will be different every time the experiment is conducted. We only expect $Q\left(\frac{i}{n+1}\right)$ to be close in value to the sample quantile $q\left(\frac{i}{n+1}\right)$ for a reasonably large data set.

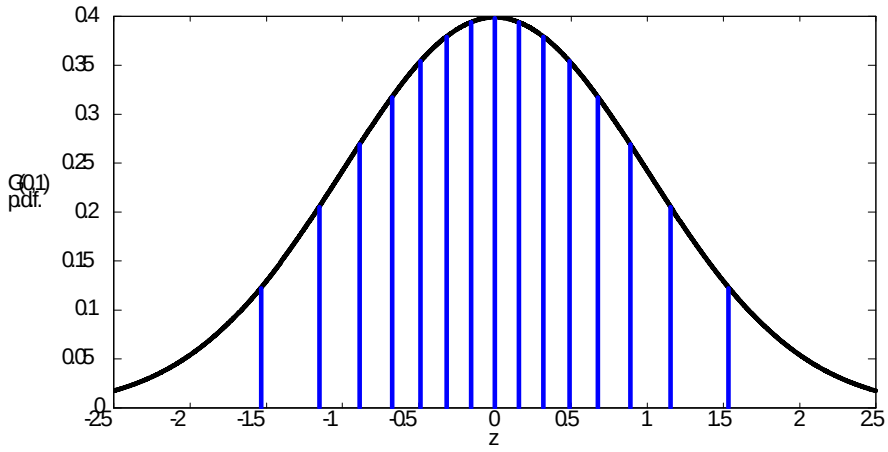


Figure 2.15: **Quantiles of the** $G(0, 1)$ **distribution for** $n = 15$

As well the points at both ends of the line can be expected to lie further from the line since the quantiles of the Gaussian distribution change in value more rapidly in the tails of the distribution. To understand this consider Figure 2.15. The area under a $G(0,1)$ probability density function, which is equal to one, has been divided into sixteen areas all of the same size equal to 1/16. The theoretical quantiles $Q\left(\frac{i}{n+1}\right)$, $i = 1, 2, \ldots, 15$ can be read from the z axis. For example, $Q\left(\frac{1}{16}\right) = -1.53$, and $Q\left(\frac{10}{16}\right) = 0.32$. Since the area under the $G(0,1)$ probability density function is more concentrated near zero, the values of the quantiles increase more quickly in the tails of the distribution. In Figure 2.15 this is illustrated by the vertical lines being closer together near $z = 0$ and further apart for $z < -1$ and $z > 1$. This means we would not expect the sample quantiles in both tails to be as close to the theoretical quantiles as compared to what we observe in the center of the distribution.

A qqplot of the female heights is given in Figure 2.16. Overall the points lie reasonably along a straight line with the points at both ends lying not as close to the line which is what we expect. As was the case for the relative frequency histogram and the empirical cumulative distribution function, the qqplot indicates that the Gaussian model is reasonable for these data. Since the heights in meters are rounded to two decimal places there are many repeated values in the dataset. The repeated values result in the qqplot looking like a set of small steps.

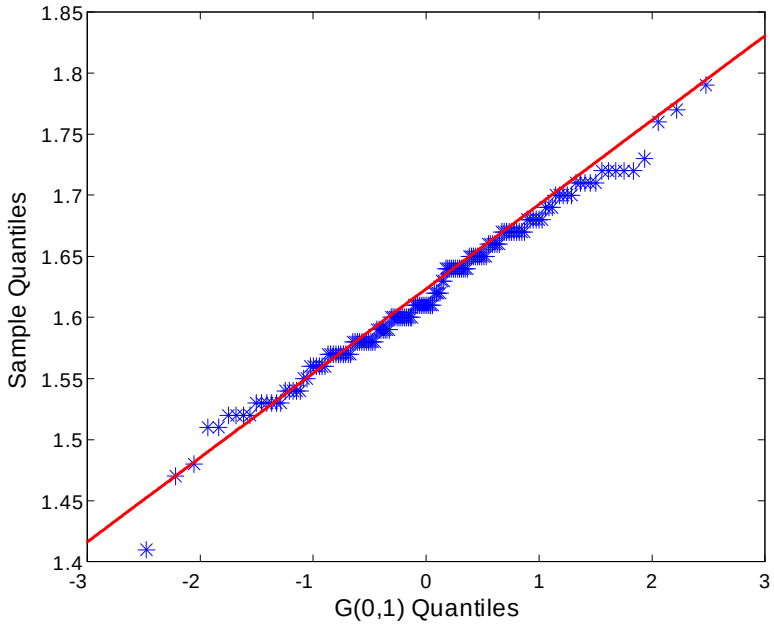


Figure 2.16: **Qqplot of heights of females**

A qqplot of 100 observations randomly generated from a *Exponential* (1) distribution is given in Figure 2.17.We notice that the points form a U-shape. This is typical of data

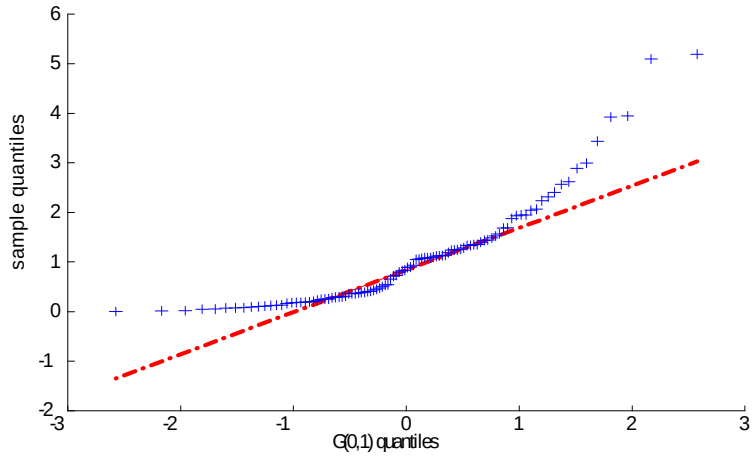


Figure 2.17: **Qqplot of a random sample of** 100 **observations from a** *Exponential* (1) **distribution**

which are best modeled by an Exponential distribution.

To understand why this happens the area under a *Exponential* (1) probability density function has been divided into sixteen areas all of the same size equal to 1/16 in Figure 2.18. The theoretical quantiles can be read from the x axis. The values of the quantiles increase more quickly in the right tail of the distribution.

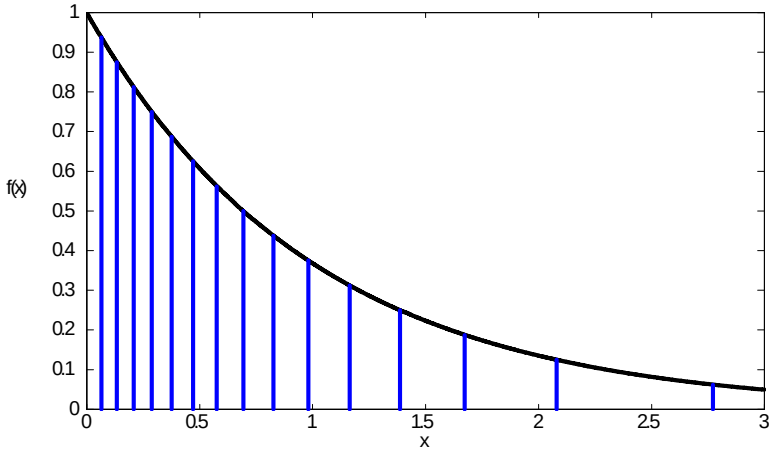


Figure 2.18: **Quantiles of the** *Exponential* (1) **distribution for** $n = 15$

If we plot the theoretical quantiles of an *Exponential* (1) distribution versus the theoretical quantiles of an $G(0, 1)$ distribution for $n = 15$ we obtain the U-shaped graph in

Figure 2.19. Since we are using the theoretical quantiles for both distributions the points lie along a curve. For real data the qqplot would look similar to the plot in Figure 2.17. In general if a dataset has a relative frequency histogram with a long right tail then the qqplot will exhibit this U-shape behaviour. Such a qqplot suggests that a Gaussian model is not reasonable for the data and a model with a long right tail like the Exponential distribution would be more suitable.

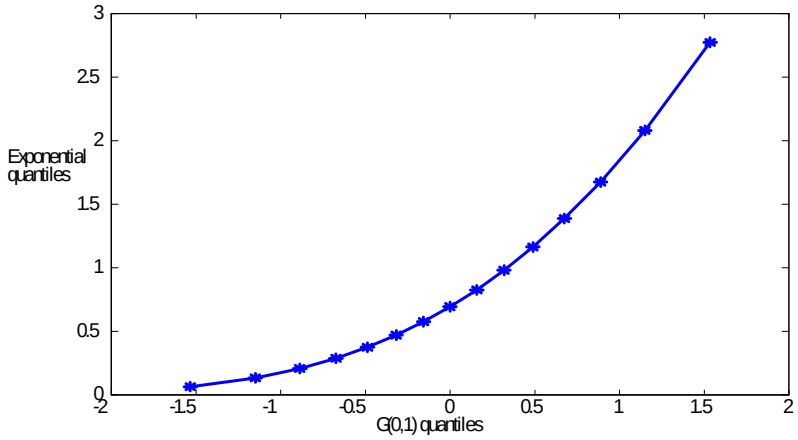


Figure 2.19: **Exponential versus Gaussian quantiles**

A qqplot of the lifetimes of brake pads (Example 1.3.4) is given in Figure 2.20. The points form a U-shaped curve. This pattern is consistent with the long right tail and positive skewness that we observed before. The Gaussian model is not a reasonable model for these data.

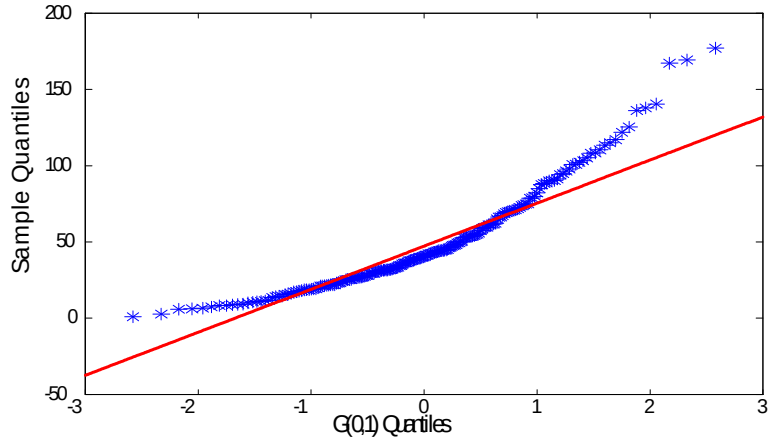


Figure 2.20: **Qqplot of lifetimes of brake pads**

A qqplot of 100 observations randomly generated from a $Uniform\,(0,1)$ distribution is given in Figure 2.21.

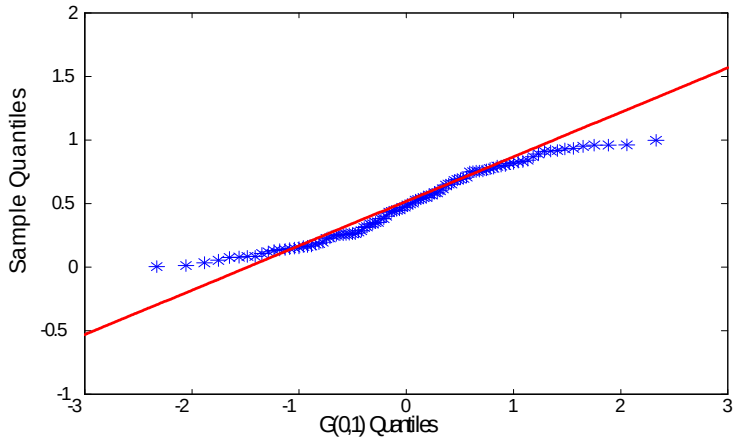


Figure 2.21: Qqplot of 100 observations

We notice that the points form an S-shape. This is typical of data which are best modeled by a Uniform distribution.

To understand why this happens the area under a $Uniform\,(0,1)$ probability density function has been divided into sixteen areas all of the same size equal to 1/16 in Figure 2.22. The theoretical quantiles can be read from the x axis. The values of the quantiles increase uniformly.

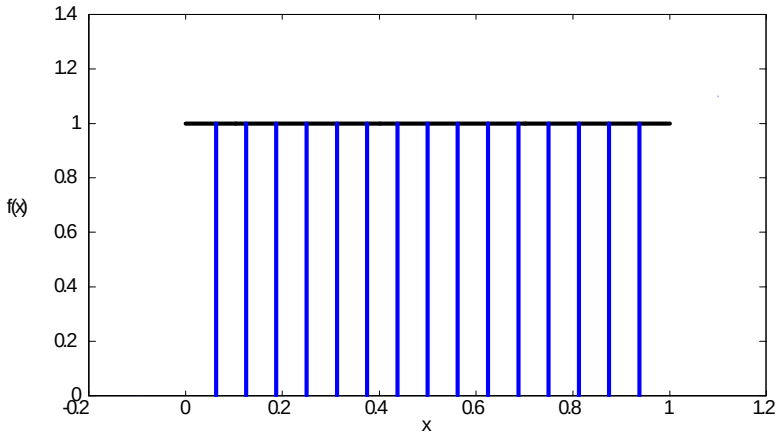


Figure 2.22: **Quantiles of the** $Uniform\,(0,1)$ **distribution for** $n = 15$

If we plot the theoretical quantiles of an $Uniform\,(0,1)$ distribution versus the theoretical quantiles of an $G\,(0,1)$ distribution for $n = 15$ we obtain the S-shaped graph in Figure 2.23. Since we are using the theoretical quantiles for both distributions the points lie along

a curve. For real data the qqplot would look similar to the plot in Figure 2.21. In general if a dataset has a relative frequency histogram which is quite symmetric and with short tails then the qqplot will exhibit this S-shape behaviour. Such a qqplot suggests that a Gaussian model is not reasonable for these data and a model such as the Uniform distribution would be more suitable.

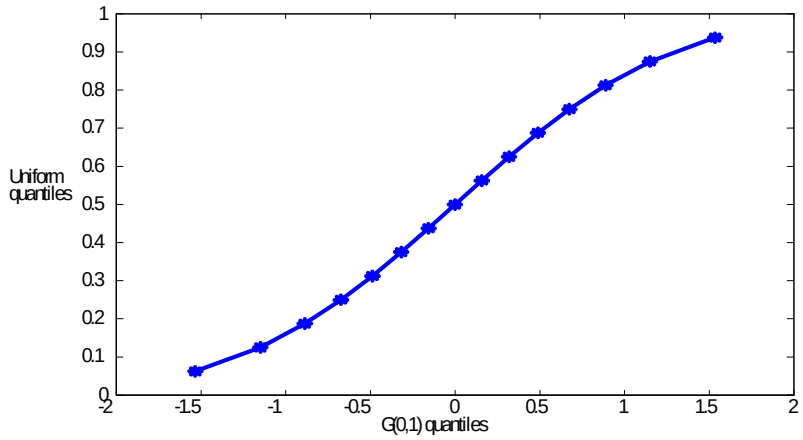


Figure 2.23: **Uniform versus Gaussian quantiles**

A qqplot of the times between eruptions of Old Faithful is given in Figure 2.24. The points do not lie along a straight line which indicates as we saw before that the Gaussian model is not a reasonable model for these data. The two places at which the shape of the points changes direction correspond to the two modes of these data that we observed previously.

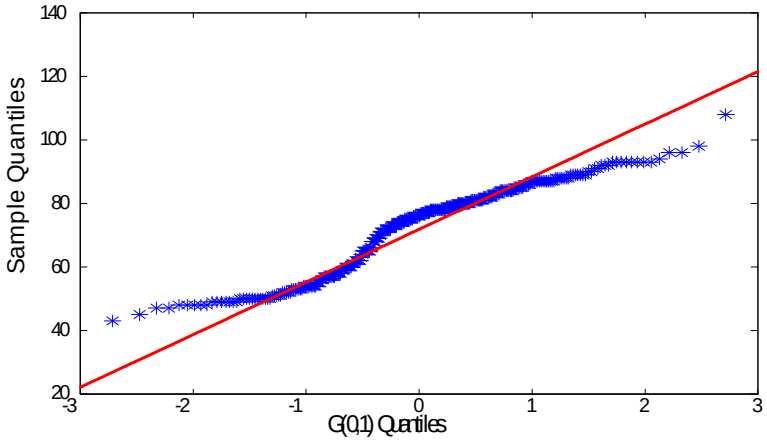


Figure 2.24: **Qqplot of times between eruptions of Old Faithful**

2.7 Chapter 2 Problems

1. To find maximum likelihood estimates we usually find θ such that $\frac{d}{d\theta}\log L(\theta) = 0$. For each of the functions $G(\theta)$ given below find the value of θ which maximizes $G(\theta)$ by finding the value of θ which maximizes $g(\theta) = \log G(\theta)$. Use the First Derivative Test to verify that the value corresponds to a maximum. Note: a and b are positive real numbers.

   (a) $G(\theta) = \theta^a (1-\theta)^b$, $0 \le \theta \le 1$

   (b) $G(\theta) = \theta^{-a} e^{-b/\theta}$, $\theta > 0$

   (c) $G(\theta) = \theta^a e^{-b\theta}$, $\theta \ge 0$

   (d) $G(\theta) = e^{-a(\theta-b)^2}$, $\theta \in \Re$

2. If y successes are observed in a Binomial experiment with n trials and $\theta = P(\text{success})$, the likelihood function for θ is

$$L(\theta) = \theta^y (1-\theta)^{n-y} \quad \text{for } 0 \le \theta \le 1$$

   If $y = 1, 2, \ldots, n-1$, the maximum likelihood estimate of θ is $\hat{\theta} = \frac{y}{n}$ which is found by solving $\frac{d}{d\theta}L(\theta) = 0$ or equivalently $\frac{d}{d\theta}\log L(\theta) = 0$. Show that if $y = 0$ or $y = n$ then the maximum likelihood estimate is not found by solving $\frac{d}{d\theta}\log L(\theta) = 0$ but the maximum likelihood estimate is still $\hat{\theta} = \frac{y}{n}$.

3. Consider the following two experiments whose purpose was to estimate θ, the fraction of a large population with blood type B.

   Experiment 1: Individuals were selected at random until 10 with blood type B were found. The total number of people examined was 100.

   Experiment 2: One hundred individuals were selected at random and it was found that 10 of them have blood type B.

   (a) Find the likelihood function for θ for each experiment and show that the likelihood functions are proportional. Show the maximum likelihood estimate $\hat{\theta}$ is the same in each case.

   (b) Suppose n people came to a blood donor clinic. Assuming $\theta = 0.10$, use the Normal approximation to the Binomial distribution (remember to use a continuity correction) to determine how large should n be to ensure that the probability of getting 10 or more donors with blood type B is at least 0.90? Use the R function `pbinom` to determine the exact value of n.

4. Specimens of a high-impact plastic are tested by repeatedly striking them with a hammer until they fracture. Let $Y =$ the number of blows required to fracture a specimen. If the specimen has a constant probability θ of surviving a blow, independently of the number of previous blows received, then the probability function for Y is

$$f(y;\theta) = P(Y = y;\theta) = \theta^{y-1}(1-\theta) \quad \text{for } y = 1, 2, \ldots;\ 0 \leq \theta < 1$$

   (a) For observed data $y_1, y_2, \ldots, y_n$, find the likelihood function $L(\theta)$ and the maximum likelihood estimate $\hat{\theta}$.

   (b) Find the relative likelihood function $R(\theta)$. Plot $R(\theta)$ if $n = 200$ and $\sum_{i=1}^{200} y_i = 400$.

   (c) Estimate the probability that a specimen fractures on the first blow using the data in (b).

5. In modelling the number of transactions of a certain type received by a central computer for a company with many on-line terminals the Poisson distribution can be used. If the transactions arrive at random at the rate of θ per minute then the probability of y transactions in a time interval of length t minutes is

$$P(Y = y;\theta) = f(y;\theta) = \frac{(\theta t)^y}{y!} e^{-\theta t} \quad \text{for } y = 0, 1, \ldots \text{ and } \theta \geq 0$$

   (a) The numbers of transactions received in 10 separate one minute intervals were 8, 3, 2, 4, 5, 3, 6, 5, 4, 1. Write down the likelihood function for θ and find the maximum likelihood estimate $\hat{\theta}$.

   (b) Estimate the probability that no transactions arrive during a two-minute interval using the data in (a).

   (c) Use the R function `rpois` with the value $\theta = 4.1$ to simulate the number of transactions received in 100 one minute intervals. Calculate the sample mean and sample variance. Are they approximately the same?
   (Note that $E(Y) = Var(Y) = \theta$ for the Poisson model.)

6. Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the distribution with probability density function

$$f(y;\theta) = \frac{2y}{\theta} e^{-y^2/\theta} \quad \text{for } y > 0 \text{ and } \theta > 0$$

   (a) Find the likelihood function $L(\theta)$ and the maximum likelihood estimate $\hat{\theta}$.

   (b) Find the relative likelihood function $R(\theta)$.

   (c) Plot $R(\theta)$ for $n = 20$ and $\sum_{i=1}^{20} y_i^2 = 72$.

7. Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(\mu, \sigma)$ distribution.

   (a) If σ is known, find the likelihood function $L(\mu)$ and the maximum likelihood estimate $\hat{\mu}$.

   (b) If μ is known, find the likelihood function $L(\sigma)$ and the maximum likelihood estimate $\hat{\sigma}$.

8. Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the distribution with probability density function

$$f(y;\theta) = (\theta+1)y^{\theta} \quad \text{for } 0 < y < 1 \text{ and } \theta > -1$$

   (a) Find the likelihood function $L(\theta)$ and the maximum likelihood estimate $\hat{\theta}$.

   (b) Find the log relative likelihood function $r(\theta) = \log R(\theta)$. If $n = 15$ and $\sum_{i=1}^{15} \log y_i = -34.5$ then plot $r(\theta)$.

9. Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the distribution with probability density function

$$f(y;\theta) = \frac{\theta}{y^{\theta+1}} \quad \text{for } y \geq 1 \text{ and } \theta > 1$$

   (a) Find the likelihood function $L(\theta)$ and the maximum likelihood estimate $\hat{\theta}$.

   (b) Find the relative likelihood function $R(\theta)$.

10. Suppose that in a population of twins, males (M) and females (F) are equally likely to occur and that the probability that a pair of twins is identical is α. If twins are not identical, their sexes are independent.

    (a) Show that

$$P(MM) = P(FF) = \frac{1+\alpha}{4} \quad \text{and} \quad P(MF) = \frac{1-\alpha}{2}$$

    (b) Suppose that n pairs of twins are randomly selected; it is found that n_1 are MM, n_2 are FF, and n_3 are MF, but it is not known whether each set is identical or fraternal. Use these data to find the maximum likelihood estimate of α. What is the value of $\hat{\alpha}$ if $n = 50$ and $n_1 = 16$, $n_2 = 16$, $n_3 = 18$?

11. When Wayne Gretzky played for the Edmonton Oilers (1979-88) he scored an incredible 1669 points in 696 games. The data are given in the frequency table below:

Number of points in a game: y	Observed number of games with y points: f_y
0	69
1	155
2	171
3	143
4	79
5	57
6	14
7	6
8	2
≥ 9	0
Total	696

(a) The Poisson(θ) model has been proposed for the random variable Y = number of points Wayne scores in a game. What does the parameter θ represent?

(b) Find the likelihood function for θ based on the Poisson model and the data in the frequency table simplifies and find the maximum likelihood estimate of θ.

(c) Determine the expected frequencies based on the Poisson model and $\theta = \hat{\theta}$. Comment on how well the Poisson model fits the data. What does this imply about the type of hockey player Wayne was during his time with the Edmonton Oilers? (Recall the assumptions for a Poisson process.)

12. Here are the data for Sidney Crosby playing for the Pittsburgh Penguins in the years 2005-2016.

Number of points in a game: y	Observed number of games with y points: f_y
0	219
1	259
2	185
3	90
4	24
5	4
6	2
≥ 7	0
Total	783

How well does the Poisson model fit these data?

13. The following model has been proposed for the distribution of Y = the number of children in a family, for a large population of families:

$$P(Y=0;\theta)=\frac{1-2\theta}{1-\theta},\ P(Y=y;\theta)=\theta^y \quad \text{for } y=1,2,\ldots \text{ and } 0\le\theta\le\frac{1}{2} \qquad (2.7)$$

(a) What does the parameter θ represent?

(b) Suppose that n families are selected at random and the observed data were

y	0	1	$\cdots$	$y_{\max}$	$> y_{\max}$	Total
f_y	f_0	f_1	$\cdots$	$f_{\max}$	0	n

where f_y = the observed number of families with y children and $y_{\max}$ = maximum number of children observed in a family. Find the probability of observing these data and thus determine the maximum likelihood estimate of θ.

(c) Consider a different type of sampling in which a single child is selected at random and then the number of offspring in that child's family is determined. Let X = the number of children in the family of a randomly chosen child. Assuming that the model (2.7) holds then show that

$$P(X=x;\theta)=cx\theta^x \quad \text{for } x=1,2,\ldots \text{ and } 0<\theta\le\frac{1}{2}$$

where

$$c=\frac{(1-\theta)^2}{\theta}$$

Hint: How do you determine then mean of a Geometric random variable?

(d) Suppose that the type of sampling in part (c) was used and that the following data were obtained:

x	1	2	3	4	> 4	Total
f_x	22	7	3	1	0	33

Find the probability of observing these data and thus determine the maximum likelihood estimate of θ. Estimate the probability a couple has no children using these data.

(e) Suppose the sample in (d) was incorrectly assumed to have arisen from the sampling plan in (b). What would $\hat{\theta}$ be found to be? This problem shows that the way the data have been collected can affect the model.

14. Radioactive particles are emitted randomly over time from a source at an average rate of θ per second. In n time periods of varying lengths $t_1, t_2, \ldots, t_n$ (seconds), the numbers of particles emitted (as determined by an automatic counter) were $y_1, y_2, \ldots, y_n$ respectively. Let Y_i = the number of particles emitted in time interval i of length t_i, $i = 1, 2, \ldots, n$. Suppose it is reasonable to assume that Y_i has a Poisson(θt_i) distribution, $i = 1, 2, \ldots, n$ independently.

(a) Show that the likelihood function for θ based on the Poisson model and the data (y_i, t_i), $i = 1, 2, \ldots, n$ can be simplified to

$$L(\theta) = \theta^{n\bar{y}} e^{-\theta n \bar{t}} \quad \text{for } \theta \geq 0$$

where $\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i$. Find the maximum likelihood estimate of θ.

(b) Suppose that the intervals are all of equal length ($t_1 = t_2 = \cdots = t_n = t$) and that instead of knowing the y_i's, we know only whether or not there were one or more particles emitted in each time interval of length t. Find the likelihood function for θ based on these data, and determine the maximum likelihood estimate of θ.

15. Run the following R code for checking the Gaussian model using numerical and graphical summaries.

```
skewness<-function(x) {(sum((x-mean(x))^3)/length(x))/
(sum((x-mean(x))^2)/length(x))^(3/2)}
kurtosis<- function(x) {(sum((x-mean(x))^4)/length(x))/
(sum((x-mean(x))^2)/length(x))^2}
library(MASS)
#  Gaussian Data Example
set.seed(456458)
yn<-rnorm(200,5,2) # 200 observations from G(5,2) distribution
summary(yn) # five number summary and mean
sd(yn) # sample standard deviation
skewness(yn) # sample skewness
kurtosis(yn) # sample kurtosis
#plot relative frequency histogram and superimpose Gaussian pdf
truehist(yn,ylab="relative frequency",las=1)
title(main="Relative Frequency Histogram of Data")
curve(dnorm(x,mean(yn),sd(yn)),col="red",add=T,lwd=2)
#plot Empirical cdf's and superimpose Gaussian cdf
plot(ecdf(yn),verticals=T,do.points=F,xlab="yn",ylab="ecdf",
main="",lwd=2,las=1)
title(main="Empirical and Gaussian C.D.F.'s")
curve(pnorm(x,mean(yn),sd(yn)),add=T,col="red",lwd=2)
#plot qqplot of the data
qqnorm(yn,xlab="Standard Gaussian Quantiles",main="Qqplot of Data",las=1)
qqline(yn,col="red",lwd=1.5) # add line for comparison
#
# Exponential Data Example
ye<-rexp(200,1/5)  # 200 observations from Exponential(5) dist'n
summary(ye) # five number summary and mean
```

```
sd(ye) # sample standard deviation
skewness(ye) # sample skewness
kurtosis(ye) # sample kurtosis
#plot relative frequency histogram and superimpose Gaussian pdf
truehist(ye,ylab="relative frequency",las=1)
title(main="Relative Frequency Histogram of Data")
curve(dnorm(x,mean(ye),sd(ye)),col="red",add=T,lwd=2)
#plot Empirical cdf's and superimpose Gaussian cdf
plot(ecdf(ye),verticals=T,do.points=F,xlab="ye",ylab="ecdf",
main="",lwd=2,las=1)
title(main="Empirical and Gaussian C.D.F.'s")
curve(pnorm(x,mean(ye),sd(ye)),add=T,col="red",lwd=2)
#plot qqplot of the data
qqnorm(ye,xlab="Standard Gaussian Quantiles",main="Qqplot of Data",las=1)
qqline(ye,col="red") # add line for comparison in red
```

For both examples assume that you don't know how the data were generated. Use the numerical and graphical summaries obtained by running the R code to assess whether it is reasonable to assume that the data have approximately a Gaussian distribution. Support your conclusion with clear reasons written in complete sentences.

16. The marks out of 30 for 100 students on a tutorial test in STAT 231 are available in the file *tutorialtestdata.txt* posted on the course website.

    (a) Use R to determine $y_{(1)}$ (the minimum observation), $y_{(n)}$ (the maximum observation), the range, $q(0.25)$ (the lower quartile), $q(0.75)$ (the upper quartile), the IQR (the interquartile range), $\hat{m}$ (the median), $\bar{y}$ (the sample mean), s (the sample standard deviation), g_1 (the sample skewness), and g_2 (the sample kurtosis).

    (b) Determine the proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$. Compare this with $P(Y \in [\mu - \sigma, \mu + \sigma])$ where $Y \sim G(\mu, \sigma)$.

    (c) Use R to plot a boxplot and qqplot of the data.

    (d) Using these numerical and graphical summaries, assess whether it is reasonable to assume a Gaussian model for these data. Be sure to support your conclusion with clear reasons. What type of variate are we modeling in this question?

17. In a study of osteoporosis, the heights in centimeters of a sample of 351 elderly women randomly selected from a community were recorded. The data are available in the file *osteoporosisdata.txt* posted on the course website.

    (a) Use R to determine $y_{(1)}$ (the minimum observation), $y_{(n)}$ (the maximum observation), the range, $q(0.25)$ (the lower quartile), $q(0.75)$ (the upper quartile),

the IQR (the interquartile range), $\hat{m}$ (the median), $\bar{y}$ (the sample mean), s (the sample standard deviation), g_1 (the sample skewness), and g_2 (the sample kurtosis).

(b) Determine the proportion of observations in the interval $[\bar{y} - s, \bar{y} + s]$ and $[\bar{y} - 2s, \bar{y} + 2s]$. Compare these proportions with $P(Y \in [\mu - \sigma, \mu + \sigma])$ and $P(Y \in [\mu - 2\sigma, \mu + 2\sigma])$ where $Y \sim G(\mu, \sigma)$.

(c) Use R to plot a relative frequency histogram and superimpose a Gaussian probability density function with $\mu = \bar{y}$ and $\sigma = s$.

(d) Use R to plot an empirical distribution function for these data and superimpose a Gaussian cumulative distribution function with $\mu = \bar{y}$ and $\sigma = s$.

(e) Use R to plot a boxplot and qqplot of the data. Do you observe anything unusual about the qqplot? Why might cause this?

(f) Using these numerical and graphical summaries, assess whether it is reasonable to assume a Gaussian model for these data. Be sure to support your conclusion with clear reasons.

18. Consider the data on heights of adult males and females from Chapter 1. The data are available in the file *bmidata.txt* posted on the course website.

(a) Assume that for **each** gender the heights in the population from which the samples were drawn can be modeled by a Gaussian distribution. Obtain the maximum likelihood estimates of the mean and standard deviation in each case.

(b) Give the maximum likelihood estimates for $Q(0.1)$ and $Q(0.9)$, the 10th and 90th percentiles of the height distribution for males and for females.

(c) Give the maximum likelihood estimate for the probability $P(Y > 1.83)$ for males and females (i.e. the fraction of the population over 1.83 m).

(d) A simpler estimate of $P(Y > 1.83)$ that does not use the Gaussian model is

$$\frac{\text{number of person in sample with } y > 1.83}{n}$$

where $n = 150$. Obtain these estimates for males and for females. Can you think of any advantages for this estimate over the one in part (c)? Can you think of any disadvantages?

(e) Suggest and try a method of estimating the 10th and 90th percentile of the height distribution that is similar to that in part (d).

19. The qqplot of the brake pad data in Figure 2.20 indicates that the Gaussian distribution is not a reasonable model for these data. Sometimes transforming the data gives a data set for which the Gaussian model is more reasonable. A log transformation is often used. Plot a qqplot of the log lifetimes and indicate whether the Gaussian

distribution is a reasonable model for these data. The data are posted on the course website.

20. In a large population of males ages $40-50$, the proportion who are regular smokers is α where $0 \leq \alpha \leq 1$ and the proportion who have hypertension (high blood pressure) is β where $0 \leq \beta \leq 1$. If the events S (a person is a smoker) and H (a person has hypertension) are independent, then for a man picked at random from the population the probabilities he falls into the four categories SH, $S\bar{H}$, $\bar{S}H$, $\bar{S}\bar{H}$ are respectively, $\alpha\beta$, $\alpha(1-\beta)$, $(1-\alpha)\beta$, $(1-\alpha)(1-\beta)$. Explain why this is true.

    (a) Suppose that 100 men are selected and the numbers in each of the four categories are as follows:

Category	SH	$S\bar{H}$	$\bar{S}H$	$\bar{S}\bar{H}$
Frequency	20	15	22	43

    Assuming that S and H are independent events, determine the likelihood function for α and β based on the Multinomial distribution, and find the maximum likelihood estimates of α and β.

    (b) Compute the expected frequencies for each of the four categories using the maximum likelihood estimates. Do you think the model used is appropriate? Why might it be inappropriate?

21. Run the following R code:

```
par(mfrow=c(2,2))
for (i in 1:4) {
y<-rnorm(30)
qqnorm(y,xlab='G(0,1) Quantiles',pch=19,col="blue",main="",las=1)
qqline(y,col="red",lwd=2)}
for (i in 1:4) {
y<-rnorm(100)
qqnorm(y,xlab='G(0,1) Quantiles',pch=19,col="blue",main="",las=1)
qqline(y,col="red",lwd=2)}par(mfrow=c(2,2))
```

    Compare the qqplots that you observe for a sample size of $n = 30$ with the qqplots for a sample size of $n = 100$.

22. Run the following R code:

```
par(mfrow=c(2,2))
y1<-rbeta(100,1,1)
qqnorm(y1,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 1')
qqline(y1,col="red",lwd=2)
y2<-rbeta(100,0.3,0.3)
qqnorm(y2,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 2')
```

```
qqline(y2,col="red",lwd=2)
y3<-c(rgamma(100,1,1),-rgamma(100,1,1))
qqnorm(y3,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 3')
qqline(y3,col="red",lwd=2)
y4<-rt(100,3)
qqnorm(y4,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 4')
qqline(y4,col="red",lwd=2)
y5<-rgamma(100,1,1)
qqnorm(y5,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 5')
qqline(y5,col="red",lwd=2)
y6<-(-rgamma(100,1,1))
qqnorm(y6,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 6')
qqline(y6,col="red",lwd=2)
y7<-rgamma(100,4,1)
qqnorm(y7,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 7')
qqline(y7,col="red",lwd=2)
y8<-rweibull(100,2,2)
qqnorm(y8,xlab='G(0,1) Quantiles',pch=19,col="blue",main='Qqplot 8')
qqline(y8,col="red",lwd=2)
```

For each qqplot determine whether the underlying distribution is symmetric. If the distribution is not symmetric indicate if the skewness is positive or negative. If the distribution is symmetric indicate if the kurtosis is larger than 3 or smaller than 3.

After you have made conclusions based on the qqplots regarding the distribution of each dataset, plot the relative frequency histograms for each dataset to see if your conclusions are consistent with the relative frequency histograms.

23. A qqplot for a dataset of 100 observations is given in Figure 2.25. Based on this qqplot, answer the following questions:

    (a) What is the approximate value of the sample median?

    (b) What is the approximate value of the IQR?

    (c) What is the approximate value of the range?

    (d) Would the frequency histogram of these data be reasonably symmetric about the sample mean?

    (e) The frequency histogram for these data would most resemble a Gaussian probability density function, an Exponential probability density function, or a Uniform probability density function?

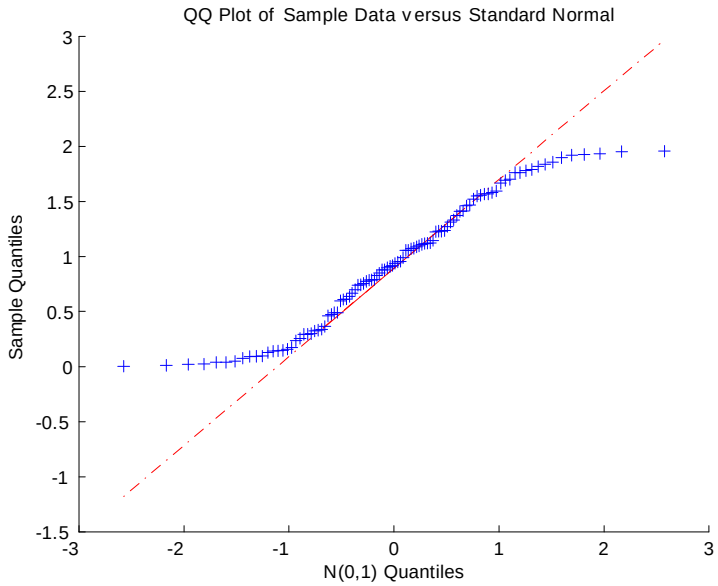


Figure 2.25: **Qqplot for** 100 **observations**

24. **Uniform data:** Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the Uniform$(0, \theta)$ distribution.

    (a) Find the likelihood function, $L(\theta)$.

    (b) Obtain the maximum likelihood estimate of θ. **Hint:** The maximum likelihood estimate is not found by solving $l'(\theta) = 0$.

25. **Challenge Problem: Censored lifetime data** Consider the Exponential distribution as a model for the lifetimes of equipment. In experiments, it is often not feasible to run the study long enough that all the pieces of equipment fail. For example, suppose that n pieces of equipment are each tested for a maximum of c hours (c is called a "censoring time"). The observed data are: k (where $0 \leq k \leq n$) pieces fail, at times $y_1, y_2, \ldots, y_k$ and $n - k$ pieces are still working after time c.

    (a) If $Y \sim$ Exponential(θ), show that $P(Y > c; \theta) = e^{-c/\theta}$, for $c > 0$.

    (b) Determine the likelihood function for θ based on the observed data described above. Show that the maximum likelihood estimate of θ is

    $$\hat{\theta} = \frac{1}{k}\left[\sum_{i=1}^{k} y_i + (n-k)c\right]$$

    (c) What is the maximum likelihood estimate of θ when $k = 0$? Explain this intuitively.

(d) A standard test for the reliability of electronic components is to subject them to large fluctuations in temperature inside specially designed ovens. For one particular type of component, 50 units were tested and $k = 5$ failed before $c = 400$ hours, when the test was terminated, with $\sum_{i=1}^{5} y_i = 450$ hours. Find the maximum likelihood estimate of θ.

26. **Challenge Problem: Estimation from capture-recapture studies** In order to estimate the number of animals, N, in a wild habitat the capture-recapture method is often used. In this scheme k animals are caught, tagged, and then released. Later on n animals are caught and the number Y of these that have tags are noted. The idea is to use this information to estimate N.

(a) Show that under suitable assumptions

$$P(Y = y) = \frac{\binom{k}{y}\binom{N-k}{n-y}}{\binom{N}{n}}$$

(b) For observed k, n and y find the value $\hat{N}$ that maximizes the probability in part (a). Does this ever differ much from the intuitive estimate $\tilde{N} = kn/y$? (**Hint:** The likelihood $L(N)$ depends on the discrete parameter N, and a good way to find where $L(N)$ is maximized over $\{1, 2, 3, \ldots\}$ is to examine the ratios $L(N+1)/L(N)$.)

(c) When might the model in part (a) be unsatisfactory?

27. **Challenge Problem: Poisson model with a covariate** Let Y represent the number of claims in a given year for a single general insurance policy holder. Each policy holder has a numerical "risk score" x assigned by the company, based on available information. The risk score may be used as an explanatory variate when modeling the distribution of Y, and it has been found that models of the form

$$P(Y = y|x) = \frac{[\theta(x)]^y}{y!} e^{-\theta(x)} \quad \text{for } y = 0, 1, \ldots$$

where $\theta(x) = e^{\alpha+\beta x}$, are useful.

(a) Suppose that n randomly chosen policy holders with risk scores $x_1, x_2, \ldots, x_n$ had $y_1, y_2, \ldots, y_n$ claims, respectively, in a given year. Determine the likelihood function for α and β based on these data.

(b) Can $\hat{\alpha}$ and $\hat{\beta}$ be found explicitly?

3. PLANNING AND CONDUCTING EMPIRICAL STUDIES

3.1 Empirical Studies

An empirical study is one which is carried out to learn about a population or process by collecting data. We have given several examples in the preceding two chapters but we have not yet considered the details of such studies. In this chapter we consider how to conduct an empirical study in a systematic way. Well-conducted empirical studies are needed to produce maximal information within existing cost and time constraints. A poorly planned or executed study can be worthless or even misleading. For example, in the field of medicine thousands of empirical studies are conducted every year at very high costs to society and with critical consequences. These investigations must be well planned and executed so that the knowledge they produce is useful, reliable and obtained at reasonable cost.

It is helpful to think about planning and conducting a study using a set of steps such as the following:

- **P**roblem: a clear statement of the study's objectives, usually involving one or more questions
- **P**lan: the procedures used to carry out the study including how the data will be collected
- **D**ata: the physical collection of the data, as described in the Plan
- **A**nalysis: the analysis of the data collected in light of the Problem and the Plan
- **C**onclusion: The conclusions that are drawn about the Problem and their limitations

We will use this set of steps, which we will refer to as PPDAC, to discuss the important ideas which must be considered when planning an empirical study. These steps, which are designed to emphasize the statistical aspects of empirical studies, are described in more detail in Section 3.2.

PPDAC can be used in two ways - first to actively formulate, plan and carry out investigations and second as a framework to critically scrutinize reported empirical investigations. These reports include articles in the popular press, scientific papers, government policy statements, and various business reports. If you see the phrase "evidence based decision" or "evidence based management", look for an empirical study. In this course we will use PPDAC most often to critically assess empirical studies reported in the media.

The following example will be used in the next section to show how PPDAC can be used to describe and critically examine how the empirical study was conducted.

Example 3.1 An empirical study on university student drinking
The following news item, published by the University of Sussex, United Kingdom on February 16, 2015, describes an empirical investigation in the field of psychology.

Campaigns to get young people to drink less should focus on the benefits of not drinking and how it can be achieved:

Pointing out the advantages and achievability of staying sober is more effective than traditional approaches that warn of the risks of heavy drinking, according to the research carried out at the University of Sussex by researcher Dr Dominic Conroy. The study, published this week in the British Journal of Health Psychology, found that university students were more likely to reduce their overall drinking levels if they focused on the benefits of abstaining, such as more money and better health. They were also less likely to binge drink if they had imagined strategies for how non-drinking might be achieved – for example, being direct but polite when declining a drink, or choosing to spend time with supportive friends. Typical promotions around healthy drinking focus on the risks of high alcohol consumption and encourage people to monitor their drinking behaviour (e.g. by keeping a drinks diary). However, the current study found that completing a drinks diary was less effective in encouraging safer drinking behaviour than completing an exercise relating to non-drinking.

Dr Conroy says: "We focused on students because, in the UK, they remain a group who drink heavily relative to their non-student peers of the same age. Similarly, attitudes about the acceptability of heavy drinking are relatively lenient among students. "Recent campaigns, such as the NHS Change4Life initiative, give good online guidance as to how many units you should be drinking and how many units are in specific drinks. "Our research contributes to existing health promotion advice, which seeks to encourage young people to consider taking 'dry days' yet does not always indicate the range of benefits nor suggest how non-drinking can be more successfully 'managed' in social situations."

Dr Conroy studied 211 English university students aged 18-25 over the course of a month. Participants in the study completed one of four exercises involving either: imagining positive outcomes of non-drinking during a social occasion; imagining strategies required to successfully not drink during a social occasion; imagining both positive outcomes and required strategies; or completing a drinks diary task.

At the start of the study, participants in the outcome group were asked to list positive outcomes of not drinking and those in the process group listed what strategies they might use to reduce their

drinking. Those in the combined group did both. They were reminded of their answers via email during the one month course of the study and asked to continue practising this mental simulation. All groups completed an online survey at various points, indicating how much they had drunk the previous week. Over the course of one month, Dr Conroy found that students who imagined positive outcomes of non-drinking reduced their weekly alcohol consumption from 20 units to 14 units on average. Similarly, students who imagined required strategies for non-drinking reduced the frequency of binge drinking episodes – classified as six or more units in one session for women, and eight or more units for men – from 1.05 episodes a week to 0.73 episodes a week on average.

Interestingly, the research indicates that perceptions of non-drinkers were also more favourable after taking part in the study. Dr Conroy says this could not be directly linked to the intervention but was an interesting additional feature of the study. He says: "Studies have suggested that holding negative views of non-drinkers may be closely linked to personal drinking behaviour and we were interested to see in the current study that these views may have improved as a result of taking part in a non-drinking exercise. "I think this shows that health campaigns need to be targeted and easy to fit into daily life but also help support people to accomplish changes in behaviour that might sometimes involve 'going against the grain', such as periodically not drinking even when in the company of other people who are drinking."

To discuss the steps of PPDAC in detail we need to introduce a number of technical terms. Every subject has its own jargon, that is, words with special meaning, and you need to learn the terms describing the details of PPDAC to be successful in this course.

3.2 The Steps of PPDAC

Problem

The Problem step describes what the experimenters are trying to learn or what questions they want to answer. Often this can be done using questions starting with "What".

- What conclusions are the experimenters trying to draw?
- What group of things or people do the experimenters want the conclusions to apply?
- What variates can be defined?
- What is(are) the question(s) the experimenters are trying to answer?

Types of problems

Three common types of statistical problems that are encountered are described below.

- *Descriptive:* The problem is to determine a particular attribute of a population or process. Much of the function of official statistical agencies such as *Statistics Canada*

involves problems of this type. For example, the government needs to know the national unemployment rate and whether it has increased or decreased over the past month.

- *Causative:* The problem is to determine the existence or non-existence of a causal relationship between two variates. For example:

"Does taking a low dose of aspirin reduce the risk of heart disease among men over the age of 50?"

"Does changing from assignments to multiple term tests improve student learning in STAT 231?"

"Does second-hand smoke from parents cause asthma in their children.

"Does compulsory driver training reduce the incidence of accidents among new drivers?"

- *Predictive:* The problem is to predict a future value for a variate of a unit to be selected from the process or population. This is often the case in finance or in economics. For example, financial institutions need to predict the price of a stock or interest rates in a week or a month because this effects the value of their investments.

In a causative problem, the experimenter is interested in whether one variate x *tends to cause* an increase or a decrease in another variate Y. Where possible this is conducted in a controlled experiment in which x is increased or decreased while holding everything else in the experiment constant and we observe the changes in Y. As indicated in Chapter 1, an *experimental study* is one in which the experimenter manipulates the values of the explanatory variates while an *observational study* is one in which the explanatory variates are only observed and not controlled. In the study of the relationship between second-hand smoke and asthma described in Chapter 1, it is unlikely that the experimenter would be able to manipulate the explanatory variate (child lives in household where adults smoke) and so the experimenter can only conduct an observational study. In Chapter 8 we will see how an experimental study can be designed to investigate a causative problem. An observational study in which the experimenter is not in control of the explanatory variates cannot usually be used to investigate a causative problem.

In the drinking study in Example 3.1, which is an experimental study, the problem is causative since the researchers wanted to study the effect of different mental exercises related to non-drinking on the drinking behaviour of university students.

Defining the problem

The first step in describing the Problem is to define the *units* and the *target population* or *target process*.

Definition 17 *The target population or target process is the collection of units to which the experimenters conducting the empirical study wish the conclusions to apply.*

In the drinking study the units are university students and the target population consists of English university students aged $18-25$ in the United Kingdom at the time of the study. Note that "all university students aged $18-25$ in the world" would not be a suitable target population since it would not make much sense to include countries in which the consumption of alcohol is not allowed. A target population of "all English university students aged $18-25$" with no time mentioned is also not a suitable target population for this study since we might expect the drinking behaviour of university students to change over time.

In Chapter 1 we considered a survey of Ontario residents aged $14-20$ in a specific week to learn about their smoking behaviour. In this study the units are young adults and the target population is all young adults aged $14-20$ living in Ontario at the time of the survey. Since smoking behaviour varies from province to province and year to year, the target population of young adults aged $14-20$ in Ontario at the time of the study is the best choice.

In Chapter 1 we considered the comparison of two can filling machines used by a manufacturer with respect to the volume of liquid in the filled cans. The units are the individual cans. The target process is all cans, which could be filled by the manufacturer using the two machines, now and into the future under current operating conditions. Note that in defining the target process the expression "under current operating conditions" has not been well defined.

Recall the following definition from Chapter 1.

Definition 1 *A variate is a characteristic associated with each unit.*

The values of the variates change from unit to unit in the population/process. There are usually many variates associated with each unit.

In the drinking study the most important variates are the weekly alcohol consumption measured over the course of a month, and which mental exercise the student was assigned to. Other variates which were collected are the age of the student and the sex of the student.

In the smoking survey, whether or not each young adult (unit) in the target population smokes is the variate of primary interest. Other variates of interest defined for each unit might be age and sex. In the can-filling example, the volume of liquid in each can (unit) is a variate. Whether the old machine or the new machine was used to fill the can is another variate.

Recall the following definition from Chapter 1.

Definition 2 *An attribute is a function of the variates over a population or process.*

The questions of interest in the Problem are specified in terms of attributes of the target population/process. In the university student drinking study the **mean** (average) alcohol consumption for the different mental exercise groups is the most important attribute. In the smoking example, one important attribute is the **proportion** of young adults in the target population who smoke. In the can-filling example, the attributes of interest were the

mean (average) volume and the **variability** (standard deviation) of the volumes for all cans filled by each machine under current conditions. Possible questions of interest (among others) are:

"Is there a difference in the mean alcohol consumption between the four different mental exercise groups?"

"What proportion of young adults aged $14-20$ in Ontario smoke?"

"Is the standard deviation of volumes of cans filled by the new machine less than that of the old machine?"

We can also ask questions about graphical attributes of the target population such as the **population histogram**, **population cumulative distribution function**, or a **scatterplot** of one variate versus another over the target population.

It is very important that the Problem step contain clear questions about one or more attributes of the target population.

Plan

The Plan step depends on the questions posed in the Problem step. The Plan step includes a description of the population or process of units from which units will be selected, what variates will be collected for the units selected, and how the variates will be measured.

In most cases, the attributes of interest for the target population/process cannot be estimated since only units from a subset of the target population/process can be considered for study or only units from another population completely can be considered for study. This may be due to lack of resources and time, as in the smoking survey in which it would be very costly and nearly impossible to create a list of all young adults aged $14-20$ in Ontario at the time of the study. It may also be a physical impossibility such as in the development of a new product where the manufacturer may wish to make conclusions about a production process in the future but only units produced in a pilot process can be examined. It may also be unethical such as in a clinical trial of a new treatment whose side effects for humans are unknown and which could be life threatening and therefore only laboratory animals such as mice can be used.

Definition 18 *The study population or study process is the collection of units available to be included in the study.*

The study population is often but not always a subset of the target population. In many surveys, the study population is a list of people defined by their telephone number. The sample is selected by calling a subset of the telephone numbers. The study population is a subset of the target population which excludes those people without telephones or with unlisted numbers. In the clinical trial example the study population only consists of the laboratory animals that are available for the study which is not a subset of any target population of humans. In the development of new products example, the units in the pilot process are not a subset of the target process which are the units produced in the future.

The news item for the drinking study does not indicate how the students in the study were recruited. To determine this information we need to check the research journal article. The more detailed article in the British Journal of Health Psychology indicated that administrators at 80 academic departments across 45 English universities were asked to forward a pre-prepared recruitment message to their students containing a URL to an online survey. No reason was given for choosing only these universities. Note that there are over 100 English universities in the United Kingdom as well as other universities in Scotland and Wales (also part of the United Kingdom) which English students could attend. The study population is therefore English university students aged $18 - 25$ at the time of the study at these 45 English universities which is a subset of the target population.

In the smoking survey, it would be difficult to create a list of all young adults aged $14 - 20$ living in Ontario at the time of the survey. Since schools must keep a list of students attending their school as well as student contact information, the researchers may decide to choose a study population of all young adults aged $14 - 20$ living in Ontario at the time of the survey who are attending school. The study population is a subset of the target population.

In the can-filling study a possible study process is all cans which are available at the time of the study and could possibly be filled by the manufacturer using the two machines under current operating conditions. In this case the study process is a subset of the target process.

The study population/process is nearly always different than the target population/process since there are always restrictions on the units which are available to be studied.

Definition 19 *If the attributes in the study population/process differ from the attributes in the target population/process then the difference is called study error.*

Study error cannot be quantified since the values of the target population/process attributes and the study population/process attributes, are unknown. (If these attributes were known then an empirical study would not be necessary!) Context experts would need to be consulted, for example, in order to decide whether or not it is reasonable to assume that conclusions from an investigation using mice are relevant to the human target population. The statistician's role it to warn the context experts of the possibility of such error, especially when the study population/process is very different from the target population/process.

In the drinking study, the study population only included English students at the 45 English universities contacted. If the mean alcohol consumption under various mental exercises at these universities was systematically different than the mean alcohol consumption under various mental exercises for students in the target population then this difference would be study error.

Suppose in the smoking survey that young adults aged $14 - 20$ living in Ontario at the time of the survey who are attending school were less likely to smoke (people with more education tend to smoke less). In this case the proportion of smokers in the target

population would be different than the proportion of smokers in the study population and this difference would be study error.

Definition 20 *The sampling protocol is the procedure used to select a sample of units from the study population/process. The number of units sampled is called the sample size.*

In Chapter 2, we discussed modeling the data and often claimed that we had a "random sample" so that our model was simple. In practice, it is exceedingly difficult and expensive to select a random sample of units from the study population and so other less rigorous methods are used. Often researchers "take what they can get".

Sample size is usually driven by cost or availability. In Section 4.4 we will see how to use the Binomial model to determine sample sizes, and in Section 4.6 we will see how to use the Gaussian model to determine sample sizes.

In the drinking study, the sampling protocol involved asking administrators at 80 academic departments across 45 English universities to forward a pre-prepared recruitment message to their students containing a URL to an online survey. Departments could decide whether or not to forward the message to their students and students who received the message could decide whether or not to take part in the study. The sample size, as reported in the news item, was 211. Although not indicated in the news item, the journal article indicates that students who agreed to participate were randomly assigned by the researchers to one of the four mental health exercises (imagining positive outcomes of non-drinking during a social occasion; imagining strategies required to successfully not drink during a social occasion; imagining both positive outcomes and required strategies; or completing a drinks diary task). The importance of randomization in making a cause and effect conclusion is discussed in Chapter 8. The students were then asked to report their alcohol consumption in units in the week before they completed the various online surveys over a period of one month.

Definition 21 *If the attributes in the sample differ from the attributes in the study population/process the difference is called sample error.*

Sample error cannot be quantified since the values of the study population/process attributes are unknown. Different random sampling protocols can produce different sample errors. We will see in Chapter 4 how models can be used to get an idea of how large this error might be.

In the university student drinking study, not all academic departments forwarded the recruitment message (only 23 according to the journal article). Suppose only departments who thought students at their university had drinking issues forwarded the message and then only students who were heavy drinkers chose to participate in the study. If the mean alcohol consumption under various mental exercises for students who received the recruitment message and decided to participate was systematically higher than the mean alcohol consumption under various mental exercises for students in the study population

then this difference is sample error. Sample error should be suspected in all surveys in which the participants are volunteers.

The experimenters must decide which variates are going to be measured or determined for the units in the sample. For any attributes of interest, as defined in the Problem step, the corresponding variates must certainly be measured. Other variates which may aid the analysis may also need to be measured. In the smoking survey, experimenters must determine whether each young adult in the sample smokes or not (this requires a careful definition). They may also determine other demographic variates such as age and sex so that they can compare the smoking rate across age groups, sex, etc. In experimental studies, the experimenters assign the value of a variate they are controlling to each unit in the sample. For example, in a clinical trial, sampled units can be assigned to the treatment group or the placebo group by the experimenters.

When the value of a variate is determined for a given unit, errors are often introduced by the measurement system which determines the value.

Definition 22 *If the measured value and the true value of a variate are not identical the difference is called measurement error.*

Measurement errors are unknown since the true value of the variate is unknown. (If we knew the true value we would not need to measure it!) In practice, experimenters try to ensure that the processes used to take the measurements, referred to as the measurement systems, do not contribute substantial error to the conclusions. They may have to study the measurement systems which are used in separate studies to ensure that this is true. See, for example, the case study in Section 3.3.

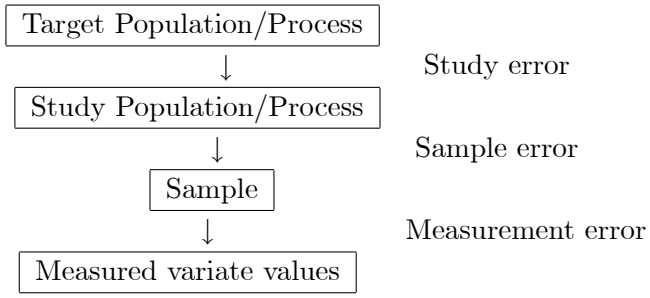
One variate which was determined for each unit (student) in the drinking study was which mental exercise group the student was assigned to. If the actual group assignment was recorded incorrectly then this is measurement error. The students were also asked to report their daily alcohol consumption in the week before they completed the online surveys. The journal article indicated that students measured their daily alcohol consumption in UK units (an alcohol unit in the United Kingdom is defined as 10 milliliters of pure ethyl alcohol) with the help of a visual aid and that the online surveys occurred at the beginning of the study, at two weeks, and at four weeks. Note that a different variate would be associated with each time that the student reported their alcohol consumption. These variates were self-reported by each student. If a student does not accurately report their alcohol consumption then this is measurement error. Measurement error should always be suspected when variates are measured by self-reporting.

Response bias and missing data

Suppose city officials wish to conduct a study to determine if ethnic residents of a city are satisfied with police service in their neighbourhood. A questionnaire is prepared. A random sample of 300 mailing addresses in a predominantly ethnic neighbourhood is chosen

and a uniformed police officer is sent to each address to interview an adult resident. Is there a possible bias in this study? It is likely that those who are strong supporters of the police are quite happy to respond but those with misgivings about the police will either choose not to change some of their responses to favour the police or not respond at all. This type of bias is called *response bias*. When those that do respond have a somewhat different characteristics than the population at large, the quality of the data is threatened, especially when the response rate (the proportion who do respond to the survey) is lower. For example in Canada in 2011, the long form of the Canadian Census (response rate around 98%) was replaced by the *National Household Survey* (a voluntary version with similar questions, response rate around 68%) and there was considerable discussion[9] of the resulting response bias. See for example the CBC story "Census Mourned on World Statistics Day"[10].

The figure below shows the steps in the Plan and the sources of error:



Steps in the plan and sources of error

A person using PPDAC for an empirical study should, by the end of the Plan step, have a good understanding of the study population/process, the sampling protocol, the variates which are to be measured, and the quality of the measurement systems that are intended for use.

In this course you will most often use PPDAC to critically examine a study done by someone else. You should examine each step in the Plan (you may have to ask to see the Plan since many reports omit it) for strengths and weaknesses. You must also pay attention to the various types of error that may occur and how they might impact the conclusions.

Data

The goal of the Data step is to collect the data according to the Plan. Any deviations from the Plan should be noted. The data must be stored in a way that facilitates the Analysis.

[9] http://www.youtube.com/watch?v=0A7ojjsmSsY

[10] http://www.cbc.ca/news/technology/story/2010/10/20/long-form-census-world-statistics-day.html

The previous sections noted the need to define variates clearly and to have satisfactory methods of measuring them. It is difficult to discuss the Data step except in the context of specific examples, but we mention a few relevant points.

- Mistakes can occur in recording or entering data into a data base. For complex investigations, it is useful to put checks in place to avoid these mistakes. For example, if a field is missed, the data base should prompt the data entry person to complete the record if possible.

- In many studies the units must be tracked and measured over a long period of time (e.g. consider a study examining the ability of aspirin to reduce strokes in which persons are followed for 3 to 5 years). This requires careful management.

- When data are recorded over time or in different locations, the time and place for each measurement should be recorded.

- There may be departures from the study Plan that arise over time (e.g. persons may drop out of a long term medical study because of adverse reactions to a treatment; it may take longer than anticipated to collect the data so the number of units sampled must be reduced). Departures from the Plan should be recorded since they may have an important impact on the Analysis and Conclusion.

- In some studies the amount of data may be extremely large, so data base design and management is important.

Analysis

The Analysis step includes both simple and complex calculations to process the data into information. Numerical and graphical methods such as those discussed in Chapter 1, as well as others, are used in this step to summarize the data.

A key component of the Analysis step is the selection of an appropriate model that describes the data and how the data were collected. As indicated in Chapter 1 variates can be of different types: continuous, discrete, categorical, ordinal, and complex. It is important to identify the types of variates collected in a study since this helps in selecting appropriate models. In the Problem step, the problems of interest were stated in terms of the attributes of interest. These attributes need to be described in terms of the parameters and properties of the model. It is also very important to check whether a proposed model is appropriate. Some methods for checking the fit of a model were discussed in Chapter 2. Other methods will be discussed in Chapter 7.

It is difficult to describe this step in more detail except in the context of specific examples. You will see many examples of formal analyses in the following chapters.

In the drinking study the researchers conducted a formal analysis to test for differences between the mean alcohol consumption for the four groups across the three times points. This type of analysis is beyond the scope of this course. However in Chapter 6 we will see how to test for a difference between means when the data consist of two independent groups and the data are assumed to arise from different Gaussian distributions.

Conclusions

The purpose of the Conclusion step is to address the questions posed in the Problem. The conclusions can only be made in relation to the study population. For example, it is not reasonable to make a conclusion about humans if the study population only consisted of laboratory animals. An attempt should also be made to quantify (or at least discuss) potential errors as described in the Plan step. Limitations to the conclusions should be discussed. The conclusions for the drinking study are given below.

Here is a PPDAC for the drinking study based on the published information:

- **Problem** The problem was to study the differences in mean alcohol consumption if different mental exercises related to non-drinking were used. The target population was English university students aged $18-25$ in the United Kingdom at the time of the study. The problem is causative since the researchers wanted to study the effect of the different mental exercises on mean alcohol consumption.

- **Plan** The study population was English university students aged $18-25$ at the time of the study at 45 English universities. The sampling protocol involved asking administrators at 80 academic departments across 45 English universities to forward a pre-prepared recruitment message to their students containing a URL to an online survey. Departments could decide whether or not to forward the message to their students and students who received the message could decide whether or not to take part in the study. The sample size was 211. Students who agreed to participate were randomly assigned by the researchers to one of the four mental health exercises (imagining positive outcomes of non-drinking during a social occasion; imagining strategies required to successfully not drink during a social occasion; imagining both positive outcomes and required strategies; or completing a drinks diary task). The age and sex of each student was also recorded. At the beginning of the study, at two weeks, and at four weeks the students self-reported, using an online survey, how much alcohol they had consumed in the previous week in UK units using a visual aid.

- **Data** The data included which mental exercise group the student was assigned to, their age, their sex, and self-reported information about their alcohol consumption at three different time points.

- **Analysis** The researchers conducted a formal analysis to test for differences between the mean alcohol consumption for the four groups across the three times points.

- **Conclusion** In the drinking study the researchers concluded that completing mental exercises relating to non-drinking was more effective in encouraging safer drinking behaviour than completing a drinks diary alone. The researchers should have indicated that the conclusion only applies to students in the study population not students in the target population and certainly not students in other countries. This is an experimental study since the researchers determined group assignment for each student by randomization. There are several serious drawbacks in this study. Students were not recruited from all English universities. This could lead to study error. Also not all contacted departments forwarded the recruitment message and participants were volunteers. Both of these issues could lead to sample error. Alcohol consumption was self-reported which could lead to measurement error.

3.3 Case Study

Introduction

This case study is an example of more than one use of PPDAC which demonstrates some real problems that arise with measurement systems. The documentation given here has been rewritten from the original report to emphasize the underlying PPDAC framework.

Background

An automatic in-line gauge measures the diameter of a crankshaft journal on 100% of the 500 parts produced per shift. The measurement system does not involve an operator directly except for calibration and maintenance. Figure 3.1 shows the diameter in question.

The journal is a "cylindrical" part of the crankshaft. The diameter of the journal must be defined since the cross-section of the journal is not perfectly round and there may be taper along the axis of the cylinder. The gauge measures the maximum diameter as the crankshaft is rotated at a fixed distance from the end of the cylinder.

The specification for the diameter is -10 to $+10$ units with a target of 0. The measurements are re-scaled automatically by the gauge to make it easier to see deviations from the target. If the measured diameter is less than -10, the crankshaft is scrapped and a cost is incurred. If the diameter exceeds $+10$, the crankshaft can be reworked, again at considerable cost. Otherwise, the crankshaft is judged acceptable.

Overall Project

A project is planned by a crankshaft manufacturer to reduce scrap/rework by reducing part-to-part variation in the diameter. A first step involves an investigation of the measurement system itself. There is some speculation that the measurement system contributes substantially to the overall process variation and that bias in the measurement system is resulting in the scrapping and reworking of good parts. To decide if the measurement

Figure 3.1: **Crankshaft with arrow pointing to "journal"**

system is making a substantial contribution to the overall process variability, we also need a measure of this attribute for the current and future population of crankshafts. Since there are three different attributes of interest, it is convenient to split the project into three separate applications of PPDAC.

Study 1

In this application of PPDAC, we estimate the properties of the errors produced by the measurement system. In terms of the model, we will estimate the bias and variability due to the measurement system. We hope that these estimates can be used to predict the future performance of the system.

Problem

The target process is all future measurements made by the gauge on crankshafts to be produced by the manufacturer. The *response variate* is the measured diameter associated with each unit. The attributes of interest are the average measurement error and the population standard deviation of these errors. We can quantify these concepts using a model (see below). A detailed *fishbone diagram* for the measurement system is also shown in Figure 3.2. In such a diagram, we list *explanatory variates* organized by the major "bones" that might be responsible for variation in the response variate, here the measured journal diameter. We can use the diagram in formulating the Plan.

Note that the measurement system includes the gauge itself, the way the part is loaded into the gauge, who loads the part, the calibration procedure (every two hours, a master

part is put through the gauge and adjustments are made based on the measured diameter of the master part; that is "the gauge is zeroed"), and so on.

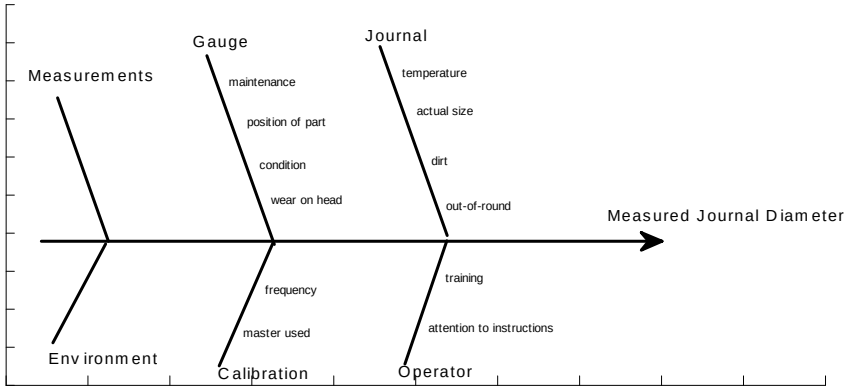


Figure 3.2: **Fishbone diagram for variation in measured journal diameter**

Plan

To determine the properties of the measurement errors we must measure crankshafts with known diameters. "Known" implies that the diameters were measured by an off-line measurement system that is very reliable. For any measurement system study in which bias is an issue, there must be a reference measurement system which is known to have negligible bias and variability which is much smaller than the system under study.

There are many issues in establishing a study process or a study population. For convenience, we want to conduct the study quickly using only a few parts. However, this restriction may lead to study error if the bias and variability of the measurement system change as other explanatory variates change over time or parts. We guard against this latter possibility by using three crankshafts with known diameters as part of the definition of the study process. Since the units are the taking of measurements, we define the study population as all measurements that can be taken in one day on the three selected crankshafts. These crankshafts were selected so that the known diameters were spread out over the range of diameters normally observed in the production of crankshafts. This will allow us see if the attributes of the system depend on the size of the diameter being measured. The known diameters which were used were: -10, 0, and $+10$. Remember the diameters have been rescaled so that a diameter of -10 is okay.

No other explanatory variates were measured. To define the sampling protocol, it was proposed to measure the three crankshafts ten times each in a random order. Each measurement involved the loading of the crankshaft into the gauge. Note that this was to be done quickly to avoid delay of production of the crankshafts. The whole procedure took

only a few minutes.

The preparation for the data collection was very simple. One operator was instructed to follow the sampling protocol and write down the measured diameters in the order that they were collected.

Data

The repeated measurements on the three crankshafts are shown below. Note that due to poor explanation of the sampling protocol, the operator measured each part ten times in a row and did not use a random ordering. (Unfortunately non-adherence to the sampling protocol often happens when real data are collected and it is important to consider the effects of this in the Analysis and Conclusion steps.)

Crankshaft 1		Crankshaft 2		Crankshaft 3	
−10	−8	2	1	9	11
−12	−12	−2	2	8	12
−8	−10	0	1	10	9
−11	−10	1	1	12	10
−12	−10	0	0	10	12

Analysis

A model to describe the repeated measurement of the known diameters is

$$Y_{ij} = \mu_i + R_{ij}, \quad R_{ij} \sim G(0, \sigma_m) \quad \text{independent} \tag{3.1}$$

where $i = 1$ to 3 indexes the three crankshafts and $j = 1, 2, \ldots, 10$ indexes the ten repeated measurements. The parameter μ_i represents the long term average measurement for crankshaft i in the study population. The random variables R_{ij} (called the *residuals*) represent the variability of the measurement system, while σ_m quantifies this variability. Note that we have assumed, for simplicity, that the variability σ_m is the same for all three crankshafts in the study.

We can rewrite the model in terms of the random variables Y_{ij} so that $Y_{ij} \sim G(\mu_i, \sigma_m)$. Now we can write the likelihood as in Example 2.3.2 and maximize it with respect to the four parameters μ_1, μ_2, μ_3, and σ_m (the trick is to solve $\partial\ell/\partial\mu_i = 0$, $i = 1, 2, 3$ first). Not surprisingly the maximum likelihood estimates for μ_1, μ_2, μ_3 are the sample averages for each crankshaft so that

$$\hat{\mu}_i = \bar{y}_i = \frac{1}{10} \sum_{j=1}^{n} y_{ij} \quad \text{for } i = 1, 2, 3$$

To examine the assumption that σ_m is the same for all three crankshafts we can calculate the sample standard deviation for each of the three crankshafts. Let

$$s_i = \sqrt{\frac{1}{9} \sum_{j=1}^{10} (y_{ij} - \bar{y}_i)^2} \quad \text{for } i = 1, 2, 3$$

The data can be summarized as:

	$\bar{y}_i$	s_i
Crankshaft 1	-10.3	1.49
Crankshaft 2	0.6	1.17
Crankshaft 3	10.3	1.42

The estimate of the bias for crankshaft 1 is the difference between the observed average $\bar{y}_1$ and the known diameter value which is equal to -10 for crankshaft 1, that is, the estimated bias is $-10.3-(-10)=-0.3$. For crankshafts 2 and 3 the estimated biases are $0.6-0=0.6$ and $10.3-10=0.3$ respectively so the estimated biases in this study are all small.

Note that the sample standard deviations s_1, s_2, s_3 are all about the same size and our assumption about a common value seems reasonable. (Note: it is possible to test this assumption more formally.) An estimate of σ_m is given by

$$s_m=\sqrt{\frac{s_1^2+s_2^2+s_3^2}{3}}=1.37$$

Note that this estimate is not the average of the three sample standard deviations but the square root of the average of the three sample variances. (Why does this estimate make sense? Is it the maximum likelihood estimate of σ_m? What if the number of measurements for each crankshaft were not equal?)

Conclusion

The observed biases -0.3, 0.6, 0.3 appear to be small, especially when measured against the estimate of σ_m and there is no apparent dependence of bias on crankshaft diameter.

To interpret the variability, we can use the model (3.1). Recall that if $Y_{ij}\sim G(\mu_i,\sigma_m)$ then

$$P(\mu_i-2\sigma_m\le Y_{ij}\le\mu_i+2\sigma_m)=0.95$$

Therefore if we repeatedly measure the same journal diameter, then about 95% of the time we would expect to see the observations vary by about $\pm2(1.37)=\pm2.74$ in the study population.

There are several limitations to these conclusions. Because we have carried out the study on one day only and used only three crankshafts, the conclusion may not apply to all future measurements (study error). The fact that the measurements were taken within a few minutes on one day might be misleading if something special was happening at that time (sample error). Since the measurements were not taken in random order, another source of sample error is the possible drift of the gauge over time.

We could recommend that, if the study were to be repeated, more than three known-value crankshafts could be used, that the time frame for taking the measurements could be extended and that more measurements be taken on each crankshaft. Of course, we would

also note that these recommendations would add to the cost and complexity of the study. We would also insist that the operator be better informed about the Plan.

Study 2

The second study is designed to estimate the overall population standard deviation of the diameters of current and future crankshafts (the target process). We need to estimate this attribute to determine what variation is due to the process and what is due to the measurement system. A cause-and-effect or fishbone diagram listing some possible explanatory variates for the variability in journal diameter is given in Figure 3.3. Note that there are many explanatory variates other than the measurement system. Variability in the response variate is induced by changes in the explanatory variates, including those associated with the measurement system.

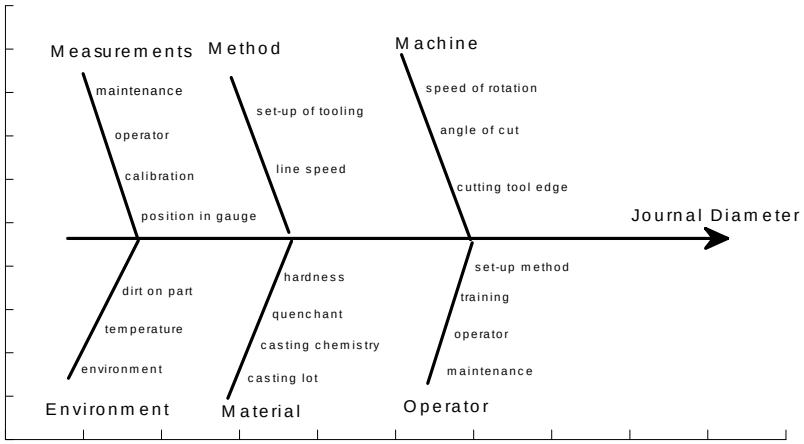


Figure 3.3: **Fishbbone diagram for cause-and-effect**

Plan

The study population is defined as those crankshafts available over the next week, about 7500 parts (500 per shift times 15 shifts). No other explanatory variates were measured.

Initially it was proposed to select a sample of 150 parts over the week (10 from each shift). However, when it was learned that the gauge software stores the measurements for the most recent 2000 crankshafts measured, it was decided to select a point in time near the end of the week and use the 2000 measured values from the gauge memory to be the sample. One could easily criticize this choice (sample error), but the data were easily available and inexpensive.

Data

The individual observed measurements are too numerous to list but a histogram of the data is shown in Figure 3.4. From this, we can see that the measured diameters vary from -14 to $+16$.

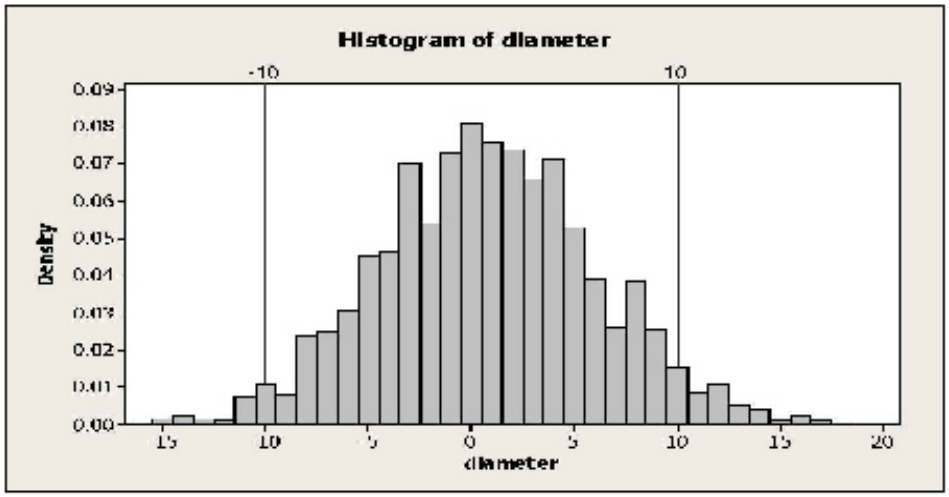


Figure 3.4: **Histogram of 2000 measured values from the gauge memory**

Analysis

A model for these data is given by

$$Y_i = \mu + R_i, \quad R_i \sim G(0, \sigma) \quad \text{independently for } i = 1, 2, \ldots, 2000$$

where Y_i represents the distribution of the measurement of the ith diameter, μ corresponds to the study population mean diameter and the residual R_i represents the variability due to sampling and the measurement system. We let σ quantify this variability. We have not included a bias term in the model because we assume, based on our results from Study 1, that the measurement system bias is small. As well we assume that the sampling protocol does not contribute substantial bias.

The histogram of the 2000 measured diameters shows that there is considerable spread in the measured diameters. About 4.2% of the parts require reworking and 1.8% are scrapped. The shape of the histogram is approximately symmetrical and centred close to zero. The sample mean is

$$\bar{y} = \frac{1}{2000} \sum_{i=1}^{2000} y_i = 0.82$$

which gives us an estimate of μ (the maximum likelihood estimate) and the sample standard

deviation is

$$s = \sqrt{\frac{1}{1999}\sum_{i=1}^{2000}(y_i - \bar{y})^2} = 5.17$$

which gives us an estimate of σ (not quite the maximum likelihood estimate).

Conclusion

The overall process variation is estimated by s. Since the sample contained 2000 parts measured consecutively, many of the explanatory variates did not have time to change as they would in the study population. Thus, there is a danger of sample error producing an estimate of the variation that is too small.

The variability due to the measurement system, estimated to be 1.37 in Study 1, is much less than the overall variability which is estimated to be 5.17. One way to compare the two standard deviations σ_m and σ is to separate the total variability σ into the variability due to the measurement system σ_m and that due to all other sources. In other words, we are interested in estimating the variability that would be present if there were no variability in the measurement system ($\sigma_m = 0$). If we assume that the total variability arises from two independent sources, the measurement system and all other sources, then we have $\sigma^2 = \sigma_m^2 + \sigma_p^2$ or $\sigma_p = \sqrt{\sigma^2 - \sigma_m^2}$ where σ_p quantifies the variability due to all other uncontrollable variates (sampling variability). An estimate of σ_p is given by

$$\sqrt{s^2 - s_m^2} = \sqrt{(5.17)^2 - (1.37)^2} = 4.99$$

Hence, eliminating all of the variability due to the measurement system would produce an estimated variability of 4.99 which is a small reduction from 5.17. The measurement system seems to be performing well and not contributing substantially to the overall variation.

Study 3: A Brief Description

A limitation of Study 1 was that it was conducted over a very short time period. To address this concern, a third study was recommended to study the measurement system over a longer period during normal production use. In Study 3, a master crankshaft of known diameter equal to zero was measured every half hour until 30 measurements were collected. A plot of the measurements versus the times at which the measurements were taken is given in the run chart in Figure 3.5.

In the first study the standard deviation was estimated to be 1.37. In a sample of observations from a $G(0, 1.37)$ distribution we would expect approximately 95% of the observations to lie in the interval $[0 - 2(1.37),\ 0 + 2(1.37)] = [-2.74,\ 2.74]$ which is obviously not true for the data displayed in the run chart. These data have a much larger variability. This was a shocking result for the people in charge of the process.

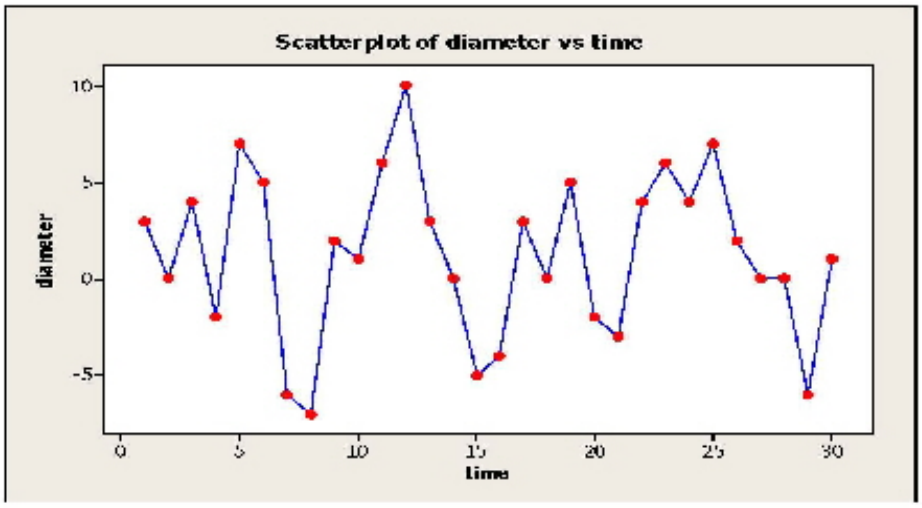


Figure 3.5: **Scatter plot of diameter versus time**

Comments

Study 3 revealed that the measurement system had a serious long term problem. At first, it was suspected that the cause of the variability was the fact that the gauge was not calibrated over the course of the study. Study 3 was repeated with a calibration before each measurement. A pattern similar to that for Study 3 was seen. A detailed examination of the gauge by a repairperson from the manufacturer revealed that one of the electronic components was not working properly. This was repaired and Study 3 was repeated. This study showed variation similar to the variation of the short term study (Study 1) so that the overall project could continue. When Study 2 was repeated, the overall variation and the number of scrap and reworked crankshafts was substantially reduced. The project was considered complete and long term monitoring showed that the scrap rate was reduced to about 0.7% which produced an annual savings of more than $\$100,000$.

As well, three similar gauges that were used in the factory were put through the "long term" test. All were working well.

Summary

- An important part of any Plan is the choice and assessment of the measurement system.
- The measurement system may contribute substantial error that can result in poor decisions (e.g. scrapping good parts, accepting bad parts).
- We represent systematic measurement error by bias in the model. The bias can be assessed only by measuring units with known values, taken from another reference

measurement system. The bias may be constant or depend on the size of the unit being measured, the person making the measurements, and so on.

- Variability can be assessed by repeatedly measuring the same unit. The variability may depend on the unit being measured or any other explanatory variates.
- Both bias and variability may be a function of time. This can be assessed by examining these attributes over a sufficiently long time span as in Study 3.

3.4 Chapter 3 Problems

1. Answer the questions below based on the following:
   A Waterloo-based public opinion research firm was hired by the Ontario Ministry of Education to investigate whether the financial worries of Ontario university students varied by sex. To reduce costs, the research firm decided to study only university students living in the Kitchener-Waterloo region in September 2012. An associate with the research firm randomly selected 250 university students attending a Laurier-Waterloo football game. The students were asked whether they agreed/disagreed with the statement "I have significant trouble paying my bills." Their sex was also recorded. The results are given below:

	Agreed	Disagreed	Total
Male	68	77	145
Female	42	63	105
Total	110	140	250

   (a) What are the units?

   (b) Define the target population.

   (c) Define the study population.

   (d) What are two variates in this problem and what is their type?

   (e) What is the sampling protocol?

   (f) What is a possible source of study error?

   (g) What is a possible source of sample error?

   (h) Describe an attribute of interest for the target population and provide an estimate based on the given data.

2. Four weeks before a national election, a political party conducts a poll to assess what proportion of eligible voters plan to vote and, of those, what proportion support the party. This will determine how they run the rest of the campaign. They are able to obtain a list of eligible voters and their telephone numbers in the 20 most populated areas. They select 3000 names from the list and call them. Of these, 1104 eligible voters agree to participate in the survey with the results summarized in the table below. Answer the questions below based on this information.

	Support	Party
Plan to Vote	YES	NO
YES	351	381
NO	107	265

   (a) Define the Problem for this study. What type of Problem is this and why?

(b) What is the target population?

(c) Identify the variates and their types for this study.

(d) What are the attributes of interest in the target population?

(e) What is the study population?

(f) What is the sample?

(g) What is a possible source of study error?

(h) Describe one possible source of sample error.

(i) Estimate the attributes of interest for the study population based on the given data.

3. **Online brain-training: does it really work?** Brain training, or the goal of improved cognitive function through the regular use of computerized tests, is a multi-million dollar industry. Lumosity, which is one of the most popular cognitive training programs, is made up of more than 40 games designed to improve cognitive abilities, including memory, attention and problem solving. Members pay a monthly membership and are supposed to play the games for 15 minutes, three to five times a week. A 2007 press release from the company calls the games "a scientifically developed online brain fitness program demonstrated to improve memory and attention with fun and effective brain workouts."

   To investigate whether regular brain training leads to any improvement in cognitive function, researchers in Britain, lead by neuroscientist Adrian Owen, conducted a study in 2010. Viewers of the BBC popular science programme 'Bang Goes The Theory' were invited to participate in a six-week online study of brain training. Of 52,617 participants aged 18–60 who initially registered, 11,430 completed both benchmarking assessments and at least two full training sessions during the six-week period. An initial 'benchmarking' assessment included a broad neuropsychological battery of four tests that are sensitive to changes in cognitive function in health and disease. Specifically, baseline scores for reasoning, verbal short-term memory, spatial working memory and paired-associates learning were acquired. Participants were then randomly assigned to one of two experimental groups or a third control group and logged on to the BBC Lab UK website to practise six training tasks for a minimum of 10 minutes a day, three times a week. In experimental group 1, the six training tasks emphasized reasoning, planning and problem-solving abilities. In experimental group 2, a broader range of cognitive functions was trained using tests of short-term memory, attention, visuospatial processing and mathematics similar to those commonly found in commercially available brain-training devices. The difficulty of the training tasks increased as the participants improved to continuously challenge their cognitive performance and maximize any benefits of training. The control group did not formally practise any specific cognitive tasks during their 'training' sessions, but answered obscure questions from six different categories using any available online

resource. At six weeks, the benchmarking assessment was repeated and the pre- and post-training scores were compared. The difference in benchmarking scores provided the measure of generalized cognitive improvement resulting from training. Similarly, for each training task, the first and last scores were compared to give a measure of specific improvement on that task.
The relationship between the number of training sessions and changes in benchmark performance was negligible in all groups for all tests. These results provide no evidence for any generalized improvements in cognitive function following brain training in a large sample of healthy adults.

Answer the questions below based on the information provided.

(a) What type of study is this? Why?

(b) Define the Problem for this study.

(c) What type of Problem is it? Why?

(d) Define a suitable target population for this study.

(e) What variates were collected in this study? Specify the type of variate for each.

(f) What are the attributes of interest in this study?

(g) Define a suitable study population for this study.

(h) What is the sampling protocol?

(i) What is a possible source of study error is?

(j) What is a possible source of sample error?

(k) What is a possible source of measurement error?

(l) Why was it important for the researchers to randomly assign the participants to the three different groups?

(m) What is the importance of the control group?

(n) Use the article and your answers to the above questions to construct a PPDAC for this empirical study in as much detail as possible.

4. **U.S. to fund study of Ontario math curriculum,** Globe & Mail, January 17, 2014, Caroline Alphonso - Education Reporter (article has been condensed)
The U.S. Department of Education has funded a $2.7-million (U.S.) project, led by a team of Canadian researchers at Toronto's Hospital for Sick Children. The study will look at how elementary students at several Ontario public schools fare in math using the current provincial curriculum as compared to the JUMP math program, which combines the conventional way of learning the subject with so-called discovery learning. Math teaching has come under scrutiny since OECD results that measured the scholastic abilities of 15-year-olds in 65 countries showed an increasing percentage of Canadian students failing the math test in nearly all provinces. Dr. Tracy Solomon

and her team are collecting and analyzing two years of data on students in primary and junior grades from one school board, which she declined to name. The students were in Grades 2 and 5 when the study began, and are now in Grades 3 and 6, which means they will participate in Ontario's standardized testing program this year. The research team randomly assigned some schools to teach math according to the Ontario curriculum, which allows open-ended student investigations and problem-solving. The other schools are using the JUMP program. Dr. Solomon said the research team is using classroom testing data, lab tests on how children learn and other measures to study the impact of the two programs on student learning.

Answer the questions below based on this article.

(a) What type of study is this? Why?

(b) Define the Problem for this study.

(c) What type of Problem is it? Why?

(d) Define a suitable target population for this study.

(e) What variates were collected in this study? Specify the type of variate for each.

(f) What are the attributes of interest in this study?

(g) Define a suitable study population for this study.

(h) What is the sampling protocol?

(i) What is a possible source of study error?

(j) What is a possible source of sample error?

(k) What is a possible source of measurement error?

(l) Why was it important for the researchers to randomly assign some schools to teach math according to the Ontario curriculum and some other schools to teach math using the Jump program?

(m) Use the information in the article and your answers to the above questions to construct a PPDAC for this empirical study in as much detail as possible.

5. **Playing racing games may encourage risky driving, study finds**, Globe & Mail, January 8, 2015 (article has been condensed)
Playing an intense racing game makes players more likely to take risks such as speeding, passing on the wrong side, running red lights or using a cellphone in a simulated driving task shortly afterwards, according to a new study. Young adults with more adventurous personalities were more inclined to take risks, and more intense games led to greater risk-taking, the authors write in the journal Injury Prevention. Other research has found a connection between racing games and inclination to risk-taking while driving, so the new results broaden that evidence base, said lead author of the new study, Mingming Deng of the School of Management at Xi'an Jiaotong University

in Xi'an, China. "I think racing gamers should be [paying] more attention in their real driving," Deng said.
The researchers recruited 40 student volunteers at Xi'an Jiaotong University, mostly men, for the study. The students took personality tests at the start and were divided randomly into two groups. Half of the students played a circuit-racing-type driving game that included time trials on a race course similar to Formula 1 racing, for about 20 minutes, while the other group played computer solitaire, a neutral game for comparison. After a five-minute break, all the students took the Vienna Risk-Taking Test, viewing 24 "risky" videotaped road-traffic situations on a computer screen presented from the driver's perspective, including driving up to a railway crossing whose gate has already started lowering. How long the viewer waits to hit the "stop" key for the manoeuvre is considered a measure of their willingness to take risks on the road. Students who had been playing the racing game waited an average of almost 12 seconds to hit the stop button compared with 10 seconds for the solitaire group. The participants' experience playing these types of games outside of the study did not seem to make a difference.

Answer the questions below based on this article.

(a) What type of study is this? Why?

(b) Define the Problem for this study.

(c) What type of Problem is this? Why?

(d) Define a suitable target population for this study.

(e) What variates were collected in this study? Specify the type of variate for each.

(f) What are the attributes of interest in this study?

(g) Define a suitable study population for this study.

(h) Describe the sampling protocol for this study.

(i) Give a possible source of study error for this study in relation to your answer to (d).

(j) Give a possible source of sample error for this study.

(k) Estimate the attribute of interest for the study population based on the given data.

(l) Use the information in the article and your answers to the above questions to construct a PPDAC for this empirical study in as much detail as possible.

6. **Higher coffee consumption associated with lower risk of early death**, European Society of Cardiology, August 27, 2017
Higher coffee consumption is associated with a lower risk of death, according to research presented today at ESC Congress. The study in nearly $20,000$ participants suggests that coffee can be part of a healthy diet in healthy people. "Coffee is one

of the most widely consumed beverages around the world," said Dr Adela Navarro, a cardiologist at Hospital de Navarra, Pamplona, Spain. "Previous studies have suggested that drinking coffee might be inversely associated with all-cause mortality but this has not been investigated in a Mediterranean country."

The purpose of this study was to examine the association between coffee consumption and the risk of mortality (death) in a middle-aged Mediterranean cohort. The study was conducted within the framework of the Seguimiento Universidad de Navarra (SUN) Project, a long-term prospective cohort study of Spanish university graduates which began in 1999 and which has recruited new Spanish university graduates to the study every year since then. This analysis included 19,896 participants of the SUN Project, whose average age at enrollment was 37.7 years old. On entering the study, participants completed a previously validated semi-quantitative food frequency questionnaire to collect information on coffee consumption, lifestyle and sociodemographic characteristics, and previous health conditions.

Patients were followed-up for an average of ten years. Information on mortality was obtained from study participants and their families, postal authorities, and the National Death Index. During the ten year period, 337 participants died. The researchers found that participants who consumed at least four cups of coffee per day had a 64% lower risk of all-cause mortality than those who never or almost never consumed coffee. In those who were at least 45 years old, drinking two additional cups of coffee per day was associated with a 30% lower risk of mortality during follow-up. The association was not significant among younger participants.

Dr Navarro said: "In the SUN project we found an inverse association between drinking coffee and the risk of all-cause mortality, particularly in people aged 45 years and above. This may be due to a stronger protective association among older participants." She concluded: "Our findings suggest that drinking four cups of coffee each day can be part of a healthy diet in healthy people."

(a) What type of study is this and why?

(b) Define the Problem for this study.

(c) What type of Problem is this? Why?

(d) Define a suitable target population/process for this study.

(e) What variates were collected in this study? Specify the type of variate for each.

(f) What are the attributes of interest in this study?

(g) Define a suitable study population/process for this study.

(h) Define study error and give a possible source of study error for this study in relation to your answers to (d) and (e).

(i) Define measurement error and give a possible source of measurement error for one of the two variates you gave in (c).

(j) Give at least one limitation to this study.

(k) Use the information in the article and your answers to the above questions to construct a PPDAC for this empirical study in as much detail as possible.

(l) Suppose you are not a coffee drinker. **On the basis of this study** do you think it would be a good idea to start drinking four cups of coffee a day. Why or why not?

7. Answer the following questions based on the study given in Chapter 1, Problem 26.

(a) Define the Problem for this study in one or two sentences.

(b) What type of Problem is this? Explain why.

(c) Define a suitable target population for this study.

(d) Define a suitable study population for this study.

(e) Describe possible sources of study error for this study.

(f) Describe the sampling protocol for this study in as much detail as possible.

(g) What is the sample and sample size for this study?

(h) Describe possible sources of sample error for this study.

(i) Describe possible sources of measurement error for this study.

(j) What is the most serious limitation to the conclusion(s) of this study?

(k) Use the information in the article and your answers to the above questions to construct a PPDAC for this empirical study in as much detail as possible.

8. Suppose you wish to study the smoking habits of teenagers and young adults, in order to understand what personal factors are related to whether, and how much, a person smokes. Briefly describe the main components of such a study, using the PPDAC framework. Be specific about the target and study population, the sample, and the variates you would collect.

9. Suppose you wanted to study the relationship between a person's "resting" pulse rate (heart beats per minute) and the amount and type of exercise they get.

(a) List some factors (including exercise) that might affect resting pulse rate. You may wish to draw a cause and effect (fishbone) diagram to represent potential causal factors.

(b) Describe briefly how you might study the relationship between pulse rate and exercise using (i) an observational study, and (ii) an experimental study.

10. A large company uses photocopiers leased from two suppliers A and B. The lease rates are slightly lower for B's machines but there is a perception among workers

that they break down and cause disruptions in work flow substantially more often. Describe briefly how you might design and carry out a study of this issue, with the ultimate objective being a decision whether to continue the lease with company B. What additional factors might affect this decision?

11. For a study like the one in Example 1.3.2, where heights x and weights y of individuals are to be recorded, discuss sources of variability due to the measurement of x and y on any individual.

4. ESTIMATION

4.1 Statistical Models and Estimation

In statistical estimation we use two models:

(1) A model which describes the variability in the variate(s) of interest in the population or process being studied.

(2) A model which takes in to account how the data were collected and which is constructed in conjunction with the model in (1).

We use these two models to estimate the unknown attributes in the population or process based on the observed data and to determine the uncertainty in these estimates. The unknown attributes are usually represented by unknown parameters in the models or by functions of the unknown parameters. We have already seen in Chapter 2, that these unknown parameters can be estimated using the method of maximum likelihood and the invariance property of maximum likelihood estimates.

Several issues arise:

(1) Where do we get our probability model? What if it is not a good description of the population or process?

We discussed the first question in Chapters 1 and 2. It is important to check the adequacy (or "fit") of the model; some ways of doing this were discussed in Chapter 2 and more formal methods will be considered in Chapter 7. If the model used is **not** satisfactory, it is not wise to use the estimates based on it. For the lifetimes of brake pads data introduced in Example 1.3.4, a Gaussian model did not fit the data well. Sometimes the data can be transformed in such a way that the Gaussian model does fit (see Chapter 2, Problem 19).

(2) The estimation of parameters or population attributes depends on data collected from the population or process, and the likelihood function is based on the probability of the observed data. This implies that factors associated with the selection of sample units or the measurement of variates (e.g. measurement error) must be included in the model. In many examples it is assumed that the variate of interest is measured without error for a random sample of units from the population. We will typically

assume that the data come from a random sample of population units, but in any given application we would need to design the data collection plan to ensure this assumption is valid.

(3) Suppose in the model chosen the population mean is represented by the parameter θ. The sample mean $\overline{y}$ is an estimate of θ, but not usually equal to it. How far away from θ is $\overline{y}$ likely to be? If we take a sample of only $n = 50$ units, would we expect the estimate $\overline{y}$ to be as "good" as $\bar{y}$ based on 150 units? What does "good" mean?

We focus on the third point in this chapter and assume that we can deal with the first two points with the methods discussed in Chapters 1 and 2.

4.2 Estimators and Sampling Distributions

Suppose that some attribute of interest for a population or process can be represented by a parameter θ in a statistical model. We assume that θ can be estimated using a random sample drawn from the population or process in question. Recall in Chapter 2 that a *point estimate* of θ, denoted as $\hat{\theta}$, was defined as a function of the observed sample $y_1, y_2, \ldots, y_n$, $\hat{\theta} = g(y_1, y_2, \ldots, y_n)$. For example

$$\hat{\theta} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is a point estimate of θ if $y_1, y_2, \ldots, y_n$ is an observed random sample from a Poisson distribution with mean θ.

The method of maximum likelihood provides a general method for obtaining estimates, but other methods exist. For example, if $\theta = E(Y) = \mu$ is the average (mean) value of y in the population, then the sample mean $\hat{\theta} = \overline{y}$ is an intuitively sensible estimate; it is the maximum likelihood estimate of θ if Y has a $G(\theta, \sigma)$ distribution but because of the Central Limit Theorem it is a good estimate of θ more generally. Thus, while we will use maximum likelihood estimation a great deal, you should remember that the discussion below applies to estimates of any type.

The problem facing us in this chapter is how to determine or quantify the uncertainty in an estimate. We do this using *sampling distributions*, which are based on the following idea. If we select random samples on repeated occasions, then the estimates $\hat{\theta}$ obtained from the different samples will vary. For example, five separate random samples of $n = 50$ persons from the same male population described in Example 1.3.2 gave five different estimates $\hat{\theta} = \overline{y}$ of $E(Y)$ as:

$$1.723 \quad 1.743 \quad 1.734 \quad 1.752 \quad 1.736$$

Estimates vary as we take repeated samples and therefore we associate a random variable and a distribution with these estimates.

More precisely, we define this idea as follows. Let the random variables $Y_1, Y_2, \ldots, Y_n$ represent potential observations in an empirical study. Associate with the estimate

$\hat{\theta} = g(y_1, y_2, \ldots, y_n)$ a random variable $\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$. The random variable $\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ is simply a rule that tells us how to process the data to obtain a numerical value $\hat{\theta} = g(y_1, y_2, \ldots, y_n)$ which is an estimate of the unknown parameter θ for a given data set $y_1, y_2, \ldots, y_n$. For example

$$\tilde{\theta} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is a random variable and $\hat{\theta} = \overline{y}$ is a numerical value. We call $\tilde{\theta}$ the *estimator* of θ corresponding to $\hat{\theta}$. We use $\hat{\theta}$ to denote an estimate, that is, a numerical value, and $\tilde{\theta}$ to denote the corresponding estimator, the random variable.

Definition 23 *A (point) estimator $\tilde{\theta}$ is a random variable which is a function $\tilde{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ of the random variables $Y_1, Y_2, \ldots, Y_n$. The distribution of $\tilde{\theta}$ is called the sampling distribution of the estimator.*

Since $\tilde{\theta}$ is a function of the random variables $Y_1, Y_2, \ldots, Y_n$ then this means that $\tilde{\theta}$ is also a random variable. If we know the distribution of $Y_1, Y_2, \ldots, Y_n$ then we can find the sampling distribution of $\tilde{\theta}$, at least in principle. In other words we can find the probability (density) function or cumulative distribution function of $\tilde{\theta}$ and use it to make probabilities statements about $\tilde{\theta}$. If we know the sampling distribution of the estimator $\tilde{\theta}$ then we can use it to quantify the uncertainty in an estimate $\hat{\theta}$, that is, we can determine *the probability that the estimator $\tilde{\theta}$ is "close" to the true but unknown value of θ.* In Examples 4.2.1 - 4.2.3 we examine ways of finding the sampling distribution, at least approximately.

Example 4.2.1

Suppose we have a variate of interest (for example, the height in meters of a male in the population of Example 1.3.2) whose distribution it is reasonable to model as a $G(\mu, \sigma)$ random variable. Suppose also that we plan to take a random sample $Y_1, Y_2, \ldots, Y_n$ to estimate the unknown mean μ where $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$. The maximum likelihood estimator of μ is

$$\tilde{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

From properties of Gaussian random variables (See Chapter 1, Problem 18) we know that $\tilde{\mu} = \overline{Y} \sim G(\mu, \sigma/\sqrt{n})$ and so the sampling distribution of $\overline{Y}$ is $G(\mu, \sigma/\sqrt{n})$.

If we knew σ we could determine how often the estimator $\tilde{\mu} = \overline{Y}$ is within a specified amount of the unknown mean μ. For example, if the variate is height and heights are measured in meters then we could determine how often the estimator $\tilde{\mu} = \overline{Y}$ is within 0.01 meters of the true mean μ as follows:

$$\begin{aligned} P\left(|\tilde{\mu} - \mu| \leq 0.01\right) &= P\left(\mu - 0.01 \leq \overline{Y} \leq \mu + 0.01\right) = P\left(\frac{-0.01}{\sigma/\sqrt{n}} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq \frac{0.01}{\sigma/\sqrt{n}}\right) \\ &= P\left(-0.01\sqrt{n}/\sigma \leq Z \leq 0.01\sqrt{n}/\sigma\right) \quad \text{where } Z \sim G(0,1) \end{aligned}$$

Suppose $\sigma = 0.07$ meters and $n = 50$ then

$$\begin{aligned} P(|\tilde{\mu} - \mu| &\leq 0.01) = P\left(-0.01\sqrt{50}/0.07 \leq Z \leq 0.01\sqrt{50}/0.07\right) = P(-1.01 \leq Z \leq 1.01) \\ &= 0.688 \end{aligned}$$

and if $n = 100$

$$\begin{aligned} P(|\tilde{\mu} - \mu| &\leq 0.01) = P\left(-0.01\sqrt{100}/0.07 \leq Z \leq 0.01\sqrt{100}/0.07\right) = P(-1.43 \leq Z \leq 1.43) \\ &= 0.847 \end{aligned}$$

This illustrates the rather intuitive fact that, the larger the sample size, the higher the probability that the estimator $\tilde{\mu} = \overline{Y}$ is within 0.01 meters of the true but unknown mean height μ in the study population. It also allows us to express the uncertainty in an estimate $\hat{\mu} = \bar{y}$ from an observed sample $y_1, y_2, \ldots, y_n$ by indicating the probability that any single random sample will give an estimate within a certain distance of μ.

Example 4.2.2

In Example 4.2.1 the distribution of the estimator $\tilde{\mu} = \overline{Y}$ could be determined exactly. Sometimes the distribution of the estimator can only be determined approximately using the Central Limit Theorem. For example, for Binomial data with n trials and y successes the estimator $\tilde{\theta} = Y/n$ has $E(\tilde{\theta}) = \theta$ and $Var(\tilde{\theta}) = \theta(1-\theta)/n$. By the Normal approximation to the Binomial we have

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim G(0,1) \quad \text{approximately}$$

This result could be used, for example, to determine how large n should be to ensure that, with probability at least 0.95, the estimator $\tilde{\theta}$ is within 0.03 of the true but unknown proportion θ in the study population. In other words, this result could be used to determine how large n should be to ensure that

$$P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) \geq 0.95$$

for all $\theta \in [0,1]$. This problem is consided in Example 4.4.4.

In some cases the sampling distribution can be approximated using a simulation study as illustrated in the next example.

Example 4.2.3

Suppose the population of interest is a finite population consisting of 500 units. Suppose associated with each unit is a number between 1 and 10 which is the variate of interest. If we wanted to estimate the mean μ of this population we could select a random sample $y_1, y_2, \ldots, y_n$ without replacement and estimate μ using the estimate $\hat{\mu} = \bar{y}$. Let us examine

how good the estimator $\tilde{\mu} = \overline{Y}$ is in the case of the population which has the distribution of variate values as indicated in Table 4.1.

Variate value	1	2	3	4	5	6	7	8	9	10	Total
No. of units	210	127	66	39	23	13	11	7	3	1	500

Table 4.1 Distribution of variate values in finite population

In Figure 4.1 a histogram of the variate values is plotted. We notice that the population of variate values is very positively skewed. The population mean and the population

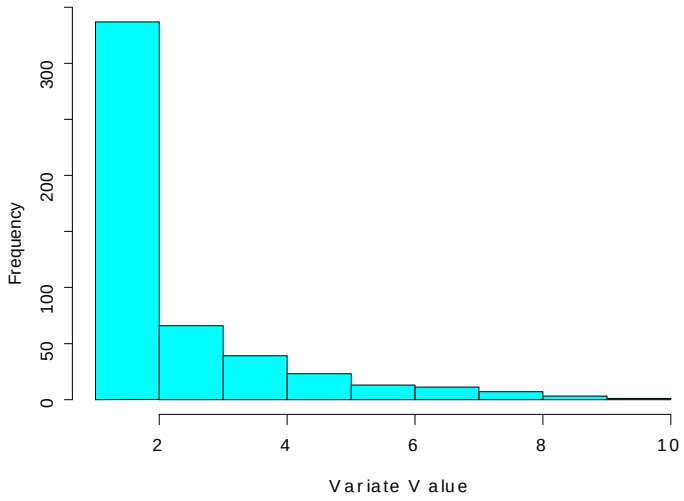


Figure 4.1: **Histogram of the variate values for the finite population of Table 4.1**

standard deviation are given respectively by

$$\mu = \frac{1}{500}\left[210\,(1) + 127\,(2) + \cdots + 1\,(10)\right] = \frac{1181}{500} = 2.362$$

and

$$\sigma = \sqrt{\frac{1}{500}\left[210\,(1)^2 + 127\,(2)^2 + \cdots + 1\,(10)^2 - 500\left(\frac{1181}{500}\right)^2\right]} = 1.7433$$

Note that the population variance is divided by 500 and not 499. To determine how good an estimator $\tilde{\mu} = \overline{Y}$ is we need the sampling distribution of $\overline{Y}$. This could be determined exactly but would require a great deal of effort. Another way to approximate the sampling

distribution is to use a computer simulation. The simulation can be done in two steps. First a random sample $y_1, y_2, \ldots, y_n$ is drawn at random without replacement from the population. Secondly the sample mean $\bar{y}$ for this sample is determined. These two steps are repeated k times. The k sample means, $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$, can then be considered as a random sample from the distribution of the random variable $\tilde{\mu} = \overline{Y}$, and we can study the distribution by plotting a histogram of the values $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$. The R code for such a simulation is given in Chapter 4, Problem 1.

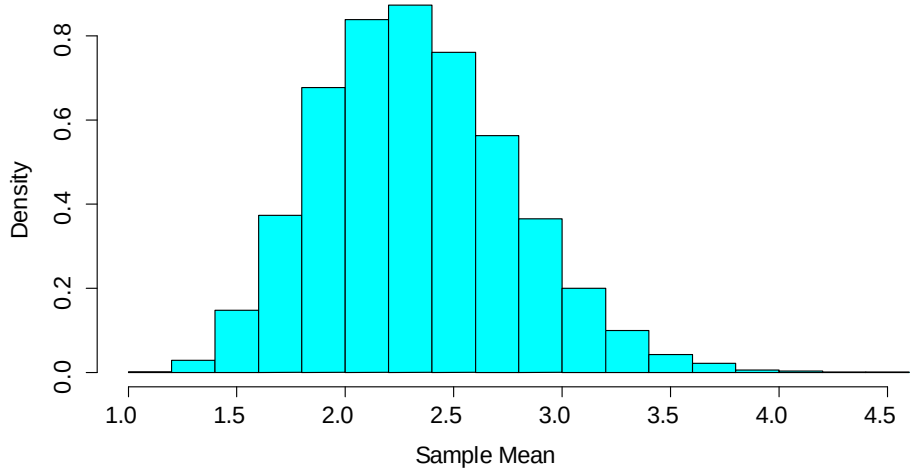


Figure 4.2: **Relative frequency histogram of sample means from 10000 samples of size 15 drawn from the population defined by Table 4.1**

The histogram in Figure 4.2 was obtained by drawing $k = 10000$ samples of size $n = 15$ from the population defined by Table 4.1, calculating the 10000 sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{10000}$ and then plotting the relative frequency histogram. The relative frequency histogram represents an approximation to the sampling distribution of the estimator $\overline{Y}$. The number of simulations k only affects how good the approximation is. We note that, although the histogram of the variate values was very skewed with a long right tail, the relative frequency histogram of the 10000 sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{10000}$ is reasonably symmetric and looks very much like a Gaussian distribution.

It can be shown[11] that the mean and standard deviation of the true sampling distribution of $\overline{Y}$ are

$$E\left(\overline{Y}\right) = \mu = 2.362 \quad \text{and} \quad sd\left(\overline{Y}\right) \approx \frac{\sigma}{\sqrt{n}} = \frac{1.7433}{\sqrt{15}} = 0.4501$$

The relative frequency histogram in Figure 4.2 indicates that the sample mean and median of the 10000 observations $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{10000}$ are both close in value to $E\left(\overline{Y}\right) = 2.362$.

[11] For a sample of size n drawn without replacement from a finite population of size N, $sd\left(\bar{Y}\right) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

The relative frequency histogram also indicates that the range of the 10000 observations $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{10000}$ is approximately 3. If $X \sim G(\mu_1, \sigma_1)$ then $P(|X - \mu_1| \leq 3\sigma_1) = 0.997$ which implies that 99.7% of the probability lies on the interval $[\mu_1 - 3\sigma_1, \mu_1 + 3\sigma_1]$ which has width $6\sigma_1$. The range which is 3 is close to the value $6(0.4501) \approx 2.7$.

Based on this simulation we can approximate $P\left(\left|\overline{Y} - 2.362\right| \leq 0.5\right)$, the probability that the sample mean $\overline{Y}$ is within 0.5 of the population mean $\mu = 2.362$, by counting the number of sample means in the simulation which are within 0.5 of the value 2.362. For the simulation in Figure 4.2 this value was 0.7422.

If samples of size $n = 30$ were drawn, how would the location, variability, and symmetry of the histogram of simulated means change? How would the estimate of $P\left(\left|\overline{Y} - 2.362\right| \leq 0.5\right)$ be affected? See Chapter 4, Problem 1.

Regardless of how the sampling distribution of an estimator $\tilde{\theta}$ is determined, the sampling distribution is important because it allows us to compute probabilities of the form $P(|\tilde{\theta} - \theta| \leq d)$ for given d so that we can quantify the uncertainty in the estimate $\hat{\theta}$.

The estimates of unknown parameters that we have discussed so far are often referred to as *point estimates*. This is because they consist of a single value or "point". Sampling distributions allow us to address the uncertainty in a point estimate. The uncertainty in a point estimate is usually conveyed by an *interval estimate*, which takes the form $[L(\mathbf{y}), U(\mathbf{y})]$ where the endpoints of the interval, $L(\mathbf{y})$ and $U(\mathbf{y})$, are both functions of the observed data $\mathbf{y}$. In the next two sections we will look at methods for constructing interval estimates and how they can be used to quantify the uncertainty in our point estimates.

4.3 Interval Estimation Using the Likelihood Function

The likelihood function can be used to obtain interval estimates for parameters in a very straightforward way. We do this here for the case in which the probability model involves only a single scalar parameter θ. Individual models often have constraints on the parameters. For example in the Gaussian distribution, the mean can be any real number $\mu \in \Re$ but the standard deviation must be positive, that is, $\sigma > 0$. Similarly for the Binomial model the probability of success must lie in the interval $[0, 1]$. These constraints are usually identified by indicating that the parameter is an element of some set Ω, called the *parameter space*.

As mentioned in Chapter 2 we often rescale the likelihood function to have a maximum value of one to obtain the relative likelihood function.

Definition 24 *Suppose θ is scalar and that some observed data (say a random sample $y_1, y_2, \ldots, y_n$) have given a likelihood function $L(\theta)$. The relative likelihood function $R(\theta)$ is defined as*

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad \text{for } \theta \in \Omega$$

where $\hat{\theta}$ is the maximum likelihood estimate and Ω is the parameter space.
Note:

$$0 \leq R(\theta) \leq 1 \quad \textit{for all } \theta \in \Omega$$

Definition 25 *A $100p\%$ likelihood interval for θ is the set $\{\theta : \ R(\theta) \geq p\}$.*

The set $\{\theta : R(\theta) \geq p\}$ is not necessarily an interval unless $R(\theta)$ is unimodal, but this is the case for nearly all models that we consider here. (What does the relative likelihood function look like for Chapter 2, Problem 2?)

To understand why a likelihood interval makes sense, suppose the model we have chosen for our data is a discrete probability model. In this case the maximum likelihood estimate $\hat{\theta}$ is the value of θ which not only maximizes $L(\theta)$ (and also $R(\theta)$), it is also the value of θ which makes the observed data most probable. Values of θ close to $\hat{\theta}$ would also make the observed data highly probable. The justification for a likelihood interval is that the values of θ that give large values of $R(\theta)$, are the most plausible in light of the data since they make the observed data most probable. Values of θ which give small values of $R(\theta)$ are less plausible in light of the observed data because they don't make the observed data very probable. A likelihood interval is therefore an interval of values for θ, which are reasonable based on how probable they make the observed data. For example, if you tossed a coin with $P(\text{Head}) = \theta$, 100 times and observed 10 heads then $\theta = 0.5$ would not make the observed data very probable and therefore $\theta = 0.5$ would not be a plausible value for θ. However values of θ close to $\hat{\theta} = 0.1$ would be plausible since they would make the observed data probable.

Notes:
(1) Usually likelihood intervals cannot be found explicitly. They can be determined approximately by plotting the relative likelihood function $R(\theta)$ using software such as R. They may be determined more accurately by solving the equation $R(\theta) - p = 0$ using the `uniroot` function in R.
(2) A likelihood interval is an interval of the form $[L(\mathbf{y}), U(\mathbf{y})]$ where $L(\mathbf{y})$ and $U(\mathbf{y})$ are functions of the observed data $\mathbf{y}$. $L(\mathbf{y})$ and $U(\mathbf{y})$ are the two solutions of the equation $R(\theta) - p = 0$ with $L(\mathbf{y}) \leq U(\mathbf{y})$. Since $R(\theta) = R(\theta; \mathbf{y})$ depends on the data $\mathbf{y}$, the solutions $L(\mathbf{y})$ and $U(\mathbf{y})$ will also depend on the data $\mathbf{y}$.

Example 4.3.1 Polls

Let θ be the proportion of people in a large study population who have a specific characteristic. Let Y be the number who have the characteristic in a random sample of size n. A reasonable model for Y is the Binomial(n, θ) distribution.[12] Suppose we select a particular set of n people and y are observed to have the characteristic of interest. As shown

[12] Polls are conducted using sampling without replacement in which case Y has a Hypergeometric distribution. Since the sampling is done from a large population a Binomial model is reasonable.

previously (see Table 2.2), the maximum likelihood estimate of θ is the sample proportion $\hat{\theta} = y/n$ and the relative likelihood function is

$$R(\theta) = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} \quad \text{for } 0 \leq \theta \leq 1$$

Figure 4.3 shows the relative likelihood functions $R(\theta)$ for two different sized polls:

$$\begin{aligned} \text{Poll 1} &: n = 200,\ y = 80,\ \hat{\theta} = 80/200 = 0.40 \\ \text{Poll 2} &: n = 1000,\ y = 400,\ \hat{\theta} = 400/1000 = 0.40 \end{aligned}$$

A 10% likelihood interval for θ based on Poll 1 can be obtained approximately from Figure 4.3 by determining the two points of intersection of the relative likelihood curve for Poll 1 (blue curve) and the dashed black line $y = 0.1$. The points of intersection are $(0.33, 0.1)$ and $(0.47, 0.1)$ and therefore the 10% likelihood interval is $[0.33, 0.47]$. In a similar manner, the 10% likelihood interval for θ based on Poll 2 is $[0.37, 0.43]$. The 10% likelihood interval for Poll 1 is wider than for Poll 2 which indicates that the uncertainty in the estimate of θ based on the data from Poll 1 is greater than the uncertainty in the estimate of θ based on the data from Poll 2. This makes sense since Poll 2 has five times as many observations as Poll 1 and therefore should contain more information about the unknown parameter θ.

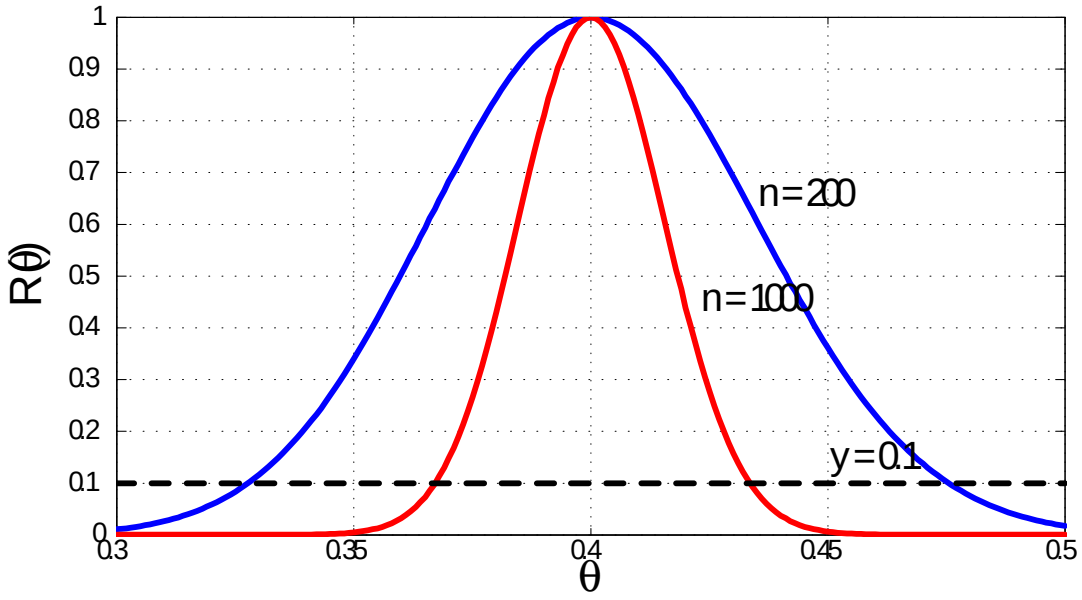


Figure 4.3: **Relative likelihood function for two polls with different sample sizes**

Table 4.2 gives guidelines for interpreting likelihood intervals. **These are only guidelines for this course.**

Table 4.2
Guidelines for Interpreting Likelihood Intervals

Values of θ inside a 50% likelihood interval are very plausible in light of the observed data.
Values of θ inside a 10% likelihood interval are plausible in light of the observed data.
Values of θ outside a 10% likelihood interval are implausible in light of the observed data.
Values of θ outside a 1% likelihood interval are very implausible in light of the observed data.

The values 1%, 10%, and 50% are typically used because they are nice round numbers and they provide useful summaries. **Other values could also be used.** In Section 4.6 we will see that 15% likelihood intervals have a connection with 95% confidence intervals which are discussed in Section 4.4. (Values inside a 15% likelihood interval are also plausible in light of the observed data.) A 10% likelihood interval is useful because it excludes parameter values for which the probability of the observed data is less than $\frac{1}{10}$ of the probability when $\theta = \hat{\theta}$. In other words a 10% likelihood interval summarizes the interval of values for the unknown parameter which are reasonably supported by the observed data in an empirical study. A 50% likelihood interval contains values of the parameter for which the probability of the observed data is at least $\frac{1}{2}$. A narrower 50% likelihood interval might be used if decisions made on the basis of the plausible values of the unknown parameter in light of the data had serious consequences in terms of money or lives of people. A 1% likelihood interval, which is wider than a 10% likelihood interval, would be used if the aim of the empirical study was to summarize all the parameter values which are supported in some way by the observed data. Which likelihood interval is used, therefore, depends very much on the goals of the empirical study that is being conducted.

A drawback of likelihood intervals (as well as confidence intervals as we will see in the next section) is that we never know whether the interval obtained contains the true value of the parameter or not. In Section 4.6 we will see that the construction of a likelihood interval ensures that we can be reasonably confident that it does.

Sometimes it is more convenient to compute the natural log of the relative likelihood function instead of $R(\theta)$.

Definition 26 *The log relative likelihood function is*

$$r(\theta) = \log R(\theta) = \log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] = l(\theta) - l(\hat{\theta}) \quad \text{for } \theta \in \Omega$$

where $l(\theta) = \log L(\theta)$ *is the log likelihood function.*

If the likelihood function $R(\theta)$ is unimodal then the log likelihood function $r(\theta)$ is also unimodal. Both $R(\theta)$ and $r(\theta)$ obtain a maximum value at $\theta = \hat{\theta}$. Note however that $R(\hat{\theta}) = 1$ while $r(\hat{\theta}) = 0$. The plots of $R(\theta)$ and $r(\theta)$ differ in terms of their shape. The plot of the relative likelihood function $R(\theta)$ (see for example, Figure 4.3) often resembles a Gaussian probability density function in shape while the plot of the log relative likelihood $r(\theta)$ resembles a quadratic function of θ (see, for example, Figure 4.4.)

The log relative likelihood function can also be used to obtain a $100p\%$ likelihood interval since $R(\theta) \geq p$ if and only if $r(\theta) \geq \log p$. In other words, a $100p\%$ likelihood interval can also be defined as $\{\theta : \ r(\theta) \geq \log p\}$. For example, $\{\theta : \ r(\theta) \geq \log(0.1) = -2.30\}$ is a 10% likelihood interval. A $100p\%$ likelihood interval can be determined approximately from a graph of the log relative likelihood function or by solving the equation $r(\theta) - \log p = 0$ using the R function `uniroot`.

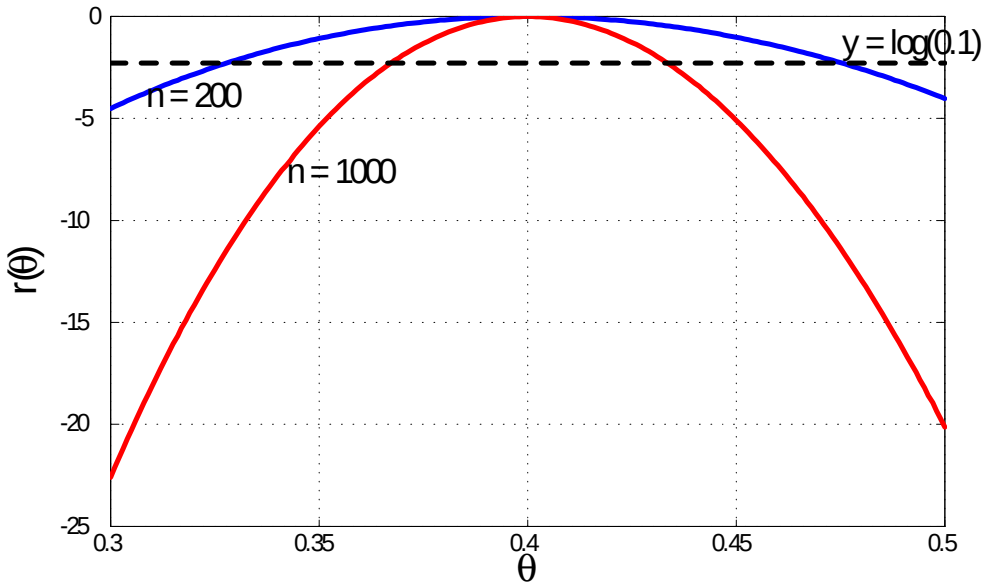


Figure 4.4: **Log relative likelihood function for two polls with different sample sizes**

The idea of a likelihood interval for a parameter θ can also be extended to the case of a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. In this case $R(\boldsymbol{\theta}) \geq p$ gives likelihood "regions" for $\boldsymbol{\theta}$.

4.4 Confidence Intervals and Pivotal Quantities

In the previous section we saw how the likelihood function could be used to obtain an interval estimate for an unknown parameter. In this section we look at another type of interval estimate called a confidence interval. We begin with an example.

Example 4.4.1 Gaussian distribution with unknown mean and known standard deviation

Suppose it is reasonable to assume that the observations $y_1, y_2, \ldots, y_n$ are a random sample from a $G(\mu, 2)$ distribution, that is, $\mu = E(Y_i)$ is unknown but $sd(Y_i) = 2$ is known. The maximum likelihood estimate of μ is $\hat{\mu} = \bar{y}$ and the maximum likelihood estimator is $\tilde{\mu} = \overline{Y}$. The sampling distribution of $\overline{Y}$ is $G(\mu, 2/\sqrt{n})$. Consider the interval estimate of μ given by

$$\left[\bar{y} - 1.96\,(2)/\sqrt{n}, \bar{y} + 1.96\,(2)/\sqrt{n}\right] \tag{4.1}$$

which is a symmetric interval about the maximum likelihood estimate $\hat{\mu} = \bar{y}$. (In this very special case this interval estimate is also a 14.7% likelihood interval.) For a given set of data this interval would be an interval of real numbers. For example, if $n = 16$ and $\bar{y} = 8.7$, then we obtain the interval

$$[8.7 - 1.96\,(2)/4,\ 8.7 + 1.96\,(2)/4] = [7.72,\ 9.78]$$

To interpret the interval estimate for μ given by (4.1) we need to consider the corresponding interval estimator for μ given by

$$\left[\overline{Y} - 1.96\,(2)/\sqrt{n}, \overline{Y} + 1.96\,(2)/\sqrt{n}\right]$$

This interval is called a random interval since both endpoints are random variables. Since $\overline{Y} \sim G(\mu, 2/\sqrt{n})$ we obtain

$$\begin{aligned} &P\left(\mu \in \left[\overline{Y} - 1.96\,(2)/\sqrt{n}, \overline{Y} + 1.96\,(2)/\sqrt{n}\right]\right) = P\left(\overline{Y} - 1.96\,(2)/\sqrt{n} \le \mu \le \overline{Y} + 1.96\,(2)/\sqrt{n}\right) \\ &= P\left(-1.96 \le \frac{\overline{Y} - \mu}{2/\sqrt{n}} \le 1.96\right) = P\left(-1.96 \le Z \le 1.96\right) \quad \text{where } Z \sim G(0,1) \\ &= 0.95 \end{aligned} \tag{4.2}$$

Can we conclude that $P\left(\mu \in [\bar{y} - 1.96\,(2)/\sqrt{n}, \bar{y} + 1.96\,(2)/\sqrt{n}]\right) = 0.95$ based on this statement? Unfortunately the answer is no. Once the data have been observed and the interval $[\bar{y} - 1.96\,(2)/\sqrt{n}, \bar{y} + 1.96\,(2)/\sqrt{n}]$ has been calculated, then either μ is in the interval or μ is not in the interval. Since μ is unknown we don't know which of these two statements is true.

So how can we use the fact that $P\left(\mu \in \left[\overline{Y} - 1.96\,(2)/\sqrt{n}, \overline{Y} + 1.96\,(2)/\sqrt{n}\right]\right) = 0.95$? Suppose a large number of independent random samples of size n were drawn from the $G(\mu, 2)$ distribution and the interval estimate $[\bar{y} - 1.96\,(2)/\sqrt{n}, \bar{y} + 1.96\,(2)/\sqrt{n}]$ was calculated for each sample. Since $P\left(\mu \in \left[\overline{Y} - 1.96\,(2)/\sqrt{n}, \overline{Y} + 1.96\,(2)/\sqrt{n}\right]\right) = 0.95$, we

would expect approximately 95% of the constructed intervals to contain the true but unknown value of μ. Of course approximately 5% of these constructed would not contain the value μ. Now we usually only have one data set and one interval and for this one interval we do not know whether it contains the true value of μ or not. We can only say that we are 95% **confident** that our interval contains the true value of μ. In other words, we hope we were one of the "lucky" 95% who constructed an interval containing the true value of μ.

As an example, suppose $n = 16$. The interval estimate for μ based on one sample with observed sample mean $\bar{y}$ is

$$\left[\bar{y} - 1.96\,(2)\,/\sqrt{16},\ \bar{y} + 1.96\,(2)\,/\sqrt{16}\right] = [\bar{y} - 0.98,\ \bar{y} + 0.98]$$

Suppose we generate 25 random samples of size $n = 16$ and we calculate this interval for each of these 25 data sets. One such simulation gave the following intervals for $\mu = 8$:

$[7.021, 8.981]$	$[7.375, 9.335]$	$[7.281, 9.241]$	$[7.059, 9.019]$	$[6.615, 8.575]$
$[7.613, 9.573]$	$[6.767, 8.727]$	$[6.645, 8.605]$	$[6.974, 8.934]$	$[7.026, 8.986]$
$[6.697, 8.657]$	$[7.716, 9.676]$	$[7.696, 9.656]$	$[7.115, 9.075]$	$[7.295, 9.255]$
$[6.772, 8.732]$	$[7.662, 9.622]$	$[7.879, 9.839]$	$[6.911, 8.871]$	$[7.061, 9.021]$
$[6.291, 8.251]$	$[5.962, 7.922]$	$[7.831, 9.791]$	$[6.868, 8.828]$	$[7.271, 9.231]$

We note that the intervals are all different and that 24 of the 25 or 96% of the generated intervals contain the value $\mu = 8$. Since $P\left(\mu \in \left[\overline{Y} - 0.98, \overline{Y} + 0.98\right]\right) = 0.95$, we would expect approximately $(25)\,(0.95) = 23.75$ of the intervals to contain the true value of μ which is very close to what we observed for this one simulation.

This example leads us to the following definition.

Definition 27 *Suppose the interval estimator* $[L(\mathbf{Y}), U(\mathbf{Y})]$ *has the property that*

$$P\left\{\theta \in [L(\mathbf{Y}), U(\mathbf{Y})]\right\} = P\left[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})\right] = p \tag{4.3}$$

Suppose the interval estimate $[L(\mathbf{y}), U(\mathbf{y})]$ *is constructed for the parameter* θ *based on observed data* $\mathbf{y}$*. The interval estimate* $[L(\mathbf{y}), U(\mathbf{y})]$ *is called a* $100p\%$ confidence interval for θ *and* p *is called the* confidence coefficient*.*

Note: $P\left[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})\right]$ in (4.3) is called the *coverage probability* of the interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$.

Based on Definition 27, the intervals in Example 4.4.1 are all 95% confidence intervals.

We can think of an interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$ which satisfies (4.3) as a rule which indicates how to construct a $100p\%$ confidence interval for θ. If we observe data $\mathbf{y}$ then the $100p\%$ confidence interval for θ based on the observed data is $[L(\mathbf{y}), U(\mathbf{y})]$. If we conduct an experiment a large number of times and each time we construct the interval $[L(\mathbf{y}), U(\mathbf{y})]$ based on the observed data $\mathbf{y}$ then based on (4.3) we know that approximately $100p\%$ of

these intervals will contain the true value of the parameter θ. This gives us some confidence that for one particular sample, the true value of the parameter is contained in the confidence interval constructed from the observed data.

Important: $P(\theta \in [L(\mathbf{y}), U(\mathbf{y})]) = p$ is an **incorrect** statement. The parameter θ is a constant, not a random variable.

How do we choose such an interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$? In practice, we choose interval estimators for which the probability in (4.3) is fairly close to 1 (values 0.90, 0.95 and 0.99 are often used) while keeping the constructed intervals $[L(\mathbf{y}), U(\mathbf{y})]$ as narrow as possible. We now consider a general method for constructing confidence intervals based on a special type of random variable called a *pivotal quantity.*

Definition 28 *A* pivotal quantity $Q = Q(\mathbf{Y}; \theta)$ *is a function of the data* $\mathbf{Y}$ *and the unknown parameter* θ *such that the distribution of the random variable* Q *is fully known. That is, probability statements such as* $P(Q \leq b)$ *and* $P(Q \geq a)$ *depend on* a *and* b *but not on* θ *or any other unknown information.*

Example 4.4.1 Revisited Gaussian distribution with unknown mean and known standard deviation

In Example 4.4.1 the parameter $\mu = E(Y_i)$ was unknown but the standard deviation $sd(Y_i) = 2$ was known. Since $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $G(\mu, 2)$ distribution, $E(\overline{Y}) = \mu$, and $sd(\overline{Y}) = 2/\sqrt{n}$, it follows that

$$Q(\mathbf{Y}; \mu) = \frac{\overline{Y} - \mu}{2/\sqrt{n}} \sim G(0, 1)$$

Since the distribution of the random variable $Q(\mathbf{Y}; \mu)$ is completely known, therefore it is a pivotal quantity.

We now describe how a pivotal quantity can be used to construct a confidence interval. We begin with the statement $P[a \leq Q(\mathbf{Y}; \theta) \leq b] = p$ where $Q(\mathbf{Y}; \theta)$ is a pivotal quantity whose distribution is completely known. Suppose that we can re-express the inequality $a \leq g(\mathbf{Y}; \theta) \leq b$ in the form $L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})$ for some functions L and U. Then since

$$\begin{aligned} p &= P[a \leq Q(\mathbf{Y}; \theta) \leq b] \\ &= P[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})] \\ &= P(\theta \in [L(\mathbf{Y}), U(\mathbf{Y})]) \end{aligned}$$

we see that, although the coverage probability for the interval estimator $[L(\mathbf{Y}), U(\mathbf{Y})]$ depends on a and b, these values are known since the distribution of $Q(\mathbf{Y}; \theta)$ is known

For observed data $\mathbf{y}$, the $100p\%$ confidence interval $[L(\mathbf{y}), U(\mathbf{y})]$ for θ, will also not depend on θ.

Example 4.4.2 Confidence interval for the mean μ of a Gaussian distribution with known standard deviation σ

Suppose $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a random sample from the $G(\mu, \sigma)$ distribution where $E(Y_i) = \mu$ is unknown but $sd(Y_i) = \sigma$ is known. (Note that the steps in this example are similar to the steps in Example 4.4.1 in which σ was assumed to be 2.) Since

$$Q = Q(\mathbf{Y};\mu) = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1) \tag{4.4}$$

and $G(0,1)$ is a completely known distribution, Q is a pivotal quantity. To obtain a 95% confidence interval for μ we first note that $P(-1.96 \leq Z \leq 1.96) = 0.95$ where $Z \sim G(0,1)$. From (4.4) it follows that

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\ &= P\left(\overline{Y} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{Y} + 1.96\sigma/\sqrt{n}\right) \end{aligned}$$

Therefore

$$\left[\bar{y} - 1.96\sigma/\sqrt{n},\ \bar{y} + 1.96\sigma/\sqrt{n}\right]$$

is a 95% confidence interval for μ based on the observed data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

Note that if a and b are values such that $0.95 = P(a \leq Z \leq b)$ where $Z \sim G(0,1)$ then the interval $[\bar{y} - b\sigma/\sqrt{n},\ \bar{y} - a\sigma/\sqrt{n}]$ is also a 95% confidence interval for μ. The interval $[\bar{y} - 1.96\sigma/\sqrt{n},\ \bar{y} + 1.96\sigma/\sqrt{n}]$ can be shown to be the narrowest possible 95% confidence interval for μ.

The interval $[\bar{y} - 1.96\sigma/\sqrt{n},\ \bar{y} + 1.96\sigma/\sqrt{n}]$ or $\bar{y} \pm 1.96\sigma/\sqrt{n}$ is often referred to as a *two-sided* confidence interval. Note that this interval takes the form

$$\text{point estimate} \pm a \times \text{ standard deviation of the estimator}$$

where a is a quantile from the $G(0,1)$ distribution. Many of the two-sided confidence intervals which we will encounter in later sections of these course notes will take a similar form.

Exercise Show that
(*a*) $\bar{y} \pm 1.6449\sigma/\sqrt{n}$ is a 90% confidence interval for μ
(*b*) $\bar{y} \pm 2.5758\sigma/\sqrt{n}$ is a 99% confidence interval for μ.

Since $P\left(\mu \in [\overline{Y} - 1.6449\sigma/\sqrt{n}, \infty)\right) = 0.95$, the interval $[\bar{y} - 1.6449\sigma/\sqrt{n}, \infty)$ is also a 95% confidence interval for μ. The interval $[\overline{\bar{y}} - 1.6449\sigma/\sqrt{n}, \infty)$ is usually referred to as a *one-sided* confidence interval. This type of interval is useful when we are interested in determining a lower bound on the value of μ. Simlarly the one-sided 95% confidence interval $(-\infty, \bar{y} + 1.6449\sigma/\sqrt{n}]$ is useful when we are interested in determining an upper bound on the value of μ.

Behaviour of confidence interval as $n \to \infty$

Confidence intervals become narrower as the size of the sample on which they are based increases. For example, note the effect of the sample size n in Example 4.4.2. The width of the 95% confidence interval was $2(1.96)\sigma/\sqrt{n}$ which decreases at the rate $\sqrt{n}$, as n increases. We noted this behaviour before for likelihood intervals. We will see in Section 4.6 that likelihood intervals are a type of confidence interval.

It turns out that for most models it is not possible to find *exact pivotal quantities* or confidence intervals for θ whose coverage probabilities do not depend on the true value of θ. However, in general we can find quantities $Q_n = Q_n(Y_1, Y_2, \ldots, Y_n, \theta)$ such that as $n \to \infty$, the distribution of Q_n ceases to depend on θ or other unknown information. We then say that Q_n is asymptotically pivotal, and in practice we treat Q_n as a pivotal quantity for sufficiently large values of n. We call Q_n an *asymptotic pivotal quantity.*

Asymptotic Gaussian Pivotal Quantities

Suppose $\tilde{\theta}$ is a point estimator of the unknown parameter θ. Suppose also that the Central Limit Theorem can be used to obtain the result that

$$\frac{\tilde{\theta} - \theta}{g(\theta)/\sqrt{n}}$$

has approximately a $G(0,1)$ distribution for large n where $E(\tilde{\theta}) = \theta$ and $sd(\tilde{\theta}) = g(\theta)/\sqrt{n}$ for some real valued function $g(\theta)$. If we replace θ by $\tilde{\theta}$ in the denominator then it can be shown that

$$Q_n(\tilde{\theta}; \theta) = \frac{\tilde{\theta} - \theta}{g(\tilde{\theta})/\sqrt{n}}$$

also has approximately a $G(0,1)$ distribution for large n. (This result is proved in STAT 330.) Therefore $Q_n(\tilde{\theta}; \theta)$ is an asymptotic Gaussian pivotal quantity which can be used to construct approximate confidence intervals for θ.

Example 4.4.3 Approximate confidence interval for Binomial model

Suppose $Y \sim \text{Binomial}(n, \theta)$. The maximum likelihood estimator of θ is $\tilde{\theta} = Y/n$ with

$$E(\tilde{\theta}) = E\left(\frac{Y}{n}\right) = \theta$$

and

$$sd(\tilde{\theta}) = sd\left(\frac{Y}{n}\right) = \sqrt{\frac{\theta(1-\theta)}{n}}$$

By the Central Limit Theorem the random variable

$$\frac{\tilde{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}$$

has approximately a $G(0,1)$ distribution for large n. In this example $g(\theta) = \sqrt{\theta(1-\theta)}$.

If we replace θ in the denominator by the estimator $\tilde{\theta} = Y/n$ then, based on the previous discussion, we have that the random variable

$$Q_n = Q_n(Y;\theta) = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$$

has approximately a $G(0,1)$ distribution for large n. Therefore Q_n is an asymptotic Gaussian pivotal quantity which can be used to construct confidence intervals for θ.

For example, since

$$\begin{aligned} 0.95 &\approx P(-1.96 \leq Q_n \leq 1.96) \\ &= P\left(\tilde{\theta} - 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}} \leq \theta \leq \tilde{\theta} + 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}\right) \end{aligned}$$

therefore

$$\left[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right] \tag{4.5}$$

$$= \hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \tag{4.6}$$

is an approximate 95% confidence interval for θ where $\hat{\theta} = y/n$ and y is the observed data.

Note Asymptotic Gaussian pivotal quantities exist for other models. See Problem 16 (Poisson), Problem 25 (Exponential), and Problem 18. See Table 4.3 in Section 4.8 for a summary of the approximate confidence intervals for these models.

Choosing the sample size for a Binomial experiment

We have seen that confidence intervals for a parameter tend to get narrower as the sample size n increases. When designing a study we often decide how large a sample to collect on the basis of (i) how narrow we would like confidence intervals to be, and (ii) how much we can afford to spend (it costs time and money to collect data). The following example illustrates the procedure.

Example 4.4.4

Suppose we want to estimate the probability θ from a Binomial experiment in which $Y \sim \text{Binomial}(n,\theta)$ distribution. We use the asymptotic pivotal quantity

$$Q_n = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$$

which was introduced in Example 4.4.3 and which has approximately a $G(0,1)$ distribution for large n, to obtain confidence intervals for θ.

Here is a criterion that is widely used for choosing the size of n: *Choose n large enough so that the width of a 95% confidence interval for θ is no wider than $2\,(0.03)$.* Let us see where this leads and why this rule is used.

From Example 4.4.3, we know that

$$\left[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right] \tag{4.7}$$

is an approximate 0.95 confidence interval for θ and that the width of this interval is

$$2\,(1.96)\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

To make this confidence interval narrower than $2\,(0.03)$, we need n large enough so that

$$1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq 0.03$$

or

$$n \geq \left(\frac{1.96}{0.03}\right)^2 \hat{\theta}(1-\hat{\theta}) \tag{4.8}$$

Of course we don't know what $\hat{\theta}$ is because we have not taken a sample, but we note that the interval (4.7) is the widest when $\hat{\theta} = 0.5$. To ensure that the inequality (4.8) holds for all values of $\hat{\theta}$, we find n such that

$$n \geq \left(\frac{1.96}{0.03}\right)^2 (0.5)\,(0.5) \approx 1067.1$$

Thus, choosing $n = 1068$ (or larger) will result in an approximate 95% confidence interval of the form $\hat{\theta} \pm c$, where $c \leq 0.03$.

If you look or listen carefully when polling results are announced, you will often hear words like "this poll is accurate to within 3 percentage points 19 times out of 20." What this really means is that the estimator $\tilde{\theta}$ (which is usually given in percentile form) approximately satisfies $P(\left|\tilde{\theta} - \theta\right| \leq 0.03) = 0.95$, or equivalently, that the actual estimate $\hat{\theta}$ is the centre of an approximate 95% confidence interval $\hat{\theta} \pm c$, for which $c = 0.03$. In practice, many polls are based on $1050 - 1100$ people, giving "accuracy to within 3 percent" with probability 0.95. Of course, one needs to be able to afford to collect a sample of this size. If we were satisfied with an accuracy of 5 percent, then we'd only need $n = 385$ (can you show this?). In many situations however this might not be sufficiently accurate for the purpose of the study.

Exercise Show that to ensure that the width of the approximate 95% confidence interval is $2\,(0.02) = 0.04$ or smaller, you need $n = 2401$. What should n be to make ensure the width of a 99% confidence interval is less than $2\,(0.02) = 0.04$?

Remark Very large Binomial polls ($n \geq 2000$) are not done very often. Although we can in theory estimate θ very precisely with an extremely large poll, there are two problems:

1. It is difficult to pick a sample that is truly random, so $Y \sim \text{Binomial}(n, \theta)$ is only an approximation.

2. In many settings the value of θ fluctuates over time. A poll is at best a snapshot at one point in time.

As a result, the "real" accuracy of a poll cannot generally be made arbitrarily high.

Census versus a random sample

Conducting a complete census is usually costly and time-consuming. This example illustrates how a random sample, which is less expensive, can be used to obtain "good" information about the attributes of interest for a population.

Suppose interviewers are hired at \$20 per hour to conduct door to door interviews of adults in a municipality of $50,000$ households. There are two choices:

(1) conduct a census using all $50,000$ households or

(2) take a random sample of households in the municipality and then interview a member of each household.

If a random sample is used it is estimated that each interview will take approximately 20 minutes (travel time plus interview time). If a census is used it is estimated that each interview will take approximately 10 minutes each since there is less travel time. We can summarize the costs and precision one would obtain for one question on the form which asks whether a person agrees/disagrees with a statement about the funding levels for higher education. Let θ be the proportion in the population who agree. Suppose we decide that a "good" estimate of θ is one that is accurate to within 2% of the true value 95% of the time.

For a census, six interviews can be completed in one hour. At \$20 per hour the interviewer cost for the census is approximately $\$20\,(50000)\,/6 = \$166,667$ since there are $50,000$ households. For a random sample, three interviews can be completed in one hour. An approximate 95% confidence interval for θ of the form $\hat{\theta} \pm 0.02$ requires $n = 2401$. The cost of the random sample of size $n = 2401$ is $\$20\,(2401)\,/3 \approx \$16,000$ as compared to $\$166,667$ for the census - more than ten times the cost of the random sample!

Of course, we have also not compared the costs of processing $50,000$ versus 2401 surveys but it is obvious again that the random sample will be less costly and time consuming.

4.5 The Chi-squared and t Distributions

In this section we introduce two new distributions, the Chi-squared distribution and the Student t distribution. These two distributions play an important role in constructing confidence intervals and the tests of hypotheses to be discussed in Chapter 5.

The χ^2 (Chi-squared) Distribution

To define the Chi-squared distribution we first recall the Gamma function and its properties:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy \quad \text{for } \alpha > 0$$

Properties of the Gamma function:
(1) $\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1)$
(2) $\Gamma(\alpha) = (\alpha-1)!$ for $\alpha = 1, 2, \ldots$
(3) $\Gamma(1/2) = \sqrt{\pi}$

The $\chi^2(k)$ distribution is a continuous family of distributions on $(0, \infty)$ with probability density function of the form

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2} \quad \text{for } x > 0 \tag{4.9}$$

where $k \in \{1, 2, \ldots\}$ is a parameter of the distribution. We write $X \sim \chi^2(k)$. The parameter k is referred to as the "degrees of freedom" (d.f.) parameter. In Figure 4.5 you see the characteristic shapes of the Chi-squared probability density functions. *For $k = 2$, the probability density function is the Exponential(2) probability density function.* For $k > 2$, the probability density function is unimodal with maximum value at $x = k - 2$. *For values of $k \geq 30$, the probability density function resembles that of a $N(k, 2k)$ probability density function.*

The cumulative distribution function, $F(x;k)$, can be given in closed algebraic form for even values of k. Probabilities for the $\chi^2(k)$ distribution are provided in the Chi-squared table at the end of these Course Notes. In R the function `dchisq(x,k)` gives the probability density function $f(x;k)$, `pchisq(x,k)` gives the cumulative distribution function $F(x;k) = P(X \leq x;k)$, and `qchisq(p,k)` gives the value a such that $P(X \leq a;k) = p$.

If $X \sim \chi^2(k)$ then

$$E(X) = k \quad \text{and} \quad Var(X) = 2k$$

This result follows by first showing that

$$E\left(X^j\right) = 2^j\frac{\Gamma\left(\frac{k}{2}+j\right)}{\Gamma\left(\frac{k}{2}\right)} \quad \text{for} \quad j = 1, 2, \ldots.$$

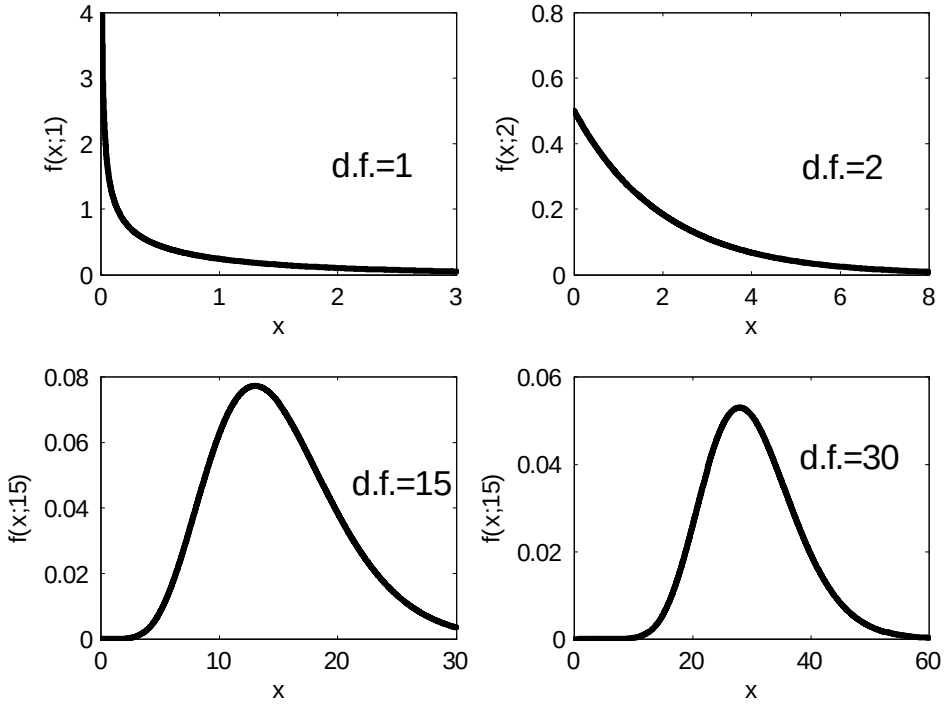


Figure 4.5: **Chi-squared probability density functions for $k = \mathbf{1}, \mathbf{2}, \mathbf{15}, \mathbf{30}$**

This is true since

$$\begin{aligned} E\left(X^j\right) &= \int_0^\infty \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)+j-1} e^{-x/2} dx \quad \text{let } y = x/2 \text{ or } x = 2y \\ &= \int_0^\infty \frac{1}{2^{k/2}\Gamma(k/2)} (2y)^{(k/2)+j-1} e^{-y} 2dy \\ &= \frac{2^j}{\Gamma(k/2)} \int_0^\infty y^{(k/2)+j-1} e^{-y} dy \\ &= 2^j \frac{\Gamma\left(\frac{k}{2}+j\right)}{\Gamma\left(\frac{k}{2}\right)} \end{aligned}$$

Letting $j = 1$ we obtain

$$E(X) = 2\frac{\Gamma\left(\frac{k}{2}+1\right)}{\Gamma\left(\frac{k}{2}\right)} = 2\left(\frac{k}{2}\right) = k$$

Letting $j = 2$ we obtain

$$E\left(X^2\right) = 2^2\frac{\Gamma\left(\frac{k}{2}+2\right)}{\Gamma\left(\frac{k}{2}\right)} = 4\left(\frac{k}{2}+1\right)\left(\frac{k}{2}\right) = k(k+2)$$

and therefore

$$Var\left(X\right)=E\left(X^{2}\right)-\left[E\left(X\right)\right]^{2}=k\left(k+2\right)-k^{2}=2k$$

The following results will also be very useful.

Theorem 29 *Let $W_1, W_2, \ldots, W_n$ be independent random variables with $W_i \sim \chi^2(k_i)$. Then $S=\sum_{i=1}^{n} W_i \sim \chi^2(\sum_{i=1}^{n} k_i)$.*

Proof. See Problem 21. ■

Theorem 30 *If $Z \sim G(0,1)$ then the distribution of $W=Z^2$ is $\chi^2(1)$.*

Proof. *Suppose $W=Z^2$ where $Z \sim G(0,1)$. Let Φ represent the cumulative distribution function of a $G(0,1)$ random variable and let ϕ represent the probability density function of a $G(0,1)$ random variable.*

For $w>0$

$$P(W\leq w)=P(-\sqrt{w}\leq Z\leq\sqrt{w})=\Phi(\sqrt{w})-\Phi(-\sqrt{w})$$

and therefore the probability density function of W is

$$\begin{aligned}\frac{d}{dw}P(W\leq w)&=\frac{d}{dw}\left[\Phi(\sqrt{w})-\Phi(-\sqrt{w})\right]\\&=\left[\phi(\sqrt{w})+\phi(-\sqrt{w})\right]\left(\frac{1}{2}w^{-1/2}\right)\\&=\frac{1}{\sqrt{2\pi}}w^{-1/2}e^{-w/2}\end{aligned}$$

which is the probability density function of a $\chi^2(1)$ random variable as required. ■

Corollary 31 *If $Z_1, Z_2, \ldots, Z_n$ are mutually independent $G(0,1)$ random variables and $S=\sum_{i=1}^{n} Z_i^2$, then $S \sim \chi^2(n)$.*

Proof. *Since $Z_i \sim G(0,1)$ then by Theorem 30, $Z_i^2 \sim \chi^2(1)$ and the result follows by Theorem 29.*

■

The following results will be useful in Chapter 5.

Useful Results:

1. If $W \sim \chi^2(1)$ then $P(W \geq w)=2\left[1-P\left(Z\leq\sqrt{w}\right)\right]$ where $Z \sim G(0,1)$.
2. If $W \sim \chi^2(2)$ then $W \sim Exponential(2)$ and $P(W \geq w)=e^{-w/2}$.

Student's t Distribution

Student's t distribution (or more simply the t distribution) has probability density function

$$f(t;k) = c_k \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2} \qquad \text{for } t \in \Re \text{ and } k = 1, 2, \ldots$$

where the constant c_k is given by

$$c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}$$

The parameter k is called the *degrees of freedom*. We write $T \sim t(k)$ to indicate that the random variable T has a t distribution with k degrees of freedom. In Figure 4.6 the probability density function $f(t;k)$ for $k = 2$ and $k = 25$ is plotted together with the $G(0,1)$ probability density function.

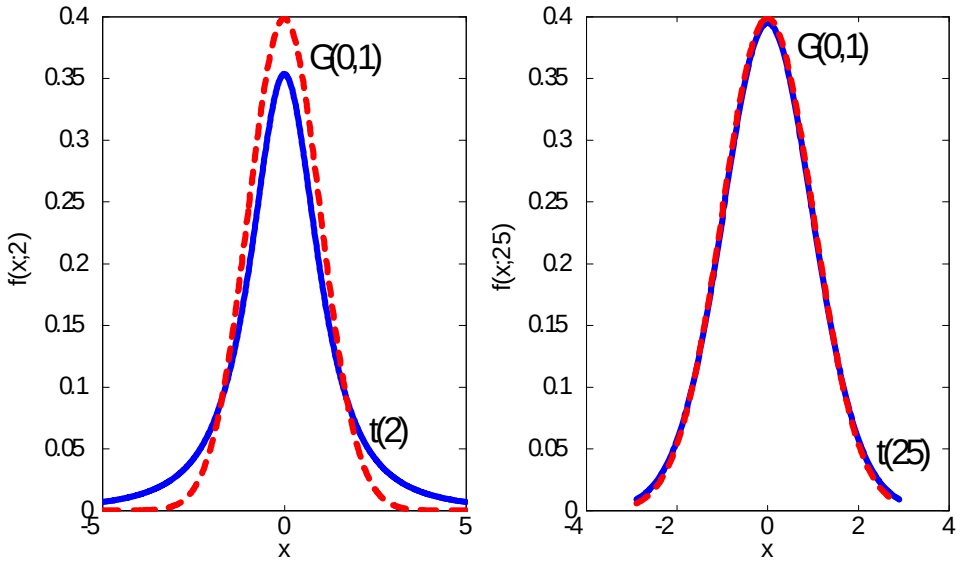


Figure 4.6: **Probability density functions for $t(k)$ (solid blue) and $G(0,1)$ (dashed red)**

The t probability density function is similar to that of the $G(0,1)$ distribution in several respects: it is symmetric about the origin, it is unimodal, and indeed for large values of k, the graph of the probability density function $f(t;k)$ is indistinguishable from that of the $G(0,1)$ probability density function. The primary difference, for small k such as the one plotted, is in the tails of the distribution. The t probability density function has fatter "tails" or more area in the extreme left and right tails. Problem 22 at the end of this chapter considers some properties of $f(x;k)$.

Probabilities for the t distribution are provided in the t table at the end of these Course Notes. In R the function `dt(t,k)` gives the probability density function $f(t;k)$, `pt(t,k)` gives the cumulative distribution function $F(t;k) = P(T \leq t;k)$, and `qt(p,k)` gives the value a such that $P(T \leq a;k) = p$.

The t distribution arises as a result of the following theorem. The proof of this theorem is beyond the scope of this course.

Theorem 32 *Suppose* $Z \sim G(0,1)$ *and* $U \sim \chi^2(k)$ *independently. Let*

$$T = \frac{Z}{\sqrt{U/k}}$$

Then T *has a* ***Student's*** t ***distribution with*** k ***degrees of freedom.***

4.6 Likelihood-Based Confidence Intervals

We will now show that likelihood intervals are also confidence intervals. Recall the relative likelihood

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \quad \text{for } \theta \in \Omega$$

is a function of the maximum likelihood estimate $\hat{\theta}$. Replace the estimate $\hat{\theta}$ by the random variable (the estimator) $\tilde{\theta}$ and define the random variable $\Lambda(\theta)$

$$\Lambda(\theta) = -2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right]$$

where $\tilde{\theta}$ is the maximum likelihood estimator. The random variable $\Lambda(\theta)$ is called the *likelihood ratio statistic*. The following theorem implies that $\Lambda(\theta)$ is an asymptotic pivotal quantity.

Theorem 33 *If* $L(\theta)$ *is based on* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$*, a random sample of size* n*, and if* θ *is the true value of the scalar parameter, then (under mild mathematical conditions) the distribution of* $\Lambda(\theta)$ *converges to a* $\chi^2(1)$ *distribution as* $n \to \infty$*.*

This theorem means that $\Lambda(\theta)$ can be used as a pivotal quantity for sufficiently large n in order to obtain approximate confidence intervals for θ. More importantly we can use this result to show that the likelihood intervals discussed in Section 4.3 are also approximate confidence intervals.

Theorem 34 *A* $100p\%$ *likelihood interval is an approximate* $100q\%$ *confidence interval where* $q = 2P\left(Z \leq \sqrt{-2\log p}\right) - 1$ *and* $Z \sim N(0,1)$*.*

Proof. A $100p\%$ likelihood interval is defined by $\{\theta; R(\theta) \geq p\}$ which can be rewritten as

$$\{\theta; R(\theta) \geq p\} = \left\{\theta : -2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right] \leq -2\log p\right\}$$

By Theorem 33 the confidence coefficient for this interval can be approximated by

$$\begin{aligned} P\left[\Lambda(\theta) \leq -2\log p\right] &= P\left\{-2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \leq -2\log p\right\} \\ &\approx P\left(W \leq -2\log p\right) \quad \text{where} \quad W \sim \chi^2(1) \\ &= P\left(|Z| \leq \sqrt{-2\log p}\right) \quad \text{where} \quad Z \sim N(0,1) \\ &= 2P\left(Z \leq \sqrt{-2\log p}\right) - 1 \end{aligned}$$

as required.

■

Example If $p = 0.1$ then

$$\begin{aligned} q &= 2P\left(Z \leq \sqrt{-2\log(0.1)}\right) - 1 \quad \text{where} \quad Z \sim G(0,1) \\ &= 2P\left(Z \leq 2.15\right) - 1 = 0.96844 \end{aligned}$$

and therefore a 10% likelihood interval is an approximate 97% confidence interval.

Exercise
(a) Show that a 1% likelihood interval is an approximate 99.8% confidence interval.
(b) Show that a 50% likelihood interval is an approximate 76% confidence interval.

Theorem 33 can also be used to find a likelihood interval which is also an approximate $100p\%$ confidence interval.

Theorem 35 *If a is a value such that $p = 2P(Z \leq a) - 1$ where $Z \sim N(0,1)$, then the likelihood interval $\left\{\theta : R(\theta) \geq e^{-a^2/2}\right\}$ is an approximate $100p\%$ confidence interval.*

Proof. The confidence coefficient corresponding to the interval $\left\{\theta : R(\theta) \geq e^{-a^2/2}\right\}$ is

$$\begin{aligned} P\left[\frac{L(\theta)}{L(\tilde{\theta})} \geq e^{-a^2/2}\right] &= P\left\{-2\log\left[\frac{L(\theta)}{L(\tilde{\theta})}\right] \leq a^2\right\} \\ &\approx P\left(W \leq a^2\right) \quad \text{where} \quad W \sim \chi^2(1) \quad \text{by Theorem 33} \\ &= 2P\left(Z \leq a\right) - 1 \quad \text{where} \quad Z \sim N(0,1) \\ &= p \end{aligned}$$

as required.

■

Example

Since

$$0.95 = 2P(Z \leq 1.96) - 1 \quad \text{where} \quad Z \sim N(0,1)$$

and

$$e^{-(1.96)^2/2} = e^{-1.9208} \approx 0.1465 \approx 0.15$$

therefore *a* 15% *likelihood interval for* θ *is also an approximate* 95% *confidence interval for* θ*.*

Exercise

(a) Show that a 26% likelihood interval is an approximate 90% confidence interval.

(b) Show that a 4% likelihood interval is an approximate 99% confidence interval.

Example 4.6.1 Approximate confidence intervals for Binomial model

For Binomial data with n trials and y successes the relative likelihood function is (see Example 4.3.1)

$$R(\theta) = \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} \quad \text{for } 0 \leq \theta \leq 1$$

Suppose $n = 100$ and $y = 40$ so that $\hat{\theta} = 40/100 = 0.4$. From the graph of the relative likelihood function given in Figure 4.7 we can read off the 15% likelihood interval which is $[0.31, 0.495]$ which is also an approximate 95% confidence interval.

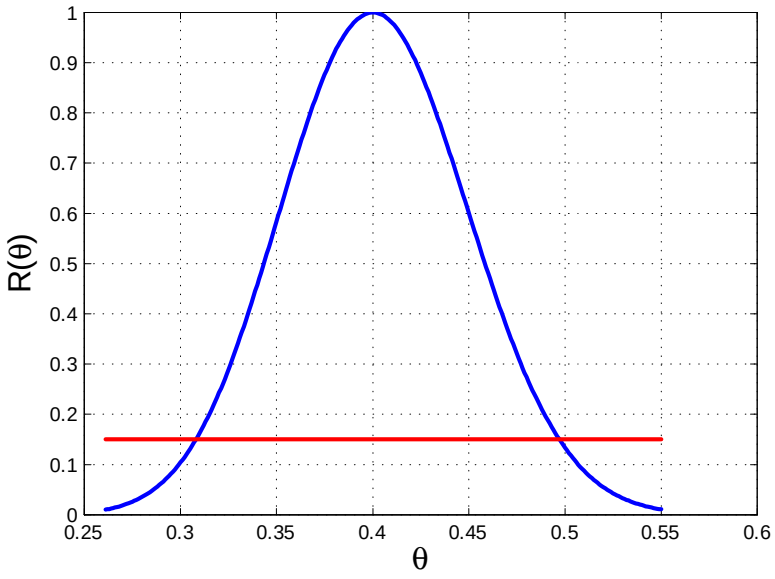


Figure 4.7: **Relative likelihood function for Binomial with** $n = 100$ and $y = 40$

The approximate 95% confidence interval

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \tag{4.10}$$

is $[0.304, 0.496]$. The two intervals differ slightly but are very close.

If $n = 30$ and $\hat{\theta} = 0.1$ then from Figure 4.8 the 15% likelihood interval is $[0.03, 0.24]$ which is also an approximate 95% confidence interval. The approximate 95% confidence

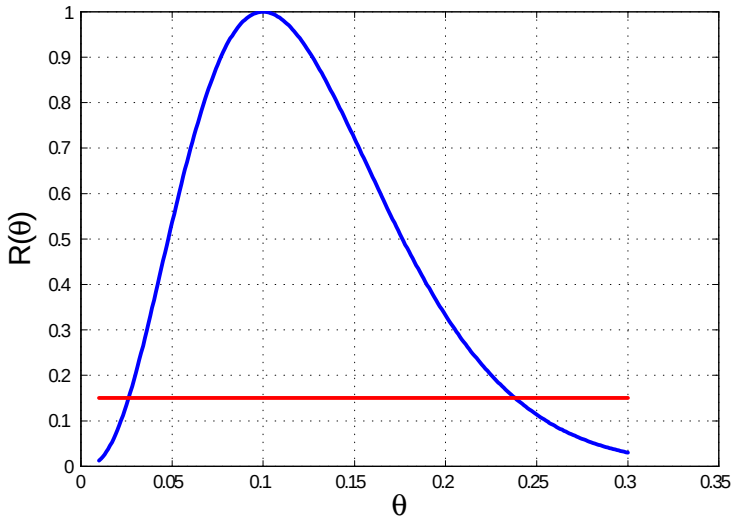


Figure 4.8: **Relative likelihood function for Binomial with** $n = 30$ and $y = 3$

interval based on 4.10 is $[-0.0074, 0.2074]$ which is quite different than the likelihood based approximate confidence interval and which also contains negative values for θ. Of course θ can only take on values between 0 and 1. This happens because the confidence interval in (4.10) is always symmetric about $\hat{\theta}$ and if $\hat{\theta}$ is close to 0 or 1 and n is not very large then the interval can contain values less than 0 or bigger than 1. The graph of the likelihood interval in Figure 4.8 is not symmetric about $\hat{\theta}$. In this case the 15% likelihood interval is a better summary of the θ values which are supported by the data.

More generally, if $\hat{\theta}$ is close to 0.5 or n is large then the likelihood interval will be fairly symmetric about $\hat{\theta}$ and there will be little difference in the two approximate confidence intervals. If $\hat{\theta}$ is close to 0 or 1 and n is not large then the likelihood interval will not be symmetric about $\hat{\theta}$ and the two approximate confidence intervals will not be similar. In this case the 15% likelihood interval will be a better summary of the θ values which are supported by the data.

4.7 Confidence Intervals for Parameters in the $G(\mu, \sigma)$ Model

Suppose we have a variate of interest (for example, the weight in kilograms of a female in the population of Example 1.3.2) whose distribution it is reasonable to model as a $G(\mu, \sigma)$ random variable. Suppose also that we plan to take a random sample $Y_1, Y_2, \ldots, Y_n$ to estimate the unknown mean μ where $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n.$ and that the standard deviation σ is also unknown. Recall that the maximum likelihood estimator of μ is

$$\tilde{\mu} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

and the maximum likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

The sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

is also an estimator of σ^2. The estimators of σ^2 differ only in the denominator. Indeed if n is large there is very little difference between $\tilde{\sigma}^2$ and S^2. Note that the sample variance has the advantage that it is an *unbiased estimator*, that is, $E(S^2) = \sigma^2$ (see Chapter 1, Problem 20).

Confidence intervals for μ

If σ were known then we have seen in Section 4.4 that

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1) \tag{4.11}$$

is a pivotal quantity that can be used to obtain confidence intervals for μ. However, σ is generally unknown. Fortunately it turns out that if we simply replace σ with either the maximum likelihood estimator $\tilde{\sigma}$ or the sample standard deviation S in Z, then we still have a pivotal quantity. We will write the pivotal quantity in terms of S. The pivotal quantity is

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \tag{4.12}$$

Since S, unlike σ, is a random variable in (4.12) the distribution of T is no longer $G(0, 1)$. The random variable T actually has a t distribution which was introduced in Section 4.5.

Theorem 36 *Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample mean $\overline{Y}$ and sample variance S^2. Then*

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t\,(n-1) \tag{4.13}$$

Note: The random variable T is a pivotal quantity since it is a function of the data $Y_1, Y_2, \ldots, Y_n$ and the unknown parameter μ whose distribution $t\,(n-1)$ is completely known.

To see how

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t\,(n-1)$$

follows from Theorem 32 let

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0,1)$$

and

$$U = \frac{(n-1)S^2}{\sigma^2}$$

We choose this function of S^2 since it can be shown that $U \sim \chi^2(n-1)$. It can also be shown that Z and U are independent random variables. The proofs of these very important results are beyond the scope of this course and are covered in a third year mathematical statistics course.

By Theorem 32 with $k = n-1$, we have

$$\frac{Z}{\sqrt{U/k}} = \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t\,(n-1)$$

In other words if we replace σ in the pivotal quantity (4.11) by its estimator S, the distribution of the resulting pivotal quantity has a $t(n-1)$ distribution rather than a $G(0,1)$ distribution. The degrees of freedom of the t distribution are determined by the degrees of freedom of the Chi-squared random variable U.

We now show how to use the pivotal quantity (4.13) to obtain a confidence interval for μ when σ is unknown. Since the t distribution is symmetric we determine the constant a such that $P(-a \le T \le a) = p$ using the t table provided in these Course Notes or R. Note that, due to symmetry, $P(-a \le T \le a) = p$ is equivalent to $P\,(T \le a) = (1+p)\,/2$ (you should verify this) and since the t table tabulates the cumulative distribution function $P\,(T \le t)$, it is easier to find a such that $P\,(T \le a) = (1+p)\,/2$. Then since

$$\begin{aligned} p &= P(-a \le T \le a) \\ &= P\left(-a \le \frac{\overline{Y} - \mu}{S/\sqrt{n}} \le a\right) \\ &= P\left(\overline{Y} - aS/\sqrt{n} \le \mu \le \overline{Y} + aS/\sqrt{n}\right) \end{aligned}$$

a $100p\%$ confidence interval for μ is given by

$$\bar{y} \pm as/\sqrt{n} = \left[\bar{y} - as/\sqrt{n},\ \bar{y} + as/\sqrt{n}\right] \tag{4.14}$$

(Note that if we attempted to use (4.11) to build a confidence interval we would have two unknowns in the inequality since both μ and σ are unknown.) As usual the method used to construct this interval implies that $100p\%$ of the confidence intervals constructed from samples drawn from this population contain the true value of μ.

We note that this interval is of the form $\bar{y} \pm as/\sqrt{n}$ or

$$\text{estimate } \pm a \times \text{estimated standard deviation of estimator.}$$

Recall that a confidence interval for μ in the case of a $G(\mu, \sigma)$ population when σ is known has a similar form

$$\text{estimate } \pm a \times \text{standard deviation of estimator}$$

except that the standard deviation of the estimator is known in this case and the value of a is taken from a $G(0, 1)$ distribution rather than the t distribution.

Behaviour of confidence interval as $n \to \infty$

As the sample size n increases, confidence intervals behave in a largely predictable fashion. Since $E(S) \approx \sigma$ for large n, the sample standard deviation s gets closer to the true standard deviation σ. Secondly as the degrees of freedom $= n - 1$ increase, the quantiles of the t distribution approach the quantiles of the $G(0, 1)$ distribution. For example, for the column labeled $p = 0.975$ in the t table we notice that as the degrees of freedom increase, the quantiles are approaching the value 1.96 since $P(Z \leq 1.96) = 0.975$. In general for large n, the width of the confidence interval gets narrower as n increases (but at the rate $1/\sqrt{n}$) so that, in the limit, the confidence interval shrinks to include only the point $\bar{y}$.

Example 4.7.1 Study on physical activity and academic performance

Researchers at a university in a large city who were interested in studying the relationship between physical activity and academic performance were given permission to randomly select 51 Grade 7 girls attending a very large senior public school to participate in the study. Parental consent for each student was also obtained. Data on age, height, weight, IQ score, and scores on a fitness test were collected for each participant. To analyse the data on heights (in centimeters) the model $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 51$ was assumed. The study population is all Grade 7 girls attending the large senior public school. The sample size is $n = 51$. The parameter μ represents the mean height in centimeters of the girls in the study population and the parameter σ represents the standard deviation of the heights in centimeters of the girls in the study population. (We assume the heights are measured without error.)

For these data

$$\bar{y} = 150.1412 \qquad s = 5.3302$$

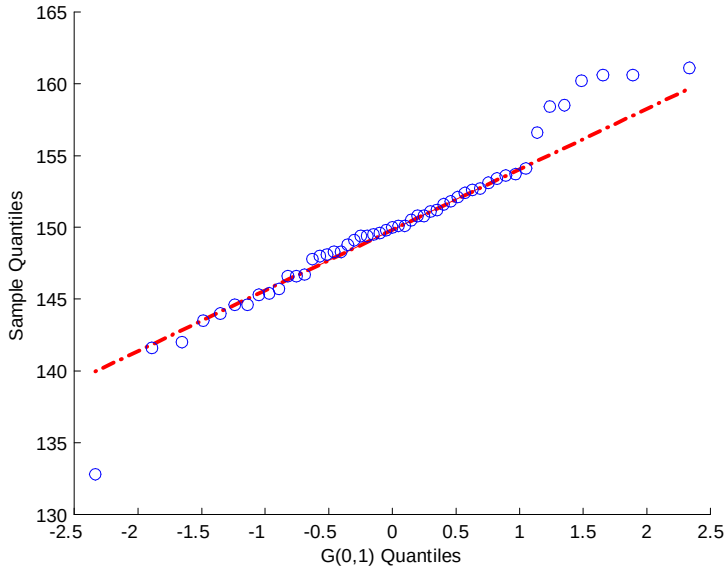


Figure 4.9: QQplot of heights for Example 4.7.1

and a qqplot of the data are given in Figure 4.9. Since the points in the qqplot lie reasonably along a straight line with more variability at both ends which is expected, we would conclude that a Gaussian model is reasonable for these data.

Since

$$P(T \leq 2.0086) = \frac{1 + 0.95}{2} = 0.975 \quad \text{for } T \sim t\,(50)$$

a 95% confidence interval for μ based on (4.14) is

$$\begin{aligned} & \bar{y} \pm 2.0086 s/\sqrt{51} \\ = \; & 150.1412 \pm (2.0086)\,(5.3302)\,/\sqrt{51} \\ = \; & 150.1412 \pm 1.4992 \\ = \; & [148.6420, 151.6403] \end{aligned}$$

Sample size required for a given width of confidence interval for μ

If we know the value of σ approximately (possibly from previous studies), we can determine the value of n needed to make a 95% confidence interval a given length. This is used in deciding how large a sample to take in a future study. A 95% confidence interval using the $G\,(0, 1)$ quantiles takes the form $\bar{y} \pm 1.96\sigma/\sqrt{n}$. If we wish a 95% confidence interval of the form $\bar{y} \pm d$ (the width of the confidence interval is then $2d$), we should choose

$$\begin{aligned} 1.96\sigma/\sqrt{n} &\approx d \\ \text{or} \quad n &\approx (1.96\sigma/d)^2 \end{aligned}$$

We would usually choose n a little larger than this formula gives to accommodate the fact that we used $G(0,1)$ quantiles rather than the quantiles of the t distribution which are larger in value and we only know σ approximately.

Confidence intervals for σ^2 and σ

Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution. We have seen that there are two closely related estimators for the population variance, $\tilde{\sigma}^2$ and the sample variance S^2. We use S^2 to build a confidence interval for the parameter σ^2. Such a construction depends on the following result.

Theorem 37 *Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the $G(\mu, \sigma)$ distribution with sample variance S^2.*

$$U = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2 \sim \chi^2(n-1) \tag{4.15}$$

Note: The random variable U is a pivotal quantity since it is a function of the data $Y_1, Y_2, \ldots, Y_n$ and the unknown parameter σ^2 whose distribution $\chi^2(n-1)$ is completely known.

While the proof of this result is beyond the scope of this course, we will try to explain the puzzling number of degrees of freedom $n-1$, which, at first glance, seems wrong since $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is the sum of n squared Normal random variables. Does this contradict Corollary 31? It is true that each $W_i = (Y_i - \overline{Y})$ is a Normally distributed random variable. However W_i does **not** have a $N(0,1)$ distribution and more importantly the W_i's are **not independent**! (See Problem 24.) One way to see that $W_1, W_2, \ldots, W_n$ are not independent random variables is to note that since $\sum_{i=1}^{n} W_i = 0$ this implies $W_n = -\sum_{i=1}^{n-1} W_i$ so the last term can be determined using the sum of the first $n-1$ terms. Therefore in the sum, $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} W_i^2$ there are really only $n-1$ terms that are linearly independent or "free". This is an intuitive explanation for the $n-1$ degrees of freedom for the pivotal quantities (4.15) and (4.13). In both cases, the degrees of freedom are determined by S^2 and are related to the dimension of the subspace inhabited by the terms in the sum for S^2, that is, the terms $W_i = Y_i - \overline{Y}$, $i = 1, 2, \ldots, n$.

We will now show how to use the pivotal quantity (4.15) to construct a $100p\%$ confidence interval for the parameter σ^2 or σ. Using the Chi-squared table or R we can find constants a and b such that

$$P(a \leq U \leq b) = p$$

where $U \sim \chi^2(n-1)$. Since

$$\begin{aligned} p &= P(a \leq U \leq b) \\ &= P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) \\ &= P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) \\ &= P\left(\sqrt{\frac{(n-1)S^2}{b}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{a}}\right) \end{aligned}$$

a $100p\%$ confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{b},\ \frac{(n-1)s^2}{a}\right] \tag{4.16}$$

and a $100p\%$ confidence interval for σ is

$$\left[s\sqrt{\frac{n-1}{b}},\ s\sqrt{\frac{n-1}{a}}\right] \tag{4.17}$$

The choice for a, b is not unique. For convenience, a and b are usually chosen such that

$$P(U \leq a) = P(U > b) = \frac{1-p}{2} \tag{4.18}$$

where $U \sim \chi^2(n-1)$. Note that since the Chi-squared table provided in these Course Notes tabulate the cumulative distribution function, $P(U \leq u)$, this means using the table to find a and b such that

$$P(U \leq a) = \frac{1-p}{2} \quad \text{and} \quad P(U \leq b) = p + \frac{1-p}{2} = \frac{1+p}{2}$$

The intervals (4.16) and (4.17) are called *equal-tailed* confidence intervals. The choice (4.18) for a, b does not give the narrowest confidence interval. The narrowest interval must be found numerically. For large n the equal-tailed interval and the narrowest interval are nearly the same.

Note that, unlike confidence intervals for μ, the confidence interval for σ^2 is **not symmetric** about s^2, the estimate of σ^2. This happens of course because the $\chi^2(n-1)$ distribution is not a symmetric distribution.

In some applications we are interested in an upper bound on σ (because small σ is "good" in some sense). In this case we take $b = \infty$ and find a such that $P(a \leq U) = p$ or $P(U \leq a) = 1 - p$ so that a one-sided $100p\%$ confidence interval for σ is

$$\left[0,\ s\sqrt{\frac{n-1}{a}}\right]$$

Example 4.7.2 Optical glass

At the Clear Eye Optical Lab Company a manufacturing process produces wafer-shaped pieces of optical glass for lenses. Pieces must be very close to 25 millimeters thick, and only a small amount of variability around this can be tolerated. From past experience it is known that if Y represents the thickness of a randomly selected piece of glass then it is reasonable to assume the model $Y \sim G(\mu, \sigma)$. (Thicknesses are assumed to be measured without error.) The parameter μ represents the mean and σ represents the standard deviation of the thicknesses in millimeters of the pieces of optical glass produced by the manufacturing process at the Clear Eye Optical Lab (the study process). For quality control purposes a random sample of size $n = 15$ is drawn every eight hours to check if the process is working properly. The values of μ and σ are estimated based on the sample to see if they are consistent with $\mu = 25$ and with σ being under 0.02 millimeters. On one such occasion the observed data were

$$\bar{y} = 25.009 \quad \text{and} \quad s = 0.013$$

To obtain a 95% confidence interval for σ we determine a and b such that

$$P(U \leq a) = \frac{1 - 0.95}{2} = 0.025 \quad \text{and} \quad P(U \leq b) = \frac{1 + 0.95}{2} = 0.975$$

where $U \sim \chi^2(14)$. From the Chi-squared table or R we obtain

$$P(U \leq 5.629) = 0.025 \quad \text{and} \quad P(U \leq 26.119) = 0.975$$

so $a = 5.63$ and $b = 26.12$. Substituting these values along with $(14)\, s^2 = 0.002347$ into (4.17) we obtain

$$\left[0.013\sqrt{\frac{14}{26.119}},\ 0.013\sqrt{\frac{14}{5.629}}\right] = [0.00952,\ 0.0205]$$

as the 95% confidence interval for σ.

It seems plausible that $\sigma \leq 0.02$, though the right endpoint of the 95% confidence interval is very slightly over 0.02. Using $P(U \leq 6.571) = 0.05$ we can obtain a one-sided 95% confidence interval for σ which is given by

$$\left[0,\ s\sqrt{\frac{n-1}{a_1}}\right] = \left[0,\ 0.013\sqrt{\frac{14}{6.571}}\right] = [0, 0.0190]$$

and the value 0.02 is not in the interval.

Why are the intervals different? As we have already noted, both intervals are 95% confidence intervals. The two-sided interval is

$$\left[s\sqrt{\frac{n-1}{b}},\ s\sqrt{\frac{n-1}{a}}\right] \quad \text{where } P(U \leq a) = 0.025 = P(U \geq b)$$

whereas the one-sided interval is

$$\left[0,\ s\sqrt{\frac{n-1}{a_1}}\right] \quad \text{where } P\left(U \leq a_1\right) = 0.05$$

Now $a_1 > a$ since $0.05 > 0.025$ so the upper endpoint of the one-sided interval is smaller than the upper endpoint of the two-sided interval. Correspondingly the lower endpoint of the two-sided interval is larger than the lower endpoint of the one-sided interval. The one-sided interval is summarizing all the small values of σ which are supported by the observed data. If we are concerned that σ is too large, then it makes sense to look at all the small values of σ that are supported by the observed data. The two-sided interval is excluding both large and small values of σ that are not supported by the observed data.

Prediction Interval for a Future Observation

In Chapter 3 we mentioned that a common type of statistical problem was a predictive problem in which the experimenter wishes to predict the response of a variate for a given unit. This is often the case in finance or in economics. For example, financial institutions need to predict the price of a stock or interest rates in a week or a month because this affects the value of their investments. We will now show how to do this in the case where the Gaussian model for the data is valid.

Suppose that $y_1, y_2, \ldots, y_n$ is an observed random sample from a $G(\mu, \sigma)$ population and that Y is a new observation which is to be drawn at random from the same $G(\mu, \sigma)$ population. We want to estimate Y and obtain an interval of values for Y. As usual we estimate the unknown parameters μ and σ using $\bar{y}$ and s respectively. The best point estimate of Y based on the data is $\hat{\mu} = \bar{y}$ with corresponding estimator $\tilde{\mu} = \overline{Y} \sim G\left(\mu, \sigma/\sqrt{n}\right)$.

To obtain an interval of values for Y we note that $Y \sim G\left(\mu, \sigma\right)$ independently of $\tilde{\mu} = \overline{Y} \sim G\left(\mu, \sigma/\sqrt{n}\right)$. Since $Y - \overline{Y}$ is a linear combination of independent Gaussian random variables then $Y - \overline{Y}$ also has a Gaussian distribution with mean

$$E\left(Y - \overline{Y}\right) = \mu - \mu = 0$$

and variance

$$Var\left(Y - \overline{Y}\right) = Var\left(Y\right) + Var\left(\overline{Y}\right) = \sigma^2 + \frac{\sigma^2}{n}$$

Since

$$\frac{Y - \overline{Y}}{\sigma\sqrt{1 + \frac{1}{n}}} \sim G\left(0, 1\right)$$

independently of

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2\left(n-1\right)$$

then by Theorem 32

$$\frac{\frac{Y - \overline{Y}}{\sigma\sqrt{1+\frac{1}{n}}}}{\sqrt{S^2/\sigma^2}} = \frac{Y - \overline{Y}}{S\sqrt{1 + \frac{1}{n}}} \sim t\left(n-1\right)$$

is a pivotal quantity which can be used to obtain an interval of values for Y.

Let a be the value such that

$$P(-a \leq T \leq a) = p \quad \text{or} \quad P(T \leq a) = \frac{(1+p)}{2} \qquad \text{where } T \sim t(n-1)$$

which is obtained from the t table or by using R. Since

$$\begin{aligned} p &= P(-a \leq T \leq a) \\ &= P\left(-a \leq \frac{Y - \overline{Y}}{S\sqrt{1+\frac{1}{n}}} \leq a\right) \\ &= P\left(\overline{Y} - aS\sqrt{1+\frac{1}{n}} \leq Y \leq \overline{Y} + aS\sqrt{1+\frac{1}{n}}\right) \end{aligned}$$

therefore

$$\left[\bar{y} - as\sqrt{1+\frac{1}{n}},\ \bar{y} + as\sqrt{1+\frac{1}{n}}\right] \tag{4.19}$$

is an interval of values for the future observation Y with confidence coefficient p. The interval (4.19) is called a $100p\%$ *prediction interval* instead of a confidence interval since Y is not a parameter but a random variable. Note that the interval (4.19) is wider than a $100p\%$ confidence interval for mean μ. This makes sense since μ is an unknown constant with no variability while Y is a random variable with its own variability $Var(Y) = \sigma^2$.

A $100p\%$ prediction interval summarizes a set of values for an unknown future observation (a random variable) based on the observed data. Confidence intervals are for unknown but fixed parameters (not random variables). The procedure for constructing the prediction interval is based on the probability statement

$$P[L(\mathbf{Y}) \leq Y \leq U(\mathbf{Y})] = p \tag{4.20}$$

where Y (a random variable) is the future observation and $\mathbf{Y}$ (a random variable possibly a vector) are the data from an experiment. Suppose you conduct the experiment once and observe the data $\mathbf{y}$. The constructed interval based on the probability statement (4.20) and the observed data $\mathbf{y}$ is $[L(\mathbf{y}), U(\mathbf{y})]$.

To interpret a prediction interval, suppose you conducted the same experiment independently a large number of times and each time you constructed the interval $[L(\mathbf{y}), U(\mathbf{y})]$ based on your observed data $\mathbf{y}$. (Of course $\mathbf{y}$ won't be the same every time you conduct the experiment.) Approximately $100p\%$ of these constructed intervals would contain the future unknown observation. Of course you usually only conduct the experiment once and you only have one interval $[L(\mathbf{y}), U(\mathbf{y})]$. You would that say you are $100p\%$ confident that your constructed interval contains the value of the future observation.

Example 4.7.2 Revisited Optical glass

Suppose in Example 4.7.2 a 95% prediction interval is required for a piece of optical glass drawn at random from the study process.

Since $\bar{y} = 25.009$, $s = 0.013$ and

$$P(T \leq 2.1448) = \frac{1 + 0.95}{2} = 0.975 \quad \text{for } T \sim t\,(14)$$

a 95% prediction interval for this new piece of optical glass is given by

$$\begin{aligned} & 25.009 \pm 2.1448\,(0.013)\sqrt{1 + \frac{1}{15}} \\ = \; & 25.009 \pm 0.0288 \\ = \; & [24.9802, 25.0378] \end{aligned}$$

Note that this interval is much wider than a 95% confidence interval for μ = the mean of the population of lens thicknesses produced by this manufacturing process which is given by

$$\begin{aligned} & 25.009 \pm 2.1448\,(0.013)\,/\sqrt{15} \\ = \; & 25.009 \pm 0.0072 \\ = \; & [25.0018, 25.0162] \end{aligned}$$

4.8 Chapter 4 Summary

Approximate Confidence Intervals based on Likelihood Intervals

A $100p\%$ likelihood interval is defined as $\{\theta : R(\theta) \geq p\}$ where $R(\theta) = R(\theta; \mathbf{y})$ is the relative likelihood function for θ based on observed data $\mathbf{y}$ (possible a vector). Likelihood intervals must usually be found using a numerical method such as the `uniroot` function in R.

A$100p\%$ likelihood interval is an approximate $100q\%$ confidence interval where $q = P(W \leq -2\log p)$ and $W \sim \chi^2(1)$. (Note: $q =$ `pchisq`$(-2 * \log p, 1)$ in R.)

An approximate $100p\%$ confidence interval is given by a $100\left(e^{-b/2}\right)\%$ likelihood interval where b is the value such that $p = P(W \leq b)$ and $W \sim \chi^2(1)$. (Note: $b =$ `qchisq`$(p, 1)$ in R.)

These results are derived from the fact that $-2\log R(\theta; \mathbf{Y})$ is an asymptotic pivotal quantity with approximately a $\chi^2(1)$ distribution.

Table 4.3
Approximate Confidence Intervals for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Observed Data	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Asymptotic Gaussian Pivotal Quantity	Approximate $100p\%$ Confidence Interval
Binomial(n, θ)	y	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
Poisson(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\sqrt{\frac{\tilde{\theta}}{n}}}$	$\hat{\theta} \pm a\sqrt{\frac{\hat{\theta}}{n}}$
Exponential(θ)	$y_1, y_2, \ldots, y_n$	$\bar{y}$	$\overline{Y}$	$\frac{\tilde{\theta}-\theta}{\frac{\tilde{\theta}}{\sqrt{n}}}$	$\hat{\theta} \pm a\frac{\hat{\theta}}{\sqrt{n}}$

Note: The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0, 1)$. In R, $a =$ `qnorm`$\left(\frac{1+p}{2}\right)$

Table 4.4
Confidence/Prediction Intervals for Gaussian and Exponential Models

Model	Unknown Quantity	Pivotal Quantity	$100p\%$ Confidence/Prediction Interval
$G(\mu,\sigma)$ σ known	μ	$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim G(0,1)$	$\bar{y} \pm a\sigma/\sqrt{n}$
$G(\mu,\sigma)$ σ unknown	μ	$\frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\bar{y} \pm bs/\sqrt{n}$
$G(\mu,\sigma)$ μ unknown σ unknown	Y	$\frac{Y-\overline{Y}}{S\sqrt{1+\frac{1}{n}}} \sim t(n-1)$	$100p\%$ Prediction Interval $\bar{y} \pm bs\sqrt{1+\frac{1}{n}}$
$G(\mu,\sigma)$ μ unknown	σ^2	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\frac{(n-1)s^2}{d}, \frac{(n-1)s^2}{c}\right]$
$G(\mu,\sigma)$ μ unknown	σ	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left[\sqrt{\frac{(n-1)s^2}{d}}, \sqrt{\frac{(n-1)s^2}{c}}\right]$
Exponential(θ)	θ	$\frac{2n\overline{Y}}{\theta} \sim \chi^2(2n)$	$\left[\frac{2n\bar{y}}{d_1}, \frac{2n\bar{y}}{c_1}\right]$

Notes: (1) The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
In R, $a = \texttt{qnorm}\left(\frac{1+p}{2}\right)$
(2) The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n-1)$. In R, $b = \texttt{qt}\left(\frac{1+p}{2}, n-1\right)$
(3) The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n-1)$.
In R, $c = \texttt{qchisq}\left(\frac{1-p}{2}, n-1\right)$ and $d = \texttt{qchisq}\left(\frac{1+p}{2}, n-1\right)$
(4) The values c_1 and d_1 are given by $P(W \leq c_1) = \frac{1-p}{2} = P(W > d_1)$ where $W \sim \chi^2(2n)$.
In R, $c_1 = \texttt{qchisq}\left(\frac{1-p}{2}, 2n\right)$ and $d_1 = \texttt{qchisq}\left(\frac{1+p}{2}, 2n\right)$

4.9 Chapter 4 Problems

1. The following R code produces a histogram similar to Figure 4.2.

```
# pop = vector of variate values for the population given in Table 4.1
pop<-c(rep(1,times=210),rep(2,times=127),rep(3,times=66),rep(4,times=39),
rep(5,times=23),rep(6,times=13),rep(7,times=11),rep(8,times=7),
rep(9,times=3),rep(10,times=1))
hist(pop,breaks=seq(1,10,1),col="cyan",main="",xlab="Variate Value")
mu<-mean(pop) # population mean
mu
(499*var(pop)/500)^0.5 # population standard deviation
k<-10000 # number of simulations
n<-15 # sample size
sim<-rep(0,k) # vector to store sample means
# Calculate k sample means for samples of size n drawn from population pop
for (i in 1:k)
sim[i]=mean(sample(pop,n,replace=F))
hist(sim,freq=F,col="cyan",xlab="Sample Mean",main="")
# percentage of times sample mean is within 0.5 of true mean mu
mean(abs(sim-mu)<0.5)
```

   (a) Run the R code and compare the output with the answers in Example 4.2.3.

   (b) Run the R code replacing `n<-15` with `n<-30` and compare the results with those for $n = 15$.

   (c) Explain how the mean, standard deviation and symmetry of the original population affect the histogram of simulated means.

   (d) Explain how the sample size n affects the histogram of simulated means.

2. **R code for plotting a Binomial relative likelihood**
   Suppose for a Binomial experiment we observe $y = 15$ successes in $n = 40$ trials. The following R code will plot the relative likelihood function of θ and the line $R(\theta) = 0.15$ which can be used to determine a 15% likelihood interval.

```
y<-15
n<-40
thetahat<-y/n
theta<-seq(0.15,0.65,0.001)
# points between 0.15 and 0.65 spaced 0.001 apart
Rtheta<-exp(y*log(theta/thetahat)+(n-y)*log((1-theta)/(1-thetahat)))
# plot Binomial relative likelihood function
plot(theta,Rtheta,xlab=expression(theta),
```

```
ylab=expression(paste("R(",theta,")")),type="l",lwd=2,las=1)
title(main="Binomial Relative Likelihood Function")
# draw a horizontal line at 0.15
abline(a=0.15,b=0,col="red",lwd=2)
title(main="Binomial Relative Likelihood Function")
```

Modify this code for $y = 75$ successes in $n = 200$ trials and $y = 150$ successes in $n = 400$ trials and observe what happens to the width of the 15% likelihood interval.

3. R **code for plotting a Poisson relative likelihood**
Suppose we have a sample $y_1, y_2, ..., y_n$ from a Poisson distribution with $n = 25$ and $\bar{y} = 5$. The following R code will plot the relative likelihood function of θ and the line $R(\theta) = 0.15$ which can be used to determine a 15% likelihood interval.

```
thetahat<-5
n<-25
theta<-seq(3.7,6.5,0.001)
Rtheta<-exp(n*thetahat*log(theta/thetahat)+n*(thetahat-theta))
# plot Exponential relative likelihood function
plot(theta,Rtheta,xlab=expression(theta),
ylab=expression(paste("R(",theta,")")),type="l",lwd=2,las=1)
# draw a horizontal line at 0.15
abline(a=0.15,b=0,col="red",lwd=2)
title(main="Poisson Relative Likelihood Function")
```

Modify this code for larger sample sizes $n = 100$ and $n = 400$, and observe what happens to the width of the 15% likelihood interval.

4. For Chapter 2, Problem 4(b) determine a 15% likelihood interval for θ. The likelihood interval can be found from the graph of $R(\theta)$ or by using the function `uniroot` in R.

5. For Chapter 2, Problem 6(b) determine a 15% likelihood interval for θ. The likelihood interval can be found from the graph of $R(\theta)$ or by using the function `uniroot` in R.

6. For Chapter 2, Problem 8(b) determine a 15% likelihood interval for θ. The likelihood interval can be found from the graph of $r(\theta)$ or by using the function `uniroot` in R.

7. (a) For Chapter 2, Problem 10 plot the relative likelihood function $R(\alpha)$ and determine a 10% likelihood interval. The likelihood interval can be found from the graph of $R(\alpha)$ or by using the function `uniroot` in R. How well can α be determined based on these data?

   (b) Suppose that we can find out whether each pair of twins is identical or not, and that it is determined that of 50 pairs, 17 were identical. Obtain the likelihood function, the maximum likelihood estimate and a 10% likelihood interval for α

in this case. Plot the relative likelihood function on the same graph as the one in (a), and compare how well α can be determined based on the two data sets.

8. For Chapter 2, Problem 13(c) determine a 15% likelihood interval for θ. The likelihood interval can be found from the graph of $R(\theta)$ or by using the function `uniroot` in R.

9. Suppose that a fraction θ of a large population of persons are infected with a certain virus. Let n and k be integers. Suppose that blood samples for $n \times k$ people are to be tested to obtain information about θ. In order to save time and money, *pooled testing* is used, that is, samples are mixed together k at a time to give a total of n pooled samples. A pooled sample will test negative if all k individuals in that sample are not infected.

   (a) Find the probability that y out of n samples will be negative, if the nk people are a random sample from the population. State any assumptions you make.

   (b) Obtain a general expression for the maximum likelihood estimate of θ in terms of n, k and y.

   (c) Suppose $n = 100$, $k = 10$ and $y = 89$. Find the maximum likelihood estimate of θ, and a 10% likelihood interval for θ.

10. Company A leased photocopiers to the federal government, but at the end of their recent contract the government declined to renew the arrangement and decided to lease from a new vendor, Company B. One of the main reasons for this decision was a perception that the reliability of Company A's machines was poor.

    (a) Over the preceding year the monthly numbers of failures requiring a service call from Company A were

    12 14 15 16 18 19 19 22 23 25 28 29

    Assuming that the number of service calls needed in a one month period has a Poisson distribution with mean θ, obtain and graph the relative likelihood function $R(\theta)$ based on the data above.

    (b) In the first year using Company B's photocopiers, the monthly numbers of service calls were

    7 8 9 10 10 12 12 13 13 14 15 17

    Under the same assumption as in part (a), obtain $R(\theta)$ for these data and graph it on the same graph as used in (a).

    (c) Determine the 15% likelihood interval for θ for each company. The intervals can be obtained from the graphs of the relative likelihood functions or by using the function `uniroot` in R. Do you think the government's decision was a good one, as far as the reliability of the machines is concerned? Justify your answer.

(d) What conditions would need to be satisfied to make the assumptions and analysis in (a) to (c) valid?

11. Suppose that a fraction θ of a large population of persons over 18 years of age never drink alcohol. In order to estimate θ, a random sample of n persons is to be selected and the number y who do not drink determined; the maximum likelihood estimate of θ is then $\hat{\theta} = y/n$. We want our estimate $\hat{\theta}$ to have a high probability of being close to θ, and want to know how large n should be to achieve this. Consider the random variable Y and the estimator $\tilde{\theta} = Y/n$.

(a) Determine $P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right)$, if $n = 1000$ and $\theta = 0.5$ using the Normal approximation to the Binomial. You do not need to use a continuity correction.

(b) If $\theta = 0.50$ determine how large n should be to ensure that

$$P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) = P\left(\left|\tilde{\theta} - \theta\right| \leq 0.03\right) \geq 0.95$$

(c) If θ is unknown determine how large n should be to ensure that

$$P\left(-0.03 \leq \tilde{\theta} - \theta \leq 0.03\right) = P\left(\left|\tilde{\theta} - \theta\right| \leq 0.03\right) \geq 0.95$$

for all $\theta \in [0, 1]$.

12. The following R code generates k approximate Binomial(n, θ) confidence intervals based on the Gaussian asymptotic pivotal quantity and determines the proportion which contain the true value of θ:

```
n<-30        # n = number of trials
theta<-0.25  # value of theta
k<-1000      # number of confidence intervals generated
a<-1.96      # a = 1.96 for approximate 95% confidence interval
that<-rbinom(k,n,theta)/n # vector of thetahat's for k simulations
pm<-a*(that*(1-that)/n)^0.5 # used to get confidence interval
# each confidence interval is stored in a row of matrix int
int<-matrix(c(that-pm,that+pm),nrow=k,byrow=F)
# Look at first 25 intervals to see how variable intervals are
int[1:25,1:2]
# proportion of intervals which contain the true value theta
mean(abs(theta-that)<pm)
```

(a) Run this code to determine the proportion of approximate 95% confidence intervals which contain the true value.

(b) Run this code for `n<-100` and `n<-1000` and observe what happens to the proportion.

(c) Run this code for `theta<-0.1` and observe what happens to the proportion.

13. The following excerpt is from a March 2, 2012 cbc.ca news article:
"**Canadians lead in time spent online:** Canadians are spending more time online than users in 10 other countries, a new report has found. The report, *2012 Canada Digital Future in Focus*, by the internet marketing research company comScore, found Canadians spent an average of 45.3 hours on the internet in the fourth quarter of 2011. The report also states that smartphones now account for 45% of all mobile phone use by Canadians."
Assume that these results are based on a random sample of 1000 Canadians.

    (a) Suppose a 95% confidence interval for μ, the mean time Canadians spent on the internet in this quarter, is reported to be $[42.8, 47.8]$. How should this interval be interpreted?

    (b) Construct an approximate 95% confidence interval for the proportion of Canadians whose mobile phone is a smartphone.

    (c) Since this study was conducted in March 2012 the research company has been asked to conduct a new survey to determine if the proportion of Canadians whose mobile phone is a smartphone has changed. What size sample should be used to ensure that the width of an approximate 95% confidence interval is less than $2\,(0.02)$?

14. Two hundred adults are chosen at random from a population and each adult is asked whether information about abortions should be included in high school public health sessions. Suppose that 70% say they should.

    (a) Obtain an approximate 95% confidence interval for the proportion θ of the population who support abortion information included in high school public health sessions.

    (b) Suppose you found out that the 200 persons interviewed consisted of 50 married couples and 100 other persons. The 50 couples were randomly selected, as were the other 100 persons. Discuss the validity (or non-validity) of the analysis in (a).

15. In the United States. the prevalence of HIV (Human Immunodeficiency Virus) infections in the population of child-bearing women has been estimated by doing blood tests (anonymously) on all women giving birth in a hospital. One study tested $29,000$ women and found that 64 were HIV positive (had the virus). Give an approximate 99% confidence interval for θ, the fraction of the population that is HIV positive. State any concerns you have about the accuracy of this estimate.

16. If $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Poisson(θ) distribution then by the Central Limit Theorem and other limit theorems the random variable

$$\frac{\overline{Y} - \theta}{\sqrt{\overline{Y}/n}}$$

has approximately a $G(0,1)$ distribution.

(a) Show how this asymptotic pivotal quantity leads to an approximate 95% confidence interval for θ given by

$$\bar{y} \pm 1.96\sqrt{\frac{\bar{y}}{n}}$$

(b) Use the result from (a) to construct an approximate 95% confidence interval for θ in Chapter 2, Problem 11.

(c) Compare the approximate 95% confidence interval for θ with a 15% likelihood interval. What do you notice?

17. Use the result from Problem 16 to determine approximate 95% confidence intervals for θ for each company in Problem 10. Compare these intervals with the 15% likelihood intervals obtained in Problem 10.

18. A manufacturing process produces fibers of varying lengths. The length of a fiber Y is a continuous random variable with probability density function

$$f(y;\theta) = \frac{y}{\theta^2}e^{-y/\theta} \quad \text{for } y \geq 0 \text{ and } \theta > 0$$

where θ is an unknown parameter.

(a) If Y has probability density function $f(y;\theta)$ show that $E(Y) = 2\theta$ and $Var(Y) = 2\theta^2$. **Hint:** Use the Gamma function.

(b) Let $y_1, y_2, \ldots, y_n$ be the lengths of n fibers selected at random. Find the maximum likelihood estimate of θ based on these data.

(c) Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random variables with probability density function $f(y;\theta)$ given above. Find $E\left(\overline{Y}\right)$ and $Var\left(\overline{Y}\right)$ using the result in (a).

(d) Justify the statement

$$P\left(-1.96 \leq \frac{\overline{Y} - 2\theta}{\theta\sqrt{2/n}} \leq 1.96\right) \approx 0.95$$

(e) Show how you would use the statement in (d) to construct an approximate 95% confidence interval for θ.

(f) Suppose $n = 18$ fibers were selected at random and the lengths were:

6.19	7.92	1.23	8.13	4.29	1.04	3.67	9.87	10.34
1.41	10.76	3.69	1.34	6.80	4.21	3.44	2.51	2.08

For these data $\sum_{i=1}^{18} y_i = 88.92$. Give the maximum likelihood estimate of θ and an approximate 95% confidence interval for θ using your result from (e).

19. The lifetime T (in days) of a particular type of light bulb is assumed to have a distribution with probability density function

$$f(t;\theta) = \frac{1}{2}\theta^3 t^2 e^{-\theta t} \quad \text{for } t > 0 \text{ and } \theta > 0$$

(a) Suppose $t_1, t_2, \ldots, t_n$ is a random sample from this distribution. Find the maximum likelihood estimate $\hat{\theta}$ and the relative likelihood function $R(\theta)$.

(b) If $n = 20$ and $\sum_{i=1}^{20} t_i = 996$, graph $R(\theta)$ and determine the 15% likelihood interval for θ which is also an approximate 95% confidence interval for θ. The interval can be obtained from the graph of $R(\theta)$ or by using the function `uniroot` in R.

(c) Suppose we wish to estimate the mean lifetime of a light bulb. Show $E(T) = 3/\theta$. **Hint:** Use the Gamma function. Find an approximate 95% confidence interval for the mean.

(d) Show that the probability p that a light bulb lasts less than 50 days is

$$\begin{aligned} p &= p(\theta) \\ &= P(T \leq 50;\theta) \\ &= 1 - e^{-50\theta}[1250\theta^2 + 50\theta + 1] \end{aligned}$$

Determine the maximum likelihood estimate of p. Find an approximate 95% confidence interval for p from the approximate 95% confidence interval for θ. For the data referred to in (b), the number of light bulbs which lasted less than 50 days was 11 (out of 20). Using a Binomial model, obtain an approximate 95% confidence interval for p. What are the pros and cons of the second interval over the first one?

20. **The Chi-squared distribution**

(a) Use the Chi-squared table provided at the end of these Course Notes to answer the following:

(i) If $X \sim \chi^2(10)$ find $P(X \leq 2.6)$ and $P(X > 16)$.

(ii) If $X \sim \chi^2(4)$ find $P(X > 15)$.

(iii) If $X \sim \chi^2(40)$ find $P(X \leq 24.4)$ and $P(X \leq 55.8)$. Compare these values with $P(Y \leq 24.4)$ and $P(Y \leq 55.8)$ if $Y \sim N(40, 80)$.

(iv) If $X \sim \chi^2(25)$ find a and b such that $P(X \leq a) = 0.025$ and $P(X > b) = 0.025$.

(v) If $X \sim \chi^2(12)$ find a and b such that $P(X \leq a) = 0.05$ and $P(X > b) = 0.05$.

(b) Use the R functions `pchisq(x,k)` and `qchisq(p,k)` to check the values in (a).

(c) Determine the following WITHOUT using the Chi-squared table:

(i) If $X \sim \chi^2(1)$ find $P(X \leq 2)$ and $P(X > 1.4)$.

(ii) If $X \sim \chi^2(2)$ find $P(X \leq 2)$ and $P(X > 3)$.

(d) If $X \sim G(3, 2)$ and $Y_i \sim Exponential(2)$, $i = 1, 2, \ldots, 5$ all independently then what is the distribution of $W = \sum_{i=1}^{5} Y_i + \left(\frac{X-3}{2}\right)^2$?

(e) If $X_i \sim \chi^2(i)$, $i = 1, 2, \ldots, 10$ independently then what is the distribution of $\sum_{i=1}^{10} X_i$?

21. **Properties of the Chi-squared distribution** Suppose $X \sim \chi^2(k)$ with probability density function given by

$$f(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad \text{for } x > 0$$

(a) Show that this probability density function integrates to one for $k = 1, 2, \ldots$ using the properties of the Gamma function.

(b) Plot the probability density function for $k = 5$, $k = 10$ and $k = 25$ on the same graph. What do you notice?

(c) Show that the moment generating function of Y is given by

$$\begin{aligned} M(t) &= E\left(e^{tX}\right) \\ &= (1-2t)^{-k/2} \quad \text{for } t < \frac{1}{2} \end{aligned}$$

and use this to show that $E(X) = k$ and $Var(X) = 2k$.

(d) Prove Theorem 29 using moment generating functions.

22. **Student's** t distribution Suppose $T \sim t(k)$.

    (a) Plot the probability density function for $k = 1, 5, 25$. Plot the $N(0,1)$ probability density function on the same graph. What do you notice?

    (b) Show that $f(t;k)$ is unimodal.

    (c) Use Theorem 32 to show that $E(T) = 0$. **Hint:** If X and Y are independent random variables then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$.

    (d) Use the t table provided at the end of these Course Notes to answer the following:

        (i) If $T \sim t(10)$ find $P(T \leq 0.88)$, $P(T \leq -0.88)$ and $P(|T| \leq 0.88)$.

        (ii) If $T \sim t(17)$ find $P(|T| > 2.90)$.

        (iii) If $T \sim t(30)$ find $P(T \leq -2.04)$ and $P(T \leq 0.26)$. Compare these values with $P(Z \leq -2.04)$ and $P(Z \leq 0.26)$ if $Z \sim N(0,1)$.

        (iv) If $T \sim t(18)$ find a and b such that $P(T \leq a) = 0.025$ and $P(T > b) = 0.025$.

        (v) If $T \sim t(13)$ find a and b such that $P(T \leq a) = 0.05$ and $P(T > b) = 0.05$.

    (e) Use the R functions `pt(x,k)` and `qt(p,k)` to check the values in (d).

23. **Limiting** t distribution Suppose $T \sim t(k)$ with probability density function

$$f(t;k) = c_k \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad \text{for } t \in \Re \text{ and } k = 1, 2, \ldots$$

where

$$c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)}$$

Show that

$$\lim_{k\to\infty} f(t;k) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \quad \text{for } t \in \Re$$

which is the probability density function of the $G(0,1)$ distribution. **Hint:** You may use the fact that $\lim_{k\to\infty} c_k = 1/\sqrt{2\pi}$ which is a property of the Gamma function.

24. Suppose $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$ independently and let $W_i = Y_i - \overline{Y}$, $i = 1, 2, \ldots, n$.

    (a) Show that W_i, $i = 1, 2, \ldots, n$ can be written as a linear combination of independent Normal random variables.

    (b) Show that $E(W_i) = 0$ and $Var(W_i) = \sigma^2\left(1 - \frac{1}{n}\right)$, $i = 1, 2, \ldots, n$. **Hint:** Show $Cov\left(Y_i, \overline{Y}\right) = \frac{\sigma^2}{n}$, $i = 1, 2, \ldots, n$. Note that this result along with the result in (a) implies that

$$W_i = Y_i - \overline{Y} \sim G\left(0, \sigma\sqrt{1 - \frac{1}{n}}\right), \quad i = 1, 2, \ldots, n$$

(c) Show that $Cov(W_i, W_j) = -\frac{\sigma^2}{n}$, for all $i \neq j$ which implies that the $W_i's$ are correlated random variable and therefore not independent random variables.

25. If $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Exponential(θ) distribution then $E(\overline{Y}) = \theta$ and $Var(\overline{Y}) = \theta^2/n$. By the Central Limit Theorem the random variable

$$\frac{\overline{Y} - \theta}{\theta/\sqrt{n}}$$

has approximately a $G(0,1)$ distribution. It also follows that

$$Q = \frac{\overline{Y} - \theta}{\overline{Y}/\sqrt{n}}$$

has approximately a $G(0,1)$ distribution. Show how the asymptotic pivotal quantity Q leads to an approximate $100p\%$ confidence interval for θ given by

$$\bar{y} \pm a\frac{\bar{y}}{\sqrt{n}}$$

where $P(Z \leq a) = (1+p)/2$ and $Z \sim G(0,1)$.

26. In an early study concerning survival time for patients diagnosed with Acquired Immune Deficiency Syndrome (AIDS), the survival times (i.e. times between diagnosis of AIDS and death) of 30 male patients were such that $\sum_{i=1}^{30} y_i = 11,400$ days. Assume that the survival times are Exponentially distributed with mean θ days.

(a) Use the result in Problem 25 to obtain an approximate 90% confidence interval for θ.

(b) Graph the relative likelihood function for these data and obtain an approximate 90% likelihood based confidence interval for θ. Compare this with the interval obtained in (a).

(c) Show that $m = \theta \ln 2$ is the median survival time. Give an approximate 90% confidence interval for m based on your interval from (b).

27. **Exact confidence intervals for θ for Exponential data**

(a) If $Y \sim$ Exponential(θ) then show that $W = 2Y/\theta$ has a $\chi^2(2)$ distribution. (**Hint:** compare the probability density function of W with (4.9)).

(b) Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Exponential(θ) distribution. Use the results of Section 4.5 to prove that

$$U = 2\sum_{i=1}^{n} \frac{Y_i}{\theta} \sim \chi^2(2n)$$

This result implies that U is a pivotal quantity.

(c) Show how the pivotal quantity U can be used to construct an exact confidence interval for θ.

(d) Refer to the data in Problem 26. Obtain an exact 90% confidence interval for θ based on the pivotal quantity U. Compare this with the approximate confidence intervals for θ obtained in Problem 26.

28. Suppose the model $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$ independently is assumed where μ is a known value. Show that

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

is a pivotal quantity. Show how this pivotal quantity can be used to construct a $100p\%$ confidence interval for σ^2 and σ.

29. A study on the common octopus (Octopus Vulgaris) was conducted by researchers at the University of Vigo in Vigo, Spain. Nineteen octopi were caught in July 2008 in the Ria de Vigo (a large estuary on the northwestern coast of Spain). Several measurements were made on each octopus including their weight in grams. These weights are given in the table below.

680	1030	1340	1330	1260	770	830	1470	1380	1220
920	880	1020	1050	1140	960	1060	1140	860	

Let y_i = weight of the i'th octopus, $i = 1, 2, \ldots, 19$. For these data

$$\sum_{i=1}^{19} y_i = 20340 \quad \text{and} \quad \sum_{i=1}^{19} (y_i - \bar{y})^2 = 884095$$

To analyze these data the model $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 19$ independently is assumed where μ and σ are unknown parameters.

(a) Use a qqplot to determine how reasonable the Gaussian model is for these data.

(b) Describe a suitable study population for this study. The parameters μ and σ correspond to what attributes of interest in the study population?

(c) The researchers at the University of Vigo were interested in determining whether the octopi in the Ria de Vigo are healthy. For common octopi, a population mean weight of 1100 grams is considered to be a healthy population. Determine a 95% confidence interval for μ. What should the researchers conclude about the health of the octopi, in terms of weight, in the Ria de Vigo?

(d) Determine a 90% confidence interval for σ based on these data.

30. Consider the data on weights of adult males and females from Chapter 1. The data are available in the file *bmidata.txt* posted on the course website.

    (a) Determine whether it is reasonable to assume a Gaussian model for the female heights and a different Gaussian model for the male heights.

    (b) Obtain a 95% confidence interval for the mean for the females and males separately. Does there appear to be a difference in the means for females and males? (We will see how to test this formally in Chapter 6.)

    (c) Obtain a 95% confidence interval for the standard deviation for the females and males separately. Does there appear to be a difference in the standard deviations?

31. Sixteen packages are randomly selected from the production of a detergent packaging machine. Let y_i = weight in grams of the i'th package, $i = 1, 2, \ldots, 16$.

287	293	295	295	297	298	299	300
300	302	302	303	306	307	308	311

    For these data

    $$\sum_{i=1}^{16} y_i = 4803 \quad \text{and} \quad \sum_{i=1}^{16} {y_i}^2 = 1442369$$

    To analyze these data the model $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 16$ independently is assumed where μ and σ are unknown parameters.

    (a) Describe a suitable study population for this study. The parameters μ and σ correspond to what attributes of interest in the study population?

    (b) Obtain 95% confidence intervals for μ and σ.

    (c) Let Y represent the weight of a future, independent, randomly selected package. Obtain a 95% prediction interval for Y.

32. Radon is a colourless, odourless gas that is naturally released by rocks and soils and may concentrate in highly insulated houses. Because radon is slightly radioactive, there is some concern that it may be a health hazard. Radon detectors are sold to homeowners worried about this risk, but the detectors may be inaccurate. University researchers in Waterloo purchased 12 detectors of the same type at Home Depot. The detectors were placed in a chamber where they were exposed to 105 picocuries per liter of radon over 3 days. The readings given by the detectors were:

91.9	97.8	111.4	122.3	105.4	95.0	103.8	99.6	96.6	119.3	104.8	101.7

    Let y_i = reading for the i'th detector, $i = 1, 2, \ldots, 12$. For these data

    $$\sum_{i=1}^{12} y_i = 1249.6 \quad \text{and} \quad \sum_{i=1}^{12} (y_i - \bar{y})^2 = 971.4267$$

    To analyze these data assume the model $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 12$ independently where μ and σ are unknown parameters.

(a) Describe a suitable study population for this study. The parameters μ and σ correspond to what attributes of interest in the study population?

(b) Obtain a 95% confidence interval for μ. Does it contain the value $\mu = 105$?

(c) Obtain a 95% confidence interval for σ.

(d) As a statistician what would you say to the university researchers about the accuracy and precision of the detectors?

(e) University researchers purchased one more radon detector. It is to be exposed to 105 picocuries per liter of radon over 3 days. Calculate a 95% prediction interval for the reading for this new radon detector.

(f) Suppose the researchers wanted to determine the mean level of radon detected by the radon detectors to "within 3 picocurie per liter". As a statistician we would interpret this as requiring that the 95% confidence interval for a μ should have width at most 6. How many detectors in total would you advise the researchers to test?

33. A chemist has two ways of measuring a particular quantity; one has more random error than the other. For method I, measurements $X_1, X_2, \ldots, X_m$ follow a Gaussian distribution with mean μ and standard deviation σ_1, whereas for method II, measurements $Y_1, Y_2, \ldots, Y_n$ have a Gaussian distribution with mean μ and variance σ_2.

(a) Assuming that σ_1 and σ_2 are known, find the combined likelihood function for μ based on observed data $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ and show that the maximum likelihood estimate of μ is

$$\hat{\mu} = \frac{w_1\bar{x} + w_2\bar{y}}{w_1 + w_2}$$

where $w_1 = m/\sigma_1^2$ and $w_2 = n/\sigma_2^2$. Why does this estimate make sense?

(b) Suppose that $\sigma_1 = 1$, $\sigma_2 = 0.5$ and $n = m = 10$. How would you rationalize to a non-statistician why you were using the estimate $(\bar{x} + 4\bar{y})/5$ instead of $(\bar{x} + \bar{y})/2$?

(c) Suppose that $\sigma_1 = 1$, $\sigma_2 = 0.5$ and $n = m = 10$, determine the standard deviation of the maximum likelihood estimator

$$\tilde{\mu} = \frac{w_1\bar{X} + w_2\overline{Y}}{w_1 + w_2}$$

and the estimator $(\bar{X} + \overline{Y})/2$. Why is $\tilde{\mu}$ a better estimator?

34. **Challenge Problem** For "two-sided" intervals based on the t distribution, we usually pick the interval which is symmetrical about $\bar{y}$. Show that this choice provides the *shortest* $100p\%$ confidence interval.

35. **Challenge Problem** A sequence of random variables $\{X_n\}$ is said to *converge in probability* to the constant c if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - c| \geq \epsilon\right) = 0$$

We denote this by writing $X_n \xrightarrow{p} c$.

(a) If $\{X_n\}$ and $\{Y_n\}$ are two sequences of random variables with $X_n \xrightarrow{p} c_1$ and $Y_n \xrightarrow{p} c_2$, show that $X_n + Y_n \xrightarrow{p} c_1 + c_2$ and $X_n Y_n \xrightarrow{p} c_1 c_2$.

(b) Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with probability density function $f(x;\theta)$. A point estimator $\tilde{\theta}_n$ based on a random sample $X_1, X_2, \ldots, X_n$ is said to be *consistent* for θ if $\tilde{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

(i) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed Uniform$(0, \theta)$ random variables. Show that $\tilde{\theta}_n = \max\left(X_1, X_2, \ldots, X_n\right)$ is consistent for θ.

(ii) Let $X \sim$ Binomial(n, θ). Show that $\tilde{\theta}_n = X/n$ is consistent for θ.

36. **Challenge Problem** Refer to the definition of consistency in Problem 32(b). Difficulties can arise when the number of parameters increases with the amount of data. Suppose that two independent measurements of blood sugar are taken on each of n individuals and consider the model

$$X_{i1}, X_{i2} \sim N(\mu_i, \sigma^2) \quad \text{for } i = 1, 2, \ldots, n$$

where X_{i1} and X_{i2} are the independent measurements. The variance σ^2 is to be estimated, but the μ_i's are also unknown.

(a) Find the maximum likelihood estimator $\tilde{\sigma}^2$ and show that it is not consistent.

(b) Suggest an alternative way to estimate σ^2 by considering the differences $W_i = X_{i1} - X_{i2}$.

(c) What does σ represent physically if the measurements are taken very close together in time?

37. **Challenge Problem Proof of Central Limit Theorem** (Special Case) Suppose $Y_1, Y_2, \ldots$ are independent random variables with $E(Y_i) = \mu$, $Var(Y_i) = \sigma^2$ and that they have the same distribution, whose moment generating function exists.

(a) Show that $(Y_i - \mu)/\sigma$ has moment generating function of the form $(1 + \frac{t^2}{2} +$ terms in $t^3, t^4, \ldots)$ and thus that $(Y_i - \mu)/\sqrt{n}\sigma$ has moment generating function of the form $\left[1 + \frac{t^2}{2n} + o(n)\right]$, where $o(n)$ signifies a remainder term R_n with the property that $R_n/n \to 0$ as $n \to \infty$.

(b) Let

$$Z_n = \sum_{i=1}^{n} \frac{(Y_i - \mu)}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma}$$

and note that its moment generating function is of the form $\left[1 + \frac{t^2}{2n} + o(n)\right]^n$. Show that as $n \to \infty$ this approaches the limit $e^{t^2/2}$, which is the moment generating function for a $G(0, 1)$ random variable. (**Hint:** For any real number a, $(1 + a/n)^n \to e^a$ as $n \to \infty$.)

5. HYPOTHESIS TESTING

5.1 Introduction

What does it mean to test a hypothesis in the light of observed data or information? Suppose a statement has been formulated such as "I have extrasensory perception." or "This drug that I developed reduces pain better than those currently available." and an experiment is conducted to determine how credible the statement is in light of observed data. How do we measure credibility? If there are two alternatives: "*I have ESP.*" and "*I do not have ESP.*" should they both be considered *a priori* as equally plausible? If I correctly guess the outcome on 52 of 100 tosses of a fair coin, would you conclude that my gift is real since I was correct more than 50% of the time? If I develop a treatment for pain in my basement laboratory using a mixture of seaweed and tofu, would you treat the claims "*this product is superior to aspirin*" and "*this product is no better than aspirin*" symmetrically?

To understand a test of hypothesis it is helpful to draw an analogy with the criminal court system used in many places in the world, where the two hypotheses "*the defendant is innocent*" and "*the defendant is guilty*" are **not** treated symmetrically. In these courts, the court assumes *a priori* that the first hypothesis, "*the defendant is innocent*" is true, and then the prosecution attempts to find sufficient evidence to show that this hypothesis of innocence is not plausible. There is no requirement that the defendant be proved innocent. At the end of the trial the judge or jury may conclude that there was insufficient evidence for a finding of guilty and the defendant is then exonerated. Of course there are two types of errors that this system can (and inevitably does) make; convict an innocent defendant or fail to convict a guilty defendant. The two hypotheses are usually not given equal weight *a priori* because these two errors have very different consequences.

A test of hypothesis is analogous to this legal example. We often begin by specifying a single "default" hypothesis ("the defendant is innocent" in the legal context) and then check whether the data collected is unlikely under this hypothesis. This default hypothesis is often referred to as the "null" hypothesis and is denoted by H_0 ("null" is used because it often means a new treatment has no effect). Of course, there is an alternative hypothesis H_A, which may not always be specified. In many cases H_A is simply that H_0 *is not true.*

We will outline the logic of a test of hypothesis in the first example, the claim that I have ESP. In an effort to prove or disprove this claim, an unbiased observer tosses a fair coin

100 times and before each toss I guess the outcome of the toss. We count Y, the number of correct guesses which we can assume has a Binomial distribution with $n = 100$. The probability that I guess the outcome correctly on a given toss is an unknown parameter θ. If I have no unusual ESP capacity at all, then we would assume $\theta = 0.5$, whereas if I have some form of ESP, either a positive attraction or an aversion to the correct answer, then we expect $\theta \neq 0.5$. We begin by asking the following questions in this context:

(1) Which of the two possibilities, $\theta = 0.5$ or $\theta \neq 0.5$, should be assigned to H_0, the null hypothesis?

(2) What observed values of Y are highly inconsistent with H_0 and what observed values of Y are compatible with H_0?

(3) What observed values of Y would lead to us to conclude that the data provide no evidence against H_0 and what observed values of Y would lead us to conclude that the data provide strong evidence against H_0?

In answer to question (1), hopefully you observed that these two hypotheses ESP and NO ESP are not equally credible and decided that the null hypothesis should be $H_0 : \theta = 0.5$ or H_0 : I do not have ESP.

To answer question (2), we note that observed values of Y that are very small (e.g. $0 - 10$) or very large (e.g. $90 - 100$) would clearly lead us to to believe that H_0 is false, whereas values near 50 are perfectly consistent with H_0. This leads naturally to the concept of a *test statistic* or *discrepancy measure.*

Definition 38 *A test statistic or discrepancy measure D is a function of the data $\mathbf{Y}$ that is constructed to measure the degree of "agreement" between the data $\mathbf{Y}$ and the null hypothesis H_0.*

Usually we define D so that $D = 0$ represents the best possible agreement between the data and H_0, and values of D not close to 0 indicate poor agreement. A general method for constructing test statistics will be described in Sections 5.3, but in this example, it seems natural to use $D(Y) = |Y - 50|$.

Question (3) could be resolved easily if we could specify a threshold value for D, or equivalently some function of D. In the given example, the observed value of Y was $y = 52$ and so the observed value of D is $d = |52 - 50| = 2$. One might ask what is the probability, when H_0 is true, that the discrepancy measure results in a value less than d. Equivalently, what is the probability, assuming H_0 is true, that the discrepancy measure is greater than or equal to d? In other words we want to determine $P(D \geq d; H_0)$ where the notation "$; H_0$" means "assuming that H_0 is true". We can compute this probability easily for this

example. If H_0 is true then $Y \sim \text{Binomial}(100, 0.5)$ and

$$\begin{aligned} P(D \geq d; H_0) &= P(|Y - 50| \geq |52 - 50|; H_0) \\ &= P(|Y - 50| \geq 2) \quad \text{where } Y \sim \text{Binomial}(100, 0.5) \\ &= 1 - P(49 \leq Y \leq 51) \\ &= 1 - \binom{100}{49}(0.5)^{100} - \binom{100}{50}(0.5)^{100} - \binom{100}{51}(0.5)^{100} \\ &\approx 0.76 \end{aligned}$$

How can we interpret this value in terms of the test of H_0? Roughly 76% of claimants similarly tested for ESP, who have no abilities at all but simply randomly guess, will perform as well or better (that is, result in at least as large a value of D as the observed value of 2) than I did. This does not prove I do not have ESP but it does indicate we have failed to find any evidence in these data to support rejecting H_0. There is no evidence against H_0 in the observed value $d = 2$, and this was indicated by the high probability that, when H_0 is true, we obtain at least this much measured disagreement with H_0.

We now proceed to a more formal treatment of a test of hypothesis. We will concentrate on two types of hypotheses:

(1) the hypothesis $H_0 : \theta = \theta_0$ where it is assumed that the data $\mathbf{Y}$ have arisen from a family of distributions with probability (density) function $f(\mathbf{y}; \theta)$ with parameter θ

(2) the hypothesis $H_0 : Y \sim f_0(y)$ where it is assumed that the data $\mathbf{Y}$ have a specified probability (density) function $f_0(y)$.

The ESP example is an example of a type (1) hypothesis. If we wish to determine if is reasonable to assume a given data set is a random sample from an Exponential(1) distribution then this is an example of a type (2) hypothesis. We will see more examples of type (2) hypotheses in Chapter 7.

A statistical test of hypothesis proceeds as follows. First, assume that the hypothesis H_0 will be tested using some random data $\mathbf{Y}$. We then adopt a test statistic or discrepancy measure $D(\mathbf{Y})$ for which, normally, large values of D are less consistent with H_0. Let $d = D(\mathbf{y})$ be the corresponding observed value of D. We then calculate the *p-value* or *observed significance level of the test.*

Definition 39 *Suppose we use the test statistic $D = D(\mathbf{Y})$ to test the hypothesis H_0. Suppose also that $d = D(\mathbf{y})$ is the observed value of D. The p-value or observed significance level of the test of hypothesis H_0 using test statistic D is*

$$p - value = P(D \geq d; H_0)$$

In other words, the $p - value$ is the probability, calculated assuming H_0 is true, of observing a value of the test statistic greater than or equal to the observed value of the

test statistic. If d, the observed value of D, is large and consequently the $p-value$ is small then one of the following two statements is correct:

(1) H_0 is true but by chance we have observed an outcome that does not happen very often when H_0 is true

or

(2) H_0 is false.

If the $p-value$ is close to 0.05, then the event of observing a D value as unusual or more unusual as we have observed happens only 5 times out of 100, that is, not very often. Therefore we interpret a $p-value$ close to 0.05 as indicating that the observed data are providing evidence against H_0. If the $p-value$ is very small, for example less than 0.001, then the event of observing a D value as unusual or more unusual as we have observed happens only 1 time out of 1000, that is, very rarely. Therefore we interpret a $p-value$ close to 0.001 as indicating that the observed data are providing strong evidence against H_0. If the $p-value$ is greater than 0.1, then the event of observing a D value as unusual or more unusual as we have observed happens more than 1 time out of 10, that is, fairly often and therefore the observed data are consistent with H_0.

Remarks

(1) Note that the $p-value$ is defined as $P(D \geq d; H_0)$ and not $P(D = d; H_0)$ even though the event that has been observed is $D = d$. If D is a continuous random variable then $P(D = d; H_0)$ is always equal to zero which is not very useful. If D is a discrete random variable with many possible values then $P(D = d; H_0)$ will be small which is also not very useful. Therefore to determine how unusual the observed result is we compare it to all the other results which are as unusual or more unusual than what has been observed.

(2) The $p-value$ is NOT the probability that H_0 is true. This is a common misinterpretation.

The following table gives a rough guideline for interpreting $p-values$. *These are only guidelines for this course. The interpretation of* $p-values$ *must always be made in the context of a given study.*

Table 5.1: Guidelines for interpreting $p-values$

$p-value$	Interpretation
$p-value > 0.10$	No evidence against H_0 based on the observed data.
$0.05 < p-value \leq 0.10$	Weak evidence against H_0 based on the observed data.
$0.01 < p-value \leq 0.05$	Evidence against H_0 based on the observed data.
$0.001 < p-value \leq 0.01$	Strong evidence against H_0 based on the observed data.
$p-value \leq 0.001$	Very strong evidence against H_0 based on the observed data.

Example 5.1.1 Test of hypothesis for Binomial for large n

Suppose that in the ESP experiment the coin was tossed $n = 200$ times and I correctly guessed 110 of the outcomes. In this case we use the test statistic $D = |Y - 100|$ with observed value $d = |110 - 100| = 10$. The $p - value$ is

$$p - value = P\left(|Y - 100| \geq 10\right) \quad \text{where } Y \sim \text{Binomial}(200, 0.5)$$

which can be calculated using R or using the Normal approximation to the Binomial since $n = 200$ is large. Using the Normal approximation (without a continuity correction since it is not essential to have an exact $p - value$) we obtain

$$\begin{aligned} p - value &= P\left(|Y - 100| \geq 10\right) \quad \text{where } Y \sim \text{Binomial}(200, 0.5) \\ &= P\left(\frac{|Y - 100|}{\sqrt{200\,(0.5)\,(0.5)}} \geq \frac{10}{\sqrt{200\,(0.5)\,(0.5)}}\right) \\ &\approx P\left(|Z| \geq 1.41\right) \quad \text{where } \; Z \sim N\left(0, 1\right) \\ &= 2\left[1 - P\left(Z \leq 1.41\right)\right] \\ &= 2\left(1 - 0.92073\right) \\ &= 0.15854 \end{aligned}$$

Since the $p - value$ is greater than 0.1 we would conclude there is no evidence against the hypothesis that I was guessing based on the observed data.

Example 5.1.2 One-sided test of hypothesis for Binomial model

Suppose that it is suspected that a 6-sided die has been "doctored" so that the number one turns up **more often** than if the die were fair. Let $\theta = P\left(\text{die turns up one}\right)$ on a single toss and consider the hypothesis $H_0 : \theta = 1/6$. To test H_0, we toss the die n times and observe the number of times Y that a one occurs. Assuming $H_0 : \theta = 1/6$ is true, $Y \sim \text{Binomial}(n, 1/6)$ distribution. If we only wanted to focus on the alternative hypothesis $H_A : \theta > 1/6$ then a reasonable test statistic would be $D = \max\left[(Y - n/6), 0\right]$. The hypothesis $H_A : \theta > 1/6$ is an example of a one-sided alternative.

Suppose that $n = 180$ tosses gave $y = 44$, then the observed value of D is $d = \max\left[(44 - 180/6), 0\right] = 14$ and the $p - value$, calculated using R, is

$$\begin{aligned} p - value &= P(D \geq 14; H_0) \\ &= P(Y \geq 44) \quad \text{where } Y \sim \text{Binomial}\left(180, \frac{1}{6}\right) \\ &= \sum_{y=44}^{180} \binom{180}{y} \left(\frac{1}{6}\right)^{y} \left(\frac{5}{6}\right)^{180-y} \\ &= 0.005 \end{aligned}$$

Since the $p - value$ is between 0.001 and 0.01, we would conclude there is strong evidence against H_0 based on the observed data. The data suggest that θ is bigger than 1/6. This is an example of a one-sided test.

Example 5.1.2 Revisited

Suppose that in the experiment in Example 5.1.2 we observed $y = 35$ ones in $n = 180$ tosses. The $p - value$, calculated using R, is now

$$\begin{aligned} p - value &= P(Y \geq 35; \theta = 1/6) \\ &= \sum_{y=35}^{180} \binom{180}{y} \left(\frac{1}{6}\right)^{y} \left(\frac{5}{6}\right)^{180-y} \\ &= 0.18 \end{aligned}$$

and this probability is not especially small. Indeed almost one die in five, though fair, would show this level of discrepancy with H_0. We conclude that there is no evidence against H_0 based on the observed data.

Note that we do **not** claim that H_0 is true, only that there is no evidence in light of the observed data that it is not true. Similarly in the legal example, if we do not find evidence against H_0 : "defendant is innocent", this does not mean we have proven he or she is innocent, only that, for the given data, the amount of evidence against H_0 was insufficient to conclude otherwise.

The approach to testing a hypothesis described above is very general and straightforward, but a few points should be stressed:

(1) If the $p - value$ is very small then we would conclude that there is **strong evidence against H_0 in light of the observed data**; this is often termed "statistically significant" evidence against H_0. We believe that statistical evidence is best measured when we interpret $p-values$ as in Table 5.1. However, it is still common in some areas of research to adopt a threshold $p - value$ such as 0.05 and **"reject H_0" whenever the p-value is below this threshold.** This may be necessary when there are only two possible decisions from which to choose. For example in a trial, a person is either convicted or acquitted of a crime. In the examples in these Course Notes we report the $p - value$ and use the guidelines in Table 5.1 to make a conclusion about whether there is evidence against H_0 or not. We emphasize the point that any decisions which are made after determining the $p - value$ for a given hypothesis H_0 must be made in the context of the empirical study.

(2) If the $p - value$ is not small, we **do not conclude that H_0 is true**. We simply say there is **no evidence against H_0 in light of the observed data**. The reason for this "hedging" is that in most settings a hypothesis may never be strictly "true". For example, one might argue when testing $H_0 : \theta = 1/6$ in Example 5.1.2 that no real die ever has a probability of exactly 1/6 for side 1. Hypotheses can be "disproved" (with a small degree of possible error) but not proved.

(3) Just because there is strong evidence against a hypothesis H_0, there is no implication about how "wrong" H_0 is. A test of hypothesis should always be supplemented with an interval estimate that indicates the magnitude of the departure from H_0.

(4) It is important to keep in mind that although we might be able to find **statistically significant** evidence against a given hypothesis, this does not mean that the differences found are of **practical significance**. For example, suppose an insurance company randomly selects a large number of policies in two different years and finds a statistically significant difference in the mean value of policies sold in those two years of \$5.21. This difference would probably not be of practical significance if the average value of policies sold in a year was greater than \$1000. Similarly, if we collect large amounts of financial data, it is quite easy to find evidence against the hypothesis that stock or stock index returns are Normally distributed. Nevertheless for small amounts of data and for the pricing of options, such an assumption is usually made and considered useful. Finally suppose we compared two cryptographic algorithms using the number of cycles per byte as the unit of measurement. A mean difference of two cycles per byte might be found to be statistically significant but the decision about whether this difference is of practical importance or not is best left to a computer scientist who studies algorithms.

(5) When the observed data provide strong evidence against the null hypothesis, researchers often have an "alternative" hypothesis in mind. For example, suppose a standard pain reliever provides relief in about 50% of cases and researchers at a pharmaceutical company have developed a new pain reliever that they wish to test. The null hypothesis is $H_0 : P(\text{relief}) = 0.5$. Suppose there is strong evidence against H_0 based on the data. The researchers will want to know in which direction that evidence lies. If the probability of relief is greater than 0.5 the researchers might consider adopting the drug or doing further testing, but if the probability of relief is less than 0.5, then the pain reliever would probably be abandoned. The choice of the discrepancy measure D is often made with a particular alternative in mind.

A drawback with the approach to testing described so far is that we do not have a general method for choosing the test statistic or discrepancy measure D. Often there are "intuitively obvious" test statistics that can be used; this was the case in the examples in this section. In Section 5.3 we will see how to use the likelihood function to construct a test statistic in more complicated situations where it is not always easy to come up with an intuitive test statistic.

For the Gaussian model with unknown mean and standard deviation we use test statistics based on the pivotal quantities that were used in Chapter 4 for constructing confidence intervals.

5.2 Hypothesis Testing for Parameters in the $G(\mu, \sigma)$ Model

Suppose that $Y \sim G(\mu, \sigma)$ models a variate y in some population or process. A random sample $Y_1, Y_2, \ldots, Y_n$ is selected, and we want to test hypotheses concerning one of the two parameters (μ, σ). The maximum likelihood estimators of μ and σ^2 are

$$\tilde{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

As usual we prefer to use the sample variance estimator

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

to estimate σ^2.

Recall from Chapter 4 that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t\,(n-1)$$

We use this pivotal quantity to construct a test of hypothesis for the parameter μ when the standard deviation σ is unknown.

Test of Hypothesis for μ

Suppose we wish to test the hypothesis $H_0 : \mu = \mu_0$, where μ_0 is some specified value, against the alternative hypothesis that $\mu \neq \mu_0$. Values of $\overline{Y}$ which are either larger than μ_0 or smaller than μ_0 provide evidence against the null hypothesis $H_0 : \mu = \mu_0$. The test statistic

$$D = \frac{|\overline{Y} - \mu_0|}{S/\sqrt{n}} \tag{5.1}$$

makes intuitive sense. We obtain the $p - value$ using the fact that

$$\frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t\,(n-1)$$

if $H_0 : \mu = \mu_0$ is true. Let

$$d = \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \tag{5.2}$$

be the observed value of D in a sample with mean $\bar{y}$ and standard deviation s, then

$$\begin{aligned} p - value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(|T| \geq d) \quad \text{where } T \sim t\,(n-1) \\ &= 2\,[1 - P\,(T \leq d)] \end{aligned}$$

Since values of $\overline{Y}$ which are larger or smaller than μ_0 provide evidence against the null hypothesis this test is called a two-sided test of hypothesis.

Example 5.2.1 Testing for bias in a measurement system

Two cheap scales A and B for measuring weight are tested by taking 10 weighings of a one kg weight on each of the scales. The measurements on A and B are

$A:$	1.026	0.998	1.017	1.045	0.978	1.004	1.018	0.965	1.010	1.000
$B:$	1.011	0.966	0.965	0.999	0.988	0.987	0.956	0.969	0.980	0.988

Let Y represent a single measurement on one of the scales, and let μ represent the average measurement $E(Y)$ in repeated weighings of a single 1 kg weight. If an experiment involving n weighings is conducted then a test of $H_0 : \mu = 1$ can be based on the test statistic (5.1) with observed value (5.2) and $\mu_0 = 1$.

The samples from scales A and B above give us

$$\begin{aligned} A: &\quad \bar{y} = 1.0061, \quad s = 0.0230, \quad d = 0.839 \\ B: &\quad \bar{y} = 0.9810, \quad s = 0.0170, \quad d = 3.534 \end{aligned}$$

The $p - value$ for A is

$$\begin{aligned} p - value &= P(D \geq 0.839;\ \mu = 1) \\ &= P\left(|T| \geq 0.839\right) \quad \text{where } T \sim t\left(9\right) \\ &= 2\left[1 - P\left(T \leq 0.839\right)\right] \\ &= 2\left(1 - 0.7884\right) \\ &\approx 0.42 \end{aligned}$$

where the probability is obtained using R. Alternatively if we use the t table provided in these notes we obtain $P\left(T \leq 0.5435\right) = 0.7$ and $P\left(T \leq 0.88834\right) = 0.8$ so

$$0.4 = 2\left(1 - 0.8\right) < p - value < 2\left(1 - 0.7\right) = 0.6.$$

In either case we have that the $p - value > 0.1$ and thus there is no evidence of bias, that is, there is no evidence against $H_0 : \mu = 1$ for scale A based on the observed data.

For scale B, however, we obtain

$$\begin{aligned} p - value &= P\left(D \geq 3.534;\ \mu = 1\right) \\ &= P\left(|T| \geq 3.534\right) \quad \text{where } T \sim t\left(9\right) \\ &= 2\left[1 - P\left(T \leq 3.534\right)\right] \\ &= 0.0064 \end{aligned}$$

where the probability is obtained using R. Alternatively if we use the t table we obtain $P\left(T \leq 3.2498\right) = 0.995$ and $P\left(T \leq 4.2968\right) = 0.999$ so

$$0.002 = 2\left(1 - 0.999\right) < p - value < 2\left(1 - 0.995\right) = 0.01$$

In either case we have that the $p-value < 0.01$ and thus there is strong evidence against $H_0 : \mu = 1$. The observed data suggest strongly that scale B is biased.

Finally, note that just although there is strong evidence against H_0 for scale B, the degree of bias in its measurements is not necessarily large enough to be of practical concern. In fact, we can obtain a 95% confidence interval for μ for scale B by using the pivotal quantity

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{10}} \sim t\,(9)$$

For the t table we have $P(T \leq 2.2622) = 0.975$ and a 95% confidence interval for μ is given by

$$\bar{y} \pm 2.2622 s/\sqrt{10} = 0.981 \pm 0.012$$

or

$$[0.969, 0.993]$$

Evidently scale B consistently understates the weight but the bias in measuring the one kg weight is likely fairly small (about $1\% - 3\%$).

Remark The function `t.test` in R will give confidence intervals and test hypotheses about μ. See Problem 3.

One-sided test of hypothesis for μ

Suppose data on the effects of a new treatment follow a $G(\mu, \sigma)$ distribution and that the new treatment can either have no effect represented by $\mu = \mu_0$ or a beneficial effect represented by $\mu > \mu_0$. In this case the null hypothesis is $H_0 : \mu = \mu_0$ and the alternative hypothesis is $H_A : \mu > \mu_0$.

To test $H_0 : \mu = \mu_0$ we would use the test statistic

$$D = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$$

so that large values of D provide evidence against H_0 in the direction of the alternative $\mu > \mu_0$.

Let the observed value of D be

$$d = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

Then

$$\begin{aligned} p - value &= P(D \geq d; H_0 \text{ is true}) \\ &= P(T \geq d) \\ &= 1 - P\,(T \leq d) \quad \text{where } T \sim t\,(n-1) \end{aligned}$$

This is another example of a one-sided test of hypothesis. What test statistic should be used if the alternative hypothesis is $H_A : \mu < \mu_0$?

Relationship between Hypothesis Testing and Interval Estimation

Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the $G(\mu, \sigma)$ distribution. Suppose we test $H_0 : \mu = \mu_0$.

Now

$$p - value \ \geq 0.05$$

$$\text{if and only if } P\left(\frac{|\overline{Y} - \mu_0|}{S/\sqrt{n}} \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}};\ H_0 : \mu = \mu_0 \text{ is true}\right) \geq 0.05$$

$$\text{if and only if } \ P\left(|T| \ \geq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \geq 0.05 \quad \text{where } T \sim t\,(n-1)$$

$$\text{if and only if } \ P\left(|T| \ \leq \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}}\right) \leq 0.95$$

$$\text{if and only if } \ \frac{|\bar{y} - \mu_0|}{s/\sqrt{n}} \leq a \quad \text{where } P\left(|T| \ \leq a\right) = 0.95$$

$$\text{if and only if } \ \mu_0 \in \left[\bar{y} - as/\sqrt{n}, \bar{y} + as/\sqrt{n}\right]$$

which is a 95% confidence interval for μ. In other words, the $p-value$ for testing $H_0 : \mu = \mu_0$ is greater than or equal to 0.05 if and only if the value $\mu = \mu_0$ is an element of a 95% confidence interval for μ (assuming we use the same pivotal quantity). Note that both endpoints of the 95% confidence interval correspond to a $p-value$ equal to 0.05 while the values inside the 95% confidence interval will have $p - values$ greater than 0.05.

More generally, suppose we have data $\mathbf{y}$ and a model $f(\mathbf{y}; \theta)$. Suppose also that we use the same pivotal quantity to construct the (approximate) confidence interval for θ and to test the hypothesis $H_0 : \theta = \theta_0$. Then the parameter value $\theta = \theta_0$ is an element of the $100q\%$ (approximate) confidence interval for θ if and only if the $p - value$ for testing $H_0 : \theta = \theta_0$ is greater than or equal to $1 - q$.

Example 5.2.1 Revisited

For the weigh scale example the $p - value$ for testing $H_0 : \mu = 1$ for scale A was greater than 0.4 and thus greater than 0.05. Therefore we know that the value $\mu = 1$ is in a 95% confidence interval for the mean μ. In fact the 95% confidence interval for the mean μ is

$$\bar{y} \pm 2.2622s/\sqrt{10} = 1.0061 \pm 0.01645 = [0.9897, 1.0226]$$

which does indeed contain the value $\mu = 1$.

For scale B, a 95% confidence interval for the mean μ was $[0.969, 0.993]$. Since $\mu = 1$ is not in this interval we know that the $p - value$ for testing $H_0 : \mu = 1$ would be less than 0.05. In fact we showed the $p - value$ equals 0.0064 which is indeed less than 0.05.

Test of Hypothesis for σ

Suppose that we have a sample $Y_1, Y_2, \ldots, Y_n$ of independent random variables each from the same $G(\mu, \sigma)$ distribution. Recall that we used the pivotal quantity

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \sim \chi^2(n-1)$$

to construct confidence intervals for the parameter σ. We may also wish to test a hypothesis such as $H_0 : \sigma = \sigma_0$ or equivalently $H_0 : \sigma^2 = \sigma_0^2$. One approach is to use a likelihood ratio test statistic which is described in the next section. Alternatively we could use the test statistic

$$U = \frac{(n-1)S^2}{\sigma_0^2}$$

for testing $H_0 : \sigma = \sigma_0$. Large values of U and small values of U provide evidence against H_0. (Why is this?) Now U has a Chi-squared distribution when H_0 is true and the Chi-squared distribution is not symmetric which makes the determination of "large" and "small" values somewhat problematic. The following simpler calculation approximates the $p - value$:

1. Let $u = (n-1)s^2/\sigma_0^2$ denote the observed value of U from the data.
2. If u is large (that is, if $P(U \leq u) > \frac{1}{2}$) compute the $p - value$ as

   $$p - value = 2P(U \geq u)$$

   where $U \sim \chi^2(n-1)$.
3. If u is small (that is, if $P(U \leq u) < \frac{1}{2}$) compute the $p - value$ as

   $$p - value = 2P(U \leq u)$$

   where $U \sim \chi^2(n-1)$.

Figure 5.1 shows a picture for a large observed value of u. In this case $P(U \leq u) > \frac{1}{2}$ and the $p - value = 2P(U \geq u)$.

Note:

Only one of the two values $2P(U \geq u)$ and $2P(U \leq u)$ will be less than one and this value is the desired $p - value$.

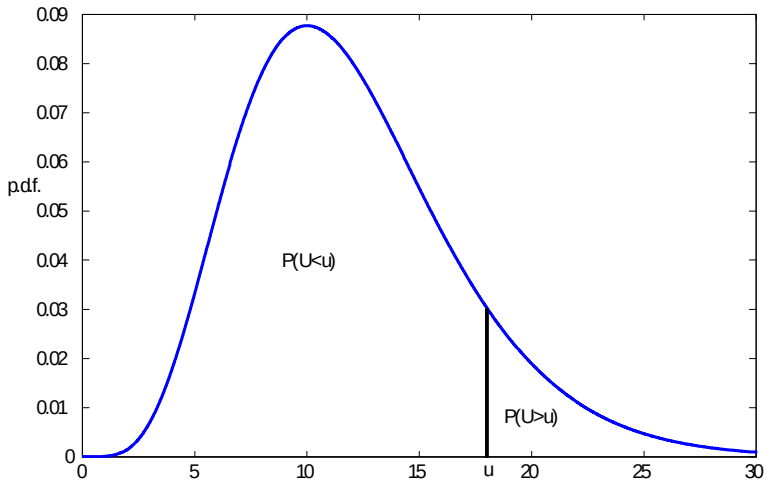


Figure 5.1: **Picture of large observed** u

Example 5.2.2

Suppose for the manufacturing process in Example 4.7.2, we wish to test the hypothesis $H_0 : \sigma = 0.008$ (0.008 is the desired or target value of σ the manufacturer would like to achieve). Since the 95% confidence interval for σ was found to be [0.0095, 0.0204] which does not contain the value $\sigma = 0.008$ we already know that the $p - value$ for a test of H_0 based on the test statistic $U = (n-1)S^2/\sigma_0^2$ will be less than 0.05.

To find the $p - value$, we use the procedure given above:

1. $u = (n-1)s^2/\sigma_0^2 = (14)\, s^2/\left(0.008\right)^2 = 0.002347/\left(0.008\right)^2 = 36.67$

2. The $p - value$ is

$$p - value = 2P(U \geq u) = 2P(U \geq 36.67) = 0.0017 \quad \text{where} \quad U \sim \chi^2\,(14)$$

where the probability was obtained using R.

Alternatively if we use the Chi-squared table provided in these Course Notes we obtain $P\left(U \leq 31.319\right) = 0.995$ so

$$p - value < 2\left(1 - 0.995\right) = 0.01$$

In either case we have that the $p - value < 0.01$ and thus there is strong evidence based on the observed data against $H_0 : \sigma = 0.008$. Both the observed value of $s = \sqrt{0.002347/14} = 0.0129$ and the 95% confidence interval for σ suggest that σ is bigger than 0.008.

5.3 Likelihood Ratio Test of Hypothesis - One Parameter

When a pivotal quantity exists then it is usually straightforward to construct a test of hypothesis as we have seen Section 5.2 for the Gaussian distribution parameters. When a pivotal quantity does not exist then a general method for finding a test statistic with good properties can be based on the likelihood function. In Chapter 2 we used likelihood functions to gauge the plausibility of parameter values in the light of the observed data. It should seem natural, then, to base a test of hypothesis on a likelihood value or, in comparing the plausibility of two values, a ratio of the likelihood values. Let us suppose, for example, that we are engaged in an argument over the value of a parameter θ in a given model (we agree on the model but disagree on the parameter value). I claim that the parameter value is θ_0 whereas you claim it is θ_1. Having some data $\mathbf{y}$ at hand, it would seem reasonable to attempt to settle this argument using the ratio of the likelihood values at these two values, that is,

$$\frac{L(\theta_0)}{L(\theta_1)} \tag{5.3}$$

As usual we define the likelihood function $L(\theta) = L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$ where $f(\mathbf{y}; \theta)$ is the probability (density) function of the random variable $\mathbf{Y}$ representing the data. If the value of the ratio $L(\theta_0)/L(\theta_1)$ is much greater than one then the data support the value θ_0 more than θ_1.

Let us now consider testing the plausibility of my hypothesized value θ_0 against an unspecified alternative. In this case it is natural to replace θ_1 in (5.3) by the value which appears most plausible given the data, that is, the maximum likelihood estimate $\hat{\theta}$. The resulting ratio is just the value of the relative likelihood function at θ_0:

$$R(\theta_0) = \frac{L(\theta_0)}{L(\hat{\theta})}$$

If $R(\theta_0)$ is close to one, then θ_0 is plausible in light of the observed data, but if $R(\theta_0)$ is very small and close to zero, then θ_0 is not plausible in light of the observed data and this suggests evidence against H_0. Therefore the corresponding random variable, $L(\theta_0)/L(\tilde{\theta})$[13], appears to be a natural statistic for testing $H_0 : \theta = \theta_0$. To determine $p - values$ we need the sampling distribution of $L(\theta_0)/L(\tilde{\theta})$ under H_0. It is actually easier to use the likelihood ratio statistic which was introduced in Chapter 4

$$\Lambda(\theta_0) = -2 \log \left[\frac{L(\theta_0)}{L(\tilde{\theta})} \right] \tag{5.4}$$

(remember $log = ln$) which is a one-to-one function of $L(\theta_0)/L(\tilde{\theta})$. We choose this particular function because, if $H_0 : \theta = \theta_0$ is true, then $\Lambda(\theta_0)$ has approximately $\chi^2(1)$ distribution.

[13] Recall that $L(\theta) = L(\theta; \mathbf{y})$ is a function of the observed data $\mathbf{y}$. Replacing $\mathbf{y}$ by the corresponding random variable $\mathbf{Y}$ means that $L(\theta; \mathbf{Y})$ is a random variable. The random variable $L(\theta_0)/L(\tilde{\theta}) = L(\theta_0; \mathbf{Y})/L(\tilde{\theta}; \mathbf{Y})$ is a function of $\mathbf{Y}$ in several places including $\tilde{\theta} = g(\mathbf{Y})$.

Note that small values of $R(\theta_0)$ correspond to large observed values of $\Lambda(\theta_0)$ and therefore large observed value of $\Lambda(\theta_0)$ indicate evidence against the hypothesis $H_0 : \theta = \theta_0$. We illustrate this in Figure 5.2. Notice that the more plausible values of the parameter θ correspond to larger values of $R(\theta)$ or equivalently, in the bottom panel, to small values of $\Lambda(\theta) = -2\log[R(\theta)]$. The particular value displayed θ_0 is around 0.3 and it appears that $\Lambda(\theta_0) = -2\log[R(\theta_0)]$ is quite large, in this case around 9. To know whether this is too large to be consistent with H_0, we need to compute the $p - value$.

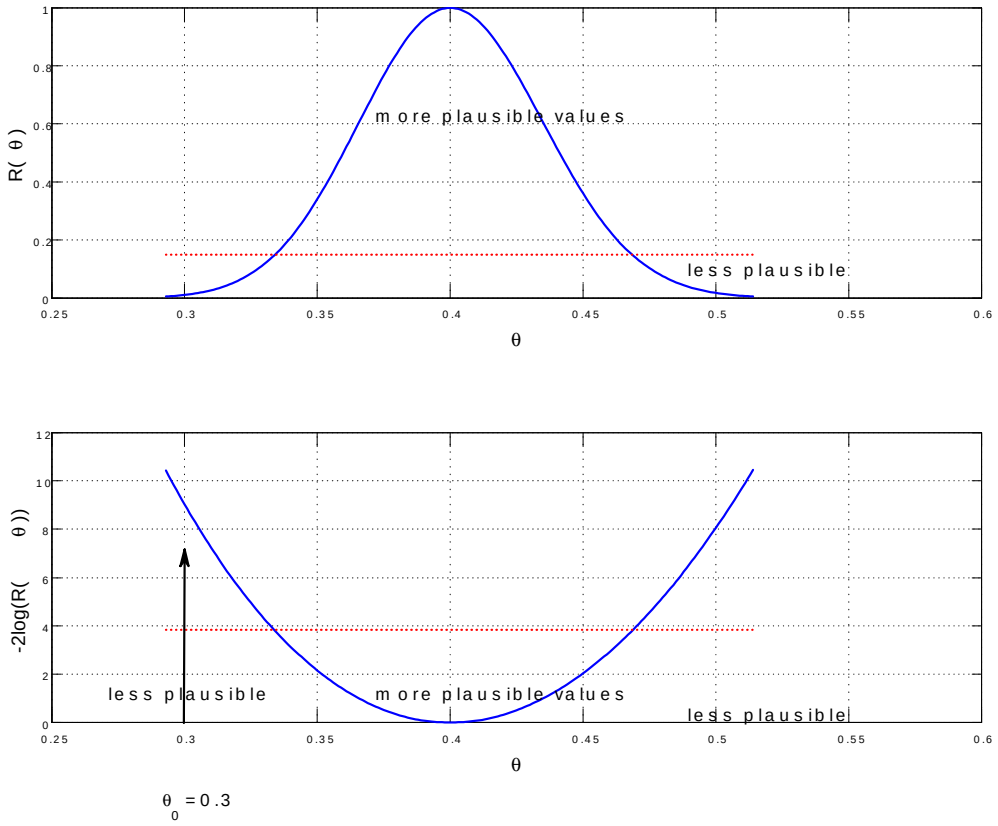


Figure 5.2: **Top panel:** Graph of the relative likelihood function
Bottom Panel: $\Lambda(\theta) = -2\log R(\theta)$
Note that $\Lambda(\theta_0)$ is relatively large when $R(\theta_0)$ is small.

To determine the $p - value$ we first calculate the observed value of $\Lambda(\theta_0)$, denoted by $\lambda(\theta_0)$ and given by

$$\lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\hat{\theta})}\right] = -2\log R(\theta_0)$$

where $R(\theta_0)$ is the relative likelihood function evaluated at $\theta = \theta_0$. The approximate

$p-value$ is then

$$\begin{aligned} p-value &\approx P\left[W \geq \lambda\left(\theta_0\right)\right] \text{ where } W \sim \chi^2(1) \\ &= P\left(|Z| \geq \sqrt{\lambda\left(\theta_0\right)}\right) \text{ where } Z \sim G(0,1) \\ &= 2\left[1-P\left(Z \leq \sqrt{\lambda\left(\theta_0\right)}\right)\right] \end{aligned} \tag{5.5}$$

Let us summarize the construction of a test from the likelihood function. Let the random variable (or vector of random variables) $\mathbf{Y}$ represent data generated from a distribution with probability function or probability density function $f(\mathbf{y};\theta)$ which depends on the scalar parameter θ. Let Ω be the parameter space (set of possible values) for θ. Consider a hypothesis of the form

$$H_0 : \theta = \theta_0$$

where θ_0 is a single point (hence of dimension 0). We can test H_0 using as our **test statistic** the **likelihood ratio test statistic** Λ, defined by (5.4). Then large observed values of $\Lambda\left(\theta_0\right)$ correspond to a disagreement between the hypothesis $H_0 : \theta = \theta_0$ and the data and so provide evidence against H_0. Moreover if $H_0 : \theta = \theta_0$ is true, $\Lambda\left(\theta_0\right)$ has approximately a $\chi^2(1)$ distribution so that an approximate $p-value$ is obtained from (5.5). The theory behind the approximation is based on a result which shows that under H_0, the distribution of Λ approaches $\chi^2(1)$ as the size of the data set becomes large.

Example 5.3.1 Likelihood ratio test statistic for Binomial model

Since the relative likelihood function for the Binomial model is

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} \\ &= \frac{\theta^y(1-\theta)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} \\ &= \left(\frac{\theta}{\hat{\theta}}\right)^y\left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-y} \quad \text{for } 0 \leq \theta \leq 1 \end{aligned}$$

the likelihood ratio test statistic for testing the hypothesis $H_0 : \theta = \theta_0$ is

$$\begin{aligned} \Lambda\left(\theta_0\right) &= -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right] \\ &= -2\log\left[\left(\frac{\theta_0}{\tilde{\theta}}\right)^y\left(\frac{1-\theta_0}{1-\tilde{\theta}}\right)^{n-y}\right] \end{aligned}$$

where $\tilde{\theta} = Y/n$ is the maximum likelihood estimator of θ. The observed value of $\Lambda\left(\theta_0\right)$ is

$$\begin{aligned} \lambda\left(\theta_0\right) &= -2\log R(\theta_0) \\ &= -2\log\left[\left(\frac{\theta_0}{\hat{\theta}}\right)^y\left(\frac{1-\theta_0}{1-\hat{\theta}}\right)^{n-y}\right] \end{aligned}$$

where $\hat{\theta} = y/n$. If $\hat{\theta}$ is close in value to θ_0 then $R(\theta_0)$ will be close in value to 1 and $\lambda(\theta_0)$ will be close in value to 0.

Suppose we use the likelihood ratio test statistic to test $H_0 : \theta = 0.5$ for the ESP example and the data in Example 5.1.1. Since $n = 200$, $y = 110$ and $\hat{\theta} = 0.55$, the observed value of the likelihood ratio statistic for testing $H_0 : \theta = 0.5$ is

$$\begin{aligned}\lambda(0.5) &= -2\log R(0.5) = -2\log\left[\left(\frac{0.5}{0.55}\right)^{110}\left(\frac{1-0.5}{1-0.55}\right)^{90}\right] \\ &= -2\log(0.367) \\ &= 2.003\end{aligned}$$

(Note that since $R(0.5) = 0.367 > 0.1$ then we already know that $\theta = 0.5$ is a plausible value of θ.) The approximate $p - value$ for testing $H_0 : \theta = 0.5$ is

$$\begin{aligned}&p - value \approx P(W \geq 2.003) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{2.003}\right)\right] \quad \text{where } Z \sim G(0,1) \\ &= 2\left[1 - P(Z \leq 1.42)\right] = 2(1 - 0.9222) \\ &= 0.1556\end{aligned}$$

and there is no evidence against $H_0 : \theta = 0.5$ based on the data. Note that the test statistic $D = |Y - 100|$ used in Example 5.1.1 and the likelihood ratio test statistic $\Lambda(0.5)$ give nearly identical results. This is because $n = 200$ is large.

Example 5.3.2 Likelihood ratio test statistic for Exponential model

Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the Exponential(θ) distribution. The likelihood function (see Example 2.3.1) is

$$L(\theta) = \frac{1}{\theta^n}\exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} y_i\right) = \theta^{-n}e^{-n\bar{y}/\theta} \quad \text{for } \theta > 0$$

Since the maximum likelihood estimate is $\hat{\theta} = \bar{y}$, the relative likelihood function can be written as

$$\begin{aligned}R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} \\ &= \frac{\theta^{-n}e^{-n\bar{y}/\theta}}{\hat{\theta}^{-n}e^{-n\bar{y}/\hat{\theta}}} \\ &= \left(\frac{\hat{\theta}}{\theta}\right)^n e^{n\left(1-\hat{\theta}/\theta\right)} \quad \text{for } \theta > 0\end{aligned}$$

The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0$ is

$$\begin{aligned}\Lambda(\theta_0) &= -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right] \\ &= -2\log\left[\left(\frac{\tilde{\theta}}{\theta_0}\right)^n e^{n\left(1-\tilde{\theta}/\theta_0\right)}\right]\end{aligned}$$

where $\tilde{\theta} = \overline{Y}$ and the observed value of $\Lambda(\theta_0)$ is

$$\begin{aligned}\lambda(\theta_0) &= -2\log R(\theta_0) \\ &= -2\log\left[\left(\frac{\hat{\theta}}{\theta_0}\right)^n e^{n\left(1-\hat{\theta}/\theta_0\right)}\right]\end{aligned}$$

If $\hat{\theta}$ is close in value to θ_0 then $R(\theta_0)$ will be close in value to 1 and $\lambda(\theta_0)$ will be close in value to 0.

The variability in lifetimes of light bulbs (in hours, say, of operation before failure) is often well described by an Exponential(θ) distribution where $\theta = E(Y) > 0$ is the average (mean) lifetime. A manufacturer claims that the mean lifetime of a particular brand of bulbs is 2000 hours. We can examine this claim by testing the hypothesis $H_0 : \theta = 2000$. Suppose a random sample of $n = 50$ light bulbs was tested over a long period and that the observed lifetimes were:

572	2732	1363	716	231	83	1206	3952	3804	2713
347	2739	411	2825	147	2100	3253	2764	969	1496
2090	371	1071	1197	173	2505	556	565	1933	1132
5158	5839	1267	499	137	4082	1128	1513	8862	2175
3638	461	2335	1275	3596	1015	2671	849	744	580

with $\sum_{i=1}^{50} y_i = 93840$. For these data the maximum likelihood estimate of θ is $\hat{\theta} = \bar{y} = 93840/50 = 1876.8$. To check whether the Exponential model is reasonable for these data we plot the empirical cumulative distribution function for these data and then superimpose the cumulative distribution function for a Exponential(1876.8) random variable. See Figure 5.3. Since the agreement between the empirical cumulative distribution function and the Exponential(1876.8) cumulative distribution function is quite good we assume the Exponential model to test the hypothesis that the mean lifetime the light bulbs is 2000 hours. The observed value of the likelihood ratio test statistic for testing $H_0 : \theta = 2000$ is

$$\begin{aligned}\lambda(2000) &= -2\log R(2000) \\ &= -2\log\left[\left(\frac{1876.8}{2000}\right)^{50} e^{50(1-1876.8/2000)}\right] = -2\log(0.9058) \\ &= 0.1979\end{aligned}$$

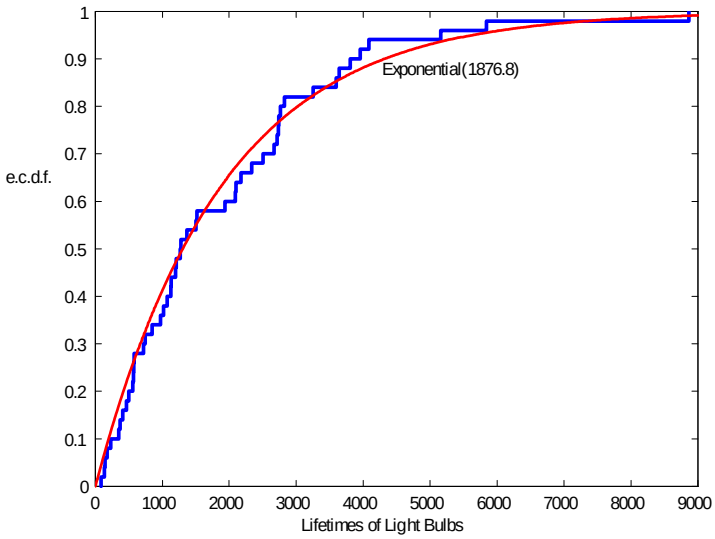


Figure 5.3: **Empirical c.d.f. and Exponential**(1876.8) **c.d.f.**

(Note that since $R(2000) = 0.9058 > 0.1$ then we already know that $\theta = 2000$ is a plausible value of θ.) The approximate $p - value$ for testing $H_0 : \theta = 2000$ is

$$\begin{aligned} p - value &\approx P(W \geq 0.1979) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq \sqrt{0.1979}\right)\right] \quad \text{where } Z \sim G(0,1) \\ &= 2\left[1 - P\left(Z \leq 0.44\right)\right] = 2\left(1 - 0.67003\right) \\ &= 0.65994 \end{aligned}$$

and there is no evidence against $H_0 : \theta = 2000$ based on the data. Therefore there is no evidence against the manufacturer's claim that θ is 2000 hours based on the data. Although the maximum likelihood estimate $\hat{\theta}$ was under 2000 hours (1876.8) it was not sufficiently under to give evidence against $H_0 : \theta = 2000$.

Example 5.3.3 Likelihood ratio test of hypothesis for μ for $G(\mu, \sigma)$, known σ

Suppose $Y \sim G(\mu, \sigma)$ with probability density function

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right] \quad \text{for } y \in \Re$$

Suppose the standard deviation σ has a known value and the only unknown parameter is μ. From the results in Example 2.3.2, we have that the likelihood function based on the observed sample $y_1, y_2, \ldots, y_n$ is

$$L(\mu) = \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2\right\} \exp\left[-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right]$$

or more simply (ignoring constants with respect to μ)

$$L(\mu) = \exp\left[-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right] \quad \text{for } \mu \in \Re$$

The corresponding log likelihood function is

$$l(\mu) = -\frac{n(\bar{y}-\mu)^2}{2\sigma^2} \quad \text{for } \mu \in \Re$$

To find the maximum likelihood estimate of μ we solve the equation

$$l'(\mu) = \frac{n(\bar{y}-\mu)}{\sigma^2} = 0$$

which gives $\hat{\mu} = \bar{y}$. The corresponding maximum likelihood estimator of μ is

$$\tilde{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The relative likelihood function can be written as

$$\begin{aligned} R(\mu) &= \frac{L(\mu)}{L(\hat{\mu})} \\ &= \exp\left[-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right] \quad \text{for } \mu \in \Re \end{aligned}$$

since $\hat{\mu} = \bar{y}$ gives $L(\hat{\mu}) = 1$.

To test the hypothesis $H_0 : \mu = \mu_0$ we use the likelihood ratio statistic

$$\begin{aligned} \Lambda(\mu_0) &= -2\log\left[\frac{L(\mu_0)}{L(\tilde{\mu})}\right] \\ &= -2\log\left\{\exp\left[-\frac{n(\overline{Y}-\mu_0)^2}{2\sigma^2}\right]\right\} \quad \text{since } \tilde{\mu} = \overline{Y} \\ &= \frac{n(\overline{Y}-\mu_0)^2}{\sigma^2} \\ &= \left(\frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \end{aligned} \tag{5.6}$$

The purpose of writing the likelihood ratio statistic in the form (5.6) is to draw attention to the fact that, in this special case, $\Lambda(\mu_0)$ has exactly a $\chi^2(1)$ distribution for all values of n since $\frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} \sim G(0,1)$.

More generally it is not obvious that the likelihood ratio test statistic has an approximate $\chi^2(1)$ distribution.

5.4 Likelihood Ratio Test of Hypothesis - Multiparameter

Let the random vector $\mathbf{Y}$ represent data generated from a distribution with probability or probability density function $f(\mathbf{y};\theta)$ which depends on the k-dimensional parameter $\boldsymbol{\theta}$. Let Ω be the parameter space (set of possible values) for $\boldsymbol{\theta}$.

Consider a hypothesis of the form

$$H_0 : \boldsymbol{\theta} \in \Omega_0$$

where $\Omega_0 \subset \Omega$ and Ω_0 is of dimension $p < k$. For example H_0 might specify particular values for $k - p$ of the components of $\boldsymbol{\theta}$ but leave the remaining parameters unspecified. The dimensions of Ω and Ω_0 refer to the minimum number of parameters (or "coordinates") needed to specify points in them. Again we test H_0 using as our **test statistic** the **likelihood ratio test statistic** Λ, defined as follows. Let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$ over Ω so that, as before,

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})$$

Similarly we let $\hat{\boldsymbol{\theta}}_0$ denote the maximum likelihood estimate of $\boldsymbol{\theta}$ over Ω_0 (i.e. we maximize the likelihood with the parameter $\boldsymbol{\theta}$ constrained to lie in the set $\Omega_0 \subset \Omega$) so that

$$L(\hat{\boldsymbol{\theta}}_0) = \max_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta})$$

Now consider the corresponding statistic (random variable)

$$\Lambda = -2\log\left[\frac{L(\tilde{\boldsymbol{\theta}}_0)}{L(\tilde{\boldsymbol{\theta}})}\right] = 2\left[l(\tilde{\boldsymbol{\theta}}) - l(\tilde{\boldsymbol{\theta}}_0)\right]$$

and let

$$\lambda = -2\log\left[\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right] = 2\left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\right] \tag{5.7}$$

denote the observed value of Λ. If λ is very large, then there is evidence against H_0 (confirm that this means $L(\hat{\boldsymbol{\theta}})$ is much larger than $L(\hat{\boldsymbol{\theta}}_0)$). It can be shown that under H_0, the distribution of Λ is approximately $\chi^2(k-p)$ as the size of the data set becomes large. Large values of λ indicate evidence **against** H_0 so the $p - value$ is given approximately by

$$p - value = P(\Lambda \geq \lambda;\ H_0) \approx P(W \geq \lambda) \tag{5.8}$$

where $W \sim \chi^2(k-p)$.

The likelihood ratio test covers a great many different types of examples, but we only provide a few examples here.

Example 5.4.3 Comparison of two Poisson means

In Chapter 4, Problem 10 data were given on the numbers of failures per month for each of two companies' photocopiers. We assume that in a given month the number of failures Y follows a Poisson distribution with probability function

$$f(y;\mu) = P(Y = y) = \frac{\mu^y e^{-\mu}}{y!} \quad \text{for } y = 0, 1, \ldots$$

where $\mu = E(Y)$ is the mean number of failures per month. (This ignores that the number of days that the copiers are used varies a little across months. Adjustments could be made to the analysis to deal with this.) Denote the value of μ for Company A's copiers as μ_A and the value for Company B's as μ_B. Let us test the hypothesis that the two photocopiers have the same mean number of failures

$$H_0 : \mu_A = \mu_B$$

Essentially we have data from two Poisson distributions with possibly different parameters. For convenience let $x_1, \ldots, x_n$ denote the observations for Company A's photocopier which are assumed to be a random sample from the model

$$P(X = x; \mu_A) = \frac{\mu_A^x e^{-\mu_A}}{x!} \quad \text{for } x = 0, 1, \ldots \quad \text{and } \mu_A \geq 0$$

Similarly let $y_1, y_2, \ldots, y_m$ denote the observations for Company B's photocopier which are assumed to be a random sample from the model

$$P(Y = y; \mu_B) = \frac{\mu_B^y e^{-\mu_B}}{y!} \quad \text{for } y = 0, 1, \ldots \quad \text{and } \mu_B \geq 0$$

independently of the observations for Company A's photocopier. In this case the parameter vector is the two dimensional vector $\boldsymbol{\theta} = (\mu_A, \mu_B)$ and $\Omega = \{(\mu_A, \mu_B) : \mu_A \geq 0, \mu_B \geq 0\}$. The note that the dimension of Ω is $k = 2$. Since the null hypothesis specifies that the two parameters μ_A and μ_B are equal but does not otherwise specify their values, we have $\Omega_0 = \{(\mu, \mu) : \mu \geq 0\}$ which is a space of dimension $p = 1$.

To construct the likelihood ratio test of $H_0 : \mu_A = \mu_B$ we need the likelihood function for the parameter vector $\boldsymbol{\theta} = (\mu_A, \mu_B)$. We first note that the likelihood function for μ_A only based on the data $x_1, x_2, \ldots, x_n$ is

$$L_1(\mu_A) = \prod_{i=1}^{n} f(x_i; \mu_A) = \prod_{i=1}^{n} \frac{\mu_A^{x_i} e^{-\mu_A}}{x_i!} \quad \text{for } \mu_A \geq 0$$

or more simply

$$L_1(\mu_A) = \mu_A^{n\bar{x}} e^{-n\mu_A} \quad \text{for } \mu_A \geq 0$$

Similarly the likelihood function for μ_B only based on $y_1, y_2, \ldots, y_m$ is given by

$$L_2(\mu_B) = \mu_B^{m\bar{y}} e^{-m\mu_B} \quad \text{for } \mu_B \geq 0$$

Since the datasets are independent, the likelihood function for $\boldsymbol{\theta} = (\mu_A, \mu_B)$ is obtained as a product of the individual likelihoods

$$\begin{aligned} L(\boldsymbol{\theta}) = L(\mu_A, \mu_B) &= L_1(\mu_A) L_2(\mu_B) \\ &= \mu_A^{n\bar{x}} e^{-n\mu_A} \mu_B^{m\bar{y}} e^{-m\mu_B} \quad \text{for } (\mu_A, \mu_B) \in \Omega \end{aligned}$$

with corresponding log likelihood function

$$l(\boldsymbol{\theta}) = -n\mu_A - m\mu_B + (n\bar{x}) \log \mu_A + m\bar{y} \log \mu_B \quad \text{for } (\mu_A, \mu_B) \in \Omega \tag{5.9}$$

The number of photocopy failures in twelve consecutive months for company A and company B are given below:

Month	1	2	3	4	5	6	7	8	9	10	11	12	Total
Company A	16	14	25	19	23	12	22	28	19	15	18	29	$\sum_{i=1}^{12} x_i = 240$
Company B	13	7	12	9	15	17	10	13	8	10	12	14	$\sum_{j=1}^{12} y_j = 140$

The log likelihood function is

$$l(\boldsymbol{\theta}) = l(\mu_A, \mu_B) = -12\mu_A + 240 \log \mu_A - 12\mu_B + 140 \log \mu_B \quad \text{for } (\mu_A, \mu_B) \in \Omega$$

The values of μ_A and μ_B which maximize $l(\mu_A, \mu_B)$ are obtained by solving the two equations

$$\frac{\partial l}{\partial \mu_A} = 0 \qquad \frac{\partial l}{\partial \mu_B} = 0$$

which gives two equations in two unknowns:

$$\begin{aligned} -12 + \frac{240}{\mu_A} &= 0 \\ -12 + \frac{140}{\mu_B} &= 0 \end{aligned}$$

The maximum likelihood estimates of μ_A and μ_B (unconstrained) are $\hat{\mu}_A = \bar{x} = \frac{240}{12} = 20.0$ and $\hat{\mu}_B = \bar{y} = \frac{140}{12} = 11.667$ and $\hat{\boldsymbol{\theta}} = (\bar{x}, \bar{y}) = (20.0, 11.667)$.

To determine

$$L(\hat{\boldsymbol{\theta}}_0) = \max_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta})$$

we need to find the (constrained) maximum likelihood estimate $\hat{\boldsymbol{\theta}}_0$, which is the value of $\boldsymbol{\theta} = (\mu_A, \mu_B)$ which maximizes $l(\mu_A, \mu_B)$ under the constraint $\mu_A = \mu_B$. To do this we merely let $\mu = \mu_A = \mu_B$ in (5.9) to obtain

$$\begin{aligned} l(\mu, \mu) &= -12\mu + 240 \log \mu - 12\mu + 140 \log \mu \\ &= -24\mu + 380 \log \mu \quad \text{for } \mu \geq 0 \end{aligned}$$

Solving $\partial l(\mu,\mu)/\partial\mu = 0$, we find $\hat{\mu} = \frac{n\bar{x}+m\bar{y}}{n+m} = \frac{380}{24} = 15.833$ $(= \hat{\mu}_A = \hat{\mu}_B)$ so $\hat{\boldsymbol{\theta}}_0 = (15.833, 15.833)$.

The observed value of the likelihood ratio statistic using (5.7) is

$$\begin{aligned}\lambda &= -2\log\left[\frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})}\right] = 2\left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\right] \\ &= 2\left[l(20.0, 11.667) - l(15.833, 15.833)\right] \\ &= 2\left(682.92 - 669.60\right) \\ &= 26.64\end{aligned}$$

and the approximate $p-value$ (5.8) is

$$\begin{aligned}p-value &= P(\Lambda \geq 26.64; H_0) \\ &\approx P(W \geq 26.64) \quad \text{where } W \sim \chi^2(1) \\ &\approx 0\end{aligned}$$

Based on the data there is very strong evidence against the hypothesis $H_0 : \mu_A = \mu_B$. The data suggest that Company B's photocopiers have a lower rate of failure than Company A's photocopiers.

Note that we could also follow up this conclusion by giving a confidence interval for the mean difference $\mu_A - \mu_B$ since this would indicate the magnitude of the difference in the two failure rates. The maximum likelihood estimates $\hat{\mu}_A = 20.0$ average failures per month and $\hat{\mu_B} = 11.67$ failures per month differ by quite a bit, but we could also give a confidence interval in order to express the uncertainty in such estimates.

Example 5.4.4 Likelihood ratio test of hypothesis for σ for $G(\mu,\sigma)$, unknown μ

Consider a test of $H_0 : \sigma = \sigma_0$ based on a random sample $y_1, y_2, \ldots, y_n$. In this case the unconstrained parameter space is $\Omega = \{(\mu,\sigma) : -\infty < \mu < \infty, \sigma > 0\}$, obviously a 2-dimensional space, but under the constraint imposed by H_0, the parameter must lie in the space $\Omega_0 = \{(\mu,\sigma_0), -\infty < \mu < \infty\}$ a space of dimension 1. Thus $k = 2$, and $p = 1$. The likelihood function is

$$\begin{aligned}L(\boldsymbol{\theta}) &= L(\mu,\sigma) = \prod_{i=1}^{n} f(Y_i;\mu,\sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\ &= (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]\end{aligned}$$

or more simply

$$L(\mu,\sigma) = \sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$

The log likelihood function is

$$l(\mu,\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

The maximum likelihood estimators of (μ,σ) in the unconstrained case are

$$\begin{aligned}\tilde{\mu} &= \overline{Y}\\ \tilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\end{aligned}$$

Under the constraint imposed by $H_0:\sigma=\sigma_0$ the maximum likelihood estimator of the parameter μ is also $\overline{Y}$ so the likelihood ratio statistic is

$$\begin{aligned}\Lambda(\sigma_0) &= 2l\left(\overline{Y},\tilde{\sigma}\right) - 2l\left(\overline{Y},\sigma_0\right)\\ &= -2n\log(\tilde{\sigma}) - \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 + 2n\log(\sigma_0) + \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\\ &= 2n\log\left(\frac{\sigma_0}{\tilde{\sigma}}\right) + \left(\frac{1}{\sigma_0^2}-\frac{1}{\tilde{\sigma}^2}\right)n\tilde{\sigma}^2\\ &= n\left[\left(\frac{\tilde{\sigma}^2}{\sigma_0^2}-1\right) - \log\left(\frac{\tilde{\sigma}^2}{\sigma_0^2}\right)\right]\end{aligned}$$

This distribution of this random variable is not easy to determine. The likelihood ration statistic is, as one might expect, a function of $\tilde{\sigma}^2/\sigma_0^2$ which is the ratio of the maximum likelihood estimator of the variance divided by the value of σ^2 under H_0. In fact the value of $\Lambda(\sigma_0)$ increases as the quantity $\tilde{\sigma}^2/\sigma_0^2$ gets further away from the value 1 in either direction.

The test proceeds by determining the observed value of $\Lambda(\sigma_0)$

$$\lambda(\sigma_0) = n\left[\left(\frac{\hat{\sigma}^2}{\sigma_0^2}-1\right) - \log\left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)\right]$$

and then obtaining and interpreting the $p-value$

$$p-value \approx P(W>\lambda(\sigma_0)) \quad \text{where } W\sim\chi^2(1)$$

Remark It can be shown that the likelihood ratio statistic $\Lambda(\sigma_0)$ is a function of $U=(n-1)S^2/\sigma_0^2$, in fact $\Lambda(\sigma_0) = U - n\log(U/n) - n$. See Problem 17(b). This is not a one-to-one function of U but $\Lambda(\sigma_0)$ is zero when $U=n$ and $\Lambda(\sigma_0)$ is large when U/n is much bigger than or much less than one (that is, when S^2/σ_0^2 is much bigger than one or much less than one). Since U has a Chi-squared distribution with $n-1$ degrees of freedom when H_0 is true, we can use U as the test statistic for testing $H_0:\sigma=\sigma_0$ and compute exact $p-values$ instead of using the Chi-squared approximation for the distribution of $\Lambda(\sigma_0)$.

Example 5.4.5 Tests of hypotheses for Multinomial model

Consider a random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ with Multinomial probability function

$$f(y_1, y_2, \ldots, y_k; \theta_1, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k} \quad \text{for } y_j = 0, 1, \ldots, n \text{ and } \sum_{j=1}^{k} y_j = n$$

Suppose we wish to test a hypothesis of the form: $H_0 : \theta_j = \theta_j(\boldsymbol{\alpha})$ where the probabilities $\theta_j(\boldsymbol{\alpha})$ are all functions of an unknown parameter (possibly vector) $\boldsymbol{\alpha}$ with dimension $\dim(\alpha) = p < k - 1$. The parameter in the original model is $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ and the parameter space $\Omega = \{(\theta_1, \theta_2, \ldots, \theta_k) : 0 \leq \theta_j \leq 1$, where $\sum_{j=1}^{k} \theta_j = 1\}$ has dimension $k - 1$. The parameter in the model assuming H_0 is $\boldsymbol{\theta}_0 = (\theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), \ldots, \theta_k(\boldsymbol{\alpha}))$ and the parameter space $\Omega_0 = \{(\theta_1(\boldsymbol{\alpha}), \theta_2(\boldsymbol{\alpha}), \ldots, \theta_k(\boldsymbol{\alpha})) : \text{for all } \boldsymbol{\alpha}\}$ has dimension p.

The likelihood function is

$$L(\boldsymbol{\theta}) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

or more simply

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_j^{y_j}$$

$L(\boldsymbol{\theta})$ is maximized over Ω (of dimension $k-1$) by the vector $\hat{\boldsymbol{\theta}}$ with $\hat{\theta}_j = y_j/n$, $j = 1, 2, \ldots, k$.

The likelihood ratio test statistic for testing $H_0 : \theta_j = \theta_j(\boldsymbol{\alpha})$ is

$$\Lambda = -2 \log \left[\frac{L(\tilde{\boldsymbol{\theta}}_0)}{L(\tilde{\boldsymbol{\theta}})} \right]$$

where $L(\boldsymbol{\theta}_0)$ is maximized over Ω_0 by the vector $\tilde{\boldsymbol{\theta}}_0$ with $\hat{\theta}_j = \theta_j(\hat{\boldsymbol{\alpha}})$.

If H_0 is true and n is large the distribution of Λ is approximately $\chi^2(k - 1 - p)$ and the $p - value$ can be calculated approximately as

$$p - value = P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda) \quad \text{where } W \sim \chi^2(k - 1 - p)$$

where

$$\lambda = 2[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)]$$

is the observed value of Λ.

We will give specific examples of the Multinomial model in Chapter 7.

5.5 Chapter 5 Summary

Test of Hypothesis based on Likelihood Ratio Statistic

Suppose $R(\theta) = R(\theta; \mathbf{y})$ is the relative likelihood function for θ based on observed data $\mathbf{y}$ (possible a vector). To test the hypothesis $H_0 : \theta = \theta_0$ we can use the likelihood ratio statistic $-2logR(\theta_0; \mathbf{Y})$ as the test statistic. Let $\lambda = -2logR(\theta_0; \mathbf{y})$ be the observed value of the likelihood ratio statistic for the data $\mathbf{y}$. The corresponding $p-value$ is approximately equal to $P(W \geq \lambda)$ where $W \sim \chi^2(1)$. In R this can be calculated as `1-pchisq(lambda,1)`.

This result is based on the fact that $-2\log R(\theta_0; \mathbf{Y})$ has approximately a $\chi^2(1)$ distribution assuming $H_0 : \theta = \theta_0$ is true.

Table 5.2
Hypothesis Tests for Named Distributions
based on Asymptotic Gaussian Pivotal Quantities

Named Distribution	Point Estimate $\hat{\theta}$	Point Estimator $\tilde{\theta}$	Test Statistic for $H_0 : \theta = \theta_0$	Approximate $p - value$ based on Gaussian approximation
Binomial(n, θ)	$\frac{y}{n}$	$\frac{Y}{n}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right)$ $Z \sim G(0,1)$
Poisson(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\sqrt{\frac{\theta_0}{n}}}\right)$ $Z \sim G(0,1)$
Exponential(θ)	$\bar{y}$	$\overline{Y}$	$\frac{\lvert\tilde{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}$	$2P\left(Z \geq \frac{\lvert\hat{\theta}-\theta_0\rvert}{\frac{\theta_0}{\sqrt{n}}}\right)$ $Z \sim G(0,1)$

Note: To find $2P(Z \geq d)$ where $Z \sim G(0,1)$ in R, use $2 * (1 -$ `pnorm`$(d))$

Table 5.3
Hypothesis Tests for Gaussian
and Exponential Models

Model	Hypothesis	Test Statistic	Exact $p-value$
$G(\mu,\sigma)$ σ known	$H_0:\mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{\sigma/\sqrt{n}}$	$2P\left(Z\geq\frac{\lvert\bar{y}-\mu_0\rvert}{\sigma/\sqrt{n}}\right)$ $Z\sim G(0,1)$
$G(\mu,\sigma)$ σ unknown	$H_0:\mu=\mu_0$	$\frac{\lvert\overline{Y}-\mu_0\rvert}{S/\sqrt{n}}$	$2P\left(T\geq\frac{\lvert\bar{y}-\mu_0\rvert}{s/\sqrt{n}}\right)$ $T\sim t(n-1)$
$G(\mu,\sigma)$ μ unknown	$H_0:\sigma=\sigma_0$	$\frac{(n-1)S^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n-1)s^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n-1)s^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n-1)$
Exponential(θ)	$H_0:\theta=\theta_0$	$\frac{2n\bar{Y}}{\theta_0}$	$\min(2P\left(W\leq\frac{2n\bar{y}}{\theta_0}\right),$ $2P\left(W\geq\frac{2n\bar{y}}{\theta_0}\right))$ $W\sim\chi^2(2n)$

Notes:

(1) To find $P(Z\geq d)$ where $Z\sim G(0,1)$ in R, use $1-$ `pnorm`(d)

(2) To find $P(T\geq d)$ where $T\sim t(k)$ in R, use $1-$ `pt`(d,k)

(3) To find $P(W\leq d)$ where $W\sim\chi^2(k)$ in R, use `pchisq`(d,k)

5.6 Chapter 5 Problems

1. A woman who claims to have special guessing abilities is given a test, as follows: a deck which contains five cards with the numbers 1 to 5 is shuffled and a card drawn out of sight of the woman. The woman then guesses the card, the deck is reshuffled with the card replaced, and the procedure is repeated several times.

   (a) Let θ be the probability the woman guesses the card correctly and let Y be the number of correct guesses in n repetitions of the procedure. Discuss why $Y \sim \text{Binomial}(n, \theta)$ would be an appropriate model. If you wanted to test the hypothesis that the woman is guessing at random what is the appropriate null hypothesis H_0 in terms of the parameter θ?

   (b) Suppose the woman guessed correctly 8 times in 20 repetitions. Using the test statistic $D = |Y - E(Y)|$, calculate the $p - value$ for your hypothesis H_0 in (a) and give a conclusion about whether you think the woman has any special guessing ability.

   (c) In a longer sequence of 100 repetitions over two days, the woman guessed correctly 32 times. Using the test statistic $D = |Y - E(Y)|$, calculate the $p - value$ for these data. What would you conclude now?

2. The accident rate over a certain stretch of highway was about $\theta = 10$ per year for a period of several years. In the most recent year, however, the number of accidents was 25. We want to know whether this many accidents is very probable if $\theta = 10$; if not, we might conclude that the accident rate has increased for some reason. Investigate this question by assuming that the number of accidents in the current year follows a Poisson distribution with mean θ and then testing $H_0 : \theta = 10$. Use the test statistic $D = \max(0, Y - 10)$ where Y represents the number of accidents in the most recent year.

3. A hospital lab has just purchased a new instrument for measuring levels of dioxin (in parts per billion). To calibrate the new instrument, 20 samples of a "standard" water solution known to contain 45 parts per billion dioxin are measured by the new instrument. The observed data are given below:

44.1	46.0	46.6	41.3	44.8	47.8	44.5	45.1	42.9	44.5
42.5	41.5	39.6	42.0	45.8	48.9	46.6	42.9	47.0	43.7

   For these data

   $$\sum_{i=1}^{20} y_i = 888.1 \quad \text{and} \quad \sum_{i=1}^{20} {y_i}^2 = 39545.03$$

   (a) Use a qqplot to check whether a $G(\mu, \sigma)$ model is reasonable for these data.

   (b) Describe a suitable study population for this study. The parameters μ and σ correspond to what attributes of interest in the study population?

(c) Assuming a $G(\mu, \sigma)$ model for these data test the hypothesis $H_0 : \mu = 45$. Determine a 95% confidence interval for μ. What would you conclude about how well the new instrument is working?

(d) The manufacturer of these instruments claims that the variability in measurements is less than two parts per billion. Test the hypothesis that $H_0 : \sigma = 2$ and determine a 95% confidence interval for σ. What would you conclude about the manufacturer's claim?

(e) Suppose the hospital lab rechecks the new instrument one week later by taking 25 new measurements on a standard solution of 45 parts per billion dioxin. If the new data give

$$\bar{y} = 44.1 \quad \text{and} \quad s = 2.1$$

what would you conclude about how well the instrument is working now? Explain the difference between a result which is statistically significant and a result which is of practical significance in the context of this study.

(f) Run the following R code which does the calculations for (c) and (d)

```
y<-c(44.1,46,46.6,41.3,44.8,47.8,44.5,45.1,42.9,44.5,
42.5,41.5,39.6,42,45.8,48.9,46.6,42.9,47,43.7)
t.test(y,mu=45,conf.level=0.95)  # test hypothesis mu=45
# and gives a 95% confidence interval
df<-length(y)-1  # degrees of freedom
s2<-var(y)   # sample variance
p<-0.95      # p=0.95 for 95% confidence interval
a<-qchisq((1-p)/2,df)  # lower value from Chi-squared dist'n
b<-qchisq((1+p)/2,df)  # upper value from Chi-squared dist'n
c(s2*df/b,s2*df/a)  # confidence interval for sigma squared
c(sqrt(s2*df/b),sqrt(s2*df/a))  # confidence interval for sigma
sigma0sq<-2^2   # test hypothesis sigma=2 or sigmasq=4
chitest<-s2*df/sigma0sq
q<-pchisq(chitest,df)
min(2*q,2*(1-q))   # p-value for testing sigma=2
```

4. In Problem 3 suppose we accept the manufacturer's claim and assume we know $\sigma = 2$. Test the hypothesis $H_0 : \mu = 45$ and determine a 95% confidence interval for μ for the original data with $\bar{y} = 44.405$.
Hint: Use the pivotal quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim G(0, 1)$$

with $\sigma = 2$.

5. For Chapter 4, Problem 32 test the hypothesis $H_0 : \mu = 105$.

6. Suppose in Problem 5 we assume that $\mu = 105$. Test the hypothesis $H_0 : \sigma^2 = 100$ and determine a 95% confidence interval for σ.
Hint: Use the pivotal quantity

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 \sim \chi^2(n)$$

with $\mu = 105$.

7. Between 10 a.m. on November 4, 2014 and 10 p.m. on November 6, 2014 a referendum on the question "Should classes start on the first Thursday after Labour Day to allow for two additional days off in the Fall term?" was conducted by the Federation of Students at the University of Waterloo. All undergraduates were able to cast their ballot online. Six thousand of the 30,990 eligible voters voted. Of the 6000 who voted, 4440 answered yes to this question.

   (a) The Federation of Students used an empirical study to determine whether or not students support a fall term break. The Plan step of the empirical study involved using an online referendum. Give at least one advantage and at least one disadvantage of using the online referendum in this context.

   (b) Describe a suitable target population and study population for this study.

   (c) Assume the model $Y \sim \text{Binomial}(6000, \theta)$ where Y = number of people who responded yes to the question "Should classes start on the first Thursday after Labour Day to allow for two additional days off in the Fall term?" The parameter θ corresponds to what attribute of interest in the study population? How valid do you think the Binomial model is and why?

   (d) Give the maximum likelihood estimate of θ. How valid do you think this estimate is?

   (e) Determine an approximate 95% confidence interval for θ.

   (f) By reference to the approximate confidence interval, indicate what you know about the approximate $p - value$ for a test of the hypothesis $H_0 : \theta = 0.7$.

8. Data on the number of accidents at a busy intersection in Waterloo over the last 5 years indicated that the average number of accidents at the intersection was 3 accidents per week. After the installation of new traffic signals the number of accidents per week for a 25 week period were recorded as follows:

4	5	0	4	2	0	1	4	1	3	1	1	2
2	2	1	1	3	2	3	2	0	2	2	3	

Let y_i = the number of accidents in week i, $i = 1, 2, \ldots, 25$. To analyse these data we assume Y_i has a Poisson distribution with mean θ, $i = 1, 2, \ldots, 25$ independently.

(a) To decide whether the mean number of accidents at this intersection has changed after the installation of the new traffic signals we wish to test the hypothesis H_0 : $\theta = 3$. Why is the discrepancy measure $D = \left|\sum_{i=1}^{25} Y_i - 75\right|$ reasonable? Calculate the exact $p - value$ for testing $H_0 : \theta = 3$. What would you conclude?

(b) Justify the following statement:

$$P\left(\frac{\overline{Y} - \theta}{\sqrt{\theta/n}} \leq c\right) \approx P\left(Z \leq c\right) \quad \text{where} \quad Z \sim N\left(0, 1\right)$$

(c) Why is the discrepancy measure $D = \left|\overline{Y} - 3\right|$ reasonable for testing $H_0 : \theta = 3$? Calculate the approximate $p - value$ using the approximation in (b). Compare this to the value in (a).

9. Use the likelihood ratio test statistic to test $H_0 : \theta = 3$ for the data in Problem 8. Compare this answer to the answers in $8\,(a)$ and $8\,(c)$.

10. For Chapter 2, Problem $6\,(b)$ test the hypothesis $H_0 : \theta = 5$ using the likelihood ratio test statistic. Is this result consistent with the approximate 95% confidence interval for θ that you found in Chapter 4, Problem 5?

11. For Chapter 2, Problem $8\,(b)$ test the hypothesis $H_0 : \theta = -0.1$ using the likelihood ratio test statistic. Is this result consistent with the approximate 95% confidence interval for θ that you found in Chapter 4, Problem 6?

12. Data from the 2011 Canadian census indicate that 18% of all families in Canada have one child. Suppose the data in Chapter 2, Problem $13\,(d)$ represented 33 children chosen at random from the Waterloo Region. Based on these data, test the hypothesis that the percentage of families with one child in Waterloo Region is the same as the national percentage using the likelihood ratio test statistic. Is this result consistent with the approximate 95% confidence interval for θ that you found in Chapter 4, Problem 8?

13. A company that produces power systems for personal computers has to demonstrate a high degree of reliability for its systems. Because the systems are very reliable under normal use conditions, it is customary to 'stress' the systems by running them at a considerably higher temperature than they would normally encounter, and to measure the time until the system fails. According to a contract with one personal computer manufacturer, the average time to failure for systems run at $70°C$ should be no less than $1,000$ hours. From one production lot, 20 power systems were put on test and observed until failure at $70°C$. The 20 failure times $y_1, y_2, \ldots, y_{20}$ were (in hours):

374.2	544.0	509.4	1113.9	1244.3	551.9	853.2	3391.2	297.0	1501.4
250.2	678.1	379.6	1818.9	1191.1	162.8	332.2	1060.1	63.1	2382.0

Note: $\sum_{i=1}^{20} y_i = 18,698.6$. Failure times are assumed to have an Exponential(θ) distribution.

(a) Check whether the Exponential model is reasonable for these data. (See Example 5.3.2.)

(b) Use a likelihood ratio test to test $H_0 : \theta = 1000$ hours. Is there any evidence that the company's power systems do not meet the contracted standard?

14. The R function `runif()` generates pseudo random Uniform$(0, 1)$ random variables. The command `y<-runif(n)` will produce a vector of n values $y_1, y_2, \ldots, y_n$.

(a) Suggest a test statistic which could be used to test that the y_i's, $i = 1, 2, \ldots, n$ are consistent with a random sample from Uniform$(0, 1)$.
(See: www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA393366)

(b) Generate 1000 y_i's and carry out the test in (a).

15. The Poisson model is often used to compare rates of occurrence for certain types of events in different geographic regions. For example, consider K regions with populations $P_1, P_2, \ldots, P_K$ and let θ_j, $j = 1, 2, \ldots, K$ be the annual expected number of events per person for region j. By assuming that the number of events Y_j for region j in a given t-year period has a Poisson distribution with mean $P_j\theta_j t$, we can estimate and compare the θ_j's or test that they are equal.

(a) Under what conditions might the stated Poisson model be reasonable?

(b) Suppose you observe values $y_1, y_2, \ldots, y_K$ for a given t-year period. Describe how to test the hypothesis that $\theta_1 = \theta_2 = \cdots = \theta_K$.

(c) The data below show the numbers of children y_j born with "birth defects" for 5 regions over a given five year period, along with the total numbers of births P_j for each region. Test the hypothesis that the five rates of birth defects are equal.

P_j	2025	1116	3210	1687	2840
y_j	27	18	41	29	31

16. Using the data from Chapter 2, Problems 11 and 12 and assuming the Poisson model holds for each dataset, test the hypothesis that the mean number of points per game is the same for Wayne Gretzky and Sidney Crosby. **Hint:** See Example 5.4.3. Comment on whether you think this is a reasonable way to compare these two great hockey players.

17. **Challenge Problem: Likelihood ratio test statistic for Gaussian model μ and σ unknown** Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent $G(\mu, \sigma)$ observations.

    (a) Show that the likelihood ratio test statistic for testing $H_0 : \mu = \mu_0$ (σ unknown) is given by

    $$\Lambda(\mu_0) = n \log\left(1 + \frac{T^2}{n-1}\right)$$

    where $T = \sqrt{n}(\overline{Y} - \mu_0)/S$ and S is the sample standard deviation. Note: you will want to use the identity

    $$\sum_{i=1}^{n}(Y_i - \mu_0)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + n(\overline{Y} - \mu_0)^2$$

    (b) Show that the likelihood ratio test statistic for testing $H_0 : \sigma = \sigma_0$ (μ unknown) can be written as $\Lambda(\sigma_0) = U - n \log(U/n) - n$ where

    $$U = \frac{(n-1)S^2}{\sigma_0^2}$$

    See Example 5.4.4.

18. **Challenge Problem: Likelihood ratio test statistic for comparing two Exponential means** Suppose that $X_1, X_2, \ldots, X_m$ is a random sample from the Exponential(θ_1) distribution and independently and $Y_1, Y_2, \ldots, Y_n$ is a random sample from the Exponential(θ_2) distribution. Determine the likelihood ratio test statistic for testing $H_0 : \theta_1 = \theta_2$.

6. GAUSSIAN RESPONSE MODELS

6.1 Introduction

A response variate Y is one whose distribution has parameters which depend on the value of other variates. For the Gaussian models we have studied so far, we assumed that we had a random sample $Y_1, Y_2, \ldots, Y_n$ from the *same* Gaussian distribution $G(\mu, \sigma)$. A Gaussian response model generalizes this to permit the parameters of the Gaussian distribution for Y_i to depend on a vector $\mathbf{x}_i$ of *covariates* (explanatory variates which are measured for the response variate Y_i). Gaussian models are by far the most common models used in statistics.

Definition 40 *A* Gaussian response model *is one for which the distribution of the response variate Y, given the associated vector of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ for an individual unit, is of the form*

$$Y \sim G(\mu(\mathbf{x}), \sigma(\mathbf{x}))$$

If observations are made on n randomly selected units we write the model as

$$Y_i \sim G(\mu(\mathbf{x}_i), \sigma(\mathbf{x}_i)) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

In most examples we will assume $\sigma(\mathbf{x}_i) = \sigma$ is constant. This assumption is not necessary but it does make the models easier to analyze. The choice of $\mu(\mathbf{x})$ is guided by past information and on current data from the population or process. The difference between various Gaussian response models is in the choice of the function $\mu(\mathbf{x})$ and the covariates. We often assume $\mu(\mathbf{x}_i)$ is a *linear function* of the covariates. These models are called *Gaussian linear models* and can be written as

$$Y_i \sim G(\mu(\mathbf{x}_i), \sigma) \text{ for } i = 1, 2, \ldots, n \text{ independently} \tag{6.1}$$
$$\text{with } \mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is the vector of known covariates associated with unit i and $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters. These models are also referred to as *linear regression models*[14], and the β_j's are called the *regression coefficients*. Linear regressions models are used in both machine learning and data science.

Here are some examples of settings where Gaussian response models can be used.

Example 6.1.1 Can filler study

The soft drink bottle filling process of Example 1.5.2 involved two machines (Old and New). For a given machine it is reasonable to represent the distribution for the amount of liquid Y deposited in a single bottle by a Gaussian distribution.

In this case we can think of the machines as acting like a covariate, with μ and σ differing for the two machines. We could write

$$\begin{aligned} Y &\sim G(\mu_O, \sigma_O) \quad \text{for observations from the old machine} \\ Y &\sim G(\mu_N, \sigma_N) \quad \text{for observations from the new machine.} \end{aligned}$$

In this case there is no formula relating μ and σ to the machines; they are simply different. Notice that an important feature of a machine is the variability of its production so we have, in this case, permitted the two variance parameters to be different.

Example 6.1.2 Price versus size of commercial building

Ontario property taxes are based on "market value", which is determined by comparing a property to the price of those which have recently been sold. The value of a property is separated into components for land and for buildings. Here we deal with the value of the buildings only but a similar analysis could be conducted for the value of the property.

A manufacturing company was appealing the assessed market value of its property, which included a large building. Sales records were collected on the 30 largest buildings sold in the previous three years in the area. The data, which are available in the file *sizepricedata.txt* posted on the course website, are plotted in Figure 6.1. The size of the building x is measured in $m^2/10^5$ and the selling price y is in \$ per m^2. The purpose of the analysis is to determine whether and to what extent we can determine the value of a property from the single covariate x so that we know whether the assessed value appears to be too high. The size of the building in question was 4.47×10^5 m^2, with an assessed market value of \$75 per m^2.

The scatterplot shows that the price y decreases linearly with size x but there is obviously variability in the price of buildings having the same area (size). In this case we might consider a model where the price of a building of size x_i is represented by a random variable Y_i, with

$$Y_i \sim G\left(\beta_0 + \beta_1 x_i, \sigma\right) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

[14] The word *regression* is an historical term introduced in the 19th century in connection with these models.

where β_0, β_1 and σ are unknown parameters and $x_1, x_2, \ldots, x_n$ are known constants. Note that, although this model assumes that the mean of the response variate Y depends on the explanatory variate x, the model assumes that the standard deviation σ does not depend on x.

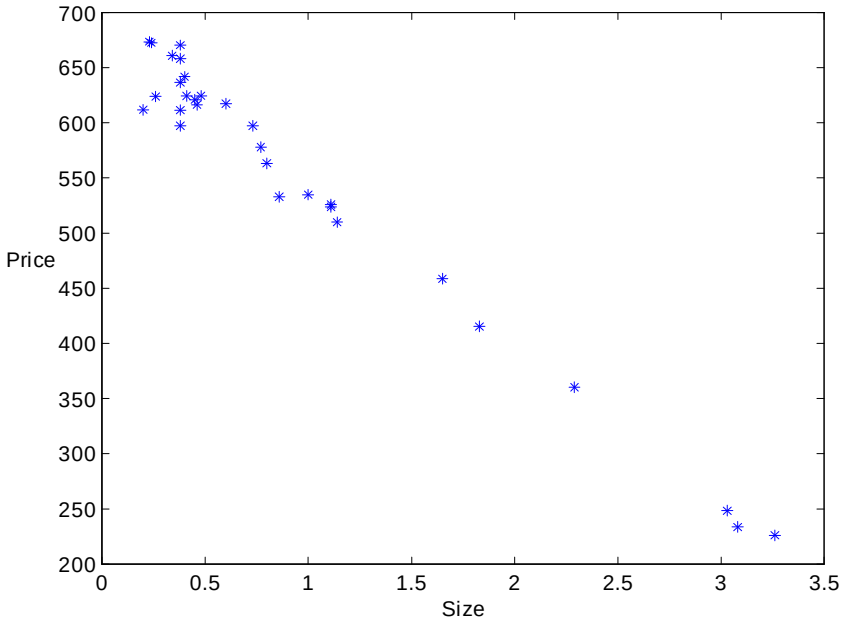


Figure 6.1: **Scatterplot of price versus building size**

Example 6.1.3 Exam mark versus midterm mark

An instructor of an online course was interested in the relationship between midterm marks and final exam marks. The data are (x_i, y_i), $i = 1, 2, \ldots, 65$ where y_i = final exam mark and x_i = midterm mark for 65 students enrolled in the online course during a particular winter term. The data, which are available in the file *midexamdata.txt* posted on the course website, are plotted in Figure 6.2. The scatterplot shows the final exam mark y increases linearly with midterm mark x. The variability in final exam marks seems greater for midterm marks between 55 and 70 however we notice that this is mostly due to the fact that there are many more observations in this range as compared to the number of observations for midterm marks below 45 and above 90. For these data it also seems reasonable to use the model

$$Y_i \sim G\left(\beta_0 + \beta_1 x_i, \sigma\right) \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

where β_0, β_1 and σ are unknown parameters and and $x_1, x_2, \ldots, x_n$ are known constants. The standard deviation σ is assumed to be the same for all x.

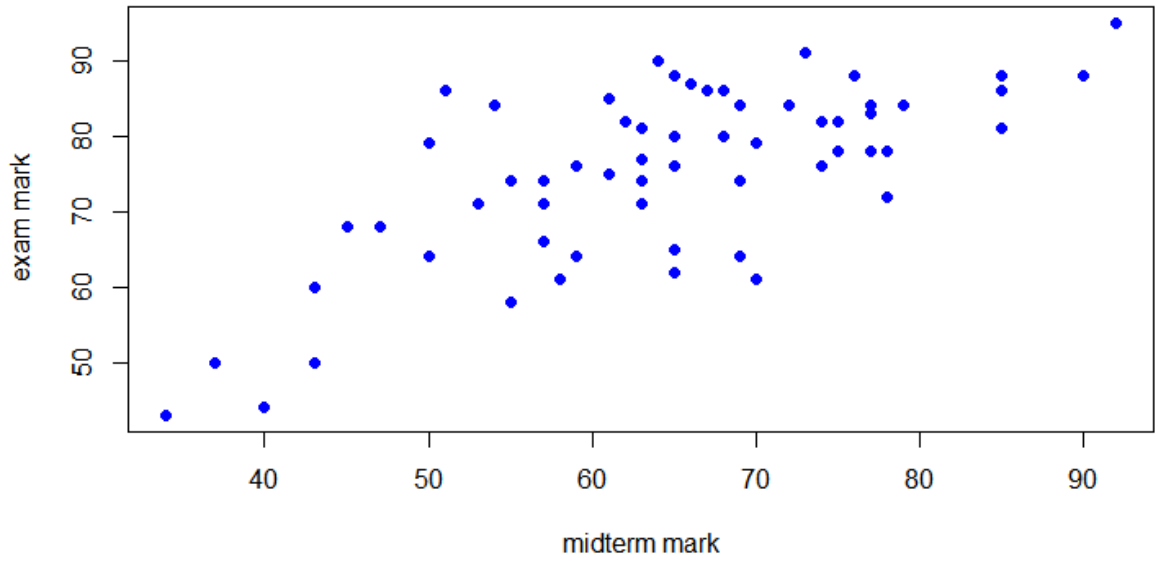


Figure 6.2: **Scatterplot of exam mark versus midterm mark**

Example 6.1.4 Breaking strength versus diameter of steel bolt

The "breaking strength" of steel bolts is measured by subjecting a bolt to an increasing (lateral) force and determining the force at which the bolt breaks. This force is called the breaking strength; it depends on the diameter of the bolt and the material the bolt is composed of. There is variability in breaking strengths since two bolts of the same dimension and material will generally break at different forces. Understanding the distribution of breaking strengths is very important in manufacturing and construction.

In a quality control experiment the breaking strengths y of six steel bolts at each of five different bolt diameters x were measured. The data, which are available in the file *diameterstrengthdata.txt*, are plotted in Figure 6.3. The scatterplot gives a clear picture of the relationship between y and x. A reasonable model for the breaking strength Y of a randomly selected bolt of diameter x would appear to be $Y \sim G(\mu(x), \sigma)$. The variability in y values appears to be about the same for bolts of different diameters which again provides some justification for assuming σ to be constant. It is not obvious what the best choice for $\mu(x)$ would be although the relationship looks slightly nonlinear so we might try a quadratic function

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where $\beta_0, \beta_1, \beta_2$ are unknown parameters.

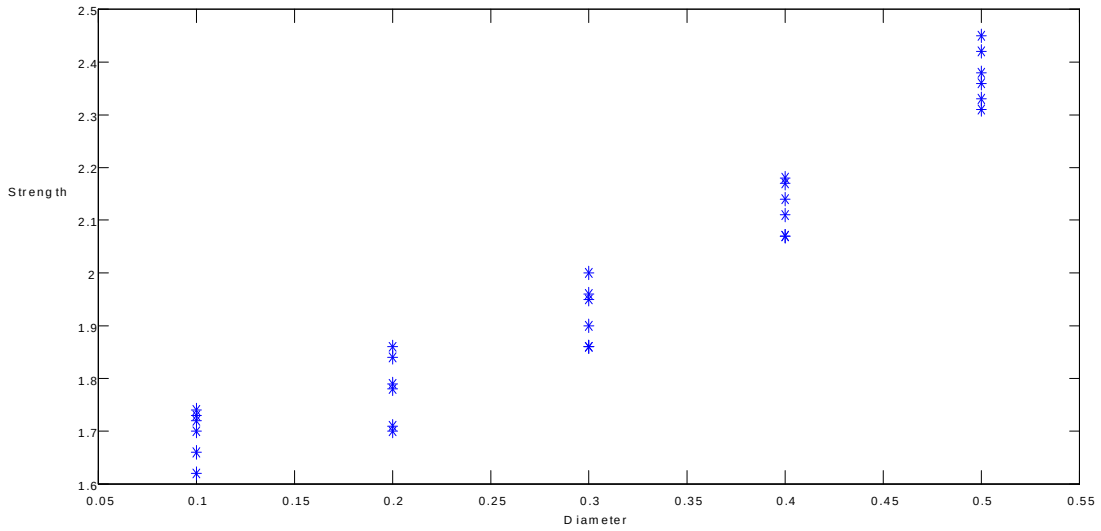


Figure 6.3: **Scatterplot of strength versus bolt diameter**

Remark Sometimes the model (6.1) is written as

$$Y_i = \mu(\mathbf{x}_i) + R_i \quad \text{where } R_i \sim G(0, \sigma)$$

In this form we can see that Y_i is the sum of a deterministic component, $\mu(\mathbf{x}_i)$ (a constant), and a stochastic component, R_i (a random variable).

We now consider estimation and testing procedures for these Gaussian response models. We begin with models which have no covariates so that the observations are all from the same Gaussian distribution.

$G(\mu, \sigma)$ Model

In Chapters 4 and 5 we discussed estimation and testing hypotheses for samples from a Gaussian distribution. Suppose that $Y \sim G(\mu, \sigma)$ models a response variate y in some population or process. A random sample $Y_1, Y_2, \ldots, Y_n$ is selected, and we want to estimate the model parameters and possibly to test hypotheses about them. We can write this model in the form

$$Y_i = \mu + R_i \quad \text{where } R_i \sim G(0, \sigma) \tag{6.2}$$

so this is a special case of the Gaussian response model in which the mean function is constant. The estimator of the parameter μ that we used is the maximum likelihood estimator $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. This estimator is also a "least squares estimator". $\overline{Y}$ has the property that it is closer to the data than any other constant, or

$$\min_{\mu} \sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

You should be able to verify this. It will turn out that the methods for estimation, constructing confidence intervals, and tests of hypothesis discussed earlier for the single Gaussian $G(\mu, \sigma)$ are all special cases of the more general methods derived in Section 6.5.

In the next section we begin with a simple generalization of (6.2) to the case in which the mean is a linear function of a single covariate.

6.2 Simple Linear Regression

Many studies involve covariates $\mathbf{x}$, as described in Section 6.1. In this section we consider the case in which there is a single covariate x. Consider the model with independent Y_i's such that

$$Y_i \sim G(\mu(x_i), \sigma) \quad \text{where} \quad \mu(x_i) = \alpha + \beta x_i \tag{6.3}$$

This is of the form (6.1) with (β_0, β_1) replaced by (α, β). The x_i's are assumed to be known constants. The unknown parameters are α, β, and σ.

The likelihood function for (α, β, σ) is

$$L(\alpha, \beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

or more simply

$$L(\alpha, \beta, \sigma) = \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right] \quad \text{for } \alpha \in \Re,\ \beta \in \Re,\ \sigma > 0$$

The log likelihood function is

$$l(\alpha, \beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \quad \text{for } \alpha \in \Re,\ \beta \in \Re,\ \sigma > 0$$

To obtain the maximum likelihood estimates we solve the three equations

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \tag{6.4}$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)\, x_i = 0 \tag{6.5}$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = 0$$

simultaneously to obtain the maximum likelihood estimates

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}\left(S_{yy} - \hat{\beta}S_{xy}\right)$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

See Chapter 6, Problems 1 and 2.

Least squares estimation

If we are given data (x_i, y_i), $i = 1, 2, \ldots, n$ then one criterion which could be used to obtain a line of "best fit" to these data is to fit the line which minimizes the sum of the squares of the distances between the observed points, (x_i, y_i), $i = 1, 2, \ldots, n$, and the fitted line $y = \alpha + \beta x$. Mathematically this means we want to find the values of α and β which minimize the function

$$g(\alpha, \beta) = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

Such estimates are called *least squares estimates.* To find the least squares estimates we need to solve the two equations

$$\begin{aligned} \frac{\partial g}{\partial \alpha} &= \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial g}{\partial \beta} &= \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0 \end{aligned}$$

simultaneously. We note that this is equivalent to solving the maximum likelihood equations (6.4) and (6.5).

In summary we have that the least squares estimates and the maximum likelihood estimates obtained assuming the model (6.3) are the same estimates. Of course the *method of least squares* only provides point estimates of the unknown parameters α and β while assuming the model (6.3) allows us to obtain both estimates and confidence intervals for the unknown parameters.

Note that the line $y = \hat{\alpha} + \hat{\beta} x$ is often called *the fitted regression line for y on x* or more simply *the fitted line.*

We now show how to obtain confidence intervals based on the model (6.3).

Distribution of the estimator $\tilde{\beta}$

The maximum likelihood estimator corresponding to $\hat{\beta}$ is

$$\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} x_i (Y_i - \overline{Y})$$

Since

$$\sum_{i=1}^{n} x_i (Y_i - \overline{Y}) = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$$

(see Chapter 6, Problem 1) we have

$$\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i = \sum_{i=1}^{n} a_i Y_i \quad \text{where } a_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

which shows that $\tilde{\beta}$ is a linear combination of the Gaussian random variables Y_i and therefore has a Gaussian distribution. To find the mean and variance of $\tilde{\beta}$ we use the identities

$$\sum_{i=1}^{n} a_i = 0, \quad \sum_{i=1}^{n} a_i x_i = 1, \quad \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{xx}}$$

(see Chapter 6, Problem 1) to obtain

$$\begin{aligned} E(\tilde{\beta}) &= \sum_{i=1}^{n} a_i E(Y_i) = \sum_{i=1}^{n} a_i (\alpha + \beta x_i) \\ &= \beta \sum_{i=1}^{n} a_i x_i \quad \text{since } \sum_{i=1}^{n} a_i = 0 \\ &= \beta \quad \text{since } \sum_{i=1}^{n} a_i x_i = 1 \end{aligned}$$

and

$$\begin{aligned} Var(\tilde{\beta}) &= \sum_{i=1}^{n} a_i^2 Var(Y_i) \quad \text{since the } Y_i \text{ are independent random variables} \\ &= \sigma^2 \sum_{i=1}^{n} a_i^2 \\ &= \frac{\sigma^2}{S_{xx}} \quad \text{since } \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{xx}} \end{aligned}$$

In summary

$$\tilde{\beta} \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

Confidence intervals for β and test of hypothesis of no relationship

Although the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \frac{1}{n} \left(S_{yy} - \hat{\beta} S_{xy} \right)$$

we will estimate σ^2 using

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \frac{1}{n-2} \left(S_{yy} - \hat{\beta} S_{xy} \right)$$

since $E\left(S_e^2\right) = \sigma^2$ where

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} x_i)^2$$

Confidence intervals for β are important because the parameter β represents the increase in the mean value of the response Y, resulting from an increase of one unit in the value of the covariate x. As well, if $\beta = 0$ then x has no effect on the mean of Y (within this model). Since

$$\frac{\tilde{\beta} - \beta}{\sigma / \sqrt{S_{xx}}} \sim G(0, 1)$$

holds independently of

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2) \tag{6.6}$$

then by Theorem 32 it follows that

$$\frac{\tilde{\beta} - \beta}{S_e / \sqrt{S_{xx}}} \sim t(n-2)$$

This pivotal quantity can be used to obtain confidence intervals for β and to construct tests of hypotheses about β.

Using the t table or R we find the constant a such that $P(-a \leq T \leq a) = p$ where $T \sim t(n-2)$. Since

$$\begin{aligned} p &= P(-a \leq T \leq a) \\ &= P\left(-a \leq \frac{\tilde{\beta} - \beta}{S_e / \sqrt{S_{xx}}} \leq a\right) \\ &= P\left(\tilde{\beta} - aS_e/\sqrt{S_{xx}} \leq \beta \leq \tilde{\beta} + aS_e/\sqrt{S_{xx}}\right) \end{aligned}$$

therefore a $100p\%$ confidence interval for β is given by

$$\begin{aligned} &\left[\hat{\beta} - as_e/\sqrt{S_{xx}}, \hat{\beta} + as_e/\sqrt{S_{xx}}\right] \\ = \ &\hat{\beta} \pm as_e/\sqrt{S_{xx}} \end{aligned} \tag{6.7}$$

To test the hypothesis of no relationship or $H_0 : \beta = 0$ we use the test statistic

$$\frac{\left|\tilde{\beta} - 0\right|}{S_e / \sqrt{S_{xx}}}$$

with observed value

$$\frac{\left|\hat{\beta} - 0\right|}{s_e / \sqrt{S_{xx}}}$$

and $p - value$ given by

$$\begin{aligned} p - value &= P\left(|T| \geq \frac{\left|\hat{\beta} - 0\right|}{s_e / \sqrt{S_{xx}}}\right) \\ &= 2\left[1 - P\left(T \leq \frac{\left|\hat{\beta} - 0\right|}{s_e / \sqrt{S_{xx}}}\right)\right] \quad \text{where} \quad T \sim t(n-2) \end{aligned}$$

Note that (6.6) can be used to obtain confidence intervals for σ^2 or σ. A $100p\%$ confidence interval for σ^2 is

$$\left[\frac{(n-2)s_e^2}{b},\ \frac{(n-2)s_e^2}{a}\right]$$

and a $100p\%$ confidence interval for σ is

$$\left[s_e\sqrt{\frac{n-2}{b}},\ s_e\sqrt{\frac{n-2}{a}}\right]$$

where

$$P\left(U \leq a\right) = P\left(U > b\right) = \frac{1-p}{2} \quad \text{and } U \sim \chi^2\left(n-2\right)$$

The parameter σ corresponds to the varibility in the response Y for each value of the covariate x. For the simple linear regression model the parameter σ can be thought of graphically as modeling the variability in the responses about the true line $y = \alpha + \beta x$ in the vertical direction for each value of the covariate x.

Remark In regression models we often "redefine" a covariate x_i as $x_i^* = x_i - c$, where c is a constant value that makes $\sum_{i=1}^{n} x_i^*$ close to zero. (Often we take $c = \bar{x}$, which makes $\sum_{i=1}^{n} x_i^*$ exactly zero.) The reasons for doing this are that it reduces round-off errors in calculations, and that it makes the parameter α more interpretable. Note that β does not change if we "centre" x_i this way, because

$$E(Y|x) = \alpha + \beta x = \alpha + \beta(x^* + c) = (\alpha + \beta c) + \beta x^*$$

Thus, the intercept α changes if we redefine x, but not β. In the examples we consider here we have kept the given definition of x_i, for simplicity.

Confidence intervals for the mean response $\mu(x) = \alpha + \beta x$

We are often interested in estimating the quantity $\mu(x) = \alpha + \beta x$ since it represents the mean response at a specified value of the covariate x. We can obtain a pivotal quantity for doing this. The maximum likelihood estimator of $\mu(x)$ obtains by replacing the unknown values α, β by their maximum likelihood estimators,

$$\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x = \overline{Y} + \tilde{\beta}(x - \bar{x})$$

since $\tilde{\alpha} = \overline{Y} - \tilde{\beta}\bar{x}$. Since

$$\tilde{\beta} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i$$

we can rewrite $\tilde{\mu}(x)$ as

$$\tilde{\mu}(x) = \overline{Y} + \tilde{\beta}(x - \bar{x}) = \sum_{i=1}^{n} b_i Y_i \quad \text{where} \quad b_i = \frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}}$$

Since $\tilde{\mu}(x)$ is a linear combination of Gaussian random variables it has a Gaussian distribution. To find the mean and variance of $\tilde{\mu}(x)$ we use the identities

$$\sum_{i=1}^{n} b_i = 1, \quad \sum_{i=1}^{n} b_i x_i = x \quad \text{and} \quad \sum_{i=1}^{n} b_i^2 = \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}$$

(see Chapter 6, Problem 1) to obtain

$$\begin{aligned} E[\tilde{\mu}(x)] &= \sum_{i=1}^{n} b_i E(Y_i) = \sum_{i=1}^{n} b_i (\alpha + \beta x_i) \\ &= \alpha \left(\sum_{i=1}^{n} b_i \right) + \beta \left(\sum_{i=1}^{n} b_i x_i \right) \\ &= \alpha + \beta x \quad \text{since } \sum_{i=1}^{n} b_i = 1 \text{ and } \sum_{i=1}^{n} b_i x_i = x \\ &= \mu(x) \end{aligned}$$

and

$$\begin{aligned} Var\left[\tilde{\mu}(x)\right] &= \sum_{i=1}^{n} b_i^2 Var(Y_i) \quad \text{since the } Y_i \text{ are independent random variables} \\ &= \sigma^2 \sum_{i=1}^{n} b_i^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \right] \end{aligned}$$

Note that the variance of $\tilde{\mu}(x)$ is smallest when x is close to $\bar{x}$ (the center of the data) and much larger when $(x-\bar{x})^2$ is large. In summary, we have shown that

$$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \right)$$

Since

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

holds independently of (6.6) then by Theorem (32) we obtain the pivotal quantity

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

which can be used to obtain confidence intervals for $\mu(x)$ in the usual manner. Using the t

table or R find the constant a such that $P(-a \leq T \leq a) = p$ where $T \sim t(n-2)$. Since

$$\begin{aligned} p &= P(-a \leq T \leq a) \\ &= P\left(-a \leq \frac{\tilde{\mu}(x) - \mu(x)}{S_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \leq a\right) \\ &= P\left(\tilde{\mu}(x) - aS_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \leq \mu(x) \leq \tilde{\mu}(x) + aS_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right) \end{aligned}$$

a $100p\%$ confidence interval for $\mu(x)$ is given by

$$\left[\hat{\mu}(x) - as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}},\ \hat{\mu}(x) + as_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right] \tag{6.8}$$

where $\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$.

Remark Note that since $\alpha = \mu(0)$, a 95% confidence interval for α, is given by (6.8) with $x = 0$ which gives

$$\hat{\alpha} \pm as_e\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}} \tag{6.9}$$

In fact one can see from (6.9) that if $\bar{x}$ is large in magnitude (which means the average x_i is large), then the confidence interval for α will be very wide. This would be disturbing if the value $x = 0$ is a value of interest, but often it is not.

Prediction Interval for Future Response

In Section 4.7 we constructed a $100p\%$ prediction interval for a future observation when the data were assumed to arise from a $G(\mu, \sigma)$ distribution. In the case of simple linear regression we also like to estimate or predict the Y value for a random unit, not part of the sample, which has a specific value x for its covariate. To obtain a pivotal quantity that can be used to construct a prediction interval for the future response Y, we note that $Y \sim G(\mu(x), \sigma)$ from (6.3) or alternatively

$$Y = \mu(x) + R, \quad \text{where } R \sim G(0, \sigma)$$

is independent of $Y_1, Y_2, \ldots, Y_n$. For a point estimator of Y it is natural to use the maximum likelihood estimator $\tilde{\mu}(x)$ of $\mu(x)$ which has distribution

$$\tilde{\mu}(x) \sim G\left(\mu(x), \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right)$$

The error in the point estimator of Y is given by

$$Y - \tilde{\mu}(x) = Y - \mu(x) + \mu(x) - \tilde{\mu}(x) = R + [\mu(x) - \tilde{\mu}(x)] \tag{6.10}$$

Since R is independent of $\tilde{\mu}(x)$ (it is not connected to the existing sample), (6.10) is the sum of independent Normally distributed random variables and is consequently Normally distributed. Since

$$\begin{aligned} E[Y - \tilde{\mu}(x)] &= E\{R + [\mu(x) - \tilde{\mu}(x)]\} \\ &= E(R) + E[\mu(x)] - E[\tilde{\mu}(x)] \\ &= 0 + \mu(x) - \mu(x) = 0 \end{aligned}$$

and

$$\begin{aligned} Var[Y - \tilde{\mu}(x)] &= Var(Y) + Var[\tilde{\mu}(x)] \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] \end{aligned}$$

we have

$$Y - \tilde{\mu}(x) \sim G\left(0, \sigma\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]^{1/2}\right)$$

or

$$\frac{Y - \tilde{\mu}(x)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1) \tag{6.11}$$

Since (6.11) holds independently of (6.6) then by Theorem (32) we obtain the pivotal quantity

$$\frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

For an interval estimate with confidence coefficient p we choose a such that $p = P(-a \leq T \leq a)$ where $T \sim t(n-2)$. Since

$$\begin{aligned} p &= P\left(-a \leq \frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \leq a\right) \\ &= P\left(\tilde{\mu}(x) - aS_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \leq Y \leq \tilde{\mu}(x) + aS_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right) \end{aligned}$$

we obtain the $100p\%$ prediction interval

$$\left[\hat{\mu}(x) - as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}},\ \hat{\mu}(x) + as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right] \tag{6.12}$$

If we compare (6.8) and (6.12), we observe that the prediction interval will be wider than the confidence interval particularly if n is large. The prediction interval is an interval for a future observation Y which is a random variable whereas the confidence interval is an interval for the unknown mean $\mu(x) = \alpha + \beta x$. The width of the confidence interval depends on the uncertainty in the estimation of the parameters α and β, that is, it depends on the variances of the estimators $\tilde{\alpha}$ and $\tilde{\beta}$. The width of the prediction interval depends on the uncertainty in the estimation of the parameters α and β as well the variance σ^2 of the random variable. In other words the uncertainty in determining an interval for a random variable Y is greater than the uncertainty in determining an interval for the constant $\mu(x) = \alpha + \beta x$.

Remark When we construct a confidence interval or a prediction interval for a value of x which lie outside the interval of observed x_i's we are assuming that the linear relationship holds beyond the observed data. This is dangerous since there are no data to support this assumption.

These results are summarized in Tables 6.1 and 6.2.

Example 6.1.3 Revisited Exam mark versus midterm mark

In Example 6.1.3, Figure 6.2 suggested that a linear regression model of the form $E(Y|x) = \alpha + \beta x$ would be reasonable for the data on final exam mark y and midterm mark x. For the given data

$$\begin{aligned} n &= 65 \\ \bar{x} &= 65.06154 \qquad \bar{y} = 75.38462 \\ S_{xx} &= 10813.75 \\ S_{xy} &= 6869.462 \\ S_{yy} &= 8665.385 \end{aligned}$$

so we find

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{6869.462}{10813.75} = 0.6352523 \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 75.38462 - (0.6352523)(65.06154) = 34.05413 \\ s_e^2 &= \frac{S_{yy} - \hat{\beta} S_{xy}}{n-2} = \frac{1}{63}\left[8665.385 - (0.6352523)(6869.462)\right] = 68.27847 \\ s_e &= 8.263079 \end{aligned}$$

Note that when calculating these values using a calculator you should use as many decimal places as possible otherwise the values are affected by roundoff error. The estimate $\hat{\beta} = 0.6352523$ indicates an increase in average exam mark of 0.6352523 for each one mark increase in midterm mark x.

Table 6.1
Confidence/Prediction Intervals for
Simple Linear Regression Model

Unknown Quantity	Estimate	Estimator	Pivotal Quantity	$100p\%$ Confidence/ Prediction Interval
β	$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$	$\tilde{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}$	$\frac{\tilde{\beta}-\beta}{S_e/\sqrt{S_{xx}}} \sim t(n-2)$	$\hat{\beta} \pm as_e/\sqrt{S_{xx}}$
α	$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$	$\tilde{\alpha} = \overline{Y} - \tilde{\beta}\bar{x}$	$\frac{\tilde{\alpha}-\alpha}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	$\hat{\alpha} \pm as_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}$
$\mu(x) = \alpha+\beta x$	$\hat{\mu}(x) = \hat{\alpha}+\hat{\beta}x$	$\tilde{\mu}(x) = \tilde{\alpha}+\tilde{\beta}x$	$\frac{\tilde{\mu}(x)-\mu(x)}{S_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	$\hat{\mu}(x) \pm as_e\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$
σ^2	$s_e^2 = \frac{S_{yy}-\hat{\beta}S_{xy}}{n-2}$	$S_e^2 = \frac{\sum_{i=1}^{n}\left(Y_i-\tilde{\alpha}-\tilde{\beta}x_i\right)^2}{n-2}$	$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2(n-2)$	$\left[\frac{(n-2)s_e^2}{c}, \frac{(n-2)s_e^2}{b}\right]$
Y			$\frac{Y-\tilde{\mu}(x)}{S_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}} \sim t(n-2)$	Prediction Interval $\hat{\mu}(x) \pm as_e\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}}$

Notes: The value a is given by $P(T \le a) = \frac{1+p}{2}$ where $T \sim t(n-2)$.
The values b and c are given by $P(W \le b) = \frac{1-p}{2} = P(W > c)$ where $W \sim \chi^2(n-2)$.

Table 6.2
Hypothesis Tests for
Simple Linear Regression Model

Hypothesis	Test Statistic	$p-value$
$H_0:\beta=\beta_0$	$\frac{\left\|\tilde{\beta}-\beta_0\right\|}{S_e/\sqrt{S_{xx}}}$	$2P\left(T\geq\frac{\left\|\hat{\beta}-\beta_0\right\|}{s_e/\sqrt{S_{xx}}}\right)$ where $T\sim t\left(n-2\right)$
$H_0:\alpha=\alpha_0$	$\frac{\left\|\tilde{\alpha}-\alpha_0\right\|}{S_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T\geq\frac{\left\|\hat{\alpha}-\alpha_0\right\|}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$ where $T\sim t\left(n-2\right)$
$H_0:\sigma=\sigma_0$	$\frac{(n-2)S_e^2}{\sigma_0^2}$	$\min\left(2P\left(W\leq\frac{(n-2)s_e^2}{\sigma_0^2}\right),2P\left(W\geq\frac{(n-2)s_e^2}{\sigma_0^2}\right)\right)$ $W\sim\chi^2\left(n-2\right)$

Figure 6.4 shows a the scatterplot of the data together with the fitted line, $y=\hat{\alpha}+\hat{\beta}x=34.05413+0.6352523x$. The fitted line passes through the points but we notice that there is a quite a bit of variability about the fitted line.

The $p-value$ for testing $H_0:\beta=0$ is

$$\begin{aligned} & 2P\left(T\geq\frac{\left|\hat{\beta}-0\right|}{s_e/\sqrt{S_{xx}}}\right) \\ = \; & 2P\left(T\geq\frac{\left|0.6352523-0\right|}{\left(8.263079\right)/\sqrt{10813.75}}\right) \\ = \; & 2P\left(T\geq 7.994522\right)\approx 0 \end{aligned}$$

where $T\sim t\left(63\right)$. Therefore there is very strong evidence against the hypothesis $H_0:\beta=0$ or the hypothesis of no relationship between exam mark and midterm mark based on the data which is consistence with what we see in Figure 6.4.

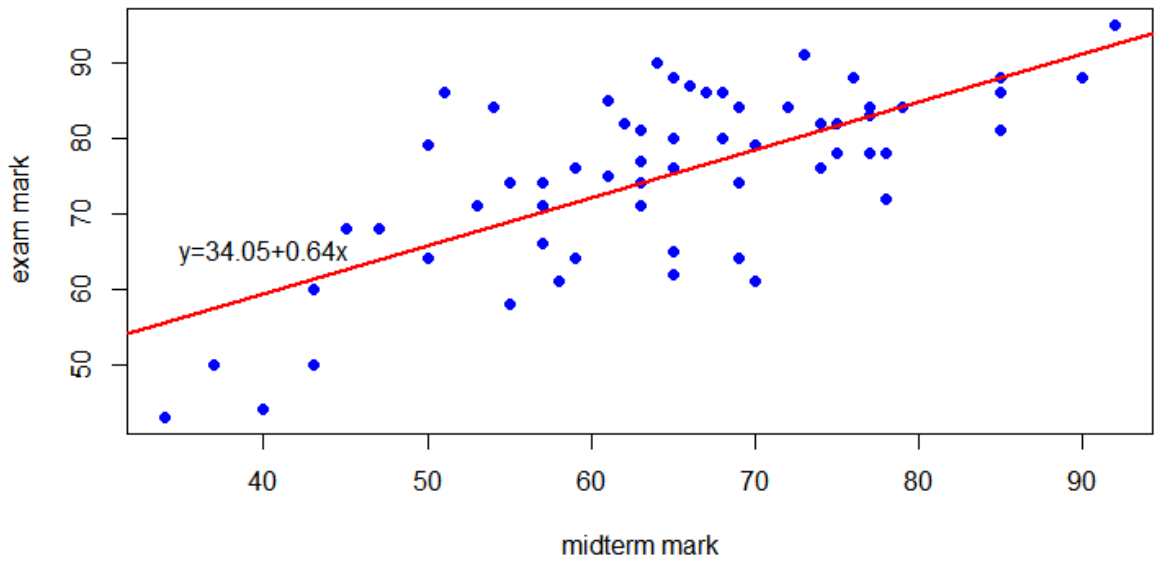


Figure 6.4: **Scatterplot and fitted line for exam mark versus midterm mark**

The $p-value$ for testing $H_0 : \alpha = 0$ is

$$\begin{aligned} & 2P\left(T \geq \frac{|\hat{\alpha}-0|}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right) \\ = \; & 2P\left(T \geq \frac{|34.05413-0|}{(8.263079)\sqrt{\frac{1}{65}+\frac{(65.06154)^2}{10813.75}}}\right) \\ = \; & 2P\left(T \geq 6.461314\right) \\ \approx \; & 0 \end{aligned}$$

where $T \sim t\,(63)$. Therefore there is very strong evidence against the hypothesis $H_0 : \alpha = 0$. Note that this $\alpha = 0$ corresponds to a midterm mark of $x = 0$ which is well outside the range of observed midterm marks. In other words we are assuming the linear relationship hold outside the range of observed x values which might not be valid. In this example the hypothesis $H_0 : \alpha = 0$ is not of particular interest.

These results can also be more easily be obtained by using the command `summary(lm(y~x))` in R. The table below gives us the parts of the output which are of interest to us for this course.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.05413	5.27046	6.461	$1.72e-08$***
x	0.63525	0.07946	7.995	$3.65e-11$***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 8.263 on 63 degrees of freedom

The values which are given in this table are:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat{\alpha}$	$s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}$	$\frac{\hat{\alpha}-\alpha_0}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}$	$2P\left(T \geq \frac{\lvert\hat{\alpha}-\alpha_0\rvert}{s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}}}\right)$
x	$\hat{\beta}$	$s_e/\sqrt{S_{xx}}$	$\frac{\hat{\beta}-\beta_0}{s_e/\sqrt{S_{xx}}}$	$2P\left(T \geq \frac{\lvert\hat{\beta}-\beta_0\rvert}{s_e/\sqrt{S_{xx}}}\right)$

where $T \sim t(n-2)$. The *Residual standard error* is equal to s_e the estimate of σ. The entry 3.65e-11 *** in the row labeled x in the table indicates that the $p-value$ for testing $H_0 : \beta = 0$ is equal to 3.65×10^{-11} which is less than 0.001.

Since $P(T \leq 1.998341) = 0.975$ where $T \sim t(63)$, a 95% confidence interval for β is

$$\begin{aligned} & \hat{\beta} \pm 1.998341 s_e/\sqrt{S_{xx}} \\ = \; & 0.6352523 \pm 1.998341\,(8.263079)/\sqrt{10813.75} \\ = \; & [0.4764622, 0.7940424] \end{aligned}$$

This interval does not contain any values of β close to zero which is consistent with the fact that the $p-value$ for testing $H_0 : \beta = 0$ was approximately zero.

A 95% confidence interval for α is

$$\begin{aligned} & \hat{\alpha} \pm 1.998341 s_e\sqrt{\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}} \\ = \; & 34.05413 \pm 1.998341\,(8.263079)\sqrt{\frac{1}{65}+\frac{(65.06154)^2}{10813.75}} \\ = \; & [23.52194, 44.58632] \end{aligned}$$

This interval does not contain any values of α close to zero which is consistent with the fact that the $p-value$ for testing $H_0 : \alpha = 0$ was approximately zero.

A 95% confidence interval for $\mu(50) = \alpha + \beta(50) =$ the mean exam grade for students with a midterm mark of $x = 50$ is

$$\begin{aligned} & \hat{\alpha} + \hat{\beta}(50) \pm 1.998341 s_e \sqrt{\frac{1}{n} + \frac{(\bar{x} - 50)^2}{S_{xx}}} \\ = \; & 65.81674 \pm 1.998341 (8.263079) \sqrt{\frac{1}{65} + \frac{(65.06154 - 50)^2}{10813.75}} \\ = \; & [62.66799, 68.96549] \end{aligned}$$

Note that this is a confidence interval for the mean or average exam mark for students who obtain a midterm mark of $x = 50$. If we want to give an interval of values for an individual student who obtained a midterm mark of $x = 50$ then we should use a prediction interval. A 95% prediction interval is

$$\begin{aligned} & \hat{\alpha} + \hat{\beta}(50) \pm 1.998341 s_e \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - 50)^2}{S_{xx}}} \\ = \; & 65.81674 \pm 1.998341 (8.263079) \sqrt{1 + \frac{1}{65} + \frac{(65.06154 - 50)^2}{10813.75}} \\ = \; & [49.00675, 82.62673] \end{aligned}$$

As we have indicated before, this interval is much wider than the confidence for the mean exam mark. Based on this interval, what advice would you give to a student who obtained a mark of 50 on the midterm?

These intervals can also be easily obtained using R. For example, the R commands

```
confint(lm(y~x),level=0.95)
predict(lm(y~x),data.frame("x"=50),interval="confidence",lev=0.95)
predict(lm(y~x),data.frame("x"=50),interval="prediction",lev=0.95)
```

give the output

	2.5 %	97.5 %
(Intercept)	23.5219441	44.5863083
x	0.4764623	0.7940423

	fit	lwr	upr
1	65.81674	62.66799	68.96549

	fit	lwr	upr
1	65.81674	49.00676	82.62672

The values in the these tables can be compared to the intervals obtained above.

Example 6.1.4 Revisited Breaking strength versus diameter of steel bolt

Recall the data given in Example 6.1.4, where Y represented the breaking strength of a randomly selected steel bolt and x was the bolt's diameter. A scatterplot of points (x_i, y_i) for 30 bolts suggested a nonlinear relationship between Y and x. A bolt's strength might be expected to be proportional to its cross-sectional area, which is proportional to x^2. Figure 6.5 shows a plot of points (x_i^2, y_i) which looks quite linear. Because of this let us assign a new variate name to x^2, say $x_1 = x^2$. We then fit a linear model

$$Y_i \sim G(\alpha + \beta x_{1i}, \sigma) \quad \text{where} \quad x_{1i} = x_i^2$$

to the data.

For these data we obtain $\hat{\alpha} = 1.6668$, $\hat{\beta} = 2.8378$, $s_e = 0.002656$ and $S_{x_1x_1} = 0.2244$. The fitted regression line is shown on the scatterplot in Figure 6.5. The model appears to fit the data well.

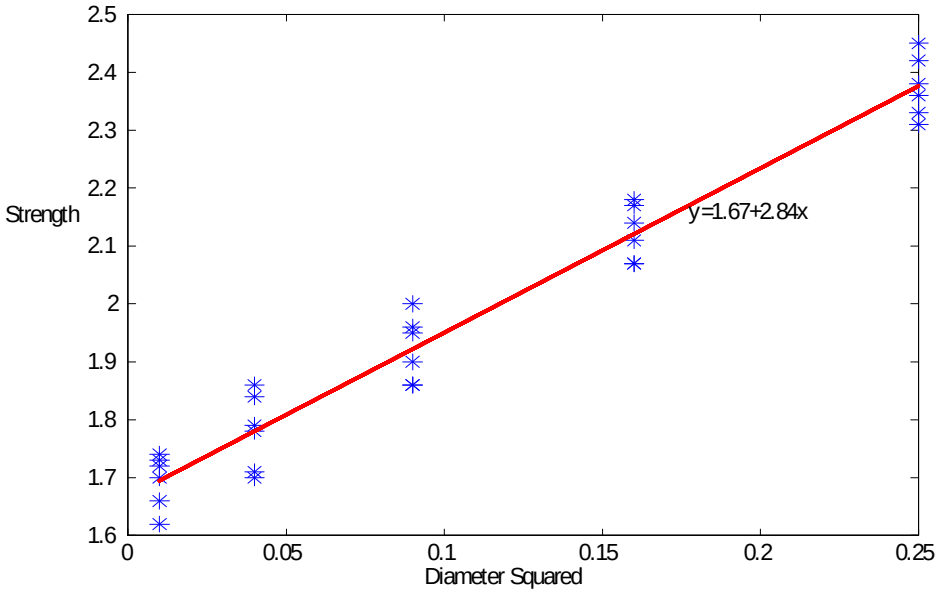


Figure 6.5: **Scatterplot plus fitted line for strength versus diameter squared**

The parameter β represents the increase in average strength $\mu(x_1)$ from increasing $x_1 = x^2$ by one unit. Using (6.7) and the fact that $P(T \leq 2.0484) = 0.975$ for $T \sim t\,(28)$, a 95% confidence interval for β is given by

$$\begin{aligned} & \hat{\beta} \pm 2.0484 s_e / \sqrt{S_{xx}} = 2.8378 \pm 0.2228 \\ = \quad & [2.6149,\ 3.0606] \end{aligned}$$

6.3 Checking the Model

There are two main components in Gaussian linear response models:

- The assumption that $E(Y_i) = \mu(x_i)$ is a linear combination of observed covariates with unknown coefficients.
- The assumption that the random variable Y_i (given any covariates x_i) has a Gaussian distribution with constant standard deviation σ.

Models should always be checked. For Gaussian linear response models we use graphical summaries to check the model asssumptions.

Scatterplot with Fitted Line

If there is only one x covariate, a scatterplot of the data with the fitted line superimposed can be used. Such a plot is checking whether the response variate can be modeled by a random variable whose mean is a linear function of the explanatory variate and whose standard deviation is constant over the range of values of the explanatory variate. If these assumptions hold then we would expect to see the observed points lying reasonably along the fitted line with the variability about the fitted line being reasonably constant over the range of values of the explanatory variate. The plots in Figures 6.4 and 6.5 exhibit such a pattern. If there are two or more covariates in the model, residual plots, which are described below, are very useful for checking the model assumptions.

Residual Plots

Consider the simple linear regression model for which $Y_i \sim G(\mu_i, \sigma)$ where $\mu_i = \alpha + \beta x_i$ and $R_i = Y_i - \mu_i \sim G(0, \sigma)$, $i = 1, 2, \ldots, n$ independently. Residuals are defined as the difference between the observed response y_i and the fitted response $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$, that is, $\hat{r}_i = y_i - \hat{\mu}_i$, $i = 1, 2, \ldots, n$. The idea behind the $\hat{r}_i$'s is that they can be thought of as "observed" R_i's. This is not exactly true since we are using $\hat{\mu}_i$ instead of μ_i, but if the model is correct, then the $\hat{r}_i$'s should behave roughly like a random sample from a $G(0, \sigma)$ distribution. Another reason why the $\hat{r}_i$'s only behave roughly like a random sample from the $G(0, \sigma)$ distribution is because $\sum_{i=1}^{n} \hat{r}_i = 0$. To see this, recall that the maximum likelihood estimate of α is $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ which implies

$$0 = \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i \right) = \frac{1}{n} \sum_{i=1}^{n} \hat{r}_i$$

or

$$\sum_{i=1}^{n} \hat{r}_i = 0$$

A random sample from a $G(0, \sigma)$ distribution does not satisfy such a restriction.

Often we prefer to use standardized residuals

$$\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} = \frac{y_i - \hat{\alpha} - \hat{\beta}x_i}{s_e} \quad \text{for } i = 1, 2, \ldots, n$$

Since the $\hat{r}_i$'s behave roughly like a random sample from the $G(0, \sigma)$ distribution, the $\hat{r}_i^*$'s should behave roughly like a random sample from the $G(0, 1)$ distribution. Since $P(-3 \leq Z \leq 3) = 0.9973$ where $Z \sim G(0, 1)$, then roughly 99.73% of the observations should lie in the interval $[-3, 3]$.

Here are three *residual plots* which can be used to check the model assumptions.

(1) Plot points $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, n$.

(2) Plot points $(\hat{\mu}_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, n$.

(3) Plot a Gaussian qqplot of the residuals $\hat{r}_i^*$.

The residual plots are checking whether the response variate can be modeled by a Gaussian random variable whose mean is a linear function of the explanatory variate and whose standard deviation is constant over the range of values of the explanatory variate. If the model is satisfactory then the points in plots (1) and (2) should lie roughly within a horizontal band of constant width between -3 and 3. Approximately half the points should lie on either side of the line $\hat{r}_i^* = 0$ should. Figure 6.6 exhibits the type of pattern we expect to see.

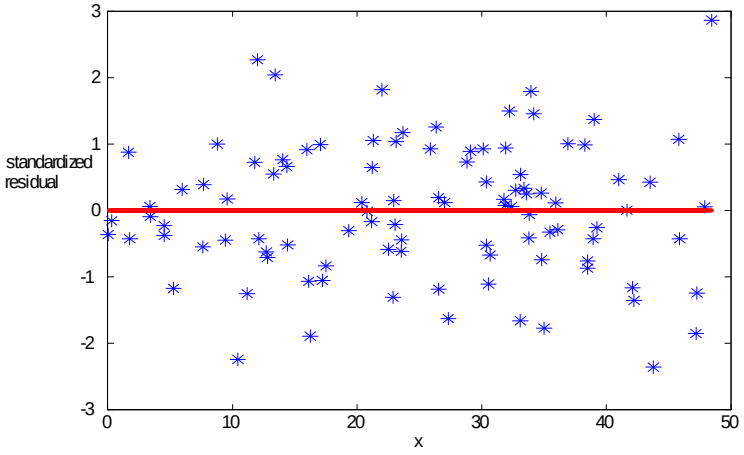


Figure 6.6: **Residual plot for example in which model assumptions hold**

If the model is satisfactory then the points in the qqplot (3) should lie roughly along a straight line with more variability in the points at both ends of the line.

Reading these plots requires some practice. That is, you should try generating many graphs for which the asssumptions hold and for different sample sizes so you can see the

type of variability that occurs even when the assumptions are true. In particular, you should try not to read too much into plots if the plots are based on a small number of points.

Systematic departures from the "expected" pattern suggest the model assumptions do not hold. For example, in Figure 6.7 the points exhibit a U-shaped pattern which suggests that the mean function $\mu_i = \mu(x_i)$ is not correctly specified. A quadratic form for the mean such as $\mu(x_i) = \alpha + \beta x_i + \gamma x_i^2$ might provide a better fit to these data.

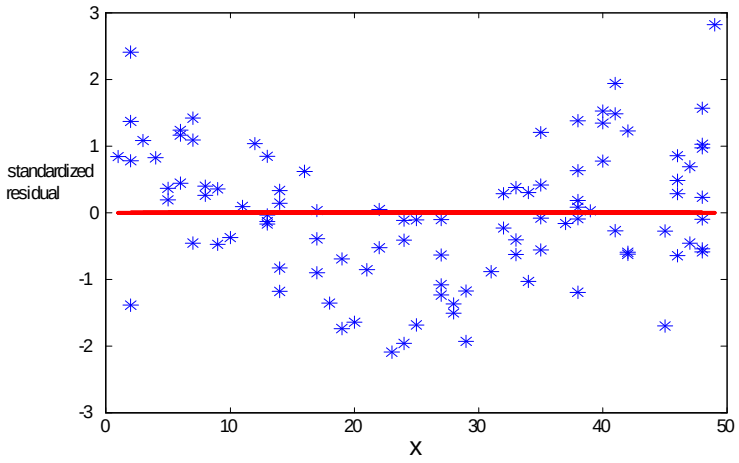


Figure 6.7: **Example of residual plot which indicates that assumption $E(Y_i) = \alpha + \beta x_i$ does not hold**

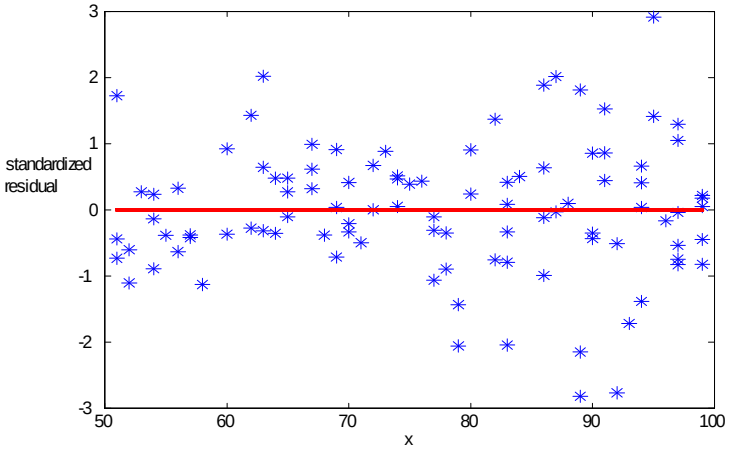


Figure 6.8: **Example of residual plot which indicates that assumption $Var(Y_i) = \sigma^2$ does not hold**

In Figure 6.8, the spread of the points about the line $\hat{r}_i^* = 0$ increases as x increases.

Such a pattern indicates that the standard deviation is not constant over the range of values of the explanatory variate. Sometimes transforming the response variate can solve the problem of non-constant variance also called heteroscedasticity. Transformations such as $\log y$ and $\sqrt{y}$ are frequently used.

Example 6.1.3 Revisited Exam mark versus midterm mark

Figure 6.9 shows a standardized residual plot for the midterm/exam data and gives a qqplot of the standardized residuals. No deviations from the expected pattern are observed in either plot which agrees with the plot in Figure 6.4. Based on all three plots we would conclude that the simple linear regression model is a reasonable model for these data.

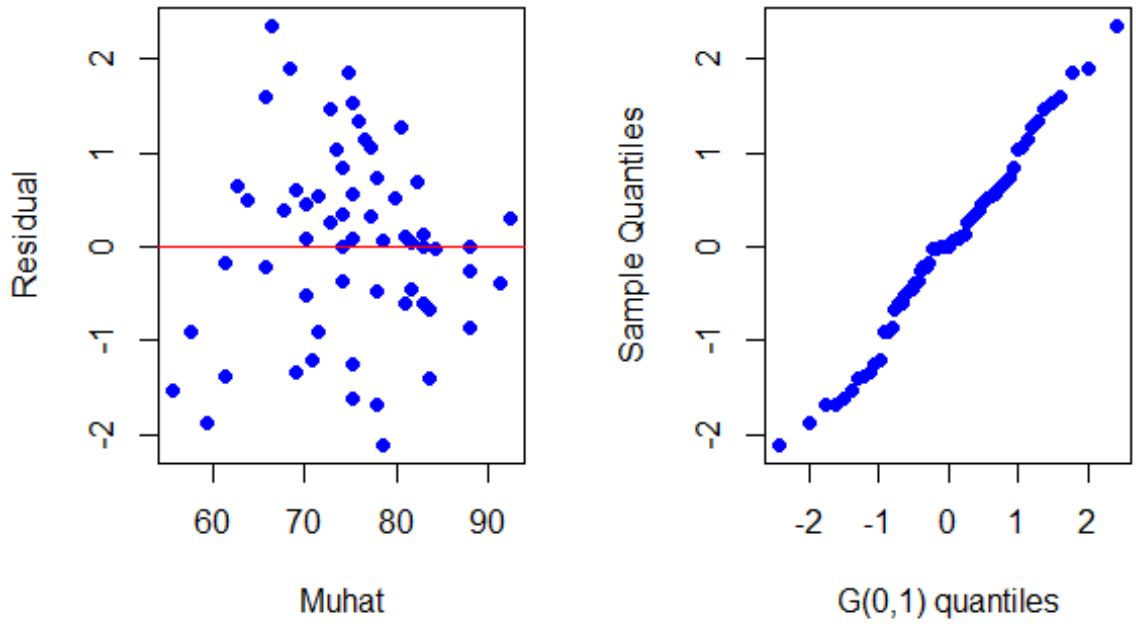


Figure 6.9: **Residual plots for exam mark versus midterm mark example**

6.4 Comparison of Two Population Means

Two Gaussian Populations with Common Variance

Suppose $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from a $G(\mu_1, \sigma)$ distribution and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from a $G(\mu_2, \sigma)$ distribution. Notice that we have assumed that both populations have the same variance σ^2. We use double subscripts for the Y's here, the first index to indicate the population from which the sample was drawn, the second to indicate which draw from that population. We could easily conform with the notation of (6.1) by stacking these two sets of observations in a vector of $n = n_1 + n_2$ observations:

$$(Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, Y_{22}, \ldots, Y_{2n_2})^T$$

and obtain the conclusions below as a special case of the linear model. Below we derive the estimates from the likelihood directly.

The likelihood function for μ_1, μ_2, σ is

$$L(\mu_1, \mu_2, \sigma) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(y_{ji} - \mu_j\right)^2\right] \quad \text{for } \mu_1 \in \Re,\ \mu_2 \in \Re,\ \sigma > 0$$

Maximization of the likelihood function gives the maximum likelihood estimates

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = \bar{y}_1 \\
\hat{\mu}_2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = \bar{y}_2 \\
\text{and } \hat{\sigma}^2 &= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2\right]
\end{aligned}$$

An estimate of the variance σ^2 called the *pooled estimate of variance* is

$$\begin{aligned}
s_p^2 &= \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2\right] \\
&= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\
&= \frac{n_1 + n_2}{n_1 + n_2 - 2}\hat{\sigma}^2
\end{aligned}$$

where

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2$$

are the sample variances obtained from the individual samples. The estimate s_p^2 can be written as

$$s_p^2 = \frac{w_1 s_1^2 + w_2 s_2^2}{w_1 + w_2}$$

to show that s_p^2 is a *weighted average* of the sample variances s_j^2 with weights equal to $w_j = n_j - 1$. With these weights the sample variance which has a larger sample size is weighted more. Why does this make sense?

We will use the estimate s_p^2 for σ^2 rather than $\hat{\sigma}^2$ since

$$E\left(S_p^2\right) = E\left[\frac{\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \overline{Y}_2)^2}{n_1 + n_2 - 2}\right] = \sigma^2$$

Confidence intervals for $\mu_1 - \mu_2$

To determine whether the two populations differ and by how much we will need to generate confidence intervals for the difference $\mu_1 - \mu_2$. First note that the maximum likelihood estimator of this difference is $\overline{Y}_1 - \overline{Y}_2$ which has expected value

$$E(\overline{Y}_1 - \overline{Y}_2) = \mu_1 - \mu_2$$

and variance

$$Var(\overline{Y}_1 - \overline{Y}_2) = Var(\overline{Y}_1) + Var(\overline{Y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

It naturally follows that an estimator of $Var(\overline{Y}_1 - \overline{Y}_2)$ from the pooled data is

$$S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and that this has $n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$ degrees of freedom. This provides at least an intuitive justification for the following:

Theorem 41 *If $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from a $G(\mu_1, \sigma)$ distribution and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from a $G(\mu_2, \sigma)$ distribution then*

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\left(n_1 + n_2 - 2\right)$$

and

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_j)^2 \sim \chi^2\left(n_1 + n_2 - 2\right)$$

Confidence intervals or tests of hypothesis for $\mu_1 - \mu_2$ and σ can be obtained using these pivotal quantities.

In particular a $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \bar{y}_2 \pm a s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.13}$$

where $P(T \leq a) = (1+p)/2$ and $T \sim t(n_1 + n_2 - 2)$.

To test $H_0 : \mu_1 - \mu_2 = 0$ we use the test statistic

$$D = \frac{\left|\overline{Y}_1 - \overline{Y}_2 - 0\right|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\left|\overline{Y}_1 - \overline{Y}_2\right|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.14}$$

with

$$p - value = P\left(|T| \geq \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) = 2\left[1 - P\left(T \leq \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)\right]$$

where $T \sim t(n_1 + n_2 - 2)$.

A $100p\%$ confidence interval for σ is

$$\left[\sqrt{\frac{(n_1 + n_2 - 2)\, s_p^2}{b}}, \sqrt{\frac{(n_1 + n_2 - 2)\, s_p^2}{a}}\right]$$

where

$$P(U \leq a) = \frac{1-p}{2}, \quad P(U \leq b) = \frac{1+p}{2}, \quad \text{and} \quad U \sim \chi^2(n_1 + n_2 - 2)$$

Two Gaussian Populations with Unequal Variances

The procedures derived above assume that the two Gaussian distributions have the same standard deviations. Sometimes this is not a reasonable assumption (it can be tested using a likelihood ratio test) and we must assume that $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ is a random sample from a $G(\mu_1, \sigma_1)$ distribution and independently $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ is a random sample from a $G(\mu_2, \sigma_2)$ but $\sigma_1 \neq \sigma_2$. If σ_1 and σ_2 are known then we could use the pivotal quantity

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim G(0,1) \tag{6.15}$$

A $100p\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\bar{y}_1 - \overline{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $P(Z \leq a) = \frac{1+p}{2}$ and $Z \sim G(0,1)$. To test $H_0 : \mu_1 - \mu_2 = 0$ we use the test statistic

$$D = \frac{\left|\overline{Y}_1 - \overline{Y}_2 - 0\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\left|\overline{Y}_1 - \overline{Y}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

with

$$p - value = P\left(|Z| \geq \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) = 2\left[1 - P\left(Z \leq \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)\right]$$

where $Z \sim G(0,1)$.

Table 6.3
Confidence Intervals for
Two Sample Gaussian Model

Model	Parameter	Pivotal Quantity	$100p\%$ Confidence Interval
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ σ_1, σ_2 known	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\sim G(0,1)$	$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 = \sigma_2 = \sigma$ σ unknown	$\mu_1 - \mu_2$	$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\sim t(n_1 + n_2 - 2)$	$\bar{y}_1 - \bar{y}_2 \pm b s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$G(\mu_1, \sigma)$ $G(\mu_2, \sigma)$ μ_1, μ_2 unknown	σ^2	$\frac{(n_1+n_2-2)S_p^2}{\sigma^2}$ $\sim \chi^2(n_1 + n_2 - 2)$	$\left[\frac{(n_1+n_2-2)s_p^2}{d}, \frac{(n_1+n_2-2)s_p^2}{c}\right]$
$G(\mu_1, \sigma_1)$ $G(\mu_2, \sigma_2)$ $\sigma_1 \neq \sigma_2$ σ_1, σ_2 unknown	$\mu_1 - \mu_2$	asymptotic Gaussian pivotal quantity $\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ for large n_1, n_2	approximate $100p\%$ confidence interval $\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Notes:
The value a is given by $P(Z \leq a) = \frac{1+p}{2}$ where $Z \sim G(0,1)$.
The value b is given by $P(T \leq b) = \frac{1+p}{2}$ where $T \sim t(n_1 + n_2 - 2)$.
The values c and d are given by $P(W \leq c) = \frac{1-p}{2} = P(W > d)$ where $W \sim \chi^2(n_1 + n_2 - 2)$.

Table 6.4
Hypothesis Tests for
Two Sample Gaussian Model

Model	Hypothesis	Test Statistic	$p-value$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ σ_1, σ_2 known	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ σ unknown	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$2P\left(T\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$ $T\sim t(n_1+n_2-2)$
$G(\mu_1,\sigma)$ $G(\mu_2,\sigma)$ μ_1, μ_2 unknown	$H_0:\sigma=\sigma_0$	$\frac{(n_1+n_2-2)S_p^2}{\sigma_0^2}$	$\min(2P\left(W\leq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right),$ $2P\left(W\geq\frac{(n_1+n_2-2)s_p^2}{\sigma_0^2}\right))$ $W\sim\chi^2(n_1+n_2-2)$
$G(\mu_1,\sigma_1)$ $G(\mu_2,\sigma_2)$ $\sigma_1\neq\sigma_2$ σ_1, σ_2 unknown	$H_0:\mu_1=\mu_2$	$\frac{\lvert\overline{Y}_1-\overline{Y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$	approximate $p-value$ $2P\left(Z\geq\frac{\lvert\bar{y}_1-\bar{y}_2-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right)$ $Z\sim G(0,1)$

In the case in which σ_1 and σ_2 are unknown then there is no exact pivotal quantity which can be used. However if we replace the quantities σ_1^2 and σ_2^2 in the pivotal quantity (6.15) by their respective estimators S_1^2 and S_2^2 to obtain the the random variable

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{6.16}$$

then it can be shown that this asymptotic pivotal quantity has approximately a $G(0,1)$ distribution if n_1 and n_2 are both large. An approximate $100p\%$ confidence interval for $\mu_1 - \mu_2$ based on this pivotal quantity is

$$\bar{y}_1 - \bar{y}_2 \pm a\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.17}$$

where $P(Z \leq a) = \frac{1+p}{2}$ and $Z \sim G(0,1)$.

These results are summarized in Tables 6.3 and 6.4.

Example 6.4.1 Durability of paint

In an experiment to assess the durability of two types of white paint used on asphalt highways, 12 lines (each 4 inches wide) of each paint were laid across a heavily traveled section of highway, in random order. After a period of time, reflectometer readings were taken for each line of paint; the higher the readings the greater the reflectivity and the visibility of the paint. The measurements of reflectivity were as follows:

Paint A	12.5	11.7	9.9	9.6	10.3	9.6	9.4	11.3	8.7	11.5	10.6	9.7
Paint B	9.4	11.6	9.7	10.4	6.9	7.3	8.4	7.2	7.0	8.2	12.7	9.2

The objectives of the experiment were to test whether the average reflectivities for paints A and B are the same, and if there is evidence of a difference, to obtain a confidence interval for their difference. (In many problems where two attributes are to be compared we start by testing the hypothesis that they are equal, even if we feel there may be a difference. If there is no statistical evidence of a difference then we stop there.)

To do this it is assumed that, to a close approximation, the reflectivity measurements Y_{1i}, $i = 1, 2, \ldots, 12$ for paint A are independent $G(\mu_1, \sigma_1)$ random variables, and independently the measurements Y_{2i}, $i = 1, 2, \ldots, 12$ for paint B are independent $G(\mu_2, \sigma_2)$ random variables. We can test $H : \mu_1 - \mu_2 = 0$ and get confidence intervals for $\mu_1 - \mu_2$ by using the pivotal quantity

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{12} + \frac{1}{12}}} \sim t(22) \tag{6.18}$$

Using this pivotal quantity means that we have assumed that the two population variances are equal, $\sigma_1 = \sigma_2 = \sigma$, and that we are using the estimator S_p for σ. If the observed sample variances differed by a great deal we would not make this assumption.

From these data we have

$$n_1 = 12 \quad \bar{y}_1 = 10.4 \quad \sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 = 14.08 \quad s_1 = 1.1314$$
$$n_2 = 12 \quad \bar{y}_2 = 9.0 \quad \sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2 = 38.64 \quad s_2 = 1.8742$$

$$s_p = \sqrt{\frac{1}{12+12-2}\left[\sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2\right]} = 1.5480$$

The observed value of the test statistic (6.14) is

$$d = \frac{|\bar{y}_1 - \bar{y}_2 - 0|}{s_p\sqrt{\frac{1}{12} + \frac{1}{12}}} = \frac{1.4}{1.5480\sqrt{\frac{1}{6}}} = 2.215$$

with

$$\begin{aligned} p - value &= P(|T| \geq 2.215) \\ &= 2\left[1 - P\left(T \leq 2.215\right)\right] \\ &= 0.038 \end{aligned}$$

where $T \sim t\,(22)$. Since $0.01 < p - value < 0.05$, there is evidence against $H_0 : \mu_1 = \mu_2$ based on the data.

Since $\bar{y}_1 > \bar{y}_2$, the indication is that paint A keeps its visibility better. Since $P(T \leq 2.074) = 0.975$ where $T \sim t\,(22)$ a 95% confidence interval for $\mu_1 - \mu_2$ based on (6.13) is

$$\begin{aligned} & 10.4 - 9.0 \pm 2.074\,(1.5480)\sqrt{\frac{1}{12} + \frac{1}{12}} \\ = \;& 1.4 \pm 1.3107 \\ = \;& [0.089,\ 2.711] \end{aligned}$$

This suggests that although the difference in reflectivity (and durability) of the paint is statistically significant, the size of the difference is not really large relative to the sizes of μ_1 and μ_2. This can be seen by noting that $\hat{\mu}_1 = \bar{y}_1 = 10.4$, and $\hat{\mu}_2 = \bar{y}_2 = 9.0$, whereas $\hat{\mu}_1 - \hat{\mu}_2 = 1.4$ so the relative difference is of the order of 10%.

Remark The R command `t.test(y1,y2,var.equal=T,conf.level=p)`, where y_1 and y_2 are the data vectors,.will carry out the test above and give a $100p\%$ confidence interval for $\mu_1 - \mu_2$.

Example 6.4.2 Scholastic achievement test scores

Tests that are designed to measure the achievement of students are often given in various subjects. Educators and parents often compare results for different schools or districts. We consider here the scores on a mathematics test given to Canadian students in the 5th grade. Summary statistics (sample sizes, means, and standard deviations) of the scores y for the students in two small school districts in Ontario are as follows:

$$\begin{array}{lllll} \text{District 1:} & n_1 = 278 & \bar{y}_1 = 60.2 & s_1 = 10.16 \\ \text{District 2:} & n_2 = 345 & \bar{y}_2 = 58.1 & s_2 = 9.02 \end{array}$$

If a likelihood ratio test of the hypothesis $\sigma_1 = \sigma_2$ is conducted there is strong evidence against the hypothesis based on the data so we cannot assume equal variances in this example.

The average score is somewhat higher in District 1, but is this difference statistically significant? We will give a confidence interval for the difference in average scores in a model representing this setting. This is done by thinking of the students in each district as a random sample from a conceptual large population of "similar" students writing "similar" tests. We assume that the scores in District 1 have a $G(\mu_1, \sigma_1)$ distribution and that the scores in District 2 have a $G(\mu_2, \sigma_2)$ distribution. We can then test the hypothesis $H_0 : \mu_1 = \mu_2$ or alternatively construct a confidence interval for the difference $\mu_1 - \mu_2$. (Achievement tests are usually designed so that the scores are approximately Gaussian, so this is a sensible procedure.)

Since $n_1 = 278$ and $n_2 = 345$ we use (6.17) to construct an approximate 95% confidence interval for $\mu_1 - \mu_2$. We obtain

$$\begin{aligned} & 60.2 - 58.1 \pm 1.96\sqrt{\frac{(10.16)^2}{278} + \frac{(9.02)^2}{345}} \\ = \; & 2.1 \pm (1.96)(0.779) \\ = \; & [0.57, 3.63] \end{aligned}$$

Since $\mu_1 - \mu_2 = 0$ is outside the approximate 95% confidence interval (can you show that it is also outside the approximate 99% confidence interval?) we can conclude there is fairly strong evidence against the hypothesis $H_0 : \mu_1 = \mu_2$ based on the data. The data suggest that $\mu_1 > \mu_2$. We should not rely only on a comparison of their means. It is a good idea to look carefully at the data. Side-by-side boxplots would be one way of doing this. The assumptions that the scores in District 1 have a $G(\mu_1, \sigma_1)$ distribution and that the scores in District 2 have a $G(\mu_2, \sigma_2)$ distribution, as always, should be checked using the usual numerical and graphical summaries.

The mean is a little higher for District 1 and because the sample sizes are so large, this gives a "statistically significant" difference in a test of $H_0 : \mu_1 = \mu_2$. Unfortunately, "significant" tests like this are often used to make claims that one group or class or school is "superior" to another. Recall that the validity of this method depends on the assumption

that the students in each district is a random sample from a conceptual large population of "similar" students writing "similar" tests. How reasonable is this assumption? How likely is it that marks in a class are independent of one another and no more alike than marks between two classes in two different years?

Comparison of Means Using Paired Data

Often experimental studies designed to compare means are conducted with *pairs of units*, where the responses within a pair are not independent. The following examples illustrate this.

Example 6.4.3 Heights of males versus females

In a study in England, the heights of 1401 (brother, sister) pairs of adults were determined. One objective of the study was to compare the heights of adult males and females; another was to examine the relationship between the heights of male and female siblings.

Let Y_{1i} and Y_{2i} be the heights of the male and female, respectively, in the i'th (brother, sister) pair, $i = 1, 2, \ldots, 1401$. Assuming that the pairs are sampled randomly from the population, we can use them to estimate

$$\mu_1 = E(Y_{1i}) \quad \text{and} \quad \mu_2 = E(Y_{2i})$$

and the difference $\mu_1 - \mu_2$. However, the heights of related persons are not independent. If we know that one sibling from a family is tall(small) then on average we would expect other siblings in the family to also be tall (small) so heights of siblings are correlated and therefore not independent. The method in the preceding section should not be used to estimate $\mu_1 - \mu_2$ since it would require **independent** random samples of males and females. In fact, the primary reason for collecting these data was to consider the joint distribution of Y_{1i}, Y_{2i} and to examine their relationship. A clear picture of the relationship could be obtained by plotting the observed points (y_{1i}, y_{2i}) in a scatterplot.

Example 6.4.4 Comparison of car fuels

In a study to compare standard gasoline with gas containing an additive designed to improve mileage (i.e. reduce fuel consumption), the following experiment was conducted. Fifty cars of a variety of makes and engine sizes were chosen. Each car was driven in a standard way on a test track for 1000 km, with the standard fuel (S) and also with the enhanced fuel (E). The order in which the S and E fuels was used was randomized for each car (you can think of a coin being tossed for each car, with fuel S being used first if a Head occurred) and the same driver was used for both fuels in a given car. Drivers were different across the 50 cars.

Suppose we let Y_{1i} and Y_{2i} be the amount of fuel consumed (in liters) for the i'th car with the S and E fuels, respectively. We want to estimate $E(Y_{1i} - Y_{2i})$. The fuel consumptions Y_{1i}, Y_{2i} for the i'th car are related, because factors such as size, weight and engine size

(and perhaps the driver) affect consumption. The assumption that the Y_{1i}'s are a random sample from a large population with mean μ_1, and independently the Y_{2i}'s are a random sample from a large population with mean μ_2 would not be appropriate in this example. The observations have been paired deliberately to eliminate some factors (like driver/ car size) which might otherwise affect the conclusion. Note that in this example it may not be of much interest to consider $E(Y_{1i})$ and $E(Y_{2i})$ separately, since there is only a single observation on each car type for each fuel.

There are two types of Gaussian models which can be used to model paired data. The first involves what is called a Bivariate Normal distribution for (Y_{1i}, Y_{2i}), and it could be used in the fuel consumption example. The Bivariate Normal distribution is a continuous bivariate model for which each component has a Normal distribution and the components may be dependent. We will not describe this model here (it is studied in third year courses), except to note one fundamental property: If (Y_{1i}, Y_{2i}) has a Bivariate Normal distribution then the difference between the two is also Normally distributed; where $\sigma^2 = Var(Y_{1i}) + Var(Y_{2i}) - 2Cov(Y_{1i}, Y_{2i})$. Thus, if we are interested in making inferences about $\mu_1 - \mu_2$ then we can do this by analyzing the *within-pair differences* $Y_i = Y_{1i} - Y_{2i}$ and using the model

$$Y_i = Y_{1i} - Y_{2i} \sim G(\mu_1 - \mu_2, \sigma) \quad i = 1, 2, \ldots, n \text{ independently}$$

or equivalently

$$Y_i \sim G(\mu, \sigma) \quad i = 1, 2, \ldots, n \text{ independently} \tag{6.19}$$

where $\mu = \mu_1 - \mu_2$. **The methods for a $G(\mu, \sigma)$ model discussed in Sections 4.7 and 5.2 can then be used to estimate and test hypotheses about the parameters μ and σ.**

The second Gaussian model used with paired data assumes

$$Y_{1i} \sim G\left(\mu_1 + \alpha_i, \sigma_1^2\right), \text{ and } Y_{2i} \sim G\left(\mu_2 + \alpha_i, \sigma_2^2\right) \text{ independently}$$

where the α_i's are unknown constants. The α_i's represent factors specific to the different pairs so that some pairs can have larger (smaller) expected values than others. This model also gives a Gaussian distribution like (6.19), since $Y_{1i} - Y_{2i}$ has a Gaussian distribution with

$$E(Y_{1i} - Y_{2i}) = \mu_1 - \mu_2 = \mu$$

(note that the α_i's cancel) and

$$Var(Y_{1i} - Y_{2i}) = \sigma_1^2 + \sigma_2^2 = \sigma^2$$

Such a model might be reasonable for Example 6.4.4, where α_i refers to the i'th car type.

Thus, whenever we encounter paired data in which the random variables Y_{1i} and Y_{2i} are adequately modeled by Gaussian distributions, we will make inferences about $\mu_1 - \mu_2$

by working with the model (6.19).

Example 6.4.3 Revisited Heights of males versus females

The data on 1401 (brother, sister) pairs gave differences $Y_i = Y_{1i} - Y_{2i}$, $i = 1, 2, \ldots, 1401$ for which the sample mean and variance were

$$\bar{y} = 4.895 \text{ inches}$$

and

$$s^2 = \frac{1}{1400} \sum_{i=1}^{1401} (y_i - \bar{y})^2 = 6.5480 \ \text{(inches)}^2$$

Assume $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, 1401$ independently. This assumption should be checked of course. Using the pivotal quantity

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(1400)$$

a 95% confidence interval for $\mu = E(Y_i)$ is given by

$$\begin{aligned} \bar{y} \pm 1.96 s/\sqrt{n} &= 4.895 \pm 1.96\sqrt{6.5480/1401} \\ &= 4.895 \pm 0.134 \\ &= [4.76,\ 5.03] \end{aligned}$$

Note that $t(1400)$ is indistinguishable from $G(0,1)$ so we use the value 1.96 from the $G(0,1)$ distribution.

Remark The method above assumes that the (brother, sister) pairs are a random sample from the population of families with a living adult brother and sister. The question arises as to whether $E(Y_i)$ also represents the difference in the average heights of all adult males and all adult females (call them μ_1' and μ_2') in the population. If $\mu_1' = \mu_1$ (that is, the average height of all adult males equals the average height of all adult males who also have an adult sister) and similarly $\mu_2' = \mu_2$, then $E(Y_i)$ does represent this difference.

Pairing and Experimental Design

In settings where the population can be arranged in pairs, the estimation of a difference in means, $\mu_1 - \mu_2$, can often be made more precise (shorter confidence intervals) by using pairing in the study. The condition for this is that the association or correlation between Y_{1i} and Y_{2i} be positive. In Examples 6.4.3 and 6.4.4 a positive correlation seems to be a reasonable assumption and the pairing in these studies is a good idea.

To see why the pairing is helpful in estimating the mean difference $\mu_1 - \mu_2$, suppose that $Y_{1i} \sim G\left(\mu_1, \sigma_1^2\right)$ and $Y_{2i} \sim G\left(\mu_2, \sigma_2^2\right)$, but that Y_{1i} and Y_{2i} are not necessarily independent $(i = 1, 2, \ldots, n)$. The estimator of $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2$$

and we have that $E(\overline{Y}_1 - \overline{Y}_2) = \mu_1 - \mu_2$ and

$$\begin{aligned} Var(\overline{Y}_1 - \overline{Y}_2) &= Var(\overline{Y}_1) + Var(\overline{Y}_2) - 2Cov(\overline{Y}_1, \overline{Y}_2) \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2\frac{\sigma_{12}}{n} \end{aligned}$$

where $\sigma_{12} = Cov(Y_{1i}, Y_{2i})$. If $\sigma_{12} > 0$, then $Var(\overline{Y}_1 - \overline{Y}_2)$ is **smaller** than when $\sigma_{12} = 0$ (that is, when Y_{1i} and Y_{2i} are independent). We would expect that the covariance between the heights of siblings in the same family to be positively correlated since they share parents. Therefore if we can collect a sample of pairs (Y_{1i}, Y_{2i}), this is better than two independent random samples (one of Y_{1i}'s and one of Y_{2i}'s) for estimating $\mu_1 - \mu_2$. Note on the other hand that if $\sigma_{12} < 0$, then pairing is a bad idea since it increases the value of $Var(\overline{Y}_1 - \overline{Y}_2)$.

The following example involves an experimental study with pairing.

Example 6.4.5 Fibre in diet and cholesterol level

In a study 20 subjects, volunteers from workers in a Boston hospital with ordinary cholesterol levels, were given a low-fibre diet for 6 weeks and a high-fibre diet for another 6 week period. The order in which the two diets were given was randomized for each subject (person), and there was a two-week gap between the two 6 week periods, in which no dietary fibre supplements were given. A primary objective of the study was to see if cholesterol levels are lower with the high-fibre diet.

Details of the study are given in the *New England Journal of Medicine*, volume 322 (January 18, 1990), pages 147-152. Here we will simply present the data from the study and estimate the effect of the amount of dietary fibre.

Subject	y_{1i} High F	y_{2i} Low F	$y_i = y_{1i} - y_{2i}$	Subject	y_{1i} High F	y_{2i} Low F	$y_i = y_{1i} - y_{2i}$
1	5.55	5.42	0.13	11	4.44	4.43	0.01
2	2.91	2.85	0.06	12	5.22	5.27	−0.05
3	4.77	4.25	0.52	13	4.22	3.61	0.61
4	5.63	5.43	0.20	14	4.29	4.65	−0.36
5	3.58	4.38	−0.80	15	4.03	4.33	−0.30
6	5.11	5.05	0.06	16	4.55	4.61	−0.06
7	4.29	4.44	−0.15	17	4.56	4.45	0.11
8	3.40	3.36	0.04	18	4.67	4.95	−0.28
9	4.18	4.38	−0.20	19	3.55	4.41	−0.86
10	5.41	4.55	0.86	20	4.44	4.38	0.06

Table 6.5: Cholesterol levels on two diets

Table 6.5 shows the cholesterol levels (in millimole per liter) for each subject, measured at the end of each 6 week period. We let the random variables Y_{1i}, Y_{2i} represent the cholesterol levels for subject i on the high fibre and low fibre diets, respectively. We assume that the differences can be modeled using

$$Y_i = Y_{1i} - Y_{2i} \sim G(\mu, \sigma) \quad \text{for } i = 1, 2, \ldots, 20$$

where the parameter μ represents the mean difference in cholesterol levels before and after a 6 week high fibre diet for people in the study population. (What is the study population?) The observed differences y_i, shown in Table 6.3, give $\bar{y} = -0.020$ and $s = 0.411$. Since $P(T \leq 2.093) = 0.975$ where $T \sim t(19)$, a 95% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned} & \bar{y} \pm 2.093 s/\sqrt{n} \\ = \; & -0.020 \pm 2.093\,(0.411)/\sqrt{20} \\ = \; & -0.020 \pm 0.192 \\ & \text{or} \quad [-0.212, 0.172] \end{aligned}$$

This confidence interval includes $\mu = 0$, and there is clearly no evidence that on average the high fibre diet gives a lower cholesterol level at least in the time frame represented in this study. A serious limitation of this study is the fact that there were only 20 observations.

Remark The results here can be obtained using the R function `t.test`.

Exercise Compute the $p - value$ for the test of hypothesis $H_0 : \mu = 0$, using the test statistic (5.1).

Final Remarks When you see data from a **comparative study** (that is, one whose objective is to compare two distributions, often through their means), you have to determine whether it involves paired data or not. Of course, a sample of Y_{1i}'s and Y_{2i}'s cannot be from a paired study unless there are equal numbers of each, but if there are equal numbers the study might be either "paired" or "unpaired". Note also that there is a subtle difference in the study populations in paired and unpaired studies. In the former it is pairs of individual units that form the population where as in the latter there are (conceptually at least) separate individual units for Y_1 and Y_2 measurements.

6.5 General Gaussian Response Models

We now consider general models of the form (6.1):

$$Y_i \sim G(\mu_i, \sigma) \text{ with } \mu(\mathbf{x}_i) = \sum_{j=1}^{k} \beta_j x_{ij} \quad \text{for } i = 1, 2, \ldots, n \text{ independently}$$

(Note: To facilitate the matrix proof below we have taken $\beta_0 = 0$ in (6.1). The estimator of β_0 can be obtained from the result below by letting $x_{i1} = 1$ for $i = 1, 2, \ldots, n$ and $\beta_0 = \beta_1$.) For convenience we define the $n \times k$ (where $n > k$) matrix X of covariate values as

$$X = (x_{ij}) \quad \text{for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, k$$

and the $n \times 1$ vector of responses $\mathbf{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)^T$. We assume that the values x_{ij} are non-random quantities which we observe. We now summarize some results about the maximum likelihood estimators of the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$ and σ.

Maximum Likelihood Estimators of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$ and of σ

Theorem 42 *The maximum likelihood estimators for $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$ and σ are:*

$$\tilde{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} \tag{6.20}$$

$$\textit{and} \quad \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \tilde{\mu}_i)^2 \quad \textit{where } \tilde{\mu}_i = \sum_{j=1}^{k} \tilde{\beta}_j x_{ij} \tag{6.21}$$

Proof. The likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right] \quad \text{where } \mu_i = \sum_{j=1}^{k} \beta_j x_{ij}$$

and the log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma) &= \log L(\boldsymbol{\beta}, \sigma) \\ &= -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 \end{aligned}$$

Note that if we take the derivative with respect to a particular β_j and set this derivative equal to 0, we obtain,

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \mu_i}{\partial \beta_j} = 0$$

or

$$\sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = 0$$

for each $j = 1, 2, \ldots, k$. In terms of the matrix X and the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ we can rewrite this system of equations more compactly as

$$\begin{aligned} X^T(\mathbf{y} - X\boldsymbol{\beta}) &= \mathbf{0} \\ \text{or} \quad X^T \mathbf{y} &= X^T X \boldsymbol{\beta}. \end{aligned}$$

Assuming that the $k \times k$ matrix $X^T X$ has an inverse we can solve these equations to obtain the maximum likelihood estimate of $\boldsymbol{\beta}$, in matrix notation as

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

with corresponding maximum likelihood estimator

$$\widetilde{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$$

In order to find the maximum likelihood estimator of σ, we take the derivative with respect to σ and set the derivative equal to zero and obtain

$$\frac{\partial l}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left[-n \log \sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\right] = 0$$

or

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \mu_i)^2 = 0$$

from which we obtain the maximum likelihood estimate of σ^2 as

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

where

$$\hat{\mu}_i = \sum_{j=1}^{k}\hat{\beta}_j x_{ij}$$

The corresponding maximum likelihood estimator σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_i)^2$$

where

$$\tilde{\mu}_i = \sum_{j=1}^{k}\tilde{\beta}_j x_{ij}$$

■

Recall that when we estimated the variance for a single sample from the Gaussian distribution we considered a minor adjustment to the denominator and with this in mind we also define the following estimator[15] of the variance σ^2:

$$S_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \tilde{\mu}_i)^2 = \frac{n}{n-k}\tilde{\sigma}^2$$

Note that for large n there will be small differences between the observed values of $\tilde{\sigma}^2$ and S_e^2.

[15] It is clear why we needed to assume $k < n$. Otherwise $n - k \leq 0$ and we have no "degrees of freedom" left for estimating the variance.

Theorem 43 *1. The estimators $\tilde{\beta}_j$ are all Normally distributed random variables with expected value β_j and with variance given by the j'th diagonal element of the matrix $\sigma^2(X^TX)^{-1}$, $j = 1, 2, \ldots, k$.*

2. The random variable

$$W = \frac{n\tilde{\sigma}^2}{\sigma^2} = \frac{(n-k)S_e^2}{\sigma^2} \tag{6.22}$$

has a Chi-squared distribution with $n - k$ degrees of freedom.

3. The random variable W is independent of the random vector $(\tilde{\beta}_1, \ldots, \tilde{\beta}_k)$.

Proof. The estimator $\tilde{\beta}_j$ can be written using (6.20) as a linear combination of the Normal random variables Y_i,

$$\tilde{\beta}_j = \sum_{i=1}^{n} b_{ji} Y_i$$

where the matrix $B = (b_{ji})_{k\times n} = (X^TX)^{-1}X^T$. Note that $BX = (X^TX)^{-1}\left(X^TX\right)$ equals the identity matrix I. Because $\tilde{\beta}_j$ is a linear combination of independent Normal random variables Y_i, it follows that $\tilde{\beta}_j$ is Normally distributed. Moreover

$$\begin{aligned} E(\tilde{\beta}_j) &= \sum_{i=1}^{n} b_{ji} E(Y_i) \\ &= \sum_{i=1}^{n} b_{ji}\mu_i \quad \text{where } \mu_i = \sum_{l=1}^{k} \beta_l x_{il} \\ &= \sum_{i=1}^{n} b_{ji}\mu_i \end{aligned}$$

Note that $\mu_i = \sum_{l=1}^{k} \beta_l x_{il}$ is the j'th component of the vector $X\boldsymbol{\beta}$ which implies that $E(\tilde{\beta}_j)$ is the j'th component of the vector $BXX\boldsymbol{\beta}$. But since BX is the identity matrix, this is the j'th component of the vector $\boldsymbol{\beta}$ or β_j. Thus $E(\tilde{\beta}_j) = \beta_j$ for all j. The calculation of the variance is similar.

$$\begin{aligned} Var(\tilde{\beta}_j) &= \sum_{i=1}^{n} b_{ji}^2 Var(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} b_{ji}^2 \end{aligned}$$

and an easy matrix calculation will show, since $BB^T = (X^TX)^{-1}$, that $\sum_{i=1}^{n} b_{ji}^2$ is the j'th diagonal element of the matrix $(X^TX)^{-1}$. We will not attempt to prove part (3) here, which is usually proved in a subsequent statistics course.

■

Remark The maximum likelihood estimate $\hat{\beta}$ is also called a **least squares estimate** of β in that it is obtained by taking the sum of squared vertical distances between the observations Y_i and the corresponding fitted values $\hat{\mu}_i$ and then adjusting the values of the

estimated β_j until this sum is minimized. Least squares is a method of estimation in linear models that predates the method of maximum likelihood. Problem 16 describes the method of least squares.

Remark[16] From Theorem 32 we can obtain confidence intervals and test hypotheses for the regression coefficients using the pivotal

$$\frac{\tilde{\beta}_j - \beta_j}{S_e\sqrt{c_j}} \sim t\,(n-k) \tag{6.23}$$

where c_j is the j'th diagonal element of the matrix $\left(X^T X\right)^{-1}$.

Confidence intervals for β_j

In a manner similar to the construction of confidence intervals for the parameter μ for observations from the $G(\mu,\sigma)$ distribution, we can use (6.23) to construct confidence intervals for the parameter β_j. For example for a 95% confidence interval, we begin by using the t distribution with $n-k$ degrees of freedom to find a constant a such that

$$P(-a < T < a) = 0.95 \quad \text{where} \quad T \sim t\,(n-k)$$

We then obtain the confidence interval by solving the inequality

$$-a \le \frac{\hat{\beta}_j - \beta_j}{s_e\sqrt{c_j}} \le a$$

to obtain

$$\hat{\beta}_j - a s_e\sqrt{c_j} \le \beta_j \le \hat{\beta}_j + a s_e\sqrt{c_j}$$

where

$$s_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \quad \text{and} \quad \hat{\mu}_i = \sum_{j=1}^{k}\hat{\beta}_j x_{ij}$$

Thus a 95% confidence interval for β_j is

$$\left[\hat{\beta}_j - a s\sqrt{c_j},\ \hat{\beta}_j + a s\sqrt{c_j}\right]$$

which takes the familiar form

$$\text{estimate} \ \pm a \times \text{estimated standard deviation of estimator.}$$

We now consider a special case of the Gaussian response models. We have already seen this case in Chapter 4, but it provides a simple example to validate the more general

[16] Recall: If $Z \sim G(0,1)$ and $W \sim \chi^2\,(m)$ then the random variable $T = Z/\sqrt{W/m} \sim t\,(m)$.
Let $Z = \frac{\tilde{\beta}_j - \beta_j}{\sigma\sqrt{c_j}}$, $W = \frac{(n-k)S^2}{\sigma^2}$ and $m = n-k$ to obtain this result.

formulae.

Single Gaussian distribution

Here, $Y_i \sim G(\mu, \sigma)$, $i = 1, 2, \ldots, n$, i.e. $\mu(\mathbf{x}_i) = \mu$ and $\mathbf{x}_i = x_{1i} = 1$, for all $i = 1, 2, \ldots, n$, $k = 1$ we use the parameter μ instead of $\boldsymbol{\beta} = (\beta_1)$. Notice that $X_{n\times 1} = (1, 1, \ldots, 1)^T$ in this case. This special case was also mentioned in Section 6.1. The pivotal quantity (6.23) becomes

$$\frac{\tilde{\beta}_1 - \beta_1}{S_e\sqrt{c_1}} = \frac{\tilde{\mu} - \mu}{S/\sqrt{n}}$$

since $(X^T X)^{-1} = 1/n$. This pivotal quantity has the t distribution with $n - k = n - 1$. You can also verify using (6.22) that

$$\frac{(n-1)S^2}{\sigma^2}$$

has a Chi-squared$(n-1)$ distribution.

6.6 Chapter 6 Problems

1. Prove the following identities which are used in this chapter.

$$\begin{aligned} \sum_{i=1}^{n} a_i &= 0, \quad \sum_{i=1}^{n} a_i x_i = 1, \quad \sum_{i=1}^{n} a_i^2 = \frac{1}{S_{xx}} \\ \text{where } a_i &= \frac{(x_i - \bar{x})}{S_{xx}} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n} b_i &= 1, \quad \sum_{i=1}^{n} b_i x_i = x \text{ and } \sum_{i=1}^{n} b_i^2 = \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}} \\ \text{where } b_i &= \frac{1}{n} + (x - \bar{x})\frac{(x_i - \bar{x})}{S_{xx}} \end{aligned}$$

2. Solve the three equations

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial l}{\partial \beta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)\, x_i = 0 \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = 0 \end{aligned}$$

simultaneously to obtain the maximum likelihood estimates

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}\left(S_{yy} - \tilde{\beta} S_{xy}\right)$$

3. Twenty-five female nurses working at a large hospital were selected at random and their age (x) and systolic blood pressure (y) were recorded. The data are:

x	y	x	y	x	y	x	y	x	y
46	136	37	115	58	139	48	134	59	142
36	132	45	129	50	156	35	120	54	135
62	138	39	127	41	132	42	137	57	150
26	115	28	134	31	115	27	120	60	159
53	143	32	133	51	143	34	128	38	127

$$\bar{x} = 43.20 \qquad \bar{y} = 133.56$$

$$S_{xx} = 2802.00 \quad S_{yy} = 3284.16 \quad S_{xy} = 2325.20$$

To analyze these data assume the simple linear regression model $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 25$ independently where the x_i's are known constants.

(a) Determine the maximum likelihood (least squares) estimates of α and β and an unbiased estimate of σ^2.

(b) Use the plots discussed in Section 6.2 to check the adequacy of the model.

(c) Construct a 95% confidence interval for β.

(d) Construct a 90% confidence interval for the mean systolic blood pressure of nurses aged $x = 35$.

(e) Construct a 99% prediction interval for the systolic blood pressure Y of a nurse aged $x = 50$.

4. This problem is designed to cover concepts in this chapter as well as previous chapters.

The data below are the STAT 230 final grades (x) and STAT 231 final grades (y) for 30 students chosen at random from the group of students enrolled in STAT 231 in Winter 2013. The data are available in the file *statgradedata.txt* posted on the course website.

x	y	x	y	x	y	x	y	x	y	x	y
76	76	60	60	87	76	65	69	83	83	94	94
77	79	81	85	71	50	71	43	88	88	83	83
57	54	86	82	63	75	66	60	52	52	51	37
75	64	96	88	77	72	90	96	75	75	77	90
74	64	79	72	96	84	50	50	99	99	77	67

$$\bar{x} = 76.7\dot{3} \qquad \bar{y} = 72.2\dot{3}$$
$$S_{xx} = 5135.8\dot{6} \quad S_{yy} = 7585.3\dot{6} \quad S_{xy} = 5106.8\dot{6}$$

(a) What type of study is this? Why?

(b) Define a possible Problem for this study. What type of Problem is it? Why?

(c) What is a unit in this study? Define a suitable target population for this study.

(d) What are the variates? What type are they?

(e) Why would it make sense to define $x =$ STAT 230 final grade as the explanatory variate and $y =$ STAT 231 final grade as the response variate?

(f) Define a suitable study population for this study. What is a possible source of study error?

(g) What is the sampling protocol?

(h) Why was it important for the students to be chosen at random from the group of students taking STAT 231? Why would it not be a good idea to chose the first 30 students in an alphabetized list of all students?

(i) How are the variates measured? What is a possible source of measurement error?

(j) Determine the sample correlation.

(k) Plot a scatterplot of the data. What do you notice?

(l) Fit the simple linear regression model to these data: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 30$ independently where the x_i's are known constants. What is the least squares estimate of α? What is the maximum likelihood estimate of β? What is an unbiased estimate of σ?

(m) Use the plots discussed in Section 6.3 to check the adequacy of the model.

(n) The parameter $\mu(x) = \alpha + \beta x$ corresponds to what attribute of interest in the study population? The parameter σ corresponds to what attribute of interest in the study population?

(o) The parameter β corresponds to what attribute of interest in the study population? Test the hypothesis that there is no relationship ($H_0 : \beta = 0$) between STAT 231 final grades and STAT 230 final grades.

(p) Test the hypothesis $H_0 : \beta = 1$. Why is this hypothesis of interest?

(q) Construct a 95% confidence interval for β. Is your confidence interval consistent with the $p - values$ determined in (o) and (p)? What is the interpretation of this interval?

(r) Construct a 95% confidence interval for the mean STAT 231 final grade for students with a STAT 230 final grade of $x = 75$.

(s) Construct a 95% prediction interval for STAT 231 final grade for a student with a STAT 230 final grade of $x = 75$. Compare this with the interval in (r). Why is this interval so wide? How could the width of the interval be reduced?

(t) Construct a 90% confidence interval for σ. What does the parameter σ represent in the context of this study?

5. Consider the data in Chapter 1 on the variates $x =$ "value of an actor" and $y =$ "amount grossed by a movie". The data are available in the file *actordata.txt* posted on the course website.

(a) Fit the simple linear regression model to these data: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 20$ independently where the x_i's are known constants.

(b) Use the plots discussed in Section 6.3 to check the adequacy of the model.

(c) What is the relationship between the maximum likelihood estimate of β and the sample correlation?

(d) Construct a 95% confidence interval for β. The parameter β corresponds to what attribute of interest in the study population?

(e) Test the hypothesis that there is no relationship between the "value of an actor" and the "amount grossed by a movie". Are there any limitations to your conclusion. (**Hint:** How were the data collected?)

(f) Construct a 95% confidence interval for the mean amount grossed by movies for actors whose value is $x = 50$. Construct a 95% confidence interval for the mean amount grossed by movies for actors whose value is $x = 100$. What assumption is being made in constructing the interval for $x = 100$?

6. Consider the price versus size of commercial building in Example 6.1.2. For these data

$$n = 30 \qquad \bar{x} = 0.9543 \qquad \bar{y} = 548.9700$$
$$S_{xx} = 22.9453 \quad S_{xy} = -3316.6771 \quad S_{yy} = 489,624.723$$

(a) Fit the simple linear regression model to these data: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 30$ independently where the x_i's are known constants.

(b) Plot the fitted line on the scatterplot of the data.

(c) How would you interpret the value of $\hat{\beta}$?

(d) These data were used to decide a fair assessment value for a large building of size $x = 4.47$ $(\times 10^5)m^2$. Determine a 95% confidence interval for the mean price of a building of this size.

(e) Determine a 95% prediction interval for a building of size $x = 4.47$ $(\times 10^5)m^2$.

(f) If you were an assessor deciding the fair assessment for a building of size $x = 4.47$ $(\times 10^5)m^2$, would you use the interval in (e) or (f)?

7. Consider the steel bolt experiment in Example 6.1.4.

(a) Construct a 95% confidence interval for the mean breaking strength of bolts of diameter $x = 0.35$, that is, $x_1 = (0.35)^2 = 0.1225$.

(b) Construct a 95% prediction interval for the breaking strength Y of a single bolt of diameter $x = 0.35$. Compare this with the interval in (a).

(c) Suppose that a bolt of diameter 0.35 is exposed to a large force V that could potentially break it. In structural reliability and safety calculations, V is treated as a random variable and if Y represents the breaking strength of the bolt (or some other part of a structure), then the probability of a "failure" of the bolt is $P(V > Y)$. Give a point estimate of this value if $V \sim G(1.60, 0.10)$, where V and Y are independent.

8. There are often both expensive (and highly accurate) and cheaper (and less accurate) ways of measuring concentrations of various substances (e.g. glucose in human blood, salt in a can of soup). The table below gives the actual concentration x (determined by an expensive but very accurate procedure) and the measured concentration y

obtained by a cheap procedure, for each of 20 units.

x	y	x	y	x	y	x	y
4.01	3.7	13.81	13.02	24.85	24.69	36.9	37.54
6.24	6.26	15.9	16	28.51	27.88	37.26	37.2
8.12	7.8	17.23	17.27	30.92	30.8	38.94	38.4
9.43	9.78	20.24	19.9	31.44	31.03	39.62	40.03
12.53	12.4	24.81	24.9	33.22	33.01	40.15	39.4

$$\bar{x} = 23.7065 \qquad \bar{y} = 23.5505$$
$$S_{xx} = 2818.946855 \quad S_{yy} = 2820.862295 \quad S_{xy} = 2818.556835$$

The data are available in the file *expensivevscheapdata.txt* posted on the course website. To analyze these data assume the regression model: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 20$ independently.

(a) Fit the model to these data. Use the plots discussed in Section 6.3 to check the adequacy of the model.

(b) Construct a 95% confidence intervals for the slope β and test the hypothesis $\beta = 1$. Construct 95% confidence intervals for the intercept α and test the hypothesis $\alpha = 0$. Why are these hypotheses of interest?

(c) Describe briefly how you would characterize the cheap measurement process's accuracy to a lay person.

(d) If the units to be measured have true concentrations in the range $0 - 40$, do you think that the cheap method tends to produce a value that is lower than the true concentration? Support your answer based on the data and the assumed model.

9. **Regression through the origin:** Consider the model $Y_i \sim G(\beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ independently.

(a) Assuming that σ is known, show that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

is the maximum likelihood estimate of β and also the least squares estimate of β.

(b) Show that

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Hint: Write $\tilde{\beta}$ in the form $\sum_{i=1}^{n} a_i Y_i$.

(c) Prove the identity

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta} x_i\right)^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_i y_i\right)^2 \left(\sum_{i=1}^{n} x_i^2\right)^{-1}$$

This identity can be used to calculate

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \hat{\beta} x_i\right)^2$$

which is an unbiased estimate of σ^2.

(d) Show how to use the pivotal quantity

$$\frac{\tilde{\beta} - \beta}{S_e / \sqrt{\sum_{i=1}^{n} x_i^2}} \sim t\,(n-1)$$

to construct a 95% confidence interval for β.

(e) Explain how to test the hypothesis $\beta = \beta_0$ using the test statistic

$$\frac{\left|\tilde{\beta} - \beta_0\right|}{S_e / \sqrt{\sum_{i=1}^{n} x_i^2}}$$

10. For the data in Problem 8

$$\sum_{i=1}^{20} x_i y_i = 13984.5554 \quad \sum_{i=1}^{20} x_i^2 = 14058.9097 \quad \sum_{i=1}^{20} y_i^2 = 13913.3833$$

Use the results from Problem 9 to do the following.

(a) Fit the model $Y_i \sim G(\beta x_i, \sigma)$, $i = 1, 2, \ldots, 20$ independently to these data.

(b) Let $\hat{\mu}_i = \hat{\beta} x_i$ and $\hat{r}_i^* = (y_i - \hat{\mu}_i)\,/s_e$.

Plot the following:

(i) a scatterplot of the data with the fitted line

(ii) the residual plot $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, 20$

(iii) the residual plot $(\hat{\mu}_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, 20$

(iv) a qqplot of the standardized residuals $\hat{r}_i^*$.

For each plot indicate what you would expect to see if the model is correct. Based on these plots, comment on how well the model fits the data.

(c) Construct a 95% confidence intervals for the slope β and test the hypothesis $\beta = 1$.

(d) Using the results of this analysis as well as the analysis in Problem 8 what would you conclude about using the model $Y_i \sim G(\alpha+\beta x_i, \sigma)$ versus the simpler model $Y_i \sim G(\beta x_i, \sigma)$ for these data?

11. The following data were recorded concerning the relationship between drinking (x = per capita wine consumption) and y = death rate from cirrhosis of the liver in $n = 46$ states of the U.S.A. (for simplicity the data have been rounded).

x	y	x	y	x	y	x	y	x	y	x	y
5	41	12	77	7	67	4	52	7	41	16	91
4	32	7	57	18	57	16	87	13	67	2	30
3	39	14	81	6	38	9	67	8	48	6	28
7	58	12	34	31	130	6	40	28	123	3	52
11	75	10	53	13	70	6	56	23	92	8	56
9	60	10	55	20	104	21	58	22	76	13	56
6	54	14	58	19	84	15	74	23	98		
3	48	9	63	10	66	17	98	7	34		

$$\bar{x} = 11.5870 \qquad \bar{y} = 63.5870$$
$$S_{xx} = 2155.1522 \quad S_{yy} = 24801.1521 \quad S_{xy} = 6175.1522$$

The data are available in the file *liverdata.txt* posted on the course website.

(a) Fit the simple linear regression model to these data: $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, 46$ independently where the x_i's are known constants.

(b) Use the plots discussed in Section 6.3 to check the adequacy of the model.

(c) Test the hypothesis that there is no relationship between wine consumption per capita and the death rate from cirrhosis of the liver.

(d) Construct a 95% confidence interval for β.

12. Skinfold body measurements are used to approximate the body density of individuals. The data on $n = 92$ men, aged 20-25, where x = skinfold measurement and Y = body density are given available in the file *SkinfoldData.txt* posted on the course website.

Note: The R function `lm`, with the command `lm(y~x)` gives the calculations for linear regression. The command `summary(lm(y~x))` gives a summary of the calculations.

```
# Import dataset skinfolddata.txt from course website using RStudio
relabel Skinfold variate as x
x<-SkinfoldData$Skinfold
# relabel Body Density variate as y
y<-SkinfoldData$BodyDensity
# run regression y = alpha+beta*x
RegModel<-lm(y~x)
# parameter estimates and p-value for test of no relationship
summary(RegModel)$coefficients
alphahat<-RegModel$coefficients[1]   # estimate of intercept
betahat<-RegModel$coefficients[2]    # estimate of slope
muhat<-RegModel$fitted.values        # fitted responses
r<- RegModel$residuals               # residuals
se<-summary(RegModel)$sigma          # estimate of sigma
# Scatterplot of data with fitted line
par(mfrow=c(2,2))
plot(x,y,xlab="Skinfold",ylab="Body Density",pch=19,las=1)
title(main="Scatterplot with Fitted Line")
abline(a=alphahat,b=betahat,col="red",lwd=2)
# Residual plots
rstar <- r/se # standardized residuals
plot(x,rstar,xlab="Skinfold",ylab="Standardized Residual",pch=19,las=1)
title(main="Standardized Residual vs Skinfold")
abline(0,0,col="red",lwd=1.5)
plot(muhat,rstar,xlab="Muhat",ylab="Standardized Residual",pch=19,las=1)
abline(0,0,col="red",lwd=1.5)
title(main="Standardized Residual vs Muhat")
qqnorm(rstar,main="")
title(main="Qqplot of Residuals")
# 95% Confidence interval for slope
confint(RegModel,level=0.95)
# 90% confidence interval for mean response at x=2
predict(RegModel,data.frame("x"=2),interval="confidence",level=0.90)
# 99% prediction interval for response at x=1.8
predict(RegModel,data.frame("x"=1.8),interval="prediction",level=0.99)
# 95% confidence interval for sigma
df<-length(y)-2
a<-qchisq(0.025,df)
b<-qchisq(0.975,df)
int<-c(se*sqrt(df/b),se*sqrt(df/a))
cat("95% confidence interval for sigma:  ",int)
```

(a) Run the given R code. What is the equation of the fitted line?

(b) What is the value of the test statistic and the $p-value$ for the hypothesis of no relationship? What would you conclude?

(c) Give an estimate of σ.

(d) What do the plots indicate about the fit of the model?

(e) What is a 95% confidence interval for β?

(f) What is a 90% confidence interval for the mean body density of males with a skinfold measurement of 2?

(g) What is a 99% prediction interval for the body density of a male with skinfold measurement of 1.8?

(h) What is a 95% confidence interval for σ?

(i) Do you think that the skinfold measurements provide a reasonable approximation to body density measurements?

13. The following data, collected by the British botanist Joseph Hooker in the Himalaya Mountains between 1848 and 1850, relate atmospheric pressure to the boiling point of water. Hooker wanted to estimate altitude above sea level from measurements of the boiling point of water. He knew that the altitude could be determined from the atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. His interest in the above modelling problem was motivated by the difficulty of transporting the fragile barometers of the 1840's. Measuring the boiling point would give travelers a quick way to estimate elevation, using both the known relationship between elevation and atmospheric pressure, and the model relating atmospheric pressure to the boiling point of water. The data in the table below are also available in the file *boilingpointdata.txt* on the course website.

(a) Let y = atmospheric pressure (in Hg) and x = boiling point of water (in °F). Fit a simple linear regression model to the data (x_i, y_i), $i = 1, 2, \ldots, 31$. Prepare a scatterplot of y versus x and draw on the fitted line. Plot the standardized residuals versus x. How well does the model fit these data?

(b) Let $z = \log y$. Fit a simple linear regression model to the data (x_i, z_i), $i = 1, 2, \ldots, 31$. Prepare a scatterplot of z versus x and draw on the fitted line. Plot the standardized residuals versus x. How well does the model fit these data?

(c) Based on the results in (a) and (b) which data are best fit by a linear model? Does this confirm the theory's model?

(d) Obtain a 95% confidence interval for the mean atmospheric pressure if the boiling

point of water is $195°F$.

Boiling Point of Water $°F$	Atmospheric Pressure Hg	Boiling Point of Water $°F$	Atmospheric Pressure Hg
210.8	29.211	189.5	18.869
210.2	28.559	188.8	18.356
208.4	27.972	188.5	18.507
202.5	24.697	185.7	17.267
200.6	23.726	186.0	17.221
200.1	23.369	185.6	17.062
199.5	23.030	184.1	16.959
197.0	21.892	184.6	16.881
196.4	21.928	184.1	16.817
196.3	21.654	183.2	16.385
195.6	21.605	182.4	16.235
193.4	20.480	181.9	16.106
193.6	20.212	181.9	15.928
191.4	19.758	181.0	15.919
191.1	19.490	180.6	15.376
190.6	19.386		

14. An educator believes that the new directed readings activities in the classroom will help elementary school students improve some aspects of their reading ability. She arranges for a Grade 3 class of 21 students to take part in the activities for an 8-week period. A control classroom of 23 Grade 3 students follows the same curriculum without the activities. At the end of the 8-week period, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data are:

Treatment Group:	24	43	58	71	43	49	61	44	67	49	53
	56	59	52	62	54	57	33	46	43	57	

Control Group:	42	43	55	26	62	37	33	41	19	54	20	85
	46	10	17	60	53	42	37	42	55	28	48	

The data are available in the file *treatmentvscontroldata.txt* posted on the course website.

Let y_{1j} = the DRP test score for the treatment group, $j = 1, 2, \ldots, 21$.

Let y_{2j} = the DRP test score for the control group, $j = 1, 2, \ldots, 23$. For these data

$$\bar{y}_1 = 51.4762 \quad \sum_{j=1}^{21} (y_{1j} - \bar{y}_1)^2 = 2423.2381$$

$$\bar{y}_2 = 41.5217 \quad \sum_{j=1}^{23} (y_{2j} - \bar{y}_2)^2 = 6469.7391$$

To analyze these data assume

$$Y_{1j} \sim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 21 \text{ independently}$$

for the treatment group and independently

$$Y_{2j} \sim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 23 \text{ independently}$$

for the control group where μ_1, μ_2 and σ are unknown parameters.

(a) The parameters μ_1, μ_2 and σ correspond to what attributes of interest in the study population?

(b) Plot a qqplot of the responses for the treatment group and a qqplot of the responses for the control group. How reasonable are the Gaussian assumptions stated in the assumed model?

(c) Calculate a 95% confidence interval for the difference in the means $\mu_1 - \mu_2$.

(d) Test the hypothesis of no difference between the means, that is, test the hypothesis $H_0 : \mu_1 = \mu_2$. What conclusion should the educator make based on these data? Be sure to indicate any limitations to these conclusions.

(e) Here is the R code for doing this analysis

```
#Import dataset treatmentvscontroldata.txt in folder S231Datasets
y<-TeatmentVsContolData$DRP
y1<-y[seq(1,21,1)]   # data for Treatment Group
y2<-y[seq(22,44,1)]   # data for Control Group
# qqplots
qqnorm(y1,main="Qqplot for Treatment Group")
qqnorm(y2,main="Qqplot for Control Group")
# t test for hypothesis of no difference in means
# and 95% confidence interval for mean difference mu
# note that R uses mu = mu_control - mu_treament
t.test(DRP~Group,data=treatmentvscontroldata,var.equal=T,
conf.level=0.95)
```

15. A study was done to compare the durability of diesel engine bearings made of two different compounds. Ten bearings of each type were tested. The following table gives the "times" until failure (in units of millions of cycles):

Type I: y_{1i}	3.03	5.53	5.60	9.30	9.92	12.51	12.95	15.21	16.04	16.84
Type II: y_{2i}	3.19	4.26	4.47	4.53	4.67	4.69	12.78	6.79	9.37	12.75

$$\bar{y}_1 = 10.693 \quad \sum_{i=1}^{10} (y_{1i} - \bar{y}_1)^2 = 209.02961 \quad \bar{y}_2 = 6.75 \quad \sum_{i=1}^{10} (y_{2i} - \bar{y}_2)^2 = 116.7974$$

To analyze these data assume

$$Y_{1j} \sim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 10 \text{ independently}$$

for the Type I bearings and independently

$$Y_{2j} \sim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 10 \text{ independently}$$

for the Type II bearings where μ_1, μ_2 and σ are unknown parameters.

(a) Obtain a 90% confidence interval for the difference in the means $\mu_1 - \mu_2$.

(b) Test the hypothesis $H_0 : \mu_1 = \mu_2$.

(c) It has been suggested that log failure times are approximately Normally distributed, but not failure times. Assuming that the $\log Y$'s for the two types of bearing are Normally distributed with the same variance, test the hypothesis that the two distributions have the same mean. How does the answer compare with that in part (b)?

(d) How might you check whether Y or $\log Y$ is closer to Normally distributed?

(e) Give a plot of the data which could be used to describe the data and your analysis.

16. To compare the mathematical abilities of incoming first year students in Mathematics and Engineering, 30 Math students and 30 Engineering students were selected randomly from their first year classes and given a mathematics aptitude test. A summary of the resulting marks y_{1i} (for the math students) and y_{2i} (for the engineering students), $i = 1, 2, \ldots, 30$, is as follows:

$$\begin{array}{lllll} \text{Math students:} & n = 30 & \bar{y}_1 = 120 & \sum_{i=1}^{30} (y_{1i} - \bar{y}_1)^2 = 3050 \\ \text{Engineering students:} & n = 30 & \bar{y}_2 = 114 & \sum_{i=1}^{30} (y_{2i} - \bar{y}_2)^2 = 2937 \end{array}$$

To analyze these data assume

$$Y_{1j} \sim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 30 \text{ independently}$$

for the Math students and independently

$$Y_{2j} \sim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 30 \text{ independently}$$

for Engineering students where μ_1, μ_2 and σ are unknown parameters.

(a) Obtain a 95% confidence interval for the difference in mean scores for first year Math and Engineering students.

(b) Test the hypothesis that the difference is zero.

17. Fourteen welded girders were cyclically stressed at 1900 pounds per square inch and the numbers of cycles to failure were observed. The sample mean and variance of the log failure times were $\bar{y}_1 = 14.564$ and $s_1^2 = 0.0914$. Similar tests on ten additional girders with repaired welds gave $\bar{y}_2 = 14.291$ and $s_2^2 = 0.0422$. Log failure times are assumed to be independent with a Gaussian distribution. Assuming equal variances for the two types of girders, obtain a 95% confidence interval for the difference in mean log failure times and test the hypothesis of no difference.

18. Consider the data in Chapter 1 on the lengths of male and female coyotes. The data are available in the file *coyotedata.txt* posted on the course website.

(a) Construct a 95% confidence interval the difference in mean lengths for the two sexes. State your assumptions.

(b) Estimate $P(Y_1 > Y_2)$ (give the maximum likelihood estimate), where Y_1 is the length of a randomly selected female and Y_2 is the length of a randomly selected male. Can you suggest how you might get a confidence interval?

(c) Give separate confidence intervals for the average length of males and females.

19. To assess the effect of a low dose of alcohol on reaction time, a sample of 24 student volunteers took part in a study. Twelve of the students (randomly chosen from the 24) were given a fixed dose of alcohol (adjusted for body weight) and the other twelve got a nonalcoholic drink which looked and tasted the same as the alcoholic drink. Each student was then tested using software that flashes a coloured rectangle randomly placed on a screen; the student has to move the cursor into the rectangle and double click the mouse. As soon as the double click occurs, the process is repeated, up to a total of 20 times. The response variate is the total reaction time (i.e. time to complete the experiment) over the 20 trials. The data are given below.

"Alcohol" Group:

1.33 1.55 1.43 1.35 1.17 1.35 1.17 1.80 1.68 1.19 0.96 1.46

$$\bar{y}_1 = \tfrac{16.44}{12} = 1.370 \quad \sum_{i=1}^{12} (y_{1i} - \bar{y}_1)^2 = 0.608$$

"Non-Alcohol" Group:

1.68 1.30 1.85 1.64 1.62 1.69 1.57 1.82 1.41 1.78 1.40 1.43

$$\bar{y}_2 = \tfrac{19.19}{12} = 1.599 \quad \sum_{i=1}^{12} (y_{2i} - \bar{y}_2)^2 = 0.35569$$

To analyze these data assume

$$Y_{1j} \sim G(\mu_1, \sigma), \quad j = 1, 2, \ldots, 12 \text{ independently}$$

for the Alcohol Group and independently

$$Y_{2j} \sim G(\mu_2, \sigma), \quad j = 1, 2, \ldots, 12 \text{ independently}$$

for the Non-Alcohol Group where μ_1, μ_2 and σ are unknown parameters. Determine a 95% confidence interval for the difference in the means $\mu_1 - \mu_2$. What can the researchers conclude on the basis of this study?

20. An experiment was conducted to compare gas mileages of cars using a synthetic oil and a conventional oil. Eight cars were chosen as representative of the cars in general use. Each car was run twice under as similar conditions as possible (same drivers, routes, etc.), once with the synthetic oil and once with the conventional oil, the order of use of the two oils being randomized.
The gas mileages were as follows:

Car	1	2	3	4	5	6	7	8
Synthetic: y_{1i}	21.2	21.4	15.9	37.0	12.1	21.1	24.5	35.7
Conventional: y_{21}	18.0	20.6	14.2	37.8	10.6	18.5	25.9	34.7
$y_i = y_{1i} - y_{2i}$	3.2	0.8	1.7	−0.8	1.5	2.6	−1.4	1

$$\bar{y}_1 = 23.6125 \quad \sum_{i=1}^{8} (y_{1i} - \bar{y}_1)^2 = 535.16875$$
$$\bar{y}_2 = 22.5375 \quad \sum_{i=1}^{8} (y_{2i} - \bar{y}_2)^2 = 644.83875$$
$$\bar{y} = 1.075 \quad \sum_{i=1}^{8} (y_i - \bar{y})^2 = 17.135$$

(a) Obtain a 95% confidence interval for the difference in mean gas mileage, and state the assumptions on which your analysis depends.

(b) Repeat (a) if the natural pairing of the data is (improperly) ignored.

(c) Why is it better to take pairs of measurements on eight cars rather than taking only one measurement on each of 16 cars?

21. The following table gives the number of staff hours per month lost due to accidents in eight factories of similar size over a period of one year and after the introduction of an industrial safety program.

Factory i	1	2	3	4	5	6	7	8
After: y_{1i}	28.7	62.2	28.9	0.0	93.5	49.6	86.3	40.2
Before: y_{2i}	48.5	79.2	25.3	19.7	130.9	57.6	88.8	62.1
$y_i = y_{1i} - y_{2i}$	−19.8	−17.0	3.6	−19.7	−37.4	−8.0	−2.5	−21.9

$$\bar{y} = -15.3375 \quad \sum_{i=1}^{8} (y_i - \bar{y})^2 = 1148.79875$$

There is a natural pairing of the data by factory. Factories with the best safety records before the safety program tend to have the best records after the safety program as well. The analysis of the data must take this pairing into account and therefore the model

$$Y_i \sim G(\mu, \sigma), \quad i = 1, 2, \ldots, 8 \quad \text{independently}$$

is assumed where μ and σ are unknown parameters.

(a) The parameters μ and σ correspond to what attributes of interest in the study population?

(b) Calculate a 95% confidence interval for μ.

(c) Test the hypothesis of no difference due to the safety program, that is, test the hypothesis $H_0 : \mu = 0$.

22. **Comparing sorting algorithms:** Suppose you want to compare two algorithms A and B that will sort a set of numbers into an increasing sequence. (The R function, `sort(x)`, will, for example, sort the elements of the numeric vector x.) To compare the speed of algorithms A and B, you decide to "present" A and B with random permutations of n numbers, for several values of n. Explain exactly how you would set up such a study, and discuss what pairing would mean in this context.

23. **Sorting algorithms continued:** Two sort algorithms as in the preceding problem were each run on (the same) 20 sets of numbers (there were 500 numbers in each set). Times to sort the sets of two numbers are shown below.

Set:	1	2	3	4	5	6	7	8	9	10
A:	3.85	2.81	6.47	7.59	4.58	5.47	4.72	3.56	3.22	5.58
B:	2.66	2.98	5.35	6.43	4.28	5.06	4.36	3.91	3.28	5.19
y_i	1.19	−.17	1.12	1.16	0.30	0.41	0.36	−.35	−.06	0.39

Set:	11	12	13	14	15	16	17	18	19	20
A:	4.58	5.46	3.31	4.33	4.26	6.29	5.04	5.08	5.08	3.47
B:	4.05	4.78	3.77	3.81	3.17	6.02	4.84	4.81	4.34	3.48
y_i	0.53	0.68	−.46	0.52	1.09	0.27	0.20	0.27	0.74	−.01

$$\bar{y} = 0.409 \quad s^2 = \tfrac{1}{19} \sum_{i=1}^{20} (y_i - \bar{y})^2 = 0.237483$$

Data are available in the file *sortdata.txt* available on the course website.

(a) Since the two algorithms are each run on the same 20 sets of numbers we analyse the differences $y_i = y_{Ai} - y_{Bi}$, $i = 1, 2, \ldots, 20$. Construct a 99% confidence interval for the difference in the average time to sort with algorithms A and B, assuming the difference have a Gaussian distribution.

(b) Use a Normal qqplot to determine if a Gaussian model is reasonable for the differences.

(c) Give a point estimate of the probability that algorithm B will sort a randomly selected list faster than A.

(d) Another way to estimate the probability p in part (c) is to notice that of the 20 sets of numbers in the study, B sorted faster on 15 sets of numbers. Obtain an approximate 95% confidence interval for p. (It is also possible to get a confidence interval using the Gaussian model.)

(e) Suppose the study had actually been conducted using two independent samples of size 20 each. Using the two sample Normal analysis determine a 99% confidence interval for the difference in the average time to sort with algorithms A and B. Note:

$$\bar{y}_1 = 4.7375 \quad s_1^2 = 1.4697 \quad \bar{y}_2 = 4.3285 \quad s_2^2 = 0.9945$$

How much better is the paired study as compared to the two sample study?

(f) Here is the R code for doing the t tests and confidence intervals for the paired analysis and the unpaired analysis:

```
# Import dataset sortdata.txt in folder S231Datasets
t.test(Time~Alg,data=sortdata,paired=T,conf.level=0.99)
t.test(Time~Alg,data=sortdata,paired=F,var.equal=T,conf.level=0.99)
```

24. **Challenge Problem** Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the $G(\mu_1, \sigma_1)$ distribution and let $X_1, \ldots, X_n$ be a random sample from the $G(\mu_2, \sigma_2)$ distribution. Obtain the likelihood ratio test statistic for testing the hypothesis $H_0 : \sigma_1 = \sigma_2$ and show that it is a function of $F = S_1^2/S_2^2$, where S_1^2 and S_2^2 are the sample variances from the y and x samples respectively.

25. **Challenge Problem** Readings produced by a set of scales are independent and Normally distributed about the true weight of the item being measured. A study is carried out to assess whether the standard deviation of the measurements varies according to the weight of the item.

(a) Ten weighings of a 10 kilogram weight yielded $\bar{y} = 10.004$ and $s = 0.013$ as the sample mean and standard deviation. Ten weighings of a 40 kilogram weight yielded $\bar{y} = 39.989$ and $s = 0.034$. Is there any evidence of a difference in the standard deviations for the measurements of the two weights?

(b) Suppose you had a further set of weighings of a 20 kilogram item. How could you study the question of interest further?

26. **Challenge Problem** Suppose you have a model where the mean of the response variable Y_i given the covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ has the form

$$\mu_i = E(Y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i; \boldsymbol{\beta})$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters. Then the **least squares estimate** of $\boldsymbol{\beta}$ based on data $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots, n$ is the value that minimizes the objective function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - \mu(\mathbf{x}_i; \boldsymbol{\beta})]^2$$

Show that the least squares estimate of $\boldsymbol{\beta}$ is the same as the maximum likelihood estimate of $\boldsymbol{\beta}$ in the Gaussian model $Y_i \sim G(\mu_i, \sigma)$, when μ_i is of the form

$$\mu_i = \mu(\mathbf{x}_i; \boldsymbol{\beta}) = \sum_{j=1}^{k} \beta_j x_{ij}$$

27. **Challenge Problem Optimal Prediction** In many settings we want to use covariates $\mathbf{x}$ to predict a future value Y. (For example, we use economic factors $\mathbf{x}$ to predict the price Y of a commodity a month from now.) The value Y is random, but suppose we know $\mu(\mathbf{x}) = E(Y|\mathbf{x})$ and $\sigma(\mathbf{x})^2 = Var(Y|\mathbf{x})$.

    (a) Predictions take the form $\hat{Y} = g(\mathbf{x})$, where $g(\cdot)$ is our "prediction" function. Show that the minimum achievable value of $E(\hat{Y} - Y)^2$ is minimized by choosing $g(\mathbf{x}) = \mu(\mathbf{x})$.

    (b) Show that the minimum achievable value of $E(\hat{Y} - Y)^2$, that is, its value when $g(\mathbf{x}) = \mu(\mathbf{x})$ is $\sigma(\mathbf{x})^2$.
    This shows that if we can determine or estimate $\mu(\mathbf{x})$, then "optimal" prediction (in terms of Euclidean distance) is possible. Part (b) shows that we should try to find covariates x for which $\sigma(\mathbf{x})^2 = Var(Y|\mathbf{x})$ is as small as possible.

    (c) What happens when $\sigma(x)^2$ is close to zero? (Explain this in ordinary English.)

7. MULTINOMIAL MODELS AND GOODNESS OF FIT TESTS

7.1 Likelihood Ratio Test for the Multinomial Model

Many important hypothesis testing problems can be addressed using Multinomial models. Suppose the data arise from a Multinomial distribution with joint probability function

$$f(y_1, y_2, \ldots, y_k; \theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k} \tag{7.1}$$

where $y_j = 0, 1, \ldots$ and $\sum_{j=1}^{k} y_j = n$. The Multinomial probabilities θ_j satisfy $0 \leq \theta_j \leq 1$ and $\sum_{j=1}^{k} \theta_j = 1$. The likelihood function based on (7.1) is

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_k^{y_k}$$

or more simply

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_j^{y_j} \tag{7.2}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. It can be shown that $L(\boldsymbol{\theta})$ is maximized by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ where $\hat{\theta}_j = y_j/n$, $j = 1, 2, \ldots, k$. Note that although $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ there are actually only $k - 1$ parameters to be estimated since $\sum_{j=1}^{k} \theta_j = 1$.

Suppose that we wish to test the hypothesis that the probabilities $\theta_1, \theta_2, \ldots, \theta_k$ are related in some way, for example, that they are all functions of a parameter $\boldsymbol{\alpha}$, such that

$$H_0 : \theta_j = \theta_j(\boldsymbol{\alpha}) \quad \text{for } j = 1, 2, \ldots, k \tag{7.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ and $p < k - 1$. In other words, p is equal to the number of parameters that need to be estimated in the model assuming the null hypothesis (7.3). For example, suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ so $k = 4$ and the null hypothesis is

$$H_0 : \theta_1 = \alpha_1,\ \theta_2 = \alpha_1 + \alpha_2,\ \theta_3 = \alpha_2,\ \theta_4 = 1 - 2(\alpha_1 + \alpha_2)$$

then $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $p = 2$.

A likelihood ratio test of (7.3) is based on the likelihood ratio statistic

$$\Lambda = -2 \log \left[\frac{L(\tilde{\boldsymbol{\theta}}_0)}{L(\tilde{\boldsymbol{\theta}})} \right] \tag{7.4}$$

where $\tilde{\boldsymbol{\theta}}_0$ maximizes $L(\boldsymbol{\theta})$ assuming the null hypothesis (7.3) is true.

The test statistic (7.4) can be written in a simple form. Let $\tilde{\boldsymbol{\theta}}_0 = (\theta_1(\tilde{\alpha}), \ldots, \theta_k(\tilde{\alpha}))$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$ under the null hypothesis (7.3). Then

$$\Lambda = 2 \sum_{j=1}^{k} Y_j \log \left[\frac{\tilde{\theta}_j}{\theta_j(\tilde{\alpha})} \right]$$

Noting that $\tilde{\theta}_j = Y_j/n$ and defining the expected frequencies under H_0 as

$$E_j = n\theta_j(\tilde{\alpha}) \quad \text{for } j = 1, 2, \ldots, k$$

we can rewrite Λ as

$$\Lambda = 2 \sum_{j=1}^{k} Y_j \log \left(\frac{Y_j}{E_j} \right) \tag{7.5}$$

Let

$$\lambda = 2 \sum_{j=1}^{k} y_j \log \left(\frac{y_j}{e_j} \right)$$

be the observed value of Λ where $e_j = n\theta_j(\hat{\alpha})$, $j = 1, 2, \ldots, k$. (Remember log = ln.) Note that the value of λ will be close to 0 if the observed values $y_1, y_2, \ldots, y_k$ are close to the expected values $e_1, e_2, \ldots, e_k$ and that the value of λ will be large if the y_j's and e_j's differ greatly.

If n is large and H_0 is true then the distribution of Λ is approximately $\chi^2(k-1-p)$. This enables us to compute $p - values$ from observed data by using the approximation

$$p - value = P(\Lambda \geq \lambda; H_0) \approx P(W \geq \lambda) \quad \text{where } W \sim \chi^2(k-1-p)$$

This approximation is accurate when n is large and none of the θ_j's is too small. In particular, the expected frequencies determined assuming H_0 is true should all be at least 5 to use the Chi-squared approximation.

An alternative test statistic that was developed historically before the likelihood ratio test statistic is the Pearson goodness of fit statistic

$$D = \sum_{j=1}^{k} \frac{(Y_j - E_j)^2}{E_j} \tag{7.6}$$

with observed value

$$d = \sum_{j=1}^{k} \frac{(y_j - e_j)^2}{e_j}$$

The Pearson goodness of fit statistic has similar properties to Λ, that is, d takes on small values if the y_j's and e_j's are close in value and d takes on large values if the y_j's and e_j's differ greatly. It also turns out that, like Λ, the statistic D has a limiting $\chi^2(k-1-p)$ distribution when H_0 is true.

The remainder of this chapter consists of the application of the general methods above to some important testing problems.

7.2 Goodness of Fit Tests

Recall from Section 2.4 that one way to check the fit of a probability distribution is by comparing the observed frequencies f_j and the expected frequencies $e_j = n\hat{p}_j$. As indicated there we did not know how close the observed and expected frequencies needed to be to conclude that the model was adequate. It is possible to test the appropriateness of a model by using the Multinomial model. We illustrate this test through two examples.

Example 7.2.1 MM, MN, NN blood types

In Example 2.4.2, n people were selected from a population and classified as being one of three blood types MM, MN, NN. Let Y_1 = number of MM types observed, Y_2 = number of MN types observed and Y_3 = number of NN types observed. If the proportions of the population that are these three types are θ_1, θ_2, θ_3 respectively, with $\theta_1 + \theta_2 + \theta_3 = 1$ then the joint probability function of Y_1, Y_2, Y_3 is Multinomial$(n; \theta_1, \theta_2, \theta_3)$ and $k = 3$.

Genetic theory indicates that the θ_j's can be expressed in terms of a single parameter α. The null hypothesis corresponding to this is

$$H_0 : \theta_1 = \alpha^2, \;\; \theta_2 = 2\alpha(1-\alpha), \;\; \theta_3 = (1-\alpha)^2 \tag{7.7}$$

There is only one unknown parameter α under (7.7), so $p = 1$.

Data collected on 100 persons gave $y_1 = 17$, $y_2 = 46$, $y_3 = 37$, and we can use this to test the hypothesis H_0. The likelihood function under (7.7) is

$$\begin{aligned} L_1(\alpha) &= L(\theta_1(\alpha), \theta_2(\alpha), \theta_3(\alpha)) \\ &= c(\alpha^2)^{17}[2\alpha(1-\alpha)]^{46}[(1-\alpha)^2]^{37} \\ &= c\alpha^{80}(1-\alpha)^{120} \quad \text{for } 0 \le \alpha \le 1 \end{aligned}$$

where c is a constant with respect to α. We easily find that $\hat{\alpha} = 0.40$. The observed expected frequencies under (7.7) are $e_1 = 100\hat{\alpha}^2 = 16$, $e_2 = 100[2\hat{\alpha}(1-\hat{\alpha})] = 48$, $e_3 = 100[(1-\hat{\alpha})^2] = 36$. The observed value of the likelihood ratio statistic (7.5) is

$$2\sum_{j=1}^{3} y_j \log\left(\frac{y_j}{e_j}\right) = 2\left[17\log\left(\frac{17}{16}\right) + 46\log\left(\frac{46}{48}\right) + 37\log\left(\frac{37}{36}\right)\right] = 0.17$$

The degrees of freedom for the Chi-squared approximation equal $k-1-p=3-1-1=1$. The approximate $p-value$ is

$$\begin{aligned} p-value &\approx P(\Lambda \geq 0.17; H_0) \approx P(W \geq 0.17) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1-P\left(Z \leq 0.41\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2(1-0.6591) = 0.6818 \end{aligned}$$

so there is no evidence against the model (7.7).

The observed values of the Pearson goodness of fit statistic (7.6) and the likelihood ratio statistic Λ are usually close when n is large and so it does not matter which test statistic is used. In this case we find that the observed value of (7.6) for these data is also 0.17.

Example 7.2.2 Goodness of fit and Poisson model

The number of service interruptions in a communications system over 200 separate days is summarized in the following frequency table:

Number of interruptions:	0	1	2	3	4	5	>5	Total
Frequency observed y_j:	64	71	42	18	4	1	0	200

Let Y_j = number of times j interruptions are observed. The joint model for the Y_j's is Multinomial.

We wish to test whether a Poisson model for Y = the number of interruptions on a single day is consistent with these data. The null hypothesis is

$$H_0 : \theta_j = \frac{\theta^j e^{-\theta}}{j!} \quad \text{for } j = 0, 1, \ldots$$

(Note that we are using θ rather than α as the parameter of interest.) The maximum likelihood estimate of θ based on the observed data in the table is

$$\hat{\theta} = \frac{1}{200}\left[0\left(64\right)+1\left(71\right)+2\left(42\right)+3\left(18\right)+4\left(4\right)+5\left(1\right)\right] = \frac{230}{200} = 1.15$$

The observed and expected frequencies assuming a Poisson(1.15) distribution are given in the table below

No. of interruptions	0	1	2	3	4	≥ 5	Total
y_i	64	71	42	18	4	1	200
e_i	63.33	72.83	41.88	16.05	4.61	1.30	

where

$$e_j = 200\left[\frac{(1.15)^j\, e^{-1.15}}{j!}\right] \quad \text{for} \quad j = 0, 1, \ldots, 4$$

and the last category is obtained by subtraction. Since the expected frequency in the last category is less than 5 we combine the last two categories to obtain

No. of interruptions	0	1	2	3	≥ 4	Total
y_i (e_i)	64(63.33)	71(72.83)	42(41.88)	18(16.05)	5(5.91)	200

The observed value of the likelihood ratio statistic is

$$\begin{aligned} & 2\left[64\log\left(\frac{64}{63.33}\right)+71\log\left(\frac{71}{72.83}\right)+42\log\left(\frac{42}{41.88}\right)+18\log\left(\frac{18}{16.05}\right)+5\log\left(\frac{5}{5.91}\right)\right] \\ = \; & 0.43 \end{aligned}$$

The collapsed table has five categories so $k = 5$ and only one parameter θ has been estimated under H_0 so $p = 1$. The degrees of freedom for the Chi-squared approximation equal $k - 1 - p = 5 - 1 - 1 = 3$. Since

$$\begin{aligned} p - value &\approx P(W > 0.43) \text{ where } W \sim \chi^2(3) \\ &= 0.93 > 0.1 \end{aligned}$$

there is no evidence against the Poisson model based on the observed data.

Example 7.2.3 Goodness of fit and Exponential model

Continuous distributions can also be tested by grouping the data into intervals and then using the Multinomial model. Example 2.6.2 previously did this in an informal way for an Exponential distribution and the lifetimes of brake pads data.

Suppose a random sample $t_1, t_2, \ldots, t_{100}$ is collected and we wish to test the hypothesis that the data come from an Exponential(θ) distribution. We partition the range of T into intervals $j = 1, 2, \ldots, k$, and count the number of observations y_j that fall into each interval. Assuming an Exponential(θ) model, the probability that an observation lies in the j'th interval $I_j = (a_{j-1}, a_j)$ is

$$p_j(\theta) = \int_{a_{j-1}}^{a_j} f(t;\theta)dt = e^{-a_{j-1}/\theta} - e^{-a_j/\theta} \quad \text{for } j = 1, 2, ..., k \tag{7.8}$$

and if y_j is the number of observations (t's) that lie in I_j, then $Y_1, Y_2, \ldots, Y_k$ follow a Multinomial$(100; p_1(\theta), p_2(\theta), \ldots, p_k(\theta))$ distribution.

Suppose the observed data are

Interval	$0-100$	$100-200$	$200-300$	$300-400$	$400-600$	$600-800$	>800
y_j	29	22	12	10	10	9	8
e_j	27.6	20.0	14.4	10.5	13.1	6.9	7.6

so $k = 7$. To calculate the expected frequencies under the null hypothesis (7.8) we need an estimate of θ which is obtained by maximizing the likelihood function

$$L(\theta) = \prod_{j=1}^{7} [p_j(\theta)]^{y_j}$$

Since there is only one unknown parameter θ under (7.8), $p = 1$. It is possible to maximize $L(\theta)$ to obtain $\hat{\theta} = 310.0$. The expected frequencies, $e_j = 100p_j(\hat{\theta})$, $j = 1, 2, \ldots, 7$, are given in the table.

The observed value of the likelihood ratio statistic (7.5) is

$$2\sum_{j=1}^{7} y_j \log\left(\frac{y_j}{e_j}\right) = 2\left[29\log\left(\frac{29}{27.6}\right) + 22\log\left(\frac{22}{20}\right) + \cdots + 8\log\left(\frac{8}{7.6}\right)\right] = 1.91$$

The degrees of freedom for the Chi-squared approximation equal $k-1-p = 7-1-1 = 5$. Since

$$\begin{aligned} p-value &= P(\Lambda \geq 1.91; H_0) \\ &\approx P(W \geq 1.91) \quad \text{where } W \sim \chi^2(5) \\ &= 0.86 > 0.1 \end{aligned}$$

there is no evidence against the model (7.8) based on the observed data.

A goodness of fit test has some arbitrary elements, since we could have used different intervals and a different number of intervals. Theory has been developed on how best to choose the intervals. For this course we only give rough guidelines which are: chose $4-10$ intervals, so that the observed expected frequencies under H_0 are at least 5.

Example 7.2.4 Lifetime of brake pads and the Exponential model

Recall the data in Example 2.6.2 on the lifetimes of brake pads. The expected frequencies are calculated using an Exponential model with mean estimated by the sample mean $\hat{\theta} = 49.0275$. The data are given in Table 7.1.

Interval	Observed Frequency: f_j	Expected Frequency: e_j
$[0, 15)$	21	52.72
$[15, 30)$	45	38.82
$[30, 45)$	50	28.59
$[45, 60)$	27	21.05
$[60, 75)$	21	15.50
$[75, 90)$	9	11.42
$[90, 105)$	12	8.41
$[105, 120)$	7	6.19
$[120, +\infty)$	8	17.3
Total	200	200

Table 7.1: Frequency table for brake pad data

The observed value of the likelihood ratio statistic (7.5) is

$$2\sum_{j=1}^{9} f_j \log\left(\frac{f_j}{e_j}\right) = 2\left[21\log\left(\frac{21}{52.72}\right) + 45\log\left(\frac{45}{38.82}\right) + \cdots + 8\log\left(\frac{8}{17.3}\right)\right] = 50.36$$

The expected frequencies are all at least five and so $k = 9$. There is only one parameter θ to be estimated under the hypothesized Exponential(θ) model so $p = 1$. The degrees of freedom for the Chi-squared approximation equal $k - 1 - p = 9 - 1 - 1 = 7$. Since

$$\begin{aligned} p - value &= P(\Lambda \geq 50.36; H_0) \approx P(W \geq 50.36) \quad \text{where } W \sim \chi^2(7) \\ &\approx 0 \end{aligned}$$

there is very strong evidence against the Exponential model based on the observed data. This conclusion is not unexpected since, as we noted in Example 2.6.2, the observed and expected frequencies are not in close agreement. We could have chosen a different set of intervals for these continuous data but the same conclusion of a lack of fit would be obtained for any reasonable choice of intervals.

7.3 Two-Way (Contingency) Tables

Often we want to assess whether two factors or variates appear to be related. One tool for doing this is to test the hypothesis that the factors are independent and thus statistically unrelated. We will consider this in the case where both variates are discrete, and take on a fairly small number of possible values. This turns out to cover a great many important settings.

Two types of studies give rise to data that can be used to test independence, and in both cases the data can be arranged as frequencies in a *two-way* table. These tables are also called *contingency* tables.

Cross-Classification of a Random Sample of Individuals

Suppose that individuals or items in a population can be classified according to each of two factors A and B. For A, an individual can be any of a mutually exclusive types $A_1, A_2, \ldots, A_a$ and for B an individual can be any of b mutually exclusive types $B_1, B_2, \ldots, B_b$, where $a \geq 2$ and $b \geq 2$.

If a random sample of n individuals is selected, let y_{ij} denote the number that have A-type A_i and B-type B_j. The observed data may be arranged in a two-way table as seen below:

$A \backslash B$	B_1	B_2	$\cdots$	B_b	Total
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}	r_1
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}	r_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_a	y_{a1}	$\cdots$	$\cdots$	y_{ab}	r_a
Total	c_1	c_2	$\cdots$	c_b	n

where $r_i = \sum_{j=1}^{b} y_{ij}$ are the row totals, $c_j = \sum_{i=1}^{a} y_{ij}$ are the column totals, and $\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} = n$. Let θ_{ij} be the probability a randomly selected individual is combined type (A_i, B_j) and

note that $\sum_{i=1}^{a}\sum_{j=1}^{b}\theta_{ij} = 1$. The $a \times b$ frequencies $(Y_{11}, Y_{12}, \ldots, Y_{ab})$ follow a Multinomial distribution with $k = ab$ classes.

To test independence of the A and B classifications, we test the hypothesis

$$H_0 : \theta_{ij} = \alpha_i \beta_j \quad \text{for } i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b \tag{7.9}$$

where $0 < \alpha_i < 1$, $0 < \beta_j < 1$, $\sum_{i=1}^{a}\alpha_i = 1$, $\sum_{j=1}^{b}\beta_j = 1$. Note that

$$\alpha_i = P(\text{an individual is type } A_i) \quad \text{and} \quad \beta_j = P(\text{an individual is type } B_j)$$

and that (7.9) is the standard definition for independent events: $P(A_i \cap B_j) = P(A_i)P(B_j)$.

We note that testing (7.9) falls into the general framework of Section 7.1, where $k = ab$, and the number of parameters estimated under (7.9) is $p = (a-1) + (b-1) = a + b - 2$. All that needs to be done in order to use the statistics (7.5) or (7.6) to test H_0 is to obtain the maximum likelihood estimates $\hat{\alpha}_i$, $\hat{\beta}_j$ under the model (7.9), and then the calculate the expected frequencies e_{ij}.

Under the model (7.9), the likelihood function for the y_{ij}'s is

$$\begin{aligned} L_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{i=1}^{a}\prod_{j=1}^{b}[\theta_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{y_{ij}} \\ &= \prod_{i=1}^{a}\prod_{j=1}^{b}(\alpha_i \beta_j)^{y_{ij}} \end{aligned}$$

The log likelihood function $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ must be maximized subject to the linear constraints $\sum_{i=1}^{a}\alpha_i = 1$, $\sum_{j=1}^{b}\beta_j = 1$. The maximum likelihood estimates can be shown to be

$$\hat{\alpha}_i = \frac{r_i}{n},\ \hat{\beta}_j = \frac{c_j}{n} \quad i = 1, 2, \ldots a; \quad j = 1, 2, \ldots b$$

and the expected frequencies are given by

$$e_{ij} = n\hat{\alpha}_i\hat{\beta}_j = \frac{r_i c_j}{n} \quad i = 1, 2, \ldots a; \quad j = 1, 2, \ldots b \tag{7.10}$$

The observed value of the likelihood ratio statistic for testing H_0 is

$$\lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} \log\left(\frac{y_{ij}}{e_{ij}}\right)$$

The degrees of freedom for the Chi-squared approximation are

$$k - 1 - p = (ab - 1) - (a - 1 + b - 1) = (a - 1)(b - 1)$$

and the approximate $p - value$ is given by

$$p - value = P\left(\Lambda \geq \lambda; H_0\right) \approx P\left(W \geq \lambda\right) \quad \text{where } W \sim \chi^2\left((a-1)(b-1)\right)$$

Example 7.3.1 Blood classifications

Human blood is classified according to several systems. Two systems are the OAB system and the Rh system. In the former a person is one of four types O, A, B, AB and in the latter system a person is Rh+ or Rh−. To determine whether these two classification systems are genetically independent, a random sample of 300 persons were chosen. Their blood was classified according to the two systems and the observed frequencies are given in the table below.

	O	A	B	AB	Total
Rh+	82	89	54	19	244
Rh−	13	27	7	9	56
Total	95	116	61	28	300

We can think of the Rh types as the A-type classification and the OAB types as the B-type classification in the general theory above. The row and column totals are also shown in the table, since they are the values needed to compute the e_{ij}'s in (7.10).

To carry out the test that a person's Rh and OAB blood types are statistically independent, we merely need to compute the e_{ij}'s by (7.10). For example,

$$e_{11} = \frac{(244)(95)}{300} = 77.27, \quad e_{12} = \frac{244(116)}{300} = 94.35 \text{ and } e_{13} = \frac{244\,(61)}{300} = 49.61$$

The remaining expected frequencies can be obtained by subtraction and these are given in the table below in brackets below the observed frequencies.

	O	A	B	AB	Total
Rh+	82 (77.27)	89 (94.35)	54 (49.61)	19 (22.77)	244
Rh−	13 (17.73)	27 (21.65)	7 (11.39)	9 (5.23)	56
Total	95	116	61	28	300

The degrees of freedom for the Chi-squared approximation equal $(a-1)(b-1) = 3\,(1) = 3$ which is consistent with the fact that, once we had calculated three of the expected frequencies, the remaining expected frequencies could be obtained by subtraction.

The observed value of the likelihood ratio test statistic is $\lambda = 8.447$. Since

$$\begin{aligned} p - value &\approx P(W \geq 8.447) \text{ where } W \sim \chi^2\,(3) \\ &= 0.0376 < 0.05 \end{aligned}$$

there is evidence against the hypothesis of independence based on the data.

Note that by comparing the e_{ij}'s and the y_{ij}'s we see that the degree of dependence does not appear large.

Testing equality of Multinomial parameters for two or more groups

A similar problem arises when individuals in a population can be one of b types $B_1, B_2, \ldots, B_b$, but where the population is sub-divided into a groups $A_1, A_2, \ldots, A_a$. In this case, we might be interested in whether the proportions of individuals of types $B_1, B_2, \ldots, B_b$ are the same for each group. This is essentially the same as the question of independence in the preceding section: we want to know whether the probability θ_{ij} that a person in population group i is B-type B_j is the same for all $i = 1, 2, \ldots, a$. That is, $\theta_{ij} = P(B_j|A_i)$ and we want to know if this depends on A_i or not.

Although the framework is superficially the same as the preceding section, the details are a little different. In particular, the probabilities θ_{ij} satisfy

$$\theta_{i1} + \theta_{i2} + \cdots + \theta_{ib} = 1 \quad \text{for each} \quad i = 1, 2, \ldots, a$$

and the hypothesis we are interested in testing is

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_a, \tag{7.11}$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ib})$. Furthermore, the data in this case arise by selecting specified numbers of individuals n_i from groups $i = 1, 2, \ldots, a$ and so there are actually a different Multinomial distributions, Multinomial$(n_i; \theta_{i1}, \theta_{i2}, \ldots, \theta_{ib})$, $i = 1, 2, \ldots, a$.

If we denote the observed frequency of B_j-type individuals in the sample from the i'th group as y_{ij} (where $y_{i1} + y_{i2} + \cdots + y_{ib} = n_i$), then it can be shown that the likelihood ratio statistic for testing (7.11) is exactly the same as (7.10), where now the expected frequencies e_{ij} are given by

$$e_{ij} = n_i \left(\frac{y_{+j}}{n} \right) \quad \text{for } i = 1, 2, \ldots, a; \ j = 1, 2, \ldots, b$$

where $n = n_1 + n_2 + \cdots + n_a$ and $y_{+j} = \sum_{i=1}^{a} y_{ij}$. Since $n_i = y_{i+} = \sum_{j=1}^{b} y_{ij}$ the expected frequencies have exactly the same form as in the preceding section, when we lay out the data in a two-way table with a rows and b columns.

Example 7.3.1 Revisited Blood classifications

The study in Example 7.3.1 could have been conducted differently, by selecting a fixed number of Rh+ persons and a fixed number of Rh− persons, and then determining their OAB blood type. Then the proper framework would be to test that the probabilities for the four types O, A, B, AB were the same for Rh+ and for Rh− persons, and so the methods of the present section apply. This study gives exactly the same testing procedure as one where the numbers of Rh+ and Rh− persons in the sample are random, as discussed.

Example 7.3.2 Aspirin and strokes

In a randomized clinical trial to assess the effectiveness of a small daily dose of aspirin in preventing strokes among high-risk persons, a group of patients were randomly assigned to get either aspirin or a placebo. A total of 240 patients were assigned to the aspirin group

and 236 were assigned to the placebo group. (There were actually an equal number in each group but four patients withdrew from the placebo group during the study.) The patients were followed for three years, and it was determined for each person whether they had a stroke during that period or not. The data were as follows (expected frequencies are given in brackets).

	Stroke	No Stroke	Total
Aspirin Group	64 (75.6)	176 (164.4)	240
Placebo Group	86 (74.4)	150 (161.6)	236
Total	150	326	476

The patients receiving aspirin and the patients receiving placebo are two independent groups. We are interested in testing the hypothesis

$$H_0 : \theta_{11} = \theta_{21}$$

where $\theta_{11} = P(\text{stroke})$ for a person in the aspirin group and $\theta_{21} = P(\text{stroke})$ for a person in the placebo group.

The expected frequencies under $H_0 : \theta_{11} = \theta_{21}$ are

$$e_{ij} = \frac{(y_{i+})(y_{+j})}{476} \quad \text{for } i = 1, 2$$

which are shown in the table in brackets. The observed value of the likelihood ratio statistic is

$$2\sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij} \log\left(\frac{y_{ij}}{e_{ij}}\right) = 5.25$$

The original model consists of two independent Binomial models each with their own unknown parameter. Under the null hypothesis $\theta_{11} = \theta_{21}$, the model is two independent Binomial models with only one unknown parameter. Therefore the degrees of freedom for the Chi-squared approximation equal $2 - 1 = 1$ and

$$\begin{aligned} p - value &= P(\Lambda \geq 5.25; H_0) \\ &\approx P(W \geq 5.25) \quad \text{where } W \sim \chi^2(1) \\ &= 2\left[1 - P\left(Z \leq 2.29\right)\right] \quad \text{where } Z \sim N(0,1) \\ &= 2(1 - 0.98899) \\ &= 0.02202 \end{aligned}$$

Since $0.01 < p - value < 0.05$, there is evidence against $H_0 : \theta_{11} = \theta_{21}$ based on the observed data. A look at the y_{ij}'s and the e_{ij}'s indicates that persons receiving aspirin have had fewer strokes than expected under H_0, suggesting that $\theta_{11} < \theta_{21}$.

This analysis can be accompanied by point and interval estimates for θ_{11} and θ_{21}. Let Y_{11} = number of patients in the aspirin group who have a stroke and Y_{21} = the number of patients in the placebo group who have a stroke. Then Y_{11} has a Binomial$(240, \theta_{11})$ distibution and Y_{21} has a Binomial$(236, \theta_{21})$ distribution. The maximum likelihood estimates are given by the sample proportions which are

$$\hat{\theta}_{11} = \frac{y_{11}}{n_1} = \frac{64}{240} = 0.267$$

and

$$\hat{\theta}_{21} = \frac{y_{21}}{n_2} = \frac{86}{236} = 0.364$$

An approximate 95% confidence interval for θ_{11} based on the asymptotic Gaussian pivotal quantity is

$$0.267 \pm 1.96\sqrt{\frac{(0.267)\,(0.733)}{240}} \quad \text{or} \quad [0.211, 0.323]$$

and an approximate 95% confidence interval for θ_{21} based on the asymptotic Gaussian pivotal quantity is

$$0.364 \pm 1.96\sqrt{\frac{(0.364)\,(0.636)}{236}} \quad \text{or} \quad [0.303, 0.425]$$

Confidence intervals for the difference in proportions $\theta_{11} - \theta_{21}$ can also be obtained from the approximate $G(0, 1)$ pivotal quantity

$$\frac{(\tilde{\theta}_{11} - \tilde{\theta}_{21}) - (\theta_{11} - \theta_{21})}{\sqrt{\tilde{\theta}_{11}(1 - \tilde{\theta}_{11})/n_1 + \tilde{\theta}_{21}(1 - \tilde{\theta}_{21})/n_2}}$$

Remark This and other tests involving Binomial probabilities and contingency tables can be carried out using the R function `prop.test` which uses the Pearson goodness of fit statistic.

7.4 Chapter 7 Problems

1. In a large STAT 231 class, each student was given a box of Smarties and then asked to count the number of each colour: red, green, yellow, blue, purple, brown, orange, pink. The observed frequencies were:

Colour:	Red	Green	Yellow	Blue	Purple	Brown	Orange	Pink
Frequency (y_i):	556	678	739	653	725	714	566	797

   Test the hypothesis that each of the colours has the same probability $H_0 : \theta_i = \frac{1}{8}$, $i = 1, 2, \ldots, 8$.

   The following R code calculates the observed values of the likelihood ratio test statistic Λ and the Pearson goodness of fit statistic D and the corresponding $p - values$.

```
y<-c(556,678,739,653,725,714,566,797) # observed frequencies
e<-sum(y)/8   # expected frequencies
lambda<-2*sum(y*log(y/e)) # observed value of LR statistic
df<-7    # degrees for freedom for this example equal 7
pvalue<-1-pchisq(lambda,df)    # p-value for LR test
c(lambda,df,pvalue)
d<-sum((y-e)^2/e)      # observed value of Pearson goodness of fit statistic
pvalue<-1-pchisq(d,df) # p-value for Pearson goodness of fit test
c(d,df,pvalue)
```

   What would you conclude about the distribution of colours in boxes of Smarties?

2. Test whether a Poisson model for Y = the number of alpha particles emitted in a time interval of 1/8 minute is consistent with the Rutherford and Geiger data of Example 2.6.1.

3. Test whether a Poisson model for Y = the number of points per game is consistent with the data for Wayne Gretzky given in Chapter 2, Problem 11.

4. Test whether a Poisson model for Y = the number of points per game is consistent with the data for Sidney Crosby given in Chapter 2, Problem 12.

5. Mass-produced items are packed in cartons of 12 as they come off an assembly line. The items from 250 cartons are inspected for defects, with the following results:

Number defective:	0	1	2	3	4	5	6	> 6	Total
Frequency observed:	103	80	31	19	11	5	1	0	250

   (a) Test the hypothesis that the number of defective items Y in a single carton has a Binomial$(12, \theta)$ distribution.

   (b) Why might the Binomial not be a suitable model?

6. In the Wintario lottery draw, six digit numbers were produced by six machines that operate independently and which each simulate a random selection from the digits $0, 1, \ldots, 9$. Of 736 numbers drawn over a period from 1980-82, the following frequencies were observed for position 1 in the six digit numbers:

Digit (i):	0	1	2	3	4	5	6	7	8	9	Total
Frequency (f_i):	70	75	63	59	81	92	75	100	63	58	736

Consider the 736 draws as trials in a Multinomial experiment and let

$$\theta_j = P(\text{digit } j \text{ is drawn on any trial}),\ j = 0, 1, \ldots, 9$$

If the machines operate in a truly "random" fashion, then we should have $\theta_j = 0.1$, $j = 0, 1, \ldots, 9$.

(a) Test this hypothesis using the likelihood ratio test. What do you conclude?

(b) The data above were for digits in the first position of the six digit Wintario numbers. Suppose you were told that similar likelihood ratio tests had in fact been carried out for each of the six positions, and that position one had been singled out for presentation above because it gave the largest observed value of the likelihood ratio statistic Λ. How would you test the hypothesis $\theta_j = 0.1$, $j = 0, 1, 2, \ldots, 9$ for all six positions simultaneously?

7. A long sequence of digits $(0, 1, \ldots, 9)$ produced by a pseudo random number generator was examined. There were 51 zeros in the sequence, and for each successive pair of zeros, the number of (non-zero) digits between them was counted. The results were as follows:

1	1	6	8	10	22	12	15	0	0
2	26	1	20	4	2	0	10	4	19
2	3	0	5	2	8	1	6	14	2
2	2	21	4	3	0	0	7	2	4
4	7	16	18	2	13	22	7	3	5

(a) Give an appropriate probability model for the number of digits between two successive zeros, if the pseudo random number generator is truly producing digits for which $P\,(\text{any digit } = j) = 0.1$, $j = 0, 1, \ldots, 9$, independent of any other digit.

(b) Construct a frequency table and test the goodness of fit of your model.

8. The table below records data on 292 litters of mice classified according to litter size and number of females in the litter. Note that $y_{n+} = \sum_j y_{nj}$.

		Number of females $= j$					Total number
	y_{nj}	0	1	2	3	4	of litters $= y_{n+}$
	1	8	12				20
Litter	2	23	44	13			80
Size $= n$	3	10	25	48	13		96
	4	5	30	34	22	5	96

   (a) For litters of size n ($n = 1, 2, 3, 4$) assume that the number of females in a litter of size n has Binomial distribution with parameters n and $\theta_n = P(\text{female})$. Test the Binomial model separately for each of the litter sizes $n = 2$, $n = 3$ and $n = 4$. (Why is it of scientific interest to do this?)

   (b) Assuming that the Binomial model is appropriate for each litter size, test the hypothesis that $\theta_1 = \theta_2 = \theta_3 = \theta_4$.

9. The following data on heights of 210 married couples were presented by Yule in 1900.

	Tall wife	Medium wife	Short wife	Total
Tall husband	18	28	19	65
Medium husband	20	51	28	99
Short husband	12	25	9	46
Total	50	104	56	210

   Test the hypothesis that the heights of husbands and wives are independent.
   The following R code determines the $p-value$ for testing the hypothesis of independence.

```
# matrix of observed frequencies
f<-matrix(c(18,28,19,20,51,28,12,25,9),ncol=3,byrow=TRUE)
row<-margin.table(f,1)     # row totals
col<-margin.table(f,2)     # column totals
e<-outer(row,col)/sum(f)   # matrix of expected frequencies
lambda<-2*sum(f*log(f/e))  # observed value of likelihood ratio statistic
df<-(length(row)-1)*(length(col)-1)  # degrees of freedom
pvalue<-1-pchisq(lambda,df)
c(lambda,df,pvalue)
```

10. A study was undertaken to determine whether there is an association between the birth weights of infants and the smoking habits of their parents. Out of 50 infants of above average weight, 9 had parents who both smoked, 6 had mothers who smoked but fathers who did not, 12 had fathers who smoked but mothers who did not, and 23 had parents of whom neither smoked. The corresponding results for 50 infants of below average weight were 21, 10, 6, and 13, respectively.

    (a) Test whether these results are consistent with the hypothesis that birth weight is independent of parental smoking habits.

    (b) Are these data consistent with the hypothesis that, given the smoking habits of the mother, the smoking habits of the father are not related to birth weight?

11. School children with tonsils were classified according to tonsil size and absence or presence of the carrier for streptococcus pyogenes. The results were as follows:

	Normal	Enlarged	Much enlarged	Total
Carrier present	19	29	24	72
Carrier absent	497	560	269	1326
Total	516	589	293	1398

    Is there evidence of an association between the two classifications?

12. A random sample of 1000 Canadians aged $25-34$ were classified according to their highest level of education and whether they were employed or not (data based on 2011 Canadian census data).

	Employed	Unemployed	Total
No certificate, diploma or degree	66	10	76
High school diploma or equivalent	185	16	201
Postsecondary certificate, diploma or degree	683	40	723
Total	934	66	1000

    Test the hypothesis that level of education is independent of whether or not a Canadian aged $25-34$ is employed.

13. In the following table, 64 sets of triplets are classified according to the age of their mother at their birth and their sex distribution:

	3 boys	2 boys	2 girls	3 girls	Total
Mother under 30	5	8	9	7	29
Mother over 30	6	10	13	6	35
Total	11	18	22	13	64

(a) Is there any evidence of an association between the sex distribution and the age of the mother?

(b) Suppose that the probability of a male birth is 0.5, and that the sexes of triplets are determined independently. Find the probability that there are y boys in a set of triples $y = 0, 1, 2, 3$, and test whether the column totals are consistent with this distribution.

14. To investigate the effectiveness of a rust-proofing procedure, 50 cars that had been rust-proofed and 50 cars that had not were examined for rust five years after purchase. For each car it was noted whether rust was present (actually defined as having moderate or heavy rust) or absent (light or no rust). The data are as follows:

	Rust-Proofed	Not Rust Proofed
Rust present	14	28
Rust absent	36	22
Total	50	50

(a) Test the hypothesis that the probability of rust occurring is the same for the rust-proofed cars as for those not rust-proofed. What do you conclude?

(b) Do you have any concerns about inferring that the rust-proofing prevents rust? How might a better study be designed?

15. Two hundred volunteers participated in an experiment to examine the effectiveness of vitamin C in preventing colds. One hundred were selected at random to receive daily doses of vitamin C and the others received a placebo. (None of the volunteers knew which group they were in.) During the study period, 20 of those taking vitamin C and 30 of those receiving the placebo caught colds. Test the hypothesis that the probability of catching a cold during the study period was the same for each group.

16. **Comparing speech recognition algorithms:** To compare the performance of two algorithms, A and B, for speech recognition, researchers presented each algorithm with a set of labeled utterances for recognition. The utterances which are syllables, words, or short phrases are assumed to be a random sample from a large population of utterances. An error is made when an algorithm does not correctly identify the utterance. On a set of 1400 utterances algorithm A made 72 errors. On a different

independent set of 1400 utterances algorithm B made 62 errors. Test the hypothesis that the probability of an error is the same for both algorithms.

17. **Challenge Problem: Comparing speech recognition algorithms - paired data** Suppose in Problem 16 that the same set of utterances is presented to both algorithms A and B. The analysis in Problem 16 cannot be used. (Both algorithms are more likely to identify simple utterances correctly and make errors more often when utterances are more complicated.) The data for this type of paired experiment can be summarized in general as

		B		
		Correct	Incorrect	
A	Correct	y_{11}	y_{12}	
	Incorrect	y_{21}	y_{22}	
				n

where, for example, y_{11} = the number of utterances correctly identified by both algorithms A and B. Since $(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \sim \text{Multinomial}(n; \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, the hypothesis that the probability of an error is the same for both algorithms is $H_0 : P(A \text{ identifies utterance correctly}) = \theta_{11} + \theta_{12} = P(B \text{ identifies utterance correctly}) = \theta_{11} + \theta_{21}$ or equivalently $H_0 : \theta_{12} = \theta_{21}$.

(a) Show that, under H_0, the maximum likelihood estimates are

$$\hat{\theta}_{11} = \frac{y_{11}}{n}, \quad \hat{\theta}_{12} = \hat{\theta}_{21} = \frac{y_{12} + y_{21}}{2n}, \quad \hat{\theta}_{22} = \frac{y_{22}}{n}$$

(b) Use the likelihood ratio test to test the hypothesis $H_0 : \theta_{12} = \theta_{21}$ for the data

		B		
		Correct	Incorrect	
A	Correct	1325	3	
	Incorrect	13	59	
				1400

Compare this result with the result in Problem 16.

8. CAUSAL RELATIONSHIPS

8.1 Establishing Causation

As mentioned in Chapters 1 and 3, many studies are carried out with causal objectives in mind. That is, we would like to be able to establish or investigate a possible cause and effect relationship between variates x and y.

We use the word "causes" often; for example we might say that "gravity causes dropped objects to fall to the ground", or that "smoking causes lung cancer". The concept of **causation** (as in "x causes y") is nevertheless difficult to define. One reason is that the "strengths" of causal relationships vary a lot. For example, on earth gravity may always lead to a dropped object falling to the ground; however, not everyone who smokes gets lung cancer.

Idealized definitions of causation are often of the following form. Let y be a response variate associated with units in a population or process, and let x be an explanatory variate associated with some factor that may affect y. Then, **if all other factors that affect y are held constant, let us change x (or observe different values of x) and see if y changes. If y changes then we say that x has a causal effect on y.**

In fact, this definition is not broad enough, because in many settings a change in x may only lead to a change in y in some probabilistic sense. For example, giving an individual person at risk of stroke a small daily dose of aspirin instead of a placebo may not necessarily lower that individual's risk. Not everyone will necessarily be helped by the daily dose of aspirin. However, on average, the effect is to lower the risk of stroke in the population.

Therefore, a better idealized definition of causation is to say that changing x should result in a change in some attribute of the variate y, for example, the proportion of the population who develop a stroke within 3 years. Thus we revise the definition above to say:

If all other factors that affect y are held constant, let us change x (or observe different values of x) and see if some specified attribute of y changes. If the specified attribute of y changes then we say x has a causal effect on y.

These definitions are unfortunately unusable in most settings since we cannot hold all other factors that affect y constant; often we don't even know what all the factors are. However, the definition serves as a useful ideal for how we should carry out studies in order to show that a causal relationship exists. We try to design studies so that alternative (to the variate x) explanations of what causes changes in attributes of y can be ruled out,

leaving x as the causal agent. This is much easier to do in experimental studies, where explanatory variates may be controlled, than in observational studies. The following are brief examples.

Example 8.1.1 Strength of steel bolts

Recall Example 6.1.4 concerning the (breaking) strength y of a steel bolt and the diameter x of the bolt. It is clear that bolts with larger diameters tend to have higher strength, and it seems clear on physical and theoretical grounds that increasing the diameter "causes" an increase in strength. This can be investigated in experimental studies like that in Example 6.1.4, when random samples of bolts of different diameters are tested, and their strengths y determined.

Clearly, the value of x does not determine y exactly (different bolts with the same diameter don't have the same strength), but we can consider attributes such as the average value of y. In the experiment we can hold other factors more or less constant (e.g. the ambient temperature, the way the force is applied; the metallurgical properties of the bolts) so we feel that the observed larger average values of y for bolts of larger diameter x is due to a causal relationship.

Note that even here we have to depart slightly from the idealized definition of cause and effect. In particular, a bolt cannot have its diameter x changed, so that we can see if y changes. All we can do is consider two bolts that are as similar as possible, and are subject to the same explanatory variates (aside from diameter). This difficulty arises in many experimental studies.

Example 8.1.2 Smoking and lung cancer

Suppose that data have been collected on $10,000$ persons aged $40-80$ who have smoked for at least 20 years, and $10,000$ persons in the same age range who have not. There is roughly the same distribution of ages in the two groups. The (hypothetical) data concerning the numbers with lung cancer are as follows:

	Lung Cancer	No Lung Cancer	Total
Smokers	500	9500	$10,000$
Non-Smokers	100	9900	$10,000$

There are many more lung cancer cases among the smokers, but without further information or assumptions we cannot conclude that a causal relationship (smoking causes lung cancer) exists. Alternative explanations might explain some or all of the observed difference. (This is an observational study and other possible explanatory variates are not controlled.) For example, family history is an important factor in many cancers; maybe smoking is also related to family history. Moreover, smoking tends to be connected with other factors such as diet and alcohol consumption; these may explain some of the effect seen.

The last example illustrates that **association (statistical dependence) between two variates x and y does not imply that a causal relationship exists.** Suppose for

example that we observe a positive correlation between x and y; higher values of x tend to go with higher values of y in a unit. Then there are at least three "explanations" for this correlation:

(1) x causes y (meaning x has a causative effect on y)

(2) y causes x

(3) some other variate(s) z cause both x and y.

We'll now consider the question of cause and effect in experimental and observational studies in a little more detail.

8.2 Experimental Studies

Suppose we want to investigate whether a variate x has a causal effect on a response variate y. In an experimental setting we can control the values of x that a unit "sees". In addition, we can use one or both of the following devices for ruling out alternative explanations for any observed changes in y that might be caused by x:

(1) Hold other possible explanatory variates fixed.

(2) Use randomization to control for other variates.

These devices are mostly simply explained via examples.

Example 8.2.1 Aspirin and the risk of stroke

Suppose 500 persons that are at high risk of stroke have agreed to take part in a clinical trial to assess whether aspirin lowers the risk of stroke. These persons are representative of a population of high risk individuals. The study is conducted by giving some persons aspirin and some a placebo, then comparing the two groups in terms of the number of strokes observed.

Other factors such as age, sex, weight, existence of high blood pressure, and diet also may affect the risk of stroke. These variates obviously vary substantially across persons and cannot be held constant or otherwise controlled. However, such studies use **randomization** in the following way: among the study subjects, who gets aspirin and who gets a placebo is determined by a random mechanism. For example, we might flip a coin (or draw a random number from $\{0, 1\}$), with one outcome (say Heads) indicating a person is to be given aspirin, and the other indicating that they get the placebo.

The effect of this randomization is to **balance** the other possible explanatory variates in the two "treatment" groups (aspirin and placebo). Thus, if at the end of the study we observe that 20% of the placebo subjects have had a stroke but only 9% of the aspirin subjects have, then we can attribute the difference to the causative effect of the aspirin.

Here's how we rule out alternative explanations: suppose you claim that its not the aspirin but dietary factors and blood pressure that cause this observed effect. I respond that the randomization procedure has lead to those factors being balanced in the two treatment groups. That is, the aspirin group and the placebo group both have similar variations in dietary and blood pressure values across the subjects in the group. Thus, a difference in the two groups should not be due to these factors.

Example 8.2.2 Driving speed and fuel consumption

It is thought that fuel consumption in automobiles is greater at speeds in excess of 100 km per hour. (Some years ago during oil shortages, many U.S. states reduced speed limits on freeways because of this.) A study is planned that will focus on freeway-type driving, because fuel consumption is also affected by the amount of stopping and starting in town driving, in addition to other factors.

In this case a decision was made to carry out an experimental study at a special paved track owned by a car company. Obviously a lot of factors besides speed affect fuel consumption: for example, the type of car and engine, tire condition, fuel grade and the driver. As a result, these factors were controlled in the study by balancing them across different driving speeds. An experimental plan of the following type was employed.

- 8 cars, each of a different type, were used; each car was used for 8 test drives
- the cars were each driven twice for 600 km on the track at each of four speeds: 80, 100, 120, and 140 km/hr.
- 8 drivers were involved, each driving each of the 8 cars for one test, and each driving 2 tests at each of the 4 speeds.
- the cars had similar initial mileages and were carefully checked and serviced so as to make them as comparable as possible; they used comparable fuels.
- the drivers were instructed to drive steadily for the 600 km. Each was allowed a 30 minute rest stop after 300 km.
- the order in which each driver did their 8 test drives was randomized. The track was large enough that all 8 drivers could be on it at the same time. (The tests were conducted over 8 days.)

The response variate was the amount of fuel consumed for each of the 64 test drives. Obviously in the analysis we must deal with the fact that the cars differ in size and engine type, and their fuel consumption will depend on that as well as on driving speed. A simple approach would be to add the fuel amounts consumed for the 16 test drives at each speed, and to compare them (other methods are also possible). Then, for example, we might find that the average consumption (across the 8 cars) at 80, 100, 120, and 140 km/hr were 43.0, 44.1, 45.8, and 47.2 liters respectively. Statistical methods of testing and estimation could

then be used to test or estimate the differences in average fuel consumption at each of the four speeds. (Can you think of a way to do this?)

Exercise Suppose that statistical tests demonstrated a significant difference in consumption across the four driving speeds, with lower speeds giving lower consumption. What (if any) qualifications would you have about concluding there is a causal relationship?

8.3 Observational Studies

In observational studies there are often unmeasured factors that affect the response variate y. If these factors are also related to the explanatory variate x whose (potential) causal effect we are trying to assess, then we cannot easily make any inferences about causation. For this reason, we try in observational studies to measure other important factors besides x.

For example, Problem 14 at the end of Chapter 7 discusses an observational study on whether rust-proofing prevents rust. It is clear that an unmeasured factor is the care a car owner takes in looking after a vehicle; this could quite likely be related to whether a person decides to have their car rust-proofed.

The following example shows how we must take note of other variates that may affect y.

Example 8.3.1 Graduate studies admissions

Suppose that over a five year period, the applications and admissions to graduate studies in Engineering and Arts faculties in a university are as follows:

	No. Applied	No. Admitted	% Admitted	
Engineering	1000	600	60%	Men
	200	150	75%	Women
Arts	1000	400	40%	Men
	1800	800	44%	Women
Total	2000	1000	50%	Men
	2000	950	47.5%	Women

We want to see if females have a lower probability of admission than males. If we looked only at the totals for Engineering plus Arts, then it would appear that the probability a male applicant is admitted is a little higher than the probability for a female applicant. However, if we look separately at Arts and Engineering, we see the probability for females being admitted appears higher in each case! The reason for the reverse direction in the totals is that Engineering has a higher admission rate than Arts, but the fraction of women applying to Engineering is much lower than for Arts.

In cause and effect language, we would say that the faculty one applies to (i.e. Engineering or Arts) is a causative factor with respect to probability of admission. Furthermore, it is related to the sex (male or female) of an applicant, so we cannot ignore it in trying to see if sex is also a causative factor.

Remark The feature illustrated in the example above is sometimes called *Simpson's Paradox*. In probabilistic terms, it says that for events A, B_1, B_2 and $C_1, \ldots, C_k$, we can have

$$P(A|B_1C_i) > P(A|B_2C_i) \text{ for each } i = 1, 2, \ldots, k$$

but have

$$P(A|B_1) < P(A|B_2)$$

(Note that $P(A|B_1) = \sum_{i=1}^{k} P(A|B_1C_i)P(C_i|B_1)$ and similarly for $P(A|B_2)$, so they depend on what $P(C_i|B_1)$ and $P(C_i|B_2)$ are.) In the example above we can take $B_1 = \{$person is female$\}$, $B_2 = \{$person is male$\}$, $C_1 = \{$person applies to Engineering$\}$, $C_2 = \{$person applies to Arts$\}$, and $A = \{$person is admitted$\}$.

Exercise Write down estimated probabilities for the various events based on Example 8.3.1, and so illustrate Simpson's paradox.

Epidemiologists (specialists in the study of disease) have developed guidelines or criteria which should be met in order to argue that a causal association exists between a risk factor x and a disease (represented by a response variate $y = I($person has the disease$)$, for example) in the case in which an experimental study cannot be conducted. These include:

- The association between x and y must be observed in many studies of different types among different groups. This reduces the chance that an observed association is due to a defect in one type of study or a peculiarity in one group of subjects.
- The association between x and y must continue to hold when the effects of plausible confounding variates are taken into account.
- There must be a plausible scientific explanation for the direct influence of x on y, so that a causal link does not depend on the observed association alone.
- There must be a consistent response, that is, y always increases (decreases) when x increases.

Example 8.3.2 Smoking and lung cancer

The claim that cigarette smoking causes lung cancer meets these four criteria. A strong association has been observed in numerous studies in many countries. Many possible sources of confounding variates have been examined in these studies and have not been found to

explain the association. For example, data about nonsmokers who are exposed to second-hand smoke contradicts the genetic hypothesis. Animal experiments have demonstrated conclusively that tobacco smoke contains substances that cause cancerous tumors. Therefore there is a known pathway by which smoking causes lung cancer. The lung cancer rates for ex-smokers decrease over time since smoking cessation. The evidence for causation here is about as strong as non-experimental evidence can be.

Similar criteria apply to other scientific areas of research.

8.4 Clofibrate Study

In the early seventies, the Coronary Drug Research Group implemented a large medical trial[17] in order to evaluate an experimental drug, clofibrate, for its effect on the risk of heart attacks in middle-aged people with heart trouble. Clofibrate operates by reducing the cholesterol level in the blood and thereby potentially reducing the risk of heart disease.

Study I: An Experimental Plan

Problem

- Investigate the effect of clofibrate on the risk of fatal heart attack for patients with a history of a previous heart attack.

The target population consists of all individuals with a previous non-fatal heart attack who are at risk for a subsequent heart attack. The response of interest is the occurrence/non-occurrence of a fatal heart attack. This is primarily a causative problem in that the investigators are interested in determining whether the prescription of clofibrate causes a reduction in the risk of subsequent heart attack. The fishbone diagram (Figure 8.1) indicates a broad variety of factors affecting the occurrence (or not) of a heart attack.

Plan

The study population consists of men aged 30 to 64 who had a previous heart attack not more than three months prior to initial contact. The sample consists of subjects from the study population who were contacted by participating physicians, asked to participate in the study, and provided informed consent. (All patients eligible to participate had to sign a consent form to participate in the study. The consent form usually describes current state of knowledge regarding the best available relevant treatments, the potential advantages and disadvantages of the new treatment, and the overall purpose of the study.)

The following treatment protocol was developed:

[17] *The Coronary Drug Research Group, New England Journal of Medicine (1980), pg. 1038.*

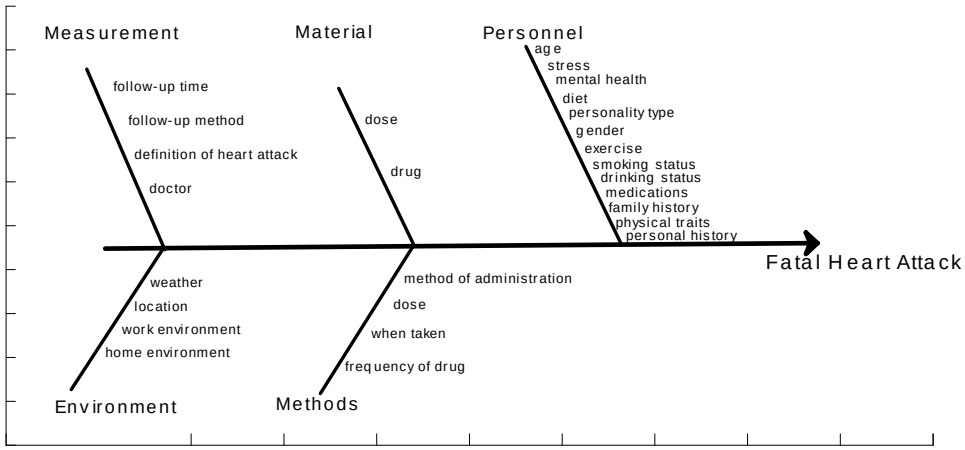


Figure 8.1: **Fishbone diagram for chlofibrate example**

- Randomly assign eligible men to either clofibrate or placebo treatment groups. (This is an attempt to make the clofibrate and placebo groups alike with respect to most explanatory variates other than the focal explanatory variate. See the fishbone diagram above.)
- Administer treatments in identical capsules in a double-blinded fashion. (In this context, *double-blind* means that neither the patient nor the individual administering the treatment knows if it is clofibrate or placebo; only the person heading the investigation knows. This is to avoid differential reporting rates from physicians enthusiastic about the new drug - a form of measurement error.)
- Follow patients for 5 years and record the occurrence of any fatal heart attacks experienced in either treatment group.

Determination of whether a fatality was attributable to a heart attack or not is based on electrocardiograms and physical examinations by physicians.

Data

- 1,103 patients were assigned to clofibrate and 2,789 were assigned to the placebo group.
- 221 of the patients in the clofibrate group died and 586 of the patients in the placebo group died.

Analysis

- The proportion of patients in the two groups having subsequent fatal heart attacks (clofibrate: $221/1103 = 0.20$ and placebo: $586/2789 = 0.21$) are comparable.

Conclusions

- Based on these data we would conclude that Clofibrate does not reduce mortality due to heart attacks in high risk patients.

  This conclusion has several limitations. For example, study error has been introduced by restricting the study population to male subjects alone. While clofibrate might be discarded as a beneficial treatment for the target population, there is no information in this study regarding its effects on female patients at risk for secondary heart attacks.

Study II: An Observational Plan

Supplementary analyses indicate that one reason that clofibrate did not appear to save lives might be because the patients in the clofibrate group did not take their medicine. It was therefore of interest to investigate the potential benefit of clofibrate for patients who adhered to their medication program.

Subjects who took more than 80% of their prescribed treatment were called "adherers" to the protocol.

Problem

- Investigate the occurrence of fatal heart attacks in the group of patients assigned to clofibrate who were adherers.
- The remaining parts of the problem stage are as before.

Plan

- Compare the occurrence of heart attacks in patients assigned to clofibrate who maintained the designated treatment schedule with the patients assigned to clofibrate who abandoned their assigned treatment schedule.
- Note that this is a further reduction of the study population.

Data

- In the clofibrate group, 708 patients were adherers and 357 were non-adherers. The remaining 38 patients could not be classified as adherers or non-adherers and so were excluded from this analysis. Of the 708 adherers, 106 had a fatal heart attack during the five years of follow up. Of the 357 non-adherers, 88 had a fatal heart attack during the five years of follow up.

Analysis

- The proportion of adherers suffering from subsequent heart attack is given by $106/708 = 0.15$ while this proportion for the non-adherers is $88/357 = 0.25$.

Conclusions

- It would appear based on these data that clofibrate does reduce mortality due to heart attack for high risk patients if properly administered.

  However, great care must be taken in interpreting the above results since they are based on an observational plan. While the data were collected based on an experimental plan, only the treatment was controlled. The comparison of the mortality rates between the adherers and non-adherers is based on an explanatory variate (adherence) that was not controlled in the original experiment. The investigators did not decide who would adhere to the protocol and who would not; the subjects decided themselves.

  Now the possibility of confounding is substantial. Perhaps, adherers are more health conscious and exercised more or ate a healthier diet. Detailed measurements of these variates are needed to control for them and reduce the possibility of confounding.

8.5 Chapter 8 Problems

1. In an Ontario study, 50267 live births were classified according to the baby's weight (less than or greater than 2.5 kg.) and according to the mother's smoking habits (non-smoker, 1-20 cigarettes per day, or more than 20 cigarettes per day). The results were as follows:

No. of cigarettes	0	$1-20$	>20
Weight ≤ 2.5	1322	1186	793
Weight > 2.5	27036	14142	5788

   (a) Test the hypothesis that birth weight is independent of the mother's smoking habits.

   (b) Explain why it is that these results do not prove that birth weights would increase if mothers stopped smoking during pregnancy. How should a study to obtain such proof be designed?

   (c) A similar, though weaker, association exists between birth weight and the amount smoked by the father. Explain why this is to be expected even if the father's smoking habits are irrelevant.

2. One hundred and fifty Statistics students took part in a study to evaluate computer-assisted instruction (CAI). Seventy-five received the standard lecture course while the other 75 received some CAI. All 150 students then wrote the same examination. Fifteen students in the standard course and 29 of those in the CAI group received a mark over 80%.

   (a) Are these results consistent with the hypothesis that the probability of achieving a mark over 80% is the same for both groups?

   (b) Based on these results, the instructor concluded that CAI increases the chances of a mark over 80%. How should the study have been carried out in order for this conclusion to be valid?

3.

   (a) The following data were collected some years ago in a study of possible sex bias in graduate admissions at a large university:

	Admitted	Not admitted
Male applicants	3738	4704
Female applicants	1494	2827

   Test the hypothesis that admission status is independent of sex. Do these data indicate a lower admission rate for females?

(b) The following table shows the numbers of male and female applicants and the percentages admitted for the six largest graduate programs in (*a*):

	Men		Women	
Program	Applicants	% Admitted	Applicants	% Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Test the independence of admission status and sex for each program. Do any of the programs show evidence of a bias against female applicants?

(c) Why is it that the totals in (a) seem to indicate a bias against women, but the results for individual programs in (b) do not?

4. To assess the (presumed) beneficial effects of rust-proofing cars, a manufacturer randomly selected 200 cars that were sold 5 years earlier and were still used by the original buyers. One hundred cars were selected from purchases where the rust-proofing option package was included, and one hundred from purchases where it was not (and where the buyer did not subsequently get the car rust-proofed by a third party). The amount of rust on the vehicles was measured on a scale in which the responses Y were assumed to have a Gaussian distribution. For the rust-proofed cars the responses were assumed to be $G(\mu_1, \sigma_1)$ and for the non-rust-proofed cars the responses were assumed to be $G(\mu_2, \sigma_2)$. Sample means and standard deviations for the two sets of cars were (higher y means more rust):

Rust-proofed cars	$\bar{y}_1 = 11.7$	$s_1 = 2.1$
Non-rust-proofed cars	$\bar{y}_2 = 12.0$	$s_2 = 2.4$

(a) Test the hypothesis that there is no difference between the mean amount of rust for rust-proofed cars as compared to non-rust-proofed cars.

(b) The manufacturer was surprised to find that the data did not show a beneficial effect of rust-proofing. Describe problems with their study and outline how you might carry out a study designed to demonstrate a causal effect of rust-proofing.

5. In Chapter 6, Problem 11 there was strong evidence against the hypothesis of no relationship between death rate from cirrhosis of the liver and wine consumption per capita in 46 states in the United States. Based on this study is it possible to conclude a causal relationship between wine consumption and cirrhosis of the liver?

6. Chapter 6, Problem 13 contained data, collected by the British botanist Joseph Hooker in the Himalaya Mountains between 1848 and 1850, on atmospheric pressure and the boiling point of water. Was this an experimental study or an observational study? Based on these data can you conclude that the boiling point of water affects atmospheric pressure?

7. In randomized clinical trials that compare two (or more) medical treatments it is customary not to let either the subject or their physician know which treatment they have been randomly assigned. These are referred to as *double blind* studies.

   Discuss why doing a double blind study is a good idea in an experimental study.

8. Public health researchers want to study whether specifically designed educational programs about the effects of cigarette smoking have the effect of discouraging people from smoking. One particular program is delivered to students in grade 9, with follow-up in grade 11 to determine each student's smoking "history". Briefly discuss some factors you would want to consider in designing such a study, and how you might address them.

9. REFERENCES AND SUPPLEMENTARY RESOURCES

9.1 References

R.J. Mackay and R.W. Oldford (2001). Statistics 231: *Empirical Problem Solving* (Stat 231 Course Notes)

C.J. Wild and G.A.F. Seber (1999). *Chance Encounters: A First Course in Data Analysis and Inference.* John Wiley and Sons, New York.

J. Utts (2003). What Educated Citizens Should Know About Statistics and Probability. *American Statistician* 57,74-79

9.2 Departmental Web Resources

See **www.watstat.ca**

10. DISTRIBUTIONS AND STATISTICAL TABLES

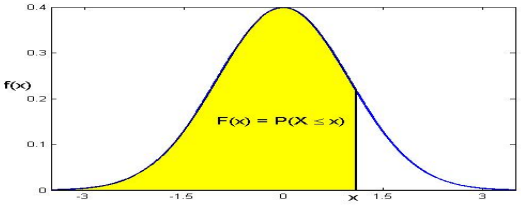
Summary of Discrete Distributions

Notation and Parameters	**Probability Function** $f(y)$	**Mean** $E(Y)$	**Variance** $Var(Y)$	**Moment Generating Function** $M(t)$
Discrete Uniform(a,b) $b \geq a$ a,b integers	$\frac{1}{b-a+1}$ $y = a, a+1, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{1}{b-a+1}\sum_{x=a}^{b} e^{tx}$ $t \in \Re$
Hypergeometric(N,r,n) $N = 1,2,\ldots$ $n = 0,1,\ldots,N$ $r = 0,1,\ldots,N$	$\frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$ $y = \max(0, n-N+r),$ $\ldots, \min(r,n)$	$\frac{nr}{N}$	$\frac{nr}{N}\left(1-\frac{r}{N}\right)\frac{N-n}{N-1}$	Not tractable
Binomial(n,p) $0 \leq p \leq 1,\ q = 1-p$ $n = 1,2,\ldots$	$\binom{n}{y}p^y q^{n-y}$ $y = 0,1,\ldots,n$	np	npq	$(pe^t + q)^n$ $t \in \Re$
Bernoulli(p) $0 \leq p \leq 1,\ q = 1-p$	$p^y q^{1-y}$ $y = 0,1$	p	pq	$pe^t + q$ $t \in \Re$
Negative Binomial(k,p) $0 < p \leq 1,\ q = 1-p$ $k = 1,2,\ldots$	$\binom{y+k-1}{y}p^k q^y$ $= \binom{-k}{y}p^k(-q)^y$ $y = 0,1,\ldots$	$\frac{kq}{p}$	$\frac{kq}{p^2}$	$\left(\frac{p}{1-qe^t}\right)^k$ $t < -\ln q$
Geometric(p) $0 < p \leq 1,\ q = 1-p$	pq^y $y = 0,1,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{p}{1-qe^t}$ $t < -\ln q$
Poisson(λ) $\lambda \geq 0$	$\frac{e^{-\lambda}\lambda^y}{y!}$ $y = 0,1,\ldots$	λ	λ	$e^{\lambda(e^t-1)}$ $t \in \Re$
Multinomial$(n;p_1,p_2,\ldots,p_k)$ $0 \leq p_i \leq 1$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} p_i = 1$	$f(y_1,y_2,\ldots,y_k) =$ $\frac{n!}{y_1!y_2!\cdots y_k!}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}$ $y_i = 0,1,\ldots,n$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} y_i = n$	$E(Y_i) = np_i$ $i = 1,2,\ldots,k$	$Var(Y_i)$ $= np_i(1-p_i)$ $i = 1,2,\ldots,k$	$M(t_1,t_2,\ldots,t_k)$ $=(p_1e^{t_1}+p_2e^{t_2}+\cdots$ $+p_{k-1}e^{t_{k-1}}+p_k)^n$ $t_i \in \Re$ $i = 1,2,\ldots,k-1$

Summary of Continuous Distributions

Notation and Parameters	**Probability Density Function** $f(y)$	**Mean** $E(Y)$	**Variance** $Var(Y)$	**Moment Generating Function** $M(t)$
Uniform(a,b) $b > a$	$\frac{1}{b-a}$ $a \le y \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$ $t \neq 0$; 1 $t = 0$
Exponential(θ) $\theta > 0$	$\frac{1}{\theta}e^{-y/\theta}$ $y \ge 0$	θ	θ^2	$\frac{1}{1-\theta t}$ $t < \frac{1}{\theta}$
$\mathrm{N}(\mu,\sigma^2) = G(\mu,\sigma)$ $\mu \in \Re,\ \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(y-\mu)^2/(2\sigma^2)}$ $y \in \Re$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$ $t \in \Re$
$\chi^2(k)$ $k = 1,2,\ldots$	$\frac{y^{(k/2)-1}e^{-y/2}}{2^{k/2}\Gamma(k/2)}$ $y > 0$ $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$	k	$2k$	$(1-2t)^{-k/2}$ $t < \frac{1}{2}$
$t(k)$ $k = 1,2,\ldots$	$\frac{c_k}{(1+\frac{y^2}{k})^{(k+1)/2}}$ $y \in \Re$ $c_k = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)}$	0 if $k = 2,3,\ldots$ DNE if $k = 1$	$\frac{k}{k-2}$ if $k = 3,4,\ldots$ DNE if $k = 1,2$	DNE

N(0,1) Cumulative Distribution Function



This table gives values of $F(x) = P(X \le x)$ for $X \sim N(0,1)$ and $x \ge 0$

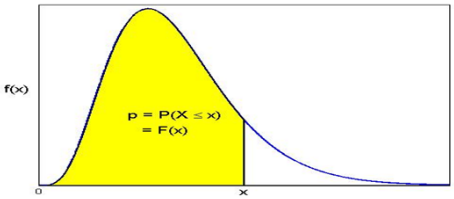
x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

N(0,1) Quantiles:

This table gives values of $F^{-1}(p)$ for $p \ge 0.5$

p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.075	0.08	0.09	0.095
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.1891	0.2019	0.2275	0.2404
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4538	0.4677	0.4959	0.5101
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7554	0.7722	0.8064	0.8239
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1503	1.1750	1.2265	1.2536
0.9	1.2816	1.3408	1.4051	1.4758	1.5548	1.6449	1.7507	1.8808	1.9600	2.0537	2.3263	2.5758

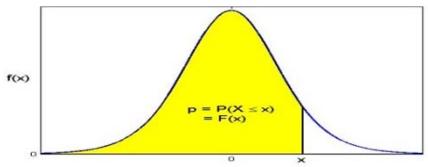
Chi-Squared Quantiles



This table gives values of x for p = P(X ≤ x) = F(x)

df\p	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.842	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	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	1.146	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	1.647	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	3.054	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	9.391	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	100.430	104.210
80	51.172	53.540	57.153	60.391	64.278	96.578	101.880	106.630	112.330	116.320
90	59.196	61.754	65.647	69.126	73.291	107.570	113.150	118.140	124.120	128.300
100	67.328	70.065	74.222	77.929	82.358	118.500	124.340	129.560	135.810	140.170

Student t Quantiles



This table gives values of x for $p = P(X \le x) = F(x)$, for $p \ge 0.6$

df \ p	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.3249	0.7265	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192
2	0.2887	0.6172	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991
3	0.2767	0.5844	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240
4	0.2707	0.5686	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.2672	0.5594	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.2648	0.5534	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.2632	0.5491	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.2619	0.5459	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.2610	0.5435	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.2602	0.5415	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.2596	0.5399	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.2590	0.5386	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.2586	0.5375	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.2582	0.5366	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.2579	0.5357	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.2576	0.5350	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.2573	0.5344	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.2571	0.5338	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
19	0.2569	0.5333	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.2567	0.5329	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
21	0.2566	0.5325	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193
22	0.2564	0.5321	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921
23	0.2563	0.5317	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676
24	0.2562	0.5314	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454
25	0.2561	0.5312	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251
26	0.2560	0.5309	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066
27	0.2559	0.5306	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896
28	0.2558	0.5304	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739
29	0.2557	0.5302	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594
30	0.2556	0.5300	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460
40	0.2550	0.5286	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069	3.5510
50	0.2547	0.5278	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614	3.4960
60	0.2545	0.5272	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602
70	0.2543	0.5268	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108	3.4350
80	0.2542	0.5265	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163
90	0.2541	0.5263	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019
100	0.2540	0.5261	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905
>100	0.2535	0.5247	0.8423	1.2832	1.6479	1.9647	2.3338	2.5857	3.1066	3.3101